The McGraw-Hill Companies

Mobile Computing

Technology, Applications and Service Creation

Second Edition

The McGraw·Hill Companies

The McGraw-Hill Companies

Mobile Computing

Technology, Applications and Service Creation

Second Edition

Asoke K Talukder, Ph.D.
Chief Scientific Officer and Director, Geschickten Solutions, Bangalore
Adjunct Professor, Department of Computer Science and Engineering, NIT, Warangal
Adjunct Faculty, ABV Indian Institute of Information Technology and Management, Gwalior
Adjunct Faculty, Department of Computer Engineering, NITK Surathkal

Hasan Ahmed
Nokia Research and Development
Bangalore

Roopa R Yavagal
Symphony Services
Bangalore

Tata McGraw Hill Education Private Limited
NEW DELHI

McGraw-Hill Offices

New Delhi New York St Louis San Francisco Auckland Bogotá Caracas
Kuala Lumpur Lisbon London Madrid Mexico City Milan Montreal
San Juan Santiago Singapore Sydney Tokyo Toronto

The McGraw-Hill Companies

Tata McGraw Hill

Published by Tata McGraw Hill Education Private Limited,
7 West Patel Nagar, New Delhi 110 008.

ISBN (13): 978-0-07-014457-6
ISBN (10): 0-07-014457-5

Vice President and Managing Director—Asia-Pacific Region: *Ajay Shukla*
Executive Publisher: *R Chandra Sekhar*
Manager—Production: *Sohan Gaur*
Manager—Sales & Marketing: *S. Girish*
Deputy Marketing Manager—Science, Technology and Computing: *Rekha Dhyani*
General Manager—Production: *Rajender P Ghansela*
Asst. General Manager—Production: *B L Dogra*

Typeset at Bukprint India, B-180A, Guru Nanak Pura, Laxmi Nagar, Delhi 110 092 and printed at
Pashupati Printers Pvt. Ltd., 1/429/16, Gali No. 1, Friends Colony, Industrial Area, G.T. Road, Shahdara, Delhi 110 095

Cover Printer: SDR Printers

Cover Designer: Kapil Gupta

RBZLCRQZDRCCD

The McGraw-Hill Companies

The McGraw·Hill Companies

To my Parents–who taught me how to step into this world;

To my Gurus–who taught me how to face this world;

To my Wife and Daughter–who taught me how to love this world;

To all my Friends–who taught me how to live in this world;

And to the Readers of this book—who will teach me how to leave a legacy behind.

–Asoke K Talukder

The McGraw·Hill Companies

Preface to the Second Edition

The first edition of this book comprising 18 chapters was published in 2005, where we introduced various technologies that form the foundation of mobile computing. The book started to be used as a text and reference book for M.Tech students across universities in India and abroad. Since the last few years, technology has witnessed several innovations, as a result the need for a revised edition was imminent.

The second edition is a consolidation of the topics covered in the first edition along with incorporation of several new innovations that have occurred in the mobile computing domain in the last five years.

Changes in the Second Edition:

- Three new chapters have been added in the book; these are Chapter 18 on Multimedia, Chapter 19 on IP Multimedia Service (IMS), and Chapter 21 on Next Generation Networks (NGN).
- In Chapter 1, many new standards bodies that have emerged and are relevant today to mobile computing have been included.
- In Chapter 2, Model View Control (MVC) has been added and its relevance highlighted in the context of mobile computing environment.
- In Chapter 3, new material on various multiplexing techniques and Satellite Communication technologies have been added.
- In Chapter 4, we have included new topics like Mobile IP, IPv6, Mobile IPv6, and IPsec.
- In Chapter 5, we have extensively covered Mobility Management and Personal Communications Service (PCS).
- In Chapter 9, we comprehensively cover IMT 2000.
- In Chapter 10, we have added new topics like Wireless in Local Loop (WLL) and HiperLAN.
- In Chapter 11, we have added content on security issues on the SS7 network with special emphasis on SS7 Security and MAP Security. We have also included Virtual Private Network (VPN).
- In Chapter 12, some recent innovations have been included.
- In Chapter 14, new content on recent developments in Symbian OS have been added.
- In Chapter 15, we have added all Java Specification Request (JSR) related to Mobility—this will help a J2ME developer to locate the right procedure for mobile computing.
- In Chapter 17, we have added Voice over Wireless LAN.
- Chapter 18 of the first edition has been renumbered as Chapter 20 in second. We have also added a few topics that deal with security issues in mobile computing.

Besides these, other notable changes include:

- More questions at the end of each chapter. This will help an instructor to use this book as a textbook and frame questions and quizzes. Some of these questions can also be used for programming across different technologies; and hence, can be used as laboratory assignments.
- New sections in quite a few chapters.
- Updated references and links to various online resources on programming, technology connectivity, add-ons, plug-ins, applications, etc.
- Ancillary material: We have developed power point slides on each chapter. These are available on http://highered.mcgraw-hill.com/sites/0070144575.

Organization of the Book

The Second Edition includes 21 chapters.

Chapter 1 introduces the book and builds the foundation for mobile computing. It covers definitions and significance of terms and technologies. This chapter introduces various frameworks and explains how they fit into the jigsaw puzzle called mobile computing. This chapter also touches upon the digital divide, services and applications required for the masses. Though this chapter covers Standards bodies and their roles, many more, relevant to mobile computing today, have been added, along with their functions and Web addresses.

Chapter 2 covers the general architecture of mobile computing and the first step towards making information ubiquitous through the public network (Internet). This chapter deals with multi-tier architecture of applications development and its significance. It discusses various types of middleware, their functions, roles, and how they can be used to implement mobile services. In this chapter, Model View Control (MVC) has been added to show how it is relevant in a mobile computing environment.

Chapter 3 introduces concepts behind the telephony system. It deals with how to access data and information using a telephone as a client device. With the proliferation of computers and mobile internet, voice-based applications may appear as legacy technology; however, the greatest advantage of voice is that it is hands-free and information can be accessed using only 12 keys. This chapter introduces the philosophy behind voice-based applications development through CTI (Computer Telephony Interface/Computer Telephony Integration) using IVR (Interactive Voice Response) medium. It also covers Voice XML and Voice browsers. New material on various multiplexing techniques and Satellite Communication technologies have been added in this chapter.

Chapter 4 discusses accessory technologies related to mobile computing. These are Bluetooth, Radio Frequency Identifiers of RFID, Wireless Broadband or WiMax, Mobile and Cellular IP, IPv6 (Internet Protocol version 6), next generation IP, and IPsec.

Chapter 5 describes GSM technologies. It starts with the basic concepts of cellular networks. It describes GSM architecture and different elements within the GSM network. It covers the role of these elements and how they interoperate to ensure seamless routing, handover, mobility management, and roaming within GSM networks. Extensive material on Mobility Management and Personal Communications Service (PCS) has been added in this chapter. Principles of mobility management as presented here can be extended for any other technology domain. This chapter concludes with details of GSM security algorithms and infrastructure.

Chapter 6 deals with SMS architecture and technology in detail. It discusses how SMS can be used as a data bearer to develop value-added services and applications where SMS is used as input/output media. It includes two-way SMS Pull technology and one-way SMS Push technology—SMS Push services are very critical for unsolicited alerts and notification services. It finishes with the SMPP (Short Message Peer-to-Peer) protocol and the GNU open source Kannel SMS gateway—SMS gateway is the critical component required for integrating a computer system with the telephony network.

Chapter 7 moves up in the hierarchy from Second Generation (2G) to Two Point Five Generation (2.5G), i.e., from GSM to GPRS. This chapter starts with GPRS architecture and various elements of the GPRS network. It describes differences between GSM and GPRS. It describes in detail how GPRS deals with mobility issues with respect to data. It also elucidates some mobile applications that are more suited for higher bandwidth GPRS networks.

Chapter 8 discusses WAP and MMS technology. It starts with the description of WAP stack and WAP Application Environment. It covers WML (Wireless Markup Language) and how to develop applications using WML. It then describes WMLScript and Wireless Telephony Application Interface (WTAI) moving on to Multimedia Messaging Service. It introduces SMIL (Synchronized Multimedia Integration Language) and describes how to develop MMS applications using SMIL. The chapter concludes with DRM (Digital Rights Management) as defined by the WAP forum.

Chapter 9 discusses CDMA and 3G technology. This chapter begins with the concept of Spread Spectrum Technology and Code Division Multiple Access. IS-85 was the first cellular technology to use CDMA multiplexing technique. Though IS-95 is a 2nd generation network, it is described in this chapter. It describes security, handoff, and roaming in the IS-95 networks. This chapter then proceeds to 3G networks and describes IMT-2000, CDMA-2000, UMTS, AND WCDMA networks. The chapter concludes with a description of 3G application framework and the concept of VHE (Virtual Home Network). Some of the new concepts of ubiquity, convergence, context awareness, etc., are described as part of the 3G framework.

Chapter 10 discusses Wireless Local Area Network (WLAN) or WiFi technology and aspects of WLAN mobility and roaming. It describes where, why, how and when to use Wireless LAN. It discusses the scope of various standards within the 802.11 family. It also describes ad hoc and infrastructure-based wireless networks. It describes how CDMA and CSMA-CA (Carrier Sense Multiple Access––Collision Avoidance) is used in WLAN. It then describes deployment of WLAN, briefly touching upon sensor networks and usage of WLAN for sensor networks. It then discusses WLAN security, its vulnerabilities and evolution of security standards to ensure higher data security over WLAN. Addition of new topics like Wireless in Local Loop (WLL) and HiperLAN have been made in this chapter. The chapter concludes with a comparison between 3G and WiFi.

Chapter 11 covers Intelligent Network (IN) and Interworking of telephony networks. This chapter introduces the SS#7 signaling network and describes how to use this network to develop Intelligent Network applications. It starts with fundamentals of call processing and routing. It then goes on to explain when and why we call a network intelligent with special emphasis on SS#7 Security and MAP Security. The chapter concludes with examples of some applications in IN.

Chapter 12 introduces the philosophy and technique to develop technology neutral applications. This chapter is the introduction to Chapter 13 through Chapter 15 where we cover various platforms, tools and techniques to develop mobile applications.

Chapter 13 describes programming the PalmOS for PDAs (Personal Digital Assistance) and wireless services. It describes PalmOS architecture and then moves forward to applications development on PalmOS. We start with the applications development environment in PalmOS and some aspects one needs to keep in mind during their development. These include graphical user interface, forms, networking and communications, databases, and security. It then describes how to program PalmOS for multimedia and telephony services.

Chapter 14 describes Symbian OS that is used in many mobile phone models from Nokia, Sony, etc. It starts with Symbian architecture and explains Symbian's application development environment. This chapter has been updated with new content on recent developments in Symbian OS. The chapter concludes with security considerations in Symbian.

Chapter 15 includes all aspects of J2ME starting from simple applications to multimedia, tickers, database, networking, and communication using sockets. We have added Java Specification Request (JSR) related to mobility––this will help a J2ME developer to locate the right procedure for mobile computing. The chapter concludes with security considerations and the latest in J2ME.

Chapter 16 talks about Windows CE and how to use it for application development for small devices starting from mobile phones to PDAs. It describes the architecture of Windows CE and various development platforms available for application development in Windows CE.

Chapter 17 talks about Voice over Internet Protocol (VoIP). It deals with different protocols and technology for convergence of Internet with telecommunication network. It starts with H.323 and describes different elements and components within H.323. It then describes SIP (Session Initiation Protocol) followed by all the technologies and protocols required to create a convergence between telecommunications and IP. It also focuses on application development techniques and the SIP environment. Voice over Wireless LAN has been added in this chapter.

Chapter 18 describes Multimedia. It starts with compression and decompression technologies that form the foundation of coders and decoders (codecs). It covers technology related to multimedia applications and their transmission. It includes delivery of multimedia content over the Internet. It also includes best effort delivery and intserv and diffserv protocols besides matter on how to develop applications using multimedia libraries.

Chapter 19 talks about IP Multimedia Service (IMS) or IP Multimedia Subsystem. It starts with the architecture of IMS and the way it works through different technology planes. It addresses technologies and protocols used in IMS. In any telecommunication service, accounting and charging are at its core; therefore, it includes the charging principles and architecture for charging in IMS as well as how to develop IMS services.

Chapter 20 deals with security issues in mobile computing. Majority of the information security challenges in desktop and mobile computing are common. Therefore, this chapter starts with vulnerability, exploits, and attacks. It then describes symmetric and public key encryption techniques followed by security protocols specific to SSL (Secured Socket Layer), TLS (Transport Layer Security) and WTLS (Wireless TLS). It also covers the security framework for mobile environment and 3GPP (3rd Generation Partnership Program). The chapter ends with details about virus and worms in a wireless environment.

Chapter 21 is an introduction to Next Generation Networks or NGN. NGN is about convergence––the journey of convergence that was started with IMS would be completed by NGN. As NGN is yet to mature, this chapter addresses trends and emerging technologies and services. It covers

different technologies that would be used by NGN. It also covers the services that are expected to dominate NGN like multiple plan and multimedia broadcast services.

Target Audience

The book targets a large cross-section of readers.

To the Professional

This book is designed to cover a broad range of topics suitable for an ICT professional. We believe that this book will be an excellent handbook for professionals who need to understand mobile computing. Each chapter covers one topic and they are more or less standalone. Therefore as a professional you can focus on a topic of your interest. If you need to understand some of the wireless technologies to manage a project, this is the book for you. Even if you have not been in touch with some of the recent technologies, this book can help you jumpstart. The book is suitable for application programmers, telecommunication professionals, or managers who may want to know about technology and services in mobile computing.

To the Teacher

The book is designed to provide a broad spectrum of subjects. Topics covered can be used for graduate and postgraduate studies in mobile computing. We have tried to keep the chapters connected; however, they can be approached independently as well. There are many questions at the end of each chapter that can be used as quizzes and projects.

To the Student

We hope this textbook will trigger your interest in the field of mobile computing and wireless services. This book can serve as a handbook, text as well as reference book. This book is suitable for a computer science student interested in developing a mobile application for the computer savvy, or a social science student who wants to develop an application to eradicate the digital divide. Organization of the book has evolved from wire to wireless and telecom, Internet to convergence.

The Table of Contents and Index may be used as reference to locate the updates to the chapters.

Though we have tried our best, it is possible that there could be errors in the book that have escaped our attention. We would like to hear from you––your suggestions, feedback, criticism, or any other comments that will enable us to improve the next edition are welcome. Please send in your comments and suggestions to asoket@gmail.com.

ASOKE K TALUKDER

The McGraw·Hill Companies

Preface to the First Edition

It has been known for centuries that knowledge is power. *Pundits* knew how to transform data into information and then into knowledge. They used data, information, and knowledge in a different context. What was not known for centuries is how to store data, information and knowledge in such a way that it could be made available to everybody! The convergence of information and communication technology (ICT–Information and Communication Technology) has created avenues to address all these challenges. With the help of ICT, it is possible to render information and knowledge to anybody, anywhere, anytime.

The last decade of the 20th century witnessed a lot of activities in ICT. The GSM was launched; www (World Wide Web) became popular; the telecom industry saw a lot of promise. Investors started putting their money on ICT; The Dot com boom happened. As an ICT person, I could not shy away from all this excitement. In 2000, we started a WAP (Wireless Application Protocol) portal company called Cellnext in the ICT space. Within a few months we realized that technology is an enabler and does not sell on its own. Within Cellnext, we quickly moved from WAP to GSM (Global System for Mobile Communications) and SMS (Short Message Service). Later we embraced IVR (Interactive Voice Response), GPRS (General Packet Radio Service), 3G, CDMA (Code Division Multiple Access) and MMS (Multimedia Messaging Service). To put all these things together, we had to do a lot of research. Also, there were interests from academics to offer courses on some of these emerging technologies. I met Professor Dinesha and Professor Sadagopan, captains from IIIT-B (International Institute of Information Technology Bangalore). In 2001, I offered a one-semester elective subject to second year M.Tech students of IIITB on WAP and SMS. Like technology, the course content changed over a period of time. In 2003, the elective WAP course evolved into a core course on mobile computing for third semester M.Tech students. IIITB was one of the very few institutes offering mobile computing as a full semester course. This book is an outcome of lecture notes and topics of the mobile computing course that I offered to M. Tech students of IIIT-B in the last few years.

Making information available from anywhere anytime is one set of challenge. Making information available all the time when the user is mobile and may be traveling by train or car is another set of challenge! Mobile computing technology will address these challenges and enable the realization of a global village with ubiquitous information, where people can seamlessly access any information from anywhere through any device while stationary or when mobile. There are a few books which cover wireless and mobile communications, but there are not many books that cover the service aspect of mobile applications. This book covers all the communication technologies starting from first generation to third generation cellular technologies, wired telecommunication technology, wireless LAN (WiFi), and wireless broadband (WiMax). It covers intelligent networks (IN) and emerging technologies like mobile IP, IPv6, and VoIP (Voice over IP).

The book is targeted to address a large cross-section of audience. This book can be used either as a text or a reference book. As this book gives a big picture of all the technologies from CTI (computer telephony interface) to 3G (third generation) including Bluetooth, IN, WiFi and WiMax, it can be used by non-technical people as a primer.

To the Professional

This book is designed to cover a broad range of topics suitable for an ICT professional. We believe that this book will be an excellent handbook for professionals who need to understand mobile computing. Each chapter covers one topic and they are more or less standalone. Therefore, as a professional you can focus on any topic that interests you. If you need to understand some of the wireless technologies to manage a project, this is the book for you. Even if you have not been in touch with some of the recent technologies, this book can help you to jumpstart. The book is suitable for application programmers, telecommunication professionals, or managers who may want to know about technology and services in mobile computing.

To the Teacher

The book is designed to provide a broad spectrum of subjects. Topics covered can be used for postgraduate studies in mobile computing. We have tried to keep the chapters connected; however, they are self-contained. For an undergraduate program, part of every chapter can be included. You may like to read through 'Organization of the Book' to be able to decide on the optimal mix of chapters in your course module.

To the Student

We hope that this textbook will trigger your interest in the field of Mobile Computing and wireless services. We tried to organize this book as a textbook. However, we organized the chapters, topics, and the content in a fashion that it can also serve as a reference book or a handbook. This book is suitable for a computer science student who wants to develop a mobile application for computer savvy people, or a social science student who wants to design an application to eradicate the digital divide. The organization of the book has evolved from wire to wireless and telecom, Internet to convergence. Therefore, you may like to read the book in the sequence the chapters have been organized. However, we tried to keep the content of every chapter as independent as possible so that you can select a particular topic and understand it.

Organization of the Book

The book is organized into 18 chapters:

Chapter 1 is the general introduction to the book. It builds the foundation for mobile computing. It also covers various definitions and significance of terms and technologies. As part of the foundation,

the chapter introduces various frameworks and explains how they fit into the complex jigsaw puzzle called mobile computing. How can mobile computing help businesses to use information in an effective way! Also, it touches upon the digital divide, services, and applications required for the masses at the "bottom of the pyramid". At the end, this chapter covers Standards bodies and their roles. It also goes into details of the scope of different Standards bodies and their relevance in the context of mobile computing.

Chapter 2 captures the general architecture of mobile computing and the first step towards making information ubiquitous through the public network (Internet). This chapter deals with multi-tier architecture of applications development and its significance. It discusses various types of middleware, their functions, roles, and how can they be used to implement mobile services. This chapter describes the philosophy behind context in a mobile application. It also goes into details of how to determine context and develop a context aware system.

Chapter 3 introduces the concepts behind telephony system. It deals with how to access information using a telephone as a client device. As telephones are available throughout the world, voice-based applications can be considered as legacy technology for today's mobile ubiquitous computing. This chapter introduces the philosophy behind voice-based applications development through CTI (Computer Telephony Interface/Computer telephony Integration) using IVR (Interactive Voice Response) medium. It also covers Voice XML and Voice browsers.

Chapter 4 discusses many technologies. All these technologies are related to mobile computing; however, they are yet to become mainstream technologies. These technologies are: Bluetooth, Radio Frequency Identifiers or RFID, Wireless Broadband or WiMax, Mobile and Cellular IP, IPv6 (Internet Protocol version 6) or the next generation IP. These technologies are very important and have every possibility to become mainstream technologies of tomorrow. Therefore, we have introduced these technologies as related technologies, but have not covered them in detail in this book.

Chapter 5 describes GSM technologies. It starts with the basic concepts of cellular networks. It describes the GSM architecture and different elements within the GSM network. It covers the role of these elements and how they interoperate to ensure seamless routing, handover, mobility management, and roaming within GSM networks. The chapter finishes with various security algorithms and security infrastructure within GSM.

Chapter 6 deals with SMS. It describes the SMS architecture in detail. It describes the SMS data technology over the air and over SMS gateways. It describes how SMS can be used as a data bearer to develop value-added services and applications for the masses through SMS. It finishes with the SMPP (Short Message Peer-to-Peer) protocol and the GNU open source Kannel SMS gateway.

Chapter 7 moves up in the hierarchy from Second Generation (2G) to Two point Five Generation (2.5G), GSM to GPRS. This chapter starts with the GPRS architecture and various elements of the GPRS network. It describes the differences between GSM and GPRS. It describes in detail how GPRS deals with mobility issues with respect to data. It describes some mobile applications that are more suited for higher bandwidth GPRS networks.

Chapter 8 discusses the WAP and MMS technology. It starts with the description of WAP stack and WAP Application Environment. It covers WML (Wireless Markup Language) and how to develop applications using WML. It then describes WMLScript and Wireless Telephony Application Interface

(WTAI) moving on to Multimedia Messaging Service. It introduces SMIL (Synchronized Multimedia Integration Language) and describes how to develop MMS applications using SMIL. The chapter finishes with DRM (Digital Writes Management) as defined by WAP forum.

Chapter 9 moves one step further to discuss CDMA and 3G technology. This chapter begins with the concept of Spread Spectrum Technology and Code Division Multiple Access. IS-95 was the first cellular technology to use CDMA multiplexing technique. Though IS-95 is a 2nd generation network, it is described in this chapter. It then describes security, handoff, and roaming in the IS-95 network. It then goes into the 3G networks. It describes IMT-2000, CDMA-2000, UMTS, and WCDMA networks. The chapter concludes with description of 3G application framework and the concept of VHE (Virtual Home Network). Some of the new concepts of ubiquity, convergence, context awareness, etc. are described as a part of 3G framework.

Chapter 10 discusses Wireless Local Area Network (WLAN) or WiFi technology. It describes where, why, how, and when to use Wireless LAN. It describes the scope of various standards within the 802.11 family. It describes adhoc and infrastructure based wireless networks. It describes how CDMA and CSMA-CA (Carrier Sense Multiple Access-Collision Avoidance) is used in WLAN. The chapter covers aspects of mobility and roaming in WLAN. It then describes the deployment of WLAN. It very briefly touches upon sensor networks and how WLAN can be used for sensor networks. It then discusses WLAN security, what is available today, its vulnerabilities, and evolution of security standards to ensure higher data security over WLAN. The chapter concludes with a comparison between 3G and WiFi.

Chapter 11 covers Intelligent Network and Interworking. This chapter introduces the SS#7 signaling network and describes how to use this network to develop Intelligent Network (IN) applications. It starts with the fundamentals of call processing and routing. It then explains when and why we call a network intelligent. It then goes into details of the SS#7 network and different elements within a SS#7 network. It touches upon different application parts at the application layer of the SS#7 stack. This chapter finishes with some examples of applications in IN.

Chapter 12 introduces the philosophy and technique to develop technology neutral applications. In Chapters 3 through Chapter 10, we discussed various techniques for mobile applications. These techniques are specific to the specific technologies. This Chapter is the introduction to Chapters 13 through Chapter 15 where we cover various platforms, tools and techniques to develop mobile applications.

Chapter 13 describes programming the PalmOS for PDAs (Personal Digital Assistance) and wireless services. In this chapter, we cover the tools and techniques to develop mobile applications for PalmOS based PDAs. It starts with PalmOS architecture and then moves forward on applications development on PalmOS. We start with applications development environment in PalmOS. It describes some of the aspects one needs to bear in mind while developing PalmOS applications. These include graphical user interface, forms, networking and communications, databases, and security. It then describes how to program PalmOS for multimedia and telephony interfaces.

Chapter 14 deals with what goes in a Symbian OS, which is becoming quite popular in communicators or high-end mobile phones. This chapter starts with Symbian architecture and explains the applications development environment in Symbian. It then describes how to develop applications in a Symbian environment. It finishes the chapter with security considerations in Symbian.

Chapter 15 is all about Java for wireless and mobile devices. It starts with Java 2 Micro Edition (J2ME). It describes the architecture of J2ME and goes in detail of how to develop applications using J2ME. This includes all aspects of J2ME starting from simple applications to multimedia, tickers, database, networking, and communication using sockets. The chapter finishes with security considerations for J2ME.

Chapter 16 is about WindowsCE and using it for applications development for small devices starting from mobile phones to PDAs. It describes the architecture of WindowsCE. It then describes the various development platforms available for applications development in WindowsCE.

Chapter 17 talks about the Voice over Internet Protocol (VoIP). It deals with different protocols and technology for convergence of Internet with the telecommunication network. It starts with H.323 and describes different elements and components within H.323. It then describes SIP (Session Initiation Protocol) followed by all the technologies and protocols required to make the convergence of telecommunication and IP. It then describes applications development techniques and the environment within SIP.

Chapter 18 deals with security issues in mobile computing. Majority of the information security challenges in desktop computing and mobile computing are common. Therefore, this chapter starts with vulnerability, exploits, and attacks. It then describes symmetric and public key encryption techniques. It then describes security protocols specific to SSL (Secured Socket layer) and TLS (Transport Layer Security) and WTLS (Wireless TLS). It then covers the security framework for mobile environment and 3GPP (3rd Generation Partnership Program). The chapter ends with virus and worms in wireless environment.

Though we have tried our best, it is possible that there could be errors that have escaped our attention in the book. We would like to hear from you–your suggestions, feedback, criticism, or any other comments that will enable us improve the next edition of Mobile Computing. Please send your comments and suggestions to:asoke.talukder@iiitb.ac.in.

ASOKE K TALUKDER

The McGraw·Hill Companies

The McGraw-Hill Companies

Acknowledgements

This book would not have been possible without the guidance and support—technical and personal—of a number of people. We would like to thank all those individuals, researchers, professionals and technicians who are working in the domain of ICT for contributing to this book directly or indirectly. We went through many problems and challenges during our tenure with the industry. Many of the challenges that I faced have helped me to learn new things. Some of my former colleagues who have directly contributed to my work and whose names I would like to mention are: Nikhil Nanda, Rajan Swaroop, Nityananda, Atish Dasgupta, Rishi Pal, Manish Chaitanya, Rajkumar, Gururaj, Debmalya, Ashish Tara, Kurian John, Saugat Maitra, Ayush Sharma, Arun Kumar Pandey, Rithesh Swami, Varun, Pradeep and Srihari P Mule.

I have read several books and many articles on the web and in various journals and magazines. Also I had referred many recommendations proposed by various forums and standards making bodies. These organizations contributed to this work in a major way. I have acknowledged most of them as references and further readings. The omissions, if any, are inadvertent and not deliberate.

All trademarks and registered trademarks used in the book are the properties of their respective owners/companies.

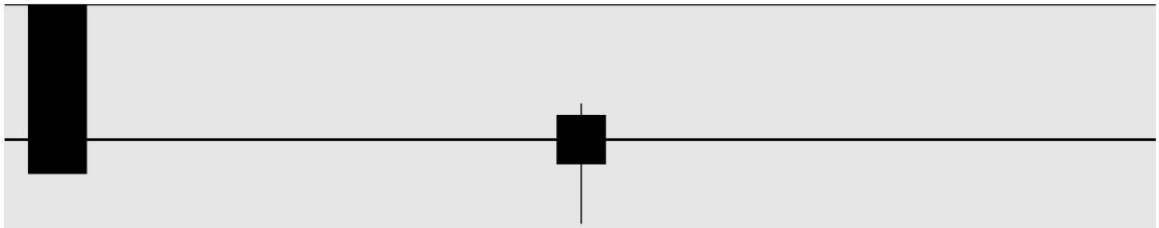
I would like to acknowledge the contribution of Hasan Ahmed for writing Chapters 18, 19 and 21 and reviewing the entire manuscript. I would also like to acknowledge the contribution of Ms Roopa Yavagal for contributing to Chapters 12 to 15 of this book. In addition, I would like to thank Prof S. Sadagopan, Prof Prabhu, Prof Dinesha, Prof Debabrata Das and Mr C.M. Abraham of IIIT-B for their encouragement and support while I was the Daimler Chrysler Chair professor at IIIT-B and was writing this book. I would like to thank Tata McGraw-Hill for publishing this book.

Finally, I would like to thank my wife Kalyani and daughter Debi, without whose support this book would not have been complete. Not only did they inspire me to write this book but also encouraged me to complete it.

Asoke K Talukder

The McGraw·Hill Companies

Contents



List of Abbreviations

1G	First Generation
2.5G	2.5 Generation
2G	Second Generation
3G	Third Generation
3GPP	Third Generation Partnership Project
3GPP LTE	3GPP Long Term Evolution
3GPP2	Third Generation Partnership Project 2
4G	Fourth Generation communications

A

AAA	Authentication, Authorization and Accounting
AABS	Automatic Alternative Billing Service
AAS	Adaptive Antenna System
AC	Admission Control
AC	Authentication Centre
ACELP	Algebraic Code Excited Linear Prediction
ACK	Acknowledgement
ACL	Access Control List
ACL	Asynchronous Connectionless Linkz
ACM	Address Complete Message
ACM	Audio Compression Manager
AD	Access Device
ADC	Analog to Digital Converter
ADPCM	Adaptive Differential (or Delta) PCM
aDSL	asynchronous Digital Subscriber Line
AES	Advanced Encryption Standard
AI	application Interface (prefix to interface class method)
AIFF	Audio Interchange File Format
AIPN	All Internet Protocol Network
ALAC	Apple Lossless Audio Codec
ALS	Audio Lossless Coding
AM	Amplitude Modulation
AMBE	Advanced Multi-band Excitation
AMC	Adaptive Modulation and Coding

AMPS	Advanced Mobile Phone System
AMR	Adaptive Multi Rate
ANM	Answer Message
ANSI	American National Standards Institute
AoC	Advice of Charge
AP	Access Point
API	Application Programming Interface
APN	Access Point Name
APN-NI	APN Network Identifier
APPUI	Application User Interface
AR	Access Requestor
ARDOR	Adaptive Rate-distortion Optimized sound codeR
ARFCN	Absolute Radio Frequency Channel Numbers
ARIB	Association of Radio Industries and Businesses
ARP	Address Resolution Protocol
ARPA	Advance Research Project Agency
ARPU	Average Revenue Per User
ARQ	Automatic Repeat Request
ASCII	American Standard Code for Information Interchange
aSi-TFT	amorphous Silicon TFT
ASP	Active Server Page
ASP	Application Service Provider
AT	Attention
ATD	Absolute Time Difference
ATM	Asynchronous Transfer Mode
ATN	Automated Trust Negotiatin Systems
ATRAC	Adaptive Transform Acoustic Coding
AUC	Authentication Center
AVP	Attribute Value Pairs
AVS	Audio Video Standard

B

B2B	Business to Business
B3G	Beyond 3rd Generation
BASIC	Beginners All purpose Symbolic Instructional Code
BCCH	Broadcast Control Channel
BER	Bit Error Rate
BG	Border Gateway
BGCF	Breakout Gateway Control Function
BIB	Backward Indicator Bit
BMP	Bit Map
BPSK	Binary Phase Shift Keying
BS	Base Station
BSA	Basic Station Area
BSC	Base Station Controller
BSN	Backward Sequence Number
BSS	Base Station Subsystem
BSS	Basic Service Set
BSSAP	BSS Application Part
BT	Busy Tone
BTS	Base Transceiver Station
BTS	Base Transceiver System
BWA	Broadband Wireless Access

C

CA	Certification Authority
CA	Content Aggregator
CAC	Channel Access and Control
CAMEL	Customized Application for Mobile Network Enhanced Logic
CAP	CAMEL Application Part

CAS	Call Associated Signalling
CAS	Conditional Access System
CC	Country Code
CC/PP	Composite Capabilities/ Preference Profiles
CCETT	Centre Commun d'études de Télévision et Telecommunications
CCK	Complementary Code Keying
CCSA	China Communications Standards Association
CCSSO	Common Channel Signaling Switching Office
CDC	Connected Device Configuration
CDF	Charging Data Function
CDMA	Code Division Multiple Access
CDN	Content Distribution Network
CDPD	Cellular Digital Packet Data
CDR	Call Detail Record
CDR	Charging Data Records
CE devices	Customer Edge devices
CEK	Content Encryption Key
CELP	Code Excited Linear Prediction
CEPT	Conference of European Posts and Telegraphs
CE-r	Customer Edge routers
CE-s	Customer Edge switches
CF	Contention Free
CFB	Call Forwarding Busy
CFNA	Call Forwarding Not Answered
CFNR	Call Forwarding Not Reachable
CFP	Contention-Free Period
CFU	Call Forwarding Unconditional
CGF	Charging Gateway Function
CGI	Computer Gateway Interface
cHTML	compact Hyper Text Markup Language
CI	Call Identifier
CICS	Customer Information Control System
CID	Cell ID
CIMD	Computer Interface to Message Delivery
CLDC	Connected Limited Device Configuration
CLI	Caller Line Identification
CM	Connection Management
CMOS	Complementary Metal Oxide Semiconductor
CN	Core Network
CO	Central Office
CODEC	Coder and Decoder
COPS	Common Open Policy Service
COPS-PR	COPS for Policy Provisioning
CORBA	Common Object Request Broker Architecture
CoS	Class of Service
CP	Content Provider
CP	Contention Period
CPE	Customer Premises Equipment
CPI	Capability and Preference Information
CPU	Central Processing Unit
CRC	Cyclic Redundancy Code
CRP	Customer Routing Point
CS	Capability Set
CS	Carrier Sense
CSCF	Call Session Control Function
CSD	Circuit Switched Data
CSE	CAMEL Service Environment
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CSMA/CD	Carrier Sense Multiple Access with Collision Detection
CSP	Communication Service Provider

CT	Communication Technology
CTI	Computer Telephony Interface/Computer
CTS	Clear To Send
CUG	Closed User Group
CVSD	Continuously Variable Slope Delta Modulation
CWTS	China Wireless Telecommunication Standard group
Cyborg	Cyber Organism

D

DAB	Digital Audio Broadcast
DAC	Digital to Analog Converter
DAMA	Demand Assignment Multiple Access
DC	Data Center
DCE	Data Circuit terminating Equipment
DCF	Distributed Coordination Function
DCF	DRM Content Format
DCT	Discrete Cosine Transform
DECT	Digital Enhanced Cordless Communications
DECT	Digital Enhanced Cordless Telecommunications
DFP	Distributed Functional Plane
DFRD	Device Family Reference Designs
DHCP	Dynamic Host Configuration Protocol
DIFS	Distributed Inter Frame Space
DL	Downlink
DLB	Dynamic Label Segment
DLC	Digital Loop Carrier
DLL	Data Link Layer
DLL	Dynamic Link Library
DLNA	Digital Living Network Alliance
DMA	Direct Memory Access
DMH	Data Message Handler
DNS	Domain Name Server
DNS	Domain Name Service
DoCoMo	DO (Everywhere) + COMO (Communication)
DoD	Department of Defense
DPC	Destination Point Code
DPRMA	Dynamic Packet Reservation Multiple Access
DR TDMA	Dynamic Reservation Time Division Multiple Access
DRM	Digital Rights Management
DRNC	Drift RNC
DS	Direct Sequence
DS	Distribution System
DSL	Digital Subscriber Line
DSM CC	Digital Storage Media Command and Control
DSP	Digital Signal Processing
DSP	Digital Signal Processor
DSS	Digital Speech Standard
DSSS	Direct Sequence Spread Spectrum
DST	Direct Stream Transfer
DT	Dial Tone
DTE	Data Terminal Equipment
DTMF	Dual Tone Multi Frequency
DUP	Data User Part
DV Codec	Digital Video Codec
DVB	Digital Video Broadcasting
DVD	Digital Video Disc
DWDM	Dense Wavelength Division Multiplexing

E

EAP	Extensible Authentication Protocol

EDGE	Enhanced Data rate for GSM Evolution
EGPRS	Enhanced GPRS
EIA	Electronic Industries Alliance
EIFS	Extended Inter Frame Space
EIR	Equipment Identity Register
EMS	Extended Message Service
ENIAC	Electronic Numerical Integrator and Computer
ENUM	Electronic Numbering
E-OTD	Enhanced Observed Time Difference
ERP	Enterprise Resource Planning
ESME	External Short Message Entity
ESN	Electronic Serial Number
ESS	Electronic Switching System
ESS	Extended Service Set
ETSI	European Telecommunication Standards Institute
EU	End User
EVRC	Enhanced Variable Rate Codec
EY-NPMA	Elimination-Yield Non-Preemptive Multiple Access

F

FAMA	Fixed Assignment Multiple Access
FCAPS	Fault, Configuration, Accounting, Performance, Security
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FE	Functional Entity
FEA	Functional Entity Action
FEC	Forward Error Correction
FER	Frame Error Rate
FH	Frequency Hopping
FHSS	Frequency Hopping Spread Spectrum
FIB	Forward Indicator Bit
FISU	Fill-In Signal Units
FLAC	Free Lossless Audio Codec
FM	Frequency Modulation
FNC	Federal Networking Council
FOFDM	Flash Orthogonal Frequency Division Multiplexing
FRA	Fixed Radio Access
FSN	Forward Sequence Number
FSNP	Full Service and Network Provider
FSU	Fixed Subscriber Unit
FTP	File Transfer Protocol
FWA	Fixed Wireless Access

G

GERAN	GSM EDGE Radio Access Network
GFP	Global Functional Plane
GGSN	Gateway GPRS Support Node
GIF	Graphics Interchange Format
GIWU	Gateway Inter Working Unit
GMLC	Gateway MLC
GMSC	Gateway MSC
GPRS	General Packet Radio Service
GPS	Global Positioning System
GSM	Global System for Mobile communications
GT	Global Title
GTT	Global Title Translation
GUI	Graphical User Interface

H

HC SDMA	High Capacity Spatial Division Multiple Access

HCI	Human Computer Interface
HCPDU	HiperLAN CAC Protocol Data Unit
HCSAP	HiperLAN CAC Service Access Point
HCSDU	HiperLAN CAC Service Data Unit
HDLC	High level Data Link Control
HDML	Handheld Device Markup Language
HDTP	Handheld Device Transport Protocol
HDTV	High Definition Television
HE AAC	High Efficiency Advance Audio Coding
HE	Home Environment
HFC	Hybrid Fiber Coaxial
HiperLAN	High Performance Radio Local Area Network
HiperMAN	High Performance Radio Metropolitan Area Network
HLF	Home Location Function
HLR	Home Location Register
HMPDU	HiperLAN MAC Protocol Data Unit
HPLMN	Home Public Land Mobile Network
HSDPA	High Speed Downstream Packet Availability
HSOPA	High Speed OFDM Packet Access
HSPA	High Speed Packet Access
HSS	Home Subscriber Server
HSUPA	High Speed Uplink Packet Access
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
HVXC	Harmonic Vector Excitation Coding

I

I/O	Input/Output
IAM	Initial Address Message
IAPP	Inter-Access Point Protocol
IBSS	Independent Basic Service Set
IC	Integrated Circuit
ICAP	Internet Content Adaptation Protocol
ICCC	International Computer Communication Conference
ICL	International Computers Limited
ICMP	Internet Control Message Protocol
I-CSCF	Interrogating Call Session Control Function
ICT	Information and Communication Technology
IDE	Interactive Development Environment
IED	Information Element Data
IEDL	Information Element Data Length
IEEE	Institute of Electrical and Electronics Engineers
IEI	Information Element Identifier
IETF	Internet Engineering Task Force
IGMP	Internet Group Management Protocol
IHF	Integrated Hands Free
I-HSPA	Internet HSPA
iLBC	internet Low Bit rate Codec
IM SSF	IP Multimedia Services Switching Function
IMAP	Internet Message Access Protocol
IMBE	Improved Multi-Band Excitation

IMEI	International Mobile Equipment Identity
IMS GWF	IMS Gateway Function
IMS	IP Multimedia Subsystems
IMSI	International Mobile Subscriber Identity
IMT DS	IMT Direct Spread
IMT FT	IMT FDMA/TDMA
IMT FT	IMT Frequency Time
IMT MC	IMT Multi Carrier
IMT SC	IMT Single Carrier
IMT TC	IMT TDD Carrier
IMT	International Mobile Telecommunications
IMT	International Mobile Telecommunications
IN	Intelligent Networks
INAP	Intelligent Network Application Part
INCM	IN Conceptual Model
IP	Internet Protocol
IPCP	Internet Protocol Control Protocol
IPDL	Idle Period Downlink
IPDR	IP Data Record
IPNG	Next Generation Internet Protocol
IPsec IKE	IPsec Internet Key Exchange
IPsec	Internet Protocol Security
IPTV	Internet Protocol Television
IR	Infra Red
IrDA	Infrared Data Association
IrMC	Infrared Mobile Communication
IRT	Institutefür Rundfunktechnik GmbH
iSAC	internet Speech Audio Codec
ISDN	Integrated Services Digital Network
ISI	Inter Symbol Interference
ISIM	IP multimedia Subscriber Identity Module
ISM	Industrial, Scientific, and Medical
ISO	International Organization for Standardization
ISP	Internet Service Provider
ISUP	ISDN User Part
ISV	Independent Software Vendor
IT	Information Technology
ITTP	Intelligent Terminal Transfer Protocol
ITU	International Telecommunication Union
ITU-T	International Telecommunication Union–Telecommunication Standardization
IVR	Interactive Voice Response
IWF	Inter Working Function
IWMSC	Inter Working MSC

J

J2EE	Java 2 Enterprise Edition
J2ME	Java 2 Micro Edition
J2SE	Java 2 Standard Edition
JDBC	Java Data Base Connector
JFIF	JPEG File Interchange Format
JPEG	Joint Photographic Experts Group
JSP	Java Server Pages
JSR	Java Specification Request

K

KVCD	K Video Compression Dynamics

L

L2CAP	Logical Link Control and Adaptation Protocol

L2TP	Layer 2 Tunneling Protocol
LA	Location Application
LA	Location Area
LA	Lossless Audio
LAF	Location Application Function
LAI	Location Area Identifier
LAN	Local Area Network
LAP	LAN Access Point
LAP	Link Access Procedure
LAPD	Link Access Procedure-D
LBS	Location Based Services
LCAF	Location Client Authorization Function
LCCF	Location Client Control Function
LCCTF	Location Client Coordinate Transformation Function
LCD	Liquid Crystal Diode
LCD	Liquid Crystal Display
LCF	Location Client Function
LCP	Link Control Protocol
LCS	LoCation Services
LCZTF	Location Client Zone Transformation Function
LDR	Location Deferred Request
LED	Light Emitting Diodes
LEO	Low Earth Orbit
LIR	Location Immediate Request
LLC	Logical Link Control
LMP	Link Manager Protocol
LMSI	Local Mobile Subscriber Identity
LMU	Location Measurement Unit
LNP	Local Number Portability
LPAC	Lossless Predictive Audio Compression
LPC	Linear Prediction Coding
LPCM	Linear Pulse Code Modulation
LSAF	Location Subscriber Authorization Function
LSBcF	Location System Broadcast Function
LSBF	Location System Billing Function
LSCF	Location System Control Function
LSOF	Location System Operation Function
LSP	Label Switched Path
LSPF	Location Subscriber Privacy Function
LSR	Label Switching Router
LSSU	Link Status Signal Unit
LTAC	Lossless Transform Audio Compression
LTE	Long Term Evolution
LTPS-TFT	Low Temperature Poly Silicon TFT
LZW	Lempel Ziv Welch

M

M2M	Machine to Machine
MAC	Media Access Control
MAN	Metropolitan Area Network
MAP	Mobile Application Part
Mbone	Multicast back bone
MCC	Mobile Country Code
MCU	Master Controller Unit
ME	Mobile Equipment
MEGACO	Media Gateway Control Protocol
MELP	Mixed Excitation Linear Prediction
MEO	Medium Earth Orbit
MExE	Mobile Execution Environment

MGCP	Media Gateway Control Protocol
MGW	Media Gateway
MIB	Management Information Base
MIDP	Mobile Information Device Profile
MIMO	Multiple Input Multiple Output
MIN	Mobile ID Number
MIPS	Millions of instructions per second
MIS	Management Information Systems
mITF	mobile IT Forum
MLC	Mobile Location Center
MLME	MAC sub Layer Management Entity
MM	Mobility Management
MMI	Man Machine Interface
MMS	Multimedia Message Service
MMSC	MMS Controller
MMSE	MMS Environment
MMTel	Multimedia Telephony
MMU	Memory Management Unit
MNC	Mobile Network Code
MNG	Multiple image Network Graphics
MO	Mobile Originated
MO-LR	Mobile Originated Location Request
MOM	Message Oriented Middleware
MP2	MPEG-1 layer-2 audio coding
MP3	MPEG 1 Part 3 Layer 3
MPDU	MAC Protocol Data Unit
MPEG	Moving Pictures Expert Group
MPEG-4 ASP	MPEG 4 Advanced Simple Profile
MPLS	Multiprotocol Label Switching
MPLS	Multiprotocol Label Switching
MRF	Media Resource Function
MRFC	Media Resource Function Controller
MRFP	Media Resource Function Processor
MS	Mobile Station
MSAP	Media Service Access Point
MSC	Mobile Switching Centre
MSDU	MAC Service Data Unit
MSDU	MAC Service Data Unit
MSIN	Mobile Subscriber Identification Number
MSISDN	Mobile Station ISDN
MSP	Mobile Service Provider
MSRN	Mobile Station Roaming Number
MSU	Message Signal Units
MT	Mobile Terminated
MTAS	Multimedia Telephony Application Service
MT-LR	Mobile Terminated Location Request
MTP	Message Transfer Part
MVC	Model-View-Controller
MVNO	Mobile Virtual Network Operator

N

NA-ESRD	North American Emergency Service Routing Digits
NA-ESRK	North American Emergency Service Routing Key
NAT	Network Address Translator
NAV	Network Allocation Vector
NDC	National Destination Code
NDP	Network Decision Point
NFC	Near Field Communications
NGN	Next Generation Network
NIC	Network Interface Card

NID	Network Identification
NI-LR	Network Induced Location Request
NNI	Network to Network Interface
NO	Network Operator
N-PE devices	Network-facing PE devices
NSF	National Science Foundation
NSS	Network and Switching Subsystem
NTT	Nippon Telegraph and Telephone Corporation

O

OBEX	Object Exchange Protocol
OCC	Occasionally Connected Computing
OCR	Optimal Call Routing
OCS	Online Charging System
ODBC	Open Data Base Connectivity
OEM	Original Equipment Manufacturer
OFC	Optical Fiber Cable
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OFR	OptimFROG
OMA	Open Mobile Alliance
OMAP	Operations, Maintenance and Administration Part
OMC	Operation and Maintenance Center
OOPS	Object Oriented Programming
OPC	Originating Point Code
OPL	Organiser Programming Language
OS	Operating System
OSA	Open Service Access
OSA	Open Service Architecture
OSA-SCS	Open Service Access Service Capability Server
OSS	Operation and Support Subsystem
OSS	Operations Support System
OTA	Over-The-Air
OTDOA	Observed Time Difference Of Arrival

P

P3P	Platform for Privacy Preference Project
PAM	Pulse-amplitude Modulation
PAN	Personal Area Network
PBX	Private Branch Exchange
PBX	Private Business Exchange
PC	Point Coordinator
PC	Power Control
PCF	Point Coordination Function
PCF	Power Calculation Function
PCH	Paging Channels
PCI	Peripheral Component Interface
PCM	Pulse Coded Modulation
PCMCIA	Personal Computer Memory Card International Association
PCN	Personal Communication Networks
PCS	Personal Communications Service
P-CSCF	Proxy Call Session Control Function
PDA	Personal Digital Assistant
PDC	Personal Digital Cellular
PDF	Policy Decision Function
PDN	Packet Data Network
PDP	Packet Data Protocol

PDP	Policy Decision Point
PDTCH	Packet Data Traffic Channel
PDU	Protocol Data Unit
PE devices	Service Provider Edge devices
PE	Physical Entity
PEAP	Protected EAP
PEP	Policy Enforcement Point
PE-r	Provider Edge routers
PE-rs	Provider Edge devices that are capable of both routing and switching
PE-s	Provider Edge switches
PHP	Hypertext Preprocessor
PHS	Personal Handyphone System
PHY	Physical (layer)
PIB	Policy Information Bases
PICS	Platform for Internet Content Selection
PIFS	Point (coordination function) Inter Frame Space
PIM	Personal Information Management
PKI	Public Key Infrastructure
PLCP	Physical Layer Convergence Procedure
PLL	Physical Link Layer
PLMN	Public Land Mobile Network
PLW	PSDU Length Word
PMD	Physical Medium Dependent
PN	Pseudo random Noise
PNG	Portable Network Graphics
POI	Point Of Initiation
POI	Privacy Override Indicator
PoP	Points of Presence
POP	Post Office Protocol
POR	Point Of Return
POS	Point Of Sale
POTS	Plain Old Telephone Service
PP	Physical Plane
PPDU	PLCP Protocol Data Unit
PPG	Push Proxy Gateway
PPP	Point-to-Point Protocol
PR	Policy Repository
PRCF	Positioning Radio Co-ordination Function
PRMA HS	Packet Reservation Multiple Access Hindering States
PRMA	Packet Reservation Multiple Access
PRNG	Pseudo-Random Number Generator
PRRM	Positioning Radio Resource Management
PS	Power Save (mode)
PSDN	Public Switched Data Network
PSDU	Physical sublayer Service Data Unit
PSE	Personal Service Environment
PSF	PLCP Signaling Field
PSMF	Positioning Signal Measurement Function
PSPDN	Public Switched Packet Data Networks
PSTN	Public Switched Telephone Network
PTM	Point-To-Multipoint
PTP	Point-To-Point

Q

QCELP	Qualcomm Pure Voice
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keyed

R

RA	Routing Area
RACH	Random Access Channel

RADIUS	Remote Authentication Dial In User Service
RAM	Random Access Memory
RAN	Radio Access Network
RANAP	RAN Application Part
RAND	Random Number
RAS	Remote Access Service
RASP	Reliability, Availability, Security, and Performance
RCELP	Relaxed Code Excited Linear Prediction
RDF	Resource Description Framework
REL	Release
RF	Radio Frequency
RFC	Request For Comments
RFCOMM	Radio Frequency Communication
RFID	Radio Frequency Identifiers
RFL	Radio Frequency Layer
RGB	Red Green Blue
RIFF	Resource Interchange File Format
RIP	Routing Information Protocol
RIS	Radio Interface Synchronization
RISC	Reduced Instruction Set Computer
RKAU	RK Audio
RLC	Radio Link Control
RLC	Release Complete
RLE	Run Length Encoding
RLL	Radio Local Loop
RLP	Radio Link Protocol
RNC	Radio Network Controller
ROM	Read Only Memory
RPC	Remote Procedure Call
RPCU	Radio Port Control Unit
RPE-LPC	Regular Pulse Excited-Linear Predictive Coder
RRM	Radio Resource Management
RSA	Rivest, Shamir, Adelman
RSACi	Recreational Software Advisory Council internet
RSM	Remote Switching Modules
RSU	Radio Subscriber Unit
RSVP	Resource reSerVation Protocol
RT	Ring Tone
RTCP	RTP Control Protocol
RTD	Real-time Difference
RTP	Real-time Transfer Protocol
RTS	Request To Send
RTSP	Real-time Streaming Protocol
RTT	Radio Transmission Technology

S

SA	Security Associations
SAP	Service Access Point
SAT	SIM Application Toolkit
SBS	Switched Beam System
SC	Service Centre
SCCP	Signalling Connection Control Part
SCF	Service Capability Feature
SCO	Synchronous Connection Oriented link
SCP	Service Control Point
SCS	Service Capability Servers
S-CSCF	Serving Call Session Control Function
SCUA	Service Control User Agent
SDH	Synchronous Digital Hierarchy
SDK	Software Development Kit
SDMA	Space Division Multiple Access
SDP	Service Discovery Protocol
SDP	Session Description Protocol

sDSL	synchronous Digital Subscriber Line
SDTV	Standard Definition TV
SF	Service Feature
SGSN	Serving GPRS Support Node
SGW	Signaling Gateway
SHN	Shorten
SI	Service Interface (prefix to interface class method)
SIB	Service Independent Building block
SIBO	Single Board Organizer
SID	System Identification
SIF	Service Information Field
SIFS	Short Inter Frame Space
SIM	Subscriber Identity Module
SIO	Service Indicator Octet
SIP AS	SIP Application Server
SIP	Session Initiation Protocol
SIR	Signal Interference Ratio
SIS	Symbian OS Installation
SLF	Subscriber Location Function
SLPP	Subscriber LCS Privacy Profile
SME	Short Message Entity
SME	Station Management Entity
SMG	Special Mobile Group
SMIL	Synchronization Multimedia Integration Language
SMLC	Serving Mobile Location Centre
SMMO	Short Message Mobile Originated point-to-point
SMMT	Short Message Mobile Terminated point-to-point
SMPP	Short Message Peer-to-Peer
SMS	Service Management System in SMS/800
SMS	Short Message Service
SMSC	SMS Centre
SMTP	Simple Mail Transfer Protocol
SMV	Selectable Mode Vocoder
SN	Service Node
SN	Subscriber Number
SNA	System Network Architecture
SNDCF	Sub Network Dependent Convergence Function
SNDCP	Sub Network Dependent Convergence Protocol
SNR	Signal to Noise Ratio
SOAP	Simple Object Access Protocol
SoD	Session oriented Dialogue
SP	Service Plane
SP	Service Point
SPC	Signaling Point Code
SPI	Service Provider Interface
SQL	Structured Query Language
SRC	Short Retry Count
SRES	Signature Response
SRNC	Serving RNC
SS	Signalling System
SS	Station Service
SS7	Signaling Stack 7
SS7	Signaling System No 7
SSID	Service Set Identifier
SSL	Secured Socket Layer
SSO	Single Sign On
SSP	Service Switching Point
STA	Station
STB	Set Top Boxes
STP	Signaling Transfer Point
SVG	Scalable Vector Graphics
SWAP	Shared Wireless Access Protocol

T

TA	Timing Advance
TA	True Audio

TACS	Total Access Communication System
TCAP	Transaction Capabilities Application Part
TCP	Transmission Control Protocol
TCP/IP	Transmission Control Protocol/ Internet Protocol
TCS	Telephony Control Specification
TD-CDMA	Time Division Code Division Multiple Access
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
	Telephony Integration
TFT	Thin Film Transistor
THIG	Topology Hiding Interworking Gateway
TIA	Telecommunication Industries Association
TINA	Telecommunications Information Network Architecture consortium
TKIP	Temporal Key Integrity Protocol
TLS	Transport Layer Security
TMSI	Temporary Mobile Subscriber Identity
TN3270	Telnet protocol for IBM 3270
TN5250	Telnet protocol for IBM 5250
TOA	Time Of Arrival
TP	Transaction Processing
TPMS	Transaction Processing Management System
TRC	Triple Rate CODER
TSCC	Tech Smith Screen Capture Codec
TSP	Telephone Service Provider
TT	Trouble Ticket
TTA	Telecommunications Technology Association–Korea
TTC	Telecommunication Technology Committee–Japan
TTL	Time To Live
TTML	Tagged Text Mark-up Language
TTS	Text To Speech
TUP	Telephone User Part

U

UAProf	User Agent Profile
UDH	User Data Header
UDHI	User Data Header Indicator
UDP	User Datagram Protocol
UDT	Unit Data message
UE	User Equipment
UICC	Universal Integrated Circuit Card
UL	Uplink
UMTS	Universal Mobile Telecommunication System
UNI	User to Network Interface
URI	Universal Resource Identifier
URL	Universal Resource Locator
USIM	Universal Subscriber Identity Module
USSD	Unstructured Supplementary Service Data
UTRAN	Universal Terrestrial Radio Access Network

V

VAS	Value Added Service
VASP	Value Added Service Provider
VCD	Video Compact Disc
VCR	Video Cassette Recorder
VDU	Visual Display Unit
VHE	Virtual Home Environment
VHS	Video Home System

VLR	Visitor Location Register
VLSI	Very Large Scale Integration
VME	Virtual Machine Environment
VoD	Video on Demand
VOFDM	Vector Orthogonal Frequency Division Multiplexing
VoIP	Voice over IP
VPLS	Virtual Private LAN Service
VPN	Virtual Private Network
VPNC	Virtual Private Network Consortium
VPS	Voice Processing System
VRML	Virtual Reality Markup Language
VRU	Voice Response Unit
VSAT	Very Small Aperture Terminal
VSELP	Vector Sum Excited Linear Prediction
VT3K	Visual Terminal for HP 3000

W

W3C	WWW Consortium
WAE	Wireless Application Environment
WAFU	Wireless Access Fixed Unit
WAP	Wireless Application Protocol
WATM	Wireless Asynchronous Transfer Mode
WBMP	Wireless BMP
WCDMA	Wideband Code Division Multiple Access
WDP	Wireless Datagram Protocol
Wi-Fi	Wireless Fidelity
WiLL	Wireless in Local Loop
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless LAN
WLL	Wireless Local Loop
WM	Wireless Medium
WMA	Windows Media Audio
WML	Wireless Markup Language
WMV	Windows Media Video
WOFDM	Wideband Orthogonal Frequency Division Multiplexing
WSP	Wireless Session Protocol
WTA	Wireless Telephony Applications
WTAI	Wireless Telephony Application Interface
WTLS	Wireless Transport Layer Security
WTP	Wireless Transaction Protocol
WV	WavPack
WWAN	Wireless Wide Area Network
WWW	World Wide Web

X

XML	eXtensible Markup Language

CHAPTER 1

Introduction

1.1 MOBILITY OF BITS AND BYTES

Information is power. But for a long time people did not know how to store information and knowledge which could be easily accessible. Convergence of information and communication technology has created ways to address these challenges. Today even when we are on the move, we can access information from anywhere, any time.

In the last two centuries, mobility has been redefined. Both physical and virtual objects are now mobile. Mobility of physical objects relate to movement of matters, whereas movements of virtual objects relate to movements of bits and bytes.

> The foundation of mobility of information was laid by Joseph Henry, (1797–1878), who invented the electric motor and techniques for distant communication. In 1831, Henry demonstrated the potential of using an electromagnetic phenomenon of electricity for long distance communication. He sent electric current over one mile of wire to activate an electromagnet, which caused a bell to ring. Later, Samuel F. B. Morse used this property of electricity to invent the telegraph. Morse transmitted his famous message "What hath God wrought?" from Washington to Baltimore over 40 miles in 1844. Then on March 10, 1876, in Boston, Massachusetts, Alexander Graham Bell laid the foundation of telephone by making the first voice call over wire—"Mr. Watson, come here, I want to see you".

On October 4, 1957, the USSR (Union of Soviet Socialist Republic, now mainly Russia) launched the Sputnik. It was the first artificial earth satellite launched from Baikonur cosmodrome in Kazakhstan. This demonstrated the technological superiority of USSR. In response to this, the US formed the Advanced Research Projects Agency (ARPA) within the Department of Defense (DoD). The mandate for ARPA was to establish the US as a leader in science and technology. ARPA funded different research projects to help conduct research in computer networks. This laid the

foundation of packet switched data networks. There were multiple flavors of packet switched networks in the US and in Europe. The important ones are TCP/IP and X.25. TCP/IP was driven by education and defense in the US whereas X.25 was driven by European telecommunication industry and governments. With the evolution of computers and packet switched networks, movement of bits and bytes progressed to a new state of maturity. Over the last 175 years, virtual reality evolved from ringing an electric bell to mobile computing.

1.1.1 The Convergence Leading to ICT

The first step towards the convergence between telecommunication and IT happened in 1965 when AT&T used computers to do the switching in an electronic switching system (ESS). On the other hand, packet switch network was bringing communication closer to computers. The World Wide Web (WWW), which was started by Tim Berners-Lee in 1989 as a text processing software, brought these two faculties of technology together and established Internet as a powerful media. The Internet meets four primary needs of the society: communication, knowledge sharing, commerce, and entertainment. This convergence is called Information and Communications Technologies (ICT). Through ICT we are now moving towards an information-based society. ICT will address the need to access data, information, and knowledge from anywhere, anytime.

1.2 WIRELESS—THE BEGINNING

In 1947, researchers in AT&T Bell Labs conceived the idea of cellular phones. They realized that by using small service areas or cells they can reuse the frequency. This in turn can enhance the traffic capacity of mobile phones. AT&T requested the Federal Communication Commission (FCC) to allocate a large number of radio-spectrum frequencies so that widespread mobile telephone services would become feasible. FCC is a government agency in the US that regulates the usage and licensing of frequency bands. Every country has its regulatory agencies like FCC. In India the regulatory authority is Telecom Regulatory Authority of India (TRAI). FCC in the US is charged with regulating interstate and international communications by radio, television, wire, satellite and cable. Initially, FCC agreed to license a very small band to AT&T. This small frequency range made only 23 simultaneous phone conversations possible in one service area. With 23 channels there was no market incentive for either research or commercial deployment for AT&T. Though the idea of cellular telephony was very much there in the late forties, it did not take off.

1.2.1 Evolution of Wireless Networks

The first wireless network was commissioned in Germany in 1958. It was called A-Netz and used analog technology at 160 MHz. Only outgoing calls were possible in this network. That is to say that connection set-up was possible from the mobile station only. This system evolved into B-Netz operating at the same 160 MHz. In this new system, it was possible to receive an incoming call from a fixed telephone network, provided that location of the mobile station was known. This system was also available in Austria, the Netherlands, and Luxemburg. A-Netz was wireless but not a cellular network. Therefore, these systems (A-Netz and B-Netz) did not have any function,

which permitted handover or change of base station. The B-Netz had 13,000 customers in West Germany and needed a big transmitter set, typically installable in cars.

In 1968, in the US, the FCC reconsidered its position on the cellular network concept. FCC agreed to allocate a larger frequency band for more number of mobile phones provided the technology to build a better mobile service be demonstrated. AT&T and Bell Labs proposed a cellular system to the FCC with many small, low-powered, broadcast towers, each covering a hexagonal 'cell' of a few kilometers in radius. Collectively these cells could cover a very large area. Each tower would use only a few of the total frequencies allocated to the system. As the phones traveled across the area, calls would be passed from tower to tower.

Besides AT&T and Bell Labs, other enterprises were also engaged in research in the wireless domain. In April 1973, Martin Cooper of Motorola invented the first mobile phone handset and made the first call from a portable phone to Joel Engel, his rival in AT&T and Bell Labs. By 1977, AT&T and Bell Labs constructed a prototype of a public cellular network. In 1978, public trials of the cellular telephony system started in Chicago with over 2000 trial customers. In 1982, FCC finally authorized commercial cellular service for the US. A year later in 1983, the first American commercial analog cellular service AMPS (Advanced Mobile Phone Service) was made commercially available in Chicago. This was the first cellular mobile network in the world.

While the US was experiencing the popularity of cellular phones, Japan and Europe were not lagging behind. In 1979, the first commercial cellular telephone system began operations in Tokyo. During the early 1980s, cellular phone experienced a very rapid growth in Europe, particularly in Scandinavia and the United Kingdom. There was decent growth of cellular phones in France and Germany as well. The message was quite clear by then that mobile technology was here to stay.

To take advantage of this growing market, each country in Europe developed its own analog mobile system and joined the bandwagon. These cellular systems developed by each country in Europe were mutually incompatible. These incompatibilities made the operation of the mobile equipment limited to national boundaries. Also, a mobile subscriber of one network cannot use the same device in another network in another country. Though the market was growing, these incompatible systems made the market very limited for equipment manufacturers. This became an increasingly unacceptable situation in a unified Europe.

To cope with these problems Europeans decided to evolve a standard for mobile phone technology. In 1982, the Conference of European Posts and Telegraphs (CEPT) formed a study group called the Groupe Spécial Mobile (GSM) to develop a standard for the pan-European mobile system. In 1989, GSM responsibility was transferred to the European Telecommunication Standards Institute (ETSI), and GSM became a technical committee within ETSI. In 1990, Phase I of the GSM specifications were published. Commercial services of GSM started in mid-1991. Although standardized in Europe, GSM became popular outside Europe as well. Therefore, to give a global flavor, GSM was renamed as 'Global System for Mobile communications'. This has grown to more than 1 billion by the end of February 2004 in over 200 countries. At the beginning of 2009 the number of mobile phones in the world crossed the 4 billion mark—two phones in every three persons in the world. In the beginning of 1994, there were 1.3 million subscribers worldwide. In October 2004, the number of mobile subscribers in India crossed the number of fixed phones.

If we look at the critical success factor of GSM, we find quite a few technical and non-technical reasons for its tremendous success. These are:

- The developers of GSM sat together to arrive at a standard before they built the system. The advantage of standards is that they provide enough standardization to guarantee proper interoperability between different components of the system. GSM standards also facilitate the interworking between different vendors. Over 8000 pages of GSM recommendations ensure competitive innovation among suppliers.
- International roaming between networks. A subscriber from one network can seamlessly roam in another network and avail of services without any break.
- Emergence of SMS (Short Message Service) has spawned several applications within the GSM framework.
- The developers of GSM took considerable technological risks by choosing an unproven (in 1980s) digital system. They had the confidence that advancements in compression algorithms and digital signal processing would allow the continual improvement of the system in terms of quality and cost.

1.2.2 Evolution of Wireless Data

Like computers, the evolution of wireless technology has also been defined in generations. The first generation or 1G wireless technology uses analog technology. It uses FDMA (Frequency Division Multiple Access) technology for modulation; for example, AMPS (Advanced Mobile Phone Service) in the US. The second generation or 2G technology uses digitized technology. It uses a combination of TDMA (Time Division Multiple Access) and FDMA technologies. An example is GSM. In 2G technology, voice is digitized over a circuit. In 1G and 2G networks, data is transmitted over circuits. This technology is called Circuit Switched Data or CSD in short. Using modems, a data connection is established between the device and the network. This is similar to what happens in a dial-up network over analog telephones at home. The next phase in the evolution is 2.5G. In 2.5G technology, voice is digitized over a circuit. However, data in 2.5G is packetized. 2.5G uses the same encoding techniques as 2G. GPRS networks is an example of 2.5G. The Third Generation or 3G wireless technology makes a quantum leap from a technology point of view. 3G uses Spread Spectrum techniques for media access and encoding. In 3G networks, both data and voice use packets. UMTS and CDMA2000 are examples of 3G networks.

While 1G, 2G, or 3G were making their mark in the metropolitan area wireless networks (MAN), wireless technology has been getting popular in local area networks (LAN) and personal area networks (PAN). Wireless technology offers convenience and flexibility. With the success of wireless telephony and messaging services like paging, wireless communication is beginning to be applied to the realm of personal and business computing in the domain of local area networks. Wireless LANs are being deployed in homes, campuses, and commercial establishments. Wireless LANs are also being deployed in trains and commercial vehicles. The domain of wireless data networks today comprises Wireless PAN (Bluetooth, Infrared), Wireless LAN (IEEE 802.11 family) and Wireless WAN (Wide Area Networks) (GSM, GPRS, 3G).

1.2.3 Evolution of Wireless LAN

In late 1980s, vendors started offering wireless products, which were to substitute the traditional wired LAN (Local Area Network) ones. The idea was to use a wireless local area network to avoid

the cost of installing LAN cabling and ease the task of relocation or otherwise modifying the network's structure. When Wireless LAN (WLAN) was first introduced in the market, the cost per node was higher than the cost of its counterpart in the wired domain. However, as time progressed, the cost per node started dropping, making wireless LAN quite attractive. Slowly WLAN started becoming popular and many companies started offering products. The question of interoperability between different wireless LAN products became critical. IEEE Standards committee took the responsibility to form the standard for WLAN. As a result the IEEE 802.11 series of standards emerged.

WLAN uses the unlicensed Industrial, Scientific, and Medical (ISM) band that different products can use as long as they comply with certain regulatory rules. These rules cover characteristics such as radiated power and the manner in which modulation occurs. The ISM bands specified by the ITU-R are: 6.765–6.795 MHz, 13.553–13.567 MHz, 26.957–27.283 MHz, 40.66–40.70 MHz, 433.05–434.79 MHz, 902–928 MHz, 2.400–2.500 GHz, 5.725–5.875 GHz, 24.00–24.25 GHz, 61.00–61.5 GHz, 122–123 GHz, 244–246 GHz. WLAN uses 2.4 GHz and 5.8 GHz ISM bands. WLAN works both in infrastructure mode and ad hoc mode. WLAN is also known as Wireless Fidelity or WiFi in short. There are many products which use these unlicensed bands along with WLAN; examples could be cordless telephone, microwave oven, etc.

1.2.4 Evolution of Wireless PAN

Wireless technology offers convenience and flexibility. Some people will call this freedom from being entangled with the wire. The success of wireless technology in cellular telephones or Wireless MAN (Metropolitan Area Network) made people think of using the technique in Wireless LAN and Wireless Personal Area Network (WPAN). Techniques for WPANs are infrared and radio waves. Most of the laptop computers support communication through infrared, for which standards have been formulated by IrDA (Infrared Data Association—www.irda.org). Through WPAN, a PC can communicate with another IrDA device like another PC or a Personal Digital Assistant (PDA) or a Cellular phone.

The other best known PAN technology standard is Bluetooth. Bluetooth uses radio instead of infrared. It offers a peak over the air speed of about 2.1 Mbps over a short range of about 100 meters (power dependent). The advantage of radio wave is that unlike infrared it does not need a line of sight. WPAN works in ad hoc mode only.

1.3 Mobile Computing

Mobile computing can be defined as a computing environment of physical mobility. The user of a mobile computing environment will be able to access data, information, or other logical objects from any device in any network while on the move. A mobile computing system allows a user to perform a task from anywhere using a computing device in the public (the Web), corporate (business information) and personal information spaces (medical record, address book). While on the move, the preferred device will be a mobile device, while back at home or in the office the device could be a desktop computer. To make the mobile computing environment ubiquitous, it is necessary that the communication bearer is spread over both wired and wireless media. Be it for the mobile

workforce, holidayers, enterprises, or rural population, access to information and virtual objects through mobile computing is absolutely necessary for optimal use of resource and increased productivity.

Mobile computing is used in different contexts with different names. The most common names are:

- *Mobile Computing*: This computing environment moves along with the user. This is similar to the telephone number of a GSM (Global System for Mobile communication) phone, which moves with the phone. The offline (local) and real-time (remote) computing environment will move with the user. In real-time mode the user will be able to use all his remote data and services online.
- *Anywhere, Anytime Information*: This is the generic definition of ubiquity, where the information is available anywhere, all the time.
- *Virtual Home Environment*: Virtual Home Environment (VHE) is defined as an environment in a foreign network such that the mobile users can experience the same computing experience as they have in their home or corporate computing environment. For example, one would like to keep the room heater on when one has stepped outside for about 15 minutes.
- *Nomadic Computing*: The computing environment is nomadic and moves along with the mobile user. This is true for both local and remote services.
- *Pervasive Computing*: A computing environment, which is pervasive in nature and can be made available in any environment.
- *Ubiquitous Computing*: A (nobody will notice its presence) everyplace computing environment. The user will be able to use both local and remote services.
- *Global Service Portability*: Making a service portable and available in every environment. Any service of any environment will be available globally.
- *Wearable Computers*: Wearable computers can be worn by humans like a hat, shoe or clothes (these are wearable accessories). Wearable computers need to have some additional attributes compared to standard mobile devices. Wearable computers are always on; operational while on the move; hands-free, context-aware (with different types of sensors). Wearable computers need to be equipped with proactive attention and notifications. The ultimate wearable computers will have sensors implanted in the body and supposedly integrate with the human nervous system. These are part of a new discipline of research categorized by "Cyborg" (Cyber Organism).

1.3.1 Mobile Computing Functions

We can define a computing environment as mobile if it supports one or more of the following characteristics:

- *User Mobility*: The user should be able to move from one physical location to another and use the same service. The service could be in a home or remote network. For example, a user moves from London to New York and uses Internet to access the corporate application the same way the user uses it in the home office.
- *Network Mobility*: Network mobility deals with two types of use-cases. In one use-case, the user is moving from one network to another and uses the same service seamlessly. An example could be a user moving from a WiFi network within the university campus and changing to

3G network outside while using the same online service.

In other use-case of network mobility, the network itself is mobile like in a Mobile Ad hoc Network (MANET). In MANET, each node in the network is a combination of a host and a router. As the nodes move, the routers within the network also move changing the routing table structure. These types of networks are used in battlefields or sensor networks, where routers/nodes are constantly moving.

- *Bearer Mobility*: The user should be able to move from one bearer to another and use the same service. An example could be a user using a service through WAP bearer in his home network in Bangalore. He moves to Coimbatore where WAP is not supported and switches over to the voice or SMS (short message service) bearer to access the same application.
- *Device Mobility*: The user should be able to move from one device to another and use the same service. An example could be sales representatives using their desktop computer in their home office. During the day while they are on the street they would like to use their Palmtop to access the application.
- *Session Mobility*: A user session should be able to move from one user-agent environment to another. An example could be a user using his service through a CDMA (Code Division Multiple Access) 1X network. The user entered into the basement to park the car and got disconnected from his CDMA network. He goes to his home office and starts using the desktop. The unfinished session in the CDMA device moves from the mobile device to the desktop computer.
- *Agent Mobility*: The user-agent or the applications should be able to move from one node to another. Examples could be aglets, crawler software, or even a malicious worm or virus software that moves from one node to another. There is another use-case of mobile agent in the Cloud Computing paradigm, where applications will be moving from platform to platform and infrastructure to infrastructure depending on temporal and economic considerations. In Cloud Computing, there will not be any fixed association between the application and the host running it—software agents in the cloud will constantly be mobile.
- *Host Mobility*: The user device can be either a client or server. When it is a server or host, some of the complexities change. In case of host mobility, mobility of the IP needs to be taken care of.

The mobile computing functions can be logically divided into the following major segments (Fig. 1.1):

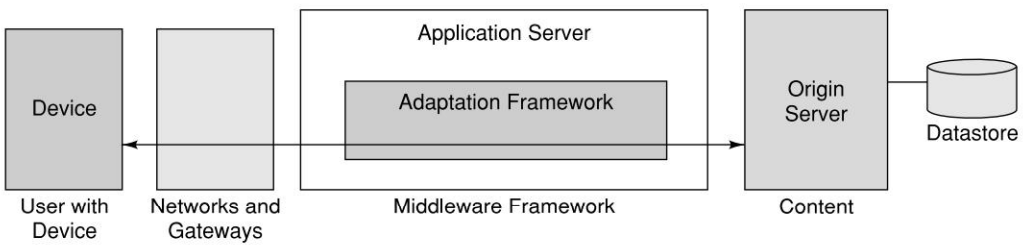


Figure 1.1 Mobile Computing Functions

1. *User with device*: This means that this could be a fixed device like a desktop computer in an office or a portable device like mobile phone. Example: laptop computers, desktop computers, fixed telephone, mobile phones, digital TV with set-top box, palmtop computers, pocket PCs, two-way pagers, handheld terminals, etc.
2. *Network*: Whenever a user is mobile, he will use different networks at different locations at different times. Example: GSM, CDMA, iMode, Ethernet, Wireless LAN, Bluetooth, etc.
3. *Gateway*: This acts as an interface between different transport bearers. These gateways convert one specific transport bearer to another. Example: From a fixed phone (with voice interface) we access a service by pressing different keys on the telephone. These keys generate DTMF (Dual Tone Multi Frequency) signals. These analog signals are converted into digital data by the IVR (Interactive Voice Response) gateway to interface with a computer application. Other examples will be WAP gateway, SMS gateway, etc.
4. *Middleware*: This is more of a function rather than a separate visible node. In the present context, middleware handles the presentation and rendering of the content on a particular device. It may optionally also handle the security and personalization for different users.
5. *Content*: This is the domain where the origin server and content is. This could be an application, system, or even an aggregation of systems. The content can be mass market, personal or corporate content. The origin server will have some means of accessing the database and storage devices.

1.3.2 Mobile Computing Devices

The device for mobile computing can be either a computing or a communication device. In the computing device category it can be a desktop, laptop, or a palmtop computer. On the communication device side it can be a fixed line telephone, a mobile telephone or a digital TV. Usage of these devices are becoming more and more integrated into a task flow where fixed and mobile, computing and communication functions are used together. The device is a combination of hardware and software; the hardware is technically called the User Equipment (UE) with software inside, which functions as an agent to connect to the remote service—this software is called a User Agent (UA). One of the most common UA today is a Web browser. When computing technology is embedded into equipment, Human-Computer Interaction (HCI) plays a critical role in effectiveness, efficiency, and user experience. This is particularly true as mobile information and communication devices are becoming smaller and more restricted with respect to information presentation, data entry and dialogue control. The human computer interface challenges are:

1. Interaction must be consistent from one device to another.
2. Interaction must be appropriate for the particular device and environment in which the system is being used.

Note: The requirement does not call for identical metaphors and methods. The desktop computer allows for different interaction techniques than a palmtop computer or a digital TV. Using the keyboard and a mouse may be appropriate for the desktop computer. Using the pen may be appropriate for the palmtop or Tablet PC. Microphones and speakers may be appropriate for a fixed or mobile phone. A remote control on the other hand will be more desirable for a digital TV.

1.4 DIALOGUE CONTROL

In any communication there are two types of user dialogues. These are long session-oriented transactions and short sessionless transactions. An example of a session-oriented transaction is: Reading a few pages from one chapter of a book at a time. Going to a particular page directly through an index and reading a particular topic can be considered a short sessionless transaction. Selection of the transaction mode will depend on the type of device we use. A session may be helpful in case of services offered through computers with large screens and mouse. For devices with limited input/output like SMS for instance, short sessionless transactions may be desired.

For example, consider enquiring about your bank balance over the Internet. In case of Internet banking through a desktop computer, the user has to go through the following minimum dialogues:

1. Enter the URL of the bank site.
2. Enter the account number/password and login into the application.
3. Select the balance enquiry dialogue and see the balance.
4. Logout from Internet banking.

This example is a session-oriented transaction. Using short sessionless transactions, the same objective can be met through a single dialogue. In a short sessionless transaction, the user sends an SMS message, 'mybal' to the system and receives the information on balance. The application services all the five dialogue steps as one dialogue. In this case steps like authentication and selection of transactions need to be performed in smarter ways. For example, user authentication will be done through the user's mobile number. It can be assumed that mobile devices are personal, therefore, authenticating the mobile phone implies authenticating the user account.

1.5 NETWORKS

Mobile computing will use different types of networks. These can be fixed telephone networks, GSM, GPRS, ATM (Asynchronous Transfer Mode), Frame Relay, ISDN (Integrated Service Digital Network), CDMA, CDPD (Cellular Digital Packet Data), DSL (Digital Subscriber Loop), Dial-up, WiFi (Wireless Fidelity), 802.11, Bluetooth, Ethernet, Broadband, etc.

1.5.1 Wireline Networks

This is a network, which is designed over wire or tangible conductors. This network is called fixedline or wireline network. Fixed telephone networks over copper and fiber-optic will be part of this network family. Broadband networks over Digital Subscriber Line (DSL) or cable will also be part of wireline networks. Wireline networks are generally public networks and cover wide areas. Though microwave or satellite networks do not use wire, when a telephone network uses microwave or satellite as part of its longhaul transmission infrastructure, it is considered part of wireline networks. When we connect to Internet Service Providers (ISP), it is generally a wireline network. The Internet backbone is a wireline network as well.

1.5.2 Wireless Networks

Mobile networks are called wireless network. These include wireless networks used by radio taxis, one-way and two-way pager, cellular phones. Examples will be PCS (Personal Cellular System), AMPS (Advanced Mobile Phone System), GSM, CDMA, DoCoMo, GPRS, etc. WiLL (Wireless in Local Loop) networks using different types of technologies are part of wireless networks as well. In a wireless network the last mile is wireless and works over radio interface. In a wireless network, other than the radio interface, rest of the network is wireline and is generally called the PLMN (Public Land Mobile Network).

1.5.3 Ad hoc Networks

In Latin, *ad hoc* means "for this purpose only". An ad hoc (or spontaneous) network is a small area network, especially one with wireless or temporary plug-in connections. In these networks some of the devices are part of the network only for the duration of a communication session. An ad hoc network is also formed when mobile or portable devices operate in close proximity to each other or with the rest of the network. When we beam a business card from our PDA (Personal Digital Assistant) to another, or use an IrDA port to print documents from our laptop, we have formed an ad hoc network. The term ad hoc has been applied to networks in which new devices can be quickly added using, for example, Bluetooth or wireless LAN (802.11). In these networks, devices communicate with the computer and other devices through wireless transmission. Typically based on short-range wireless technology, these networks don't require subscription services or carrier networks.

1.5.4 Bearers

For different type of networks, there are different types of transport bearers. These can be TCP/IP, HTTP, protocols or dial-up connection. For GSM it could be SMS, USSD (Unstructured Supplementary Service Data) or WAP. For mobile or fixed phone, it will be Voice.

1.6 Middleware and Gateways

Any software layered between a user application and operating system is a middleware. Middleware examples are communication middleware, object-oriented middleware, message-oriented middleware, transaction processing middleware, database middleware, behavior management middleware, Remote Procedure Call (RPC) middleware, etc. There are some middleware components like behavior management middleware, which can be a layer between the client device and the application. In a mobile computing context we need different types of middleware components and gateways at different layers of the architecture (Fig. 1.2). These are:

1. Communication middleware.
2. Transaction processing middleware.
3. Behavior management middleware.
4. Communication gateways.

1.6.1 Communication Middleware

The application will communicate with different nodes and services through different communication middleware. Different connectors for different services will fall in this category. Examples could be TN3270 for IBM mainframe services, or Javamail connector for IMAP or POP3 services.

1.6.2 Transaction Processing Middleware

In many cases a service will offer session-oriented dialogue (SoD). For a session we need to maintain a state over the stateless Internet. This is done through an application server. The user may be using a device, which demands a sessionless dialogue (SlD) made of short sessionless transactions whereas the service at the backend offers a SoD. In such cases a separate middleware component will be required to convert a SoD to a SlD. Management of the Web components will be handled by this middleware as well.

1.6.3 Behavior Management Middleware

Different devices deliver differently. We can have applications which are developed specially to deliver in a certain manner. For example, we can have one application for the Web, another for WAP, and a different one for SMS. On the contrary, we may choose to have a middleware, which will manage device-specific rendering at run-time. This middleware will identify the device properly and handle all device-specific rendering independent of the application. The system may be required to have some context awareness, which will be handled by the behavior management middleware.

1.6.4 Communication Gateways

Between the device and the middleware there will be a system of networks. Gateways are deployed when there are different transport bearers or networks with dissimilar protocols. For example, we need an IVR gateway to interface Voice with a computer, or a WAP gateway to access Internet over a mobile phone.

Figure 1.2 presents a schematic diagram of services in a mobile computing environment with different devices providing different services.

1.7 Application and Services (Contents)

Data and information, through mobile computing services, are required by all people regardless of their mobility. Mobile users include people like mobile executives, sales people, service engineers, farmers in the field, milkmen, newspaper boys, courier or pizza delivery boy. Logically, everyone is a mobile user at some time or the other in life. For people who are stationary, mobile computing becomes necessary outside office hours. For example, we may need to do a bank transaction from home at night or respond to an urgent mail while at home.

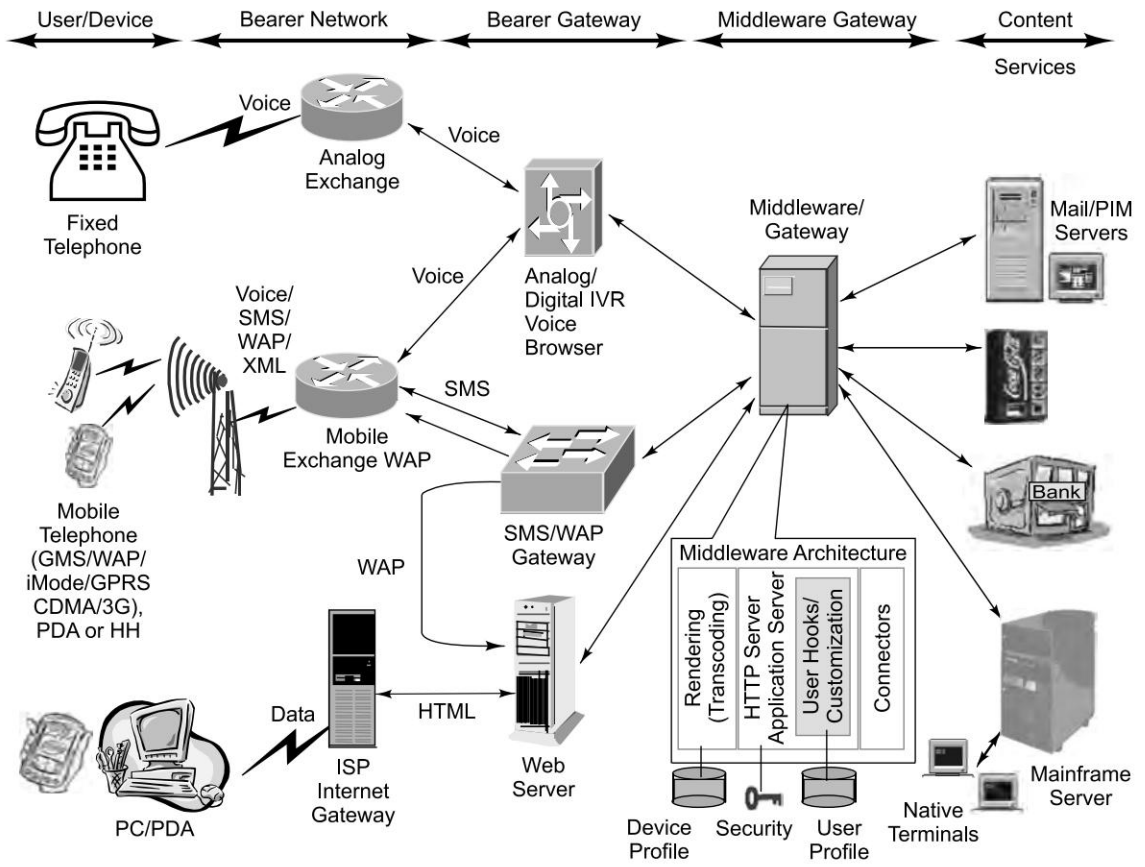


Figure 1.2 Schematic Representation of a Mobile Computing Environment

There can be many applications and services for the mobile computing space. These applications or services run on the origin server. These are also known as content servers. Content will primarily be lifestyle-specific. An individual has different lifestyles in different social environments. Also, lifestyles change during the day. One individual can be an executive needing the corporate MIS (Management Information System) application during the day while at home the same individual can use applications related to lifestyle or entertainment. The list of possible mobile applications can never be complete. On the basis of life styles, they can be grouped into different categories, such as:

Personal: Belongs to the user (wallet, medical records, diary).

Perishable: Time-sensitive and of relevance and passes quickly (general news, breaking news, weather, sports, business news, stock quotes).

Transaction-oriented: Transactions need to be closed (bank transactions, utility bill payment, mobile shopping).

Location-specific: Information related to current geographical location (street direction map, restaurant guide).

Corporate: Corporate business information {mail, Enterprise Requirements Planning (ERP), inventory, directory, business alerts, reminders}.

Entertainment: Applications for fun, entertainment. Social networking sites like Facebook can be part of this category.

Here are some examples:

News: This is a very big basket of applications having different types of news. News could be political, current affairs, breaking news, business news, sports news, community news, etc. While people are on the move, they can always be connected to their culture and community through news, using mobile computing.

Youth: This is a very high growth market with different applications to suit the lifestyles of the youth. These are primarily message-based applications like person-to-person messaging, chat, forums, dating, etc.

Weather: There are different types of applications and services where mobile computing can make a difference. Notification services on weather is a very sought after application. If we have access to information related to the weather while on vacation or while driving from one location to another then the global positioning system (GPS) can help locate a person or sometimes save lives in case of a natural calamity.

Corporate application: Standard corporate information is an important piece of information for mobile workers and includes corporate mail, address book, appointments, MIS applications, corporate Intranet, corporate ERP, etc.

Sales force automation: This group will offer many applications. It will cater to the large population of sales personnel. Applications will include sales order bookings, inventory enquiry, shipment tracking, logistics related applications, etc. These applications will be very effective over wireless devices.

m-broker: Getting correct and timely information related to different stocks are very important. Also, online trading of stocks while on the move is quite critical for certain lifestyles. Stock tickers, stock alerts, stock quotes, and stock trading can be made ubiquitous so that users can check their portfolio and play an active role in the market.

Telebanking: We need to access our bank information for different transactions. Earlier, people used to go to the bank, but things are changing. Banks are coming to customers through telebanking. If telebanking can be made ubiquitous it helps the customer as well as the bank. Many banks in India are today offering banking over Internet (web), voice and mobile phones through SMS.

m-shopping: This mobile application is used to shop with the help of mobile devices like Palm top, Pocket PC, mobile phones, etc. You can use this application to pay for a soft drink or soda from a vending machine in an airport or a movie theatre using a mobile phone, especially when you do not have sufficient cash.

Micropayment-based application: Micropayments involve transactions where the amount of money involved is not very high—it could be a maximum of Rs 1000 ($ 25) or so. Micropayment through mobile phones can help rural people to do business effectively.

Interactive games: Many mobile network operators have started offering different types of contests and interactive games that can be played through mobile phones. The applications could be similar to any quiz, housie, etc.

Interactive TV shows: Many TV companies around the world use email, SMS and Voice as a bearer for interactive TV or reality TV shows. In these shows viewers are encouraged to participate by asking questions, sharing opinions or even answering different quizzes. Nowadays viewers vote for their favorite TV stars using SMS.

Digital/Interactive TV: These are interactive TV programs through digital TV using set-top boxes and Internet. Video-on-demand, community programs, healthcare, and shopping applications are quite popular under this media category.

Experts on call: This is an application system for experts. Experts use these services to schedule their time and business with clients; clients use this to schedule business with the expert. A typical example could be to fix up an appointment with the tax consultant.

GPS-based systems: Applications related to location tracking come under this category. This could be a simple service like tracking a vehicle. Another example could be tracking an individual who got stuck due to bad weather while on a trekking trip. Fleet management companies and locations-aware software systems need GPS-based applications.

Remote monitoring: This is important for children at home where parents monitor where their children are or what are they doing. Also, monitoring and controlling of home appliances will be part of this application.

Entertainment: This contains a very large basket of applications starting from horoscope to jokes. Many people in some parts of Asia decide their day based on the planetary positions and horoscope information.

Directory services: This includes information related to movies, theatre, public telephones, restaurant guide, public information systems and Yellow pages.

Sports: This service offers online sports updates. In India live cricket score is the most popular mobile computing application. Getting scores of a live cricket match is the most popular mobile computer application. This service is available in India through Web, Voice, SMS, and WAP.

Maps/navigation guide: This is an application which has a lot of demand for traveling individuals. These services need to be location-aware to guide the user to use the most optimum path to reach a destination. The directions given by these applications also take traffic congestion, one way, etc., into consideration. GPS-based driving is becoming very popular in the US and advanced countries, where a user enters the postal address of the destination. The GPS-based system calculates the route, loads the right map and helps the driver navigate in real-time.

Virtual office: There are many people who are self-employed and do not have a physical office. Thus mobile and virtual office where they can check their mails, schedules, appointments, etc., while they are on the move are a must for them. Insurance agents and many other professions need these types of services.

m-exchange for industries: Manufacturing industry exchange from a mobile device can be a very cost effective solution for small/cottage industries. It may not be possible for a cottage industry to invest in a computer. However, accessing an exchange for a manufacturing company through a SMS may be affordable.

m-exchange for agricultural produce: Exchange for farmers on different type of agricultural products can be very useful for countries like India. If farmers can get information about where to get a good price for their product, it helps both farmers and consumers. There is a system *www.echoupal.com* to do exactly this. Think of this available over mobile phones.

Applications for speech/hearing challenged people: Telecommunication always meant communicating through Voice. There are people who cannot speak or hear. These include people with disabilities and senior citizens who lost their speech due to old age or after suffering a stroke. Text-based communication can help rehabilitate some of these disabled individuals.

Agricultural information: Think about a case where a farmer receives an alert in his local language through his mobile phone and immediately knows that the moisture content in air is 74%. He can then decide how much to water his harvest. This can save his money, the harvest (excess water is sometimes harmful), and the scarce water resource. Portable devices with voice interface can change the economics of rural India with this kind of application.

Corporate knowledge-based applications: Many corporates today have a knowledge base. Making this ubiquitous can reduce cost and increase productivity.

Community knowledge-based applications: Knowledge is equally important for a community. Making knowledge ubiquitous always help society.

Distance learning: Applications related to distance learning are a must for countries with limited or no access to digital and information technology. For virtual schools in Asia or Africa, it is possible to have access to good faculty through the distance learning mode.

Digital library: These are libraries which can be accessed from anywhere anytime because of the Internet. Digital libraries can go a long way in shortening the digital divide as they also have support of local language and are easy and cheaper to commission.

Telemedicine and healthcare: Making telemedicine and healthcare easily available can save many lives. For example, a person complains of chest pain while traveling and requires immediate medical attention. He has to be taken to a doctor in a remote town. In this case, access to the patient's record can help expedite diagnosis. Reminder services for medicines or checkups can be very useful. In rural India, virtual clinics can help those who otherwise do not have access to medical care.

Micro-credit schemes: Micro-credit has a distinct role to play for a country's microeconomy. Grameen Bank with all its applications in Bangladesh is the best example of micro-credit.

Environmental protection and management: Ubiquity is a must for applications on environmental protection and management. Applications related to industrial hygiene will be part of this category.

e-governance: These applications are very important to bridge the digital divide. The Bhoomi project of Karnataka government has computerized 20 million land records of 0.67 million farmers living in 30,000 villages in the state. Many such projects of the government can be made electronic, resulting in better and faster access to information managed by the government.

Virtual laboratories: There are many laboratories and knowledge repositories around the world which are made accessible to various cultures and countries through digital and information technology.

Community forums: There are different social and community meetings. In the case of India, panchayats can be made electronic. These may help increase the involvement of more people in community development work.

Law enforcements: Most of the time law enforcement staff are on the streets and need access to different types of services through wireless methods. These may be access to criminal records, information related to vehicles, or even a picture of the accident site taken through a MMS phone. This information can help insurance companies to resolve the claim faster.

Job facilitator: These could be either proactive alerts or information related to jobs and employment opportunities.

Telemetric applications: Almost every industry and sphere of life has the need for telemetric applications. Examples could be monitoring and control in manufacturing industry; vehicle tracking; meter reading; health care and emergency services; vending machine monitoring; research (telemetric orthodontic); control and service request for different emergency services for utilities like power plants, etc.

Downloads: Different types of downloads starting from ring tones to pictures are part of this category. In many countries this type of application is very popular. It is estimated that the market for ring tone downloads is more than 1 billion dollars.

Alerts and notifications: This can be either business or personal alerts. Simple examples could be breaking news alerts from a newspaper. Complex examples of alert could be for a doctor when the patient is in critical condition. In India many mobile operators are offering cricket alerts. In this service, subscribers receive score information every 15 minutes, about every wicket fall!

1.8 Developing Mobile Computing Applications

Any portal system today supports user mobility. If I have an Internet mail account like Google-mail or Yahoo-mail, I can access my mail from anywhere. I need a desktop or laptop computer to access my mailbox. I may not be able to access the same mail through some other device like a fixed phone. There are a number of factors that make mobile computing different from desktop computing. As a result of mobility, the attributes associated with devices, network, and users are constantly changing. These changes imply that the context and behavior of applications need to be adapted to suit the current environment. Context and behavior adaptation are required to provide a service that is tailored to the user's present situation. There are several ways in which

context and behavior can be adapted. One way is to build applications without any context or behavior awareness. Context and behavior adaptation will be handled by a behavior management middleware at runtime. Another option is to build different applications specific to different context and behavior patterns. There could be some system in the organization, which was originally developed 15 years ago for some direct connected terminals like VT52. Due to change in market expectation these systems need to be made mobile. Complexities involved in making an existing application mobile versus developing a new mobile system will be different. For a new application it is possible to embed the behavior within the application. However, for a long-life system or a legacy application the content behavior adaptation will need to be done externally.

1.8.1 New Mobile Applications

Let us assume that in a bank, some new applications need to be built for e-Commerce. The bank wants to offer banking through Voice (telephone) and Web (Internet). Assuming that the bank already has a computerized system in place, the bank will develop two new applications. One will handle the telephone interface through Interactive Voice Response (IVR) and the other through Web. At a later point in time, if the bank decides to offer SMS and WAP, they will develop two new applications to support SMS and WAP interfaces respectively. To protect the investment and quick adaptation, the bank may decide to use transaction processing middleware and RPC middleware. All these are possible only if it is a fresh applications development.

1.8.2 Making Legacy Application Mobile

How do we make an existing legacy application which has been functioning for long mobile? We define an application as legacy if it has one or more of the following characteristics:

1. It has moved into the sustenance phase in the software development lifecycle.
2. It cannot be modified. This could be due to unavailability of the original development platforms, unavailability of original source code or unavailability of expertise to make the necessary changes.
3. It is a products and packaged software where enterprise does not have any control. This could be due to high cost of ownership for the new upgrade or the vendor does not have any plan to support the new requirement.

Let us assume that an enterprise has licensed an ERP system from an external vendor. The enterprise wants to offer a notification of yesterday's sales figures to some select executives at 9:30 a.m. every morning through SMS. The ERP vendor plans to offer a similar function in its next release six months down the line. The license fee for the next upgrade will be very expensive. Another example is that a wireless network operator wants to offer enterprise mails through its network. In all such cases the adaptation will be done without changing the base product. This requires a framework that attempts to perform most of the adaptation dynamically. Content and behavior management will be managed real-time through a behavior management middleware.

1.9 Security in Mobile Computing

Security issues in mobile computing environment pose a special challenge. This is because we have to offer services over the air using networks over which we do not have any control. All the infrastructure and technology designed by GSM and other forums are primarily to increase the revenue of the network operators. This makes the technology complex and very much dependent on the network operator. For example, the SMS technology is operator centric; WAP requires WAP gateway. These gateways are installed in the operator's network and managed by the operator. The security policy implemented by the network operator depends on the operator's priority and revenue generation potential and not on the need of the content provider.

In a mobile computing environment, the user can move from one network to another, one device to another or one bearer to another. Therefore, theoretically the security implementations need to be device independent, network independent, bearer independent, and so on. The requirement is to arrive at a security model, which can offer homogenous end-to-end security.

1.10 Standards—Why Are they Necessary?

Standards are documented agreements containing technical specifications or other precise criteria to be used consistently as rules, guidelines or definitions of characteristics. Standards ensure that materials, products, processes and services are fit for their defined and agreed purpose. A standard begins as a technical contribution, which is supported by a number of interested parties to the extent that they indicate their willingness to participate in the standard's development. Standards are available for experts to challenge, examine and validate. No industry in today's world can truly claim to be completely independent of components, products, and rules of application that have been developed in other sectors. Without standards, interoperability of goods and services will not be possible.

When the proposed standard or technical document is near completion, the formulating engineering committee circulates the draft of the document for a ballot. The purpose of this ballot is to identify any unresolved issues and to establish consensus within the formulating group. Every effort is made to address and resolve the comments received.

The opposite of standard is proprietary. Proprietary systems for similar technologies are seen as technical barriers to trade and competition. Today's free-market economies increasingly encourage diverse sources of supply and provide opportunities for expanding markets. On the technology front, fair competition needs to be based on identifiable and clearly defined common references that are recognized from one country to another, and from one region to the next. An industry-wide standard, internationally recognized, developed by consensus among trading partners, serves as the language of trade.

There are some fundamental differences between how the US and Europe adapt technology. In the US, market force and time to market drive the technology. Interoperability has always been the primary issue in Europe. Therefore, in Europe, standards drive the adaptation of technology. This is one of the reasons why the US has more proprietary systems compared to Europe.

1.10.1 Who Makes the Standards?

There are many institutes that generate and provide standards across the world. There are standard bodies at the regional, country as well as international level. Based on the area of operations, standard bodies are formed by the governments, professional institutes or industry consortiums. These standard bodies sometimes also function as regulators. In India there is a standard body under the Government of India, which is called Bureau of Indian Standard or simply BIS (www.bis.org.in). A standards process include the following steps:

1. Consensus on a proposed standard by a group or "consensus body" that includes representatives from materially affected and interested parties.
2. Broad-based public review and comments on draft standards.
3. Consideration of and response to the comments submitted by voting members of the relevant consensus body and by public review commenters.
4. Incorporation of approved changes in a draft standard.
5. Right to appeal by any participant who believes that due process principles were not sufficiently respected during the standards development in accordance with the ANSI-accredited procedures of the standards developer.

1.11 Standards Bodies

The International Organization for Standardization (ISO) (http://www.iso.ch) is a worldwide federation of national standards bodies from more than 140 countries, one from each country. ISO is a non-governmental organization established in 1947. The mission of ISO is to promote the development of standardization and related activities in the world with a view to facilitating the international exchange of goods and services, and to developing cooperation in the spheres of intellectual, scientific, technological and economic activity. Though ISO is commonly believed as the acronym for International Standard Organization, in fact the word "ISO" is derived from the Greek isos, meaning "equal". Sometimes ISO makes its own standard or it adapts standards from its member organizations. The adapted standard is then made an international one by ISO. One of the most widely known standard from ISO is ISO 9000. ISO 9000 relates to software quality. The famous 7-layer model for Open System Interconnection (OSI) is ISO standard (ISO 7498). For information security ISO has come up with the recommendation ISO 17799.

Internet Engineering Task Force (IETF) (http://www.ietf.org) is the standard-making body for Internet and related technologies. IETF is an open international community of network designers, operators, vendors and researchers concerned with the evolution of Internet architecture and smooth operation of the Internet. It is open to any individual. The actual technical work of the IETF is done in its working groups. Working groups are

organized into several areas by topic (e.g., routing, transport, security, etc.). The Internet Assigned Numbers Authority (IANA) is the central coordinator for the assignment of a unique IP address. The IANA is chartered by the Internet Society (ISOC) to act as the clearing house to assign and coordinate the use of numerous Internet protocol parameters. Standards defined by IETF are called Request For Comment or RFC. The standard for email is defined in RFC821 (Simple Mail Transfer Protocol or SMTP); RFC2616 describes the version 1.1 of Hypertext Transfer Protocol (HTTP/1.1).

ETSI (the European Telecommunications Standards Institute) (http://www.etsi.org) is an organization whose mission is to produce telecommunications standards that will be used for decades to come throughout Europe and possibly beyond. ETSI unites members from countries inside and outside of Europe, and represents regulators, network operators, manufacturers, service providers, research bodies and users. ETSI plays a major role in developing a wide range of standards and other technical documentation as Europe's contribution to world-wide standardization in telecommunications, broadcasting and information technology. ETSI's prime objective is to support global harmonization by providing a forum in which all the key players can contribute actively. ETSI is officially recognized by the European Commission. GSM Standard is created, maintained and managed by a committee within ETSI. GSM standards document GSM 01.04 (ETR 350): 'Digital cellular telecommunications system (Phase 2+); Abbreviations and acronyms'. GSM 12.13 standard defines the interface digital cellular telecommunications system (Phase 2+); Man-Machine Interface (MMI) of the Mobile Station (MS) (GSM 02.30 version 7.1.0 Release 1998).

OMA and WAP Forum (http://www.wapforum.org) (http://www.openmobilealliance.org): The Open Mobile Alliance (OMA) has been established by the consolidation of the WAP Forum and the Open Mobile Architecture initiative. It intends to expand the market for the entire industry by removing barriers to interoperability and supporting a seamless and easy-to-use mobile experience for end users. The Open Mobile Alliance encourages competition through innovation and differentiation, while ensuring the interoperability of mobile services through the entire value chain. The supporters of the Open Mobile Alliance recognize the significant industry benefits of creating a standards organization that will include all elements of the wireless value chain, and contribute to timely and efficient introduction of services and applications to the market. WAP and MMS standards are created, maintained, and managed by OMA.

ITU (International Telecommunication Union) (www.itu.int) is an organization within the United Nations System. It was founded on the principle of cooperation between governments and the private sector. With a membership encompassing telecommunication policy-makers and regulators, network operators, equipment manufacturers, hardware and software developers, regional standards-making organizations and financing institutions, ITU's activities, policies and strategic direction are determined and shaped by the industry it serves. ITU has three sectors of the Union; they are Radio communication (ITU-R), Telecommunication Standardization (ITU-T), and

Telecommunication Development (ITU-D). Their activities cover all aspects of telecommunication, from setting standards that facilitate seamless interworking of equipment and systems to adopting operational procedures for the wireless services, and designing programs to improve telecommunication infrastructure. ITU Telecommunication Standardization Sector (ITU-T)'s mission is to ensure an efficient and on-time production of high quality standards (recommendations) covering all fields of telecommunications. ITU-T was founded in 1993, replacing the former International Telegraph and Telephone Consultative Committee (CCITT) whose origins go back to 1865. Any telephone in this world has a unique number (technically known as global title). These numbering schemes are defined through the ITU-T standards E.164.

IEEE Standards Association (IEEE-SA) (http://standards.ieee.org) is an organization that produces standards, which are developed and used internationally. While the IEEE-SA focuses considerable resources on the long-respected full consensus standards process carried out by the standards committees and IEEE societies, the IEEE-SA pioneers new and innovative programs to increase the value of IEEE standards to members, industry, and the global society. IEEE-SA members continue to set the pace for the development of standards products, technical reports and documentation that ensure sound engineering practices worldwide. IEEE-SA demonstrates strong support of an industry-led consensus process for the development of standards and operating procedures and guidelines. Standards for Wireless LAN are created, maintained and managed by IEEE. These are defined through different 802.11 standards.

The Electronic Industries Alliance (EIA) (http://www.eia.org) is a national trade organization within the US that includes the full spectrum of its electronics industry. The Alliance is a partnership of electronic and high-tech associations and companies whose mission is promoting the market development and competitiveness of the US high-tech industry through domestic and international policy efforts. EIA comprises companies whose products and services range from the smallest electronic components to the most complex systems used by defense and space and industry, including the full range of consumer electronic products. The progressive structure of the Alliance enables each sector association to preserve unique autonomy while uniting for a common cause under EIA. One of the most commonly used EIA standard is EIA RS-232. This is a standard for the 25-pin connector between a computer and a modem.

World Wide Web Consortium (W3C) (http://www.w3.org) develops interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential. W3C is a forum for information, commerce, communication and collective understanding. By promoting interoperability and encouraging an open forum for discussion, W3C is commited to leading the technical evolution of the Web. To meet the growing expectations of users and increasing power of machines, W3C is already laying the foundations for the next generation of the Web. W3C's technologies will help make the Web a robust, scalable and adaptive infrastructure for a world of information. W3C contributes to efforts to standardize Web technologies by producing specifications (called

recommendations) that describe the building blocks of the Web. W3C recommendations include HTML, XML, CSS (Cascading Style Sheet), Web Services, DOM (Document Object Model), MathML (Maths Markup Language), PNG (Portable Network Graphics), SGV (Scalable Vector Graphics), RDF (Resource Description Framework), P3P (Platform for Privacy Preferences), etc.

3GPP (http://www.3gpp.org) is to produce globally applicable technical specifications and technical reports for 3rd Generation Mobile System based on evolved GSM core networks and radio access technologies that they support, i.e., Universal Terrestrial Radio Access (UTRA) both Frequency Division Duplex (FDD) and Time Division Duplex (TDD) modes. The scope was subsequently amended to include maintenance and development of the Global System for Mobile communication (GSM) technical specifications and technical reports including evolved radio access technologies (e.g., General Packet Radio Service (GPRS) and Enhanced Data rates for GSM Evolution (EDGE)).

The American National Standards Institute (ANSI) (www.ansi.org) is the national standard organization in the United States. In many instances, the US standards are taken forward to ISO and IEC (International Electrotechnical Commission), where they are adapted in whole or in part as international standards. For this reason, ANSI plays an important part in creating international standards that support the worldwide sale of products, which prevent regions from using local standards to favor local industries. ANSI Standard X3.4-1968 defines the 'American National Standard Code for Information Interchange (ASCII)' character set. ASCII character set is used in almost every modern computer today. The same standard has also been adapted as ISO 8859-1 standard.

Universal Mobile Telecommunications System (UMTS) (www.umts-forum.org) represents an evolution in terms of services and data speeds from today's second-generation mobile networks like GSM. As a key member of the global family of third generation (3G) mobile technologies identified by the ITU, UMTS is the natural evolutionary choice for operators of GSM networks. Using fresh radio spectrum to support increased numbers of customers in line with industry forecasts of demand for data services over the next decade and beyond, UMTS is synonymous with a choice of WCDMA radio access technology that has already been selected by many licensees worldwide. UMTS-Forum is the standards-making body for WCDMA (Wideband Code Division Multiple Access) and UMTS technology.

Bluetooth (http://www.bluetooth.com) wireless technology is a worldwide specification for a small-form factor, low-cost radio solution that provides links between mobile computers, mobile phones, other portable handheld devices, and connectivity to the Internet. The standards and specification for Bluetooth are developed, published and promoted by the Bluetooth Special Interest Group.

The CDMA Development Group (CDG) (http://www.cdg.org) is an international consortium of companies who have joined together to lead the adoption and evolution of CDMA wireless systems around the world. The CDG comprises the world's leading CDMA service providers and manufacturers. By working together, the

members will help ensure interoperability among systems, while expediting the availability of CDMA technology to consumers.

The Public-Key Cryptography Standards (PKCS) (http://www.rsasecurity.com/rsalabs/pkcs/) are specifications produced by RSA Laboratories in cooperation with secure systems developers worldwide for the purpose of accelerating the deployment of public-key cryptography. First published in 1991 as a result of meetings with a small group of early adopters of public-key technology, the PKCS documents have become widely referenced and implemented. Contributions from the PKCS series have become part of many formal and de facto standards, including ANSI X9 documents, PKIX, SET, S/MIME and SSL (Secure Sockets Layer).

PAM Forum (http://www.pamforum.org). In the world of ubiquitous computing, knowing the position and context of a device is very important. The Presence and Availability Management (PAM) Forum is an independent consortium with a goal to accelerate the commercial deployment of targeted presence and availability applications and services that respect users' preferences, permissions and privacy. Working in partnership with industry participants, the PAM Forum defines a framework for the various standards and specifications needed for context/location aware applications.

Parlay Group (http://www.parlay.org). The Parlay Group is a multi-vendor consortium formed to develop open, technology-independent application programming interfaces (APIs). Parlay integrates intelligent network (IN) services with IT applications via a secure, measured, and billable interface. By releasing developers from underlying code, networks and environments, Parlay APIs allow for innovation within the enterprise. These new, portable, network-independent applications are connecting the IT and telecom worlds, generating new revenue streams for network operators, application service providers (ASPs), and independent software vendors (ISVs). Using Parlay APIs, one will be able to develop applications, which rely on network-related data like service provisioning. Parlay will also help develop location/context aware applications and services.

DECT (www.dect.org) stands for Digital Enhanced Cordless Communications. It is an ITSI standard for portable phones. DECT is known in ITU as a 3G system and is commonly referred to as IMT-FT (IMT Frequency Time).

WiMAX Forum (www.wimaxforum.org) is Worldwide Interoperability for Microwave Access Forum dedicated to certifying the operations of interconnecting devices. WiMAX aims to provide wireless data over long distances in different forms ranging from point-to-point links to full scale mobile access networks for wireless broadband communication. WiMAX Forum is the industry body that does the job of certification, promotion and producing interoperability specifications for WiMAX.

TTA (www.tta.or.kr/English/new/main/index.htm) is Telecommunications Technology Association. TTA is an IT standards organization catering to development of new standards based in Korea. It provides one-stop services for comprehensive IT standards.

Wi-Fi (www.wi-fi.org) owns trademark to Wi-Fi alliance. It was previously known as Wireless Ethernet Compatibility Alliance. It is focused on interoperability and compatibility of Wireless LAN devices and committed to continuous improvements in design and better user experience.

Association of Radio Industries and Businesses (ARIB) (www.arib.or.jp/english/) is an institution, based in Japan, dedicated to efficient use of radio spectrum and its implications in businesses.

China Communications Standards Association (CCSA) (www.ccsa.org.cn/english/) is an attempt of Chinese Ministry of IT to reform telecommunications industry and market. It aims to become a nationally unified standards organization in China.

Digital Living Network Alliance (DLNA) (www.dlna.org) is a cross-industry association of consumer electronics, computing industry and mobile device companies. The objective of DLNA is to establish a conglomeration of wired and wireless interoperable network of personal computers, consumer electronics and mobile devices in the home and outside in order to enable a seamless environment for sharing digital media content.

1.12 Players in the Wireless Space

In a wireless network there are many stakeholders. These are:

1. Regulatory authorities.
2. The operator or service provider.
3. The user or the subscriber.
4. Equipment vendors (network equipment and user device).
5. Research organizations.

In most parts of the world, the radio spectrums are regulated. Generally, a license is required to use a part of this spectrum. There are certain bands like ISM (Industry, Scientific and Medical) used by cordless telephones or microwave ovens or 802.11 which are unregulated. This means that if one develops equipment which needs to use these bands, then one need not go to the government and ask for permission to use them. However, for the regulated bands, one has to get a license before it can be used. GSM, CDMA etc., use frequency bands, which are regulated. Therefore, a network company offering these services, needs to get clearance from the government. Governments generally auction these spectrums to different network operators. The spectrums for 3G networks were auctioned in Europe for about 100 billion US Dollars. In India, the whole country was divided into metros and circles. The average license fee for GSM for these circles was in the tune of hundreds of crores of rupees.

Once the license is obtained, the network operator needs to conduct a detailed survey of the region with a plan for the cell sites. Cell site survey is very important and critical for any wireless network. Cell site survey is logically the design of the architecture of the network. During this phase the network operator determines the location of the base station and positioning of the cell. The location of a wireless base station tower will be determined by many factors; examples could be subscriber density, hills, and other obstacles. Similarly, for wireless LAN or WiFi, the location of AP (Access Points) will be determined by the layout of the building floor, concrete, glass walls, etc. A site survey is necessary for a wireless LAN as well before the APs are installed.

Cellular network operators need to create the infrastructure. There are a few equipment manufacturers who supply the hardware to the network operators. These hardware will be MSC (Mobile Switching Centre), BSC (Base Station Controller), BTS (Base Transceiver Station), and the Cells. Cells and BTSs are spread across the region; however, MSCs and BSCs are generally installed under one roof commonly known as switching room. Some of the leading manufacturers of these hardware are Ericsson and Nokia in Europe; Motorola and Lucent in the US; Samsung in Asia.

To use a cellular network, we need a handset or device. There are different types of handsets available in the market today. All these devices offer voice and SMS as minimum, and range up to fancy handsets which offer WAP, MMS, J2ME or even digital cameras. Some of the leading suppliers of these handsets are Nokia, Sony Ericsson, Motorola, Samsung, LG, etc.

In GSM world, all these handsets contain a small piece of card known as SIM (Subscriber Identity Module). These are technically Smart Cards or processor cards with a small memory and an independent processor. The size of the memory ranges from 8K bytes to 64K bytes. They contain some secured data installed by the network operator related to the subscriber and the network. Some of the leading suppliers of SIM card are Gemplus, Schlumberger, Orga, etc.

When a person wants to subscribe to a cellular phone, he contacts a cellular operator. The subscriber is then registered with the network as a prepaid or postpaid subscriber. In a GSM network, the operator issues a SIM card to the subscriber that contains all relevant security information. The subscriber buys a handset and installs the SIM card inside the handset; however, in Europe and the US, in certain subscription plans the handset is bundled with the plan. A provisioning and activation needs to be done within the network for this new subscriber. During the provisioning, some of the databases within the operator will be updated. Once the databases for authentication and billing are completed, the subscriber is activated. Following the activation, the subscriber can use the network for making or receiving calls.

REFERENCES/FURTHER READING

1. 'An Investment Perspective' *UMTS Report*, Durlacher Research, www.durlacher.com.
2. 'Africa: The Impact of Mobile Phones, Moving the debate forward', *The Vodafone Policy Paper Series*, 2, March 2005.
3. Banks, Ken and Richard Burge, (2004), *Mobile Phones: An Appropriate Tool for Conservation and Development*, Fauna & Flora International, UK: Cambridge.
4. 'Enabling UMTS Third Generation Services and Applications', *UMTS Forum Report* # 11, October 2000, http://www.umts-forum.org.

5. Eyers, Dean, 'Telecom Markets and the Recession: An Imperfect Storm', *Gartner Report*, AV-14-9944, 27 November 2001.
6. Horrigan, John B., Senior Researcher, Lee Rainee, Director, (2002), 'Getting Serious Online', *Pew Internet & American Life Project Report,* March 2002.
7. Jhunjhunwala, Ashok, Bhaskar Ramamurthi, and Timothy A. Gonsalves, (1998), 'The Role of Technology in Telecom Expansion in India', *IEEE Communications Magazine*, November 1998.
8. Lewin David and Susan Sweet, 'The economic benefits of mobile services in India: A case study for the GSM Association', *OVUM*, CLM28, Version 1, January 2005.
9. McGraw, Alistair, Christophe de Hauwer, Tim Willey, and Adam Mantzos, 'Industry Analysis Wireless Data: The World in Your Hand, Arthur Andersen Technology.' *Media and Communications*, October 2000.
10. Milojicic, D., F. Douglis, and R. Wheeler, (Eds), (1999), *Mobility Processes, Computers, and Agents*, Addison-Wesley.
11. Talukder Asoke K., (2002), *Mobile Computing–Impact in Our Life, Harnessing and Managing Knowledge*, Chakravarthy, C.R., L.M. Pathak, T. Sabapathy, M.L. Ravi (Eds), 13, Tata McGraw-Hill.
12. 'The Future Mobile Market Global Trends and Developments with a Focus on Western Europe, *UMTS Forum Report* # 8, March 1999, http://www.umtsforum.org.
13. 'The Path Towards UMTS: Technologies for the Information Society', *UMTS Forum Report* # 2, 1998, http://www.umts-forum.org.

REVIEW QUESTIONS

Q1: What are the essential functional differences between 1st generation, 2nd generation, and 3rd generation of networks?
Q2: Describe what do you understand by Wireless PAN, Wireless LAN and Wireless MAN.
Q3: What is an ISM band? Why is it called a free band?
Q4: What are the characteristics of a mobile computing environment?
Q5: Give examples for five mobile computing applications.
Q6: What are the advantages and disadvantages of standards? Name the standard committees responsible for 3G?
Q7: Describe the variants of Mobile Computing.
Q8: Describe the various aspects of mobility with respect to Mobile Computing.
Q9: What should be the characteristics of Mobile Computing devices?
Q10: How should dialogues be controlled for communication in a Mobile Computing environment?
Q11: Briefly describe the following networks with example and applications:
(a) Wired networks
(b) Wireless networks
(c) Ad hoc networks

Q12: What are the differences between middleware and gateways? Enunciate with examples in the context of Mobile Computing?

Q13: How would you broadly classify Mobile Computing applications?

Q14: Describe the design of Mobile Computing applications using at least two transport communication bearers. Make assumptions, if required.

Q15: How could one achieve the migration of legacy application to it being mobile?

Q16: Who are the players in wireless space? Explain the role of each of them assuming you are going to launch a mobile computing application in the commercial domain and what should you be prepared with when tackling them.

CHAPTER 2

Mobile Computing Architecture

2.1 HISTORY OF COMPUTERS

Nothing has changed the world around us the way digital technology and computers have. Computers have entered every aspect of our life and the environment around us. The origin of computers can be traced back to thousands of years. Though different forms of computers were in existence for centuries, the real transformation happened with electronic or digital computers. Development of the electronic computer started during the Second World War. In 1941, German engineer Konrad Zuse developed a computer called Z3 to design airplanes and missiles. In 1943, the British developed a computer called Colossus for cryptanalysis to decode encrypted messages transmitted by Germans. With a team of engineers in 1944, Howard H. Aiken developed the Harvard–IBM Automatic Sequence Controlled Calculator Mark I, or Mark I for short. This is considered as the early general-purpose computer. In 1945, John von Neumann introduced the concept of stored program. Another general-purpose computer development spurred by the war was the Electronic Numerical Integrator and Computer, better known as ENIAC, developed by John Presper Eckert and John W. Mauchly in 1946. In 1947, the invention of the transistor by John Bardeen, Walter H. Brattain, and William Shockley at Bell Labs changed the development scenario of digital computers. The transistor replaced the large, energy-hungry vacuum tube in first generation computers. Jack Kilby, an engineer with Texas Instruments, developed the integrated circuit (IC) in 1958. IC combined all the essential electronic components (inductor, resistor, capacitor, etc.) on to a small silicon disc, which was made from quartz. By the 1980s, very large scale integration (VLSI) squeezed hundreds of thousands of components on to a chip. VLSI led the development of third generation computers. All these early computers contained all the components we find today in any modern-day computers like printers, persistent storage, memory, operating systems and stored programs. However, one aspect of modern-day computers was missing in these machines—that was the networking aspect of today's computers.

2.2 History of Internet

Following the successful launch of Sputnik in 1957 by the Russians, the US felt the need of research in certain focused areas. Therefore, Advance Research Project Agency (ARPA) was formed to fund Science and Technology projects and position the US as a leader in technology. Internet represents one of the best examples of the benefits of sustained investment on research and development through ARPA. Beginning with the early research in packet switching, the government, industry and academia have been partners in evolving and deploying the exciting Internet technology. People in almost all parts of life starting from education, IT, telecommunications, business, and society have felt the influence of this pervasive information infrastructure. Today, almost everybody uses terms like "faculty@iiitb.ac.in" or "http://www.isoc.org".

In the early sixties, Leonard Kleinrock developed the basic principles of packet switching at Massachusetts Institute of Technology (MIT). During the same period Paul Baran in a series of RAND Corporation reports recommended several ways to accomplish packet switch network as well. In 1965, Lawrence G. Roberts in association with Thomas Merrill, connected the TX-2 computer in Massachusetts to the Q-32 in California with a low speed dial-up telephone line creating the first computer network. In 1971, Ray Tomlinson at BBN wrote the software to send and read simple electronic mail. In October 1972, demonstration of the ARPANET was done at the International Computer Communication Conference (ICCC). This was the first public demonstration of this new network technology to the public. It was also in 1972 that the initial 'hot' application, electronic mail, was introduced. In 1973, work began on the Transmission Control Protocol (TCP) at a Stanford University laboratory headed by Vincent Cerf.

In 1986, the US NSF (National Science Foundation) initiated development of the NSFNET which provides a major backbone communication service for the Internet. In Europe, major international backbones such as NORDUNET and others provide connectivity to a large number of networks. Internet slowly evolved as the universal network of networks, which connects almost every data network of the world with a reach spread across earth. It can be debated as to what the definition and scope of this global network is. On 24 October 1995, the Federal Networking Council (FNC) unanimously passed a resolution to officially define the term Internet. According to this resolution, the definition of Internet is "Internet refers to the global information system that (i) is logically linked together by a globally unique address space based on the Internet Protocol (IP) or its subsequent extensions/follow-ons; (ii) is able to support communications using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite or its subsequent extensions/follow-ons, and/or other IP-compatible protocols; and (iii) provides, uses or makes accessible, either publicly or privately, high level services layered on the communications and related infrastructure described herein."

Vannevar Bush through his July 1945 essay "As We May Think", described a theoretical machine he called a "memex", which was to enhance human memory by allowing the user to store and retrieve documents linked by associations. This can be considered as the early hypertext. During the 1960s, Doug Engelbart prototyped an 'oNLine System' (NLS) that does hypertext browsing, editing, etc. He invented the mouse for this purpose. In 1991, Tim Berners-Lee invented HTML (Hyper Text Markup Language) and HTTP (Hyper Text Transport Protocol). Tim wrote a client program and named it 'WorldWideWeb', which finally became the 'www' (World Wide Web),

almost synonymous with Internet. We would like to differentiate all these technologies by different names. We will use Web for the HTTP (WWW technology), Internet for the interworking with the network of networks, and Internet for the Internet managed by IETF (Internet Engineering Task Force).

2.3 INTERNET—THE UBIQUITOUS NETWORK

For any content to be available anywhere, we need a ubiquitous network that will carry this content. As of today, there are two networks which are ubiquitous. One is the telecommunication network and the other is the Internet network. Both these networks are in real terms the network of networks. Different networks have been connected together using a common protocol (glue). In simple terms it can be stated that SS#7 is the glue for telecommunication network whereas TCP/IP is the glue for Internet. We need one of these networks to transport content from one place to another.

We have three types of basic content: audio, video and text. Some of these content can tolerate little delays in delivery whereas some cannot. Packet switched networks like Internet are better suited for content which can tolerate little delays. Telecommunication or circuit switch networks are better suited for real-time content that cannot tolerate delays. A ubiquitous application needs to use these networks to take the content from one place to another. A network can be divided into three main segments, viz., Core, Edge and Access.

Core: As the name signifies, core is the backbone of the network. This is the innermost part of the network. The primary function of the core network is to deliver traffic efficiently at the least cost. Core looks at the traffic more from the bit stream point of view. Long-distance operators and backbone operators own core networks. This part of the network deals with transmission media and transfer points.

Edge: As the name suggests, this is at the edge of the network. These are generally managed and owned by ISPs (Internet Service Providers) or local switches and exchanges. Edge looks at the traffic more from the service point of view. It is also responsible for the distribution of traffic.

Access: This part of the network services the end point or the device by which the service will be accessed. This deals with the last mile of transmission. This part is either through a wireline or the wireless. From the mobile computing point of view, this will be mostly through the wireless.

Internet is a network of networks and is available universally. In the last few years, the popularity of web-based applications has made more and more services available through the Internet. This had a snowball effect encouraging more networks and more content to be added to the Web. Therefore, Internet is the preferred bearer network for audio, video or text content that can tolerate delays. Internet supports many protocols. However, for ubiquitous access, web-based applications are desirable. A web-based application in the Internet uses HTTP protocol and works like a request/response service. This is similar to the conventional client/server application. The fundamental difference between a web application and a conventional client/server paradigm is that in the case of conventional client/server application, the user facing the client interface contains part of the business logic. However, in the case of web applications, the client will be a thin client without any business logic. The thin client or the agent software in the client device will relate only to the

rendering functions. Such user agents will be web browsers like Mozilla, Internet Explorer or Netscape Navigator.

The types of client devices that can access the Internet are rapidly expanding. These client devices are networked either through the wireless or through a wireline. The server on the contrary, is likely to be connected to the access network through wired LAN. In addition to standard computers of different shapes and sizes, client devices can be Personal Digital Assistants (PDA) such as the PalmPilot, Sharp Zaurus, or iPaq; handheld personal computers such as the EPOC, Symbian, Psion and numerous Windows- CE machines; mobile phones with GPRS/WAP and 3G capability such as Nokia, Sony Ericsson, etc.; Internet-capable phones such as the Smartphone (cellular) and Screenphone (wired); set-top boxes such as WebTV, etc. Even the good old voice-based telephone can be used as the client device. Voice-activated Internet browsers will be very useful for visually challenged people. To fulfill the promise of universal access to the Internet, devices with very diverse capabilities need to be made available. For the wireless, devices range from the small footprint mobile phone to the large footprint laptop computers.

2.4 Architecture for Mobile Computing

In mainframe computers many mission critical systems use a Transaction Processing (TP) environment. At the core of a TP system, there is a TP monitor software. In a TP system, all the terminals—VDU (Visual Display Terminal), POS (Point of Sale Terminal), printers, etc., are terminal resources (objects). There are different processing tasks, which process different transactions or messages; these are processing resources (objects). Finally, there are database resources. A TP monitor manages terminal resources, database objects and coordinates with the user to pick up the right processing task to service business transactions. The TP monitor manages all these objects and connects them through policies and rules. A TP monitor also provides functions such as queuing, application execution, database staging, and journaling. When the world moved from large expensive centralized mainframes to economic distributed systems, technology moved towards two-tier conventional client/server architecture. With growth in cheaper computing power and penetration of Internet-based networked systems, technology is moving back to centralized server-based architecture. The TP monitor architecture is having a reincarnation in the form of three-tier software architecture.

In the early days of mainframes, the TP monitor and many other interfaces were proprietary. Even the networked interfaces to different terminals were vendor-specific and proprietary. The most successful early TP system was the reservation system for the American Airlines. This was over a Univac computer using U100 protocol. For IBM TP environment, which runs on OS/390 known as CICS (Customer Information Control System), the network interface was through SNA. In India DoT (Department of Telecommunication; currently BSNL and MTNL) launched the 197 telephone directory enquiry system in 1986, it used TPMS (Transaction Processing Management System) on ICL mainframe running VME operating system. The network interface was over X.25 interface.

The network-centric mobile computing architecture uses three-tier architecture as shown in Figure 2.1. In the three-tier architecture, the first layer is the User Interface or Presentation Tier. This layer deals with user facing device handling and rendering. This tier includes a user system interface where user services (such as session, text input, dialog and display management) reside. The second tier is the Process Management or Application Tier. This layer is for application programs or process management where business logic and rules are executed. This layer is capable of accommodating hundreds of users. In addition, the middle process management tier controls transactions and asynchronous queuing to ensure reliable completion of transactions. The third and final tier is the Database Management or Data Tier. This layer is for database access and management. The three-tier architecture is better suited for an effective networked client/server design. It provides increased *performance*, *flexibility*, *maintainability*, *reusability*, and *scalability*, while hiding the complexity of distributed processing from the user. All these characteristics have made three-tier architectures a popular choice for Internet applications and net-centric information systems. Centralized process logic makes administration and change management easier by localizing changes in a central place and using them throughout the system.

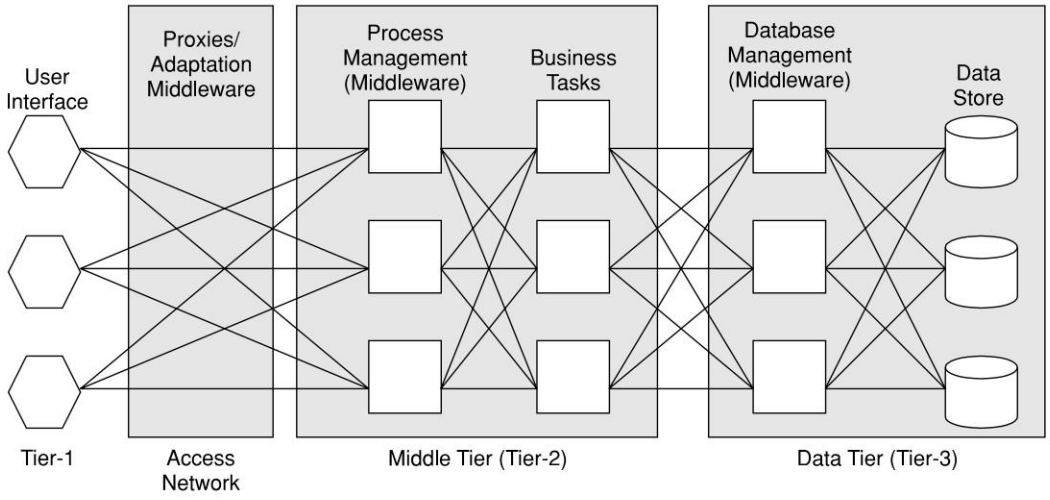


Figure 2.1 Three-tier Architecture for Mobile Computing

2.5 THREE-TIER ARCHITECTURE

To design a system for mobile computing, we need to keep in mind that the system will be used through any network, bearer, agent and device. To have universal access, it is desirable that the server is connected to a ubiquitous network like the Internet. To have access from any device, a web browser is desirable. The reason is simple; web browsers are ubiquitous, they are present in any computer. The browser agent can be Internet Explorer or Netscape Navigator or Mozilla or any other standard agent. Also, the system should preferably be context aware. We will discuss context awareness later.

We have introduced the concept of three-tier architecture. We have also discussed why it is necessary to go for Internet and three-tier architecture for mobile computing. The important question is what a mobile three-tier application actually should consist of. Figure 2.2 depicts a three-tier architecture for a mobile computing environment. These tiers are presentation tier, application tier and data tier. Depending upon the situation, these layers can be further sublayered.

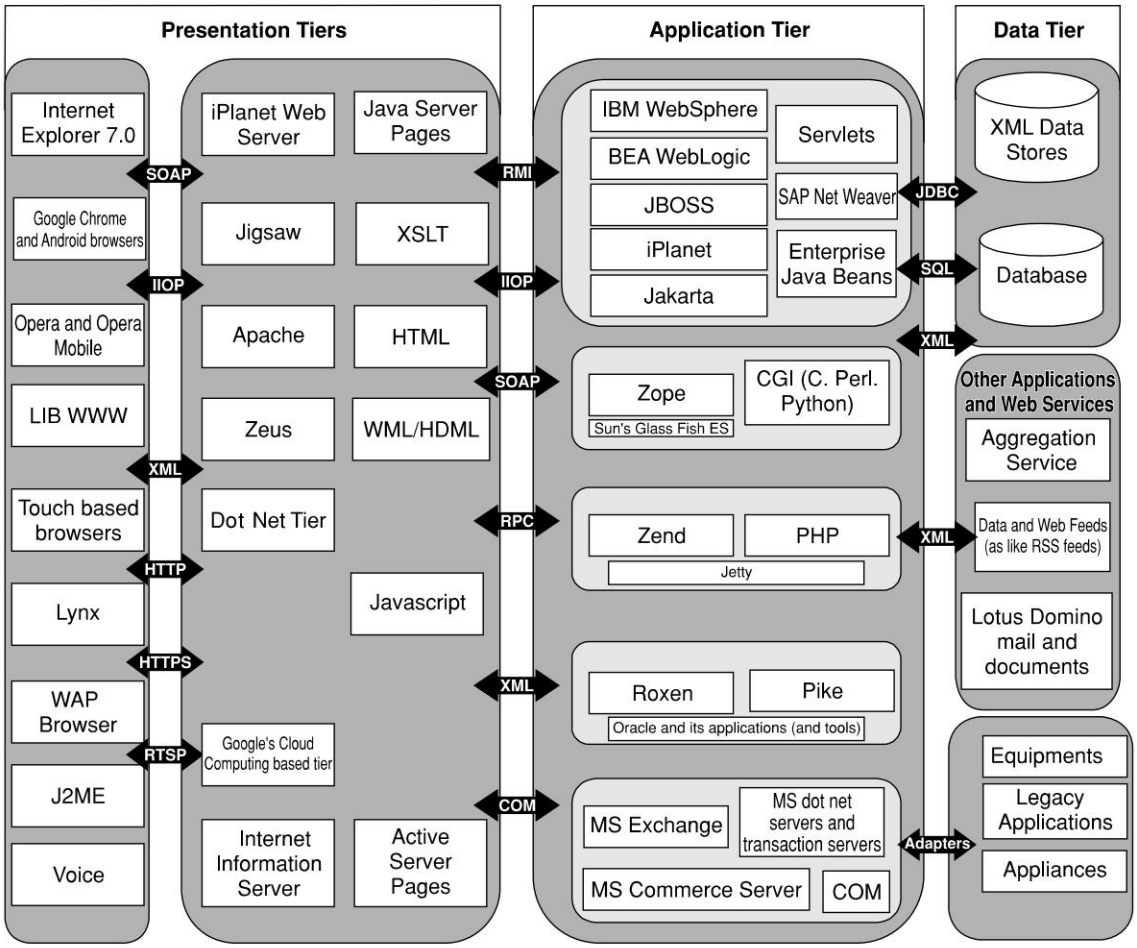


Figure 2.2 The Mobile Computing Architecture

2.5.1 Presentation (Tier-1)

This is the user facing system in the first tier. This is the layer of agent applications and systems. These applications run on the client device and offer all the user interfaces. This tier is responsible for presenting the information to the end user. Humans generally use visual and audio means to

receive information from machines (with some exceptions like vibrator in mobile phones). Humans also use keyboard (laptop computers, cell phones), pen (tablet PC, palmtops), touch screen (kiosks), or Voice (telephone) to feed the data to the system. In the case of the visual, the presentation of information will be through a screen. Therefore, the visual presentation will relate to rendering on a screen. 'Presentation Tier' includes web browsers (like Mozilla, Lynx, Internet Explorer and Netscape Navigator), WAP browsers and customized client programs. A mobile computing agent needs to be context-aware and device independent.

In general, the agent software in the client device is an Internet browser. In some cases, the agent software is an applet running on a browser or a virtual machine (Java Virtual Machine, for example). The functions performed by these agent systems can range from relatively simple tasks like accessing some other application through HTTP API, to sophisticated applications like real-time sales and inventory management across multiple vendors. Some of these agents work as web scrapers. In a web scraper, the agent embeds functionality of the HTTP browser and functions like an automated web browser. The scraper picks up part of the data from the web page and filters off the remaining data according to some predefined template. These applications can be in Business to Business (B2B) space, Business to Consumer (B2C) space or Business to Employee (B2E) space, or machine to machine (M2M) space. Applications can range from e-commerce, workflow, supply chain management to legacy applications.

There are agent software in the Internet that access the remote service through telnet interface. There are different flavors of telnet agents in use. These are standard telnet for UNIX servers; TN3270 for IBM OS/390; TN5250 for IBM AS/400 or VT3K for HP3000. For some applications, we may need an agent with embedded telnet protocol. This will work like an automated telnet agent (virtual terminal) similar to a web scraper. These types of user agents or programs work as M2M interface or software robots. These kinds of agents are used quite frequently to make legacy applications mobile. Also, such systems are used in the telecommunication world as mediation servers within the OSS (Operation and Support Subsystem).

2.5.2 Application (Tier-2)

The application tier or middle tier is the "engine" of a ubiquitous application. It performs the business logic of processing user input, obtaining data, and making decisions. In certain cases, this layer will do the transcoding of data for appropriate rendering in the Presentation Tier. The Application Tier may include technology like CGIs, Java, JSP, .NET services, PHP or ColdFusion, deployed in products like Apache, WebSphere, WebLogic, iPlanet, Pramati, JBOSS or ZEND. The application tier is presentation and database-independent.

In a mobile computing environment, in addition to the business logic there are quite a few additional management functions that need to be performed. These functions relate to decisions on rendering, network management, security, datastore access, etc. Most of these functions are implemented using different middleware software. A middleware framework is defined as a layer of software, which sits in the middle between the operating system and the user facing software. Stimulated by the growth of network-based applications and systems, middleware technologies are gaining increasing importance in net-centric computing. In case of net-centric architecture, a middleware framework sits between an agent and business logic. Middleware covers a wide range of software systems, including distributed objects and components, message-oriented

communication, database connectors, mobile application support, transaction drivers, etc. Middleware can also be considered as a software gateway connecting two independent open objects.

It is very difficult to define how many types of middleware are there. A very good description of middleware is available in Carnegie Mellon University Software Engineering Institute (http://www.sei.cmu.edu/str/descriptions/middleware.html), which readers can refer to.

We can group middleware into the following major categories:

1. Message-oriented Middleware.
2. Transaction Processing Middleware.
3. Database Middleware.
4. Communication Middleware.
5. Distributed Object and Components.
6. Transcoding Middleware.

Message-oriented Middleware (MOM)

Message-oriented Middleware is a middleware framework that loosely connects different applications through asynchronous exchange of messages. A MOM works over a networked environment without having to know what platform or processor the other application is resident on. The message can contain formatted data, requests for action, or unsolicited response. The MOM system provides a message queue between any two interoperating applications. If the destination process is out of service or busy, the message is held in a temporary storage location until it can be processed. MOM is generally asynchronous, peer-to-peer, and works in publish/subscribe fashion. In the publish/subscriber mode one or many objects subscribe to an event. As the event occurs, it will be published by the loosely coupled asynchronous object. The MOM will notify the subscribers about this event. However, most implementations of MOM support synchronous (request/response) message passing as well. MOM is most appropriate for event-driven applications. When an event occurs, the publisher application hands on to the messaging middleware application the responsibility of notifying subscribers that the event has happened. In a net-centric environment, MOM can work as the integration platform for different applications. An example of MOM is Message Queue from IBM known as MQ Series. The equivalent from Java is JMS (Java Message Service).

Transaction Processing (TP) Middleware

Transaction Processing Middleware provides tools and an environment for developing transaction-based distributed applications. An ideal TP system will be able to input data into the system at the point of information source and the output of the system is delivered at the point of information sink. In an ideal TP system, the device for input and output can potentially be different (Fig. 2.3). Also, the output can be an unsolicited message for a device. TP is used in data management, network access, security systems, delivery order processing, airline reservations, customer service, etc., to name a few. TP systems are generally capable of providing services to thousands of clients in a distributed client/server environment. CICS (Customer Information Control System) is one of the early TP application systems on IBM mainframe computers.

TP middleware maps numerous client requests through application-service routines to different application tasks. In addition to these processing tasks, TP middleware includes numerous management features, such as restarting failed processes, dynamic load balancing and ensuring

consistency of distributed data. TP middleware is independent of the database architecture. TP middleware optimizes the use of resources by multiplexing many client functions on to a much smaller set of application-service routines. This also helps in reducing the response time. TP middleware provides a highly active system that includes services for delivery-order processing, terminal and forms management, data management, network access, authorization, and security. In the Java world and net-centric systems, transaction processing is done through the J2EE application server with the help of entity and session beans.

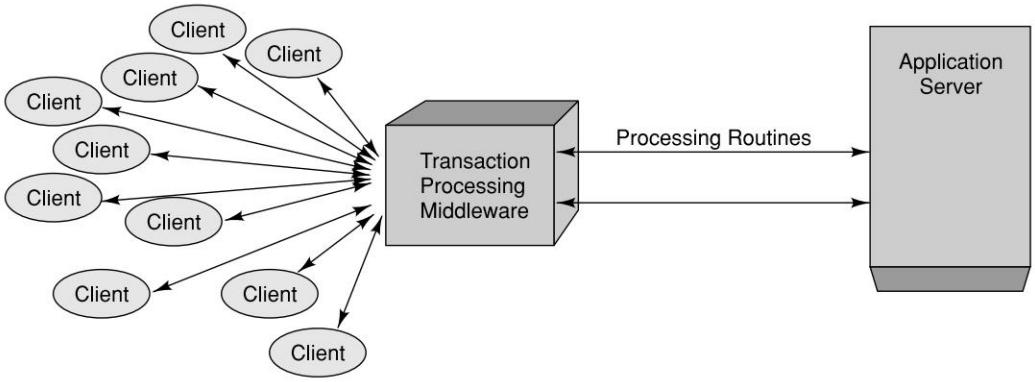


Figure 2.3 Transaction Processing Middleware

Model View Controller (MVC): Java uses the MVC architectural pattern which is an example of transaction processing system. It splits an application into separate layers, viz., presentation, domain logic, and data access. *Model* is the domain-specific representation of the information on which the application operates. Domain logic manipulates and adds meaning to the raw data. MVC does not specifically mention the data access layer because it is assumed to be encapsulated by the model. *View* is responsible for rendering the model into a form suitable for interaction and understood by the user, typically a user interface element. *Controller* manages processes and responds to events, typically user actions, and may invoke changes on the model. In the context of Web applications and J2EE, the MVC pattern is widely used. In Web applications, where the view is the actual HTML page, and the controller is the code which gathers dynamic data and generates the content within the HTML, the model is represented by the actual content, usually stored in a database.

Communication Middleware

Communication Middleware is used to connect one application to another through some communication middleware, like connecting one application to another through telnet. These types of middleware are quite useful in the telecommunication world. There are many elements in the core telecommunication network where the user interface is through telnet. A mediation server automates the telnet protocol to communicate with these nodes in the network. Another example could be to integrate legacy applications through proprietary communication protocols like TN5250 or TN3270.

Distributed Object and Components

An example of distributed objects and components is CORBA (Common Object Request Broker Architecture). CORBA is an open distributed object computing infrastructure being standardized by the Object Management Group (http://www.omg.org). CORBA simplifies many common network programming tasks used in a net-centric application environment. These are object registration, object location, and activation; request demultiplexing; framing and error-handling; parameter marshalling and demarshalling; and operation dispatching. CORBA is vendor-independent infrastructure. A CORBA-based program from any vendor on almost any computer, operating system, programming language and network, can interoperate with a CORBA-based program from the same or another vendor, on almost any other computer, operating system, programming language and network. CORBA is useful in many situations because of the easy way that CORBA integrates machines from so many vendors, with sizes ranging from mainframes through minis and desktops to hand-helds and embedded systems. One of its most important, as well as the most frequent uses is in servers that must handle a large number of clients, at high hit rates, with high reliability.

Transcoding Middleware

Transcoding Middleware is used to transcode one format of data to another to suit the need of the client. For example, if we want to access a web site through a mobile phone supporting WAP, we need to transcode the HTML page to WML page so that the mobile phone can access it. Another example could be accessing a map from a PDA. The same map, which can be shown in a computer, needs to be reduced in size to fit the PDA screen. Technically transcoding is used for content adaptation to fit the need of the device. Content adaptation is also required to meet the network bandwidth needs. For example, some frames in a video clip need to be dropped for a low bandwidth network. Content adaptation used to be done through proprietary protocols. To allow interoperability, IETF has accepted the Internet Content Adaptation Protocol (ICAP). ICAP is now standardized and described in RFC3507.

Internet Content Adaptation Protocol (ICAP)

Popular web servers are required to deliver content to millions of users connected at ever-increasing bandwidths. Progressively, content is being accessed through different devices and agents. A majority of these services have been designed keeping the desktop user in mind. Some of them are also available for other types of protocols. For example, there are a few sites that offer contents in HTML and WML to service desktop and WAP phones. However, the model of centralized services that are responsible for all aspects of every client's request seems to be reaching the end of its useful life. ICAP, the Internet Content Adaptation Protocol, is a protocol aimed at providing simple object-based content vectoring for HTTP services. ICAP is a lightweight protocol to do transcoding on HTTP messages. This is similar to executing a "remote procedure call" on a HTTP request. The protocol allows ICAP clients to pass HTTP messages to ICAP servers for some sort of transformation. The server executes its transformation service on messages and sends back responses to the client, usually with modified messages. The adapted messages may be either HTTP requests or HTTP responses. For example, before a document is displayed in the agent, it is checked for virus.

There are two major components in ICAP architecture:

1. What are the semantics for the transformation? How do I ask for content adaptation?

2. How is policy of the transformation managed? What kind of adaptation do I ask for and from where? How do I define and manage the adaptation?

ICAP works at the edge part of the network as depicted in Figure 2.4. It is difficult, if not impossible, to define the devices users may like to use to access content from within the Internet. Customized edge delivery of Internet content will help to improve user experience. When applications are delivered from an edge device, end users find that the applications execute more quickly and are more reliable. Typical data flow in an ICAP environment is depicted in Figure 2.4 and described here.

1. A user agent makes a request to an ICAP-capable surrogate (ICAP client) for an object on an origin server.
2. The surrogate sends the request to the ICAP server.
3. The ICAP server executes the ICAP resource's service on the request and sends the possibly modified request, or a response to the request back to the ICAP client.
4. The surrogate sends the request, possibly different from the original client's request, to the origin server.
5. The origin server responds to the request.
6. The surrogate sends the reply (from either the ICAP or the origin server) to the client.

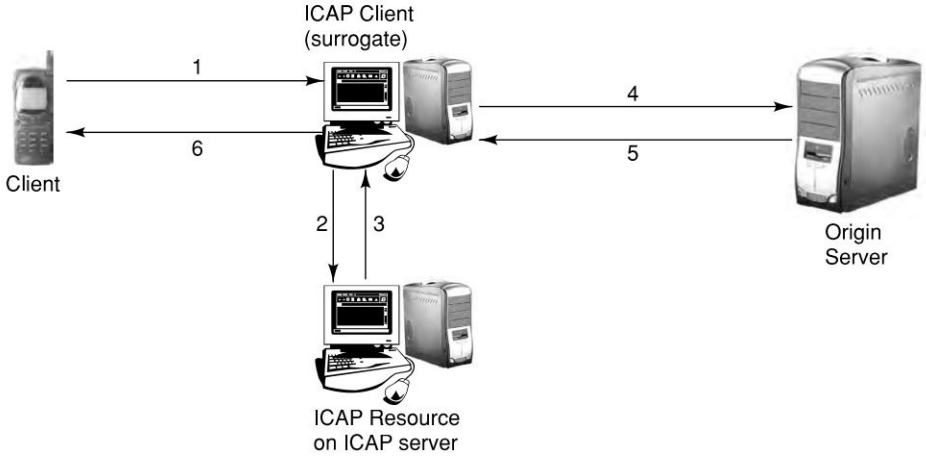


Figure 2.4 Typical Data Flow in an ICAP Environment

It is envisioned that in future, ICAP servers may be available to provide some of the following services:

- Suit content delivery based on network bandwidth.
- Suit content delivery based on device characteristics.
- Language translation based on the user's preference.
- Virus checking for the requested content.
- Content filtering based on sensor rating like PG (parental guidance), R (restricted).
- Local real-time advertisement insertion like television.
- Local real-time advertisement elimination for premium subscribers.

- Wireless protocol translation.
- Anonymous Web usage profiling for a dating service.
- Transcoding or image enhancement.
- Image magnification for the elderly.
- Image size reduction based on device display characteristics.
- Intelligent video condensation by dropping frames.
- Digest production/batch download of Web content.
- Content filtering based on copyright or digital signature.
- Peer-to-Peer compression and encryption of data.

Web Services

As the need for peer-to-peer, application-to-application communication and interoperability grows, the use of Web services on the Internet will also grow. Web services provide a standard means of communication and information exchange among different software applications, running on a variety of platforms or frameworks**.** Web service is a software system identified by a URI, whose public interfaces and bindings are defined using XML (eXtensible Markup Language). Its definition can be discovered by other software systems connected to the network. Using XML-based messages these systems may then interact with the Web service in a manner prescribed by its definition.

The basic architecture includes Web service technologies capable of:

- Exchanging messages.
- Describing Web services.
- Publishing and discovering Web service descriptions.

The Web services architecture defines the standards for exchange of messages between the service requester and service provider. Service providers are responsible for publishing a description of the services they provide. Requesters must be able to find and discover descriptions of the services.

Software agents in the basic architecture can take on one or all of the following roles:

- Service requester–requests the execution of a Web service.
- Service provider–processes a Web service request.
- Discovery agency–agency through which a Web service description is published and made discoverable.

The interactions involve the publish, find and bind operations. A service is invoked after the description is found, since the service description is required to establish a binding.

2.5.3 Data (Tier-3)

The Data Tier is used to store data needed by the application and acts as a repository for both temporary and permanent data. The data can be stored in any form of datastore or database. These can range from sophisticated relational database, legacy hierarchical database, to even simple text files. The data can also be stored in XML format for interoperability with other systems and datasources. A legacy application can also be considered as a data source or a document through a communication middleware.

Database Middleware

We have discussed that for a mobile computing environment, the business logic should be independent of the device capability. Likewise, though not essential, it is advised that business logic should be independent of the database. Database independence helps in maintenance of the system better. Database middleware allows the business logic to be independent and transparent of the database technology and the database vendor. Database middleware runs between the application program and the database. These are sometimes called database connectors as well. Examples of such middleware will be ODBC, JDBC, etc. Using these middleware, the application will be able to access data from any data source. Data sources can be text files, flat files, spreadsheets, or a network, relational, indexed, hierarchical, XML database, object database, etc., from vendors like Oracle, SQL, Sybase, etc.

SyncML

SyncML protocol is an emerging standard for synchronization of data access from different nodes. When we moved from the conventional client/server model of computing to the net-centric model of computing, we moved from distributed computing to centralized computing with networked access. The greatest benefit of this model is that resources are managed at a centralized level. All the popular mobile devices like handheld computers, mobile phones, pagers and laptops work in an occasionally connected computing mode and access these centralized resources from time to time. In an occasionally connected mode, some data are cached in the local device and accessed frequently. The ability to access and update information on the fly is key to the pervasive nature of mobile computing. Examples are emails and personal information like appointments, address book, calendar, diary, etc. Storing and accessing phone numbers of people from the phone address book is more user-friendly compared to accessing the same from a server. However, managing the appointments database is easier in a server, though caching the same on the mobile client is critical. Users will cache emails into the device for reference. We take notes or draft a mail in the mobile device. For workflow applications, data synchronization plays a significant role. The data in the mobile device and server need to be synchronized. Today vendors use proprietary technology for performing data synchronization. SyncML protocol is the emerging standard for synchronization of data across different nodes. SyncML is a new industry initiative to develop and promote a single, common data synchronization protocol that can be used industry-wide.

The ability to use applications and information on a mobile device, then to synchronize any updates with the applications and information back at the office or on the network, is key to the utility and popularity of mobile computing. The SyncML protocol supports naming and identification of records and common protocol commands to synchronize local and network data. It supports identification and resolution of synchronization conflicts. The protocol works over all networks used by mobile devices, both wireless and wireline. Since wireless networks employ different transport protocols and media, a SyncML will work smoothly and efficiently over:

- HTTP 1.1 (i.e., the Internet).
- WSP (the Wireless Session Protocol, part of the WAP protocol suite).
- OBEX (Object Exchange Protocol, i.e., Bluetooth, IrDA and other local connectivity).
- SMTP, POP3 and IMAP.
- Pure TCP/IP networks.
- Proprietary wireless communication protocols.

2.6 Design Considerations for Mobile Computing

The mobile computing environment needs to be context-independent as well as context-sensitive. Context information is the information related to the surrounding environment of an actor in that environment. The term "context" means, all the information that helps determine the state of an object (or actor). This object can be a person, a device, a place, a physical or computational object, the surrounding environment or any other entity being tracked by the system. In a mobile computing environment, context data is captured so that decisions can be made about how to adapt content or behavior to suit this context. Mobility implies that attributes associated with devices and users will change constantly. These changes mean that content and behavior of applications should be adapted to suit the current situation. There are many ways in which content and behavior can be adapted. Following are some examples:

1. **Content with context awareness:** Build each application with context awareness. There are different services for different client context (devices). For example, a bank decides to offer mobile banking application through Internet, PDA and mobile phone using WAP. These services are different and are http://www.mybank.com/inet.html, http://www.mybank.com/palm.html and http://www.mybank.com/wap.wml, respectively. The service http://www.mybank.com/inet.html assumes that the user will use computers to access this service. Therefore it is safe to offer big pages with text boxes and drop down menus. Also, it is fine to add a few animated pictures for the new product the bank is launching. We know that http://www.mybank.com/palm.html is a service for a PalmOS PDA. As the display size is small, we design the screen to be compact for the PDA and do not offer the same product animation. For the WAP service at http://www.mybank.com/wap.wml, we do a completely different user interface; we make all drop down options available through the option button in the mobile phone and remove all the graphics and animations.
2. **Content switch on context:** Another way is to provide intelligence for the adaptation of content within the service. This adaptation happens transparent to the client. In this case the service is the same for Internet, PDA and WAP. All access the bank's service through http://www.mybank.com/. An intelligent piece of code identifies the agent to decide what type of device or context it is. This intelligent code does the adaptation at runtime based upon the agent in hand. The simplest way to do this is to look at the user-agent value at the HTTP header and decide whether to route the request to http://mybank.com/inet.html or http://www.mybank.com/palm.html or http://www.mybank.com/wap.wml.
3. **Content transcoding on context:** Another way is to provide an underlying middleware platform that performs the adaptation of the content based on the context and behavior of the device. This adaptation happens transparent to the client and the application. The middleware platform is intelligent enough to identify the context either from the HTTP or additional customized parameters. In this case the service may be in html or XML, the middleware platform transcodes the code from html (or XML) to html, and wml on the fly. It can also do the transcoding based on policy so that the html generated for a computer is different from a PDA.

Following sections describe different types of context that can enhance the usability, reliability and security of the service. Figure 2.5 depicts the old web and web of the future for mobile computing.

2.6.1 Client Context Manager

When we humans interact with other persons, we always make use of the implicit situational information of the surrounding environment. We interpret the context of the current situation and react appropriately. For example, we can go close to a lion in a zoo, but definitely not in the wild. Or, a person discussing some confidential matter with another person observes the gestures and tone of the other person and reacts in an appropriate manner or changes the subject if someone shows up suddenly. When we use content through a PC within the four walls of an organization, we do not have any problem. A majority of the applications can safely assume that the context is the enterprise LAN. It can be assumed that the environment is secured; it can also be assumed that the user will be using the systems in a particular fashion using the browser standardized by the company. These applications are developed keeping the large screen (for mainly PC) and browsers in mind. A mobile computing application, on the other hand, needs to operate in dynamic conditions. This is due to various device characteristics and network conditions. This demands a reactive platform that can make decisions about how to respond to changes to device capability, user preferences, enterprise policy, network policy and many other environmental factors. Context can be used as the basis by which an adaptation manager or algorithm decides to modify content or application behavior. We therefore need a Client Context Manager to gather and maintain information pertaining to the client device, user, network and the environment surrounding each mobile device. All these information will be provided by a set of Awareness Modules. Awareness

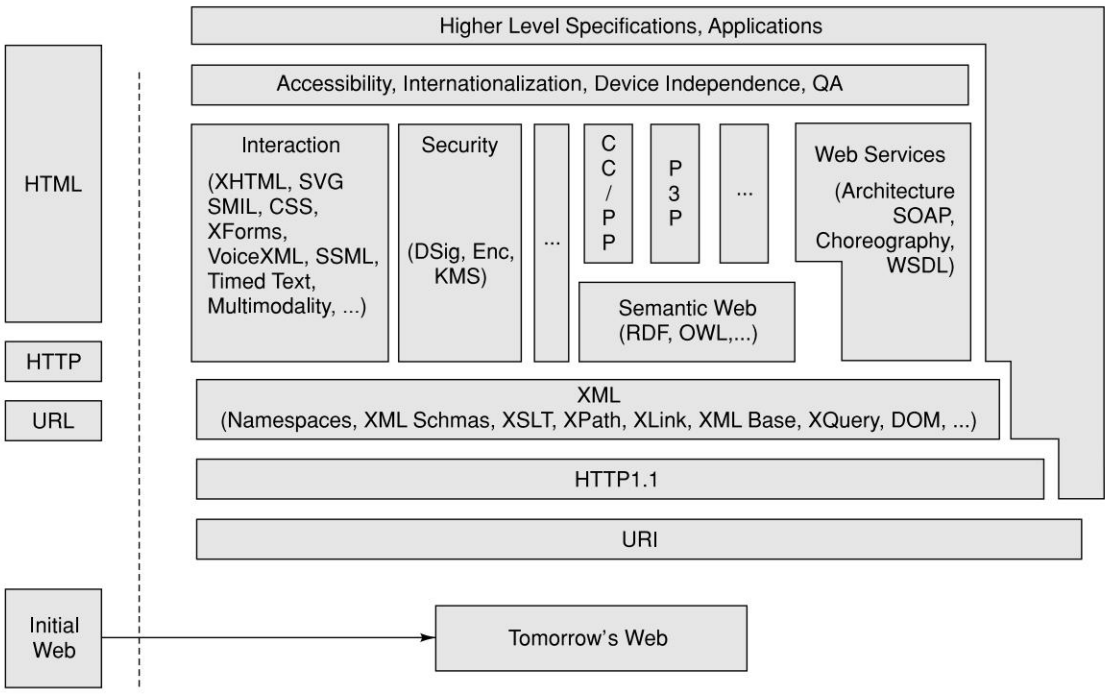


Figure 2.5 The Content Architecture with Respect to Mobile Computing

modules are sensors of various kinds. These sensors can be hardware sensors or software sensors or a combination of these. A hardware sensor can be used to identify the precise location of a user; whereas, a software sensor can be used to determine the type of the user agent. These awareness modules can be in the device, network, or even in the middleware. We use the term middleware in a very generic context. A middleware can be a functional module in the content server, a proxy or an independent system. For example, an awareness module in the device will provide information about its capabilities. Another example could be a location manager that tracks the location and orientation of the mobile device.

Almost any information available at the time of an interaction can be seen as context information. Some examples are:

1. **Identity:** The device will be in a position to communicate its identity without any ambiguity.
2. **Spatial information:** Information related to the surrounding space. This relates to location, orientation, speed, elevation and acceleration.
3. **Temporal information:** Information related to time. This will be time of the day, date, time zone and season of the year.
4. **Environmental information:** This is related to the environmental surroundings. This will include temperature, air quality, moisture, wind speed, natural light or noise level. This also includes information related to the network and network capabilities.
5. **Social situation:** Information related to the social environment. This will include who you are with, and people that are nearby; whether the user is in a meeting or in a party.
6. **Resources that are nearby:** This will relate to the other accessible resources in the nearby surroundings like accessible devices, hosts or other information sinks.
7. **Availability of resources:** This will relate to information about the device in use. This will include battery power, processing power, persistence store, display, capabilities related to I/O (input/output) and bandwidth.
8. **Physiological measurements:** This relates to the physiological state of the user. This includes information like blood pressure, heart rate, respiration rate, muscle activity and tone of voice.
9. **Activity:** This relates to the activity state of the user. This includes information like talking, reading, walking and running.
10. **Schedules and agendas:** This relates to the schedules and agendas of the user.

A system is context-aware if it can extract, interpret and use context-related information to adapt its functionality to the current context. The challenge for such systems lies in the complexity of capturing, representing, filtering and interpreting contextual data. To capture context information generally some sensors are required. This context information needs to be represented in a machine-understandable format, so that applications can use this information. In addition to being able to obtain the context-information, applications must include some 'intelligence' to process the information and deduce the meaning. These requirements lead us to three aspects of context management:

1. **Context sensing:** The way in which context data is obtained.
2. **Context representation:** The way in which context information is stored and transported.
3. **Context interpretation:** The way in which meaning is obtained from the context representation.

W3C has proposed a standard for context information. This standard is called Composite Capabilities/Preference Profiles (CC/PP), for describing device capabilities and user preferences. All these context information are collated and made available to the management components.

Composite Capabilities/Preference Profiles (CC/PP)

Composite Capabilities/Preference Profiles (CC/PP) is a proposed W3C standard for describing device capabilities and user preferences. Special attention has been paid to wireless devices such as mobile phones and PDAs. In practice, the CC/PP model is based on RDF (resource description framework) and can be serialized using XML.

A CC/PP profile contains a number of attribute names and associated values that are used by an application to determine the appropriate form of a resource to deliver to a client. This is to help a client or proxy/middleware to describe their capabilities to an origin server or other sender of resource data. It is anticipated that different applications will use different vocabularies to specify application-specific properties within the scope of CC/PP. However, for different applications to interoperate, some common vocabulary is needed. The CC/PP standard defines all these. Following is an example of a device RDF in CC/PP terminology.

```
<?xml version="1.0"?>
<!— Checked by SiRPAC 1.16, 18-Jan-2001 —>
<rdf:RDF  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
             xmlns:ccpp="http://www.w3.org/2000/07/04-ccpp#">
   <rdf:Description  rdf:about="MyProfile">
      <ccpp:component>
         <rdf:Description  rdf:about="TerminalHardware">
            <rdf:type  rdf:resource="HardwarePlatform" />
            <display>320x200</display>
         </rdf:Description>
      </ccpp:component>
      <ccpp:component>
         <rdf:Description  rdf:about="TerminalSoftware">
            <rdf:type  rdf:resource="SoftwarePlatform" />
            <name>EPOC</name>
            <version>2.0</version>
            <vendor>Symbian</vendor>
         </rdf:Description>
      </ccpp:component>
      <ccpp:component>
         <rdf:Description  rdf:about="TerminalBrowser">
            <rdf:type  rdf:resource="BrowserUA" />
            <name>Mozilla</name>
            <version>5.0</version>
            <vendor>Symbian</vendor>
            <htmlVersionsSupported>
               <rdf:Bag>
                  <rdf:li>3.0</rdf:li>
                  <rdf:li>4.0</rdf:li>
               </rdf:Bag>
      </htmlVersionsSupported>
            </rdf:Description>
```

```
        </ccpp:component>
    </rdf:Description>
</rdf:RDF>
```

CC/PP is designed in such a way that an origin server or proxy can perform some sort of content to device matching. CC/PP is designed to suit an adaptation algorithm. The sequence of steps in the general case would look something like the following (Fig. 2.6):

1. Device sends serialized profile model with request for content.
2. Origin server receives serialized RDF profile and converts it into an in-memory model.
3. The profile for the requested document is retrieved and an in-memory model is created.
4. The device profile model is matched against the document profile model.
5. A suitable representation of the document is chosen. At this stage the document to be returned can be chosen from a number of different versions of the same document (content switch on context) or it can be dynamically generated (content transcoding on context).
6. Document is returned to device and presented.

If a document or application is specific about how it should be displayed, or if there are several versions of the document or application for different devices, then the adaptation manager can ask the client context manager for detailed context information. The client context manager will enquire with the relevant awareness module and extract the necessary context information. This fine-grained approach allows a high level of adaptation to take place. In cases where the document does not provide profile information, or the profile is limited in description, the adaptation manager can obtain a general context class from the context manager and perform some limited adaptation. For example, some adaptation can still take place where the location of the user is important. The policy manager can specify some rules about how adaptation should take place when a user is at a certain location, regardless of the information provided in an application or document profile.

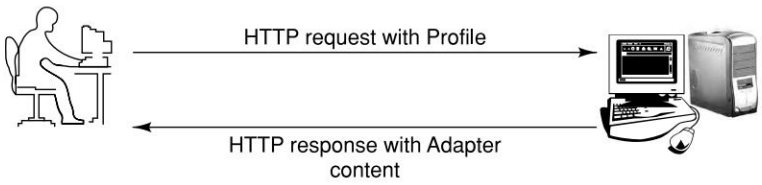


Figure 2.6 The Simplest Use of CC/PP

Policy Manager

The policy manager is responsible for controlling policies related to mobility. A policy is a set of rules; these rules need to be followed under different conditions. Introduction of mobility within an enterprise brings with it different types of challenges that are not normally seen in traditional computing environments. When we consider mobility, it is assumed that the data or information will be visible from outside the four walls of the enterprise. Organizations generally have policies regarding the disclosure of information. For example, documents from certain systems can be

printed only on certain printers in the organization. Some hard copy documents may be viewed only at the office of the CEO. These kinds of policies must be transferable to a mobile computing environment. Mobile computing policy manager will be able to define policy for documents/services and assign roles to users. Each role will have permissions, prohibitions and obligations associated with it. Each policy will have access rights associated with respect to read, write, execute. A policy in combination with role and current context information will be able to determine what actions a user is allowed to perform, or what actions a user is obligated to perform.

Semantic Web

As mentioned earlier, policies are sets of rules. When we drive in the street we are expected to follow the right of way. In a party there are some etiquettes to be followed. We humans learn these rules, policies, laws, and etiquettes from documents or experienced people. This is to help us to behave correctly in the society. The question is how to make a machine understand policies and make them behave in the expected fashion? Data in the Web is generally hidden away in HTML files, how do we determine which content is useful in some contexts, but often not in others. Facilities to put machine understandable data on the Web are becoming a necessity. The Semantic Web is targeted to address this need. The idea is of having data on the Web defined and linked in a way that it can be used by machines not just for display, but for automation, security, filtering, integration and reuse of data across various applications.

Semantic Web technologies are still very much in their infancy. It is believed that a large number of Semantic Web applications can be used for a variety of different tasks, increasing the modularity of applications on the Web. The Semantic Web is generally built on syntaxes which use URIs to represent data, usually in tuple-based structures, i.e., many tuples of URI data that can be held in databases, or interchanged on the World Wide Web using a set of particular syntaxes developed especially for the task. These syntaxes are called RDF (Resource Description Framework) syntaxes.

Security Manager

The Security Manager provides a secure connection between the client device and the origin server. Depending on the security policies of an organization, if the security requirements are not met or some content is not be viewable the security manager will ensure security with respect to:

- *Confidentiality:* The message being transacted needs to be confidential. Nobody will be able to see it.
- *Integrity:* The message being transacted needs to be tamper-resistant. Nobody will be able to change any part of the message.
- *Availability:* The system will be available. Nobody will be able to stop the service.
- *Non-repudiation:* Users of the system can be identified. Nobody after using the system can claim otherwise.
- *Trust:* There are complex issues of knowing what resources, services or agents to trust. The system will be trusted.

Confidentiality is managed by encryption. Using encryption techniques we change the message to some other message so that it cannot be understood. There are different types of encryption algorithms and standards. In a defined environment like enterprise LAN or a VPN (Virtual Private Network), we can standardize some encryption algorithm like 128 bits AES to be used. However,

in a ubiquitous environment, the environment is unpredictable with ad-hoc groups of devices. Also, the networks and their security level cannot be guaranteed all the time. Integrity can be managed using different hashing algorithms. Availability relates to peripheral security related to Web server, firewall, etc. Non-repudiation can be managed with digital signatures. For trust we may need to establish some sort of third-party recommendation system. Third-party rating system can also help establish trust. The security manager needs to manage all these aspects.

Platform for Privacy Preference Project (P3P)

The Platform for Privacy Preference Project (P3P) is an emerging standard defined by W3C. P3P enables web sites to express their privacy practices in a standardized format so that they can be retrieved and interpreted by user agents. With P3P, users need not read the privacy policies they visit; instead, key information about the content of the web site can be conveyed to the user. Any discrepancies between a site's practices and the user's preferences can be flagged as well. The goal of P3P is to increase user trust and confidence in the Web.

P3P provides a technical mechanism to inform users about privacy policies about the site. This will help users to decide whether to release personal information or not. However, P3P does not provide any mechanism for ensuring that sites act according to their policies. P3P is intended to be complementary to both legislative and self-regulatory programs that can help enforce web site policies.

Adaptability Manager

The Adaptability Manager is responsible for adapting content, behavior and other aspects according to context and policy. The adaptability manager may take any number of actions depending on the information passed to it by the context manager. This information may or may not be in the form of RDF. The most obvious action to perform is to transcode content so that it may be viewed on a particular device. Other actions might include appending location-specific information to documents.

Content Adaptation and Transcoding

In a ubiquitous situation, services are used from any device through any network. Therefore, the content should be able to adapt to these dynamic situations. The adaptation may be static or dynamic.

Content adaptation can be performed either at the content level at the server end or at the agent level in the client device. Content adaptation can be done at an intermediate level in a middleware framework as well. To do a good job of content adaptation, we need to go beyond the header. We need to consider the requirements of the entire Web page or relationships between its various components in different media. It also needs to look at adaptation within the scope of the same and a different modality. Modes can be audio, video, voice, image or text. We are differentiating between audio and voice by the characteristics that audio is a sound clip as an object like the audio part of a multimedia lecture, whereas voice is real-time and synthesized from some other form or representation. Content adaptation needs to consider the following attributes.

1. **Physical capabilities of the device:** Screen size, i.e., width and height in pixels, color and bits/pixel.
2. **Logical capabilities of the device:** Required for displaying video, image and playing audio.

3. **Effective network bandwidth**.
4. **Payload:** The total amounts of bits that can be delivered to the agent for the static parts. For streaming media this will be the initial buffer space required before the media starts playing. For storage constrained devices, the payload will be defined as the storage space.

Transcoding can be classified as the following:

- *Spatial transcoding* is transcoding in space or dimension. In this transcoding technique a standard frame is downscaled and reduced. The frame is changed from one size to a different size to suit the target device.
- *Temporal transcoding* copes with a reduction of number of frames in the time scale. This technique downscales the number of transferred frames to suit the target device and network bandwidth.
- *Color transcoding* is sometimes requested for monochrome clients. Using less bits for pixel can reduce bandwidth and sometimes modify the perception of images.
- *Code transcoding* is used to change coding from one standard to another. One such example could be compression of the data or transcode a BMP file to WBMP for wireless device.
- *Object or semantic transcoding* comprises some different techniques based on computer vision techniques. The goal is to extract semantically valuable objects from the scene and transfer them with the lower amount of compression in order to maintain both details and speed.

Server side content adaptation can be achieved through the concept of InfoPyramid. InfoPyramid creates context-aware content through static transcoding. The transcoding is done off-line at the content creation time. InfoPyramid is used to store multiple resolutions and modalities of the transcoded content, along with any associated meta-data. For server side adaptation, each atomic item of the document is analysed to determine its resource requirements. The types of resources considered are those that may differentiate different client devices. The resource requirement is determined by the following attributes.

1. Static content size in bits.
2. Display size such as height, width and area.
3. Streaming bit-rate.
4. Color requirements.
5. Compression formats.
6. Hardware requirements, such as display for images, support for audio and video.

This is very useful for enterprises whose users are likely to use the service from different networks and devices. For example, a bank or a courier company which has its customer base across the world and is likely to use the service from any device from any network. When the Web server receives a user request, it determines the capabilities of the requesting client device. A customization module (context-sensitive content switch) dynamically selects the page from the InfoPyramids. The selection is based on the resolutions or modalities that best meet the client capabilities. This selected content is then rendered in a suitable delivery format for delivery to the client. This type of transcoding is most suitable for enterprises where the content type is known.

In case of client-side adaptation, the adaptation is done by the agent application. The agent application does the adaptation based on its capabilities. For example, let us assume that the client device does not support color: therefore, a color image received by the agent will be displayed as a black and white image. Client-side adaptation can be quite effective for static images. However, it may not be very effective for streaming payload delivery.

The other technique of transcoding is through a middleware. One big benefit of the middleware approach is that it is totally transparent to the device and the content. Content providers do not have to change the way they author or serve content. However, there are a number of drawbacks to this approach:

1. Content providers have no control over how their content will appear to different clients.
2. There may be legal issues arising from copyright that may preclude or severely limit the transcoding by proxies.
3. HTML tags mainly provide formatting information rather than semantic information.
4. Transcoding sometimes could be difficult to apply to many media types such as video and audio.
5. Developing a general purpose transcoding engine is very difficult if not impossible.

Transcoding through middleware is transparent to both device and content. Therefore, this transcoding technique has to be very robust and universal. That is why this transcoding technique is the most difficult to engineer. It is most desirable for content aggregators and value-added service providers.

Content Rating and Filtering

Any city in the world has regions well marked like business district, residential area, shopping complex, so on and so forth. In Bangalore, for example, Commercial Street, Koramangala, and Shivaji Market signify commercial/shopping area, residential area and market place respectively. By looking at the name of a web site or the document header, can we make some judgement about the content? This is necessary for content filtering and personalization. If we want to make sure that children at home are not accessing some restricted material, how do we do this? In a bookstore, adult magazines are displayed on the topmost shelf so that children cannot reach them. Children below 18 are not allowed to buy cigarettes or alcohol from a shop. In Internet, everything is freely accessible. How do we enforce such social discipline in the electronic world?

W3C has proposed a standard called PICS (Platform for Internet Content Selection) for rating of web content. Filtering of the content can take place depending on this rating. PICS specification is a set of technical specifications for labels (meta-data) that help software and rating services to work together. Rating and labeling services choose their own criteria for proper identification and filtering of the content. Since rating will always involve some amount of subjective judgement, it is left to the service provider to define the ratings. Rating can be through self-labeling or third-party labeling of content. In third-party labeling some independent rating agency can be used. The rating of Internet sites was originally designed to help parents and teachers control what children access on the Internet, but it also facilitates other uses for labels, including code signing and privacy.

The RSACI (Recreational Software Advisory Council Internet) has a PICS-compliant rating system called Resaca. Web pages that have been rated with the Resaca system contain labels recognized by many popular browsers like Netscape and Internet Explorer. Resaca uses four categories—violence, nudity, sex, and language—and a number for each category indicating the

degree or level of potentially offensive content. Each number can range from 0, meaning the page contains no potentially offensive content, to 4, meaning the page contains the highest levels of potentially offensive content. For example, a page with a Resaca language level of 0 contains no offensive language or slangs. A page with a language level of 4 contains crude, vulgar language or extreme hate speech. When an end-user asks to see a particular URL, the software filter fetches the document but also makes an inquiry to the label bureau to ask for labels that describe that URL. Depending on what the labels say, the filter may block access to that URL. PICS labels can describe anything that can be named with a URL. That includes FTP and Gopher. E-mail messages do not normally have URLs, but messages from discussion lists that are archived on the Web do have URLs and can thus be labeled. A label can include a cryptographic signature. This mechanism lets the user check that the label was authorized by the service provider.

While the motivation for PICS was concern over children accessing inappropriate materials, it is a general "meta-data" system, meaning that labels can provide any kind of descriptive information about Internet material. For example, a labeling vocabulary could indicate the literary quality of an item rather than its appropriateness for children. Most immediately, PICS labels could help in finding particularly desirable materials, and this is the main motivation for the ongoing work on a next generation label format that can include arbitrary text strings. More generally, the W3C is working to extend Web meta-data capabilities generally and is applying them specifically in the following areas:

1. Digital Signature: Coupling the ability to make assertions with a cryptographic signature block that ensures integrity and authenticity.
2. Intellectual Property Rights Management: Using a meta-data system to label Web resources with respect to their authors, owners and rights management information.
3. Privacy (P3): Using a meta-data system to allow sites to make assertions about their privacy practices and for users to express their preferences for the type of interaction they want to have with those sites.
4. Personalization: Based on some policy, the content can be personalized to suit the need of the user and the service.

Regardless of content control, meta-data systems such as PICS are going to be an important part of the Web, because they enable more sophisticated commerce (build and manage trust relationships), communication, indexing, and searching services. Content filtering can take place either at the client end or at the middleware proxy end.

Content Aggregation

Over a period, the dynamics associated with the content has changed considerably. Earlier, there was a requester requesting for content and a responder responding to the content requested. The game was simple with only two players, the requester and the responder. These contents were corporate content or content for the mass (primarily web sites). There was no concept of charging for the content. Today there is a concept of OEM (Original Equipment Manufacturer) in content. There are some organizations which create content like an OEM. There are other ASPs (Application Service Providers), MVNOs (Mobile Virtual Network Operators), and content aggregators who source content from these OEMs and provide the content as a value added service to different individuals, content providers, and network operators.

In the current scenario, there are primarily four parties involved; they are end user (EU), the content provider (CP), the content aggregator (CA), and the ISP (Internet Service Provider) or the wireless or wireline network operator (NO). The network operator will have routers, cache, gateways and other nodes to offer the service. In this scheme anybody can become a requester or a responder. There could be different parameters, which will determine the content. These parameters are of two types, static and dynamic. The static adaptation parameters are those which can be received before the service begins. The content is adapted, based on this parameter. The dynamic adaptation parameters are those which are required with every request. For example, a user may initiate a request for a MPEG stream. The NO will transcode the stream to suit the bandwidth of the end user and delivers the same to the user. However, through a dynamic parameter, the user can specify a different parameter for transcoding.

From the content aggregator's perspective we may classify the service into two categories:

1. Single service request: This works at the user level and works for only one user. For example, a user may request the proxy server at the NO to translate the page into Hindi and then deliver the same to the user. In this case, the end user buys the content and the translation service.
2. Group service request: This works for a group of users. This type of request is initiated either at the CA level or the NO level. For example, the content aggregator has some arrangement for advertisement. The content aggregator examines all the HTML pages and inserts an advertisement at an appropriate place.

Seamless Communication

The basic premise of a ubiquitous system is that the system will be available and accessible from anywhere, anytime and through any network or device. A user will be able to access the system after moving from one place to another place (foreign place). The user will also be able to access the system while on the move (traveling mode). Mobile healthcare professionals, for example, may need to seamlessly switch between different modes of communication when they move from indoors to outdoors. A corporate user requires a similar kind of facility as well. Also, what is necessary is, during the movement, the session needs to continue. If we take the example of healthcare sector, some data and information are exchanged between the patient and the hospital. While the patient is moved from home, to ambulance, to a helicopter, to the hospital, the information exchange has to continue without any interruption.

Seamless communication will combine seamless handoffs and seamless roaming. Handoff is the process by which the connection to the network (point of attachment) is moved from one base station (access point) to another base station within the same network. Whereas, roaming will involve the point of attachment moving from one base station of one network to another base station of another network. The basic challenge in handoff is that it has to work while a session is in progress. Cellular technology with respect to voice has reached a level of maturity where a seamless voice communication is possible through handoff and roaming. The data technology is yet to mature to provide a similar level of service. In some parts of the world, handoff is termed as handover.

Seamless communication offers users freedom to roam across different wireless networks. Roaming works within homogeneous networks, like GSM to GSM or CDMA2000 to CDMA2000.

Nowadays, roaming is also possible from GSM to CDMA2000 network and vice-versa provided the user device is dual band and can connect to both these networks. True seamless roaming will include handoff and roaming in a heterogeneous hybrid network. The user will move from a WiFi to 3G to wired LAN to GSM while the session is in progress. Users will be able to communicate using whatever wireless device is currently at hand. Thus, GPRS-enabled cell phones, PDAs and laptops will be able to roam and communicate freely and access the Internet across both WLANs and WWANs.

In seamless roaming, the following aspects need to be maintained and managed in a seamless fashion without any disruption of service:

1. Authentication across network boundaries.
2. Authorization across network boundaries.
3. Billing and charging data collection.
4. End-to-end data security across roaming.
5. Handoff between wireless access points.
6. Roaming between networks.
7. Session migration.
8. IP mobility.

The task of managing authentication between client devices and networks, often involving multiple login names and passwords, will become automatic and invisible to the user, as will the configuration of various settings and preferences that accumulate with client devices.

Autonomous Computing

The world is heading for a software complexity crisis. Software systems are becoming bigger and more complex. Systems and applications cover millions of lines of code and require skilled IP professionals to install, configure, tune and maintain. New approaches are needed to provide flexible and adaptable software and hardware, both for mobile devices and the intelligent environment. Ease of use will have some effect on acceptance of a ubiquitous system. The scale of these ubiquitous systems necessitates "autonomic" systems. The purpose of autonomous system is to free users and system administrators from the details of system operation and maintenance complexity. Also, the system will run 24×7. The essence of autonomous system is self-management, which is a combination of the following functions:

1. **Self-configurable:** An autonomous system will configure itself automatically in accordance with high-level policies. This will suit the functional requirement of the user.
2. **Self-optimizing:** An autonomous system will continuously look for ways to improve its operation with respect to resource, cost and performance. This will mean that an autonomous system will keep on tuning hundreds of tunable parameters to suit the user and the environment.
3. **Self-healing:** An autonomous system will detect, diagnose and repair localized problems resulting from bugs or failures. These failures could be the result of either software or hardware failure.
4. **Self-protecting:** An autonomous system will be self-protecting. This will be from two aspects. It will defend itself from external attacks; also, it will not propagate or cascade failure to other parts of the system.

5. **Self-upgradable:** An autonomous system will be able to grow and upgrade itself within the control of the above properties.

Design tools and theories may be needed to support large-scale autonomic computing for small devices.

2.6.2 Context Aware Systems

The role of a context manager is to maintain information pertaining to location, mobile devices, network, users, the environment around each mobile device and any other context information deemed relevant. Following is a description of these information and relevance in the mobile computing environment.

- *Location information:* This feature helps us to identify the location of the user/device. This can be achieved in either of the two ways. One is through the device and the other is through the network. From the device, the best way to find the location is through GPS (Global Positioning Systems). GPS-based systems can offer location information to a precision of 10 feet radius. Also, the location of the base station with which the device is associated can help us to get the location information. In certain networks, GSM for example, the base station location can be obtained from the device through the CID (Cell ID) value. From the network side the location of the device can be determined through timing advance technology. However, this information relates to a point when a successful call was made. Base-station-based location information is likely to be correct to the precision of 100 feet radius.
- *Device information:* This feature helps us to know the characteristics of the device. This is required to determine the resource capability and the user interface capability. In a mobile computing environment the user will move from device to device. Therefore, it is essential to know the device context. Device information can be obtained from the device and from the network. Through the User-Agent parameter of HTTP protocol we can get some information about the device. As this information is provided by the browser in the device, the information is very generic. This does not give the device properties like color, pixel capability, display size, etc. From the network side, the information about the device can be obtained from the EIR (Equipment Identity Register) database of the network. In all the wireless networks (GSM, GPRS, UMTS, 3G) we have the EIR. However, we do not have any concept of EIR in wireless LAN or WiFi.
- *Network information:* In a mobile computing environment, the user moves from network to network. Sometime they are even heterogeneous in nature. Network information is required to identify the capability of the network. Capability information will include security infrastructure, services offered by the networks, etc. For example, while roaming a user moves from a GPRS network to a GSM network. Therefore, the rendering may need an adaptation from WAP to SMS. In the future, some of these will be done through programmable networks.
- *User information:* This information is required to identify the user correctly. From the security point of view, the system needs to ensure that the user is genuine and is who he claims to be. We need to ensure that nobody else is impersonating. This information can be validated

through authentication independent of device or network. However, user preferences' information need to be obtained from the network. For charging the user properly we need to refer to some subscriber information available in the network.

- *Environment information:* This includes ambient surrounding awareness. We need to know the temperature, elevation, moisture, and other ambient-related information which are necessary for sensor-based networks.

For a general mobile-computing environment we need information related to location, network, user, and device. We also notice that for a majority of the parameters we need to access the information available in different databases within the network. These information are being available through different network interfaces of intelligent networks. These interfaces are Softswitch (http://www.softswitch.org), JAIN (Java API for IN http://java.sun.com/products/jain), Parlay (http://www.parlay.org), and TINA (www.tinac.com). These are explained in Chapter 11.

GPS

Global Positioning System (GPS) is a system that gives us the exact position on the Earth. GPS is funded by and controlled by the US Department of Defense. There are GPS satellites orbiting the Earth, which transmit signals that can be detected by anyone with a GPS receiver. Using the receiver, we can determine the location of the receiver. GPS has three parts: the space segment, the user segment, and the control segment.

The space segment consists of 24 satellites, each in its own orbit 11,000 nautical miles above the Earth. Each GPS setellite takes 12 hours to orbit the Earth. Each satellite is equipped with an accurate clock to let it broadcast signals coupled with a precise time message.

The user segment consists of receivers, which can be in the users' hand, embedded in a mobile device or mounted in a vehicle. The user segment receives the satellite signal which travels at the speed of light. Even at this speed, the signal takes a measurable amount of time to reach the receiver. The difference between the time the signal is sent and the time it is received, multiplied by the speed of light, enables the receiver to calculate the distance to the satellite. To measure precise latitude, longitude and altitude, the receiver measures the time it took for the signals from four separate satellites to get to the receiver. If we know our exact distance from a satellite in space, we know we are somewhere on the surface of an imaginary sphere with radius equal to the distance to the satellite radius. If we know our exact distance from four satellites, we know precisely where we are on the surface of the each.

2.7 Mobile Computing through Internet

We discussed that a network can be divided into three major functional areas, namely, core, edge and access. Likewise, we can divide a ubiquitous network into three functional areas. Out of the three, the core and the edge are likely to be Internet and internet. By internet we define a network which is a combination of various networks and interworks with one another, whereas Internet with the uppercase I is the Internet we know. For mobile and ubiquitous computing, the access network will be both wireless and wired networks. In the case of wireless access network, it could range from infrared, Bluetooth, WiFi, GSM, GPRS, IS-95, CDMA, etc. For wired, it is expected to

be some kind of LAN. In the case of wired network the bandwidth is higher, stable and the device is likely to be a workstation with a large memory and display. Also, such devices are not constrained by the limited battery power.

When the user-facing device is a wired device, the complexity and challenges are far less. However, some of the constraints for wireless can still apply in the case of wired devices and networks. Therefore, from the mobile computing client point of view, consideration for a wired device will be the same as a wireless client.

2.8 Making Existing Applications Mobile-enabled

There are many applications that are now being used within the intranet or the corporate network, that need to be made ubiquitous. These are different productivity tools like e-mail or messaging applications, workflow systems, etc. Information systems for partners and vendors and employees like sales force automation, etc. will also fall within this category. These applications need to be made ubiquitous and mobile-computing capable. There are many ways by which this can be achieved.

1. **Enhance existing application:** Take the current application and enhance it to support mobile computing.
2. **Rent an application from an ASP:** There are many organizations which develop ubiquitous application and rent the same at a fee.
3. **Write a new application:** Develop a new application to meet the new business requirement of mobile computing.
4. **Buy a packaged solution:** There are many companies which are offering packaged solutions for various business areas starting from manufacturing to sales and marketing. Buy and install one of these which will also address the mobile computing needs of the enterprise.
5. **Bridge the gap through middleware:** Use different middleware techniques to face-lift and mobile-computing-enable the existing application.

One of these techniques, or any combinations can be used to make an application ubiquitous. If the enterprise has a source code for the application, enhancement of the existing application may be a choice. Writing a new application by taking care of all the aspects described above may also be a possibility. Buying a package or renting a solution from an ASP can also be a preferred path for some business situations.

Many of these applications might have been developed in-house, but may not be in a position to be enhanced. Some might have been purchased as products. A product developed by an outside agency cannot be enhanced or changed as desired. In many such situations, mobile computing enabling can be done through middleware. The combination of communication and application middleware can be used to make an application mobile. Let us assume that the enterprise has its sales and distribution application running in SAP in IBM AS/400 system. The enterprise wants this system to be wireless-enabled for its mobile sales force. Using TN5250 communication middleware, the application can be abstracted as an object. Through a transaction processing middleware and APIs, the SAP application can be used as a document. By using a transcoding middleware, the application can be wireless-enabled and used through WAP, J2ME or even SMS (Short Message Service). Through middleware, some additional security features can be added.

REFERENCES/FURTHER READING

1. *Brief History of Internet*: http://www.isoc.org/internet/history/brief.shtml.
2. Bush, Vannevar, VB45, 'As we may think', *Atlantic Monthly*: http://www.theatlantic. com-/unbound/flashbks/computer/bushf.htm.
3. *History of Internet*: http://www.isoc.org/internet/history/.
4. *Internet Content Rating Association*: http://www.icra.org.
5. *Internet Timeline*: http://www.zakon.org/robert/internet/timeline/.
6. Kephart, Jeffry O. and David M. Chess, 'The Vision of Automatic Computing,' *IEEE Computer Magazine*, pp 41–50, January 2003.
7. Khan, Javed I. and Yihua He, *Ubiquitous Internet Application Services on Sharable Infrastructure: Technical Report 2002-03-02*, Internetworking and Media Communications Research Laboratories, Deptt. of Computer Science, Kent State University; http:// medianet.kent.edu/technicalreports.html.
8. Korkea-aho, Mari, *Context-Aware Applications Survey*, http://www.hut.fi/~mkorkeaa/ doc/context-aware.html.
9. Milojicic D., F. Douglis, and R. Wheeler (Eds) (1999), *Mobility Processes, Computers, and Agents*, Addison-Wesley.
10. Mohan Rakesh, John R. Smith, and Chung-Sheng Li, (1999), 'Adapting Multimedia Internet Content for Universal Access', *IEEE Transactions on Multimedia*, Vol. 1, No. 1, March 1999, pp. 104–114.
11. *Mosaic*: http://archive.ncsa.uiuc.edu/SDG/Software/Mosaic/NCSAMosaicHome.html.
12. *Platform for Internet Content Selection*: http://www.w3.org/2000/03/PICS-FAQ/.
13. *SyncML*: http://www.openmobilealliance.org/syncml/.

REVIEW QUESTIONS

Q1: Describe the significance of core, edge, and access network. Explain their functions as well.

Q2: What are the different tiers in three-tier architecture? Describe the functions of these tiers?

Q3: What is middleware? Describe its significance in handling the context of any application.

Q4: Explain how can an ISP implement a system using ICAP where some web sites are inaccessible during certain times of the day.

Q5: What do you understand by context? Why is context important? To develop a navigational system for a car, what types of context information will be necessary?

Q6: Write brief notes on:

(a) Client Context Manager

(b) CC/PP

(c) Policy Manager
(d) Security Manager
(e) Semantic Web
(f) P3P

Q7: What is content adaptation? What are the various classifications of transcoding? How can the two be helpful for various device classes?

Q8: How is content rating and filtering helpful in classifying content? What is the role of RSACI and PICS in classifying content?

Q9: You have been asked to develop a location aware restaurant guide system for the Restaurant Foundation of India. Describe four main functions of this system. Describe how will you implement these four functions?

Q10: What is seamless communication? How can seamless communication help in an emergency service rescue operation?

Q11: What is a context aware system? What all can be the types of information needed for developing a fully context aware system?

Q12: Write brief notes on:
(a) GPS
(b) Mobile Computing through Internet

Q13: Discuss with examples how existing applications can be made mobile.

CHAPTER 3

Mobile Computing through Telephony

3.1 EVOLUTION OF TELEPHONY

The first telephone system developed by Alexandra Graham Bell allowed two-way voice communication between two individuals in two locations on either side of a wire. We (known as calling or A party, the person who makes the call) speak into one unit of the phone at one end of the wire and someone else hears our voice at another location (known as called or B party, the person who responds to the call) at the other end of the wire, instantly in real-time. During the long era of analog telephony, the purpose of interconnecting two subscribers was to establish a physical connection between their respective telephone devices. This is achieved by establishing a physical circuit between two parties (A party and B party). In early days, each telephone was connected to a central place (the exchange) and from this exchange the operator would manually connect the call to another subscriber. Whenever a subscriber turned the crank of the telephone, a ringing signal sounded at the operator's switchboard. Upon answering the signal, the operator was asked to connect the call to the other subscriber, which the operator did manually. The operator was required to make a note of who placed the call, whom the call was for, and when it started and ended. This information made it possible to charge the caller for the call, the classic billing and charging information. If we wanted to make a call to someone outside our own local exchange, say to the neighboring exchange, an operator at our exchange would call an operator at the adjacent exchange and then ask the other operator to connect through to the desired subscriber. If we wanted to call someone much further away we had to book a trunk call. In the case of a trunk call, the call would have to be set up with a whole chain of operators, each one calling the next, and so on.

We can say that the market forces of the early 1890s prompted the development of the first automatic telephone exchange. It was called the 'Strowger switch', after its originator Almon B. Strowger. Strowger did not invent the idea of automatic switching; it was first invented in 1879 by Connolly and McTigthe. Strowger was the first person to put it to commercial use. Almon B. Strowger was an undertaker in Kansas City in the US. The story goes that there was another local

competing undertaker whose wife was a telephone operator at the local (manual) telephone exchange. Whenever any caller used to request to be connected to Strowger, calls were deliberately put through to his competitor, who in fact was the husband of this operator. This made Strowger devise a system to eliminate the human factor of the whole equation! Strowger developed a system of automatic switching using an electromechanical switch based around electromagnets and pawls. The first version of automatic exchange was installed in 1892 at La Porte, Indiana in the US.

In 1912, the Swedish engineer Gotthief Betulander patented an automatic switching system based on a grid. This type of exchange was also electromechanical and called crossbar exchange. In 1960, the first Electronic Switching System (ESS) was developed by AT&T and commissioned for testing. Finally, on 30 May 1965, the first commercial electric central office was put into operation at Succasunna, New Jersey. The ESS required a staggering four thousand man-years of work at Bell Labs. In 1976, Bell Labs developed the 4ESS toll switch for the long-distance voice network. This was the first digital circuit switch. The idea behind a digital switch was that the analog voice is digitized before it is given to a switch for switching. The 1960s and 1970s saw the advent of telephone exchanges that were controlled by processors and software (digital computers). These were called stored program control exchanges. The primary objective of a sophisticated telephone exchange is still the same as that of the manual exchange a century ago. These are to detect the A-subscriber's (calling party) call attempt, connect him to the correct B-subscriber (called party), and to save data about the call for the purpose of billing.

The digital revolution in telephony started with the introduction of electronic switches. The next major milestone was achieved in 1962 when the carrier system was made digital. Work on digital transmission began back in the 1920s, when the Bell System researcher Harry Nyquist determined that it was possible to encode an analog signal in digital form if the analog signal was sampled at twice its frequency. Each sample could then be encoded and transmitted. There would be enough information in the encoded signal for the original voice signal to be reconstructed into an understandable analog signal at the receiving end. Assuming that the audio voice band is 0 to 4000 Hzs, we start with a 4 KHz analog voice channel. Then we take a snapshot of the voice signal's amplitude at 1/8000th of a second (every snapshot at double the frequency of 4 KHz). Then we convert the measured amplitude to a number (the quantization process) that is represented by 8 bits. This type of digitization is called Pulse Code Modulation (PCM). Thus, PCM requires 64 Kb/s of digital bandwidth (8 KHz * 8 bits). Alex H. Reeves, working for Western Electric Company in Paris, first conceived PCM in 1937. Bell Laboratory scientists first introduced digital transmission using PCM in 1962. The Bell Laboratory system was named T1 with a transmission rate of 1.544 Mb/s carrying 24 channels of 64 Kb/s each. In Europe, a similar system was called E1, where it had a bandwidth of 2 Mb/s and carried 32 channels of 64 Kb/s. The developments in transmission techniques have been advancing largely to reduce network costs. We have witnessed an evolution from systems employing open-wire lines to multiplexed, analog systems using coaxial or radio links, on to digital fiber-optic systems with a capacity of tens of Gbit/s per fiber pair. The first commercial optical systems came on the scene in 1980.

In a manual switching system, an operator would be able to inform the caller of the current status of the call. In manual exchanges, the operator's intelligence was a control system separate from the switching mechanism. An operator, alerted to an incoming call:

- listens to and remembers the desired number.

- finds the right way to connect the caller's line to the line being called.
- checks if the desired line is free.
- makes the connection, and
- notes down the call details: time of call, duration of call, calling number and called number.

Having removed the need for an operator in the automated exchange, a system was necessary to indicate progress of the call to the caller. A series of distinct tones were generated by a machine called Ring Generator. The tones produced were as follows:

- *Dial Tone (DT):* This is a signal applied to the line after the calling party (A party) has lifted his handset and the switching equipment has allocated him an available outlet (a circuit) for this call to proceed.
- *Busy Tone (BT):* Busy tone indicated either that the called subscriber (B party) is already off-hook (busy) or that the route to the called subscriber is congested.
- *Ring Tone (RT):* When a circuit between A party and the B party is established, the telephone rings at B party's end and a ring tone is generated for the A party.

A normal telephone system is called Public Switched Telephone Network (PSTN). PSTN nodes can be subdivided into three main categories: local exchanges (also known as end office), transit exchanges (also known as local access tandem) and international exchanges (also known as interexchange carrier). Local exchanges are used for the connection of subscribers. Transit exchanges switch traffic within and between different geographical areas. International exchanges, and other gateway-type exchanges switch traffic to telecommunication networks in foreign countries and other networks. A physical wire (also known as local loop) is laid from the local exchange to the telephone device at each subscriber's place. This is traditionally also known as the last mile. In case of a wireless network like GSM or WiLL (Wireless in Local Loop), there is no wire from the local exchange to the telephone. The communication between the local exchange and the telephone device is managed over the wireless radio interface. In India, there are network operators who are offering basic or fixed telephone, WiLL, and GSM.

3.2 Multiple Access Procedures

In a PSTN network, a separate physical wire is used to connect the subscriber's telephone with the switch. Therefore, multiple users can have speech communication at the same time without causing any interference to each other. The scene is different in the case of wireless communication. Radio channel, used in a wireless network, is shared by multiple subscribers. Unless we control simultaneous access of the radio channel (by multiple users), collisions can occur. In a connection-oriented communication, a collision is undesirable. Therefore, every mobile subscriber must be assigned a dedicated communication channel on demand. This is achieved by using different multiplexing techniques (Fig. 3.1).

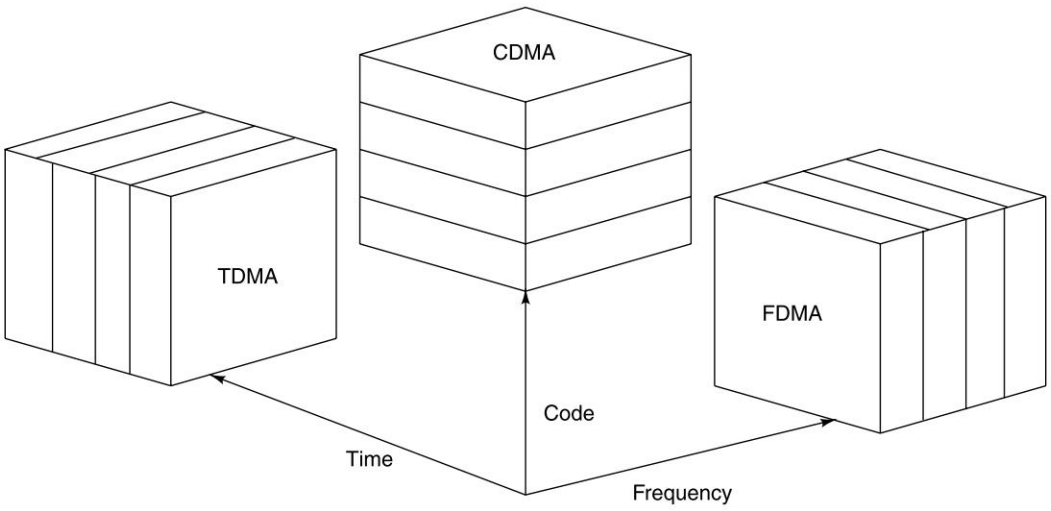


Figure 3.1 Multiple Access Procedures

3.2.1 Frequency Division Multiple Access

Frequency Division Multiple Access (FDMA) is one of the most common multiplexing procedures. The available frequency band is divided into channels of equal bandwidth so that each communication is carried on a different frequency. This multiplexing technique is used in all the first generation analog mobile networks like Advanced Mobile Phone System (AMPS) in the US and Total Access Communication System (TACS) in the UK.

3.2.2 TDMA Variants

Time Division Multiple Access (TDMA) is a multiplexing technique where multiple channels are multiplexed over time. Assuming that we have a 64Kbps channel to transmit one voice channel; but we are having a high speed carrier of 2Mbps (32 * 64Kbps), we can transmit this voice channel in 1/32 second. This implies that we can theoretically divide the 2Mbps channel into a 32 time-slot and use one of them for transmission of our voice channel. In TDMA, several users share the same frequency channel of higher bandwidth by dividing the signal into different time slots. Users transmit their data using their own respective time slots in rapid succession; to synchronize, the transmitter and the receiver need to synchronize using a global clock.

Figure 3.2 shows that a TDMA system divides its transmission medium into frames which are repeated indefinitely (one after another). Each TDMA frame is then divided into time slots of same temporal width that are allotted to individual users. TDMA is a very common multiplexing technique and used in many digital transmissions like GSM, IS-136, SS7, satellite systems, etc.

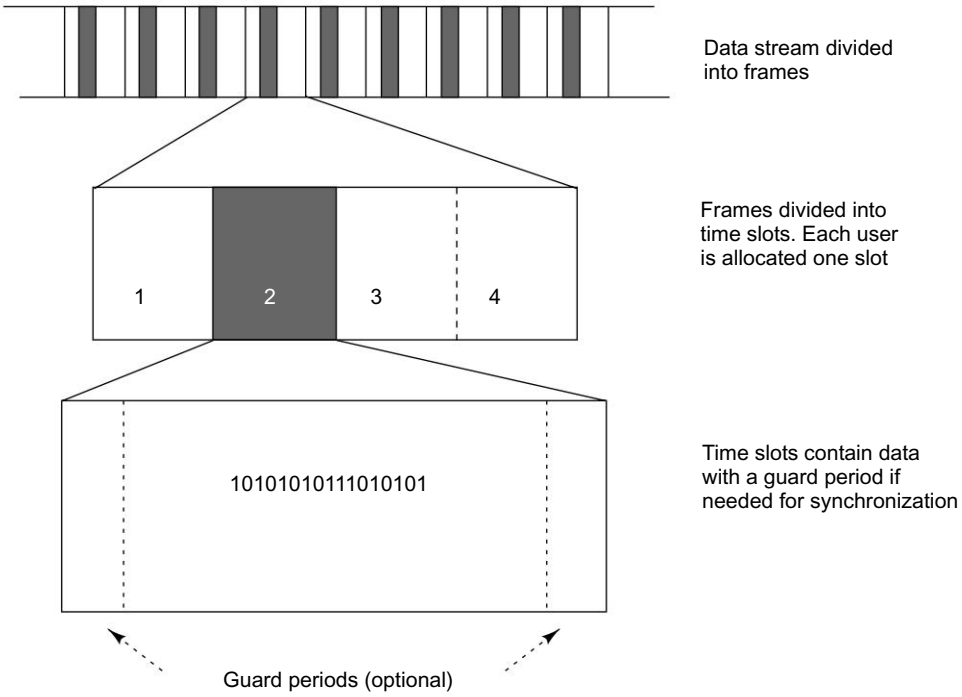


Figure 3.2 TDMA Frames and Time Slots

Fixed TDMA

In Fixed TDMA, connections between time slots in each frame and data streams assigned to a user remain static and switched only when large variations in traffic are required. In this variant of TDMA, the slot sizes are fixed at T/N (where T is time in seconds and N is the number of users). If a station does not transmit during its assigned slot, then the corresponding bandwidth is wasted. This variant is simple to implement but performs poorly since the entire bandwidth is not used.

Dynamic TDMA

In Dynamic TDMA or Dynamic Reservation TDMA (DR TDMA), a scheduling algorithm is used to dynamically reserve a variable number of time slots in each frame to variable bit-rate data streams. This reservation algorithm is based on the traffic demand of each data stream. The fixed length DR TDMA frame is time-duplexed into an uplink and downlink channel and the boundary between these two parts is dynamically adjusted as a function of the traffic load.

Packet Reservation Multiple Access

Packet Reservation Multiple Access (PRMA) is a packet based TDMA where the users contend for the time slots. In PRMA, a user can reserve a time slot in advance for future use and optimize the bandwidth in radio transmission. The two prominent variants of PRMA are Dynamic PRMA (DPRMA) and PRMA Hindering States (PRMA HS). In DPRMA, each mobile station is responsible

for making a reasonable estimate of its bandwidth requirements and then request for resource allocation to the base station. It aims to closely match each user's transmission rate with its packet generation rate. PRMA HS in contrast allows a terminal to transmit during the time interval needed to receive the outcome of a reservation attempt. PRMA HS is used for the uplink of Low Earth Orbit (LEO) satellite mobile communications.

3.2.3 Code Division Multiple Access

Code Division Multiple Access (CDMA) is a broadband system. CDMA uses spread spectrum technique where each subscriber uses the whole system bandwidth. Unlike the FDMA or TDMA where a frequency or time slot is assigned exclusively to a subscriber, in CDMA all subscribers in a cell use the same frequency band simultaneously. To separate the signals, each subscriber is assigned an orthogonal code called "chip".

3.2.4 Space Division Multiple Access

Along with TDMA, FDMA, and CDMA, we need to make use of the space effectively. Space division multiple access (SDMA) is a technique where we use different parts of the space for multiplexing. SDMA is used in radio transmission and is more useful in satellite communications to optimize the use of radio spectrum by using directional properties of antennas. In SDMA, antennas are highly directional, allowing duplicate frequencies to be used at the same time for multiple surface zones on earth. SDMA requires careful choice of zones for each transmitter, and also requires precise antenna alignment.

3.3 Satellite Communication Systems

On October 4, 1957 Union of Soviet Socialist Republics (USSR) scientists placed the first manmade artificial satellite name **Sputnik** in earth's orbit. Four months later, the US matched it by sending **Explorer 1** into earth's orbit. Because a satellite orbits around the earth, part of the earth is always within the satellite's range; therefore, technically it is possible to use satellites for communications, military, or even spying functions.

In 1960, the world's first communication satellite **Echo 1** was launched by the US. Echo 1 was a passive communication satellite to communicate across the US and across the Atlantic Ocean by reflecting signals using a large aluminized plastic balloon (100 feet in diameter). It reflected Radio and TV signals transmitted to the satellite back to the earth station within view of the satellite. Being a low orbit satellite, Echo 1 circled the earth every 90 minutes; therefore, everybody on earth could eventually see it sometime, but, no single person, ever saw it for more than 10 minutes out of every 90 minute orbit—this means, it could not be used for long communication. In 1958, the **Score** satellite was put into earth's orbit that worked as a relay agent—it carried a tape recorder that would record messages as it passed over an originating station and then rebroadcast them as it passed over the destination station.

The major limitation of a passive satellite was that it needed high transmission power to overcome transmission loss. **Telstar 1** launched on July 10, 1962 was the first active communication satellite. It was also a low orbit satellite but it could see Europe and the US simultaneously during one part of its orbit and Japan and the US during another part of its orbit; as a result, it provided real-time communications between the US and those two regions—for a few minutes out of every hour.

3.3.1 Communicating through Satellite

Every communications satellite involves the transmission of information from an originating ground station to the satellite followed by a retransmission of the information from the satellite back to the ground called the uplink and the downlink respectively (Fig. 3.3). Hence, the satellite must have a receiver with receive antennas, and a transmitter with transmit antennas. It must also have some methods for connecting the uplink to the downlink for retransmission with amplification; also, it must have electrical power through solar energy to run all of the electronics. The downlink may either be to a select number of ground stations or may be broadcast to everyone over a large area.

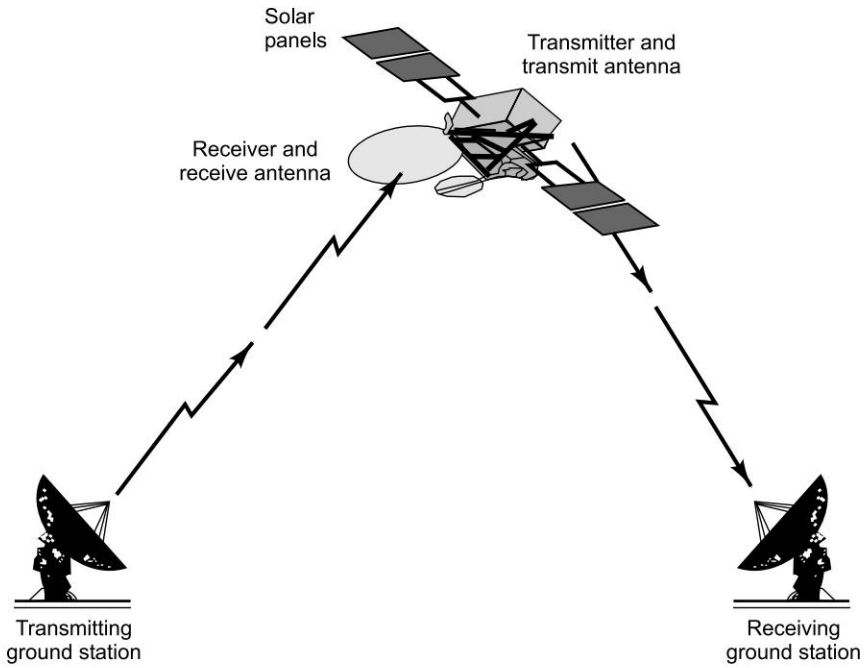


Figure 3.3 Satellite Communications System

A properly designed satellite antenna will concentrate most of the transmitter power within a designated area using space division multiplexing. One of the biggest differences between a low earth satellite and a geosynchronous satellite is in their antennas. All antennas in use today radiate energy preferentially in some direction. The most important application for communication satellites was in intercontinental long distance telephony. The fixed Public Switched Telephone Network relays telephone calls from land line telephones to an earth station, where they are then transmitted to a geostationary satellite. The downlink follows an analogous path.

3.3.2 Low Orbit Satellite

A Low Earth Orbit (LEO) satellite typically orbits around the earth about 400 kilometers above the earth's surface with a time period of about 90 minutes. These satellites are only visible from within a radius of roughly 1000 kilometers from the sub-satellite point. Sub-satellite point is the point of intersection of earth's surface with the straight line from the satellite to the center of earth. The greatest advantage of LEO satellite is that it does not need high powered rockets—making it less expensive to launch. Also, due to its proximity to the ground, LEO does not require high signal strength.

For uninterrupted communication services a large number of satellites are needed so that they communicate with each other and one of the satellites is in touch with the user. Unlike in a mobile telephone system where the user is mobile and transceivers are static, in LEO satellite system, user is relatively static compared to the transceiver, which is mobile.

3.3.3 Medium Orbit Satellite

Medium Earth Orbit (MEO), sometimes called Intermediate Circular Orbit (ICO), is the region of space around the earth above low earth orbit of 2,000 kilometres and below geostationary orbit of 35,786 kilometers. The most common use for satellites in this region is for navigation, such as the GPS (with an altitude of 20,200 kilometers), Communications satellites that cover the North and South Pole are also put in MEO. The orbital periods of MEO satellites range from about 2 to 24 hours. The MEO orbit has a moderate number of satellites.

3.3.4 Geostationary Satellite

In geostationary satellite the orbit of the artificial satellite is such that the orbital speed of the satellite is same as the speed of earth's rotation. Though the satellite is moving at a high speed, from earth it will always appear to be stationary—this is the reason for calling it geo-stationary. A Geostationary Earth Orbit (GEO) can be achieved only very close to the ring 35,786 km directly above the equator. This equates to an orbital velocity of 3.07 km/s or a period of 1436 minutes, which equates to almost exactly one sidereal day or 23.934461223 hours. The idea of a geostationary orbit was first proposed by Arthur C. Clarke in 1945; therefore, a geostationary orbit is also known as the Clarke Orbit. The GEO satellite could view approximately 42% of the earth. Therefore, a system of three GEO satellites, with the ability to relay messages from one GEO to the other could interconnect virtually all of the earth except the polar regions.

Unlike LEO or MEO, the GEO orbit is much higher—demanding high power rockets. In 1963, the necessary rocket booster power was available for the first time and the first geosynchronous satellite, **Syncom 2** was launched by the US into earth's orbit. Geosynchronous orbit is so far that the time to transmit a signal from earth to the satellite and back is approximately ¼ of a second—the time required to travel 36,000 km up and 36,000 km down at the speed of light. For telephone conversations, this delay can sometimes be annoying.

3.3.5 Satellite Phones

Initially satellite communication was being used for broadcast to stationary TV receivers, and transmission of telephone channels. However, demand on mobile phone made some companies to look into satellite phones that will connect a subscriber directly through the communication satellite, where the satellite will function as the transceiver station connecting the mobile phone. There are few companies that offer such facility; we describe some of them as case study.

- *Iridium:* Iridium comprises a group satellites working in concert as a satellite constellation. The Iridium constellation is used to provide voice and data communication to satellite phones, pagers and integrated transceivers over the entire surface of the earth. The constellation uses 66 active satellites in orbit along with spare in-orbit satellites to serve in case of failure. The satellites orbit the earth in roughly 100 minutes. Each satellite can support up to 1100 concurrent phone calls. Iridium satellites are in low earth orbit at a height of approximately 780 km with spare satellites at 667 km storage orbit. Spare satellites will be boosted to the correct altitude and put into service in case of failure of active the satellite. Satellites communicate with neighboring satellites via Ka band inter-satellite links. For more information on Iridium, please refer to www.iridium.com and www.satphoneusa.com/iridium/network.html.
- *Globalstar:* Globalstar is another mobile satellite voice and data services provider offering services to subscribers around the world. Globalstar uses 52 LEO satellites—48 satellites for communication with four satellites as spare. Globalstar's products include mobile and fixed satellite telephones, simplex and duplex satellite data modems and satellite airtime packages. Many land based and maritime industries make use of the various Globalstar products and services from remote areas beyond the reach of cellular and landline telephone service. For more information on Globalstar, please refer to www.globalstar.com.
- *Thuraya:* Thuraya is another satellite phone company that mainly services in Asia and Africa. Unlike Globalstar or Iridium that uses LEO, Thuraya uses three geostationary satellites. In all practical purpose Thuraya is a GSM (see Chapter 5) cellular telephone network with a satellite BTS (Base Transceiver System). All Thuraya phones use the same GSM SIM cards and can roam in any terrestrial GSM network around the world. Thuraya phones have a dual-mode feature that allows them to operate in the Thuraya satellite network or GSM terrestrial mobile networks while outside the satellite coverage. Thuraya subscribers can also switch to roaming GSM network if it is available. For more information on Thuraya, please refer to www.thuraya.com.

3.4 Mobile Computing through Telephone

One of the early examples of mobile computing was accessing applications and services through voice interface. This technology was generally referred to as Computer Telephony Interface (CTI). Different banks around the world were offering telephone banking for quite sometime using this technology. In a telephone banking application, the user calls a number and then does his banking

transaction through a fixed telephone. In this application the telephone does many functions of a bank teller. Input to this system is a telephone keyboard and output is a synthesized voice. These applications can be used from anywhere in the world. The only issue in this case is the cost of a call. Let us take the example of a bank, which has branches only in Bangalore (like some co-operative banks in India). Let us assume that this bank offers telephone banking facility only in Bangalore city in India. The service number for the bank is +91(80)2692265 (+91 80 2MYBANK). Assuming I am in Bangalore, it costs me a local call to check the balance in my account. When I am traveling and want to check my account detail from elsewhere, say Delhi, I make a call to the Bangalore number +91802692265 from Delhi and pay a long distance charge. Let us now assume that the bank has gone for a VPN (Virtual Private Network) between Delhi and Bangalore. The bank now offers telephone banking services in Delhi through a service number +91 (11) 26813241. This will enable me to use the same service in Delhi at the cost of a local call. The only challenge is that the number in Delhi is different from that in Bangalore. In such cases the bank customers are required to remember multiple service numbers for mobile computing.

The telephone companies soon came up with a brilliant idea to solve this problem of multiple numbers by offering 800 services using Intelligent Networks (IN) technology. This is also commonly known as Toll Free numbers. In this technology only one number like 1-800-2MYBANK is published. This number is not attached to any specific exchange or any specific city. When a subscriber calls this number an optimal routing is done and the call is connected to the nearest service center. The advantage is that users remember only one number. They can call the same number from anywhere. They also need not worry about the distance of the call as these numbers are generally toll free. Toll free means that the call is charged to the B party instead of the A party, the caller. In India this service was introduced as 1-600 service and later changed to 1-800. For example, to shop through TV one had to dial 1-600-117247. If you dial 1-800-111100 from anywhere in India you will be connected to Microsoft office in Delhi.

To make this type of mobile computing work through voice interfaces, we use Interactive Voice Response (IVR). In the US and Japan, IVRs are commonly known as Voice Response Unit (VRU). The technical name for this technology is CT (Computer Telephony) or CTI (Computer Telephony Interface or Computer Telephony Integration). IVR software can be hosted on a Windows-NT, Linux, or other computers with voice cards. There are many companies which manufacture voice cards; however, one of the most popular card vendors is from Intel/Dialogic. IVR works as the gateway between a voice-based telephone system and a computer system. Multiple telephone lines are connected to the voice card through appropriate telecom interfaces (E1 or an analog telephone extension). When a caller dials the IVR number, a ring tone is received by the voice card within the IVR. The voice card answers the call and establishes a connection between the caller and the IVR application. The caller uses the telephone keyboard to input data. Figure 3.4 depicts an IVR infrastructure. The switch can be either a PSTN exchange or a local PBX in the office. For PSTN switch, the voice card will have E1 interface whereas for a PBX, the voice card will have analog interface. The IVR will have all the gateway-related functions. The server will host the business application.

A telephone keyboard has 12 keys (viz., 1, 2, 3, 4, 5, 6, 7, 8, 9, 0, *, and #). The English alphabetic characters are also mapped on these 12 keys. They are mapped as follows.

1. Alphabet A, B, C on key 2
2. Alphabet D, E, F on key 3
3. Alphabet G, H, I on key 4
4. Alphabet J, K, L on key 5
5. Alphabet M, N, O on key 6
6. Alphabet P, Q, R, S on key 7
7. Alphabet T, U, V on key 8
8. Alphabet W, X, Y, Z on key 9

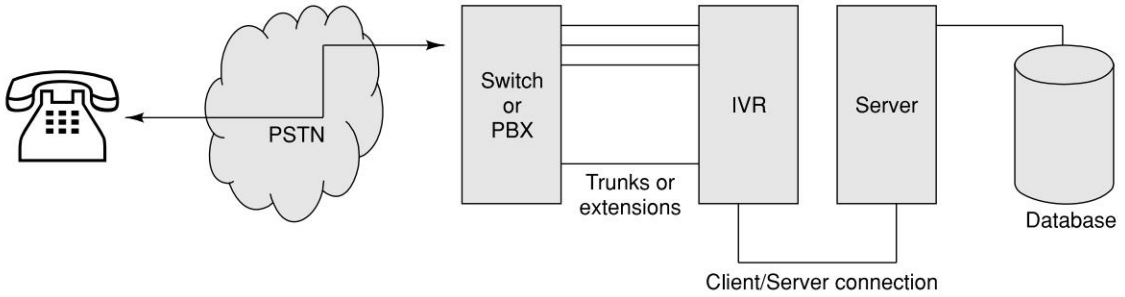


Figure 3.4 The IVR Architecture

It is possible to enter alphabetic data through the telephone keyboard by pressing a key in multiple successions. For example Delhi will be entered as 3-3 (D), 3-3-3 (E), 5-5-5-5 (L), 4-4-4 (H), 4-4-4-4 (I). These key inputs are received by the voice card as DTMF (Dual Tone Multi Frequency) inputs generated through a combination of frequencies. Following is the table (Table 3.1) of these frequencies:

If we press key 1, it will generate a frequency 697 + 1209 Hz. Likewise 0 will be 941+1336 Hz. These DTMF signals are different audio frequencies interpreted by the voice card and passed to the IVR program as numbers through appropriate APIs. For example, the user presses '2' three times. The voice card will receive 697+1336 Hz- 697+1336 Hz-697+1336 Hz. This will be interpreted by a program as 2-2-2. Looking at the time interval between the numbers, the program can decide whether the user entered '222' or 'B'. When the application needs to send an output to the user, the standard data is converted into voice either through synthesizing voice files or through TTS (Text To Speech). In a cheque-printing software we print the amount in both words and figure. For example, an amount of 'Rs. 320,145.00' will be printed on a cheque as 'Rupees three lacs twenty thousand one hundred forty-five only'. Within the cheque-printing application, one function converted the numeric number 320145 into text. Likewise in the case of IVR application, we assemble a series of prerecorded voice prompts to generate the equivalent sound response. In this case we assemble voice data 'three' 'lacs' 'twenty' 'thousand' 'one' 'hundred' 'forty' 'five' 'only' and then give the voice card to play. We can generate the same voice response by giving the number 320,145 to the TTS interface to convert the text into speech and play through the IVR. TTS is a interface software which takes text and numbers as input and generates equivalent sounds at runtime. There are different TTS available for different languages. In India there are companies that have Hindi TTS software. Hindi TTS software takes Devanagari text stream as input and generates the voice as if someone is reading the same text.

Table 3.1 DTMF Frequencies

	1209 Hz	*1336 Hz*	*1477 Hz*
697 Hz	1	2/ABC	3/DEF
770 Hz	4/GHI	5/JKL	6/MNO
852 Hz	7/PQRS	8/TUV	9/WXYZ
941 Hz	*	0	#

3.4.1 Overview of the Voice Software

Voice technology encompasses the processing and manipulation of an audio signal in a Computer Telephony (CT) system. It supports filtering, analyzing, recording, digitizing, compressing, storing, expanding and replaying of audio voice. A CT system also includes the ability to receive, recognize and generate specific telephone and network tones. This fundamental technology is at the core of most IVR systems. Voice products also offer Digital Signal Processing (DSP) technology and signal processing algorithms, for building the core of any converged communications system. Most of the voice cards come with industry-standard Peripheral Component Interface (PCI) bus expansion boards. The PCI interface makes it possible to integrate these voice products into Windows or Linux systems quite easily (Fig. 3.5).

One of the most popular voice cards used for small office interface is D/41JCT-LS from Dialogic. Dialogic (part of Intel) products are the de-facto industry standard. The D/41JCT-LS board is a four-port analog converged communications voice, fax, and software-based speech recognition board. This board is ideal for building enterprise unified messaging and interactive voice response (IVR) applications. The D/41JCT-LS provides four telephone line interface circuits for direct connection to analog loop start lines through RJ11 (the standard telephone jack used in homes) interface. D/41JCT-LS possesses dual-processor architecture, comprising a digital signal processor (DSP) and a general-purpose microprocessor, which handles all telephony signaling and performs DTMF (touchtone) and audio/voice signal processing tasks. A voice card also has some on-board memory and voice store-and-forward feature.

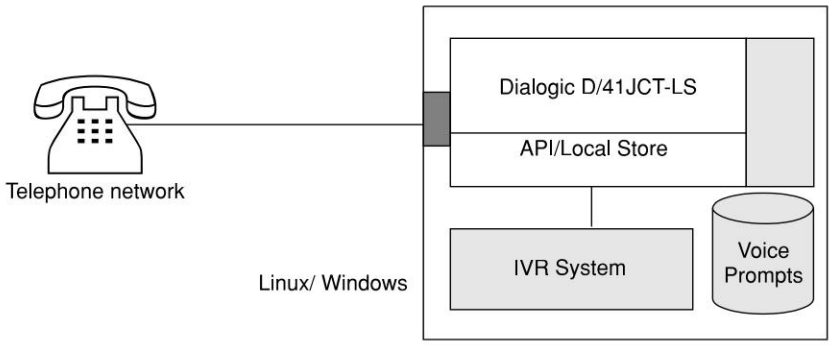


Figure 3.5 Inside an IVR

3.4.2 Voice Driver and API

In this section we describe Dialogic Voice Driver APIs. Dialogic is now part of Intel and one of the leading vendors on voice-based hardware. Many IVR vendors around the world use Dialogic cards from Intel in their IVR systems. Voice driver in an IVR system is used to communicate and control the voice hardware on the IVR system. This section describes Dialogic APIs. Voice card from some other vendor will have similar type of APIs. A voice driver can make calls, answer calls, identify caller ID, play and record sound from the phone line, detect DTMF signals (touch-tones) dialed by the caller. It can tear down a call and detect when the caller has hung up. It also offers APIs to record the transaction details. Transaction information is required for audit trail and for charging. Voice boards are treated as board devices, channels within a board are treated as channel devices or board sub-devices by the voice driver.

3.4.3 IVR Programming

There are different voice libraries provided by Dialogic to interface with the voice driver. The voice libraries for single-threaded and multi-threaded applications include:

- libdxxmt.lib–the main Voice Library
- libsrlmt.lib–the Standard Run-time Library

These C function libraries can be used to:

- Utilize all the voice board features of call management.
- Write applications using a Single-threaded Asynchronous or Multi-threaded paradigm.
- Configure devices.
- Handle events that occur on the devices.
- Return device information.
- Gather call transaction details.

The Standard Run-time Library provides a set of common system functions that are device independent and are applicable to all Dialogic devices.

3.4.4 Single-threaded Asynchronous Programming Model

Single-threaded asynchronous programming enables a single program to control multiple voice channels within a single thread. This allows the development of complex applications where multiple tasks must be coordinated simultaneously. The asynchronous programming model supports both polled and callback event management.

3.4.5 Multi-threaded Synchronous Programming Model

The multi-threaded synchronous programming model uses functions that block application execution until the function completes. This model requires that the application controls each

channel from a separate thread or process. The operating system can put individual device threads to sleep while allowing threads that control other Dialogic devices to continue their actions unabated. When a Dialogic function is completed, the operating system wakes up the function's thread so that processing continues. This model enables the IVR system to assign distinct applications to different channels dynamically in real-time.

Voice APIs

To use the voice board, Dialogic provides different APIs. All Dialogic APIs are prefixed with dx_; this helps to identify them easily. APIs are available for device management, configuration function, input output functions, play and record functions, tone detection functions, tone generation functions, call control functions, etc. Following are some of the important functions in Dialogic voice card, which are used quite often to develop a mobile computing application.

- `dx_open( )` – opens a voice channel
- `dx_close( )` – closes a voice channel
- `dx_wtcallid( )` – waits for rings and reports Caller ID
- `dx_getdig( )` – gets digits from channel digit buffer (for reading the key input)
- `dx_play( )` – plays recorded voice data
- `dx_playvox( )` – plays a single vox file
- `dx_playwav( )` – plays a single wave file
- `dx_rec( )` – records voice data
- `dx_recvox( )` – records voice data to a single vox file
- `dx_recwav( )` – records voice data to a single wave file
- `dx_dial( )` – dials an ASCII string of digits.

3.5 Developing an IVR Application

Like any other application development, computer telephony/IVR application development also requires definition of the user interface. The user interface in IVR application is called the Call Flow. In a call flow we define how the call will be managed. Also, we note down the precise prompts that are played as an output. As described earlier, these prompts are generally pre-recorded by people with professional voice. Let us take a simple example of ticket booking in a theatre. In this application, the user dials a service number and enters a phone number. The operator calls the user back and accepts the booking request. The extra step of call back is done for security reasons. Figure 3.6 depicts the call flow for this application.

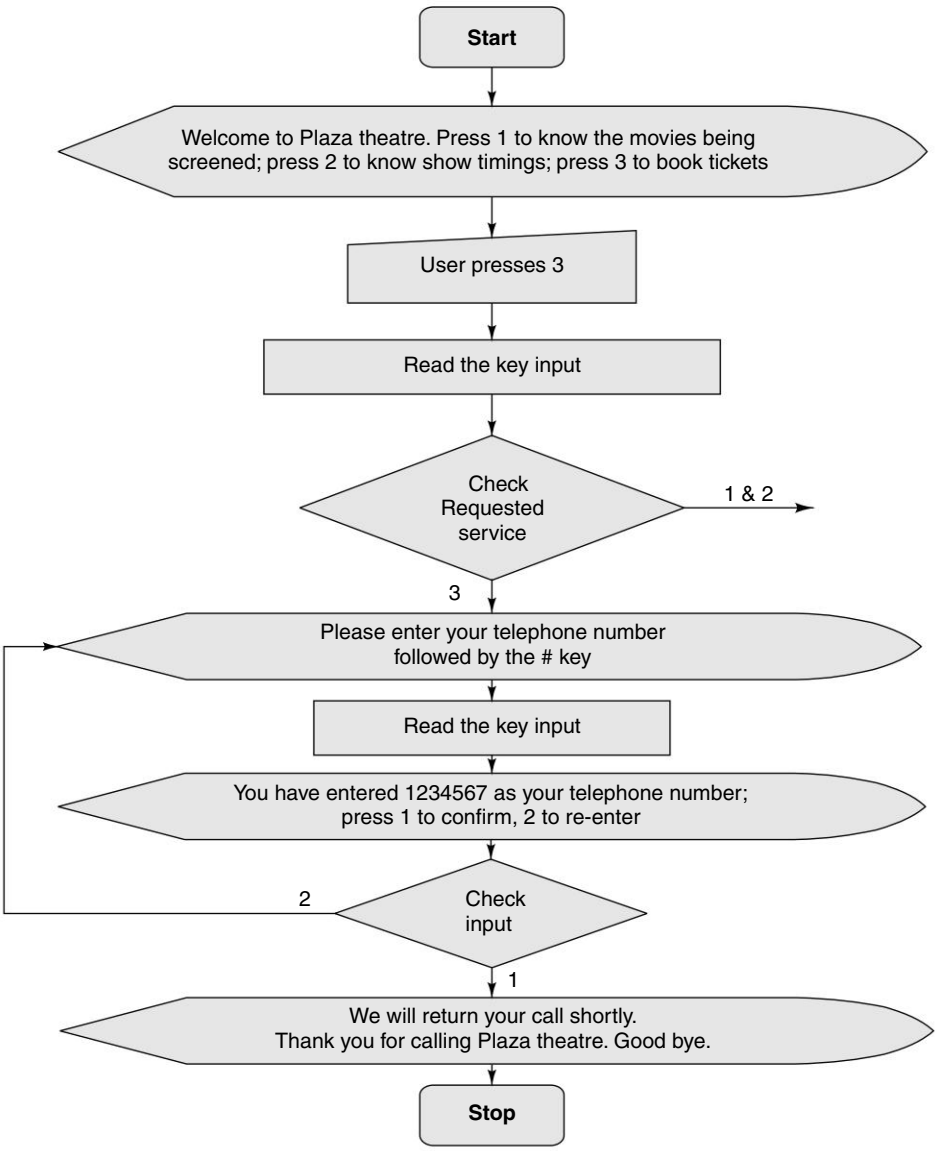


Figure 3.6 Call Flow for a Theatre Ticket Booking

Example of IVR Application

Like any other program, in an IVR program we need to open the user interface device (voice board in the case of IVR) and the data store. Table 3.2 lists an example for one such IVR program.

Table 3.2 An IVR example

```
1  /* Play a voice file. Terminate on receiving 4 digits or at end of file*/
2  #include <fcntl.h>
3  #include <srllib.h>
4  #include <dxxxlib.h>
5  #include <windows.h>
6  main()
7  {
8     int chdev;
9     DX_IOTT iott;
10    DV_TPT tpt;
11    DV_DIGIT dig;
12    .
13    .
14    /* Open the device using dx_open( ).
15    Get channel device descriptor in * chdev. */
16    if ((chdev = dx_open("dxxxB1C1",NULL)) == -1)
17       {
18       /* process error */
19       }
20    /* set up DX_IOTT */
21    iott.io_type = IO_DEV|IO_EOT;
22    iott.io_bufp = 0;
23    iott.io_offset = 0;
24    iott.io_length = -1; /* play till end of file */
25    if ((iott.io_fhandle =
26    dx_fileopen("prompt.vox", O_RDONLY|O_BINARY)) == -1)
27    {
28    /* process error */
29    }
30    /* set up DV_TPT */
31    dx_clrtpt(tpt,3);
32    tpt[0].tp_type = IO_CONT;
33    tpt[0].tp_termno = DX_MAXDTMF; /* Maximum number of digits */
34    tpt[0].tp_length = 4; /* terminate on 4 digits */
```

```
35    tpt[0].tp_flags = TF_MAXDTMF; /* terminate if already in buf. */
36    tpt[1].tp_type = IO_CONT;
37    tpt[1].tp_termno = DX_LCOFF; /* LC off termination */
38    tpt[1].tp_length = 3; /* Use 30 ms (10 ms resolution * timer) */
39    tpt[1].tp_flags = TF_LCOFF|TF_10MS; /* level triggered, clear
40       history, * 10 ms resolution */
41    tpt[2].tp_type = IO_EOT;
42    tpt[2].tp_termno = DX_MAXTIME; /* Function Time */
43    tpt[2].tp_length = 100; /* 10 seconds (100 ms resolution * timer) */
44    /* clear previously entered digits */
45    if (dx_clrdigbuf(chdev) == -1)
46          {
47          /* process error */
48          }
49       /* Now play the file */
50    if (dx_play(chdev,&iott,&tpt,EV_SYNC) == -1)
51          {
52          /* process error */
53          }
54    /* get digit using dx_getdig( ) and continue processing. */
55    /* Set up the DV_TPT and get the digits */
56    if ((numdigs = dx_getdig(chdev,tpt, &digp, EV_SYNC))== -1)
57          {
58          /* process error */
59          }
60       for (cnt=0; cnt < numdigs; cnt++)
61       {
62          printf("\nDigit received = %c, digit type = %d",
63          digp.dg_value[cnt], digp.dg_type[cnt]);
64       }
65       /* go to next state */
66       .
67       .
68       .
69    }
```

Line 16 is to open a channel for use in a Dialogic card. It is necessary to open the channel before any type of access of the same.

Line 50 is to play the voice file, which was pre-recorded with voice "Hello World". Pre-recorded voice files are recordings of normal voice and stored in digitized form. This can be done using normal telephone speaker and Dialogic card. However, in a majority of cases, this is done in a professional studio using professional people with a good voice.

In line 56 we read the digits entered through the telephone keypad.

3.6 VOICE XML

In mobile computing through telephone, the IVR is connected to the server through client/server architecture. It is also possible to host the IVR and the application on the same system. In the last few years, mobile computing through voice has come a long way. Today Internet (HTTP) is used in addition to client/server interface between the IVR and the server. This increases the flexibility in the whole mobile-computing architecture. HTTP is used for voice portals as well. In the case of a voice portal, a user uses an Internet site through voice interface. For all these advanced features, VoiceXML has been introduced. Recent IVRs are equipped with DSP (Digital Signal Processing) and are capable of recognizing voice. The output is synthesized voice through TTS (Text to Speech).

The Voice eXtensible Markup Language (VoiceXML) is an XML-based markup language for creating distributed voice applications. VoiceXML is designed for creating audio dialogs that feature synthesized speech, digitized audio, recognition of spoken voice and DTMF key input. Using VoiceXML, we can create Web-based voice applications that users can access through telephone.

VoiceXML supports dialogs that feature:

- Spoken input
- DTMF (telephone key) input
- Recording of spoken input
- Synthesized speech output (text-to-speech)
- Recorded audio output
- Dialog flow control
- Scoping of input

Architectural Model

The architectural model for VoiceXML is depicted in Figure 3.7. It has the following components:

A **Document Server** (e.g., a Web server) services requests from a client application.The client side of the application runs on a **VoiceXML Interpreter**, and is accessed through the **VoiceXML interpreter context**. The server delivers VoiceXML documents, which are processed by the VoiceXML Interpreter. The VoiceXML Interpreter Context is responsible for special actions on voice escape phrases.

For instance, in an interactive voice response application, the VoiceXML interpreter context may be responsible for detecting an incoming call, acquiring the initial VoiceXML document, and answering the call, while the VoiceXML interpreter manages the dialog after answer. The implementation platform generates events in response to user actions (e.g., spoken or character input received, disconnect) and system events (e.g., timer expiration).

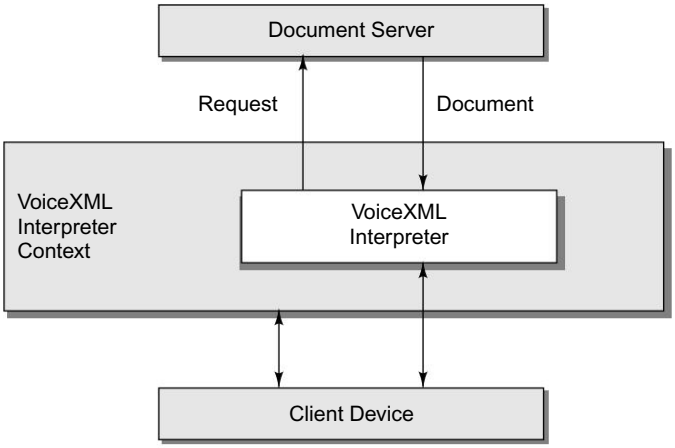


Figure 3.7 Voice XML Architectural Model

3.6.1 How Voice XML Fits into Web Environment

All of us are familiar with the web as it works today. We use a visual GUI web browser (such as Netscape Communicator or Internet Explorer), which renders and interprets HTTP requests to present information to the user (text, graphics, audio, multimedia, etc.). When the user makes a selection (for example, a click on a hyperlink), the web browser sends an HTTP request to the web server. The web server responds by locating the new page and returns the page to the user. The content server may also have to interact with a back-end infrastructure (database, servlets, etc.) to obtain and return the requested information.

The Voice Browser extends this paradigm. In Figure 3.8, a telephone and a Voice Server have been added to the web environment. The Voice Server manages several Voice Browser sessions. Each Voice Browser session includes one instance of the Voice Browser, the speech recognition engine, and the text-to-speech engine.

VoiceXML introduces a new way of presenting the web information. Instead of presenting the information visually (through HTML, graphics and text), the Voice Browser presents the information to the caller in audio using VoiceXML. When the caller says something (which is the voice equivalent of clicking on something to make a selection), the Voice Browser sends an HTTP request to the web server, which accesses the same back-end infrastructure, to return information this time in audio. This type of portal is known as voice portal. Voice portal is very useful in a hands-free situation like when you are driving.

The Voice Browser

An audio Voice Browser is similar to a visual web browser like Netscape Communicator or Microsoft Internet Explorer. Through voice browser, we interact with a web server using our voice and a telephone. Instead of rendering and interpreting a HTML document (like a GUI browser), the

Voice Browser renders and interprets VoiceXML documents. Instead of clicking a mouse and using keyboard, we use our voice and a telephone (and even the phone keypad) to access web information and services.

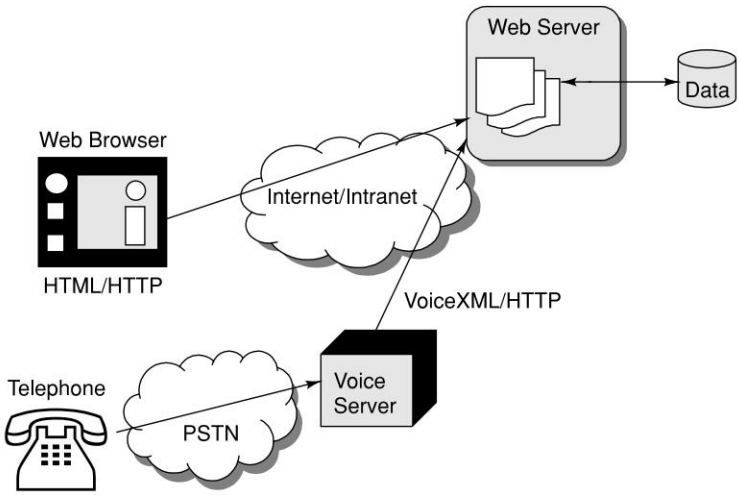


Figure 3.8 Voice Browser and Voice Portal over VoiceXML Architecture

Dialogs

A VoiceXML application defines a series of dialogs between a user and a computer. Each VoiceXML document forms a conversational finite state machine. The user is always in one conversational state, or dialog, at a time. Each dialog determines the next dialog to which to transition. Transitions are specified using URIs, which define the next document and dialog to use.

There are two types of dialogs that can be implemented in VoiceXML:

- forms
- menus

Forms define an interaction that collects values for a set of fields. Menus, on the other hand, present the user with choices or options and then transition to another dialog based on the choice.

Essential Elements of Voice XML Documents

The first line of any VoiceXML application must contain the <?xml> element. The second line must contain the <vxml> element. And each VoiceXML <tag>, must have an associated </tag>. The very last line of VoiceXML document must be the </vxml> tag. So, at a minimum, a VoiceXML document looks like this:

```
<?xml version="1.0"?>
<vxml version="1.0">
.
.
```

```
.
Interesting stuff goes here
.
.
.
</vxml>
```

In between the <vxml> and the </vxml> tags we put all the really interesting stuff, the VoiceXML code that defines the dialog with the user.

Prompts

In a VoiceXML application, we present information to the user through audio prompts. These prompts can either be prerecorded audio, or they can be synthesized speech (TTS). We use the <prompt> element in VoiceXML to generate TTS. Any text within the body of a <prompt> element is spoken. In the following example, the text-to-speech software will read 'Would you like coffee, tea, milk or nothing?' to the caller:

```
<prompt>
   Would you like coffee, tea, milk or nothing
</prompt>
```

Grammars

Each dialog has one or more speech and/or DTMF grammars associated with it.

In VoiceXML, we use the <grammar> element to define what the caller can say to the application at any given time. There are three different types of grammars supported in VoiceXML:

- Inline
- External
- Built-in

Inline grammars are those that are defined right in the VoiceXML code. For example,

```
<grammar>
   credit card | credit | tuition | tuition bill
</grammar>
```

In inline grammar, the words and phrases that a caller is allowed to say are defined within the body of the <grammar> element. Each word or phrase is separated by a vertical bar ("|") symbol. This symbol essentially means "or." So, in the previous example, the caller can say either "credit card" or "credit" or "tuition" or "tuition bill".

External grammars are those that are specified outside of the VoiceXML document in another file and are referenced from within the VoiceXML code. We use the `<grammar>` element to specify an external grammar, too. For example,

```
<grammar>src="names.gram"  type="application/x-jsgf"</grammar>
```

In this example, the grammar is defined in "names.gram" file.

Form

Form is one of the ways of developing a dialog with the caller in VoiceXML. Forms are central to VoiceXML. A VoiceXML form is a process to present information and gather input from the caller. A form is, basically, a collection of one or more fields that the caller fills in by saying something. A VoiceXML form is a very similar concept to a paper or online form, except that in the case of VoiceXML, we cannot see the field and instead of typing or writing in a field, we say something to fill it in.

In VoiceXML, we define a form using the `<form>` element and fields within the form using the `<field>` element. Here is a simple Voice Form.

```
<?xml version="1.0"?>
<vxml version="1.0">
<form id="add_funds">
   <field name="amount" type="currency">
      <prompt>How much?</prompt>
   </field>
   <field>
      <prompt>Charge to credit card or tuition
      bill?</prompt>
      <grammar> credit card | credit | tuition | tuition
      bill</grammar>
   </field>
</form>
</vxml>
```

This form has an ID of "add_funds", and it contains two fields. The first field asks the user how much money to add to the meal account "How much" and is expecting the user to say an amount as currency (e.g., "one thousand rupees"). The second field asks for the type of transaction ("charge to credit card or tuition bill") and is expecting the caller to say either "credit card", "credit", "tuition", or "tuition bill".

As we can see, fields define the information the application needs from the caller. Fields tell the caller what to say, and they also define the words and phrases that the caller can say (or the keys that can be pressed). Based on the caller's input—in other words, what the caller says or which keys were pressed—the application takes an appropriate action. When the user provides a valid response, the field is considered FILLED and the application can then do something with this information.

Events

VoiceXML provides a form-filling mechanism for handling "normal" user input. In addition, VoiceXML defines a mechanism for handling events not covered by the form mechanism. Events are thrown by the platform under a variety of circumstances, such as when the user does not respond, doesn't respond intelligibly, requests help, etc.

Links

A *link* supports mixed initiative. It specifies a grammar that is active whenever the user is in the scope of the link. If user input matches the link's grammar, control transfers to the link's destination URI. A <link> can be used to throw an event to go to a destination URI.

VoiceXML Elements

Element	Purpose Page
`<assign>`	Assign a variable a value.
`<audio>`	Play an audio clip within a prompt.
`<block>`	A container of (non-interactive) executable code.
`<break>`	JSML element to insert a pause in output.
`<catch>`	Catch an event.
`<choice>`	Define a menu item.
`<clear>`	Clear one or more form item variables.
`<disconnect>`	Disconnect a session.
`<div>`	JSML element to classify a region of text as a particular type.
`<dtmf>`	Specify a touch-tone key grammar.
`<else>`	Used in < if > elements.
`<elseif>`	Used in < if > elements.
`<emp>`	JSML element to change the emphasis of speech output.
`<enumerate>`	Shorthand for enumerating the choices in a menu.
`<error>`	Catch an error event.
`<exit>`	Exit a session.
`<field>`	Declares an input field in a form.
`<filled>`	An action executed when fields are filled.
`<form>`	A dialog for presenting information and collecting data.
`<goto>`	Go to another dialog in the same or different document.
`<grammar>`	Specify a speech recognition grammar.
`<help>`	Catch a help event.
`<if>`	Simple conditional logic.
`<initial>`	Declares initial logic upon entry into a (mixed-initiative) form.
`<link>`	Specify a transition common to all dialogs in the link's scope.
`<menu>`	A dialog for choosing amongst alternative destinations.
`<meta>`	Define a meta data item as a name/value pair.
`<noinput>`	Catch a no input event.
`<nomatch>`	Catch a no match event.
`<object>`	Interact with a custom extension.
`<option>`	Specify an option in a <field>.
`<param>`	Parameter in <object> or <subdialog>.
`<prompt>`	Queue TTS and audio output to the user.
`<property>`	Control implementation platform settings.
`<pros>`	JSML element to change the prosody of speech output.
`<record>`	Record an audio sample.

`<reprompt>`	Play a field prompt when a field is re-visited after an event.
`<return>`	Return from a subdialog.
`<sayas>`	JSML element to modify how a word or phrase is spoken.
`<script>`	Specify a block of ECMAScript client-side scripting logic.
`<subdialog>`	Invoke another dialog as a subdialog of the current.
`<submit>`	Submit values to a document server.
`<throw>`	Throw an event.
`<transfer>`	Transfer the caller to another destination.
`<value>`	Insert the value of a expression in a prompt.
`<var>`	Declare a variable.
`<vxml>`	Top-level element in each VoiceXML document.

3.7 Telephony Application Programming Interface (TAPI)

In the previous sections we have discussed how to program a Dialogic card and develop voice-based applications and services. However, there are quite a few higher level frameworks available where a developer can develop voice-based services without going too deep into it. TAPI (Telephony Application Programming Interface) is one such example. There is another related standard for speech called Speech Application Programming Interface (SAPI). Developed jointly by Intel and Microsoft, TAPI and SAPI are two standards that can be used when developing voice telephony applications. Using TAPI, programmers can take advantage of different telephone systems, including ordinary PSTN, ISDN, and PBX (Private Branch Exchange) without having to understand all their details. Use of these API will save the programmer the pain of trying to program hardware directly. Through TAPI and SAPI a program can "talk" over telephones or video phones to people or phone-connected resources. Through TAPI one will be able to:

- Use simple user interface to set up calls. This can be calling someone by clicking on their picture or other images.
- Use simple graphical interface to set up a conference call and then attend the call at the scheduled time.
- See who the user is talking to.
- Attach voice greeting with an email. This will allow the receiver to listen to this greeting while opening the email.
- Set groups and security measures such that a service can receive phone calls from certain numbers (but not from others).
- Send and receive faxes.
- Use same set of TAPI APIs which are available in many smart phones. This facilitates accessing telephony interfaces from a mobile phone and a desktop computer.

In addition to the interface for applications, TAPI includes an interface for convergence of both traditional PSTN telephony and IP telephony. IP telephony or VoIP (Voice over IP) is an emerging set of technologies that enables voice, data, and video collaboration over Internet protocol. VoIP is discussed in detail in Chapter 17.

3.8 Computer Supported Telecommunications Applications

ECMA-269 is a standard for CTI. It specifies application protocol data units (APDUs) for the services for Computer Supported Telecommunications Applications (CSTA) Phase III. The field of application of this standard is the interconnection of telephone switches and computers in a private telecommunication environment. This standard also provides a Protocol Implementation Conformance Statement (PICS) Proforma to assist implementers. It specifies interfaces and standards for interoperability for,

- Call Control features (making call, answering call, etc.).
- Call Associated features (sending user data, etc.).
- Logical Device features (do not disturb, forwarding, etc.).
- Physical Device features (writing to device display, etc.).
- Capability Exchange features (feature discovery, etc.).
- Snapshot features (query existing calls at a device, etc.).
- Monitor features (subscribing to event reports, etc.).
- Voice Services (for Listener, DTMF, Prompt and Message resources).
- Routing services.
- Media Attachment services.
- Maintenance services.
- Data Collection services.
- Accounting services, etc.

Standard ECMA-285 Protocol for Computer Supported Telecommunications Applications (CSTA) Phase III, 2nd edition (June 2000) can be downloaded from http://www.ecma-international.org/publications/standards/Ecma-285.htm

References/Further Reading

1. Aronsson's Telecom History Timeline, http://www.aronsson.se/hist.html.
2. Cole, Marion, (2001), *Introduction to Telecommunications Voice, Data, and the Internet*, Pearson Education Asia.
3. C. Y. Lee, (2000), William, *Mobile Cellular Telecommunications Analogue and Digital Systems,* McGraw-Hill.
4. Dialogic card references: http://www.intel.com.
5. Enterprise Computer Telephony Forum (ECTF): http://www.ectf.org.
6. Fundamentals of Telecommunications, The International Engineering Consortium, http://www.iec.org.
7. IBM WebSphere Voice Server, Software Developers Kit (SDK), Programmer's Guide, 2000.
8. IBM, WebSphere Voice Server for DirectTalk, User's Guide, 2001.
9. IEEE Communication Society, (2002), *A Brief History of Communication.*
10. Leonard, Regis, *Satellite Communications: A Short Course*, http://ctd.grc.nasa.gov/rleonard/regs1i.html
11. *Results on the Application of the PRMA Protocol in Low Earth Orbit Mobile Communications Systems* (http://marconi.ltt.dii.unisi.it/~giambene/short_prma.html)

12. Schiller Jochen, (2001), *Mobile Communications*, Pearson Education Asia.
13. TAPI: http://www.microsoft.com.
14. *Voice-enabled e-Business*: *Unlocking e-Business Opportunities*, Intel Corporation.
15. Voice XML Forum: http://www.voicexml.org.
16. Wikipedia–www.wikipedia.org

REVIEW QUESTIONS

Q1: Describe what is multiple access? Why is multiple access important?
Q2: Describe FDMA, TDMA, CDMA and SDMA with their application areas and examples.
Q3: What are the steps you need to follow during the design of an application development using voice?
Q4: You have been asked by a bank to develop an account enquiry system over telephone. Design the architecture of such a system.
Q5: Draw a call flow diagram for authentication of a user in a bank using CTI.
Q6: Describe the architectural model of Voice XML. How does VXML fit into the web environment?
Q7: Design a hands-free voice based email client application for the sales persons of the company.
Q8: Write short notes on:
(a) IVRS
(b) Voice APIs
(c) VXML elements
(d) TAPI
Q9: What is TDMA? Explain with the help of a diagram.
Q10: Explain Dynamic and Fixed TDMA systems.
Q11: Explain DPRMA.
Q12: What is PRMA HS?
Q13: How does communication take place through satellites? Explain in detail.
Q14: What is LEO? How is it different from MEO? Analyze the usage of LEO for a firm having multiple offices across more than three continents.
Q15: Discuss ways of establishing satellites in Earth's orbit and respective power consumption issues.
Q16: Write short notes on :
(a) History of Satellite Communication Systems
(b) Iridium
(c) Globalstar
(d) Thuraya

CHAPTER 4

Emerging Technologies

4.1 INTRODUCTION

The mainstream wireless technologies have been discussed in Chapters 5 through 10. We will discuss wireless networks and application development using cellular networks like GSM, SMS, GPRS, WAP, CDMA, and 3G in Chapters 5 through 9. We will also discuss wireless local area networks (WLAN or WiFi) in Chapter 10. In this chapter, however, will we discuss some technologies which are not yet in the mainstream but are potential candidates for the same. These technologies are included here to make the mobile computing story complete. These include technologies like Bluetooth (802.15.1a), Radio frequency identifier (RFID), Wireless metropolitan area network or wireless broadband (WiMax-802.16), Mobile IP, IPv6, and Java Card. Bluetooth is a technology in the personal area network (PAN). RFID is emerging as a leading technology in the logistics, manufacturing, and retail industry. Wireless broadband is expected to be a mainstream technology very soon. Mobile IP allows data hand-off over different sub-networks. IPv6 is the next generation Internet protocol. Java Card technology is emerging as a forerunner in the security and personal identity domain. Therefore, we introduce all these technologies in this chapter.

4.2 BLUETOOTH

Bluetooth was the nickname of a Danish king Harald Blåtand, who unified Denmark and Norway in the 10th century. The concept behind Bluetooth wireless technology was unifying the telecom and computing industries. Bluetooth technology allows users to make ad hoc wireless connections between devices like mobile phones, desktop or notebook computers without any cable. Devices carrying Bluetooth-enabled chips can easily transfer data at a speed of about 1 Mbps in basic mode within a 50 m (150 feet) range or beyond through walls, clothing and even luggage bags.

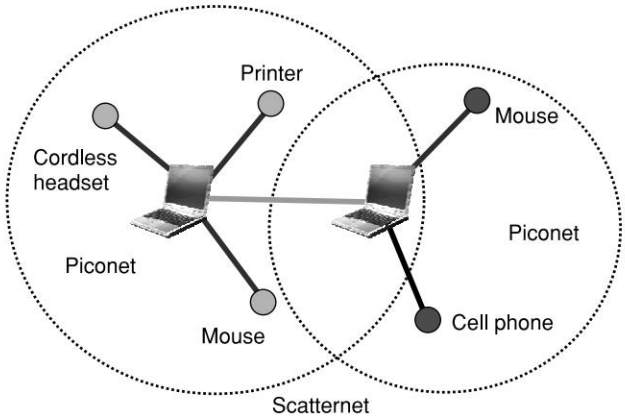


Figure 4.1 Bluetooth Scatternet as a Combination of Piconets

4.2.1 Bluetooth Protocol

The Bluetooth radio is built into a small microchip and operates in a globally available frequency band ensuring interoperability worldwide. Bluetooth uses the unlicensed 2.4 GHz ISM (Industrial Scientific and Medical) frequency band. There are 79 available Bluetooth channels spaced 1 MHz apart from 2.402 GHz to 2.480 GHz. The Bluetooth standard is managed and maintained by Bluetooth Special Interest Group (www.bluetooth. com). IEEE has also adapted Bluetooth as the 802.15.1a standard. Bluetooth allows power levels starting from 1mW covering 10 cm to 100 mW covering upto 100 meters. These power levels are suitable for short device zone to personal area networks within a home. Bluetooth supports both unicast (point-to-point) and multicast (point-to-multipoint) connections. Bluetooth protocol uses the concept of master and slave. In a master-slave protocol a device cannot talk as and when they desire. They need to wait till the time the master allows them to talk. The master and slaves together form a piconet. Up to seven "slave" devices can be set to communicate with a "master". Several of these *piconets* can be linked together to form a larger network in an ad hoc manner. The topology can be thought as a flexible, multiple piconet structure. This network of piconets is called *scatternet* (Fig. 4.1). A scatternet is formed when a device from one piconet also acts as a member of another piconet. In this scheme, a device being master in one piconet can simultaneously be a slave in the other one.

Bluetooth protocol is a combination of different protocols. The Bluetooth Core protocols plus the Bluetooth radio protocols are required by most of the Bluetooth devices, while the rest of the protocols are used by different applications as needed. At the physical layer Bluetooth uses spread spectrum technologies. It uses both direct sequence and frequency hopping spread spectrum technologies. Bluetooth uses connectionless (ACL–Asynchronous Connectionless Link) and connection-oriented (SCO–Synchronous Connection-oriented Link) links. Together, the Cable Replacement layer, the Telephony Control layer, and the Adopted protocol layer form application-oriented protocols enabling applications to run over the Bluetooth Core protocols.

4.2.2 Bluetooth Protocol Stack

Bluetooth protocol stack can be thought of as a combination of multiple application specific stacks as depicted in Figure 4.2. Different applications run over one or more vertical slices from this protocol stack. These are RFCOMM (Radio Frequency COMMunication), TCS Binary (Telephony Control Specification), and SDP (Service Discovery Protocol). Each application environmente uses a common data link and physical layer. RFCOMM and the TCS binary (Telephony Control Specification) protocol are based on the ETSI TS 07.10 and the ITU-T Recommendation Q.931 respectively. Some applications have some relationship with other protocols, e.g., L2CAP (Logical Link Control and Adaptation Protocol) or TCS may use LMP (Link Manager Protocol) to control the link manager.

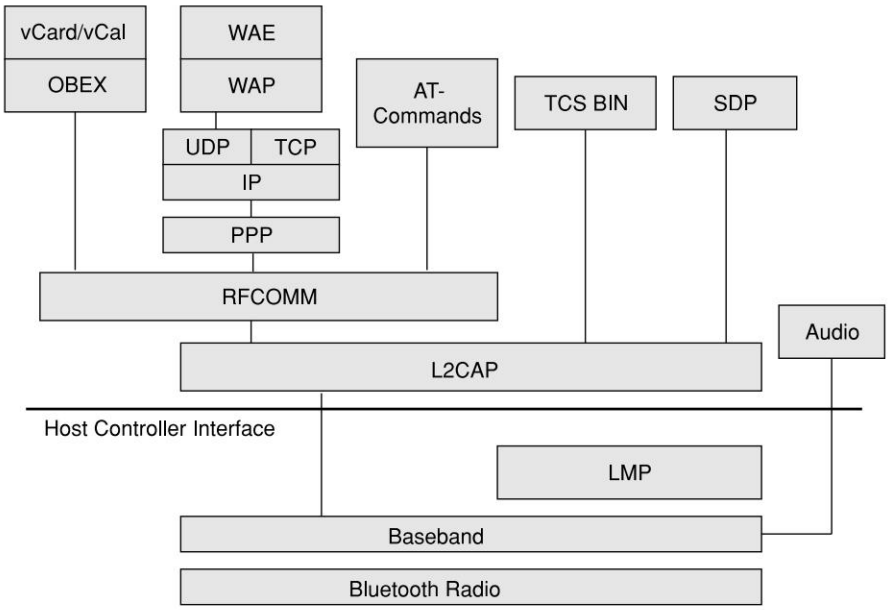


Figure 4.2 Bluetooth Protocol Stack

Bluetooth protocol stack can be divided into four basic layers according to their functions. These are:

- *Bluetooth Core Protocols*: This comprises Baseband, Link Manager Protocol (LMP), Logical Link Control and Adaptation Protocol (L2CAP), and Service Discovery Protocol (SDP).
  - **Baseband:** The Baseband and Link Control layer enables the physical RF link between Bluetooth units forming a piconet. This layer uses inquiry and paging procedures to synchronize the transmission with different Bluetooth devices. Using SCO and ACL link different packets can be multiplexed over the same RF link. ACL packets are used for data only, while the SCO packet can contain audio only or a combination of audio and data. All audio and data packets can be provided with different levels of CRC (Cyclic Redundancy Code) or FEC (Forward Error Correction) for error detection/correction.

- **Link Manager Protocol (LMP):** When two Bluetooth devices come within each other's radio range, link managers of either device discover each other. LMP then engages itself in peer-to-peer message exchange. These messages perform various security functions starting from authentication to encryption. LMP layer performs generation and exchange of encryption keys as well. This layer performs the link setup and negotiation of baseband packet size. LMP also controls the power modes, connection state, and duty cycles of Bluetooth devices in a piconet.
- **Logical Link Control and Adaptation Protocol (L2CAP):** This layer is responsible for segmentation of large packets and the reassembly of fragmented packets. L2CAP is also responsible for multiplexing of Bluetooth packets from different applications.
- **Service Discovery Protocol (SDP):** The Service Discovery Protocol (SDP) enables a Bluetooth device to join a piconet. Using SDP a device inquires what services are available in a piconet and how to access them. SDP uses a client-server model where the server has a list of services defined through service records. One service record in a server describes the characteristics of one service. In a Bluetooth device there can be only one SDP server. If a device provides multiple services, one SDP server acts on behalf of all of them. Similarly, multiple applications in a device may use a single SDP client to query servers for service records. A Bluetooth device in an inquiry mode broadcasts ID packets on 32 frequency channels of the Inquiry Hopping Sequence. It sends two ID packets every 625 μs and then listens for responses the following 625 μs. At this stage the unique identity of the devices called Bluetooth globalID is exchanged. A globalID indicates a device's profile along with capability functions. Upon matching of the device profile a connection is set up and devices exchange data. When a connection is set up, the paging device becomes the master and the paged device becomes the slave. A Bluetooth device may operate both as a server and as a client at the same time forming a scatternet. They can also switch from master to slave and vice versa. The master slave switch can take between 4:375 and 41:875 ms. In a piconet, a master device can be a laptop or PDA, while slaves' devices could be printers, mouse, cellular phones, etc.

- *Cable Replacement Protocol*: This protocol stack has only one member, viz., Radio Frequency Communication (RFCOMM).
  - RFCOMM is a serial line communication protocol and is based on ETSI 07.10 specification. The "cable replacement" protocol emulates RS-232 control and data signals over Bluetooth baseband protocol.
- *Telephony Control Protocol*: This comprises two protocol stacks, viz., Telephony Control Specification Binary (TCS BIN), and the AT-Commands.
  - **Telephony Control Protocol Binary:** TCS Binary or TCS BIN is a bit-oriented protocol. TCS BIN defines the call control signaling protocol for set up of speech and data calls between Bluetooth devices. It also defines mobility management procedures for handling groups of Bluetooth TCS devices. TCS Binary is based on the ITU-T Recommendation Q.931.
  - **AT-Commands:** This protocol defines a set of AT-commands by which a mobile phone can be used and controlled as a modem for fax and data transfers. AT (short form of attention) commands are used from a computer or DTE (Data Terminal Equipment) to control a modem or DCE (Data Circuit terminating Equipment). AT-commands in Bluetooth are based on ITU-T Recommendation V.250 and GSM 07.07.

- *Adopted Protocols*: This has many protocol stacks like Point-to-Point Protocol (PPP), TCP/IP Protocol, OBEX (Object Exchange Protocol), Wireless Application Protocol (WAP), vCard, vCalendar, Infrared Mobile Communication (IrMC), etc.
  - **PPP Bluetooth:** This offers PPP over RFCOMM to accomplish point-to-point connections. Point-to-Point Protocol is the means of taking IP packets to/from the PPP layer and placing them onto the LAN.
  - **TCP/IP:** This protocol is used for communication across the Internet. TCP/IP stacks are used in numerous devices including printers, handheld computers, and mobile handsets. Access to these protocols is operating system independent, although traditionally realized using a socket programming interface model. TCP/IP/PPP is used for the all Internet Bridge usage scenarios. UDP/IP/PPP is also available as transport for WAP.
  - **OBEX Protocol:** OBEX is a session protocol developed by the Infrared Data Association (IrDA) to exchange objects. OBEX, provides the functionality of HTTP in a much lighter fashion. The OBEX protocol defines a folderlisting object, which can be used to browse the contents of folders on remote devices.
  - **Content Formats:** vCard and vCalendar specifications define the format of an electronic business card and personal calendar entries developed by the Versit consortium, These are now maintained by the Internet Mail Consortium. Other content formats, supported by OBEX, are vMessage and vNote. These content formats are used to exchange messages and notes. They are defined in the IrMC (IrDA Mobile Communication) specification. IrMC also defines a format for synchronization of data between devices.

4.2.3 Bluetooth Security

In a wireless environment where every bit is on the air, security concerns are high. Bluetooth offers security infrastructure starting from authentication, key exchange, to encryption. In addition to encryption, a frequency-hopping scheme with 1600 hops/sec is employed. All of this make the system difficult to eavesdrop. At the lowest levels of the protocol stack, Bluetooth uses the publicly available cipher algorithm known as SAFER+ to authenticate a device's identity. In addition to these basic security functions, different application verticals use their own security infrastructure at the application layer.

4.2.4 Bluetooth Application Models

Each application model in Bluetooth is realized through a profile. Profiles define the protocols and protocol features supporting a particular usage model.

- *File Transfer*: The file transfer usage model offers the ability to transfer data objects from one device (e.g., PC, smart-phone, or PDA) to another. Object types include .xls, .ppt, .wav, .jpg, .doc files, folders or directories or streaming media formats. Also, this model offers a possibility to browse the contents of the folders on a remote device.
- *Internet Bridge*: In this usage model, a mobile phone or cordless modem acts as modem to the PC, providing dial-up networking and fax capabilities without need for physical connection to the PC.

- *LAN Access*: In this usage model multiple data terminals use a LAN access point (LAP) as a wireless connection to an Ethernet LAN. Once connected, the terminals operate as if they were connected directly to the LAN.
- *Synchronization*: The synchronization usage model provides a device-to-device (phone, PDA, computer, etc.) synchronization of data. Examples could be PIM (personal information management) information, typically phonebook, calendar, message, and note information.
- *Headset*: The headset can be wirelessly connected for the purpose of acting as a remote device's audio input and output interface. This is very convenient for hands-free cellular phone usage in automobiles.

4.3 Radio Frequency Identification (RFID)

RFID is a radio transponder carrying an ID (Identification) that can be read through radio frequency (RF) interfaces. These transponders are commonly known as RFID tags or simply tags. To assign an identity to an object, a tag is attached to the object. Data within the tag provides identification for the object. The object could be an entity in a manufacturing shop, goods in transit, item in a retail store, a vehicle in a parking lot, a pet, or a book in a library. Biologists had been using RFID for sometime to track animals for the purpose of studying animal behavior and conservation. The earliest use of RFID was for tracking farm animals. A RFID system comprises different functional areas like:

1. Means of reading or interrogating the data in the tag.
2. Mechanism to filter some of the data.
3. Means to communicate the data in the tag with a host computer.
4. Means for updating or entering customized data into the tag.

RFID tags are categorized on three basic criteria. These are based on frequency, application area and the power level.

- *On Frequency*: There are six basic frequencies on which RFID operates. These are 132.4 KHz, 13.56 MHz, 433 MHz, 918 MHz, 2.4 GHz and 5.8 GHz. Low frequency (30 KHz to 500 KHz) systems have short reading ranges and lower system costs. Tags in this frequency range are slow in data transfer and suitable for slow-moving objects. They are most commonly used in security access, asset tracking and animal identification applications. High-frequency (850 MHz to 950 MHz and 2.4 GHz to 2.5 GHz) systems offer long read ranges and high data transfer speeds. High reading speed is required for fast moving objects like railway wagon tracking and identification of vehicles on freeways for automated toll collection. The higher the frequency the higher the data transfer rates.
- *On Application*: RFIDs are also grouped according to application and usage. Speed of the object and distance to read determines the type of tag to be used. RFID used for livestock will be different from the tag used in railroad. The significant advantage of all types of RFID systems is the contactless, nonline-of-sight nature of the technology. Tags can be read through a variety of substances such as snow, fog, paint, plastic, wall, container and other challenging conditions, where barcodes or other optical means of reading are not effective. RFID tags can also be read at high speeds. In these cases RFIDs can respond within 100 milliseconds. A RFID tag contains two segments of memory. One segment is a factory-set and used to uniquely

identify a tag. The other segment is usable by the application. Application specific data can be written or stored in this portion of the tag. The read/write capability of a RFID system is an advantage in interactive applications such as work-in-process or maintenance tracking. Compared to the barcode, RFID is a costlier technology. However, RFID has become indispensable for a wide range of automated data collection and identification applications that would not be possible otherwise.

- *Power-based Grouping*: RFIDs can be grouped into two types based on power requirements. These are active and passive tags. Passive tags are generally in low frequency range, whereas tags at higher frequency range can be either active or passive.
- *Active RFID Tags*: Active tags are powered by an internal battery and are typically read/write. The life of an active tag is limited by the life of the battery. The data within an active tag can be rewritten or modified. An active tag's memory can vary from a few bytes to 1MB. The battery-supplied power of an active tag generally gives it a longer read range. The trade off is, greater the size greater the cost, and a limited operational life. Depending upon the battery type and temperatures, the life of such tags could be 10 years. Some active tags can also be smart and do not send their information all the time. In a typical read/write RFID system, a tag might give a machine a set of instructions, and the machine would then report its performance to the tag. This encoded data would then become part of the tagged part's history. This data can be details about the port of transit with dates.
- *Passive RFID Tags*: Passive tags operate without a power source of its own. A passive tag obtains operating power from the reader's antenna. The data within a passive tag is read only and generally cannot be changed during operation. Passive tags are lighter, less expensive and offer a virtually unlimited operational lifetime. The trade off is that they have shorter read ranges than active tags and require a higher-powered reader. Passive tags contain data usually 32 to 128 bits long.

RFID tags are of different shapes and sizes. Animal tracking tags are inserted beneath the skin and are as small as a pencil lead. Tags can be screw-shaped to identify trees or wooden logs. In stores, plastic tags are attached to merchandise and used as anti-theft devices. Heavy-duty large tags are used to track containers or heavy machinery. The reader emits radio waves in any range from one centimeter to 25 meters or more. When an RFID tag passes through the electromagnetic zone of the reader, it detects the reader's activation signal. The reader decodes the data encoded in the tag's integrated circuit and the data is passed to the host computer for processing. A basic RFID system consist of three components:

- A transponder programmed with unique information (RFID tag).
- A transceiver with decoder (a reader).
- An antenna or coil.

The antenna emits radio signals to read data from or write data into the tag. Antennas control data acquisition and communication. An antenna is fitted with the transceiver to become a reader. Close proximity passive tags rely on electromagnetic or inductive coupling techniques (Fig. 4.3a). Whereas, active tags are based upon propagating electromagnetic waves techniques (Fig. 4.3b). Coupling is via "antenna" structures forming an integral feature in both tags and readers. While the term antenna is generally considered more appropriate for propagating systems, it is also loosely applied to inductive systems.

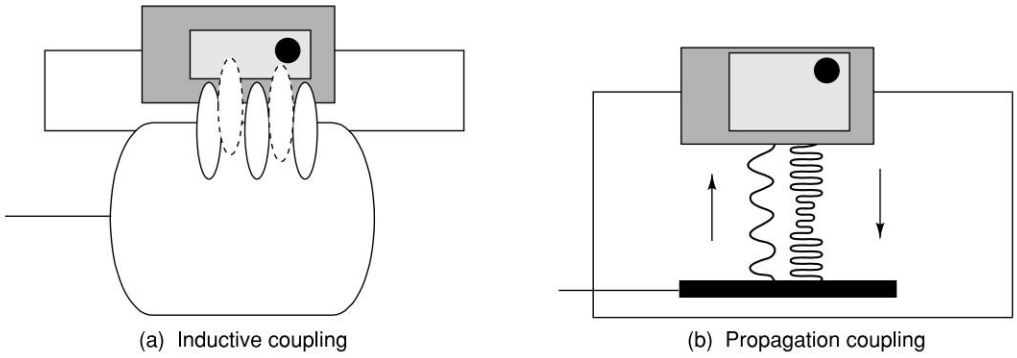


Figure 4.3 Passive and Active RFID

4.3.1 Areas of Application for RFID

Potential applications for RFID may be identified in virtually every sector of industry, commerce and services where data is to be collected. The attributes of RFID are complementary to other data-capture technologies and therefore able to satisfy particular application requirements that cannot be adequately accommodated by alternative technologies. Principal areas of application for RFID that can be currently identified include:

- Transportation and logistics.
- Manufacturing and processing.
- Security.
- Animal tagging.
- Store in an enterprise.
- Retail store.
- Community library.
- Time and attendance.
- Postal tracking.
- Airline baggage reconciliation.
- Road toll management.

The lack of standards has been a deterrent for the growth of the RFID industry. Standards are essential for interoperability and growth of a technology. Standardization helped GSM and barcode technology to be widely accepted. Therefore, a number of organizations in US and Europe are working to address this issue. ANSI's X3T6 group is currently developing a draft document-based systems' operation at a carrier frequency of 2.45 GHz. ISO has already adopted international RFID standards for animal tracking, and they are ISO 11784 and 11785.

4.4 Wireless Broadband (WiMAX)

Wireless technologies are proliferating in a major way into the first-mile (as computer people call it) or last-mile (as communication people call it) subscriber access, as opposed to twisted-pair local loop. These technologies are generally referred to as (WLL–wireless local loop) or WiLL (wireless in local loop). Wireless local loop is also known as fixed-wireless system. The world is moving towards

a convergence of voice, data and video. This convergence will demand interoperability and high data rate. Keeping this in mind, the IEEE 802 committee set up the 802.16 working group in 1999 to develop wireless broadband or WirelessMAN (wireless metropolitan area network) standards. WirelessMAN offers an alternative to high bandwidth wired access networks like fiber optic, cable modems and DSL (Digital Subscriber Line). WirelessMAN is popularly known as WiMAX (Worldwide Interoperability for Microwave Access). WiMAX provides wireless transmission of data using a variety of transmission modes, from point-to-multipoint links to portable and fully mobile Internet access. This technology provides up to 10 Mbps bandwidth without the need for cables. Figure 4.4 illustrates the WiMAX Architecture; whereas, Fig. 4.5 illustrates a typical. WirelessMAN deployment scenario.

The release of WirelessMAN (IEEE 802.16) standards in April 2002 has paved the way for the entry of broadband wireless access as a new bearer to link homes and businesses with core telecommunications networks. WirelessMAN provides network access to buildings through exterior antennas communicating with radio base stations. The technology is expected to provide less expensive access with more ubiquitous broadband access with integrated data, voice and video services. One of the most attractive aspects of wireless broadband technology is that networks can be created in just weeks by deploying a small number of base stations on buildings or poles to create high-capacity wireless access systems. In a wired set up, one physical wire will connect the device with the network. Also, we need to keep many wires reserved for future growth. Therefore, the initial investment in wired infrastructure is very high. Wireless network can grow as the demand increases. At any point in time the number of active users are always a fraction of the number of subscribers. In a wireless environment the number of channels is always low compared to the number of subscribers. This makes wireless technologies very attractive to the service providers.

IEEE 802.16 standardizes the air interface and related functions associated with WLL. Three working groups have been chartered to produce the following standards:

- IEEE 802.16.1–Air interface for 10 to 66 GHz.
- IEEE 802.16.2–Coexistence of broadband wireless access systems.
- IEEE 802.16.3–Air interface for licensed frequencies, 2 to 11 GHz.
- Extensive radio spectrum is available in frequency bands from 10 to 66 GHz worldwide. In a business scenario, 802.16 can serve as a backbone for 802.11 networks. Other possibilities are using 802.16 within the enterprise along with 802.11a, 802.11b or 802.11g.

IEEE 802.16 standards are concerned with the air interface between a subscriber's transceiver station and a base transceiver station. The 802.16 standards are organized into a three-layer architecture.

- The physical layer: This layer specifies the frequency band, the modulation scheme, error-correction techniques, synchronization between transmitter and receiver, data rate and the multiplexing structure.
- The MAC (Media Access Control) layer: This layer is responsible for transmitting data in frames and controlling access to the shared wireless medium through media access control (MAC) layer. The MAC protocol defines how and when a base station or subscriber station may initiate transmission on the channel.
- Above the MAC layer is a convergence layer that provides functions specific to the service being provided. For IEEE 802.16.1, bearer services include digital audio/video multicast, digital telephony, ATM, Internet access, wireless trunks in telephone networks and frame relay.

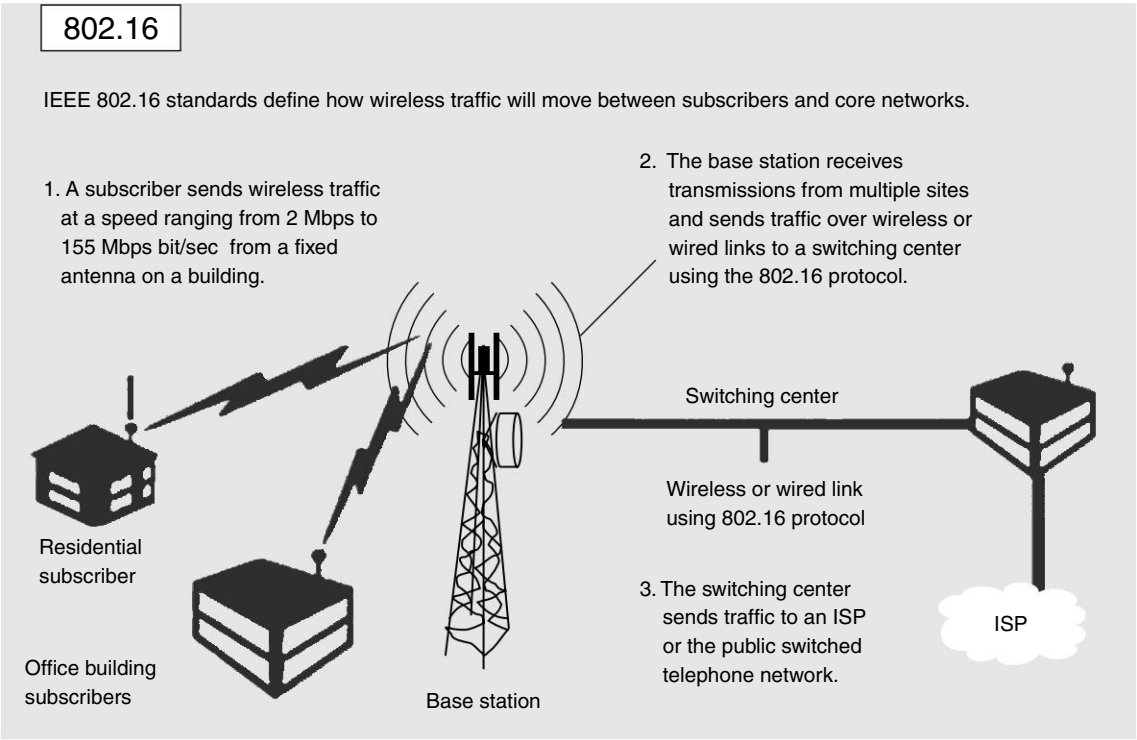


Figure 4.4(a) WiMAX (Wireless MAN) Deployment Architecture

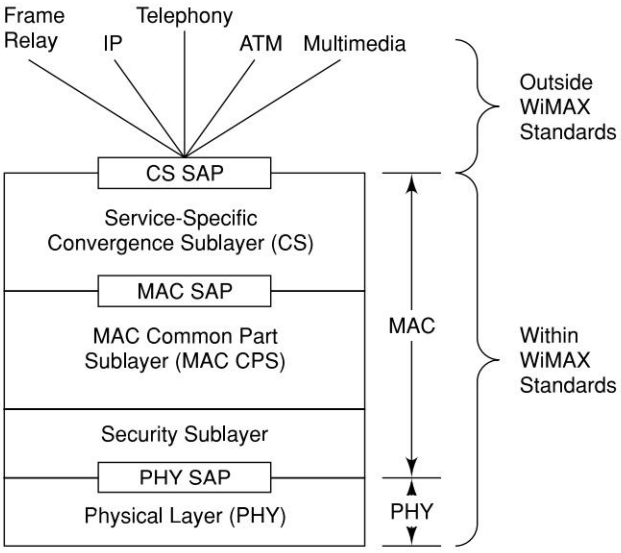


Figure 4.4(b) WiMAX Protocol Stack

4.4.1 Physical Layer

To support duplexing, 802.16 adapted a burst design that allows both time-division duplexing (TDD) and frequency-division duplexing (FDD). In TDD the uplink and downlink share a channel but do not transmit simultaneously. In the case of FDD the uplink and downlink operate on separate channels and sometimes simultaneously. Support for half-duplex FDD subscriber stations is also supported in 802.16. Both TDD and FDD alternatives support adaptive burst profiles in which modulation and coding options may be dynamically assigned on a burst-by-burst basis.

The 2–11 GHz bands, both licensed and unlicensed, are used in 802.16. Design of the 2–11 GHz physical layer is driven by the need for non-line-of-sight operation. The draft currently specifies that compliant systems implement one of three air interface specifications, each of which provides for interoperability. The 802.16 standard specifies three physical layers for services:

- WirelessMAN-SC2: This uses a single-carrier modulation format. This is to support existing networks and protocols.
- WirelessMAN-OFDM: This uses orthogonal frequency-division multiplexing with a 256-point transform. Access is by TDMA. This air interface is mandatory for license-exempt bands.
- WirelessMAN-OFDMA: This uses orthogonal frequency-division multiple access with a 2048-point transform. In this system, multiple access is provided by addressing a sub-set of the multiple carriers to individual receivers.

4.4.2 802.16 Medium Access Control

The IEEE 802.16 MAC protocol was designed for point-to-multipoint broadband wireless access. It addresses the need for very high bit rates, both uplink (to the base station) and downlink (from the base station). To support, a variety of services like multimedia and voice, the 802.16 MAC is equipped to accommodate both continuous and bursty traffic. To facilitate the more demanding physical environment and different service requirements of the frequencies between 2 and 11 GHz, the 802.16 project is upgrading the MAC to provide automatic repeat request (ARQ) and support for mesh, rather than only point-to-multipoint, network architectures.

4.4.3 Broadband Applications

Wireless broadband allows higher data rates in homes, offices, and even mobile environment. Therefore, all the user applications in home and offices are potential candidates for wireless broadband. These include standard Ethernet LAN or WiFi indoor using 802.16d and outdoor mobile using 802.16e. The IEEE 802.16 Broadband Wireless Metropolitan Area Network Standards can be found at IEEE site (http://standards.ieee.org/getieee802/802.16.html).

Along with the existing applications a new brand of applications are also being thought about. One such system is mobile cellular system.

4.4.4 Broadband Mobile Cellular System

During different discussions on systems and architecture of mobile computing, we talked about mobility with the network being static. In mobile cellular system the cellular network itself will be mobile. A cellular system like 3G can provide high data rate. WirelessMAN is also geared up to support high

data rate. However, these high data rates are possible with low speed mobility. Scientists are now thinking in terms of high-speed mobility specially designed for high-speed telematics application.

Figure 4.5 depicts one such mobile communication system to support high-speed mobility. This is achieved by installing moving base stations and fixed radio ports uniformly distributed along the median of the roadway. The moving base stations allow communication links to be established between the mobile units traveling on the roadway and a fixed communication network through the fixed radio ports. The small-cell (picocell) architecture of the proposed system enables the use of extremely lightweight low-power mobile units that can be used almost anywhere. In this architecture the picocell will move in the direction of the moving vehicle so that the relative speed between them is low. This proposed infrastructure is suitable for high-speed multilane highways in cities. The proposed system will be able to communicate to devices traveling at speeds up to and in excess of 150 kmph.

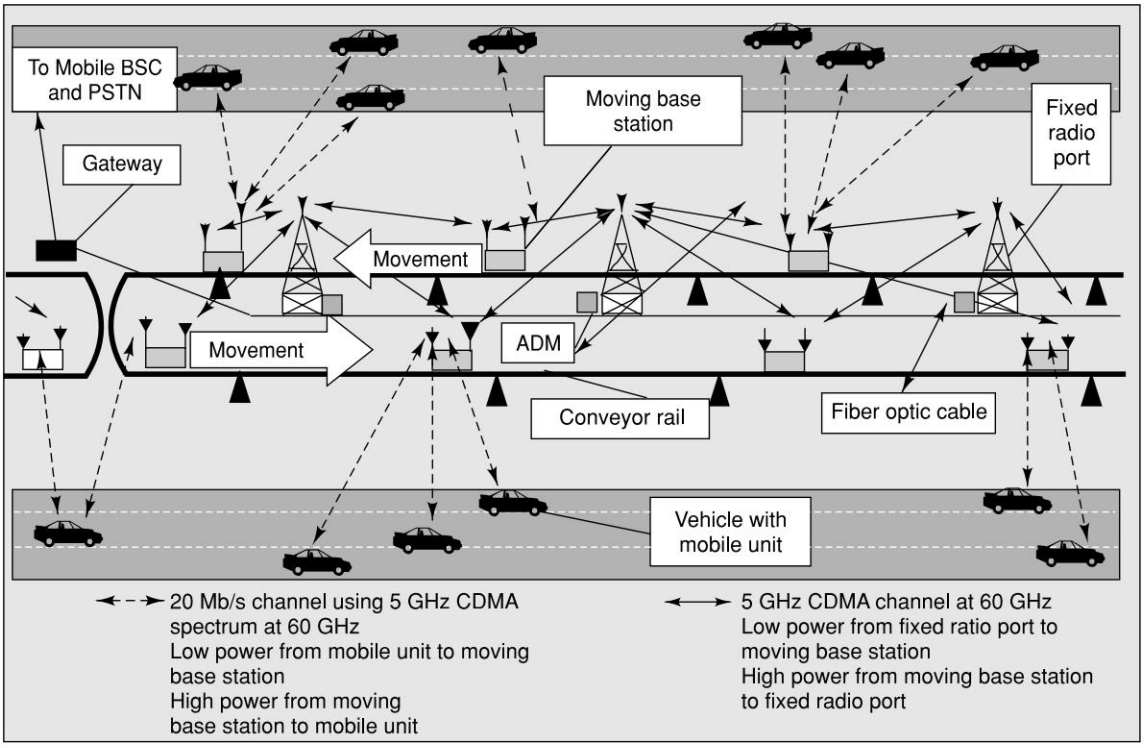


Figure 4.5 Mobile Broadband Communication System with Moving BTS

4.5 Mobile IP

Mobile computing in the true sense will be able to provide an environment where a user will be able to continuously access data and services in a state of mobility. Mobile computing should not be confused with portable computing. In a portable computing environment, we move with the computing device from one location to another and use the network while stationary. For example,

while I am in office with my laptop computer, I use the company Ethernet LAN; and, when I am back home, I use the broadband at home. In this portable computing environment I use the network only when stationary and disconnect from one network before movement. Mobile computing on the other hand offers seamless computing and data networking facility even if the user is in a state of mobility and changes the network. Mobility Management (MM) deals with a situation where the user is at a vehicular state and accessing the network. Vehicular state generally means moving at a speed 60 kmph or higher. We will discuss the mobility management for voice network in Chapter 5. Here in Mobile IP, we will discuss the mobility management in TCP/IP data networks; Mobile IP standards are specified in RFC3344.

A data connection between two end-points through TCP/IP network requires a source IP address, source TCP port and a target IP address with a target TCP port. The combination of the IP address of the node (client or server device) system combined with the TCP port as the identification of a service becomes a point of attachment for an end-point. TCP port number is application-specific and remains constant. IP address, on the other hand, is network-specific and varies from network to network. IP addresses are assigned to a node from a set of addresses assigned to a network. This structure works well as long as the client is static and is using a desktop computer where the point of attachment is fixed. Let us assume that the user is mobile and is using a laptop with WiFi. As the user moves, the point of attachment will change from one subnet to another resulting in a change of IP address. This will force the connection to terminate. Therefore, the question is how do we allow mobility while a data connection is alive. The technology to do so is "Mobile IP". The term "mobile" in "Mobile IP" signifies that, while a user is connected to applications across the Internet and the user's point of attachment changes dynamically, all connections are maintained despite the change in underlying network properties including the point of attachments. This is similar to the handoff/roaming scenario in cellular networks. In a cellular network, when a user is mobile, the point of attachment (base station) changes. However, in spite of such changes the user is able to continue the conversation without any break in service.

4.5.1 How does Mobile IP Work?

IP routes packets from a source endpoint to a destination endpoint through various routers. An IP address of a node can be considered to be a combination of network address (most significant 24 bits) and the node address (least significant 8 bits). Let us assume a "C" class IP address 75.126.113.230 to be the mail server of Geschickten (mail.geschickten.com). We can assume that the first 24 bits 75.126.113 is the address of the network and the last 8 bits containing 230 is the address of the node. The network portion of an IP address is used by routers to deliver the packet to the last router in the chain to which the target computer is attached. This last router then uses the host portion (230 in this example) of the IP address to deliver the IP packet to the destination computer. In addition to the IP addresses of the nodes, for meaningful communication we need the TCP or UDP (User Datagram Protocal) port of the applications. The port number is used by the host to deliver the packet to the appropriate application.

A TCP connection is identified by a quadruplet that contains the IP address and port number of the sender endpoint along with the IP address and port number of the receiving endpoint. To ensure that an active TCP connection is not terminated while the user is mobile, it is essential that all of these four identities remain constant—physically or virtually. The TCP ports are application specific and generally constant—they do not change after an end-to-end connection is established.

However, the IP address will change when a node moves from one subnet to another. Therefore, to fix this problem mobile IP allows the mobile node to use two IP addresses. These IP addresses are called home address and care-of address. Home address is the original static IP address of the node and known to everybody as the identity of the node. The care-of address changes at each new point of attachment and can be thought of as the mobile node's location specific address. These are similar to MSISDN (Mobile Station ISDN) number and the MSRN (Mobile Station Roaming Number) respectively as in GSM network (see Chapter 5).

In addition to home address and care-of address there are two network elements in Mobile IP that play a very significant role in routing of the packets as part of mobility management; these are home agent and foreign agent. A home agent is a router on a mobile node's home network which forwards datagrams for delivery to the mobile node through a tunnel when it is away from home. The home agent also maintains current location information of the mobile node. In contrast, a foreign agent is a router on a mobile node's visited network which provides routing services to the mobile node while registered. The foreign agent detunnels and delivers datagrams to the mobile node that were tunneled by the mobile node's home agent. For datagrams sent by a mobile node, the foreign agent may serve as a default router for registered mobile nodes. This is similar to the concept of HLR (Home Location Register) and VLR (Visitor Location Register) in cellular networks (Chapter 5).

When the mobile node is located on its home network, it operates without mobility services. When the mobile node detects that it has moved to a foreign network, it registers with the foreign agent and obtains a care-of address on the foreign network. The care-of address can either be determined from a foreign agent's advertisements, or by some external assignment mechanism such as DHCP. The mobile node registers its new care-of address with its home agent informing its new location and new care-of address. The home agent forwards all incoming data packet to the foreign network using the care-of address. The delivery requires that the packet header is modified so that the care-of address becomes the destination IP address. This new header (Fig. 4.8) encapsulates the original packet, causing the mobile node's home address to have no impact on the encapsulated packet's routing. Figure 4.6 shows in general terms how Mobile IP deals with the problem of dynamic IP addresses. On returning to its home network from being registered elsewhere, the mobile node deregisters with its foreign agent, through exchange of a Registration Request and Registration Reply message.

Let us take an example of IP datagrams being exchanged over a TCP connection between the mobile node (A) and another host (server X in Fig. 4.6). The following steps occur:

1. Server X wants to transmit an IP datagram to node A. The home address of A is advertised and known to X. X does not know whether A is in the home network or somewhere else. Therefore, X sends the packet to A with A's home address as the destination IP address in the IP header. The IP datagram is routed to A's home network.
2. At the A's home network, the incoming IP datagram is intercepted by the home agent. The home agent discovers that A is in a foreign network. A care-of address has been allocated to A by this foreign network and available with the home agent. The home agent encapsulates the entire datagram inside a new IP datagram, with A's care-of address in the IP header. This new datagram with the care-of address as the destination address is retransmitted by the home agent.
3. At the foreign network, the incoming IP datagram is intercepted by the foreign agent. The foreign agent is the counterpart of the home agent in the foreign network. The foreign agent strips off the outer IP header, and delivers the original datagram to A.

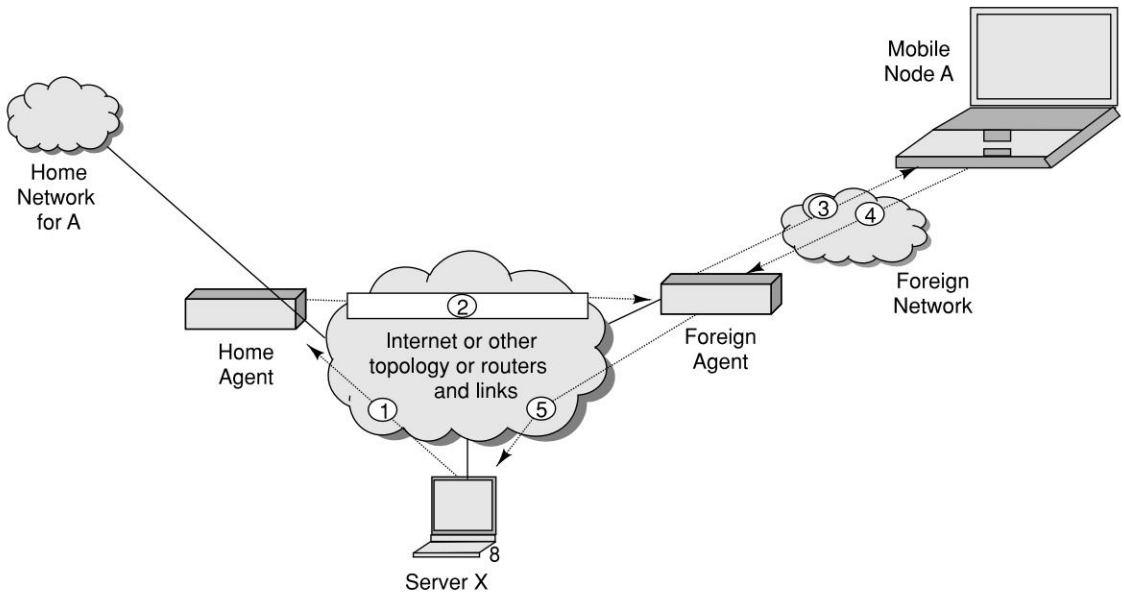


Figure 4.6 Mobile IP Architecture

4. A intends to respond to this message and sends traffic to X. In this example, X is not mobile; therefore X has a fixed IP address. For routing A's IP datagram to X, each datagram is sent to some router in the foreign network. Typically, this router is the foreign agent. A uses X's IP static address as the destination address in the IP header.
5. The IP datagram from A to X travels directly across the network, using X's IP address as the destination address.

To support the operations illustrated in fig. 4.6, mobile IP needs to support three basic capabilities:

- *Discovery*: A mobile node uses a discovery procedure to identify prospective home and foreign agents.
- *Registration*: A mobile node uses a registration procedure to inform its home agent of its care-of address.
- *Tunneling*: Tunneling procedure is used to forward IP datagrams from a home address to a care-of address.

4.5.2 Discovery

Agent advertisements are transmitted by both home and foreign agents to advertise their services on a link. Mobile nodes use these advertisements to determine their current point of attachment to the Internet. The Mobile IP discovery procedure has been built on top of an existing ICMP router discovery, router advertisement, and router solicitation procedure as specified for ICMP Router Discovery in RFC 1256. Mobile IP uses control messages that are sent to and from UDP port number 434. Mobile IP needs extensions to current messages formats. Extensions to ICMP router Discovery include:

0 One-byte Padding (encoded with no Length nor Data field);
16 Mobility Agent Advertisement; and
19 Prefix-Lengths.

Mobile IP control messages, however, include extensions like:

1 Registration Request;
3 Registration Reply;
32 Mobile-Home Authentication;
33 Mobile-Foreign Authentication; and
34 Foreign-Home Authentication.

Using the discovery procedure, the mobile node determines whether it is in a foreign network. For the purpose of discovery, a router or an agent periodically issues a router advertisement ICMP message. The mobile node on receiving this advertisement packet compares the network portion of the router IP address with the network portion of its own IP address allocated by the home network (home address). If these network portions do not match, then the mobile node knows that it is in a foreign network. A router advertisement can carry information about default routers and information about one or more care-of addresses. If a mobile node needs a care-of address without waiting for the agent advertisement, the mobile node can broadcast a solicitation that will be answered by any foreign agent.

4.5.3 Registration

Once a mobile node obtained a care-of address from the foreign network, the same needs to be registered with the home agent. The mobile node sends a registration request to the home agent with the care-of address information. When the home agent receives this request, it updates its routing table and sends a registration reply back to the mobile node.

Authentication: As a part of registration, the mobile node needs to be authenticated. Each mobile node, foreign agent, and home agent support a mobility security association (SA) for mobile entities, indexed by their security parameters index (SPI) and IP address. In the case of the mobile node, this must be its Home Address. Registration messages between a mobile node and its home agent MUST be authenticated with an authorization-enabling extension, e.g., the Mobile-Home Authentication Extension. This extension MUST be the first authentication extension; other foreign agent-specific extensions MAY be added to the message after the mobile node computes the authentication. Using 128-bit secret key and the HMAC-MD5 hashing algorithm, a digital signature is generated. Each mobile node and home agent shares a common secret. This secret makes the digital signature unique and allows the agent to authenticate the mobile node. At the end of the registration a triplet containing the home address, care-of address and registration lifetime is maintained in the home agent. This is called a binding for the mobile node. The home agent maintains this association until the registration life expires. The registration process involves the following four steps:

- The mobile node requests for forwarding service from the foreign network by sending a registration request to the foreign agent.
- The foreign agent relays this registration request to the home agent of that mobile node.
- The home agent either accepts or rejects the request and sends a registration reply to the foreign agent.
- The foreign agent relays this reply to the mobile node.

We have assumed that the foreign agent will allocate the care-of address. However, it is possible that a mobile node moves to a network that has no foreign agents or on which all foreign agents are busy. It is also possible that the care-of address is dynamically acquired as a temporary address by the mobile node such as through DHCP (Dynamic Host Configuration Protocol) as explained in RFC2131, or may be owned by the mobile node as a long-term address for its use only while visiting some foreign network. As an alternative therefore, the mobile node may act as its own foreign agent by using a co-located care-of address. A co-located care-of address is an IP address obtained by the mobile node that is associated with the foreign network. If the mobile node is using a co-located care-of address, then the registration happens directly with its home agent.

4.5.4 Tunneling

Figure 4.7 shows the tunneling operations in Mobile IP. In the mobile IP, an IP-within-IP encapsulation mechanism is used. Using IP-within-IP, the home agent, adds a new IP header called tunnel header. The new tunnel header uses the mobile node's care-of address as the tunnel destination IP address. The tunnel source IP address is the home agent's IP address. The tunnel header uses 4 as the protocol number (Fig. 4.8), indicating that the next protocol header is again an

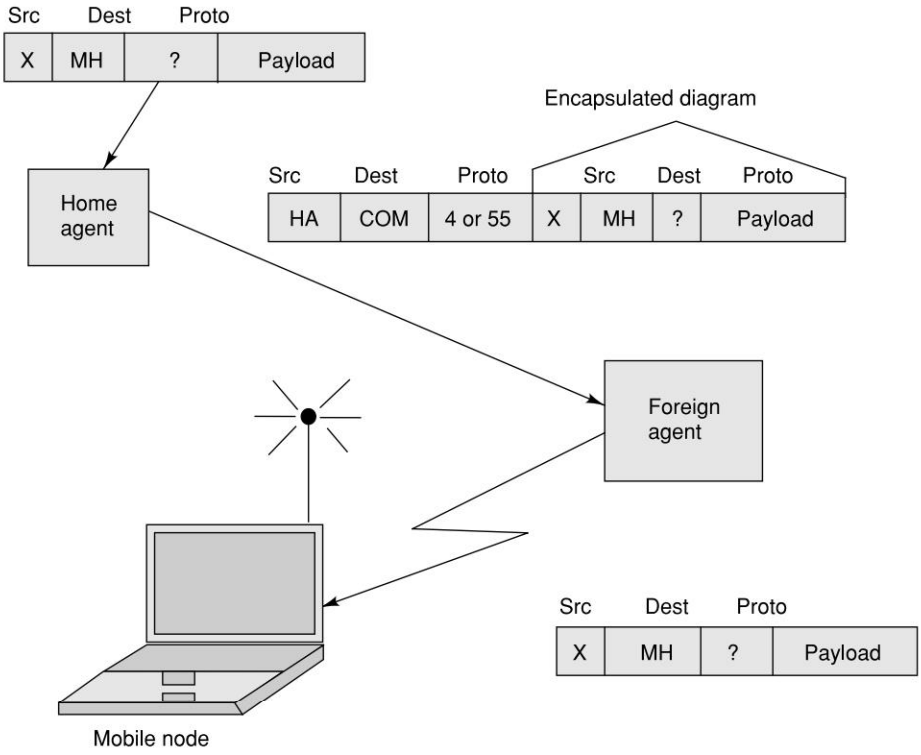


Figure 4.7 Tunneling Operations in Mobile IP

IP header. In IP-within-IP, the entire original IP header is preserved as the first part of the payload of the tunnel header. The foreign agent after receiving the packet, drops the tunnel header and delivers the rest to the mobile node.

When a mobile node is roaming in a foreign network, the home agent must be able to intercept all IP datagram packets sent to the mobile node so that these datagrams can be forwarded via tunneling. The home agent, therefore, needs to inform other nodes in the home network that all IP datagrams with the destination address of the mobile node should be delivered to the home agent. In essence, the home agent steals the identity of the mobile node in order to capture packets destined for that node that are transmitted across the home network. For this purpose ARP (Address Resolution Protocol) is used to notify all nodes in the home network.

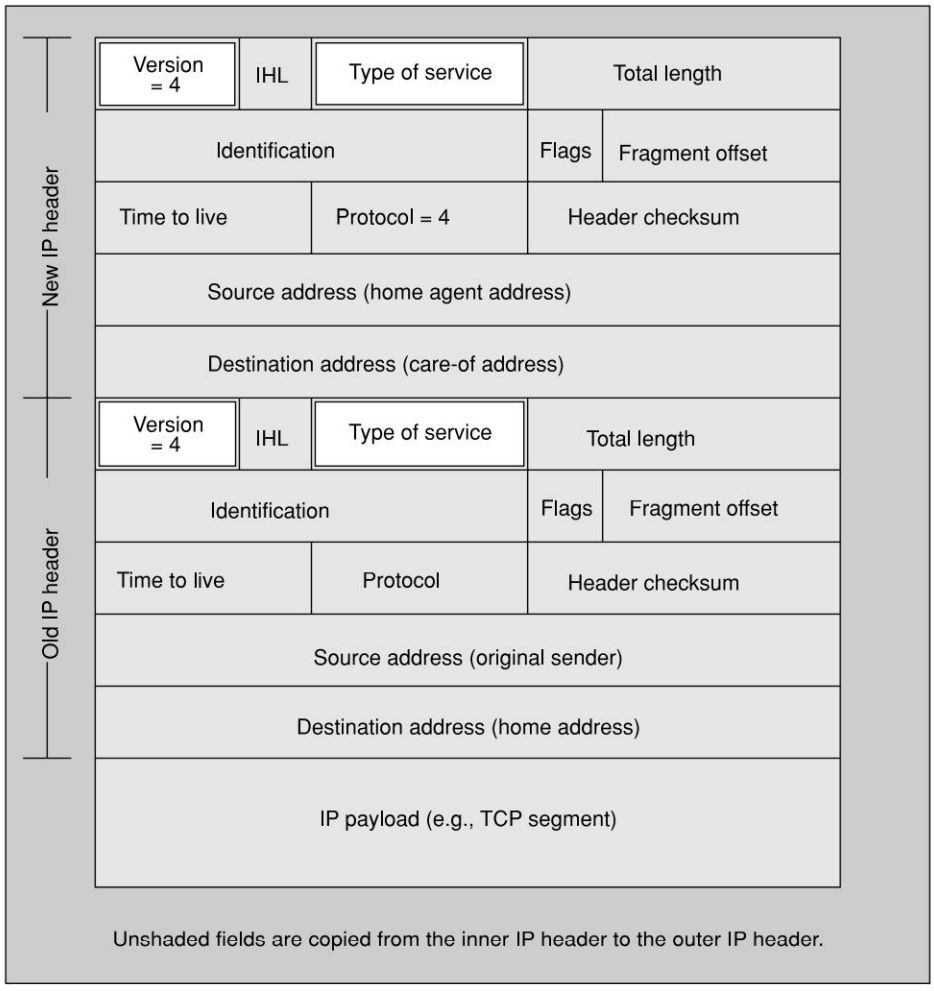


Figure 4.8 The IP Headers in Mobile IP (IP Encapsulation)

Let us take the example of Fig. 4.6. The original IP datagram from X to A has a source address as IP address of X and a destination address as the home IP address of A. The datagram is routed through the Internet to A's home network, where it is intercepted by the home agent. The home agent encapsulates the incoming datagram with an outer IP header. This outer header includes a source address same as the IP address of the home agent and a destination address equal to the care-of address. As the care-of address has the network portion of the foreign network, the packet will find its way directly to the mobile host. When this new datagram reaches the host in the foreign network, it strips off the outer IP header to extract the original datagram. From this stripped off packet it also finds out the original sender. This is necessary for the host to know who has sent the packet so that the response reaches the right destination.

In any IP data packet, the source and destination IP address must be topologically correct. The forward tunnel in Mobile IP complies with this, as its endpoints (home agent address and care-of address) are properly assigned addresses for their respective locations. On the other hand, the source IP address of a packet transmitted by the mobile node does not correspond to the network prefix from where it emanates. To mitigate this risk, IETF proposed reverse tunnelling that is specified in RFC 2344.

4.5.5 Cellular IP

The primary design goal for mobile IP protocols is to allow a host to change its point of access during data transfer without being disconnected or needing to be reconfigured. An important design goal for mobile host protocols is to support handoffs without significant disturbance to ongoing data transmission. A change of access point while connectivity is maintained is called a handoff.

To manage mobility, generally a "two tier addressing" scheme is used. One address is for a fixed location which is known to all; other one is for a dynamic location which changes as the user moves. In case of GSM this is done through Home Location Register and Visitor Location Register. Same is true in Mobile IP, where a mobile host is associated with two IP addresses: a fixed home address that serves as the host-identifier; and a care-of address that reflects its current point of attachment. The mobile IP architecture comprises three functions:

1. A database that contains the most up-to-date mapping between the two address spaces (home address to care-of address).
2. The translation of the host identifier to the actual destination address.
3. Agents ensuring that the source and destination packets for arriving and outgoing packets are updated properly so that routing of packets is proper.

Whenever the mobile host moves to a new subnet managed by a different foreign agent, the dynamic care-of-address will change. This changed care-of address needs to be communicated to the home agent. This process works for slowly moving hosts. For a high speed mobile host, the rate of update of the addresses needs to match the rate of change of addresses. Otherwise, packets will be forwarded to the wrong (old) address. Mobile IP fails to update the addressed properly for high speed mobility. Cellular IP (Fig. 4.9), a new host mobility protocol has been designed to address this issue.

In a Cellular IP, none of the nodes know the exact location of a mobile host. Packets addressed to a mobile host are routed to its current base station on a hop-by-hop basis where each node only needs to know on which of its outgoing ports to forward packets. This limited routing information

(referred to as mapping) is local to the node and does not assume that nodes have any knowledge of the wireless network topology. Mappings are created and updated based on the packets transmitted by mobile hosts.

Cellular IP uses two parallel structures of mappings through Paging Caches (PC) and Routing Caches (RC). PCs maintain mappings for stationary and idle (not in data communication state) hosts; whereas, RC maintains mappings for mobile hosts. Mapping entries in PC have a large timeout interval, in the order of seconds or minutes. RCs maintain mappings for mobile hosts currently receiving data or expecting to receive data. For RC mappings, the timeout is in the packet time scale. Figure 4.10 illustrates the relationship between PCs and RCs. While idle at location 1, the mobile host X keeps PCs up-to-date by transmitting dummy packets at a low frequency (Step 1 in Fig. 4.10). Let us assume that the host is mobile and moved to location 2 without transacting any data. The PC mapping for X now points to location 2. While at location 2, there are data packets to be routed to the mobile host X, the PC mappings are used to find the host (Step 2). As there is data transmission, the mapping database to be used will be the RC. As long as data packets keep arriving, the host maintains RC mappings, either by its outgoing data packets or through the transmission of dummy packets (Step 3).

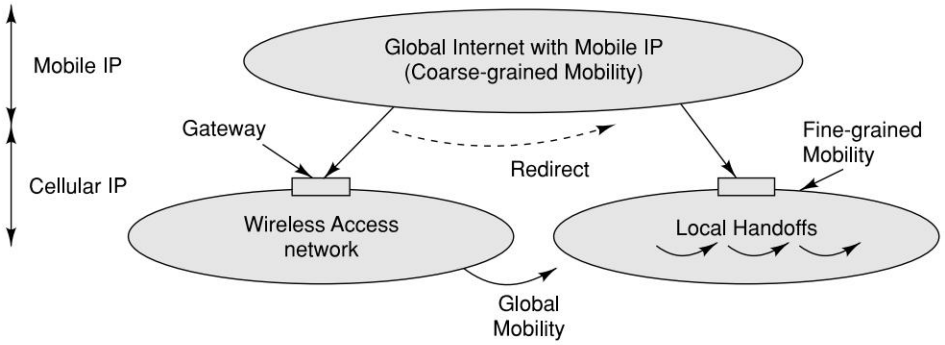


Figure 4.9 Relationships between Mobile IP and Cellular IP

Idle mobile hosts periodically generate short control packets, called paging-update packets. These are sent to the nearest available base station. The paging-update packets travel in the access network from the base station toward the gateway router, on a hop-by-hop basis. Handoff in Cellular IP is always initiated by the mobile host. As the host approaches a new base station, it redirects its data packets from the old to the new base station. First few redirected packets will automatically configure a new path of RC mappings for the host to the new base station. For a time equal to the timeout of RC mappings, packets addressed to the mobile host will be delivered at both old and new base stations.

4.6 Internet Protocol Version 6 (IPV6)

Internet offers access to information sources worldwide. We access Internet through increasing variety of wireless devices offering IP connectivity, such as PDAs, palmtops, handhelds, laptops,

and digital cellular phones. The explosion in the number of devices connected to the Internet, combined with projections for the future, made scientists think seriously whether the 32-bit address space of TCP/IP is sufficient. IP version 6 (IPv6), the successor to today's IP version 4 protocol (IPv4), dramatically expands the available address space. Internet Engineering Task Force (IETF) has produced a comprehensive set of specifications (RFC 1287, 1752, 1886, 1971, 1993, 2292, 2373, 2460, 2473, etc.) that define the next-generation IP protocol originally known as "IPNg," now renamed as "IPv6". IPv6 addresses both a short-term and long-term concern for network owners, service providers and users.

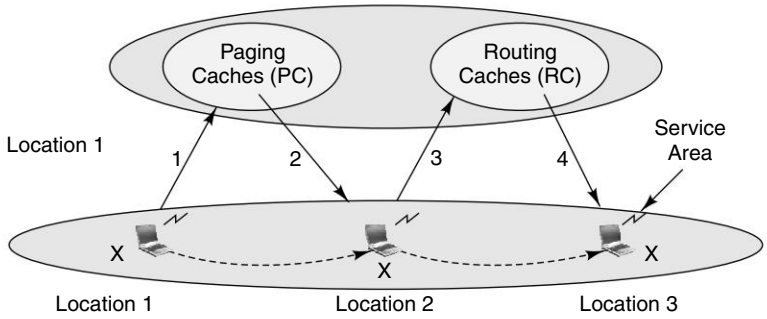


Figure 4.10 Cellular IP Paging and Routing

4.6.1 Address Space

IPv6 uses 128 bit addresses for each packet, creating a virtually infinite number of IP addresses (approximately 3.4*10^38 IP addresses), as opposed to 3758096384 IPv4 addresses (2^31 A Class address + 2^30 B Class + 2^29 C Class address). This also means that if we set the world population at 10 billion in 2050, there will be 3.4*10^27 addresses available per person.

In IPv6, there are global addresses and local addresses. Global addresses are used for routing of global Internet. Link local addresses are available within a subnet. IPv6 uses hierarchical addressing with three-level of addresses (Fig. 4.11). This includes a Public Topology (the 48 bit external routing prefix), a Site Topology (typically a 16 bit subnet number), and an Interface Identifier (typically an automatically generated 64 bit number unique on the local LAN segment).

End-user-sites get their address prefix from an ISP that provides them the IPv6 service. General IPv6 host is given a linklocal address such as fe80::EUI-64 and more than one global address such as global-prefix::EUI-64. It has 64 bit length and made by IEEE EUI-64 format. Interface ID is used to specific Interface in the same link. Interface ID is generated to use Interface's link layer address. An Ethernet MAC address for a device is 48 bits long, Interface ID is created by adding 2 octet "0xfffe" in it's center. Like 02:60:8c:de:7:79 becomes 260:8cff:fede:779.

4.6.2 IPv6 Security

One of the biggest differences between IPv6 and IPv4 is that all IPv6 nodes are expected to implement strong authentication and encryption features to improve Internet security. IPv6 comes

native with a security protocol called IP Security (IPsec). Many vendors adapted IPsec as a part of IPv4. IPsec protocol is a standards-based method of providing confidentiality, integrity, and authenticity to information transferred across IP networks. IPsec combines several different security technologies into a complete system to provide confidentiality, integrity and authenticity.

The fundamental components of the IPsec security architecture are discussed in terms of their underlying functionalities as described below:

(a) Security Associations—What they are and how they work, how they are managed, associated processing specified in RFC4301.

(b) Security Protocols—Authentication Header (AH) and Encapsulating Security Payload (ESP), specified in RFC4302 and RFC4303 respectively.

(c) Cryptographic algorithms for authentication and encryption specified in RFC4305.

(d) Key Management—manual and automated (the Internet Key Exchange (IKE)) specified in RFC4306.

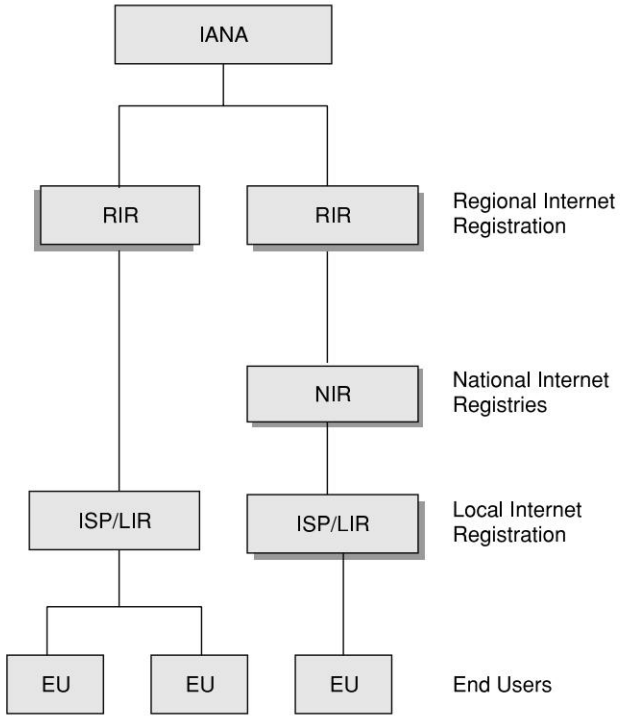


Figure 4.11 Hierarchical Addressing of IPv6

The concept of Security Association (SA) is fundamental to IPsec. IPsec uses SA to track all the particulars concerning a given IPsec communication session—it is a logical uni-directional (simplex) connection that can be defined as relationships between entities (hosts, gateways, firewalls, routers) that describe security policies amongst them. To secure a bi-directional secure communication between two IPsec-enabled systems, a pair of SAs (one in each direction) is required. The protection offered by IPsec is based on requirements defined by a Security Policy Database (SPD) established and maintained by a user or system administrator. When a security service is chosen, the two IPsec peers must determine exactly which algorithms to use (for example, AES-CBC for encryption; SHA-1 for integrity). IKE explicitly creates SA pairs in recognition of this common usage requirement. The Peer Authentication Database (PAD), provides a link between an SA management protocol (such as IKE).

In an entity there will be many SA that are stored in a SA Database (SAD). To identify a particular SA within a SAD, there has to be a pointer to the database that is called Security Parameters Index (SPI). Security services are afforded to an SA by the use of AH, or ESP, but not both. AH is used to provide integrity and data origin authentication and to provide protection against replays. ESP on contrast, offers confidentiality, integrity, authentication, and anti-replay. If both AH and ESP

protection are applied to a traffic stream, then two SAs must be created and coordinated to effect protection through iterated application of the security protocols. Each SA consists of values such as destination address, a security parameter index (SPI), the IPsec transforms used for that session, security keys, and additional attributes such as IPsec lifetime as illustrated in Figure 4.12.

In particular, IPsec uses:

- Diffie-Hellman key exchange mechanism for deriving key between peers on a public network.
- Public key cryptography to guarantee the identity of the two parties and avoid man-in-the-middle attacks.

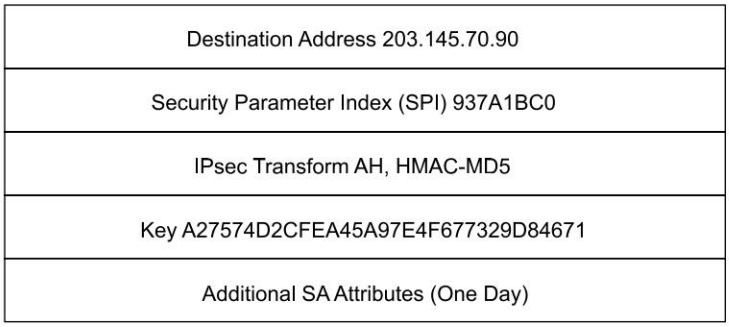


Figure 4.12 Example of a Security Association

- Bulk encryption algorithms, such as 3DES, for encrypting the data.
- Keyed hash algorithms, such as HMAC, combined with traditional hash algorithms such as MD5 or SHA for providing packet authentication.
- Digital certificates signed by a certificate authority to act as digital ID cards.
- IPsec provides IP network-layer encryption.

In addition, IPsec uses following cryptographic algorithms for bulk data (payload) encryption and authentication algorithms for the IPsec ESP protocol.

Requirement Encryption Algorithm (notes)

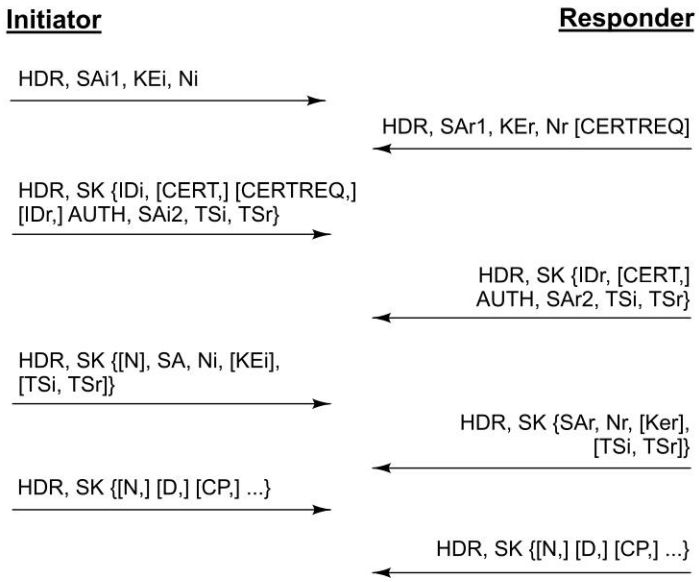
MUST	NULL (1)
MUST-	TripleDES-CBC [RFC2451]
SHOULD+	AES-CBC with 128-bit keys [RFC3602]
SHOULD	AES-CTR [RFC3686]
SHOULD	NOT DES-CBC [RFC2405] (3)

Requirement Authentication Algorithm (notes)

MUST	HMAC-SHA1-96 [RFC2404]
MUST	NULL (1)

SHOULD+	AES-XCBC-MAC-96 [RFC3566]
MAY	HMAC-MD5-96 [RFC2403] (2)

The protocol for key exchange for IPsec is similar to Transport Layer Security or TLS. The IPsec Internet Key Exchange version 2 (IKEv2) is shown in Fig. 4.13.



Requirement Encryption Algorithm (notes)

AUTH	Authentication
CERT	Certificate
CERTREQ	Certificate Request
CP	Configuration
D	Delete
E	Encrypted
EAP	Extensible Authentication
HDR	IKE Header
IDi	Identification-Initiator
IDr	Identification-Responder
KEi	Key Exchange-Initiator
KEr	Key Exchange-Responder
N	Notified
Ni	Nonce-Initiator
Nr	Nonce-Responder
Ni	Notify-Initiator
SAi	Security Association-Initiator
SAr	Security Association-Responder

TSi	Traffic Selector-Initiator
TSr	Traffic Selector-Responder
V	Vendor ID

Figure 4.13 IPsec IKEv2 Procedure

In the protocol above, HDR contains the Security Parameter Indexes (SPIs), version numbers, and flags of various sorts. The SAi1 payload states the cryptographic algorithms the initiator supports. The KE payload sends the initiator's Diffie-Hellman value. Ni is the initiator's nonce. Payloads such as [CERTREQ] that may optionally appear are shown in brackets, indicate that optionally a certificate request payload can be included. At this point in the negotiation, each party can generate SKEYSEED, from which all keys are derived for that IKE_SA. Everything except the headers of all the messages that follow are encrypted and integrity protected.

4.6.3 Packet Payload

Each IPv6 packet payload is attached a tag which can be customized to enable a better quality in the packet flow, or by a price of other class, such as non-real-time quality of service or "real-time" service. This feature does not exist natively in IPv4, although a part of payload could be used for the same, reducing unique information amount carried by the packet.

Information is packetized into IPv6 packets, with the corresponding levels of control. A neighbor discovery feature (care-of address, and stateless Prefix or Stateful DHCPv6) will in principle allow the device carrying these packets to configure itself for a consistent dialogue with other devices or software interfaces. The same can be done with IPv4 packets, but with the intervention of humans or specific tools and services and only for selected information and software architectures.

4.6.4 Migrating from IPv4 to IPv6

The migration from IPv4 to IPv6 is quite an involved task. This includes the following:

1. Migration of the network components to be able to support IPv6 packets. As there is no change at the physical layer between IPv4 and IPv6, network components like hub or switch need not change. As there is a change in the packet header the routers need to be upgraded. However, using IP tunneling IPv6 packets can propagate over an IPv4 envelope. Existing routers can support IP tunneling.
2. Migration of the computing nodes in the network: this will need the operating system upgrades so that they support IPv6 along with IPv4. Upgraded systems will have both IPv4 and IPv6 stacks. Therefore, both the IPv4 and IPv6 applications can run without any difficulty.
3. Migration of networking applications in both client and server systems: this requires porting of the applications from IPv4 to IPv6 environment.

Migration of Windows System

The Microsoft Windows 9x families do not support IPv6. Windows XP and Windows Server 2003 support IPv6 natively. Windows 2000 Professional can be upgraded to support IPv6. IPv6 in Windows support different tools and dlls. These are:

wship6.dll: The Winsock helper dynamically linked library for the INET6 address family.

wininet.dll: Winsock INET6 libraries.

ftp.exe: This is the IPv6 ftp client and server application.
telnet.exe: This is the IPv6 telnet client application.
tlntsvr.exe: Telnet server.

ipv6.exe: This tool retrieves and displays configuration information about the IPv6 protocol. This tool is used to view the state of interfaces, the neighbor caches, the binding cache, the destination cache, and the route table. This utility can also be used to manually configure interfaces, addresses, and route table entries.

ping6.exe: This tool is equivalent to the current IPv4 ping.exe tool. It sends ICMPv6 Echo Request messages, waiting for the corresponding ICMPv6 Echo Reply messages and then displaying information on round trip times.

tracert6.exe: This tool is equivalent to the current IPv4 tracert.exe tool. It sends

ICMPv6 Echo Request messages with monotonically increasing values of the Hop Limit field to discover the path traveled by IPv6 packets between a source and destination.

ttcp.exe: This tool is used to send TCP segment data or UDP messages between two nodes. ttcp.exe supports both IPv4 and IPv6.

6to4cfg.exe: This tool is used to configure IPv6 connectivity over an IPv4 network.

ipsec6.exe: This tool is used to configure policies and security associations for IPv6 IPsec traffic.

checkv4.exe: This tool is used to scan source code files to identify code that needs to be changed to support IPv6. This is similar to the *lint* command in UNIX.

Migration of Linux System

Linux kernel 2.4.x either supports IPv6 directly or can be upgraded to support IPv6. All versions after Red Hat Linux 7.1 support IPv6 directly. The kernel needs to be built and properly configured for IPv6. Different tools are available in Linux as a part of v6 installation. These are:

#**ping6**: equivalent of ping in IPv4.

#**traceroute6**: equivalent of tracerout in IPv4.

#**tracepath6**: equivalent of tracepath in IPv4.

#**tcpdump**: equivalent of tcpdump in IPv4.

#proto: displays only tunneled IPv6-in-IPv4 traffic.

#ip6: displays all native IPv6 traffic including ICMPv6.

#icmp6: displays only native ICMPv6 traffic.

4.6.5 Migration of Applications

There are many networking software and systems around the world that use TCP/IP socket. One of the changes we need to do in all these applications is to allow the larger address space for the destination endpoint. This is similar to the classic Y2K problem of last century, where the space provided for a date field had to be enlarged. Moreover, in the case of IPv6, the header has been

changed according to the practical need of the present-day applications as well as to facilitate high speed routing. The difference is that Y2K was a time-bound project and had to be finished before 1 January 2000. However, in case of IPv6, there is no such compulsion other than all applications in 3G network will have to be IPv6 as 3G will not support IPv4.

Figures 4.14a and 4.14b are the snapshots of a telnet application after migration to IPv6 environment.

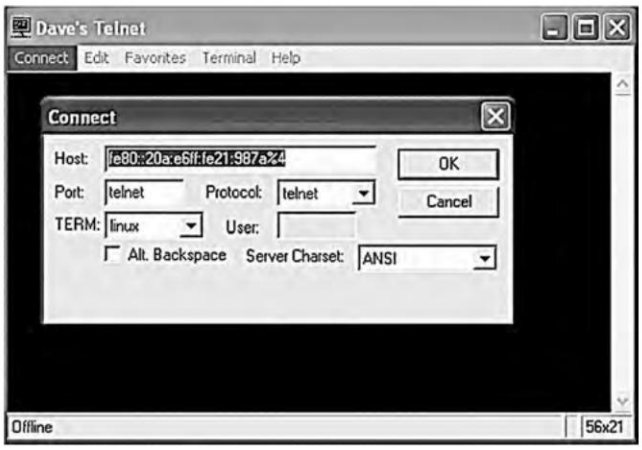


Figure 4.14a Snapshot of Telnet Connection Request in IPv6 (Windows XP through dtelnet)

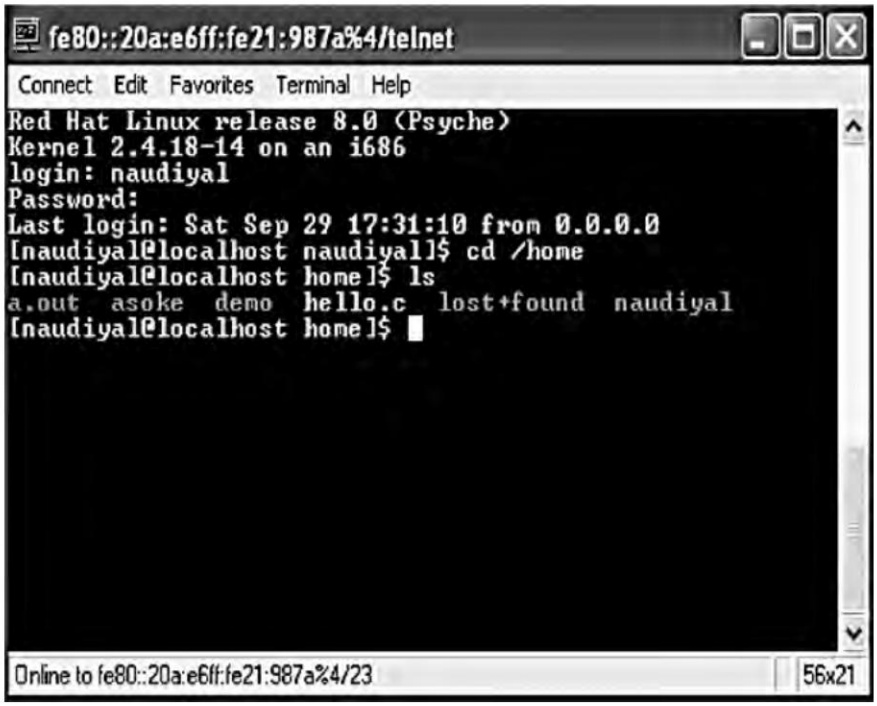


Figure 4.14b Snapshot of Telnet Client in IPv6

4.6.6 Interconnecting IPv6 Networks

Till all the routers/system become IPv6-compatible, the interconnection between IPv6 networks can be accomplished by tunneling. Tunneling is one of the key deployment strategies for both service providers as well as enterprises during the period of IPv4 and IPv6 co-existence. Tunneling service providers can offer an end-to-end IPv6 service without major upgrades to the infrastructure and without impacting current IPv4 services.

A variety of tunnel mechanisms are available. These mechanisms include:

1. Manually created tunnels such as IPv6 manually configured tunnels (RFC 2893).
2. IPv6 over IPv4 tunnels.
3. Semiautomatic tunnel mechanisms such as that employed by tunnel broker services.
4. Fully automatic tunnel mechanisms such as IPv4-compatible and 6 to 4.

The dual-stack routers run both IPv4 and IPv6 protocols simultaneously and thus can interoperate directly with both IPv4 and IPv6 end systems and routers.

4.6.7 Mobile IP with IPv6

In Section 4.5 we have discussed Mobile IP as originally specified for IPv4; and, in this section we will discuss Mobile IP for IPv6 specified in RFC3775 that includes many additional features. IPv6 with hierarchical addressing scheme will be able to manage IP mobility much efficiently. IPv6 in addition, attempts to simplify the process of renumbering, which could be critical to the future routability of the Internet traffic. It retains the ideas of a home network, home agent and the use of encapsulation to deliver packets from the home network to the mobile node's current point of attachment. While discovery of a care-of address is still required, a mobile node can configure its care-of address by using Stateless Address Autoconfiguration and Neighbor Discovery. Thus, foreign agents are not required to support mobility in IPv6.

Basic differences between Mobile IPv4 and Mobile IPv6 are,

- Mobile IPv6 operates in any location without any special support required from the local router; therefore, "foreign agents" are not required for Mobile IPv6.
- Route optimization is a fundamental part of the Mobile IPv6 protocol.
- Mobile IPv6 route optimization can operate securely even without pre-arranged security associations (SA is described in Section 4.6.2).
- Mobile IPv6 will allow route optimization to coexist efficiently with routers that perform "ingress filtering".
- The IPv6 Neighbor Unreachability Detection assures symmetric reachability between the mobile node and its default router in the current location.
- Most packets sent to a mobile node while away from home in Mobile IPv6 are sent using an IPv6 routing header rather than IP encapsulation.
- Mobile IPv6 is decoupled from any particular link layer, as it uses IPv6 Neighbor Discovery instead of ARP.
- The use of IPv6 encapsulation (and the routing header) removes the need in Mobile IPv6 to manage "tunnel soft state".
- The dynamic home agent address discovery mechanism in Mobile IPv6 returns a single reply to the mobile node. The directed broadcast approach used in IPv4 returns separate replies from each home agent.

4.7 JAVA CARD

Java Card is a smart card with Java framework. Smart card was developed in 1974, by Roland Moreno. Smart card is a plastic card with intelligence and memory. Smart cards are becoming popular as identity module and wireless security devices. In many countries driving licenses are being issued on smart cards. The SIM card on a GSM mobile phone is a smart card as well. The importance of smart card made ISO standardize all its interfaces. These are done through ISO 7816 standards. These ISO standards define the physical characteristic of the card (ISO 7816-1: Physical Characteristics), locations and dimensions of the contacts (7816-2:: Dimensions and Locations of the Contacts), signals and transmission interfaces (7816-3:: Electronic Signals and Transmission Protocols), and command interfaces (7816-4:: Interindustry Commands for Interchange). A smart card is embedded with either (i) a microprocessor and a memory chip or (ii) only a memory chip with non-programmable logic. A microprocessor card can have an intelligent program resident within the card which can add, delete, and otherwise manipulate information on the card. A memory card on contrast, can store some information for some pre-defined operation. Smart cards are capable of carrying data, functions, and information on the card. Therefore, unlike memory strip cards, they do not require access to remote databases at the time of the transaction.

Microprocessor based smart cards which were used for some specific application areas are becoming quite common. Smart cards have now emerged as multi-function cards. To allow interoperability, Java was chosen. All the microprocessor based smart cards now offer Java API framework on them. This is why smart cards with Java framework are also called Java Cards. 3GPP has decided to use Java Card as the standard for USIM and ICC (Integrated Circuit cards). Java Card technology preserves many of the benefits of the Java programming languages such as: productivity, security, robustness, tools, and portability. For Java card, the Java Virtual Machine (JVM), the language definition, and the core packages have been made more compact to bring Java technology to the resource constrained smart cards.

A smart card of a GSM SIM card supporting Java Card functionalities may typically have an 8 or 16 bit microprocessor running at speeds between 5 MHz to 40 MHz with 32K to 128K bytes of EEPROM (Electronically Erasable Programmable Read Only Memory). Though Java card works in a master/slave mode, using the proactive SIM technology of GSM Phase 2+, it is possible for the application on the SIM card to get activated in an automated fashion. Also, Java card technology supports OTA (Over the Air) downloads. In OTA download, a Java applet (through SMS) can be downloaded by the network operator proactively or by the user interactively over the wireless media. Applications written for the Java Card platform are referred to as applets.

The development framework in Java card is different from that on a desktop computer. The major challenge of Java Card technology on smart card is to fit Java system software in a resource constraint smart card while conserving enough space for applications. Java Card supports a subset of the features of Java language available on desktop computers. The Java Card virtual machine on a smart card is split into two parts (Fig. 4.15): one that runs off-card and the other that runs on-card. Many processing tasks that are not constrained to execute at runtime, such as class loading, bytecode verification, resolution and linking, and optimization, are dedicated to the virtual machine that is running off-card where resources are usually not a concern. The on-card components of Java Card include components like the Java Card virtual machine (JCVM), the Java Card Runtime Environment (JCRE), and the Java API. Task of the compiler is to convert a Java source into Java

class files. The converter will convert class files into a format downloadable into the smart card. Converter ensures the byte code validity before the application is installed into the card. The converter checks the classes off-card for,

- How well it is formed?
- Java Card subset violations.
- Static variable initialization.
- Reference resolution.
- Byte code optimization.
- Storage allocation.
- The Java Card interpreter.
- Applet execution.
- Controls run time resources.
- Enforces runtime security.

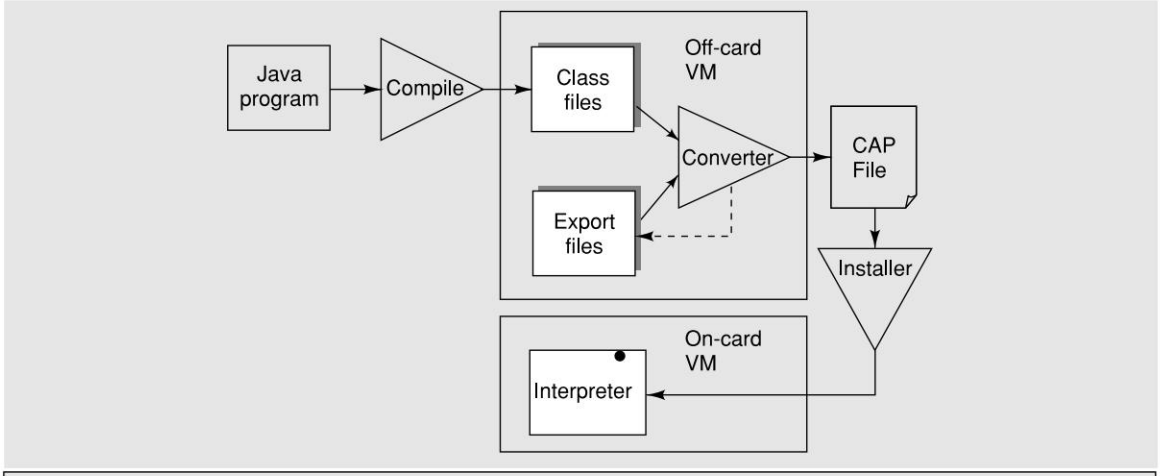


Figure 4.15 Architecture of Java Card Applications Development Process

Following conversion by the off-card VM into CAP (Converted APlet) format, the applet is transferred into the card using the installer. The applet is selected for execution by the JCRE. JCRE is made up of the on-card virtual machine and the Java Card API classes. JCRE performs additional runtime security checks through the applet firewall. Applet firewall partitions the objects stored into separate protected object spaces, called contexts. Applet firewall controls the access to shareable interfaces of these objects. The JCVM is a scaled down version of standard JVM (Java Virtual Machine). Elements of standard Java not supported in JCVM are,

- Security manager.
- Dynamic class loading.
- Bytecode verifier.
- Threads.
- Garbage collection.
- Multi-dimensional arrays.

- Char and strings.
- Floating point operation.
- Object serialization.
- Object cloning.

As mentioned above, Java applications for Java Cards are called Applets. Java Card applets should not be confused with Java applets on the Internet. A Java Card applet is not intended to run within an Internet browser environment. The reason for choosing the name applet is that Java Card applets can be loaded into the Java Card runtime environment after the card has been manufactured. That is, unlike applications in many embedded systems, Java Card applets do not need to be burned into the ROM during manufacture.

REFERENCES/FURTHER READING

1. A Technical Overview of the WirelessMAN Air Interface for Broadband Wireless Access, *IEEE Communications Magazine*, June 2002.
2. *Association for Automatic Identification and Mobility*: http://www.aimglobal.org.
3. Chen Zhiqun, (2000), *Technology for Smart Cards: Architecture and Programmer's Guide*, Sun, Addison-Wesley.
4. Cong, D., M. Hamler and C. Perkins, (2006), 'The Definitions of Managed objects for IP Mobility Support', *RFC*: 2006.
5. Eklund Carl, Roger B. Marks, Kenneth L. Stanwood and Stanley Wang, *IEEE Standard*, 802.16:
6. Gavrilovich, Charles D. Jr., Gray Cary Ware and Freidenrich L.L.P., "Broadband Communication on the Highways of Tomorrow", *IEEE Communications Magazine*, April 2001.
7. *Guidelines For 64-Bit Global Identifier (EUI-64)*. Registration Authority: http://standards.ieee.org/regauth/oui/tutorials/EUI64.html.
8. Held Gil, (2001), *Data Over Wireless Networks Bluetooth, WAP, & Wireless LANs*, McGraw-Hill.
9. *IEEE 802.16 for Broadband*: http://www.nwfusion.com/news/tech/2001/0903-tech.html.
10. *Java card Forum*: http://www.javacardforum.org/.
11. Karagiannis Georgios, 'Mobile IP', *Ericsson Open Report* # 3/0362-FCP NB 102 88 Uen, 13 July, 1999.
12. Karagiannis Georgios, (1999), 'Mobile IP,' *Ericsson State of the Art Report* # 3/0362-FCP NB 102 88 Uen.
13. Muller Nathan J., (2001), *Bluetooth Demystified*, Tata McGraw-Hill.
14. Official Bluetooth site: http://www.bluetooth.com.
15. Perkins, C., 'Mobile IP,' *IEEE Communications Magazine*, May 1997.
16. Perkins, C., 'Mobile Networking through Mobile IP,' *IEEE Internet Computing*, January-February 1998, p 58.
17. Perkins, C., (1998), *Mobile IP: Design Principles and Practices*, Prentice-Hall PTR.

18. Prabhu, C.S.R. and A. Prathap Reddi, (2004), *Bluetooth Technology and Its Applications with Java and J2ME*, Prentice-Hall of India.
19. RFC 2002: IP Mobility Support–http://www.faqs.org/rfcs/rfc2002.html.
20. RFC 2005, Applicability Statement for IP Mobility Support.
21. Stallings William, IEEE 802.16 for broadband wireless, Network World, 09/03/01, RFC1825: Security.
22. Andras G. Valko, "Cellular IP: A New Approach to Internet Host Mobility," *ACM SIGCOMM Computer Communication Review*, pp 50–65.

REVIEW QUESTIONS

Q1: Describe the protocol stack of Bluetooth.

Q2: How does a new Bluetooth device discover a Bluetooth network? Describe the security principles in Bluetooth.

Q3: What is active RFID? Describe two applications of active RFID. How is active RFID different from passive RFID? Describe two applications of passive RFID.

Q4: What is WiMax (Wireless broadband)? How is it different from WiFi? Explain the WiMax physical layer.

Q5: What is Mobile IP? Explain tunneling in the context of Mobile IP.

Q6: How does Mobile IP work? What are the challenges with mobile IP with respect to high speed mobility? How does Cellular IP solve some of these challenges?

Q7: What is Cellular IP?

Q8: In what ways is IPv6 better than IPv4? Briefly enunciate the migration issues from IPv4 to IPv4 in the context of different operating systems.

Q9: You have a communication application that uses sockets in IPv4, what are the steps you need to follow to port this application from IPv4 to IPv6?

Q10: Write short notes on:
- (a) Java Card
- (b) Security Association in IPv6
- (c) IPsec

Q11: You need to develop a secured healthcare application. What information will you keep in the Java card and what will be in the backend server? How will you secure such information on the Java Card?

CHAPTER 5

Global System for Mobile Communications (GSM)

5.1 GLOBAL SYSTEM FOR MOBILE COMMUNICATIONS

GSM is much more than just an acronym for Global System for Mobile Communication. It signifies an extremely successful technology and bearer for mobile communication system. GSM today covers 71% of all the digital wireless market. The mobile telephone has graduated from being a status symbol to a useful appliance. People use it not only in business but also in personal life. Its principal use is for wireless telephony, and messaging through SMS. It also supports facsimile and data communication.

GSM is based on a set of standards, formulated in the early 1980s (see Table 5.1 for the GSM timeline). In 1982, the Conference of European Posts and Telegraphs (CEPT) formed a study group called the Groupe Spécial Mobile (GSM) to study and develop a pan-European mobile system, which was later rechristened as Global System for Mobile Communication. See Chapter 1 for cellular network evolution and standards. The proposed GSM system had to meet certain business objectives. These are:

- Support for international roaming.
- Good speech quality.
- Ability to support handheld terminals.
- Low terminal and service cost.
- Spectral efficiency.
- Support for a range of new services and facilities.
- ISDN compatibility.

Due to its innovative technologies and strengths, GSM rapidly became truly global. Many of the new standardization initiatives came from outside Europe. Depending on locally available frequency bands, different air interfaces were defined. Of these prominent ones are 900 MHz, 1800 MHz and 1900 MHz. However, architecture, protocols, signaling and roaming are identical in all networks independent of the operating frequency bands.

Table 5.1 GSM history timeline

Year	*Event*
1982	Groupe Spécial Mobile (GSM) established
1987	Essential elements of wireless transmission specified
1989	GSM becomes an ETSI technical committee
1990	Phase 1 GSM 900 specification (designed 1987 through 1990) frozen
1991	First GSM network launched
1993	First roaming agreement came into effect
1994	Data transmission capability launched
1995	Phase 2 launched. Fax and SMS roaming services offered
2002	SMS volume crosses 24 billion/year, 750 million subscribers

GSM uses a combination of FDMA (Frequency Division Multiple Access) and TDMA (Time Division Multiple Access). See Section 3.2 for definition of these multiple access procedures. The GSM system has an allocation of 50 MHz (890–915 MHz and 935–960 MHz) bandwidth in the 900 MHz frequency band. Using FDMA, this band is divided into 124 (125 channels, 1 not used) channels each with a carrier bandwidth of 200 KHz. Using TDMA, each of these channels is then further divided into eight time slots. Therefore, with the combination of FDMA and TDMA we can realize a maximum of 992 channels for transmitting and receiving. In order to be able to serve hundreds of thousands of users, the frequency must be reused. This is done through cells.

The frequency reuse concept led to the development of cellular technology as originally conceived by AT&T and Bell Labs way back in 1947. The essential characteristics of this reuse are as follows:

- The area to be covered is subdivided into radio zones or cells (Fig. 5.1). Though in reality these cells could be of any shape, for convenient modeling purposes these are modeled as hexagons. Base stations are positioned at the center of these cells.
- Each cell i receives a subset of frequencies fbi from the total set assigned to the respective mobile network. To avoid any type of co-channel interference, two neighboring cells never use the same frequencies.
- Only at a distance of D (known as frequency reuse distance), the same frequency from the set fbi can be reused. Cells with distance D from cell i, can be assigned one or all the frequencies from the set fbi belonging to cell i.
- When moving from one cell to another during an ongoing conversation, an automatic channel change occurs. This phenomenon is called handover. Handover maintains an active speech and data connection over cell boundaries.

The regular repetition of frequencies in cells result in a clustering of cells. The clusters generated in this way can consume the whole frequency band. The size of a cluster is defined by k, the number of cells in the cluster. This also defines the frequency reuse distance D. Figure 5.1 shows an example of a cluster size of 4.

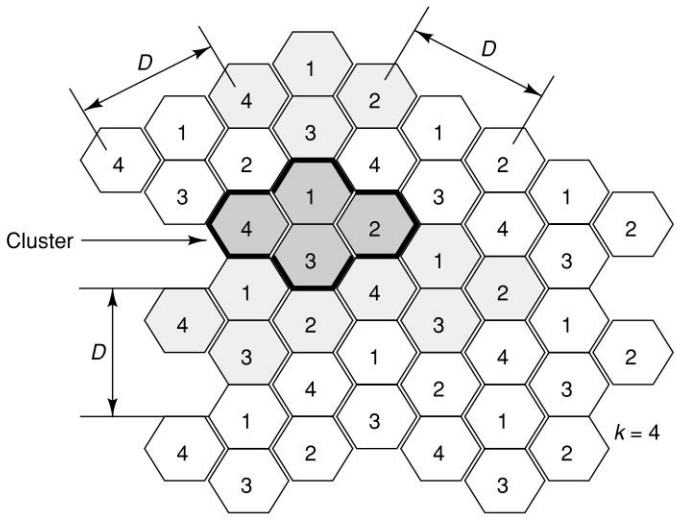


Figure 5.1 Cell Clusters in GSM

5.2 GSM ARCHITECTURE

GSM networks are structured in hierarchic fashion (Fig. 5.2). It consists at the minimum one administrative region assigned to one MSC (Mobile Switching Centre). The administrative region is commonly known as PLMN (Public Land Mobile Network). Each administrative region is subdivided into one or many Location Area (LA). One LA consists of many cell groups. Each cell group is assigned to one BSC (Base Station Controller). For each LA there will be at least one BSC. Cells in one BSC can belong to different LAs.

Cells are formed by the radio areas covered by a BTS (Base Transceiver Station) (Fig. 5.3). Several BTSs are controlled by one BSC. Traffic from the MS (Mobile Station) is routed through MSC. Calls originating from or terminating in a fixed network or other mobile networks is handled by the GMSC (Gateway MSC). Figure 5.3 depicts the architecture of a GSM PLMN from technology point of view, whereas Figure 5.4 depicts the same architecture from the operational point of view.

For all subscribers registered with a cellular network operator, permanent data such as the service profile is stored in the Home Location Register (HLR). The data relate to the following information:

- Authentication information like International Mobile Subscriber Identity (IMSI).
- Identification information like name, address, etc., of the subscriber.
- Identification information like Mobile Subscriber ISDN (MSISDN), etc.
- Billing information like prepaid or postpaid customer.
- Operator selected denial of service to a subscriber.

- Handling of supplementary services like for CFU (Call Forwarding Unconditional), CFB (Call Forwarding Busy), CFNR (Call Forwarding Not Reachable) or CFNA (Call Forwarding Not Answered).

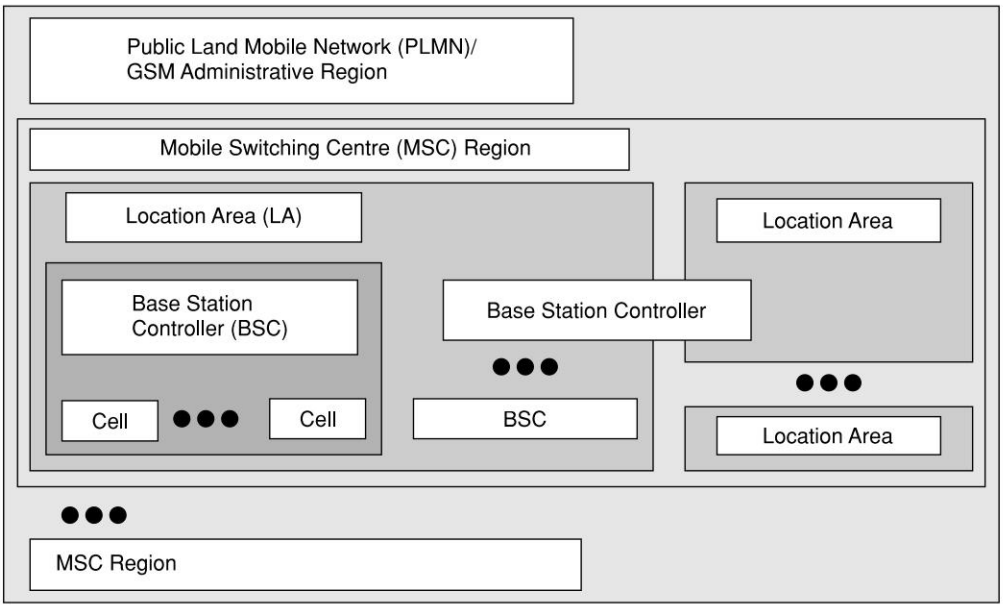


Figure 5.2 GSM System Hierarchy

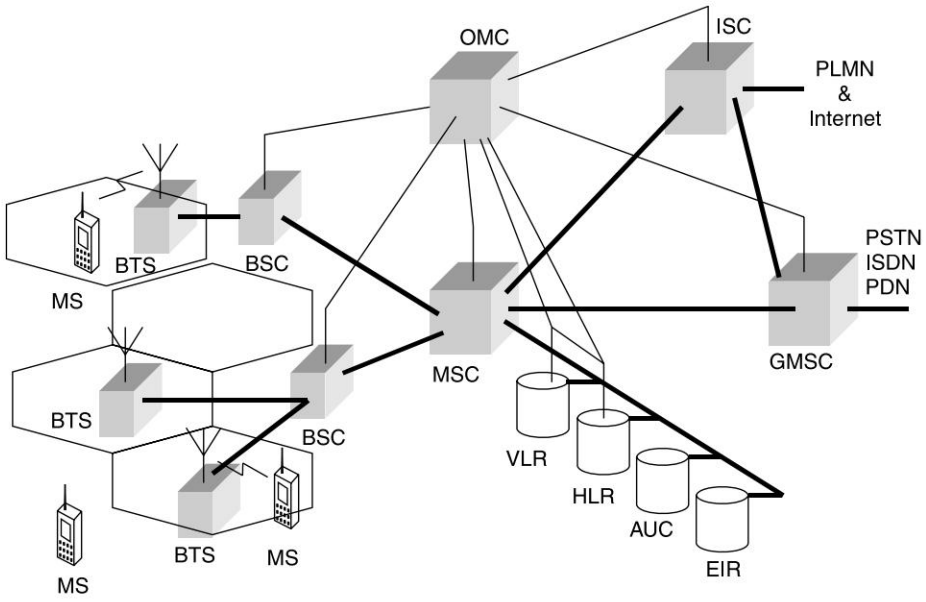
- Storage of SMS Service Center (SC) number in case the mobile is not connectable so that whenever the mobile is connectable, a paging signal is sent to the SC.
- Provisioning information like whether long distance and international calls are allowed or not.
- Provisioning information like whether roaming is enabled or not.
- Information related to auxiliary services like Voice mail, data, fax services, etc.
- Information related to auxiliary services like CLI (Caller Line Identification), etc.
- Information related to supplementary services for call routing. In GSM network, one can customize the personal profile to the extent that while the subscriber is roaming in a foreign PLMN, incoming calls can be barred. Also, outgoing international calls can be barred, etc.

There is some variable information, which could also be part of the HLR. This includes the pointer to the VLR, location area of the subscriber, Power OFF status of the handset, etc.

5.3 GSM Entities

The GSM technical specifications define different entities that form the GSM network by defining their functions and interface requirements. The GSM network can be divided into five main groups (Fig. 5.4):

- The Mobile Station (MS). This includes the Mobile Equipment (ME) and the Subscriber Identity Module (SIM).



AUC	Authentication Center
BSC	Base Station Controller
BTS	Base Transceiver Station
EIR	Equipment Identity Register
GMSC	Gateway MSC
HLR	Home Location Register
ISC	International Switching Center
ISDN	Integrated System Digital Network
MS	Mobile Station
MSC	Mobile Switching Center
OMC	Operation and Maintenance Center
PDN	Packet Data Network
PLMN	Public Land Mobile Network
PSTN	Public Switched Telephone Network
VLR	Visitor Location Register

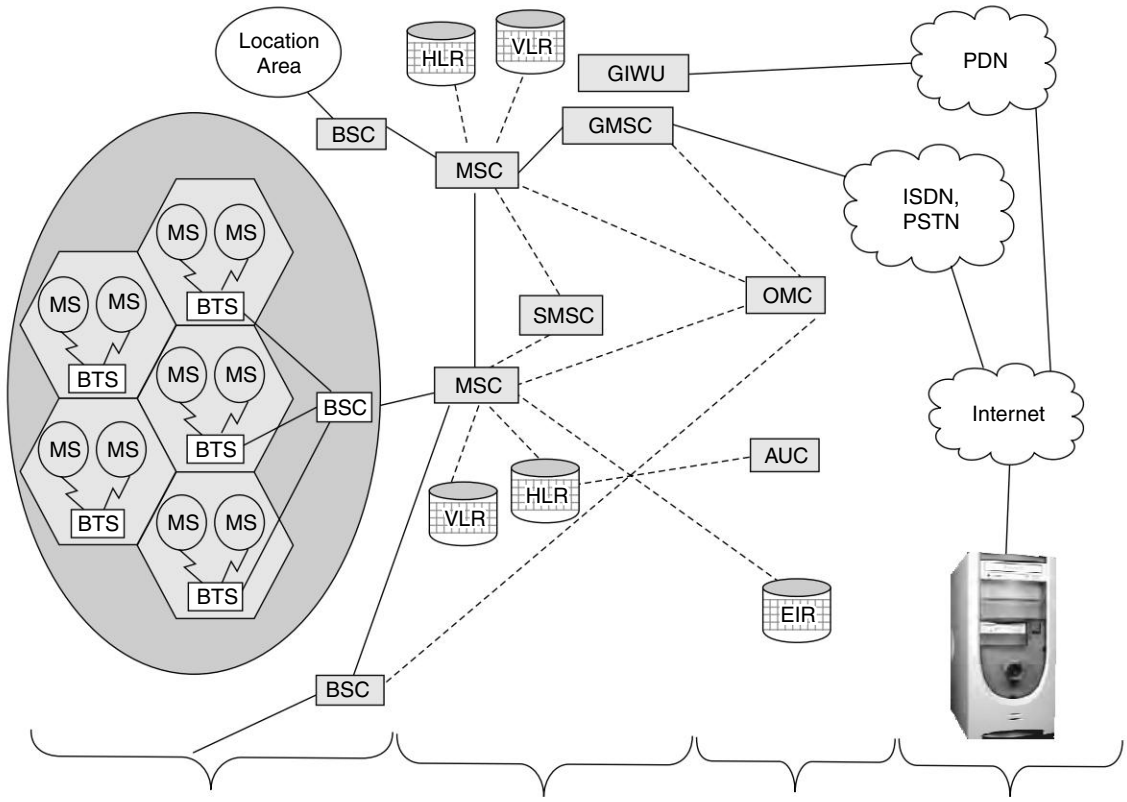
Figure 5.3 Architecture of GSM

- The Base Station Subsystem (BSS). This includes the Base Transceiver Station (BTS) and the Base Station Controller (BSC).
- The Network and Switching Subsystem (NSS). This includes Mobile Switching Center (MSC), Home Location Register (HLR), Visitor Location Register (VLR), Equipment Identity Register (EIR), and the Authentication Center (AUC).
- The Operation and Support Subsystem (OSS). This includes the Operation and Maintenance Center (OMC).
- The data infrastructure that includes Public Switched Telephone Network (PSTN), Integrated System Digital Network (ISDN), and the Public Data Network (PDN).

5.3.1 Mobile Station

Mobile Station is the technical name of the mobile or the cellular phone. In early days mobile phones were a little bulky and were sometimes installed in cars like other equipment. Even the

handheld terminals were quite big. Though the phones have become smaller and lighter, they are still called Mobile Stations. MS consists of two main elements:



AUC	Authentication Center	MS	Mobile Station
BSC	Base Station Controller	MSC	Mobile Switching Center
BTS	Base Transceiver Station	OMC	Operation and Maintenance Center
EIR	Equipment Identity Register	PDN	Public Data Network
GIWU	Gateway Interworking Unit	PSTN	Public Switched Telephone Network
GMSC	Gateway MSC	SC	Service Center/SMS Controller
HLR	Home Location Register	SMSC	Short Message Service Center
ISDN	Integrated System Digital Network	VLR	Visitor Location Register

Figure 5.4 System Architecture of GSM

- The mobile equipment or the mobile device. In other words, this is a phone without the SIM card.
- The Subscriber Identity Module (SIM).

There are different types of terminals distinguished principally by their power and application. The handheld GSM terminals have experienced the highest evolution. The weight and volume of

these terminals are continuously decreasing. The life of a battery between charging is also increasing. The evolution of technologies allowed decrease of power requirement to less than 1 W.

The SIM is installed in every GSM phone and identifies the terminal. Without the SIM card, the terminal is not operational. The SIM cards used in GSM phones are smart processor cards. These cards possess a processor and a small memory. By inserting the SIM card into the terminal, the user can have access to all the subscribed services. The SIM card contains the International Mobile Subscriber Identity (IMSI) used to identify the subscriber to the system, a secret key for authentication, and other security information. Another advantage of the SIM card is the mobility of the users. In fact, the only element that personalizes a terminal is the SIM card. Therefore, the user can have access to its subscribed services in any terminal using his or her SIM card. The SIM card may be protected against unauthorized use by a password or personal identity number. Typically, SIM cards contain 32 K bytes of memory. Part of the memory in the SIM card is available to the user for storing address book and SMS messages. Applications are developed and stored in SIM cards using SAT (SIM Application Toolkit). SAT is something similar to Assembly languages of computers and is proprietary to the SIM vendor. Nowadays Java Smart cards are coming to the market. In Java Smart card, the applications are written in Java language and are portable across SIM cards from different vendors.

5.3.2 The Base Station Subsystem

The BSS (Base Station Subsystem) connects the Mobile Station and the NSS (Network and Switching Subsystem). It is in charge of the transmission and reception for the last mile. The BSS can be divided into two parts:

- The Base Transceiver Station (BTS) or Base Station in short.
- The Base Station Controller (BSC).

The Base Transceiver Station corresponds to the transceivers and antennas used in each cell of the network. In a large urban area, a large number of BTSs are potentially deployed. A BTS is usually placed in the center of a cell. Its transmitting power defines the size of a cell. The BTS houses the radio transmitter and the receivers that define a cell and handles the radio-link protocols with the Mobile Station. Each BTS has between one and 16 transceivers depending on the density of users in the cell.

Base Station Controller is the connection between the BTS and the Mobile service Switching Center (MSC). The BSC manages the radio resources for one or more BTSs. It handles handovers, radio-channel setup, control of radio frequency power levels of the BTSs, exchange function, and frequency hopping.

5.3.3 The Network and Switching Subsystem

The central component of the Network Subsystem is the Mobile Switching Center (MSC). It does multiple functions. They are:

- It acts like a normal switching node for mobile subscribers of the same network (connection between mobile phone to mobile phone within the same network).
- It acts like a normal switching node for the PSTN fixed telephone (connection between mobile phone to fixed phone).

- It acts like a normal switching node for ISDN.
- It provides all the functionality needed to handle a mobile subscriber, such as registration, authentication, location updating, handovers and call routing.
- It includes databases needed in order to store information to manage the mobility of a roaming subscriber.

These different services are provided in conjunction with several functional entities, which together form the Network Subsystem. The signaling between functional entities in the Network Subsystem uses Signaling System Number 7 (SS7). SS7 is used for trunk signaling in ISDN and widely used in today's public networks. SS7 is also used for SMS, prepaid, roaming and other intelligent network functions.

The MSC together with Home Location Register (HLR) and Visitor Location Register (VLR) databases, provide the call-routing and roaming capabilities of GSM. The HLR is considered a very important database that stores information of subscribers belonging to the covering area of a MSC. Although a HLR may be implemented as a distributed database, there is logically only one HLR per GSM network. The HLR contains all the administrative information of each subscriber registered in the corresponding GSM network. This includes information like current location of the mobile, all the service provisioning information and authentication data. When a phone is powered off, this information is stored in the HLR. The location of the mobile is typically in the form of the signaling address of the VLR associated with the mobile station. HLR is always fixed and stored in the home network, whereas the VLR logically moves with the subscriber.

The VLR can be considered a temporary copy of some of the important information stored in the HLR. VLR is similar to a cache, whereas HLR is the persistent storage. The VLR contains selected administrative information borrowed from the HLR, necessary for call control and provisioning of the subscribed services. This is true for each mobile currently located in the geographical area controlled by a VLR. GSM standards define interfaces to HLR; however, there is no interface standard for VLR. Although each functional entity can be implemented as an independent unit, all manufacturers of switching equipment implement the VLR as an integral part of the MSC, so that the geographical area controlled by the MSC corresponds to that controlled by the VLR.

Note: MSC contains no information about a particular mobile station—this information is stored in location registers.

When a subscriber enters the covering area of a new MSC, the VLR associated with this MSC will request information about the new subscriber from its corresponding HLR in the home network. For example, if a subscriber of a GSM network in Bangalore is roaming in Delhi, the HLR data of the subscriber will remain in Bangalore with the home network, however, the VLR data will be copied to the roaming network in Delhi. The VLR will then have enough information in order to assure the subscribed services without needing to refer to the HLR each time a communication is established. Though the visiting network in Delhi will provide the services, the billing for the services will be done by the home network in Bangalore.

Within the NSS there is a component called Gateway MSC (GMSC) that is associated with the MSC. A gateway is a node interconnecting two networks. The GMSC is the interface between the mobile cellular network and the PSTN. It is in charge of routing calls from the fixed network towards a GSM user and vice versa. The GMSC is often implemented in the same node as the MSC. Like the GMSC, there is another node called GIWU (GSM Interworking Unit). The GIWU corresponds to an interface to various networks for data communications. During these communications, the transmission of speech and data can be alternated.

5.3.4 The Operation and Support Subsystem (OSS)

As the name suggests, Operations and Support Subsystem (OSS) controls and monitors the GSM system. The OSS is connected to different components of the NSS and to the BSC. It is also in charge of controlling the traffic load of the BSS. However, the increasing number of base stations, due to the development of cellular radio networks, has resulted in some of the maintenance tasks being transferred to the BTS. This transfer decreases considerably the costs of maintenance of the system. Provisioning information for different services is managed in this subsystem.

Equipment Identity Register (EIR) is a database that contains a list of all valid mobile equipment within the network, where each mobile station is identified by its International Mobile Equipment Identity (IMEI). EIR contains a list of IMEIs of all valid terminals. An IMEI is marked as invalid if it has been reported stolen or is not type approved. The EIR allows the MSC to forbid calls from this stolen or unauthorized terminals.

Authentication Center (AUC) is responsible for the authentication of a subscriber. This is a protected database and stores a copy of the secret key stored in each subscriber's SIM card. These data help to verify the user's identity.

5.3.5 Message Centre

Short Message Service or SMS is one of the most popular services within GSM. SMS is a data service and allows a user to enter text message up to 160 characters in length when 7-bit English characters are used. It is 140 octets when 8-bit characters (some European alphabets or binary data) are used, and 70 characters in length when non-Latin alphabets such as Arabic, Chinese or Hindi are used (70 characters of 16-bit Unicode). SMS is a proactive bearer and is an always ON network. Message center is also referred to as Service Centre (SC) or SMS Controller (SMSC). SMSC is a system within the core GSM network, which works as a store and forward system for SMS messages. Refer to Figure 5.5 for SMS architecture.

There are two types of SMS, SMMT (Short Message Mobile Terminated Point-to-Point), and SMMO (Short Message Mobile Originated Point-to-Point). SMMT is an incoming short message from the network and is terminated in the MS (phone or Mobile Station). SMMO is an outgoing message, originated in the MS, and forwarded to the network for delivery. For an outgoing message, the SMS is sent from the phone to SC via the VLR and the Interworking MSC (IWMSC). For incoming SMS message the path is from SC to the MS via the HLR and the Gateway MSC (GMSC). Please see Chapter 6 for SMS and related technologies.

5.4 Call Routing in GSM

Human interface is analog. However, the advancement in digital technology makes it very convenient to handle information in digital fashion. In GSM there are many complex technologies used between the human analog interface in the mobile and the digital network (Fig. 5.6).

Digitizer and source coding: The user speech is digitized at 8 KHz sampling rate using Regular Pulse Excited–Linear Predictive Coder (RPE–LPC) with a Long Term Predictor loop. In this technique, information from previous samples is used to predict the current sample. Each sample

is then represented in signed 13-bit linear PCM value. This digitized data is passed to the coder with frames of 160 samples. The encoder compresses these 160 samples into 260-bits GSM frames resulting in one second of speech compressed into 1625 bytes and achieving a rate of 13 Kbits/sec.
Channel coding: This step introduces redundancy information into the data for error detection and possible error correction. The gross bit rate after channel coding is 22.8 kbps (or 456 bits every 20 ms). These 456 bits are divided into eight 57-bit blocks, and the result is interleaved amongst eight successive time slot bursts for protection against burst transmission errors.

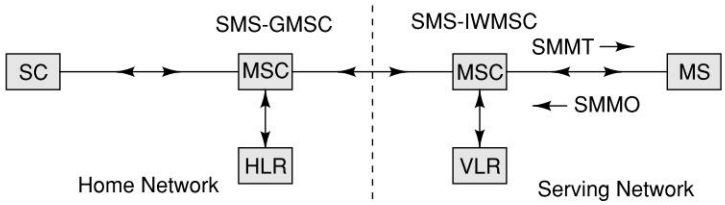


Figure 5.5 The Network Structure for the Short Message Transfer

Interleaving: This step rearranges a group of bits in a particular way. This is to improve the performance of the error-correction mechanisms. The interleaving decreases the possibility of losing whole bursts during the transmission, by dispersing the errors.

Ciphering: Encrypts blocks of user data using a symmetric key shared by the mobile station and the BTS.

Burst formatting: Adds some binary information to the ciphered block. This additional information is used for synchronization and equalization of the received data.

Modulation: The modulation technique chosen for the GSM system is the Gaussian Minimum Shift Keying (GMSK). Using this technique the binary data is converted back into analog signal to fit the frequency and time requirements for the multiple access rules. This signal is then radiated as radio wave over the air. Each time slot burst is 156.25 bits and contains two 57-bit blocks, and a 26-bit training sequence used for equalization (Fig. 5.6). A burst is transmitted in 0.577 ms for a total bit rate of 270.8 Kbps.

Multipath and equalization: At the GSM frequency bands, radio waves reflect from buildings, cars, hills, etc. So not only is the "right" signal (the output signal of the emitter) received by an antenna, but many reflected signals, which corrupt the information, with different phases are also received. An equaliser is in charge of extracting the "right" signal from the received signal. It estimates the channel impulse response of the GSM system and then constructs an inverse filter. In order to extract the "right" signal, the received signal is passed through the inverse filter.

Synchronization: For successful operation of a mobile radio system, time and frequency synchronization are needed. Frequency synchronization is necessary so that the transmitter and receiver frequency match (in FDMA). Time synchronization is necessary to identify the frame boundary and the bits within the frame (in TDMA).

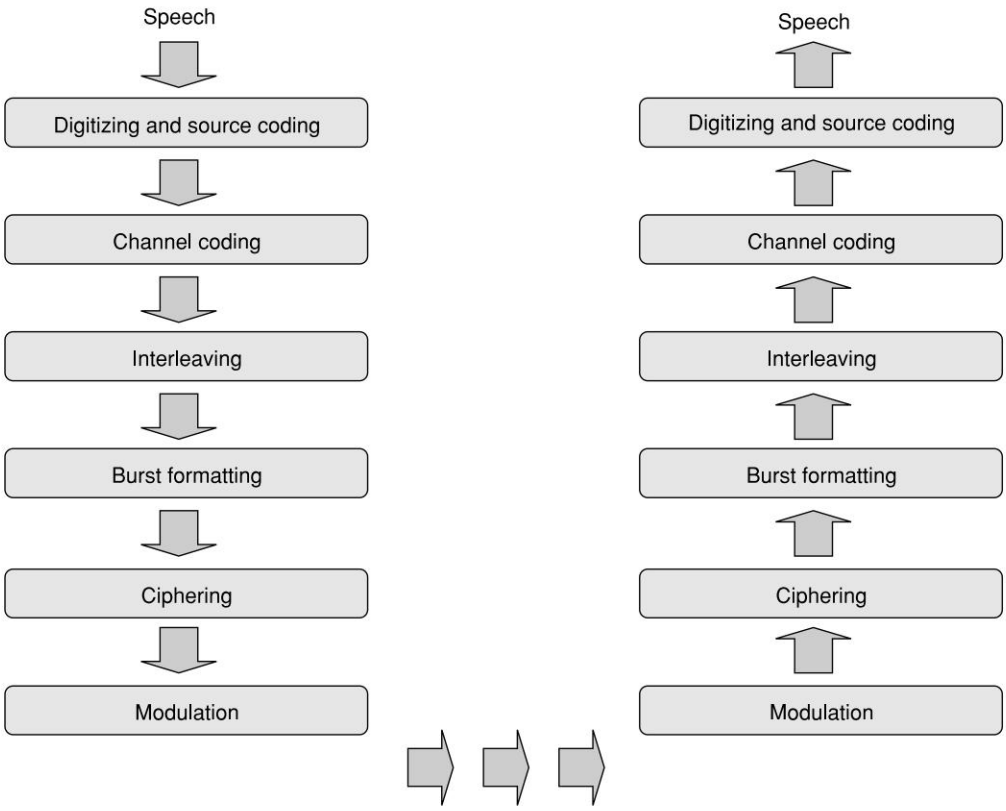


Figure 5.6 Sequence of Operation from Speech to Radio Wave

The mobile station can be anywhere within a cell. Also, the distance between the base station and the mobile station vary. Due to mobility of the subscriber, the propagation time between the base station and the mobile keeps varying. When a mobile station moves further away, the burst transmitted by this mobile may overlap with the time slot of the adjacent time slot. To avoid such collisions, the Timing Advance technique is used. In this technique, the frame is advanced in time so that this offsets the delay due to greater distance. Using this technique and the triangulation of the intersection cell sites, the location of a mobile station can be determined from within the network.

5.4.1 An Example

In this section let us take an example of how and what happens within the GSM network when someone from a fixed network calls someone in a GSM network. Let us assume that the called party dialed a GSM directory number +919845052534. Figure 5.7 depicts the steps for this call processing.

The directory number dialed to reach a mobile subscriber is called the Mobile Subscriber ISDN (MSISDN), which is defined by the E.164 numbering plan. This number includes a country code and a National Destination Code, which identifies the subscriber's operator. The first few digits of the remaining subscriber number may identify the subscriber's HLR within the home PLMN. For example, the MSISDN number of a subscriber in Bangalore associated with Airtel network is +919845XYYYYY. This is a unique number and understood from anywhere in the world. In this example + means the prefix for international dialing like 00 in UK/India or 011 in USA. 91 is the country code for India (404 as defined in GSM). 45 is the network operator's code (Airtel in this case). X is the level number managed by the network operator ranging from 0 to 9. YYYYY is the subscriber code managed by the operator as well.

The call first goes to the local PSTN exchange. The PSTN exchange looks at the routing table and determines that it is a call to a mobile network. It forwards the call to the Gateway MSC (GMSC) of the mobile network. The MSC enquires the HLR to determine the status of the subscriber. It will decide whether the call is to be routed or not. If the user has not paid the bills, the call may not be routed. If the phone is powered off, a message may be played or forwarded to the voice mail. However, if MSC finds that the call can be processed, it will find out the address of the VLR where the mobile is expected to be present. If the VLR is that of a different PLMN, it will forward the call to the foreign PLMN through the Gateway MSC. If the VLR is in the home network, it will determine the Location Area (LA). Within the LA it will page and locate the phone and connect the call.

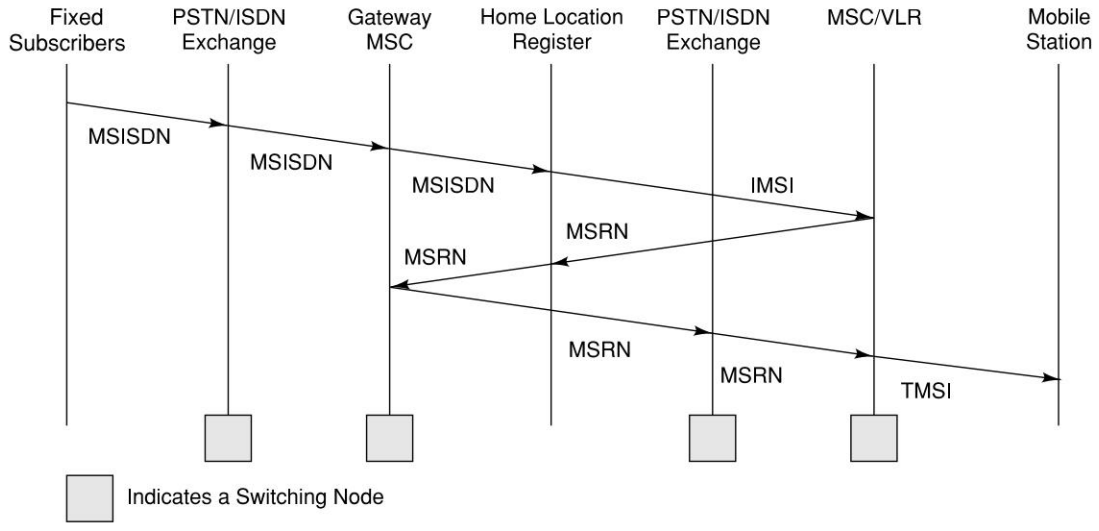
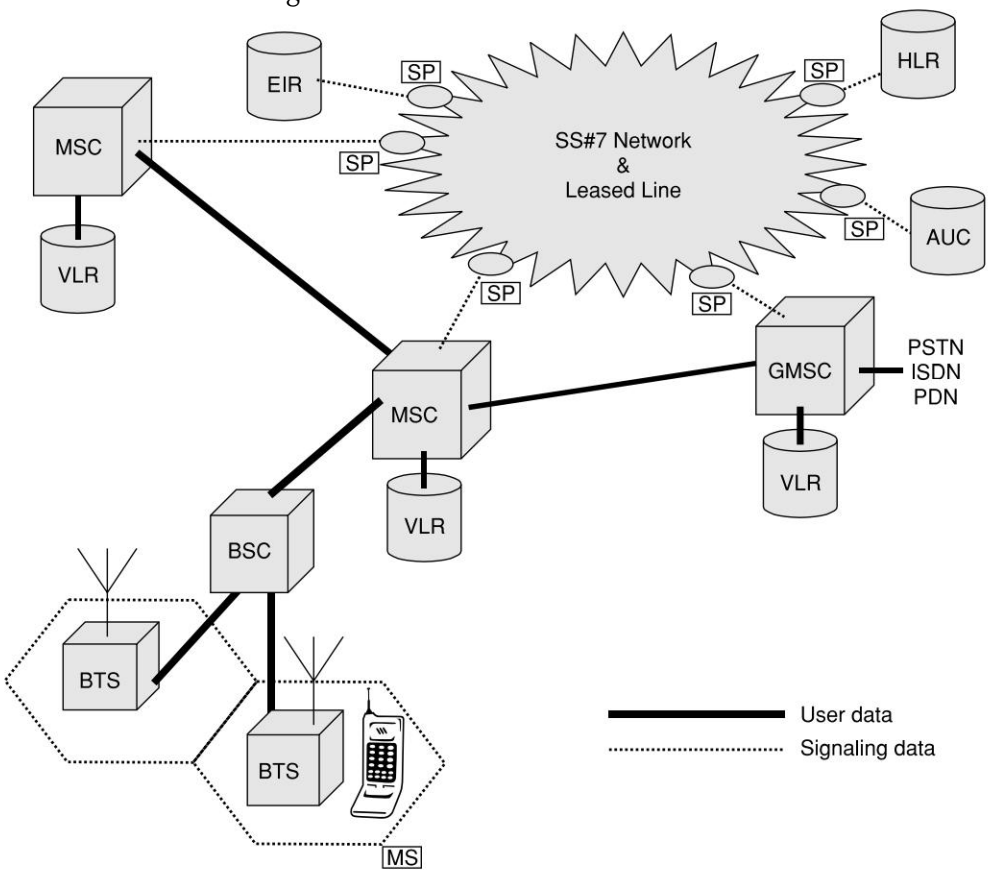


Figure 5.7 Call Routing for a Mobile Terminating Call

5.5 PLMN INTERFACES

The basic configuration of a GSM network contains a central HLR and a central VLR. HLR contains all security, provisioning and subscriber-related information. VLR stores the location information and other transient data. MSC needs subscriber parameter for successful call set-up. Figure 5.8 shows a basic configuration of a GSM mobile communication network.



AUC	Authentication Center	ISDN	Integrated System Digital Network
BSC	Base Station Controller	MS	Mobile Station
BTS	Base Transceiver Station	MSC	Mobile Switching Center
EIR	Equipment Identity Register	PDN	Public Data Network
GMSC	Gateway MSC	PSTN	Public Switched Telephone Network
HLR	Home Location Register	SP	Signaling Point
		VLR	Visitor Location Register

Figure 5.8 Configuration of a GSM PLMN

Within the switching and management system, the transmission rate is 2 Mbits/s. This 2 Mbits/ interface is called E1 interface in India and in Europe. These are realized typically through

microwave or leased lines. Any data related to user call (connection, teardown, etc.) are processed with SS7 protocol for signaling using ISUP (ISDN User Part) stack between network nodes. For mobile specific signaling a protocol stack called MAP (Mobile Application Part) is used over the SS7 network. All database transactions (enquiries, updates, etc.) and handover/roaming transactions between the MSC are performed with the help of MAP. For this purpose, each MSC uses registers known as SP (Signaling Point). These SPs are addressable through a unique code called Signaling Point Code (SPC). Signaling between MSC and BSS uses Base Station System Application Part (BSSAP) over SS7. Within BSS and at the air interface, signaling is GSM proprietary and does not use SS7.

5.6 GSM Addresses and Identifiers

GSM distinguishes explicitly between the user and the equipment. It also distinguishes between the subscriber identity and the telephone number. To manage all the complex functions, GSM deals with many addresses and identifiers. They are:

- *International Mobile Station Equipment Identity (IMEI):* Every mobile equipment in this world has a unique identifier. This identifier is called IMEI. The IMEI is allocated by the equipment manufacturer and registered by the network operator in the Equipment Identity Register (EIR). In your mobile handset you can type *#06# and see the IMEI.
- *International Mobile Subscriber Identity (IMSI):* When registered with a GSM operator, each subscriber is assigned a unique identifier. The IMSI is stored in the SIM card and secured by the operator. A mobile station can only be operated when it has a valid IMSI. The IMSI consists of several parts. These are:
  - ❏ Three decimal digits of Mobile Country Code (MCC). For India the MCC is 404.
  - ❏ Two decimal digits of Mobile Network Code (MNC). This uniquely identifies a mobile operator within a country. For Airtel in Delhi this code is 10.
  - ❏ Maximum 10 decimal digits of Mobile Subscriber Identification Number (MSIN). This is a unique number of the subscriber within the home network.
- *Mobile Subscriber ISDN Numbers (MSISDN):* The MSISDN number is the real telephone number as is known to the external world. MSISDN number is public information, whereas IMSI is private to the operator. This is a number published and known to everybody. In GSM a mobile station can have multiple MSISDN numbers. When a subscriber opts for fax and data, he is assigned a total of three numbers: one for voice call, one for fax call and another for data call. The MSISDN categories follow the international ISDN (Integrated Systems Data Network) numbering plan as the following:
  - ❏ Country Code (CC): One to three decimal digits of country code.
  - ❏ National Destination Code (NDC): Typically 2 to 3 decimal digits.
  - ❏ Subscriber Number (SN): Maximum 10 decimal digits.

The CC is standardized by the ITU-T through the E.164 standard. There are CCs with one, two, or three digits. For example, the CC for USA is 1, for India it is 91, and for Finland it is 358.The national regulatory authority assigns the NDC. In India it is 94 for BSNL and 98 for all other operators. In India the subscriber number SN is eight decimal digits. SN consists of two decimal

digits of operator code, followed by one decimal digit level number with a five decimal digit subscriber number. In India, a MSISDN number looks like 919845062050. In this number 91 is the CC, 98 is the NDC, and 45062050 is the SN. In India, the SN is subdivided into operator code and subscriber code (45 is the operator code and 062050 is the subscriber code). Sometimes subscriber code is also subdivided into one digit level number (0 in this case) followed by five digit subscriber ID (62050).

- *Location Area Identity:* Each LA in a PLMN has its own identifier. The Location Area Identifier (LAI) is structured hierarchically and unique. LAI consists of three digits of CC, two digits of Mobile Network Code and maximum five digits of Location Area Code.
- *Mobile Station Roaming Number (MSRN):* When a subscriber is roaming in another network a temporary ISDN number is assigned to the subscriber. This ISDN number is assigned by the local VLR in charge of the mobile station. The MSRN has the same structure as the MSISDN.
- *Temporary Mobile Subscriber Identity (TMSI):* This a temporary identifier assigned by the serving VLR. It is used in place of the IMSI for identification and addressing of the mobile station. TMSI is assigned during the presence of the mobile station in a VLR and can change (ID hopping). Thus, it is difficult to determine the identity of the subscriber by listening to the radio channel. The TMSI is never stored in the HLR. However, it is stored in the SIM card. Together with the current location area, a TMSI allows a subscriber to be identified uniquely. For an ongoing communication the IMSI is replaced by the 2-tuple LAI, TMSI code.
- *Local Mobile Subscriber Identity (LMSI):* This is assigned by the VLR and also stored in the HLR. This is used as a searching key for faster database access within the VLR.
- *Cell Identifier:* Within a LA, every cell has a unique Cell Identifier (CI). Together with a LAI a cell can be identified uniquely through Global Cell Identity (LAI+CI).
- *Identification of MSCs and Location Registers:* MSCs, Location Registers (HLR, VLR), SCs are addressed with ISDN numbers. In addition, they may have a Signaling Point Code (SPC) within a PLMN. These point codes can be used to address these nodes uniquely within the Signaling System number 7 (SS#7) network.

5.7 Network Aspects in GSM

Transmission of voice and data over the radio link is only a part of the function of a cellular mobile network. A GSM mobile can seamlessly roam nationally and internationally. This requires that registration, authentication, call routing and location updating functions are standardized across GSM networks. The geographical area covered by a network is divided into cells of small radius. When a call is in progress and the user is on the move, there will be a handover mechanism from one cell to another. This is like a relay race where one athlete passes on the baton to another. Though both roaming and handover functions are the basic characteristic of mobility, there is a difference between these functions. These functions are performed by the Network Subsystem, mainly using the Mobile Application Part (MAP) built on top of the Signalling System # 7 (SS7) protocol (Fig. 5.9).

The signaling protocol in GSM is structured into three general layers, depending on the interface, as shown in Figure 5.9. Layer 1 is the physical layer, which uses the channel structures over the air interface. Layer 2 is the data link layer. Across the Um interface, the data link layer is a modified version of the LAPD protocol used in ISDN or X.25, called LAPDm. Across the A interface, the Message Transfer Part layer 2 of Signaling System Number 7 is used. Layer 3 of the GSM signaling protocol is itself divided into three sublayers.

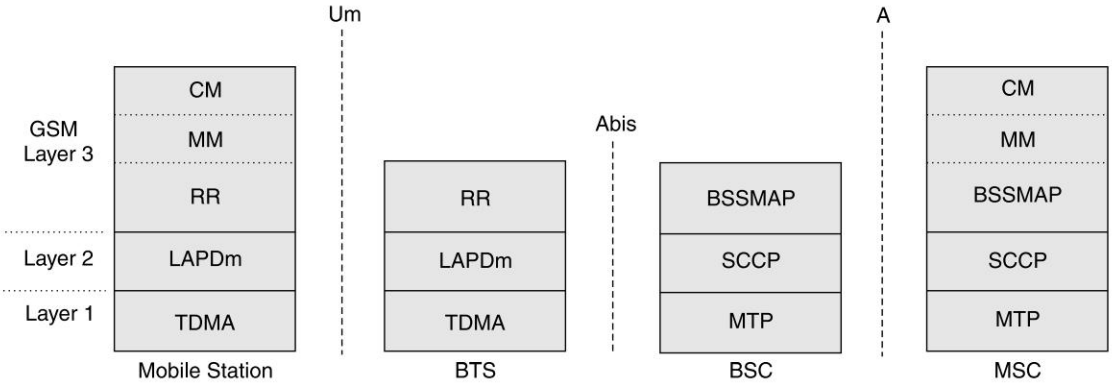


Figure 5.9 Signaling Protocol Structure in GSM

- *Radio Resources Management:* Controls the set-up, maintenance, and termination of radio and fixed channels, including handovers.
- *Mobility Management:* Manages the location updating and registration procedures as well as security and authentication.
- *Connection Management:* Handles general call control, similar to CCITT Recommendation Q.931 and manages Supplementary Services and the Short Message Service.

Signaling between the different entities in the fixed part of the network, such as between the HLR and VLR, is accomplished through the Mobile Application Part (MAP). MAP is built on top of the Transaction Capabilities Application Part (TCAP, the top layer of SS7). SS7 is also used for many other Intelligent Network services within the GSM. The specification of the MAP is quite complex, and at over 500 pages, it is one of the longest documents in the GSM recommendations.

5.8 Mobility Management

Think about a world in 1970s—a person traveling by train without having a GSM mobile phone or wireless communication device. If he wants to talk to someone (call the doctor, for example, for some critical healthcare help) or someone wants to talk (incoming call) to this person—how can this conversation take place? If the train stops at a station for a long duration, the person could try to locate a public call office in the station and make the call; but he must finish the call before the train departs. Unfortunately there was no way for someone to call this person (incoming call). Mobility

Management (MM) solves all these challenges. Using MM one can make outgoing calls and receive incoming calls while in motion; even at the vehicular state where the speed is higher than 60 kmph. The MM function handles all functions that arise from the mobility of the subscriber.

In a wired network where the device is stationary, the point of attachment to the network is fixed; here, address of the device is sufficient to locate the device in the routing table and establish a connection. However, in a wireless or mobile environment the point of attachment constantly changes, the device moves from one location to another location making the old routing table invalid; therefore, establishing and maintaining a connection is complex. As long as there is a wireless network with available channels, mobile originated outgoing calls are relatively easy to handle; for mobile terminated incoming calls, however, Paging and Location Updates are necessary. Also, Handover and Roaming are two important aspects in mobility. We will discuss mobility management in the following sections.

5.8.1 Paging

For a mobile terminated call, the MS needs to be traced, located, and then the call connected. The MS is traced through the Paging process within a location. Using the BSS signaling channel the Paging message for an MS is sent that includes the IMSI as the identifier of the MS. The message may also include an indication of which combination of channels will be needed for the subsequent transaction related to the paging. A single paging message across the MSC to BSS interface contains information of the cells in which the page shall be broadcast.

In Paging, the most difficult part of the decision is—which cell to start the paging from; because a cellular network may be spread over thousands of square-kilometres with thousands of cells. If we cannot locate the mobile quickly, the call cannot be connected resulting in lost revenue. For example, it can start at the center of the network and keep on searching each and every cell for a long time. However, such global paging is very expensive in terms of backbone and radio signaling channels. Also, global paging will take enormous amount of time. To optimize the cost and response time, paging starts at the location where the MS was present last. The location of the MS is recorded in the HLR and updated through Location Update. The MS is searched in these cells where it has the highest probability of being present. There are various algorithms for paging so that the MS can be located quickly with minimum effort and cost.

5.8.2 Location Update

Location update is concerned with the procedures that enable the network to know the current location of a powered-on MS so that the mobile terminated call routing can be completed. If the location of the MS is not known, tracking the MS through paging costs in terms of radio and backbone SS7 signalling (see Chapter 11) bandwidth. To optimize this, location information is regularly updated within the core network. Through location update, the presence and location information is kept up-to-date within the VLR and the HLR. Presence deals with willingness and availability of an MS for communication. Assuming that the MS is willing to communicate, the MS must be powered-on and attached to the network. If the MS is attached to the network, it must be located through Paging before a successful communication can take place for mobile terminated calls and mobile terminated SMS.

When an MS is powered-off, the HLR is updated with an explicit IMSI detach. IMSI detach is equivalent to HLR data for the particular IMSI being unavailable; in this case logically the MS is not available and a connection cannot be established. However, the MS may be powered-on but may not be successfully connected to the network for an operator defined interval; this could be due to MS being out of coverage area—in such a case an implicit IMSI detach occurs.

To complete a call, IMSI must stay attached with a VLR and HLR. IMSI attach is accomplished through location updates. Location update can be initiated by either MS or the network. Frequent location update costs in terms of power in the MS and the SS7 signalling traffic; therefore, location update is restricted to the following conditions:

- When there is a mobile originated outgoing call, the location information is updated in the VLR and the HLR.
- When the MS moves from one Location Area (LA) to another LA the location information in the VLR and the HLR is updated. In Figure 5.2, we illustrated the hierarchical GSM networks with at least one administrative region, which is assigned to an MSC. Each administrative region is made up of multiple location area (LA) and each location area consists of several cell groups.
- The MS updates the location co-ordinates when its location is more than "k" cells away in distance from the location information of the last update.
- The MS updates the location information when it crosses exactly "k" cell boundaries irrespective of the distance.
- In addition, there is an explicit periodic location update by the MS. This time period is defined by the GSM network operator and is within the range of 1 deci-hour (6 minutes) to 240 deci-hours (24 hours) with a granularity of 1 deci-hour.

5.8.3 Handover

In a cellular network, while a call is in progress, the relationship between radio signal and the user is dynamic. User movements may make a user move away from a wireless tower, causing the radio signal strength to reduce, and ultimately break. Therefore, the user needs to be moved to another cell where the signal strength is higher. This will result in changing the association of resources to another channel within the same cell or a different cell altogether. This procedure of changing the resources is called handover. The handover needs to be very fast without any disruption to the service at the higher layer. This handover procedure is called "handoff" in North America. The handover can be initiated either by the MS or the network. A mobile initiated handover is based on radio subsystem criteria of Radio Frequency (RF) level, quality of the radio signal, or the distance from the tower; whereas, the network initiated handover that is assisted by a mobile device is based on the current traffic loading per cell, maintenance requests, etc.

There are four different types of handover in the GSM system, which involve transferring a call between:

- Channels (time slots) in the same cell.
- Cells (BTS) under the control of the same BSC.
- Cells under the control of different BSCs, but belonging to the same MSC.
- Cells under the control of different MSCs.

The first two types of handover, called internal handovers, involve only one BSC. To save SS7 signaling bandwidth, they are managed by the BSC without involving the MSC, except to notify it at the completion of the handover. The last two types of handover, called external handovers, are handled by the MSC. In order to determine whether a handover is required, due to RF criteria, the mobile shall take radio measurements from neighboring cells. These measurements are reported to the serving cell to determine a need for a handover. Additionally, the handover decision by the network may take into account both the measurement results from the MS and network directed criteria. The same decision process is used to determine when to perform both the intra-MSC and inter-MSC handover.

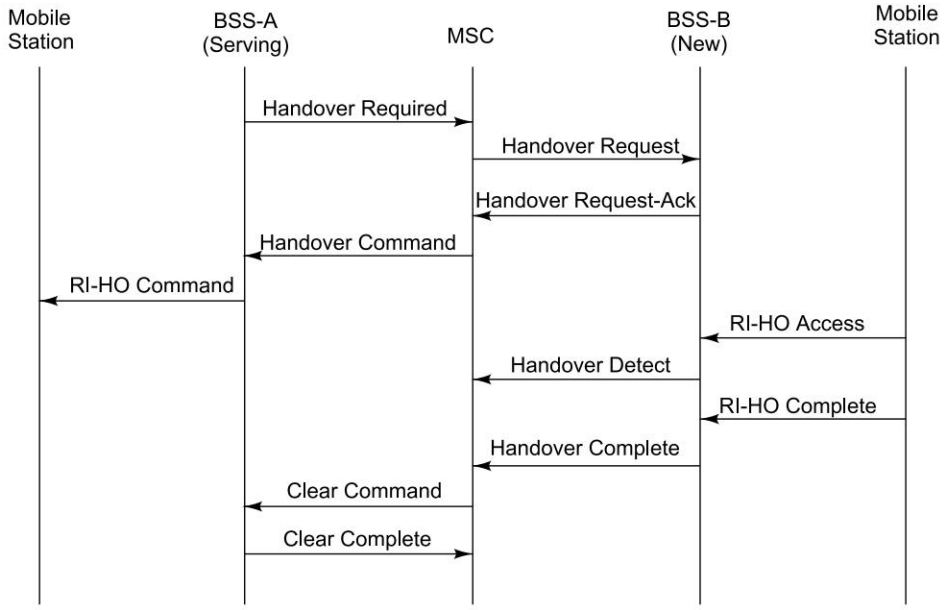


Figure 5.10 Handover Procedure

Figure 5.10 illustrates the handover procedure in GSM. The currently serving BSS sends a Handover_Required message to the MSC; the MSC sends a Handover_Request message to the new BSS from which it requires radio resources. This message contains details of the resource that is required. The message may also specify the channel in use. On receipt of this message the new BSS shall choose a suitable idle radio resource. This information is passed by the new BSS to the MS through MSC and old BSS using Handover_Request_Acknowledgement, Handover_Command, and Radio_Interface_Handover_Command respectively. The MS changes its association from the old BSS to the new BSS with a Handover_Access burst which contains the received handover reference number. The new BSS checks the handover reference number to ensure that it is the same as expected, and that there is a high probability that the correct MS has been captured. When the MS is successfully in communication with the network, i.e., the Radio Resource (RR) message Handover_Complete has been received from the MS, then the new BSS will immediately send a BSSMAP message Handover_Complete to the MSC and terminate the

procedure. The MSC in this case will terminate the procedure with the old BSS by sending a Clear_Command with cause "Handover successful". When the MS is successfully in communication with the network, i.e., the RR message Handover_Complete has been received from the MS, then the new BSS will immediately send a BSSMAP message Handover_Complete to the MSC and terminate the procedure. The MSC in this case will terminate the procedure with the old BSS by sending a Clear_Command with cause "Handover successful".

5.8.4 Authentication and Security Issues during Handover

GSM uses A3, A8, and A5 algorithms (see Section 5.10) for security. A3 algorithm is used to authenticate the subscriber; A8 algorithm is used to generate the ciphering key K_c; and, A5 algorithm is used to cipher everything that is transmitted over the air that include both signal and traffic. Security issues in GSM network are covered in detail in GSM standard 03.20.

When a handover occurs, the necessary information (e.g., key K_c, initialization data) is transmitted within the system infrastructure to enable the communication to proceed from the old BSS to the new BSS, and the synchronization procedure is resumed. The key K_c remains unchanged at handover.

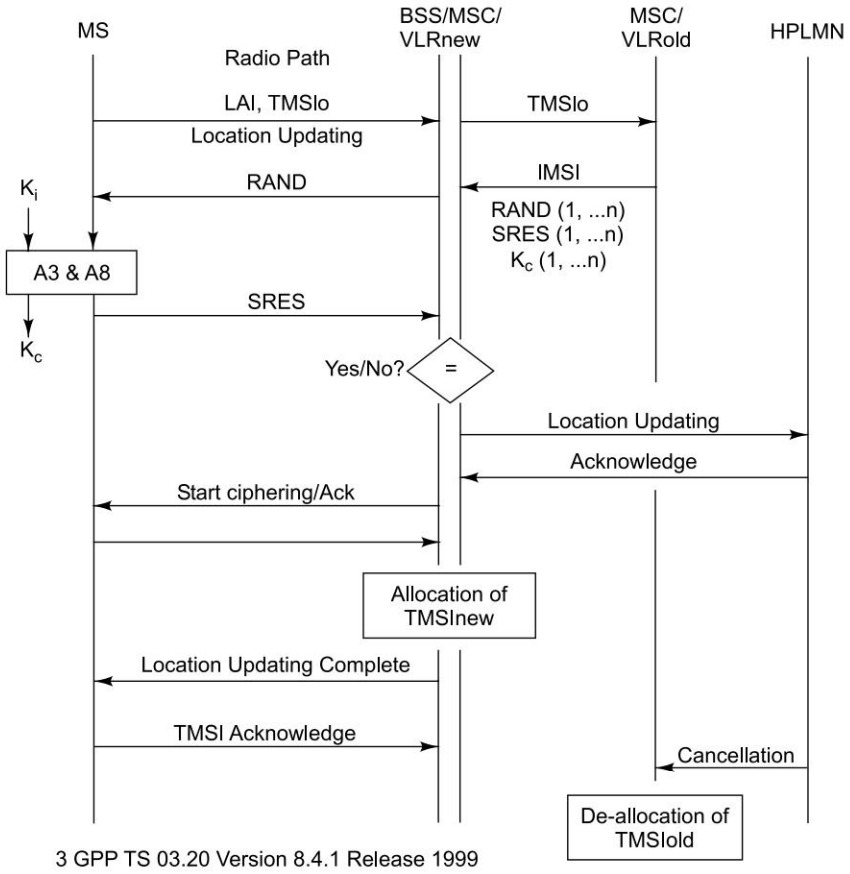


Figure 5.11 Normal Location Updating Procedure (Fig. 5.1)

Figure 5.11 illustrates the normal location updating procedure with all elements pertaining to security functions, i.e., TMSI (Temporary Mobile Subscriber Identity) management, authentication and K_c management. Here it is assumed that during the handover the MSC/VLR is changed from old VLRo to new VLRn. During the Location Update the MS sends the LAI (Location Area Identifier) and the old TMSIo to the old VLRo. The VLRo sends the series of challenges RAND (1, ...n), and their respective answers SRES (1, ...n) of challenges, and the respective ciphering keys Kc (1, ...n) with the IMSI of the MS. VLRo receives all these RAND challenges from the HPLMN (Home Public Land Mobile Network). If the authentication is successful, the HLR is updated with new location; the ciphering starts with new Kc and a new TMSI is allocated. As part of housekeeping, the new VLRn is registered in the HLR; the HLR also informs the VLRo to de-register the IMSI. The VLRo deletes all entries related to this IMSI including the TMSIo.

5.8.5 Roaming

Handover relates to moving from one point of attachment to another point of attachment within the same network operator; when this movement happens between two different networks it is called roaming. Different networks imply two separate billing and charging domains.

When a mobile station is powered-off, it performs an IMSI detach procedure in order to tell the network that it is no longer connected. When a mobile station is switched on in a new network (for example, the user has disembarked from an aircraft in a new country) or the subscriber moves to a different operator's PLMN (Public Land Mobile Network), the subscriber must register with the new network to indicate its current location. The first location update procedure is called the IMSI attach procedure where the MS indicates its IMSI to the new network. Normally, a location update message is sent to the new MSC/VLR, which records the location area information, and then sends the location information to the subscriber's HLR. If the mobile station is authenticated and authorized in the new MSC/VLR, the subscriber's HLR cancels the registration of the mobile station with the old MSC/VLR. A location update is also performed periodically. If after the updating time period, the mobile station has not registered, it is then deregistered.

Roaming is a killer application in GSM that allows users to seamlessly move around nationally and internationally and remain connected. Unlike routing in the fixed network, where a terminal is semi-permanently wired to a central office, GSM allows roaming around the world. When there is an incoming call for a subscriber, the mobile phone needs to be located, a channel needs to be allocated and the call connected. A powered-on mobile is informed of an incoming call by a paging message sent over the paging channel of the cells within the current location area. The location updating procedures, and subsequent call routing, use the MSC and both HLR and the VLR. The information sent to the HLR is normally the SS7 address of the new VLR. If the subscriber is entitled to service, the HLR sends a subset of the subscriber information needed for call control to the new MSC/VLR, and sends a message to the old MSC/VLR to cancel the old registration.

An incoming mobile terminating call is directed to the Gateway MSC (GMSC) function. The GMSC is basically a switch, which is able to interrogate the subscriber's HLR to obtain routing information and thus contains a table linking MSISDNs to their corresponding HLR. A simplification is to have a GSMC handle one specific PLMN. Though the GMSC function is distinct from the MSC function, it is usually implemented within an MSC. The routing information that is returned to the GMSC is the Mobile Station Roaming Number (MSRN), which is also

defined in the E.164 numbering plan that includes CC (Country Code), NDC (National Destination Code), and SN (Subscriber Number) (Fig. 5.7).

MSRN is a temporary location-dependent MSISDN number. It is assigned by the serving VLR for each MS in its area. MSRNs are numbers reserved by a PLMN only for roaming use; and, not assigned to subscribers, nor are they visible to subscribers. The allocation of MSRN is done in such a fashion that the currently responsible MSC in the visited network (CC+NDC) can do routing of the call quite easily.

The most general routing procedure begins with the GMSC querying the called subscriber's HLR for an MSRN (Fig. 5.7). The HLR typically stores only the SS7 address of the subscriber's current VLR. The VLR temporarily allocates an MSRN from its pool for the call. This MSRN is returned to the HLR and back to the GMSC, which can then route the call to the new MSC. At the new MSC, the IMSI corresponding to the MSRN is looked up, and the mobile is paged in its current location area. As a rule of thumb, HLR is referred for incoming call; whereas VLR is referred for outgoing call.

Roaming is of two types. These are:

- *Horizontal Roaming:* Horizontal roaming is between two networks from same family. For example, GSM to GSM roaming or GSM to UMTS roaming will be considered horizontal roaming.
- *Vertical Roaming:* Vertical roaming is between two networks from different families. For example, GSM to CDMA roaming or GPRS to WiFi will be considered vertical roaming. When vertical roaming happens without any disruption of session or service, it is called Seamless Roaming.

5.8.6 Roaming Example

Let us assume that the user's mobile number is +919844012345. This is a number in Spice network in Bangalore. The mobile subscriber is roaming in Mumbai. Somebody from a fixed phone in Mumbai wants to talk to this Spice subscriber. The caller (also known as 'A' party) dials 09844012345 from Mumbai. This call will be switched at the PSTN network in Mumbai and will be routed to Spice network in Bangalore. The Spice MSC will look at the HLR and know that the subscriber (called 'B' party) is now within the coverage of a mobile operator (Vodafone) in Mumbai—this is done using the MSRN. The call will be routed to the Mumbai MSC at Vodafone. The Vodafone MSC at Mumbai will look at its VLR to locate the Spice subscriber and route the call. Also, when the call is over, the charging information will be forwarded to the Spice network. Please note that for the incoming call, the routing always happens via the home network resulting in the call routing from Mumbai PSTN to Bangalore PLMN to Mumbai PLMN. The calling party (person in Mumbai) pays long distance tariff for Mumbai PSTN to Bangalore PLMN; the called party (Spice subscriber) pays for Bangalore PLMN to Mumbai PLMN long distance tariff in addition to roaming airtime charges. For outgoing call, the home network is not referred (other than the first time authentication), resulting in the call being directly routed by the visiting network. Let us consider the opposite scenario; the Spice subscriber from Bangalore is still roaming in Mumbai and wants to call someone in Mumbai. The Spice subscriber dials the Mumbai number, the Vodafone MSC looks at the VLR and routes the call directly to the Mumbai number. In this case, the Spice subscriber pays a local Mumbai to Mumbai call charge in addition to the airtime charges.

Let us now look at a completely different scenario where both the caller and the called party are roaming in a foreign network. Let us assume that two subscribers 'A' and 'B' from Airtel Bangalore are visiting Kolkata. When 'A' calls 'B', 'A' dials the number of 'B' which is a Bangalore number. Therefore, the call will be routed to Airtel in Bangalore. In Bangalore it is found that 'B' is roaming in Kolkata, therefore the call will be routed back to Kolkata. If you notice, though both subscribers are in Kolkata, the call is routed through Bangalore and both of them pay the long distance charges. To avoid this, some network operators came up with something called Optimal Call Routing (OCR). OCR will work only when the called party's VLR and the calling party's VLR are within the same MSC/VLR. Let us take the previous example and assume that both 'A' and 'B' are roaming at Kolkata's Airtel network. While 'A' makes a call to 'B', he prefixes a # in front of the number like #09845050505. This being an outgoing call, the Airtel MSC in Kolkata will look at the Kolkata VLR first. As the number is prefixed with #, it assumes that the other number is roaming in the same network as well. Therefore it looks at its own VLR once again to see whether 'B' is available in its database. If yes, it routes internally without forwarding the call to the home network. In case of 'B', though it is an incoming call, it is routed directly through the VLR without referring to the HLR at Bangalore.

5.9 GSM Frequency Allocation

GSM in general uses 900 MHz band; out of this, 890-915 MHz are allocated for the uplink (mobile station to base station) and 935–960 MHz for the downlink (base station to mobile station). Each way the bandwidth for the GSM system is 25 MHz (Fig. 5.12), which provides 125 carriers uplink/downlink each having a bandwidth of 200 kHz. The ARFCN (Absolute radio frequency channel numbers) denotes a forward and reverse channel pair which is separated in frequency by 45 MHz.

Mobile-to-base: $Ft(n) = 890.2 + 0.2(n-1)$ MHz

Base-to-mobile: $Ft(n) = Fr(n) + 45$ MHz

In practical implementation, a guard band of 100 kHz is provided at the upper and lower end of the GSM 900MHz spectrum, and only 124 (duplex) channels are implemented. Since 1995, new bands have been added to the basic 900MHz GSM. These bands are 1800 MHz and 1900MHz. 1800MHz band is licensed to the fourth GSM operator in India. The 1800 MHz band uses 1710–1785MHz and 1805–1880MHz (three times as much as primary 900MHz) with a total of 374 duplex channels. GSM 900 uses the four-cell repeat pattern for the frequency reuse cell sets. In most cases, each cell is divided into 120 degree sectors, with three base transceiver subsystems in each cell. Each base transceiver has a 120 degree antenna. These 12 sectors (called cells in GSM system) share the 124 channels.

To share the bandwidth for multiple users, GSM uses a combination of Time-Division Multiple Access (TDMA) and Frequency-Division Multiple Access (FDMA) encoding. One or more carrier frequencies are assigned to each base station. Each of these carrier frequencies is then divided in time, using a TDMA scheme. The fundamental unit of time in this TDMA scheme is called a burst period and it lasts approximately 0.577 ms. Eight burst periods are grouped into a TDMA frame for approximately 4.615 ms, which forms the basic unit for the definition of logical channels. One physical channel is one burst period per TDMA frame. Channels are defined by the number and

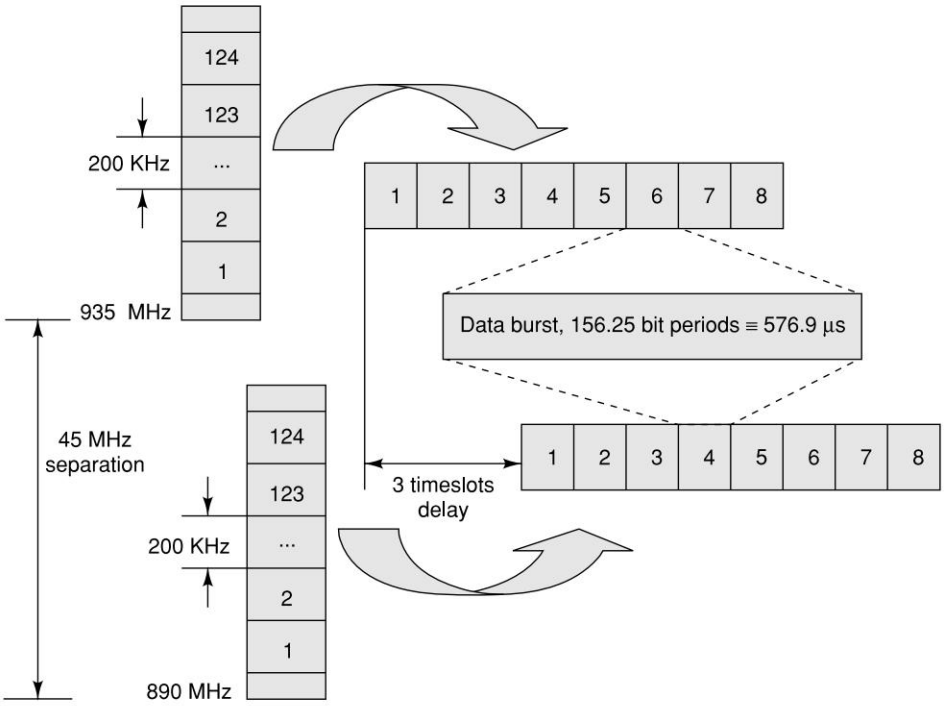


Figure 5.12 Carrier Frequencies and TDMA Frames

position of their corresponding burst periods. A traffic channel (TCH) is used to carry speech and data traffic. Traffic channels are defined using a 26-frame multiframe, or group of 26 TDMA frames (Fig. 5.13). Out of the 26 frames, 24 are used for traffic, one is used for the Slow Associated Control Channel (SACCH) and one is currently unused.

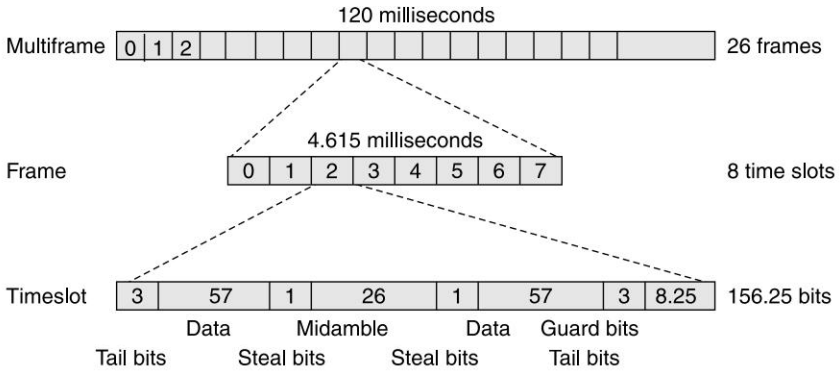


Figure 5.13 Organization of Bursts and TDMA Frames

5.10 Personal Communications Service

Personal Communications Service (PCS) technology is a flavor of GSM technology for digital cellular phone services that use frequency bands of 1800 and 1900 MHz; though, PCS is also used to signify one-to-one personal communications over the wireless media. PCS is used to denote Digital Enhanced Cordless Telecommunications (DECT) in the 1920 MHz to 1930 MHz Satellite Personal Communications. In general, PCS signifies the 1900 MHz radio band digital cellular mobile phone system in North American countries like Canada, Mexico and the United States. In Hong Kong, however, PCS is used to refer to GSM-1800. Like the GSM, PCS offers services like voice, data, SMS and roaming.

5.10.1 PCS Switching Center

Like the GSM, the PCS switching center represents a collection of network elements. It is a service which supports access technology independent call control/service control and a connection control switching functions. It also facilitates interconnection of access and network systems to support end-to-end services.

5.10.2 Supportability for PCS Frequencies

As such, CDMA, GSM and D-AMPS systems can also be used on PCS frequencies. In spite of the fact that Dual-band GSM phones can work in both the 850 and 1900 MHz bands, they are incompatible with 900 and 1800 MHz European and Asian variants. However, GSM tri-band and quad-band phones offered by North American carriers usually support both European and other domestic frequencies.

5.11 Authentication and Security

The radio medium is open to everybody and anybody. Anybody who can get hold of a radio receiver can access GSM signal or data. Therefore, it is necessary and important that the communication over the wireless radio media is secured. The first step to GSM security is the authentication. Authentication of a user is done to ensure that the user is really the person he claims to be. Authentication involves two functional entities, the SIM card in the mobile phone, and the Authentication Center (AUC). Authentication is done by using an algorithm by name A3. Following the authentication, a key is generated for encryption. An algorithm by the name A8 is used to generate the key. A different algorithm called A5 is used for both ciphering and deciphering procedures. The ciphering is done on both signaling, voice and data. This in other words means that SS7 signal, voice, data, and SMS within GSM are ciphered over the wireless radio interface.

The GSM specifications for security were designed by the GSM Consortium in secrecy and are distributed only on a need-to-know basis to hardware and software manufacturers and to GSM network operators. The specifications were never exposed to the public. The GSM Consortium relied on Security by Obscurity, i.e., the algorithms would be harder to crack if they were not publicly available.

5.11.1 The MS Authentication Algorithm A3

During the authentication process the MSC challenges the MS with a random number (RAND). This is illustrated in Figure 5.14. The SIM card uses this RAND received from the MSC and a secret key Ki stored within the SIM as input. Both the RAND and the Ki secret are 128 bits long. Using the A3 algorithm with RAND and Ki as input a 32-bit output called signature response (SRES) is generated in the MS. This SRES is then sent back to the MSC as the response to the challenge. Using the same set of algorithms, the AUC also generates a SRES. The SRES from MS (SIM) and the SRES generated by the AUC are compared. If they are the same, the MS is authenticated. The idea is that no keys will be transacted over the air. However, if the SRES values calculated independently by the SIM and the AUC are the same, the Ki has to be same. If Ki is same, the SIM card is genuine.

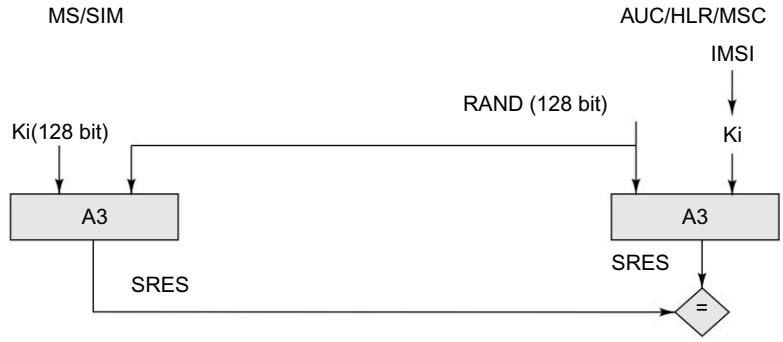


Figure 5.14 The Workflow of Authentication

5.11.2 The Voice-Privacy Key Generation Algorithm A8

For any type of cipher, we need a key. If the key is random and difficult to guess, the cipher is relatively secured. In the GSM security model, A8 algorithm is the key generation algorithm (Fig. 5.15). A8 generates a session key, Kc, from the random challenge, RAND, received from the MSC and from the secret key Ki. The inputs for A8 are the same set of 128-bit Ki and RAND as used in A3. The A8 algorithm takes these inputs and generates a 64-bit output. The keys are generated at both the MS (SIM) and the network end. The BTS received the Kc from the MSC. The session key Kc, is used for ciphering, till the time the MSC decides to authenticate the MS once again. This might sometimes take days.

5.11.3 The Strong Over-the-Air Voice-Privacy Algorithm A5/1

In the GSM security model, A5 is the stream cipher algorithm used to encrypt over-the-air transmissions (Fig. 5.16). The stream cipher is initialized all over again for every frame sent. The stream cipher is initialized with the session key, Kc, and the number of the frame being encrypted or decrypted. The same Kc is used throughout the call, but the 22-bit frame number (F_n) changes during the call, thus generating a unique keystream for every frame (Fig. 5.17).

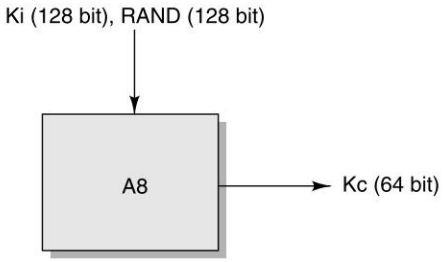


Figure 5.15 Session Key (Kc) Calculation

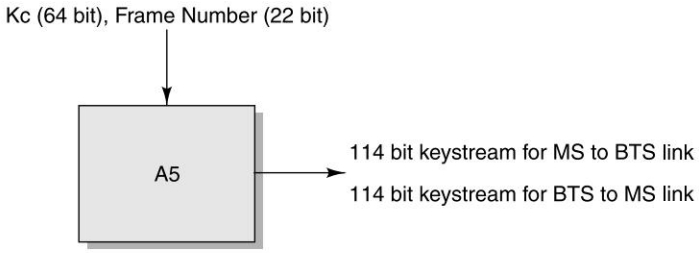


Figure 5.16 Keystream Generation

All these security algorithms put together make the entire GSM security protocol as shown in Figure 5.17.

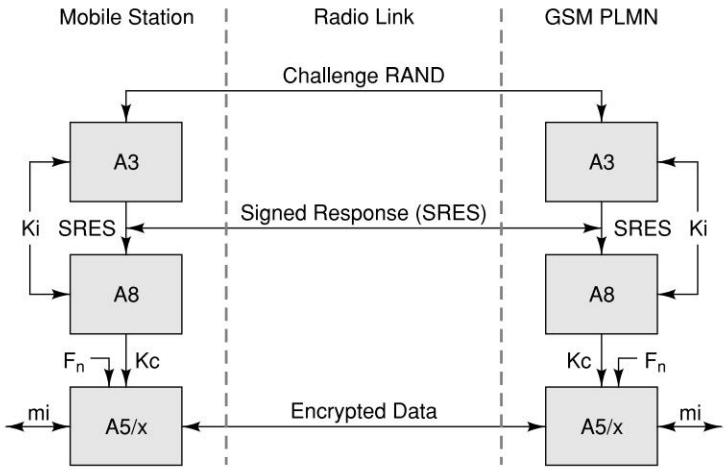


Figure 5.17 The GSM Security Protocol

Since the first GSM systems, many variations of A5 algorithms have been designed and implemented. The main contention has been that the original A5 encryption algorithm was too strong. A strong and difficult encryption algorithm always attracts export restrictions. The original A5 algorithm was not allowed to be used outside Europe. Therefore, the first "original" A5 algorithm was renamed A5/1. Other algorithms including A5/0, which means no encryption at all, and A5/2, a weaker over-the-air privacy algorithm were developed. The A5 algorithms after A5/1 have been named A5/x. Most of the A5/x algorithms are considerably weaker than the original A5/1.

REFERENCES/FURTHER READING

1. Eberspacher, Jorg and Hans-Jorg Vogel, (1999), *GSM Switching, Services and Protocols,* John Wiley & Sons.
2. Garg, Vijay K. and Joseph E. Wilkes, (2002), *Principles & Applications of GSM*, Pearson Publication Asia.
3. GSM Cloning: Reference: http://www.isaac.cs.berkeley.edu/isaac/gsm.html.
4. GSM Standard 03.01: Digital cellular telecommunications system (Phase 2+); Network functions, Release 1998, www.etsi.org.
5. GSM Standard 03.02: Digital cellular telecommunications system (Phase 2+); Network architecture, Release 1998, www.etsi.org.
6. GSM Standard 03.03: Digital cellular telecommunications system (Phase 2+); Numbering, addressing and identification, Release 1998, www.etsi.org.
7. GSM Standard 03.09: Digital cellular telecommunications system (Phase 2+); Handover procedures, Release 1998, www.etsi.org.
8. GSM Standard 03.20: Digital cellular telecommunications system (Phase 2+); Security related network functions, Release 1998, www.etsi.org.
9. GSM 03.68 Digital cellular telecommunications system (Phase 2+) (GSM);Voice Group Call Service (VGCS); Stage 2.
10. GSM World: http://www.gsmworld.com.
11. Heine Gunnar, (1999), *GSM Networks: Protocols, Terminology, and Implementatio*, Artech House.
12. Lee, William C.Y. (2000), *Mobile Cellular Telecommunications Analogue and Digital Syatems*, McGraw-Hill.
13. Redl, Siegmund M., Matthias K. Weber and Malcom W. Oliphant, (1998), *GSM and Personal Communications Handbook*, Artech House.
14. Sempere Javier Gozálvez, *An overview of the GSM system*, http://www.telcor.-gob.ni/BCS/nd/gsm.html.
15. Sivagnanasundaram Suthaharan (1997), *GSM Mobility Management Using an Intelligent Network Platform*, Ph.D Thesis, University of London.
16. 3GPP TS 03.20 version 8.4.1 Release 1999, Digital cellular telecommunications system (Phase 2+); Security-related network functions.
    - www.berkeley.edu

- www.gsmworld.com
- www.wikipedia.org

17. 3GPP TS 08.06, Digital cellular telecommunications system (Phase 2+); Signalling Transport Mechanism Specification for the Base Station System - Mobile Services Switching Centre (BSS-MSC) Interface.
18. 3GPP TS 08.08, Digital cellular telecommunications system (Phase 2+);Mobile-services Switching Centre–Base Station System (MSC-BSS) interface; Layer 3 specification.
19. 3GPP TS 09.02 Digital cellular telecommunications system (Phase 2+); Mobile Application Part (MAP) specification.
20. 3GPP TS 23.122, 3rd Generation Partnership Project; Technical Specification Group Core Network; NAS Functions related to Mobile Station (MS) in idle mode.
21. 3GPP TR 23.908, Digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); Technical Report on Pre-paging.
22. 3GPP TS 29.018: 3rd Generation Partnership Project; Technical Specification Group Core Network; General Packet Radio Service (GPRS); Serving GPRS Support Node (SGSN) - Visitors Location Register (VLR) Gs interface layer 3 specification.
23. 3GPP TS 29.198-14, Universal Mobile Telecommunications System (UMTS); Open Service Access (OSA) Application Programming Interface (API); Part 14: Presence and Availability Management (PAM) Service Capability Feature (SCF).

REVIEW QUESTIONS

Q1: Describe the GSM architecture with its constituent elements.

Q2: In GSM network, there are some databases used for various purposes. What are they? What are their functions?

Q3: Explain the following in brief in the context of GSM networks:

(a) Mobile Station (b) BSS
(c) NSS (d) OSS
(e) IMSI (f) IMEI
(g) TMSI (h) MSRN

Q4: Explain mobile terminating call in the context of GSM networks.

Q5: What is handover/handoff? How is handoff different from roaming?

Q6: What is the role of AuC? How is authentication done in a GSM network?

Q7: What are the different algorithms used for security in GSM?

Q8: What are HLR and VLR? Describe the functions of HLR and VLR in call routing and roaming?

Q9: What is a PLMN? How is PLMN connected to PSTN and PDN?

Q10: What is PCS? Which areas of the world is it presently used in?

Q11: What is the role of PCS switching center?

Q12: Discuss the supportability for PCS frequencies.

CHAPTER 6

Short Message Service (SMS)

6.1 MOBILE COMPUTING OVER SMS

GSM supports data access over CSD (Circuit Switched Data). GSM is digitized but not packetized. In case of CSD, a circuit is established and the user is charged based on the time the circuit is active and not on the number of packets transacted. GPRS (General Packet Radio Service), also known as 2.5G, which is the next phase within the evolution of GSM, supports data over packets. WAP is a data service supported by GPRS and GSM to access Internet and remote data services. WAP has been covered in Chapter 8. Other data services in GSM include Group 3 facsimile, which is supported by use of an appropriate fax adaptor. A unique data service of GSM, not found in older analog systems, is the Short Message Service (SMS). SMS enables sending and receiving text messages to, and from, GSM mobile phones. In this chapter we discuss SMS and developing applications using SMS bearer.

6.2 SHORT MESSAGE SERVICE (SMS)

Like many other eccentric technologies, SMS was also allegedly the right idea at the wrong time. On December 3, 1992, a scientist named Neil Papworth at Sema, a British technology company, sent the first text message "Merry Christmas" to the GSM operator Vodafone. It was sent to Vodafone director Richard Jarvis in a room at Vodafone's HQ in Newbury in southern England. The message was an overly premature seasonal greeting, some three weeks ahead of the festivities. Vodafone offered this service as a text messaging service with a brand name TeleNotes service targeted for the business community. The service was not at all popular in its early days. SMS was almost forgotten and became an unwanted child until seven years later in 1999 when other mobile phone operators started to allow customers to swap SMS. Today SMS is the most popular data bearer/service within GSM with an average of one billion SMS messages (at the end of 2002) transacted every day around the world, with a growth of on an average half a billion every month. The SS7

signaling channels are always physically present but mostly unused, be it during an active user connection or in the idle state. It is, therefore, quite an attractive proposition to use these channels for transmission of used data. SMS uses the free capacity of the signaling channel. Each short message is up to 160 characters in length when 7-bit English characters are used. It is 140 octets when 8-bit characters (some European alphabets) are used, and 70 characters in length when non-Latin alphabets such as Arabic, Chinese or Hindi are used (70 characters of 16 bit Unicode).

6.2.1 Strengths of SMS

Following is the list of unique characteristics of SMS, which make this an attractive bearer for mobile computing.

Omnibus nature of SMS: SMS uses SS7 signaling channel, which is available throughout the world. SMS is the only bearer that allows a subscriber to send a long distance SMS without having long distance subscription. For example, you cannot make a voice call to a mobile phone in UK unless you have an international calling facility. However, you can send a SMS to a subscriber in UK, without having an international call facility.

Stateless: SMS is sessionless and stateless. Every SMS message is unidirectional and independent of any context. This makes SMS the best bearer for notifications, alerts and paging. SMS can be used for proactive information dissemination for "unsolicited response" and business triggers generated by applications (referred as "Push" in Fig. 8.4).

Asynchronous: In HTTP, for every command (e.g., GET or POST) there is a request and a response pair making it synchronous at the transaction level. Unlike HTTP, SMS is completely asynchronous. In case of SMS, even if the recipient is out of service, the transmission will not be abandoned. Therefore, SMS can be used as message queues. In essence, SMS can be used as a transport bearer for both synchronous (transaction oriented) and asynchronous (message queue and notification) information exchange.

Self-configurable and last mile problem resistant: SMS is self-configurable. In case of Web or WAP, it is no trivial task to connect to a service from a foreign network without any change in the configuration or preference setting. The device needs to be configured interactively by the user or system administrator to access the network. This makes the access dependent on the last mile. SMS has no such constraints. While in a foreign network, one can access the SMS bearer without any change in the phone settings. The subscriber is always connected to the SMS bearer irrespective of the home and visiting network configurations. While roaming in a foreign network, even if the serving network does not have an SMSC (SMS Center) or SC (Service Center), SMS can be sent and received.

Non-repudiable: SMS message carries the SC and the source MSISDN as a part of the message header. Unlike an IP address it is not easy to handcraft an MSISDN address in the SMS. It is possible for an application connected to an SMS to handcraft an MSISDN address like "999" or even alphabetic addresses like "MYBANK". However, an application can not handcraft the SC address. Therefore, an SMS can prove beyond doubt the origin of itself.

Always connected: As SMS uses the SS7 signaling channel for its data traffic, the bearer media is always on. User cannot SWITCH OFF, BAR or DIVERT any SMS message. When a phone is busy and a voice, data or FAX call is in progress, SMS message is delivered to the MS (Mobile Station) without any interruption to the call.

6.2.2 SMS Architecture

SMS are basically of two types, SM MT (Short Message Mobile Terminated Point-to-Point), and SM MO (Short Message Mobile Originated Point-to-Point). SM MT is an incoming short message from the network side and is terminated in the MS. SM MO is an outgoing message, originated in the user device (MS), and forwarded to the network for delivery. For outgoing message, the path is from MS to SC via the VLR and the IWMSC function of the serving MSC, whereas for incoming message the path is from SC to the MS via HLR and the GMSC function of the home MSC (Fig. 6.1).

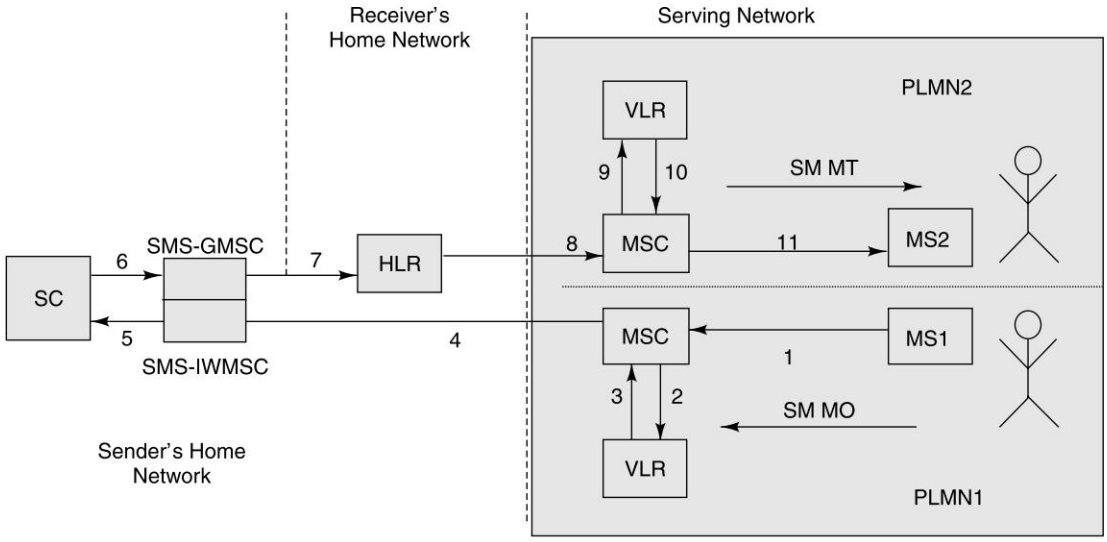


Figure 6.1 Flow of SMS between Two MS

To use SMS as a bearer for information exchange, the Origin server or the Enterprise server needs to be connected to the SC through a short message entity (SME) as in Figure 6.2. The SME in this case works as an SMS gateway, which interacts to the SC in one side, and the enterprise server on the other side.

6.2.3 Short Message Mobile Terminated (SM MT)

For an SM MT message, the message is sent from SC to the MS. This whole process is done in one transaction (Fig. 6.2). For the delivery of MT or incoming SMS messages, the SC of the serving

network is never used. This implies that an SMS message can be sent from any SC in any network to a GSM phone anywhere in the world. This makes any SM MT message mobile operator independent.

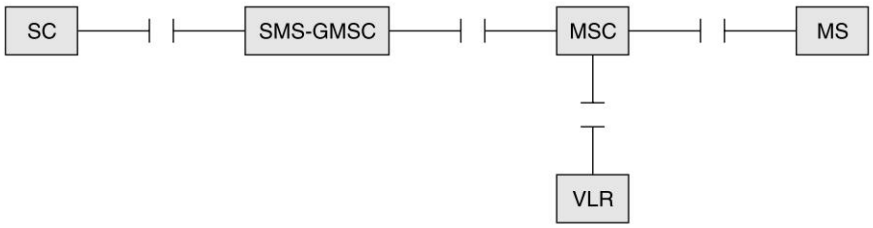


Figure 6.2 Interface Involved in the SM MT Procedure

6.2.4 Short Message Mobile Originated (SM MO)

SM MO is an outgoing message originated in the MS where generally the user types in a message and sends it to another MSISDN number or an application. For an MO message, the MSC forwards the message to the home SC of the sender. The SC is an independent computer in the network and works as a store and forward node with a large database. The database is used to store the SMSs. In SS7 terminology SC is an SCP (Service Control Point) within the SS7 cloud. MO message works in two asynchronous phases. In the first phase, the message is sent from the MS to the home SC as an MO message (Fig. 6.3). In the second phase, the message is sent from the home SC to the receiving MS as an MT message (Fig. 6.2). It is possible to attempt to send an SMS message to an invalid MSISDN number. In such a case, the message will be sent successfully from the MS to the SC. However, it will fail during the SC to the MS transfer. The user will see SM MO message sent successfully but SM MT message delivery would fail.

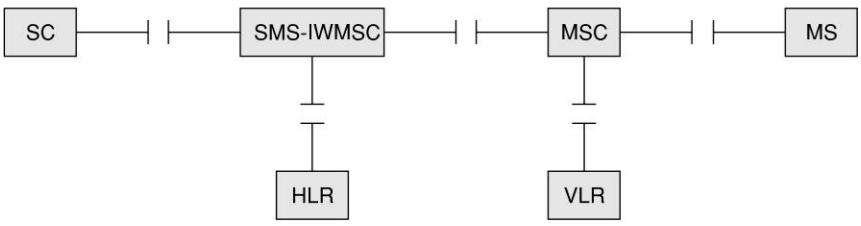


Figure 6.3 Interface Involved in the Short Message Mobile Originated (SM MO) Procedure

6.2.5 SMS as an Information Bearer

SMS is a very popular bearer in the person-to-person, mobile-to-mobile or point to point messaging domain. However, it is gaining popularity in other verticals like enterprise applications, services provided by independent service providers as ASP (Application Service Provider), and notification services, where one endpoint is a mobile phone but the other endpoint is a mobile application. Here SMS functions as an input-output media for information exchange for a mobile application (Fig. 6.4).

To use SMS as a bearer for any information service, we need to connect the services running on the Enterprise Origin server to the SC through an SME (Short Message Entity) or ESME (External Short Message Entity). SME in any network is generally a SMS gateway. With respect to SMS, a GSM subscriber is always in control of the SC in the home network irrespective of the serving network. Thus, if there is any SMS-based data service in the home network, it will be available to the subscriber from any foreign network.

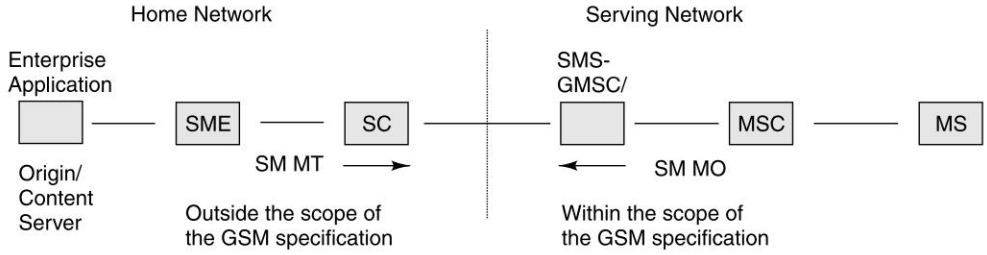


Figure 6.4 SMS as an Information Bearer/Medium for Mobile Applications

6.2.6 Operator-centric Pull

For an SMMO to work it is mandatory that an SC is used. As a part of SMS value added services, operators offer different information on demand and entertainment services. These are done through connecting an Origin server to the SC via an SMS gateway. In different parts of the world a new industry vertical has emerged to address this market. These service providers are known as MVNO (Mobile Virtual Network Operators). Virtual operators develop different systems, services, and applications to offer data services using SMS. Many enterprises use these MVNOs to make their services available to mobile phone users. There are quite a few banks in India which offer balance enquiry and other low security banking services over SMS. For example, if a HDFC customer wants to use these services, he needs to register for the service. During the registration, the HDFC customer needs to mention the MSISDN of the phone which will be used for this service. Once a user is registered for the service, he enters "HDFCBAL" and sends the message to a service number (like 333 for example in the case of Escotel) as an MO message. SC delivers this MO message to the SMS gateway (technically known as SME–Short Message Entity) connected to this service number. The SMS gateway then forwards this message to the enterprise application. The response from the enterprise application is delivered to the MS as an MT message from the SME. Even if the subscriber is in some remote region of a foreign network within GSM coverage, he can send the same SMS to the same service number in his home network. This makes the home services available in the foreign network. This also implies that an operator-centric SMS pull service is completely ubiquitous.

The connectivity between SC to SME and SME to Enterprise Origin server is not defined by GSM. However, there are a few de facto standard protocols for this communication. The most popular protocol is Short Message Peer to Peer (SMPP). There are certain other protocols like CIMD from Nokia as well. The connectivity between SME and Origin server could be anything like SOAP (Simple Object Access Protocol), or direct connection through TCP socket. However, common practice is through HTTP. HTTP helps user to get information from the Internet via SMS. There is an open source for SMS gateway called Kannel, which supports a multitude of

protocols and forwards the SMS enquiry as an HTTP request and gets information from the Internet. This is how an SMS can be converted into a simple Internet access. Conventionally SMS queries are keywords driven like "CRI" for live cricket score, or "RSK 2627 3 03" to get the availability of seat/berth in Indian Railways train number 2627 (Karnataka Express) for March 3. There are applications where SMS is used in session-oriented transactions. Applications like "SMS chat" and "SMS contests" need to remember the user context over multiple transactions.

6.2.7 Operator-independent Push

We have seen that it is possible to send an SMS to any phone in any network. For example, an MT message can be delivered from a network in India to an MS of UK roaming in Germany (Fig. 6.5). Which in other words means that any push, which may be an alert, notification or even response from a pull message generated by an application, can be serviced by any network and delivered to any GSM phone in any network without any difficulty.

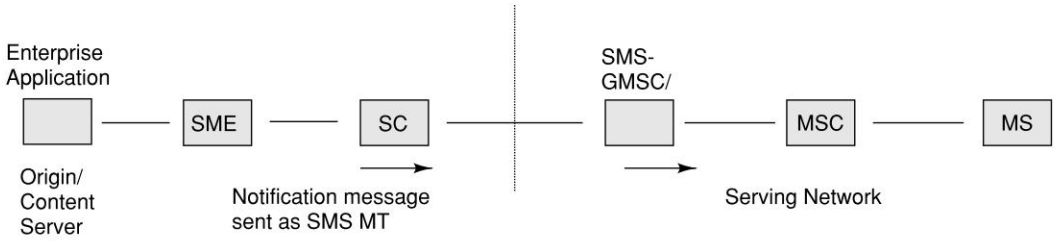


Figure 6.5 The Basic Network Structure of the SMS Push

Assuming that appropriate roaming tie-ups are in place, an enterprise can use SMS to send business alerts or proactive notifications to its customer anywhere, anytime on any GSM phone. With roaming tie-ups, operators reach an agreement on revenue share and call forwarding mechanism. Roaming tie-ups are a commercial issue rather than technical. Some credit card companies in India send SMS notifications to its cardholders in different networks using operator-independent push.

6.2.8 Challenge for SMS as a Mobile Computing Bearer

When it comes to offering enterprise services using SMS, the scene becomes difficult to manage. Let us take the example of Indian Bank. In Delhi, a customer of this bank who is a subscriber of operator "A" (Airtel) sends "HDFCBAL" to 300 to know the balance in his account. In the same city of Delhi another customer of the same bank who happens to be a subscriber of a different operator "B" (Essar) sends "HDFCBAL" to 1234 to get the balance information. HDFC bank has a sizable population of customers in the Middle East. The same banking services, which are available in India, are not available in the Middle East. The reason being both cellular operators "A" and "B" connect to the bank's application through their private SC and SME, whereas the operators in Middle East do not have an SME to connect to the bank's application. This is like in the early days of telephony when an enterprise used to announce different customer care numbers for different cities. If the enterprise did not have an office in a city, the customers had to make long distance

calls to customer care in some other city. All these changed with the introduction of the 1–800 service. Enterprises need something similar to 1–800 in SMS. Also, this gives some identity to the enterprise. My Inc for example may like to publish a number like +9198375MYINC for any of its customer anywhere in the world.

The major challenge for implementing ubiquitous service through SMS requires operator independent SM MO messages or operator independent pull services. The SMS routing needs to work exactly in the same fashion as 1–800 services.

6.2.9 Operator-independent Pull

As the SME is always connected to the home network's SC, with the conventional framework, it is not possible to route mobile originated SMS messages to any application or any SME of choice. There are ways by which an SMS message can be routed to some enterprise SME connected to external SC. This is achieved through SAT, where the SAT application running on the SIM card changes the SC number during the transmission of the SMS and forces the SMS to recognize a different SC of a different network as its home SC. In this case also, technically the SMS is sent to the SME connected to the home SC. SMS has always been considered a revenue generating tool for cellular operators. Therefore, the current framework suits a cellular operator very well. If a SMS service is operator dependent, the cellular operator can use this to its advantage. In today's global scenario an enterprise or a MVNO has its customers around the world subscribing to different GSM networks. To make this possible, enterprises need operator-independent pull as well. Operator-independent pull services can be achieved using GSM modem technology described in the following sections. Also, the same can be done using Intelligent Network Technologies.

6.3 Value Added Services through SMS

Value Added Services (VAS) can be defined as services, which share one or more of the following characteristics:

- Supplementary service (not a part of basic service) but adds value to total service offering.
- Stimulates incremental demand for core services offering.
- Stands alone in terms of profitability and revenue generation potential.
- Can sometimes stand-alone operationally.
- Does not cannibalize basic service unless clearly favorable.
- Can be an add-on to basic service, and as such, may be sold at a premium price.
- May provide operational and/or administrative synergy between or among other services and not merely for diversification.

A GSM operator's primary business goal is to offer the network infrastructure. Voice, SMS are basic services provided by a GSM operator. However, offering different other services using SMS as a bearer will be a VAS. There are various flavors and variations of VAS over SMS. We will give some examples and discuss how to develop them. The most popular VAS over SMS are entertainment and information on demand. Information on demand has three categories as described below.

1. **Static information.** This type of information does not change frequently. A good example is a restaurant guide. It is sufficient to update this type of information once in a fortnight, or even less frequently. These contents generally fall in mass market category.
2. **Dynamic information.** This type of information changes in days. For example, the daily horoscope needs to be updated on daily basis. Mass market contents fall in this category as well.
3. **Real-time information.** This type of information changes continually. Third-party contents fall in this category. For example, scores in a live cricket match or stock quote undergo continual change. All the enterprise contents will fall in this category. For enterprise content, the content will be obtained directly from the enterprise.

6.3.1 User Interface in SMS Value Added Services

We have already seen that SMS is sessionless. In Chapter 1 also we have discussed session-oriented transaction and short transaction (Section 1.4). Majority of services over SMS will use the short transaction model. For a SMS-based service, the user interface is always keyword-based. This is something similar to the character-based command interfaces, where the first word is the keyword (command) and rest are the parameters for the command. For example, I want to know the latest news. For this I enter News and send it to the VAS service. If I want business news, I enter News Biz. News is the keyword. Another example could be **RSA 2627 Bangalore New Delhi 20 01.** This example is for finding out the seat availability on the Indian Railways train number 2627 from Bangalore to New Delhi for 20 January. For Indian Railways, the tickets are available only 60 days in advance. Therefore, we do not need the year. The response for this enquiry will be **Date: 20-1 Train: 2627 KARNATAKA EXP Class:2A Status: WL 31/WL 14 Class: 3A Status: WL 63/WL 51 Class: SL Status: WL 59/WL 29**. Please note that the response from a service can sometimes be more than 160 characters. If it is more than 160 characters, we need to split the response into multiple message responses. It is advised that while the message is broken into multiple messages, it is broken at the word boundary. It is also advised that a sequence number like … **1/3,** … **2/3**, and … **3/3** is added in the first, second and third messages, respectively.

6.3.2 VAS Examples

In this section we describe some of the popular value-added services.

News/Stock Quotes Service

In a service like News or Stock Quote, we get the latest news or stock information. This will be a Short transaction. The keyword for news will be News, whereas the keyword for stock quote can be BSE. BSE Infosys will give the stock price of Infosys at the Bombay Stock Exchange. These are examples for real-time information on demand. For services like News and Stock Quote we need to have a relationship with some content provider who will supply us the up-to-date information. For example, we could tie up with *CNN* for international news, *The Indian Express* for general news, *weather.com* for weather news, etc. For stock quote, we may need to tie up with a stock exchange like Bombay Stock Exchange or National Stock Exchange. We will receive live feed from these content providers and update the content database on a real-time basis. As and when a subscriber wants these information, we supply the latest information from the live database.

Session-based Chat Application

A chat service is essentially a session-oriented transaction. In a chat service the user needs to log in. The user needs to explicitly log out or will be logged out implicitly following a period of inactivity. Every time the user sends a chat keyword, we need to know the previous transactions. Every SMS message carries the unique MSISDN number. This unique MSISDN number of the phone can be used as the session key. In the chat software we remember the state of the transaction using this MSISDN.

Email through SMS

This is a very useful service and is a transaction-oriented dialogue. To send an email through SMS, the user message will be **mail roopa@iitb.ac.in we will meet tomorrow 6:00 pm**. This VAS will send a mail to Roopa with mail id roopa@iitb.ac.in. The body of the mail will be "we will meet tomorrow 6:00 pm." The mail will be sent to Roopa by a SMTP server.

Health Care Services

Health care applications need both pull and push. A typical health care application could be ICU (Intensive Care Unit) system. The system will include alerts to doctor. In status monitoring service, a doctor or a nurse can enquire the status of a patient in the ICU. A limited enquiry facility will be provided to one MSISDN outside the hospital staff. This could be for someone in the family. This enquiry will be a short transaction. We will have alert services as well. In the alert service, nurses and doctors are notified periodically about the status of the patients in the ICU. The alerts can also be integrated with medical equipment.

Micro-payment Services

Let us take an example of micro-payment for a vending machine. This will be a session-oriented dialogue. In this application there will be some number of identifier (ID) pasted on the vending machine. The customer enters this identification number and sends a request to purchase a merchandise to the service provider. The service provider will authenticate the user and check whether the user has sufficient money in credit. Based upon the credit the transaction will either be approved or rejected. If approved, an authorization message will be sent back to the vending machine. The vending machine will ask the user to select merchandise. The user selects the merchandise, a soft drink, for example. Once the merchandise is dispensed, the vending machine will send back a message to the VAS indicating that the merchandise has been dispensed. The price for the merchandise is debited from the user account.

6.3.3 Alert Services

These are proactive alert services. For a stock quote the alert services can be of the following kind.

Time-based: In this service, proactive alerts are sent to the mobile phone at a pre-assigned time of the day. The alert contains the stock quote of different scripts of the portfolio.

Watermark-based: In this service whenever the stock price goes up or falls down to a certain level, alerts are sent. This information will help the subscriber to decide whether to buy or sell some particular stock.

For other services, like cricket score, it can be a periodic alert (every 10 minutes) during the match. There can be other alerts like inform the live score whenever a player is out etc.

6.3.4 Location-based Software

Location-based services could be road direction, restaurant guide, etc. Some location-aware VAS services provide shopping alerts as well. In location-based services only the information relevant to the current location of the mobile phone (or the subscriber) is provided. In a shopping service, the user will receive alerts on discount or sale information when they pass through close the proximity of the shopping malls. In the case of a restaurant guide, let us assume that the subscriber is in a office on M.G. Road in Bangalore and sends Res to the VAS. Only the restaurants in, and around, M.G. Road will be provided as response to this request. When the same user asks for the same information in Mumbai, restaurants in Mumbai will be given as response. For location-aware software, the precise location of the user needs to be determined. The location of a mobile phone can be determined either from the network or from the device. Using Time Advancing techniques within the BTS, the location of the mobile phone can be determined. This technique however requires the support of the network. The other option is to find out the location from the device. Device-specific location awareness requires either of the following technologies:

1. Cell ID (CID)-based system.
2. Global Positioning System (GPS)-based system.

In a CID-based system, the CID of the current BTS is determined. The CID-based system needs a mapping of the cell identifier to the geographical location. To handle the growing subscribers, new cell sites are added and the CIDs reconfigured. In such cases the mapping between locations versus CIDs need to be synchronized. For CID-based system, the signal strength from all the different CIDs are extracted from the device and sent to the server through a SMS. The location of the user is determined using the signal strength and triangulation algorithms. In a GPS-based system, the location is detertmined through a GPS receiver installed within the phone. GPS provides facility to compute position, velocity and time of a GPS receiver. To offer a travel direction through GPS, the GPS system will inform the application about the exact location of the phone. From the velocity it will also know the direction the user is moving. Based on the location and direction, the direction will be provided. Please note that sometimes it may not be a trivial task to take a U turn on a freeway or motorway. GPS-based system is not dependent on the network operator.

In Figure 6.6 we describe the basic value added service provisioning architecture for SMS. The reader should try to map an application scenario and get a feeling of how information travels across in the case of pull/push .

6.4 Accessing the SMS Bearer

There are two ways the SMS bearer can be accessed:

1. Use a mobile phone as a GSM modem and connect it to the computer.
2. Use the SMSC of an operator through SMPP or similar interface.

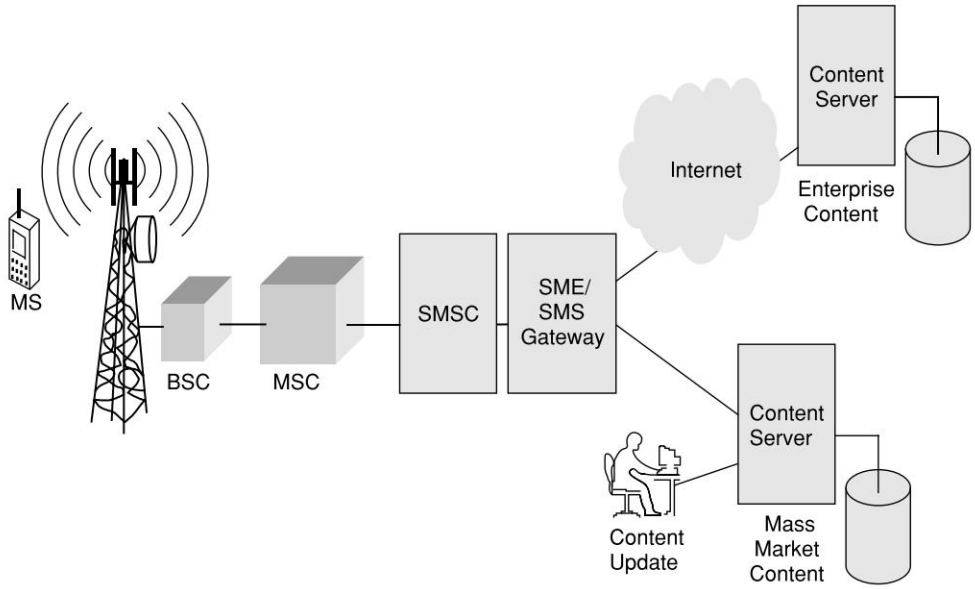


Figure 6.6 SMS Value Added Services Architecture

6.4.1 GSM Modem (Over the Air)

This is an operator independent quick-fix solution. GSM modem does not have scalability. In a GSM modem, we use a normal cell phone as a data modem. The cell phone will have a SIM card and will be in a position to access the network as a normal GSM phone. It will be able to send and receive SMS messages. To convert this cell phone into a GSM modem, we need to connect the phone to a computer. This connectivity is established through either IrDa or direct cable. We can use data cables available from Nokia to use a Nokia phone as modem. Nokia manufactures different data cables for different models of Nokia phones. We need to install the software (device drivers) associated with the phone model into the computer. One end of the data cable will be connected to the cell phone and the other end will be connected to the COM port of the computer. This is similar to steps involved in using an external modem in a computer. Once all these are ready, one can use the cell phone as an external GSM modem and issue **AT** commands to transact data over the GSM/SMS bearer as done in case of any other Hayes compatible modems.

One can use the Hyper terminal (Go to **Start** → **Programs** → **Accessories** → **Communications** → **Hyper Terminal**) software and try sending SMS from a PC. Following is a very simple example of sending SMS from a hyper terminal.

Sent: `AT`
Recv: `OK`
Sent: `AT+CMGF=1`

```
Recv: OK
Sent: AT+CMGS="9810080856"
> This SMS message is being sent from a computer using
                              hyper-terminal and my Nokia phone<ctrl-Z>
Recv: +CMGS: 122
OK
```

In this example, we use the standard Hayes Modem command sets. We send the AT (In Hayes terminology this is known as attention) command to the modem from the computer. The GSM modem responds by saying "OK". This means that the modem is ready and can take instructions. We then set the message format to text mode through CMGF command. In the next request we send AT+CMGS="9810080856". This is to send a SMS message to a mobile with MSISDN 9810080856. The GSM modem accepts the request and responds with a '>' sign. This is a prompt from the modem requesting for the user input. The user enters the data followed by a control "Z". "^Z" (control z, 0x1A) is used to indicate the end of message. When the message is sent, the GSM modem responds with a number 122. This number is the message identifier of the message successfully sent.

AT command can also be used for other functions of the phone. Most of the functions available as a part of MMI (Man Machine Interface), are available through AT command. Examples could be sending an SMS, read an SMS; check battery power or write a phone book entry. Following is a list of the AT commands supported for SMS.

SMS Text Mode

AT+CSMS	Select Message Service
AT+CPMS	Preferred Message Storage
AT+CMGF	Message Format
AT+CSCA	Service Center Address
AT+CSMP	Set Text Mode Parameters
AT+CSDH	Show Text Mode Parameters
AT+CSCB	Select Cell Broadcast Message Types
AT+CSAS	Save Settings
AT+CRES	Restore Settings
AT+CNMI	New Message Indications to TE
AT+CMGL	List Messages
AT+CMGR	Read Message
AT+CMGS	Send Message
AT+CMSS	Send Message from Storage
AT+CMGW	Write Message to Memory
AT+CMGD	Delete Message

SMS PDU Mode

AT+CMGL	List Messages

AT+CMGR	Read Message
AT+CMGS	Send Message
AT+CMGW	Write Message to Memory

In a text mode the user sends the message as a text. In this mode the message is ASCII encoded. However, during transmission this is converted in TPDU (Transfer Protocol Data Unit) encoding. The message can alternatively be encoded in the TPDU format as well. For ringing tones or picture messages, we need to send binary data; it is therefore mandatory that these data be encoded in TPDU mode. In TPDU mode, a text will also be encoded in binary. For example, if you want to send "Mobile Computing" in text mode, you send a string of 8 bit unsigned integers with values 0x4D, 0x6F, 0x62, 0x69, 0x6C, 0x65, 0x20, 0x43, 0x6F, 0x6D, 0x70, 0x75, 0x74, 0x69, 0x6E, 0x67. This in bit streams will look like the following:

01001101-01101111-01100010-01101001-01101100-01101001-00100000-01000011-01101111-01101011-01110000-01110101-01110100-01101001-01101110-01100111

It is sufficient to use 7 bits for any ASCII character. Therefore, in TPDU mode we use only 7 out of 8 bits. In TPDU mode we remove the most significant bit of every character to convert the above bit stream into:

1001101-1101111-1100010-1101001-1101100-1101001-0100000-1000011-1101111-1101011-1110000-1110101-1110100-1101001-1101110-1100111

Each byte now has only 7 bits of information. To make an 8-bit byte, we borrow bits from following characters. The least significant bit from the following 7-bits character is taken and made as the most significant bit of the preceding character. This will make the bit stream look like:

11001101-10110111-00111000-11001101-01001110-10000011-10000110-11101111-00110110-10111100-01001110-01001111-10111011-11001111

Therefore, the same text "Mobile Computing" in TPDU mode will look like 0xCD 0xB7 0x38 0xCD 0x2E 0x83 0x86 0xEF 0x36 0xBC 0x4E 0x4F 0xBB 0xCF. To send a message in text mode the message format needs to be set by using AT+CMGF command. The value of 1 for CMGF is for text mode whereas the value 0 is for TPDU mode. In case of text mode we enter ^Z as the end of string (like the null character or the ^@ in case of 'C' string). In case of TPDU mode the length of the string needs to be explicitly given along with the ^Z.

Let us take an example of sending Mobile Computing in TPDU format to a MSISDN +919845062050 using a SMSC whose service center number is +919810051914.

```
AT+CMGF=0                  // set SMS PDU mode on
OK
AT+CMGS=28                 // length of the SMS PDU,
                           // The RP layer SC address octets are not counted in the length.
>07911989015091411100 0C91198954600205000 0A711CDB738CD2E8386EF3
6BC4E4FBBCF<ctrl-z)
+CMGS: 212                 // message reference is shown
OK
```

Table 6.1 explains the fields of the above TPDU.

Table 6.1 RP SC Address-value field followed by a TPDU in hexadecimal form

RP SC address (optional)	*Value*	*Description*	*Status*
	07	Address length	Length of the address is 7. Including the type of numbering plan indication.
	91	Type of address	International address using ISDN telephone numbering plan.
	19 89 01 50 91 41	Short message service center address	The short message service center number. For example, 198901509141 is encoded as 91 98 10 05 19 14. In this case the address takes 6 octets. +919810051914 (Airtel in Delhi) encoded as 198901509141
TPDU Octet 1 bits	*Value (hex11)*	*Description*	*Status*
7	0	TP-Reply-Path	Reply path no set
6	0	TP-User-Data-header-indicator	Indication that user data doesn't contain additional header.
5	0	TP-Status-Report-Request	Not requested
4	1	TP-Validity-Period-Format	Relative format (bits 4 and 3)
3	0	TP-Validity-Period-Format	Relative format (bits 4 and 3)
2	0	TP-Rejected-Duplicates	Do not reject duplicates in SC
1	0	TP-Message-Type-Indicator	Type: SMS-SUBMIT (from phone to network), (bits 1 and 0)
0	1	TP-Message -Type-Indicator	Type: SMS-SUBMIT (from phone to network), (bits 1 and 0)
TPDU Octet 2	*Value*	*Description*	*Status*
	00	TP-Message-Reference	Given by the phone, application/ user does not need to fill this octet.
TPDU Octet 3	*Value*	*Description*	*Status*
	0C	Address length in semi-octets.	Length of the address is 12 in semioctets. Length of the type of numbering plan indication.

(*Contd*)

TPDU Octet 4	*Value*	*Description*	*Status*
	91	Type of address	International address using ISDN telephone numbering plan.
TPDU Octet 5–10	*Value*	*Description*	*Status*
	91	TP-Destination-Address	The destination telephone number.
	98		For example, +919845062053 encoded as 198954600235.
	45		The address can be 2
	06		to 12 octets long.
	20		+919845062053
	50		encoded as 198954600235
TPDU Octet 11	*Value*	*Description*	*Status*
	00	TP-Protocol-Identifier, consist one octet. For the details, see GSM 03.40 specification, version 7.2.0, page 53.	Parameter identifying the above layer protocol, if any. Note that for the straightforward case of simple MS-to-SC short message transfer, the TP-Protocol-Identifier is set to the value 00.
TPDU Octet 12 bits	*Value (hex00)*	*Description*	*Status*
7	0	TP-Data-Coding-Scheme used in TPUser-Data, consist one octet. See GSM 3.38	Functionality (bits 7 and 6) related to usage of bits 4-0.
6	0		Functionality (bits 7 and 6) related to usage of bits 4-0.
5	0		Indicates that text is uncompressed.
4	0		Indicated that bits 1 and 0 have no message class meaning.
3	0	Alphabet being used (bits 3 and 2)	7-bit message
2	0	Alphabet being used (bits 3 and 2)	7-bit message
1	0	Reserved	No meaning, indicated by bit 4
0	0	Reserved	No meaning, indicated by bit 4

(*Contd*)

TPDU Octet 13	*Value*	*Description*	*Status*
	A7	TP-Validity-Period (Relative format). See GSM 03.40, version 7.2.0, page 55 for details	A7 -> 24 hours
TPDU Octet 14	*Value*	*Description*	*Status*
	11	TP-User -Data-Length	Parameter indicating the length of the TP-User-Data field to follow. Represented as amount of septets (integer). 11 hex -> 17 septets. This is because of 7-bit user data. User data is coded to seven databits, because SMS have to be sent to air in 7 bit format. Length includes the user data header (not included in this example) and data itself.
TPDU Octet 15–29	*Value*	*Description*	*Status*
	CDB73 8CD2E 8386EF 36BC4 E4FBB CF	TP-User-Data	The user data. Format of the user data depends on what kind of message is sent. This example includes text string Mobile Computing. 16 septets + fill bits = 14 octets.

Encode the Short Message for Ringing Tone (8 bit, User-dataheader)

Nokia phones can be customized with personalized ringing tone. Ringing tones can be sent as SMS messages. These are also called EMS (Extended Message Service). Following is an example of sending a ringing tone to a Nokia phone with MSISDN +919845062050. Ringing tone is a 8-bit binary message in TPDU format. The name of the tone is "test". This name will be displayed in the menu.

```
AT+CMGF=0                    // set SMS PDU mode on
OK
AT+CMGS=50                   // length of the SMS PDU
                             // The RP layer SC address is not included in this example.
>0051000C91198954600205F515A7240605041581158102 4A3A51D195CDD00
8001B2055059061056055855054854082084990 00<ctrl-z>
+CMGS: 214                   // message reference is shown
OK
```

Following is the note for Indian National Anthem *Jana-gana-mana*. If one wants to send *Jana-gana-mana* as a ringing tone, one will compose a SMS with the following code as the TP-User-Data. Also, we need to modify the MSISDN accordingly.

```
02  4A  3A  69  8C  E9  71  B5  E4  81  91  BD  8D  D4  04  00
3A  D9  34  91  41  34  15  41  54  15  41  54  15  41  54  15
21  54  15  41  34  15  41  62  10  81  52  15  41  54  13  21
34  13  42  0E  21  24  D0  45  00  00
```

Sending a Picture (bitmap) OTA (Over The Air)

A picture message can occupy the complete display area of the phone. For Nokia phones, the maximum size of the picture message is 72×28 pixels. The maximum size of the operator logo and the CLI logo is 72×14 pixels.

Each semi-octet in the OTA bitmap presents 4 pixels in the original bitmap. Because one row takes 18 semi-octets, the whole 72×14 size (operator logo and CLI logo) bitmap takes 18×14 = 252 semi-octets = 126 octets. In the case of the picture message the whole 72×28 size bitmap takes 18×28 = 504 semi-octets = 252 octets (as it must be = sent using concatenate message). For details please refer to Nokia site at www.nokia.com.

Reading a Message through GSM Modem

In previous examples we discussed how to send SMS or EMS messages through GSM modem. When you send a SMS from a computer, the GSM modem always acts as a pass-through. However, for input message it behaves differently. When a message is received by a GSM modem, by default it is routed to the phone inbox. Therefore, to read a SMS from the GSM modem, we need to ensure that the SMS is forwarded to the computer rather than the phone's local store. For this we use the CNMI command. The value 2,1,0,0,0 signifies that whenever a SMS is received, an interrupt will be raised on the COM port followed by the messages flushed on the COM port. This transformation is described in detail in GSM 03.38 standard. Following is an example of reading a message "hellohello"

```
AT+CMGF=0                 // set SMS PDU mode on
AT+CNMI=2,1,0,0,0  // set the modem-computer interface
AT+CMGR                                  // read the message
```

The data we read from the port will be:
07917283010010F5040BC87238880900F10000993092516195800 3C16010

In the above octet sequence there are three parts: An initial octet indicating the length of the SMSC information ("07"), the SMSC information itself ("917283010010F5"), and the SMS_DELIVER part (specified by ETSI in GSM 03.40).

All the octets above are hexa-decimal 8-bit octets, except the service center number, the sender number and the timestamp; they are decimal semi-octets. The message part in the end of the PDU string consists of hexadecimal 8-bit octets, but these octets represent 7-bit data (Table 6.2). Semi-octets are binary coded decimals, e.g., the sender number is obtained by performing internal swapping within the semi-octets from "72 38 88 09 00 F1" to "27 83 88 90 00 1F". The length of the phone

number in this example has 11 digits (odd). Note that a proper octet sequence cannot be formed by this number. This is the reason why a trailing F has been added. The time stamp, when parsed, equals "99 03 29 15 16 59 80", where the first 6 octets represent date in YYMMDD format; the following 6 octets represents time in HHMMSS format; and the last two represents timezone related to GMT. Timezone 1 signifies 15 minutes. For all operators in India timezone will be 22 (GMT+5.5).

Table 6.2 A received TPDU

RP SC address (optional)	*Value*	*Description*	*Status*
	07	Address length	Length of the SMSC information (in this case 7 octets).
TPDU Octet 1			
	91	Type-of-Address of the SMSC.	(91 means international format of the phone number)
TPDU Octet 2–7			
	72 83 01 00 10 F5	Service center number (in decimal semi-octets).	The length of the phone number is odd (11), so a trailing F has been added to form proper octets. The phone number of this service center is "+27381000015".
TPDU Octet 1 bits	*Value (hex 51)*	*Description*	*Status*
7	0	TP-Reply-Path	Parameter indicating that reply path exists.
6	0	TP-User-Data-Header-Indicator	Indication that user data contains an additional header.
5	0	TP-Status-Report-Request	This bit is set to 1 if a status report is going to be returned to the SME
4	0	TP-Validity-Period-Format	Relative format (bits 4 and 3)
3	0	TP-Validity-Period-Format	Relative format (bits 4 and 3)
2	1	TP-More-Message to-Send	This bit is set to 0 if there are more messages to send
1	0	TP-Message-Type-Indicator	type:SMS-DELIVER (from network to phone), (bits 1 and 0 both set to 0)
0	0	TP-Message-Type-Indicator	type:SMS-DELIVER (from network to phone), (bits 1 and 0 both set to 0)

(*Contd*)

TPDU Octet 3	*Value*	*Description*	*Status*
	0B	Address length	Length of the address is 11 in decimal.
TPDU Octet 4	*Value*	*Description*	*Status*
	91	Type of address	International address using ISDN telephone number plan.
TPDU Octets 5–10	*Value*	*Description*	*Status*
	72 38 88 09 00 F1	TP-Destination-Address	The destination telephone number +27838890001 is encoded as 72 38 88 09 00 F1. In this case the address is 11 digit, therefore a F is added to make it occupy 6 octets. The address field can be anywhere between 2 to 12 octets long.
TPDU Octet 11	*Value*	*Description*	*Status*
	00	TP-Protocol-Identifier, consists of one octet.	Short Message Type 0. This means that the ME must acknowledge receipt of the short message but may discard its contents.
TPDU Octet 12 bits	*Value (hex 15)*	*Description*	*Status*
7	0	TP-Data-Coding-Scheme used in TPUser-Data, consists of one octet. See GSM 3.38	Functionality (bits 7 and 6) related to usage of bits 4-0.
6	0		Functionality (bits 7 and 6) related to usage of bits 4-0.
5	0		Indicates that text is uncompressed
4	0		Indicated that bits 1 and 0 have message class meaning.
3	0	Alphabet being used (bits 3 and 2)	8-bit data
2	0	Alphabet being used (bits 3 and 2)	8-bit data
1	0	Message class (bits 1 and 0)	Class 1, Default meaning: MEspecific
0	0	Message class (bits 1 and 0)	Class 1, Default meaning: MEspecific

(*Contd*)

TPDU Octet 13	*Value*	*Description*	*Status*
	99 30 92 51 61 95 80	TP-Service-Center-Time-Stamp	Format Year, Month, Day; Hour, Minute, Second; Timezone relative to GMT with 1 unit as 15 minutes. 0x99 0x30 0x92 0x51 0x61 0x95 0x80 means 29 March 1999 15:16:59 GMT+2
TPDU Octet 14	*Value*	*Description*	*Status*
	0A	TP-User-Data-Length	Parameter indicating the length of the TP-User-Data field to follow. Represented as amount of octets (integer). 0A hex -> 10 octets. Length includes the user data header and data itself. User data length, length of message. The TP-DCS field indicated 7-bit data, so the length here is the number of septets (10). If the TP-DCS field were set to indicate 8-bit data or Unicode, the length would be the number of octets (9).
TPDU Octets 15–50	*Value*	*Description*	*Status*
	E8329BF D4697D 9EC37	TP-User-Data	The user data. Format of the user data depends on what kind of message is sent. This message includes user data header and the user data itself. The example is 8-bit octets representing 7-bit data. The user data is "hellohello" message.

6.4.2 Example Code for GSM Modem

Following is an example code for GSM modem. This is written in Visual Basic. The code uses Microsoftmscomm controls. The mscomm controls use the COM1 port for communication.

In the example, line 13 is for setting of the communication port and the interface between the computer and the modem.

Line 14–24 is for initialization of the GSM phone as modem.

Line 28–46 is to send a SMS.

Line 50–52 is for reading SMS from the modem.

6.4.3 SMPP

```
 1   ' (c) 2002
 2   ' GSM Modem implementation using MSCOMM and Nokia phone
 3   .
 4   .
 5   .
 6   ' Set up the communications port
 7   MSComm1.CommPort = 1 ' Set COM1 for MSCOMM
 8   ' Set for 9600 baud, no parity, 8 data, and 1 stop bit.
 9   MSComm1.Settings = "9600,N,8,1"
10   ' Tell the control to read entire buffer when Input is used
11   MSComm1.InputLen = 0
12   ' Open the port
13   MSComm1.PortOpen = True
14   ' AT commands are terminated by Carriage Return & Line feed
15   ' Send an initial 'AT' command to the phone
16   MSComm1.Output = "AT" ' Write AT on COM1
17   MSComm1.Output = Chr$(13) ' Write Carriage Return
18   MSComm1.Output = Chr$(10) ' Write Line Feed
19   ' The phone will respond with an 'OK'
20   ' Set the GSM modem so that all SMSs are forwarded to our program
21   MSComm1.Output = "AT+CNMI=1,2,0,1,0" ' Write AT on COM1
22   MSComm1.Output = Chr$(13) ' Write Carriage Return
23   MSComm1.Output = Chr$(10) ' Write Line Feed
24   ' The phone will respond with an 'OK'
25   .
26   .
27   .
28   ' Set up the phone for a text message
29   MSComm1.Output = "AT+CMGF=1" & Chr$(13) & Chr(10)
30   ' The phone will respond with an 'OK'
31   ' Prep for SMS, give destination type and destination address.
32   ' Enter the destination type & address to prep for SMS
33   ' e.g., AT+CMGS="+919845170882",^Z
34   MSComm1.Output = "AT+CMGS="
35   MSComm1.Output = Chr$(34) ' The start quote character
36   MSComm1.Output = "+919845170882" ' Mobile number with country code
37   MSComm1.Output = Chr$(34) ' The end quote character
38   MSComm1.Output = Chr$(13) ' Write Carriage Return
39   MSComm1.Output = Chr$(10) ' Write Line Feed
40   ' The phone will return a'>' prompt, and await entry of the SMS
     message text.
41   ' Now send the text to the phone and terminate with (Ctrl-Z)
```

```
42   MSComm1.Output = "This is a test message" ' Frame the message
43   MSComm1.Output = Chr$(26) ' Add the ^Z
44   ' The phone will respond with a conformation containing the
45   ' Close the port
46   MSComm1.PortOpen = False
47   .
48   .
49   .
50   ' Read the input buffer
51   buffer = MSComm1.Input
52   InpStr = StrConv(buffer, vbUnicode)
53   .
54   .
```

This is an operator dependent solution with high scalability. The SMS data speed is about 300 bits/sec. Using a GSM modem, it takes about 4 seconds to read a message, and about 8 seconds to send a message, resulting in about 5 message pairs/minute. GSM modem can be used for operator independent SMS application. The only limitation of GSM modem is that it cannot scale. The theoretical limit is about 300 message pairs in an hour. If the transaction rate is low, GSM modem can work out to be a convenient and economical way of using SMS for mobile computing. However, if we need reliability and scalability, GSM modem technology is not the best answer. For such cases we need carrier grade solution. For this we need connection directly to the SMSC of the operator. This is achieved through SMPP (Short Message Peer to Peer) protocol.

The SMPP protocol is an open, industry standard protocol designed to provide a flexible data communications interface for transfer of short message data between a Message Center (SC or SMSC) and a VAS application, such as a WAP Proxy Server, Voice Mail Server, EMail Gateway or other Messaging Gateway (Fig. 6.7). An SMPP client is termed a External Short Message Entity (ESME), and is connected to the SC.

SMPP release v3.4 presently supports Digital Cellular Network technologies which include the following.

- GSM
- IS-95 (CDMA)
- CDMA 1X/CDMA 2000
- ANSI-136 (TDMA)
- IDEN

Using the SMPP protocol, an SMS application system called the ESME may initiate an application layer connection with an SC over a TCP/IP or X.25 connection and send and receive short messages. The ESME may also query, cancel or replace short messages using SMPP.

SMPP supports a full featured set of two-way messaging functions such as the ones described below:

- Transmit messages from an ESME to single or multiple destinations via the SMSC.
- An ESME may receive messages via the SMSC from other SMEs (e.g., mobile stations).
- Query the status of a short message stored on the SMSC.

- Cancel or replace a short message stored on the SMSC.
- Send a registered short message (for which a "delivery receipt" will be returned by the SMSC to the message originator).
- Schedule the message delivery date and time.
- Select the message mode, i.e., datagram or store and forward.
- Set the delivery priority of the short message.
- Define the data-coding type of the short message.
- Set the short message validity period.
- Associate a service type with each message, e.g., voice mail notification.

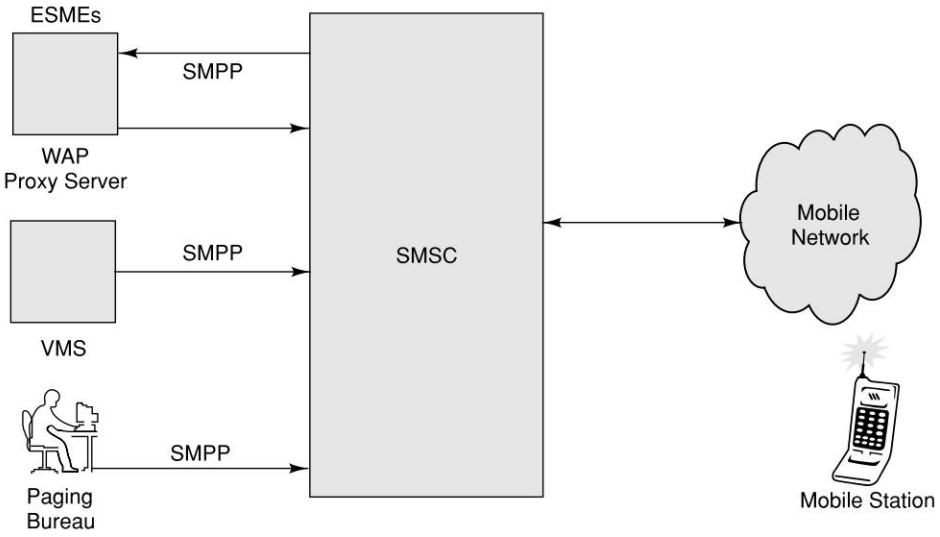


Figure 6.7 Communication between SMSC and SME

SMPP Protocol Overview

SMPP protocol is an open message-transfer protocol that enables Short Message Entities (SMEs) outside the mobile network to interface with an SC. Non-mobile entities that submit messages to, or receive messages from an SMSC are known as External Short Message Entities (ESMEs).

The SMPP protocol defines operations and data as described below.

- Set of operations for the exchange of short messages between an ESME and an SMSC.
- Data that an ESME application must exchange with an SMSC during SMPP operations.

Subscribers to an SMS-capable cellular network may receive short messages on a Mobile Station (MS) from one or more ESMEs. The examples of such ESME applications include the following.

- Voice mail alerts originating from a VMS (Voice Messaging System), indicating voice messages at a customer's mailbox.
- Numeric and alphanumeric paging services.
- Information services. For example, an application that enables mobile subscribers to query

currency rates or share-price information from a database or the WWW and have it displayed as a short message on the handsets.
- Calls directly dialed or diverted to a message-bureau operator, who forwards the message to the SMSC, for onward delivery to a subscriber's handset.
- A fleet management application that enables a central station to use the SMSC to determine the location of its service vehicles and notify the closest vehicle of a service request in their area.
- Telemetry applications. For example, a household meter that transmits a short message to a utility company's billing system to automatically record customer usage.
- WAP Proxy Server. A WAP Proxy Server acts as the WAP gateway for wireless Internet applications. A WAP Proxy Server may select an SMS or USSD bearer for sending WDP (Wireless Data Protocol) datagrams to and receiving WDP datagrams from a mobile station.

There is an open source SMS gateway available from www.kannel.org**.** Along with SMS, Kannel gateway supports WAP and MMS (Multi Media Messaging) interfaces. Kannel offers HTTP interface for message transfer and administrating of the gateway.

Kannel divides its various functions into different kinds of processes (Figure 6.8), called boxes, mostly based on what kinds of external agents it needs to interact with.
- The bearerbox implements the bearer level of SMS. As part of this, it connects to the SMS centers. Definitions of different TCP/IP ports, usernames, passwords, etc., are required to be defined for this connection.
- The smsbox implements the rest of the SMS gateway functionality. It receives textual SMS messages from the bearerbox, and interprets them as service requests, and responds to them in the appropriate way. All the services will be handled and managed by this box.

There can be only one bearerbox, but any number of smsboxes in a single Kannel instance. Duplicating the bearerbox is troublesome. Also, each SMS center can be connected only to one client. While it is possible to have each SMS center served by a different process, it has been deemed not to give enough extra reliability or scalability to warrant the complexity. Having multiple smsboxes can be beneficial when the load is very high. Although the processing requirements as such are fairly low per request, network bandwidth from a single machine, or at least operating system limits regarding the number of concurrent network connections are easier to work around with multiple processes, which can, if necessary, be spread over several hosts. Each box is internally multithreaded. For example, in the bearerbox, each SMS center connection is handled by a separate thread. The thread structures in each of the boxes are fairly static, i.e., the threads are mostly spawned at startup, instead of spawning a new one for each message.

In Kannel we configure the SMS services like:

```
# SMSC Configuration
group = smsc
smsc = smpp
smsc-id = hssup
host = 203.168.14.69
port = 15009
smsc-username = smppabc
```

```
smsc-password = abcocse
address-range = "^333$"
# SMSBOX Configuration
group = smsbox
bearerbox-host = localhost
sendsms-port = 333
global-sender = 333
##########################################################
# Service Configuration
# News
group = sms-service
keyword = news
url = "http://sms.apps.com/SMSService.jsp?keyvalue=%a&mobileno=%p"
max-messages = 4
```

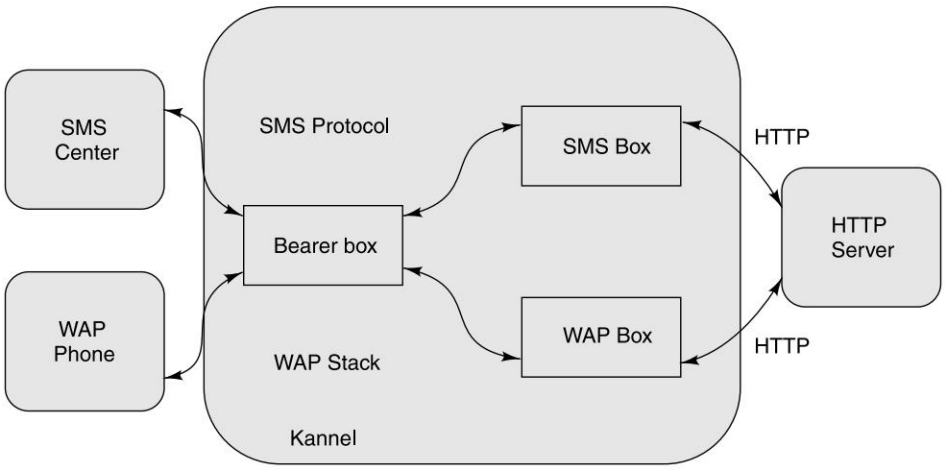


Figure 6.8 Boxes of Kannel

We install Kannel on a Linux system and then connect the same to the SMSC. The connection to SMSC will be through TCP/IP. Kannel also support proprietary protocol like CIMD2 from Nokia SC. In either case, a part of configuration shall define the behavior of bearerbox; some will define the smsbox.

Pull messages

In case of pull transaction, the user enters a message with a keyword and then sends the same to a service number (333 in this example). Let us assume that the network offers a news service. To access the breaking news, the user needs to enter News brk and then send it to 333. During binding of the SMS gateway we intimate the SC that we are listening for 333. Therefore, all the messages sent to 333 will be routed to our SMPP gateway. In the Kannel configuration file, we mention that whenever there is a message with keyword news, it should be forwarded to an HTTP

URL; like, http://sms.apps.com/SMSService.jsp. To service the user with appropriate response, we need to know the request with all the parameters and the MSISDN number of the phone. These are transferred from Kannel gateway to the URL through %a and %p. The complete URL would appear as follows.

```
sms.apps.com/SMSService.jsp?keyvalue=news%20brk&mobileno=91 9845052534.
```

The response of the http request will be forwarded directly to the user (919845052534) by Kannel gateway. If the response from the content/origin server is more than 160 characters, Kannel splits the message into multiple messages. During splitting, Kannel ensures that the messages are split at word boundary. The max-messages parameter defines the limit of maximum number of messages as response. If we set the max-messages to 0, no response will be sent to the user, though there could be some response coming from the http request.

Push messages

In case of push messages, the message is sent through http interface as well. An application uses an HTTP URL to communicate with the Kannel gateway and to send SMS messages. Kannel delivers these messages to the SC. Following are few examples of sending a text, ring tone (binary), picture (binary), and Hindi Unicode messages to a mobile phone 919811557988. To offer a certain level of security Kannel allows the user authentication through user-ID and a password to access these URLs.

```
# TEXT
http://kannel.apps.com:31333/cgi-bin/sendsms?user=
tester&pass=foobar&to=919811557988&text=hello
# BINARY (Ring Tone)
http://kannel.apps.com:31333/cgi-bin/sendsms?user=
   tester&pass=foobar&to=919811557988&udh=%06%05%04%15%81%00%
00&text=%02%4A%3A%71%5D%85%B1%AC%81%BD%98%81%31%A5%99%94%0
4%00%4F%20%CA%E8%38%93%89%20%82%0C%2E%C3%0C%38%83%0C%2C%C2
%A9%2A%92%08%20%42%08%2C%C3%0C%38%93%89%20%82%0C%2C%C3%0C%
38%83%0C%2C&coding=2
```

The significance of UDH in the above example is as follows:

06 – Length of the UDH,

05 – IEI, Information Element Identifier (Application port addressing scheme, 16 bit port address)

04 – IEDL, Information Element Data Length

1581 – Information Element Data (Destination Port)

0000 – Information Element Data (Originator Port)

BINARY (Picture - 3 Messages)

```
http://kannel.apps.com:31333/cgi-bin/sendsms?user=
tester&pass=foobar&to=919811557988&udh=%0B%05%04%15%8A%00%
00%00%03%72%03%01&text=%FF%FF%FF%FA%FE%03%FF%FF%FF%FF%FF%F
```

```
F%F7%FE%01%FF%FF%FF%FF%FF%FF%EB%FC%02%FF%FF%FF%FF%FF%FF%F1
%F0%01%FF%FF%FF%FF%FF%FF%F8%E0%00%FF%FF%FF%FF%FF%FF%F0%00%
01%FF%FF%FF%FF%FF%FF%E0%00%00%FF%FF%FF%FF%FF%FF%E0%00%00%7
F%FF%FF%FF%FF%FF%AB%80%28%EF%FF%FF%FF%FF%FF%45%F1%FC%5F%FF
%FF%FF%FF%FF%A2%EA%E8%CF%FF%FF%FF%FF%FF%F1%11%10%5F%FF%FF%F
F%FF%FF%80%88%A0%BF%FF%FF%FF%FF%FF%C0%00%01%1F%FF%FF%FF%FF"
```

The significance of UDH is as follows:

0B – Length of the UDH

05 – IEI, Information Element Identifier (Application port addressing scheme, 16 bit port address)

04 – IEDL, Information Element Data Length

158A – IED, Information Element Data (Destination Port)

0000 – IED (Originator Port)

00 – IEI (Concatenated Short Message, 8 bit reference number)

03 – IEDL (Information Element Data Length)

03 – IED (Total number of concatenated messages 0-255)

01 – IED (Sequence number of current short message)

```
# Hindi UNICODE
http://kannel.apps.com:31332/cgi-bin/sendsms?user=
tester&pass=foobar&to=919811557988&text=%09%50&coding=3
```

Text is a single character OM in Hindi. The range of Hindi/Devanagari, as decided by Unicode Consortium is within the range 0900–097F.

REFERENCES/FURTHER READING

1. GSM Standard 03.19: Digital cellular telecommunications system (Phase 2+); Subscriber Identity Module Application Programming Interface (SIM API); SIM API for Java Card (TM); Stage 2, Release 1999, http://www.etsi.org.
2. GSM Standard 03.39: Digital cellular telecommunications system (Phase 2+); Interface protocols for the connection of Short Message Service Centres (SMSCs) to Short Message Entities (SMEs), Release 1998, http://www.etsi.org.
3. GSM Standard 03.40: Digital cellular telecommunications system (Phase 2+); Technical realization of the Short Message Service (SMS) Point-to-Point (PP), Release 1997, http://www.etsi.org.
4. GSM Standard 03.47: Digital cellular telecommunications system (Phase 2+); Example protocol stacks for interconnecting Service Centre(s) (SC) and Mobile services Switching Centre(s) (MSC), Release 1998, http://www.etsi.org.
5. GSM Standard 03.48: Digital cellular telecommunications system (Phase 2+); Security mechanisms for SIM application toolkit; Stage 2, Release 1999, http://www.etsi.org.

6. GSM Standard 11.11: Digital cellular telecommunications system (Phase 2+); Specification of the Subscriber Identity Module–Mobile Equipment (SIM–ME) interface, Release 1999, http://www.etsi.org.
7. GSM Standard 11.14: Digital cellular telecommunications system (Phase 2+); Specification of the SIM Application Toolkit (SAT) for the Subscriber Identity Module–Mobile Equipment (SIM-ME) interface, Release 1999, http://www.etsi.org.
8. Introduction to SMS: http://www.gsmworld.com/technology/sms/intro.shtml.
9. Kannel WAP and SMS Gateway: http://www.kannel.org.
10. Kurian, John and Asoke K. Talukder, 2002. 'Ubiquity and Virtual Home Environment for Legacy Applications and Services', *Proceedings of the 15th International Conference in Computer Communication*, pp 371–381, 12–14 August 2002, Mumbai, India.
11. Nokia, AT Command Set for Nokia GSM Products, Nokia Mobile Phones 2000.
12. Pal, Rishi, and Asoke K. Talukder, 2002. 'Global Service Portability and Virtual Home Environment through Short Message Service (SMS) and GSM,' *Proceedings of the 15th International Conference in Computer Communication (ICCC2002)*, pp 8–18, 12–14 August 2002, Mumbai, India.
13. Short Message Peer to Peer Protocol Specification v3.4, Document Version:- 12- Oct-1999 Issue 1.2, http://smsforum.net/.
14. Talukder, Asoke K., Bijendra Singh and Debabrata Das, 'Ubiquitous-Trustworthy-Secure Data Communication in Hybrid Cellular-Internet Networks', *Proceedings of IEEE INDICON* 2004, December 20–22, 2004, pp 135–138. IIT Kharagpur.
15. Talukder, Asoke K., Siddhartha Chhabra, T.S. Sudarsan, K.A. Gowrishankar and Debabrata Das (2005), 'Ubiquitous Rescue Operation at Affordable Cost through Location-Aware SMS', *Proceedings of National Conference on Communication*, IIT-Kharagpur, 28–30 January 2005, pp 651–655.
16. Venkatraman Jesudoss, Vijay Raghavan, Debabrata Das, and Asoke K. Talukder, 'Trust and Security Realization for Mobile Users in GSM Cellular Networks', *Proceedings of Asian Applied Computer Conference*, Kathmandu, October 29–31, 2004, pp 302–309, LNCS 3285.

REVIEW QUESTIONS

Q1: What are various strengths of SMS? Explain each of them. Also, state the various applications areas where these strengths can be used?
Q2: Explain the architecture of SMS.
Q3: Explain the difference between SM MT and SM MO.
Q4: How is SMS used as an information bearer?
Q5: How do you develop location-based applications using SMS?
Q6: Explain each of the following in brief:
(a) Operator Centric Pull

(b) Operator Independent Push
(c) Operator Independent Pull
(d) Value Added Services through SMS
(e) SMS gateway

Q7: How do you use SMS in an application using GSM Modem (Over-The-Air)? Explain different phases in this technology.

Q8: What is SMPP protocol? Why and when is it used? What is the primary difference between SMPP-based technology and GSM Modem technology?

CHAPTER 7

General Packet Radio Service (GPRS)

7.1 INTRODUCTION

People love freedom. In Hollywood movies of yesteryears, we see actors moving around the room and talking on the telephone holding the phone in the left hand and the handset in the right. Today all this has changed. We have cordless phones and wireless mobile phones. People can move around (inside and outside their homes or even inside vehicles) and still talk. As the world is changing, people's expectations are also changing. People are looking for freedom from wire with respect to data. Tech Pundits believe that the trend taking place in fixed networks whereby the growth of data traffic is overtaking that of voice traffic, will also influence the wireless networks. GSM started with voice in mind and offered whatever a wireless voice user wanted. The popularity of GSM, Internet, and digital communication forced GSM to look for wireless data with higher band-width. General Packet Radio Service (GPRS) is a step to efficiently transport high-speed data over the current GSM and TDMA-based wireless network infrastructures.

7.2 GPRS AND PACKET DATA NETWORK

GPRS will thrive in both vertical and horizontal markets where high-speed data transmission over wireless networks is necessary. The deployment of GPRS networks allows a variety of new applications ranging from mobile e-commerce to mobile corporate VPN access. Deployments of GPRS networks has already taken place in several countries in Europe and the Far East. In Mumbai and Delhi GPRS was launched quite sometime ago.

7.2.1 Capacity and Other End-user Aspects

GPRS has the ability to offer data speeds of 14.4 Kbps to 171.2 Kbps, which allow for comfortable Internet access. It allows for short "bursty" traffic, such as e-mail and web browsing, as well as large

volumes of data. To support GPRS operations, new protocols and new network devices are required. By allowing information to be transmitted more quickly, immediately and efficiently across the mobile network, GPRS may well be a relatively less costly mobile data service compared to SMS and Circuit Switched Data. For GPRS, no dial-up modem connection is necessary. It offers fast connection set-up mechanism to offer a perception of being "always on". This is why GPRS users are sometimes referred to as being "always connected". This is like SMS, which is an always-on service. Immediacy is one of the advantages of GPRS compared to Circuit Switched Data.

7.2.2 Quality of Service (QoS)

The Quality of Service (QoS) requirements of typical mobile packet data applications are very diverse. For example the QoS for real-time multimedia content is different from web browsing or email transfer. GPRS allows definition of QoS profiles using the parameters of service precedence, reliability, delay and throughput.

- **Service precedence** is the priority of a service in relation to another service. There exist three levels of priority: high, normal, and low.
- **Reliability** indicates the transmission characteristics required by an application. Three reliability classes are defined, which guarantee certain maximum values for the probability of loss, duplication, mis-sequencing and corruption (an undetected error) of packets.
- **Delay** parameters define maximum values for the mean delay and the 95-percentile delay. The delay is defined as the end-to-end transfer time between two communicating mobile stations or between a mobile station and the signaling interface to an external packet data network.
- **Throughput** specifies the maximum/peak bit rate and the mean bit rate.

Using these QoS classes, QoS profiles can be negotiated between the mobile user and the network for each session.

7.2.3 Integral Part of the Future 3G Systems

The different approaches to third generation (3G) wireless systems (IMT-2000, UMTS, CDMA, WCDMA, 3GPP, 3GPP2 etc.) were intended to address the challenge of voice-to-data crossover and integration. The complexities of new and exciting wireless technologies have slowed down progress in their development and widespread deployment. To lessen the impact of the delay in implementing 3G wireless systems, GPRS was introduced as an intermediate step to efficiently transport high-speed data over the current GSM and TDMA-based wireless network infrastructures. GPRS is therefore called the 2.5G (two and half G or two and half generation) in the evolution process of wireless cellular networks.

7.3 GPRS Network Architecture

GPRS uses the GSM architecture for voice. In order to offer packet data services through GPRS, a new class of network nodes need to be introduced as an upgrade to the existing GSM network.

These network nodes are called GPRS support nodes (GSN). GPRS support nodes are responsible for the delivery and routing of data packets between the mobile stations and the external packet data networks (PDN). There are two types of support nodes, viz., SGSN (Serving GSN) and GGSN (Gateway GSN). Figure 7.1 depicts GPRS system components for data services.

Serving GPRS Support Node (SGSN): A serving GPRS support node (SGSN) is at the same hierarchical level as the MSC. Whatever functions MSC does for voice, SGSN does the same for packet data. SGSN's tasks include packet switching, routing and transfer, mobility management (attach/detach and location management), logical link management, and authentication and charging functions. SGSN processes registration of new mobile subscribers and keeps a record of their location inside a given service area. The location register of the SGSN stores location information (e.g., current cell, current VLR) and user profiles of all GPRS users registered with this SGSN. SGSN sends queries to Home Location Register (HLR) to obtain profile data of GPRS subscribers. The SGSN is connected to the base station system with Frame Relay.

Gateway GPRS Support Node (GGSN): A gateway GPRS support node (GGSN) acts as an interface between the GPRS backbone network and the external packet data networks. GGSN's function is similar to that of a router in a LAN. GGSN maintains routing information that is necessary to tunnel the Protocol Data Units (PDUs) to the SGSNs that service particular mobile stations. It converts the GPRS packets coming from the SGSN into the appropriate packet data protocol (PDP) format for the data networks like Internet or X.25. PDP sends these packets out on the corresponding packet data network. In the other direction, PDP receives incoming data packets from data networks and converts them to the GSM address of the destination user. The readdressed packets are sent to the responsible SGSN. For this purpose, the GGSN stores the current SGSN address of the user and his or her profile in its location register. The GGSN also performs authentication and charging functions related to data transfer.

7.3.1 GPRS Network Enhancements

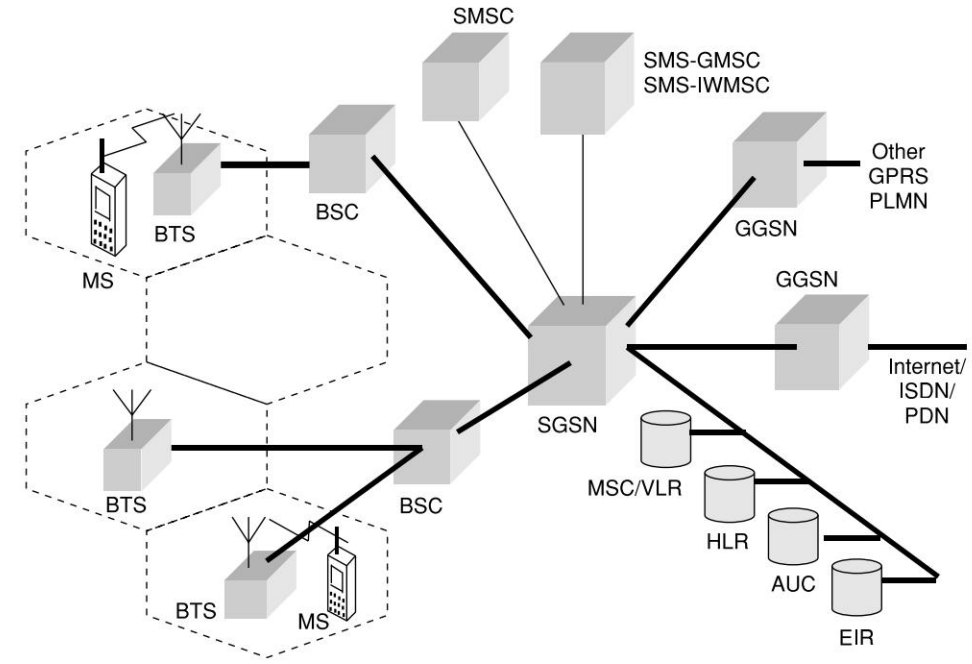
In addition to the new GPRS components (SGSN and GGSN), some existing GSM network elements must also be enhanced in order to support packet data. These are:

Base Station System (BSS): BSS system needs enhancement to recognize and send packet data. This includes BTS upgrade to allow transportation of user data to the SGSN. Also, the BTS needs to be upgraded to support packet data transportation between the BTS and the MS (Mobile Station) over the radio.

Home Location Register (HLR): HLR needs enhancement to register GPRS user profiles and respond to queries originating from GSNs regarding these profiles.

Mobile Station (MS): The mobile station or the mobile phone for GPRS is different from that of GSM.

SMS Nodes: SMS-GMSCs and SMS-IWMSCs are upgraded to support SMS transmission via the SGSN. Optionally, the MSC/VLR can be enhanced for more efficient coordination of GPRS and non-GPRS services and functionality.



AUC	Authentication Centre	MS	Mobile Station
BSC	Base Station Controller	MSC	Mobile Switching Centre
BTS	Base Transceiver Station	PDN	Packet Data Network
EIR	Equipment Identity Register	PLMN	Public Land Mobile Network
GGSN	Gateway GPRS Support Node	SMSC	Short Message Service Centre
GPRS	General Packet Radio Service	SMS-GMSC	SMS Gateway MSC
HLR	Home Location Register	SMS-IWMSC	SMS Inter-Working MSC
ISDN	Integrated System Digital Network	SGSN	Serving GPRS Support Node

Figure 7.1 GPRS System Architecture

7.3.2 Channel Coding

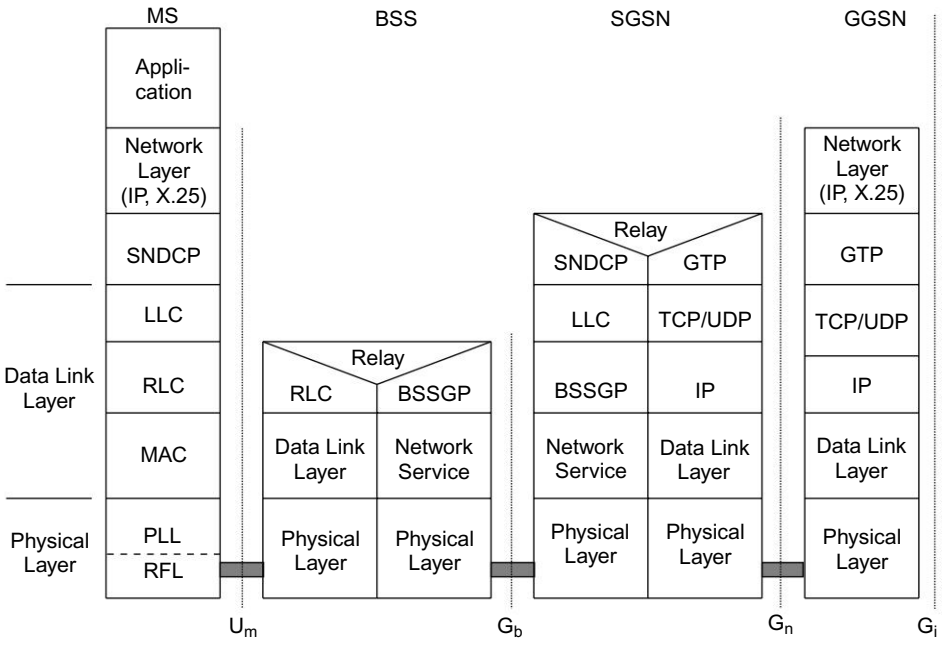
Channel coding is used to protect the transmitted data packets against errors. The channel coding technique in GPRS is quite similar to the one employed in conventional GSM. Under very bad channel conditions, the reliable coding scheme is used. In reliable coding scheme many redundant bits are added to recover from burst errors. In this scheme a data rate of 9.05 Kbps is achieved per time slot. Under good channel conditions, no encoding scheme is used resulting in a higher data rate of 21.4 Kbps per time slot. With eight time slots, a maximum data rate of 171.2 Kbps can be achieved.

7.3.3 Transmission Plane Protocol Architecture

Figure 7.2 illustrates the protocol architecture of the GPRS transmission plane, providing transmission of user data and its associated signaling.

Signaling Plane

The protocol architecture of the signaling plane comprises protocols for control and support of the functions of the transmission plane. This includes GPRS attach and detach, PDP context activation, control of routing paths, and allocation of network resources. The signaling architecture between SGSN and the registers like HLR, VLR, and EIR uses the same protocols as GSM. However, they are extended to support GPRS-specific functionality. Between SGSN and HLR as well as between SGSN and EIR, an enhanced MAP (Mobile Application Part) is employed. MAP is a mobile network-specific extension of the Signaling System SS7 used in GSM. It transports the signaling information related to location updates, routing information, user profiles, and handovers. The exchange of MAP messages is accomplished over the transaction capabilities application part (TCAP) and the signaling connection control part (SCCP). The base station system application part (BSSAP+) is an enhancement of GSM's BSSAP. It is used to transfer signaling information between the SGSN and the VLR.



BSSGP	BSS GPRS Application Protocol	RFL	Physical RF Layer
GTP	GPRS Tunneling Protocol	RLC	Radio Link Control
IP	Internet Protocol	SNDCP	Subnetwork Dependent Convergence Protocol
LLC	Logical Link Control	TCP	Transmission Control Protocol
MAC	Medium Access Control	UDP	User Datagram Protocol
PLL	Physical Link Layer		

Figure 7.2 Transmission Plane and GPRS Protocol Stack

GPRS Backbone

GPRS backbone includes the transmission plane between SGSN and GGSN. User data packets and related signaling information within the GPRS network are encapsulated using the GPRS

Tunneling Protocol (GTP). The GTP protocol is used in both intra- PLMN (between SGSN and GGSN within one PLMN) and inter-PLMN (between SGSN and GGSN of different PLMNs). In the transmission plane, GTP protocol tunnels the user data packets through the GPRS backbone by adding GPRS specific routing information. GTP packets carry the user's data packets from both IP and X.25 data networks. Below GTP, the standard protocols TCP or UDP are used to transport the GTP packets within the backbone network. X.25 expects a reliable data link; therefore TCP is used for tunneling X.25 data. For IP based user data, UDP is used as it does not expect reliability in the network layer or below. Ethernet, ISDN, or ATM-based protocols may be used in the physical layer in the IP backbone. In essence, in the GPRS backbone we have an IP/X.25-over-GTP-over-UDP/TCP-over-IP transport architecture.

BSS–SGSN Interface

The BSS and SGSN interface is divided into the following layers**:**

Sub-Network Dependent Convergence Protocol (SNDCP): The SNDCP is used to transfer data packets between SGSN and MS. Its functionality includes:

- Multiplexing of several connections of the network layer on to one virtual logical connection of the underlying LLC layer.
- Segmentation, compression, and decompression of user data.

Logical Link Control (LLC): A data link layer protocol for GPRS which functions similar to Link Access Procedure–D (LAPD). This layer assures the reliable transfer of user data across a wireless network.

Base Station System GPRS Protocol (BSSGP): The BSSGP delivers routing and QoS-related information between BSS and SGSN.

Network Service: This layer manages the convergence sublayer that operates between BSSGP and the Frame Relay Q.922 Core by mapping BSSGP's service requests to the appropriate Frame Relay services.

Air Interface

The air interface of GPRS comprises the physical and data link layer.

Data Link Layer

The data link layer between the MS and the BSS is divided into three sublayers: the logical link control (LLC) layer, the radio link control (RLC) layer and the medium access control (MAC) layer.

Logical Link Control (LLC): This layer provides a reliable logical link between an MS and its assigned SGSN. Its functionality is based on HDLC (High-level Data Link Control) protocol and includes sequence control, in-order delivery, flow control, detection of transmission errors, and retransmission (automatic repeat request, ARQ). Encryption is used in this interface to ensure data confidentiality. Variable frame lengths are possible. Both acknowledged and unacknowledged data transmission modes are supported. This protocol is an improved version of the LAPDm protocol used in GSM.

Radio Link Control (RLC): The main purpose of the radio link control (RLC) layer is to establish a reliable link between the MS and the BSS. This includes the segmentation and reassembly of LLC frames into RLC data blocks and ARQ of uncorrectable data.

Medium Access Control (MAC): The medium access control (MAC) layer controls the access attempts of an MS on the radio channel shared by several MSs. It employs algorithms for contention resolution, multiuser multiplexing on a packet data traffic channel (PDTCH), and scheduling and prioritizing based on the negotiated QoS.

Physical Layer

The physical layer between MS and BSS is divided into two sublayers: the physical link layer (PLL) and the physical RF Layer (RFL).

Physical Link Layer (PLL): This layer provides services for information transfer over a physical channel between the MS and the network. These functions include data unit framing, data coding, and the detection and correction of physical medium transmission errors. The Physical Link layer uses the services of the Physical RF layer.

Physical RF Layer (RFL): This layer performs the modulation of the physical waveforms based on the sequence of bits received from the Physical Link layer above. The Physical RF layer also demodulates received wave forms into a sequence of bits that are transferred to the Physical Link layer for interpretation.

Multiple Access Radio Resource Management

On the radio interface, GPRS uses a combination of FDMA and TDMA. As in GSM (Fig. 5.10), GPRS uses two frequency bands at 45 MHz apart; viz., 890–915 MHz for uplink (MS to BTS), and 935–960 MHz for downlink (BTS to MS). Each of these bands of 25 MHz width is divided into 124 single carrier channels of 200 kHz width. Each of these 200 kHz frequency channels is divided into eight time slots. Each time slot of a TDMA frame lasts for a duration of 156.25 bit times and contains a data burst.

On top of the physical channels, a series of logical channels are defined to perform functions like signaling, broadcast of general system information, synchronization, channel assignment, paging or payload transport. As with GSM, these channels can be divided into two categories: traffic channels and signaling channels. Traffic channel allocation in GPRS is different from that of GSM. In GSM, a traffic channel is permanently allocated for a particular user during the entire call period (whether any data is transmitted or not). In contrast, in GPRS traffic, channels are only allocated when data packets are sent or received. They are released after the transmission of data. GPRS allows a single mobile station to use multiple time slots of the same TDMA frame for data transmission. This is known as multislot operation and uses a very flexible channel allocation. One to eight time slots per TDMA frame can be allocated for one mobile station. Moreover, uplink and downlink are allocated separately, which efficiently supports asymmetric data traffic like Internet where the bandwidth requirements in uplink and downlink are different.

In GPRS, physical channels to transport user data packet is called data traffic channel (PDTCH). The PDTCHs are taken from a common pool of all channels available in a cell. Thus, the radio resources of a cell are shared by all GPRS and non-GPRS mobile stations located within the cell. The mapping of physical channels to either packet switched data (in GPRS mode) or circuit switched data (in GSM mode) services are performed dynamically depending on demand. This is done depending on the current traffic load, the priority of the service and the multislot class. A load supervision procedure monitors the load of the PDTCHs in the cell. According to the demand, the

number of channels allocated for GPRS can be changed. Physical channels not currently in use by GSM can be allocated as PDTCHs to increase the bandwidth of GPRS.

7.3.4 Security

GPRS security functionality is similar to the existing GSM security. The SGSN performs authentication and cipher-setting procedures based on the same algorithms, keys and criteria as in GSM. GPRS uses a ciphering algorithm optimized for packet data transmission. Like its predecessor, a GPRS device also uses SIM card.

7.4 GPRS Network Operations

Data transmission in a GPRS network requires several steps as described below in the context of the protocol layers described in the previous section. Once a GPRS mobile station is powered on, it "introduces" itself to the network by sending a "GPRS attach" request. Network access can be achieved from either the network side or the MS side of the GPRS network.

7.4.1 Attachment and Detachment Procedure

In order to access the GPRS services, an MS needs to make its presence known to the network. It must register itself with an SGSN of the network. This is done through a GPRS attach. This operation establishes a logical link between the MS and the SGSN. The network checks if the MS is authorized to use the services; if so, it copies the user profile from the HLR to the SGSN, and assigns a packet temporary mobile subscriber identity (P-TMSI) to the MS. In order to exchange data packets with external PDNs after a successful GPRS attach, a mobile station must apply for an addressee. If the PDN is an IP network, it will request for an IP address; for a X.25 network it will ask for a X.25 DTE (Data Terminal Equipment) address. This address is called PDP (Packet Data Protocol) address. For each session, a PDP context is created. It contains the PDP type (e.g., IPv4), the PDP address assigned to the mobile station (e.g., 129.187.222.10), the requested QoS, and the address of the GGSN that will function as the access point to the PDN. This context is stored in the MS, the SGSN and the GGSN. With an active PDP context, the MS is "visible" to the external PDN. A user may have several simultaneous PDP contexts active at a given time. User data is transferred transparently between the MS and the external data networks trough GTP encapsulation and tunneling. User data can be compressed and encrypted for efficiency and reliability.

The allocation of the PDP address can be static or dynamic. In case of static address, the network operator permanently assigns a PDP address to the user. In the other case, a PDP address is assigned to the user upon activation of a PDP context. The PDP address can be assigned by the home network (dynamic home–PLMN PDP address) or by the visited network (dynamic visited–PLMN PDP address). In case of dynamic PDP address assignment, the GGSN is responsible for the allocation and the activation/deactivation of the PDP addresses. This function is similar to the DHCP (Dynamic Host Configuration Protocol) function.

Figure 7.3 shows the PDP context activation procedure. Using the message "activate PDP context request," the MS informs the SGSN about the requested PDP context. If the request is for dynamic

PDP address assignment, the parameter PDP address will be left empty. In following steps security functions (e.g., authentication of the user) are performed. If authentication is successful, the SGSN will send a "create PDP context request" message to the GGSN. The GGSN creates a new entry in its PDP context table, which enables the GGSN to route data packets between the SGSN and the external PDN. The GGSN returns a confirmation message "create PDP context response" to the SGSN, which contains the PDP address. The SGSN updates its PDP context table and confirms the activation of the new PDP context to the MS ("activate PDP context accept").

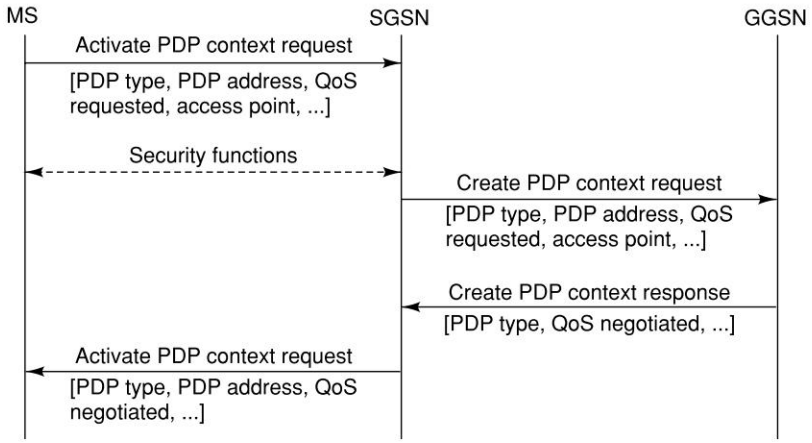


Figure 7.3 PDP Context Activation

The disconnection from the GPRS network is called GPRS detach. All the resources are released following a GPRS detach. Detach process can be initiated by the mobile station or by the network.

7.4.2 APN—Access Point Name

GPRS/EDGE cellular data networks use a mechanism called an Access Point Name (APN) to determine how a Mobile Station (MS), communicates via the GPRS network to a host site; in other words, how does the carrier network passes IP traffic to the host network through the SGSN and GGSN. A mobile device connects to a GPRS network by setting up a PDP (Packet Data Protocol) context. Following this, an Access Point Name (APN) is chosen according to the settings in the mobile device and the SIM card. Next, the chosen APN is used to query the network operator's Domain Name Server (DNS) server. This process gives the IP address of a GGSN allowing the PDP connection to be activated. An APN provides routing information for SGSN and GGSN and defines how users can access the data network at that entry point, what IP addresses are assigned to the mobile station, what security methods are used. APNs are general-purpose and are available to multiple customers or can be customized for particular customers to address unique requirements. For example, some APN by default will not allow mobile terminated connections while others use RADIUS servers and require user name/password authentication in addition to SIM authentication.

7.4.3 Mobility Management

As a mobile station moves from one area to another, mobility management functions are used to track its location within each PLMN. SGSNs communicate with each other to update the MS's location in the relevant registers. The mobile station's profiles are preserved in the VLRs that are accessible to SGSNs via the local MSC. A logical link is established and maintained between the mobile station and the SGSN at each PLMN. At the end of transmission or when a mobile station moves out of the area of a specific SGSN, the logical link is released and the resources associated with it can be reallocated.

7.4.4 Routing

Figure 7.4 depicts an example of how packets are routed in GPRS. The example assumes two intra-PLMN backbone networks of different PLMNs. Intra-PLMN backbone networks connect

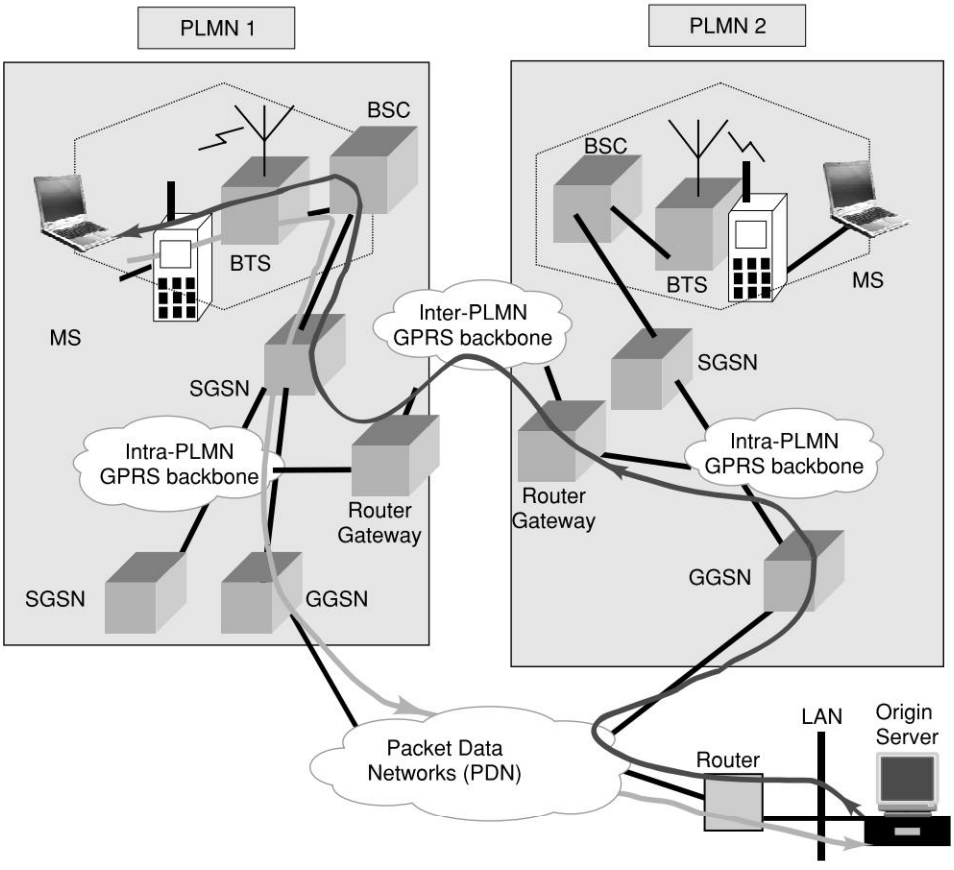


Figure 7.4 GPRS System Architecture and Routing Example

GSNs of the same PLMN or the same network operator. These are private packet-based networks of the GPRS network provider; for example, Airtel GSNs in Bangalore connecting to Airtel GSNs in Delhi through a private data network. In the diagram, these intra-PLMN networks are connected with an inter-PLMN backbone. An inter-PLMN backbone network connects GSNs of different PLMNs and operators. To install such a backbone, a roaming agreement is necessary between two GPRS network providers. For example, Airtel GSNs in Bangalore connect to Vodafone GSNs in Delhi. The gateways between the PLMNs and the external inter-PLMN backbone are called border gateways. Among other things, they perform security functions to protect the private intra-PLMN backbones against unauthorized users and attacks.

We assume that the packet data network is an IP network. A GPRS mobile station located in PLMN1 sends IP packets to a host connected to the IP network, e.g., to a Web server connected to the Internet. The SGSN that the mobile station is registered with encapsulates the IP packets coming from the mobile station, examines the PDP context and routes them through the intra-PLMN GPRS backbone to the appropriate GGSN. The GGSN decapsulates the packets and sends them out on the IP network, where IP routing mechanisms are used to transfer the packets to the access router of the destination network. The latter delivers the IP packets to the host.

Let us assume the home-PLMN of the mobile station is PLMN2. An IP address has been assigned to the mobile by the GGSN of PLMN2. Thus, the MS's IP address has the same network prefix as the IP address of the GGSN in PLMN2. The correspondent host is now sending IP packets to the MS. The packets are sent out onto the IP network and are routed to the GGSN of PLMN2 (the home-GGSN of the MS). The latter queries the HLR and obtains the information that the MS is currently located in PLMN1. It encapsulates the incoming IP packets and tunnels them through the inter-PLMN GPRS backbone to the appropriate SGSN in PLMN1. The SGSN decapsulates the packets and delivers them to the MS.

The HLR stores the user profile, the current SGSN address, and the PDP addresses for every GPRS user in the PLMN. For example, the SGSN informs the HLR about the current location of the MS. When the MS registers with a new SGSN, the HLR will send the user profile to the new SGSN. The signaling path between GGSN and HLR may be used by the GGSN to query a user's location and profile in order to update its location register.

7.4.5 Communicating with the IP Networks

A GPRS network can be interconnected with Internet or a corporate intranet. GPRS supports both IPv4 and IPv6. From an external IP network's point of view, the GPRS network looks like any other IP sub-network, and the GGSN looks like a usual IP router. Figure 7.5 shows an example of how a GPRS network may be connected to the Internet. Each registered user who wants to exchange data packets with the IP network gets an IP address. The IP address is taken from the address space of the GPRS operator maintained by a DHCP server (Dynamic Host Configuration Protocol). The address resolution between IP address and GSM address is performed by the GGSN, using the appropriate PDP context.

Moreover, a domain name server (DNS) managed by the GPRS operator or the external IP network operator is used to resolve host names. To protect the PLMN from unauthorized access, a firewall is installed between the private GPRS network and the external IP network. With this

configuration, GPRS can be seen as a wireless extension of the Internet all the way to a mobile station or mobile computer. The mobile user has direct connection to the Internet.

7.5 Data Services in GPRS

A wide range of corporate and consumer applications are enabled by GPRS services. A user is likely to use either of the two modes of the GPRS network. These are *Application mode* or *Tunneling mode.*

Application mode: In this mode the user will be using the GPRS mobile phone to access the applications running on the phone itself. The phone here acts as the end user device. All GPRS phones have WAP browser as an embedded application. This browser allows browsing of WAP sites. Some GPRS devices support mobile execution environment (MExE classmark 3). These devices support development of client application that can run on the device. The device operating execution environments supported are Symbian and J2ME. Applications can be developed in C/C++ or Java.

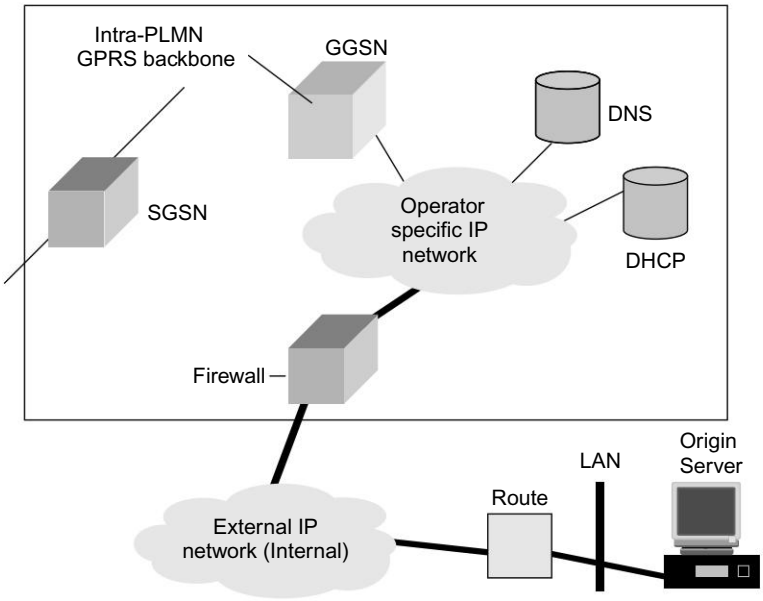


Figure 7.5 Example of a GPRS–Internet Connection

Tunneling mode: This mode is for mobile computing where the user will use the GPRS interface as an access to the network. The end user device will be a large footprint device like laptop computer or small footprint device like PDAs. The mobile phone will be connected to the device and used as a modem to access the wireless data network. For these devices, access can be gained via a PC Card (PCMCI) or via a serial cable to a GPRS capable phone. These "black-box" devices do not have display, keypad and voice accessories of a standard phone.

7.5.1 GPRS Handsets

A GPRS terminal can be one of three classes: A, B or C. A Class A terminal supports GPRS data and other GSM services such as SMS and voice simultaneously. This includes simultaneous attach, activation, monitor, and traffic. As such, a Class A terminal can make or receive calls on two services simultaneously. In the presence of circuit switched services, GPRS virtual circuits will be held or placed on busy rather than being cleared. SMS is supported in Class A terminal. Like GSM, a SMS can be received while a voice or data call is in progress.

A Class B terminal can monitor GSM and GPRS channels simultaneously, but can support only one of these services at any time. Therefore, a Class B terminal can support simultaneous attach, activation, and monitor but not simultaneous traffic. As with Class A, the GPRS virtual circuits will not be closed down when circuit-switched traffic is present. Instead, they will be switched to busy or held mode. Thus, users can make or receive calls on either a packet or a switched call type sequentially but not simultaneously. SMS is supported in Class B terminal. Like GSM, a SMS can be received while a voice or data call is in progress.

A Class C terminal supports only non-simultaneous attach. The user must select which service to connect to. Therefore, a Class C terminal can make or receive calls from only the manually selected network service. The service that is not selected is not reachable. The GPRS specifications state that support of SMS is optional for Class C terminals.

7.5.2 Device Types

In addition to the three types of terminals, each handset will have a unique form factor. Terminals will be available in the standard form factor with a numeric keypad and a relatively small display. Other types of phones and different form factors, color displays, with cameras are common. Smart phones with built-in voice, non-voice and Web-browsing capabilities are common too. Smart phones have various form factors, which may include a keyboard or an icon drive screen.

7.5.3 Bearers in GPRS

The bearer services of GPRS offer end-to-end packet switched data transfer. GPRS is planned to support two different kinds of data transport services. These are the point-to-point (PTP) service and the point-to-multipoint (PTM) service. The PTP service offers transfer of data packets between two users.

GPRS supports the following types of data services:

SMS: Short message service was originally designed for GSM network. GPRS will continue to support SMS as a bearer. Please refer to Chapter 6 for details on SMS and application development using SMS.

WAP: WAP is Wireless Application Protocol. It is a data bearer service over HTTP protocol.

WAP uses WML (Wireless Markup Language) and a WAP gateway. Please refer to Chapter 8 for details of WAP and application development using WAP.

MMS: MMS is Multimedia Messaging Service. This is the next generation messaging service.

SMS supports text messages whereas MMS supports multimedia messages. MMS uses WAP and SMS as its lower layer transport. Video, audio pictures or clips can be sent through MMS. Please refer to Chapter 8 for details of WAP and application development using WAP.

7.6 Applications for GPRS

In this section we describe some applications which need higher data bandwidth and are suitable for GPRS.

7.6.1 Generic Applications

There are many applications suitable for GPRS. Many of them are of generic type, some of them are specific to GPRS. Generic applications are applications like information services, Internet access, email, Web Browsing, which are very useful while mobile. These are generic mass market applications offering content like sports scores, weather, flight information, news headlines, prayer reminders, lottery results, jokes, horoscopes, traffic information and so on. Using Circuit Switched Data (CSD as in GSM), user experience for using these applications have never been enduring. Due to higher bandwidth, mobile Internet browsing will be better suited to GPRS. Access to corporate Intranet can add a new dimension to mobile workers. Mobile commerce is another generic application people may like to use while mobile. Banking over wireless is another generic application. Some Indian banks are offering banking over GPRS/WAP.

7.6.2 GPRS-Specific Applications

Chat: Chat is a very popular service in Internet and GSM (over SMS). Groups of like-minded people use chat services as a means to communicate and discuss matters of common interest. Generally people use different chat services; one, through Internet (offered by Yahoo, ICQ, Google, etc.) and the other, using SMS (offered by mobile operators). GPRS will offer ubiquitous chat by integrating Internet chat and wireless chat using SMS and WAP.

Multimedia Service: Multimedia objects like photographs, pictures, postcards, greeting cards and presentations, static web pages can be sent and received over the mobile network. There are many phones available in the marketplace where a digital camera is integrated with the phone. These pictures can be sent as an electronic object or a printed one. Sending moving images in a mobile environment has several vertical market applications including monitoring parking lots or building sites for intruders or thieves. This can also be used by law enforcement agents, journalists, and insurance agents for sending images of accident site. Doctors can use these applications to send pictures of patients from a health center for expert help.

Virtual Private Network: GPRS network can be used to offer VPN services. Many Bank ATM machines use VSAT (Very Small Aperture Terminal) to connect the ATM system with the banks server. As the bandwidth in GPRS is higher, many banks in India are migrating from VSAT to GPRS-based networks. This is expected to reduce the transaction time by about 25%.

Personal Information Management: Personal diary, address book, appointments, engagements, etc. are very useful for a mobile individual. Some of these are kept in the phone, some in the organizer and some in the Intranet. Using J2ME and WTAI (Wireless Telephony, Application Interface), the address book, the diary of the phone can be integrated with the diary at the home office. GPRS and other bearer technology will help achieve this.

Job Sheet Dispatch: GPRS can be used to assign and communicate job sheets from office-based staff to mobile field staff. Customers typically telephone a call center whose staff takes the call and categorize it. Those calls requiring a visit by field sales or service representative can then be escalated to those mobile workers. Job dispatch applications can optionally be combined with vehicle positioning applications so that the nearest available suitable personnel can be deployed to serve a customer.

Unified Messaging: Unified messaging uses a single mailbox for all messages, including voice mail, fax, e-mail, SMS, MMS, and pager messages. With the various mailboxes in one place, unified messaging systems then allow for a variety of access methods to recover messages of different types. Some will use text-to-voice systems to read e-mail and, less commonly, faxes over a normal phone line, while most will allow the interrogation of the contents of the various mailboxes through data access, such as the Internet. Others may be configured to alert the user on the terminal type of their choice when messages are received.

Vehicle Positioning: This application integrates GPS (Global Positioning System) that tell people where they are. GPS is a free-to-use global network of 24 satellites run by the US Department of Defense. Anyone with a GPS receiver can receive their satellite position and thereby find out where they are. Vehicle-positioning applications can be used to deliver several services including remote vehicle diagnostics, ad hoc stolen vehicle tracking and new rental car fleet tariffs. In India this application is becoming popular in logistics industry.

Location-based Services and Telematics: Location-based services provide the ability to link push or pull information services with a user's location. Examples include hotel and restaurant finders, roadside assistance, and city-specific news and information. All systems developed for Intelligent Transportation System (ITS) are built around GPRS and GPS technology. Location can be determined either through GPS or cell identification from the operator. This technology also has vertical applications such as workforce management and vehicle tracking.

7.7 LIMITATIONS OF GPRS

There are some limitations with GPRS, which can be summarized as:

Limited Cell Capacity for All Users: There are only limited radio resources that can be deployed for different uses. Both Voice and GPRS calls use the same network resources. Use for one data precludes simultaneous use for voice. If tariffing and billing are not done properly, this may have impact on revenue.

Speed Lower in Reality: Achieving the theoretical maximum GPRS data transmission speed of 172.2 Kbps would require a single user taking over all eight time slots without any error protection.

It is unlikely that a network operator will allow all time slots to be used by a single GPRS user. Additionally, the initial GPRS terminals are expected to be supporting only one, two or maximum three time slots. Normally, GPRS provides data rates of 56–114 Kbps in 2G systems.

7.8 Billing and Charging in GPRS

There is a saying in the wireless business community, "Data sells, voice pays." Tariffing of data in wireless network has always been a challenge. For voice networks tariffs are generally based on distance and time. This in other words means that users pay more for long distance calls. They also pay more if they keep the circuit busy by talking for a longer period of time. In a voice system, charging is the fundamental part of the architecture. On the other hand, data services have evolved from research and education without any concept of charging. In packet network keeping the circuit busy does not have any meaning. Also, charging a customer by the distance traversed by a packet does not make any sense. Many times due to congestions packets traverse much longer distance than the optimum distance.

7.8.1 Tariffing

The main challenge for a network operator is to integrate these two models and charge the customer. Decisions on charging for GPRS by packet or simply a flat monthly fee are contentious but need to be made. Charging different packets at different rates can make things complicated for the user, whilst flat rates favor heavy users more than occasional ones. It is believed that the optimal GPRS pricing model be based on two variables, time and packet. Network operators levy a nominal per packet charge during peak times plus a flat rate. There will be no per packet charge during non-peak times. Time and packet-related charging will encourage applications such as remote monitoring, social network and chat to use GPRS at night when spare network capacity is available. Simultaneously, a nominal per packet charge during the day will help to allocate scarce radio resources, and charge radio heavy applications such as file and image transfer more than applications with lower data intensity. It has the advantage of automatically adjusting customer charging according to their application usage.

7.8.2 Billing

GPRS is essentially a packet switching overlay on a circuit switching network. The GPRS specifications stipulate that the minimum charging information that must be collected:

- Destination and source addresses.
- Usage of radio interface.
- Usage of external packet data networks.
- Usage of the packet data protocol addresses.
- Usage of general GPRS resources and location of the mobile station.

Since GPRS networks break the information to be communicated down into packets, at a minimum, a GPRS network needs to be able to count packets to charging customers for the volume of packets they send and receive. Today's billing systems have difficulties in handling the charging

process for today's data services. It is unlikely that circuit switched billing systems will be able to process a large number of new variables created by GPRS.

GPRS call records are generated in the GPRS Service Nodes. The incumbent billing systems are often not able to handle real time Call Detail Record flows. As such, an intermediary charging platform is a good idea to perform billing mediation by collecting the charging information from the GPRS nodes and preparing it for submission to the billing system. Packet counts are passed to a Charging Gateway that generates Call Detail Records that are sent to the billing system.

The billing of the services can be based on the transmitted data volume, the type of service, and the chosen QoS profile. It may well be the case that the cost of measuring packets is greater than their value. The implication is that there will not be a per packet charge since there may be too many packets to warrant counting and charging for. For example, a single traffic monitoring application can generate tens of thousands of packets per day. Thus the charging gateway function is more a policing function than a charging function since network operators are likely to tariff certain amounts of GPRS traffic at a flat rate and then need to monitor whether these allocations are far exceeded. The billing of roaming GPRS subscribers from one network to another is still a challenge.

In GPRS architecture, the Charging Gateway Functionality (CGF) provides a mechanism to transfer charging information from the SGSN and GGSN nodes to the network operator's chosen Billing Systems (BS). The Charging Gateway concept enables an operator to have just one logical interface between the CGF and the BS. Details of the GPRS charging function can be found in 3GPP TS 12.15 standards.

7.9 Enhanced Data Rates for GSM Evolution (EDGE)

Enhanced Data rates for GSM Evolution (EDGE), is an upgrade to GSM/GPRS networks to offer higher capacity and performance in data transmission. EDGE is an improvement to the GPRS air interface that enables higher user bit-rates and greater system capacity by enhancing capabilities of the physical layer. Core nodes in GPRS, like SGSN and GGSN are not affected by these higher bit-rates; and hence, no new hardware is required for EDGE; only base station transceivers need to be EDGE-capable. Also, the BSS and BSC software needs to be updated. EDGE is also called Enhanced GPRS (EGPRS) or IMT Single Carrier (IMT-SC).

EDGE has improved steadily since its introduction and, today it offers user bit-rates of around 250 Kbit/s, with end-to-end latency of less than 150 ms. To achieve this, EDGE uses 8 phase shift keying (8PSK) for the upper five of its nine modulation and coding schemes. In effect, EDGE makes a 3-bit word for every change in carrier phase. This results in tripling of the gross data rate offered by GSM. Like GPRS, EDGE too, uses a rate adaptation algorithm that adapts the Modulation and Coding Scheme (MCS) according to the current quality of the radio channel. EDGE introduces Incremental Redundancy which (instead of retransmitting disturbed packets) sends more redundancy information to be combined in the receiver. This results in increased probability of correct decoding. 8-PSK, it is a high level linear modulation method that carries

three times more information through an extended signal constellation. Figure 7.6 shows GMSK in comparison to 8-PSK.

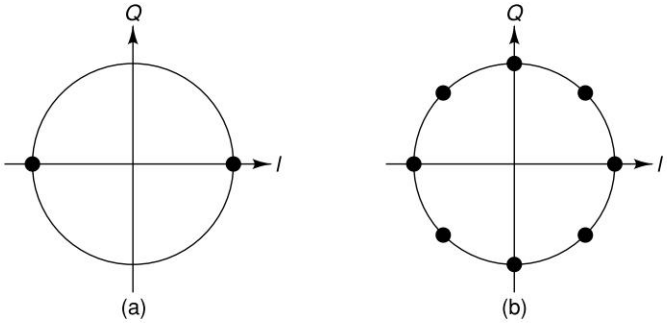


Figure 7.6 (a) GMSK–1bit/symbol and (b) 8PSK–3 bits/symbol

Table 7.1 DTMF frequencies

Technology	*Download (Kbit/s)*	*Upload (Kbit/s)*	*TDMA Time slots allocated*
GPRS	80.0	20.0 (Class 8 & 10 and CS-4)	4+1
GPRS	60.0	40.0 (Class 10 and CS-4)	3+2
EDGE	236.8	59.2 (Class 8, 10 and MCS-9)	4+1
EDGE	177.6	118.4 (Class 10 and MCS-9)	3+2

7.9.1 Evolved EDGE

The next generation of EDGE is called Evolved EDGE that is capable of offering data rates close to 1 Mbit/s. In this EDGE Evolution, latencies would be reduced by lowering the Transmission Time Interval (TTI) from 20 ms to 10 ms. This would be through using dual carriers, higher symbol rate and higher-order modulation (32 QAM and 16 QAM instead of 8-PSK) and turbo codes to improve error correction. Though it is still in the test phase, it is being hailed as a cheaper alternative to full-fledged deployment to 3G networks.

REFERENCES/FURTHER READING

1. Andersson Christoffer, *GPRS and 3G Wireless Applications* (2001), John Wiley & Sons.
2. Bettstetter Christian, Hans-Jorg Vogel and Jorg Eberspacher, "GSM Phase 2+ General Packet Radio Servie GPRS: Architecture, Protocols, and Air Interface," *IEEE Communications Surveys*, Vol. 2, No. 3, Third Quarter 1999.

3. Engadget Mobile – www.engadgetmobile.com.
4. General Packet Radio Service (GPRS); GPRS Charging, 3GPP TS 12.15 version 7.7.0.
5. GSM Standard 03.01: Digital cellular telecommunications system (Phase 2+); Network functions, Release 1998, http://www.etsi.org.
6. GSM Standard 03.02: Digital cellular telecommunications system (Phase 2+); Network architecture, Release 1998, http://www.etsi.org.
7. GSM Standard 03.03: Digital cellular telecommunications system (Phase 2+); Numbering, addressing and identification, Release 1998, http://www.etsi.org.
8. GSM Standard 03.09: Digital cellular telecommunications system (Phase 2+); Handover procedures, Release 1998, http://www.etsi.org.
9. GSM Standard 03.20: Digital cellular telecommunications system (Phase 2+); Security related network functions, Release 1998, http://www.etsi.org.
10. GSM Standard 03.60: Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS); Service description; Stage 2, Release 1998, http://www.etsi.org.
11. *Introduction to GPRS*: http://www.gsmworld.com/technology/gprs/intro.shtml.
12. Kurian, John and Asoke K. Talukder, "Ubiquity and Virtual Home Environment for Legacy Applications and Services", *Proceedings of the 15th International Conference in Computer Communication*, pp 371–381, 12–14 August 2002, Mumbai, India.
13. www.ericsson.com
14. Wikipedia – www.wikipedia.org

REVIEW QUESTIONS

Q1: What is the difference between GSM and GPRS? What are the network elements in GPRS that are different from GSM?

Q2: Explain the GPRS architecture with its constituent elements.

Q3: How is data handled in GPRS?

Q4: How is data routing done in GPRS? In what respect is data routing different from voice routing?

Q5: Explain each of the following in the context of GPRS network:
- (a) QoS parameters
- (b) SGSN
- (c) GGSN
- (d) Channel Coding
- (e) GPRS Data Link Layer
- (f) GPRS Physical Layer

Q6: Explain PDP context activation with respect to GPRS networks.

Q7: Explain call routing in the context of GPRS networks.
Q8: Describe applications suitable for GPRS.
Q9: Describe the various limitations of GPRS.
Q10: What is the motivation behind EDGE?
Q11: Explain the changes needed for implementing an EDGE system.
Q12: What is Evolved EDGE?

CHAPTER 8

Wireless Application Protocol (WAP)

8.1 INTRODUCTION

We are moving towards a net-centric world, where Internet is becoming part of our environment. Along with the physical environment, we also acquire information and knowledge from the Internet. The appetite for data and information over communication networks are growing day by day and has made people look at cellular networks for data access as well. In 2G cellular networks, data access was possible over mobile phones using SMS (Short Message Service) (Chapter 6) and WAP (Wireless Application Protocol) over circuit (circuit switched data—CSD). In a circuit, the user pays for the circuit even during the idle period when there is no data transmission. Also, the data speed supported by CSD is in the range of 9.6K bits per second. For Internet access, 9.6 Kbps speed is unlikely to offer a good user experience. GPRS (Chapter 7) is designed to overcome some of these constraints of GSM and offer a higher data rate.

In Chapter 7 we have discussed how General Packet Radio Service (GPRS) is designed to efficiently transport high-speed data over the current GSM and TDMA-based cellular networks. GPRS is a packet network with higher bandwidth compared to CSD in GSM. Deployments of GPRS networks have already begun in several countries including India. GPRS allows for short "bursty" traffic, such as e-mail and web browsing, including large volumes of data. GPRS has the ability to offer data speeds from 14.4 Kbps to 171.2 Kbps. Internet traffic is asymmetric in the sense that the traffic volume from network to the user agent (client device) is significantly higher compared to the traffic in reverse direction. GPRS manages asymmetric traffic quite well by dynamically configuring bandwidth. Being a packet network, GPRS may well be a relatively less expensive mobile data service compared to SMS or CSD.

Wireless application protocol (WAP) is designed for access to Internet and advanced telephony services from mobile phones. WAP pays proper sensitivity to the constraints of these devices like small display, limited keys on the keypad, no pointer device like mouse, etc. Independent of their network, bearer and terminals, a user will be able to access Internet and corporate intranet services

while mobile. Net–net using WAP, a mobile user will be able to access the same wealth of information from a pocket-sized device as s/he can from a desktop. Though WAP can be used from a variety of networks, GPRS and 3G networks are more suited for these applications. In this chapter we will discuss application development using the WAP and MMS (Multimedia Messaging Service).

8.1.1 Evolution of Wireless Data and WAP

In 1992, Nippon Telegraph and Telephone Corporation (NTT), a leading telephone company in Japan spun off a wireless division and named it DoCoMo. DoCoMo was derived combining two syllables Do and Como. Do in Japanese mean "everywhere" and como was for Communications. DoCoMo's success has been due to i-mode, its mobile Internet service. DoCoMo developed a language called cHTML (Compact Hyper Text Markup Language) and a gateway. Using these frameworks an i-mode user can use the Internet.

In 1994, in the US a company named Unwired Planet was founded to develop and market a platform for Internet access through wireless devices like PDA. Unwired Planet developed a comprehensive framework including browser, gateway and markup language. The markup language was called HDML (Handheld Device Markup Language). Unwired Planet also developed HDTP (Handheld Device Transport Protocol). Unwired Planet launched all these technologies in 1995.

In 1995, another leading wireless company in Europe Ericsson, began work on a protocol known as ITTP or Intelligent Terminal Transfer Protocol. ITTP was designed with the intent of making it easy for call control and add services to mobile telephony platforms.

In 1997 Nokia, the other major wireless company in Europe developed the TTML (Tagged Text Mark-up Language). TTML was designed to allow a mobile phone to communicate with a World Wide Web site via gateway. Following TTML, Nokia also introduced Narrowband Sockets (NBS) to create wireless messaging applications for PCs communicating with GSM "smart phone".

Out of all these different technologies, i-mode in Japan was the most successful service. It started growing like a bushfire. In the US and Europe, though there were different services and protocols offered by different companies, none of them could match the popularity of DoCoMo. Realizing that these competing wireless protocols could fragment and possibly destroy the potential market, in June of 1997, Ericsson, Motorola, Nokia, and Unwired Planet (now known as Phone.com) joined hands to launch the WAP Forum (www.wapforum.com). WAP Forum is now known as Open Mobile Alliance. The goal of this effort was to produce a refined, license-free protocol, which is independent of the underlying airlink standard. The WAP inherited its main characteristics and functionality from HDML and HDTP developed by Unwired Planet, the Smart Messaging specification based on TTML and NBS developed by Nokia, and the ITTP specification developed by Ericsson. The first release of the WAP 1.0 specifications was released in the spring of 1998.

8.1.2 Networks for WAP

As part of the WAP Forum's goals, WAP is accessible from (but not limited to) the following networks:

- GSM-900, GSM-1800, GSM-1900
- GPRS
- CDMA IS-95, cdma2000

- TDMA IS-136
- i-mode
- 3G systems—IMT-2000, UMTS, W-CDMA, Wideband IS-95.

WAP can be used through 2G, 2.5G, and 3G networks. There is a perception that WAP or MMS requires GPRS networks. This is not correct. WAP and MMS can be accessed technically from a 2G network using CSD. However, high-speed data networks like GPRS are more suitable for WAP and MMS applications.

8.2 WAP

WAP forum develops standards for application deployment over wireless devices like PDAs and mobile phones. WAP is based on layered architecture. The WAP Protocol Stack is similar to the OSI network model (Fig. 8.1). These layers consist (from top to bottom) of:

- Wireless Application Environment (WAE).
- Wireless Session Protocol (WSP).
- Wireless Transaction Protocol (WTP).
- Wireless Transport Layer Security (WTLS).
- Wireless Datagram Protocol (WDP).

The application environment of WAE comprises multiple components to provide facilities like:

- User agent: the browser or a client program.
- Wireless Markup Language (WML): a lightweight markup language, similar to HTML, but optimized for use in wireless devices.

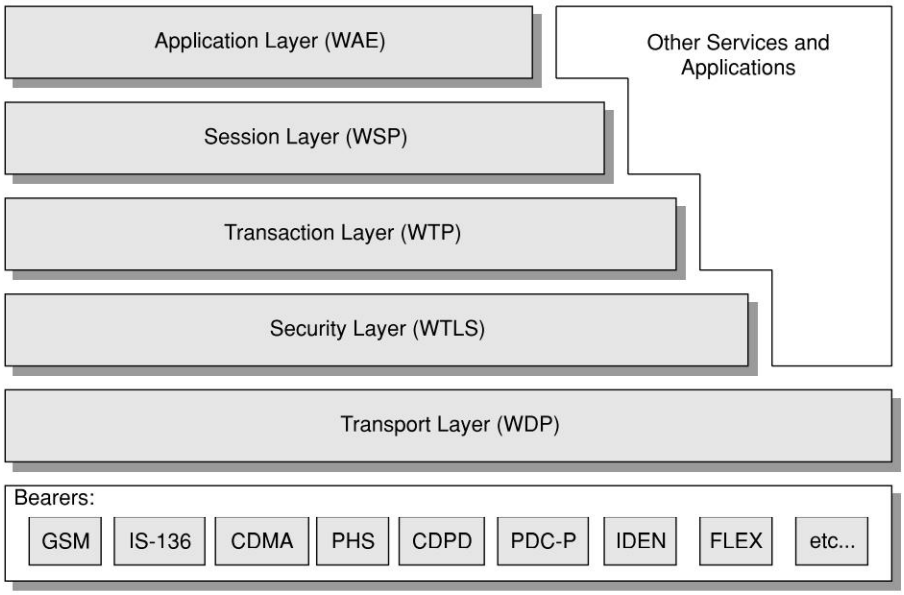


Figure 8.1 WAP Layered Architecture and Protocol Stack

- WMLScript: A lightweight client side scripting language, similar to JavaScript in Web.
- Wireless Telephony Application: Telephony services and programming interfaces.
- WAP Push Architecture: Mechanisms to allow origin servers to deliver content to the terminal without the terminal requesting for it.
- Content Formats: A set of well-defined data formats, including images, phone book records and calendar information.

WAP supports different types of bearer networks. These are GSM, IS-136, CDMA, PHS (Personal Handyphone System), CDPD (Cellular Digital Packet Data), etc.

8.2.1 WAP Application Environment (WAE)

The primary objective of the WAP application environment (WAE) is to provide an interoperable environment to build services in wireless space. It covers system architecture relating to the user agents, networking schemes, content formats, programming languages and shared services based on World Wide Web (WWW) technologies. Content is transported using standard protocols in the WWW domain and an optimized HTTP-like protocol in the wireless domain. WAE architecture allows all content and services to be hosted on standard Web servers. All contents are located using WWW standard URLs. WAE enhances some of the WWW standards to reflect some of the telephony network characteristics.

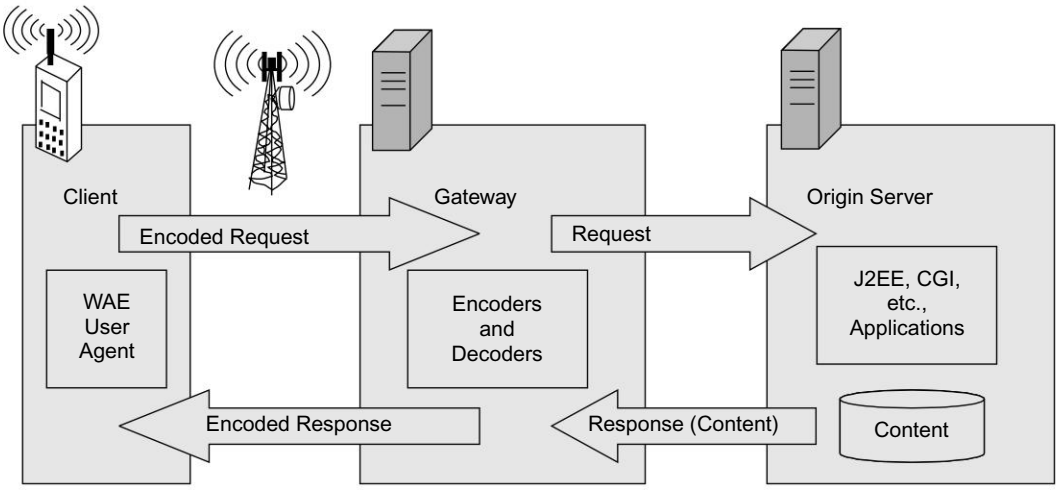


Figure 8.2 WAE Logical Model

A WAP request from the browser (user agent) is routed through a WAP gateway (Fig. 8.2). The gateway acts as an intermediary between the client and network through a wireless last mile (GSM, GPRS, CDMA, etc.). The gateway does encoding and decoding of data transferred from and to the mobile user agent. The purpose of encoding is to minimize the size of data transacted over-the-air. Reduced data size reduces the computational power required by the client to process that data. In most cases the WAP gateway resides on the TCP/IP network. The gateway processes the request,

retrieves content from the server using Java servlets, J2EE, CGI scripts, or some other dynamic mechanism. The data is formatted as WML (Wireless Markup Language) and returned to the client. The client device can employ logic via embedded WMLScript for client-side processing of WML.

The major elements of the WAE model include:

- *WAE User Agents:* User facing client software (browser). User agents are integrated into the WAP architecture. They interpret network content referenced by a URL. WAE includes user agents for two primary standard contents: encoded WML and compiled WMLScript.
- *Content Generators:* Applications on origin servers that extract standard content in response to requests from user agents. Content servers are typically HTTP servers as used in WWW.
- *Standard Content Encoding:* A set of well-defined content encoding, allowing a WAE user agent to navigate web content. Standard content encoding includes compressed encoding for WML, bytecode encoding for WMLScript, standard image formats, business, and calendar data formats.
- *Wireless Telephony Applications (WTA):* A collection of telephony specific extensions for call and telephony feature control.

WAE defines a set of user agent capabilities that is exchanged between the client and the server using WSP. These capabilities include global device characteristics as WML version, WMLScript version, floating-point support, image formats, and so on.

8.2.2 User Agent

Technically, a user agent signifies someone who works on behalf of the user. In the WWW and WAE context, user agent is the user facing browser software. In WAE this is generally referred to as micro-browser. WAE does not formally specify the functionality of any user agent. WAE only defines fundamental services and formats that are needed to ensure interoperability among implementations and different layers. Features and capabilities of a user agent are left to the implementers. WAE is not limited to a WML user agent. WAE allows the integration of domain-specific user agents as well. A Wireless Telephony Application (WTA) user agent has been specified as an extension to the WAE specification for the mobile telephony environments. This covers features like call control as well as other applications in the telephones, such as phonebooks and calendar applications. Following sections provide details on WML, and WMLScript.

8.2.3 User Agent Profile (UAProf)

The User Agent Profile (UAProf) specification allows WAP to notify the content server about the device capability. UAProf is also referred to as ***Capability and Preference Information*** (CPI). CPI is passed from the WAP client to the origin server through intermediate network points. It is compatible with Composite Capability/Preference Profile (CC/PP) of the W3C (WWW Consortium). This CPI may include, hardware characteristics (screen size, color capabilities, image capabilities, manufacturer, etc.), software characteristics (operating system vendor and version, support for MExE, list of audio and video encoders, etc.), application/user preferences (browser manufacturer and version, markup languages and versions supported, scripting languages supported, etc.), WAP characteristics (WML script libraries, WAP version, WML deck size, etc.). In a WSP

response, it transmits information about the client, user, and network that will be processing the content.

Devices that support UAProf architecture provide a URL in the WAP or HTTP session header. This URL points to a XML file that describes the profile of that device. Many vendors have their own public HTTP-servers where service providers can download device profiles as standardized XML documents. In case of MMS (Multimedia Message Service), the MMSC (MMS Controller) is able to pick the profile address from the protocol header and fetch the respective device profile. Device profile information is used by the MMSC to format the content to best suit the terminal's capabilities. Content can be adapted based on the device capabilities. For example, the scaling of a bitmap and adjusting its color map may be required to fit the display size or reduce the size of an image or a music file.

8.2.4 Wireless Markup Language (WML)

WML is a tag-based document manipulation language. It shares a heritage with HTML of W3C and HDML of Unwired Planet. WML is designed to specify presentation and user interaction on mobile phones and other wireless devices. These devices suffer from different constraints like small displays, limited user-input facilities, narrow band network connections, limited memory resources, limited computational resources, and absence of pointer devices like mouse.

WML implements a deck and card metaphor. A *deck* is a logical representation of a document. Decks are made up of multiple *cards*. Each WML card, in a deck, performs a specific task for a particular user interaction. To access a document, a user navigates to a card; reviews its contents, makes a choice or enters requested information, and then moves to another card. WML decks can be stored in "static" files and fetched by CGI, JSP or ASP scripts. It can also be dynamically generated by a Java servlets running on an origin server.

WML has a wide variety of features, including:

- *Support for Text and Images:* WML provides the facility to render text and images to the user. WML provides a limited set of text mark-up elements. These include:
  - ❏ *Emphasis* elements like bold, italic, big, etc.
  - ❏ *Line breaks* models like line wrapping, line wrapping suppression, etc.
  - ❏ *Tab columns* that support simple tabbing alignment.
- *Support for User Input:* WML supports several elements to solicit user input. WML includes a *text entry* control that supports text and password entry. Text entry fields can also be masked to prevent the user from entering unwanted character types. WML includes an *option selection* control that allows the author to present the user with a list of options that can set data, navigate among cards, or invoke scripts. WML supports client-side validation by allowing the author to invoke scripts to check the user's input.
- *Task invocation Controls:* These controls initiate a navigation or a history management task such as traversing a link to another card (or script) or popping the current card off of the history stack. WML also allows several navigation mechanisms using URLs. Navigation includes HTML-style hyperlinks, inter-card navigation elements, as well as history navigation elements.
- *International Support:* The character set for WML document is the Universal Character Set. Currently, this character set is identical to Unicode 2.0.

- *MMI Independence:* WML's specification of layout and presentation enables terminal and device vendors to control the MMI design for their particular products.
- *Narrow-band Optimization:* WML specification includes different technologies to optimize traffic on a narrow-band device. This includes the ability to specify multiple user interactions (cards) in one document transfer (a deck).
- *State and Context Management:* Each WML input control can introduce variables. The lifetime of a variable state can extend beyond a single deck. The state can be shared across multiple decks without having to use a server to save intermediate state between deck invocations.

WML is mostly about rendering text. Many tags that are available in HTML and demand a high resource are generally not supported in WML. The use of tables and images are restricted. Since WML is an XML application, all tags are case sensitive (`<wml>` is not same as `<WML>`). In WML all tags must be properly closed. Cards within a deck can be related to each other with links. A card element can contain text, input-fields, links, images etc. When a WML page is accessed from a mobile phone, all the cards in the page are downloaded from the WAP server. Navigation between the cards is done inside the phone without any extra access trips to the server. Following is an example of Hello World in WML.

```
<?xml version="1.0"?>
<!DOCTYPE wml PUBLIC "-//WAPFORUM//DTD WML 1.1//EN"
"http://www.wapforum.org/DTD/wml_1.1.xml">
<wml>
   <card id="no1" title="Card 1">
      <p>Hello World!</p>
   </card>
   <card id="no2" title="Card 2">
      <p>Welcome to WML!</p>
   </card>
</wml>
```

In the above example, the WML document is an XML document. The DOCTYPE is defined to be wml, and the DTD is accessed at www.wapforum.org/DTD/wml_1.1.xml.

The document content is inside the `<wml>...</wml>` tags. Each card in the document is inside `<card>...</card>` tags, and actual paragraphs are inside `<p>...</p>` tags. Each card element has an identifier and a title.

When the above WML deck is executed, the result will look like this (Figure 8.3):

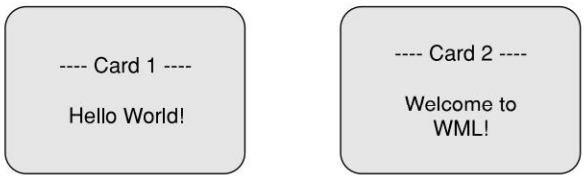


Figure 8.3 Output from the Example Hello World Application

8.2.5 WMLScript

WMLScript is an extended subset of JavaScript and forms a standard means for adding procedural logic to WML decks. WMLScript is used to do client side processing. Therefore, it can be used very effectively to add intelligence to the client and enhance the user interface. Using WMLScript, it is possible to access the device resources. WMLScript provides the application programmer with a variety of interesting capabilities. These are as follows:

- The ability to do local validation of user input before it is sent to the content server.
- The ability to access device resources, functions, and peripherals.
- The ability to interact with the user without reference to the origin server.

Key WMLScript features include:

- *JavaScript-based scripting language:* WMLScript is based on industry standard JavaScript solution and adapts it to the narrow-band environment.
- *Procedural Logic:* WMLScript adds the power of procedural logic to WML.
- *Compiled implementation:* WMLScript can be compiled to a more space efficient bytecode that is transported to the client.
- *Event-based:* WMLScript may be invoked in response to certain user or environmental events.
- *Integrated into WAE:* WMLScript is fully integrated with the WML browser. WMLScript has access to the WML state model and can set and get WML variables.
- *International Support:* WMLScript supports Unicode 2.0.
- *Efficient extensible library support:* WMLScript can be used to expose and extend device functionality without changes to the device software.
- *Data types:* Following basic data types are supported in WMLScript: *boolea*n, *intege*r, *floating-poin*t, *string and invali*d. WMLScript attempts to automatically convert between the different types as needed.

8.2.6 Wireless Telephony Application (WTA, WTAI)

WAP offers WTAI (Wireless Telephony Application Interface) functions to create Telephony Applications. This is achieved through a wireless telephony application (WTA) user-agent using the appropriate WTAI function. For example, let us say that we want to book a table for a lunch meeting in a restaurant. From the WAP application, we go to the restaurant site and get the telephone number. In normal case we note down the telephone number on a piece of paper, exit from the browser session, and then make a voice call to book the table. In case of WTAI that is not required. We can display an action item call in the WAP screen and make a call straight from the WAP page. The WTAI function libraries are accessed from server side using URL's; or at the client side through WMLScript.

There are different library functions to do different telephony functions:

- *Voice Call Control:* This library handles call set-up and control of device during an ongoing Call. The call may be either outgoing or incoming.
- *Network Text:* Using this library, SMS text messages can be integrated with the WML, WMLScript functions.
- *Phonebook:* Using this library, the phonebook entries in the device can be manipulated.
- *Call Logs:* Using this library, call logs in the device can be accessed.

8.2.7 WAP Push Architecture

The WAP Push framework allows information to be sent to a client device without a previous user action. In a normal client/server model, a client requests for a service or information from a server. The server then responds to this request by transmitting information back to the client. This is referred to as pull technology (Fig. 8.4), where the client pulls information from the server. In addition to this type of synchronized request response transaction, WAP offers push technology (Fig. 8.5). Push is also based on the client/server model, but there is no explicit request from the client before the server transmits its content. This can be termed as unsolicited response. In other words, "pull" transactions are always initiated from the client, whereas, "push" transactions are server-initiated. Push technology is helpful to implement alerts and notification.

8.2.8 The Push Framework

The push content generally is originated in a server in the Internet that needs to be delivered to a mobile phone. The Push Initiator contacts the Push Proxy Gateway (PPG) from the Internet side, delivering content for the destination client (Fig. 8.6). The PPG then forwards the content to the mobile network to be delivered to the destination client over-the-air. In addition to providing simple proxy gateway services, the PPG is capable of notifying the Push Initiator about the final outcome of the push operation. It may even wait for the client to accept or reject the content in two-way mobile networks (MMS uses this function).

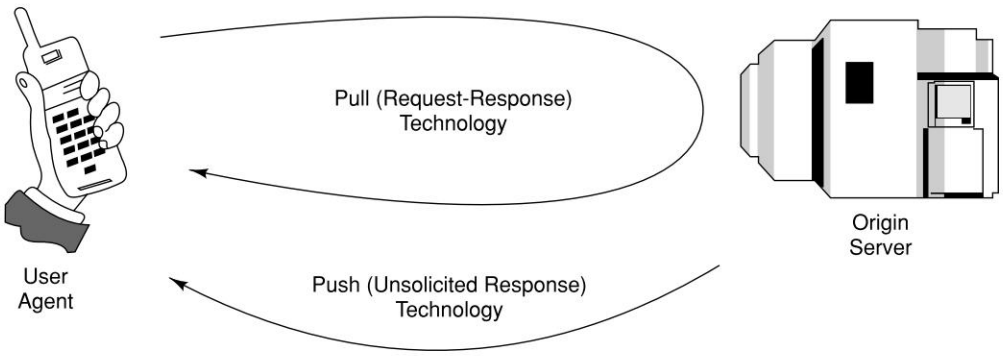


Figure 8.4 Pull versus Push Technology

The Internet-side PPG access protocol is called the *Push Access Protocol*. The WAP-side (OTA) protocol is called the *Push Over-The-Air Protocol*.

8.2.9 Wireless Session Protocol (WSP)

The Wireless Session Protocol (WSP) provides a consistent interface between two session services (client and server). It provides the cooperating client/server applications to:

(a) Establish a reliable session from client to server and close it in an orderly manner.
(b) Agree on a common level of protocol functionality using capability negotiation.

(c) Exchange content between client and server using compact encoding.
(d) Suspend and resume the session.

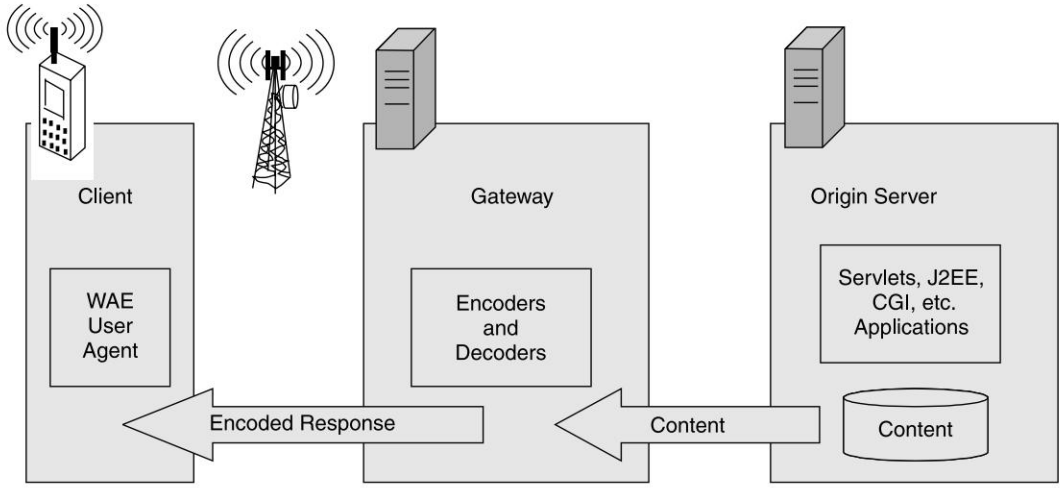


Figure 8.5 WAE Push-based Model

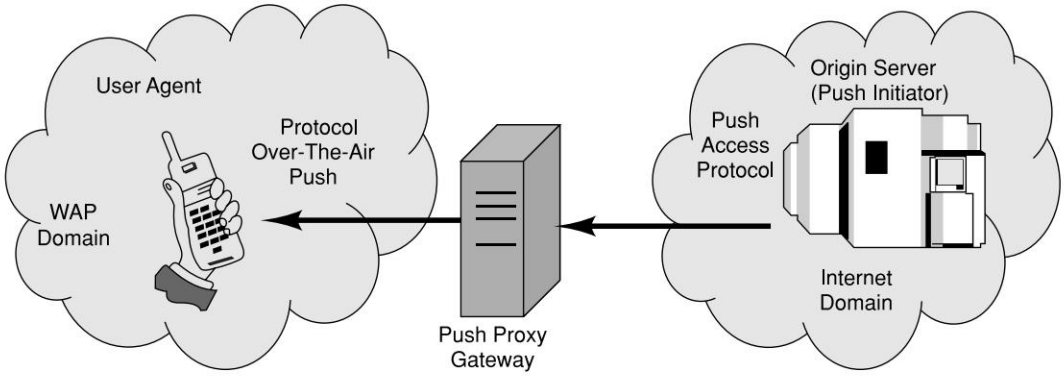


Figure 8.6 Push Framework with PPG (Push Proxy Gateway)

Currently the scope of WSP is suited mostly for browsing applications. It offers both connection-oriented and connectionless service. The connectionless service is most suitable, when applications do not need reliable delivery of data and do not care about confirmation.

The connection-oriented session services are divided into the following categories:

- Session Management facility.
- Method Invocation facility.
- Exception Reporting facility.
- Push facility.

- Confirmed Push facility.
- Session Resume facility.

WSP is designed to function on the transaction and datagram services between WAE and the WTP. WSP itself does not require a security layer; however, applications that use WSP may require it. The transaction, session or application management entities are assumed to provide the additional support that is required to establish security contexts and secure connections.

8.2.10 Wireless Transaction Protocol (WTP)

The Wireless Transaction Protocol (WTP) runs on top of a datagram service and provides a lightweight transaction-oriented protocol that is suitable for implementation in "thin" clients. WTP allows for interactive browsing (request/response) applications and supports three transaction classes: unreliable with no result message, reliable with no result message, and reliable with one reliable result message. WTP provides the following features:

- Three classes of transaction service are:
  - ❏ Unreliable one-way requests.
  - ❏ Reliable one-way requests.
  - ❏ Reliable two-way request-reply transactions.
- Optional user-to-user reliability: WTP user triggers the confirmation of each received message;
- Optional out-of-band data on acknowledgements;
- PDU concatenation and delayed acknowledgement to reduce the number of messages sent; and
- Asynchronous transactions.

8.2.11 Wireless Transport Layer Security (WTLS)

WTLS is a security protocol based upon the Transport Layer Security (TLS) protocol. WTLS and TLS are derived from the Secure Sockets Layer (SSL) protocol. WTLS is intended for use with the WAP transport protocols and has been optimized for use over narrow-band communication channels. WTLS provides the following features:

- *Data Integrity:* WTLS contains facilities to ensure that data sent between the terminal and an application server is unchanged and uncorrupted.
- *Privacy:* WTLS contains facilities to ensure that data transmitted between the terminal and an application server is private and cannot be seen by any intermediate parties that may have intercepted the data stream.
- *Authentication:* WTLS contains facilities to establish the authenticity of the terminal and application server.
- *Denial-of-service Protection:* WTLS contains facilities for detecting and rejecting data that is replayed or not successfully verified. WTLS makes many typical denial-of-service attacks harder to accomplish and protects the upper protocol layers.

8.2.12 Wireless Data Protocol (WDP)

The Transport layer protocol in the WAP architecture is referred to as the Wireless Datagram Protocol. The WDP layer operates above the data capable bearer services supported by the various

network type general transport service. WDP offers a consistent service to the upper layer protocols of WAP and communicates transparently over one of the available bearer services. While WDP uses IP as the routing protocol, unlike the Web, WAP does not use TCP. Instead, it uses UDP (User Datagram Protocol), which does not require messages to be split into multiple packets, and sent out only to be reassembled on the client. Due to the nature of wireless communications, the mobile application must be talking directly to a WAP gateway, which greatly reduces the overhead required by TCP.

Since the WDP protocols provide a common interface to the upper layer protocols the security, session, and application layers are able to function independently of the underlying wireless network. This is accomplished adapting the transport layer to specific features of the underlying bearer. By keeping the transport layer and the basic features consistent, global interoperability can be achieved using mediating gateways.

8.2.13 WAP Gateway

WAP gateway acts as a middleware which performs coding and encoding between the cellular device and web server. The WAP gateway can be located either in a telecom network or within a computer data network (an ISP). A user from a WAP device requests for a WAP page using a URL, the gateway establishes a connection to the target WAP site. It collects the document from the site. Then the WAP page is "compiled" and converted into binary code. Binary code takes far less space compared to the WML source. This realizes quicker delivery. The code is then sent across to the phone or the wireless device over the air. When the phone receives the stream of octets, it "de-compiles" it. The client browser does the reverse operation of compilation by decompiling the binary code. This will allow the client to regenerate the normal WML page and then displays it on the device. We talked about Kannel SMS gateway in Chapter 6. Kannel also has a free opensource WAP gateway. The WAP phones all have a maximum allowed size for a compiled WAP page. Basic functions of a WAP gateway (Fig. 8.7) are:

- Implementing WAP protocol stack.
- Protocol translation between phone and server.
- Compress WML pages to save bandwidth.
- User authentication and billing.

The WAP protocols are designed to operate over a variety of different bearer services, including short message, circuit-switched data, and packet data. The bearers offer different levels of quality of service with respect to throughput, error rate, and delays. The WAP protocols are designed in such a fashion that it can compensate for or tolerate this varying level of QoS (quality of service). External Interfaces of a gateway are:

- SMS center, using various protocols.
- HTTP servers, to fetch WML pages.
- WAP devices using WAP protocol stack.

Many WAP gateways include additional functions. These relate to user authentication and charging. For charging, it captures the usage data. The gateway does not actually include a billing system itself but it provides the user and the service provider the usage data. The usage data is given to the billing system of the operator. From the user's point of view, the gateway is also

responsible for optimizing WAP usage as far as possible. The gateway keeps the number of packets small to keep costs down and make the best use of available bandwidth.

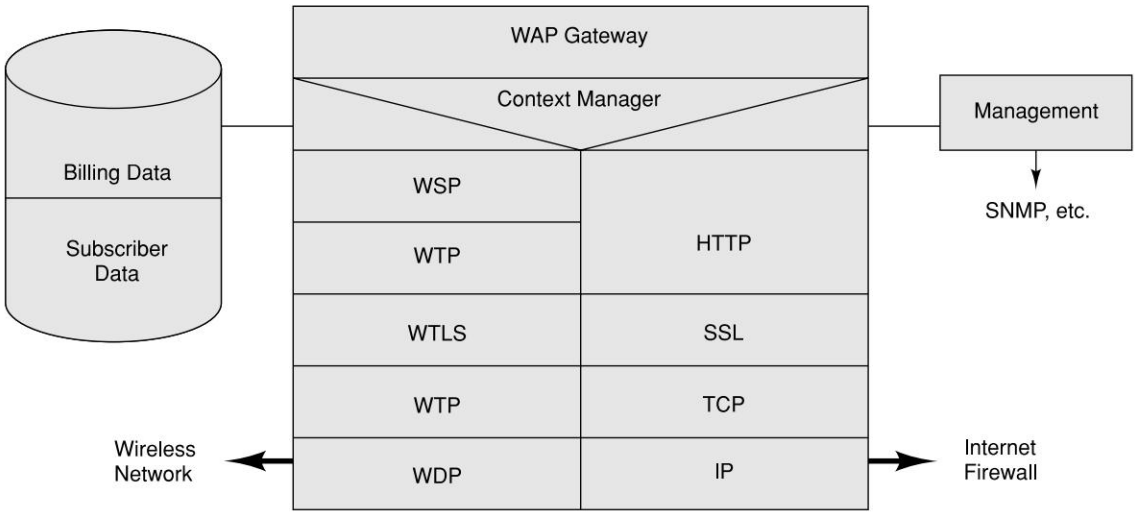


Figure 8.7 Architecture of a WAP Gateway

Generally Internet is not self-configurable. The same is true with WAP. This means that when we move from one network to another, there may be a need to configure the client device to suit the network parameters of the serving network. This also may depend on the type of the network. The configuration of WAP will require an IP address of the WAP gateway. Though the WAP gateway can be from home network, due to security and charging reasons, service providers do not allow the usage of external WAP gateways. Therefore, at a minimum, two parameters need to be changed. These are the telephone number for the WAP dial-up connection and the IP address of the WAP gateway. The rest can be configured once only. There could also be some dependence of the settings on other security parameters.

8.3 MMS

In the GSM world, the popularity of SMS (Short Messaging Service) surprised everybody. SMS was launched in 1992 and has become the most successful wireless data service to date. SMS was originally designed to carry text message. Vendors started thinking about adding more life to the text message. The result was enhanced SMS (EMS). EMS offered a combination of text and simple pixel-image (pictures) and melody (ringing tone). The technology for SMS and EMS is already discussed in Chapter 6. We can refer to SMS as the first generation of messaging, whereas EMS can be considered as second generation messaging. SMS was person-to-person, whereas EMS was content-to-person. The popularity of the first and the second generation messaging made device vendors think of next generation messaging. In the third generation of messaging, the

message content would be multimedia objects. This is called Multimedia Messaging Service or MMS in short. An MMS message can contain formatted text, graphics, data, animations, images, audio clips, voice transmissions and video sequences.

Though MMS is targeted for the 3G networks, it can work under a 2G or 2.5G network as well. All it needs is a MMS handset and the MMS infrastructure. There are two standards bodies producing specifications relating to MMS messages. These are WAP Forum and the 3GPP (third Generation Partnership Project). The standards produced by these two bodies in turn use existing specifications from two Internet standards bodies: the W3C (World Wide Web Consortium) and the IETF (Internet Engineering Task Force). The standards from the WAP Forum specify how messages are composed and packaged whereas the standards from the 3GPP specify how messages are sent, routed, and received.

Figure 8.8 shows a generalized view of the Multimedia Message Service architecture for a 3G messaging system. It combines different networks and network types. It integrates existing messaging systems within these networks. A user takes a picture using his mobile phone and sends it to another person using the MMS functionality (person-to-person). The picture can be sent directly as a MMS message or can be sent as an attachment to an email message. MMS messages can also be automatically generated and sent through software (content-to-person). For example, a user could ask for the day's weather forecast to be sent to her phone each morning complete with animated maps and audio of the weatherman.

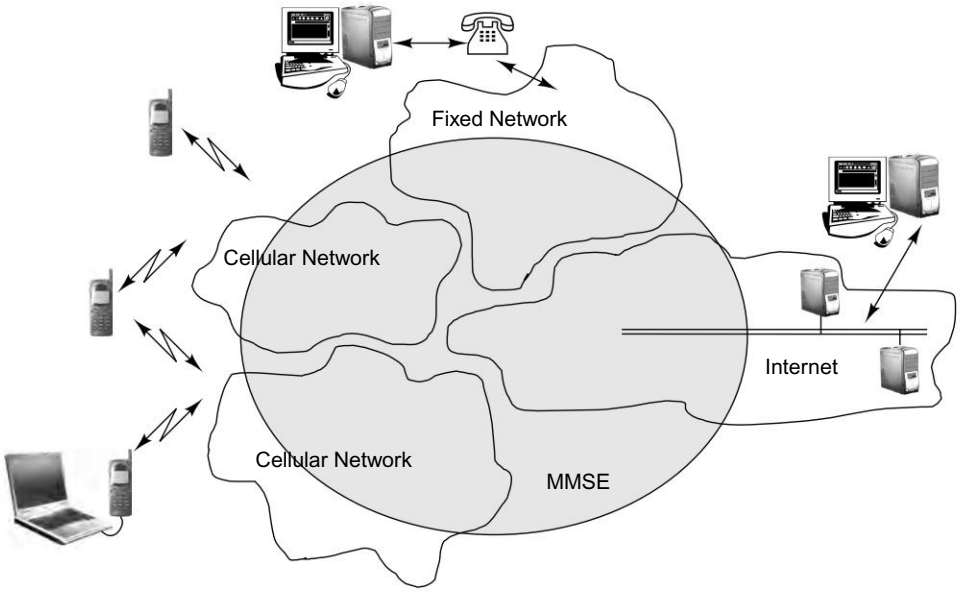


Figure 8.8 General View of MMS Provision within Different Networks

In the first phase of the MMS, users can create presentation slides through software. The layout and ordering of the slides are specified through a language called SMIL (Synchronization

Multimedia Integration Language). This may be a slide show with multiple or even a single slide. The slide show may combine a number of still pictures or animations into one MMS message. The display areas of these slides are divided into different sections. Currently, there are only two sections per slide—one for an image and one for the text. It is also acceptable to have either just an image or a text region. In the second phase, users can record their own video content (10 second video clip) and send it via MMS. The contents of the slides—the actual images, text, and audio—are separate pieces that are sent along with the slides. The maximum size of the entire packaged message that first generation devices could support was 50 kB.

8.3.1 MMS Architecture

The connection between different networks in Figure 8.8 is provided by the Internet protocol and its associated set of messaging protocols. This approach enables messaging in wireless networks to be compatible with messaging systems found on the Internet. Multimedia Message Service Environment (MMSE) encompasses various elements required to deliver a MMS (Fig. 8.9). This includes:

- *MMS Client:* This is the entity that interacts with the user. It is an application on the user's wireless device.
- *MMS Relay:* This is the system element that the MMS client interacts with. It provides access to the components that provide message storage services. It is responsible for messaging activities with other available messaging systems. The SMS relay along with the MMS content server is referred to as MMSC (MMS Controller).
- *WAP Gateway:* It provides standard WAP services needed to implement MMS.
- *MMS Server:* This is the content server, where the MMS content is generated
- *Email Server:* MMS can integrate seamlessly to the email system of Internet.

The messages that transit between the MMS Client and MMS Relay pass through WAP Gateway. Data is transferred between the MMS client and WAP gateway using WAP Session Protocol (WSP). Data is transferred between the WAP gateway and the MMS Relay using HTTP.

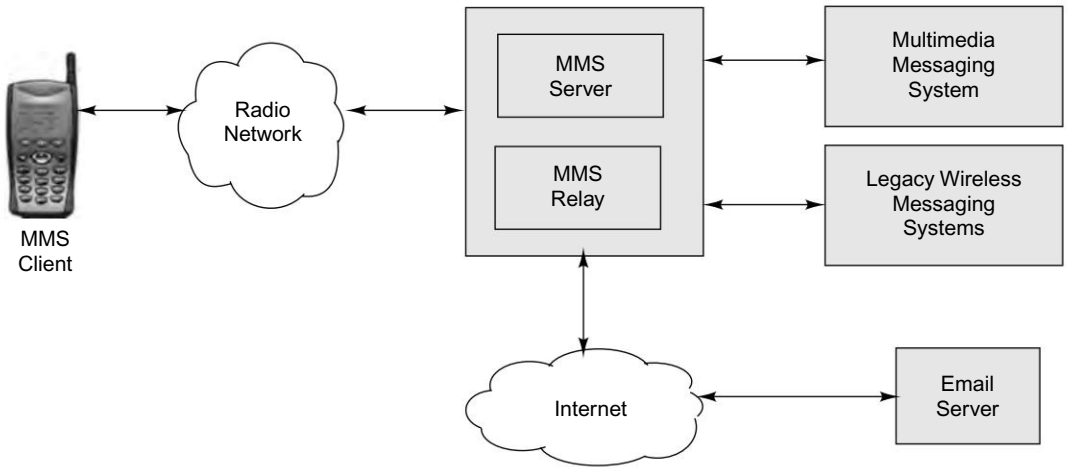
8.3.2 MMS Transaction Flows

As mentioned earlier, the MMS service is realized by the invocation of transactions between the MMS Client and the MMS Relay. The general transactions of sending and retrieving messages do not depend on what type of client the message is sent to or received from. The other endpoint for the message may be another MMS Client or a client on a legacy wireless messaging system or it may even be an email server.

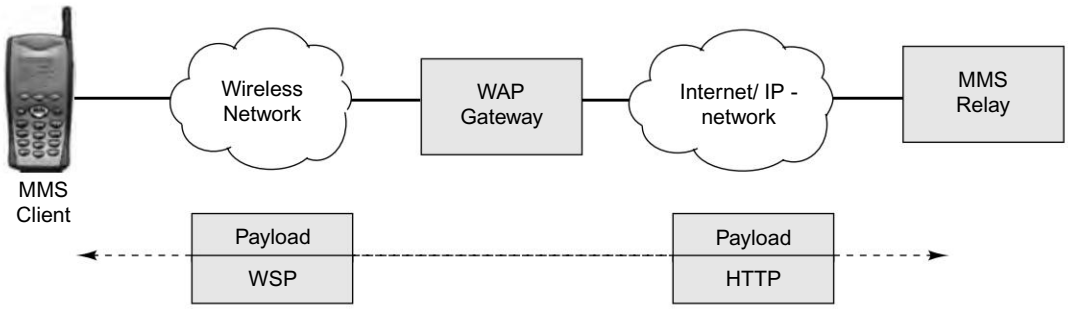
The above message exchanges can be considered to form the following logically separate transactions (Fig. 8.10):

- MMS Client (sender) sends a message to MMS Relay (M-Send.req, M-Send.conf).
- MMS Relay notifies MMS Client (recipient) about a new message arrival (M-Notification.ind, M-NotifyResp.ind).
- MMS Client fetches (recipient) a message from MMS Relay (WSP GET.req, M-Retrieve.conf).

- MMS Client (recipient) sends a retrieval acknowledgement to MMS Relay (M-Acknowledge.req).
- MMS Relay sends a delivery report about a sent message to MMS Client (sender) (M-Delivery.ind).



(a) MMS networks



(b) Client to MMS Relay Link

Figure 8.9 MMS Environment

From this list it is clear that MMS uses eight types of messages to perform messaging transactions. The M-Notification.ind function is generally done through SMS. This is a special type of SMS. This SMS is not forwarded to the SMS inbox; it is forwarded to the MMS client. Client notifies the user about the arrival of a MMS message. The user then fetches the message from the relay. Some terminals allow the facility to configure the MMS client so that the message is fetched automatically.

The multimedia messaging PDUs (Protocol Data Units) consist of MMS headers and a message body. The message body may contain any content type such as text, image, audio, and video. The message body is used only when the multimedia message is sent or retrieved. All other PDUs contain only the MMS-headers part.

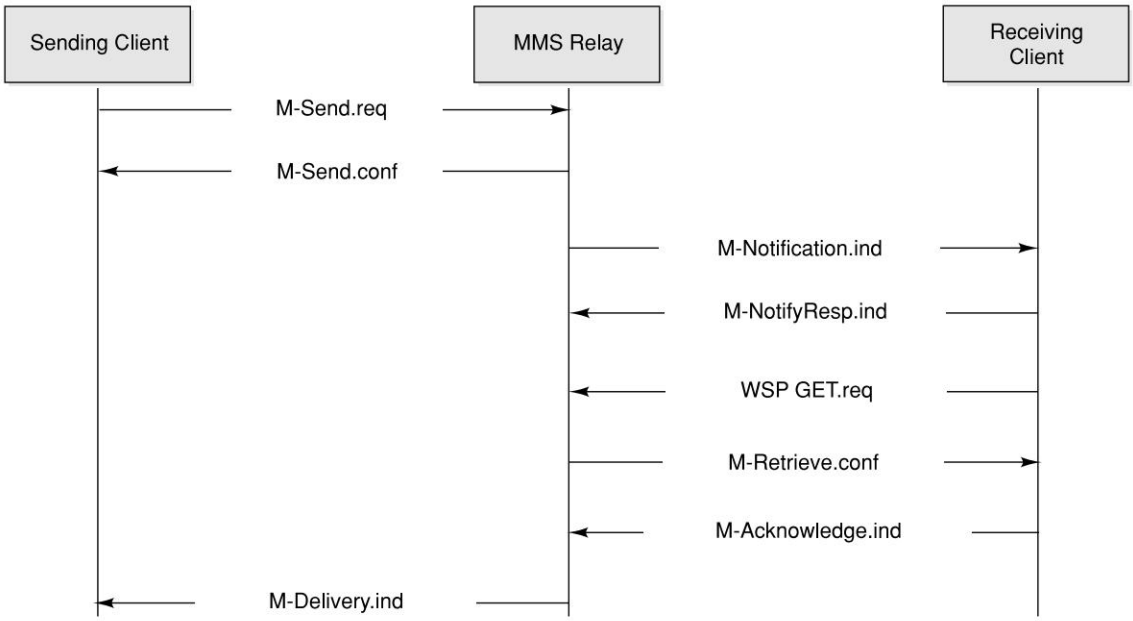


Figure 8.10 Example of MMS Transaction Flow-Delayed Retrieval

8.3.3 SMIL (Synchronized Multimedia Integration Language)

Synchronized Multimedia Integration Language (SMIL) is an XML-based language specified by the W3C. It is a markup language for specifying how and when multimedia clips will play. SMIL integrates streaming audio and video with images, text or any other media type. MMS messages are sent using SMIL as the presentation language. The presentation part specifies how the various other parts of the message should be presented to the user—at what time and in which place in relation to the other parts. MMS adapted a limited subset of SMIL, often referred to as "MMS SMIL". Links can be to various places inside the current presentation to allow the user to jump around from place to place within the timeline of the presentation. They can also point to a website, for example, letting the user get more information, downloads, etc., using the terminal's browser.

Example

Following is a simple example of the SMIL for an MMS message.

```
<smil>
   <head>
```

```
        <meta name="title" content="vacation photos" />
        <meta name="author" content="Radha Krishna" />
        <layout>
           <root-layout width="160" height="120"/>
           <region id="Image" width="100%"
           height="80" left="0" top="0" />
           <region id="Text" width="100%"
           height="40" left="0" top="80" />
        </layout>
     </head>
     <body>
        <par dur="8s">
           <img src="FirstImage.jpg" region="Image" />
           <text src="FirstText.txt" region="Text" />
           <audio src="FirstSound.amr"/>
        </par>
        <par dur="7s">
           <img src="SecondImage.jpg" region="Image" />
           <text src="SecondText.txt" region="Text" />
           <audio src="SecondSound.amr" />
        </par>
     </body>
  </smil>
```

The example above is for a terminal whose screen will displays the slide in a portrait orientation, where the height is greater than the width. On a PC screen, the SMIL slides are all displayed and are exactly 160 pixels wide and 120 pixels tall. The total slide area is divided into two smaller areas. The image region will be 80 pixels tall and always appears above the 40 pixel tall text area. On an MMS client, however, this will be different. The screen may not be large enough to accommodate the layout. Each slide in turn contains at least two elements: one for the image region and one for the text region. Two of the slides also contain an audio element that will be played when the slide is viewed. In normal SMIL, the names of the layout regions (image and text in our MMS message) are just handy names for generic regions that can contain any type of content. In MMS SMIL, however, the image region must contain an image element and the text region, a text element.

As we can see, the SMIL markup is very similar to HTML or WML markup language. The entire message body is enclosed within `<smil></smil>` tags and the message (or document) itself has both head and body sections. The head section contains information that applies to the entire message. The title and author meta fields here correspond to the From and Subject fields of the message. These meta fields are optional. Under MMS implementations of SMIL, a client is free to re-format the layout in a way best suited to the client's display. Actual slides are within the body section of the message. These slides are denoted with the `par`–for parallel tag. Parallel denotes that all the elements within the tag are to be displayed simultaneously. The `dur` attribute for each slide is the duration of the slide in the slide show. Again, the receiving client is free to modify or ignore this, replacing duration with a button for the next slide, for example.

The following are the specific media formats that will be supported in the first generation of MMS systems.

Images: Image formats supported are baseline JPEG with JFIF exchange format, GIF87a, GIF89a, and WBMP. The maximum guaranteed image resolution is 160 pixels wide by 120 pixels high. Larger images are supported, but need to be converted for the target device. The browser safe color palette (256 colors) is recommended for color image. JPEG is better suited for rendering photographs; whereas, GIF is a better choice for line drawings.

Text: The text of the message may use us-ascii, utf-8, or utf-16 character encoding. The supported character sets on any client will always be at least all of ISO 8859-1.

Audio: Audio should be encoded as AMR (Adaptive Multi Rate), a codec used for voice in GSM and 3G networks. Many clients will also support iMelody for ring tones.

8.3.4 MMS Interconnection, Interoperability and Roaming

Like any other service, MMS also has to meet the challenges of interoperability and roaming. Interoperability of MMS means the ability of terminals to exchange mutually acceptable messages between terminals from different vendors, or network components like MMSCs, and with WAP gateways. This includes the end-to-end exchange of formats and protocols. MMS roaming means that a subscriber can send and receive MMS messages when roaming in another network.

The main method for GPRS roaming is PLMN roaming where the home PLMN GGSN is used. The other method is ISP roaming where the visited PLMN GGSN is used. When the user is roaming, MMS messages are sent via normal packet data traffic between the home and roaming operator network. In addition to this, the roaming customer must be able to receive SMS from the home SMSC. To achieve MMS roaming, both GPRS and SMS roamings are required. Participating operators need to have a packet data roaming agreement and SMS roaming agreement in place. A roaming agreement means the technical and commercial agreement between operators on interoperability and charging. Charging includes functions like exchange of charging data, billing the subscriber, and sharing the revenue. Operators must solve the problem of handling the interconnection charge, first within one country and then globally. They both collect statistics from traffic volumes, with clearing based on statistics and agreements. In practice, there are three ways for operators to arrange MMS interconnection: using GRX (GPRS Roaming Exchange), VPN over Internet, or VPN over leased lines. Figure 8.11 depicts MMS sender roaming, whereas Figure 8.12 shows MMS receiver roaming.

OMA and 3GPP have defined three domains for multimedia messages. The first of these is the Core MM Content Domain where full interoperability is guaranteed. The second is the Standard MM Content Domain, where terminals and multimedia messages are still compliant with MMS standards but terminals have certain freedoms. The third domain is the unclassified MM Content Domain, giving full freedom to create multimedia messages.

8.3.5 MMS Device Management and Configuration

MMS services sometime require complex configuration. For example, the settings required for MMS include MMSC IP address, connection type and about 10 other parameters. Therefore,

there is a need to be able to configure the users' devices, by providing device settings over the air. The OMA device management architecture consists of two components: OMA Client Provisioning and a continuous management technology that is based on the SyncML Device Management specification. Client Provisioning is a messaging based provisioning technology that sends settings over the air to the device and configure. All the user has to do is to accept the sent settings and the device will be correctly configured and ready for use.

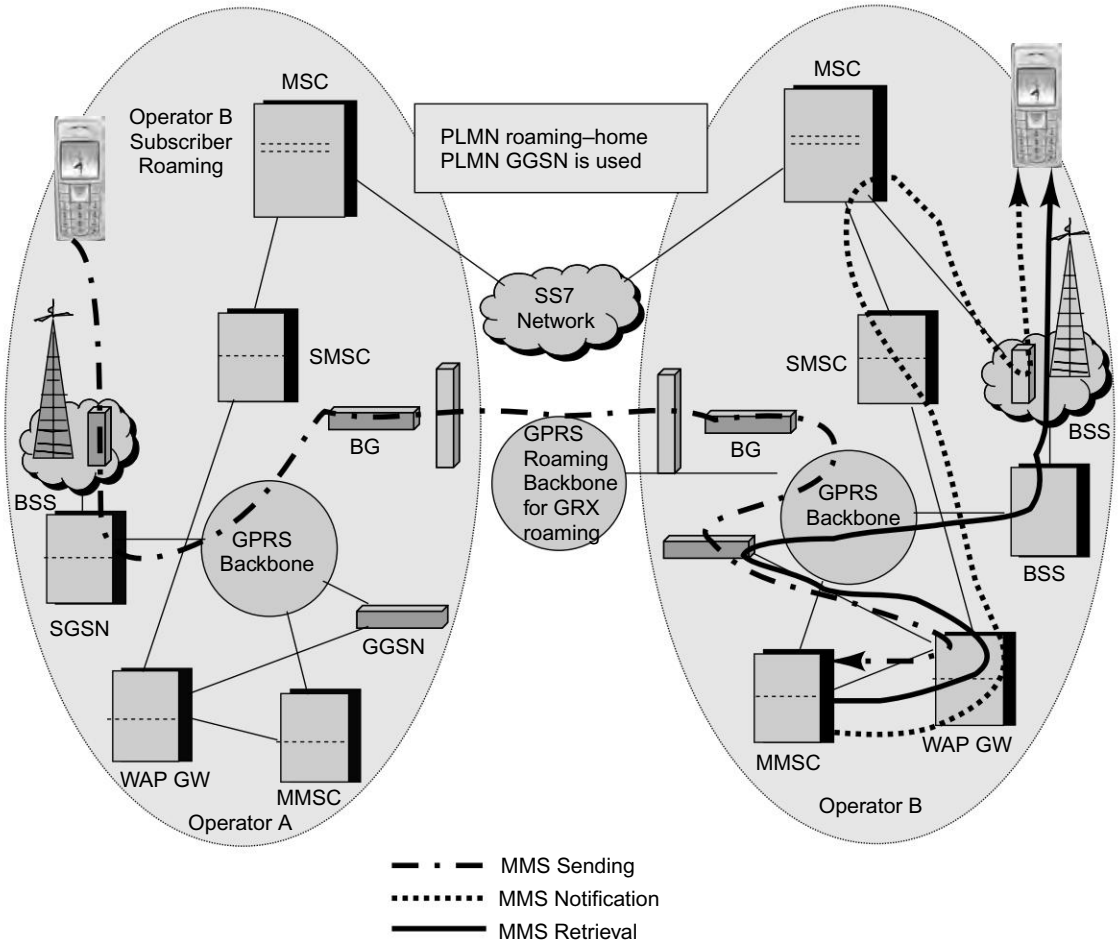


Figure 8.11 MMS Sender Roaming

8.4 GPRS APPLICATIONS

For GPRS or WAP there are no specific services. These are the same services over the wireless media that can run either on GSM or 3G. However, as GPRS offers a higher bit rate, the user

experience is better. Though, it is not mandatory, MMS based contents are better suited for GPRS networks. Some of the applications that are better suited for GPRS network are:

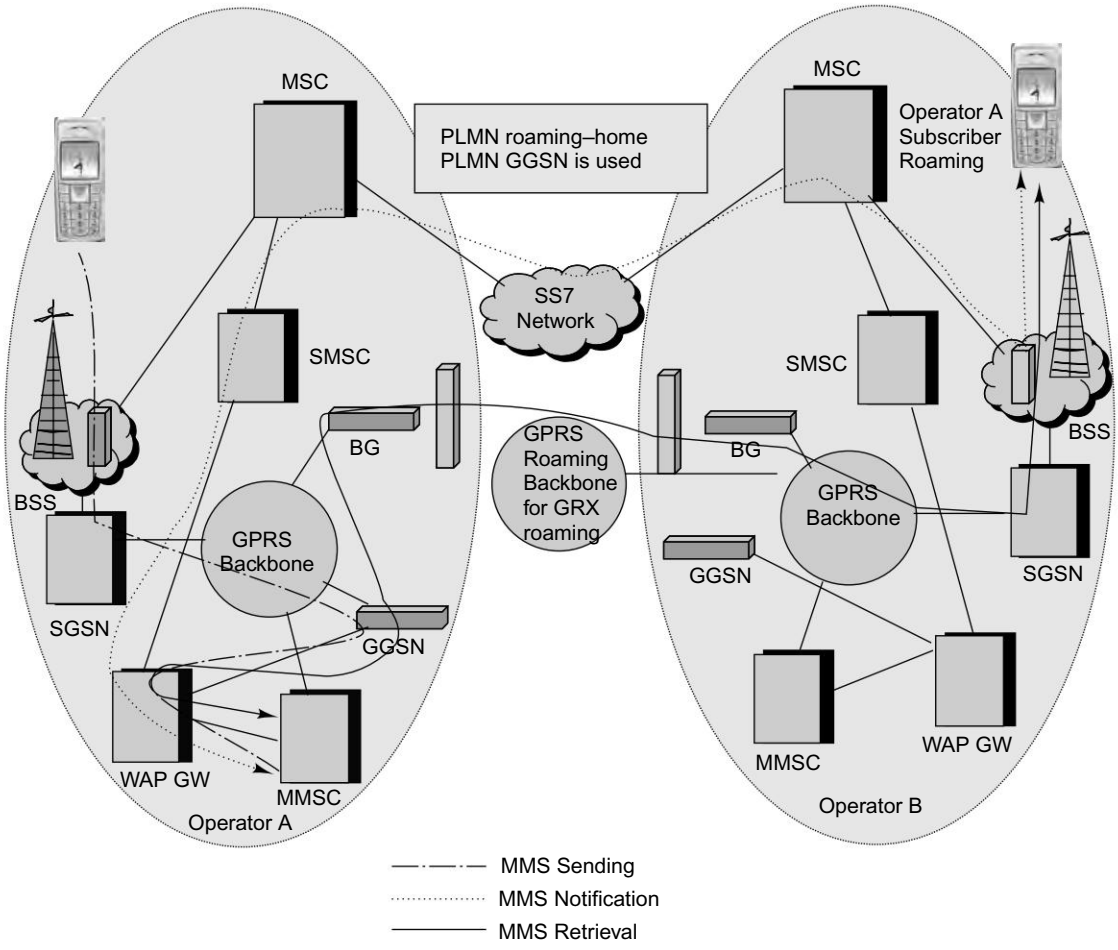


Figure 8.12 MMS Receiver Roaming

- *Dating:* This is an interactive dating system that exploits data bearer rather than the SMS bearer.
- *Games:* Online games, cartoons can be better suited for GPRS network.
- *Sending/Receiving Fax/Email on the Handset:* Different mobile office applications are quite well suited for GPRS network.
- *Location-aware Applications:* Location-aware applications combined with MAP and other type of travel related applications will have a better reach in GPRS network.
- *VPN:* It is possible to deploy wireless virtual private networks over GPRS network.
- *Multimedia Downloads:* Various picture, music, video clips downloads are part of this.

8.4.1 Digital Rights Management

MMS opens up lots of new possibilities for content providers to offer innovative premium mobile services, such as comic services, news services, sports updates, movie and music clips. As these services take off, it is crucial to take into account Digital Rights Management (DRM) for the distribution and consumption of mobile content. With DRM, the content owner or service provider can determine if and how his content can be distributed if at all from person to person. Of course, some content will be available even without DRM. All the content created by mobile users themselves such as capturing and sending of photo and video clips, or advertising content like sponsored movie trailers that promote new films in the cinema, will be free. Companies involved in mobile content services such as ring tones, wallpapers, or games where the business plan is based on being able to collect the rightful payment, will be interested in implementing DRM technology.

8.4.2 OMA Digital Rights Management

OMA has proposed Digital Rights Management and OMA Download standards. By implementing OMA DRM, service providers can allow end users to preview content before making a purchase decision. OMA DRM also allows end users to distribute content to other users via superdistribution. The OMA DRM standard governs the use of mobile-centric content types. The first DRM standard, OMA DRM version 1.0, was officially approved in October 2002. The standard provides three DRM methods as following:

- Forward-lock is intended for the delivery of subscription-based services. The device is allowed to play, display or execute the MMS, but it cannot forward the MMS object. The content itself is hidden inside the DRM message that is delivered to the terminal. Examples will be news, sports, information and images that should not be sent on to others.
- Combined Delivery enables usage rules to be set for the media object. This method extends Forward-lock by adding a rights object to the DRM Message. Rights define how the device is allowed to render the content and can be limited using both time and count constraints. This method allows previews.
- Separate Delivery protects higher value media and enables superdistribution. This allows the device to forward the media, but not the rights. This is achieved by delivering the media and rights via separate channels, which is more secure than combined delivery. The media is encrypted into DRM Content Format (DCF) using symmetric encryption, while the rights hold the Content Encryption Key (CEK), which is used by the DRM User Agent in the device for decryption. Recipients of superdistributed content must contact the content retailer to obtain rights to either preview or purchase media.

REFERENCES/FURTHER READING

1. Arehart, Charles et al (2000), *Professional WAP*, Wrox Press.
2. Multimedia Messaging Service: Service Aspects; Stage 1, 3GPP 3G TS 22.140 Release 1999.
3. Multimedia Messaging Service: Functional Description; Stage 2, 3GPP 3G TS 23.140 Release 1999.

4. MMS Technology Tutorial: http://www.nokia.com/support/tutorials/MMS/en/.
5. MMS: http://www.gsmworld.com/technology/mms/index.shtml.
6. Nokia white paper on "MMS Entering into the Next Phase".
7. Synchronized Multimedia Integration Language: http://www.w3.org/TR/REC-smil/.
8. WAP MMS Architecture Overview, WAP-205-MMSArchOverview.
9. WAP MMS Client Transactions, WAP-206-MMSCTR.
10. WAP MMS Encapsulation Protocol, WAP-209-MMSEncapsulation.
11. Wireless Application Protocol Architecture Specification, WAP Forum, April 1998.
12. Wireless Application Protocol Wireless Markup Language Specification, Version 1.2, WAP Forum, November 1999.
13. Wireless Application Protocol WMLScript Language Specification Version 1.1, WAP Forum, November 1999.
14. Wireless Application Protocol Wireless Application Environment Specification Version 1.2, WAP Forum, November 1999.
15. Wireless Application Protocol Wireless Application Environment Overview, WAP Forum, November 1999.
16. Wireless Application Group User Agent Profile Specification, WAP Forum, November 1999.
17. Wireless Application Protocol Push OTA Protocol Specification, WAP Forum, November 1999.
18. Wireless Application Protocol Push Architectural Overview, WAP Forum, November 1999.
19. Wireless Transport Layer Security Version 06-Apr-2001 Wireless Application Protocol WAP-261-WTLS-20010406-a, WAP Forum.
20. Wireless Application Protocol Wireless Telephony Application Interface Specification, WAP Forum, November 1999.
21. Wireless Application Protocol WAP 2.0, Technical White Paper, WAP Forum, January 2002.

REVIEW QUESTIONS

Q1: Describe the WAP protocol stack while enumerating the functions of different layers.

Q2: Describe the WAP Application Environment.

Q3: What is WTAI (Wireless Telephony Application Interface)? Why is it important to have such a function? Describe an application where WTAI can make the user experience better.

Q4: Describe the following with respect to WAP:

(a) WML
(b) UAProf
(c) WML Scipt
(d) WSP
(e) WTP
(f) WTLS
(g) WDP

Q5: What is WAP Push? How is push different from pull?
Q6: What is a WAP Gateway? What are its functions?
Q7: What is MMS? Describe the MMS architecture.
Q8: How is MMS different from Short Message Service and Extended Message Service?
Q9: What is SMIL? How is SMIL used in MMS?
Q10: What is digital rights management? Discuss the OMA proposal on digital rights management.

CHAPTER 9

CDMA and 3G

9.1 INTRODUCTION

The popularity and growth of cellular phones is keeping technology and business people on their toes. Technologists are busy developing even newer technologies to offer better user experience. Operators and service providers, on the other hand, are coming up with innovative applications and services to get a share of this market. Users today expect better quality of voice and data services while on the move. Not too long ago, hardly any one would have imagined mobile phone being used not only for voice communication, but also for watching a video clip or as a network interface for a laptop. CDMA and 3G expressly support such versatile usage.

Many of these opportunities and challenges made the scientific and business community look at the Spread-Spectrum technology as an option for wireless communication. Mobile phone technology had a reincarnation from first generation analogue (using FDMA) to second generation digital (using TDMA). The next incarnation is from second generation digital TDMA to third generation packet (using CDMA). CDMA is a specific modulation technique of Spread-Spectrum technology. Third generation or 3G is more of a generic term to mean mobile networks with high bandwidth. Looking at the success of second generation GSM (using TDMA and roaming) and also the potential of second generation cdmaOne (IS-95 using CDMA), it was quite apparent that the next generation networks would have to be a combination of the best of these two technologies with amalgamation of some of the recent technology innovations.

9.1.1 How it Started

Let us tell you an interesting story about the origin of the Spread-Spectrum technology. Have you heard the name of the famous Hollywood actress Hedy Lamarr? She was born in Vienna in 1914 as Hedwig Eva Maria Kiesler (note the full name). In 1933, Hedy Kiesler married the Austrian industrialist Fritz Mandl, CEO of the Hirtenberger Patronenfabrik, then one of the world's leading arms producers. Fritz was interested in control systems and conducted research in that field. Hedy

was so beautiful that he was obsessed with keeping Hedy at his side all the time. He would take her even to business meetings and parties. Hedy received an education in munitions manufacturing from her husband and other Nazi officials through these meetings. Hedy escaped to London in 1937 and later traveled to the US to become one of the better known actresses in Hollywood. What many people may not know is that Hedy Lamarr helped the Allies win World War II, and she was the original patent holder of Spread-Spectrum technology, which is at the foundation of today's CDMA (Code Division Multiple Access), Wireless LAN, IMT-2000 (International Mobile Telecommunication–2000), 3G (Third Generation), and GPS (Global Positioning System) technology.

With the help of an electrical engineering professor from MIT, Hedy Lamarr and George Antheil, a film music composer patented "Secret Communication System" in 1942. Like many other great technologies, the idea of "Secret Communication System" was ahead of its time. Electronic technologies were beginning to develop and in the 1950s, engineers from Sylvania Electronic Systems Division began to experiment with the ideas in the Secret Communication System patent, using digital components. They developed an electronic spread-spectrum system that handled secure communications for the US during the Cuban Missile Crisis in 1962. It was in the early 1960s that the term "spread-spectrum" began to be used. Today it refers to digital communications that use a wide frequency spreading factor (much wider than typical voice telephone communications), and are not dependent on a particular type of tonality (such as a human voice) in the transmitting waveform.

In the mid-1980s, the US military declassified Spread-Spectrum technology. Immediately, the commercial sector began to develop it for consumer electronics. Qualcomm was the first to use this technology for commercial deployment of CDMA. 3G has been in gestation since 1992, when the International Telecommunications Union (ITU) began work on a standard called IMT-2000. IMT stands for International Mobile Telecommunications; the number 2000 initially had three meanings: the year that services should become available (year 2000), the frequency range in MHz that would be used (2000 MHz or 2 GHz), and the data rate in Kbits/sec (2000 Kbps or 2 Mbps).

9.2 Spread-Spectrum Technology

In a conventional transmission system, the information is modulated with a carrier signal and then transmitted through a medium. When transmitted, all the power of the signal is transmitted centered around a particular frequency. This frequency represents a specific channel and generally has a very narrow band. In spread-spectrum we spread the transmission power over the complete band as shown in Figure 9.1.

In spread-spectrum the transmission signal bandwidth is much higher than the information bandwidth. There are numerous ways to cause a carrier to spread; however, all spread-spectrum systems can be viewed as two steps modulation processes. First, the data to be transmitted is modulated. Second, the carrier is modulated by the spreading code, causing it to spread out over a large bandwidth. Different spreading techniques are:

- **Direct Sequence (DS):** DS spread spectrum is typically used to transmit digital information. A common practice in DS systems is to mix the digital information stream with a pseudo random code.

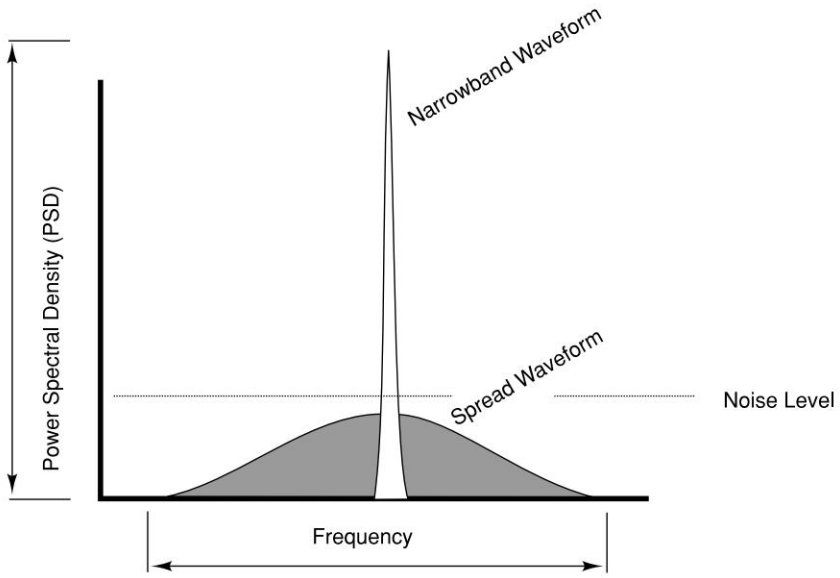


Figure 9.1 Narrow Band and Spread Spectrum

- **Frequency Hopping (FH):** Frequency hopping is a form of spreading in which the center frequency of a conventional carrier is altered many times within a fixed time period (like one second) in accordance with a pseudo-random list of channels.
- **Chirp:** The third spreading method employs a carrier that is swept over a range of frequencies. This method is called chirp spread spectrum and finds its primary application in ranging and radar systems.
- **Time Hopping:** The last spreading method is called time hopping. In a timehopped signal, the carrier is on-off keyed by the pseudo-noise (PN) sequence resulting in a very low duty cycle. The speed of keying determines the amount of signal spreading.
- **Hybrid System:** A hybrid system combines the best points of two or more spread-spectrum systems. The performance of a hybrid system is usually better than can be obtained with a single spread-spectrum technique for the same cost. The most common hybrids combine both frequency-hopping and direct-sequence techniques.

Amateurs and business community are currently authorized to use only two spreading techniques. These are frequency hopping and direct sequence techniques. Rest of the Spread-Spectrum technologies are classified and used by military and space sciences.

9.2.1 Direct Sequence Spread Spectrum (DSSS)

Direct Sequence Spread Spectrum (DSSS) is often compared to a party, where many pairs are conversing, each in a different language. Each pair understands only one language and therefore, concentrates on his or her own conversation, ignoring the rest. A Hindi-speaking couple just homes

on to Hindi, rejecting everything else as noise. Its analogous to DSSS is when pairs spread over the room conversing simultaneously, each pair in a different language. The key to DSSS is to be able to extract the desired signal while rejecting everything else as random noise. The analogy may not be exact, because a roomful of people all talking at once soon becomes very loud. In general, Spread-Spectrum communications is distinguished by three key elements:

1. The signal occupies a bandwidth much larger than what is necessary to send the information.
2. The bandwidth is spread by means of a code, which is independent of the data.
3. The receiver synchronizes to the code to recover the data. The use of an independent code and synchronous reception allows multiple users to access the same frequency band at the same time.

In order to protect the signal, the code used is pseudo-random, which makes it appear random while being actually deterministic, which enables the receivers to reconstruct the code for synchronous detection. This pseudo-random code is also called pseudo-noise (PN). DSSS allows each station to transmit over the entire frequency all the time. DSSS also relaxes the assumption that colliding frames are totally garbled. Instead, it assumes that multiple signals add linearly.

DSSS is commonly called Code Division Multiple Access or CDMA in short. Each station is assigned a unique m-bit code. This code is called the CDMA chip sequence. To transmit a 1 bit, the transmitting station sends its chip sequence, whereas to send 0, it sends the complement chip sequence. Thus if station A is assigned the chip sequence 00011011, it sends bit 1 by sending 00011011 and bit 0 by sending 11100100. Using bipolar notations, we define bit 0 as +1 and bit 1 as –1. The bit 0 for station A will now become (–1 –1 –1 +1 +1 –1 +1 +1) and 1 becomes (+1, +1, +1, –1, –1, +1, –1, –1). Figure 9.2 depicts this with 6 chips/bit (011010). For manipulation of bits, we XOR (addition with modulo 2) the input bits, in bipolar notations we multiply to get the desired result:

```
0 XOR 0 = 0 ⇒ +1 × +1 = +1
1 XOR 1 = 0 ⇒ -1 × -1 = +1
1 XOR 0 = 1 ⇒ -1 × +1 = -1
0 XOR 1 = 1 ⇒ +1 × -1 = -1
```

Each station has its unique chip sequence. Let us use the symbol S to indicate the m-chip vector for station S, and $\overline{S}$ is for its negation. All chip sequences are pair-wise orthogonal, by which we mean that the normalized inner product of any two distinct chip sequences, S and T (written as $S \cdot T$) is 0. In mathematical terms,

$$S \cdot T = \frac{1}{m} \sum_{i=1}^{m} S_i \cdot T_i = 0$$

$$S \cdot S = \frac{1}{m} \sum_{i=1}^{m} S_i \cdot S_i = 1$$

This orthogonality property is very crucial for mobile communication. Note that if $S \cdot T = 0$ then $S \cdot \overline{T}$ is also 0. The normalized inner product of any chip sequence with itself is 1. This follows because each of the m terms in the inner product is 1, so the sum is m. Also note that $S \cdot \overline{S} = -1$.

When two or more stations transmit simultaneously, their bipolar signals add linearly. For example, if in one chip period three stations output +1 and one station outputs –1, the result is +2.

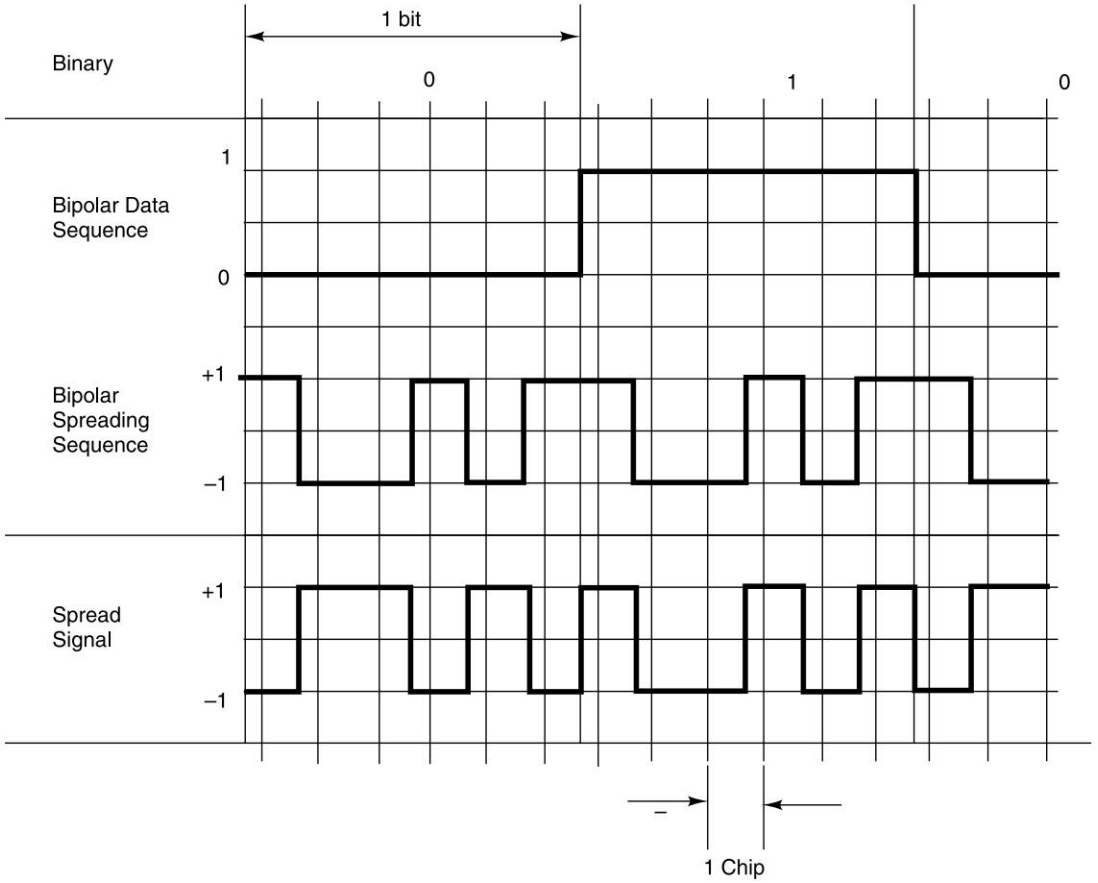


Figure 9.2 The CDMA Chip Sequence

One can think of this as adding voltages: three stations outputting +1 volt and one station outputting –1 volts gives 2 volts.

In Figure 9.3, we see there are four stations A, B, C, and D with their chip sequences. In this example, we have taken eight chips. Figure 9.3(a) is the bit sequence of the chips whereas Figure 9.3(b) is the bipolar notations of the same. In Figure 9.3(c) we assume that there are six cases of four stations transmitting at the same time. In the first example, Figure 9.3(c), we assume that only C is transmitting bit 1. In the second example, B transmits a bit 1, and C transmits a bit 1. Therefore, we get:

$(-1\ +1\ -1\ +1\ +1\ +1\ -1\ -1) = S_1$

$(-1\ -1\ +1\ -1\ +1\ +1\ +1\ -1) + (-1\ +1\ -1\ +1\ +1\ +1\ -1\ -1) = (-2\ 0\ 0\ 0\ +2\ +2\ 0\ -2) = S_2$

In the third example, station A transmits a 1 and station B transmits a 0, others are silent. In the fourth example, A and C transmit a 1 while B sends a 0. In the fifth example, all four stations transmit a 1. Finally, in the last example, A, B, and D transmit a 1, while C sends a 0. The result of these transmissions are different sequences S_1 through S_6 as given in Figure 9.3(d). All these examples represent only one bit time.

To recover the bit stream of any station, the receiver must know that station's chip sequences in advance. This is similar to the example of the party where different couples are conversing in different languages. We know someone is speaking in Hindi and may be someone else is speaking in French. The listener who knows Hindi can only understand the message from the partner speaking in Hindi. Someone knowing French can extract the French message.

```
A: 00011011
B: 00101110
C: 01011100
D: 01000010
```

9.3(a)

```
A: (-1,-1,-1,+1,+1,-1,+1,+1)
B: (-1,-1,+1,-1,+1,+1,+1,-1)
C: (-1,+1,-1,+1,+1,+1,-1,-1)
D: (-1,+1,-1,-1,-1,-1,+1,-1)
```

9.3(b)

```
--1-    C
-11-    B + C
10--    A + B̄
101-    A + B̄ + C
1111    A + B + C + D
1101    A + B + C̄ + D
```

9.3(c)

```
S1 = (-1,+1,-1,+1,+1,+1,-1,-1)
S2 = (-2, 0, 0, 0,+2,+2, 0,-2)
S3 = ( 0, 0,-2,+2, 0,-2, 0,+2)
S4 = (-1,+1,-3,+3,-1,-1,-1,+1)
S5 = (-4, 0,-2, 0,+2, 0,+2,-2)
S6 = (-2,-2, 0,-2, 0,-2,+4, 0)
```

9.3(d)

```
S1.C = (+1+1+1+1+1+1+1+1)/8 = 1
S2.C = (+2+0+0+0+2+2+0+2)/8 = 1
S3.C = (+0+0+2+2+0-2+0-2)/8 = 0
S4.C = (+1+1+3+3+1-1+1-1)/8 = 1
S5.C = (+4+0+2+0+2+0-2+2)/8 = 1
S6.C = (+2-2+0-2+0-2-4+0)/8 =-1
```

9.3(e)

Figure 9.3 CDMA Code Arithmetic

DSSS does the recovery by computing the normalized inner product of the received chip sequence (the linear sum of all the stations that transmitted) and the chip sequence of the station whose bit stream it is trying to recover. Let us assume that we are interested in recovering the bit sequence of station C. If the received chip sequence is S (S_1, S_2, … S_6) we compute the normalized inner product, S C. From each of the six sums S_1 through S_6, we calculate the bit by summing the pairwise products of the received S and the C vector of Figure 9.3(d) and then take 1/8 of the result. As shown in Figure 9.3(e), the product extracts the correct bit. Note that $S_3 \cdot C = 0$; this means that in the third example of 9.3(c) station C did not transmit. Also, note that $S_6 \cdot C = -1$; this means that in the sixth example station C transmitted a 0.

Walsh Function

The CDMA orthogonal codes are generated through Walsh function. Walsh functions are generated by code-word rows of special square matrices called Hadamard Matrices. These matrices contain one row of all 0s, with the remaining rows having an equal number of 1s and 0s. Walsh function can be constructed for block length $N = 2^j$, where j is an integer. The TIA IS-95 CDMA system uses a set of 64 orthogonal functions generated by using Walsh functions. The modulated symbols are numbered from 0 through 63.

The 64×64 matrix can be generated by using the following recursive procedure:

$$H_1 = [0];\ H_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix};\ H_8 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \\ \phi_6 \\ \phi_7 \\ \phi_8 \end{pmatrix};\ H_{2N} = \begin{pmatrix} H_N & H_N \\ H_N & \bar{H}_N \end{pmatrix}$$

Where N is a power of 2 and $\bar{H}_N$ is the complement of H_N.

The period of time required to transmit a single modulation symbol is called a Walsh symbol interval and is equal to 1/4800 seconds (203.33 μs). The period of time associated with 1/64 of the modulation symbol is referred as a Walsh chip and is equal to 1/307,200 seconds (3.255 μs). Within a Walsh symbol, Walsh chips are transmitted in the order 0, 1, 2, … 63.

For the forward channel (base station to mobile station), Walsh functions are used to eliminate multiple access interference among users within the same cell. Steps followed are:

- The input user data of individual user is multiplied by orthogonal Walsh functions.
- All the data of all the users are combined.
- The combined data is then spread by the base station (BS) pilot pseudo-random (PN) code.
- This spread signal is then transmitted on a radio carrier.
- At the receiver, the mobile removes the coherent carrier and gets the spread signal.
- The mobile receiver multiplies the signal by the synchronized PN code associated with the base station to get the spread data.
- The multiplication by the synchronized Walsh function for the i^{th} user will eliminate the interferences due to transmission from BS to other users.

In IS-95 or cdmaOne system different techniques are used for forward channel (BS to MS (Mobile Station)) and reverse channel (MS to BS) encoding. cdmaOne is the brand name of the service introduced by Qualcomm for digital mobile communication. The same technology was adopted by TIA (Telecommunication Industry Association) as IS-95 standard for second generation digital mobile communication in the US. Channelization in the forward link is accomplished through the use of orthogonal Walsh codes, while channelization in the reverse link is achieved using temporal offsets of the spreading sequence.

Different base stations are identified on the downlink based on unique time offsets utilized in the spreading process. Therefore, all base stations must be tightly coupled to a common time reference. In practice, this is accomplished through the use of the Global Positioning System (GPS), a satellite broadcast system that provides information on Greenwich Mean Time and can be used to extract location information about the receiver. This common time reference is known as system time.

There are two types of PN spreading sequences used in IS-95: the long code and the short code. Both the PN sequences are clocked at 1.2288 MHz, which is the chipping rate. Two short code PN sequences are used since IS-95 employs quadrature spreading. These two codes are the in-phase sequence

$$P_I(x) = x^{15} + x^{13} + x^9 + x^8 + x^7 + x^5 + 1$$

and the quadrature sequence

$$P_Q(x) = x^{15} + x^{12} + x^{11} + x^{10} + x^6 + x^5 + x^4 + x^3 + 1$$

These two sequences are generated using 15-bit shift register sequences; although they are nominally $2^{15} - 1 = 32767$ chips, a binary '0' is inserted in each sequence after a string of 14 consecutive 0's appears in either sequence to make the final length of the spreading sequence an even 32768 chips.

The long code is given by the polynomial

$$P(x) = x^{42} + x^{35} + x^{33} + x^{31} + x^{27} + x^{26} + x^{25} + x^{22} + x^{21} + x^{19} + x^{18} + x^{17} + x^{16} + x^{10} + x^7 + x^6 + x^5 + x^3 + x^2 + x^1 + 1$$

It is of length $2^{42} - 1$ chips as it is generated by a 42-bit shift register. It is primarily used for privacy, as each user of the mobile network may be assigned a unique temporal offset for the long code with reference to system time. Since the long code has a period of 41 1/2 days, it is nearly impossible to blindly detect a user's temporal offset. The offset is accomplished with the use of a long code mask, which is a 42-bit value that is combined with the shift.

BPSK and QPSK

The simplest form of a DSSS communications system employs coherent Binary Phase Shift Keying (BPSK) for both the data modulation and spreading modulation. But the most common form of DSSS uses BPSK for data modulation and QPSK (Quadrature Phase Shift Keyed) modulation for spreading modulation. QPSK modulation can be viewed as two independent BPSK modulations with 180 degree phase difference.

The input binary bit stream $\{d_k\}$, $d_k = 0, 1, 2, \ldots$ arrives at the modulator input at a rate 1/T bits/sec and is separated into two data streams $d_I(t)$ and $d_Q(t)$ containing odd and even bits respectively like,

$$d_I(t) = d_0, d_2, d_4, \ldots$$
$$d_Q(t) = d_1, d_3, d_5, \ldots$$

QPSK can be viewed as two independent BPSK modulations. Figure 9.4 depicts an example of QPSK for a bit stream 00111000.

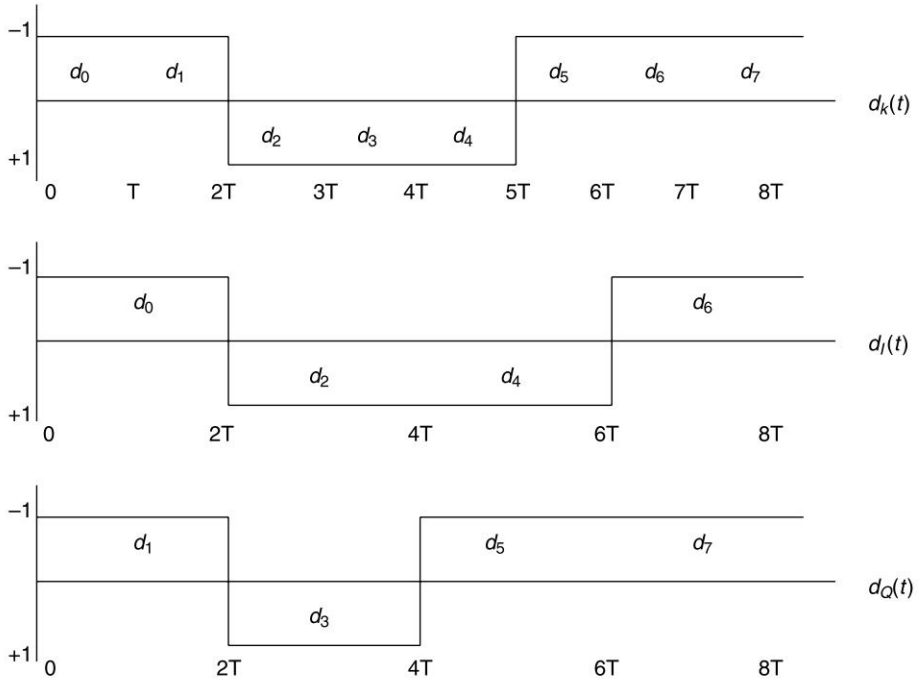


Figure 9.4 The QPSK Modulation

9.3 IS-95

Prof. George Cooper of Purdue University, in the US did some work in the late 1970s in the field of spread-spectrum communications and its usage for mass commercial deployment. In a commercial cellular system we need to increase the transmission power when the mobile user moves further and reduce the power when the user comes closer to the base station. Cooper recognized the need for some type of power control system to overcome this near–far effect prevalent in CDMA systems, but could not achieve this. Qualcomm overcame some of these challenges and in the mid-1980s developed the first commercial spread-spectrum-based system for use in the cellular band. This system was considered an attractive alternative to the analogue FDMA technologies (AMPS, primarily) and TDMA systems (IS-54, IS-136, GSM). As mentioned earlier, this resulted in the Telecommunications Industry Association (TIA) developing the IS-95 standard. This standard formed the basis for the first CDMA systems deployed in the cellular band (from 800 to 900 MHz) in North America. This development eventually led to the TIA working with the T1P1 to develop the J-STD-008 standard for the PCS band (from 1800 to 1900 MHz). Since then, there has been some effort to enhance symmetric data rates for IS-95, resulting in the formation of a new standard in 1998, IS-95-B. The IS-95 family of standards is known as cdmaOne. It is a second generation digital mobile communication system.

9.3.1 Speech and Channel Coding

The normal audio range of human beings is between 20 Hz to 20 KHz. However, this range is normally used for high fidelity CD quality music. For telephonic communications where generally human voice is used, the frequency range of 300 to 3300 Hz is sufficient. For digitizing the speech, it is sufficient to sample at 8000 samples per second (assuming information bandwidth up to 4000 Hz). Therefore, to achieve telephone quality speech, 12 bits are sufficient to encode each sample. By using logarithmic sampling system 12 bits can be reduced to 8 bits per sample. This results in the PCM encoding of the speech and digitization of the voice at 64 Kbps. This digitized voice is then passed through a coding scheme using Code-Excited Linear Prediction (CELP) algorithm. Linear Prediction Coding (LPC) is a combination of waveform coding and vocoder. Vocoder emulates the human vocal cord functions electronically and generates synthesized voice. In this process the analog voice is converted into 9.6 Kbps digitized data.

In a mobile telecommunication environment, signal strength varies with location and movement of the mobile transmitter/receiver. Signal strength influences error rates, which in turn affects the quality of communication. Due to varying signal strengths, mobile telecommunication systems are susceptible to burst errors. Burst errors are groupings of errors in adjacent bits as compared to errors that are dispersed over the whole data block. IS-95 addresses the problem of burst errors by utilizing an error correction scheme based on encoding and interleaving. Generally, interleaving is used in conjunction with encoding (e.g., error-correcting codes) in order to lower the error rates. Interleaving is a technique in which encoded digital data is reordered before transmission in such a manner that any two successive digital data bits in the original data stream are separated by a predetermined distance in the transmitted data stream. Deinterleaving is the reverse of interleaving where data bits are reordered back to their original sequence.

9.3.2 IS-95 Architecture

The Key to the North American systems is the use of a common reference model from the cellular standards group TR-45. Different network entities within IS-95 are very similar to the network elements within a GSM network.

cdmaOne or IS-95 uses CDMA for its radio or last mile communication. Other than the radio interface, the rest of the network and especially the core are very similar to GSM.

The main elements of IS-95 (Fig. 9.5) reference model are:

- *Mobile Station (MS)*: This is the mobile phone unit with the user. The MS terminates the radio path on the user side and enables the user to gain access to services from the network. The MS can be a stand-alone device. It can have other devices (e.g., personal computers, fax machines) connected to it where it works as a pass through.
- *Base Station (BS)*: The BS terminates the radio path and connects to the mobile switching center (MSC). BS is a system between the MS and the MSC. The BS is segmented into the BTS and BSC.
  - ❏ Base Transceiver Station (BTS): BTS consists of one or more transceivers placed at a single location and terminates the radio path on the network side.
  - ❏ Base Station Controller (BSC): The BSC is the control and management system for one or more BTSs. The BSC exchanges messages with both the BTS and the MSC. Some signaling messages may pass through BSC transparently.

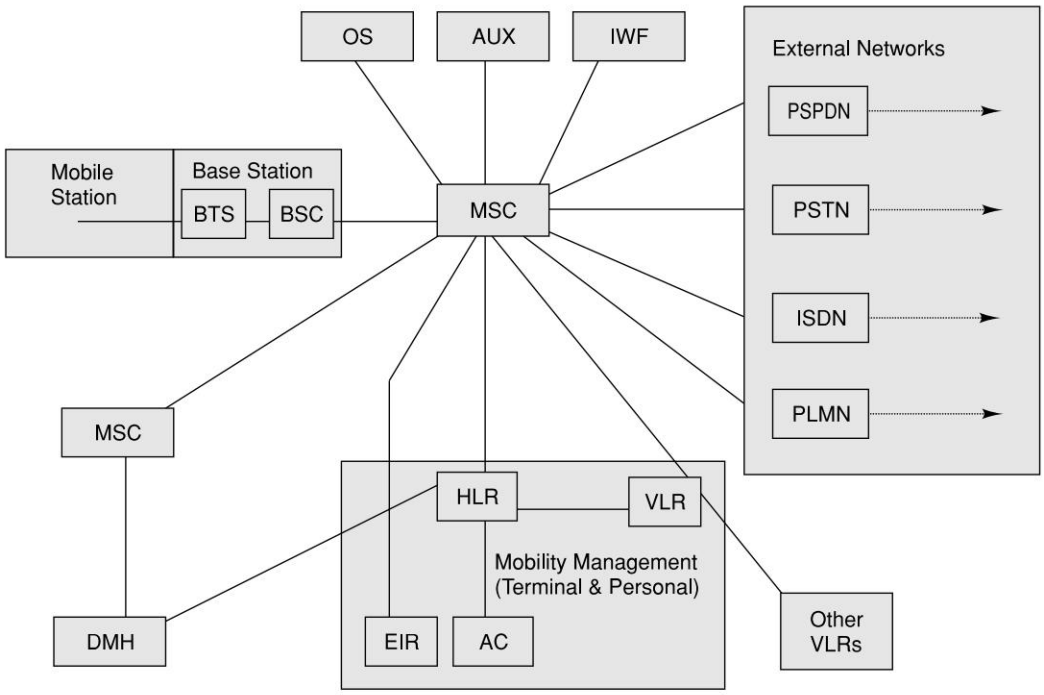


Figure 9.5 The IS-95 Architecture Model

- *Mobile Switching Center (MSC)*: This is the main switching center equivalent to the telephone exchange in a fixed network. The MSC is an automatic system that interfaces the user traffic from the wireless network with the wireline network or other wireless networks. The MSC does one or more of the following functions:
  - ❏ Anchor MSC: First MSC providing radio contact to a call
  - ❏ Border MSC: An MSC controlling BTSs adjacent to the location of the mobile station
  - ❏ Candidate MSC: An MSC that could possibly accept a call or a handoff
  - ❏ Originating MSC: The MSC directing an incoming call towards a mobile station
  - ❏ Remote MSC: The MSC at the other end of an intersystem trunk
  - ❏ Serving MSC: The MSC currently providing service to a call
  - ❏ Tandem MSC: An MSC providing only trunk connections for a call in which a handoff has occurred
  - ❏ Target MSC: The MSC selected for a handoff
  - ❏ Visited MSC: The MSC providing service to the mobile station
- *Home Location Register (HLR)*: HLR is the functional unit that manages mobile subscribers by maintaining all subscriber-related information. The HLR may be collocated with an MSC as an integral part of the MSC or may be independent of the MSC. One HLR can serve multiple MSCs or an HLR may be distributed over multiple locations.
- *Data Message Handler (DMH)*: The DMH is responsible for collating the billing data.

- *Visited Location Register (VLR)*: VLR is linked to one or more MSCs and is the functional unit that dynamically stores subscriber information obtained from the subscriber's HLR data. When a roaming MS enters a new service area covered by the MSC, the MSC informs the associated VLR about the MS by querying the HLR after the MS goes through a registration procedure. VLR can be considered as cache whereas HLR is similar to a persistent storage.
- *Authentication Center (AC)*: The AC manages the authentication associated with individual subscriber. The AC may be located within an HLR or MSC or may be located independent of both.
- *Equipment Identity Register (EIR)*: The EIR provides information about the mobile device for record purposes. The EIR may be located with the MSC or may be located independent of it.
- *Operations System (OS)*: The OS is responsible for overall management of the wireless network.
- *Interworking Function (IWF)*: The IWF enables the MSC to communicate with other networks.
- *External Networks*: These are other communication networks and can be a Public Switched Telephone Networks (PSTN), an Integrated Services Digital Network (ISDN), a Public Land Mobile Network (PLMN) or a Public Switched Packet Data Network (PSPDN).

9.3.3 IS-95 Channel Structure

IS-95 system operates on the same frequency band as the first generation AMPS (Advanced Mobile Phone System). It uses Frequency Division Duplex (FDD) with 25 MHz in each direction. It uses 824 to 849 MHz for forward link (base station to mobile station) and 869 to 894 MHz for reverse link (mobile station to base station). In digital communication, one data path maps on to one communication channel. In FDMA system one channel occupies a distinct frequency band. In TDMA, it is a distinct timeslot within a frequency. In CDMA a channel is defined in terms of a code sequence and frequency. This results in offering a higher channel capacity, which translates into an overall higher bandwidth.

As mentioned earlier, IS-95 uses a different modulation and spreading technique for forward link and reverse link. In forward link 64 Walsh codes are used to map 64 logical channels. The base station simultaneously transmits the user data for all mobiles in the cell by using different Walsh codes for each mobile. This is then spread using a PN of length 2^{15} chips. The user data is spread to a channel chip rate of 1.2288 Mchips. On the reverse link, channels are identified by long PN sequence.

For forward channels, base stations transmit information in four logical channel formats: pilot channels, sync channels, paging channels (PCH), and traffic channels (Code) (Fig. 9.6). On the reverse link, all mobiles respond in an asynchronous fashion. The user data is encoded, interleaved, and then blocks of 6 bits are mapped to one of the 64 orthogonal Walsh functions. Finally, the data is spread by a user specific code of 42 bits (channel identifier). The reverse channel is organized in access and traffic channels.

At both the base station and the terminal, Rake receivers are used to resolve and combine multipath components, in order to improve the link quality. In IS-95, a three-finger Rake receiver is used at the base station.

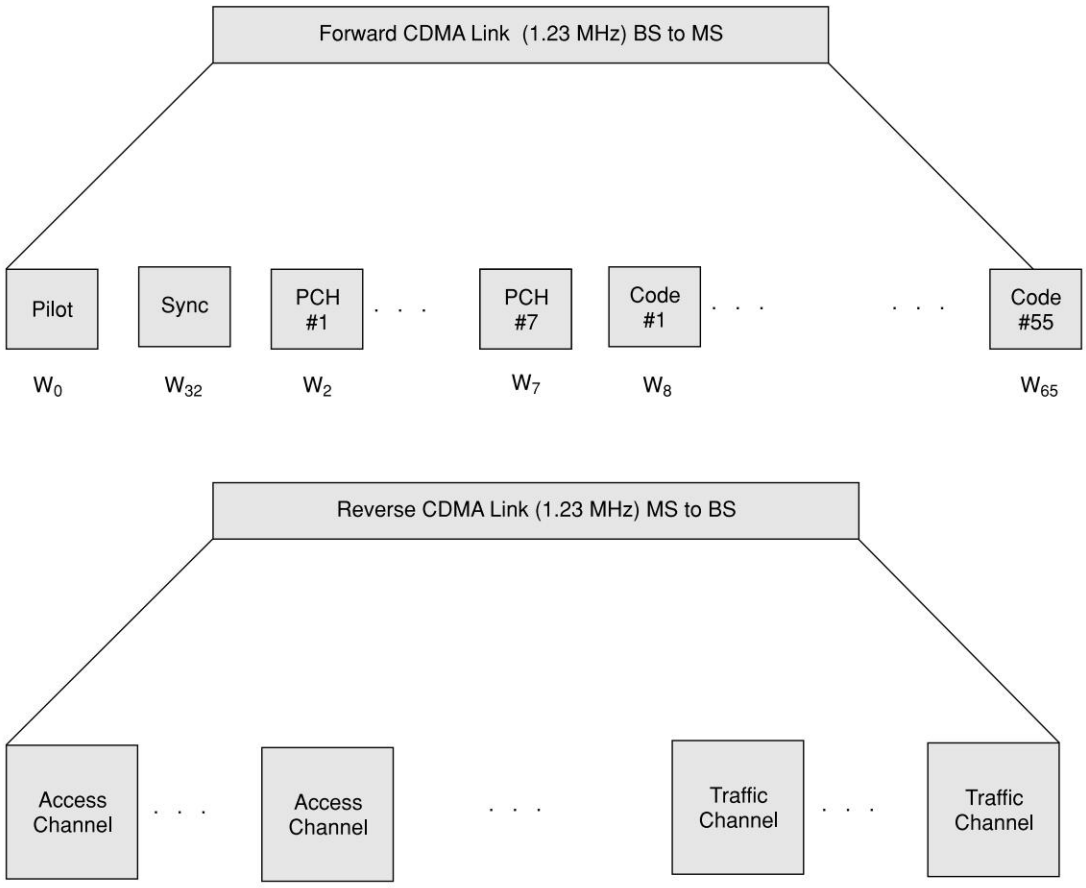


Figure 9.6 IS-95 Forward and Reverse Link Channel Structures

In FDMA and TDMA multipath causes signal interference. In 1958, Price and Green proposed a method of resolving multipath problem in CDMA. In CDMA the time shifted version of the same signal appears like a noise and is almost uncorrelated. A signal that propagates from transmitter to receiver over multiple paths can be resolved into separately fading signals by cross-correlating the received signal with multiple time shifted version of the same sequence. In the received three signals with maximum power are time shifted and added. The block diagram of this technique looks like a garden rake, hence the name rake receiver. A three-finger rake receiver takes three multipath signals.

Pilot Channel: The pilot CDMA signal transmitted by a base station provides a reference for all mobile stations. It is assigned the Walsh code W0 (all 0). The pilot signal level for all base stations is kept about 4 to 6 dB higher than the traffic channel with a constant signal power. This is because

the MS at the cell boundaries should be able to receive the pilot signal from other cells to decide when to perform handoff. The pilot signals from all base stations use the same PN sequences, but each base station is identified by a unique time offset. These offsets are in increments of 64 chips to provide 512 unique offsets.

Sync Channel: Sync channel is assigned the Walsh function W32 and is used with the pilot channel to acquire initial time synchronization. W32 has a pattern of 32 consecutive 0s and 32 consecutive 1s, which is ideal for synchronization. The Sync channel message parameters are: System Identification (SID), Network Identification (NID), Pilot short PN sequence offset index, Long-code state, System time, Offset of local time, Daylight saving time indicator and Paging Channel data rate (4.8 Kbps or 9.6 Kbps).

Paging Channel: There are up to seven paging channels that transmit control information to the terminals that do not have calls in progress. The paging channels are assigned the Walsh functions W1 to W7. Some of the messages carried by the paging channel include:

- System Parameter Message: such as base station identifier, the number of paging channels and the page channel number.
- Neighbor List Message: information about neighbor base station parameters, such as the PN Offset.
- Access Parameters Message: parameters required by the mobile to transmit on an access channel.
- Page Message: provides a page to the mobile station.
- Channel Assignment Message: to inform the mobile station to tune to a new carrier frequency.
- Data Burst Message: data message sent by the base station to the mobile.
- Authentication Challenge: allows the base station to validate the mobile identity.

Access Channel: Access channel is used by a terminal without a call in progress to send messages to the base station for three principal purposes: to originate a call, to respond to a paging message, and to register its location. Each base station operates with up to 32 access channels. The messages carried by the access channel include:

- *Order Message*: The transmits information such as base station challenge, mobile station acknowledgement, local control response and mobile station reject.
- *Registration Message*: Sends to the base station information necessary to page the mobile such as location, status and identification.
- *Data Burst Message*: User-generated data message sent by the mobile station to the base station.
- *Origination Message*: Allows the mobile station to place a call sending dialed in digits.
- *Authentication Challenge Response Message*: Contains necessary information to validate the mobile stations identity.

Forward Traffic Channel: Channels not used for paging or sync can be used for traffic. Thus, the total number of traffic channels at the base station is 63 minus the number of paging and sync channels in operation at the base station. Information on the forward traffic channel includes the primary traffic (voice or data) secondary traffic (data) and signaling. When the forward link is used for signaling, some of the typical messages would be:

- *Order Message*: Similar to the order message in forward traffic channel.
- *Authentication Challenge Message*: Used to prove the identity of the mobile when the base station suspects its validity.
- *Alert with Information Message*: Allows the base station to validate the mobile identity.
- *Handoff Direction Message*: Provides the mobile with information needed to begin the handoff process.
- *Analog Handoff Direction Message*: Tells the mobile to switch to the analog mode and begin the handoff process.
- *In-traffic System Parameters Message*: Updates some of the parameters set by the System Parameters message in the paging channel.
- *Neighbor List Update Message*: Updates the neighbor base station parameters set by the Neighbor List message in the paging channel.
- *Data Burst Message*: Data message sent by the base station to the mobile.
- *Mobile Registration Message*: Informs the mobile that it is registered and supplies the necessary system parameters.
- *Extended Handoff Direction Message*: One of several handoff messages sent by the base station.

Reverse Traffic Channels: This channel can multiplex primary (voice) and secondary (data) or signaling traffic. Some of the typical messages that the reverse traffic channel carries are:

- *Order Messages*: Include base station challenge, parameter update confirmation, mobile station acknowledgement, service option request and response, release, connect, DTMF (Dual Tone Multi Frequency) tone, etc.
- *Authentication Challenge Response Message*: Information to validate the mobile station.
- *Pilot Strength Measurement Message*: Information about the strength of other pilot signals that are not associated with the serving base station.
- *Data Burst Message*: A user-generated data message sent by the mobile to the base station.
- *Handoff Completion Message*: Is the mobile response to a Handoff Direction message.
- *Parameter Response Message*: Is the mobile response to the base station to a Retrieve Parameters message.

9.3.4 IS-95 Call Processing

To set up a call or to transmit data, a data path needs to be established through a traffic channel. To establish a traffic channel, a mobile station in IS-95 goes through several states. They are:

- System initialization.
- System idle state.
- System access.
- Traffic channel state.

In the system initialization state the mobile acquires a pilot channel by searching all the PN offsets possibilities and selecting the strongest pilot (W0) signal. Once the pilot is acquired, the sync channel is acquired using the W32 Walsh function and the detected pilot channel. Then the mobile obtains the system configuration and timing information.

Next the mobile enters the system idle state where it monitors the paging channel. If a call is being placed or received, the mobile enters the system access state where the necessary parameters are exchanged. The mobile transmits its response on the access channel and the base station transmits its response on the paging channel. When the access attempt is successful the mobile enters the traffic state. In the traffic state voice or data is transacted.

CDMA Registration

The registration process is used by the mobile device to notify its location, status, identification and other characteristics. Location information is required to page the mobile for an incoming mobile terminated call. When the MS does power on or power off it goes through the registration process as well. These functions are similar to GSM. Registration information is stored in HLR.

9.3.5 Authentication and Security

The Electronic Serial Number (ESN) of a IS-95 mobile station is a 32 bits binary number that identifies the mobile. It is factory-set and is not alterable in the field. All mobiles are assigned a unique ESN when manufactured. A mobile station also has a unique 15-digit number called Mobile Identification Number (MIN). This is the mobile's 10 digit directory number similar to the MSISDN number in GSM. The difference is that in IS-95, the mobile is assigned a number similar to the North American numbering scheme. For example, a fixed line number in Bangalore may have a directory number of 080-2593-2137 whereas a mobile phone may have a directory number of a GSM phone like 98450-62050. On the contrary in the US if a fixed line number is 1-630-858-7131, the mobile number can be 11-630-240-8900.

Whenever a mobile is turned on, it registers with the network. During the authentication process the network throws a global challenge to the mobile. The AC (Authentication Center) transmits a random number to the mobile station. The mobile station encrypts it using a key shared between the mobile station and the AC. The encrypted random message is sent back to the network. The AC checks this result with the result it calculates using the same random number and shared key. If these match, the mobile is authenticated. The global challenge in IS-95 is more frequent than in GSM. In IS-95, authentications even take place following successful handoffs. Following successful authentication, the VLR assigns a TMSI (Temporary International Mobile Subscriber Identity). The TMSI provides anonymity since it is a transient identity only the mobile and the network are aware of. The spreading PN (Pseudorandom Noise) sequence also play a role in security. For anybody to impersonate the CDMA traffic, the eavesdropper needs to know the PN sequence.

9.3.6 Handoff and Roaming

A handover in GSM is called a handoff in IS-95. When a subscriber moves away from a base station, the signal power reduces resulting in a potential drop in connection. To ensure that the call does not break, some other base station closer to the mobile station needs to attach the mobile to it and let the call continue without any interruption.

In CDMA, handoffs are handled differently compared to GSM. GSM handover is a hard handover. In a hard handover, the attachment with the current cell is broken first and then a new connection is set up with another cell. In GSM it is "break before make". In CDMA, the spectrum

is spread and everybody gets the same signal. Logically a mobile station in CDMA is always connected to different base stations at the same time. Therefore, handoff is managed by changing the attachment. There are three types of handoffs in CDMA. These are Soft handoff, Hard handoff, and Softer handoff.

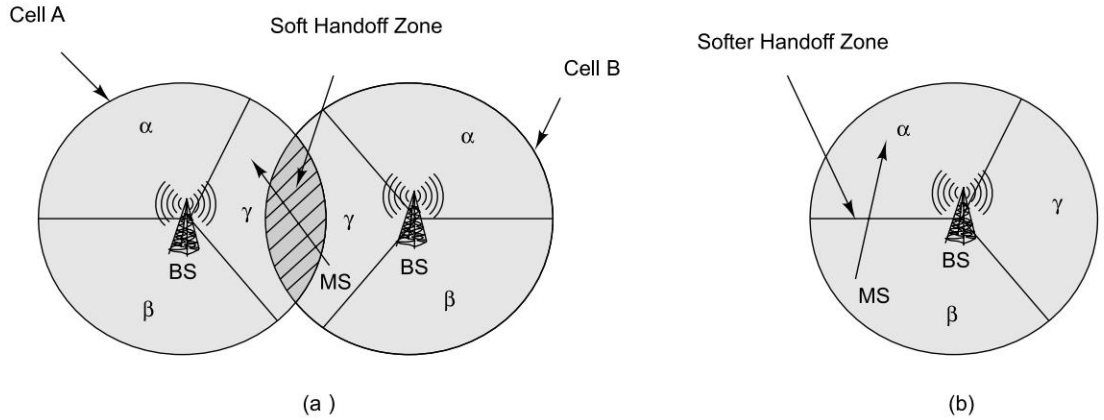


Figure 9.7 (a) Soft Handoff; (b) Softer Handoff

In CDMA a cell is divided into sectors. Like in GSM it is normally divide into three sectors each covering 120°. CDMA antennas are either Switched Beam System (SBS) or an Adaptive Antenna System (AAS). The SBS uses multiple fixed beams in a sector and a switch to select the best beam to receive a signal. In an AAS, the receiving signals by multiple antennas are weighted and combined to maximize the signal to noise ratio (SNR).

- *Soft Handoff*: This is the case of intercell handoffs (Fig. 9.7(a)). Soft handoff is a process in which the control of a mobile station is assigned to an adjacent cell or an adjacent sector (in the same frequency) without dropping the original radio link. The mobile keeps two radio links during the soft handoff process. Once the new communication link is well established, the original link is dropped. This process is also known as "make before break", which guarantees no loss of voice during handoff. In Figure 9.7(a), as the user moves, a soft handoff takes place from Cell B to Cell A.
- *Hard Handoff*: This is the case of interfrequency handoffs. CDMA to CDMA hard handoff is the process in which a mobile is directed to handoff to a different frequency assigned to an adjacent cell or a sector. The mobile drops the original link before establishing the new link. This is similar to a GSM handover. The voice is muted momentarily during this process. This handoff is completed very fast and cannot be noticed.
- *Softer Handoff*: A mobile communicates with two sectors of the same cell (Fig. 9.7(b)). A rake receiver at the base station combines the best version of the voice frame from the diversity antennas of the two sectors into a single traffic frame. This is a logical handoff where signals from multiple sectors are combined instead of switching from one sector to another.

9.3.7 IS-95 Channel Capacity

In first generation mobile networks the frequency channels were fixed and hence the capacity too. This is true with GSM as well. In GSM we multiply 125 frequencies with eight time slots to get 1000 channels. Therefore TDMA and FDMA capacities are bandwidth limited and hard-limited. The capacity of CDMA has a soft limit in the sense that we can add one additional user and tolerate a slight degradation of the signal quality. This is similar to a room full of people. Let us assume that people are talking to each other using a loudspeaker. In such a case not many people will get a chance to talk or to listen. However, more people can talk to each other if they converse in low voices. Another conclusion that can be drawn from this fact is that, any reduction in the multiple access interference converts directly and linearly into an increase in the capacity. The capacity of a CDMA system depends on the following criteria:

- *Voice Activity Detection (VAD)*: The human voice activity cycle is 35 percent. This means that during a conversation people talk about 35% of the time. When users assigned to a cell are not talking, VAD will allow all other users to benefit due to reduced mutual interference. Thus interference is reduced by a factor of 65 percent. CDMA is the only technology that takes advantage of this phenomenon. It can be shown that the capacity of CDMA is increased by about three times due to VAD.
- *Sectorization for Capacity*: In FDMA and TDMA systems, sectoring is done to reduce the co-channel interference. In GSM there are in total 1000 channels distributed between multiple operators, sectors and cells. The trunking efficiency of these systems decreases due to sectoring. This in turn reduces the capacity. On the other hand, sectorization increases the capacity of CDMA systems. Sectoring is done by simply introducing three (similar) radio equipments in three sectors. The reduction in mutual interference due to this arrangement translates into a three-fold increase in capacity (in theory). In general, any spatial isolation through the use of multibeam or multisector antennas provides an increase in the CDMA capacity.
- *Frequency Reuse Considerations*: The previous comparisons of CDMA capacity with those of conventional systems primarily apply to mobile satellite (single cell) systems. In the case of terrestrial cellular systems, the biggest advantage of CDMA over conventional systems is that it can reuse the entire spectrum over all the cells since there is no concept of frequency allocation in CDMA. This increases the capacity of the CDMA system by a large percentage (related to the increase in the frequency reuse factor).

As a rule of thumb the CDMA capacity is about four times that of TDMA and eight times that of FDMA.

9.4 CDMA VERSUS GSM

GSM is a relatively mature technology, now several years in existence with a huge installation base. GSM has many experienced operators and equipment manufacturers. Interoperability within GSM is well proven. GSM is complete, open and has proven standards. GSM includes all the specifications from the handset other over the air, switch, interconnect it with switching, and every-aspect of mobile telecommunication. On the other hand, IS-95 is mainly a single vendor (Qualcomm cdmaOne) specification. IS-95 only covers the air interface making it incomplete. Though there are many claims

and counter claims, it is generally believed that CDMA has high potential to address some of the difficult challenges of the past quite effectively. These are described in Table 9.1.

Table 9.1 GSM versus 3G

Functions	*GSM*	*IS-95*
Frequency	900 MHz; 1800 MHz (DCS180); 1900 MHz (PCS 1900)	800 MHz; 1900 MHz
Channel bandwidth	Total 25 MHz bandwidth with 200 KHz per channels, 8 timeslots per channel with frequency hopping	Total 12 MHz with 1.25 MHz for the spread spectrum
Voice codec	13 Kbits/second	8 Kbits/sec or 13 Kbps
Data bit rate	9.6 Kbits/second and expandable	9.6 Kbits
Short message service	160 characters of text Supports	120 characters
SIM card	Yes	No
Multipath	Causes interference and destruction to service	Used as an advantage
Radio interface	TDMA	CDMA
Handoff	Hard Handover (handoff)	Soft Handoff (handover)
System Capacity	Fixed and limited	Flexible and higher than GSM
Economics	Expensive	Due to many technological advantages, dimension of investment per subscriber is expected to be lower than GSM

9.5 Wireless Data

Data transmission over wireless networks like CDMA or GSM is always a challenge. Typically raw channel data error rates for cellular transmission are 10^{-2}. This means that one in every 100 bits has an error. This is an error rate, which can be tolerated for voice transmission. This is because; our perception of hearing cannot detect it. Even if our ear is sometime able to detect it, our mind is able to correct it from the context. This error rate of 10^{-2} is too high for data transmission. An acceptable BER (Bit Error Rate) for data transmission is 10^{-6}. This means that one bit in a million can be tolerated as an error. In order to achieve this high level of reliability, it requires a design of effective error correction code and Automatic Repeat Request (ARQ). The CDMA protocol stack (Fig. 9.8) for data and facsimile has the following layers.

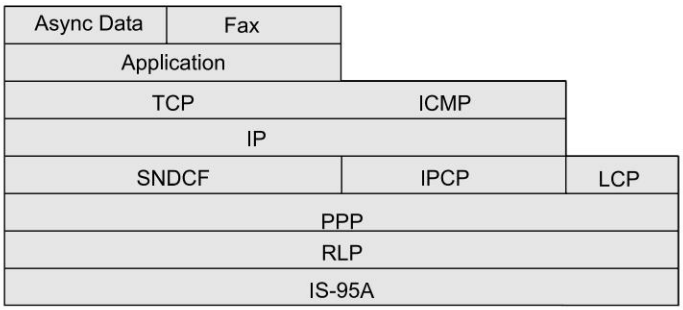
Application Interface Layer: This layer includes an application interface between the data source in the mobile station and the transport layer. The application interface provides functions like modem control, AT (Attention) command processing, data compression, etc.

Transport Layer: The transport layer for CDMA asynchronous data and fax is based on TCP. TCP has been modified for IS-95.

Network Layer: The network layer for CDMA asynchronous data and fax services is based on IP. The standard IP protocol has been enhanced for IS-95.

Sub-network Dependent Convergence Function: The SNDCF performs header compression on the header of the transport and network layers. Mobile station supports Van Jacobson TCP/IP header compression algorithm. Negotiation of the parameters for header compression is carried out using IPCP (Internet Protocol Control Protocol). The SNDCF sublayer accepts the network layer datagram packets from the network layer, performs header compression and passes that datagram to the PPP (Point to Point Protocol) layer. In the reverse operation, it receives network layer datagrams with compressed header from the PPP layer and passes it to the network layer.

Data Link Layer: This layer uses PPP. The PPP Link Control Protocol (LCP) is used for initial link establishment and for the negotiation of optional link capabilities.



ICMP : Internet Control Message Protocol
IP : Internet Protocol
IPCP : Internet Protocol Control Protocol
LCP : Link Control Protocol
PPP : Point-to-Point Protocol
RLP : Radio Link Protocol
SNDCF : Subnetwork Dependent Convergence Function
TCP : Transmission Control Protocol

Figure 9.8 CDMA Data Protocol Stack

Internet Protocol Control Protocol Sublayer: This sublayer supports negotiation of the IP address and IP compression protocol parameters. In general, a mobile station does not have a permanent IP address. Therefore, the IP address needs to be negotiated and obtained from the network. IPCP does this job of leasing an IP address when the transport connection is established. The IP address is discarded when the connection is closed. This is similar to obtaining the IP address from a DHCP (Dynamic Host Configuration Protocol) server in a LAN environment.

Radio Link Protocol Layer: This layer provides octet stream service over the air. This service is responsible for reducing the error rate over the forward and reverse channels. There is no direct

relationship between PPP packet and the traffic channel frame. A large packet may span multiple traffic channel frames. A single traffic channel frame may contain multiple PPP packets. RLP frames may be transported as traffic or signaling via data burst message.

9.5.1 Short Message Service

SMS in IS-95 is similar to SMS in GSM. Unlike GSM, the maximum size of a SMS in IS-95 is 120 octets. The SMS in IS-95 work the same way as in GSM. It supports SMPP protocol and other features as in GSM. Like in GSM, the SMS in IS-95 uses the signaling channel for data transfer. SMS administration features include storage, profiling, verification of receipt and status enquiry capabilities.

9.6 Third Generation Networks

The telecommunications world is changing due to trends in media convergence and industry consolidation. The perception of mobile phone has changed significantly over the last few years. More changes predicted for the future are:

- The mobile devices will be used as an integral part of our lives.
- Data ("non-voice") usage of 3G will become important and different from the traditional voice business.
- A great deal of convergence will take place between information and communication technology.
- The look of the phone will be as important as its usage.
- Mobile communications will be similar in its social positioning. People will have only a mobile device.

To address these challenges and opportunities, the mobile telecommunication technology needs to adapt new techniques, facilities and services. The 3G system will offer a plethora of telecommunication services including voice, multimedia, video and high speed data. With 3G mobile Internet technology significant changes will be brought about in the day-to-day life of the people.

CDMA is the preferred approach for the third generation networks and systems. In North America cdma2000 is the version of 3G. cdma2000 standards are being driven by Telecommunication Industries Association (TIA). It uses the CDMA air interface, which is based on IS-95 and cdmaOne. In Japan 3G standard uses (Wideband Code Division Multiple Access) WCDMA (DoCoMo) version. This standard is being driven by ARIB. In Europe, Asia, Australia and many parts of the world 3G has been accepted as UMTS and WCDMA. UMTS/WCDMA is being driven by ETSI, and is the normal evolution from GSM/GPRS.

The main goal of UMTS (Universal Mobile Telecommunications System) is to offer a much more attractive and richer set of services to the users.

- *Universal Roaming*: Any user will be able to move across the world and access the network.
- *Higher Bit Rate*: More speed would open the path toward multimedia applications.

In the beginning of 1998 six partners—ARIB (Association of Radio Industries and Businesses), T1, TTA (Telecommunications Technology Association, Korea), ETSI in Europe, CWTS (China Wireless Telecommunication Standard group), TTC (Telecommunication Technology Committee, Japan) started discussions to cooperate for creating a standards for a third generation mobile system with a core network based on evolution for GSM and an access network based on all the radio access technologies supported by the different partners. This project was called the Third Generation Partnership Project (3GPP). About a year later ANSI decided to establish 3GPP2, a 3G partnership project for evolved ANSI/Telecommunications Industry Association (TIA)/ Electronics Industry Association (EIA)-41 networks. There is also a strategic group called International Mobile Telecommunication Union-2000 (IMT-2000) within the International Telecommunication Union (ITU), which focuses its work on defining interface between 3G networks evolved from GSM on one hand and ANSI-41 on the other, in order to enable seamless roaming between 3GPP and 3GPP2 networks. 3GPP started referring to 3G mobile system as Universal Mobile Telecommunication System (UMTS).

- *Mobile-Fixed Convergence*: There is a need to offer users cross-domain services. An example is the tracking of a user's location in the mobile, fixed and Internet domain and automatically adapting the content of his incoming messages to SMS, voice message, fax or email. VHE (Virtual Home Environment) is the enabler to this service portability across networks and terminals in different domains.
- *Flexible Service Architecture*: By standardizing not the services themselves but the building blocks that make up services, UMTS shortens the time for marketing services from GSM and enhances creativity/flexibility when inventing new services.

9.6.1 International Mobile Telecommunications–2000

2G mobile networks were mainly built for digital voice; data was available only over circuits. The first major step towards packet data in the evolution to 3G occurred with the introduction of GPRS, which came to be known as 2.5G. GPRS offered a moderate data bandwidth that was sufficient for services like Wireless Application Protocol (WAP) access, Multimedia Messaging Service (MMS) and low bandwidth Internet. GPRS networks evolved into Enhanced Data rates for GSM Evolution (EDGE) networks that offered high bandwidth packet data capable of multimedia video; however, it fell slightly short of 3G and is often referred to as 2.75G. Then, finally IMT 2000/3G evolved.

International Mobile Telecommunications-2000 (IMT-2000) is the global standard for third generation (3G) wireless communications, defined by a set of interdependent ITU Recommendations. IMT-2000 will provide a framework for worldwide access of services by linking the diverse systems of terrestrial and/or satellite based networks through the synergy between digital mobile telecommunications technologies and systems for fixed and mobile wireless access systems. IMT-2000 was, originally, envisioned to be launched in the year 2000 with a bandwidth of 2000Kbits/second (2Mbps). It is also popularly known as 3G or 3rd Generation that includes EDGE, CDMA 2000, UMTS, DECT and WiMAX (which was added in 2007) standards. These

standards are both evolutionary and revolutionary. They are evolutionary standards in the sense that they are backward compatible to interoperate with pre-existing 2G networks while they are revolutionary as they require all-new networks and frequency allocations. Various independently developed standards like DECT and WiMAX were included because they fit the IMT-2000 definition. The services provisioned by IMT-2000 set of standards include wide area wireless voice telephone, video calls and wireless data for a mobile user. Therefore, 3G networks enable network operators to offer users a wide range of more sophisticated services while achieving greater network capacity through improved spectral efficiency (which is possible through simultaneous use of speech and data services and better data rates). Table 9.2 shows an overview of IMT-2000 standards.

Table 9.2 Overview of 3G/IMT 2000 standards

ITU IMT-2000	*Common name(s)*	*High-speed data*	*Pre-4G Duplex*	*Channel*	*Description*	*Geographical areas*
CDMA Single-Carrier (IMT-SC)	EDGE (UWT-136)	EDGE Evolution	None	TDMA	Evolutionary upgrade to GSM/GPRS	Worldwide except Japan and Korea
CDMA Multi-Carrier (IMT-MC)	CDMA 2000	EV-DO	UMB	FDD CDMA	Evolutionary upgrade to cdmaOne (IS-95)	America, Asia and some others
CDMA Direct Spread (IMT-DS) CDMA TDD (IMT-TC)	W-CDMA UMTS TD-CDMA TD-SCDMA	HSPA	LTE		Family of revolutionary standards.	Worldwide esp. Europe and China
FDMA/TDMA (IMT-FT)	DECT	None	TDD	FDMA/ TDMA	Short-range; standard for cordless phone	Europe, USA
IP-OFDMA	WiMAX (IEEE 802.16)		OFDMA		Late addition	Worldwide

Evolution beyond IMT-2000 and towards 4G

3GPP and 3GPP2 are researching on further extensions to current 3G standards, namely Long Term Evolution (LTE) and Ultra Mobile Broadband, respectively. As these technologies would be fully based on an all-IP network infrastructure, they have started displaying characteristic features for IMT-Advanced (4G) standards. Practically, these standards fall short of the speed requirements for 4G (which is set to be 1 Gbit/s for stationary and 100 Mbit/s for mobile operations). Therefore, these standards are classified as 3.9G or Pre-4G. Progress is all the more accelerated by valuable contributions from mobile IT Forum (mITF) Japan and other professional bodies in Europe.

9.6.2 CDMA-2000

cdma2000 is the third generation version of cdmaOne or IS-95. The cdma2000 Radio Transmission Technology (RTT) is a spread spectrum, wideband radio interface. It uses CDMA technology as its underlying modulation technology. cdma2000 meets the specification for ITU (International Telecommunication Union) and IMT-2000. It addresses the specification for indoor, indoor-to-outdoor, pedestrian and vehicular environment. cdma2000 can operate in wide range of environments, viz.,

- Indoor/Outdoor picocell (< 50 meter radius; e.g., one office floor)
- Indoor/Outdoor microcell (up to 1 km radius; e.g., a shopping mall)
- Outdoor macrocell (1–35 km radius)
- Outdoor megacell (> 35 km radius)
- Wireless in Local Loop (WiLL).

cdma2000 supports chip rates of N × 1.2288 Mcps (where N = 1, 3, 6, 9, 12). For N = 1, the spreading is similar to IS-95. However, for forward link QPSK modulation is used before the spread. There are two option for chips rate for N > 1. These are multicarrier and direct spread (Fig. 9.9). In the multicarrier procedures for N > 1, the modulation symbols are demultiplexed on to N separate 1.25 MHz carriers where N = 3, 6, 9, 12. Each of these carriers is then spread with 1.2288 M chips. For direct spread procedures for N > 1, the modulation symbols are spread on a single carrier with a chip rate of N × 1.2288 M chips where N = 3, 6, 9, 12.

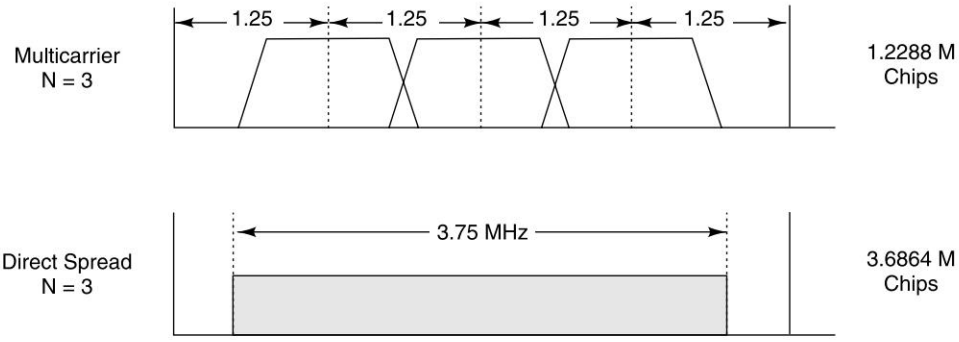


Figure 9.9 Multicarrier and Direct Spread in cdma2000

Two types of data services are currently under consideration for cdma2000. These are packet data and high speed circuit switched data. Packet data will be used for asymetric bursty traffic like Internet browsing or mails. The circuit switched data can be used for delay sensitive real-time traffic. Video applications are potential candidates for circuit switch data as they need a dedicated channel for the duration of the call.

The cdma2000 will have phased development. Phase 1 of the cdma2000 effort, branded as CDMA 1x, employs 1.25 MHz of frequency bandwidth and delivers a peak data rate of 144 Kbps for stationary or mobile applications. In India some of the WiLL operators (Tata Telecom and Reliance Infocomm) are using this technology for WiLL and mobile services. Reliance Infocomm in India is also offering data services with multimedia applications. Phase 2 of cdma2000

development branded as CDMA 3x will use 5 MHz bandwidth. CDMA 3x is expected to support 144 Kbps data for mobile and vehicular applications and up to 2 Mbps data for fixed applications. The primary difference between second generation CDMA (cdmaOne or IS-95) and third generation CDMA (cdma2000) is bandwidth and peak data rate capability.

9.6.3 UMTS/WCDMA

The standards body for ETSI for 3G is called UMTS and 3GPP. Some of the CDMA encoding techniques are patented by Qualcomm. To avoid copyright issues, ETSI in Europe and ARIB in Japan have devised a different flavor of CDMA. This is branded as Wideband CDMA or WCDMA. WCDMA is also known as UTRAN (UMTS Terrestrial Radio Access Network) FDD (Frequency Division Duplex). Their responsibilities are similar with overlapping functions and responsibilities.

The physical layer of the universal mobile telecommunications system (UMTS) wideband code division multiple access (WCDMA) standard uses direct sequence spread spectrum (DSSS) modulation with a chip rate of 3.84 Mcps. The channel bandwidth is 5 MHz, this wider bandwidth has benefits such as higher data rates and improved multipath resolution. The data rates supported ranges from a few kb/s to 2 Mb/s. The physical layer supports two modes of operation: FDD (Frequency Division Duplex) and TDD (Time Division Duplex).

FDD and TDD Operational Modes

The frequency-division duplex (FDD) mode carries the uplink and downlink channels on separate frequency bands of 5 MHz each. This mode is typically used for large outdoor cells because it can support a larger number of users. The uplink and downlink transmissions in FDD mode are assigned fixed and equal frequency bands. This assignment works well when carrying voice traffic since such traffic tends to have uplink and downlink transmissions of approximately equal size. In time-division duplex (TDD) mode, the transmissions share the same frequency band by sending the uplink and downlink channels during different time slots. The TDD mode does not support as many users as the FDD mode, and hence the TDD mode is more suitable for smaller cells. In Internet the traffic pattern is asymmetric as the bandwidth requirement for download is more than the upload. Therefore, the TDD mode is more suited for carrying asymmetric data traffic like Internet. In TDD mode the uplink and downlink bandwidths can be modified by assigning more or fewer time slots to each link as and when necessary.

The Walsh codes in WCDMA are generated using the code tree as shown in Figure 9.10. If we look at these codes carefully, we will find that the WCDMA codes are same as IS-95. The spreading codes used in WCDMA are called orthogonal variable spreading factor (OVSF) codes, and the spreading factor can vary from SF=4 to SF=512. Since the 3.84 Mcps chip rate is held constant, higher data rates are obtained by using shorter spreading codes and lower data rates are obtained using longer spreading codes. Decreasing the spreading factor increases the data rate but reduces the number of users that can be supported because fewer codes are available at the shorter spreading factors.

9.6.4 Fixed Wireless

3G is commonly associated with mobile phones. However, the 3G specification includes the fixed wireless as well. Presently, we use separate links for data and voice. A fixed wireless will make it

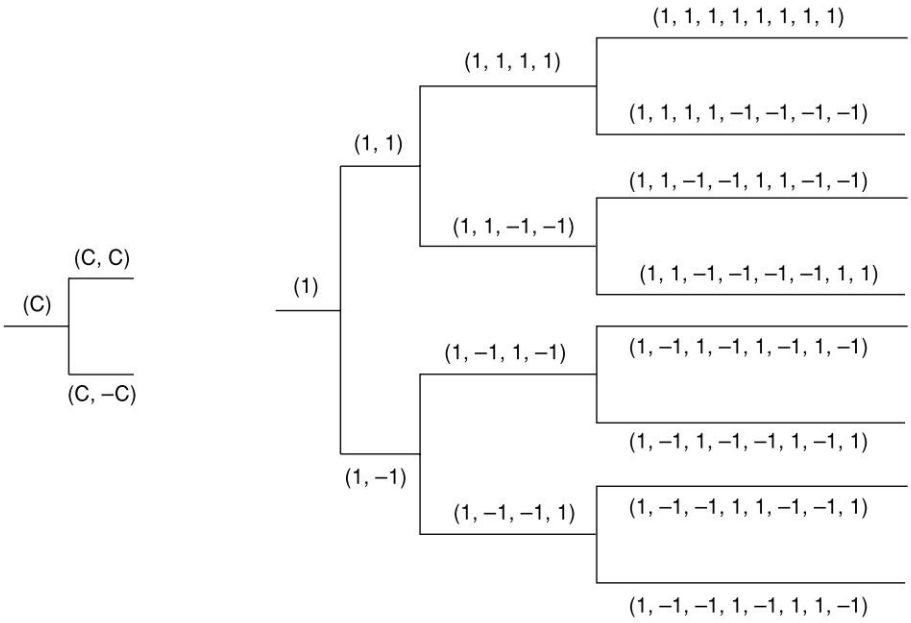


Figure 9.10 The Code Tree (Walsh Code) in WCDMA

only one common link. The IMT-2000 specification makes specific provisions for 3G Fixed Wireless Access (FWA).

In most emerging economies, and developing countries, the wired infrastructure is inadequate. Fixed Wireless Access is expected to become the mainstream technology in such geographies. In developed countries, however, 3G residential wireless represents a new horizon for competitive access providers. Users can expect a wireless connection to provide somewhere between 1.5 Mbps and 2 Mbps data at home. In India some operators are using CDMA 1X technology for Wireless in Local Loop (WiLL). They are offering both mobile and fixed phones. These fixed phone lines are examples of fixed wireless access.

Fixed wireless 3G is a converged, multimedia-driven technology. In fixed mode, 3G utilizes a point-to-multipoint network architecture that can transmit data and voice simultaneously at high speeds across core wireless infrastructure. Potential applications for 3G fixed services include business and home networking which creates a high-speed interface/gateway between an in-building "network of networks" (e.g., wireless interworking of telephony, data, video, home energy monitoring, and security networks) and the outside world (e.g., the Internet and the PSTN).

9.7 Applications on 3G

Devices in 3G can works in multiple ways. They can run in a tunneling mode or in an application mode. In tunneling mode the device works more as a pass through device or a modem. In this

mode, the mobile phone is connected to another device like a laptop and functions as a wireless media interface. The intelligence of the phone is not used, only the communication interface of the phone is used.

In an application mode, applications run on the phone itself. A 3G mobile phone will support, SMS, WAP, Java, etc. (MExE classmark 3). A MExE classmark 3 mobile device will have an execution environment that will allow application development for the client device. This application platform can be Java (through JavaPhone, PersonalJava, or J2ME, Java virtual machine), C/C++ (through Symbian, Brew or PalmOS) or Visual Basic (through Windows CE).

MExE classmark 3 devices will offer API to access device resource. These device resources will be SMS, messaging, diary, address book, etc. In future, network related information will also be available to the MExE environment through API (Application Programming Interfaces). WTAI (Wireless Telephony Application Interface) can also be used in a WAP environment to access the telephone resource.

In 3G, there will be different types of client applications. These are:

1. Local
2. Occasionally connected
3. Online
4. Real-time.

Games, cartoons and similar applications are examples of local applications. These applications can be downloaded over the air and used offline. In an occasionally connected computing (OCC) environment, the user will connect to the network occasionally. Downloading and uploading of emails are the best examples of OCC. Online applications will be the corporate applications. Examples of such applications will be online order booking or updating of inventory status. Real-time applications could be real-time stock updates or applications for law-enforcement agents for real-time tracking or navigational systems.

9.7.1 3G Specific Applications

There will be different types of applications in 3G networks. These will be for both fixed wireless and mobile. Different types of applications are candidates for 3G. These include

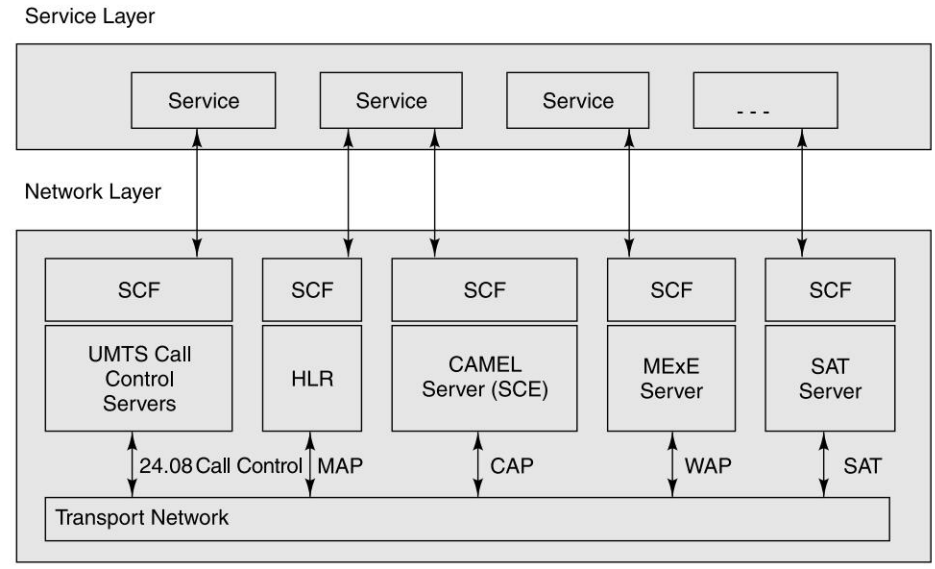
- Personal Applications.
- Content Applications.
- Communication Applications.
- Productivity Applications.
- Business Applications.

Majority of these applications were discussed in previous chapters. We will discuss some of the applications, which are new and specific to 3G.

Virtual Home Environment

Conceptually, the Virtual Home Environment can be defined as a concept where an environment is created in a foreign network (or home network outside the home environment) so that the

mobile users can experience the same computing experience as they have in their home or corporate computing environment while they are mobile and roaming. VHE is therefore aimed at roamers 3GPP defined VHE as "a system concept for personalisation service portability across network boundaries and between terminals". The aim is to enable end users to access the services of their home network/service provider even when roaming in the domain of another network provider, thus making them feel "virtually at home". VHE will allow a user to personalize the set of services for which he/she has a subscription with his/her home network and provide these home services with the user's personalized "look and feel" across different types of networks—mobile, public switched telephone network (PSTN), and Internet. VHE will offer the same user experience over varied terminals—mobile, laptop, fixed phone, PDA, PC he/she might be using. Therefore, in a VHE environment device identification, location awareness and content adaptation are going to be major challenges. An example of one of the personal service settings could be "from 9:00 am to 7:00 pm I want to be alerted for incoming messages from my boss." The VHE will automatically adapt the type of messaging the user is using at that time. If the user is using WAP terminal but not roaming in a network that supports WAP, the VHE will convert the message into another format (may be SMS). Figure 9.11 depicts the VHE architecture.



CAMEL	Customized Application for Mobile Enhancement Logic	MExE	Mobile Execution Environment
CAP	CAMEL Application Part	SAT	SIM Application Toolkit
HLR	Home Location Register	SCF	Service Capability Feature
MAP	Mobile Application Part	UMTS	Universal Mobile Telecom System
		WAP	Wireless Application Protocol

Figure 9.11 The Virtual Home Environment Architecture Over Open Service Architecture

The VHE specification introduces some new concepts related to open service architecture (OSA). This is an open interface between network layer and the service layer. There will be service

capability servers (SCS), which will provide functionality to create new services in an object-oriented fashion. If we take the MSC as an example of an SCS, call control will be a class consisting of several call control related functions. For example, "create a new call leg" or "If time is between 9:00 am to 6:00 pm connect call leg A to call leg B and charge B, else connect call leg A to call leg C and charge B". To establish a call from point X to point Y, the telecom network has to establish many connections between networks; these are called legs. The classes of OSA are called service capability features (SCF). Examples of SCFs are call control, location/positioning information and notification.

As identified by 3GPP VHE specification, the SCSs and their roles in service provisioning are:

- *UMTS call control Servers*: As SCS servers they offer mechanisms for applications to access basic bearer and call control capabilities.
- *Home Location Register (HLR) Servers*: The HLR is an intelligent database that contains location and subscriber information including the tariff and service provisioning details. The MAP (Mobile Application Part) protocol allows the exchange of location and subscriber information between different networks and services.
- *Mobile Execution Environment (MExE) Servers*: These servers will service MExE services with Java, WAP and WTAI.
- *SIM Application Toolkit*: Applications based on Smart card technologies. This will be STA (SIM Toolkit Application), Java card, or USIM applications.
- *Customized Application for Mobile Networks Enhancements Logic (CAMEL) Servers*: Camel extends the scope of IN (Intelligent Networks) service provisioning to the mobile environment. International roaming on prepaid cards is implemented using CAMEL.

Personal Communication Networks

Personal Communication Networks (PCN) are digital telephone networking infrastructures, which supports personal numbering, individual service selection, moves towards unified billing and call anytime, anywhere through wireless digital telephony. Personal communication networks are centered around the twin concepts of deploying extensive digital cellular networks and the notion of a unique identification number called a Universal Personal Telecommunication (UPT) number.

At the core of the PCN concept is the idea that each subscriber is assigned a personal identification number. This number identifies the subscriber to the network and enables them to receive or initiate phone calls, regardless of their respective location. All universal personal telecommunication numbers are held in a networked database. As a PCN mobile phone moves from one micro-cell to another it uses the signalling network to notify the network that its location has changed. Alternately, as a call for a given number enters a switching exchange, the exchange triggers a signalling system request across the network to look up in the database how to handle the call. The database enquiry returns information on how the call should be handled. Let us assume that Mr. A's mobile telephone number at Delhi is +91-9811083712. He goes to London and rents a mobile phone for a week with

number +44-77896-56872. PCN will offer a facility by which when someone calls Mr. A's Delhi number, the call will automatically be routed to the mobile phone in London.

USIM

USIM (Universal Subscriber Identity Module) is the smart card for third generation mobile phones. A SIM card in the mobile phone offers portability, security and individuality. Some standards such as the Personal Digital Cellular (PDC) standard in Japan that NTT DoCoMo uses for i-mode previously never had SIM cards. However, they will now have SIM cards with their 3G offerings. A SIM card helps to makes a device independent from the network. The USIM is the next generation of smart card based subscriber identity module (SIM). SIM card was initially designed as a simple security device for subscriber/network authentication including the ability to roam across networks. SIM cards soon became a platform for storing SMS, user's phonebook and preferences. The USIM smart card will continue to perform basic subscriber/network authentication functions but in a more flexible way. For example, it will employ contextual mechanisms that are dependent on the type of network detected.

The USIM will also provide enhanced personalization in the form of comprehensive phonebooks. These are similar to palmtop organizers and include e-mail and Web addresses alongside phone numbers. In addition, more sophisticated USIMs with high performance processors and cryptography capabilities are likely to be available by the time 3G networks start to roll out.

The USIM has the following features:

- 64 Kbytes memory.
- Card operating system based on either Java or MULTOS (a popular smart card OS standard).
- Backwards compatibility with GSM. USIMs will not work in GSM phones, but existing GSM SIM cards will work in 3G/UMTS devices.
- A number of security features from PKI (Public Key Infrastructure) to WIM (Wireless Information Module) to security algorithms will be incorporated into different vendor's USIMs.
- The 3GPP is committed to open interfaces for USIM cards with defined Application Programming Interfaces (APIs) making it possible for application developers and network operators to develop new services.

Audio/Video

Audio or video over the Internet will be either downloaded or streamed. In a downloaded environment the content is transferred, stored and played offline (local application). In a streamed environment the content is played as it is being downloaded, often in a burst, but not stored (online application). Downloaded content is generally of better quality. Audio/video contents are transferred using various different compression algorithms such as those from Microsoft or Real Networks or the MPEG-1 Audio Layer 3 (better known as MP3) protocol. MP3 is a compression/decompression algorithm. MP3 was invented in 1987 in Germany and approved by the Moving Pictures Experts Group, in 1992. With 3G, MP3 files will be downloadable over the air directly to the phone.

Third generation applications will be used to download music, multimedia, news, etc. In India, Reliance Infocomm is offering services where consumers can download popular Hindi, Tamil, Kanada, Bengali and other regional language songs and plays. Many of these also include video clips from the film. One can also download news clips from popular TV channels like Star or CNN.

Voice over Internet Protocol (VoIP/Voice over Packet Network)

Another audio application for 3G is Voice over IP (VoIP). In 3G, VoIP is a data application where normal voice calls will use Internet or other packet networks.

Electronic Agents

Electronic agents will play an important role in the future. Electronic agents will be dispatched to carry out searches and tasks on the Internet and report back to their owners. This is an efficient way to get things done on the move. In fact, one would be able to bid for items one may have wanted. The mobile agent will find out when, where and how the auction is proceeding.

Electronic agents are defined as "mobile programs that go places in the network to carry out their owners' instructions. They can be thought of as extensions of the people who dispatch them." Agents are "self-contained programs that roam communication networks delivering and receiving messages or looking for information or services." One example of agents could be in a manufacturing industry where an agent will move from one vendor's system to another and finally make the bill of material ordered in hours as opposed to weeks. This will help implement the just in time manufacturing system.

Downloading of Software and Content

As we move into the future, more and more content will be digital. Today, software is increasingly downloaded electronically from the Internet rather than purchased as boxed products in stores. In future people will be able to borrow a book from a digital library sitting at home.

ENUM

ENUM is a protocol that is emerging from work of Internet Engineering Task Force's (IETF's) Telephone Number Mapping working group. The charter of this working group is to define a Domain Name System (DNS)-based architecture and protocols for mapping a standard telephone number to a Uniform Resource Identifier (URI). This URI can be used to contact a resource associated with that telephone number. The protocol is defined in RFC 2916 "E.164 number and DNS". E.164 defines the syntax for the international public telecommunication telephony numbering plan and URI defines the syntax for Uniform Resource Identifiers (URIs defined in RFC 2396).

Using as an example the 10 digit phone number (and country code) +1-440-951-7997, the ENUM process for converting this phone number into a DNS address is as follows:

1. Remove all characters, save the +, to read: +14409517997.
2. All characters are removed and dots are placed between these digits: 1.4.4.0.9.5.1.7.9.9.7 (in DNS terms, each digit between the dots can then become a defined and distributed zone. For this example, delegation to North America at the country code zone designation of '1'. The same can be accomplished at the area code zone.
3. The order of the digits is reversed: 7.9.9.7.1.5.9.0.4.4.1.
4. The ENUM domain e164.arpa is put at the end: 7.9.9.7.1.5.9.0.4.4.1.e164.arpa.

REFERENCES/FURTHER READING

1. Andersson Christoffer (2001), *GPRS and 3G Wireless Applications*, John Wiley & Sons.
2. Dornan Andy, CDMA and 3G Cellular Networks, *Network Magazine*, http://www.networkmagazine.com/article/NMG20000831S0006.
3. 'Enabling UMTS Third Generation Services and Applications', *UMTS Forum Report* # 11, October 2000.
4. ETSI ETS 300 779, Network Aspects (NA); Universal Personal Telecommunication (UPT); Phase 1—Service description, 1997.
5. *Fundamentals of Wireless Communications & CDMA Qualcomm*, Student Guide CDMA-050 80-13127-1 X6, January 24, 2000.
6. Garg Vijay K. (2003), *IS-95, CDMA and cdma2000*, Pearson Education.
7. Lamarr Hedy: http://www.inventions.org/culture/female/lamarr.html.
8. *Loading Java into USIM*: http://forum.java.sun.com/thread.jspa?threadID-=611101&messageID=3360952.
9. Meel Ir. J. (1999), 'Spread Spectrum' *Sirius Communications*, October.
10. Muratone Flavio (Editor) (2000), *UMTS Mobile Communications for the Future*, John Wiley & Sons.
11. Tanenbaum Andrew S. (1999), *Computer Networks*, Prentice-Hall of India.
12. The Future Mobile Market Global trends and developments with a focus on Western Europe UMTS Forum Report # 8, March 1999.
13. www.itu.int
14. www.3gpp.org
15. www.3gpp2.org
16. www.umts-forum.org
17. www.ietf.org
18. www.openmobilealliance.org
19. www.mitf.org
20. www.etsi.org
21. www.scribd.com
22. www.gsmworld.com
23. www.wikipedia.org
24. 1xEV: 1x EVolution IS-856 TIA/EIA Standard, Airlink Overview, QUALCOMM, November 2001.
25. 3GPP TR 22.970: 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects Service aspects; Virtual Home Environment (VHE), 1999.

REVIEW QUESTIONS

Q1: What is Direct Sequence Spread Spectrum Technology? How does it works in CDMA technology?

Q2: Describe the IS-95 architecture. Compare its architecture with the GSM architecture.

Q3: Describe CDMA data protocol stack.

Q4: Describe different types of handoffs. What are the differences between Hard handoff, Soft handoff and Softer handoff?

Q5: Describe each of the following in brief:

(a) IS-95 Channel Structure (b) IS-95 Call Processing

(c) IS-95 Channel Capacity

Q6: Give six functional differences between CDMA and GSM.

Q7: Describe 3G networks. How is a 3G network different from a 2G network?

Q8: Describe Virtual Home Environment (VHE). How is VHE realized in 3G networks?

Q9: Describe each of the following:

(a) UMTS (b) USIM (c) ENUM

Q10: What are IMT-2000 set of standards? Explain their evolution from 2G networks.

Q11: What are the characteristic features of IMT-2000?

Q12: How is IMT-2000 set of standards expected to evolve towards 4G? What all would be the necessary conditions for that?

CHAPTER 10

Wireless LAN

10.1 INTRODUCTION

Wireless Local Area Network (LAN) is a local area data network without wires. Wireless LAN is also known as WLAN in short. Mobile users can access information and network resources through wireless LAN as they attend meetings, collaborate with other users, or move to other locations in the premises. Wireless LAN is not a replacement for the wired infrastructure. It is implemented as an extension to a wired LAN within a building or campus.

10.2 WIRELESS LAN ADVANTAGES

Schools, campuses, manufacturing plants, hospitals and enterprises install wireless LAN systems for many reasons. Some of these are:

- **Mobility:** Productivity increases when people have access to data and information from any location. The decision-making capability based on real-time information can significantly improve work efficiency. Wireless LAN offers wire-free access to information within the operating range of the WLAN.
- **Low Implementation Costs:** WLANs are easy to set up, relocate, change and manage. Networks that frequently change, both physically and logically, can benefit from WLAN's ease of implementation. WLANs can operate in locations where installation of wiring may be impractical.
- **Installation Speed and Simplicity:** Installing a wireless LAN system can be fast and easy and can eliminate the need to install cable through walls and ceilings.
- **Network Expansion:** Wireless technology allows the network to reach where wires cannot.
- **Reduced Cost-of-Ownership:** While presently the initial investment required for Wireless LAN hardware is higher than the cost of wired LAN hardware, overall installation expenses

and life-cycle costs are expected to be significantly lower. Long-term cost benefits are the greatest in dynamic environments requiring frequent moves, adds and changes.

- **Higher User to Install Base Ratio:** Wireless environment offers a higher user to capacity ratio. For example in a wired network like telephone, physical wire needs to be laid for each and every subscriber. Whereas, for a cellular network the ratio between subscribers and available channel is from 10 to 25 or even more. This means that if there is capacity for 100 channels, the network operator can safely have 2500 subscribers. Likewise in a wireless LAN, the network can offer a very high level of return on investment.
- **Reliability:** One of the common causes of failure in wired network is downtime due to cable fault. WLAN is resistant to different types of cable failures.
- **Scalability:** Wireless LANs can be configured in a variety of topologies to meet the needs of specific applications and installations. Configurations are easily changeable and range from peer-to-peer networks suitable for a small number of users to full infrastructure networks of thousands of users that allow roaming over a broad area.
- **Usage of ISM band:** Wireless LAN operates in the unregulated **ISM** (Industrial Scientific and Medical) band (2.40 GHz to 2.484 GHz, 5.725 GHz to 5.850 GHz) available for use by anyone. A user need not go to the government to get a license to use the wireless LAN. In India 2.4 GHz band is made free for use in WLANs. The 5.7 GHz band is not yet unregulated as it may conflict with the C-band of satellite.

Wireless LAN is also commercially known as WiFi or Wi-Fi. Wi-Fi is an acronym for Wireless Fidelity. The Wi-Fi™ logo is a registered trademark of the Wireless Ethernet Compatibility Alliance (http://wi-fi.org), a group founded by many companies that develop 802.11 based products.

10.2.1 Wireless LAN Evolution

Wireless LAN development started in an unstructured way. Some vendors started offering wireless communication between the corporate LAN and mobile devices (like laptop computers). This is like using a wireless keyboard or a wireless mouse. Protocols and interfaces were proprietary. However, within a short period of time many vendors started offering products in this space. These products were incompatible and soon interoperability became an issue. As IEEE is responsible for maintaining Ethernet LAN standards, IEEE assumed the responsibility of defining the wireless Ethernet LAN standards. The initial standard was published in June 1997. All these early 802.11 systems are first generation systems.

It was not until the introduction of the 11-Mbps 802.11b standard in September 1999 that the horizontal WLAN market achieved some semblance of legitimacy. Also, standards like 802.11a and 802.11g offered much higher bandwidth. All these are second generation WLANs. Second generation WLANs extended the security through 802.1x specifications and offered horizontal roaming. In horizontal roaming, a user can move from one AP to another AP seamlessly.

In third generation WLANs, vertical roaming will be possible. Vertical roaming will provide seamless roaming between different networks. Third generation WLANs will integrate with third generation (3G) telecom networks. These WLANs will eliminate the boundaries between enterprise LAN (both wireline and wireless) systems and the public wireless systems for seamless roaming. It

will extend the application of IP mobility standards. The security system is also being extended. These will be achieved through standards like 802.11f and 802.11i.

10.2.2 Wireless LAN Applications

There are many areas and applications of wireless LAN. Wireless LAN is best suited for dynamic environments. Following are some of the examples.

Office/Campus Environment

WLAN is very useful in office environments and buildings with a big campus. In big buildings or in campuses people move between floors, rooms, indoors and outdoors. In an office environment, a person can move with his laptop to the meeting room and continue working. In a university campus, a student can move from the library to the cafeteria and continue working. In a hotel, a guest can move to the pool and work. In a hospital, a doctor can carry the patient information with him while on a regular round.

Factory Shop Floor

This includes environments like factory shop floor, warehouse, exhibition sites, retail shops, labs, etc. These are very dynamic environments, where floor layouts change very frequently; objects within the building are constantly moving. Laying cables and setting up a wired LAN in these kinds of facilities are almost impossible. Wireless LAN can be very useful in such situations.

Homes

In homes WLAN can be used for convergence applications. These will include networking of different home devices like phones, computers and appliances.

Workgroup Environment

WLAN can be very useful for any set-up where small workgroups or teams need to work together, be it within a building or in the neighborhood. This may include a survey team on top of a hill or rescue members after a natural disaster or an accident site. WLAN can be very useful in civil construction sites as well.

Heritage Buildings

There are many buildings of national heritage, where a data network needs to be set up. In a very old church for example, if we need to setup a virtual reality show, it is difficult to install a wired LAN. Wireless LAN can solve the problem.

Public Places

This includes airports, railway stations or places where many people assemble and need to access information.

War/Defense Sites

When there is a war or war game, access to networks help. There is some major research going on in the US on mobile ad hoc networks for defense establishments.

10.3 IEEE 802.11 STANDARDS

The IEEE 802 committee was set up in February 1980 (that is the origin of the name) to set the standard for local area networks. From time to time, IEEE came up with different standards in the LAN domain. This includes all the layers from physical, media access, and data link layer. When IEEE deliberated the standards for WLAN, it was clear that wireless LAN will be different only at the physical and media access layer.

There were many WLAN technologies developed by researchers and industry driven by different motivations; some of them were even standardized (Table 10.1). However, WiFi or IEEE 802.11 became the most popular WLAN protocol world over. When we refer to 802.11 or IEEE 802.11, we generally mean the generic IEEE 802.11 WLAN family of standards. The 802.11 standardization originally published in 1997 with the goal to support 1 Mbit/s and 2 Mbit/s, 2.4 GHz RF and infrared transmission. All others standards released following that were amendments to this original standard with almost all the letters from the English alphabet starting from 'a' to 'z' like IEEE 802.11a, IEEE 802.11b or IEEE 802.11z. Different standards covered different aspects of WLAN like bandwidths, modulation techniques, physical media, security, roaming etc. Table 10.2 is a list of these standards.

Table 10.1 The IEEE Wireless LAN Standards

Standard	*Description*	*Publication*
IEEE 802.11	Standard for Wireless LAN operations at data rates up to 2 Mbps in the 2.4-GHz Industrial, Scientific and Medical (ISM) band.	1997
IEEE 802.15.1	Wireless Personal Area Network standard based on the Bluetooth specification, operating at the 2.4-GHz ISM band.	2002
IEEE802.1x	Port-based network access control defines infrastructures in order to provide a means of authenticating and authorizing devices attached to a LAN port that has point-to-point connection characteristics.	2001

To avoid confusion, on June 12, 2007 IEEE published the consolidated IEEE Std 802.11-2007 standard entitled "IEEE Standard for Information Technology–Telecommunications and information exchange between systems–Local and metropolitan area networks–Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications". This standard gives users, in one document, the entire IEEE 802.11 set of specifications for wireless local area networks with many amendments that have been published till 2007. This standard includes all amendments of a, b, d, e, g, h, i and j. Theoretically, today standards like IEEE 802.11a or IEEE 802.11g do not exist. The next consolidated standard is expected to be released in 2011 when standards like IEEE 802.11k, IEEE 802.11y and many other may be merged into IEEE Std 802.11-2011.

Table 10.2 The IEEE 802.11 Wireless LAN standards

Standard	*Description*	*Publication*
IEEE 802.11a	54 Mbit/s, 5 GHz standard	1999
IEEE 802.11b	Enhancements to 802.11 to support 5.5 and 11 Mbit/s	1999
IEEE 802.11c	Bridge operation procedures; included in the IEEE 802.1D standard	2001
IEEE 802.11d	International (country-to-country) roaming extensions	2001
IEEE 802.11e	Enhancements QoS including packet bursting	2005
IEEE 802.11f	Inter-Access Point Protocol	2003
IEEE 802.11g	54 Mbit/s, 2.4 GHz standard (backwards compatible with b)	2003
IEEE 802.11h	Spectrum Managed 802.11a (5 GHz) for European compatibility	2004
IEEE 802.11i	Enhanced security	2004
IEEE 802.11j	Extensions for Japan	2004
IEEE 802.11k	Radio resource measurement enhancements	2008
IEEE 802.11n	Higher throughput improvements using Multiple Input, Multiple Output (MIMO) antennas	2009 (target)
IEEE 802.11p	WAVE–Wireless Access for the Vehicular Environment	2010 (target)
IEEE 802.11r	Fast roaming	2008
IEEE 802.11s	Mesh Networking, Extended Service Set (ESS)	2010 (target)
IEEE 802.11t	Wireless Performance Prediction (WPP)–test methods and metrics recommendation	
IEEE 802.11u	Interworking with non-802 networks (for example, cellular)	2010 (target)
IEEE 802.11v	Wireless network management	2010 (target)
IEEE 802.11w	Protected Management Frames	2009 (target)
IEEE 802.11y	3650–3700 MHz Operation in the U.S.	2008
IEEE 802.11z	Extensions to Direct Link Setup (DLS)	2011 (target)
IEEE 802.11aa	Robust streaming of Audio Video Transport Streams	2011 (target)
IEEE 802.11mb	Maintenance of the standard	2011 (target)
IEEE 802.11ac	Very High Throughput < 6 GHz; potential improvements over 802.11n:	2012 (target)
IEEE 802.11ad	Very High Throughput 60 GHz	2012 (target)

10.4 WIRELESS LAN ARCHITECTURE

10.4.1 Types of Wireless LAN

There are different types and flavors of wireless local area networks. Some of the most popular ones are:

- **802.11:** In June 1997, the IEEE finalized the initial specification for wireless LANs: IEEE 802.11. This standard specifies a 2.4 GHz frequency band with data rate of 1 Mbps and 2 Mbps. This standard evolved into many variations of the specification like 802.11b, 802.11a, 802.11g, etc., using different encoding technologies. Today these standards offer a local area network of bandwidths going up to a maximum of 54Mbps.
- **HyperLAN:** HyperLan began in Europe as a specification (EN 300 652) ratified in 1996 by the ETSI Broadband Radio Access Network group. HyperLAN/1, the current version works at the 5 GHz band and offers up to 24 Mbps bandwidth. Next version HyperLAN/2 (http://www.hyperlan2.com) will support a bandwidth of 54 Mbps with QoS support. This will be able to carry Ethernet frames, ATM cells, IP packets and support data, video, voice and image.
- **HomeRF:** In 1998, the HomeRF Working Group (http://www.homerf.org) offered to provide an industry specification to offer Shared Wireless Access Protocol (SWAP). This standard will offer interoperability between PC and consumer electronic devices within the home. SWAP uses frequency hopping spread spectrum modulation and offers 1 Mbps and 2 Mbps at 2.4 GHz frequency band.
- **Bluetooth:** Bluetooth was promoted by big industry leaders like IBM, Ericsson, Intel, Lucent, 3Com, Microsoft, Nokia, Motorola, and Toshiba. It was named after Danish king Harold Bluetooth during 952 to 995 A.D., who had a vision of a world with cooperation and interoperability. Bluetooth is more of a wireless Personal Area Network (PAN) operating at 2.4 GHz band and offers a peak 1Mbps data rate. Bluetooth uses frequency hopping spread-spectrum modulation with relatively low power and smaller range.
- **MANET:** Manet (http://www.ietf.org/html.charters/manet-charter.html) is a working group within the IETF to investigate and develop the standard for Mobile ad hoc NETworks.

10.4.2 Ad hoc versus Infrastructure Mode

Wireless Networks are of two types, infrastructure mode and ad hoc mode. In an infrastructure mode, the mobile stations (MS) are connected to a base station or Access Point (AP). This is similar to a star network where all the mobile stations are attached to the base station. Through a protocol the base station manages the dialogues between the AP and the MS. Figure 10.1 depicts a wireless LAN in infrastructure mode.

In an ad hoc mode, there is no access point or infrastructure. A number of mobile stations form a cluster to communicate with each other. Figure 10.2 depicts wireless LAN in adhoc mode.

In an Infrastructure mode, 802.11 LAN is based on a cellular architecture where the system is subdivided into small clusters or cells (see Fig. 10.1). Each cell is called Basic Service Set, or BSS. Depending on the topology one BSS is connected to other BSS or other infrastructure. In an ad hoc network, the BSS is completely independent. Therefore, technically an ad hoc network is

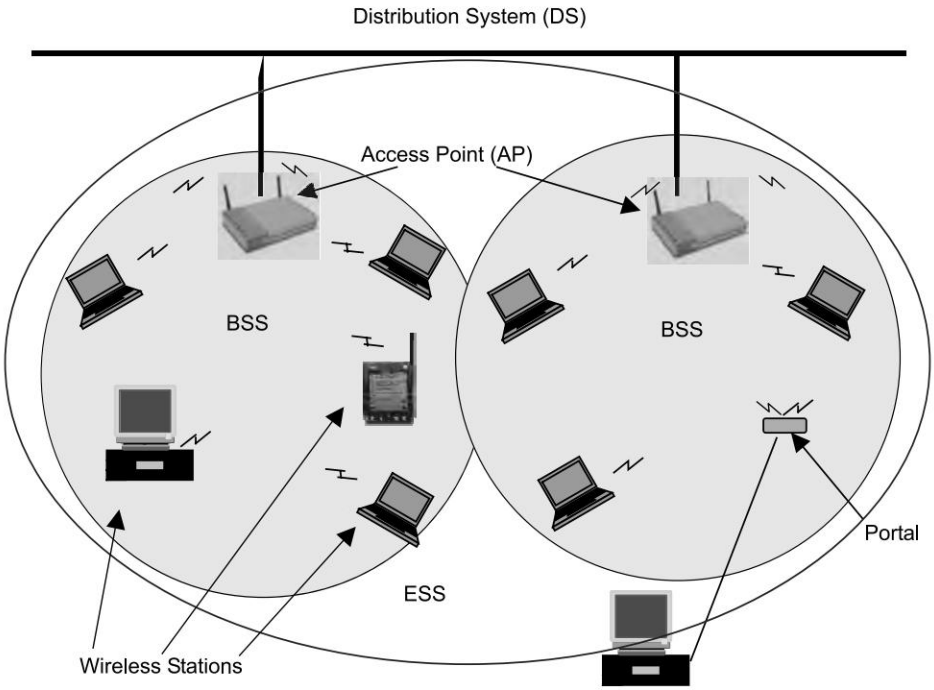


Figure 10.1 Wireless LAN in Infrastructure Mode

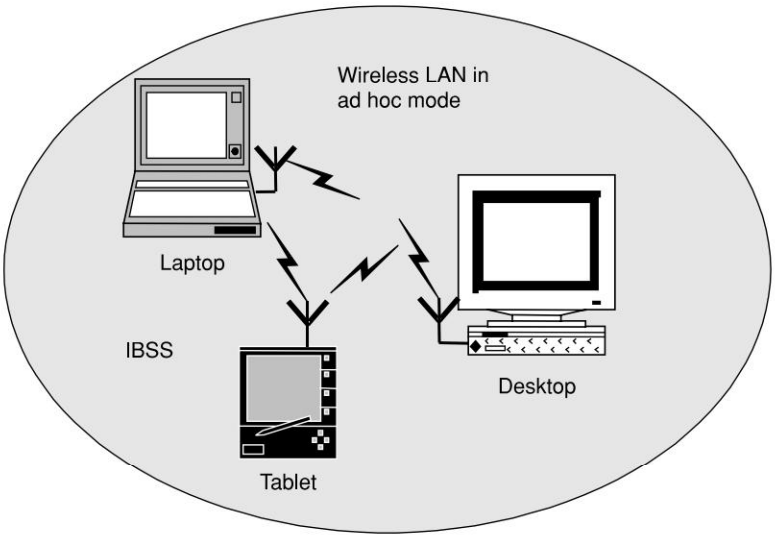


Figure 10.2 Wireless LAN in Ad hoc Mode

termed as Independent BSS or IBSS. Whereas, in infrastructure mode the mobile stations form a cluster with an AP. Multiple such BSS form an Extended Service SET (ESS). The ESS is connected to the backbone LAN or the distribution system.

10.4.3 802.11 Architecture

In the 802.11 nomenclatures one cell or a BSS is controlled by one Base Station. This base station is called Access Point or AP in short. In some literature, access points are referred to as Hot Spots.

Although a wireless LAN may be formed by a single cell, with a single Access Point, most installations will be formed by several cells, where the access points are connected through some kind of backbone. This backbone is called Distribution System or DS (Fig. 10.1). This backbone is typically Ethernet and, in some cases, wireless itself (see Fig. 10.3).

The whole interconnected Wireless LAN, including the different cells, their respective Access Points and the Distribution System, is seen as a single 802 network to the upper layers of the OSI model and is called Extended Service Set (ESS).

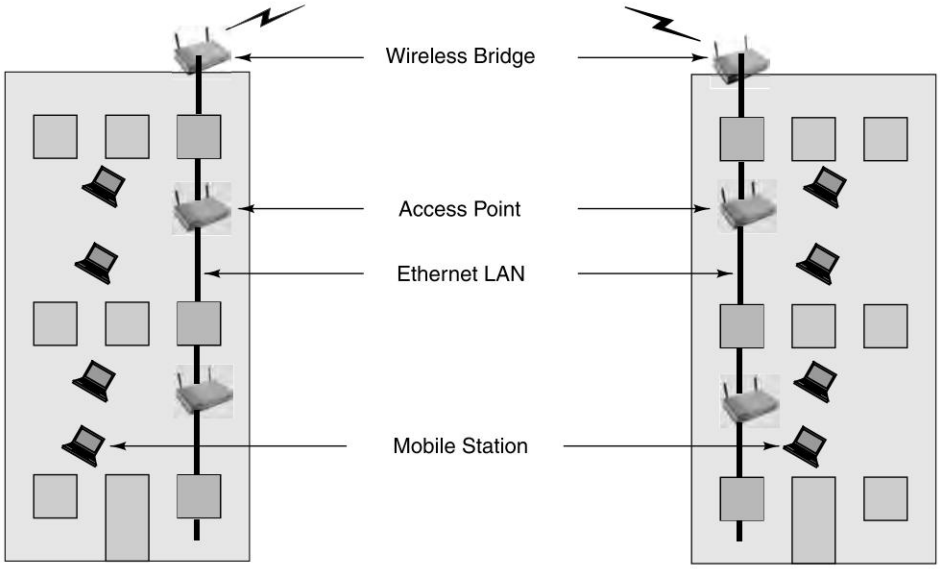


Figure 10.3 Two Access Points (Wireless Bridges) as Part of the Distribution System

The 802.11 standard also defines the concept of a "portal" (see Fig. 10.1). A portal is a device that interconnects between an 802.11 and another 802 LAN. This concept is an abstract description of part of the functionality of a bridge.

Cell Design in Wireless LAN

For proper functioning of wireless LAN, neighboring cells (BSS) are set up on different frequencies, so that wireless LAN cards in each cell do not interfere with one another when they transmit signals. In order for these cells to work without interference, the DSSS standards define 13 different

frequencies or channels (Table 10.3). For Frequency Hopping Spread Spectrum (FHSS) there are 79 channels. These frequencies are typically "non-overlapping". This means that they operate in different sections of the radio spectrum or band.

In a design where there are many cells, effective use of these non-overlapping frequencies are very important.

In the following design (see Fig. 10.4) there are four wireless cells (cells A through D). A cell is defined by the space and area the radio wave of a wireless LAN access point is able to cover. Cells A, B and C all use non-overlapping frequencies, while cell D uses the same frequency as cell A. The use of two frequencies by two cells will not have any effect on each other, as long as distance "x" is great enough to ensure effective radio isolation from each other. Radio isolation or radio separation means that a device in cell A will not be able to detect the signal transmitted by any device in cell D. This is because the air distance between two cells attenuates or weakens the radio signal so much that they cannot detect the radio signal of each other.

Table 10.3 Channels within the 2.4GHz band

Channel No	*Frequency (GHz)*
1	2.412
2	2.417
3	2.422
4	2.427
5	2.432
6	2.437
7	2.442
8	2.447
9	2.452
10	2.457
11	2.462
12	2.467
13	2.472

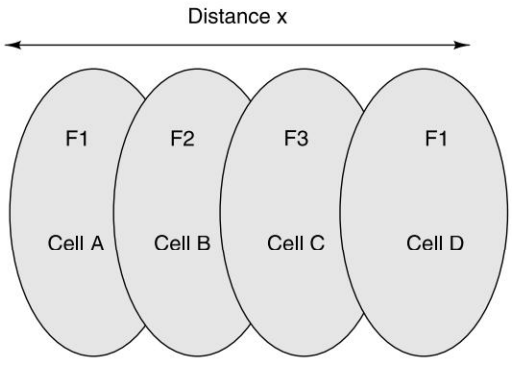


Figure 10.4 Cell Design in a WLAN

IEEE 802.11 Layers Description

The 802.11 standards cover definitions for both MAC (Medium Access Control) and Physical Layer. The standard currently defines a single MAC, which interacts with three PHYs (Fig. 10.5) as follows:

- Frequency Hopping Spread Spectrum
- Direct Sequence Spread Spectrum, and
- InfraRed.

Beyond the standard functionality usually performed by media access layers, the 802.11 MAC performs other functions that are typically done by upper layer protocols, such as fragmentation, packet retransmissions, and acknowledgements.

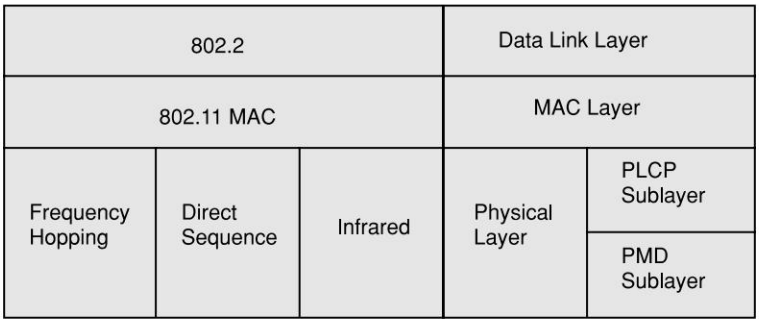


Figure 10.5 The 802.11 Stack

Physical Layer (Layer 1) Architecture

The architecture of the physical layer (see Fig. 10.5) comprises the two sublayers for each station:

- **PLCP (Physical Layer Convergence Procedure):** PLCP sublayer is responsible for the Carrier Sense (CS) part of the Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) protocol. PLCP layer prepares the MAC Protocol Data Unit (MPDU) for transmission. The PLCP also delivers the incoming frames from the wireless medium to the MAC layer. PLCP appends fields to the MPDU that contains information needed by the physical layer transmitter and receiver. This frame is called PLCP Protocol Data Unit (PPDU). The structure of PLCP provides for asynchronous transfer of MPDU between stations. The PLCP header contains logical information that allows the receiving station's physical layer to synchronize with each individual incoming packet.
- **PMD (Physical Medium Dependent):** The PMD provides the actual transmission and reception of physical layer entities between stations through the wireless media. This sublayer provides the modulation/demodulation of the transmission.

FHSS (Frequency Hopping Spread Spectrum) Physical Layer

In FHSS mode, this layer carries the clocking information to synchronize the receiver clock with the clock of the transmitted packet. Figure 10.6 depicts the FHSS PPDU packet.

The fields in the FHSS PLCP are as follows:

1. **SYNC.** This field is made up of alternate zeroes and ones. This bit pattern is to synchronize the clock of the receiver.
2. **Start Frame Delimiter.** This field indicates the beginning of the frame and the content of this field is fixed and is always 0000110010111101.
3. **PSDU Length Word (PLW).** This field specifies the length of the PSDU in octets.
4. **PLCP Signaling (PSF).** This field contains information about the data rate of the fields from whitened PSDU. The PLCP preamble is always transmitted at 1Mbps irrespective of the data rate of the wireless LAN. This field contains information about the speed of the link. For example, 0000 means 1 Mbps and 0111 signifies 4.5 Mbps bandwidth.
5. **Header Error Check.** This field contains the CRC (Cyclic Redundancy Check) according to CCITT CRC-16 algorithm.

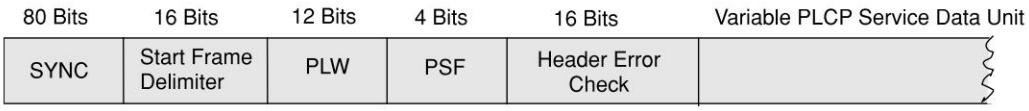


Figure 10.6 Frequency Hopping Spread Spectrum PLCP

FHSS PMD is responsible for converting the binary bit sequence into analog signal and transmit the PPDU frame into the air. FHSS PDM does this using the frequency hopping technique. The 802.11 standard defines a set of channels within the ISM band for frequency hopping. For the US and Europe there are 79 1MHz channels within 2.402 to 2.480 GHz band. The FHSS PMD transmits PPDU by hopping from channel to channel according to a particular pseudo-random hopping sequence. Once the hopping sequence is set in the access point, stations automatically synchronize to the correct hopping sequence.

Direct Sequence Spread Spectrum (DSSS) Physical Layer

DSSS PLCP is responsible for synchronizing and receiving the data bits correctly. Figure 10.7 depicts the DSSS PPDU packet.

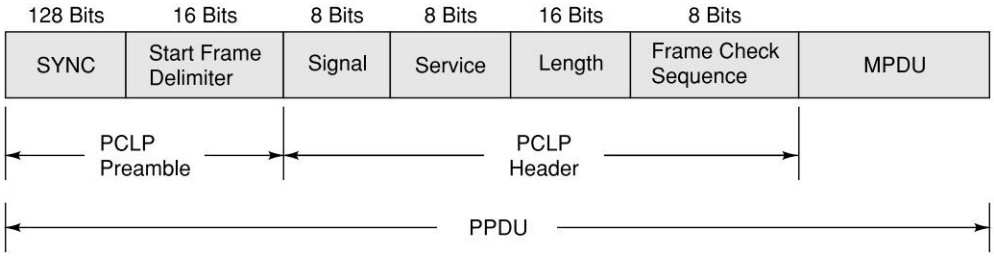


Figure 10.7 Direct Sequence Spread Spectrum PLCP Protocol Data Unit

The fields in the DSSS PLCP are as following:

1. **SYNC.** This field is made up of alternate zeroes and ones. This bit pattern is to synchronize the clock of the receiver with the received frame.

2. **Start Frame Delimiter.** This field indicates the beginning of the frame and the content of this field is fixed and is always 1111001110100000.
3. **Signal.** This field defines the type of modulation the receiver must use to demodulate the signal. When the value of this field is multiplied by 100 Kbps we get the bandwidth of the transmission. For 11 Mbps bandwidth this field will have a value of 01101110 (decimal 110). The PLCP preamble and the header are always transmitted at 1 Mbps. The bandwidth defined by this field applies to MPDU field.
4. **Service.** This field is not used and is usually 0.
5. **Length.** This field contains an unsigned 16-bit integer indicating the length of the frame. However, unlike the FHSS, this is not in octets. It is rather in microseconds. The receiver will use this to synchronize with the clock to determine the end of frame.
6. **Frame Check Sequence.** This is a 16-bit checksum based on CCITT CRC-16 algorithm.

DSSS PMD translates the binary digital sequence into analog radio signals and transmits the PPDU frame into the air. The DSSS physical layer operates within the ISM band. If we take the 2.4 GHz band, then it is between 2.4 GHz and 2.8435 GHz (802.11b and 802.11g) frequency band divided into multiple channels with 22 MHz width.

In DSSS the data is spread with a pseudo random noise (PN) code. This PN code is referred to as chip or spreading sequence. For 1 Mbps and 2 Mbps 802.11, the PN code is called the 11-bit Barker sequence. It is an 11 bit sequence of positive and negative 1s like +1, –1, +1, +1, –1, +1, +1, +1, –1, –1, –1 (Fig. 10.8). 5.5 Mbps and 11 Mbps versions of 802.11b do not use the Barker sequence. They use the Complementary Code Keying (CCK) technique instead. CCK is a set of 64 eight-bit code words used to encode data for 5.5 Mbps and 11 Mbps data rates. All these codes have unique mathematical properties that allow them to be correctly distinguished from one another by a receiver even in the presence of substantial noise and multipath interference.

Every bit in the data stream of the PPDU is modulated with this PN sequence. For example, in case of 802.11, a 1 in the data bit will be represented as +1,–1,+1,+1, –1,+1,+1,+1,–1,–1,–1; whereas a 0 in the data bit will be represented as –1,+1,–1,–1,+1, –1,–1,–1,+1,+1,+1 (Fig. 10.8).

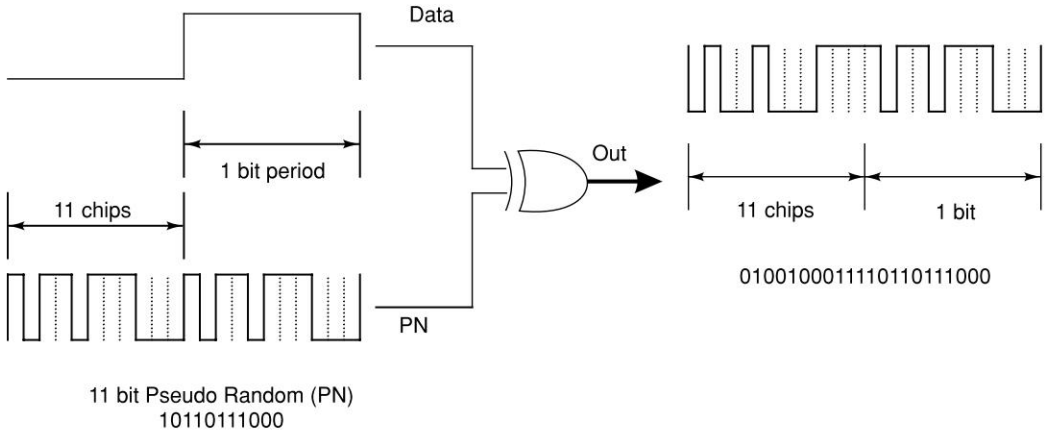


Figure 10.8 DSSS Modulator

The DSSS used in wireless LAN and the DSSS used in the CDMA (IS-94 or CDMA- 2000) for wireless MAN (Metropolitan Area Network) used in CDMA phones operate in similar fashion but with some difference. In wireless LAN the chip used for each and every mobile station is the same. However, in case of wireless MAN the chip used for each different mobile station (for uplink or reverse path) are different.

The MAC Layer (Layer 2) Architecture

The MAC Layer defines two different access methods, the Distributed Coordination Function and the Point Coordination Function:

The Basic Access Method: CSMA/CA

The basic access mechanism, called the Distributed Coordination Function by IEEE standard, is Carrier Sense Multiple Access with Collision Avoidance mechanism (CSMA/CA). CSMA protocols are well known in the industry, the most popular being the Ethernet, which is a CSMA/CD protocol (CD stands for Collision Detection). In a wired environment (Ethernet for example) every station connected to the wire can sense the signal in the wire. In a wired LAN, if there is no activity or a collision of messages, every station connected to the LAN will be able to sense the collection almost instantly. This is not true in the case of wireless media. In the case of wireless LANs, a Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) protocol is used, as it is not possible to detect a collision of data packets in mid air.

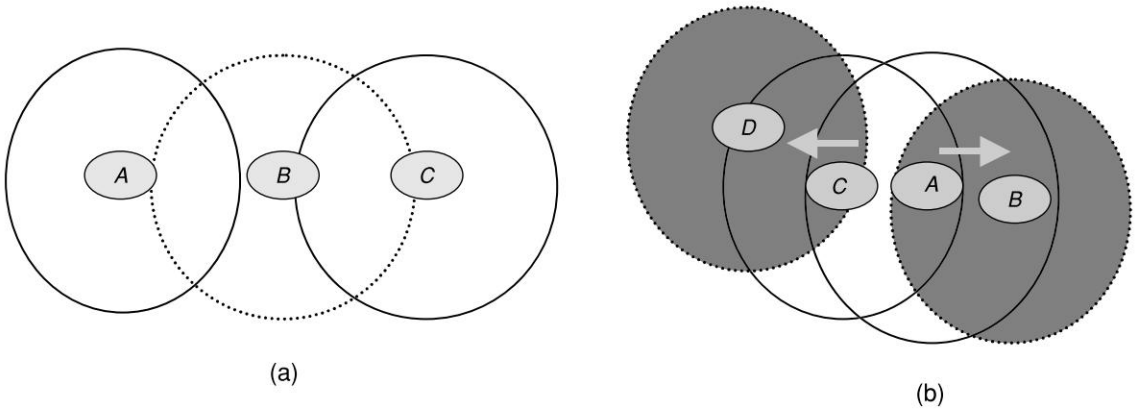


Figure 10.9 (a) Hidden Terminal; (b) Exposed Terminal

Consider the scenario with three mobile nodes as shown in Figure 10.9(a). The transmission of *A* reaches *B*, but not *C*. The transmission of *C* reaches *B*, but not *A*. However, the radio signal of *B* reaches both *A* and *C* making *A* and *C* both in the range of *B*. The net effect is *A* cannot detect *C* and vice versa.

A starts sending to *B*; *C* does not receive this transmission. *C* also wants to send to *B* and senses the medium. To *C* the medium appears to be free. Thus *C* starts sending causing collision at *B*. But now *A* cannot detect the collision and continues with its transmission. *A* is "hidden" for *C* and vice versa.

Consider another case as shown in Figure 10.9(b). The radio transmission signal of *A* reaches *C* and *B*. The radio signal of *C* reaches both *A* and *D*. A wants to communicate to *B*, *A* starts sending signals to *B*. *C* wants to communicate with *D*, *C* senses the carrier and finds that *A* is talking to *B*. *C* has to wait till the time *A* finishes with *B*. However, *D* is outside the range of *A*, therefore waiting is not necessary. In fact *A*, *B* and *C*, *D* can communicate with each other in parallel without any collision, but according to the protocol that is not possible. *A* and *C* are "exposed" terminals.

While Collision Detection mechanisms are a good idea on a wired LAN, they cannot be used on a Wireless LAN environment for two main reasons:

- Implementing a Collision Detection mechanism requires the implementation of a Full Duplex radio capable of transmitting and receiving at the same time. This increases the cost significantly.
- In a wireless environment we cannot assume that all stations will be able to receive radio signals from each other (which is the basic assumption of the Collision Detection scheme). The fact that a station wants to transmit and senses the medium as free (not able to sense signal from another station) does not necessarily mean that the medium is free (like the case of the hidden terminal) around the receiver area.

The mechanism behind CSMA/CA is as follows:

- When a wireless station (a wireless LAN device) wants to communicate, it first listens to its media (radio spectrum) to check if it can sense radio waves from any other wireless station.
- If the medium is free for a specified time then the station is allowed to transmit. This time interval is called Distributed Inter Frame Space (DIFS).
- If the current device senses a carrier signal of another wireless device on the same frequency, as it wants to transmit on, it backs off (does not transmit) and initiates a random timeout.
- After the timeout has expired, the wireless station again listens to the radio spectrum and if it still senses another wireless station transmitting, it continues to initiate random timeouts until it does not detect or sense another wireless station transmitting on the same frequency.
- When it does not sense another wireless station transmitting, the current wireless station starts transmitting its own carrier signal to communicate with the other wireless station, and once synchronized, transmits the data.
- The receiving station checks the CRC of the received packet and sends an acknowledgment packet (ACK). Receipt of the acknowledgment indicates to the transmitter that no collision occurred. If the sender does not receive the acknowledgment then it retransmits the fragment until it receives acknowledgment or is abandoned after a given number of retransmissions.

It can be seen from the above that the more times a wireless station has to back off or go into a random timeout, the less opportunity it has to transmit its data. This reduced opportunity for data transmission leads to less effective access to wireless bandwidth. This reduces the speed of the operation. In a worse case scenario the system would, after a number of retries, completely timeout and the wireless connection would be lost.

Virtual Carrier Sense

In order to reduce the probability of two stations colliding because they cannot sense each other's presence, the standard defines a Virtual Carrier Sense mechanism: A station wanting to transmit a packet first transmits a short control packet called RTS (Request To Send), which includes the source, destination, and the duration of the following transaction (the data packet and the respective

ACK). The destination station after receiving this request packet responds with a response control packet called CTS (Clear to Send), which includes the same duration information.

All stations receiving either the RTS and/or the CTS, set their Virtual Carrier Sense indicator called Network Allocation Vector or NAV, for the given duration, and use this information together with the Physical Carrier Sense when sensing the medium. This mechanism reduces the probability of a collision on the receiver side by a station that is "hidden" from the transmitter to the short duration of the RTS transmission because the station senses the CTS and "reserves" the medium as busy until the end of the transaction. The duration information on the RTS also protects the transmitter area from collisions during the ACK (from stations that are out of range of the acknowledging station). It should also be noted that, due to the fact that the RTS and CTS are short frames, the mechanism also reduces the overhead of collisions, since these are recognized faster than if the whole packet was to be transmitted. The diagrams (Fig. 10.10) show a transaction between stations A and B, and the NAV setting of their neighbors:

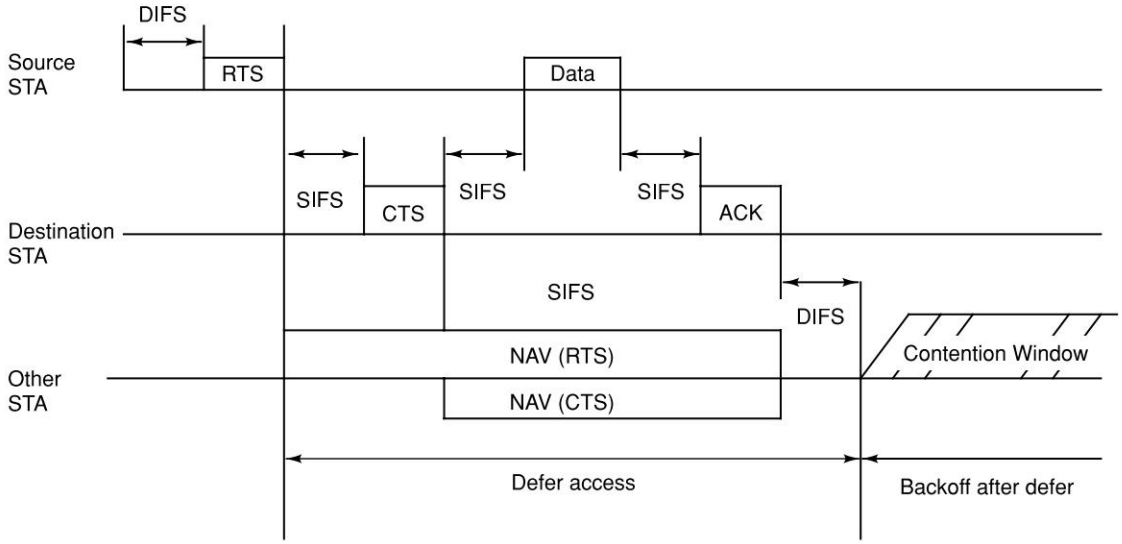


Figure 10.10 The CSMA/CA Protocol

Fragmentation and Reassembly

Typical LAN protocols use packets several hundred bytes long (the longest Ethernet packet could be up to 1518 bytes long). There are several reasons why it is preferable to use smaller packets in a Wireless LAN environment:

- Due to the higher Bit Error Rate of a radio link, the probability of a packet getting corrupted increases with the packet size.
- In case of packet corruption (either due to collision or noise), the smaller the packet, the less overhead it causes to retransmit it.
- On a Frequency Hopping system, the medium is interrupted periodically for hopping, so, the smaller the packet, smaller the chance that the transmission will be postponed after dwell time.

In a majority of cases, the wireless LAN uses standard Ethernet LAN as the backbone. Therefore, it is necessary that wireless LAN is able to handle Ethernet packets of 1518 bytes long. Also, any change in the protocol for wireless LAN may cause a major change in the protocol of the higher layers. Therefore, the IEEE committee decided to solve the problem by adding a simple fragmentation/reassembly mechanism at the MAC Layer of the wireless LAN. The mechanism is a simple Send-and-Wait algorithm, where the transmitting station is not allowed to transmit a new fragment until one of the following conditions happens:

1. Receives an ACK for the said fragment, or
2. Decides that the fragment was retransmitted too many times and drops the whole frame.

It should be noted that the standard does allow the station to transmit to a different address between retransmissions of a given fragment. This is particularly useful when an AP has several outstanding packets to different destinations and one of them does not respond. Figure 10.11 shows a frame (MSDU) being divided to several fragments (MPDUs).

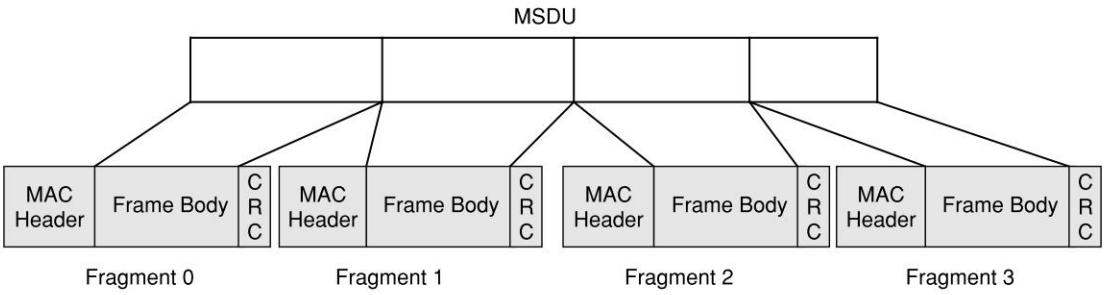


Figure 10.11 Frame Fragmentation

Inter Frame Spaces

The standard defines four types of spacing intervals. These are called Inter Frame Spaces (IFS). IFSs are used to defer a station's access to the medium and provide various levels of priorities:

- **SIFS (Short Inter Frame Space)**, is the shortest Inter Frame Space with the highest priority. RTS, CTS use SIFS intervals. SIFS value is a fixed value per PHY and is calculated in such a way that the transmitting station will be able to switch back to receive mode and be capable of decoding the incoming packet.
- **PIFS (Point Coordination IFS)**, is used by the Access Point (or Point Coordinator), to gain access to the medium before any other station. This value of PIFS is SIFS plus a Slot Time, i.e., 78 microseconds.
- **DIFS (Distributed IFS)**, is the Inter Frame Space used for a station willing to start a new transmission, which is calculated as PIFS plus one slot time, i.e., 128 microseconds.
- **EIFS (Extended IFS)**, is a longer IFS used by a station that has received a packet that it could not understand. This is needed to prevent the station (which could not understand the duration information for the Virtual Carrier Sense) from colliding with a future packet belonging to the current dialog.

Maintaining Synchronized Stations

Stations need to maintain synchronization. This is necessary to keep hopping and other functions like Power Saving synchronized. On an infrastructure BSS, synchronization is achieved by all the stations updating their clocks according to the AP's clock. The AP periodically transmits frames called Beacon Frames. These frames contain the value of the AP's clock at the moment of transmission. This is the time when physical transmission actually happens, and not when the packet was put in the queue for transmission.

The receiving stations check the value of their clocks the moment the signal is received, and correct it to be synchronized with the AP's clock. This prevents clock drifting which could cause loss of synchronization after a few hours of operation.

Power Saving

Wireless LANs are typically related to mobile applications. In this type of application, battery power is a scare resource. That is why 802.11 standard directly addresses the issue of power saving. Power saving enables stations to go into sleep mode without losing information. The AP maintains a continually updated record of all stations currently in Power Saving mode. AP buffers the packets addressed to these stations until either the stations specifically request the packets by sending a polling request, or until the stations change their operation mode.

As part of Beacon Frames, the AP periodically transmits information about which power saving stations have frames buffered at the AP. If there is an indication that there is a frame stored at the AP waiting for delivery, then the station stays awake and sends a polling message to the AP to receive these frames.

10.5 Mobility in Wireless LAN

When a station wants to access an existing BSS (either after power-up, sleep mode, or physically entering into the BSS area), the station needs to get synchronization information from the AP (or from the other stations when in ad hoc mode).

The station can get this information by one of two means:

- **Passive Scanning.** In this case the station just waits to receive a Beacon Frame from the AP, or
- **Active Scanning.** In this case the station tries to locate an Access Point by transmitting Probe Request Frames, and waits for Probe Response from the AP.

The Authentication Process

Once a wireless station has located an AP and decides to join its BSS, it goes through the authentication process. This is interchange of authentication information between the AP and the station, where the WLAN device proves its identity.

The Association Process

Once the station is authenticated, it then starts the association process which is the exchange of information about the stations and BSS capabilities, and which allows the DSS (the set of APs) to know about the current position of the station. A station is capable of transmitting and receiving data frames only after the association process is completed.

Roaming

Roaming is the process of moving from one cell (or BSS) to another without losing connection. This function is similar to the cellular phones' handover, with two main differences:

1. On a packet-based LAN system, the transition from cell to cell may be performed between packet transmissions, as opposed to telephony where the transition may occur during a phone conversation.
2. On a voice system, a temporary disconnection during handoff does not affect the conversation. However, in a packet-based environment it significantly reduces performance because retransmission is performed by the upper layer protocols.

The 802.11 standard does not define how roaming should be performed, but defines the basic tools. These include active/passive scanning, and a re-association process, where a station that is roaming from one AP to another becomes associated with the new AP. The Inter-Access Point Protocol (IAPP) specification addresses a common roaming protocol enabling wireless stations to move across multivendor access points. IAPP is the scope of IEEE standard 802.11f.

IAPP defines two basic protocols, viz., Announce protocol and Handover protocol. The Announce protocol provides coordination information between access points. This information relates to network wide configuration information about active APs. The Handover protocol allows APs to coordinate with each other and determine the status of a station. When a station associates with a different AP, the old AP forwards buffered frames for the station to the new AP. The new AP updates the necessary tables in the MAC layer to ensure that the MAC level filtering will forward frames appropriately. This type of roaming is called horizontal roaming.

Mobile IP is another protocol that is used to allow application layer roaming. Using Mobile IP, a mobile station can move from one type of network to another type of network. For example, in an IMT-200 situation, the station moves from wireless LAN environment to a 3G wireless MAN environment. Mobile IP is described in Chapter 4.

10.6 Deploying Wireless LAN

10.6.1 Network Design

The first step in designing a wireless network is to identify the areas that need to be covered, the number of users and the types of devices they will use. From these requirements we need to determine how many access points (AP) are required and where they must be placed. The goal is to ensure adequate RF coverage to users. AP placement is typically determined using a combination of theoretical principles and a thorough site survey. Site survey is necessary to determine the required coverage; number, density, and location of APs. In office environments with walls (including cube walls) and other impediments, a typical range is 75 to 80 feet (23 to 24 meters).

In addition, the site survey can identify conditions that inhibit performance through path and multipath loss, as well as RF interference. Path loss refers to the loss of signal power experienced between the AP and the client system due to distance—walls, ceilings, and furniture—and the frequency of the transmission. Multipath loss occurs as an RF signal bounces off objects in the environment such as furniture and walls while en route to its destination. Use of APs with "antenna

diversity" help to correct multipath loss. In 802.11b networks, antenna diversity is implemented using two antennas with supporting circuitry to improve signal reception.

RF interference is caused by other RF sources that also operate in the 2.4-GHz frequency band. These sources can include microwave ovens and cordless phones. In addition, emerging Bluetooth personal area network devices operating in this frequency band can interfere with 802.11 transmissions.

Scaling Capacity and Bandwidth

Figure 10.12 shows how "aggregate bandwidth" in a localized coverage area helps to service a more dense population of wireless clients or to increase the bandwidth available to each wireless client in a coverage area. Channels 1, 6 and 11 are completely non-overlapping. To achieve 33-MHz, channels 1, 6 and 11 and can be used as overlapping channels.

In the example shown in Figure 10.12(a), one AP provides up to 11 Mbps of bandwidth, which is shared by all wireless clients in the coverage area. As shown in Figure 10.12(b), two more APs can be installed next to the original AP. Each provides an additional 11 Mbps of bandwidth to the same coverage area, for an aggregate bandwidth of up to 33 Mbps. This solution does not provide an individual wireless client with 33 Mbps of bandwidth. In 802.11 networks, each client associates with only one AP at a time and shares its bandwidth with other clients associated with the AP. Capacity and bandwidth can also be scaled by reducing the size of the coverage areas.

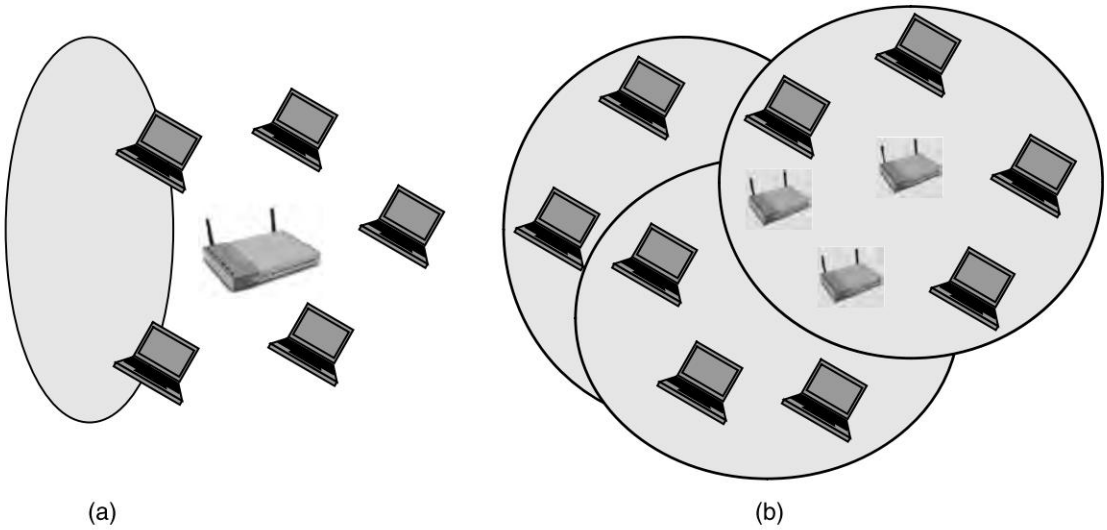


Figure 10.12 Scaling Aggregate Bandwidth (a) of 11 and (b) of 33 Mbps in a Localized Area by Co-location

Channel Selection

Within the 2.4-GHz frequency band, the 802.11 standard defines 13 "center frequency channels" (see Table 10.3). Channel 1 (2.412 GHz), channel 6 (2.437 GHz), and channel 11 (2.462 GHz) are

non-overlapping with large radio isolation bands. Therefore, these channels are commonly used to minimize the complexity of configuring and managing channels.

Figure 10.13 shows a three-storey building serviced by nine APs configured with channels 1, 6 and 11. This arrangement minimizes interference between APs located on the same floor as well as APs in the neighboring floors. It also eliminates the bandwidth contention that occurs when two APs with overlapping coverage are configured with the same channel. When this happens, 802.11 wireless Ethernet carrier sense multiple access/collision avoidance (CSMA/CA) mechanisms ensure that users in both coverage areas can access the network. However, instead of providing two separate 11-Mbps channels and an aggregate bandwidth of 22 Mbps, the two APs provide only one 11-Mbps channel.

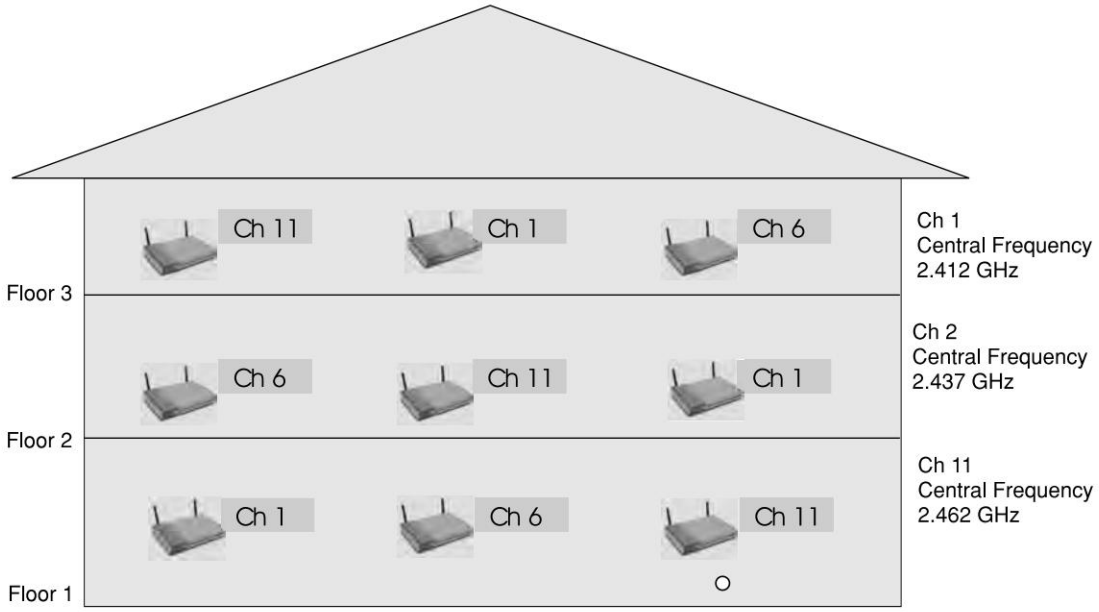


Figure 10.13 Sample Frequency Topology Using Channels 1, 6, and 11

AP Transmission Power

The transmission power of most APs ranges from 1 mw up to 100 mw. Transmission power affects the effective range of the radio signal. The higher the transmission power, the longer the range of the signal (that is, the larger the coverage area). Higher power settings are appropriate in many large enterprise installations with cube-wall offices and a lot of open space. Lower settings are appropriate in environments such as test labs or small offices where the longer range is not required. Because lowering the transmission power reduces the range of an AP, lower power settings can also enable the wireless network to provide higher aggregate throughput. At lower power settings, more APs can be installed to serve a particular area than is possible at higher power levels.

10.6.2 Configuring the Wireless LAN

Configuration of a wireless LAN includes configuration of both the access point and the mobile station. The first level of configuration is to assign an IP address to the AP. The WEP (Wired Equivalent Privacy) security, the shared key needs to be set both in the AP and the mobile station. The AP can also be configured as a DHCP (Dynamic Host Configuration Protocol) server where the AP will supply the IP address to the connecting client. Depending on the situation, security parameters for 802.1x (discussed later in this chapter) or WEP are configured in the AP. This will include configuring the RADIUS (Remote Authentication Dial In User Service) server or other authentication servers like Kerberos, etc. Other parameters like Service Set Identifier (SSID), channel selection, beacon interval, etc., will be set on the AP.

In the client we need to define the network type. Network types can be either in infrastructure mode or ad hoc mode. The SSID needs to be defined in the client for the network identification and attachment. The shared WEP key needs to be installed in the client.

10.6.3 Managing 802.11 Networks

Two key components to a successful wireless network deployment are good management and monitoring tools. Providing a stable and manageable network infrastructure with effective support, problem detection, and problem resolution is dependent upon a good foundation of network products and tools. For the 802.11 wireless network, this includes utilities on the client computer that allow the user to monitor the health of their radio connection, and the infrastructure tools used by IT to manage and monitor the wireless network. Most of the clients provide tools to check the health of the link.

Managing Access Points

The task of managing APs can be broken down into management and monitoring/reporting. Management tools are typically provided with the AP. Management tools allow IT staff to perform initial set-up and overall administration of an AP. Initial set-up includes tasks such as configuring the device name, channel selection, SSID settings, IP addressing, security settings, and Ethernet settings. Administration includes tasks such as changing IP addresses and WEP settings, upgrading firmware, performing AP remote re-boots, and analyzing AP network interfaces and AP client connections.

Monitoring and reporting tools can provide real-time monitoring and alerting as well as trend reporting for wireless network devices. These tools can allow IT staff to track network device health and receive alerts of critical events or outages.

Client Tools

For best results, choose client Network Interface Cards (NICs) with Wi-Fi certification of interoperability, as well as easy-to-use client utilities for diagnostics and determining the RF signal strength and quality. The user interface should provide pertinent information on link status, network statistics, configuration options such as SSIDs, WEP keys, and so forth. It should also allow users to easily maintain multiple profiles and to switch between them as required. Following are some free tools that can be used for this purpose.

Aida32 (http://www.aida32.hu/aida32.php) Aida32 provides direct and rapid access to a server's Event Viewer, User Manager, live lists (of Dynamic Link Libraries), open files, services in use and other hidden terminals. Developed, upgraded and maintained by Unlimited Possibilities in Budapest,

Hungary, Aida32 works specifically on Win32 platforms and can be used to perform diagnostics and benchmarking.

Network Probe (http://www.objectplanet.com/probe/) is a protocol analyzer developed by ObjectPlanet of Oslo, Norway. This tool is designed to provide a real-time view of network traffic. The software can track and isolate traffic problems and congestion on network lines. It monitors conversations between hosts and applications, and shows network managers from and to where the network traffic is traveling.

Cflowd (http://www.objectplanet.com/probe/) and flowscan (http://www.caida. org/tools/utilities/flowscan/) Cflowd collects and correlates data from NetFlow, a part of Cisco's IOS that collects and measures data as it enters router or switch interfaces. The data can be used to monitor key applications, including accounting, billing and network planning, for corporate or service provider customers.

Kismet (http://www.kismetwireless.net/) is a network sniffer. It can spot unauthorized wireless access points. Unlike a standard sniffer, Kismet can identify and separate wireless use on the IP network. Kismet works with any 802.11b wireless card that can report raw wireless packets.

There are several excellent tools that are capable of mapping the access points in the area. They are also useful tools for installation/deployment and detection of rogue access points.

- NetStumbler (http://www.netstumbler.com) for Windows.
- MiniStumbler (http://www.netstumbler.com) for Pocket PC.
- PocketWarrior (http://www.pocketwarrior.org) for Pocket PC.
- Kismet (http://www.kismetwireless.net) for Linux.
- Dstumbler (http://www.dachb0den.com/projects/dstumbler.html) for NetBSD, FreeBSD, OpenBSD.
- Wellenreiter (http://www.remote-exploit.org) for Linux, experimental BSD.
- 802.11 Network Discovery Tools (http://sourceforge.net/projects/wavelan-tools/) for Linux.
- iStumbler (http://homepage.mac.com/alfwatt/istumbler/) for Mac.
- AirMagnet (http://www.airmagnet.com) for Windows, PDA.
- THC-WarDrive (http://www.thc.org) for Linux.
- PrismStumbler (http://prismstumbler.sourceforge.net) for Linux.
- WaveStumbler (http://www.cqure.net/tools.jsp?id=08) for Linux.
- WaveMon (http://www.jm-music.de/projects.html) for Linux.

Using such tools for driving around town discovering access points is called "WarDriving" after "WarDialing" where one used a modem to call a phone number at random or sequence in hope of finding an open or insecure server.

10.7 Mobile Ad hoc Networks and Sensor Networks

A mobile ad hoc network (MANET) is an autonomous system of mobile stations connected by wireless links to form a network. This network can be modeled in the form of an arbitrary graph. Ad hoc networks are peer-to-peer, multihop networks where data packets are transmitted from a source to a destination via intermediate nodes. Intermediate nodes serve as routers in this case. In

an ad hoc network there will be situations when some of the nodes could be out of range with respect to some other nodes. When this happens, the network needs to reconfigure itself and ensure that the paths between two nodes are available. In an ad hoc network, communication links could be either symmetric (bidirectional) or asymmetric (unidirectional). Figure 10.14 shows a wireless ad hoc network.

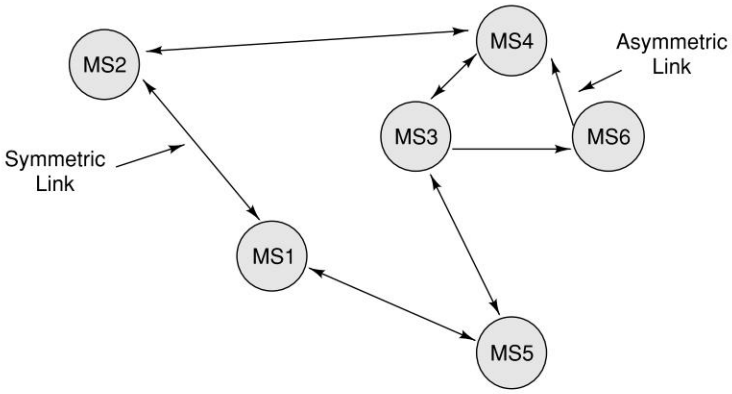


Figure 10.14 Mobile Ad hoc Network

To design a good wireless ad hoc network be it a sensor network or an information network, we need to account for various challenges. These are:

Dynamic topology. Nodes are free to move in an arbitrary fashion resulting in the topology changing arbitrarily. This characteristic demands dynamic configuration of the network.

Limited security. Wireless networks are vulnerable to attack. Mobile ad hoc networks are more vulnerable as by design any node should be able to join or leave the network any time. This requires flexibility and higher openness.

Bandwidth limited. Wireless networks in general are bandwidth limited. In an ad hoc network it is all the more so because there is no backbone to handle or multiplex higher bandwidth.

Routing. Routing in a mobile ad hoc network is complex. This depends on many factors, including finding the routing path, selection of routers, topology, protocol, etc.

10.7.1 Wireless Sensor Networks

Wireless sensor networks are a class of ad hoc networks. Sensor networks are very useful in unpredictable, unreliable environments. Sensor networks are primarily data collection points. They are widely used in defense, environment, meteorology, and study of nature. A wireless sensor network is a collection of low-cost, low-power disposable devices. Each of these devices holds sensing, memory, and communication modules. Study of the movement of glaciers is done through wireless ad hoc networks. Sensor networks are generally unmanned. Sensors may not have any power source other than small batteries. Therefore power control is a major challenge in sensor networks to ensure long life of the network.

10.8 Wireless LAN Security

In a wired network one has to be physically connected to transfer or receive data. This implies that it is possible to control the users in the network by controlling the physical access. Using a wireless network means using a radio transmitter and receiver. With varying degrees, radio signals will penetrate most building materials. Therefore, it is not possible to set up absolute physical boundary and expect that no outsider will be able to intrude into the network. With wireless networks we have no control of who might be receiving and listening to the transmissions. It could be someone in the building across the road, in a van parked in the parking lot or someone in the office above. Therefore, it is important that we understand the vulnerabilities of the wireless LAN and take necessary precautions.

As a part of the original specification, IEEE 802.11 included several security features, such as open system and shared key authentication modes; the Service Set Identifier (SSID); and Wired Equivalent Privacy (WEP). Each of these features provides varying degrees of security.

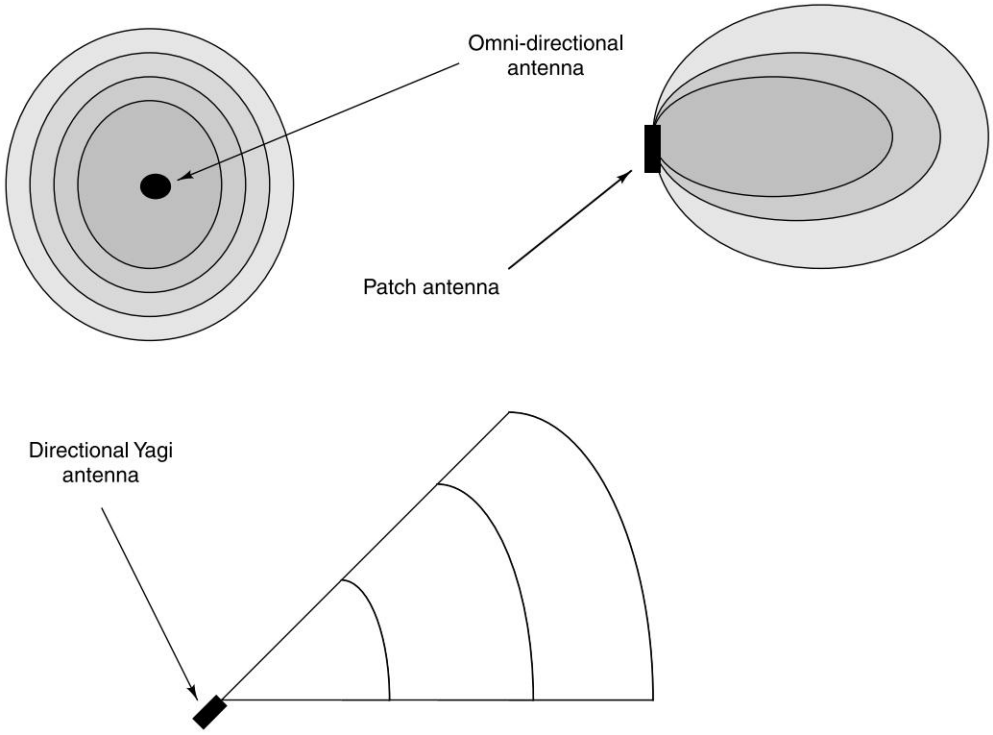


Figure 10.15 RF Transmission Pattern of Antennas

10.8.1 Limiting RF Transmission

It is important to consider controlling the range of RF transmission by an access point. It is possible to select proper transmitter/antenna combination that will help transmission of the wireless signal

only to the intended coverage area. Antennas can be characterized by two features—directionality and gain. Omni-directional antennas have a 360-degree coverage area, while directional antennas limit coverage to better-defined areas (see Fig. 10.15).

10.8.2 Service Set Identifier (SSID)

According to the 802.11 standard, a mobile station has to use the SSID of the access point for association between the NIC (Network Interface Card) in the client and the AP. The SSID is a network name (Id of the BSS or Cell) that identifies the area covered by an AP. The AP periodically broadcasts its SSID as a part of the management frame (beacon packet). As discussed earlier, the broadcast of beacon packet is necessary for clock synchronization. Unfortunately, as management frames of 802.11 are always sent in the clear, an attacker can easily listen on the wireless media for the management frames and discover the SSID to connect to the AP. The SSID can be used as a security measure by configuring the AP to broadcast the beacon packet without its SSID. The wireless station wishing to associate with the AP must have its SSID configured to that of the AP. If the SSID is not known, management frames sent to the AP from the wireless station will be rejected. It is also advised that the SSID of the AP is changed from the factory set defaults to some name, which is difficult to guess.

10.8.3 MAC Address Access Control

Many access points support MAC address filtering. This is similar to IP Filtering. The AP manages a list of MAC addresses that are allowed or disallowed in the wireless network. The idea is that the MAC address of the network card is unique and static. By controlling the access from known addresses, the administrator can allow or restrict the access of network only to known clients.

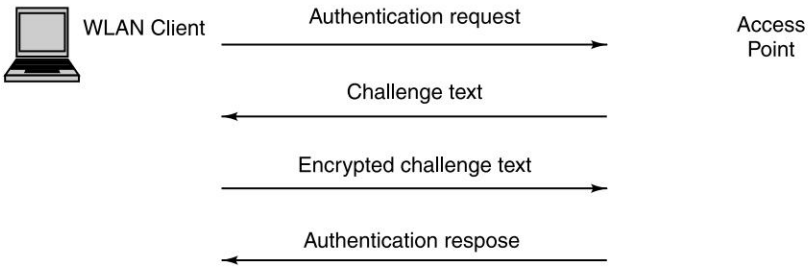


Figure 10.16 Shared Key Authentication

10.8.4 Authentication Modes

Two types of client authentication are defined in 802.11: Open System Authentication and Shared Key Authentication. Open system authentication is no authentication at all. Shared key

authentication on the other hand (Fig. 10.16) is based on the fact that both stations taking part in the authentication process have the same "shared" key.

It is assumed that this key has been transmitted to both stations through some secure channel other than the wireless media itself. In typical implementations, this is set manually on the client station and the AP. The authenticating station receives a challenge text packet (created using the WEP Pseudo Random Number Generator (PRNG)) from the AP. The station encrypts this PRNG using the shared key, and sends it back to the AP. If, after decryption, the challenge text matches, then one-way authentication is successful. To obtain mutual authentication, the process is repeated in the opposite direction.

10.8.5 WEP (Wired Equivalent Privacy)

WEP was designed to protect users of a WLAN from casual eavesdropping and was intended to offer following facilities:

- **Reasonably strong encryption.** It relies on the difficulty of recovering the secret key through a brute force attack. The difficulty grows with the key length.
- **Self-synchronizing.** Each packet contains the information required to decrypt it. There is no need to deal with lost packets.
- **Efficient.** It can be implemented in software with reasonable efficiency.
- **Exportable.** Limiting the key length leads to a greater possibility of export beyond the US.

The WEP algorithm is the RC4 cryptographic algorithm from RSA Data Security. RC4 uses stream cipher technique. It is a symmetric algorithm and uses the same key for both enciphering and deciphering the data. For each transmission, the plaintext is bitwise XORed with a pseudorandom keystream to produce ciphertext. For decryption the process is reversed.

The algorithm operates as follows:

1. It is assumed that the secret key has been distributed to both the transmitting and receiving stations by some secure means.
2. On the transmitting station, the 40-bit secret key is concatenated with a 24-bit Initialization Vector (IV) to produce a seed for input into the WEP PRNG (Pseudo Random Number Generator).
3. The seed is passed into the PRNG to produce a stream (keystream) of pseudorandom octets.
4. The plaintext PDU is then XORed with the pseudo-random keystream to produce the ciphertext PDU.
5. This ciphertext PDU is then concatenated with the 24-bits IV and transmitted on the wireless media.
6. The receiving station reads the IV and concatenates it with the secret key, producing the seed that it passes to the PRNG.
7. The receiver's PRNG produces identical keystream used by the transmitting station. When this PRNG is XORed with the ciphertext, the original plaintext PDU is produced.

It is worth mentioning that the plaintext PDU is also protected with a CRC to prevent random tampering with the ciphertext in transit.

10.8.6 Possible Attacks

The possible security attacks on wireless LAN are:

- **Passive attacks** to decrypt traffic based on statistical analysis.
- **Active attacks** to inject new traffic from unauthorized mobile stations, based on known plaintext.
- **Active attacks** to decrypt traffic, based on tricking the access point.
- **Dictionary-building attack** that, after analysis of about a day's worth of traffic, allows real-time automated decryption of all traffic.
- **Hijacking a session:** Following successful authentication, it is possible to hijack the session.

Analysis suggests that though these attacks are not common, it is possible to perform them using inexpensive off-the-shelf equipment.

10.8.7 802.1X Authentication

To prevent attacks on wireless LAN, the IEEE specification committee on 802.11 included the 802.1x authentication framework. The 802.1x framework provides the link layer with extensible authentication, normally seen in higher layers.

802.1x requires three entities (Fig. 10.17):

- **The supplicant:** Resides on the wireless LAN client.
- **The authenticator:** Resides on the access point.
- **The authentication server:** Resides on the server authenticating the client (e.g., RADIUS Kerberos, or other servers).

These are logical entities on different network elements. In a single network there could be many points of entry. These entries are through access points. Once the link between a supplicant (wireless station) and an authenticator (AP) is achieved, the connection is passed to the authentication server. The AP authenticates the supplicant through the authentication server. If the authentication is successful, the authentication server instructs the authenticator to allow the supplicant to access the network services. The authenticator works like a gatekeeper.

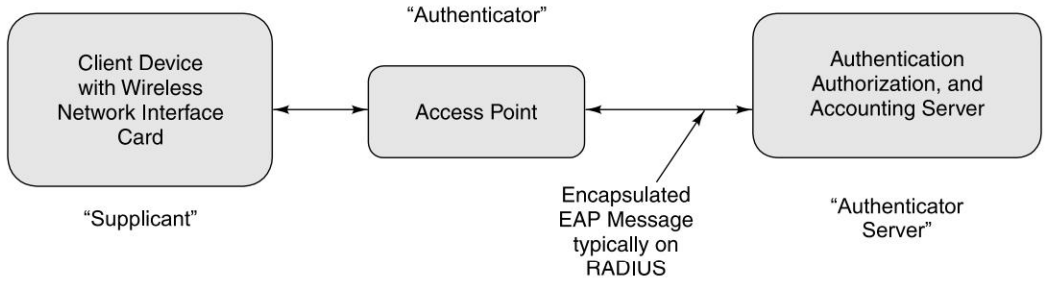


Figure 10.17 802.1x Setup

The authenticator creates one logical port per client, based on the client's association ID. This logical port has two data paths. The uncontrolled data path allows network traffic through to the network. The controlled data path requires successful authentication to allow network traffic through (Fig. 10.18). In order to obtain network connectivity, a wireless client must associate with the AP. Complete association with an AP involves three states:

1. Unauthenticated and unassociated
2. Authenticated and unassociated
3. Authenticated and associated

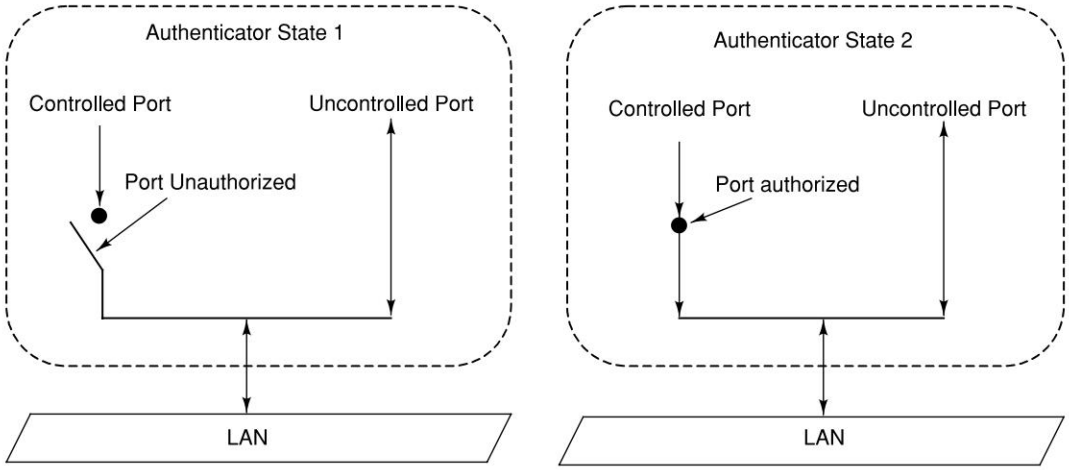


Figure 10.18 Effect of Authorization State on Controlled Ports

IEEE 802.1x offers flexibility in authentication and possible encryption. After the link has been established, PPP (Point-to-Point Protocol) provides for an optional authentication phase before proceeding to the network layer protocol phase. This is called EAP (Extensible Authentication Protocol). Through the use of EAP, support for a number of authentication schemes may be added, including Smart cards, Kerberos, Public Key, One Time Passwords, CHAP (Challenge Handshake Authentication Protocol), or some other user-defined authentication systems.

There are still some vulnerabilities in the EAP. To overcome this, a new standard is being proposed in IETF to override the EAP proposal. This new standard is called PEAP (Protected EAP). PEAP uses an additional phase of security over and above EAP.

10.8.8 Wireless VPN

Virtual Private Network technology (VPN) has been used to secure communications among remote locations via the Internet since the 1990s. It is now being extended to wireless LAN. VPNs were traditionally used to provide point-to-point encryption for long Internet connections between remote users and the enterprise networks. VPNs have been deployed in wireless LANs as well. When a wireless LAN client uses a VPN tunnel, communications data remains encrypted until it reaches

the VPN gateway, which sits behind the wireless AP. Thus, intruders are effectively blocked from intercepting all network communications.

10.8.9 802.11i

Task Group "i" within IEEE 802.11, is developing a new standard for WLAN security. The proposed 802.11i standard is designed to embrace the authentication scheme of 802.1x and EAP while adding enhanced security features, including a new encryption scheme and dynamic key distribution. Not only does it fix WEP, it takes wireless LAN security to a higher level.

The proposed specification uses the Temporal Key Integrity Protocol (TKIP) to produce a 128-bit "temporal key" that allows different stations to use different keys to encrypt data. TKIP introduces a sophisticated key generation function, which encrypts every data packet sent over the air with its own unique encryption key. Consequently, TKIP greatly increases the complexity and difficulty of decoding the keys. Intruders will not have enough time to collect sufficient data to decipher the key.

802.11i also endorses the Advanced Encryption Standard (AES) as a replacement for WEP encryption. AES has already been adopted as an official government standard by the US Department of Commerce. It uses a mathematical ciphering algorithm that employs variable key sizes of 128-, 192- or 256-bits, making it far more difficult to decipher than WEP. AES, however, is not readily compatible with today's Wi-Fi Certified WLAN devices. It requires new chipsets, which, for WLAN customers, means new investments in wireless devices. Those looking to build new WLANs will find it attractive. Those with previously installed wireless networks must justify whether AES security is worth the cost of replacing equipment.

10.9 Wireless Access in Vehicular Environment

Starting from guided navigation to security, inter vehicle network has many applications. IEEE 802.11p is a draft amendment to the IEEE 802.11 standard to add Wireless Access in Vehicular Environments (WAVE). This includes data exchange between high-speed vehicles and the roadside infrastructure in the licensed ITS band of 5.9 GHz (5.85-5.925 GHz). 802.11p will be used as the groundwork for Dedicated Short Range Communications (DSRC), a US Department of Transportation project based on the ISO Communications, Air-interface, Long and Medium range (CALM) architecture standard looking at vehicle-based communication networks. This will include services like toll collection, vehicle safety services, locating restaurants, booking train or bus tickets, e-commerce transactions, etc., from the vehicle. 802.11p defines enhancements to 802.11 required to support Intelligent Transportation Systems (ITS) applications. ITS is a mobile computing application domain where all types of transportation and hospitality services will be integrated.

802.11p will ensure interoperability between wireless devices attempting to communicate in potentially rapidly changing communications environments between moving vehicles or a moving vehicle and a roadside base station or access point where transactions must be completed in time frames much shorter than the minimum possible with infrastructure or ad hoc 802.11 networks. In particular, time frames that are shorter than the amount of time required to perform standard authentication and association to join a BSS are accommodated in this amendment. The P802.11p specification accomplishes the following:

- Describes the functions and services required by WAVE-conformant stations to operate in a rapidly varying environment and exchange messages either without having to join a BSS or within a WAVE BSS.
- Defines the WAVE signaling technique and interface functions that are controlled by the IEEE 802.11 MAC.

10.10 WIRELESS LOCAL LOOP

In fixed line telephone networks, the wire from the subscriber's telephone equipment to the local exchange has been called the local loop (or the subscriber loop). This connection has also been called the last mile (or first mile) of end-to-end transmission. When the network operator uses wireless means to connect the fixed line telephone equipment with the local exchange, it is called Wireless in Local Loop (WLL). This is also known as Radio Local Loop (RLL), Fixed Wireless Access (FWA) or Fixed Radio Access (FRA). WLL technology which delivers high speed broadband Internet to customers' premises is often known as Broadband Wireless Access (BWA).

WLL has high value in rural (remotely accessible) areas that cannot be economically served by conventional wired access technologies. Presently, WLL systems are based on a wide range of radio access technologies including satellite technology, cellular technology and microcellular technology; however, in India they are mostly CDMA (see Chapter 9) based wireless technology. In India and in many other parts of the world, though WLL uses the same technology and same mobile infrastructure, the call tariff of WLL is cheaper and usually the same as fixed line phones instead of cellular phone calls.

10.10.1 WLL Architecture

As WLL systems are fixed, the requirements of interoperability of a subscriber's equipment with different base stations is lesser stringent than that for mobile stations (like cellular phones). So, there are many scientific and commercial standards wherein each has its own air interface specification, network elements and system architecture.

The Fixed Subscriber Unit (FSU) is the serving interface between subscriber's wired devices and the wireless interface of the WLL network (see Fig. 10.19). FSUs are also known as Wireless Access Fixed Unit (WAFU) or Radio Subscriber Unit (RSU). The subscriber's wired devices can be PCs, telephones, fax machines, or any other device. FSU does channel coding/decoding, modulation/demodulation and transmission/reception of radio signals (apart from providing source coding/decoding occasionally) according to the air interface specification. The basic function of a FSU is similar to that of a mobile phone handset except that it does not have a rich set of functions for mobility management.

The Base Station Controller (BSC) controls one or more Base Transceiver Stations (BTSs). BSC provisions an interface to the local exchange in the central office apart from transcoding between the source codes used in the wired network and that at the air interface. That is why a BSC is often called Radio Port Control Unit (RPCU). Since WLL systems do not need to provide full fledged mobile services, there is no need of Home Location Register (HLR)/Visitor Location Register (VLR) as in the case of cellular networks.

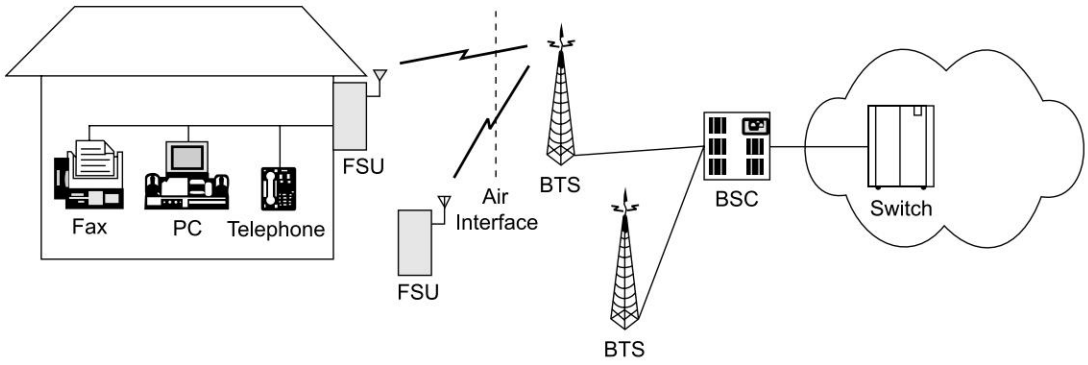


Figure 10.19 Generalized WLL Architecture

10.11 HIPERLAN

HiperLAN is a European counterpart for the wireless local area network—it stands for HIgh PErformance Radio Local Area Network. HiperLAN is a family of standards developed in Europe by BRAN project (Broadband Radio Access Networks) of ETSI. It defines interoperability standards which specify a common air interface MAC and Physical layers in OSI model. Figure 10.20 shows the stack relationship of HiperLAN with OSI layers. It is equivalent to Wireless LAN defined by IEEE 802.11 standards.

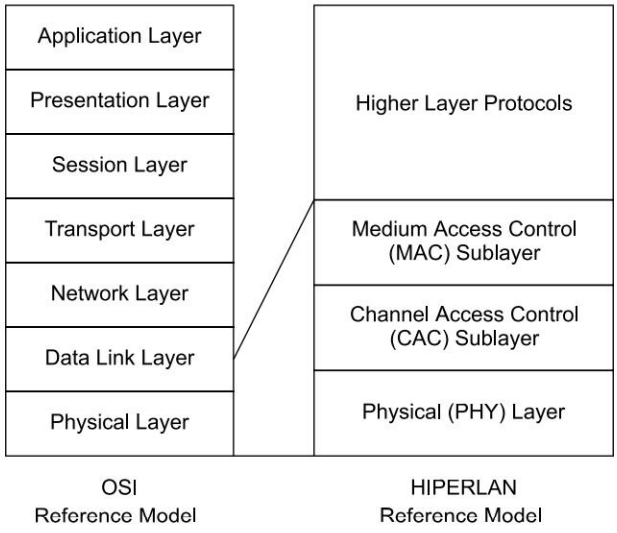


Figure 10.20 OSI and HiperLAN Reference Models

The physical layer and the Media Access Control part of the HiperLAN Data link layer are like 802.11 standards. However, there is a new sublayer called Channel Access and Control (CAC) sublayer which deals with the access requests to the channels. The request is served depending upon the usage of the channel and the priority of the request. CAC layer provisions hierarchical independence with Elimination-Yield Non-Preemptive Multiple Access mechanism (EY-NPMA). EY-NPMA codes priority choices and other functions into one variable length radio pulse preceding the packet data. EY-NPMA helps network to function with lesser collisions even when there are a large number of users. Transitively, multimedia applications work better in HiperLAN because of EY-NPMA priority mechanism.

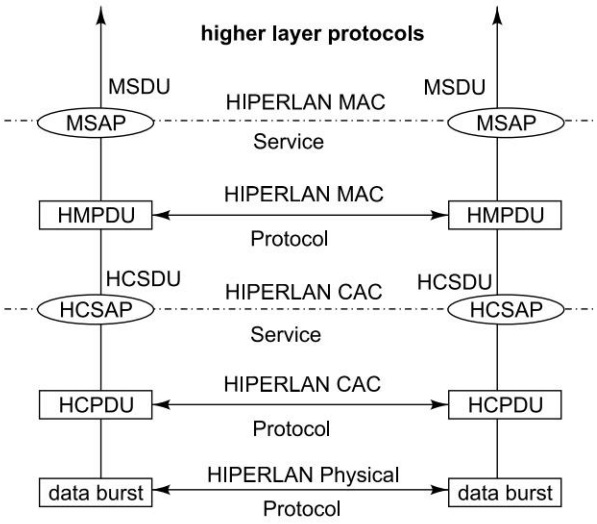


Figure 10.21 HiperLAN Communication Models

The HiperLAN communication model is shown in Figure 10.21 The HiperLAN MAC service is compatible with the ISO MAC service definition; it

- defines the communication service over a single HiperLAN;
- allows the timing requirements of the MAC Service Data Unit (MSDU) transfer to be specified; and
- allows the exploration of available HiperLANs for dynamic HiperLAN access.

The HiperLAN CAC service defines the communication service over a single shared communication channel; it

- allows the channel access priority requirements of the HiperLAN CAC Service Data Unit (HCSDU) transfer to be specified; and
- frees the HCS-user from the concerns of the characteristics peculiar to any particular communication channel.

The HiperLAN MAC protocol:

- provides the HiperLAN MAC service;
- specifies the behavior of a HM-entity in a given HiperLAN;
- is compatible with the ISO MAC bridges specification;
- uses the HiperLAN CAC service;

The HiperLAN CAC protocol provides the HiperLAN CAC service; it

- specifies, for a particular set of one or more shared radio channels, the appropriate hierarchically independent channel access mechanism used by a HiperLAN CAC-entity in a given HiperLAN; and
- uses the transmission and reception facilities specified by the HiperLAN physical layer.

The HiperLAN physical layer protocol provides the transmission and reception facilities to the HiperLAN CAC sublayer; and specifies, for a particular set of one or more shared radio channels, the techniques of transmission, reception and channel assessment in a given channel.

HiperLAN versions

HiperLAN has gone through versions like HiperLAN/1 and HiperLAN/2. HiperLAN/1, started in 1991, was approved in 1996. HiperLAN/2 specifications, completed in early 2000, were designed as a fast wireless connection for many kinds of networks (like UMTS, ATM and IP networks). HiperLAN/2 operates in the 5 GHz band and offers upto 54 Mbit/s data transfer rate. The media access control (MAC protocol), in HiperLAN/2, is Dynamic TDMA (which is used in the broadband wireless access standards like WiMax). HiperLAN/2 offers better security measures than HiperLAN/1 as the data are secured with DES or Triple DES algorithms.

HiperMAN

A variant of HiperLAN is High Performance Radio Metropolitan Area Network (HiperMAN), which is also from ETSI and similar to WiMAX. HiperMAN aims at provisioning broadband Wireless Internet access while covering a large geographic area. HiperMAN stands highly optimized for packet switched networks and readily supports fixed and nomadic applications.

10.12 WIFI VERSUS 3G

3G offers a vertically integrated, top-down, service-provider approach for delivering wireless Internet access; while WiFi offers an end-user-centric, decentralized approach to service provisioning. Table 10.4 highlights characteristics of 3G versus WiFi:

Table 10.4 3G versus WiFi

Functions	*3G*	*WiFi*
Genesis	Evolved from voice network (real-time traffic) where QoS is a critical success factor.	Evolved from data network (store and forward) where QoS is not a critical success factor.
Radio Interface	Use Spread Spectrum as the modulation technique.	Use Spread Spectrum as the modulation technique.
Access technologies	Access or edge-network facility. Offers alternatives to the last-mile wireline network. The wireless link is from the end-user device to the cell base station, which may be at a distance of up to a few kilometers.	Access or edge-network facility. Offers alternatives to the last-mile wireline network. The wireless link is a few hundred feet from the end-user device to the base station.
Bandwidth	3G supports broadband data service of up to 2 Mbps. 3G will support "always on" connectivity.	WiFi offers broadband data service of up to 54 Mbps. WiFi offers "always on" connectivity.
Business models/deployment	Service providers own and manage the infrastructure (including the	Users' organization owns the infrastructure. Following the initial

(*Contd*)

Functions	*3G*	*WiFi*
are different	spectrum). End customers typically have a monthly service contract with the 3G service provider to use the network.	investment, the usage of the network does not involve an access fee.
Spectrum policy and management	3G uses licensed spectrum. This has important implications for: (a) Cost of service. (b) Quality of Service (QoS). (c) Congestion Management. (d) Industry structure. The upfront cost of acquiring a spectrum license represents a substantial share of the capital costs of deploying 3G services. However, with licensed spectrum, the licensee is protected from interference from other service providers.	WiFi uses unlicensed, free, shared spectrum. Therefore, it does not involve any additional costs to acquire the spectrum. Unlicensed spectrum can be used by anybody and can be used for any purpose. This may cause interference in each other's right to use the spectrum. Therefore, WiFi imposes strict power limits on users (i.e., responsibility not to interfere with other users) and forces users to accept interference from others.
Status of Standards	The formal standards picture for 3G is perhaps clearer than for WiFi. For 3G, there is a relatively small family of internationally sanctioned standards, collectively referred to as IMT-2000.	WiFi protocol is one of the family of continuously evolving 802.11x wireless Ethernet standards, which itself is one of many Wireless LAN technologies that are under development.
Roaming	3G will offer well coordinated continuous and ubiquitous coverage. This offers seamless roaming. To support this service, mobile operators maintain a network of interconnected and overlapping infrastructure that allows customers to roam.	WiFi network growth is unorganized. Therefore seamless ubiquitous roaming over WiFi cannot be guaranteed. Also, WiFi technology has not been designed to support high-speed handoff associated with users moving between base station coverage areas.

REFERENCES/FURTHER READING

1. Brenner Pabio (1997), "A Technical Tutorial on the IEEE 802.11 Protocol", *Breezecom Wireless Communications.*
2. Deploying 802.11B (Wi-Fi) In the Enterprise Network, Dell White Paper, 2001.
3. EN 300 652, European Standard (Telecommunications series) Broadband Radio Access Networks (BRAN); HIgh PErformance Radio Local Area Network (HIPERLAN) Type 1; Functional specification.

4. Ergen Mustafa (2002), *IEEE 802.11 Tutorial*, UC Berkeley.
5. ETSI TR 102 079 Technical Report, Electromagnetic compatibility and Radio spectrum Matters (ERM); System Reference Document for licence-exempt Fixed Wireless Access (HIPERMAN) for band C (5,725 GHz to 5,875 GHz)
6. IEEE 802.11 Working Group: http://grouper.ieee.org/groups/802/11/.
7. "Introduction to Wireless LAN", *The Wireless LAN Alliance*, www.wlana.com.
8. Lehr, William and Lee W. McKnight, (2002), Wireless Internet Access: 3G vs. WiFi, http://itc.mit.edu/itel/docs/2002/LehrMcKnight.
9. Li Guoliang, "Physical Layer Design for a Spread Spectrum Wireless LAN", MS Thesis, Virginia Polytechnic Institute and State University.
10. Lin Y.B. (1997), "Wireless Local Loop," IEEE Potentials, Aug/Sept.
11. Little S. (1996), "Going Wireless on the Transmission Side," *Telecom Asia*, Vol. 7, No. 9, Oct. pp. 30–40.
12. Noerpel A.R. (1997), "WLL: Wireless Local Loop – Alternative Technologies," Proc. IEEE PIMRC, Helsinki, Finland.
13. Schreiber Eric and Daniel Sigg (2002), "Clock Synchronization for Wireless LAN", Thesis DA- 2002.18 Summer Term.
14. Varma V. and V. Panday (1997), "Functional Architecture for PACS Wireless Local Loop System," ATIS, *PACS Providers Forum.*, T1P1.3/96- 246R1, Jan. 27.
15. Wireless LAN protocol and its effect on cell design, Integrity Data Systems Pty Ltd. White paper, A.B.N. 17 148 989 654, 2000.
16. www.etsi.org
17. www.palowireless.com
18. www.wikipedia.org
19. www.wikipedia.org
20. Wikipedia – the free encyclopedia, http://www.wikipedia.org
21. Research and Innovative Technology Administration (RITA) Intelligent Transportation Systems IEEE 1609—Family of Standards for Wireless Access in Vehicular Environments (WAVE) http://www.standards.its.dot.gov/fact_sheet.asp?f=80
22. ASTM E2213 – 03, http://www.astm.org/Standards/E2213.htm

REVIEW QUESTIONS

Q1: What are the advantages and disadvantages of Wireless LAN? Under what situation is a WLAN desirable over LAN?

Q2: Describe the IEEE 802.11 family of standards.

Q3: In an Ethernet LAN CDMA-CD is used, whereas in Wireless LAN CDMACA is used. What is CDMA-CA? Why is CDMA-CA used instead of CDMA-CD in Wireless LAN?

Q4: How is WLAN configured and managed?

Q5: How are mobility and handoffs managed in Wireless LAN?
Q6: What is WEP? Why is it considered unsafe? What are the mechanisms advised to ensure security in Wireless LAN?
Q7: How does 802.1x overcome the security vulnerabilities in WEP?
Q8: Describe the contrast between 3G and WiFi technologies.
Q9: What is the motivation for using WLL?
Q10: Explain implementation of a WLL system.
Q11: What are the main devices used in implementing a WLL system?
Q12: Explain the business prospects of using a WLL system. How would that evolve in the future?
Q13: How is WLL different from a cellular phone technology?
Q14: What is HiperLAN? Explain its stack relationship with OSI protocol stack.
Q15: What are the prominent standards in HiperLAN? Explain each of them.
Q16: Explain HiperMAN.

CHAPTER 11

Intelligent Networks and Interworking

11.1 INTRODUCTION

A communication network provides the service of transportation of payload of its subscribers. This payload can be voice, data or other kind of payloads. Switches are at the heart of these networks routing the traffic from source to sink. Also, these work out procedures to charge the subscriber for the network service it provides. Switches used by the network operators are expensive and are designed to do certain limited functions.

Intelligence network (IN) is a concept where intelligence is taken out of the central switch and distributed within the network. Through IN, intelligence is added into the network and placed in computer nodes that are distributed throughout the network. IN is an architecture, which separates the service logic from the switching service of the telephone exchanges. IN enables the establishment of an open platform for uniform service creation, implementation and management. Once introduced, services are easily customized to meet customer's need. It aims at rapid and economical service provisioning and facilitates customer control of network services. This makes telecom networks different from data networks. In a data network, endpoints are intelligent with no intelligence in the network, whereas, in telecom networks, endpoints are dumb with intelligence in the network.

11.2 FUNDAMENTALS OF CALL PROCESSING

To understand IN, we need to understand the basic function of a switch, i.e., the processing of a call. To connect a caller to a called number, the switch performs a series of functions and makes a series of decisions. Decisions are like:

- Is this subscriber allowed to place this call?
- Where is the call to be connected?
- How will the call be connected? What path will it take?

- What should be the cost of the call?
- Who should be charged for the call?

To make a successful call, there are many steps. Figure 11.1 depicts various steps in a telecommunication call set up and tear down. However, if they are simplified, they will look like this:

Step 1 The A-subscriber lifts the handset. The moment a subscriber lifts the handset of a telephone, the switch in an exchange is able to detect that a subscriber has lifted the handset. At this moment the subscriber or the calling party (called the A-party) is welcome to enter the number of the destination or called party (called the B-party). The welcome message is a dial tone. The caller dials the telephone number of the called party.

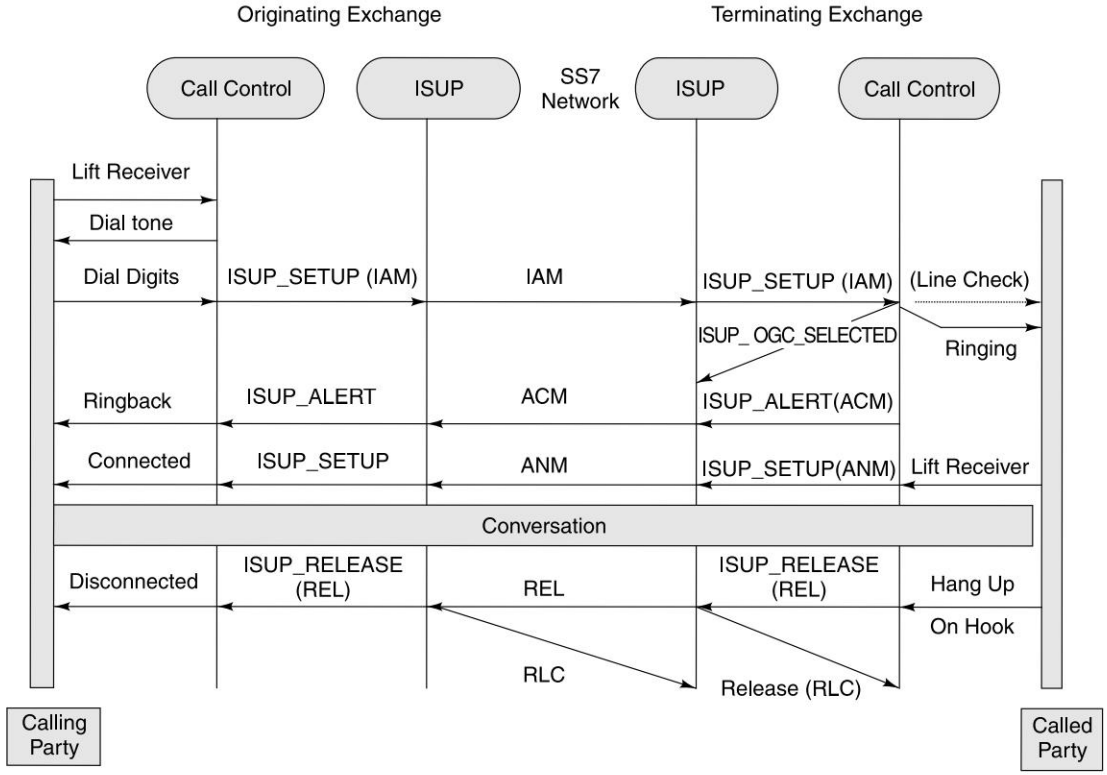


Figure 11.1 ISUP Call Establishment Procedure

Step 2 The exchange receives B-subscriber's number. B's number is received by the switch. The switch analyses B's number to determine whether B is within the same exchange or a different exchange. If B is in a different exchange, it is called out-of-switch number.

Step 3 The exchange sets up the outgoing call. If B is within the same exchange, the line interface circuit of B is obtained and A's line is physically connected to that of B's. If B is from

another exchange, a routing analysis is performed. The switch will make an attempt to connect to B using the routing path. At this stage ISUP messages move through the SS7 network from switching node to switching node. While the ISUP message is moving through the network, a voice path parallel to the ISUP path is being reserved for circuit establishment. If it is unable to set up a path due to the fact that B is busy, a busy tone is played for A. If B's phone is free, and the switch is able to establish a path between A and B, the circuit in the trunk is established. The telephone of B rings and A hears a ring tone. The ring tone of A is issued by A's exchange, whereas the ringing current (signal) to B's phone is sent by B's exchange.

The ISUP message sent from A's switch is called the IAM (Initial Address Message). This message contains all the information necessary for each switch to be able to consult its routing table and to select circuits that will result in connecting the circuit from end to end. A's switch receives a confirmation message called the ACM (Address Complete Message). When A's exchange receives this ACM message, it issues the ring tone to A. Once the phone is ringing, there is no further signalling being exchanged for a time.

A charging analysis is performed on the call request. Charging depends on subscriber's category, tariff plan, time of day, distance between the called and caller party, etc. One of the registered tariffs is selected for the billing of the call.

Following the ring at B's telephone, B decides to answer the call. B picks up the phone. When B's phone comes 'off hook' the switch at B's end sends an ANM (ANswer Message) backward into the SS7 network to A's exchange. Each switch is thus notified that the full circuit must now exist. The circuit reserved from A to B is now connected. The system switches from signal mode to traffic mode.

A and B enter into conversation mode. They talk as long as they want. The switches monitor the connection, primarily to enable the call to be charged.

Step 4 The subscribers conclude their conversation. The switch continues to scan the subscribers' lines even during the conversation. Eventually, of course, someone hangs up. The phone line once again goes "on hook" and that is sensed at the subscriber interface of the switches serving the customer who hung up. Whichever office detects the hang up, that office sends a REL (release) message on to the previous switch in the circuit. Upon receiving the REL, each switch releases the circuit connection. At the same time, it returns an RLC (release complete) back to the switch that sent the REL. Switch by switch, the scenario continues until each switch in the circuit has released its circuitry and confirmed that action to the previous switch. Charging of the call stops, and a CDR (Call Detail Record) is produced for charging.

11.3 Intelligence in the Networks

In Chapter 3 we have mentioned that we dial 1-800-111100 to talk to Microsoft customer service from anywhere (main cities) in India. We have also mentioned that this call is free. This is an example of IN. Let us now understand how it works. When a user dials this number from anywhere in India, the local exchange knows that this not a local number. It tries to find out the routing path for the number. It discovers that this is a virtual number and needs to refer to an IN node to obtain the routing path. The switch asks the IN node for the routing path. The IN node looks at its

database and finds that this virtual number is mapped to a real number in Delhi (011-2629-2640). The IN node also informs the local switch (A's exchange) that A should not be billed; instead the B-party (Microsoft in this case) will be billed. The switch gets the routing path through another routing path enquiry on 011-2629-2640. Though the user has dialed 1-800-111100, the user gets connected to 011-2629-2640, and is not charged for a STD call.

If we dial this number between 9:00 am to 6:00 pm Monday through Friday, we can talk to a Microsoft customer service representative for free. If it is at any other time of the day, one will hear a recorded message, "Welcome to Microsoft connect customer services. For product activation please dial 1, for all other services we work between 9:00 am to 6:00 pm Monday through Friday. To leave a voice mail message please dial 2." Let us make this example slightly more interesting. Let us assume that Microsoft has a few premium customers for whom the service is available 24 hours a day, 7 days a week. In this case the IN node will also check the telephone number of the A-party and the time of the call. If A's telephone number matches with the telephone number of the premium customers and the time is between 6:00 pm to 9:00 am, the call will be diverted to a service engineer's home number. Also, please note that in this case the IN node has to make some even smarter decisions. It cannot blindly divert the call to a telephone of an engineer at Delhi. If the call has originated from the premium customer's Mumbai office, the IN node needs to divert the call to the service engineer in Mumbai. If the call is from Bangalore, the call needs to be diverted to the service engineer at Bangalore. Depending on the conditions, the IN node gives the routing path. If the call is connected to a service engineer's number, the IN node tells A's exchange not to bill A, not even to bill B (service engineer in this case), but to bill a different entity (Microsoft).

If we did not have IN to implement this 800 service, we would need to add all the above complex logic into all the switches in all the exchanges in the country, which is an impossible task. However, it is easy to implement it in a separate node and just implement some logic in the exchanges that any number staring with 1-800 is a virtual number. The routing path for a virtual number is obtained from an IN node. IN services address the following requirements:

- Remove the service data from the switching network and locate it in a centralized database. This database and service will be accessible from all the switching nodes.
- Separate the service logic from the switch and put it into an independent intelligent node. Whenever a new service is added to this node, it becomes available throughout the network. These intelligent nodes are called Service Control Point or SCP.
- A real-time connection is needed between the switching nodes "service switching points" (SSPs), and the "service control points" (SCPs). This fast and standardized interconnection forms the basis for the IN architecture. Figure 11.2 shows the relationship between these network elements.

11.3.1 Standards for Intelligent Networks

In 1989, ITU and ETSI began work on IN standards. A phased approach of development was adapted to define the target IN architecture. Each phase of development intended to define a particular set of IN capabilities, known as a capability set (CS). Each capability set defines the requirements for one or more of the following areas:

- Service creation.
- Service management.

- Service interaction.
- Service processing.
- Network management.
- Network interworking.

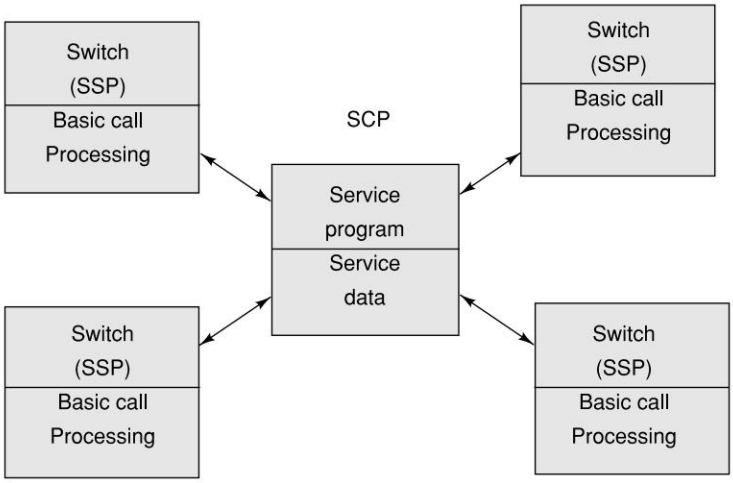


Figure 11.2 Intelligent Network Approach

In March 1992 the first capability set (CS-1) of standard was approved. Work on CS-2 was started in 1994 that addressed basic aspects that were excluded from CS-1, such as IN management. Furthermore, work on CS-3 was started in 1995.

11.4 SS#7 SIGNALING

In a telecommunications network if the switch is like the heart, signaling is like the nerve. All the information related to command, control, and monitoring of the network activities are transmitted through signaling channels. Signaling channels carry data and information for the purposes of management of subscriber traffic. Signaling System number Seven or SS#7 or SS7 in short has been designed for signaling in telecommunication networks. SS7 is a digital packet network. SS7 signaling is also called common channel or out-of-band signaling. SS7 defines the procedures and protocol by which network elements in the public switched telephone networks (PSTN) exchange information to effect wireless (cellular) and fixed line call set-up, routing, control, charging as well as network management and maintenance. Switches are the "glue" that holds the PSTN network together. Likewise, SS7 is held together by a digital counterpart of the switch known as a Signalling Transfer Point (STP).

SS7 comprises a series of interconnected network elements such as switches, databases, and routing nodes. Each of these elements is interconnected with links, each of which has a specific purpose. Main elements in SS7 network are STP, SSP, SCP, etc. These nodes are depicted in

Fig. 11.3 and has the following functions. The SS7 requires the use of continuously available transmission links. These links are 64 Kbps channels. Channel 23 is the SS7 signal channel within a 1.544 Mbps T1 line in the US and channel 16 within 2 Mbps E1 lines in India and Europe. The job of the STP is to examine the destination address of messages it receives, consult a routing table, and send the messages on their way using the links that are selected from the routing tables. In a PSTN network, a telephone handset is an end point. Likewise, SEP (Signaling End Point) is an end point in the SS7 network. Similar to a telephone number, Signaling End Points use an address known as a Signaling Point Code.

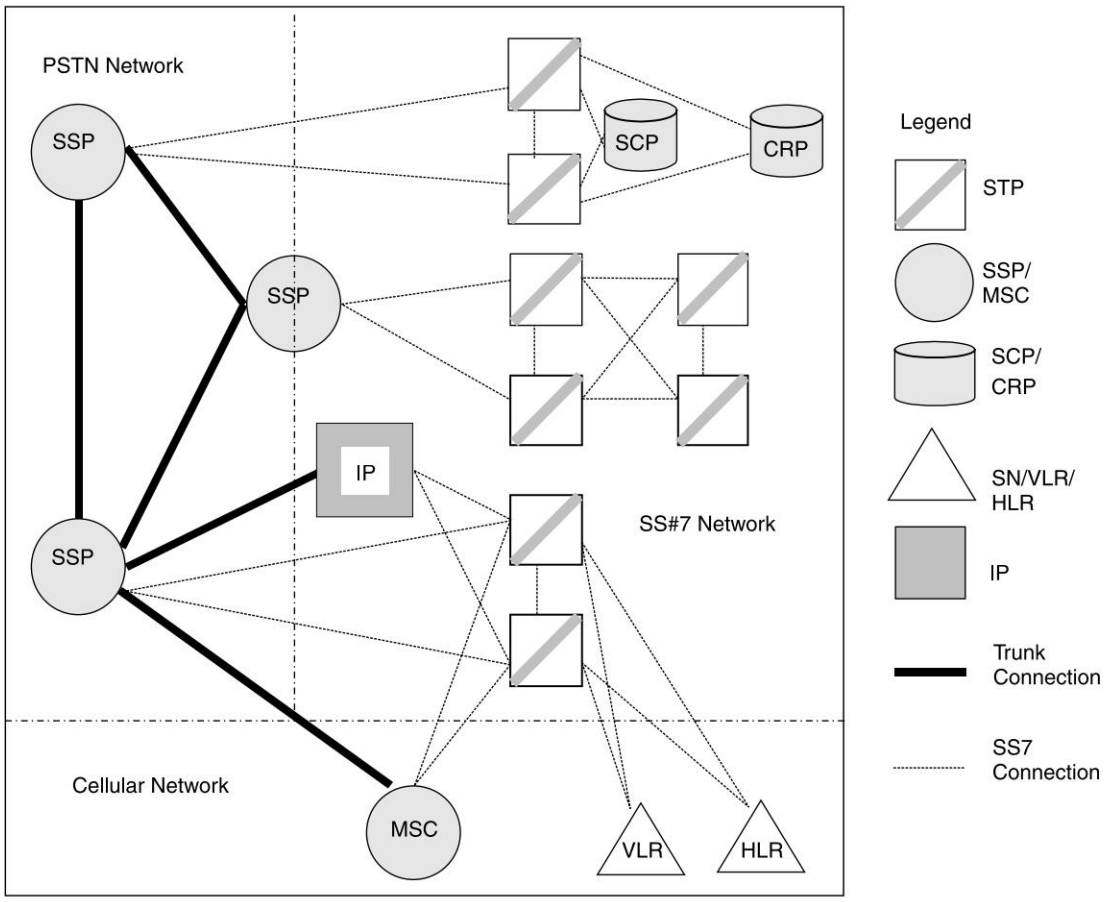


Figure 11.3 SS#7 Network Architecture

The STP (Signaling Transfer Point)

STP is like a switch or a node in the SS7 network doing the basic routing functions. For example, if an STP has links heading off toward the four compass points, it might be more "appropriate" to

direct a message addressed to Mumbai to a west-leading link than to an east-leading link. STP routing decisions are based on geography, distance, congestion, and least cost criteria. Once a SS7 message is delivered from a source to destination, a circuit on the same path is reserved for traffic. For fault tolerance, STPs are always installed in pairs with cross connections.

The SSP (Service Switching Point)

Service Switching Point is a switch in the SS7 network that can handle call set-up. The SSP has the ability to stop call processing, make queries to even unknown databases, and perform actions appropriate to the response. SSP is equipped with all the intelligence required to handle numerous feature capabilities. Example of a SSP will be a MSC in a cellular network. There is another switching point called a CCSSO (Common Channel Signaling Switching Office). These are end or tandem offices which have the capability to use the SS7 in what is referred to as a trunk signaling mode for call set-up. CCSSO is a limited version of the SSP.

The SCP (Service Control Point)

One of the first purely digital uses for the SS7 network was to provide a service to translate from one form of data to another. For example, switches need to maintain tables to translate dialed digits into routing information consistent with the international numbering system (for example iiitb number +91808410628). It is that plan that breaks India (country code 91) down into city code (80), exchange code (841), and finally to the line (0628) serving individual telephone. Let us take another example where a person has moved his home from one part of the town to another. When we dial the telephone number, it plays a recorded message like "The number you have dialed is 28670203, the number has changed to 25320203. You may dial the new number or wait for a while to be connected automatically."

When a virtual number like 1-800 in India is dialed, there is no way for the switch to determine how to route this call. This is because such prefixes have no reference to the international numbering plan. In fact, a 1-800 number dialed in Mangalore may be connected to a number in Bangalore, while the same number dialed in Pune may result in a connection to Mumbai. When that translation is returned to the switch, the number can be connected exactly as it would have been if it had been dialed in the first place. This database is located at an SS7 address called Signaling Point Code. SCPs are used for a variety of applications such as Calling Card verification, toll-free calls, televoting, premium tariff (1-900) calls, etc. Such intelligent nodes make the network intelligent. This also frees the switch from trying to maintain ever larger routing tables, and enables the use of a broad range of services which depend on translations or digital data services of a variety of types.

SCP provides the access mechanism required for a service. These services may reside in the same location as the SCP or the SCP may serve as a "front end" for services located elsewhere. In either case the SCP controls many services. To identify a service in a SS7 network, two parameters are required. These are SCP address and the service within the SCP.

CRP (Customer Routing Point)

The CRP provides on-premises control of the routing information requested by switches for translation of 800 type dialing. The operator of the CRP is a customer who requires rapid update and control of the translation of their own numbers.

Intelligent Peripheral (IP): This is a peripheral process that deals with the requests made of it through the SCP by providing the services of a variety of equipments. IP includes database functionality of the SCP along with additional capabilities such as voice interaction and control of voice resources. Generally speaking, SCPs work well with requirements that call for voluminous data transactions. IPs, on the other hand, are best suited for special circumstance call processing involving voice resources and/or interaction. Example of an IP is a voice-activated system, where instead of dialing a number we can simply say "call iiitb". Here the call will be routed to an IP, which in turn will provide the iiitb telephone number.

Services Node (SN): A programmable IP is called Service Node (SN). Still, what one network calls an IP might be called a Services Node in another network.

Services Management System (SMS): This is a node in the SS7 network that provides a human interface to the database. This also provides the facility to update the database. SMS provides GUI/command line interfaces to update and manage services and the network. Service Operators configure the SMS to manage such mission-critical tasks as billing or access authorization.

11.4.1 SS#7 Protocol Stack

The SS7 protocol stack is depicted in Fig. 11.4. In this model the upper four layers are called user parts whereas the lower are the message transfer parts.

Message Transfer Part

Message Transfer Part–Level 1: The Message Transfer Part Level 1 (MTP L1) is the "physical layer". It deals with hardware and electrical configuration. MTP level 1 is a part that deals with physical issues at the level of links, interface cards, multiplexors etc.
Message Transfer Part–Level 2: It is the last layer to handle messages being transmitted and the first layer to handle messages being received. It monitors the links and reports on their status. It checks messages to ensure their integrity (both incoming and outgoing). It acknowledges good messages; it discards bad messages and requests copies of discarded messages. It provides sequence numbering for outgoing messages.

Message Transfer Part–Level 3: The MTP Level 3 provides the functions and procedures related to Message Routing (or Signaling Message Handling) and Signaling Network Management. MTP L3 handles these functions assuming that signaling points are connected with signaling links. The message routing provides message discrimination and distribution. Signaling Network Management provides traffic, link and routing management, as well as, congestion (flow) control.

SS7 User Parts

The layers above layer 3 of the SS7 stack consist of several different protocols. These protocols are called user parts and applications parts. They are the following.

ISDN User Part (ISUP)

The Integrated Services Digital Network User Part is used throughout the PSTN to provide the control information necessary for the set-up and tear-down of all circuits, both voice and data. Wireless networks also make use of ISUP to establish the necessary switch connections into the PSTN. ISUP offers two types of services, known as Basic and Supplementary services. Basic Services

consist of services for setting up and tearing down of a call to the target number. Supplementary Services consist of services employed in passing all messages that may be necessary to maintain and/or modify the call.

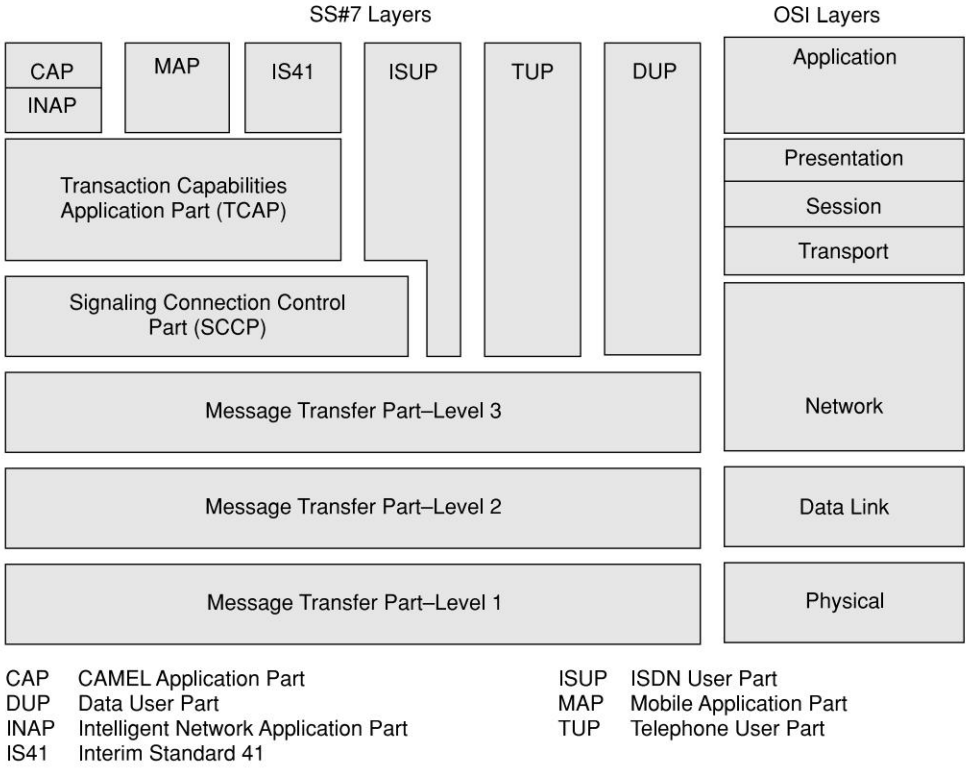


Figure 11.4 SS#7 Protocol Stack

Supplementary Services

The basic service provides the functionality for establishing circuits within the network. Supplementary services are all the other circuit-related services. These typically are those services where the messages are transported after the call path is established. Following are examples of supplementary services.

Call Forwarding: This supplementary service enables incoming calls to be automatically redirected to some other telephone depending upon the following conditions. This facility is also known as "call transfer or follow me" service. When the terminating switch receives the ISUP IAM message, it can forward the call depending upon conditions set in the SSP. Four different call forwarding services are available; all may be active at any one time.

- Unconditional: All calls will be forwarded to another telephone number unconditionally.
- On No reply: Calls will be forwarded when the subscriber does not reply. This can also be qualified by how many ringing tones will be supplied before forwarding. For example, if set

to 5, then the telephone will ring five times and then the call will be forwarded to the number specified in the switch.
- On Busy: Calls will be forwarded to the specified number when the subscriber line is busy.
- On Not Reachable: Calls will be forwarded to the specified number when the subscriber is not reachable or out of coverage area. This condition is valid only for mobile networks.

The target number can be any valid telephone number and can be a mobile phone (GSM, CDMA), a fixed phone (BSNL, Touchtel) or even a SCP offering some service like Voice Mail service. In many exchanges by default calls are diverted to voice prompt. For example if a mobile phone is powered off, the exchange plays a message like "number you have dialed is switched off": or from a fixed phone exchange "the number you have dialed has been changed". When a telecom company plays such voice prompts, the call is generally not charged. However, whenever a call is forwarded to a different number based on the conditions set by the subscriber, calls are always charged. In a GSM network, forwarding of calls can be set/reset from the subscriber's phone. For example, I am traveling through a highway where I may not get GSM coverage all through. Therefore, I may like to forward a call to my office number when I am out of coverage. I enter * * 62 * 08025531234 # . Here 62 is the command for forwarding and 08025531234 is the number the call will be forwarded to. In GSM, commands for unconditional forward is 21, forward for no reply is 61 and forward for busy is 67.

In a fixed line exchange, the follow me facility can be invoked by dialing different commands from the phone. I can register for call transfer on no reply by dialing "1228 26632245". In such a case when a call comes in my fixed line phone, and I do not reply to the call within 30 seconds, the call will be diverted to 26632245.

Call Barring: This service allows a user to bar various categories of outgoing/or incoming calls based on the following conditions.
- All Outgoing International Calls.
- All Incoming Calls.
- All Incoming Calls while Roaming.

Like call forwarding, call barring can also be set from the subscriber's phone. If I want to bar all outgoing calls from my phone, I enter * 33 * 0000 #. To release the barring I need to enter # 33 * 0000 #. In India call barring is possible for STD and ISD lines. Also, passwords can be set for calls.

Voice mail: Voice mail is a supplementary service where the caller is forwarded to a voice mail service when the call does not mature. For mobile networks this can be due to one of the four conditions as explained in call forwarding. However, for fixed networks only the first three conditions are valid.

Multiparty call conferencing: In a multiparty service the served mobile subscriber is in control of one active call and one call on hold, both calls having been answered. In this situation the served mobile subscriber can request the network to begin the multiparty service. Notification will be sent towards the served mobile subscriber and all the remote parties in a multiparty call. A notification will always be sent to all remote parties every time a new party is added to the multiparty call. Notifications shall also be sent to remote parties when they are put on hold and when they are retrieved in accordance with normal Call Hold procedures.

Caller line ID: This supplementary service offers the facility by which the telephone number of the calling party is displayed on the phone. Mobile phones generally display the name of the caller. This is possible if the caller's telephone number is stored in the address book of the mobile phone. Some phone manufacturers also offer a facility to associate icons, ringing tones specific to caller, user groups. The ISUP IAM message contains the caller's telephone number. The same is passed on to the phone.

Alternate line service: Alternate Line Service offers a subscriber the convenience of two phone numbers for one mobile phone. These are useful when a subscriber would like to have different numbers for different reasons, but would not like to carry multiple phones. For example, one number is for business calls and the other one is for personal calls. Each number may have its own associated ring tone.

Closed user group: The Closed User Group (CUG) Supplementary Service enables subscribers, connected to a PLMN and possibly also other networks, to form closed user groups (CUGs) to and from which access is restricted. Members of a specific CUG can communicate between each other but not, in general, with users outside the group. The network shall provide a subscription option in order to enable the user to specify a preferential CUG, for each basic service group included in at least one of the CUG(s).

Call Waiting: This feature allows a customer who is already in conversation with another to be informed by a Call Waiting tone that another call is waiting. The calling party will hear a ringing tone instead of an engage tone.

Telephone User Part (TUP) Data User Part (DUP)

TUP handles analog circuits whereas digital circuits and data transmission capabilities are handled by the Data User Part. These services are no longer in use worldwide. In some countries (e.g., China, Brazil), the Telephone User Part (TUP) is used to support basic call setup and tear-down. In most parts of the world, ISUP is used for voice and data call management.

Signaling Connection Control Part (SCCP)

The SCCP provides connectionless (class 0) and connection-oriented (class 1) network services and extended functions including specialized routing (GTT-global title translation) and subsystem management capabilities above MTP Level 3. A global title is an address (e.g., a dialed 800 number, calling card number, or mobile subscriber identification number), which is translated by SCCP into a destination point code and subsystem number. A subsystem number uniquely identifies an application at the destination signaling point. SCCP is used as the transport layer for TCAP-based services.

Transaction Capabilities Application Part (TCAP)

The TCAP offers its services to user-designed applications as well as to OMAP (Operations, Maintenance and Administration Part) and to IS41-C (Interim Standard 41, revision C) and GSM MAP (Global Systems Mobile). TCAP supports the exchange of non-circuit related data between applications across the SS7 network using the SCCP connectionless service. Queries and responses sent between SSPs and SCPs are carried in TCAP messages. TCAP is used largely by switching

locations to obtain data from databases (e.g., an SSP querying into an 800 number database to get routing and personal identification numbers) or to invoke features at another switch (like Automatic Callback or Automatic Recall). In mobile networks (IS-41 and GSM), TCAP carries Mobile Application Part (MAP) messages sent between mobile switches and databases to support user authentication, equipment identification, and roaming.

Intelligent Network Application Protocol (INAP)

The INAP specifies the information flows between different entities of the IN functional model in terms of protocol data units (PDUs). The PDUs represent remote operations in the scope of TCAP.

Customized Applications for Mobile Network Enhanced Logic

The CAMEL (Customized Applications for Mobile network Enhanced Logic) is a mechanism to help the network operator to provide the subscribers with the operator specific services when roaming. To support prepaid roaming services, CAMEL will be required.

CAMEL Application Part

CAMEL Application Part (CAP) is the application part based on CAMEL version 2. CAP is based on a sub-set of the Capability Set–1 (CS1) core INAP.

Mobile Application Part (MAP)

MAP is a protocol that enables real-time communication between nodes in a mobile cellular network. This application part defines protocols to exchange subscriber related information from one cellular network to another cellular network. A typical usage of the MAP protocol would be for the transfer of location information from the VLR (Visitor Location Register) to the HLR (Home Location Register).

IS-41

Interim service, IS-41 is the counterpart of GSM-MAP for the US cellular networks. IS-41 is used to offer seamless roaming to the subscribers. One area in which GSM-MAP and ANSI-41 transport differ is in the area of roamer administration. GSM-MAP networks rely on an International Mobile Station Identifier (IMSI), as opposed to the Mobile ID Number (MIN) used in IS-41.

11.4.2 SS7 Signal Unit

Within a SS7 network, signaling information is passed in the form of messages. These messages are called signal units. Signal units are continuously transmitted on any link in both directions. SS7 uses three different types of signal units:

- Fill-In Signal Units (FISUs)
- Link Status Signal Units (LSSUs)
- Message Signal Units (MSUs)

Figure 11.5 depicts different types of signal units.

Fill-In Signal Unit

FISUs is at the lowest level of service and does not carry any information. FISUs are transmitted when the link is idle and there is no payload (LSSUs or MSUs) to be transmitted.

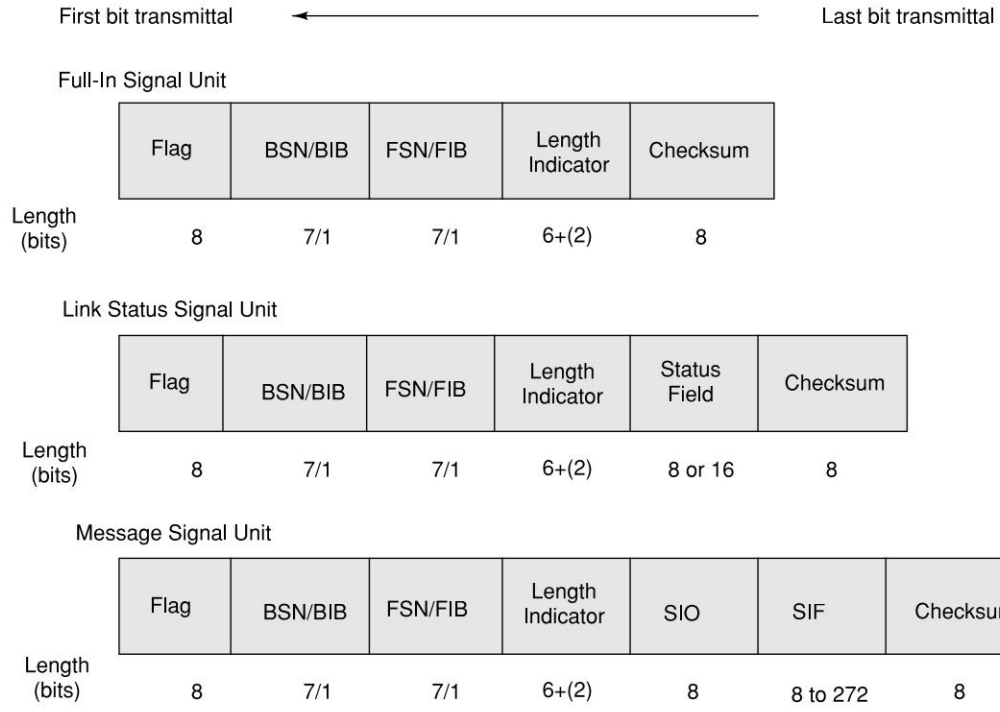


Figure 11.5 Different Types of Signal Units

Link Status Signal Unit

LSSUs transact information about the SS7 signaling link between nodes on either end of the link. When a link is identified to have failed, the signaling point that detects the error is responsible for alerting its neighboring signaling points. This information is contained in the status field of the signal unit. LSSUs signal the initiation of link alignment, quality of received traffic, and status of processors at either end of the link. LSSUs do not require any addressing information because they are only sent between signaling points. LSSU never get broadcast throughout the network.

Message Signal Units (MSU)

MSUs are the real workhorses within the SS7 network. All signaling related to call setup call and tear-down, all signaling related to database query and response use MSU. Management functions of SS7 use MSUs. MSUs provide MTP protocol fields, service indicator octet (SIO) and service information field (SIF). The SIO identifies the type of protocol (ISUP, TCAP) and standard (ITU-TS, ANSI). The SIF transfers control information and routing label. The functionality of the MSU is defined through the contents of the service indicator octet (SIO) and the service information

fields (SIF). The functionality of the MSU is defined through service indicator octet (SIO) and the service information fields (SIF). The SIO is an 8-bit field that contains three types of information: 4 bits to indicate service indicator (0–signaling network management; 1–signaling network testing and maintenance; 2–SCCP; 3–ISUP), 2 bits to indicate national (proprietary within a country) or international (ITU standard) message; remaining 2 bits to indicate priority with 3 being the highest priority. The service information field (SIF) defines the information necessary for routing and decoding of the message. SIF transfers control information and the routing label used in MTP Level 3. The routing label consists of the destination point code (DPC), originating point code (OPC) and signaling link selection (SLS) fields.

The common fields in all the signal units are as follows:

Flag: This contains a fixed pattern 01111110 and is used for clock synchronization

BSN: Backward Sequence number.

BIB: Backward indicator bit.

FSN: Forward Sequence number.

FIB: Forward indicator bit.

Length indicator: Out of 8 bits only 6 bits are used to indicate the length.

11.5 IN Conceptual Model (INCM)

INCM was developed to provide a framework for the design and description of each capability set and the target IN architecture. In an IN scenario there are four main actors. The first, the service user, is the end-user of the service. For example, this is the person who calls a free-phone or utilizes his calling card to call a friend while he is roaming. The second, the service subscriber, is the actor who subscribes to an IN feature. He can use it for himself or provide it to his customers. This could be an individual, a corporation, or a virtual service provider. The third, the service provider creates, deploys and supports IN services. This actor has contracts with service subscribers. These contracts specify the billing and the subscribed features. The fourth and last one is the network operator. This actor provides the infrastructure needed to support IN services. This actor has contracts with service providers. INCM captures the whole engineering process of the IN.

The INCM is structured into four planes (Fig. 11.6) as follows:

- Service plane.
- Global functional plane.
- Distributed functional plane.
- Physical plane.

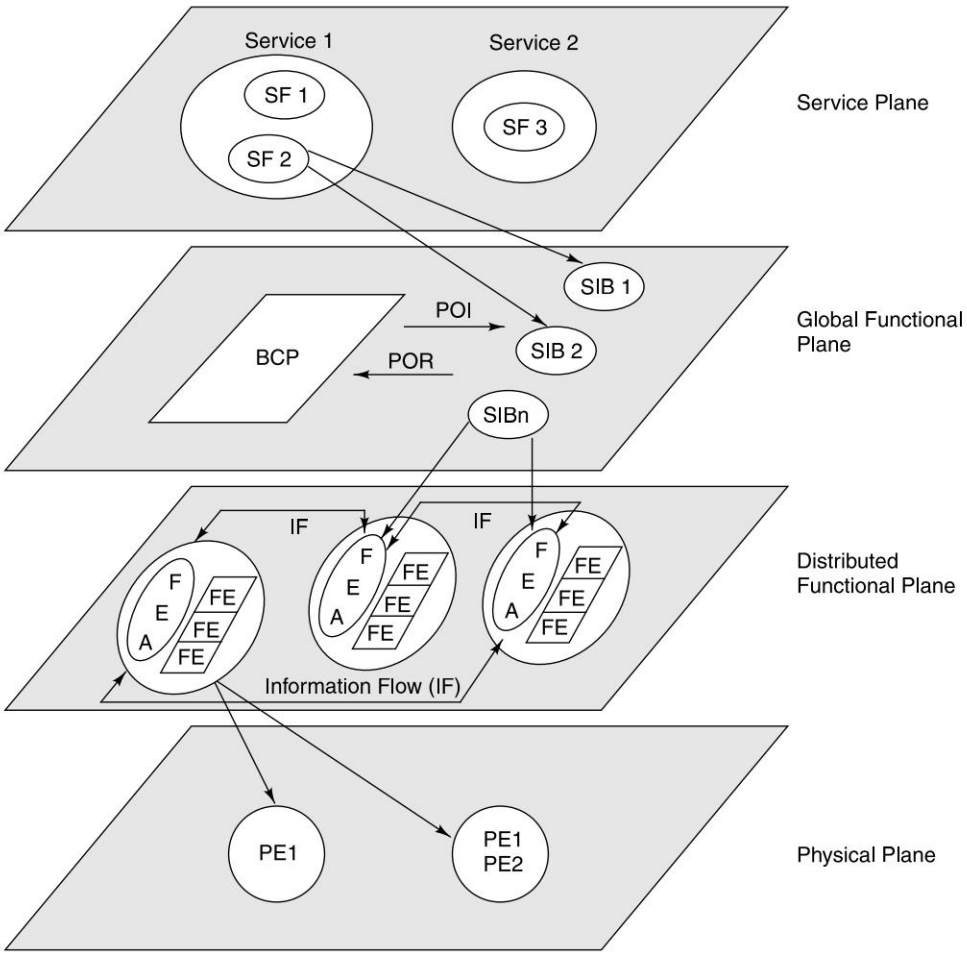
The upper two planes focus on service creation and implementation, whereas the lower two planes address the network and physical needs.

Service Plane (SP)

This plane is of primary interest to service users and providers. It describes services and service features from a user perspective, and is not concerned with how the services are implemented within the network.

Global Functional Plane (GFP)

The GFP is of primary interest to the service designer. It describes units of functionality, known as service independent building blocks (SIBs) and it is not concerned with how the functionality is distributed in the network. Services and service features can be realized in the service plane by combining SIBs in the GFP.



BCP	Basic Call Process
EF	Elementary Function
FE	Functional Entity
FEA	Functional Entity Action
IF	Information Flow
P	Protocol

PE	Physical Entity
POI	Point of Initiation
POR	Point of Return
SF	Service Feature
SIB	Service Independent Building Block
⟶	Pointer

Figure 11.6 IN Framework

Distributed Functional Plane (DFP)

This plane is of primary interest to network providers and designers. It defines the functional architecture of an IN-structured network in terms of network functionality, known as Functional Entities (FEs). SIBs in the GFP are realized in the DFP by a sequence of Functional Entity Actions (FEAs) and their resulting information flows.

Physical Plane (PP)

The PP is of primary interest to equipment providers. It describes the physical architecture for an IN-structured network in terms of Physical Entities (PEs) and the interfaces between them. The functional entities from the DFP are realized by physical entities in the physical plane.

11.5.1 Examples of IN Services

In this section we consider two IN services. First one is calling card service defined within the scope of Capability Set 1 (CS-1). The second one is the ultimate test for IN functionality called local number portability.

Virtual Calling Card Service

In a wireline telephone network, calling card services offer the possibility of making a call from any phone and charge the call to the user who makes the call, rather than charging the subscriber of the telephone line. The call needs to be set up without charging any of the lines involved in the call. This service is also known as Automatic Alternative Billing Service (AABS), account card calling, virtual card calling or credit card calling services. The distinction is made depending on whether a real physical card is being used (thus requiring a card reading terminal), or whether a "virtual" card is being used (requiring the user to dial an account/PIN number). In a cellular network the prepaid system also uses similar principles within the network; however, in the user interface the user need not enter the account information or PIN for every call that is made. In a cellular network, authentication is done using the MSISDN number of the phone.

Consider a user who wants to make a call using a virtual calling card. The processing steps for this call is depicted in Fig. 11.7 and will be as follows:

1. The user dials the service access code for the AABS (e.g., 1-600-234123). When connected to the service, there are voice prompts helping the user to navigate through the service menu. Then it asks the user to enter the destination phone number the user wants to talk to. The user enters the number of the target telephone (e.g., 011-2549-9229).
2. The switch recognizes that the call is an IN call and the service switching point (SSP) sends an INAP query containing call information to the corresponding SCP. On receipt of the query the SCP starts the corresponding service logic program. The SCP determines an appropriate Intelligent Peripheral (IP) to query for the account code and PIN of the user for validation.
3. The SCP returns to the SSP a routing number of an appropriate IP and instructs the SSP to establish the connection to the IP.
4. The SSP routes the call to the IP and instructs the IP to start an appropriate dialogue with the user.
5. The IP asks the user for PIN.
6. The user enters the PIN, and the IP collects the response digits.

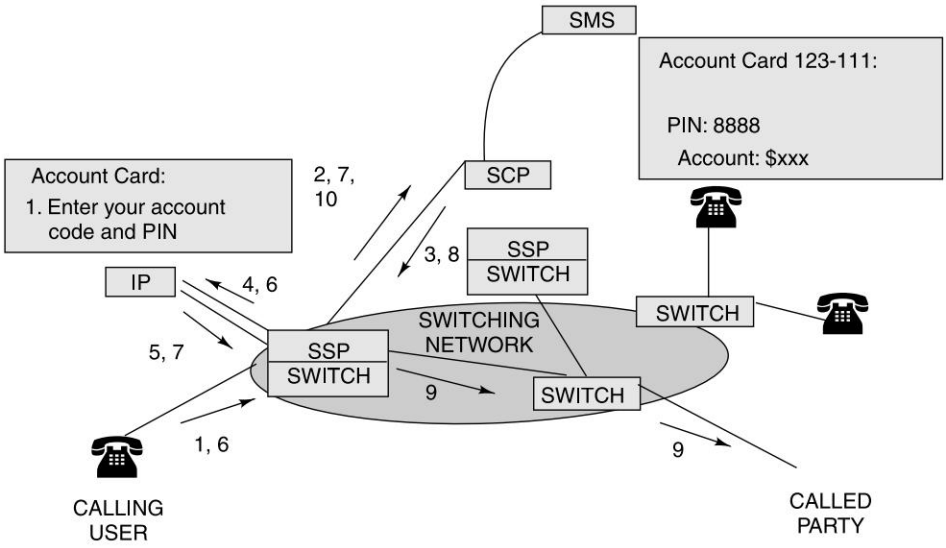


Figure 11.7 Automatic Alternative Billing Service Processing

7. The IP returns the information to the SSP. The SSP relays the account code and the PIN to the SCP. The SCP examines the PIN and checks that the account limit has not yet been exceeded.
8. The SCP instructs the SSP to disconnect the IP and to establish the connection to the destination number.
9. The SSP disconnects the IP and instructs the switch to establish a connection to the desired destination.
10. After the call is terminated, the SSP informs the SCP about the call charges. The SCP subtracts the charges of the call from the subscriber's balance.

Local Number Portability

The intent of Local Number Portability (LNP) is to open up local telephone service to competition. LNP was defined in the US in the Telecommunication Act of 1996 as the "ability of users of telecommunication services to retain, at the same location, an existing telecommunication number without impairment of quality, reliability or convenience when switching from one telecommunication carrier to another". The regulators think that the biggest roadblock to competition in the telecom vertical is the ownership of the telephone numbers. Subscribers are reluctant to switch to a new service provider because they have to give up their existing telephone number when they switch to a new service provider. There are three phases of LNP.

Phase one is service provider portability. This allows a subscriber to select a new service provider while keeping their existing telephone number.

Phase two is service portability. This allows subscribers to change the type of service they have while keeping their telephone number. For example, if the subscriber changes from POTS (Plain Old Telephone Service) to ISDN (Integrated Service Digital Network) he or she has to obtain a new telephone number, because the switching equipment used to provide the ISDN service supports a different block of numbers. With LNP the subscriber does not have to give up the telephone number when changing the type of service.

Phase three is the third and most difficult phase, location portability. This will allow a subscriber to move from city to city or even state to state within the US while maintaining the same telephone number.

11.5.2 Wireless Intelligent Network

Wireless intelligent network (WIN) is a concept being developed by the Telecommunications Industry Association (TIA) Standards Committee TR45.2. The charter of this committee is to drive intelligent network capabilities, based on interim standard IS-41, into wireless networks.

Intelligent networks principles and functions are very much woven with almost all the services mobile networks offer. This starts from as simple as supplementary services like caller ID to complex functions like roaming. Many new IN services are proposed as part of the CS-1 service sets. One such service is voice-activated services. This is a hands-free service targeted for vehicular conditions. In a vehicle we need hands-free eyes-free service. Therefore, if we just say "John Smith", the call will do speech recognition, look up a database, get the telephone number of John Smith and establish a call. Another IN example is selection of long distance carrier. Based upon the time of day, country, etc., we can select the long distance carrier for routing the call.

11.6 SOFTSWITCH

The word Softswitch is derived from the combination of software and switch. A Softswitch is an API framework that is used to bridge a traditional telecommunication network and IP networks. A softswitch will manage traffic that contains a mixture of voice, fax, data and video. Softswitch can be considered as a telecommunication switch where the intelligence is outside the switch and driven by software. Softswitch will be used in future for VoIP and unified call control which will have voice, data, multimedia and instant message. This is conceptually an extension of intelligent network at a much broader sense. Softswitch is expected to address all the shortcomings of traditional local exchange switches. The various elements in a softswitch architecture network are:

- Call agent (media gateway controller, softswitch).
- Media gateway.
- Signaling gateway.
- Feature server.
- Application server.
- Media server.
- Management, provisioning and billing interfaces.

Protocols that are supported by softswitches are MGCP, H.248 (Megaco), SIP, H.323, and Sigtran's suite of protocols and adaptation layers which include SCTP, IUA, M2UA, M3UA, SCUA,

etc. In general, many of the telephony protocols and data protocols are supported by softswitches including CAS, SS7, TCAP, INAP, ISDN, TCP/IP, etc.

As viewed by the IP network, a media gateway is an endpoint or a collection of endpoints. Its primary role is to transform media from one transmission format to another, most often from circuit to packet formats, or from analog/ISDN circuit to packet as in a residential gateway. It is always controlled by a media gateway controller. Media server operates as a slave to a media gateway controller to perform media processing on media streams. The signaling gateway and media gateway must be deployed at the boundary between the PSTN and the Softswitch. All other components may be located anywhere within the network that makes sense with regard to latency of access, co-location of control, and other operational considerations.

11.7 Programmable Networks

In a world where everything is networked, there is a need to be able to program network components to adapt to application requirements. We need to have better control on the quality of service, security, application-dependent routing, intelligent caching, utilization of bandwidth, support mobility and sophisticated management functionality. It is therefore necessary to be able to dynamically program the resources within a network. These types of application-specific functions need to be dynamically programmed within the network components in order to support flexible and adaptive networks. Such networks are called programmable networks. Programmable networks will also address the need to "open" the network up and accelerate its programmability in a controlled and secure manner for the deployment of new architectures, services and protocols. The separation of communications hardware (i.e., switching fabrics, routing engines) from control software is fundamental to making the network programmable. A programmable network is distinguished from other networking environments by the fact that it can be programmed from a minimal set of APIs to provide a wide array of higher level services.

11.8 Technologies and Interfaces for IN

We now know what IN is. We have gone into details of some IN applications. We have also learnt various layers and functions of user parts of SS7 stack. To develop an IN application we need to access one or more user parts in the SS7 stack. These applications will use native APIs supplied by the SS7 stack vendor. For some application, if we want to build the stack ourselves, we need to use APIs supplied by the SS7 hardware vendor. All these APIs are proprietary to the vendor. At a later time, if we want to use a stack from other vendor, or a different SS7, interface card, we need to change our application to suite a new set of vendor-specific APIs. This makes interoperability and the development cycle complex and expensive. To address these challenges, some standardization is required. Using these standards, an application can use some universally supported APIs. Interfaces like class name, function name, function parameters for these APIs will be same and will be supported by all vendors across the board. This will make an IN application independent of the lower layer vendor. In the following sections we shall discuss some of these interfaces and standards.

11.8.1 Parlay

The Parlay Group is an open multi-vendor consortium formed to develop open technology-independent APIs to access resources within a telecommunication network. Parlay integrates intelligent network (IN) services with IT applications via a secure, measured, and billable interface. Parlay is focused in defining umbrella architecture and API for Open Service Access (OSA). We have mentioned OSA in Section 7.1 as a part of 3G (Chapter 9). The OSA specifications define an architecture that enables service application developers to make use of network functionality through an open standardized interface.

Founded in 1998, The Parlay Group focused initial development of its APIs on functions such as call control, messaging and security. The Parlay Group was formed by a group of companies (BT, Microsoft, Nortel Networks, Siemens, and Ulticom, formerly DGM&S Telecom). The group first demonstrated a Parlay service in the UK and the US in December 1998.

In today's network, applications and services are part of the network operator's domain. This network-centric approach was good for specific applications. With the growth of Internet and new services, what is now needed is a solution that combines the benefits of the network-centric approach of economies of scale and reliability of telecommunication networks with the creativity and power of the IT industry. The Parlay APIs will allow services to be developed outside of the network that use the wide range of common functions at the center of the network. As networks of the future evolve, there will be a growing need to harmonize intelligence in the center of the network with intelligent devices at the network edge. This means that it should be possible to build applications, test and operate outside the network domain.

11.8.2 JAIN

JAIN is a community of companies led by Sun Microsystems under the Java Community Process that is developing Java APIs for next-generation systems consisting of integrated Internet Protocol (IP) or asynchronous transport mode (ATM), public switched telephone network (PSTN), wireless networks and intelligent networks (IN). These APIs include interfaces at the protocol level, for different protocols such as Media Gateway Control Protocol (MGCP), Session Initiation Protocol (SIP), and Transactional Capabilities Application Part (TCAP), as well as at higher layers of the telecommunications software stack. The JAIN APIs bring service portability, convergence, and secure network access to telephony and data networks. This is referred to as Integrated Networks. Furthermore, by allowing Java applications to have secure access to resources inside the network, the opportunity is created to deliver various services.

11.8.3 TINA

The IN is believed to be an essential revolutionary step towards the optimum restructuring of public networks. The Telecommunications Information Network Architecture Consortium (TINA-C) was founded to define a telecommunications information networking architecture (TINA) which would enable the efficient introduction, delivery and management of telecommunication services. Also, due to the rapid convergence of telecommunications and computing, the focus of attention moved away from the physical network to a software-based system. The kind of services to be supported by the TINA ranges from voice-based services to multimedia, multiparty services, the

latter of which cannot be properly supported by current IN architectures. All these software-based applications will run on a distributed hardware platform, which hides any distribution concerns from applications.

11.9 SS7 Security

SS7 network is a data network used privately by the telecommunication operators. Though it is a global network, it can be considered to be private. Unlike the Internet, anybody cannot connect a device into this network. As it is private, it is considered secured. There has not been any record of attack on a SS7 network. However, this does not imply that the security in SS7 network is robust and impossible to break. With deregulation of the telecommunication industry, network operators are obliged to allow private operators to install their IN nodes in the network. Also, with VoIP, Internet and other packet networks will be connected to the SS7 network. This makes it vulnerable to security attacks. However, as of today; it is quite secured.

11.10 MAPSec

In GSM and UMTS networks, MAP protocol plays a central role in the signaling communications between the Network Elements (NEs). User profiles, authentication, and mobility management in these networks are performed using MAP. Due to its critical role in the authentication process, operators are concerned about lack of security in MAP. Therefore, MAPSec security protocol has been designed to secure MAP messages. MAPSec has borrowed the notion of a security association (SA) from IPsec. The SA contains cryptographic keys but in addition it contains other relevant information such as key lifetimes and algorithm identifiers. The plaintext MAP message is encrypted and the result is put into a "container" in another MAP message. At the same time a message authentication code covering the original message, is included in the new MAP message. MAPSec has three protection modes: no protection, integrity protection only, and encryption with integrity protection.

11.11 Virtual Private Network (VPN)

For a long time enterprises used to lease circuits (through dedicated telephone links) from telephone companies to setup private communications channels between its offices. It used to take long time and high cost to setup such circuits through national and international exchanges; sometime these even needed drawing a physical pair of wire. As these are dedicated point-to-point circuits, they were reliable and secure. A customer could use these leased circuits in the same way they used physical cables in their premises.

Virtual private network (VPN) can be defined as connectivity between sites on a shared infrastructure with similar access and security policies as a private network—it provides a user same capabilities as private leased lines using the shared public infrastructure like the internet. VPN technology can be classified into two categories, viz., Trusted VPN, and Secure VPN. Some vendors even offer a VPN that is a combination of these two VPN technologies as Hybrid VPN.

In IP, one packet is independent of the previous packet and may follow a completely different route compared to its predecessor with risks of spoofing or sniffing. However, in VPN it is not so; in trusted VPN routing paths for each packet is similar to a circuit in private networks. This is sometimes achieved by configuring private routing paths through trusted backbone switches. A VPN also allows customers to have their own IP addressing and security policies. These are called trusted VPNs that protect the traffic at layer 1 to 3 of OSI reference model to offer a trustworthy and reliable QoS with following technologies:

- Optical based VPNs – based on Synchronous Digital Hierarchy (SDH) and Dense Wavelength Division Multiplexing (DWDM) technologies. Modern SDH/DWDM optical networks provision users with fixed bandwidth channels.
- Transport of layer 2 frames (L2F) over MPLS; this is also referred as "layer 2 VPN".
- MPLS (Multi Protocol Label Switching) with constrained distribution of routing information through BGP (Border Gateway Protocol – RFC1771); this is also referred as "layer 3 VPN" Trusted VPN can further be divided into two categories, viz.,
  - Overlay VPN that is deployed via private trunks across a service provider's shared infrastructure. Here the service provider provides the customer with a set of simulated virtual leased lines that are generally called virtual circuits (VC) that could either be constantly available as private virtual circuits (PVC) or set up on demand as switched virtual circuits (SVC).
  - Peer-to-peer VPN is a simple routing scheme where both provider and customer network use the same network protocol and all the customer routes are carried within the core network. The PE (Provider Edge) routers exchange routing information with the CE (Customer Edge) routers to setup a virtual circuit between the CE and PE.

Trusted VPNs offered reliability and QoS, but do not offer data security like confidentiality or integrity. Secured VPN offers an end-to-end secured channel over a untrusted network. This type of VPN allowed traffic to be encrypted between endpoints similar to a secured tunnel between networks and endpoints. Even if an attacker can sniff the traffic, he cannot interpret it or change it without being detected by the receiving party. For secure VPNs, technologies are

- IPsec (see Chapter 4) with encryption
- L2TP (Layer 2 Tunneling Protocol) inside of IPsec. L2TP (RFC3193) is a protocol that tunnels PPP (Point-to-Point Protocol) traffic over a variety of networks. For end-to-end security, IPsec or TLS (RFC4346) can be used inside the tunnel.

VPN devices and implementation

VPN devices in the customer network fall into one of the following two categories:

- Customer (C) devices – C devices are simple devices such as routers and switches which are located within the customer network. These devices do not have direct connectivity to the service provider network. Actually, C devices are not aware of the existing VPN.
- Customer Edge (CE) devices – CE devices are located at the edge of the customer network and connect to the provider network {via Provider Edge (PE) devices}. In CE based VPNs, CE devices are always aware of the existing VPN. In PE based VPNs, CE devices are unaware of the existing VPN. CE devices are either categorized as Customer Edge routers (CE-r) or Customer Edge switches (CE-s).

For a multi-site VPN, devices in the service provider network, also, fall into one of the two categories as below:

- Service Provider (P) devices – P devices are devices such as routers and switches within the provider network that do not directly connect to customer networks. P devices are always unaware of customer VPNs.
- Service Provider Edge (PE) devices – PE devices connect directly to customer networks via CE devices. PE devices are aware of the VPN in PE based VPNs, but are unaware of the VPN in CE based VPNs. There are three types of PE devices: Provider Edge routers (PE-r), Provider Edge switches (PE-s) and Provider Edge devices that are capable of both routing and switching (PE-rs).

Figure 11.8 illustrates Customer and Provider Network Devices. Note the presence of the VPN backbone serving three different sites.

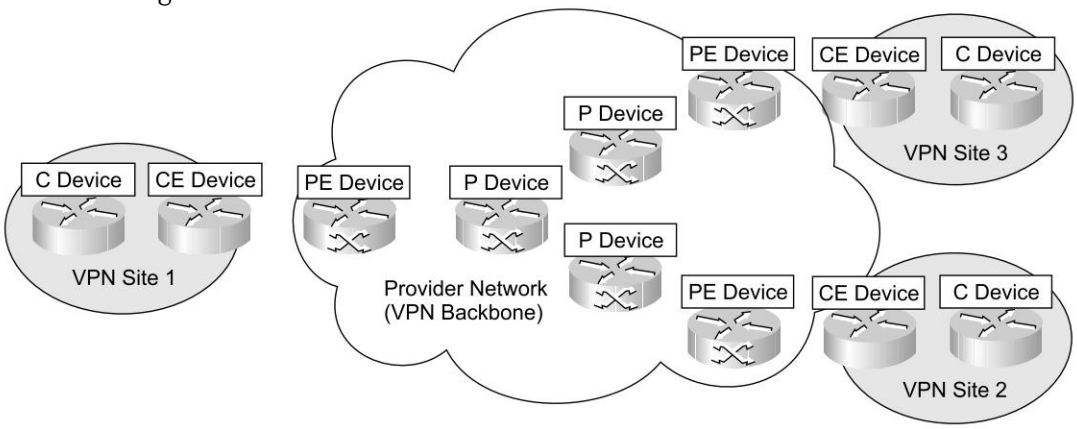


Figure 11.8 Customer and Provider Network Devices

There is one Open VPN project (http://openvpn.net) which is an open source SSL VPN solution that accommodates a wide range of configurations. Open VPN is an open source VPN program for creating point-to-point or server-to-multi-client secure tunnels. It has the capability of establishing direct links between computers across Network Address Translators (NATs) and firewalls.

Telecom VPN

In telecommunications, especially in cellular networks another type of VPN is available. In this VPN, a part of the mobile network functions like a private network where additional functionalities are available; for instance, four digit numbers to call each other within the network like calling an extension within the private company network. Even the billing principles within members of this network is different. Mobile VPNs are also designed for wireless environments and provide an access solution for mobile users who require secure access to information and applications over diverse wired and wireless networks.

REFERENCES/FURTHER READING

1. American National Standard T1.111.8-2001 (T1.111a-2002), Numbering of Signalling Point Codes.
2. ETSI 300 374-1, Intelligent Network (IN); Intelligent Network Capability Set 1 (CS1); Core Intelligent Network Application Protocol (INAP); Part 1: Protocol specification, 1994.
3. ETSI TS 100 907 V7.1.0 (1999-08): Digital cellular telecommunications system (Phase 2+); Man-Machine Interface (MMI) of the Mobile Station (MS) (GSM 02.03 version 7.1.0 Release 1998).
4. ETSI TS 100 517 V7.0.0 (1999-08): Digital cellular telecommunications system (Phase 2+); MultiParty (MPTY) Supplementary Services Stage 1 (GSM 02.84 version 7.0.0 Release 1998).
5. ETSI TS 100 518 V7.0.0 (1999-08): Digital cellular telecommunications system (Phase 2+); Closed User Group (CUG) Supplementary Services Stage 1 (GSM 02.85 version 7.0.0 Release 1998).
6. ETSI EG 201 781, Intelligent Networks (IN); Lawful Interception, 2000.
7. GPS: http://www.aero.org/publications/GPSPRIMER/.
8. *Intelligent Network (IN)*, The International Engineering Consortium.
9. Lorenz G., T. Moore, G. Manes, J. Hale and S. Shenoi, 'Securing SS7 Telecommunications Networks'; *Proceedings of the 2001 IEEE, Workshop on Information Assurance and Security United States Military Academy*', West Point, NY, 5 6 June 2001, ISBN 0-7803-9814-9.
10. Martikainen Olli, Juha Lipiäinen and Kim Molin, *Tutorial on Intelligent Networks*, IFIP IN Conference 1995.
11. Nicoll Stéphane (2001), Overview of Intelligent Networks.
12. Parlay: Draft ETSI ES 202 915-1 V0.0.4 (2003-03): Open Service Access (OSA); Application Programming Interface (API); Part 1: Overview.
13. *Prepaid Wireless Service Billed in Real Time*, The International Engineering Consortium.
14. Travis, Russell (2000), Signaling System#7, McGraw-Hill.
15. *Simply SS7*, ADC NewNet Products, 2001.
16. Standards Coordination Document No. 4, Intelligent Networks, Working Group Standards Coordination Permanent Consultative Committee I, 2000.
17. Sivagnanasundaram Suthaharan (1997), 'GSM Mobility Management Using an Intelligent Network Platform', Ph.D Thesis, University of London.
18. VocalTec Softswitch Architecture Series 3000.
19. *Wireless Intelligent Network (WIN)*, The International Engineering Consortium.
20. www.cisco.com
21. www.ja.net
22. www.openvpn.net
23. www.wikipedia.org
24. W3C CC/PP Working Group: http://www.w3.org/Mobile/CCPP/.

REVIEW QUESTIONS

Q1: Describe the steps involved in a telecom call setup.

Q2: What is an intelligent network? When does a telecommunication network becomes "intelligent"?

Q3: Describe the applications of intelligent networking.

Q4: What are the elements in a SS7 signaling network? If we compare a SS7 signaling network with the IP network, which elements are unique and which elements are similar?

Q5: Describe SS7 signaling protocol stack.

Q6: What is an SCP? What are its functions? Explain how do we use SCP to implement virtual calling card facility?

Q7: What is number portability? How is number portability different from telephone portability?

Q8: What are supplementary services? Explain with few examples.

Q9: Describe each of the following with respect to their prime functionalities:

(a) IN Conceptual Model
(b) Wireless Intelligent Network
(c) Softswitch
(d) Parlay
(e) JAIN
(f) TINA

Q10: What is Virtual Private Network? Why is it called virtual? How is it different from a private network?

Q11: What are the advantages and disadvantages of using Private Network against Virtual Private Network?

Q12: What is the motivation for using VPN? What are the tradeoffs involved in using a solely private network?

Q13: How can the VPNs be classified? Explain in detail each of them.

Q14: Explain implementing a VPN in a L2TP manner.

Q15: What are the main devices used in implementing a VPN? Explain with the help of an example.

CHAPTER 12

Client Programming

12.1 INTRODUCTION

We would like to begin by reminding ourselves that "information is not knowledge, knowledge is not wisdom and wisdom is not foresight. Each grows out of the other and we need them all", Arthur C. Clark (1997). In the same vein hardware, software and networks together make powerful applications possible. Each is not more or less important than the other.

12.2 MOVING BEYOND THE DESKTOP

This chapter aims to give us a quick overview of the current handheld computing landscape, its evolution and profitable usage. Mobile devices have traditionally been classified as phones, pagers, and personal data assistants (PDAs). Initially each had well-defined roles: cell phones provided communication capabilities similar to wired phones; pagers provided text messaging; PDAs provided portable data applications such as contacts, calendars and notes. What is interesting, however, is the way technology is emerging at the present time. These seemingly independent streams are now merging and we have device integration. Most cell phones now include address books and SMS (Short Message Services). Some pagers include e-mail access and almost all contemporary PDAs include communication capabilities. Devices today offer many permutations and combinations of these features leading to, "one-size-fits-all" devices. A historical perspective of evolution helps to comprehend the growth and direction of technology. We shall, therefore, briefly review two independent streams of developments, one leading to mobile phones and the other to PDAs. The emphasis here will be on programming aspects but to understand the environment we will take a quick peek at the underlying OS.

Let us begin with mobile phones. Briefly, wireless communications are enabled by packet radio, spread spectrum, cellular technology, satellites, infrared line of sight and microwave towers, and can be used for voice, data, video, and images. As we have already seen, a cell phone is an extremely

sophisticated radio. To fully comprehend the present-day technology and features, it helps to first understand how these evolved over time.

Time	1920–1960	1960–1970	1970–1980	1980–1990	1990–2000	2000–2010 and onwards
Devices	Marine radio and vehicle mounted telephone	• Shoe phone • Briefcase cell phone • Bat mobile phone	• First hand held cell phone • NMT hand held • Car phone Tokyo	Nokia Ericsson Motorola NMT/PCS/GSM hand sets	Nokia Communicator	iphone, Nokia N97, Nokia E72, Nokia N 900, Blackberry, Samsung Omnia
Carrier Technology	Analog radio	IMTS	Cellular Analog Systems	PCS/GSM	GPRS/3G/ UMTS/CDMA	3G/3.9G/4G/IMS/ HSDPA/Mobile TV
Key features	Based on tubes, large and bulky basic voice-only devices	Transistors allowed for miniaturization but still voice-only	Cellular concepts were deployed but large heavy voice-only systems	GSM brought digital systems and limited data capability	Sleek, light weight digital systems, cap-able of carrying multimedia data	Sleek, high battery life, huge memory capacity, large secondary storage through SD cards, 24 x 7 Internet connected, camera, runs PC applications, desktop approaching OS

Figure 12.1 Evolution of Mobile Technology

As seen in Figure 12.1, radios and vehicle-mounted telephones were the first communication devices which continued to evolve over time spanning from the 1920s to 1960s of the 20th century. The break through came with the invention of the transistors, which brought down the size, weight and subsequently the cost of the handsets. The real growth, however, started with the implementation of cellular concepts. The shift to digital technologies enabled data communications, leading to the devices of today. From huge monsters measuring 20" × 11" × 8.5" and weights close to 40 pounds that allowed only voice in the 1960s to 4.3" × 0.9" × 1.8" which allow multiple applications, multimedia, video conferencing facilities and even built-in cameras, cell phones have come a long way today. Most current phones offer built-in phone directories, calculators and even games. Many of the phones incorporate some type of PDA or Web browser.

It is but natural that the evolution of the handheld be closely mapped to the development of the networks. Thus today's mobile phones are a combination of a mobile and data device rolled into one. The evolution of the cellular networks as we have already seen can be summarized below. The move has been from voice communication to increasingly data-based communications. This is depicted in Figure 12.2.

Data over telecomm matured with the introduction of the world's first all-in-one communicator, the Nokia 9000 by Nokia in 1996. We will break at this time to see an alternate device, the PDA has that grew from a totally different requirement but have become an essential part of today's mobile computing environment.

Not only have PDAs come up as desktop replacement devices, they are also great multimedia devices with a distinction of being online always. In fact, PDAs are one of the most sought after devices and a part of the style signatures starting from the student and business community to that of the rich and flamboyant.

A PDA as the name suggests is a personal digital assistant. Its major functionality is related to storing, accessing and manipulating data. None other than the team of Startrek conceived the idea of PDA in the early 1960s. But till the later half of the 1980s this remained mostly a concept. Apple demonstrated the first device. This device was the Newton. Figure 12.3 shows the evolution of the PDA.

From a stand-alone organizer the PDA has come a long way. Most of this significant evolution is due to the progress in hardware capabilities and development of underlying bearer networks. We now take a look at the current offerings. Most handheld devices can be classified as smart phones, PocketPCs or smart communicators (a combination of the two). Figure 12.4 gives the features based on which this classification is made. However, note that the lines dividing these kinds of devices are greatly blurred and it may be difficult to identify each of them in the future.

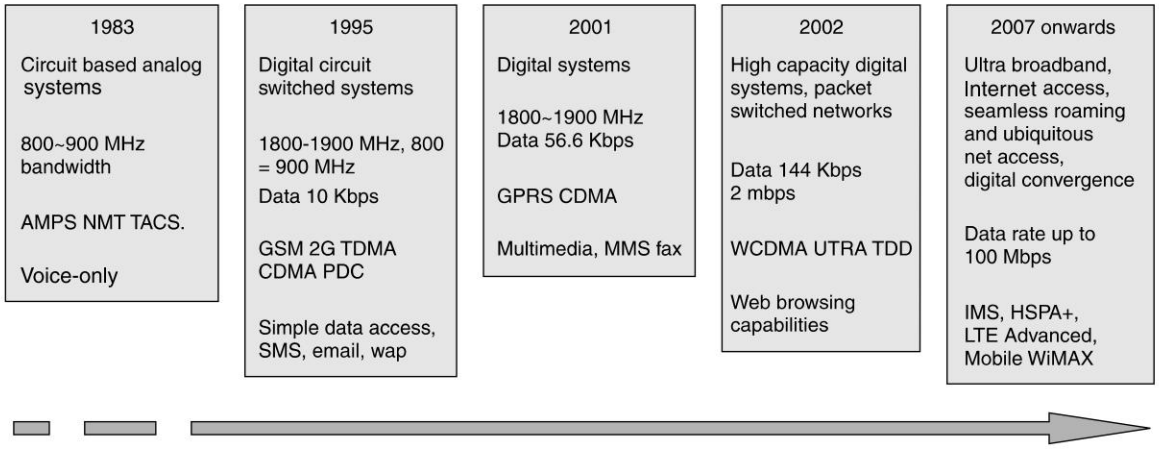


Figure 12.2 Evolution of Cellular Technology

Time	1970s–1987	1988–1992	1993	1994–1996	1997–1999	2000–2005	2005 onwards
Devices	Pison Organizer	Grid Pad, Atari, Sharp	Message Pad, Zoomer, PenPad, Envoy	Palm, Marco, MagicLinc	Sharp, Palm VII	Nokia 9210, iPaq	Nokia E72, Nokia N97, Nokia N 900, Blackberry, Treo Pro, i-mate ULTIMATE 8502
Key features	Stand alone data organizer	Handwriting recognition	Telephony application	Synchronization	Wireless link	Personal organizer cum wireless data communicator	Could run almost all desktop applications while being connected to Internet 24 x 7; Great multimedia capabilities

Figure 12.3 Evolution of PDA

Having seen where we came from, we will now see where we are going i.e., the applications. All applications have some basic requirement. The least they require is:

- some input to provide some useful output.
- memory for runtime and persistent storage.
- some communication capabilities.

Most of these are device capabilities but programming at a device level is cumbersome and generally not advisable. So we need some level of abstraction on the bare hardware. This layer of abstraction is the device operating system. Most devices provide some kind of programming environment. Palm programming is mostly done in C/C++, symbian in C++/Java and WinCE or PocketPC in embedded VB or embedded VC. Roughly the architecture is shown in Figure 12.5. We will explore each of these layers one by one.

Mobile phones	PDA	Communicator
• Mostly voice using telephone network (PCS/CDMA/GSM/GPRS/UMTS, etc.)	• PIM (personal information manager).	• A combination of the two.
• Provides phone book, CLI, messaging (SMS/EMS/MMS), limited data services.	• Network access using some external modern.	• Allows capabilities of both.
• Simple applications like email, ring tones, picture messages, etc.	• Synchronization with PC using serial cable or IR.	• Can support powerful enterprise grade applications.

Figure 12.4 Comparison of Capabilities of Different Devices

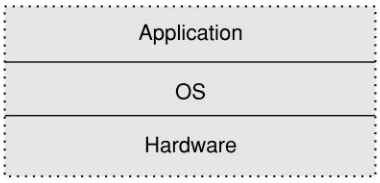


Figure 12.5 Structure of a Mobile Device

12.3 A Peek Under the Hood: Hardware Overview

As we have said before the cell phones/PDAs are among the most intricately built devices. Modern digital devices have a MIPS (million instructions per second) capability to process information (voice/data) stream. Both classes of devices essentially consist of:

- A microprocessor.

- A power source.
- A signal converter.
- An I/O unit.
- Some memory (both persistent and volatile).

Figure 12.6 shows a simplified view. As we can see, here the power supply unit (generally lithium battery) provides the required power to all the components. The processor is the brain of the device; it handles all the processing in conjecture with the memory and the Digital Signal Processor (DSP), it interfaces to the I/O unit and also the external signaling system. There is also a RF to digital converter, which is the interface to the communication channel/network. It is responsible for converting the RF (Radio Frequency) signals to digital and vice-versa. We will now take the cell phone and PDA separately.

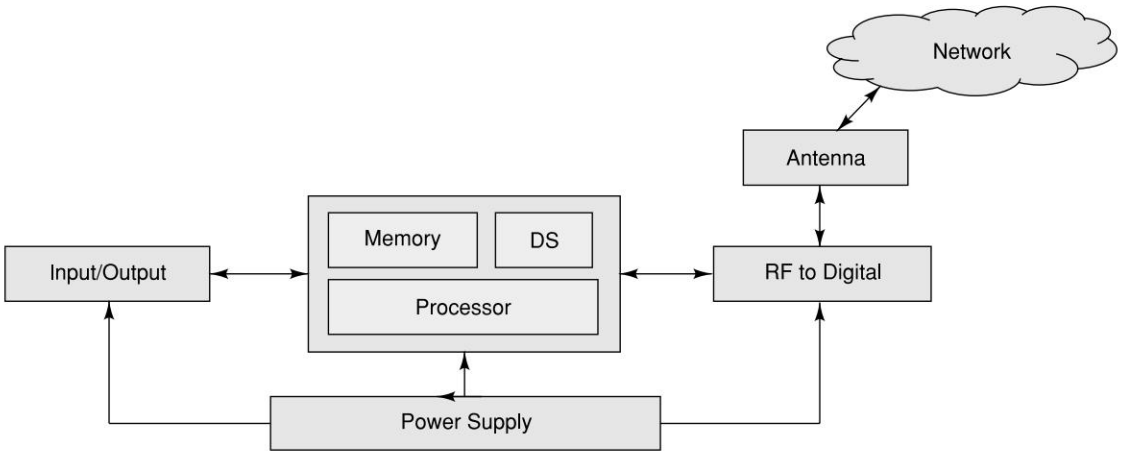


Figure 12.6 Overview of a Digital Communication Device

12.4 MOBILE PHONES

As shown in Figure 12.7, internally a cell phone consists of the following parts:

- Antenna: This is the signal reception unit.
- Circuit board: This is the control unit of the system. It has several important chips mounted on it. The prominent ones amongst these are the following.

1. The analog-to-digital and digital-to-analog conversion chips represented by the RF to Digital block in Figure 12.6. This is the entry and exit point to the phone. As the name implies, this chip is designed to convert analog network signals to digital for the phone to process and reverse.
2. The microprocessor controls and co-ordinates the handset functions. The most important of these are user input/output and network interfaces that include communication with the base station.

3. The ROM and Flash memory chips provide storage for the phone's storage requirements, which include system memory and application memory. Additional memory is also provided through external detachable memory cards.
4. The DSP (Digital Signal Processor) is a highly sophisticated chip that manages the signal manipulations.
5. The radio frequency (RF) chip manages the signal channels while the power section is responsible for power management and recharging.

- Display unit is the output unit, generally a Liquid Crystal Display (LCD) panel.
- Keyboard (qwerty or T): This is the input unit.
- Microphone: To facilitate speech transmission.
- Speaker: The microphone's listener counter part.
- Battery: The source of electrical energy.

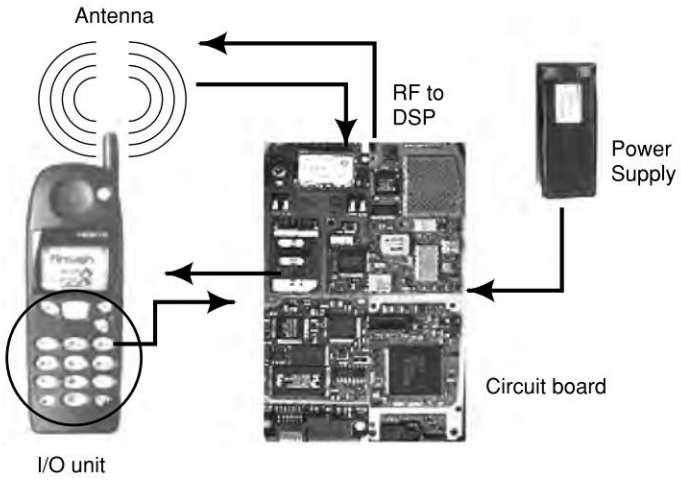


Figure 12.7 Components of a Mobile Phone

12.5 Features of Mobile Phone

Features that we will discuss in this section are the baseband architecture, processor, audio/video components, data storage and displays. With the phones coming with more and more advanced mechanism and features, it is good to understand few components in some detail and have a functional overview of them.

Although the baseband functionality is architecturally a larger concept, it is helpful in provision of cellular control logic, signal processing and user interface management while also contributing to memory and energy management. Normally, the baseband architecture in mobile phones is a conglomeration of tradeoffs between electromagnetic capability, performance, power consumption and cost. Electromagnetic capability is determined by placement of high speed circuits and RF transmitter locations along with other mechanical constructions. Performance is calibrated in terms of user requirements (like user interfaces, mobile phone execution engine, etc.), instructions

processed per second (like MIPS) and other system requirements. Power consumption is highly dependent upon system activity (like clock speed, types of ICs) and the ratio power/instruction execution speed. Cost is a function of number of components and their types or sizes, complexity involved and integration and testing time. Mobile phones now support different access technologies like GSM and CDMA in the same device while supporting multiple bands for GSM.

The mobile phone processor masters nearly all other individual functional components present in the set. There can be a set of processors in the mobile phone as well like the Master Controller Unit (MCU), Digital Signal Processor (DSP), hardware processors, etc. MCU is a general purpose processor which handles upper layers of cellular protocols apart from supervising functionalities for a general purpose system. DSP primarily handles processing intensive physical layer operations. Hardware processors generally act as an add-on for processing cellular logic, specific encryptions, 3D and system control. In the upcoming mobile devices, there can also be dedicated processors for various peripherals present and functionalities supported.

The audio/video support is becoming a most sought after feature in mobile phones nowadays. For audio, hardware needs to be supportive of listening to different radio bands. Audio hardware should also support polyphonic ring tones, speech manipulations suitable for different formats, stereo music support (as music needs different compression/decompression techniques), etc. Such hardware should manage voice coder tradeoffs and have good acoustic design with support for microphone and Integrated Hands Free (IHF) speaker while matching cellular connection speeds. For video/still images, hardware should at least have CMOS (Complementary metal–oxide–semiconductor) image sensors and optical lens assembly with logic. Then, it can also provide the mobile phone with the added functionality of view finder for more convenient imaging. The most common formats supported for still images are JPEG (Joint Photographic Experts Group) and BMP (Bit Map) while the common formats for video are H.263 and MPEG 4 (Moving Picture Experts Group version 4—described in Chapter 18). However, there are various issues for video/image capture like the following:

1. Cost optimization with respect to resolution of the camera.
2. Dust protection and operating temperature.
3. Mechanical design of the camera.
4. Object mobility.
5. Sensitivity limit with respect to available glare.
6. Noise.

Data storage in mobile phones is increasing in capacity and decreasing in terms of cost while enriching the multimedia experience for the user. The multimedia card is available in multiples of 32 MB going up to 32 GB with high speed of data storage and retrieval. There are multimedia cards available in various sizes, speeds and form factors. Such devices are optionally available in the market and generally taken as add-ons for mobile phones depending upon the need and features. Nowadays, miniature hard disks are also being seen as a mobile peripheral which can get directly attached to the phone through the cord (or Bluetooth) and used. Such hard disks are smaller in size, consume less power and are shockproof, while supporting up to 8 GB of memory capacity.

The displays in mobile phones are now supporting more than 65536 true colors. The pixel count and aspect ratio are generally locked for a particular series of mobile phones of a company. The dominant technology is Liquid Crystal Display (LCD) which is relatively thin as slab, consumes

less power, supports different optical modes (like being usable under strong sunlight and in dark room ambience) and is cheap. However, it does require a backlight as it is non-emissive. Other upcoming technologies are amorphous Silicon TFT display (aSi-TFT) and Low Temperature Poly Silicon TFT (LTPS-TFT). Thin Film Transistor (TFT) is a type of LCD which uses thin film transistor technology to improve image quality. The illumination system can be either of Light Emitting Diodes (LEDs) or Lightguide mechanism. In some mobile devices, there can be secondary displays as well as displays which can even take inputs from the user.

12.6 PDA

The PDA unlike the phone evolved from the PCs. They have a similar architecture as most desktops have, namely a microprocessor, an I/O mechanism, memory, and additionally a wireless or IR port and of course a power source, generally a battery. They typically have an operating system and mimic the applications on the desk tops, e.g., address-book e-mail, etc., and hence need to be able to synchronize with the applications on the PC. As we see, PDAs are considered a portable extension to the desktop. We will now look at the internals of a typical PDA.

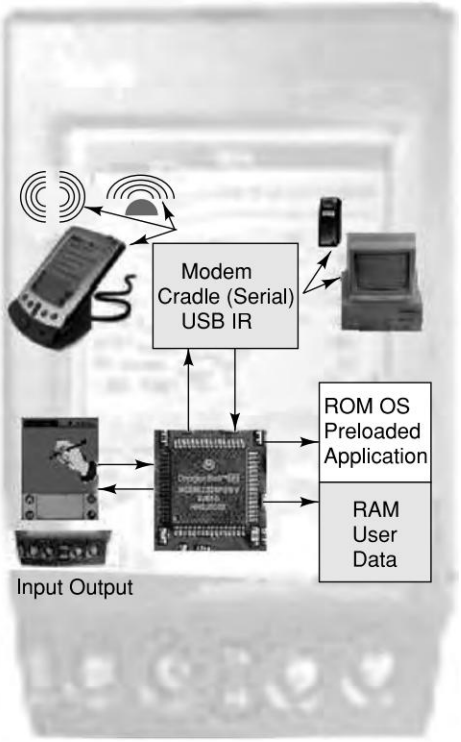


Figure 12.8 Components of PDA

Microprocessor

This is the control unit, and is responsible for initiating and coordinating the PDA's functions. Owing to intrinsic device limitations, PDAs use cheap low-end processors such as the Motorola Dragonball. These processors though very basic (about 16-75 MHz capacities) are sufficient for the requirements of a PDA. However, this is an important consideration for application developers who consistently have to ensure that they do not overload the processor leading to performance degradation and poor user experience. As we will see later, this is one of the factors which decides which part of processing should be performed on the device and what should be shifted to a backend server.

Operating System

The OS has a multitude of responsibilities very similar to those in the desktop environment. The PDA operating systems, however, are fairly simple having fewer instructions and smaller footprints. The major market players are, Palm OS, Symbian and PocketPC. We will speak of these in more details in subsequent chapters.

Memory

The PDA has no concept of a hard drive. Instead, it has solid-state memory like Static RAM or Flash memory and these days we get removable memory cards and detachable memory sticks. It is common to have about 2 MB of memory. It is used to store the OS and application programs; these remain intact even when the machine shuts down. The user data, however, is stored in RAM. We will refer to it again when we explore the details of the OS.

Batteries

Most PDAs today use rechargeable lithium, nickel-cadmium or nickel-metal hydride batteries that last for about two hours on an average. Major culprits that drain batteries are:

- Additional memory.
- Color LCD display.
- High-end features like voice recording, camera, MP3 player, etc.

PDAs use advanced power management systems to extend the battery life. It is important for application developers not to start infinite loops or polling as this can potentially interfere with the system's power management. As a last word of caution it is essential to back up data to the PC whenever possible so that even if all the data on the device gets wiped out it can be restored from the PC.

LCD Display

The Liquid Crystal Display screen of a PDA is its I/O interface and is used for both input and output. The display may be grayscale (16) or color (65,536) has resolution of 160 × 160, 240 × 320, etc. While backlight affords better reading in the dark, they consume more power, and hence, drain the battery faster.

Input Device—Buttons in Combination with Touch-Screen or Keyboard

PDAs mostly use a stylus and touch screen in combination with a handwriting recognition program. Most also have buttons to bring up frequently used applications. This is similar to using keyboard

short-cuts. Some high-end devices may actually boast of a miniature "qwerty" keyboard but these are exceptions and not the rule.

> The touch screen is an interesting piece. The set-up consists of multiple layers; on top is a thin plastic or glass sheet with a resistive coating on its bottom. This layer floats on a thin layer of nonconductive oil, which rests on a layer of glass coated with a similar resistive finish. Thin bars of silver ink line the horizontal and vertical edges of the glass by sending current first through the vertical bars and then the horizontal ones; the touch screen obtains the X and Y coordinates of the touchdown point. When the stylus touches the screen, the plastic pushes down through the gel to meet the glass (called a "touchdown"). This causes a change in the voltage field, which is recorded by the touch screen's driver software, which determines the point of contact. The driver scans the touch screen thousands of times each second and sends this data to the application that is listening for it. How does handwriting recognition work? Using a stylus, we draw on the device's touch screen. Software inside the PDA converts the characters to letters and numbers. However, these machines don't really recognize handwriting. Instead, we must print letters and numbers one at a time. On Palm devices, the software that recognizes these letters is called Graffiti. In case one finds the graffiti difficult to use, it is possible to use an onscreen keyboard. It looks just like a regular keyboard, except that the letters are tapped with the stylus.

Input/Output Ports

A PDA must be able to communicate with a PC. This communication is called data synchronization or syncing and is typically done through a serial or USB port. This can be through a cable or the cradle. These days most come with an infrared port and offer telephone modem accessories to transfer files to and from a network.

12.7 Design Constraints in Applications for Handheld Devices

We have looked at both cellphones and PDAs while communicators are a cross between the two. These vary greatly across vendors and models from the same vendor. We shall not go into the hardware for communicators.

Another breed of device that is growing very fast is the Java-enabled phone. These phones have Sun's Java virtual machine (J2me/Personal Java) embedded in them. The biggest advantage that these devices enjoy is size, OS independence and application portability. This has caused an exponential increase in the number of applications that can run on these phones.

However, what is in all these devices for application developers? Common across these devices are the following characteristics:

- Low-end processors.
- Small screen size.
- Cumbersome input device.
- Limited battery power.
- Memory limitations.

Processing Power: The processing speeds start at about 16 MHz. Heavy-duty computation like encryption key generation is a heavy drain on the device's resources. Most devices that support security provide a special processor designed for the purpose. Computation on the device should be done judiciously. Offloading computation to a backend server is always a good idea. While designing a device resident client, architects and developers always have to walk the tight rope. Too many computations on the client while allowing for faster response eat into valuable memory. Shifting everything to the server leads to poor response, which is a frustrating experience for the user. So extreme caution has to be exercised. Based on the functionality and specificity of the software, clients are of three types:

1. Thin clients: These are generic in nature and cater to a wide range of sources. They are similar to the web browsers. Communication from the server is mainly based on some flavor of Markup Language (ML). Note, however, that this requires a local parser. The server needs to send large amount of display information to the client, to adequately represent data, chocking up the transmission channel. Thus the thin client offers generality at the cost of bandwidth. This is a major consideration especially in networks where the users pay for the data and not call time.
2. Thick clients: The intelligence resides in the device and a call is made to the server only for data. Computation is done locally. While this approach resolves the bandwidth problem it introduces two more problems, namely, the size of the application on the device and distribution of the application when an update or fix happens.
3. Thin plus or semi-thick clients: These lie somewhere in between. How thick or thin is decided based on the functionality of the application, the bandwidth availability and cost constraints.

Screen Size: Most handhelds have a small screen, limiting the information that we can display at one time. Hence, screens should be designed very carefully taking care to remove all extraneous information. It is a good practice to limit scrolling to two screens below. Navigation should be easy using a single click as far as possible. A general rule of thumb is that the depth of navigation should be at most four clicks.

Cumbersome Input Devices: Most devices sport a stylus or T keypads or a very small "qwerty" keyboard where keying in long strings is a pain. Hence, inputs should be kept to the minimum. As far as possible we should try to device inputs as a single click kind of option. (Yes or No can be substituted by radio buttons as scroll and click in most devices is easier than keying.)

Application Load Time: Generally, users switch on their handhelds for short durations. For example, often it is to retrieve a contact number. The load time for applications should be low, as the users would not like to wait for the application to load each time.

Battery: Batteries are a scarce resource. One of the keys to the success of the application is long battery life. Activities like serial or IR communications, sound extended animation, and other tasks that use the CPU for long periods tend to consume large amounts of power.

Memory: All handheld devices have limited storage space, from 512K to 8MB, and a dynamic heap in the range of 32K to 256K (newer devices tend to have 8MB and 256K, respectively). Under such circumstances optimization is crucial. The optimization mantra is heap first, speed second, and code size third.

Data Storage: All devices provide some amount of persistent storage. However, this is very limited. Different devices organize this in different ways. Palm OS-based devices, for example, treat all storage as database blocks. We shall see more about it in the subsequent chapters. We will also discuss the techniques to optimize storage under different OSs.

Backward Compatibility: Backward compatibility is a key issue as the devices have evolved rather rapidly in a very short span of time and users are not expected to upgrade their system every time an advanced version becomes available. Hence, to gain wider acceptance we have to write applications in such a manner that they will run on all versions of the device and at the same time give the users of a later version better functionality.

Application Size: Applications need to be stored on the devices. Most devices will have an upper limitation on the size of executables. This needs to be taken into account during the design phase itself. Each feature should be visited multiple times with questions like "Is it necessary?", "Can it be done differently?", "Can it be clubbed with some other feature?" "Should it be done here or offloaded to a backend server?". We should try to ensure that only absolutely necessary stuff resides on the device.

Finally, there is no ideal way to design a device resident client. All clients, which are done well, evolve over a period of time involving changes and undergo several iterations before a satisfactory mix is obtained.

To conclude, in the past few years, wireless technology has seen a phenomenal and dynamic growth. This has mirrored in the growth of the device segment as well. The new technologies are enabling mobile phones to be a combined camera, video camera, computer, stereo, radio and a host of other things. This has lead to an enormous potential for applications. The development of software applications and other key technologies has enabled subscribers to use the handset to connect to the Web to receive stock quotes, check e-mail, transmit data, and send faxes and much more. Some phones support personal digital assistant software that offers the convenience of a calendar, address book, calculator, and voice recorder. We saw the evolution of these devices and looked beneath the hood to get an overview of how things work. We also saw the major concerns in writing applications for these devices. In the next chapter we will explore programming for the Palm OS. We will briefly look at the operating system to the extent required by us to write good applications. Covering all the nuances of the system may require more exploration on part of the reader.

12.8 Recent Developments in Client Technologies

There have been constant advancements of technology and capabilities of handsets, devices, and associated network bandwidths. Almost at the same pace mobile devices are becoming popular—at the end of 2009 the mobile users grew to 3.5 billion. On the other hand, the dividing line between mobile phones, PDAs or mobile computers are slowly disappearing. Even though it is the beginning, devices like Nokia N 900 and RIM's Blackberry have started supporting enterprise wide desktop applications while having huge memory capacity, automatic maintenance of Internet connectivity (be it with WLAN, 2.5G/3G, WiMAX, etc.), host of multimedia features and also, intelligent applications. There are services and portals having specific affinity to support these devices. Mobile handset major Nokia launched www.ovi.com that offers a multitude of applications, network-centric features and online support for Nokia phones. Likewise, there are many devices

available in the market that have associated websites to help the user almost plug and play. So, the players in the domain of service creation are not just the handset manufacturers and network providers; but, also independent service providers catering to various sets of use cases.

Mobile phones' OS are next only to desktop OS in terms of revenue generation. Ranging from hitherto unknown and simple operating systems, there are hugely complex OS and costly OS for phones like Symbian, Windows Mobile, Samsung's Bada, Palm OS, Blackberry OS, IPhone OS, Google's Android, etc. Google's Android which is a descendent of Linux and Open Source has gained popularity. Google, in association with Open Handset Alliance, announced the opening up of the whole source code of Android (including network and telephony stack) under an Apache license. This has led to a huge developer's community directly participating in making it more robust and user intuitive. Any kind of development with respect to Android is well supported at www.developer.android.com. Table 12.1 roughly captures its features as of early 2010.

Table 12.1 Features of Android OS as of early 2010

Handset layouts	The platform is adaptable to larger, VGA, 2D graphics library, 3D graphics library based on OpenGL ES 1.0 specifications and traditional smartphone layouts.
Storage	The Database Software SQLite is used for data storage purposes.
Connectivity	Android supports connectivity technologies including GSM/EDGE, CDMA, EV-DO, UMTS, Bluetooth and Wi-Fi.
Messaging	SMS and MMS are available forms of messaging including threaded text messaging.
Web browser	The web browser available in Android is based on the open-source Webkit application framework.
Java support	Software written in Java can be compiled to be executed in the Dalvik virtual machine, which is a specialized VM implementation designed for mobile device use, although not technically a standard Java Virtual Machine.
Media support	Android supports the following audio/video/still media formats: H.263, H.264 (in 3GP or MP4 container), MPEG-4 SP, AMR, AMR-WB (in 3GP container), AAC, HE-AAC (in MP4 or 3GP container), MP3, MIDI, OGG Vorbis, WAV, JPEG, PNG, GIF and BMP.
Additional hardware support	Android can use video/still cameras, touchscreens, GPS, accelerometers, magnetometers, accelerated 2D bit blits (with hardware orientation, scaling, pixel format conversion, etc.) and accelerated 3D graphics.
Development environment	Includes a device emulator, tools for debugging, memory and performance profiling and a plug-in for the Eclipse IDE.
Market	Like many phone-based application stores, the Android Market is a catalog of applications that can be downloaded and installed to target hardware over-the-air, without the use of a PC. Originally, only freeware applications were supported.
Multi-touch	Android has native support for multi-touch which is available in newer handsets such as the HTC Hero. The feature was initially disabled at the kernel level. Google has since released an update for the Nexus One and plans to release an update for the Motorola Droid which enables multi-touch natively.

Android is a software stack for mobile devices that includes an operating system, middleware and key applications that is shown as Figure 12.9. Some of popular handsets using Android are Dell Mini3i, Sciphone N19, HTC Dream, Samsung 17500 and XPeria X10.

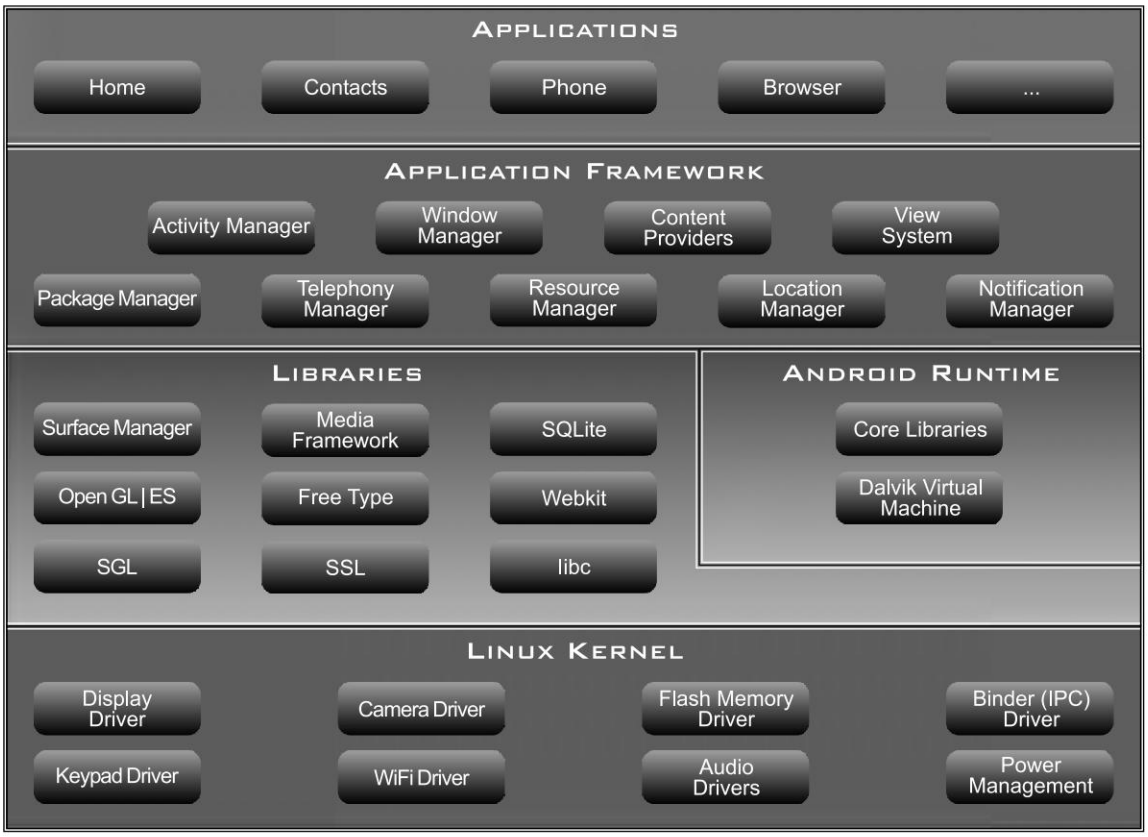


Figure 12.9 Components of Android OS

With so many non-conventional and new players getting in the realm of mobile technologies (hardware and software), it can be said, at the least, that the competition has just begun. It is going to be a competition creating and breaking lock-in and network effects. At one end is open source and inter-mingling OS (like Android) and at the other end is Apple's IPhone and RIM's Blackberry.

REFERENCES/FURTHER READING

1. A detailed technical discussion on the birth and growth of mobile telephony http://www.international-phone-card.info/mobile_telephone_history.htm.
2. An excellent collection of articles and resources for mobile computing is available at http://www.bitpipe.com.

3. Good introduction and bird's eye view on many topics covered in this book http://www.peterindia.net/MobileComputing.html.
4. Information on the cell phone internals and transmission technologies http://biz.howstuffworks. com/cell-phone.htm.
5. Information on the PDA internals and other details http://electronics.howstuffworks. com/pda.htm.

REVIEW QUESTIONS

Q1: What are the important differences between a desktop computer and a portable computer like PDA?

Q2: Write short notes on:
- (a) Evolution of mobile technology
- (b) Evolution of cellular technology

Q3: Describe the structure of a mobile handset giving insight into its primary hardware components.

Q4: What are the challenges one needs to keep in mind while designing a small footprint wireless device? Explain with an example.

Q5: What are the design constraints for applications targeted for handheld devices?

CHAPTER 13

Programming for the Palm OS

13.1 INTRODUCTION

It is interesting to note that the origin of the concept of the PDA can be traced back to 1960 when Gene Roddenberry, the late creator of Star Trek, decreed that paper and pencils were debarred from the sets of starship enterprises. All communication and data collection would be through tricoders and communicators (the name is today retained to refer to smart communication and data devices). The term PDA, however, was coined by John Sculley, then the CEO of Apple Computer during his speech at the Winter Consumer Electronics Show. Apple was the first to ship a PDA under its Newton product line. Through the second half of the 1990s various companies including IBM, Sony, Samsung, NEC and others entered the market with their own variations of a PDA, but none achieved the spectacular success of Palm.

Various comparisons and reviews both commercial and technical, favor one over the other depending on the evaluation criteria and times when these were compiled. What everyone, however accepts is that it was Palm that heralded the age of PDAs with the introduction of its PalmPilot 1000. This chapter introduces us to Palm OS and building applications for the Palm OS. The approach here as it is elsewhere is to give our readers an overview of the Operating System concepts and a guide to important APIs. Detailed discussions of the OS and programming intricacies are beyond the scope of this book. We begin by tracing Palm from its birth in 1996 and follow its turbulent journey spanning glory like none other down to its present; follow it up with a discussion about the Operating System architecture, the basics of programming and conclude with a look at the future directions planned for Palm OS.

13.2 HISTORY OF PALM OS

The travails of Palm OS closely follow the fortunes of Palm Computing, the company under whose aegis Palm OS was developed. The story of Palm's conception is legendary. Right from when its

founding father Jeff Hawkins carried a block of wood to every meeting to taking a practical approach that the user should learn the hieroglyphics of graffiti rather than putting together software that understands all nuances of human handwriting. These were based on learnings from costly mistakes made earlier and feedback from customers of an earlier product Zoomer. The Palm was designed with three commandments:

- Handwriting recognition to be limited to simplified hieroglyphics.
- Size should be small enough to fit into the pocket.
- A cradle to synchronize data with a PC.

The result was a blueprint of what the PDA should be. But by 1994, successive debacles in the PDA market ensured that financers for the project were scarce. The lucky break came in the form of U.S. Robotics which lent its might to Palm. The prototype, till then known as touchdown was christened PalmPilot 1000 and officially released in February 1996. Within the year they sold an unprecedented 350 thousand units. The volumes were growing and the tempo was built. At the same time certain management changes were happening, U.S. Robotics was acquired by 3Com in 1997. The next milestone was March 1998 when the fruits of investment and innovation saw PalmPilot III being released by 3Com. The device had double the RAM, supported infrared connections, had a better character recognition algorithm, more fonts, a handy cradle, an improved Palm OS 3.0 and stylish design. But much of it was only filling up the gaps in the earlier products. By then competition was also born in the form of WinCE still in its infancy.

At the dawn of the fateful year of 1999, everything was great; sales was at its peak, employees were motivated and competition non-existent. Trouble, however, was afoot; foresight was slow in coming. Two new models—the PalmV and PalmVII that were minor variations of the PalmIII were released but failed to make a mark. The founding parents of Palm Computing, Jeff Hawkins and Donna Dubinsky, parted ways with 3Com to set up a competitor, namely, Handspring. By September they came out with Visor. The year 2000 was quite uneventful except for the release of WinCE 3.0. The curtains, however, had come down the golden age of Palm. The year 2001 saw both Palm and Handspring slowly rolling downhill. A global market recession leading to very low demand and fierce competition both within and without, saw the blue-eyed boy of the industry getting into trouble. Some good things too happened: new high-end models m500, m505 and m515, were released, these support for Secure Digital flash cards, a new version of operating system (Palm OS 4.0), new batteries with longer life and even a color display. The company was split giving birth to two divisions, one working on and licensing the Palm OS called PalmSource and the other manufacturing Palm devices called palmOne. The acquisition of Be, markedthe move from dragonball processor to a more advanced ARM core based one. On the whole, though things were not rosy; the PDA market at the close of 2001 was clearly in favor for Palm Computing and its allies.

The year 2002 saw the effort to counter this downslide with the release of wireless capable devices like i705 and Treo series from Handspring. A series of belt-tightening measures both technically and financially in the previous year yielded some bright moments. They also had to counter a patent infringement suite from Xerox on the graffiti. This led to the licensing and adoption of jot. But now there was a need for a keyboard for applications like e-mail leading to more licensing from Blackberry Rim. On the downslide was the release of PocketPC 2002 which already enjoyed substantial support from the elite corporate and of course the global reduction of the PDA market.

The year 2003 was a continuation of the innovative efforts with lots of releases from the newly formed PalmOne and PalmSource and Handspring in the form of tungstone and zire. The best of course was the two versions of upgraded Operating System. The Chinese version of Palm OS gave it a presence in the huge Chinese market. The shift was now towards communication. So, later versions of Palm are tailored towards connectivity and communications. Some of the features supported include, WiFi, a 400MHz XScale processor, 64MB of RAM, an integrated keyboard, smaller size, built-in camera, longer battery life and more. As of now, Palm OS 6 with all its glorious enhancements is already released as Palm OS Cobalt (6.0). The focus of Palm OS 6 is mostly on wireless capabilities, security and more. We will look at some details in a later section.

At the dawn of 2004 Palm has its Dream Team in place; the founding fathers, whose creativity and experience is unparalleled, and the management who dragged Palm out of the abyss of 2001. In fact it looks like Palm is well poised to finally move from a defensive stance to that of an offensive take on the mammoth Microsoft. However, another threat to Palm comes from the now mature and more advanced Symbian, with the likes of Nokia and Ericsson backing it.

What, however, decides the success of any operating system is the applications it supports. In this case there is no beating the Palm. The success of Palm is reflected in the large number of applications for the environment. Thousands of applications are available for the Palm OS serving various purposes from spread sheets, documents and presentations, database managers to trade information, e-books, business tools for CRM, order processing, surveying, records management, data collection, and inventory management, games, messaging applications like e-mail, fax, SMS, EMS, instant messengers, web and wap browsers, enhanced PIM tools, photo editors, audio and video tools and lots more.

Currently the field is all set for the battle of the giants, all placed favorably.

13.3 PALM OS ARCHITECTURE

We will now briefly look into the OS. Like most PDAs the high level view of the Palm OS device consists of three layers (Fig. 13.1): Application, Operating System, and Hardware. Palm OS occupies approximately 300K bytes and can run in 32K bytes of RAM.

As shown in Figure 13.1, the heart of the OS is the kernel. Essentially the kernel handles all low-level communication with the processor, interrupts, multitasking facilities and messaging to the OS atop it. The kernel interfaces to the hardware via the hardware abstraction layer. On top of the kernel there are the system services. Each service has a manager. For example, event manager, graffiti manager, resource manager, feature manager, sound manager, etc. To achieve faster execution these are generally mapped ROM commands. Then there are the system libraries and independent third party libraries. Later versions of the OS also contain a PACE (Palm Application Compatibility Environment) which is an emulator for the older application ensuring backward compatibility. This essentially ensures backward compatibility. The topmost layer contains the applications that use the underlying library to perform certain tasks. We will now see these components in some detail. **Note**: A complete compendium of the OS intricacies is available at the Palm source site. What follows is mostly a summary of the information provided by Palm.

The kernel used in Palm OS is the AMX real-time, multitasking kernel, a product from KADAK Products Ltd. The kernel itself supports a lot of features but not all of these are available to the applications. Also since the devices features available in various models are different, we urge the

readers to explore the features supported by the model on which the application is to be deployed of the users are advice. Some important features supported by the kernel are listed below.

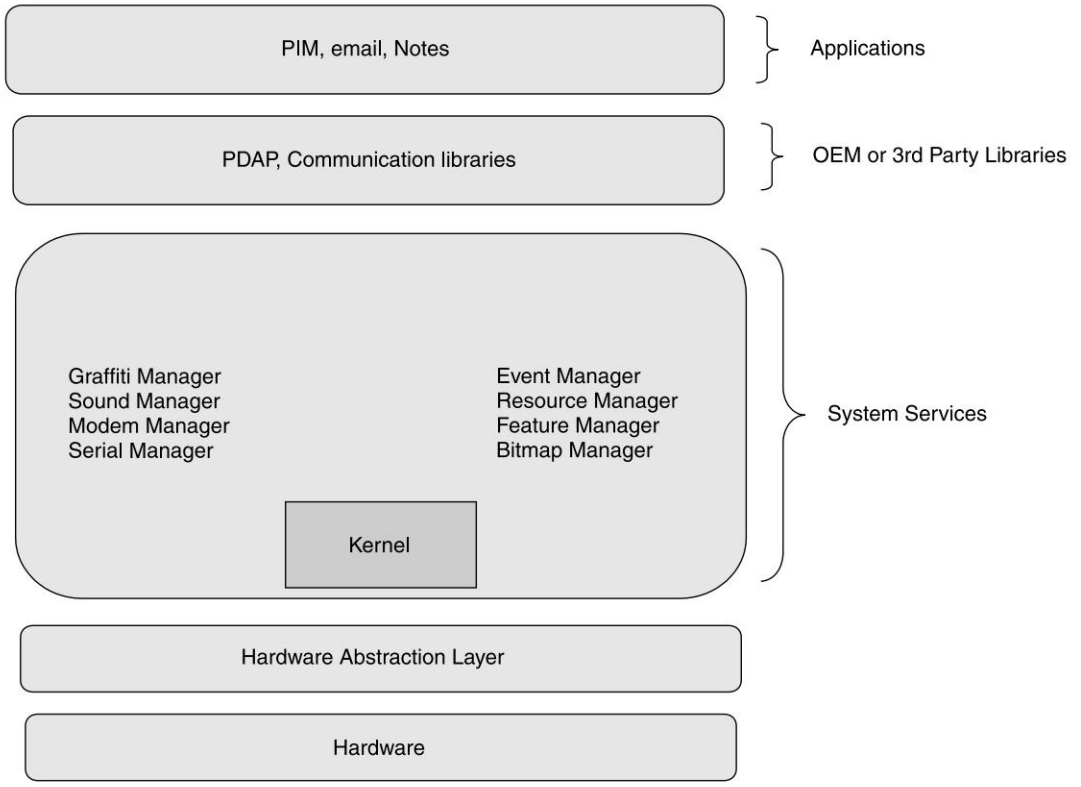


Figure 13.1 Architecture of Palm OS

13.3.1 Kernel Features

Multitasking: The kernel itself supports advanced multitasking, including semaphores. But certain licensing limitations cause these features to be available only to the system functions and not the applications. So, for our purpose the OS is essentially single-tasked.

Interrupts: The kernel supports both maskable and non-maskable interrupts in normal and nested modes. The handling is done through an interrupt specially written for it. It supports a mechanism to trap errors and is able to handle hardware interrupts. Interrupts can also initiate other tasks.

Time slicing and scheduling: This essentially allows the execution of several tasks according to their priority thereby supporting timers and time procedure. There are three types of triggers for task switching:

- Context switching: An application task requesting an implicit context switching.
- Hardware interrupt: There is an interrupt controller inside the Palm hardware system.
- Timer expiration: Each networking function has a timeout value to prevent the system from being idle in waiting state forever.

13.3.2 Memory

A Palm device has ROM and RAM. **Note:** There is no hard disk. The ROM contains the OS and some other static data. Most new models use a flash ROM and hence allow OS updates. Since the onboard memory is pretty restricted, different versions of the OS support various types of extended memory in the form of memory cards. Theoretically there can be 256 such cards.

Internally the memory contains an identifier, a list of heaps and a database directory. Palm OS 3.x and 4.x used 16-bit addresses. The newer emersion 5 uses 32-bit bus. The main memory has an ID of 0. As shown in Figure 13.2 the memory is divided into three logical heaps namely Dynamic Heap, Storage Heap and ROM Heap. As depicted in Figure 13.3, each heap has

- a header containing the unique heap ID, status flags, and the heap size,
- a master pointer table that is functionally similar to a page table and holds pointers to the beginning of each chunk,
- a variable size chunk,
- a terminator indicating the end of each chunk, and
- additional reserved space for global variables.

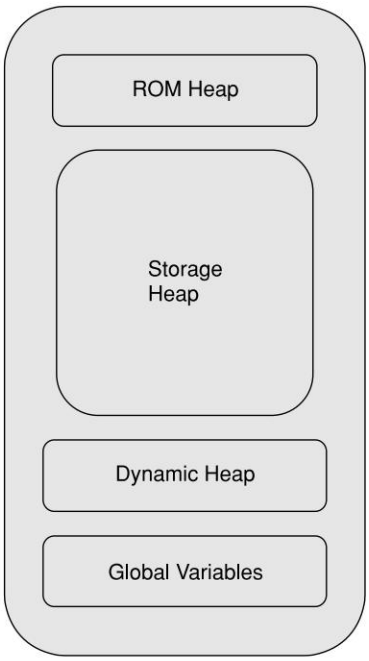


Figure 13.2 Memory Architecture

Dynamic Heap contains the operating system's global variables and data objects, user interface components, buffers, application data and an application stack, which also services system functions. The size varies between 32 kb to 256 kb depending on the Operating System version.

Storage Heap holds all the applications, user data, system patches, and any other persistent data in the system. The point to be noted is that data is not stored in files but rather in databases.

ROM Heap holds the operating system kernel. Interestingly its physical address in memory is at the higher end of the address space and helps to make compaction of the dynamic heap faster.

Database: There are two types of databases—record and resource databases. As the names suggest Record databases store user data while Resource databases store applications and free-form data. Record databases can have 64K records where resource databases can hold over 200 trillion records. Palm OS supports segmentation to overcome size restrictions on storage. It is the responsibility of the memory manager to handle logical and physical segmentation. Palm OS maps the databases to physical addresses and is analogous to the paging. Indexing is used to locate and retrieve data. Figure 13.4 shows a typical database architecture.

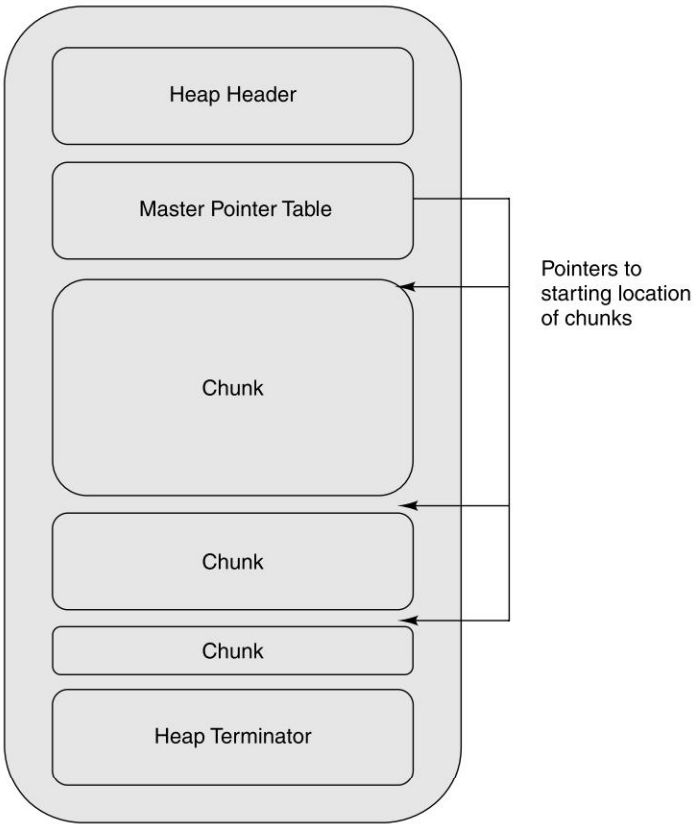


Figure 13.3 Heap Architecture

Cross card support is implemented since Version 4 onwards in the form of VFS or virtual file system. All the application data is held in the RAM and updated in place. This leads to an important consideration "Memory integrity" which is the responsibility of the memory manager. Presiding over all three memory heaps is the Palm memory manager.

The memory manager is responsible for

- Memory allocation,
- Manipulating the main data structures used to organize data within memory, and
- Providing programming APIs that ensures standard memory access across different versions of Palm OS.

For faster response there is provision of using the system stack. This is also called the Feature Memory. But application developers need to note that this taxes the system heap for very large data structures.

13.3.3 System Managers

The next layer is the system services. To obtain any system service the developers have to use the respective managers. All OS functionality the programmer wants to access must be accessed via the managers. The main language for programming Palm is C.

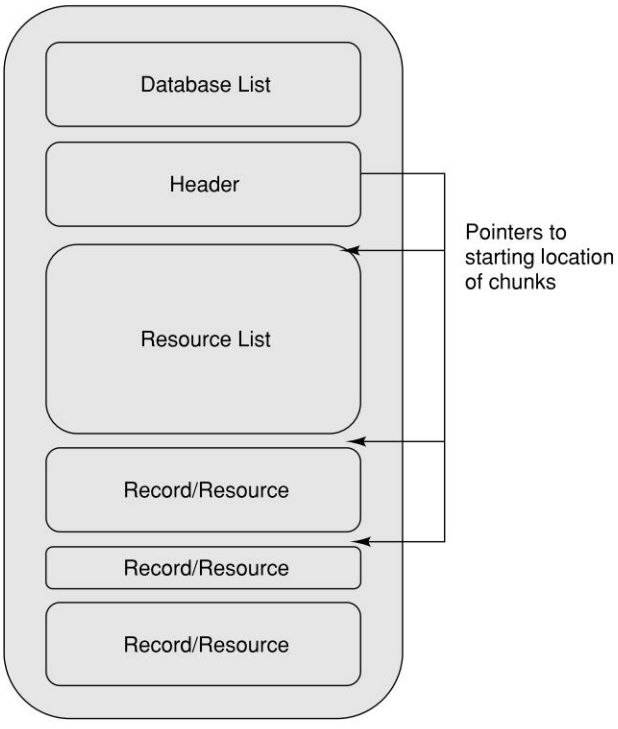


Figure 13.4 Database Memory Architecture

- Event Manager: Responsible for casting events handy for performing a global find or if the system should enter sleep mode.
- Attention Manager: This is the manager responsible for alarms. Available from OS 4.0 onwards. it serves as a central modification point for all alarms set in the system.
- Data and Resource Manager: Responsible for record and resource creation, modification and deletion.
- Exchange Manager: Responsible for all data transfer between several Palm OS devices. This includes IR, TCP/IP and Bluetooth.
- Feature Manager: Responsible for feature memory mentioned in the section on memory.
- Graffiti Manager: All graffiti is handled via the graffiti manager
- Memory Manager: This manager is responsible for all memory allocation and handling. It also ensures the integrity of the system's memory by validating every write call.
- Sound Manager: The sound manager allows synchronous and asynchronous sound (one of the few threads additionally available to the OS) and MIDI playback.
- Telephony Manager: This manager was added in OS version 4.0, and as the name signifies it allows access to telephony API.
- VFS Manager: This manager also was added in OS 4.0.

13.4 Application Development

Palm operating system is event-based. Typical events are pen taps, button presses, menu selections, and so on. Like Windows Programming, there is an event-loop. Users' actions generate a series of events. The events are FIFO queued in event queue. We query the OS for these events and perform actions based on these events.

13.4.1 Event Loop

The event loop, as the name suggests is, control excuting continuously in a loop through the executing eventQ of the current application. This means that each running application has an event queue which receives related events from the Operating System. The application fetches the event from this loop and processes it appropriately. The first step is to get an event using the EvtGetEvent function, it is then dispatched to its respective event handler. An event handler returns true if the event was successfully handled. Starting at the system level each event is successively sent to the next lower level handler for processing. The loop continues until the appStopEvent is received and the application closes. The steps are summed up below. A typical event loop control flow is shown in Figure 13.5.

1. Fetch an event from the event queue EvtGetEvent( ).
2. An optional PreprocessEvent.
3. The system is given the first chance to handle the event SysHandleEvent.
4. If Step 3 above fails, then call MenuHandleEvent.
5. If Step 4 above fails call applicationHandleEvent, a function provided by the application itself.
6. If Step 5 fails call FrmDispatchEvent.

Figure 13.5 outlines the flow chart for a typical Palm eventloop.

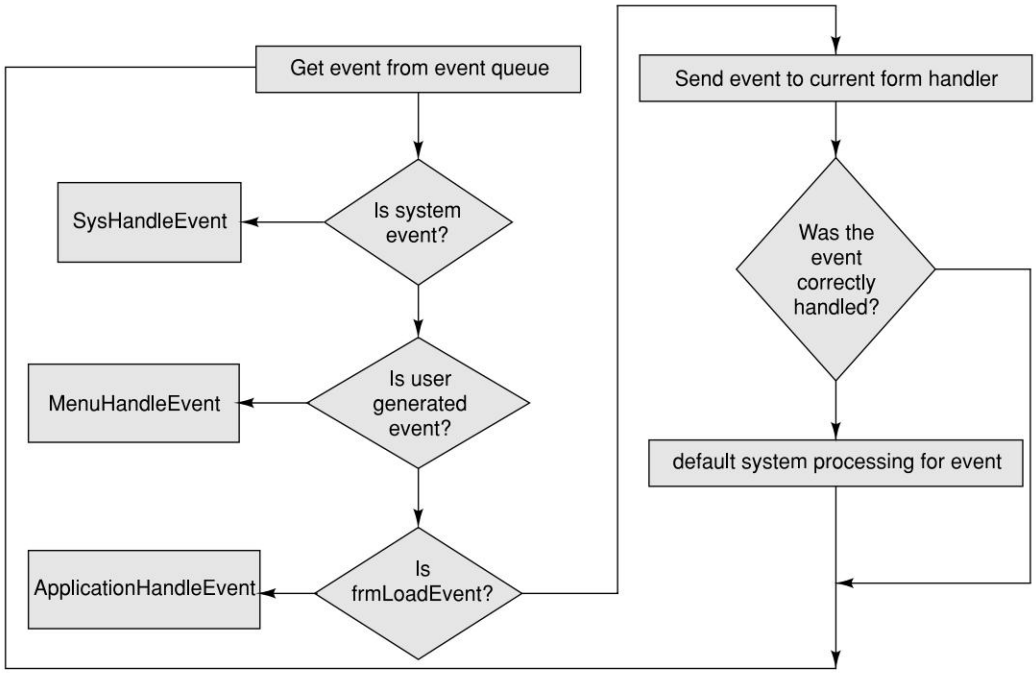


Figure 13.5 Event Handling

Events data structures are defined in Event.h, SysEvent.h, and INetMgr.h. The system passes this to the application when the user interacts with the graphical user interface.

Some of the possible events are AppStopEvent,ctlEnterEvent, ctlExitEvent, ctlRepeatEvent, ctlSelectEvent, daySelectEvent, fldChangedEvent, fldEnterEvent, etc.

Some of the data associated with an event are:

- type specifies the type of the event.
- screenX xco-ordinate or the distance from the left margin of the window.
- screenY yco-ordinate or the distance from top left margin of the window.
- data event specific data.

The system handles events like power on/power off, Graffiti® or Graffiti® 2 input, tapping input area icons, or pressing buttons. During the call to SysHandleEvent, the user may also be informed about low-battery warnings or may find and search another application.

MenuHandleEvent handles two types of events:
- brings up the menu.
- puts the events that result from the command on to the event queue.

FrmDispatchEvent first sends the event to the application's event handler for the active form
Thus the application gets to process events that pertain to the current form
If successful it returns true
Else call FrmHandleEvent to provide the system's default processing for the event.

ApplicationHandleEvent handles the frmLoadEvent it loads and activates application form resources and sets the event handler for the active form.

A sample event loop is given below.

```
static void AppEventLoop(void)
{
   UInt16 error;
   EventType event;
   do
   {
      EvtGetEvent(&event, evtWaitForever);
      // the handler for Palm system events
      if (SysHandleEvent(&e))
         continue;
      // the menu event handler
      if (MenuHandleEvent((void *)0, &e, &err));
         continue
      if (ApplicationHandleEvent(&event))
         continue;
      else
      FrmDispatchEvent(&event);
   } while (event.eType != appStopEvent);
}
```

We will now proceed to see some aspects of programming for Palm.

Get:

Palm programming tools that we have used for our development purposes

- CodeWarrior Development Environment

It contains an IDE (Interactive Development Environment) for managing and building projects. Its main components include:

- Building project (compiling and linking)
- Debugging

- User interface resource design

- Palm Emulator

Palm Emulator is a desktop software product to emulate the execution of real Palm devices. It is normally used to test applications downloading them into a real Palm device.

- ROM image

If you own a device you can extract the ROM image from there or download one from the Palm site. Instructions for both are available at www.palmos.com.

Set:
Now we need to set up our environments first. Details of setting up the environment are available at the PalmSource site and are not duplicated here.

Go:
In most cases the beginning is the toughest part. Here is where we provide the user some insights and tips. My first application "Hello Palm".

13.4.2 First Palm

Each application has a PilotMain function equivalent to Main function in C. As shown in Figure 13.6, applications start at PilotMain function. Each application also has startApplication and stopApplication functions to load and save user preference settings, such as font size. The main body is the eventLoop whose function is to detect events and to perform corresponding actions. The eventLoop function loops forever until the user exits the application. We will see more details in a little while.

The first step is to launch an application. An application can be launched from the main application menu or another application. The operating system sends a **launch code** to PilotMain function. Launch code specifies how an application responds to the event.

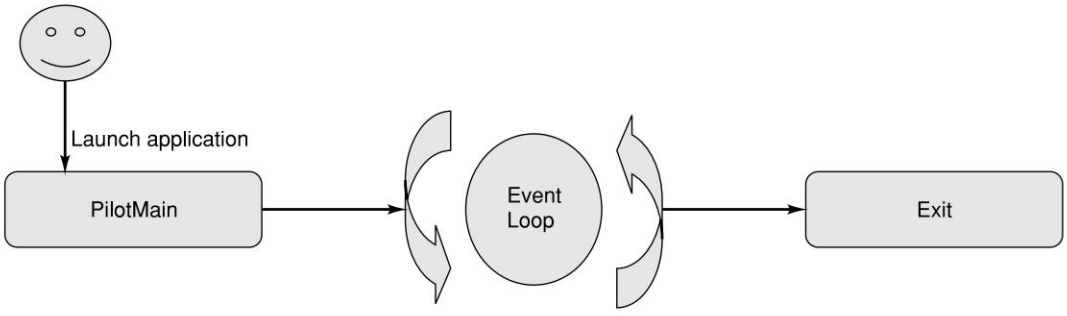


Figure 13.6 Application Life Cycle

During the launch, the application does the following:

- Initializes global variables from the values contained in the system and application preferences (like date and time formats).
- Finds the application database by creator type or create and initializes one if it does not exist.

As in any C application, the first entry are the "include files" that contain the application definitions system files and related header definition files.

```
#include "firstPalm.h"
#include <System/SystemPublic.h>
#include <UI/UIPublic.h>
```

In our example we will look at launch codes, the main event loop and event handlers. When a user performs an action, for example, taps an application icon on the application launcher screen, the system generates the launch code *sysAppLaunchCmdNormalLaunch*, which tells the application to perform a normal launch and display its user interface. An application launches when it receives a launch code. Launch codes are a means of communication between the Palm OS and the application (or between two applications). Launch codes are declared in the header file SystemMgr.h. Different launch codes specify different actions. The launch code parameter is a 16-bit word value. All launch codes with values 0 are reserved for use by the system and for future enhancements. Launch codes 32768 are available for private use by applications. User-defined launch codes are also possible. Each launch code may be accompanied by two types of information:

- A parameter block, a pointer to a structure that contains several parameters. These parameters contain information necessary to handle the associated launch code. Typical Parameter blocks are declared in AppLaunchCmd.h, AlarmMgr.h, ExgMgr.h, and Find.h.
- Launch flags indicate how the application should behave. For example, a flag could be used to specify whether the application should display UI or not.

For a complete listing and details of all launch codes refer to *"Palm OS Programmer's API Reference"* available at the Palm source site.

If an application can't handle a launch code, it exits without failure. Otherwise, it performs the action immediately and returns. When an application receives the launch code *sysAppLaunchCmdNormalLaunch*, it begins with a startup routine, then goes into an event loop, and finally exits with a stop routine.

As mentioned above, the entry point to every palm application is the PilotMain function.

UInt32 PilotMain(UInt16 cmd, MemPtr cmdPBP, UInt16 launchFlags)

cmd is the launch code for the application *sysAppLaunchCmdNormalLaunch*. In this case

cmdPBP is a pointer to a structure containing any launch-command-specific parameters, or NULL if the launch code has none.

launchFlags indicates the availability of the application's global variables, the application's state ready active, etc. Launch flag values could be

- *SysAppLaunchFlagNewGlobals*. The system has created and initialized new global values.
- *SysAppLaunchFlagUIApp*. Launch the UI.
- *SysAppLaunchFlagSubCall* . The application is calling itself indicating that the application is actually the current application.

Refer to Figure 13.6 for application control flow. A skeletal pilot main could look like

```
UInt32 PilotMain(UInt16 cmd, MemPtr cmdPBP, UInt16 launchFlags)
{
   switch (cmd)
```

```
    {
       case sysAppLaunchCmdNormalLaunch:
       EventLoop();
       break;

       default:
       break;
    }
    return 0;
}
```

The PiltoMain function is essentially an event loop that continually handles the events. For convenience of understanding we have a separate function to illustrate the event loop. In this function, we continually process events for our application and the system.

Most applications will be using the *ApplicationHandleEvent()* and *FrmDispatchEvent()* to process user's instructions. What we do in the last two we will see in a little while. But first, let us see a working sample. This application just displays a message "Hello Palm" and has an OK button. Clicking OK dismisses the application.

We need three files:

A **header file**: *firstPalm.h*

```
#define Form1   100
#define Ok          99
```

A **source file**: *firstPalm.c*

```
UInt32 PilotMain(UInt16 cmd, MemPtr cmdPBP, UInt16 launchFlags)
{
         unsigned short err;
         EventType e;
         FormType *pfrm;
         if (cmd == sysAppLaunchCmdNormalLaunch) // We will only
         handle Normal Launch.
         {
            FrmGotoForm(Form1); //Form1 has a code 100 and is
            defined in firstPalm.h .This will send a frmCloseEvent
            to the current form; send a frmLoadEvent and a
            frmOpenEvent to the specified form.
         do
   {
         EvtGetEvent(&e, 100);//poll for events every 100 millisecs
         if (SysHandleEvent(&e))
               continue;
            if (MenuHandleEvent((void *)0, &e, &err))
               continue;
            switch (e.eType) //Which type of event is it?
      {
            case ctlSelectEvent: //A control object
```

```
                on the form has been selected.
                   if (e.data.ctlSelect.controlID ==
                   Ok) // Is it the OK button.?
                   FrmCloseAllForms(); //Close all
                   forms and return to main screen.
                   break;

                case frmLoadEvent: //A new form
             FrmSetActiveForm(FrmInitForm(e.data.
             frmLoad.formID));
                   break;
             case frmOpenEvent://Set current focus to this form
                   pfrm = FrmGetActiveForm();
                   FrmDrawForm(pfrm);
                   break;
             default:
                if (FrmGetActiveForm())//If form has current
                focus handle //event
                FrmHandleEvent(FrmGetActiveForm(), &e);
             }
          } while (event.eType != appStopEvent);
       }
       return 0;
    }
```

A **resource file**: *FirstPalm.rcp* Defines the resources used by the application.

Note: It is always easier to use a resource editor to create/edit complex UI. But for our first application we will manually create it.

#include "hello.h"

*FORM ID Form1 AT (0 0 140 140) //***Where on the screen to display the form.**
*USABLE //***The form can respond to events.**
*MODAL //***Is it modal?**
BEGIN

*TITLE "Hello Palm" //***Title of the form**
LABEL "WOW My First Palm!" ID 200 AT (CENTER PREVBOTTOM+1) FONT *1//***A label to be displayed**
*BUTTON "Ok" ID Ok AT (CENTER 100 AUTO AUTO) //***There is a button too**

END
VERSION 1 "1.0.0"
LAUNCHERCATEGORY ID 200 "Examples"

You might like to insert the following before END and see what happens
CHECKBOX "Unchecked" ID 2021 AT (CENTER PREVBOTTOM+2 AUTO AUTO)

13.4.3 Form

We will now see more details on forms. Forms are the Palm's equivalent of Windows in Windows operating system. They essentially act as containers for user-interface elements/widgets (e.g., buttons, lists, fields, checkboxes).

- Each form has a form event handler function, which contains the code to handle response for UI elements within the form. A form event handler routine is of the form: *Boolean FormEventHandlerType (EventType *eventP).*
- The *FrmDispatchEven( )* routine provides indirect form-specific event handling by calling the form's event handler (*formEventHandler( )* below).

The *ApplicationHandleEvent()* function is where we set the event handlers for all of our forms using

*void FrmSetEventHandler (FormType *formP, FormEventHandlerType *handler).*

The skeleton of the sample above is then as shown below.

```
static Boolean formEventHandler(EventPtr eventP)
{
   Boolean handled = false;
   FormPtr frmP = FrmGetActiveForm(); //Get the currently active form.
   switch (eventP->eType) //What kind of an event is it?
   {
      case frmOpenEvent: //Create a form defined by frmP
      case frmLoadEvent: //Load the form
      case ctlSelectEvent: //An object on the form has been clicked.
      default:
      break;
   }
   return handled;
}
```

The resulting app event handler would now look like:

```
static Boolean AppHandleEvent(EventPtr eventP)
{
   UInt16 formId;
   FormPtr frmP;
   Boolean bRetVal = false;
   if (eventP->eType == frmLoadEvent) // is a form loading?
   {
      // Load the form resource.
      formId = eventP->data.frmLoad.formID;
      frmP = FrmInitForm(formId);
      FrmSetActiveForm(frmP); // make this the active form

      // Set the event handler for the form. The handler of the currently
```

```
            // active form is called by FrmHandleEvent each time is receives an
            // event.
            switch (formId)
                {
                    case From1: //Is this our form? Form1
                        FrmSetEventHandler(frmP, MainFormHandleEvent);
                        break;
                    default:
                        break;
                }
                bRetVal = true;
            }
            return bRetVal;
}
```

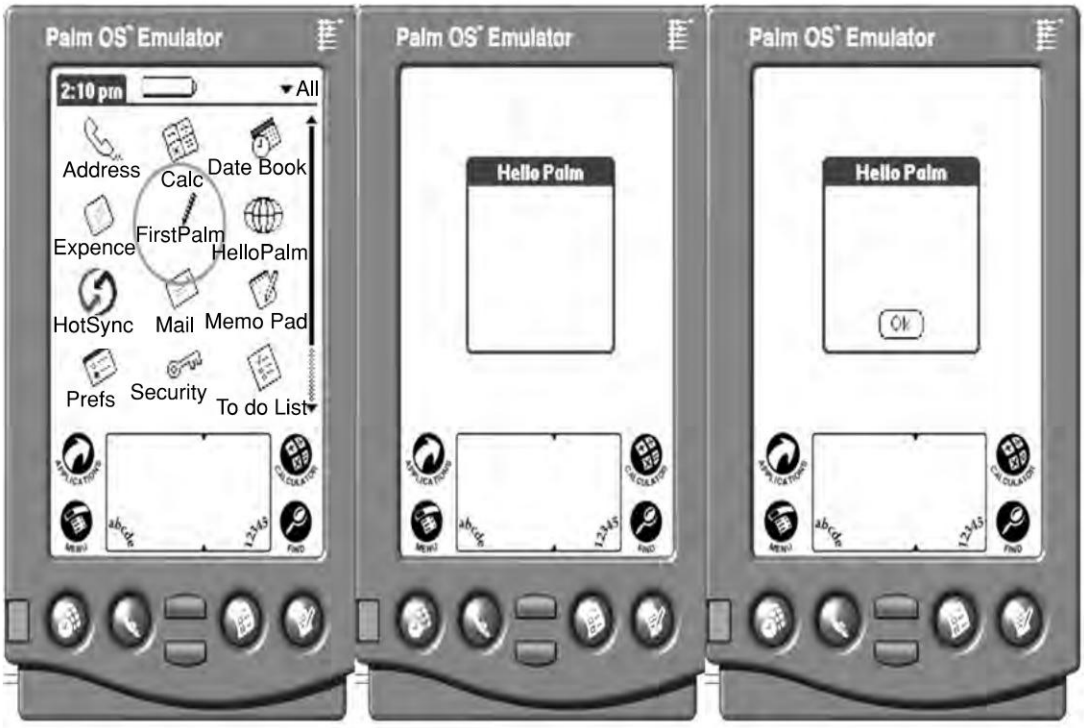


Figure 13.7a Output of Sample Code

Figure 13.7b Output of Sample Code

We have seen three components "form" "label" and "button". The resulting output for the sample code above is shown in Figure 13.7a and Figure 13.7b. Palm OS provides us a comprehensive set of widgets for creating rich UI. A listing of the most commonly used ones is given below. The list is in no way complete or detailed. The *Palm OS Programmer's Companion* and the *Palm OS Programmer's API Reference* give a detailed description of all the APIs along with the sample code to use. These are freely available at the PalmSource site. Depending on the requirements you can use one or more of the following.

13.4.4 GUI in Palm

To display a collection of objects: **Form**
- To display a menu for the user to choose from: **Menu**
- To display a series of items: **List**
- To have sub-choice in list based on the first option similar to a drop down: **Pop-up list**
- To display a non-editable string: **Label**
- To display one or more lines of editable strings: **Text field**
- To pop-up messages to the user: **Alert** (warning, error, or confirmation)
- To use built-in keyboard item: **Keyboard Dialogue**
- A submit or execute command action: **Command button**
- Selection/choice widgets
  - Select deselect options: **Check box**
  - Select a value: **Push button**

Movement items

To move control up or down: **Shift Indicator**

A progress control: **Slider**

Scrolling control: **Scroll bar**

Increment/decrement values: **Repeating button**

To display structured data: **Table**

And a custom control: **Gadget**

To handle user interfaces Palm OS provides certain low-level dedicated managers:

- The Graffiti Manager provides an API to the Palm OS Graffiti.
- The Key Manager manages the device buttons and key events.
- The Pen Manager handles pen or stylus inputs.

Most applications do not need to access these managers directly; instead, applications receive and respond to events from the Event Manager. However, if the developers so desire they can use these managers.

Creating and Handling Custom Events: At times we may want to have custom events for our applications. Custom events are similar to system events. To post a custom event we use EvtAddEventToQueue( ) or EvtAddUniqueEventToQueue( ), and retrieve them by calling EvtGetEvent( ). However, due to the event handling mechanism of Palm OS, custom events have the drawback of not reaching the application. Refer to the Palm OS programmer's manual for details.

13.5 COMMUNICATION IN PALM OS

Any discussion on handhelds is incomplete without describing its communications capability. Palm OS has a rich set of communication features that include TCP/IP, Bluetooth, serial connections, infrared beaming facilities, telephone interface, etc. Of course not all devices will support all the features. The onus is on the developer to ensure that the device supports the required connection feature and or provide an alternate path for the same. As mentioned before, the Palm source site gives excellent detailed references for all Palm programming requirements. What follows is a brief introduction.

All forms of communication begin with establishing a connection between the sender and recipient. The Connection Manager manages at a high level all connections from the Palm handheld to external devices. In the Palm OS architecture individual connection protocols are described by profiles. A profile describes the sequence of plug-ins required to implement the protocol specified. A plug-in is the counterpart of a driver in the PC world. It is essentially a piece of code responsible for configuring, establishing, maintaining, and terminating connections at the low level. For example, a Profile for TCP/IP could contain PPP (point-to-point protocol) and a dialer plug-in. We will see the details later. The main functions of the Connection Manager are listed below.

- Managing communications plug-ins.
- Managing communications profiles.
- Establishing and managing connections using profiles.

The process of using these profiles and configurations is simplified by the user interface framework, allowing editing, progress information and error display.

Figure 13.8 shows a top level view of the Connection Manager and its interaction with the plug-ins and profiles.

As shown in the figure, the connection manager interacts with two subsystems, the Connection Manager database and the Plug-ins.

The Connection Manager database stores profiles and interfaces. All objects in the database are identified by unique IDs. We will now see some details about these objects.

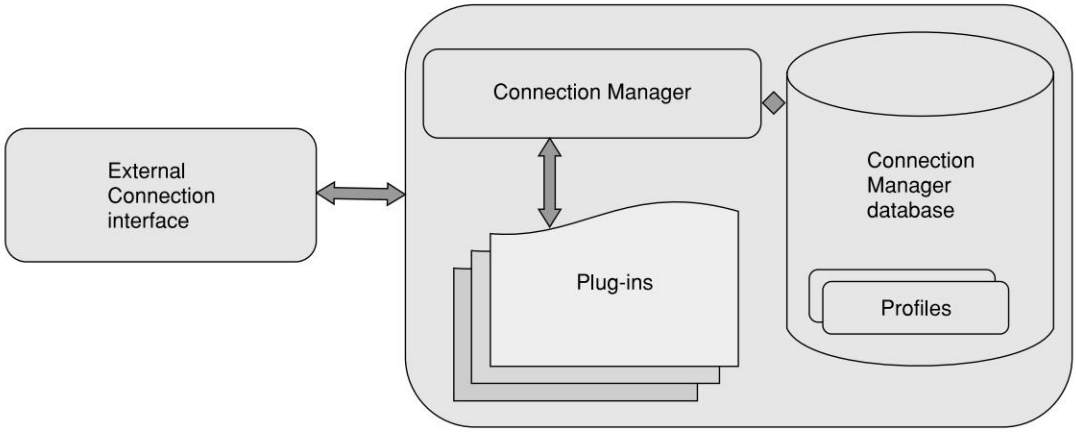


Figure 13.8 Connection Manager Architecture

13.5.1 Profiles

As described earlier, the profile is a sequential listing of the plug-ins required to make a connection. Profiles can be created by the user via the Connection Manager's UI, programmatically, or during software installation. The Connection Manager also allows modification and management of existing profiles both through the connection application or programmatically. When the application requests a connection, the connection manager will search through its database for the corresponding profile. All profiles are associated with a priority and status. The manager returns the profile with the highest priority and status equal to "available". A profile can reference other profiles. These are called sub-profiles. Sub-profiles are expanded at runtime to produce the complete sequence. A special kind of sub-profile called macro profile can be used if inline expansion is required. Each entry in a profile is called a node. The analogy is to a graph where plug-ins represent nodes. Each node has an associated list of configuration parameters. Some of these parameters are mandatory and need to be set at the time of creating the profile. Profiles can be automatic or manual. In an automatic mode, the connection manager makes a best-effort choice based on availability and priority while in a manual mode the user is asked to select a profile of their choice. For example, if we consider an Internet connection the user may have multiple ISP accounts. The same ISP in a different city may have different dialing codes. So when the users are in their home town they

might set it to automatic mode. While accessing from outstation they may want to choose the profile to be used. Or if a particular plan has better off-peak rates they may want to shift to that. kCncManualModeOption flag is used to set the connection mode. Aliases can be created using link objects. It is possible to represent a profile as a string. The general format is "node no: properties"; successive node entries are separated by "/". For more information the reader is referred to the programmers manual from PalmSource.

Here is an example profile string for PPP over infrared:
'NetOut/IPIF/PPP:User='foo', Pass='bar'/IRIF'

> The application generally interacts only with the topmost component, which defines the connection type TCP/IP in the example above and hence need not know about the lower level connections.

13.5.2 Interfaces

An interface is conceptually similar to its Java counterpart. It is essentially an abstraction of similar plug-ins that logically group together. For example, the IrComm and serial plug-ins both provide the same RS-232 connection, so they can be abstracted by a general RS-232 interface. **Note:** An interface has no associated code. It is simply an object in the Connection Manager database that relates to other plug-in objects. Typically, interfaces are used by plug-in developers. Applications normally don't need to create them.

13.5.3 Plug-ins

These are the actual workhorses. These are the code modules that implement the respective protocols. Plug-ins are built on the IOS (Input/Output Subsystem) framework and use lower-level IOS drivers and modules to interact with the hardware. Most plug-ins are configurable and provide adequate user interface to set these configuration parameters. Plug-ins are categorized based on the functionality that they implement. A general connection implementation consists of a top-level interface with the actual plug-ins lying below it.

Network Plug-ins

As the name suggests these provide connection to the network. The most commonly supported protocol, the TCP/IP, falls in this category. All network profiles start with the NetOut interface. The plug-ins lie below this in the hierarchy. Some of the plug-ins included are:

- **IPIF Plug-in**
  Manages IP configuration, domain name resolution, networking interfaces, network routes, DHCP and related configuration entries.
- **ILL Plug-in** (IP Link Layer)
  Implements a Data Link Provider Interface (DLPI).
- **PPP Plug-in**
  Implements a point-to-point (PPP) link and performs PPP negotiation. Configuration parameters to the PPP plug-in include user name, password, timeout, Maximum Receive

Unit (MRU) size, authentication type, etc. The Script Plug-in works in association with PPP and provides login script capability.

- **DLE Plug-in**
  The DLE plug-in resides directly above the network hardware and provides the Ethernet framing interface.

Serial Plug-ins

Serial plug-ins are an interface to serial communication hardware and is managed by the Connection Manager. Configuration parameters are device name, baud rate, number of data and stop bits, parity, etc. The serial plug-ins are essentially two components—a serial interface the SerialMgr interface and the various plug-ins as given below:

- USB Plug-in is an interface to USB hardware.
- Infrared Plug-in handles infrared (IR) hardware.
- Bluetooth Plug-in manages Bluetooth connections.
- Telephony Plug-ins: Most of the current Palm OS devices provide telephony access both for voice and data. Two telephony plug-ins abstract the phone hardware.
- The phone Plug-in which is the the phone driver.
- The DataCall for handling data calls.

13.5.4 Using the Connection Manager

We will now see how to use the Connection Manager. For a developer the Connection Manager is a shared system library managed by the OS. Essentially, we use the connection manager to perform the following:

- **Creating a Profile**
  All connections need a profile to connect to. We can either use an existing profile or create a new one. We can create an empty profile with CncProfileNew( ). Next, we use CncProfileInsertItem( ) to insert items into the previously created functions. Items can be plug-ins, interfaces, sub-profiles, macros, links, etc. Once the complete profile is ready, it can be submitted to the connection manager by a call to CncProfileSubmit( ).
  Another way to create a profile would be by passing the profile string to CncProfileDecode( ).
- **Changing a Profile**
  Changes to a profile happen when new component(s) are added, existing node (s) are deleted or configuration parameters are changed.
  Components can be added using CncProfileInsertItem( ).
  To delete we use CncProfileDeleteItem( ). **Note:** all deletes are cascaded meaning that all referencing profiles will also be removed.

To delete a profile from the Connection Manager database, call CncObjectDelete( ).

- **Finding Profiles**
  To find an existing profile we use CncObjectGetIndex( ) which is a name-based search or CncObjectFindAll( ) to retrieve all associated profiles and then iterate through the array returned above using CncProfileFindNext( ).

- **Managing Profiles**
  Profile management involves querying, addition, deletion and updating of the items/node. Locking and unlocking also fall in its preview.
  The process typically begins with CncObjectGetIndex( ) that returns the ProfileID.
  This is followed by CncProfileGetItemIndex( ) to retrieve the index of the item/node to on which we want to operate, CncProfileGetItemId( ). The number of items in a profile can be obtained using CncProfileGetLength( ). Finally to read the iteminfo we use CncObjectGetInfo( ) and CncObjectSetInfo( ) to modify the values.
- **Configuring Components**
  As mentioned above, most plug-ins and nodes of a profile have configurable parameters. These can be accessed using CncProfileGetParameters( ) and
  CncProfileSetParameters( ). We can also call the managers UI.
- **Invoking a Function in a Profile Plug-In**
  A plug-in can define requests that it will respond to. From within the application we can then send a request using a request parameter. This parameter tells the plug-in which parameter block to execute. While a plug-in may not respond to any user request it is mandatory for all plug-ins to implement kCncControl Availability( ) request, which returns the availability information about a particular plug-in.
- **Making a Connection**
  To make a connection from an existing stored profile, call CncProfileConnect(). It takes the ID of a stored profile and attempts to make the connection. On successful completion it returns an IOS file descriptor for the connection. This descriptor can now be used for read and write and close operations. An application may also choose to make a new profile dynamically and connecting from it.
  Note: The developer is advised to make a judicious choice here.
- Canceling or Disconnecting a Profile
  To cancel a connection we use CncProfileDisconnect( ).
  The connection Manager is also responsible for any clean-ups required in the process.

The connection manager only provides a connection. Data I/O is the responsibility of the Exchange Manager, which we will see a little later.

13.5.5 Security Considerations

Handhelds are personal devices and are likely to hold sensitive information. When these devices are exposed to the external world, security concerns are paramount. Some notable steps to ensure security are listed below.

- The Connection Manager server, which is responsible for all task runs in the system process.
- Access to the system process is restricted, through plug-ins.
- Sensitive parameters, can be designated as write-only by plug-ins disabling read ensuring that malicious applications residing on the device have no access to them.
- The users have the choice not to store the password (they must enter it each time).

- Plug-ins are system processes; hence, they must be signed to guarantee authenticity. The installation manager will not allow users to install unsigned plug-ins.

13.5.6 Object Exchange

As mentioned earlier the Connection Manager can only establish and manage a connection. But communication is all about sending and receiving data. Connection to the transport media is only one part of it. The actual reading and writing of data is the domain of the exchange manager. Palm OS supports read and write operations on default objects. A typed data object contains a stream of bytes plus some information about its contents. The content information includes any of these: a creator ID, a MIME data type, or a filename. An Address Book vCard object is a good example. It is identified by text/x-vCard MIME type.

The Exchange Manager is essentially an interface that provides APIs to send and receive typed data objects. The Exchange Manager is independent of the transport mechanism, providing transparent connectivity through the use of an exchange library.

Each protocol has its own specific exchange library that performs the actual communication with the remote device. The Exchange maintains a registry of libraries along with the information regarding the protocol that it implements and the data object it supports. When an application makes a call to the Exchange Manager, in looks through its registry and forwards the request to the appropriate exchange library. Figure 13.9 shows the interactions between the application, the exchanger manager and the libraries. The library available on a device is dependent on the hardware capabilities. Some typically available libraries include: IR Library (IrDA), Local Exchange Library, SMS (Short Messaging System) Library, Bluetooth Library, and HotSync Exchange Library.

Now we will see how to send/receive data.

1. Create and initialize the exchange socket structure. The most important parameters to this structure are the library to be used and the data type to respond to.
2. At a high level, to send data we use: the ExgPut( ) function to establish the connection with the exchange library, and the ExgSend( ) function to send the actual data. (Can be called in a loop to send large chunks of data or multiple objects).
3. Finally, we use the ExgDisconnect( ) function to end the connection.
   To be able to receive data, the application should further register for the corresponding MIME type.

As explained in the sections above, each application receives launch codes that it may choose to respond to. In this case the first one received is sysAppLaunchCmdExgAskUser which allows the developer to handle user inputs if any. Now launch the application with sysAppLaunchCmdExgReceiveData. This will actually receive the data.

1. Call ExgAccept( ) to accept the connection.
2. Call ExgReceive( ) receive the data (To receive multiple objects call within a loop).
3. Call ExgDisconnect( ) to end the connection.
   A zero (0) return value indicates a successful transmission.

The only difference when sending and receiving databases is the use of ExgDBWrite( ) and ExgDBRead( ) with callbacks to ExgSend( ) and ExgReceive( ) respectively.

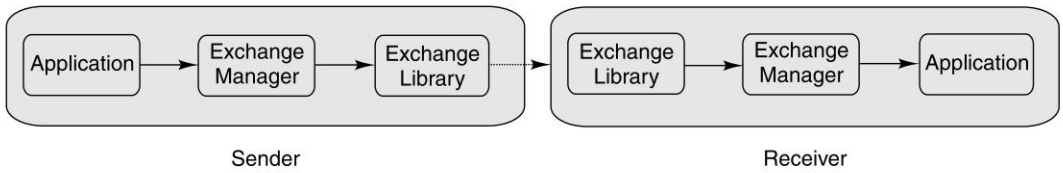


Figure 13.9 Communication using Exchange Manager

To implement a two-way communication, ExgGet( ) and ExgPut( ) can be used in combination with the ExgConnect( ).

Handling attachments and custom data types are beyond the scope of this book.

The last item in this section deals with the hot sync feature of Palm OS. Hot sync is the mechanism to synchronize values on the PC and the handheld. The advantage of using HotSync Exchange comes from its use of native format. This eliminates the need for custom conduits.

The HotSync exchange library supports two mode of operation:

- **The desktop scheme:** This is a direct exchange of a file to a HotSync desktop.
- **The send scheme:** Uses an exchange library that supports this scheme to send data.

The steps to communicate are the same as above, namely:

1. Initialize the ExgSocketType.
2. Call ExgPut( ).
3. Call ExgSend( ).
4. Call ExgDisconnect( ).

A newly installed application can use the steps in the previous section to receive data.

Another feature of the Exchange Manager is the PDI or Personal Data interchange which is essentially the exchange of personal information like business cards using a communication medium like IRDA. The Palm OS provides a PDI Library to facilitate this exchange. For more information about the PDI standards refer to the PDI consortium's web site at http://www.imc.org/pdi.

More information regarding the Palm OS PDI library and APIs is available at the Palm source site.

13.6 MULTIMEDIA

Multimedia, as the word implies, is the capability to handle multiple modes of presentation, the various modes being text graphics, sound video, etc. A combination of these can be used to produce powerful user applications. But these features require high-end computing power that the dragonball processor did not have. With the move to an ARM core things have changed drastically for Palm OS. The ARM infrastructure allows Palm OS to support multimedia features such as games, streaming video and MP3. A popular use of this capability is to take pictures and e-mail them. The new OS also lets mobile professionals handle multiple applications at once, such as Microsoft's Excel spreadsheets and Word documents, office e-mail and peer-to-peer wireless applications for data conferencing.

For the purpose of the application, multimedia is a data stream to/from the device. While recording/storing this data is encoded into an appropriate format like MIDI/MP3. While plying/retrieving the data it is decoded and sent to the hardware. This process of encoding and decoding is the job of a component called codec. Functionally, a codec translates media data from one format to another. Individual algorithms have specific codecs. Palm OS has several built-in codecs**.** A stream works within the preview of a session and represents data in a particular multimedia format. A session comprises a transaction which begins with opening a connection, transmission of data streams and finally closing the connection. Some of the parameters associated with a session are the source (where the data resides it could even be a camera or microphone) and the destination (the output might be a device like a speaker or screen, a file, or even a network stream). Each source and destination is connected by at least one stream representing the format of the data it carries. The stream further connects to a track. A track implements a codec to translate the data. Depending on the format there could be multiple tracks; for example, a movie session could contain two tracks, one audio and one video.

As with other features, multimedia has its own managers. Here we will cover the sound manager and multimedia library. The sound manager is an easy-to-use set of APIs for simple requirements like a system buzzer or alarm. For higher capability we need to use the multimedia API.

As seen above, the essence of multimedia programming is to create/read/write streams. Sound Manager can handle two types of sounds.

Simple sounds that can be played using

- SndPlaySystemSound( ) for a pre-defined system sound representing values like info alert, warning, etc.
- SndDoCmd( )to play a single tone, the parameters of this function being pitch, amplitude, and duration.
- To play a standard MIDI File (SMF) Level 0 we can use SndPlaySmf( ) [or information on MIDI and the SMF format, go to the official MIDI website, http://www.midi.org].

The following APIs can be used to play sound streams.

- SndStreamCreate( )/SndStreamCreateExtended( ) to open and configure a new sound "stream" from/into which we can record/playback buffers of "raw" data; the second method is used if we need buffers of variable length. quantization, sampling rate, channel count, etc., which are some of the parameters of this function. A pointer to a callback function (SndStreamBufferCallback( ) or SndStream Variable BufferCallback( )) is the key to the stream being created. This callback function is where the application implements its logic. The stream starts running with a call to SndStreamStart( ), the callback function is called automatically, once per buffer of data. Timing is very important in this scheme of call back mechanisms for recording and playing.
- SndPlayResource( ) plays sound data that's read from a (formatted) sound file. The function configures the playback stream, based on the format information in the sound file header. Currently, only uncompressed WAV and IMA ADPCM WAV formats are recognized.
- The Sound Manager also provides functions to set the volume, namely, SndStreamSetVolume( ). This function takes sound preference constant as its parameter. The constant can take the following values
  - ❑ prefSysSoundVolume default system volume:

- ❑ prefGameSoundVolume for game sounds.
- ❑ prefAlarmSoundVolume for alarms.

To play MIDI Files: MIDI data is typically stored in a MIDI database.

- The database type sysFileTMi1di identifies MIDI recorddatabases.
- The system MIDI database is further identified by the creatorsysFileCSystem.

The database by default holds system alarm sounds. Application MIDI records can either be stored in the same or in a separate database. The MIDI record header, the MIDI name, and the MIDI data are concatenated to form a MIDI record.

13.6.1 Multimedia Applications

We will now see the Multimedia Library. This is what covers the meat of the rich features provided by the device. Applications include the playback and recording of audio-visual media. Playback or recording sessions may be configured to run in the background while the user uses other applications. Multiple components may interact with a session simultaneously. **Note:** the Multimedia Library does not provide a means for developers to write file format handlers or codecs.

As shown in Figure 13.10, a typical multimedia application consists of

- A client application, the Media Player, for example.
- The Multimedia Library, which provides the public APIs needed by the client applications to access multimedia features.
- A Server that runs as system process and spawns sessions that usually run as background process.
- Device drivers that do the actual data communication with hardware components.

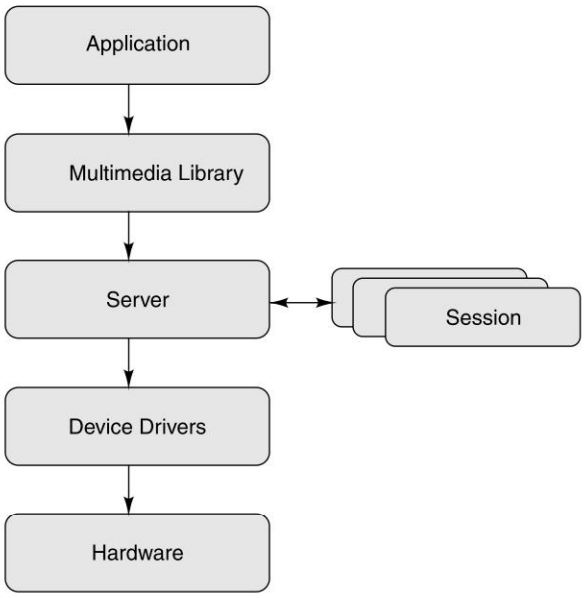


Figure 13.10 Multimedia Manager Architecture

Multimedia Library

The Multimedia Library is a shared library that the system automatically loads and initializes. This section lists the Multimedia Library API used to play multimedia content. A detailed description can be found in the Multimedia documentation of the Exploring Palm OS series. The series also documents an example recording session.

1. The first step is to create a session MMSessionCreate( ).
2. Now to provide source and destination for the session MMSessionAddSource( ) MMSessionAddDest( ).
3. Then finalize the sources and destinations by calling MMSourceFinalize( ) and MMDestFinalize( )
4. Next we need to associate the session with a track MMSessionAddTrack( ).
5. Finalize the Session MMSessionFinalize( ).
6. Now we can control the operations including start, stop, pause, capture, refresh, etc., MMSessionControl( ).

As always the last step before close is clean-up.

Telephony

The last item on the agenda is Telephony API. The telephony API support is handled by the Telephony Manager. The application can open connections and make calls to it. As of now it supports only GSM/GPRS.

Testing the Telephony Environment

Before any application uses a device facility it is a good practice to verify if the given environment supports those facilities. The Telephony Manager API TelIsServiceAvailable( ), allows us to check for the availability of the required service. The manager also provides macros that do the same work, e.g., (TelIs*ServiceName*ServiceAvailable( )).

Opening the Telephony Manager Library

After verifying that the service is available, we need to initiate a connection to the Telephony Manager library; we can use TelOpen( )to use a system-selected phone profile or TelOpenPhoneProfile( )to use a specific profile. The manager allows the application to verify support for function calls with TelIsFunctionSupported( ), the corresponding macro being (TelIs*FunctionName*Supported( )). The actual interface to the phone drivers is through the telephony server. The open call to the Telephony Manager library causes this server to be opened. Telephony Server uses the Connection Manager profile to retrieve the ID of the relevant drivers.

Making Synchronous and Asynchronous Calls

Calls are either synchronous (that will block the system till successful completion or an error causes the process to exit) or asynchronous (runs in the background and notifies the application when events such as an incoming SMS message, a call connection, battery status change, etc., occurs).

Closing the Telephony Manager Library

All sessions with the manager should end with a call to TelClose( ) function, which does the cleanup and frees any associated resources.

There are lots of other managers and libraries that support the Palm OS features. But a detailed discussion of all those is beyond the scope of this book. For more information and details about the API we urge the readers to refer to the PalmSource site.

13.7 Enhancements in the Current Release

While we are reading this, the main excitement in the Palm world is its latest version **Palm OS® Cobalt**. We summarize the major enhancements of this version below. For details please refer to the documentation at the PalmSource site.

13.7.1 System Level

- A Frameworks-based architecture, allowing for component modularity, leading to plug and play opportunities, higher scalability and easy customization. Multitasking and multi-threading features allow for concurrent applications.
- A Protected Memory scheme allowing for greater stability and protection. Increase Memory of 256MB each for ROM and RAM allowing for rich multimedia and enterprise-grade applications.
- A Compatibility layer, PACE for backward compatibility.

13.7.2 Security

- The new OS comes with a CPM (Cryptographic Provider Manager) that features 128-bit encryption pluggable default algorithms like RC4, SHA1, 3DES, MD5, SSL/TLS, etc. Provision is also available for customized algorithms if needed. This has given a good boost for building end-to-end enterprise applications.
- Two new security modules the Authorization Manager for access restrictions and Authentication Manager for additional authentication mechanisms added by third parties.

13.7.3 Multimedia

- Supports both audio and video for various standards like MP3, MPEG.

13.7.4 Database

- The updated Data Manager allows a standard database abstraction for third party application developers.
- New Database Access Control to create and control access to secure databases.

13.7.5 Communication

- Cobalt comes with a modular, flexible, industry-standard STREAMS based communications framework. A generally improved look and feel and higher integration for Telephony and Bluetooth-based applications have been provided.
- The Multi-tasking feature supports concurrent communication applications.

13.7.6 Display and UI

- Cobalt provides support for 320×320 pixel high-density displays, Graphics Rendering, Scalable Font APIs fonts for Latin locales and Multi process UI support.

13.7.7 Synchronization

- A new Sync Architecture with a new schema and secure databases, additional sync clients supporting more protocols.
- HotSync Exchange Provides access support for, many standard file formats. Drag and drop facility for sync has also been provided.

We approached the Palm OS from view of its past. We have climbed up the tech tree until we reached the version 5.0 and 6.0, along with their added possibilities and support. We have looked at the things that have made the Palm OS so strong compared to the other Operating Systems in the handheld market. After the discussion on the technical design of the operating system, starting at the Kernel, we investigated the way the Palm handles memory, how it stores and modifies data and applications, and again we had a short look at the limitations the memory of the Palm OS devices have. After a short overview over the system's managers, we moved on to the design of an application. How it can be started and how it handles events. We had a brief look at the different types of resources. The last stop was the new features in Cobalt, the official name for Palm OS 6. With this we conclude this chapter. In the next chapter we will look at another heavy weight Operating System the Symbian OS.

13.8 Latest in Palm OS

Palm has had many flavors starting from Palm OS 1.0 through Palm OS 5 and including Palm OS Cobalt. This made Palm remain in sync with the technological advancements and making the operations of both the OS and applications better. Apart from this, there has been a lot of third party enhancements. Additionally, the freeware world has made immense contributions in terms of Palm OS application, games, emulators, PDA add-ons/plug-ins, etc. Palm OS Emulator is a very powerful emulator used for writing, testing and debugging Palm OS applications. To make the UNIX side of story complete, Xcopilot is a UNIX-based Palm Pilot emulator that also runs under X11.

Herewith, we list few websites and their offerings of Palm technologies and development forums.

- www.palminfocenter.com – This has the IDE, based on the open-source Eclipse IDE, which enables developers to create ARM-Native Palm OS Protein Powered applications for Palm OS Cobalt 6.0, as well as 68K applications for all shipping versions of Palm OS. The IDE combines compilers, debuggers, simulators, device emulators and related tools into a comprehensive, integrated development suite known as Palm OS Developer Platform.
- www.developer.palm.com – This is a very good resource for tools with respect to Palm OS development.
- www.freeware.palm.com – This is a one-stop website which connects to a huge array of freeware associated with Palm OS. It has downloads and installation/development support for items like business, database, astrology, adventure, games, Protocol-converters, shopping, sports, wirelessly delivered information, etc. It pays good attention to Palm Treo devices.

- www.nsbasic.com – This offers a complete development environment for all skill levels especially for Treo and Centro.
- www.freeware-palmos.com – This is a good site for diverse application support for Palm OS.
- http://www.palmopensource.com – This offers good web support for a host of requirements for development environment of Palm OS.

REFERENCES/FURTHER READING

1. A collection of all documents from PalmSource is available for download at http://www.palmos.com/dev/support/docs/protein_books.html#devsuite.
2. Beaming "Using The Palm OS Exchange Manager" by Alex Gusev can be found at http://www.developer.com/ws/palm/article.php/3088941.
3. Download and install instructions for ROM is available at http://www.palmos.com/dev/dl/dl_tools/dl_emulator/generic_roms.html.
4. Foster Lonnon R. (2000), *Palm OS (Wireless + Mobile) Programming Bible*, IDG Books.
5. Interesting articles on development using Palm OS can be found at http://www.asptechinc.com/posdevelop.asp.
6. Sunit Katkar has a very good introduction to programming for Palm OS at http://www.vidyut.com/sunit/palmpage.asp.
7. Mykland Robert (2000), *Palm OS Programming from Ground Up*, Tata McGraw-Hill.
8. The RTOS information from Kadak can be found at http://www.kadak.com/html/prls_x86.htm.

REVIEW QUESTIONS

Q1: Describe the architecture of Palm OS.

Q2: What are the basic considerations one has to keep in mind while developing applications for Palm OS?

Q3: Describe the application life cycle in Palm OS with an example.

Q4: What are the different types of communication mechanisms available on a Palm OS?

Q5: Describe each of the following in brief:

(a) History of Palm OS

(b) GUI handling in Palm OS

(c) Plug-ins in Palm OS

(d) Multimedia handling in Palm OS

Q6: What are the different security considerations in Palm OS?

Q7: Describe the telephony interfaces available in Palm OS. Why are these interfaces important?

Q8. Enlist the steps for designing a Palm OS application for checking a duplicate entry of reminder(s) for calendar items.

Q9. How would you design a Sudoku application for IBM workpad c505? Enumerate the steps. Also, the applications should allow the user various levels of Sodoku complexity.

Q10. What shall be the design considerations linking an application of your PC to a Palm OS device through Bluetooth?

Q11. What should be a Palm OS application like to remain robust across the upcoming flavors of Palm OS?

Q12. How would you design a distributed application (say, a car racing game) so that it can run over at least four Palm devices using Bluetooth? Which programming language will be more suitable for this? What shall be the considerations in adding more players (i.e., devices) and later upgradation of the application?

CHAPTER 14

Wireless Devices with Symbian OS

Next in the series of environments for handhelds is the Symbian OS. We will begin with a short introduction to this OS that evolved out of the original SIBO (**SIBO**) and later known as EPOC (**EPOC**). We will follow it up with a brief discussion of the OS architecture. We will then introduce ourselves to application development. Developing for Symbian OS is primarily of two kinds, one being OS programming which includes porting the OS to various target machines, device drivers, OS specific OEM enhancements, etc. The second is application development that we will see a little later. There are excellent books written to cover different topics, some of these are listed in our references section. Covering all nuances of programming is beyond the scope of this chapter. What we attempt here is to give a beginner's view. We assume an understanding of C/C++ programming languages and OOPS concepts. As before, a fore note to our readers: the OS has evolved through various versions, and OEMs will provide enhancements for their devices, hence it is advisable to check out the features available for their specific target devices before embarking on the journey of application building.

14.1 INTRODUCTION TO SYMBIAN OS

This OS traces its journey back to 1981, profits from Psion's flight simulator for the Sinclar ZX Spectrum bankrolled the creation of a database orientated pocket computer, the Organiser, (1984). It boasted of "32 Kb" of combined ROM and RAM and applications included a diary, database, clock, alarm, calculator and a simple programming language called OPL (Organiser Programming Language).This device, a success for Psion, formed the basis for the growth and evolution of EPOC.

The Organiser originally ran on an OS called the SIBO. It had an extremely small footprint of "384 Kb" but supported a large number of applications including a spreadsheet. Its hardware requirements were minimal, at "128 Kb" of ROM and it provided full multitasking, a feature even some of the more advanced Palm OS doesn't support. Psion shifted their focus to a smaller device resulting in Series 3. We give here a brief timeline of the growth of Series 3.

- 1993, the 3a introduced the missing spreadsheet application;
- During 1995, PC synchronization software PsiWin was introduced allowing data to be exchanged with PC applications;
- In 1996, the 3c introduced a built-in IR port; the Siena introduced a calculator pad along side a smaller screen.

The USP for the OS was robustness, its low power usage rich set of user friendly applications reliable with a functional hardware. There were limitations, amongst the most obvious ones being a 16 bit architecture. This prompted Psion to build a new 32 bit OS. The effort started in 1994 and was named Protea. The Protea had on its plate GUI, Communication and PcSync and a very pertinent switch to the ARM processor.

In 1997, for a cost of a little over £6 million, Psion completed the project and launched Series 5. During the time that EPOC was evolving, fax, Internet and e-mail were added. Finally, in 1998, the 3mx was introduced with a significantly faster processor.

Nokia had by this time launched the 9000 Communicator. This device packed in a mobile phone and PDA, and ran on the GEOS 16 bit.

A viable hardware and a superior OS indicated a predestined alliance. A Joint Venture emerged and Symbian was announced on 24 June 1998. The then shareholder of this enterprise are Ericsson, Nokia, Motorola, Panasonic, Sony Ericsson, and Siemens, Samsung. Symbian was finally acquired by Nokia in 2008.

EPOC efforts now continued under Symbian and in June 1999 the first Symbian release of EPOC took place in parallel with the release of the Psion Series 5mx. The new and much improved OS, included new architecture for messaging, telephony, application access to contact and calendar data, and more. The next logical step was a full WAP and Bluetooth stack support.

Symbian learned from Palm's experiences and tried to avoid infighting where all licensees created Psion clones and fought for market share on price considerations alone. This was made possible by the creation of DFRDs (Device Family Reference Designs). These provide the implementation blueprints, which ensure compatibility, while allowing OEMs to customize their products. The DFRDs also allow for the OS to be used across a large range of devices with varying capabilities. Currently there are three DFRDs:

Quartz, for Data and Voice devices where telephony and data are tightly integrated.

Pearl, for Voice with Data devices.

Crystal, for Data with Voice devices, similar to the Psion Series 5 or 7, that provides full wireless access to the Internet.

And finally GT (Generic Technology) is the set of core technologies common to all the DFRDs. telephony components. For example, the year 2000 saw Symbian consolidating its position, expanding its licensee base and moving ahead with support for UMTS and GPRS. The year 2001 saw the introduction of the version 6.x. Officially OS versions 6.x + are referred to as Symbian OS. Two versions and many more licensees later, embedded in devices spread world over, the latest Symbian OS is Version 8.0.

The OS provides a rich set of facilities for smart mobile devices that combine the power of a PDA with mobile telephony and networked data services. Built, around an open and flexible

architecture, it supports applications developed in a range of Programming Languages and environments, the core set of APIs as defined by GT are exposed to all devices. The OS further expands its usability by supporting most key standards including CLDC (MIDP), wap, Bluetooth, IPv4/v6, EDGE, EGPRS, IS-95, cdma2000 1x, and WCDMA, SyncML DM 1.1.2, Unicode Standard version 3.0, etc. The current version supports the specific requirements of 2G, 2.5G and 3G mobile phones. Supported features include but are not limited to the following.

Hardware support includes a full qwerty keyboard, a 0–9*# a T-Key pad, voice, handwriting recognition and predictive text input.

OS Features include a hard real-time, multithreaded kernel working with-state-of art CPUs, peripherals and memories. Software support is provided through dedicated application engines for contacts, schedule, messaging, browsing, OBEX, etc., support for multimedia hardware acceleration, direct access to screen and keyboard for high performance.

Application Development Environments include C++, Java (J2ME) MIDP 2.0, WAP. Application engines for audio and video recording, playback ,streaming, graphics, Unicode Standard Version 3.0 for international support, etc. The system provides comprehensive security through encryption and certificates, secure protocols (HTTPS, SSL and TLS), WIM framework and authentication for installation of third party applications.

A rich set communication applications in the form of SMS multimedia messaging (MMS), enhanced messaging (EMS), e-mail and SyncML DM 1.1.2. OTA and PC-based synchronization support is also provided. Supported protocols include TCP/IP, WAP, infrared (IrDA), Bluetooth and USB. Supported networks include GSM, GPRS/UMTS 3G networks CDMA (IS-95, cdma2000 1x, and WCDMA).

Table 14.1 shows the support for various features as the OS developed through the versions.

Table 14.1 Symbian versions

Version	*Supported Features*
Symbian OS v3.	Eikon UI, Agenda and office applications
Symbian OS v5.	Contacts application, Media Server, Messaging, Java (JDK 1.1.4), Telephony
Symbian OS v5.1	Uikon UI layer, Unicode
Symbian OS v6.0	Advanced GSM telephony, PersonalJava and JavaPhone, Quartz and Crystal UIs, WAP
Symbian OS v6.1	Bluetooth, GPRS
Symbian OS v7.0	IPv6, Metrowerks CodeWarrior support, MIDP Java, Multimedia Messaging, Multimode Telephony, Opera web browser, SyncML

14.2 SYMBIAN OS ARCHITECTURE

We will now briefly discuss the OS architecture. The strength of Symbian OS lies in its small footprint (the kernel is less than 200 Kb), adaptability to limited memory devices, a powerful

power management model, a robust software layer conforming to industry standards, and support for integration with a plethora of peripheral hardware. The foundation for this is a fast, low power, low cost CPU core. The Symbian OS works atop the ARM architecture RISC processors (with V4 instruction set or higher). Supported processors including ARMv4T, ARMv5T, ARMv5TJ and Intel x86 (for the emulator). The CPU is expected to be equipped with an integrated memory management unit (MMU) and a cache.

As in any other OS, the main objective of the OS is to provide hardware abstraction and manage system resources. A Symbian system can be divided into three layers (Fig. 14.1) where the bottom most ayer interacts with the underlying hardware/hardware abstaction layer as the case maybe. This layer includes the kernel, memory, device drivers and file services. On top of this are the network and security support components. Also included are multimedia and communication protocol implimentations. The third layer is the application framework and applications support mechanism for PC synchronization, Bluetooth and USB support. The topmost layer of course is the development environment and the applications themselves.

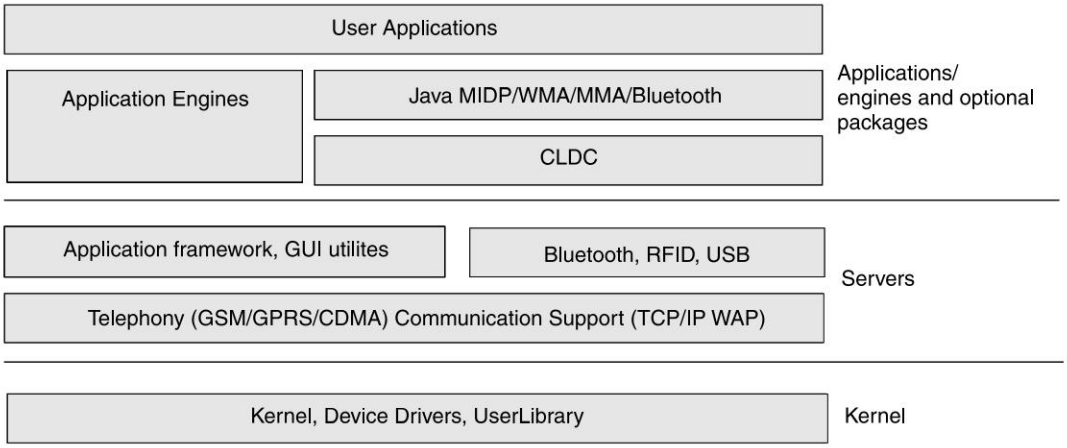


Figure 14.1 Symbian OS Architecture

Symbian OS supports pre-emptive multitasking. All system services run in privileged mode, while user applications run in a user context. When an application requests a system service it is temporarily given privileged access through a context switch. This mechanism is conceptually similar to that in UNIX/WindowsNT. The MMU is designed for interrupt handling and privileged access modes. The CPU, MMU and cache along with timers and hardware drivers, all reside on the chip.

The kernel encapsulates the system services like multitasking, file services, power management, memory management and the various device drivers. It includes support for the telephony services for GSM, GPRS, CDMA and the security features. Version 8.0 boasts of powerful kernel architecture, with hard real-time capabilities. It provides the programming framework in the form of an abstraction making it easier to port Symbian OS.

Symbian OS v8.0 is provided in application compatible two variants. The first variant, v8.0a uses the legacy kernel (EKA1) as per Symbian OS v6.1, v7.0 and v7.0s. The second variant v8.0b adopts the new hard real-time kernel (EKA2).

As most other mobile operating systems, it resides in the flash memory and executes in place. In the true spirit of plug and play, the OS is almost entirely implemented as DLLs. This helps in keeping the size small as there is a single copy of the library and everybody links to it.

Symbian also makes extensive use of a special type of DLL called polymorphic DLL. Conceptually these are similar to factory classes providing a public interface or function that applications can use. A notable example is the application DLL, which exports NewApplication( ), to create an instance of an application.

14.2.1 Hardware Interfaces

All applications need access to hardware for functions like I/O media control. Optional hardware can be plugged in and the OS is expected to recognize and service it. As for other things the hardware support is also implemented as DLLS. All hardware access is restricted to the privileged mode, i.e., all calls have to pass through the kernel.

Hardware is differentiated by the kernel's dependence on it. For example, timers, UART, DMA, etc., are essential for the OS. These are packaged along with the OS code and have direct kernel access.

Support for additional hardware is provided through a separate DLL, referred to as a kernel extension. Examples for kernel extension include keyboard, media devices and a lot more. Kernel extensions are detected and initialized at boot. Applications use the user library API to access kernel extensions.

And finally we have device drivers for optional or non essential hardware. Device drivers are true plug and play entities. Each device driver has two parts: a user side library for applications to link to and a kernel-side counterpart, for the actual hardware access. **Note**: However the screen buffer is an exception, here information is directly copied to the LCD display. The avoidance of frequent context switches allows for higher speeds.

Interestingly, a file system is also considered a device and access is provided through a device driver. The two components of this driver are a file system (generally FAT) and a media driver. The media driver is the kernel side library performing the actual operations.

14.2.2 Memory Management

Resource allocation and process life cycle are the prerogative of the kernel. It keeps track of events like thread death so that resources can be freed. As mentioned in the Palm OS chapter, memory is required for three purposes: to store the OS itself, persistent application, and user data and runtime requirements. The volatile memory is provided through a RAM. The persistent memory is provided through Flash Memory. A flash memory though more expensive than ROM is the preferred option as it can be reprogrammed.

Symbian OS uses page memory architecture. It implements a two-level page table using 4 KB pages. This allows for efficient memory usage.

Owing to the multi-tasking capability of the OS, security becomes an important consideration. Symbian OS addresses security concerns in two ways. First, by using privileged and non-privileged mode execution it takes care of restricting access to kernel and hardware access. Second, it requires all applications to run in a virtual machine (VM) thereby protecting the applications from each other. Each application executes as a single process. (A single process is likely to have multiple threads). At launch or initialization, the application is allocated memory for its data; the outer page table stores a reference to this. When a context switch causes this process to be activated, all the pages are moved to a pre-defined location in the virtual memory map, hence execution continues in the appropriate thread. Applications cannot make direct calls to the call to hardware drivers; they have to use user library APIs which in turn use system services through the kernel.

As mentioned earlier the primary concern of the MMU is to provide a protected mode system. Other functions of the MMU include

- Restriction on access to process data.
- Protection of application and OS code.
- Isolation of the peripheral hardware.

More information on the MMU and CPU architecture can be found at the Symbian site.

All handsets have limited runtime memory. Out of memory exceptions are quite likely. One way that Symbian uses to counter this is by having a clean-up stack, all partially constructed objects are placed here until their construction has been completed. If the phone does not have sufficient memory to complete object creation then it simply deletes the contents of this stack. By not allowing partially constructed objects it avoids memory leaks as well as protects applications from potential data loss.

14.2.3 System Software

All applications require system services of one type or the other. The Symbian OS System services framework operates in a client server mode where in most of the system services are provided as servers for example a file server, font and bitmap server, a media server etc. An application is a client that connects to these servers and requests their services. The client connects to the server using the kernel interfaces and uses a message passing mechanism for interaction. The server however runs in an unprivileged mode and will use other backend device drivers or kernel extensions to perform its tasks. From an application perspective we need to concentrate on the user library. The user library provides APIs to application framework and controlled access to the kernel. We will see the different frameworks and the applicable APIs as we proceed.

14.3 Applications for Symbian

The open architecture of Symbian enables Independent Software Vendors (ISVs) to focus on developing new applications for mobile phones. Third-party vendors provide software in the form of an installation (SIS) file. This file contains the libraries and resources of the application, secured by a certification system. This mechanism ensures a secure application where the vendor is identified

as a trusted source. The installer is responsible for updating the file system with the files from the SIS file. Once installed, the user can launch the application.

Applications are generally divided into two parts—the engine that implements the functionality and UI. This mechanism provides maximum opportunity for innovation to OEMs as the actual user interface is left to the manufacturers.

14.3.1 Development Environment

Symbian offers two development environments, C++ and Java (another language called OPL used to be available for EPOC). The SDKs, instruction to install and develop applications for each of these are available from the Symbian site. As mentioned above, the output of the development effort is an installable .sis file.

14.3.2 Java

As always it is advisable to check the supported versions on the target device. More so in case of Symbian Java environment because at various times in its journey from EPOC to Symbian V8.0 Symbian has supported various flavors of Java including Personal Java, Java Phone APIs and MIDP. The latest one V8.0 supports J2ME MIDP 2.0 and CLDC 1.1. Additional libraries provided include Bluetooth 1.0 (JSR082), FileGCF and PIM (JSR075), Wireless Messaging 1.0(WMA) (JSR120), Mobile media (JSR 135), 3D graphics (JSR184). We will see MIDP 2.0 in our next chapter, on J2ME. Here we will concentrate on C++.

Note: V8.0 does not support PersonalJava and JavaPhone. External support for CDC is likely in future.

14.3.3 C++

Developing using C++ involves a Windows emulator running on a PC that maps Symbian OS calls to Win32 APIs. To develop in C++, users will also require VC++. Details are provided at the Symbian site**.**

All environments have their own limitations and benefits. It is ultimately a design issue to choose one or the other. Most SDKs have an emulation environment that can mimic the exact target environment in terms of stack and heap sizes, runtime memory utilization, etc. We encourage our readers to explore and utilize this feature where ever it is available.

14.3.4 HelloSymbian

First we will take a look at the structure and life cycle of a Symbian program. For our readers wishing to move from Palm OS to Symbian OS programming there is an interesting article that compares and contrasts the two.

OPL

Organiser Programming Language (OPL) used to be quite popular during the EPOC times. It is a BASIC like language originally meant for the PISON organizer. OPL has its pros and

cons. While it supports a compile and run and even sometimes a build-on device kind of facility it has no direct access to Symbian services. (some vendors provided OPXs to overcome this limitation). Later versions of Symbian OS support only C++ and Java environments.
Owing largely to its long presence we are likely to come across a lot of applications ranging from games and graphics packages to database applications developed using OPL.

Symbian categorizes user applications into "Applications" that have application logic as well as UI for the end user, and "executables" or exe that perform certain tasks or services for others and require no user involvement. Generally, servers and engines fall into the second category. Our mandatory hello program also falls into this, as it does not have any UI and simply prints "Symbian Hello" on the console. As we will see Symbian programming is a lot more complex and involved than the others seen so far. The code below lists out "Symbian hello".

```
#include <e32base.h>
#include <e32cons.h>
LOCAL_D CConsoleBase* gConsole;
GLDEF_C int E32Main()
{
//Obtain a console
gConsole = Console::NewL(_L("SYMBIAN"),TSize(KConsFullScreen,
KConsFullScreen));
//Create a descriptor for the message
_LIT(SymbianHello,"Hello Symbian\n");
//print to the console
gConsole->Printf(SymbianHello);
//pause
User::After(1000000);
//exit
return 0;
}
```

The kernel and UserLibrary together constitute the E32. As is quite evident e32Cons.h contains the header information for the console. The e32base.h contains a few basic classes used by most Symbian applications, e.g., CBase class which is the base for all objects. The CConsoleBase also inherits from this. E32Main( ) is our main( ) equivalent of "c"_LIT that we see converts a "c" string into a Symbian descriptor. Descriptors are the Symbian's way of handling strings and binary data. Though a detailed discussion of descriptors is not possible in this text we encourage our readers to explore more about them. printf( ) becomes gconsole->Printf because in Symbian prinf( ) is a method of CBase class from which the CConsoleBase inherits.

We saw the source file but a Symbian project has other files too. We will now proceed to examine these in detail.

The first is the project specification file. Symbian applications can be built for different targets, the information for each target is a different file. For our purpose here we will only see the emulator or wins sample. This file has an .mmp extension and contains the following information:

TARGET HelloSymbian.exe //: The name of the application to be generated.

TARGETTYPE exe //: Type of the target. Two possible types are app and exe

SOURCEPATH . //: Location of the source files.

UID 0 // A unique identifier for the application. Used by the system to check for application integrity.

SOURCE hellotext.cpp //: Name of the source files to included in the application.

USERINCLUDE.

SYSTEMINCLUDE \epoc32\include

//: The locations of project-specific and system headers files are specified by the USER-INCLUDE and SYSTEMINCLUDE statements respectively

LIBRARY euser.lib //: Libraries required for the project.

Another very important file is the component definition file; this file is always called the bld.inf. This file contains information about the project specification files. In our case this is very simple but in a live application this file should list all specification files.

Applications can be built through various IDEs that you may be using or from the command line. Whatever the case the respective build commands should be documented in the product. The Symbian OS build environment is designed to minimise the complexity to developers of working with multiple program types. Every project is fully specified by its project file, and makefiles are generated from project files by the toolchain. Correctly declaring the target type in the project file will ensure the correct build process and generate an appropriate target. Thus, GUI applications are built as app type targets, ECom plug-ins as ecomiic type targets, and static interface DLLs as dll type targets.

In our case we are using the Nokia series 9200 communicator sdk, you could also be using the UIQ. The output of the program above is shown in Figure 14.2.

Figure 14.2 Hello Symbian

Applications of the first category, i.e., those that have both UI component and business logic, are structured differently. Symbian OS is an event-based system that responds to events generated either by the system or through user interaction; it implements its functionality through event handlers. UI based **Symbian programs follow a MVC model**. Since all applications must follow this, a brief discussion of the MVC pattern is included here.

The MVC pattern provides for a separation of the data and its presentation. An MVC object has three components—models for data, views for display, and controllers that manage the flow of control between various views in response to events affecting them. The MVC abstraction can be graphically represented as shown in Figure 14.3.

The system generates an event that is delivered to the controller. The controller is responsible for redirecting it to the appropriate view. The model encapsulates the application functionality; it handles the data persistence and algorithms. The view interacts with the model to obtain the relevant data. Any change or update to the data is returned back to the model for appropriate action.

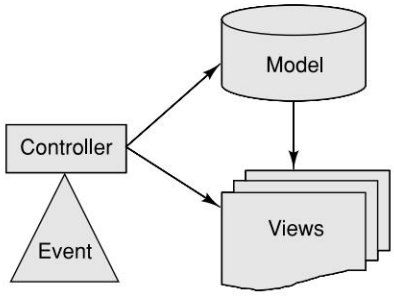


Figure 14.3 MVC Overview

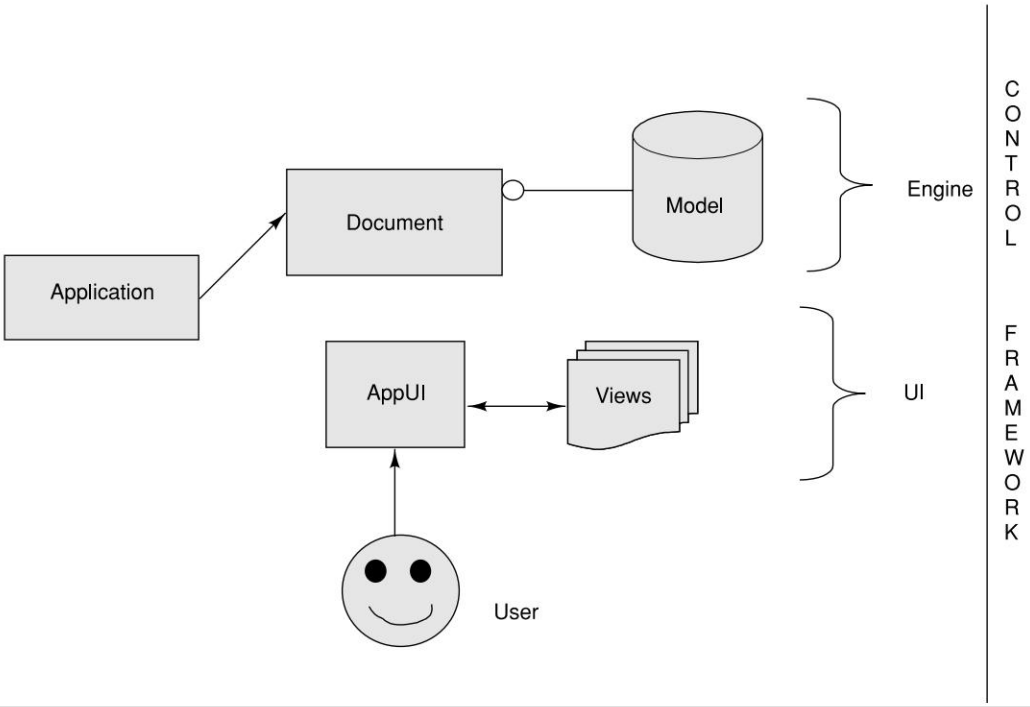


Figure 14.4 MVC in Symbian

The application is also structured and the various stages of application activation are shown in Figure 14.4. An application creates a document which contains the data model. All business logic including file and database operations are included here. This is the engine that we mentioned previously; it communicates with the system via the application and control frameworks (CF) and event-handlers. The other component is the user interface or UI where all the presentation logic resides. The event handlers are also implemented in the user interface class.

In the true spirit of an object-oriented system, Symbian OS provides most services as frameworks. A framework can be viewed as a collection of abstract base classes and some concrete classes. To use the framework, a programmer extends the abstract base classes, and provides new behavior. Traditionally frameworks were provided using polymorphicdlls. A polymorphic dll exports a single function. This creates a new instance of the newly derived framework class. We will see a practical usage of this when we discuss the HelloSymbian GUI application. Application developers will almost always be working with frameworks. Examples of frameworks include the application architecture, the UIKON etc.

Symbian OS v7.0 onwards an alternative to polymorphic dlls (dynamic linked libraries) is supplied by the *ECom plug-in architecture*. This defines a generic framework that specifies plug-in interfaces, and can be extended for developing new plug-ins. SyncML framework and transport architecture are examples of newer frameworks that require plug-ins to be written using ECom. Another advantage of the ECom comes from the fact that the responsibility of finding and instantiating suitable plug-in objects is now delegated to ECom, where polymorphic dlls were required to do these themselves.

Symbian OS includes a host of frameworks that implement its functionality. While it is beyond the scope of this book to discuss all, we will briefly look at two of the most commonly used ones.

14.3.5 Application Framework

The *Application Framework* defines the application structure and its basic user interface handling (Fig. 14.5). It also includes reusable frameworks for handling such things as text layout, user interface controls, and front end processors. It provides base classes for these and many of the key application concepts. This enables a licensee product to add its own specialist components that provide user interface elements suitable for its particular screen and input mechanisms. For example, the UIQ. A key component of the application framework is Uikon, a standard framework common to all Symbian OS platforms. It not only provides the framework for launching applications but also provides a rich array of standard control components (for example, dialog boxes, number editors, and date editors) that, applications can make use of, at runtime.

CONE

Once the application is running, "events" are channeled to it via another part of the Symbian OS framework—the Control Environment (CONE). Events could be triggered by users, things such as key presses, or system events like the machine being turned off, the application coming to the foreground, etc.

UIKON/QUIKON

Cone itself doesn't provide any concrete controls or widget, that's the job of the system GUI like UIKONE or QUIKONE. Support for UI starts with the windows server, which manages the screen, pointer ad other navigation device on behalf of the all GUI programs within the system. It

is a single server process that provides a basic API for client applications to use. CONE the control environment runs in each application process and works with the windows server client side API to allow different parts of an application to share windows and pointer events. A fundamental class delivered by CONE is CCoeControl, a control which is a unit of user interaction that uses some combination of screen, keyboard, and pointer. Many controls can share a single window. Other controls are derived from CColeControl. Together they specify a standard look and feel, and provide reusable controls and other classes that implement the look and feel.

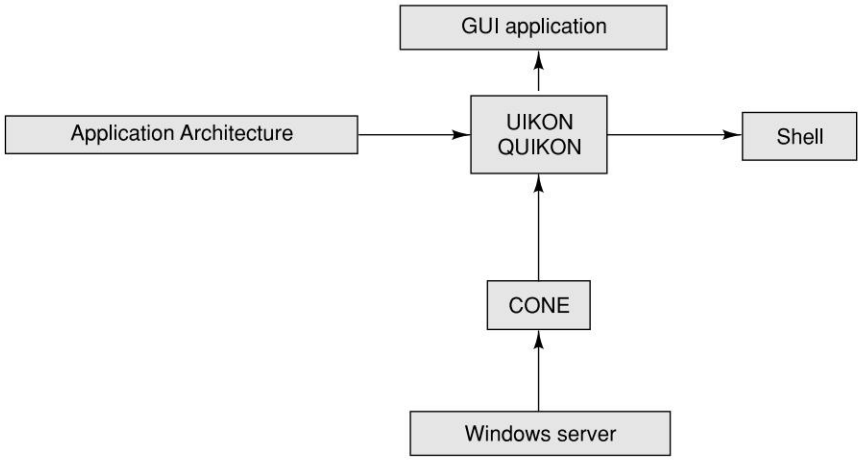


Figure 14.5 Overview of a Symbian Application Framework

All Symbian OS applications are required to have the components shown in Figure 14.6. The four main application framework classes are described below.

Readers should refer to Figure 14.7 as we go along.

- The application class is a startup object and defines the application's properties. The base class for the application class is CEikApplication which in turn uses the CONE CApaApplication class. This class bootstraps the application and returns its system unique application ID. It is the responsibility of the application framework to create the application. The application constructs the document, which in turn creates the AppUi. The AppUi constructs and owns the View(s).

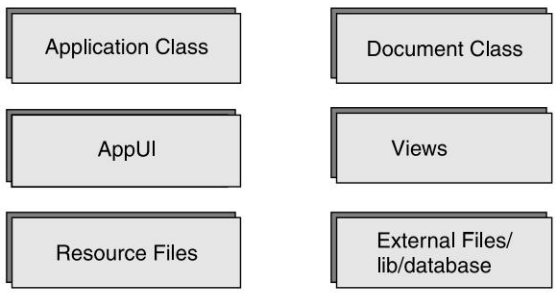


Figure 14.6 Components of a Symbian Application

- The document object is the engine of the application and is used to store the application's persistent state. An application must have an instance of the Document class even if it has no persistent data, as it is required to launch the AppUi. The base class for documents is CApaDocument, which the UIKone CEikDocument uses. In addition, the document class also contains and initializes the application's data model. The ways by which the application will save data and close down are coded in the document class.
- The AppUi is responsible for creating the application's user interface controls, views, and handles UI commands and events such as options menu commands, opening/closing files, and the application losing focus. It typically has no screen presence, i.e., it provides the foundation for the UI, but does not actually implement it. Instead drawing and screen-based interaction is delegated to the View (s) owned by the AppUI. It is also responsible for switching between Views. The base class for AppUI is CCoeAppUi.
- The View is a control, which displays data on the screen that the user can interact with. Typically, Views are notified of updates in the model's state by an observer mechanism; they also pass user commands back to the AppUi. A View is derived from CCoeControl or CEikDialog. This, in turn, springs from the Control framework base class, CCoeControl and from the view architecture class, MCoeView.

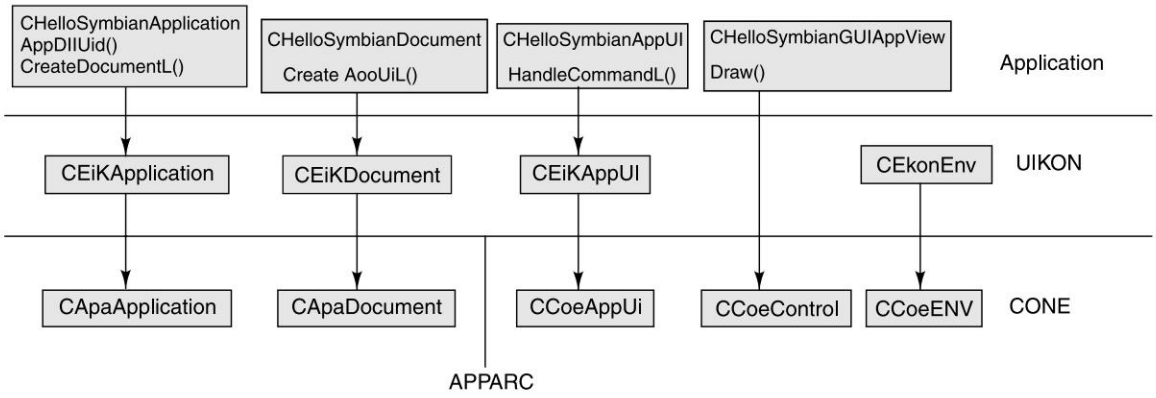


Figure 14.7 Application Classes

- Resource files are handled in almost the same way as they are in Palm OS. Symbian OS Resource files usually use the same names as their applications, with the extension .rss. When an application starts, it reads in the resource file. Further on the views are responsible for rendering it.

The application structure will become clearer once we walk through a sample program.

14.3.6 Source Code

The example is a simple application containing a single view with the text "Hello Symbian!" drawn

on it. This source file contains the single exported function required by all UI applications and the E32Dll function which is also required but is not used here.

NewApplication( ) is the starting point of any UI based application. This is the main purpose of the HelloSymbian_Main.cpp. The example is very simple containing a single view with the text "Hello World !" drawn on it.

```
#include "HelloSymbian.h"
// this source file contains the single exported function
required by all UI applications and the E32Dll function which
is also required but is not used here.
EXPORT_C CApaApplication* NewApplication( )
   {
   return new CHelloSymbianApplication;
   }
This function is required by all EPOC32 DLLs. In this example,
it does nothing.
GLDEF_C TInt E32Dll(TDllReason)
   {
   return KErrNone;
   }
```

Except in the rare case of insufficient memory the function NewApplication( ) will return an object of CHelloSymbian. The next step is the AppDllUid( ), a system level security check to ensure the identity of the application DLL. The code for HelloSymbian_Application.cpp is given below.

```
#include "HelloSymbian.h"
TUid CHelloSymbianApplication::AppDllUid( ) const
   {
   return KUidHelloSymbian;
   }
//This function is called by the UI framework at application
start-up. It creates an instance of the document class.
CApaDocument* CHelloSymbianApplication::CreateDocumentL( )
   {
   return new (ELeave) CHelloSymbianDocument(*this);
   }
```

Notice that we have used KUidHelloSymbian, but where is this value defined? It is defined in the HelloSymbian_Application.h as shown below.

```
const TUid KUidHelloSymbian = { 0x01000000 };
```

Note: This value must match the second value defined in the project definition file. This is only a sample Uid that we are using for illustration purpose. For a commercial application you will need to obtain a Uid from Symbian

```
class CHelloSymbianApplication: public CEikApplication
```

```
    {
private:
    // inherited from class CApaApplication
    CApaDocument* CreateDocumentL( );
    TUid AppDllUid( ) const;
    };
```

So now our application is created. As we saw above the main purpose of the application is to convey some information about the capabilities of the application and also act as a factory for a default document.

Now we come to the document class. In a file based application, this class handles all data operations. It also creates the appUI. The appUI is relevant when the application deals with modifiable data, contacts and notes, for example. But irrespective of the need all applications have to define an appUI. The CHelloSymbianDocument.cpp source is given below

```
#include "HelloSymbian.h"
CHelloSymbianDocument::CHelloSymbianDocument(CEikApplication&
aApp)
    : CEikDocument(aApp)
{
}
```

This is called by the UI framework as soon as the document has been created. It creates an instance of the ApplicationUI. The Application UI class is an instance of a CEikAppUi derived class.

```
CEikAppUi* CHelloSymbianDocument::CreateAppUiL( )
{
return new(ELeave) CHelloSymbianAppUi;
}
```

The corresponding header file containing the definitions for the constructor and destructors.

```
class CHelloSymbianDocument : public CEikDocument
    {
public:
    static CHelloSymbianDocument* NewL(CEikApplication& aApp);
    CHelloSymbianDocument(CEikApplication& aApp);
    void ConstructL( );
private:
    // Inherited from CEikDocument
    CEikAppUi* CreateAppUiL( );
    };
```

In a nutshell the document class

- Provides functionality for persistant data handling.
- Creation of the appUI.

Note: Symbian has a two-phase construction that includes definition and initialization. So far we have only seen the first phase.

The next step is the application UI which we created in the last phase. The applicationUI has two main functions:

> Symbian OS and Series 90 Developer Platform 2.0 applications are required to handle events, such as key presses, generated by the system. The CONE environment provides the event-handling framework to an application. Events can be key presses, menu commands, screen-redraw events, or events from other controls. UI Controls and Application Views need to handle them in a manner consistent with the Series 90 UI Style Guide.

- To capture commands to the application.
- To redirect user actions (e.g., keystrokes, pointer tap, etc.) to the controls. We will come to controls later.

The concept of commands is similar in almost all environments. These are system commands and user defined commands. In Symbian OS command is a 32bit integer ID, which is defined in the resource file. Commands normally originate from user actions like menu selection button/key press, etc. Irrespective of the source the application just needs to trap the commands that it is interested in and execute the appropriate code. We will now proceed to examine the CSymbianHelloAppUI.cpp code to understand the command handler loop.

```
#include "HelloSymbian.h"
```

This is also the second phase constructor of the application UI class. The application UI creates and owns the one and only view for this example. This reads the resource file and constructs the menu and shortcut keys for the application.

```
void CHelloSymbianAppUi::ConstructL( )
{
   // BaseConstructL( ) completes the UI framework's construction of
the App UI.
   BaseConstructL( );
Create the single application view in which to draw the text
"Hello Symbian!", passing into it the rectangle available to it.
   iAppView = CHelloSymbianAppView::NewL(ClientRect( ));
   }
```

Since the appUi owns the view, destruction too is its responsibility

```
CHelloSymbianAppUi::~CHelloSymbianAppUi( )
   {
   delete iAppView;
   }
```

The most interesting method here is the command handler. Conceptually this is similar to the MIDP command handler. This method is called by UI framework when a command has been issued.

The command Ids are defined in the HelloSymbian.hrh file, however, for this example we are using only the EEikCmdExit which is defined by the UI framework and is pulled in by including eikon.hrh. In further examples we will add a user defined command so that the process becomes clearer. For the sake of simplicity we have just added a button for exit. The code for this is in the view class which we will see next.

```
void CHelloSymbianAppUi::HandleCommandL(TInt aCommand)
{
   // UI environment
   CEikonEnv* eikonEnv=CEikonEnv::Static( );
   switch (aCommand)
{
   case ECbaButton0:
   case EEikCmdExit:
      Exit( );
      break;
      }
}
```

The ui header looks as below.

```
class CHelloSymbianAppUi : public CEikAppUi
   {
public:
   void ConstructL( );
      ~CHelloSymbianAppUi( );
private:
   // Inherited from class CEikAppUi
      void HandleCommandL(TInt aCommand);
private:
      CCoeControl* iAppView;
      };
```

So far we have created an application but now we need to associate a view with it and show it to the user. This is the equivalent of our Display in MIDP. The appUI owns one or more controls. Controls are the equivalents of MIDP displayable. Controls can draw to the screen and handle events. HelloSymbian is a very simple application that only draws a text on the screen and accepts the exit command.

Source file for the implementation of the application view class—CHelloSymbianAppView

```
#include "HelloSymbian.h"
CHelloSymbianAppView::CHelloSymbianAppView( )
   {
   }
Static NewL( ) function to start the standard two phase construction.
//
CHelloSymbianAppView* CHelloSymbianAppView::NewL(const TRect&
aRect)
```

```
    {
    CHelloSymbianAppView* self = new(ELeave) CHelloSymbianApp
    View( );
    CleanupStack::PushL(self);
    self->ConstructL(aRect);
    CleanupStack::Pop( );
    return self;
    }
// Destructor for the view.
CHelloSymbianAppView::~CHelloSymbianAppView( )
    {
    delete iHelloSymbianText;
    }
// Second phase construction.
void CHelloSymbianAppView::ConstructL(const TRect& aRect)
{
    // UI environment
    CEikonEnv* eikonEnv=CEikonEnv::Static( );
    // Fetch the text from the resource file.
    iHelloSymbianText = eikonEnv- >AllocReadResourceL(R_HELLOSYMBIAN_
    TEXT_HELLO);
    // Control is a window owning control
    CreateWindowL( );
    // this is the whole rectangle available to application.
    The rectangle is passed to us from the application UI.
    SetRect(aRect);
// At this stage, the control is ready to draw so we tell the
UI framework by activating it.
    ActivateL( );
    }
```

//Drawing the view—in this example, consists of drawing a simple outline rectangle and then drawing the text in the middle. Those familiar with windows programming will see the similarities.

It involves obtaining the graphics context, setting it to the rect supplied from the header and drawing the text. Draw is actually an abstract class in CoeControl; we have to override this and implement our drawing here. This is similar to the low level UI or canvas in MIDP. Most of the code is quite straight forward.

```
void CHelloSymbianAppView::Draw(const TRect& /*aRect*/) const
      {
    // Window graphics context
      CWindowGc& gc = SystemGc( );
      // Area in which we shall draw
      TRect drawRect = Rect( );
         // Font used for drawing text
      const CFont* fontUsed;
         // UI environment
```

```
        CEikonEnv* eikonEnv=CEikonEnv::Static( );
           // Start with a clear screen
        gc.Clear( );
           // Draw an outline rectangle (the default pen
           // and brush styles ensure this) slightly
           // smaller than the drawing area.
           drawRect.Shrink(10,10);
           gc.DrawRect(drawRect);
     // Use the title font supplied by the UI
        fontUsed = eikonEnv->TitleFont( );
        gc.UseFont(fontUsed);
           // Draw the text in the middle of the rectangle.
        TInt baselineOffset=(drawRect.Height( ) - fontUsed->Height
        InPixels( ))/2;
        gc.DrawText(*iHelloSymbianText,drawRect,baselineOffset,CG
        raphicsContext::
        ECenter, 0);
     // Finished using the font
        gc.DiscardFont( );
        }
```

The corresponding header file definition is:

```
class CHelloSymbianAppView : public CCoeControl
   {
public:
      static CHelloSymbianAppView* NewL(const TRect& aRect);
      CHelloSymbianAppView( );
      ~CHelloSymbianAppView( );
void ConstructL(const TRect& aRect);
private:
      // Inherited from CCoeControl
      void Draw(const TRect& /*aRect*/) const;
private:
      HBufC* iHelloSymbianText;
   };
```

Before we close the discussion, we will see two more important files, the resource file and the project definition file.

The resource file as in Palm OS contains the definitions for the resources. IDE would most likely provide a WYSIWYG resource editor which will provide a drag and drop facility for common controls and internally generate the resource file. However, for our purpose we will manually edit this resource file. HelloSymbian resource file contains definitions for our text, the button and the exit hotkey. The button is actually defined in the .hrh file which is then included here.

```
RESOURCE RSS_SIGNATURE { }
RESOURCE TBUF { buf=“”; }
RESOURCE EIK_APP_INFO
```

```
        {
        hotkeys=r_HelloSymbian_hotkeys;
        cba=r_HelloSymbian_cba;
        }

RESOURCE CBA r_HelloSymbian_cba
    {
    breadth=80;
    buttons=
        {

        CBA_BUTTON
        {
        id=ECbaButton0;
        txt="bye";
        bmpfile="";
        bmpid=0xffff;
        }
    };
}
```

```
RESOURCE HOTKEYS r_HelloSymbian_hotkeys
    {
        control=
            {
            HOTKEY { command=EEikCmdExit; key='e'; }
            };
}
```

RESOURCE TBUF r_HelloSymbian_text_hello { buf="Hello Symbian!"; }

Finally we get the application specification or mmp file. Except the Unique Uid everything else is the same as before in the text hello.

```
TARGET HelloSymbian.app
TARGETTYPE app
UID 0x100039CE 0x01000000

TARGETPATH \system\apps\HelloSymbian

SOURCEPATH .
SOURCE HelloSymbian_Main.cpp
SOURCE HelloSymbian_Application.cpp
SOURCE HelloSymbian_Document.cpp
SOURCE HelloSymbian_AppUi.cpp
SOURCE HelloSymbian_AppView.cpp

USERINCLUDE .
SYSTEMINCLUDE \epoc32\include
```

```
RESOURCE HelloSymbian.rss
LIBRARY euser.lib apparc.lib cone.lib eikcore.lib
```

Now it is time to build and execute this application. The output on the 9200 emulator looks as in Figure 14.8. In a different emulator such as the UIQ, it will look different.

Figure 14.8 Output of Hello Symbian GUI

To summarize:
We first create the application using NewApplication( ). This then creates the Document class which further creates the AppUi. The AppUI creates and owns the views. The views are responsible for creating controls and rendering them on the screen. Symbian OS works in an event driven mode, where user interaction generates commands or creates events. The appUI implements the Command Handler code. The sequence of events is given in Figure 14.9.

Adding Menu and other simple controls is simple. We have seen how a button is implemented in the previous example. We will see how to define controls and handle a command and attach a control. Other examples will be available as samples from the SDK vendor. We encourage our readers to explore them.

14.4 Controls and Compound Controls

A control is an area of the screen that responds to user input events. It is similar to the MIDP Form. Application developers can also implement their own concrete controls by deriving from CCoeControl; OEMs provide device specific implementations of controls, for example, the Ckon (Series 9000 UI layer) user interface library from Nokia. This provides the OEMs with a facility to implement a standard look and feel across all the applications for their devices.

Controls can be of two types—window-owning and non-window owning controls. Window owning controls are akin to the top level UI elements like form in J2ME. They contain other

controls and typically own the entire window which they occupy. The following example shows the construction of a window-owning control:

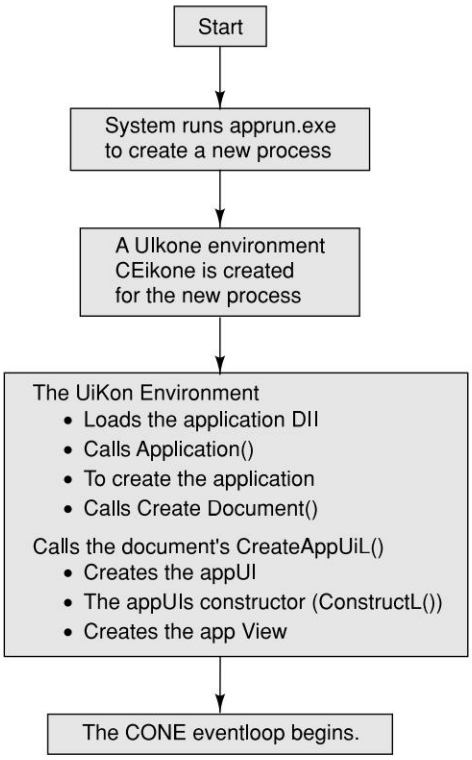


Figure 14.9 UIKone Life Cycle

```
CWOControl::ConstructL( )
{
// Create the control
CreateWindowL( );
// Set size
SetRectL(ClientRect( ));
// When created the control is blank we can then add things to it.
SetBlank( );
// Defin borders
SetBorder(TGulBorder::EFlatContainer);
// Set appView note the similarity to MIDP display.setCurrent
   (displayable); ESkinAppViewWithCbaNoToolband is the default option.
CknEnv::Skin( ).SetAppViewType(ESkinAppViewWithCbaNoToolband);
// Finally activate the window. It is now ready to be drawn.
ActivateL( );
}
```

Non-window-owning controls are similar to MIDP items and must be contained within a window-owning control. The window-owning control is also referred to as the container in this context. As in MIDP high-level items non-window owning controls are faster and require fewer resources than window-owning controls.

To construct a non-window-owning control:

```
CNWOControl::ConstructL(CCoeControl*  aParentWOControl)
{
SetContainerWindowL(aParentWOControl);
}
```

In the sample above aParentWOControl is the window-owning container for CNWO Control.

14.4.1 Compound Controls

A compound control is a group of associated controls. These are helpful in cases where the application requires the same set of controls—OK and Exit is a simple example of such a combination. A compound control will provide information about the contained controls via two virtual methods: CountComponentControls( ), the number of component controls, and ComponentControl(TInt aIndex), which returns each of the controls by index (zero-based). Figure 14.10 shows a comparison between simple and compound controls. The API reference on the Symbian site has more details on the same.

14.4.2 The Control Stack

Application UI contains controls for interacting with the users. Controls receive user commands through events like key-press navigation text in control area, etc. A control has to register itself to receive these events. The control stack is the structure that holds all controls registered to receive events. Typically, only a window-owning control the container will add itself to the control stack. The container then distributes events to its component controls. Controls on the control stack are notified of events by the control framework calling the respective event listener. For example, the key press events are notified via the OfferKeyEventL( ). Details of constructing dialogs are beyond the scope of the current text.

14.5 Active Objects

As discussed Symbian OS is a preemptive multitasking operating system and hence supports multithreading. However, multithreading is expensive in terms of resource usage and is generally not recommended. Active Objects are Symbian OS's way to enable a single thread to handle multiple asynchronous requests. Each active object implements two pure virtual functions: RunL( ) to handle request completion, and DoCancel( ) for cancellation of outstanding requests. Note: All user input events are handled through the same mechanisms, hence developers should ensure that the handler methods return fast and do not block. Long-running handler tasks should be broken up to run within the active object structure.

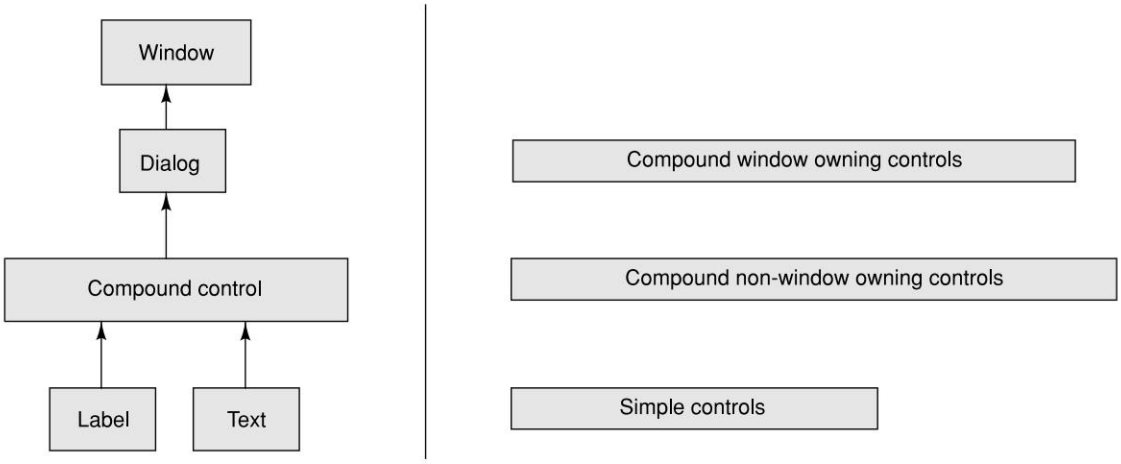


Figure 14.10 Simple and Compound Controls

14.6 LOCALIZATION

In today's ubiquitous environment it is difficult to envisage an application that is restricted to a specific geographic location. Internationalization and localization are two important considerations in application design. Most UI style guides advise a flexible attitude towards item/control/widget layouts. This is especially important in custom or application defined controls, as the responsibility of presentation here lies with the developer and not the device. In System/OME defined object the risks are lesser. Internationalization or I18N should only be a matter of changing the resource file. Each language will define its own compiled resource file. The default compiled resource file has the extension .rsc while individual languages have the extensions: .r01, .r02, .r03, etc. The following is a simple example of how localization can be achieved in a project. The project file (.mmp) contains the language information in the attribute LANG. The resource file (.rss) includes the localization file (.loc). The localization file contains the strings for the different languages.

Sample code from a loc file:

```
//Encoding information
CHARACTER_SET UTF8
//Default language specification

#ifdef LANG_01
#define str_sampleapp_string1 "String1 in default language"
#endif
// A different language resource
#ifdef LANG_02
#define str_sampleapp_string1 "String1 in language 2"
#endif
```

```
sample source code from resource file:
#include "loc file"
RESOURCE CBA r_softkeys_string1
{
   buttons =
{
      CBA_BUTTON
      {
         id = ESampleAppFunction; //note: should be defined in
         sampleapp.hrh
         txt = str_sampleapp_string1;
      }
   };
}
```

To use the language 2 resources, we just change the mmp file's LANG attribute to LANG 01 02 03, and then build the resource files. This will build the resource files for default (01), (02), (03) where 01, 02, 03 could be English, German, Italian, etc.

Resource files take care of strings but also to be considered in I18N are dates, time zones, currency conversion and other locale-sensitive data. These are handled through Tlocale and related classes. These are defined in E32STD.H.

14.7 Security on the Symbian OS

As mentioned in Chapter 12, security considerations are paramount in handheld devices. Symbian OS provides a robust security mechanism that enables confidentiality, integrity and authentication. Similar to the J2ME provisioning mechanism, Symbian OS also provides a mechanism for secure installation. Security in Symbian OS include APIs for standard cryptography algorithms, hash key generation, random number generation and certificate management, Public Key Cryptography, Certificates and Digital Signatures, secure communication using SSL/TLS, WTLS, etc. Detailed explanations of the algorithms are beyond the scope of this book.

The Symbian OS security architecture fundamentally consists of two high level components:

- Certificate management (Certman).
- Cryptography (CryptAlg).

We do not see these directly, however, all security related applications make use of these. Important among them are:

- Certificate management control panel item. This allows the user to access the certificate management.
- Software installation (authentication/digital signatures).
- Secure comms (SSL/TLS, WTLS, IPsec, etc.).

Certman as the name suggests handles the certificate management issues including checking validity of signed applications, key stores and PKI for secure communication. CryptAlg contains the algorithms for hash functions, random number generation and many more. Certman depends on CryptAlg for its algorithm requirements. Figure 14.11 shows the security architecture.

With this we conclude this chapter. We began with a background of the birth and growth of this OS and followed it with an overview of the operating system architecture. We then looked at programming for Symbian OS with a couple of very simple examples. Programming for Symbian is a long and interesting journey. Our attempt here has been to give an overview of what is involved. We encourage our readers to explore further.

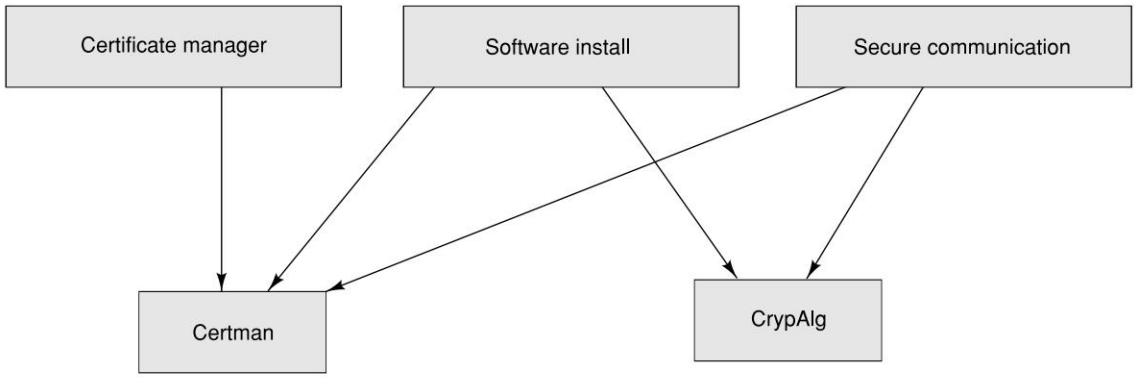


Figure 14.11 Security Architecture

14.8 Latest in Symbian

Until now, there have been major Symbian versions which were adopted by the industry for various categories of mobile phones. Being the popular mobile phone operating system and having the support for data enabled 2G, 2.5G and 3G technologies, Symbian OS has had several major revisions, the latest one being Symbian 9.5. Normally, the Symbian OS kit included nearly all Symbian OS source code with enough documentation, TechView—which is a GUI framework for testing OS and middleware components and Symbian OS emulator—which supported prompt development and debugging of all Symbian OS-based code (except kernel or device drivers) on Windows-hosted PCs and ROM building tools to build ROMs for hardware development boards. The first version in the name of Symbian was Symbian 5. Since then, Symbian had been through several revisions. We will direct our discussion through the three major Symbian versions: Symbian 8.1, 9.1 and 9.5.

14.8.1 Symbian 8.1

Symbian 8.1 offered a rich platform combined with the following basic features:

1. Mobile telephony—Symbian OS 8.1 is ready for the 3G market with support for GSM circuit switched voice and data (CSD and EDGE ECSD) and packet-based data (GPRS and EDGE EGPRS); CDMA circuit switched voice, data and packet-based data (CDMA IS-95 / IxRTT and WCDMA); SIM, RUIM and UICC Toolkit; other standards can be implemented by licensees through extensible APIs of the telephony subsystem.

2. Rich suite of application engines—the suite includes engines for contacts, schedule, messaging, browsing, utility and system control; OBEX (see Chapter 4) for exchanging objects such as appointments (using the industry standard vCalendar) and business cards (vCard); integrated APIs for data management, text, clipboard and graphics.
3. Hard real-time—this is a real-time multithreaded kernel provided with this release.
4. Hardware support—supports latest CPU architectures, peripherals and internal and external memory types.
5. Messaging—multimedia messaging (MMS), enhanced messaging (EMS) with GSM and CDMA, SMS; Internet mail using POP3, IMAP4, SMTP and MHTML; attachments, email filtering for IMAP and POP3.
6. Multimedia—framework for audio and video support for recording, playback and streaming, MIDI and speaker dependant speech recognition; framework and plug-ins for image conversion.
7. Graphics—direct access to screen and keyboard for high performance; graphics accelerator API. OpenGL ES 1.0 APIs and a non-shipable reference implementation.
8. Communications protocols—wide-area networking stacks including TCP/IP (dual mode IPv4/IPv6) and WAP, personal area networking support include infrared (IrDA), Bluetooth and USB; support is also provided for multihoming capabilities and link layer Quality-of-Service (QoS) on GPRS/UMTS networks.
9. International support—supports the Unicode Standard version 3.0.
10. Data synchronization—over-the-air (OTA) synchronization support using SyncML; PC-based synchronization over serial, Bluetooth, Infrared and USB; a PC Connectivity framework providing the ability to transfer files and synchronize PIM data.
11. Device Management/OTA provisioning—SyncML DM 1.1.2 compliant.
12. Security—full encryption and certificate management, secure protocols (HTTPS, and SSL and TLS), WIM framework and certificate-based application installation.
13. Developing for Symbian OS—content development options include: C++, Java (J2ME) MIDP 2.0, and WAP; tools are available for building C++ and Java applications and ROMs with support for on-target debugging.
14. User Inputs—generic input mechanism supporting full keyboard, 0-9*# (numeric mobile phone keypad), voice, handwriting recognition and predictive text input.

The new features which came up with Symbian 8.1 were its ability to deliver extensions to CDMA IS-95/1xRTT Telephony, Networking and SMS technology and provisioning of new customization and configurability options for multiple displays and scalable user interfaces. Symbian 8.1 made its continued alignment with standards like Java PIM, Bluetooth 1.2, Bluetooth PAN and USB Mass Storage. Symbian 8.1 came in two application compatible variants: 8.1a and 8.1b for legacy and real-time kernels, respectively.

14.8.2 Symbian 9.1

Symbian 9.1 was in the market early in 2005 while provisioning many new security-related features. Most of the features of Symbian 9.1 came from the earlier Symbian versions (like Symbian 8.1). Some of the new features in Symbian 9.1 are:

1. RTP—Symbian OS 9.1 provides a native RTP (Real-time Transfer Protocol) stack.
2. Device Management—Symbian OS 9.1 provides features which give network operators and enterprises new capabilities to manage phones in the field. This also includes OMA Device Management.
3. Bluetooth—Symbian OS 9.1 included support for Bluetooth eSCO and Bluetooth Stereo headset profiles.
4. EABI tooling—Symbian OS 9.1 allows compatibility with the latest ARM compliers and reduces the Symbian OS footprint while enhancing performance.
5. Platform Security—Symbian OS 9.1 provides a proactive defense mechanism against malware.
6. Data Caging—Symbian OS 9.1 allows applications to have their own private data partition guaranteeing a secure data store which can be used for e-commerce, location applications, etc.

Symbian 9.1 includes a controversial platform security module facilitating mandatory code signing on the pretext that applications and content are much more protected than ever. Symbian 9.1 became very popular for high-end sophisticated users and went on to provide the base for Nokia S60 platform and Sony Ericsson's M600 platform. Symbian 9.1 had a popular defect where the phone used to hang temporarily after the owner sent some hundreds of SMS which was later fixed by Nokia.

14.8.3 Symbian 9.5

Symbian 9.5 is the most recent Symbian OS update with the following new features with respect to Symbian 9.1:

1. Support for OMA device management.
2. Improved memory management.
3. Support for Wifi and High-Speed Downlink Packet Access (HSDPA).
4. Demand paging.
5. SQL support.
6. Various digital television formats.

It is claimed that applications based on Symbian 9.5 would be 75 % faster than what they are now. Announced in March 2007, Symbian 9.5 is still to meet users' expectations.

14.8.4 Symbian Platform and Developer's Suite

Symbian is a open source operating system and software platform for mobile devices promoted by Nokia, NTT DoCoMo, Sony Ericsson and Symbian Ltd; through Symbian Foundation (the official launch of the Symbian Foundation happened in April 2009). This platform includes Symbian OS assets as its core, the S60, UIQ and MOAP(S) user interfaces. It is being actively developed by the Symbian Foundation.

Although, Symbian^1 was the first release (and formed the basis for the platform), it was not open source. The first open source version of Symbian is Symbian^2, though no devices with this version were heard of. Other upcoming versions of the Symbian platform under development include Symbian^3 and Symbian^4.

Since Symbian has a lot of open source support now, many developers and interested people have sprung in action and formed online communities, weblogs and websites supporting and integrating the latest technologies with Symbian and advancing it as well. One such popular web address is http://developer.symbian.org/. Herein, the forum is well supported in multiple languages and across an array of latest Symbian affiliated technologies like C++, Java ME, Python, .NET, etc. In all, it is a rich web companion for anyone interested in development on Symbian OS. This is in addition to http://www.symbian.org/ which provides complete development support starting from scratch, for any level of beginner. The support encompasses Symbian's past to all what is cooking for its future. All what's on with Symbian can be followed on *Flickr, Facebook* and *YouTube.*

REFERENCES/FURTHER READING

1. A detailed architecture of Symbian architecture is available at http://www.symbian. com/technology/create-symb-OS-phones.html.
2. An excellent article for developers moving from Palm to Symbian http://www.symbian.com/developer/techlib/papers/SymbianOS_for_Palm/Sym bian_OS_for_Palm_Developers.html.
3. A good book for beginners is Symbian OS C++ for Mobile Phones, From Symbian Press by Richard Harrison 2004. The Indian edition is published by Wiley Dream Tech India.
4. A good resource for java on Symbian http://www.symbian.com/books/wjsd/support/ wjsd-links-downloads.html.
5. Ericsson in Symbian venture http://www.symbian.com/partners/ericsson.html.
6. Motorola in Symbian venture http://www.symbian.com/partners/motorola.html.
7. Nokia in Symbian venture http://www.symbian.com/partners/nokia.html.
8. Panasonic in Symbian venture http://www.symbian.com/partners/panasonic.html.
9. Samsung in Symbian venture http://www.symbian.com/partners/samsung.html.
10. Siemens in Symbian venture http://www.symbian.com/partners/siemens.html.
11. Sony in Symbian venture http://www.symbian.com/partners/sony-eric.html.
12. Symbian tech Library is here http://www.symbian.com/developer/techlib.
13. The Nokia forum is a great place for Nokia handsets http://forum.nokia.com/Forum.
14. The official Symbian OS developer zone for C++ is here. http://www.symbian.com/developer/development/cppdev.html.

REVIEW QUESTIONS

Q1: Describe the Symbian OS architecture. What are the functions of different layers in this architecture?

Q2: Describe memory management in Symbian OS. How is it useful for memory constrained devices?

Q3: What are the different application environments available for Symbian OS?

Q4: Describe the Symbian OS application components.

Q5: Describe each of the following in brief:

(a) HelloSymbian

(b) MVC in Symbian OS

(c) UIKON/QUIKON

Q6: What are the different types of control available in Symbian OS? Explain each of them with respect to their functional capabilities.

Q7: What are the different security procedures and protocols available on Symbian OS?

Q8: Discuss the prime offerings of Symbian OS 8.1.

Q9: How would you design a Symbian application for putting your music files directory in sync with the one available on your phone? Also, sync should incorporate the playlists and order of the songs most regularly played. Enumerate the steps for such a design.

Q10: How would you design a Symbian C++ Sudoku application? Enumerate the steps. Also, the application should allow the user various levels of Sudoku complexity and saving the game context so that user can resume the game later at his will.

Q11: What shall be the design considerations linking a Java ME application of your Symbian Handset to a Java application on your PC?

Q12: Which Symbian handsets are the most popular in the market? Can you elucidate on the correlation between Symbian's openness and popularity of such handsets?

Q13: How would you design a distributed application (say, a car racing game) so that it can run over at least for Symbian devices using WLAN?

CHAPTER 15

J2ME

15.1 JAVA IN THE HANDSET

In this chapter we take a look at the Java technologies available for mobile applications. Before we move into the specifics and language constructs it would be helpful to familiarize ourselves with the Java kingdom and its governors. The chapter begins by briefly tracing the birth and growth of Java. We will take a quick peek the specifications, authors, approvers and at implementations. With that background we shall look at the various programming aspects in mobile computing. We assume that the reader is familiar with Java programming and OO (Object Oriented) concepts. There are a number of Java books available for developers at all levels. The Sun site "http://java.sun.com" is an excellent point of reference regarding all information related to Java.

15.1.1 Why JAVA?

The contemporary languages, both procedural and Object Oriented (C/C++ for example) had inherent drawback of platform dependency; for example, a program compiled for Windows could not run on UNIX. Overcoming this barrier has been one of the most touted reasons for the success of Java. This was made possible because Java is a cross between a compiled and an interpreted language. The process is divided into two steps—first, Java compiles and generates intermediate codes called byte codes which are interpreted by the virtual machine in the second step. Of course this requires that the VM be ported to all the environments on which the program has to run. As this is being written, most commercial OS (Operating Systems) have a JVM (java virtual machine) port for their environment, and hence this requirement is no longer a deterrent. Java, however, went a step ahead and propagated the idea that the same software should run on many different kinds of computers, platforms, consumer gadgets, and other devices.

What this means is that at least in theory, Java programs are device independent, i.e., a Java program will run on any Java virtual machine irrespective of the computer or operating system the virtual machine is running on. For example, a program written and compiled (to byte code) on a

Windows 95 PC will run on a Symbian mobile phone having a JVM. The ground reality, however, is far from this, the limitations being not those of Java but of the underlying hardware that comprise a wide variety of configurations and capabilities. We have already visited these disparities in Chapter 12 Client Programming. As we will see later in this chapter, this particular problem has been overcome by Sun using different flavors of Java for machines with similar capacities.

The philosophy of Java can be very neatly summed up in the famous statement by Scott McNealy "the network is the computer". Which means it should be possible to "pull" services whenever needed and "push or offer" services wherever required. Essentially it means one single comprehensive language that can be used to write programs for all devices. This ties in neatly with Sun's new mantra "Everyone, Everything, Everywhere, Every time".

Let's step back a few years and trace the birth and growth of Java. To know more, please access Sun's official record "http://web2.java.sun.com/features/1998/05/birthday.html".

It was originally called "Green Project" aimed at developing an operating environment for ubiquitous consumer appliance and devices. James Gosling felt that the contemporary standard C++ was inappropriate, as it lacked the standard interfaces and reliability essential to its environment. In 1991 he had the first avatar of Java called "Oak" ready. Oak incorporated some cool new features in terms of UI, security, encryption, etc. But it took the World Wide Web or Internet to make Oak successful. Oak saw commercial implementation in 1995 and was renamed Java. Under the efforts of Bill Joy and Patrick Naughton, the first Java "killer app" saw daylight in the form of HotJava, an interpreter for Netscape's applets. Sun and Netscape made history by giving away the software as well as the source to the developers online. A host of new applications that sent pieces of executable code to run on the client saw Java firmly established in the world of software.

> "We've got 1.2 billion JVMs running in the world in an extraordinary diversity of devices ... and who's driving it? You and three million other Java developers across the planet." Jonathan Schwartz, EVP of Sun's Software Group, San Francisco, at the Eighth JavaOne. To read the entire excerpt go to "http://www.sys-con.com/Java/article.cfm?id=2070"

As the use of the language grew, so did the clamor for new features and capabilities. Though just 10 years old, its popularity is overwhelming and still growing. Today, millions of developers worldwide are writing Java applications ranging from embedded mobile systems, to desktop computers, to large servers. Figure 15.1 depicts a brief time line. To read a blow by blow account refer to "http://www.wired.com/wired/archive/3.12/java.saga.html".

15.2 The Three-Prong Approach to Java Everywhere

The mobile market worldwide estimated at a whopping $80 billion (http://wireless.java. sun.com/developers/business/articles/opportunities) has numerous stakeholders. Starting from the device manufacturers going down to the end users, it includes a host of others like network operators, application and service providers and others, all held together by the underlying technologies and tools. So we see companies like the Lucent, Motorola, Sun, Nokia and a host of third-party vendors and service providers, all wanting a piece of the pie.

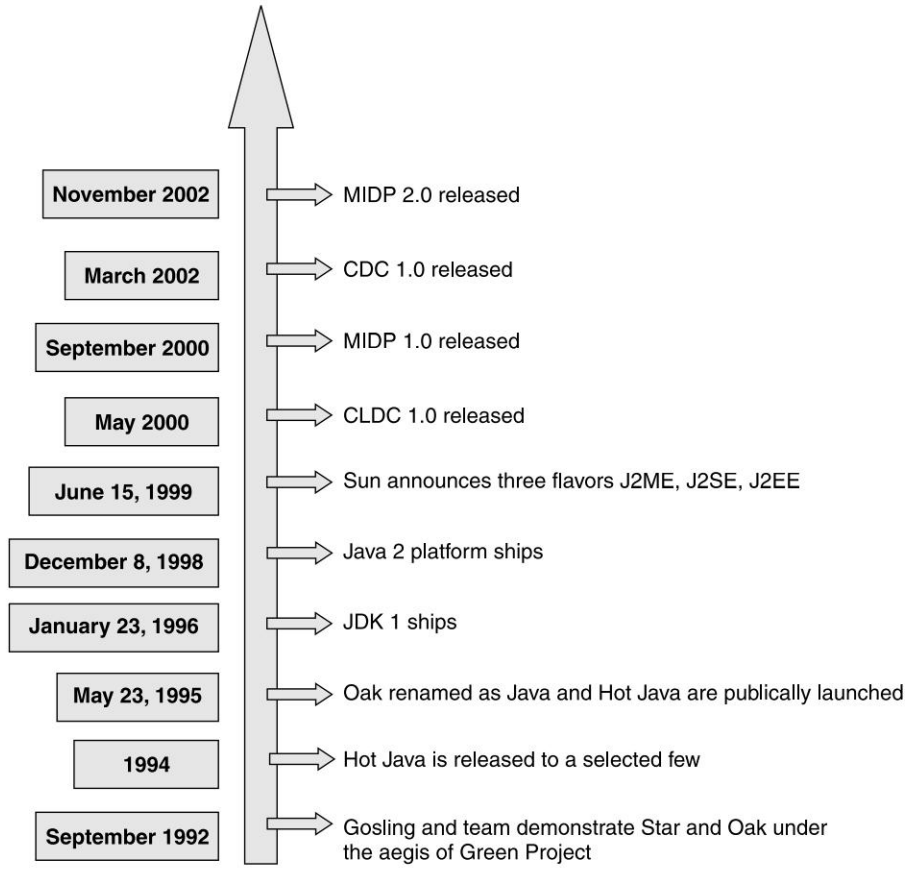


Figure 15.1 Timeline for Java and J2ME

According to researcher John Jackson, "The Yankee Group estimates a current installed base of 97.7 million Java technology-enabled handsets worldwide, and expects continued rapid growth for J2ME penetration globally." With researchers referring to Java technology as "the common language of the wireless world," and deployment of it by 53 network operators and more than 20 handset manufacturers, Java technology is cemented into the international mobility landscape http://wireless.java.sun.com/developers/business/articles/opportunities.

But all these depend on a seamless integration of all pieces starting from the application server which is hosting the application to the end user hand set. A typical top view of a mobile application will look as shown in Figure 15.2. The white papers at "http://java.sun.com/blueprints/wireless" are a wealth of information on building end-to-end mobile applications.

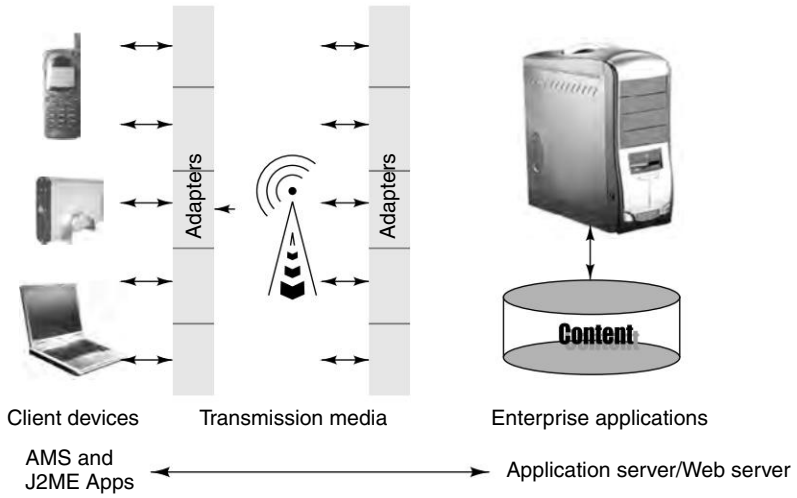


Figure 15.2 A Typical Mobile Application Architecture

In brief, there are two types of mobile applications: those that are device resident, i.e., those that utilize exclusively the device resources and do not interact with any other applications outside except maybe while being downloaded or installed. Games are an excellent example. But such applications are very restrictive and do not generate much in terms of revenue for the providers. The second type are the network-enabled applications.

The real potential of this kind of application is derived from the fact that it can connect to other devices across networks and perform tasks ranging from simple e-mails to complex monetary transactions. A typical J2ME transaction is shown in Figure 15.2. The device resident client app initiates a request to a server application using any of the various bearers like GSM, CDMA, etc. The protocols may vary from HTTP to WAP. The request received by the server is processed and the response is returned to the application. A session may be maintained over a series of transactions if required.

As Figure 15.1 shows, initially everything was Java, but Sun soon realized that one size does not fit all. So with the release of Java2 they identified Java for three different platforms—the low end or limited devices, the desktops and the high-end servers. Figure 15.3 shows a detailed diagram of the Java spectrum.

Thus we have the Hotspot-powered Java2 Enterprise Edition (J2EE) for the server side applications, the JVM-powered J2SE desktop environments and the Kilobyte Virtual Machine (KVM)/CVM powered J2ME for handheld devices and set-top boxes. A contemporary of the J2ME is the Java Card Technology for Smart cards. It is powered by a card-VM and is characterized by an extremely small footprint.

Again in a significant departure from the traditional product, Sun evolved a process called "Sun Community Process" or JCP to define the functionality of Java. It is responsible for the development and approval of Java technical specifications. It's an open community where everybody is free to

join and be involved in its process. Since most stakeholders are actively involved, the JCP procedures ensure Java technology's versatility, stability and cross-platform compatibility. The open approach of JCP makes industry adoption very fast, in fact as fast as six months from the release of the final specs. For more details on how the process works, visit "http://www.jcp.org".

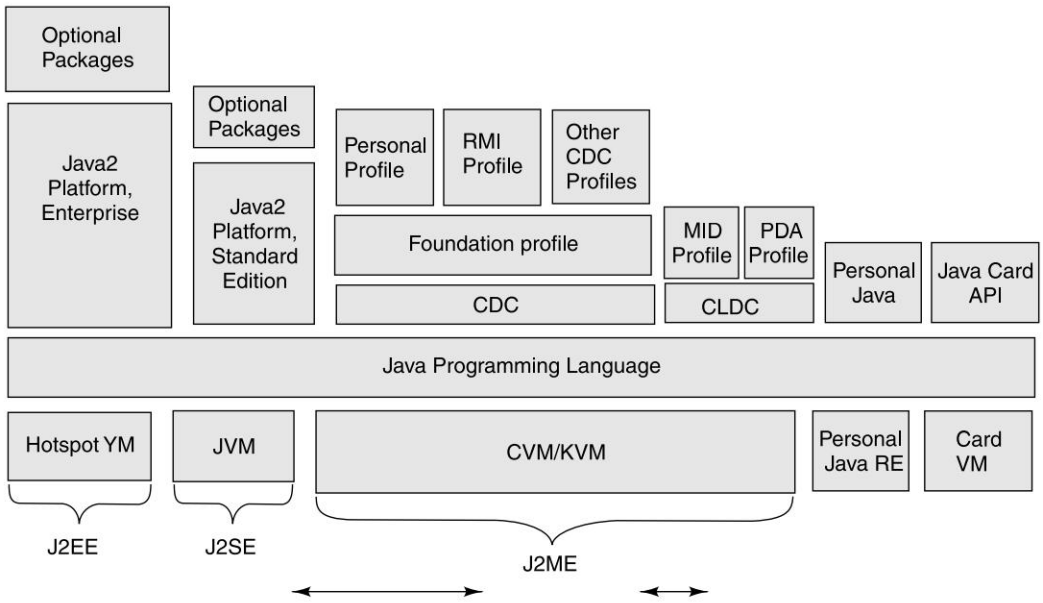


Figure 15.3 A Three Prong Approach to Java

It is interesting to note that as early as 1998 just three years from inception, Java very nearly broke out into different flavors. Read more about how it was resolved at "http://cgi.cnn.com/TECH/computing/9901/18/javasplit.idg"

The fact that there are 55 JCPs for J2ME itself speaks volumes about the magnitude of the task at hand. We will not cover any details in the section but do so as the chapter progresses.

We begin with an introduction to J2ME. This is followed by sample exercises in application development.

15.3 JAVA2 MICRO EDITION (J2ME) TECHNOLOGY

The coffee cup for the small devices is christened Java2 Micro Edition (J2ME). J2ME was conceived from the need to define a computing platform that could accommodate consumer electronics and embedded devices. As we saw in the chapter on "Client Programming (Chapter 12)" the handheld gadgets comprise a whole gamut of devices that come in varied configurations in terms of resources

and capabilities. The low-end PDAs may offer only offline data storage with a serial cable to sync with the PC while the high-end communicators would be microcomputers. Mobile phones are likely to have low bandwidth intermittent connectivity while the set-top boxes would have uninterrupted connectivity. It was not practical to attempt to define a single J2ME platform for all of these. The biggest challenge for J2ME was to specify a platform that could support a consistent set of services across a broad spectrum of devices with a large multitude of capabilities. To be able to support the large brood of devices, a modular structure was essential. The designers of J2ME came up with a concept of configurations and profiles towards achieving this goal.

> In practice, the primary differentiators are computing power, power supply and I/O capabilities. Moore's law, however, makes this differentiation quite fuzzy, because technology enables additional capability to be placed in smaller devices.

A configuration defines the lowest common denominator or the minimum capabilities that will be available across a given range of devices. It is a complete Java runtime environment, consisting of:

- A JVM (Java Virtual Machine).
- A set of Core Java runtime classes.
- A set of supported API (Application Programming Interface).

Configurations specify classes and methods that are inherited from Java2 Standard Edition (J2SE) classes. That means, J2ME is a subset of J2SE. However, they are generally not complete subsets of J2SE. Configurations also include additional classes to adapt to device capabilities and constraints.

In short, a configuration can be defined as a specification that identifies the system level facilities available; for example, the characteristics and features of the virtual machine present and the minimum Java libraries that are supported. Members of a given genera are considered to have similar capabilities in memory and processing power and can be expected to provide a certain level of system support. Although a configuration does provide a complete Java environment, the set of core classes is normally quite small and must be enhanced with additional classes supplied by J2ME profiles or by the configuration implementer. In particular, configurations do not define any user interface classes.

To avoid fragmentation and a deluge of incompatible platforms, J2ME defines only two configurations. They represent the two distinct categories of devices.

The first category is devices that have superior UI facilities such as higher computing power and are constantly connected. These implement the Connected Device Configuration (CDC), e.g., set-top boxes, Internet TVs, Internet-enabled screen phones, high-end communicators, and car entertainment/navigation systems.

The second being personal, mobile information devices that are capable of intermittent communication. These implement the Connected, Limited Device Configuration (CLDC) e.g., mobile phones, two-way pagers, personal digital assistants (PDAs), and organizers.

While the configuration concept is helpful, it's still broad and incomplete. A profile takes the configuration a step further by defining the libraries used to create applications. The profile specifies the application-level interface for a particular class of devices representing a vertical market. Profiles are built on top of and utilize the J2ME configurations. Event handling, I/O functions, user interface

APIs (for specific device categories, like mobile phones, and PDAs) and application lifecycle management lie in the purview of the profile. Profiles are also a means to guarantee interoperability—between all devices of the same category or vertical market.

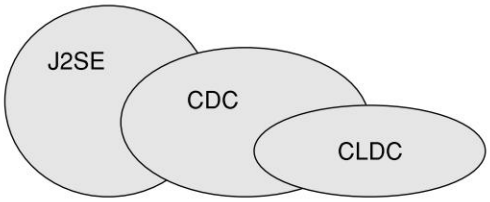


Figure 15.4 How the Three Fit In Together

Applications are built on top of the profile; they can use only the class libraries provided by these two lower-level specifications. Profiles can be built on top of one another. A J2ME platform however, can contain only one configuration.

Together the configurations and profiles enable creation of "customized flavors" of Java for different device categories. Java has already been accepted as the de-facto standard for devices. A large number of manufacturers are adopting J2ME technology, and as the range of devices using J2ME increases, newer profiles will need to be implemented. By providing a common minimum platform J2ME configurations and profiles enable creation of these custom editions in a structured way.

15.3.1 CDC

CDC is a configuration for high end devices with larger memory, in the range of 2 MB or more (at least 512K for the runtime environment, plus another 256K for applications), providing connectivity through some kind of network and some UI support. The CDC is a superset of the Connected Limited Device Configuration and is quite close to a conventional Java2 Standard Edition (J2SE) runtime environment.

The CDC specification defines

- A full-featured JVM, called the CVM.
- A subset of the J2SE 1.3 classes.
- APIs introduced in the CLDC—the Generic Connection Framework.

Eric Giguere, an undisputed guru on the subject has a lot of useful articles on his site "http://www.ericgiguere.com".

The CDC supports the following standard Java packages:

File I/O	java.io
Core Java system classes	java.lang
Networking support	java.net
Security framework	java.security
Text, dates, numbers, internationalization	java.text
Utility classes	java.util

The CDC defines three profiles. They are the Foundation Profile which is essentially the worker that provides support to other profiles, the Personal Profile which is personal Java retrofitted to J2ME and the Remote Method Invocation (RMI) profile for RMI support. We will now briefly describe these.

Foundation Profile

The Foundation Profile is targeted at devices supporting a strong network, but does not provide any UI (User Interface). Essentially it does the "useful" work. It is also used as a base by other profiles, which then build on its functionality by adding UI and other components. Typically Foundation Profile has the following requirements:

- Memory minimum 1024 KB ROM and 512 KB RAM for the profile and configuration.
- Stable Network Connectivity.

Personal Basis Profile

The J2ME Personal Profile is a reincarnation of the Personal Java Application Environment suitably modified to fit into the J2ME environment. Built on top of the Foundation Profile, it is backward compatible with Personal Java 1.1 and 1.2. It caters to devices that enjoy reliable and constant Internet connectivity and rich GUI. These devices are usually characterized by:

- A minimum of 2.5 MB of ROM and 1 MB RAM.
- Robust Internet connectivity.
- Rich (GUI) with browser like Internet support.

J2ME RMI Profile

The J2ME RMI (Remote Method Invocation) Profile, as the name suggests provides support for RMI across applications. It is built on top of the Foundation Profile and uses TCP/IP as the underlying connection protocol. It is interoperable with the Java2 Standard Edition (J2SE) RMI API 1.2 and higher. Details of the RMI Profile specifications can be found at JSR-66 home.

15.3.2 CLDC

As noted earlier, the Connected Limited Device Configuration (CLDC) is meant for low end, intermittently connected, battery-operated devices. A CLDC device is characterized by the following capabilities:

- A minimum of 128 to 512 KB for the platform.
- A 16-bit or 32-bit low end processor.
- A low bandwidth network with intermittent connectivity (mostly wireless GSM/GPRS/ CDMA).

CLDC runs on KVM, which is a highly optimized JVM for resource constrained devices. It includes just the basic classes from the java.lang, java.io and java.util packages, with a few additional classes from the new javax.microedition.io package. In particular, the CLDC specification does not support the following Java language features:

- Floating point calculations
- Object finalization
- Custom class loader

- Error classes

For a detailed index of available classes please refer to the CLDC specifications which can be freely downloaded from the Sun's site.

A notable difference between the CLDC and J2SE is in the process of class verification. Class files are required to be verified by a class verifier, through a process called preverification. This is done before the application can be loaded to the device. The J2ME VM is similar to the JavaCard VM in this respect. This allows the VM to be leaner. At runtime, the VM uses information inserted into the class files by the preverifier to perform the final verification steps.

Though CLDC is a subset of the CDC, yet the two are quite different. The two configurations are independent of each other, and cannot be used together to define a platform. Figure 15.4 shows the relationship between the two configurations and the parent J2SE platform.

Sun has released two profiles that sit atop the CLDC. They are the Mobile Information Device Profile (MIDP) and the Personal Digital Assistant Profile.

MIDP

J2ME Mobile Information Device Profile is by far the most popular and widely supported one. It was also the first profile to be released. Currently, its a 2.0 version. The MIDP specs were written by an expert group, the Mobile Information Device Profile Expert Group, an international forum represented by several leading companies in the mobile device industry.

It provides classes for writing downloadable applications and services that are of interest to the consumer for example, games, commerce applications, personalization services, etc.

The MIDP profile requires the devices to have the following capabilities:

- A minimum of 512 KB for the platform.
- Intermittent connectivity to some type of wireless network.
- Limited user interfaces.
- Some kind of input mechanism.

For further details please refer to the specification document available at JSR 37/JSR 118 home.

The MIDP specifies the following APIs:

Application control	javax.microedition.midlet
User interface	javax.microedition.lcdui
Persistent storage	javax.microedition.rms
Networking and local IO	javax.microedition.io, java.io
System classes and interfaces	java.lang
Utility objects	java.util

PDAP

J2ME Personal Digital Assistant (PDA) Profile is a recently released profile catering specifically to the PDA market. It lies on top of the CLDC specification. It provides user interface and data storage APIs for devices with the following resource constraints:

- Minimum 1000 KB for the platform.
- Battery powered low power devices.
- Good UI capability with a resolution of 128x128 pixels and depth of 1 pixel, a touch screen, and or character input mechanism in the form of a T keypad or a full "qwerty" keyboard.
- An intermittent and low bandwidth two-way connectivity.

For further details please refer to the specification document available at JSR 000075 home.

15.4 Programming for CLDC

Covering the entire gamut of profiles mentioned above is beyond the scope of this book. We will take one profile from the CLDC group, the MIDP. MIDP has been chosen as it is the most widely used and has also created a good amount of noise with the release of its version 2.0. Here unless specifically referred to as introduced by MIDP 2.0 they are applicable to both the versions. The reader, however, is encouraged to study the vendor specific features too as they allow the developer to use the support of the underlying native environment.

A MIDP application is called a MIDlet. It is a take-off on the traditional applet, as the two are similar in some ways. We now look at the MIDlet Model.

15.4.1 The MIDlet Model

A typical MIDP application or MIDlet sits atop the MIDP which in turn requires the services of the CLDC and the VM below it. Finally it is the device hardware that executes instructions on behalf of the software layers above. Figure 15.5 shows a top down view of a MIDP application.

Like an applet a MIDlet needs an execution environment. The browser's equivalent in the MIDP world is called Application Management Software or AMS. It is a device resident software (normally provided by the device vendor) and all MIDlets run within the context of an AMS. This is also required because the handset needs to respond to events outside the MIDlet scope; for example, a call might need to be answered while reading an e-mail. All MIDlets are registered with the AMS during installation. A set of MIDlets can be grouped together into a MIDletsuite. Other than managing the MIDlet, the AMS is also responsible for Application Provisioning and Application Removal.

15.4.2 Provisioning

Provisioning is the process of application discovery, download and installation. PDAs allowed this by downloading the applications on to a PC and then using a serial cable to transfer it to the device. The solution defeats the very purpose of mobility.

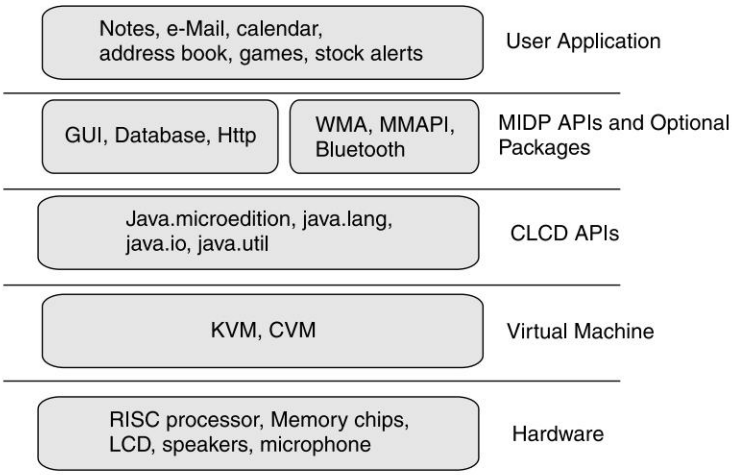


Figure 15.5 A Top Down View of a Typical MIDP Application

Provisioning includes

- **Search:** Which can be performed by the user manually entering the URL where the application is hosted or using a device resident browser step 1 in Figure 15.6.

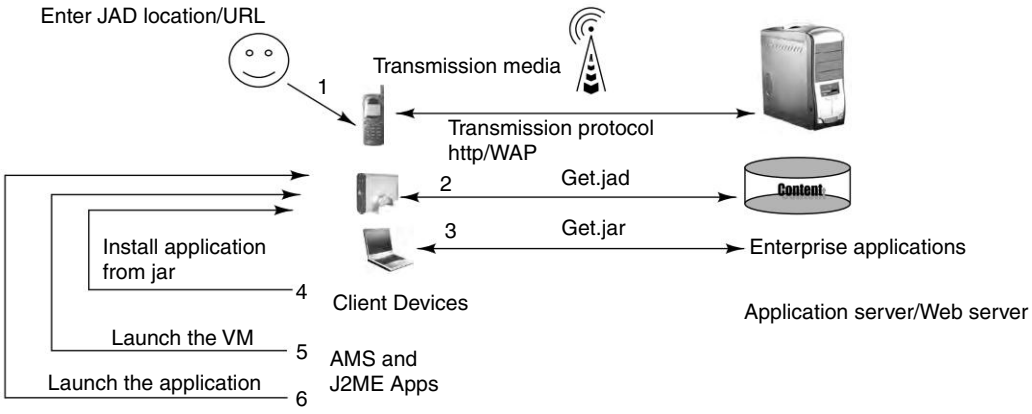


Figure 15.6 Provisioning a MIDP Application

- **Retrieve:** The descriptor file, which includes the application details, checks for version, and compatibility issues, is retrieved (step 2).
- **Download:** The appropriate jar file is downloaded. All middle suites are packaged into a jar file (step 3).
- **Install:** Once the download is completed the AMS is called to install (step 4).
- Finally the VM is launched (step 5) and the
- **Applications** are launched (step 6).

The process includes appropriate user interactions, including payment authorizations. Figure 15.6 depicts the entire process.

MIDP 2.0 has introduced a more interesting and unique capability in the form of a push registry. (MIDP 1.0 specifications allowed only pull applications, i.e., the applications had to initiate the transaction.) These features are discussed further later.

15.4.3 The MIDlet Lifecycle

As shown in Figure 15.7, a MIDlet has three states. A MIDlet class extends javax.microedition.midlet. MIDlet defines the corresponding life-cycle notification methods. These lifecycle methods allow the AMS to notify and request MIDlet state changes.

- Applications are launched either by a user selection or in response to an external event from the push registry. On being activated by the AMS, the MIDlet is constructed but is still inactive. This is the reason why resource allocation is not advisable in the constructor. Now the MIDlet is in a *paused* state.
- Once constructed, the AMS initializes and activates the MIDlet by invoking its startApp() method. The MIDlet now changes to active state. If the initialization fails a javax.microedition.midlet.MIDletStateChangeException is thrown and the state is changed to *destroyed*.
- A transition from *active* state to the paused state is initiated by the AMS by calling the MIDlet's pauseApp() method or by the MIDlet itself through the MIDlet context. In this state a MIDlet should release all its resources.
- A MIDlet can be *destroyed* from either *active* or *paused* state, If destruction is initiated by the AMS, the MIDlet's destroyApp( ) method is invoked with a boolean parameter true/false for optional or forced destruction. The optional destruction can result in a MIDletStateChangeException. The possible state changes and the transitions are shown in Figure 15.7.

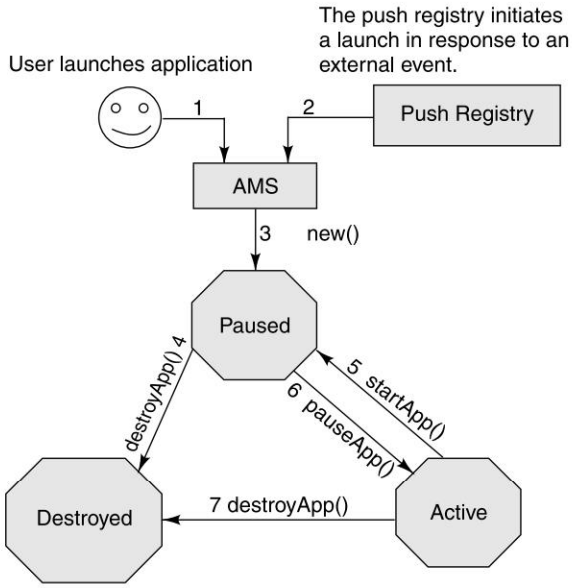


Figure 15.7 MIDlet Lifecycle

Briefly then the three states of a MIDlet are:

- *paused:* The MIDlet is constructed but inactive; transition occurs by a call to pauseApp( ).
- *active:* The MIDlet is active and can process requests; transition occurs by a call to startApp( ).
- *destroyed:* The MIDlet has been destroyed and is ready for garbage collection; transition occurs by a call to destroyApp( ).

Apart from the lifecycle methods javax.microedition.midlet.MIDlet also defines some MIDlet Context methods. These are as follows.

- getAppProperty( ) which retrieves properties from the *application descriptor*. Properties are name-value pairs in the JAD file. The contents of a sample JAD and manifest files are shown in Figure 15.7.
- resumeRequest( ) is a request to the AMS to reactivate the MIDlet. However, it is the prerogative of the AMS to decide if and when to reactivate the MIDlet. Reactivation makes a call to the MIDlet's startApp ( ) method.
- notifyPaused( ) is an alert to the AMS that the MIDlet is transitioning to a paused state.
- notifyDestroyed( ) informs the AMS that the MIDlet will now destroy itself.

Sample JAD file contents

MIDlet-1: GUI, GUI.png, FirstMIDlet
MIDlet-2: SimpleTextBox, SimpleTextBoxMIDlet
MIDlet-3: TextBoxWithCommandListener,
TextBoxWithCommandListenerMIDlet
MIDlet-4: CompleteTextBox, CompleteTextBoxMIDlet
MIDlet-5: SimpleList, SimpleListMIDlet
MIDlet-6: CompleteList, CompleteListMIDlet
MIDlet-Jar-Size: 5698
MIDlet-Jar-URL: GUI.jar
MIDlet-Name: GUI
MIDlet-Vendor: Sun Microsystems
MIDlet-Version: 1.0
MicroEdition-Configuration: CLDC-1.0
MicroEdition-Profile: MIDP-2.0

JAD file is a text file that lists important information about a set of MIDlets packaged together into a single JAR file (*a MIDlet suite*)

Sample MF file contents

MIDlet-1: GUI, GUI.png, FirstMIDlet
MIDlet-2: SimpleTextBox, SimpleTextBoxMIDlet
MIDlet-3: TextBoxWithCommandListener
TextBoxWithCommandListenerMIDlet
MIDlet-4: CompleteTextBox, CompleteTextBoxMIDlet
MIDlet-5: SimpleList, SimpleListMIDlet
MIDlet-6: CompleteList, CompleteListMIDlet
MIDlet-Name: GUI
MIDlet-Vendor: Sun Microsystems
MIDlet-Version: 1.0
MicroEdition-Configuration: CLDC-1.0
MicroEdition-Profile: MIDP-2.0
The *manifest* is the standard JAR manifest packaged with the MIDlet suite.

15.4.4 First MIDLet

The code for our FirstMIDlet would look as below:

```
import javax.microedition.MIDlet.*;
import javax.microedition.lcdui.*;
/**
* @author ryavagal
* @version
*/
public class FirstMIDlet extends MIDlet
{
   private Display display;

   public void startApp( ) throws MIDletStateChangeException
{
         if( display == null ){
            init( ); // one-time initialization
      }
}
public void pauseApp( ) {
   }
public void destroyApp(boolean unconditional) throws
MIDletStateChangeException {
      exit( );
   }
```

```
    private void init( ){
        display = Display.getDisplay( this );
            // initialization stuff goes here
    }
        public void exit( ){
            // It is a good practice to release all
resources and cleanup
            notifyDestroyed( );
        }
}
```

But before you can see this in action you will need to set up your development environment. The easiest way to set up the development environment is to visit http://java.sun.com/j2me/index.jsp where detailed download and installation instructions are given. Integration with the IDE of our choice will involve further investigation on our part. For our purpose we will use the WTK2.0 running on a Windows platform. Details are excluded here as it will only be duplication of content from the Sun's site.

Once we have successfully installed the WTK we can start it from the start menu. This will look like as shown in Figure 15.8. Click on "Open Project" to view any of the sample programs. To run the applications click "Run".

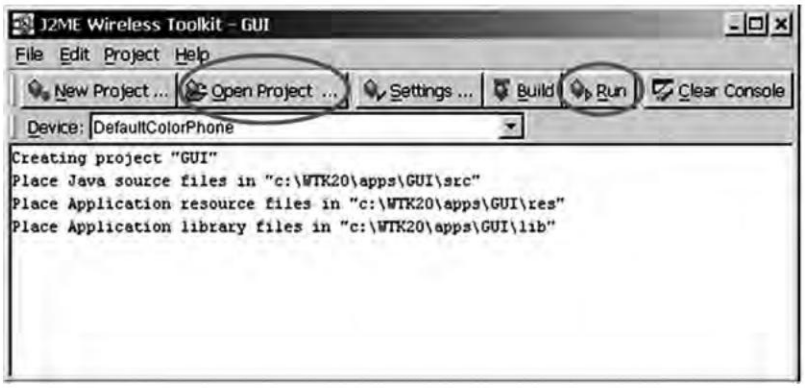


Figure 15.8 Running Samples from WTK

15.4.5 Creating a New Application

To create our own application we need to do the following.

1. Click on "New Project".
2. Enter appropriate names for the application and the application class and click "Create Project". The resulting window is shown in Figure 15.9.
3. By default WTK will create the folder Structure under the Apps folder of the WTK root (Fig. 15.10).

4. Use any editor, (even Notepad will do) to enter the program above and save it under the source folder of the application we created in step 2 .
5. Go back to the WTK and click "Build". If everything is OK we will get a Build complete message (Fig. 15.11).

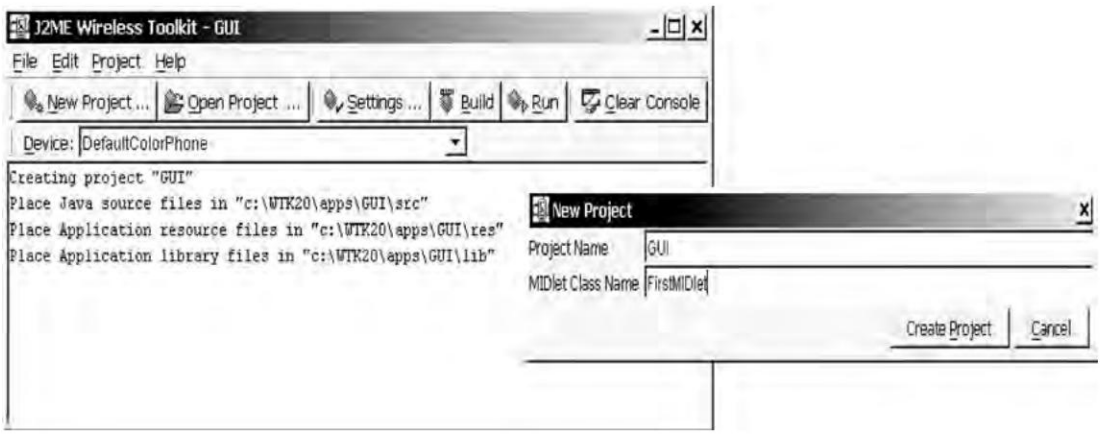


Figure 15.9 Creating a New Application

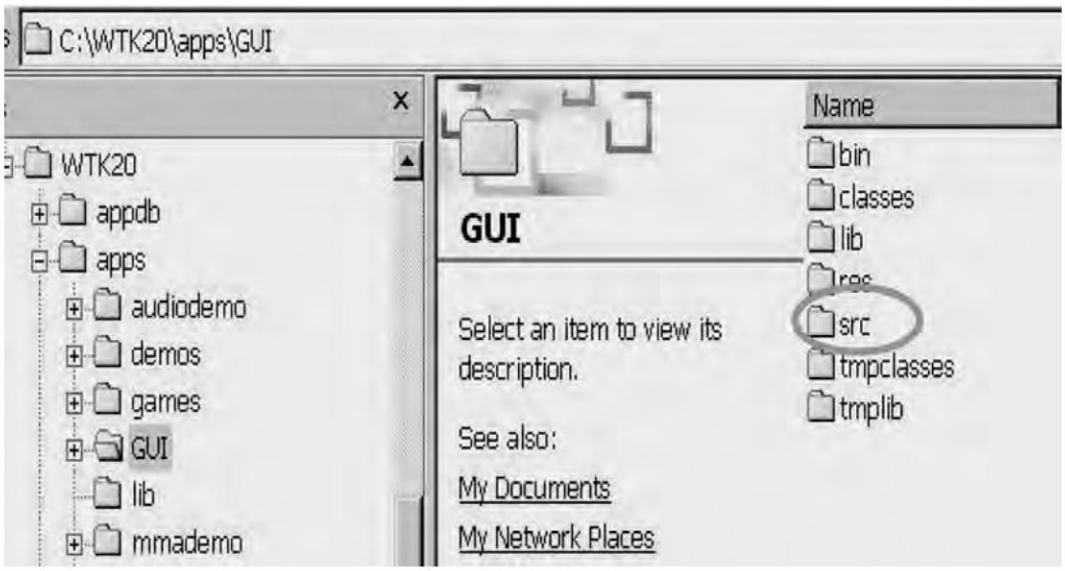


Figure 15.10 Application Project Directory Structure

6. Clicking on run will launch the emulator which lists the Midlet currently registered with the WTK AMS. For steps 4, 5 and 6 refer to Figure 15.11.

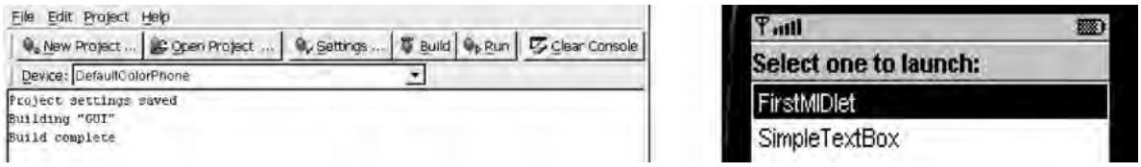


Figure 15.11 Building the First MIDlet

But selecting the FirstMidlet and clicking Launch does nothing! That is because we have not done anything yet. If we take a closer look at the code above there is variable called Display. What is it? And what does it do? And the lcdui package?

We guessed right. We will be using it to display elements on the screen. GUI in MIDP has two core concepts, the *Display* and *Displayable*. In short the MIDP's display is represented by the *Display* class. All displayable elements are called *Displayable*. To show an element we use the setCurrent method of the *Display* class. The lifecycle of GUI interactions is shown in Figure 15.12.

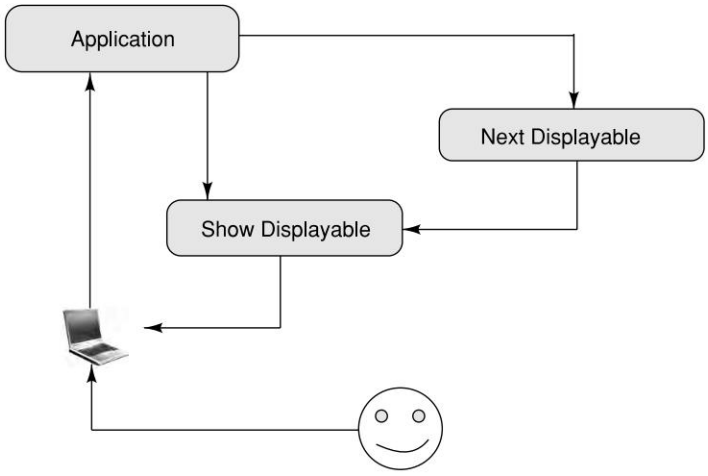


Figure 15.12 Application Control Flow

15.4.6 MIDlet Event Handling

User interactions generate events. These could be:

- Screen inputs.
- Item state change.
- Handset data update.

MIDP event handling mechanism is based on a listener model. It provides interfaces for each of the events mentioned above. These interfaces implement callback methods, which in turn invocate application-defined methods. These methods perform the desired functions in response to events. The three interfaces provided are: ItemStateListener, CommandListener, and RecordListener. Let us look at each in some detail.

Command Listener

The CommandListener as the name implies is responsible for notifying the MIDlet of any commands or events generated by the user. Objects extending it, implement the commandAction method. This method takes two parameters—a Command object and a displayable (Command c, Displayable d). This method implements the functionality that needs to be executed in response to the command event on the associated Displayable.

Displayable.setCommandListener (CommandListener L) sets the listener L to a Displayable.

Item State Listener

An ItemStateListener informs the MIDlet of changes in the state of an interactive item. It calls the itemStateChanged(Item I) method in response to an internal state change.

- Form.setItemStateListener(ItemStateListener L)sets the item state listener L for the given displayable.

Record Listener

RecordListener is related to database events which are discussed in the later section on RMS.

15.5 GUI IN MIDP

As shown in Figure 15.13 the "*Displayable*" has two main subclasses Screen and Canvas. The screen is a super class for a set of predefined UI elements. The predefined UIs are called High Level UI elements and the canvas elements are called Low Level elements. It is always preferable to use these as they involve less coding and are more portable. The Canvas allows the developer to have low level control on the screen. Games normally use the Canvas. Figure 15.13 shows the subclasses of the "*Displayable*".

15.5.1 High Level UI

We will begin with the High Level UIs and then proceed to the Low Level UIs. Let us begin by adding some of the displayables to our applications. Copy the FirstMidlet into a new program called SimpleTextBoxMIDlet.

To the init method, add the following code:

```
Display display = Display.getDisplay(this);
TextBox text = getTextBox( );
display.setCurrent(text);
```

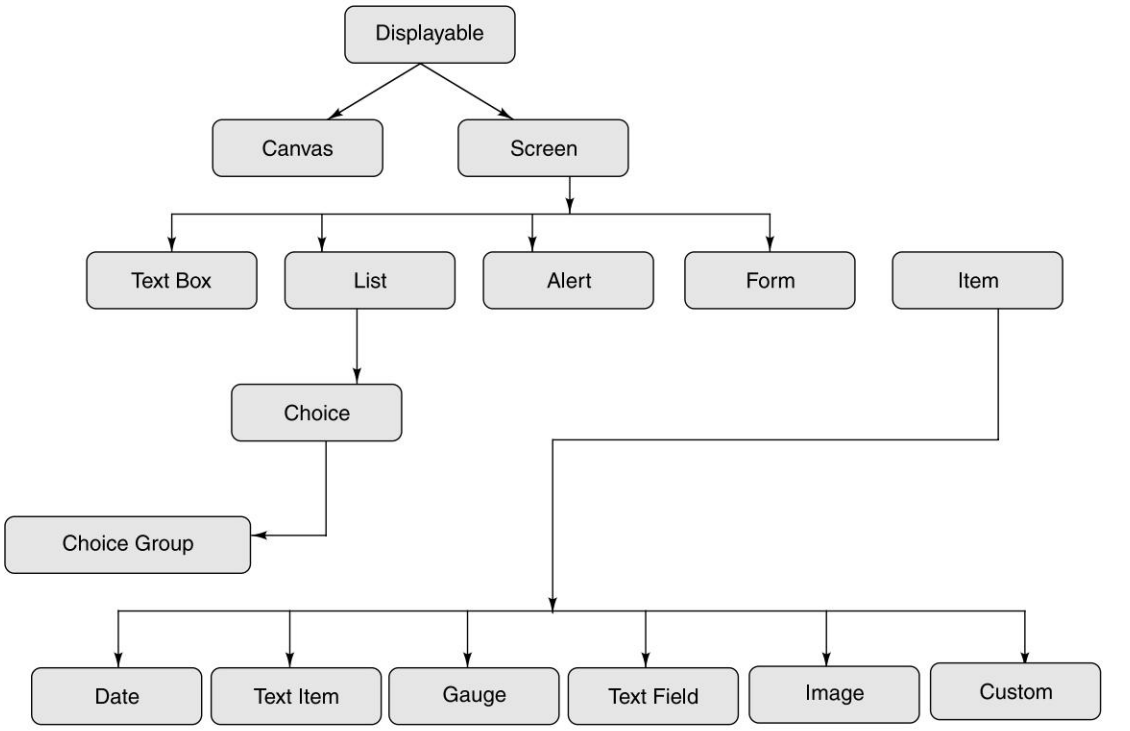


Figure 15.13 UI Elements in MIDP

Now add this new Method

```
public static TextBox getTextBox( )
{
   TextBox textBox = new TextBox("Hello Midlet", "", 50, 0);
   return textBox;
}
```

Save the file Build and Run. Wait a minute, why can't we see it in the list? We need to add the MIDlet to our suite first. Click on settings. We see the DialogBox showing the currently available MIDlets. Refer to Figure 15.14.

- Click on Add in this dialog to add a new MIDlet to our suite.
- Enter The Name to be Displayed and associate it with the Midlet class.
- Click OK.

Now select Run from the WTK toolbar and the newly created SimpleTextBox should appear (Fig. 15.15).

Select SimpleTextBox and Click Launch. We should see the TextBox (Fig. 15.16).

Lets revisit the code above. We first obtain the applications Display, Create the element we want to use and then set the Current to Display to our Element.

Figure 15.14 Adding a New MIDlet to the Suite

Figure 15.15 New Simple Text Box Appears

Figure 15.16 Text Box Appears

Forms

A form is similar to its html counterpart. It is used to hold other items and elements. In its simplest form the code for a form might look like

```
display = Display.getDisplay( this );
Form myform = new Form("Hello Form");
display.setCurrent(myform);
```

If we put it in the init method of the MIDlet above build and run we should get the window or form shown in Figure 15.17.

Figure 15.17 Form

Items

To do some useful work we have put some items into the form we created above. MIDP provides a set of predefined items. Table 15.1 shows a listing of all items available in MIDP.

ChoiceGroups

The ChoiceGroup Item allows users to select one or more elements from a group. These groups are similar to the "radio button", "check boxes" and "drop down" elements in the html parlance. AchoiceGroup item contains a simple string and an optional image per member in the group. ChoiceGroups are of three different types. ChoiceGroup and List have a lot of similarity in the options they support. The three types are listed in Table 15.2.

The ChoiceGroup constructor takes a label and a type value. Optionally images and hover text can also be added. Members can also be added after its creation, using the append( ) method.

Table 15.1 Subclasses of Items

Item	*Description*
ChoiceGroup	Allows user to select one or more elements from a group.
DateField	Counterpart of Java date field. Used for date and time values.
Gauge	A bar graph representation used for integer values.
ImageItem	To display an image.
StringItem	Equivalent of html's label widget and used for non-interactive text.
TextField	For text input.
CustomeItem	A user defined item (MIDP2.0).

Table 15.2 Choice Type Constants

Constant	*Value*
EXCLUSIVE	Allows only one element to be selected at a time.
MULTIPLE	Allows the selection of multiple elements.
POPUP	Is a dropdown like construct that allows a single option selection.

```
ChoiceGroup grp = new ChoiceGroup("Select One",Choice.POPUP);
   grp.append("Implicit", null);
   grp.append("Explicit", null);
   grp.append("Multiple", null);
   grp.setLayout(Item.BUTTON);
```

grp.setLayout (Item.BUTTON) allows us to specify the layout, in this case in a button format. To add this to the form use myform.append(grp).

The resulting output is shown in Figure 15.18. The other options of exclusive and multiple are common with another component the List and will be discussed in the section that pertains to LIST.

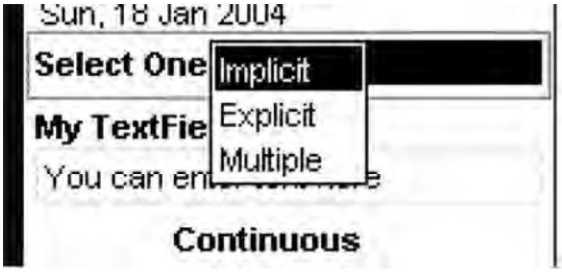


Figure 15.18 ChoiceGroup of Type POPUP

DateField

The DateField is an interactive item to enter/retrieve date and time information. DateField extends the Item class so it can be placed on Form objects. As usual we will create the date component and append it to the form. Display option for a DateField is specified using input mode constants in the constructor. Possible DateField mode constants are listed in Table 15.3.

Table 15.3 DateField Constants

Constant	*Value*
`DATE`	For date only.
`DATE_TIME`	For both date and time information.
`TIME`	For time only.

So here goes:

```
DateField dtf = new DateField("Today",DateField.DATE);
   Date date = new Date( );
   dtf.setDate(date);
   dtf.setLayout(Item.HYPERLINK);
```

And say myform.append(dtf); we should get the output as shown in Figs 15.19 and 15.20.

TimeZone and locale issues complicate date and time formatting. MIDP does not include the Java DateFormat, only a subset of the Calendar, Date and TimeZone classes. The two argument constructors above use the default time zone of the device.

To explicitly specify a time zone MIDP provides a constructor with the following format.

public DateField(String label, int mode, java.util.TimeZone zone); where, *zone* specifies the TimeZone. What we saw above was an un-initialized date:

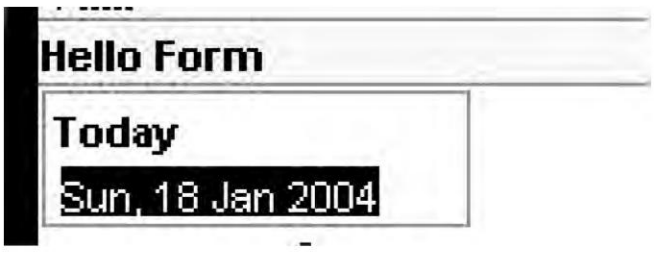


Figure 15.19 Date Object

To initialize the field to a particular date or time, we will have to use setDate and pass in a java.util.Date object initialized to the correct value:

```
Calendar cal = Calendar.getInstance( );
   cal.set( Calendar.MONTH, Calendar.FEBRUARY );
```

```
cal.set( Calendar.DAY_OF_MONTH, 06 );
c.set( Calendar.YEAR, 1977 );
cal.set( Calendar.HOUR_OF_DAY, 04 );
cal.set( Calendar.MINUTE, 00 );
cal.set( Calendar.SECOND, 00 );
cal.set( Calendar.MILLISECOND, 0 );
Date currenttime = cal.getTime( );
DateField df = new DateField( null, DateField.DATE_TIME
);
df.setTime(currenttime);
```

Figure 15.20 Date Selection Dialog

A Date object represents the number of milliseconds since midnight, 1 January 1970.

Note: It is always wiser to use Calendar class for date manipulation and not the raw milliseconds value stored in a Date object.

To obtain or update date/time, we use `getDate` method:

```
DateField df = ....;
Date editedDate = df.getDate( );
```

Gauge

The Gauge item can be thought of as a progress bar. The constructor takes a label, a Boolean flag, where true indicates interactive and false indicates non-interactive, an upper limit and an initial or starting value. An interactive Gauge as the name implies, allows the user to change the value using some device-dependent input method.

The following code snippet shows the construction of different types of Gauge. The current value of the Gauge can be set using the method `setValue( )` and read using the method `getValue( )`. Analogous `setMaxValue( )` and `getMaxValue( )` methods let we access the maximum value of the Gauge.

We will take a look at the timers in a later section; the code is otherwise self explanatory.

```
Gauge gug = new Gauge("Continuous Running Guage",false,
Gauge.INDEFINITE,Gauge.CONTINUOUS_RUNNING);
   gug.setLayout(Item.LAYOUT_CENTER);
   Gauge gug1 = new Gauge("I do nothing. Use a timer
to upgrade me",false,10,0);
   gug1.setLayout(Item.LAYOUT_EXPAND);
   //Task and Timer for a progress bar.
   MyTask task = new MyTask( );
   timer.schedule(task,500,1000);
   Gauge gug2 = new Gauge("Interactive Guage Click to
   inc or dec me",true,10,0);
   gug2.setLayout(Item.LAYOUT_EXPAND);
```

The output of the code is shown in Figures 15.21 and 15.22.

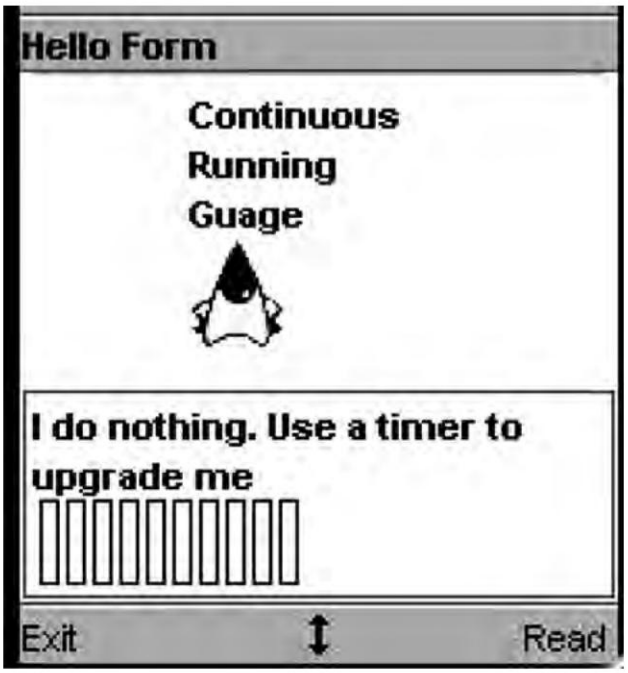


Figure 15.21 Non Interactive Gauge

Figure 15.22 Interactive Gauge

StringItem

`StringItems` are similar to labels in functionality. They have a name or label and the display string. These are extremely important for I18N or internationalization. The following code snippet shows the creation of a simple text label.

```
StringItem stringitm = new StringItem("StringItem","This is
a StringItem followed by a spacer");
StringItem stringitm1 = new StringItem("StringItem","This
is a StringItem after a spacer");
```

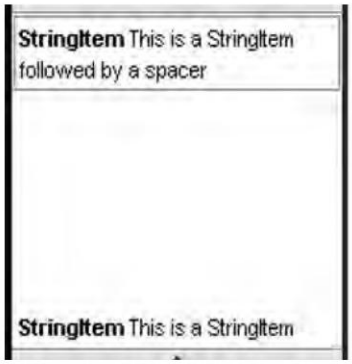


Figure 15.23 String Item

The label of the StringItem can be accessed using the `setLabel()` and `getLabel()` methods inherited from `Item`. To access the text, we can use the methods `setText()` and getText().

```
myform.append(stringitm);
myform.append(stringitm1);
```

The resulting UI is shown in Figure 15.23. Now where did that big space come from? This is a spacer item introduced in MIDP2.0. It is a non-interactive item used to correctly position the elements on the screen. Normally we define the size of the empty space as `minWidth and minHeight`, both integers. Its usage is very simple.

```
Spacer spacer = new Spacer(100,100);
myform.append(spacer);
```

ImageItem

The ImageItem is also a non-interactive item. Its constructor takes an image object, a layout parameter, and an alternative text string to be displayed if the image cannot be displayed. The image source is the location of the image file. The layout parameter can take one of the following constant values `LAYOUT_CENTER, LAYOUT_DEFAULT, LAYOUT_LEFT, LAYOUT_NEWLINE_AFTER, LAYOUT_NEWLINE_BEFORE, LAYOUT_RIGHT. The values are self explanatory.`

The following code snippet shows how a center aligned ImageItem is added to the sample MIDlet: The Duke comes with the WTK, but we could use any image of our choice

```
Image img = Image.createImage("/Duke.png");
   ImageItem imgItem =
   new ImageItem("I am an ImageItem", img,
      Item.LAYOUT_CENTER, null,
      Item.BUTTON);
   imgItem.setDefaultCommand( new Command("Image Command",
Command.ITEM, 1));
   imgItem.addCommand(CMD_IMG);
   imgItem.setItemCommandListener(this);
```

The output is as shown in Fig. 15.24 and Fig. 15.25.

We have set the command listener. So what happens when we click the corresponding Image Command shown in Figure 15.25.

We should see the following lines printed on our console.

I am here now Image Command

I am out

We will study more about the command listener soon.

Figure 15.24 Image Item

Figure 15.25 Image Command

TextFields

The `TextField` class handles all text input. The constructor takes four parameters—a label or the name of the Item, initial text (this can be an empty string if we don't want to display anything), the input length, and constants that indicate the type of input allowed.

```
TextField txtfld = new TextField("My TextField","You can enter text
here",50,TextField.ANY);
```

Valid constraint values are listed in Table 15.4.

Table 15.4 TextField Constraint Constant Values

Constant	*Value*
ANY	Any text can be entered.
EMAILADDR	A valid e-mail id of type userl@domainname.com.
NUMERIC	Numeric values.
PASSWORD	A masked text is allowed. It displays all "*".
PHONENUMBER	A valid phone number.
URL	A valid http URL.

Refer to the Java docs for more options.

A textfield is added to a form using the append function. Example: myform.append(txtfld); sample output is shown in Figure 15.26.

To read the values from an updated textfield we first need to attach a command listener to it and then check in our command action class. A sample code block might look like

```
if(itm instanceof TextField)
{
   TextField txt = (TextField)itm;
   buf.append(delim);
   buf.append(txt.getString( ));
}
```

The output of a read command will be the newly entered string displayed on the console.

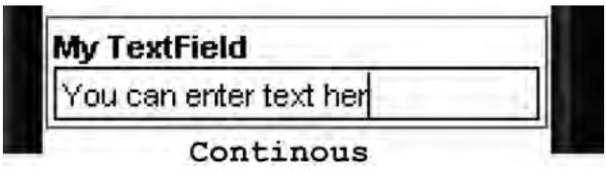


Figure 15.26 TextField

Custom Element

The new user interface API in MIDP 2.0 includes a class specifically intended for building custom UI elements. CustomItem extends the Item class, offers a simple template for developing custom items. The bonus is, its extensibility that allows for building on top of the existing Item base class features. We will now look at how to use this exciting new item.

Our custom Item is the Directory Item which most of us are familiar with. In the most simplified form this will display a + symbol. When selected, it will expand to display a sub-level item. The process involves two steps. First we need to extend the item class and create our new item. Second we will use this item in a form.

Figure 15.27 shows the output of our custom item which we have titled the DirListing.

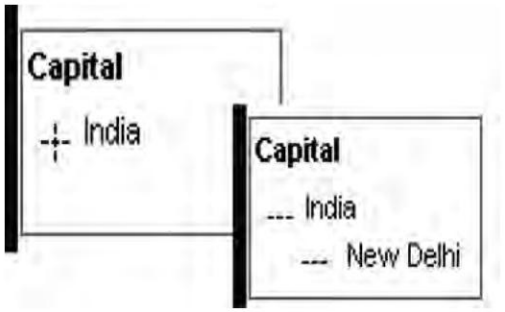


Figure 15.27 Custom Item (Directory Item)

The code for the DirListing begins by extending the CustomItem class. A new feature introduced in MIDP 2.0 is the `ItemCommandListener` class. This allows a command listener to be specified for an item rather than for the whole form. We will use it here to implement the expand function.

```
public class DirListing
extends CustomItem
implements ItemCommandListener {
```

Custome Item is an abstract class, hence we have to implement all the methods which are given here. The values returned are optimized based on a trial-error method for My Display Settings and may not hold good for ours.

```
protected int getMinContentHeight( ) { }
protected int getMinContentWidth( ) { }
protected int getPrefContentHeight(int width) { }
protected int getPrefContentWidth(int height) { }
protected void sizeChanged(int w, int h){ }
```

In a real world application these functions use the values based on the screen co-ordinates. We will see the details of that in the section on Low level APIs.

The sizeChanged( ) method take height and width parameters that are used by the application to redraw an item. Size changes occur due to user actions that change the screen size such as minimizing or reducing it.

paint( ) as the name implies does the actual job of painting the item.

Cursor movements would result in a call to traverse( ) method. It will return true if the traversal is internal which will then activate the item, or false when another item in the form is selected. Other information provided by traverse( ) is the direction of traversal, available area, etc. We will now see the implementation details of the CheckItem.

The `DirListing` item has two basic functions: it responds to the movement of the cursor and expands the selected value (clicking on the country gives the capital). The listing below gives us the constructor of the item.

```
private final static Command CMD_EXPAND = new Command("Expand",
Command.ITEM, 1);

public DirListing(String root, String Country, String Capital) {
   super(root);
   this.Country = Country;
   this.Capital= Capital;
   setDefaultCommand(CMD_EXPAND);
   setItemCommandListener(this);
}
```

The constructor takes the country and capital and sets the Expand command. The next piece is the paint( ) method which is responsible for drawing the item on the screen.

```
protected void paint(Graphics g, int w, int h) {
   g.setStrokeStyle(Graphics.DOTTED);
   if (UnExpanded==1) // The item is unexpanded so
draw the + and the associated text is the country.
      {
         g.drawLine(10,10,10,20);
         g.drawLine(5,15,15,15);
   g.drawString(Country, 22, 4, Graphics.TOP|Graphics.LEFT);
      }
      else  { // The state is expanded so the symbol
is - and both the country and capital need to be shown
            g.drawLine(5,15,15,15);
```

```
                g.drawString(caption1, 22, 4, Graphics.TOP|Graphics. LEFT);
                g.drawLine(20,30,30,30);
                g.drawString(Capital, 40, 20, Graphics.TOP|Graphics. LEFT);
                    }
        }
```

Note, for the sake of simplicity we have used absolute values. But in a live application it is advisable to use relative coordinates. Finally we reach the Command handler and the traverse( ) method.

protected boolean traverse(int dir, int viewportWidth, int viewportHeight, int[ ] visRect_inout)

```
{
        switch (dir) {
           case Canvas.DOWN:
              repaint( );
              break;
           case Canvas.UP:
              repaint( );
              break;
        }
        return false;
    }
```

The traverse is quite simple as we just repaint the item. In the command we first toggle the state between UnExpanded=1 and 0 and call repaint( ).

```
    public void commandAction(Command c, Item i) {
        if (c == CMD_EDIT) {
           if(UnExpanded ==1 ) UnExpanded =0;//location;
           else UnExpanded =1;
           repaint( );
           notifyStateChanged( );
        }
    }
```

Now our item is ready to be used. We have a new class, the DirListingDemo. The following code gives a sample form that includes the DirListing item.

```
Form mainForm = new Form("DirListingDemo");
mainForm.append(new DirListing("Capital", "Pakistan",
"Islamabad"));
mainForm.append(new DirListing("Capital", "India", "New
Delhi"));
```

List

We will continue with our discussion of the GUI components. We take up a list. A list, as the name suggests, is a group of options that the user can select from. Lists are basically of three types.

- IMPLICIT: This type is associated with the select command. When selected the event can be trapped in the command listener.

- EXCLUSIVE: This allows the user to select any one option and is similar to the choice group item.
- MULTIPLE: As the name suggests it allows the user to select multiple values.

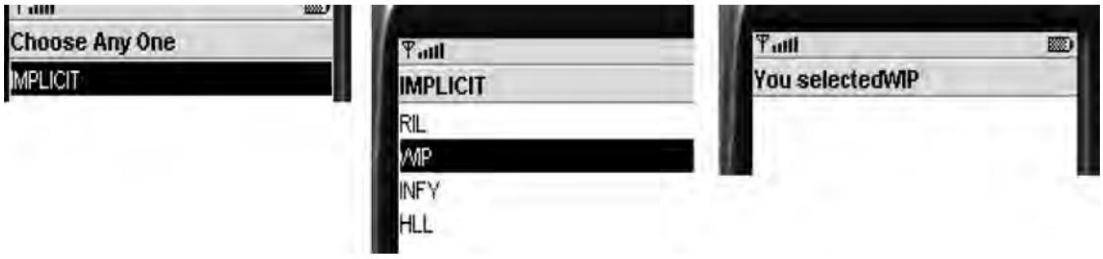


Figure 15.28 Choice Group (Implicit)

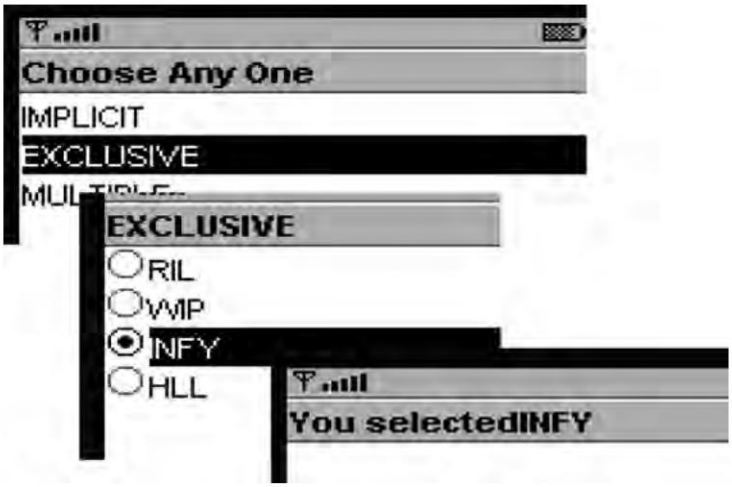


Figure 15.29 Choice Group (Exclusive)

The constructor would be given by List list = new
List("IMPLICIT",List.IMPLICIT); where the arguments to the constructor are the label and the type. To read the selected value from a list we can use the following piece of code.

```
   List list = ((List)display.getCurrent( )); //Obtain the
   currently displayed list.
// For both IMPLICIT and EXCLUSIVE only one option can be selected.
if(list.getTitle( ).equals("IMPLICIT") ||list.getTitle( ).
equals("EXCLUSIVE"))
{
String msg = list.getString(list.getSelectedIndex( )); //
```

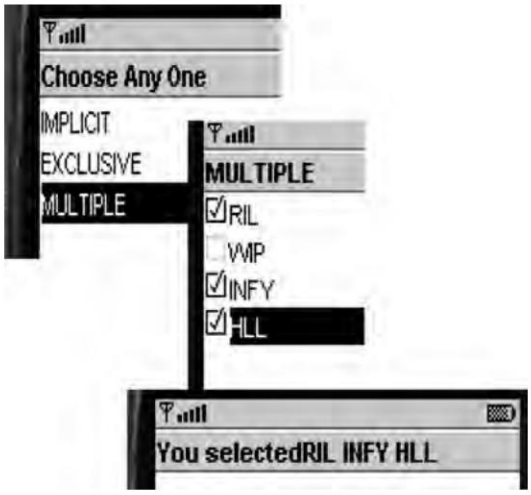


Figure 15.30 Choice Group (Multiple)

```
Get the label from the index of the value selected.
   Alert alrt = new Alert("You selected" + msg);
   display.setCurrent(alrt);
}
//For MULTIPLE more than one options can be selected the use age
//therefore is slightly different.
else
{
   boolean[] sel = new boolean[list.size( )];
   list.getSelectedFlags(sel); //Read all the selected
   values into a Boolean array
   StringBuffer buf = new StringBuffer( );
   for(int i=0;i<list.size( );i++)
   {
      if(sel[i]) //if the option has been selected
      store it some where
      {
         buf.append(list.getString(i));
         if(i!= list.size( ))
         buf.append(" ");
      }
   }
   String msg = buf.toString( );
   Alert alrt = new Alert("You selected" + msg);
//Display the values selected.
      display.setCurrent(alrt);
   }
```

The output of the sample code is shown in Figures 15.28, 15.29 and 15.30. What we saw so far were high-level APIs. What that means is, the users only had to instantiate and use these objects. The actual implementation and on screen rendering was the responsibility of the OEM or the MIDP implementer. Though easy to use and highly portable they are very restrictive for applications like games that need more control on the display. For these we have a set of low-level APIs derived from the canvas class.

15.5.2 Low-level GUI Components

Low-level APIs, as the name suggests, allow the application to control Displayable on the screen and allow the program to directly draw on the screen using the screen coordinates. This is specifically needed for applications like games and business tools that need to show graphs, bars, pie charts, etc. The package to be used is the javax.microedition.lcdui. To implement such control we need another class, mentioned above—the canvas class. The canvas is a blank screen. We can draw lines, text and shapes on this screen. To show the low-level capability of J2ME we will take two examples here.

Canvas

The first is our good old custom Item DirListing. Yes, we know we discussed it in the section on High-Level UI APIs. The noticeable difference between the CustomeItem and other items in that section is the fact the custom item is defined by the item developer and the item code is responsible for handling its display response to cursor movements and other user interactions; while in the case of other items, all the responsibility of display and user interaction was the responsibility of the system. In essence, a CustomItem uses some of the low-level APIs but this implementation is abstracted from the user of the item. The low-level implementation of DirListing will use Canvas and not the item class.

The second is a simple application that will move the Duke across the screen. Obviously, after implementing these examples we cannot stake claim to fame as game developers. We, however, will have made a beginning. The WTK comes with some excellent examples for game implementations. We urge the readers to work through them.

The first sample is very simple. The DirListing class subclasses the Canvas class, which is an abstract class that extends Displayable, and overrides the paint( ) method. The traverse methods are used to detect cursor movements. The paint( ) method uses the drawing methods of the javax.microedition.lcdui.Graphics class.

For the purpose of this example we will enhance the paint( ) method to show unexpanded values in red color and the expanded values in blue. The updated code is listed below.

```
public void paint(Graphics g) {
      g.setColor(255, 255, 255);
      g.fillRect(0, 0, getWidth( ), getHeight( )); // Gives the
      coordinates of the screen and paints it white.
      g.setColor(255, 0, 255); if (checked==1)
      {
         g.drawLine(10,10,10,20);
         g.drawLine(5,15,15,15);
         g.drawString(Country, 22, 4, Graphics.TOP|Graphics.LEFT);
```

```
        }
    else { // The state is expanded so the symbol is - and both
    the country and capital need to be shown
        g.setColor(0, 0, 255);
        g.drawLine(5,15,15,15);
        g.drawString(Country, 22, 4, Graphics.TOP|Graphics.LEFT);
        g.drawLine(20,30,30,30);
        g.drawString(Capital, 40, 20, Graphics.TOP|Graphics.LEFT);
    }
```

As seen above, setColor takes parameters (red, green, blue) in that order. An important difference here is the event handler. The canvas class provides for different events like pointer movements, keyPressed events, notify events, etc. Here we show a simple implementation of the keyPressed event.

```
protected void keyPressed(int keyCode)
{
    if (this.getKeyName(keyCode).equals("DOWN")) // Each key is
    associated with a unique id and a corresponding name.
        {
            checked = 0;
            System.out.println(" down");
            repaint( );
        }
    if  (this.getKeyName(keyCode).equals("UP"))
        {
            checked = 1;
            System.out.println("up");
            repaint( );
        }
}
```

The code to invoke the canvas implementation is as follows.

```
public void startApp( )
{
    Canvas canvas = new MyCanvas( );
    Display display = Display.getDisplay(this);
    display.setCurrent(canvas);
}
```

An interesting technique that is often used in the gaming context is Double Buffering. It is usually used where smooth animation effects are desirable. The idea is to create an offscreen buffer maintained in memory. This buffer is used in lieu of the screen. Once the buffer is ready, it is copied to the display. As is evident this process is faster and hence provides smoother effects.

To implement double buffering, first create a image buffer equal to the size of the screen:

```
int width = getWidth( );
int height = getHeight( );
```

```
    Image buffer = Image.createImage(width, height);
    Next, attach a graphics context to the buffer:
    Graphics gc = buffer.getGraphics( );
    Now, we can draw to the buffer:
    // normally an animation
    // ..
Finally copy it to the screen by overriding the paint( ) method
    public void paint(Graphics g)
    {
       g.drawImage(buffer, 0, 0, 0);
    }
```

Note, some MIDP implementations implicitly support double buffering. To check we can use the isDoubleBuffered( ) method.

Game API

An interesting enhancement to the low-level API in the MIDP 2.0 is the Game APIs. We will use these for our second example while briefly describing some of them. The intention is to introduce the process of game development. All game APIs are in the package javax.microedition.lcdui.game

The first class in this package is the GameCanvas the a subclass of javax.microedition.lcdui. Canvas. As implied it inherits from canvas and basically provides the screen for a game. The GameCanvas provides an additional facility to query the current state of the game keys and synchronous refresh. There is a provision for a unique buffer. (The concept is analogous to an implicit double buffering.) A game class using the GameAPI would typically be defined as

```
class GameTest extends GameCanvas{ ….. }
```

A typical game logic works in a loop till EOG. Within the loop the application checks for user response, updates the game appropriately and displays the new game configuration. In addition to the methods inherited from the canvas, GameCanvas provides methods to obtain the buffer and flush it. It also provides for game keys up, down, right, and left. The implementation, however, is device/OEM dependent. The GameAPI provides constructs that help in drawing and rendering images. The first of such a construct is a layer.

Layer

A Layer can be considered as an application element's visual part, a player or a field, for example. It provides basic attributes like location, size, and visibility. As the above statement implies, a game application can use several layers. A LayerManager is a construct that can automate the rendering process for multiple layers. The layerManager uses the concept of a user view which represents the current user window.

```
LayerManager layer;
Create a new layer: Layer = new LayerManager( );
Insert a new layer obj into the layer: layers.insert(layer);
Remove an existing layer from the layer.layers.remove(layer, index);
Transfer the layer image to the game graphics: layers.paint (g, 0, 0);
To render the layer on to the screen: flushGraphics( );
```

Sprite

Another important construct is the Sprite. This is mainly used for animation purpose. The Sprite provides various operations including flip, rotation, collision detection and more. In short it simplifies the implementation of a game's logic.

We will now see a simple example of an implementation of a Sprite to move and rotate the Duke. The output is shown in Figure 15.31. More complex examples are available from Sun's site within the samples with WTK.

```
Image img = Image.createImage("/Duke.png");
   sprite = new Sprite(img);
   g = getGraphics( );
   sprite.paint(g);
flushGraphics( );
sprite.move(Xcord, Ycord);
   sprite.setTransform(Sprite.TRANS_ROT180);
   sprite.paint(g);
flushGraphics( );
```

Figure 15.31 Low-level UI

TiledLayer

The last of the constructs is a Tiledlayer. A Tiledlayer can be pictured as a brick wall in which each brick represents a tile. Multiple tiles together make up a single image object. Tiles can also be filled with animated objects whose corresponding pixel data can be changed very rapidly. Imagine a puppet dancing, where multiple tiles make up the puppet, we can change only the corresponding tile to say, shake its hand. This allows the application to render large images without actually using the resources required for one.

```
TiledLayer tiles = new TiledLayer(xcolumns, yrows, Image, tileW, tileH);
```

The call above creates a Tiledlayer that is a (xcolumns * tileW) by (yrows * tileH) grid. This grid consists of cells or tiles that are tileW by tileH each. Note that the image dimensions must be within the tile dimensions or its multiples therein. Initially, the layer, is empty. To modify it we use

```
setCell(int, int, int) or fillCells(int, int, int, int, int).
```

We have used tiles.fillCells(0, 0, Image.getWidtht, Image.getHeight, 1); to provide a background filler. The background tileset can now be initialized using updateTile(x, y); Where x and y are the cell coordinates. The final task is to load this on to the layer we created earlier.

```
layers.append(tiles);
```

Both Sprite and Tiledlayer extend layer and hence can be added to and removed from a layer. With this we come to the end of our discussion on GUI.

At this point it is appropriate to include a discussion on certain important considerations in mind while designing application UI.

15.6 UI Design Issues

- Entering text through the T keypad is not very attractive nor is filling out long forms. UI should be small, simple and easy to use.
- High-level API should be used whereever possible as these are portable across different handsets.
- While using low-level API, it is advisable to remain restricted to elements defined in the Canvas class.
- Device capabilities like screen width, height, and resolution vary from device to device and hence applications should adapt to the Low-level objects accordingly.

15.7 Multimedia

Another new entry in MIDP 2.0 is the Media API, a direct subset of the Mobile Media API (JSR-135) specification. This library provides support for audio capability. Note, there is no support for video or graphics formats. As Figure 15.32a shows, a J2ME audio application consists of three main parts.

- A manager that is responsible for creating and controlling the audio resources.
- A player, the workhorse that does the actual job of playing music.
- A control to regulate the features of the player. The manager also provides support information on property, content and protocol support. To see how the manager can be used to play a simple tone, just include the following code in any MIDlet.

```
Manager.playTone(ToneControl.C4, 100, 100)
```

ToneControl is an interface that applications implement to play a user-defined tone sequence, ranging from C1 to G9. We recommend that the user refer to the javadocs for more information. The ToneControl interface defines two constants Middle C (C4) and SILENCE. The second parameter is the duration for which the tone is to be played and the last is volume control.

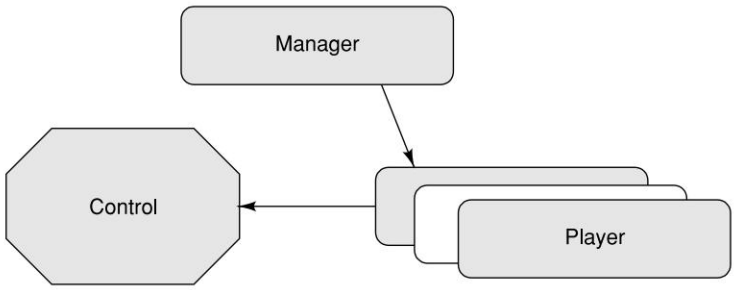


Figure 15.32a Media Player

Of course the major purpose of a manager is to create a player. This can be done through a call to createPlayer. But before we can make this player do something we need to identify what we want it to play. Sound inputs can be in three forms.

- Make a request to a .wav file over the network.
- Play a file stored in the .jar.
- Create a tune or a sequence of tunes.

A tune is essentially a byte array that contains information about the tune. A simple tune can be created using the following:

```
Private byte[ ] MyTune =
{
ToneControl.VERSION, 1,
C4,D4,E4,G4
};
Where
byte C4 = ToneControl.C4;;
byte D4 = (byte)(C4 + 2); // a whole step
byte E4 = (byte)(C4 + 4); // a major third
byte G4 = (byte)(C4 + 7); // a fifth
```

Creating a tune sequence is left as an exercise for our audience. To play this tune we first create a player

```
Player player = Manager.createPlayer(Manager.TONE_DEVICE_LOCATOR);
```

Alternate calls can be of the form

```
Player player = Manager.createPlayer("http://webserver/file.wav");
```

`A Player` plays the tune/tune sequence. *A Player's lifecycle has five states.* As shown in Figure 15.32b, a player begins its life in the *UNREALIZED* state. Here it is still unformed and requires additional information to acquire resources. At this point a call to realize( ) transitions it to the next state which is *REALIZED*. This method essentially allows the player to collect information by

which it can obtain resources. For example, which media device—audio or video or both would be required to play the content.

In the *REALIZED* state the player is ready to acquire the required resources. From here a call to *prefetch( )* transitions to the *PREFETCHED* State.

In the *PREFETCHED* state all initializations, for example, that acquire an audio device, happen.

A call to the *start( )* method will now start the player. The *start( )* method also causes a *STARTED* event to be fired. It is now in the *STARTED* state.

From here once a player reaches the end of the file it will automatically stop. Alternatively if the *stop( )* method is called it will stop and transition to the *STOPPED* state. A corresponding *STOPPED* or *END_OF_MEDIA* event will be triggered.

When a Player stops, it returns to the *PREFETCHED* state.

The following code snippet shows the transition through all five states.

```
player.realize( );
ToneControl toneControl =
(ToneControl)tonePlayer.getControl("javax.microedition.media.
control.ToneControl");
toneControl.setSequence(MyTune);
player.start( );
```

The last state in the lifecycle is the CLOSED state. Here it releases all the resources and must not be used again.

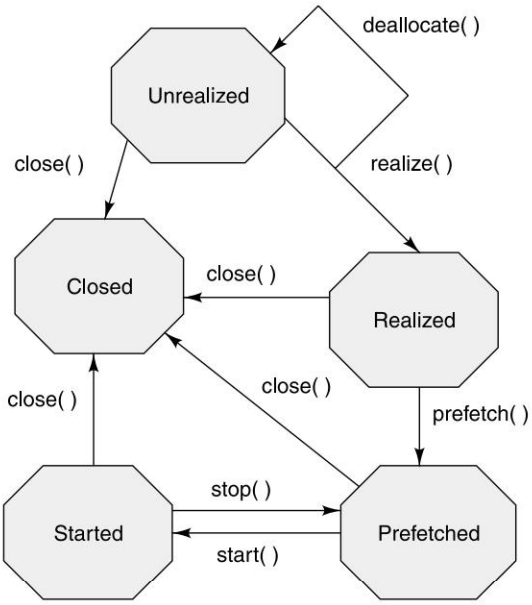


Figure 15.32b States in a Player's Lifecycle

With this we conclude our discussion of media API. Our next discussion concerns the need for persistent storage.

15.8 RECORD MANAGEMENT SYSTEM

Most application data needs to be stored for future use. For example, a PIM application needs to store the appointments information on the device. All devices provide some kind of persistent storage. However, heavy duty APIs to manage persistent data provided in Java2 Standard Edition (J2SE), such as JDBC, are not suitable for handheld devices. MIDP provides a lightweight persistent storage in the form of RMS (Record Management System). Record stores are essentially flat files in the binary format. Similar to the Palm OS RMS is also a record store consisting of a collection of records. In the WTK you will find these under the folder appdb. Record stores are platform-dependent and their definition lies in the domain of the OEM. Record stores are identified by names that are

- Unique to the MIDlet Suite(no two record stores in the same application can have the same name).
- All characters.
- Case sensitive.
- Cannot be more than 32 characters long.
- Stored as *"DBNAME.db*".

Note, records associated with the application are deleted when the application is uninstalled or deleted from the device.

Each record store has a version number as well as a date/time stamp. Both values are updated whenever an operation is performed on the store.

Till MIDP1.0 there was no mechanism to share the record stores across MIDlet suites. MIDP2.0 provides new API to share record stores between MIDlet suites. This is accomplished by

- Explicit permission granted by MIDlet that owns (creates) the store.

> "RecordStores are uniquely named using the unique name of the MIDlet suite plus the name of the Record Store. MIDlet suites are identified by the MIDlet-Vendor and MIDlet-Name attributes from the application descriptor. Sharing is accomplished through the ability to name a RecordStore created by another MIDlet suite. Access controls are defined when Record Stores to be shared are created. Access controls are enforced when RecordStores are opened. The access modes allow private or shared use with any other MIDlet suite".
>
> **Java Docs API documentation WTK2.0**

Within the Record Store, individual records can be considered to be organized as rows, with each row consisting of two columns: the first, a unique integer row identifier and the second, a series of bytes representing the data in the record. The row identifier is referred to as the *Record ID* and is the primary key for the record. RecordIDs are sequential values, i.e., the first entry will have the ID of 1, the next of 2, and so on.

The RMS APIs provide the following functionality:
- Create and delete record stores.
- Mechanism for sharing data across MIDlet suites.
- Allow application to add update and delete records within a record store.

The RMS are defined in the 'javax.microedition.rms' package. This package contains:

Interfaces:
- *RecordComparator* : To create a comparator object used for searching and sorting.
- *RecordEnumeration* : An enumeration object to traverse through the records.
- *RecordFilter*: To create a filter object to retrieve data based on filter criteria.
- *RecordListener*: To trap modify operations on the record store.

Classes:

RecordStore: Functionality is to
- Create/delete a record store.
- Add/update/delete records in the store.

Exceptions:
- *InvalidRecordIDException*: Thrown when the application tries to access a recordID that does not exist.
- *RecordStoreException*: A general exception.
- *RecordStoreFullException*: Thrown when there is no more storage space for RMS.
- *RecordStoreNotFoundException*: Thrown when the record store with the specified name does not exist.
- *RecordStoreNotOpenException*: Thrown when a modify operation is tried on a store that is closed.

We will see these in detail now. Any class wishing to implement DBfunctions has to implement the RecordListner interface. As a good coding practice this class should be separate from the main MIDlet class. In our case the RMS class implements the DB functions needed by the TestStore class.

```
public class RMS implements RecordListener {
   public RMS( ) throws RecordStoreException
   {
      MakeNewRecordStore(DefaultStoreName);
   }
public RMS(String RecStoreName) throws RecordStoreException
   {
      MakeNewRecordStore(RecStoreName);
   }
```

15.8.1 Create a New Record Store

Before any operation on a record store is possible, we need to create and open it. This is the functionality implemented in the MakeNewRecordStore. The sample code is

```
recordStore = RecordStore.openRecordStore(fileName, true);
   recordStore.addRecordListener(this);
```

MIDP defines a dual purpose API

```
"recordStore = RecordStore.openRecordStore(fileName, create
IfNecessary);"
```

The first argument to this call is the name of the store to be created or opened. The second is a *boolean* parameter which, if set to true, will cause the record store to be created if it does not exist. The call returns a reference to the RecordStore. If an application tries to open a store that is already open then a reference to the store is returned.

MIDP2.0 has introduced two new API for sharing record stores across the MIDlet Suites.

15.8.2 To Create a Sharable Store

"openRecordStore(fileName, createIfNecessary, authmode, writable)"

The first two parameters have already been discussed in the previous section while *authmode* defines the *authorization mode or permissions* on the store. Currently two modes are defined

- AUTHMODE_PRIVATE: Only the creator can access it. It behaves the same way as API.
- AUTHMODE_ANY: Any MIDlet can access the RecordStore. This should be used carefully as it could have privacy and security implications.

The last is a Boolean parameter writable which defines the access permissions on the store as read-only if false, else read/write.

Note, both the arguments above are ignored if the RecordStore exists.

The second API is

```
"openRecordStore(fileName, vendorName, suiteName)"
```

This call is used to open a shared store belonging to another MIDlet/MIDlet suite, where

```
vendorName is the vendor of the owning MIDlet suite and suiteName,
the name of the MIDlet suite that created and owns the store.
```

In response to this call the following may happen.

- Access is granted if the authorization mode of the RecordStore allows access, i.e., authorization mode set to AUTHMODE_ANY.
- If the record store is already open, a reference to the same RecordStore is returned.
- If the caller is also the owner of the record store, the call is same as openRecordStore(recordStoreName, false).

Before the application quits, it should close the RecordStore to avoid data corruption. This is accomplished by a call to

```
recordStore.closeRecordStore( );
```

15.8.3 To Delete a RecordStore

We use `recordStore.deleteRecordStore(`*recordStoreName*`)`;

The code snippet implementing the method above is given below. Note the function throws RecordStoreNotOpenException exception.

```
public void deleteRecordStore( )
   {
      try
      {
         if (recordStore.getNumRecords( ) == 0)
         {
            String fileName = recordStore.getName( );
            recordStore.closeRecordStore( );
            recordStore.deleteRecordStore(fileName);
         }
         else
         {
            recordStore.closeRecordStore( );
         }
   }
catch(RecordStoreNotOpenException rsNopEx){ rsNopEx.printStack
Trace( );}
catch(RecordStoreException rsEx){ rsEx.printStackTrace( );}
```

Now we will see how to add update and delete records to the RecordStore.

15.8.4 To Add Records

```
recordStore.addRecord(b, 0, b.length); // where b is the record data
in a byte array.
```

public synchronized void addNewRecord(String record)

```
{
   byte[] b = record.getBytes( );
   try
   {
      recordStore.addRecord(b, 0, b.length);
   }
   catch (RecordStoreException rse)
   {
      System.out.println(rse);
      rse.printStackTrace( );
   }
}
```

The corresponding call to this function would look like

```
System.out.println("Adding records to the store");
testRMS.addNewRecord("Pakistan";);
testRMS.addNewRecord("India";);
testRMS.addNewRecord("China";);
```

Since our RMS implementation has registered a record listener that prints the event triggered to the console we should see the following

MSG From RecordListener: Record added Store Name " Countries" ID 1 Value Pakistan
MSG From RecordListener: Record added Store Name " Countries" ID 2 Value India
MSG From RecordListener: Record added Store Name " Countries" ID 3 Value China

15.8.5 To Update Record

```
"recordStore.setRecord(id,byteInp,0,byteInp.length);"
```

Where

- id is the RecordID of the record to be updated.
- byteInp is the byte array for the input.
- *startat* is the offset in the byte array from where the data starts.
- *byteInp.length* is the length of the input.

```
public synchronized void UpdateRecord(int id, String record)
   {
      byte[] byteInp = record.getBytes( );
      try
      {
         recordStore.setRecord(id,byteInp,0,byteInp.length);
      } catch (RecordStoreException e)
      {
         e.printStackTrace( );
      }
   }
```

Note: There is no way to modify part of the data in a record. It can only be replaced with a new byte array.

A call to the update method above is given below.

```
System.out.println("Now let us update a record");
testRMS.UpdateRecord(1,"Bangaladesh");
```

Our RecordListner now tells us

MSG From RecordListener: Record changed Store Name " Countries" ID 1 Value Bangladesh

To read records in the RecordStore

```
resp = recordStore.getRecord(Id)
```

where *Id* is the unique RecordID of the record and response is the data in the record returned as a byte array. We have written a small utility called dumpRecordStore that reads all the records in the store and prints the contents to the console. The method is listed below.

```
public void dumpRecordStore(RMS rms )
   {
      System.out.println("******** This is a record store Dump
      *******");
```

```
            System.out.println( rms.getRecord(1));
            System.out.println( rms.getRecord(2));
            System.out.println( rms.getRecord(3));
        }
```

The output to this would be
******** This is a record store Dump *******
Bangladesh
India
China

As is very apparent, this is bad programming. A read utility should be able to traverse the entire length of the store and print all the contents. This is where enumeration comes into the picture.

15.8.6 To Delete a Record

```
"recordStore.deleteRecord(id);"
```

where *Id* is the unique RecordID of the record and response is the data in the record returned as a byte array. It is interesting to note that there is no method to obtain the RecordID of a record. Developers will have to device their own work around.

```
public boolean deleteRecord(int id)
    {
        try
    {
        recordStore.deleteRecord(id);
    }
    catch (RecordStoreException e)
    {
        e.printStackTrace( );
        return false;
    }
        return true;
    }
```

To call this method we use the following listing

```
System.out.println("Trying to delete a record");
if(testRMS.deleteRecord(1))
    {
        System.out.println("Successfully deleted record 1");
        dumpRecordStore(testRMS);
    }
        else
            System.out.println("Couldnot delete record 1");
```

Output is
Trying to delete a record

MSG From RecordListener: Record deleted Store Name "Countries" ID 1 Value
Successfully deleted record 1

Note: Some points about the delete options that need to be remembered are:

- RecordID of the deleted record is not reused (so the last RecordID of the store does not indicate the number of records in the store).
- Even if the record is deleted, the space is not reused. i.e., the space is not freed for reuse by another record. This is possible only by deleting RecordStore. So if our application is deleting very frequently we might want to perform a clean-up during init/destroy by writing a new and current store. Of course our RecordID will also be updated.

To work around the missing getRecordID we need to use the other constructs RecordEnumerator, RecordComparator and RecordFilter.

We can test this by calling a delete record for an ID that has already been deleted

```
System.out.println('Now lets try to delete the same record
again');
   if(testRMS.deleteRecord(1))
   {
      System.out.println("Successfully deleted record 1");
      dumpRecordStore(testRMS);
   }
   else
      System.out.println("Couldnot delete record 1");
```

The output this time will show.
******** This is a record store Dump *******
Now let us try to delete the same record again
javax.microedition.rms.InvalidRecordIDException
Could not delete record 1

To see if the ID of the deleted record is reused, we can add a new record and loop through to see the ID associated with it. To get a list of all the records in a record store, we use an enumeration.

15.8.7 The Enumerator

The enumerateRecords method returns a RecordEnumeration. This allows a bidirectional movement through the set of records.

```
"recordStore.enumerateRecords(null, null, false)"
```

The first two are used to filter and sort the records in an enumeration (we will discuss filtering and sorting shortly). The last argument if set to true, causes the enumeration to update itself as the record store is modified. It is, however, important to remember that it is an expensive operation and should be used judiciously.

To traverse the resulting enumerator we can use the following:

```
while( enum.hasNextElement( ))
   {
      int id = enum.nextRecordId( );
      // Do something here whit the ID
      System.out.println(testRMS.getRecord(id));
   }
The output of this loop is
```

Now we will loop through the store and print them out
China
India

Another method defined by `RecordEnumeration is getNextRecord()` to obtain the record in bytearray format. To traverse backwards we use `getPrevious RecordId()` and `getPreviousRecord()`. An application can use hasNextElement( ) or hasPreviousElement( ) to break the loop. An enumeration can be reset to the beginning by calling `reset`.

Note: It is a good idea to `destroy` the enumeration after use. This frees up any resources used by the enumeration.

Passing null as first two arguments causes all the records in the record store to be returned in an undefined order. But it is not prudent to extract all the records for every operation. RMS provides two additional methods to get a subset of the records or to return the records in a specific order. They are filter and comparator. Developers can use one or both depending on their requirement.

15.8.8 Filter

We will first see a filter. For example, to create a filter for a string that has length > 0 and begins with the alphabet 'I', the filter defined will resemble

```
public class MyFilter implements RecordFilter
   {
      public boolean matches( byte[] recordData )
      {
      return( recordData.length > 0 && recordData[0] == 'I' );
      }
   }
```

Code listing below shows the use of this filter

```
System.out.println('Let's try a filter Filtering for records begining
with I');
   try
      {
         enum = testRMS.enumerate(new MyFilter( ));
         while( enum.hasNextElement( ))
      {
         int id = enum.nextRecordId( );
         System.out.println(testRMS.getRecord(id));
      }
   }
```

```
        catch( RecordStoreException e )
        {
           e.printStackTrace( );
        }
        finally
        {
           enum.destroy( );
        }
```

The corresponding output is
Let's try a filter Filtering for records begining with 'I'
India

What happens when we have multiple entries that match the filter? Let us try and see. So here is the repeat code.

```
System.out.println("Will it work for multiple record begining
with 'I'");
System.out.println("Add IceLand");
testRMS.addNewRecord("IceLand");
System.out.println("Filtering again for records begining
with 'I'");
try
   {
      enum = testRMS.enumerate(new MyFilter( ));
      while(enum.hasNextElement( ))
      {
         int id = enum.nextRecordId( );
         System.out.println(testRMS.getRecord(id));
      }
   }
   catch( RecordStoreException e )
   {
      e.printStackTrace( );
   }
      finally
   {
      enum.destroy( );
   }
```

The output now is as below.
Will it work for multiple records begining with 'I'
Add IceLand

MSG From RecordListener: Record added Store Name 'Countries' ID 4 Value IceLand.
Filtering again for records begining with 'I'
India
IceLand

As we can see the filter class has to implement the RecordFilter interface. This interface has one method *match* that performs the matching operation on the input bytearray. The filter's matches method is called for each record in the record store. If the matches method returns true, the record is included in the enumeration. Use the filter like this:

recordStore.enumerateRecords(fltr, null, false);
Where fltr is an instance of Myfilter defined above.
Note, that the filter only gets the contents of a record.
To compare my Record with other records in the RecordStore or to search for a particular record the application must implement the Comparator.

15.8.9 Comparator

The Comparator interface is used to compare two records. The return value indicates the ordering of the two records. RecordComparator defines the following constants

- FOLLOWS: Value = 1 parameter on LHS follows the right parameter in terms of search or sort order.
- PRECEDES: Value = –1 parameter on LHS precedes the right parameter in terms of search or sort order.
- EQUIVALENT: Value = 0 Both parameters are the same.

The RecordComparator interface again defines a single method: compare. We will now define a comparator that will, when used with an enumerator, return a list of strings in lexical sorted order.

```
public class MyComparator implements RecordComparator
   {
      public int compare( byte[] rec1, byte[] rec2 )
         {
            String s1 = new String(rec1);
            String s2 = new String(rec2);
            int cmp = s1.compareTo(s2);
            System.out.println( s1 + " compares to " + s2 +"
            gives " + cmp );
            if( cmp != 0 ) return ( cmp < 0 ? PRECEDES : FOLLOWS );
         cmp = s2.compareTo(s2);
         if( cmp != 0 ) return ( cmp < 0 ? PRECEDES : FOLLOWS );
         return EQUIVALENT;
      }
```

The call is made as below:

```
"recordStore.enumerateRecords(null, cmptr, false);"
```

Where cmptr is an instance of MyComparator. Again, only the contents of the records to compare are passed to the comparator.
The following snippet shows the use of the comparator above.

```
System.out.println('How does Does a comparator work?');
   testRMS.addNewRecord("Bangaladesh");
   try
```

```
{
    enum = testRMS.enumerate(new MyComparator( ));
    System.out.println('Now what does the resulting enumarator
    look like?')
    while( enum.hasNextElement( ))
    {
        //int id = enum.nextRecordId( );
        System.out.println( new String(enum.nextRecord( )));
    }
}
catch( RecordStoreException e )
    {
        e.printStackTrace( );
    }
    finally
    {
        enum.destroy( );
    }
```

The output is

How does a comparator work?
Bangladesh compared to China gives –1
China compared to China gives 0
Iceland compared to China gives 6
India compared to China gives 6
China compared to China gives 0
India compared to India gives 0
Iceland compared to India gives –11

Now what does the resulting enumerator look like?:
Bangladesh
China
Iceland
India

Note: An enumeration can take both a filter and a comparator. The filter is called first to determine which records are in the enumeration. The comparator is then called to sort those records. If tracking is set to true, filtering and sorting occurs whenever the record store is modified. To refresh data selectively we have a `keepUpdated` method. This is used to enable or disable an enumeration's tracking of the underlying record store. We can keep tracking disabled but "refresh" the enumeration from the record store, by calling the Enumeration `rebuild` method.

15.8.10 The RecordListener

The RecordListener interface defines event handlers for record change events from a recordstore associated with the application. An application overrides the following methods to trap and use these events.

- `recordAdded( )`: When a new record is added to the record store.
- `recordDeleted( )`: When a record in the record store is deleted.
- `recordChanged( )`: When a record in the store has been updated.

Any application that wants to listen to the RecordListener has to implement the RecordListener interface and override the three functions above. The next step is to associate listener with the object. The sample implementation just shows the methods and a simple print statement for the function activated. We have already seen the output from these methods above.

```
public void recordAdded(RecordStore recordStore,int recordId)
   {
      try
      {
            System.out.println("MSG From RecordListener: Record
            added " + "Store Name \" " + recordStore.getName( ) + "\"
            ID " + recordId + " Value " + getRecord(recordId) );
         }  catch (RecordStoreNotOpenException e)
         {
            e.printStackTrace( );
      }
   }
public void recordChanged(RecordStore recordStore, int
recordId)
   {
      try
         {
            System.out.println("MSG From RecordListener: Record
            changed " + "Store Name \" "+ recordStore.getName( ) +
            "\" ID " +recordId + " Value " + getRecord
            (recordId));
         } catch (RecordStoreNotOpenException e)
         {
            e.printStackTrace( );
         }
   }
public void recordDeleted(RecordStore recordStore,int
recordId)
   {
      try
      {
         System.out.println("MSG From RecordListener: Record
         deleted " + "Store Name \" " + recordStore.getName( )
         + "\" ID " + recordId + " Value " + getRecord
         (recordId));
      }  catch (RecordStoreNotOpenException e)
      {
```

```
                e.printStackTrace( );
            }
        }
```

With this we will conclude our discussions of RMS. Before we continue, however, we will take a quick peek at good programming practices.

> ***A warning about using RMS with threads****: No locking operations are provided in this API. Record store implementations ensure that all individual record store operations are atomic, synchronous, and serialized, so no corruption will occur with multiple accesses. However, if a MIDlet uses multiple threads to access a record store, it is the MIDlet's responsibility to coordinate this access or unintended consequences may result. Similarly, if a platform performs transparent synchronization of a record store, it is the platform's responsibility to enforce exclusive access to the record store between the MIDlet and synchronization engine.*
>
> ***JavaDocs API MIDP2.0***

15.8.11 Good Programming Practices

Our mantra is:

- Don't create an object unless necessary.
- REUSE as far as possible.
- Every byte saved is a byte gained.

Specific to RMS:

- Read and write operations should be optimized such that they are minimal.
- Writes are expensive and hence should be kept to minimal.
- Know our data well and use the storage optimally.
- Keep functionality in mind, flush and recreate the RecordStore periodically.
- Handle exceptions properly so that users do not end up with frustrating blank or frozen screens.

An excellent article titled *Record Management System Basics* by Eric Giguere discusses optimization techniques for RMS in greater detail. The article can be found at "http://developers.sun.com/techtopics/mobility/midp/ttips/rmsefficient/"

> WTK20 provides a monitoring facility. Among other options is the possibility of observing the runtime usage of memory. Also important is the VM emulation feature which allows the developer to limit the resources available to the application and simulate a device like environment. This is very important to ensure proper performance for applications.

15.9 Communication in MIDP

One of the unique selling points of J2me was its claim of supporting multiple modes of connections while at the same time achieving a small footprint. Since its inception J2Me has mandated that

HTTP must be supported on all J2ME handsets. There are other protocols too that are supported and some new introductions have been made to the MIDP 2.0. All connection APIs come under the broad scope of a Generic Connection Framework (GCF) provided by CLDC. The GCF provides a single consistent set of abstractions for the developer. Figure 15.33 provides a look at the structure of the GFC. At the top we have the connection or the base interface as represented by the connector class. (A call to static method defines creation of all types of connections.) If a call to open is successful it returns the object that implements the desired protocol. This is in essence a factory class, however, we will not get into discussions on design here.

In Figure 15.33 all blocks numbered 1 represent CLDC 1.0 interfaces. It is important to remember that the CLDC only defines the classes and interfaces that make up the GCF. It is the profiles built on top of the CLDC which define the specific protocols an implementation must support. The HttpConnection interface (labeled 2) was added by MIDP 1.0, while SeverSocketConnection UDPDatagramConnection and Socket Connection (numbered 3) were added by MIDP2.0.

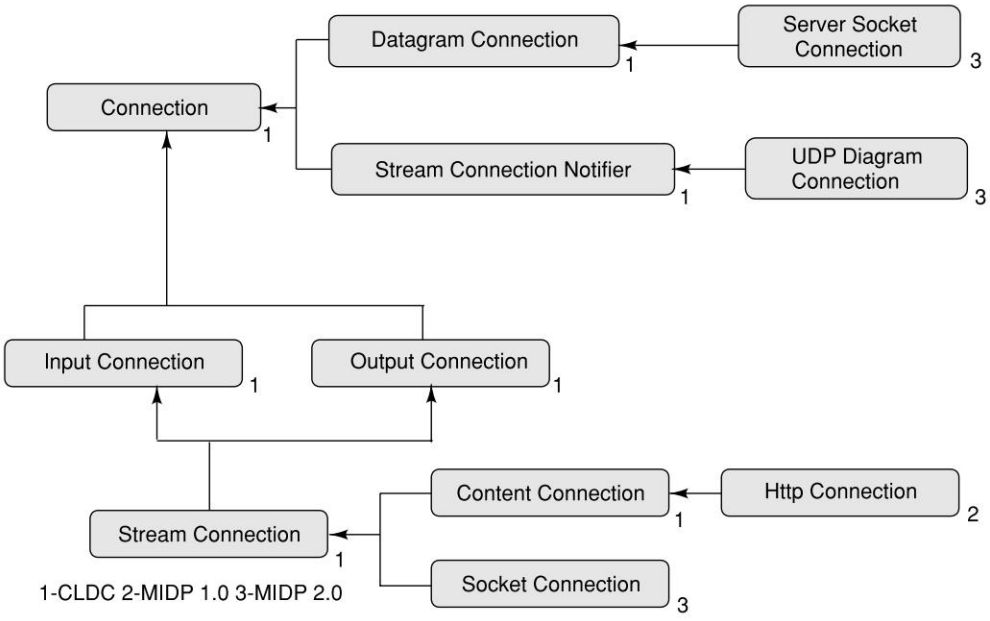


Figure 15.33 Generic Connection Framework

A call to the connection interface would look like

```
ConnectionType conn = (ConnectionType) Connector.open( URI );
```

Where the URl is typically of the form : <protocol> :< address> :< parameters>:

- `protocol` specifies the protocol to be used it is simply string-like while `'http'`, `'mailto'`, `'ftp'`, etc. are the names used to identify specific protocols. As mentioned above, MIDP mandates the implementation of some protocols, for example, the HTTP protocol (MIDP

1.0). While OEM implementations can add support for other protocols these should fit into the framework.
- `address` specifies the destination of the connection.
- `parameters are` connection parameters as required by the specified protocol.

The Connector class uses the protocol parameter to:
- Find and load the appropriate connection class.
- Make the connection.
- Return the newly created connection object.

This design ensures that the application code need not be changed even if the underlying protocol implementation changes. We will look at some details below.

Sample URI and the resulting Connection calls are given below.

CLDC interfaces;
- url = "file:e:/myj2me/readme.txt";

InputConnection conn = (InputConnection) Connector.open(url, Connector.READ);
- url = "file:e:/myj2me/readme.txt append=true";

OutputConnection conn = (OutputConnection) Connector.open(url, Connector. WRITE);
- url = "socket://www.mysite.com:80";

StreamConnection conn = (StreamConnection) Connector.open(url);
- url = "http://www.mysite.com/index.html";

ContentConnection conn = (ContentConnection) Connector.open(url);

Further examples will however concentrate only on MIDP implementations, namely HTTP, Socket, ServerSocket and Datagram protocols. A discussion about the various protocols though helpful is beyond the scope of this text. Links for articles dealing with details are provided at the end of this chapter.

MIDP 2.0 supports four protocols: HTTP, Socket, ServerSocket and Datagram.

1. HTTP Connection
   url = http://www.yahoo.com ;
   HttpConnection conn = (HttpConnection) Connector.open(url);
2. url = 'datagram://mysite.com:79' ;
   DatagramConnection conn = Connector.open(url)
3. url = "socket://mysite.com:port";
   SocketConnection client = (SocketConnection) Connector.open('socket://' +
   hostname + : + port);
4. url= socket://:2500;
   ServerSocketConnection server = (ServerSocketConnection) Connector.open(url);

The open method throws a ConnectionNotFoundException if the requested protocol handler is unavailable. It is interesting to note that while MIDP 2.0 mandates support for HTTP 1.1 as well as HTTPs connections, support for low-level IP networking is optional which means that the decision is finally in the hands of the OEMs. However, for the applications to be portable it is advisable to stick to http(s) and do the heavyweight stuff on the server end. A chat or multi user game using a

server on one of the handsets is fun, but GSM's data bandwidth makes the response time so poor that users soon lose interest. The decision regarding the constructs to be used, vary depending on the application's requirements and should be made judiciously.

From an application perspective, a communication mechanism conceptually has three parts to it: Open a channel, Read/Write in the channel and close the channel. We have already studied the first. We will now study the remaining two. As discussed HTTP inherits from both the stream and content connection while socket connections are inherited from stream connection.

Note: Though we say MIDP supports the HTTP protocol, its implementation is not restricted to IP protocols. Non-IP protocols such as WAP and i-Mode are also supported. The non-TCP/IP mechanisms usually involve an edge gateway that does the translation between the client and the HTTP server. The control flow is as depicted in Figure 15.34. In such cases security lies in the domain of the network provider and hence needs to be accounted for.

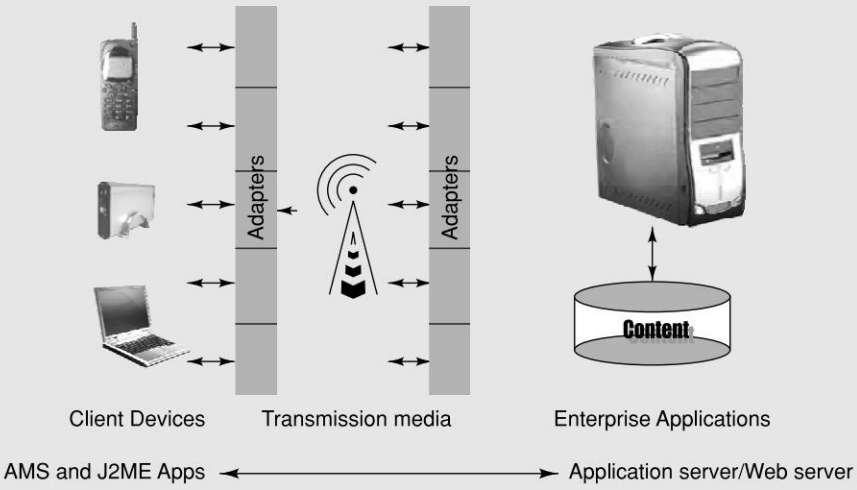


Figure 15.34

15.9.1 HTTP Connections

HTTP connections leverage on the protocol implementation to provide information regarding the headers. This allows us to set the user-agent, the user language, the version of MIDP and CLDC supported, and also the application specific variables. This feature is extremely useful when applications use server processing capabilities of a server that support various formats including web wap, J2ME etc. For this purpose MIDP provides set RequestProperty and getRequestProperty *methods*. An example might look like:

User-Agent: Profile/MIDP-2.0 Configuration/CLDC-1.0 Accept-Language: en-US

All connections are associated with an underlying InputStream OutputStream. To read, we open an input stream using the connection we obtained earlier and then continue to read and store

the contents into a buffer till we receive a –1 that indicates a EOS. The loop would look something like below.

```
InputStream is = conn.openInputStream( );
// Get the length and process the data
//Read the data.
while ((ch = s.read( )) != -1)
   {
      //Store into a buffer that can be used later.
   }
//A better way of doing it would be read chunks of data
instead of just one character something like this.
int len = (int)conn.getLength( );
byte[] data = new byte[len];
//Read the entire content of the buffer.
dis.readFully(data);
//Sometimes depending on their implementation the webservers
do not return the complete buffer in one go but
return multiple chunks of fixed lengths or less as may be
available. In such cases we need to read chunks of data. The
logic then looks something like
int datalen = (int)conn.getLength( );
int read = 0;
int totalread = 0;
byte[] data = new byte[datalen];
while ((totalread != len) && (read != -1))
{
   read = is.read(data, totalread, datalen - totalread);
   totalread += read;
}
```

To write to the channel we use the output stream. The sample implementation might look as below.

//Open an output connection

```
OutputStream os = conn.openOutputStream( );
String content = "My first connection";
//Write to the connection
os.write(content.getBytes( ));
//flush it so that it is sent out.
os.flush( );
```

Finally we use conn.close( ) to close the channel.

```
HTTP supports some constants as defined in the protocol. These include
```

- `HTTP_OK to indicate a 200 or correct response,`
- HTTP_NOT_FOUND indicates a 404 condition of server not found etc.

The Java docs accompanying the WTK give an extensive explanation and details of the API available.

The only difference between HTTP and HTTPs from the application perspective is the implementation class and the url which now becomes

url = https://mysite.com
HttpsConnection conn = (HttpsConnection)Connector.open(url);

The implementation details are abstracted from us.

15.9.2 The SocketConnection Interface

The SocketConnection interface as shown above returns a socket connection and is normally used to access TCP/IP servers. The difference between a server socket and an inbound client is the host name. If a host name is specified, we get a client; else we get a server. Also if the port number is omitted then by default an available port number is assigned to the socket. This can then be retrieved by a call to `getLocalPort( )` and `getLocalAddress( )`. The following snippet shows a simple example.

```
SocketConnection client = (SocketConnection) Connector.open
("socket://" + myhost + ":" + 7001);
   InputStream is = client.openInputStream( );
   OutputStream os = client.openOutputStream( );
   // send a request to server
   os.write(content.getBytes( ));
   // read server response
   int c = 0;
   while((c = is.read( )) != -1)
   {
   // store for later use.
   }
   // close all
   is.close( );
   os.close( );
   client.close( );
```

We can set user defined options on the socket using Call setSocketOption. SocketConnection provide constants like DELAY, KEEPALIVE, etc.

15.9.3 The ServerSocketConnection Interface

The `ServerSocketConnection` interface defines the server socket stream connection. It is used when we need to establish a server on the device. Multiuser game applications are ideal candidates here.

```
// create a server to listen on port 7001
ServerSocketConnection server = (ServerSocketConnection)
Connector.open ("socket://:7001");
```

```
// wait for a connection
SocketConnection client = (SocketConnection) server.
acceptAndOpen( );
// open streams
DataInputStream dis = client.openDataInputStream( );
DataOutputStream dos = client.openDataOutputStream( );
int datalen =0;
// read client request refer to the WTK javadocs for other
related APIs
while(datalen)
{
datalen= dis.available( )
c = dis.read( );
if( c == -1) break;
}
// process request and send response
os.write(...);
// close streams and connections
is.close( );
os.close( );
client.close( );
server.close( );
```

Similar to HTTPs, we also have a secure socket connection which is the same as the socket connection except in the class to be used and the URL.

```
urI = ssl://host.com:79;
SecureConnection sc = (SecureConnection)Connector.open(urI)
```

Both secure protocols provide information regarding the security implementation in terms of the protocol, version, certificate information and the cipher suite used.

With this we conclude our discussion of the communication facilities provided by MIDP. The last topic is security concerns which are enumerated in the section on security.

Till MIDP 2.0 came along, applications were all pull based, i.e., the application had to be running and connected to receive any input from the external world. No messages or network activity could be pushed to the application, essentially forcing applications to function only in a synchronous or polling mode. This was a major restriction on network aware applications. MIDP 2.0 rectified this by introducing the push registry. Our next topic deals with the details of this.

15.9.4 Push Registry

The push registry is actually an addition to the AMS. As we have seen above, the AMS is responsible for managing the application lifecycle. Given the Sandbox Model followed in MIPD, it would not have been possible to implement it elsewhere. This module allows applications to register for incoming connections or timer-bound requests. The registry itself is a list of all-timer events and inbound connections. To utilize this facility the applications need to register to the AMS. It can be

done in two ways: one, statically by defining the connections they need to access in the app descriptor. While installing the AMS, check if it supports all the mentioned events; if not, the user is informed and the application will not install. Two, dynamically using registerConnection( ) API of the Class javax.microedition.io. PushRegistry. Once registered, it is the responsibility of the AMS to launch the application in response to the event that it has registered for. This is made available through Serial Port Communications, a CommConnection newly available in the MIDP 2.0.

Static registrations are defined by listing one or more MIDlet-Push attributes in the JAD file or JAR manifest. When the MIDlet suite is uninstalled, the AMS automatically unregisters all its associated push registrations. This entry is the

```
MIDlet-MyPushReg-1: socket://:7001, PushMIDlet, * in the sample jad
file above.
// MIDlet-Permissions: javax.microedition.io.PushRegistry,
// javax.microedition.io.Connector.serversocket
```

The format of the MIDlet-Push attribute is:
MIDlet-Push-<n>: <ConnectionURL>, <MIDletClassName>, <AllowedSender>

- **MIDlet-Push-*<n>*** a unique name number combination to identify the registration e.g., MIDlet-MyPushReg-1
- ***<ConnectionURL>*** string that identifies the url. It is the same URL that we have defined in the section on connections above.
- ***<MIDletClassName>*** fully qualified class name of the MIDlet that is requesting the registration.
- ***<Allowed-Sender>*** is a filter that defines the origin server of this registration. This is in fact a security mechanism to regulate external access to the application.

Note: The MIDlet-Permissions property is used to request permissions for the MIDlet suite. This example requests permission to use the push registry and socket APIs. The Security Considerations section will cover permissions in more detail.

In the dynamic mode the application makes a call to registerConnection( ). A sample code snippet would look like:

```
// Open the connection.
   ServerSocketConnection conn = (ServerSocketConnection)
   Connector.open(urI);
// Register the connection to the AMS
   PushRegistry.registerConnection(url, MIDletClassName,
   filter); //filter is a string variable that can include
   wild characters.
// The rest of the connection code is as before
// wait for a connection
   SocketConnection client = (SocketConnection) server.
   acceptAndOpen( );
```

Push implementation is the combined responsibility of the MIDlet and the AMS. Post registration we have two scenarios.

1. When the MIDlet is *not* active, the AMS monitors registered push events on behalf of the MIDlet. When a push event occurs, the AMS activates the appropriate MIDlet to handle it. Figure 15.35 illustrates this sequence for a network activation.
2. If the MIDlet is active (running), the MIDlet itself is responsible for all push events. It must set up and monitor inbound connections, and schedule any cyclic tasks it needs—basically standard networking and timer processing.

The push registry can also be used for timer-based alerts or events. With this we conclude our discussion of Push. The Push facility gives rise to many exciting applications. For those interested it is an excellent discussion on using the Push registry for various inbound and outbound connections including socket, server socket, datagram and HTTP connections while extending this idea further. Working of the WMA (Wireless Messaging API) in incorporating the SMS feature of GSM networks is available at the Sun's site under the section Technical Articles.

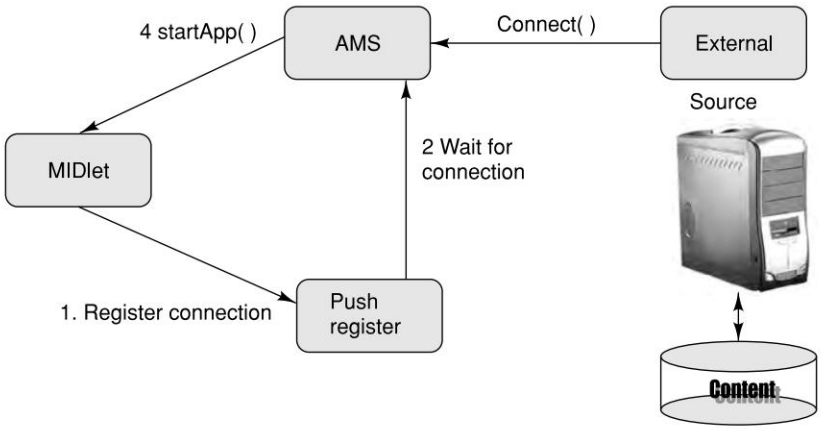


Figure 15.35 Application Activation Sequence through Push Registry

15.10 Security Considerations in MIDP

Security considerations are of different types namely, application security, data security and communication security. MIDP 1.0 implemented a strict sandbox for all applications. Thus APIs prevented access to sensitive device capability assuring application security. Only MIDlets within the same MIDlet suite could access record stores belonging to the suite, i.e., created by the MIDlets in the suite. These restrictions clubbed with the absence of any push facility ensured a very safe environment for MIDlets. MIDP 2.0 expanded the scope of MIDlets by having possibility of inter suite access to record stores and push facility. Therein lay the need for security. Following is a discussion on the security features built into MIDP 2.0 to protect the users from rogue applications while at the same time, allowing access to sensitive data.

15.10.1 Application Security

The MIDP 2.0 security model introduces the concept of security domain. A security domain defines the access permissions for an application.

Applications belong to either

- The trusted domain, i.e., permitted to use APIs that are considered sensitive.
- The un-trusted domain, i.e., having restricted access.

A MIDlet suite that is not trusted will be run in untrusted mode. Each domain is associated with a domain policy that defines the requirements for a MIDlet to be considered as trusted. A domain policy defines a set of permissions "allowed" and "user" that can be granted to the MIDlet and

- "allowed" grants the MIDlet permission to access and use the requested API.
- "user" means the MIDlet has to obtain explicit user permission through UI.

For example, if we attempt an HTTP connection the AMS throws back a screen as shown in Figure 15.36.

Another way for the MIDlet to request permissions is through its MIDlet-Permissions and MIDlet-Permissions-Opt attributes. Permissions are verified before granting access to the restricted APIs. For maximum security it is advisable to use signed MIDlet suites. Essentially this process uses the X.509 PKI security from a trusted source. The provisioning systems play a very important role in this. This ensures that only authenticated MIDlets are registered as members of the trusted domains. For further details please refer to the documentation accompanying the WTK.

Figure 15.36 Security Domain Permissions in MIDP 2.0

15.10.2 Recordstore Security

MIDP 2.0 allows for shared record stores. Sharing is defined by the MIDlet suite at the time of creation. Access controls are enforced when RecordStores are opened. We have already seen the access modes in the section on RMS. Briefly then

```
openRecordStore(recordStoreName,
   createIfNecessary,
   AUTHMODE_ANY,
   writable) throws Exception ……
Creates a RecordStore that allows other MIDlets access to this
RecordStore.
   Now openRecordStore(recordStoreName,
   vendorName,
   suiteName) throws Exception, will open the store created above.
```

But these must be used very carefully as this exposes all the application data to the caller and no further verification will be done.

15.10.3 Communication Channel Security

Secured networking was introduced in MIDP 2.0 with the introduction of HTTPs (secure HTTP). Another level of protection is the secure domains discussed above. All protocols have to initiate a request through a call to javax.microedition.io.Connector .open(…). The permissions are granted individually to protocols. It is not mandatory for all devices to implement all protocols, if implemented, the security framework specifies the naming of permissions based on the package and class name.

15.10.4 Security of PushRegistry

The PushRegistry also uses the security framework and permissions. Only MIDlet suites having the javax.microedition.io.PushRegistry permission can register for a Push based launch. A detailed discussion on the protection domains and related development is beyond the scope of this book.

15.11 Optional Packages

To enhance J2ME capability we have optional packages. An optional package is a set of APIs to support additional features that don't really belong in one specific configuration or profile. Bluetooth support, for example, is defined as an optional package.

Optional packages prescribe their own minimum requirements and are dependent on a particular configuration and one or more profiles There are many optional packages in development, including:

- JDBC Optional Package which is a subset of JDBC (database access APIs).
- Wireless Messaging API (WMA), for sending and receiving Short Message Service (SMS) messages.
- Mobile Media API (MMAPI), for the capture and playback of multimedia content.
- RMI Optional Package, for remote method invocation.

- APIs, for Bluetooth to integrate into a Bluetooth environment.
- Mobile Game API, for game development.
- Location API, for build location-aware applications.
- Mobile 3D Graphics API, for interactive 3D manipulations.
- Event Tracking API, for tracking of application events and the submission of these event records (through an event-tracking server).
- Advanced Graphics and User Interface to migrate the core Swing, Java 2D Graphics and Imaging, Image I/O, and Input Method Framework for advanced graphics and user interface facilities.
- Content Handler API to handle multi-media and web content.

This is by no means either a complete or exhaustive list. For more details readers can access "http://www.jcp.org/en/jsr/tech".

15.12 MOBILE RELATED JSR

As technology advances, the Java community is making sure that Java remains up-to-date and supports these advancements. This is done through Java Specification Request (JSR). JSR extends the Java framework to support both current practices and advanced applications development. Being the offspring of Java Community Process (JSP), JSRs mainly cater to areas like J2EE (Java2 Enterprise Edition), J2SE (Java2 Standard Edition), OSS, XML, JAIN and J2ME (Java2 Micro Edition). There are more than 180 released JSRs and many are still in the development phase.

There are 82 JSRs in Java that can be directly related to J2ME. These JSRs address mobile computing and wireless technologies. The list below summarizes the Java Micro Edition JSRs. The names and descriptions appearing in the list is the official description of JSRs in Java Micro Edition arena.

JSR #	*Name of JSR*	*Description*
1	Real-time Specification for Java	The Real-time Specification for Java extends the Java platform to support both current practice and advanced real-time systems application programming.
30	J2ME Connected, Limited Device Configuration	This specification will define a standard platform configuration of the Java2 platform, Micro Edition (J2ME) for small, resource-limited, connected devices.
36	Connected Device Configuration	The Connected Device Configuration (CDC) provides the basis of the Java2 platform, Micro Edition for devices that have a sufficient 32-bit microprocessor and ample memory.

(*Contd*)

JSR #	*Name of JSR*	*Description*
37	Mobile Information Device Profile for the J2ME Platform	This specification will define a profile that will extend and enhance the "J2ME Connected, Limited Device Configuration" (JSR-000030), enabling application development for mobile information appliances and voice communication devices.
46	Foundation Profile	The Foundation Profile is a set of APIs meant for applications running on small devices that have some type of network connection.
50	Distributed Real-time Specification	The Distributed Real-time Specification for Java extends RMI in the Real-time Specification for Java, to provide support for predictability of end-to-end timeliness of trans-node activities.
62	Personal Profile Specification	The J2ME Personal Profile provides the J2ME environment for those devices with a need for a high degree of Internet connectivity and web fidelity.
66	RMI Optional Package Specification Version 1.0	The J2ME RMI Optional Package (RMI OP) provides Java platform to Java platform remote method invocation for Java devices and interoperates with J2SE RMI.
68	J2ME Platform Specification	This specification will define the next major revision of the Java2 platform, Micro Edition.
75	PDA Optional Packages for the J2ME Platform	This JSR produces two separate optional packages for features commonly found on PDAs and other J2ME mobile devices: one for accessing PIM data and one for accessing file systems.
80	Java USB API	This specification provides a Java API for communicating with devices attached via the Universal Serial Bus (USB). It allows Java applications to discover, read, write, and manage USB devices.
82	Java APIs for Bluetooth	Bluetooth is an important emerging standard for wireless integration of small devices. The specification standardizes a set of Java APIs to allow Java-enabled devices to integrate into a Bluetooth environment.
113	Java Speech API 2.0	This JSR extends the work of the 1.0 Java Speech API, specifying a cross-platform interface to support speech recognizers and synthesizers.

(*Contd*)

JSR #	Name of JSR	Description
118	Mobile Information Device Profile 2.0	This specification will define a profile that will extend and enhance the "J2ME Mobile Information Device Profile" (JSR-000037).
120	Wireless Messaging API	This defines a set of optional APIs which provides standard access to wireless communication resources, designed to run on J2ME configurations and to enhance J2ME profiles with unique functionality.
129	Personal Basis Profile Specification	The J2ME Personal Basis Profile provides a J2ME application environment for network-connected devices supporting a basic level of graphical presentation.
133	JavaTM Memory Model and Thread Specification Revision	The proposed specification describes the semantics of threads, locks, volatile variables and data races. This includes what has been referred to as the Java memory model.
134	Java Game Profile	Defines a Java2 Micro Edition Profile for the purposes of game development targeting high-end consumer game devices and desktops.
135	Mobile Media API	This specifies a small-footprint multimedia API for J2ME, allowing simple, easy access and control of basic audio and multimedia resources while also addressing scalability and support of more sophisticated features.
138	Performance Metric Instrumentation	Specifies standard APIs for performance metric instrumentation of Java programs.
139	Connected Limited Device Configuration 1.1	This specification will define a revised version of the J2ME Connected, Limited Device Configuration (CLDC).
143	JavaDesk	JavaDesk provides a standard desktop API across platforms using an MVC model. Applications can control and enhance the desktop using the JavaDesk API.
164	SIMPLE Presence	SIMPLE Presence provides a standard portable and secure interface to manipulate presence information between a SIMPLE client (watcher) and a presence server (presence agent).

(*Contd*)

JSR #	Name of JSR	Description
165	SIMPLE Instant Messaging	SIMPLE Instant Messaging provides a standard portable and secure interface to exchange messages between SIMPLE clients. SIMPLE is an extension of SIP to support presence and instant messaging.
169	JDBC Optional Package for CDC/Foundation Profile	The proposed specification will define a JDBC Optional Package for Java2 Micro Edition (J2ME), Connected Device Configuration (CDC) Foundation Profile.
172	J2ME Web Services Specification	The purpose of this specification is to define an optional package that provides standard access from J2ME to web services.
177	Security and Trust Services API for J2ME	This specification will provide J2ME applications with APIs for security and trust services through the integration of a Security Element.
178	Mobile Game API	Defines an optional package that will facilitate the emergence of the market for the development of compelling games on mobile phones. The API shall work with MIDP1.0.
179	Location API for J2ME	An Optional Package that enables developers to write mobile location-based applications for resource-limited devices. The API works on the J2ME CLDC v1.1 and CDC configurations.
180	SIP API for J2ME	SIP API for J2ME defines a multipurpose SIP API for J2ME clients. It enables SIP applications to be executed in memory limited terminals, especially targeting to mobile phones.
184	Mobile 3D Graphics API for J2ME	This proposed JSR will provide a scalable, small-footprint, interactive 3D API for use on mobile devices.
185	Java Technology for the Wireless Industry	This JSR will provide an overall architectural description as well as an integrated TCK and RI to coordinate selected JCP efforts for the wireless industry.
186	Presence	Presence is a generic and protocol-agnostic API for Presence, providing a standard portable and secure interface to control, manage and manipulate Presence information between Presence clients and servers.
187	Instant Messaging	A protocol-agnostic API for Instant Messaging, this provides a standard portable and secure interface to control, manage and manipulate instant messages between clients through the use of presence servers.

(*Contd*)

JSR #	Name of JSR	Description
190	Event Tracking API for J2ME	This defines an optional code package that standardizes application event tracking on a mobile device and the submission of these event records to an event-tracking server via a standard protocol.
195	Information Module Profile	This JSR will define J2ME profile targeting embedded networked devices that wish to support a Java runtime environment, but do not have graphical display capabilities.
201	Extending the Java Programming Language with Enumerations, Autoboxing, Enhanced for loops and Static Import	This JSR proposes four new Java programming language features: enumerations, autoboxing, enhanced for loops and static import.
205	Wireless Messaging API 2.0	This JSR will extend and enhance the "Wireless Messaging API" (JSR-000120).
209	Advanced Graphics and User Interface Optional Package for the J2ME Platform	The Advanced Graphics and User Interface (AGUI) Optional Package will migrate the core APIs for advanced graphics and user interface facilities from the J2SE platform to the J2ME platform.
211	Content Handler API	Enabling J2ME applications to handle multi-media and web content can give developers and users a seamless and integrated user environment on mobile phones and wireless devices.
213	Micro WSCI Framework for J2ME	Effort to define another layer of the J2ME Web Service stack, implementing the "observable" behavior of a choreographed Web Service on the device, relative to the message exchange requiring support.
214	Micro BPSS for J2ME Devices	This JSR is to provide a standard set of APIs for J2ME Devices for representing and manipulating Collaboration Profile and Agreement information described by ebXML CPP/A (Collaboration Protocol Profile/Agreement) documents.
216	Personal Profile 1.1	This JSR will update the existing Personal Profile (JSR-62 specification to reflect the J2SE 1.4 APIs).
217	Personal Basis Profile 1.1	This JSR will update the existing Personal Basis Profile (JSR-129) specification to reflect the J2SE 1.4 APIs.

(*Contd*)

JSR #	Name of JSR	Description
218	Connected Device Configuration (CDC) 1.1	This JSR defines a revision to the J2ME CDC specification. This JSR provides updates (based on J2SE, v1.4) to the existing core, non-graphical Java APIs for small electronic devices.
219	Foundation Profile 1.1	This JSR defines a revision to the J2ME Foundation Profile. This JSR provides updates (based on J2SE, v1.4) to the existing core, non-graphical Java APIs for small electronic devices.
226	Scalable 2D Vector Graphics API for J2METM	This specification will define an optional package API for rendering scalable 2D vector graphics, including image files in W3C Scalable Vector Graphics (SVG) format.
228	Information Module Profile-Next Generation (IMP-NG)	This specification will define a profile that will extend and enhance the "J2ME Information Module Profile" (JSR-195).
229	Payment API	Enabling application developers to initiate mobile payment transactions in J2ME applications.
230	Data Sync API	Enabling J2ME applications to access native data synchronization implementation.
232	Mobile Operational Management	Create a predictable management environment for mobile devices capable of installing, executing, profiling, updating, and removing Java and associated native components in the J2ME Connected Device Configuration.
234	Advanced Multimedia Supplements	This specification will define an optional package for advanced multimedia functionality which is targeted to run as an supplement in connection with MMAPI (JSR-135) in J2ME/CLDC environment.
238	Mobile Internationalization API	This JSR defines an API that provides culturally correct data formatting, sorting of text strings and application resource processing for J2ME MIDlets running in MIDP over CLDC.
239	JavaTM Binding for the OpenGL® ES API	Java bindings to the OpenGL ES (Embedded Subset) native 3D graphics.
242	Digital Set Top Box Profile-"On Ramp to OCAP"	The requested specification will define a J2ME profile based on the Connected Limited Device Configuration (CLDC) that is appropriate for use by small-footprint cable television set-top boxes.

(*Contd*)

JSR #	Name of JSR	Description
246	Device Management API	Enabling J2ME applications to access device management implementations.
248	Mobile Service Architecture	This JSR creates a mobile service architecture and platform definition for the high volume wireless handsets continuing the work started in JSR-185 and enhancing the definition with new technologies.
249	Mobile Service Architecture Advanced	This JSR creates a platform definition for the advanced mobile handsets and builds on the J2ME Foundation Profile specification and other JSRs as will be selected.
253	Mobile Telephony API (MTA)	This JSR creates a mobile telephony API and platform definition which utilizes common telephony features and is small and simple to suite high volume devices with limited resources.
256	Mobile Sensor API	The API provides general Sensor API that extends the usability and choice of sensors for J2ME applications. It defines generic sensor functionality optimized for the resource-constrained devices like mobile devices.
257	Contactless Communication API	This specification will define J2ME Optional Packages for contactless communication, one package for bi-directional communication and the other for accessing read-only information.
258	Mobile User Interface Customization API	The Mobile User Interface Customization API provides a way to query and modify the user interface customization properties of a mobile device or platform.
259	Ad Hoc Networking API	The purpose of this JSR is to define an API that enables communication between mobile devices in a peer-to-peer ad-hoc network environment.
266	Unified Message Box Access API (UMBA-API)	The purpose of this JSR is to define an API to access and manage the message boxes of the mobile device and their content.
271	Mobile Information Device Profile 3	This JSR will specify the 3rd generation Mobile Information Device Profile, expanding upon the functionality in all areas as well as improving interoperability across devices.

(*Contd*)

JSR #	Name of JSR	Description
272	Mobile Broadcast Service API for Handheld Terminals	This specification will define an optional package in J2ME/MIDP/CLDC environment to provide functionality to handle broadcast content, for example, to view digital television and to utilize its rich features and services.
278	Resource Management API for Java ME	RM API will provide a simple interface for resource reclamation, accounting, and monitoring in a Java ME platform that requires resource management for multiple applications.
279	Service Connection API for Java ME	A new high-level API for connection services via frameworks supporting identity-based services, discovery, and authentication. The API supports Service Oriented Architectures (SOA) and other similar network service application models.
280	XML API for Java ME	This JSR provides a common general purpose XML API for the next generation of mobile devices.
281	IMS Services API	This JSR provides a high-level API to access IP Multimedia Subsystem (IMS) services. This API hides IMS technology details and exposes service-level support to enable easy development of IMS applications.
282	RTSJ version 1.1	Fill some minor gaps in the RTSJ.
287	Scalable 2D Vector Graphics API 2.0 for Java ME	This specification will define an optional package for rendering enhanced 2D vector graphics and rich media content based on select features from SVG Mobile 1.2, with primary emphasis on MIDP.
288	Adaptive Java ME System API	This specification will define a mechanism that enables a systems developer to include multiple configurations and profiles on a single device, using one set of developed components.
290	Java Language & XML User Interface Markup Integration	This JSR enables creation of Java ME applications which combine Web UI markup technologies with Java code. The intent is to leverage the W3C Compound Document Format (CDF) specification.
293	Location API 2.0	This specification defines an optional package that enables the developers to use new enhanced location-based features on the Java ME devices.
297	Mobile 3D Graphics API 2.0	This new revision of M3G (JSR-184) will expose the latest graphics hardware features on high-end devices, while improving performance and memory usage on the low end.

(Contd)

JSR #	Name of JSR	Description
298	Telematics API for Java ME	This JSR defines the API set for Telematics Service on mobile devices.
300	DRM API for Java ME	This specification will define an optional package for developing Java ME applications which utilize or interoperate with DRM agents that separately exist in devices.
302	Safety Critical Java Technology	This specification creates a J2ME capability, based on the Real-time Specification for Java (JSR-1), containing minimal features necessary for safety critical systems capable of certification, for example, DO-178B.
304	Mobile Telephony API version 2	This JSR extends the interfaces defined in JSR253 (Mobile Telephony API) to cover additional use cases and features not covered in that JSR.
307	Network Mobility and Mobile Data API	This JSR provides APIs for initiating and controlling data sessions in a mobile device and providing applications control over wireless network selection.
927	Java TVTM API 1.1	The maintenance of the Java TV specification.

15.13 Latest in J2ME

J2ME's latest offering is JavaME 3.0 SDK. This flavor of J2ME is highly integrable with third-party applications and framework. This ensures J2ME truly inheriting Java's "High Performance–Cross Platform" legacy. Adding to this, freeware world makes its contribution in respect of applications, games, emulators, embedded devices add-ons, plug-ins, APIs, tutorials, etc. One recent and striking capability of J2ME is its integration with Android. This helps developers to write managed code in Java, controlling the device via third-party-developed Java libraries.

One very recent development is the availability of JavaFX application platform, also known as JavaFX Mobile, and its high degree of affinity of integration with J2ME. This helps such applications (developed in JavaFX Script) to have access to capabilities of the underlying handset, such as the file system, camera, GPS, Bluetooth, etc.

We list few new arenas (and related websites) and their offerings in J2ME.

- www.java.sun.com – The official source for J2ME development and latest releases.
- www.j2mepolish.org – J2ME Polish is a suite of tools and technologies aimed at mobile developers and companies within the mobile space. Its toolkit includes features like Lush, Janus, Touch, Trunk and Marjory. Lush is a highly flexible UI toolkit and Janus provides the toolset for porting mobile application to different handsets and different technology platforms. Touch helps accessing the server side content and communicating with remote parties. Trunk and Marjory provide persistence and device database support. The website also provides support with respect to documentation, product roadmapping and licensing, etc.
- http://java.sun.com/javafx/1/tutorials – This is a good resource for JavaFX and a nice stopsite for developers with provision of tutorials.

- http://www.netmite.com/android/ – This provides a lot of developer's material for integrating J2ME applications with Android.
- http://netbeans.org/features/javame/index.html – A lot of developer's stuff can be found when you go access NetBeans IDE.

15.14 CONCLUSION

J2ME capabilities are quite restrictive from a Java developer's perspective, but to someone who has worked with Sim Tool Kit (STK) or JavaCard technologies it is a luxury. The configurations have been formed keeping in mind the limitations of the devices. Before we move over to other technologies let us take a quick look at the list of "Will and Wonts" for J2ME.

Will

- Basic Java: Object, Class, Runnable, String, System,Thread,Throwable java.lang.
- Data Types: Int, char, String, Stringbuffer.
- Utility classes: Stack, Vector, Hashtable, Enumeration, Date, Calendar, random numbers java.util.
- I/O Stream classes: java.io.
- Operations: integer math, abs, min, max java.lang.Math.
- Exceptions: Limited list of standard exception classes.

Won'ts

- Floating point: Most small CPUs don't have an on-chip floating point and MIDP doesn't include it.
- Class loader: There is no option for the developer to choose the class loader. It is the responsibility of the default loader.
- Finalization doesn't exist: The developer is responsible for cleaning up before object deletion takes place.
- JNI: There is no standardized calling convention or hardware profile and hence it doesn't make much sense to native support. This is also a security threat.
- Reflection is not included in J2ME.
- Error handling: A specific set of exceptions is generated. Other errors are left to the device manufacturer to handle as they deem fit.
- High-level APIs: Heavy weight GUI APIs, Swing and AWT are replaced by more appropriate APIs.

With this we conclude the chapter on J2ME. Next on our agenda is WINce, the Microsoft offering for handheld devices.

REFERENCES/FURTHER READING

1. *An article on MIDP push using SMS is here* http://weblogs.java.net/blog/billday/ archive/2004/02/midp_push_using.html.

2. *A detailed article for using Communication APIs in MIDP* http://developers.sun.com/techtopics/mobility/configurations/ttips/cldcconnect.
3. *A good source for all mobile technologies* http://www.devx.com/DevX/Door/16148.
4. *A great article on provisioning* http://developers.sun.com/techtopics/mobility/midp/ttips/provisioning/index.html.
5. *A historical perspective of JAVA is here* http://www.wired.com/wired/archive/3.12/java.saga.html?topic=&topic_set.
6. Feng Yu and Jun Zhu 2001. *Wireless Java Programming with J2ME*, SAMS.
7. *Introduction to various configurations and profiles* http://developers.sun.com/ techtopics/mobility/getstart/articles/intro.
8. *JCP overview* http://www.jcp.org/en/introduction/overview.
9. *JSR for CLDS J2ME is here* http://www.jcp.org/en/jsr/detail?id=30.
10. *J2me home* http://java.sun.com/j2me/index.jsp.
11. *Links to lots of free online articles and resources* http://www.j2meolympus.com/freebooks/freej2mebooks.jsp.
12. *"MIDP Game" API article* http://www.microjava.com/articles/techtalk/game_api?content_id=4271.
13. *Network programming with MIDP 2.0* http://developers.sun.com/techtopics/mobility/midp/articles/midp2network/.
14. *Official account of the first five years of java is here* http://java.sun.com/features/ 2000/06/timeline.html.
15. Piroumian Vartan 2002. *Wireless J2ME Platform Programming*, Prentice-Hall.
16. *Push registry in MIDP* http://developers.sun.com/techtopics/mobility/midp/articles/pushreg/.
17. *RMS usage* http://www-106.ibm.com/developerworks/java/library/wi-rms/.
18. *Security in J2ME* http://www-106.ibm.com/developerworks/library/j-midpds.html.
19. Sun has some excellent articles at http://www-106.ibm.com/developerworks/wireless/library/wi-midlet2.

REVIEW QUESTIONS

Q1: What are the differences between J2ME and other flavors of Java like J2SE or J2EE?
Q2: What is J2ME MIDP? Explain its various functional components.
Q3: What is CLDC? How do you program for CLDC?

Q4: Explain MIDlet lifecycle? How is provisioning done in MIDP application?
Q5: How do you program for multimedia in J2ME?
Q6: What is Record Management System in J2ME? How do you handle records in J2ME?
Q7: Write short notes on:
(a) GUI in MIDP
(b) Communication in MIDP
Q8: What are the different security considerations in J2ME? Explain the mechanisms through which they are handled.
Q9: How would you design a J2ME application for storing wallpapers in a user created directory? How can a user be allowed to edit such a wallpaper?
Q10: How would you design a J2ME Sudoku application? Enumerate the steps. Also, the application should allow the user various levels of Sudoku complexity.
Q11: What shall be the design considerations linking a Java application of your PC to a J2ME application on your mobile handset?
Q12: What are the constraints in reading a SIM using a J2ME application in a handset?
Q13: How would you design a distributed application (say, a car racing game) so that it can run over at least four J2ME devices using Infrared? Please mind the security of the application.

CHAPTER 16

Wireless Devices with Windows CE

16.1 INTRODUCTION

The current trend in electronics is to get more for less. The digital world is becoming more and more miniaturized. Starting from PDA to cellular phones, all have more power in lesser space at lesser cost. In Chapter 13, we studied Palm OS in PDAs. We studied Symbian OS for cellular phones in Chapter 14. We also studied Java (J2ME) in micro devices starting from cellular phones to PDAs in Chapter 15. One question that comes in mind is, "What about Microsoft?" Is it not present in this space? Well, Microsoft is not lagging behind. Microsoft has an OS solution for all these small devices. This OS is Windows CE, first released in early 1997. Windows CE is the Windows operating environment for small devices. The word CE does not have any full form; some people claim that CE stands for Compact Edition though. Some people also claim that Windows CE is a scaled down version of Windows 95. It does not matter what the history is, Microsoft claims that Windows CE has been developed from scratch. Some of the major differences of Windows CE (v 3.0) with respect to the embedded version of desktop Windows NT (v 4.0) are listed in Table 16.1. Please note that for smaller applications, Windows CE will be the preferred operating system and in some specialized applications (like computer telephony), Windows NT Embedded would be preferred.

Windows CE is available from various OEM (Original Equipment Manufacturer) companies. An OEM manufactures the device that will use the Windows CE operating system. This includes both PDA and cellular phone manufacturers. Following is a list of some of the Windows CE Palm Size devices:

- Casio E-15, E-100, E-105
- HP Jornada 420 and 428
- HP/Compaq 6500 Series
- IBM Workpad z50
- Samsung 1900

Table 16.1

Feature	*Windows CE*	*Windows NT Embedded*
Operating System Version	3.0	4.0 service pack 5
Supported CPUs	x86, Mips, H3, SH4, Strong ARM, ARM, Power PC	x86 (Pentium, AMD K5/6, Cyrix 5x86/6x86)
CPU Speed	Runs on as little as 80 mhz, can operate at 500 + mhz	Recommended 300 mhz, can operate at 500 + mhz
Multiprocessor	None-uniprocesor	Up to 32, server edition, 4 workstation edition
Multitasking	Preemptive–limited to 32 applications, supports threads	Preemptive, supports threads
Memory-minimum	1 MB execution, no storage	12 MB execution, 8 MB storage w/o networking, 16 Mb execution, 16 MB storage w/networking
Paging	Dynamic paging based on available internal RAM	Paging file to secondary storage (fixed or dynamic) or disable paging
Utilities	Command Shell, Pocket Internet Explorer (equivalent to Internet Explorer 4.0), Pocket Inbox, Help Engine-Client Functionality, Windoes Terminal Server Client	Command shell, text editing, Windows Explorer, Microsoft Management Console, network configuration utilities, Windows help engine, task scheduling, and others—Server or Workstation functionality
General Features	Headless support Diskless support–Boot from flash media, or CD Rom (Sega Dreamcast boots from CD)	Headless support Diskless support-Boot from flash media or CD Rom
Security	None	NTFS, application level
Display	Optimized for smaller displays, supports up to 800 x 600	640 x 480 and larger standard displays
Communications Protocols	TCP/IP, PPP, SLIP, PAP, CHAP, HTTP, IrDA	TCP/IP, IPX/SPX, Apple Talk, Netbeui, PPTP, HTTP, RPC, SNMP
Data Storage	ATA Flash, Linear Flash, PC CArd hard disk, Compact Flash hard disk, IDE hard disk-FAT format-no capacity limited, CD Rom	ATA Flash, M-Systems Disk on a Chip_max capacity is 144 MB flash, Bootable CD Rom
Data Storage Formats	FAT, FAT 32	FAT, NTFS, compression
API	Subset of Win 32	Full Win 32
Development Tools	Platform Builder; Requres Visual C++	Target Designer

16.2 Different Flavors of Windows CE

Windows CE operating system is available for the entire range of handheld devices. This includes the smallest device like a phone to slightly larger device like PDA or Pocket PC. These devices can be grouped into two major categories. One is a handheld device and the other one is an embedded device. In the case of handheld devices, the user has high interaction with the OS, whereas in the case of the embedded one, the interaction is less. An embedded version can run without a keyboard and display. Different flavors of Windows CE are available today. These are:

- **Windows Mobile for Pocket PC:** A Windows Mobile-based Pocket PC is a miniature of a PC, which can fit the size of a pocket. Windows Mobile (now called Windows Phone), is a compact mobile operating system developed by Microsoft, and is based on the Windows CE 5.2 kernel. It has a suite of basic applications development using the Microsoft Windows API. It enables mobile computing devices to run applications suitable for small devices by optimizing the user interface, applications size and corresponding feature sets around mobile personal information management and connectivity scenarios. By standardizing core resource requirements and providing a consistent set of programming APIs, Windows Mobile environment provides a consistent application development environment across devices. Windows mobile for pocket PC enables the user to store and retrieve e-mail, contacts, appointments, play multimedia files, games, exchange text messaging, browse the Web, etc. The user can also exchange or synchronize information with a desktop computer.
- **Windows CE .NET:** Windows CE .NET and Windows XP Embedded belong to the Microsoft family of embedded operating systems. Windows CE .NET combines an advanced, real-time operating system with tools for rapid development for the next generation of smart, connected and small footprint devices. Windows CE .NET provides a componentized, customizable, embedded OS. It offers rich configuration and application options for a broad range of embedded devices. Such devices range from enterprise tools such as industrial controllers, communications hubs and Windows-based thin clients to consumer products such as digital cameras, voice-over Internet protocol devices and IP-based set-top boxes. Platform Builder is the integrated development environment for building, debugging and deploying a customized embedded OS based on Windows CE .NET.
- **Windows Mobile for Phones:** This version of Windows CE comes in two flavors. These are Pocket PC Phone and Smartphone. The basic difference between the two is that Pocket PC phone is basically a PDA combined with phone features. These are also known as communicators. Whereas, a Smartphone is a phone combined with PDA features. Windows Mobile software for Pocket PC Phone offers functions like dial from contacts database, send SMS messages, identify incoming callers or take call notes. Using a Windows Mobile-based Pocket PC Phone and wireless service through GPRS, the user can access the Internet, send and receive e-mail, and access corporate intranet applications while on the move. A Windows Mobile-based Smartphone integrates PDA-type functionality into a voice-centric handset comparable in size to today's mobile phones. A Windows Mobile-based Smartphone is designed for one-handed operation with keypad access for both voice and data features. It is optimized for voice and text communication, wireless access to Outlook information, and wireless access to corporate and Internet information and services.

Table 16.2 shows evolution of Windows CE from v 1.0 through v 6.0 while citing the major add-ons. Also, Figure 16.2 effectively shows the timeline of Windows CE development.

Table 16.2

Version	*Changes*
1.0	Released in 16 November 1996. Codename "Alder". • Devices named "handheld PC" (HPC).
2.0	Released in September 1997. Codename "Birch". • Devices named "Palm-sized PC". • Real-time deterministic task scheduling. • Architectures: ARM, MIPS, PowerPC, StrongARM, SuperH and x86. • 32-bit color screens. • SSL 2.0 and SSL 3.0.
3.0	Released in June 2000. Codename "Cedar". • Major recode that made CE hard real-time down to the microsecond level. • Base for the Pocket PC 2000, Pocket PC 2002 and Smartphone 2002. • Priority levels were increased from 8 to 256. • Object store was increased from 65 536 to 4,19 million allowed objects. • Restricted access to critical APIs or restricting write access to parts of the registry.
4.x	Released in January 2002. Codename "Talisker/Jameson/McKendric. Driver structure changed greatly, new features added. • Base for "Pocket PC 2003". • Bluetooth support. • TLS (SSL 3.1), IPsec L2TP VPN, or Kerberos.
5.x	Released in August 2004. Adds lots of features. Codename "Macallan". • Automatic report of bugs to the manufacturer. • Direct3D Mobile, a COM-based version of Windows XP's DirectX multimedia API. • DirectDraw for 2D graphics and DirectShow for camera and video digitization support. • Remote Desktop Protocol (RDP) support.
6.0	Released in September 2006. Codename "Yamazaki". • Process address space is increased from 32 MB to 1 GB. • Number of processes has been increased from 32 to 32 768. • User mode and kernel mode device drivers are possible. • 512MB physically managed memory. • Device.exe, filesys.exe, GWES.exe has been moved to Kernel mode. • SetKMode and set process permissions not possible. • System call performance improved.

16.3 WINDOWS CE ARCHITECTURE

Figure 16.1 depicts the architecture for the Windows CE. It is a layered architecture. At the bottommost layer we have the hardware layer with all the hardware. Next layer is the OEM layer followed by the Operating system layer. At the top of this stack is the Application layer. Application layer interfaces with the user, whereas the hardware layer interfaces with the hardware resource. All other layers are the facilitators for the user to access the resource.

16.3.1 OEM Layer

The OEM layer is responsible for getting a Windows CE–based OS to run on a new hardware platform. OEM layer is a layer of functions which is developed and maintained by an OEM. For example, HP is an OEM which offers Windows CE with the HP Jornada. Within the OEM layer there is an OEM adaptation layer (OAL). This layer resides between the Windows CE kernel and

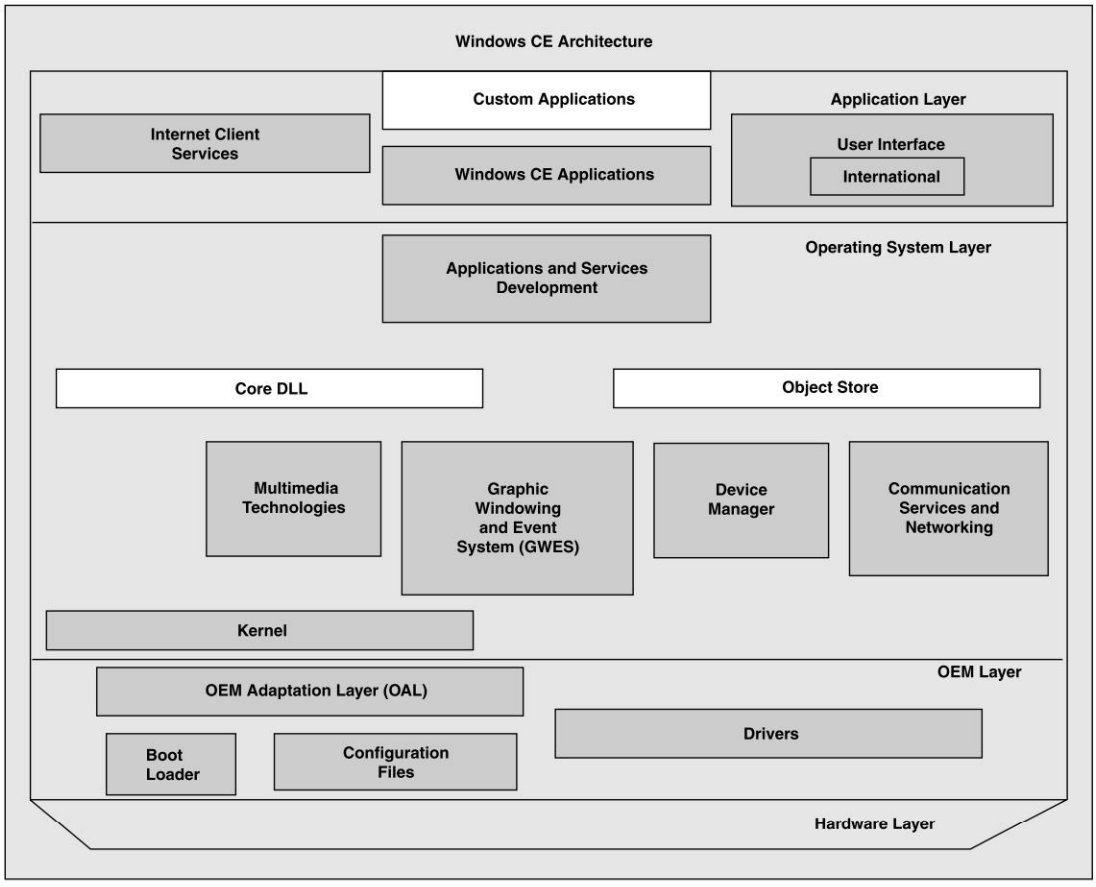


Figure 16.1 Windows CE Architecture

the hardware of the device. It facilitates communication between the operating system (OS) and the target device and includes code to handle interrupts, timers, generic IOCTLs (I/O control codes), etc. Physically, the OAL is linked with the kernel libraries to create the kernel executable file.

In any device we have a layer of software called device drivers. This is no different in the case of Windows CE. In Windows CE, different hardwares interact with the kernel through respective device drivers.

The boot loader is a piece of software that is required to boot the device. The boot loader generally resides in non-volatile storage on the device. It is executed at system power-on/reset (POR). To get the boot loader on the target device for the first time, some special program is used. However, updates of boot loaders are handled by boot loader itself flash the new OS images. The platform initialization code for the device is shared between the boot loader and the OAL. The boot loader provides a menu that allows the user to set different configuration options, such as DHCP or static IP information. These configuration parameters are stored in a configuration file.

16.3.2 Operating System Layer

The operating system layer contains all the software supplied by Microsoft as a part of the operating system. The main component in this layer is the kernel. Along with the kernel, this layer includes functions like Applications and Services Development, Core DLL, Object Store, Multimedia Technologies, Graphics Windowing and Event System (GWES), Device Manager, Communication Service and Networking. Several of these modules are divided into components. Components help Windows CE to become very compact (less than 200 KB of ROM), using only the minimum ROM, and RAM. These are explained in the following sections.

Kernel

The kernel is the core of the OS, and is represented by the Coredll.dll module. It provides the base operating system functionality that is common to all devices. Like any other operating system, Windows CE kernel is responsible for memory management, process management, and certain required file management functions. It manages virtual memory, scheduling, multitasking, multithreading and exception handling. There are some optional kernel components that are needed to include features like telephony, multimedia and graphics device interface (GDI).

Windows CE maps the bottom section of memory into 33, 32 Mb slices called “slots”. The lowest slot is used for the currently running process (the process at slot 0), and other low slots are used for system processes as:

Slot 0: current running process
Slot 1: kernel (NK.EXE)
Slot 2: File system – object store, registry, CeDB etc. (Filesys.exe)
Slot 3: Device manager (Device.exe)
Slot 4: Windows CE shell (Shell32.exe).

Five slots are used, leaving 28 remaining slots for user processes.

Being a 32-bit machine, the Windows CE address space is 4GB. The top 2GB address space is used by the operating system, which includes hardware, object store and ROM. The bottom 2GB address space is used for processes and application shared space.

FFFF FFFF: Top of memory
BF00 0000: Beginning of ROM area
AB00 0000: I/O Registers on a MIPS R4xxx
AA00 0000: Beginning of screen memory
A000 0000: Beginning of DRAM area
8000 0000: Beginning of kernel memory space
[…]
4200 0000: begin shared app memory (memory mapped files)
4000 0000: slot 32 space
[...]
0800 0000: slot4 – shell32.exe
0600 0000: slot3 – device.exe
0400 0000: slot 2 - filesys.exe
0200 0000: slot 1 - nk.exe
0000 0000: slot 0 - current process space

Graphic Windowing and Event System (GWES)

Graphic Windowing and Event System, commonly known as GWES is the graphical user interface between a user and the OS. GWES handles the user input/output. It provides controls, menus, dialog boxes and resources for devices that require a graphical display. It manages user input by translating keystrokes, stylus movements and control selections into messages that convey information to applications and the OS. GWES handles output to the user by creating and managing the windows, graphics and text that are displayed on display devices and printers. All applications need windows in order to receive messages from the OS. This is true even for those devices that lack graphical displays.

Device Manager

In desktop operating systems we load all the device drivers during the startup. However, in Windows CE we do not do so; we use the Device Manager to load the drivers as and when necessary. The Device Manager is launched from the kernel and runs continuously. Device Manager provides the following functions:

- Loads drivers by reading and updating registry keys.
- Unloads drivers when a device no longer needs them.
- Manages device interfaces and interface notifications.
- Manages resources relevant to device drivers, such as I/O space and interrupt requests.

The Device Manager searches the HKEY_LOCAL_MACHINE\Drivers*RootKey* registry key to determine the key to begin the driver loading process. Device Manager tracks loaded drivers and their interfaces. It can notify the user when device interfaces appear and disappear. Additionally, it sends power notification callbacks to device drivers and provides power management services.

Windows CE Storage

Windows CE offers different types of storage. These are registry, file system, object store, and databases. Object store is a new concept and available only in Windows CE. Most of these storages in Windows CE are RAM-based. As they are RAM-based, these storages are kept alive through

the internal battery of the device. In the event of a cold reset, all data in this storage are lost. It is therefore advisable to synchronize these data with a PC. The Windows CE file system can also use Flash-RAM cards to store data. Data on Flash-RAM are saved even during cold-reset.

The registry: This storage is similar to the registry on Windows desktop computers. The system DLL managing registry is coredll.dll. The respective include files and library files are coredll.h and coredll.lib. Windows CE supports only three root keys. These are:

- HKEY_LOCAL_MACHINE
- HKEY_CURRENT_USER
- HKEY_CLASSES_ROOT

The file system: Like any other computer system, Windows CE has a file system for persistent storage. However, this persistent storage is not on a hard disk, but on a RAM. The file system RAM is protected by battery. As the file system is not on a disk drive, the file system on Windows CE has no concept of the drive "c:\". All drives on Windows CE are mounted under the root directory. There are different APIs to handle the Windows CE file system.

The object store: This is a storage that is unique to Windows CE. This is similar to a database but for some special types of information. This database stores all PIM (Personal Information Manager) information. The system DLL handling PIM is coredll.dll. The respective include files and library files are coredll.h and coredll.lib. Accessing the object store involves four steps. These are:

- Mounting of the database volume.
- Create a new database or open an existing database function to create a database is CeCreateDatabase() and to open an existing database is CeOpenDatabase().
- Read or write a record: For reading the database, we may need two functions. These are CeSeekDatabase() and CeReadRecordPropsEx(). For writing into a database we use CeWriteDatabaseProps(). These functions look a little strange as these are not just read or write; they are read or write properties. This is because we store properties and information together in an object store. Let us explain it further. In a standard file, we define the property during the creation of the file. However, in case of object store or a PIM database, it is not so. In case of object store, every record contains the property of the records embedded as a part of the record. The number of properties for different records can be different. Even a property in one record can be a different type in the next record, though it can use the same property ID. If we need to use the CeSeekDatabase() call, all records in the database must share at least one common property. For convenience this could be the sort order on some primary information field.
- Close the database: For this we use the CloseHandle()

Object storage can be accessed through VB, C/C++ or MFC. There are three MFC classes to do these functions. These are:

- CCeDBDatabase class contains all necessary methods to access the database in the object store.
- CCeDBRecord class relates to all record-related methods like adding, deleting, reading of records.
- CCeDBProp class offers all the necessary methods to manipulate all fields and properties of a record. Properties supported by Windows CE are Short, Unsigned Short, Long, Unsigned Long, File-Time, Unicode String, Binary Blob, Boolean, and Double.

These are databases as available in desktop computers. Examples are

Database: Pocket-Access or SQL for Windows CE. These are databases that are based on ADO (Active Data Object) technology. The DLL required for this subsystem are adoce.dll, adocedb.dll, adoce30.dll, adocedb30.dll, adoceoledb30.dll, adosync.dll.

ADO is part of Microsoft's universal data access (UDA) strategy. ADO is a database middleware and offers a complete datastore abstraction. Using ADO, an application can be made completely datastore independent. For Windows CE, the ADO version is known as ADOCE. Through ADOCE, an application can access different types of datastore like:

- SQL datastore: Like any formal databases, which can be accessed through structured query language. Examples could be SQL server, Oracle, etc.
- Non-SQL datastore: These could be datastores like email, directory, text files, streams, Excel, documents, etc.
- Mainframe and legacy data: These datastores would be required by a client application on a Windows CE device to access a database on the remote mainframe.

16.3.3 Communication Services and Networking

The communications component in Windows CE provides support for various types of media access. This varies from device to device. However, the most important ones are communication through the serial port and infrared. Also there is support for Internet and remote access. Windows CE is capable of connecting to the enterprise local area network through both Ethernet and WiFi wireless LAN. In addition to built-in communication hardware, such as a serial cable or infrared (IR) transceiver, PCMCIA support permits a wide variety of aftermarket communications devices to be added to the basic package. Finally, Windows CE supports voice, SMS (Short Message Service) and other telephony services. Following is the list of communications hardware and data protocols supported by Windows CE:

- Serial I/O support
- RAS
- TCP/IP
- LAN
- Wireless Services for Windows CE
- Telephony API (TAPI)

Windows CE Communications Architecture

The underlying assumption for a Windows CE device is that it is a mobile device. Therefore, essentially a Windows CE device needs to communicate to the external world. Windows CE supports two basic types of communications. These are serial communication and communication over a network. Most devices feature built-in communications hardware, such as a serial port or an infrared (IR) transceiver. The NDIS implementation on Windows CE supports Ethernet (802.3), Token Ring (802.5), IrDA, Bluetooth, and Internet through WAN (Wide Area Network) and WLAN.

Windows CE communications model is designed to function on a variety of devices. Applications and devices in the handheld and mobile space differ in their communications needs. Therefore, Windows CE supports a diverse variety of communication options and associated APIs. It provides an OEM with a diverse set of options to choose from. For application developers, Windows CE supports most of the common types of communication. They are accessible via familiar Win32-

based APIs, allowing developers to readily implement communications capability in their applications. In many cases, existing code from other flavors of Windows CE can be used with little or no modification.

Serial I/O

Serial I/O is the most fundamental feature of the Windows CE communications model and is available virtually for all devices. The serial communication would be accessed via a cable or through the IR transceiver. A cable connection is handled with the standard API for serial and file system functions. They can be used to open, close, and manipulate COM ports and read from and write to them. The IR transceiver is also assigned a COM port. Therefore, direct serial I/O is available on an IrDA port using the usual serial communications functions.

Networking and Communication Support

Networking support includes primarily the socket programming interface. This includes different APIs and application interfaces to user programs along with WinSock for normal sockets and IrSock for infrared sockets.

- WinSock and IrSock: WinSock is the socket implementation for the standard TCP/IP. IrSock is an extension of WinSock implementation over the IrDA interface. The Windows CE TCP/IP stack is designed quite efficiently so that it can be configured to effectively support WiFi wireless networking. Windows CE also supports Secure Sockets Layer 2.0, 3.0 and PCT1.0 security protocols. IrSock enables socket-based communication via an infrared transceiver. It is designed to support the industry-standard IrDA protocols. Applications implement IrSock in much the same way as conventional WinSock, although some of the functions are used somewhat differently.
- Browser support (WinINET API): Windows CE supports subsets of the WinINET and Wnet APIs, and an SMB (Service Message Block) redirector. The WinINET API provides support for Internet browsing protocols, including FTP and HTTP 1.0. Only one proxy is supported, and there is no caching. It also provides access to two Internet security protocols, Secure Sockets Layer (SSL), and Private Communication Technology (PCT).
- Remote file access (Wnet API): The Wnet API provides access to an SMB redirector for remote file access. Currently only Microsoft Windows 95 and Windows NT operating system connections are supported.

Remote Access and Networking

This includes remote access where the Windows CE device is a client.

- Windows CE supports a remote access services (RAS) client. RAS is multi-protocol router used to connect remote devices. The Windows CE RAS client supports one point-to-point connection at a time.
- NDIS 4.0 for local area networking: For local area networks (LANs), Windows CE includes an implementation of NDIS 4.0. At present, only Ethernet miniport drivers are supported. Wide area networks (WANs) are not supported.
- Windows CE-based devices will connect to their network via a serial communications link, such as a modem. To support this type of networking, Windows CE implements the widely used Serial Line Interface (SLIP), and point to point (PPP) protocols. Authentication is provided

via password authentication protocol (PAP) and challenge authentication protocol (CHAP).

Telephony API (TAPI)

For smart phones we need various telephony supports on the device. These telephony supports are generally required for voice communication in GSM or CDMA cellular networks. These telephony interfaces are abstracted and function as an independent isolated interface within the same device. However, there will be instances when the telephony interface needs to integrate with the applications. For example, a calendar application may like to send a reminder to a few people through SMS. Also, there could be need to initiate a GPRS data call to connect to the intranet application. The connection will be established by the telephony interface within the device. These are done through the TAPI (Telephony API). TAPI is a collection of utilities that allows applications to take advantage of a wide variety of telephone and communications services. Windows CE includes TAPI service for AT command-based modems (Unimodem). AT stands for attention and a command level interface; through this interface, we can tell the telephony to do certain tasks for us. In the SMS chapter we have seen some of the AT examples. TAPI can be used with either attached or PCMCIA modems. The Windows CE TAPI supports outgoing calls with outbound dialing and addresses translation services. TAPI does not support inbound calls yet.

Multimedia support module

Windows CE supports audio and multimedia technologies. Within Windows CE this is achieved through the high-performance DirectX. DirectX provides low-level access to audio and video hardware in a device-independent manner. DirectX delivers a consistent set of capabilities across a variety of hardware configurations. This is achieved through the use of a hardware abstraction layer (HAL) and a hardware emulation layer (HEL).

Windows CE also supports Windows Media technologies. These are designed to provide audio and video playback support for a wide variety of streaming and non-streaming media formats. The Windows Media Player control allows a developer to add playback support to web pages or other applications. The following multimedia technologies are supported in Windows CE:

- Microsoft Direct 3D.
- DVD-Video API.
- Windows Media audio and video codes.
- A new unified audio model that uses waveform audio drivers for waveform audio, the audio mixer, and DirectSound.
- Microsoft DirectDraw.
- Microsoft DirectSound.
- Microsoft DirectShow.
- Windows Media technology.
- Windows Media Player control.

All the above multimedia components other than the Direct3D are supported on all Windows CE devices. Direct3D feature requires floating point support on the device.

COM Support Module

The Component Object Model or COM is available across all Windows operating systems. COM is used by applications for object-oriented inter-application and intraapplication communicates.

The desktop version of COM offers different methods for communication. However, Windows CE offers the following subsets:

- ActiveX control: Reusable code and objects, specially for user interface and other invocations can be implemented quite efficiently through ActiveX controls.
- In-process activation: In-process COM servers are practically ActiveX controls without a user interface.
- EXE to EXE communication: Using EXE to EXE communication, applications can communicate over the process boundaries. Two applications can have peer-to-peer communicate with each other through this technology without using shared memory, sockets or temporary files.
- MTS (Microsoft Transaction Server)-client only: MTS is used for application offering a transaction processing paradigm. However, a Windows CE device can be a client to a MTS-hosted application running on a desktop.
- DCOM (Distributed COM): DCOM is the implementation of COM across machine boundaries.
- Message Queue-client only: Message queues are very useful for asynchronous communication between peers. A Windows CE device can act as a message queue client.

Windows CE Shell Module and User interface

We have discussed that in Windows CE, GWES is the interface between the user, the applications and the operating system. Windows CE shell module and UI combine the Microsoft Win32 API (application programming interface), UI (user interface) and GDI (graphics device interface) libraries into the GWES module (Gwes.exe). GWES supports all the windows, controls and resources that make up the Windows CE user interface. GWES also includes support for user input and output, through support for keyboard input, fonts, text drawing, line and shape drawing, palettes and printing.

GWES supports various types of resources. Resources are objects that are used within an application but are defined outside an application. GWES includes support for the following resources:

- Keyboard accelerators
- Menus
- Dialogs boxes and message boxes
- Bitmaps
- Carets
- Cursors
- Icons
- Images
- Strings
- Timers

GWES supports various types of controls. A control is a child window that an application uses in conjunction with another window to perform input/output tasks. Common controls are a set of windows that are supported by the common control library. Common control is a DLL that is included with Windows CE. GWES includes support for the following common controls:

- Command band
- Command bar
- Header control
- Image list
- List view
- Month calendar control
- Pocket PC-style ToolTips
- Progress bar
- Property sheet
- Rebar
- Status bar
- Tab control
- Toolbar
- ToolTip
- Trackbar
- Tree view
- Up-down control
- Date and time picker

GWES supports various types of window controls. A window control is a predefined child window that allows a user to make selections, carry out commands, and perform input/output tasks. GWES includes support for the following window controls:

- Check boxes
- Push buttons
- Radio buttons
- Group boxes
- Combo boxes
- Edit controls
- List boxes
- Scroll bars
- Static controls

GWES also supports the CAPEDIT Control and SBEDIT Control for use in edit controls. Following are some additional GWES features that developers can add to displaybased platforms.

- Accessibility: Options that allow persons with disabilities to use the device more easily.
- Fonts: Provides 60 different fonts for displaying and printing text.
- Mouse/Pen: Allows users to provide input through a pointer device like mouse or pen.
- Stylus: Allows users to provide input through stylus and touch screen.
- Multiple Screens: Enables a device to connect to multiple screens.
- Printing: Supports the ability to print.
- Input Panel: Allows users to provide input through a input panel displayed on a touch screen.

Developers of applications for Windows CE can customize the UI by creating a skin. Platform developers can change the behavior of menus. In Windows CE, menus by default contain only one level, which means that menu items do not open submenus. Platform developers can choose the Overlapping Menus feature to provide support for cascading, overlapping menus to enable submenus.

The shell architecture in Windows CE allows developers to implement a wide variety of shells. All the source code for the presentation and user interface aspects of the standard shell is available to developers. This allows fully customized shells built for individual platforms to be completely integrated into the OS. Developers can add the following shell features to display-based platforms.

- Standard Shell: Provides a shell that is similar to the shell on the Windows-based desktop platforms. The source code for this shell is available for customization.
- Command Processor: This is used for a command-line-driven shell interface for console input and output and a limited number of commands.
- Windows Thin Client Shell: Provides a Windows Thin Client user interface
- API compatibility support: These extensions to the standard shell are achieved through AYGShell API extensions. AYGShell support allows most Pocket PC-based applications to run on a Windows CE-based device after being recompiled for an OEM's Windows CE-based platform.

Embedded Windows CE platforms do not require full GWES features because such platforms do not require an interactive display or an input device like a keyboard or a mouse. Therefore, for embedded devices we can use the Minimal GWES Configuration Features. These GWES supports are the glue between the embedded platforms and display-based platforms that Windows CE provides.

Following are the Minimal Configuration features in Windows CE.

- Minimal GWES Configuration: Provides basic windowing and message queue support.
- Minimal Input Configuration: Provides support for keyboard input.
- Minimal GDI Configuration: Provides GDI support, including TrueType fonts, text drawing, and palette support.
- Minimal Notifications Configuration: Provides support for notifications.
- Minimal Window Manager Configuration: Provides support for window management.

16.3.4 Application Layer

The last and final layer is the custom application layer. All the Windows CE user applications form this layer. Subsystems for user interface and internationalization are part of this layer. Windows CE applications available from Microsoft, applications developed by OEM or application developed by third parties are part of this layer.

16.4 Windows CE Development Environment

The application for Windows CE is developed on a desktop environment. It is tested through a simulator on the desktop environment. Once the application is found working on the simulator it is loaded on the physical Windows CE device for testing. Therefore, Windows CE development environment comprises different systems and subsystems in the desktop. Also, it requires some components on the Windows CE device. Following are the components:

Windows 2000 operating system on the workstation: Windows CE development environment can run on desktop OS like Windows 95, Windows 98 or other flavors of Windows. The facility available in Windows 95 and 98 are quite restricted. Therefore, it is recommended that Windows 2000, or a later edition is used as the host OS on the workstation for the Windows CE development.

- **Visual Studio:** The Windows CE development environment is an add-on on the desktop development environment. Visual Studio for Windows CE extends the Visual Studio for desktop.
- **eMbedded Visual Tools:** This set of software comprises different tools for the Windows CE development and debugging. The set includes:
  - o Remote file viewer
  - o Remote Heap Walker
  - o Remote Process Viewer
  - o Remote Registry Editor
  - o Remote Spy++
  - o Remote Zooming
  - o Control Manager
- **SDK (Software Development Kits):** For Palm-size PC and Handheld/PC SDK can also be used. This can give a better control over the device.
- **ActiveSync:** Like we archive our files in the desktop, we need to archive our files and data in the Windows CE device. Also, as the data is stored in RAM, it is advisable to save the data from time to time. ActiveSync is used to perform all these functions. All Windows CE devices come with a cradle that attaches to a desktop or a laptop PC via ActiveSync. It is a system that manages the connection between a desktop computer and the Windows CE device. ActiveSync can be configured to synchronize e-mail, calendar appointments, contacts and many more applications.

Once the software are installed, we need to connect the workstation with the Windows CE device. This is required for multiple functions starting from installing of the Windows CE software to the debugging of the Windows CE system. This will be used as the PC-Link.

16.4.1 Windows C++ Development

One of the most common platforms for developing applications for Windows CE is Visual C++. All application developed using Visual C++ will work on all Windows CE platforms. For C++ development we need Visual C++ Toolkit for Windows CE. We can develop application for Windows CE using the MFC classes. To develop applications using Visual C++ we will need the following:

- Visual C++ Professional or Enterprise edition (latest release).
- HPC or PSPC SDK, downloadable for free from the Microsoft website.
- The Visual C++ Toolkit for Windows CE (latest release).

16.4.2 VB Development

Visual Basic (VB) is another platform for developing Windows CE applications. To develop a Windows CE application using VB we need the following:

- The Visual Basic Toolkit for Windows CE (latest release).
- Visual Basic, Professional or Enterprise edition (latest release).
- The HPC SDK, HPC Pro SDK or the PSPC SDK v. 1.2.

16.4.3 Windows CE Programming

As discussed in previous sections, for Windows CE programming all the popular programming languages from Microsoft are available. These are VB (Visual Basic) and Visual C/C++ with MFC. Programming of Windows CE is no way different from standard Windows on any desktop PC. Though the underlying hardware and operating system are different in Windows CE, the programming interfaces through VB and C/C++/MFC are the same as in desktops. Also, owing to huge popularity of Windows OS, Windows CE has had a background advantage. Add to this, good availability of programmer's material in the form of websites, blogs, podcasts, etc. As Windows CE has evolved into v 6.0 (latest stable version), lot of developer portals and even Microsoft provides 24 x 7 support.

16.4.4 Windows Phone (Windows Mobile) Programming

Though Windows Mobile has been updated multiple times, while originally appearing as the Pocket PC 2000 operating system and graduating to Windows Phone 7, its market share is yet to see its full noon. Nevertheless, a lot of development support exists for it. Third-party software development is available for the Windows Phone. There are many options for developers when deploying a mobile application including writing native code with Visual C++, writing Managed code compatible with .NET Compact Framework, or even server side code that can be deployed using Internet Explorer Mobile or similar client on the user's device. Normally, Microsoft releases Windows Phone SDKs which work in conjunction with their Visual Studio development environment. And the best part is that these SDKs include emulator images for developers to test and debug their applications while writing them. This is in addition to Microsoft's regular free distribution for enthusiasts and student community.

Lazarus, Lexico, NS Basic and Basic4ppc provide an alternative desktop development environment, which is later downloadable to the device. There is also a Python port named PythonCE and a GCC port, CeGCC there is a Satellite Forms tool for Windows Phone development. Satellite Forms is a RAD tool which can produce Windows Phone compatible applications that use an RDK runtime engine. The added extension libraries can map the functionality for various hardware.

Also, there's the latest add on from Microsoft in the form of Windows Marketplace for Mobile. Windows Marketplace for Mobile is a service Windows Mobile platform that allows users to browse and download applications developed by third parties. Please note that such application are available for use directly on Windows Mobile 6.5 devices and, also, on personal computers. This is quite in competition with Apple's App Store (for its iPods). Though, both of them are proprietary and priced.

The following weblinks are worth visiting with respect to Windows Phone development:

- http://www.microsoft.com/windowsmobile/en-gb/default.aspx–Official site of Windows Mobile and development support.
- http://msdn.microsoft.com/en-gb/windowsmobile/default.aspx–MSDN Windows Mobile Development Center.
- http://developer.windowsphone.com/Default.aspx–Windows Mobile Developer's Portal with follow up on Facebook, Twitter and YouTube.

- http://forum.xda-developers.com/ – Another developer's site.
- http://blogs.msdn.com/windowsmobile/default.aspx – Blog for Windows Mobile.

REFERENCES/FURTHER READING

1. Kruglinski David J., George Shepherd and Scot Wingo: *Programming Microsoft Visual C++*, Microsoft Press.
2. Muench Chris, *The Windows CE Technology Tutorial*, Addison Wesley.
3. http://msdn.microsoft.com/library/default.asp?url=/library/enus/ wceintro5/html/ wce50conIntroducingWindowsCE.asp.
4. http://www.microsoft.com.
5. http://www.msdn.com.

REVIEW QUESTIONS

Q1: Explain different layers in the Windows CE architecture.

Q2: Explain how communications and networking are handled in Windows CE. How does one program these interfaces?

Q3: How does one program the telephony interfaces in Windows CE?

Q4: How is multimedia handled in Windows CE?

Q5: Explain the Windows CE development environment. Explain with a brief example.

Q6: Explain each of the following in brief:
- (a) Flavors of Windows CE
- (b) Windows CE programming
- (c) TAPI in Windows CE
- (d) User interface in Windows CE

Q7: Why do you think that Windows Phone is lagging behind Android in popularity?

Q8: What are the edges in using Windows Phone than any other mobile OS?

Q9: How would you develop a handheld phone loaded with Windows Phone as another device on a network with Windows NT running? The handheld should either run the application or outsource to another node with lesser load, in case it lacks the required resource capacity.

Q10: Is it possible to run a J2ME application on Windows Phone? If not, what is the best way of porting it?

Q11: How can a Windows Phone handheld connect to other handhelds not loaded with Windows Phone?

CHAPTER 17

Voice Over Internet Protocol and Convergence

17.1 VOICE OVER IP

Traditionally, for decades circuit-switched technologies were in use for voice communications. In a circuit-switched technology, a channel (a timeslot in Time Division Multiplexing, a frequency in Frequency Division Multiplexing, or a space in Space Division Multiplexing, etc.), is reserved to establish an end-to-end circuit. The channel is reserved for the connection, and users pay for the entire length of the circuit (in space and time) irrespective of the fact whether they are talking or thinking. The circuit could carry voice traffic, which could be either a digitized or analog voice. While circuit-switching provides good voice quality, it may not be efficient in channel utilization. In contrast, packet-switched networks carry data in packets from multiple sources and destinations over one channel. Such networks are better in channel utilization but suffer from delays and jitters. For real-time traffic like voice, delays and jitters are not nice qualities. IP (Internet Protocol) is one such packet-switched network protocol which is efficient for data communication but not suited for real-time voice.

In 1995 some lobbyists in Israel made an attempt to send voice over IP network between two PCs. Later in the same year, Vocaltec, Inc. released Internet Phone Software. By 1998 few companies started setting up gateways to allow PC-to-Phone and later Phone-to-Phone (over private corporate IP networks) connections. Technology to enable such voice communication over the IP network became known as Voice over Internet Protocol or VoIP in short. By 2000, VoIP traffic exceeded 3% of voice traffic. Most of these VoIP technologies were proprietary and did not interoperate. To ensure interoperability between protocols and equipment from different vendors, standards started emerging. These standards were from two major camps, viz., the telecommunication camp and the data camp. Today there are two sets of standards for VoIP switching, media, and gateways. These are H.323 from ITU (International Telecommunications Union) and SIP (Session Initiation Protocol) from IETF (Internet Engineering task Force). Figure 17.1 depicts the H.323, SIP and MGCP (Media Gateway Control Protocol) and connection among them.

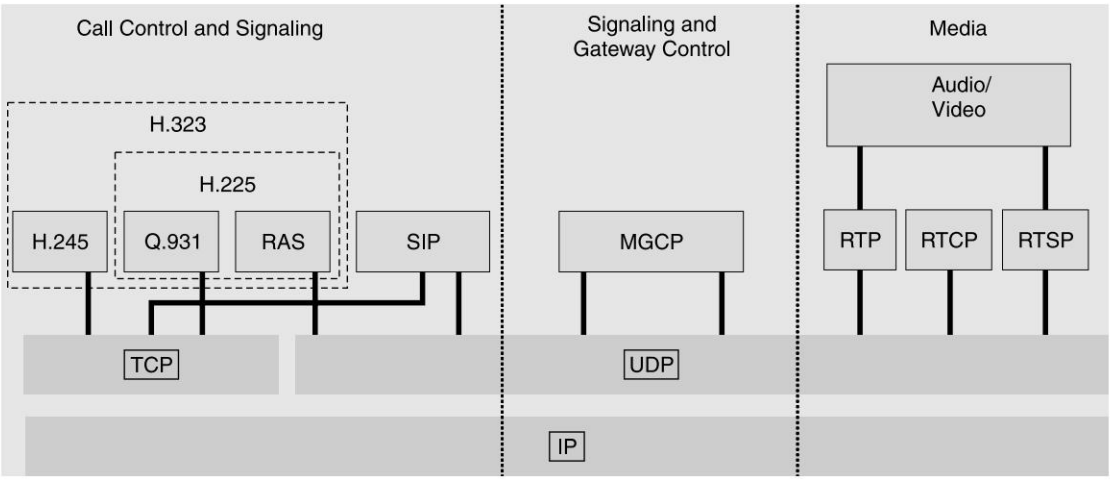


Figure 17.1 H.323, SIP and MGCP

17.2 H.323 Framework for Voice Over IP

The H.323 is a set of protocol standards that provides a foundation for multipoint conferencing of audio, video, and data communications over IP networks standardized by the ITU. It is used for peer-to-peer, two-way delivery of real-time data. The scope of H.323 (Fig. 17.2) includes parts of H.225.0–RAS, Q.931, H.245 RTP/RTCP and audio/video codecs, such as the audio codecs G.711, G.723.1, G.728, etc., and video codecs like H.261, H.263 that compress and decompress media streams. It includes codecs for data conferencing through T.120 and fax through T.38. H.235 specifies security and encryption for H.323 and H.245 based terminals. H.450.N recommendation specifies supplementary services such as call transfer, call diversion, call hold, call park, call waiting, message waiting indication, name identification, call completion, call offer, and call intrusion. H.246 specifies Internetworking of H Series terminals with circuit-switched terminals.

In a H.323 implementation, along with the end-user devices three logical entities are required as depicted in Fig. 17.3. These are Gateways, Gatekeepers and Multipoint Control Units (MCUs). Terminals, Gateways, and MCUs are collectively known as endpoints. It is possible to establish an H.323-enabled network with just terminals, which are H.323 clients. Yet for more than two endpoints, a MCU is required.

17.2.1 Gateway

The purpose of the gateway is to do the signal and media translation from IP to circuitswitch network and vice versa. This includes translation between transmission formats, translation between audio and video codecs, call setup and call clearing on both the IP side and the circuit-switched network side. The primary applications of gateways are:

- Establishing links with analog PSTN terminals.
- Establishing links with remote H.320-compliant terminals over ISDN-based switched-circuit networks.
- Establishing links with remote H.324-compliant terminals over PSTN networks.

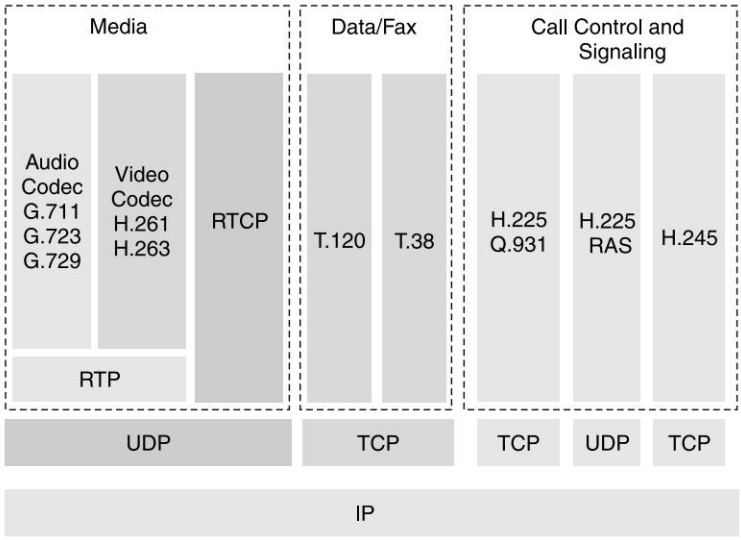


Figure 17.2 H.323 Umbrella Specification

17.2.2 Gatekeeper

A gatekeeper acts as the central point of control for all calls within its zone for all registered endpoints. A gatekeeper is not mandatory in an H.323 system. However, if a gatekeeper is present, terminals must use the services offered by gatekeepers. Gatekeepers perform functions like address translation and bandwidth management. For example, if a network has a threshold for the number of simultaneous conferences on the LAN, the gatekeeper can refuse to make any more connections once the threshold is reached. An optional feature of a gatekeeper is its ability to route H.323 calls. By routing a call through a gatekeeper, service providers can meter a call with an intention of charging. Routing of call through gateway can also be used for call forwarding to another endpoint. The gatekeeper plays a major role in multipoint connections by redirecting the H.245 Control Channel to a multipoint controller. A gateway could use a gatekeeper to translate incoming E.164 addresses into IP addresses.

17.2.3 Multipoint Control Unit

The Multipoint Control Unit (MCU) supports conferences between three or more endpoints. An MCU consists of a Multipoint Controller (MC) and a Multipoint Processor (MP). The MC handles H.245 negotiations between all terminals to determine common capabilities for audio and video

processing. An MCU optionally may have one or more MPs to deal with the media streams. MP mixes, switches, and processes audio, video, and/or data bits.

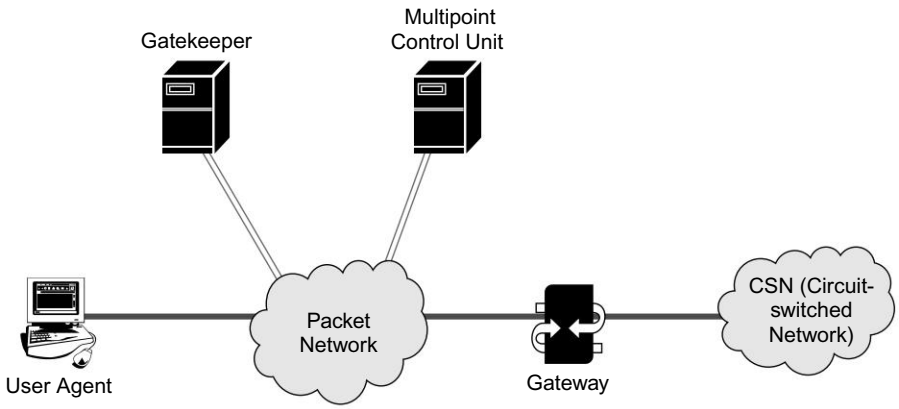


Figure 17.3 The H.323 Architecture

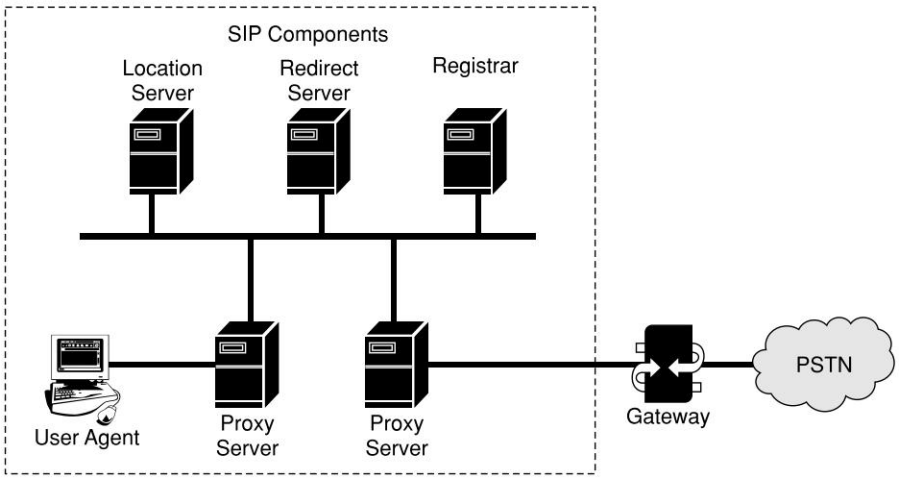


Figure 17.4 The SIP Architecture

17.3 Session Initiation Protocol (SIP)

Session Initiation Protocol (SIP) is a signaling protocol for telephone calls over IP. SIP is defined by the IETF and is gaining popularity. Unlike the H.323, SIP is designed specifically for the Internet. SIP defines interfaces for establishing, modifying and terminating sessions with one or more participants in the VoIP environment. It facilitates development of telephony application. These facilities also enable personal mobility of users. SIP supports the following facets of establishing and terminating multimedia communications:

- User location: Determination of the location and end systems to be used for communication.
- User capabilities: Determination of the media and media parameters to be used for the communication.
- User availability: Determination of the called parties' willingness to engage in communication.
- Call setup: "Ringing," establishing call parameters at both calling and called party.
- Call handling: Managing the transfer of data (voice).
- Call teardown: Terminating the call and releasing all resources.

Figure 17.4 depicts the VoIP architecture with respect to SIP. In such a VoIP setup, the end-user device can be either an IP phone or a computer in an IP network. The conversation can be IP-to-IP, PSTN-to-IP, IP-to-PSTN. In a SIP environment along with the endpoint devices, five entities are required. These are:

- Proxy server
- Registrar server
- Redirect server
- Location server
- Gateways

We describe the functions of these entities and the mode of communications in the following sections.

17.3.1 Proxy Server

According to the SIP standard (RFC3261), "SIP proxies are elements that route SIP requests to user agent servers (UAS) and SIP responses to user agent clients (UAC). A request may traverse several proxies on its way to a UAS. Each will make routing decisions, modifying the request before forwarding it to the next element. Responses will route through the same set of proxies traversed by the request in the reverse order." In the SIP context, UAC is the endpoint initiating a call and UAS is the endpoint receiving the call.

SIP proxies function similar to routers and make routing decisions, modifying the request before forwarding it to the next element. SIP standard make provision for proxies to perform actions such as validate requests, authenticate users, resolve addresses, fork requests, cancel pending calls. The versatility of SIP proxies allows the operator to use proxies for different purposes and in different locations in the network. Proxies could be deployed as edge proxy, core proxy or even enterprise proxy. This versatility also allows for the creation of a variety of proxy policies and services, such as routing calls on various intelligent rules. The 3GPP IMS architecture (Section 17.9) for example, uses proxies known as Call State Control Functions of different kinds for various purposes.

17.3.2 Registrar Server

The Registrar server in a VoIP network can be defined as the server maintaining the whereabouts of a domain. It accepts REGISTER requests from nodes in the VoIP network. It places the information it receives as a part of those requests into the location service for the domain it handles. REGISTER requests are generated by clients in order to create or remove a mapping

between their externally known SIP address and the IP address they wish to be contacted at. It uses the location service in order to store and retrieve location information. The location service may run on a remote machine and may be contacted using any appropriate protocol (such as LDAP).

17.3.3 Redirect Server

Redirect server does similar functions as in case of call forwarding in a PSTN or cellular network. A redirect server receives SIP requests and responds with redirection responses. This enables the proxy to contact an alternate set of SIP addresses. The alternate addresses are returned as contact headers in the response SIP message.

17.3.4 Presence Server

Presence is a service that allows the calling party to know the ability and willingness of the other party to participate in a call. A user interested in receiving presence information for another user (Presentity) can subscribe to his/her presence status and receive Presence status notifications from the Presence system. This is achieved through an Event Server. An Events Server is a general implementation of specific event notification, as described in RFC3265. RFC3265 provides a framework that allows an entity to subscribe for notifications on the state change of other entities. The IETF SIPMPLE Working Group is developing a set of specifications for the implementation of a Presence system using SIP. They are working within a general IETF requirements framework for Presence and Instant Messaging, which is called Common Presence and Instant Messaging (CPIM).

17.3.5 SAP/SDP

Session Announcement Protocol (SAP) is an announcement protocol that is used by session directory clients. A SAP announcer periodically multicasts an announcement packet to a known multicast address and port. The scope of multicast announcement is same as the session it is announcing. This ensures that the recipients of the announcement can also be potential recipients of the session the announcement describes.

The Session Description Protocol (SDP) describes multimedia sessions for the purpose of session announcement, session invitation and other types of multimedia session initiation. SDP communicates the existence of a session and conveys sufficient information to enable participation in the session. Many of the SDP messages are sent using SAP. Messages can also be sent using email or the WWW (World Wide Web).

17.3.6 Quality of Service and Security

In any network, quality of service and security are very important. In Internet protocols, RSVP (Resource ReSerVation Protocol) protocol is designed for quality integrated services. RSVP is used by a host to request specific quality of service (QoS) from the network. This could be a SIP service or any other service. RSVP requests the quality results in resources being reserved in each node along the data path.

COPS (Common Open Policy Service) protocol is a simple query and response protocol that can be used to exchange policy information between a policy server (Policy Decision Point–PDP) and its clients (Policy Enforcement Points–PEPs). The model does not make any assumptions about the methods of the policy server, but is based on the server returning decisions to policy requests. The policy could be related to security, authentication, or even QoS. One example of a policy client is an RSVP router that must exercise policy-based admission control over RSVP usage.

17.4 COMPARISON BETWEEN H.323 AND SIP

Functionally, SIP and H.323 are similar. Both H.323 and SIP support call control, call setup and call teardown. H.323 and SIP support basic call features such as call waiting, call hold, call transfer, call forwarding, call return, call identification, or call park. H.323 defines sophisticated multimedia conferencing like whiteboarding, data collaboration, or video conferencing. SIP supports flexible and intuitive service creation using SIP-CGI and CPL. Both H.323 and SIP support capabilities exchange. Third-party call control is currently only available in SIP. Though SIP's deployment started later, it seems to gain momentum. SIP is adopted by 3GPP. The primary factors that encourage SIP's adoption are, simplicity, scalability, and flexibility. Table 17.1 summarizes some of these features of SIP and H.323.

Table 17.1 Comparison between SIP and H323

	SIP	*H.323*
Standard Body	IETF	ITU
Relationship	Peer-to-peer	Peer-to-peer
Client	Intelligent User Agent	Intelligent H.323 terminal
Core Servers	SIP Proxy server, Redirect server, Location server, and Registration servers.	H.323 Gatekeeper, Gateway, Multipoint Control Unit.
Current Deployment	SIP is new with less installations but gaining interest.	Widespread
Capabilities Exchange	SIP uses SDP protocol for capabilities exchange; but not as extensive capabilities exchange as H.323.	H.245 provides structure for detailed and precise information on terminal capabilities.
Control Channel Encoding Type	Text-based UTF-8 encoding.	Binary ASN.1 PER encoding.
Server Processing	Stateless or stateful.	Version 1 or 2 – Stateful. Version 3 or 4 – Stateless or stateful.

(Contd)

	SIP	*H.323*
Quality of Service	SIP relies on other protocols such as RSVP, COPS to implement or enforce quality of service.	Bandwidth management/control and admission control is managed by the H.323 gatekeeper. The H323 specification recommends using RSVP for resource reservation.
Security	Registration–User agent registers with a proxy server. Authentication–User agent authentication uses HTTP digest or basic authentication. Encryption–The SIP RFC defines three methods of encryption for data privacy.	Registration–If a gatekeeper is present, end points register and request admission with the gatekeeper. Authentication and Encryption–H.235 provides recommendations for authentication and encryption in H.323 systems.
Endpoint Location and Call Routing	Uses SIP URL for addressing. Redirect or location servers provide routing information.	Uses E.164 or H323ID alias and an address mapping mechanism if gatekeepers are present in the H.323 system. Gatekeeper provides routing information.
Conferencing	Basic conferencing without conference or floor control.	Comprehensive audiovisual conferencing support. Data conferencing or collaboration defined by T.120 specification.
Service or Feature Creation	Supports flexible and intuitive feature creation with SIP using SIP-CGI and CPL. Some example features include presence, unified messaging, or find me/follow me.	H.450.1 defines a framework for supplementary service creation.
Instant Messaging	Supported	Not Supported

17.5 Real-Time Protocols

We have discussed that for good quality voice we need real-time support. To allow real-time data transmission over TCP/IP, various protocols have been developed. These include protocols for real-time data, audio, video, movie, and streaming data in a unicast or multicast situation. Examples of such protocols are Real-time Transport Protocol (RTP – RFC1889), Real-time Control Protocol (RTCP – RFC3605), and Real-time Streaming Protocol (RTSP – RFC 2326). RTP is a transport protocol for the delivery of real-time data, including streaming multimedia, audio and video. RTCP helps with lip synchronization and QoS management for RTP. RTSP is a control protocol for

managing delivery of streaming multimedia from media servers. RTSP can be considered as the "Internet VCR remote control protocol".

17.5.1 Real-time Transport Protocol

The Real-time Transport Protocol (RTP) is both an IETF and ITU standard (H.225.0). It defines the packet format for multimedia data. RTP is used by many standard protocols, such as RTSP for streaming applications, H.323 and SIP for IP telephony applications, and by SAP/SDP for pure multicast applications. It provides the data delivery format for all of these protocols.

17.5.2 Real-time Control Protocol

The Real-time Control Protocol (RTCP) is based on the periodic transmission of control packets to all participants in the session. RTCP uses the same distribution mechanism as RTP data packets. RTCP can deliver information such as the number of packets transmitted and received, the round-trip delay, jitter delay, etc., that can be used to measure Quality-of-Service in the IP network. This facility allows monitoring of the data delivery in a manner scalable to large multicast networks, to provide minimal control and identification functionality. For RTCP to work effectively, the underlying protocols must provide multiplexing of the data and control packets.

17.5.3 Real-time Streaming Protocol

The Real-time Streaming Protocol (RTSP) is a client-server protocol, designed to address the needs for efficient delivery of streamed multimedia over IP networks. Interoperability on streaming media systems involves many components. These are players in the client device, servers that store the content, encoders that transform or compress the data and tools that create the content. All these must share common mechanisms for interoperability. Encoders and tools must store data types in files in formats that will be understood by players. Encoders and content-creation tools must be able to store content in files that servers can read. Servers must be able to stream content using protocols that players in the client device can understand.

17.6 Convergence Technologies

To make convergence and interworking between PSTN and IP networks possible, three functional gateway elements are defined. Two of these are interface elements: the Media Gateway and the Signaling Gateway. The third element is the Media Gateway Controller. Signaling gateway (SG) is responsible for interfacing to the SS#7 network and forwarding the signaling message to the IP network. The Media Gateway (MG) is responsible for packetization of voice and other real-time traffic (media). The Media Gateway Controller (MGC) plays the role of the mediator to enable and control access and resource usage between the IP and PSTN network. Together, these elements form the building blocks for a distributed architecture approach to providing voice, fax and a set of digital data services over IP networks. Figure 17.5 depicts the convergence architecture. In this architecture we can see an IP SCP (Service Control Point). The functionality of the SCP is similar to those we have described in Intelligent Networks (IN) in Chapter 11. However, an IP SCP is addressable from the SS#7 network.

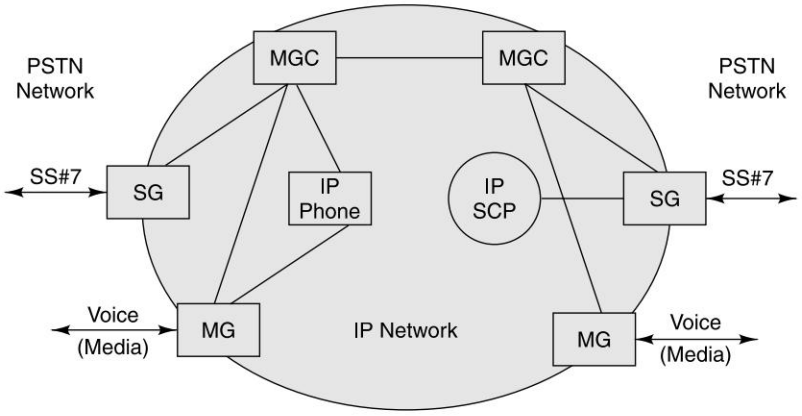


Figure 17.5 Interfaces between IP and PSTN Networks

17.6.1 Media Gateway

The primary responsibilities of the Media Gateway (MG) are to allow media of various types, e.g., voice, fax, video, and modem data to be transported from one type of network to another. These media must be transportable, both as packets in the IP network and as digital or analog streams in the circuit-switched network. They must also be able to move without loss of integrity or degradation of quality. These criteria are met through the use of various coding, compression, echo cancellation, and decoding schemes. The Media Gateway function provides a bi-directional interface between a circuit-switched network and media-related elements in an IP network. Typically, Media Gateways will interact either with IP Telephony end-user applications residing in computers attached to the IP network, or with other Media Gateways. The technology necessary to implement Media Gateways is evolving at a very rapid pace. Media Gateways can implement a variety of physical interfaces to the PSTN. For example, highly scalable Media Gateway systems can implement high speed Time Domain Multiplexing (TDM) trunk interfaces, which are commonly used between switching elements in the circuit-switched network.

17.6.2 Media Gateway Controller

The key responsibilities of the Media Gateway Controller (MGC) are to make decisions based on flow-related information, and to provide associated instructions on the interconnecting of two or more IP elements so that they can exchange information. Media Gateway Controllers maintain current status information of all media flows, and they generate the administrative records necessary for charging and billing. Typically, Media Gateway Controllers instruct Media Gateways on how to set up, handle and terminate individual media flows. A media gateway controller exchanges ISUP (ISDN User Part) messages with central office switches via a signaling gateway. They also provide the parameters associated with bandwidth allocation and, potentially, quality of service

characteristics. Media Gateway Controllers can be used by sophisticated end-user interface applications. In H.323, significant Media Gateway Controller functions are performed in network elements called Gatekeepers. Because media gateway controllers are built primary through software using off-the-shelf computer platforms, a media gateway controller is sometimes called a softswitch.

17.6.3 Signaling Gateway

The Signaling Gateway (SG) function implements a bi-directional interface between an SS#7 network and various call control-related elements in an IP network. The key responsibilities of the Signaling Gateway are to repackage SS#7 information into formats understood by elements in each network, and to present an accurate view of the elements in the IP network to the SS#7 network. Typically, the associated IP network elements will implement Media Controller functions, database storage, or query functions. The SS#7 network has stringent reliability constraints on all devices directly attached to it. By definition, Signaling Gateways need to implement reliable SS#7 messaging that obeys all the rules of the SS#7 network, while also accommodating a variety of behaviors in the IP network. Many of these behaviors are entirely appropriate in an IP world, but not acceptable by PSTN standards. To actually enable IP elements, like Media Controllers, to perform their designated administrative functions, the Signaling Gateway repackages the information contained in various high level SS#7 message protocols such as ISUP and TCAP into formats that can be understood by IP elements. It is necessary for Signaling Gateways to understand all of SS#7 protocols and messaging standards. Finally, since an IP network is a shared medium lacking physical security, Signaling Gateways must filter out the inappropriate traffic that shows up at the Signaling Gateway. It is essential that the Signaling Gateway function protect the SS#7 network from malicious intrusion or accidentally induced undesirable traffic.

17.6.4 Megaco/H.248: Media Gateway Control Protocol

Megaco or Media Gateway Control Protocol is defined in RFC 3015. It is adapted by ITU as a H.248 recommendation. Megaco defines the protocol for control of different elements in a physically decomposed multimedia gateway. There are two basic components in Megaco. These are terminations and contexts. Terminations represent streams entering or leaving the MG. Examples could be analog telephone lines, RTP streams, ATM stream, or MPEG (Moving Picture Experts Group) stream. Terminations may be placed into contexts (Fig. 17.6), which are defined as two or more termination streams being mixed and connected together. Contexts are created and released by the MG under command of the MGC. A context is created by adding the first termination, and released by removing (subtracting) the last termination. There is a special context called the null Context. It contains terminations that are not associated to any other termination.

Figure 17.6 depicts an example of a one-way call waiting scenario in a decomposed access gateway. In Fig. 17.6 (a), Terminations T1 and T2 in Context C1 are engaged in a two-way audio call. While T1 and T2 are in middle of the call, a second call arrives for T1 from Termination T3. T1 has call waiting facility; therefore, T3 stands alone (null Context) in Context C2 at waiting state. In Figure 17.6 (b), T1 accepts the call from T3. This places T2 on hold (parked call). This action results in T1 moving into Context C2.

17.6.5 Sigtran and SCTP

The Signaling Transport (SIGTRAN) group of the IETF defines Sigtran Protocol Architecture through RFC2719 and SCTP (Stream Control Transmission Protocol) standards through RFC2960. SCTP is an end-to-end, connection-oriented protocol that transports data in independent sequenced streams. SCTP was designed to provide a general-purpose transport protocol for message-oriented applications, as is needed for the transportation of signalling data. In the TCP/IP network model, SCTP resides in the transport layer, alongside TCP and UDP.

There are two main differences between SCTP and TCP. These are Multihoming and Multistreaming. In Chapter 11 we have seen that in the signalling network all nodes have sufficient redundancy. For signaling data transfer over IP, similar facility needs to be provided by the Sigtran. Through *Multihoming* SCTP supports multi-homed nodes, i.e., nodes which can be reached under several IP addresses. If we allow SCTP nodes to support more than one IP address, during network failure data can be rerouted to alternative destination IP addresses. This makes the nodes fault-tolerant. In TCP if a packet is lost, the connection effectively stops while it waits for the retransmission to happen. This phenomenon where packets are blocked by a packet in front which has been lost is known as Head-of-Line Blocking. *Multistreaming* is an effective way to limit Head-of-Line Blocking. The benefit in having multiple independent data streams is, if a packet is lost in one stream, while that stream is on wait for the retransmission, the remaining unaffected streams can continue to send data.

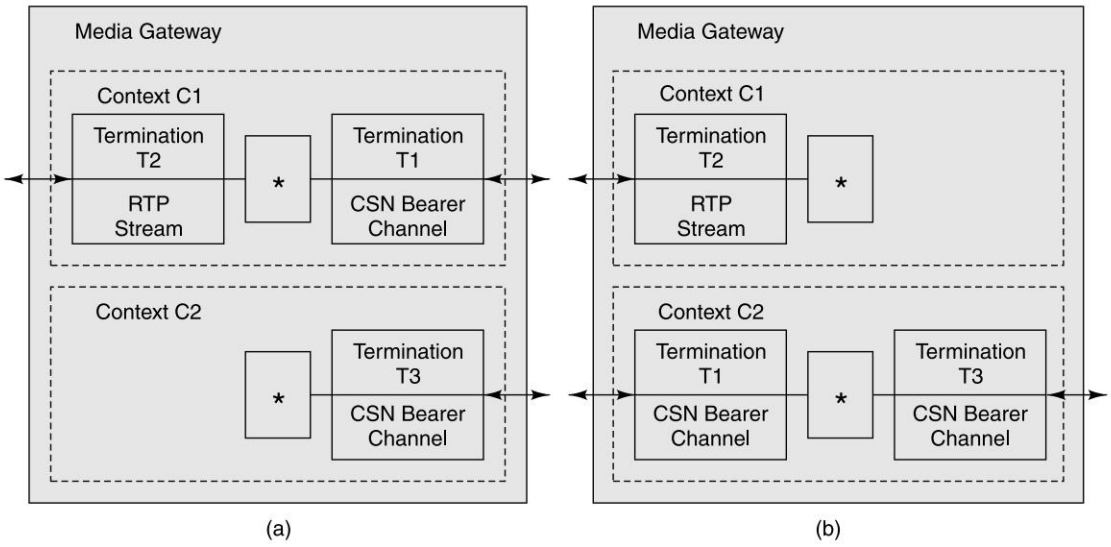


Figure 17.6 Call Waiting Scenario in Megaco

One of the main roles of Sigtran is to tunnel SS#7 signaling traffic. Therefore, to have proper convergence, all the functions of SS#7 stack as discussed in Section 11.4.1 need to be emulated in the IP network. Figure 17.7 is the protocol stack of Sigtran. This stack is the SS#7 replica in the IP

domain. In Sigtran, User Adaptation layers (UA) are defined to identify SS#7 services. They are named according to the services they replace. For example, M2UA, (MTP2 User Adaptation layer) in Sigtran will provide the same services, look and feel to its users as MTP2 (Message Transfer Part 2) does in the SS#7 network. M2PA (MTP2 Peer-to-peer Adaptation layer) on the contrary, will provide MTP2 services in a peer-to-peer manner, such as transparent SG to SG connection over the IP network. The role of M3UA is to provide MTP3 user adaptation APIs. IUS offers ISUP User Adaptation APIs. SUA (SCCP User Adaptation layer) provides the means by which an application part (such as TCAP) on an IP SCP may be accessed by a SCP in the PSTN side via an IP SG. To access a SG, we need a pointcode attached to the SG. However, to access an SUA on the IP SCP, we do not need a pointcode for the IP SCP. Functions of TCAP, SCCP, ISUP, MTP are already defined in Chapter 11.

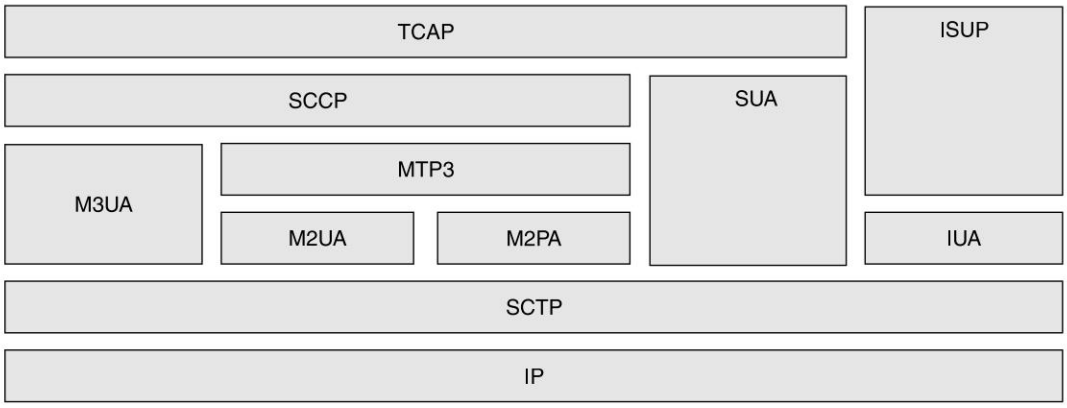


Figure 17.7 SIGTRAN Protocol Stack

17.7 Call Routing

In VoIP, call routing can be divided into four groups. These will be IP to IP, IP to PSTN, PSTN to IP and PSTN to PSTN via IP. Hardware elements, nodes, and protocols used by these groups are not the same. For example, over the LAN, VoIP (IP to IP) may not need a signaling gateway at all. In the following section we discuss the call flow for some of these protocols.

17.7.1 SIP to SIP Call flow

Figure 17.8 depicts a typical example of a SIP message exchange taken from RFC3161. This relates to a communication between two users "A" (Alice in RFC3161) and "B". (Bob in RFC3161). In this example, each message is labeled with the letter "F" and a number for reference by the text. In this example, "A" uses a softphone to call "B" over the IP network. Also there are two SIP proxy servers in the system that act on behalf of "A" and "B" to facilitate the session establishment. "A" calls "B" using B's SIP URI (Uniform Resource Identifier). It has a similar form as an email address,

typically containing a username and a host name. In this case, it is sip:bob@biloxi.com, where biloxi.com is the domain of B's SIP service provider. "A" has a SIP URI of sip:alice@atlanta.com. A's URI could be sips:bob@biloxi.com to signify a secured URI.

In this example, the transaction begins with A's softphone sending an INVITE request addressed to B's SIP URI. INVITE is a SIP method indicating a connection request. The INVITE request contains a number of header fields. The INVITE (message F1 in Fig. 17.8) might look like this:

```
INVITE SIP:BOB@BILOXI.COM SIP/2.0
VIA: SIP/2.0/UDP PC33.ATLANTA.COM;BRANCH=Z9HG4BK776ASDHDS
MAX-FORWARDS: 70
TO: BOB <SIP:BOB@BILOXI.COM>
FROM: ALICE <SIP:ALICE@ATLANTA.COM>;TAG=1928301774
CALL-ID: A84B4C76E66710@PC33.ATLANTA.COM
CSEQ: 314159 INVITE
CONTACT: <SIP:ALICE@PC33.ATLANTA.COM>
CONTENT-TYPE: APPLICATION/SDP
CONTENT-LENGTH: 142
```

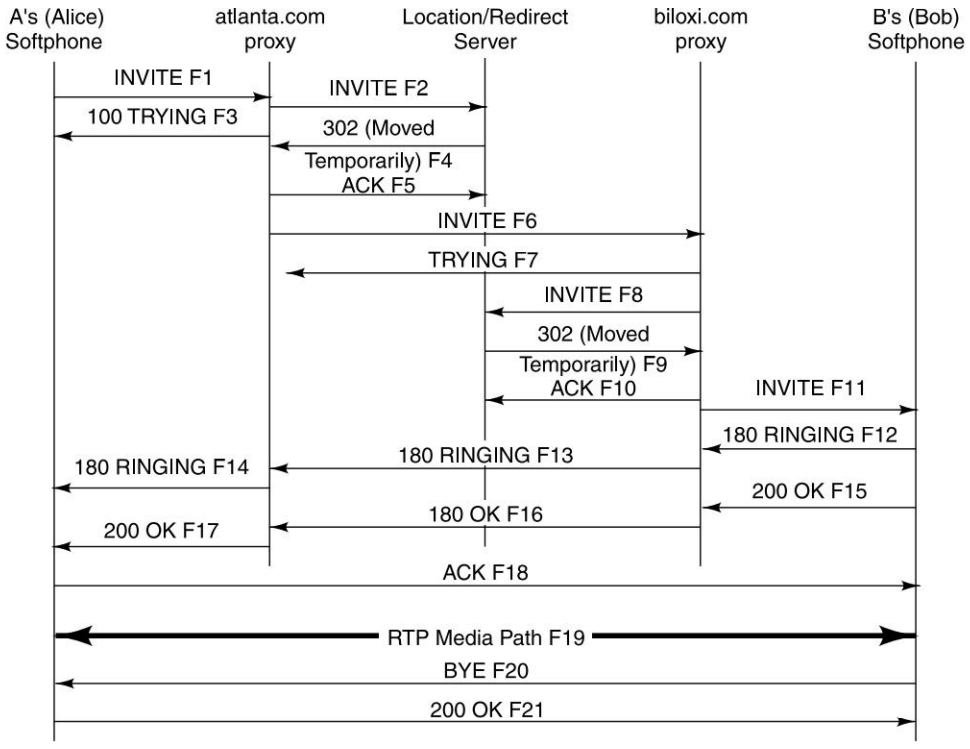


Figure 17.8 SIP Session Setup Example with SIP Trapezoid

The first line of the text-encoded message contains the method name INVITE. The lines that follow are a list of header fields. The "To" field contains a display name (Bob) and a SIP or SIPS URI (sip:bob@biloxi.com) towards which the request was originally directed. SIPS is used for secured transfer like HTTPS. The "From" field contains a display name (Alice) and a SIP URI (sip:alice@atlanta.com) that indicates the originator of the request.

Since the softphone does not know the location of "B" or the SIP server in the biloxi.com domain, the softphone sends the INVITE to the SIP proxy server that serves A's domain, atlanta.com. The proxy server receives the INVITE request and sends a 100 (Trying) response back to A's softphone. The 100 (Trying) response indicates that the INVITE has been received and that the proxy A (atlanta.com) is working on her behalf to route the INVITE to the destination. The proxy A sends the INVITE to the location server to determine whether user B is available or not. If B has moved to a new place that information will be communicated to proxy A. Proxy A then sends the INVITE proxy B (biloxi.xom) at the other end. The biloxi.com proxy server receives the INVITE and responds with a 100 (Trying) response back to atlanta.com. The proxy server consults the location service, that contains the current IP address of B. B's SIP phone receives the INVITE and as a result B's phone rings. B's SIP phone indicates this in a 180 (Ringing) response, which is routed back to A through both the proxies in the reverse direction. Each proxy uses the Via header field to determine where to send the response and removes its own address from the top. B decides to answer the call. When he picks up the handset, his SIP phone sends a 200 (OK) response to indicate that the call has been answered. A and B enter into a conversation. When they are done, they hang-up and the resources are released.

17.7.2 SIP to PSTN Call flow

In case of a SIP to PSTN call let us take the case of party A (Alice) calling party B (Bob) from an IP phone with address 192.168.3.123 to a PSTN phone with phone number 011-31313131 using a user agent from her computer. This is depicted in Fig. 17.9. She dials in 011-31313131. This number gets converted to enum e.164 format, i.e., +13131313110. Last three digits of this are used as the domain name for the SIP server to be searched to route the message to. In this case, the domain name turns out to be 1.1.0, which is the SIP proxy server. Before starting to route this message, the local SIP proxy queries the local database. The database can be an equivalent of HSS (Home Subscriber Subsystem) or HLR. The database will have routing information along with personalization and provisioning information of the user. The proxy finds out if the user has the facility of calling the person or not.

Foreign SIP server on receiving this request, queries its database and finds out that the number belongs to the PSTN. Then the SIP server, triggers a call agent/*media gateway controller*. The type of trigger/event will depend upon the type of message the SIP server has received. Depending on the type of trigger received by the call agent, it will either contact the signaling or media gateway. If the trigger is corresponding to some call establishment/maintenance/tearing then the signaling gateway will be queried. This signaling gateway is responsible for converting the message in to PSTN understandable format. On reaching the destination SSP (Service Switching Point), this request will get replied to and travel back to the originating MGC. The MGC's Signaling Gateway again converts this PSTN signal to SIP signal in the opposite direction.

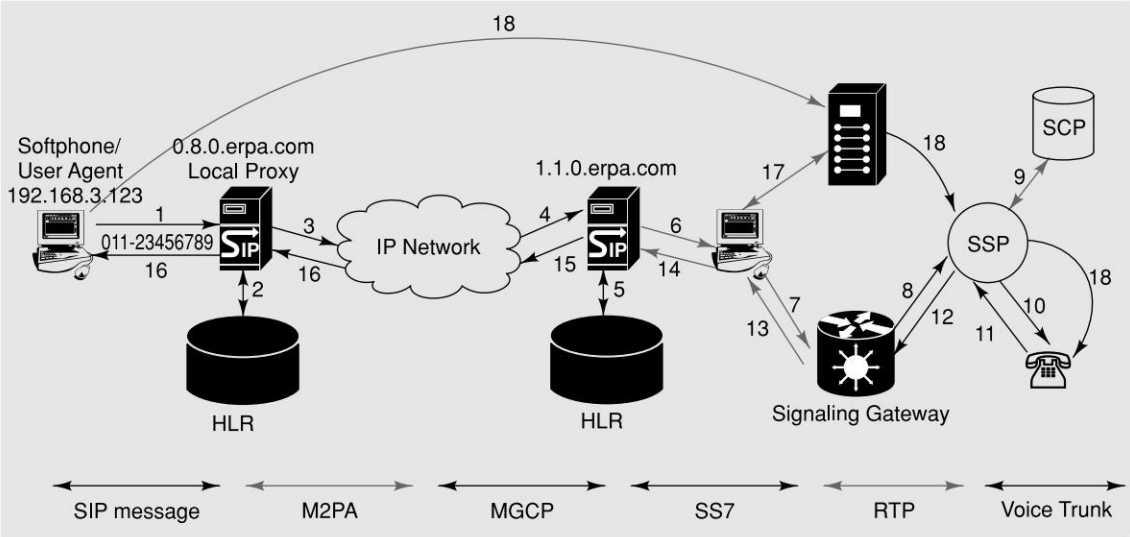


Figure 17.9 SIP to PSTN Call Flow

The Call Agent/Media Gateway control also contacts Media Gateway using MGCP for establishing an RTP path and for Codec Conversion. So a RTP path is established between caller (Party A) in IP and callee (Party B) in PSTN (through media gateway).

17.7.3 PSTN to IP Call flow

In this section we look at the PSTN to IP call flow (Fig. 17.10). When Party B (PSTN) tries to call party A (IP), assume that party A has a number which can be dialled from a normal DTMF telephone. Now as the idea is to have maximum traffic transferred on IP network, the SSP to which the calling party is linked or associated should transfer SS#7 messages on to the IP network via the signaling gateway associated with it. But for this every SSP should have a Media/Signaling gateway linked to itself. However, this doesn't seem practical. To make the design a little easier and practical, in case the SSP doesn't have a media/signaling gateway associated with it, it handles those messages to some SSP which has a media/signaling gateway.

Let us suppose that the number dialed by the person is 011-23567890, and then at the first signaling gateway, the SIP proxy would know that the destination SIP proxy is in Delhi (following the same convention as the earlier case). So as the normal SIP message routes through the IP network, this message also routes till it reaches the SIP proxy at Delhi.

After the SIP invitation has been received by the destination, acknowledgement follows the same way back. Now an RTP path is established between the source and destination and voice packets flow by. Here again protocol to control Media Gateway is MGCP and the protocol to carry SIP messages is M2PA.

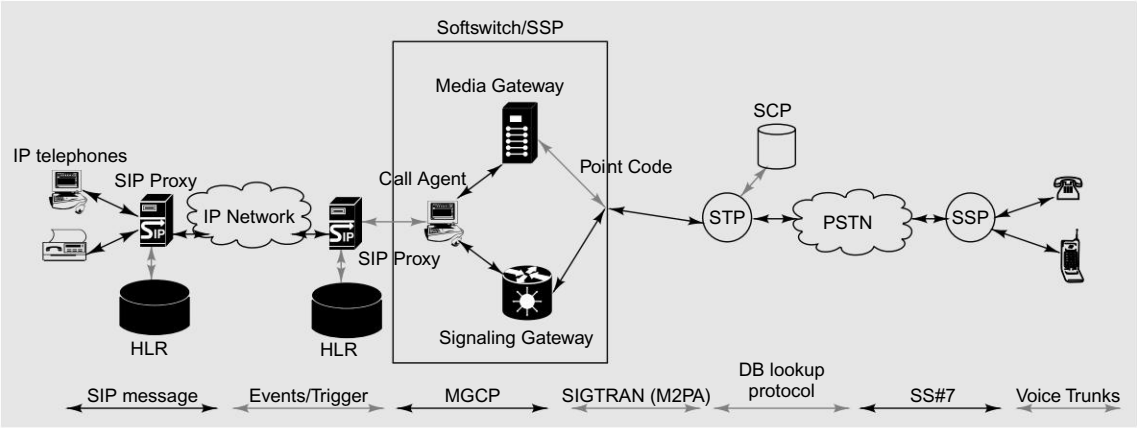


Figure 17.10 PSTN to SIP Call Flow

17.8 Voice Over IP Applications

In addition to voice communication, PSTN offers many other services like Caller Identification, 800 number translation, 900 premium services, and various Intelligent Networks services. These types of services are not offered by Internet telephony. If these Intelligent Network (IN) services are integrated within the Internet telephony then a whole new range of services can be offered. Therefore, to create new applications Internet telephony needs to be programmed.

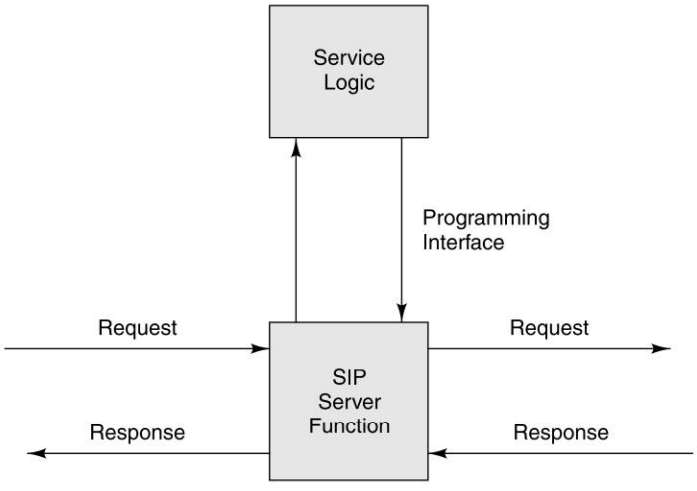


Figure 17.11 Model for Programming SIP Services

The key to programming Internet telephony services with SIP is adding logic that controls behavior at each of the system elements. In a SIP proxy server, this logic determines how the packet should be formatted, where the requests will be proxied to, how the responses should be processed, so on and so forth. A simple service such as call forwarding based on time of day will require logic in the SIP server to obtain the time when a call setup request arrives. Based on the time, the proxy will forward a request to a particular destination. The logic can direct the server's actions based on many parameters defined by the subscriber. These could be time of day, caller, call subject, session type, call urgency, location of the subscriber, media composition, data obtained from directories, data obtained from Web pages, etc. Based on some conditions, the logic may also instruct the server to generate new requests or responses. The basic model for providing logic for SIP services is shown in Figure 17.11.

17.8.1 SIP CGI

In the Web, CGI (Computer Gateway Interface) is the most flexible mechanism for creating dynamic content. As SIP's functionality is similar to that of HTTP, some researchers applied CGI interfaces to Internet telephony. This is because CGI possesses the following characteristics:

- Language Independence: CGI works with Perl, C, VisualBasic, Tcl, and many other languages.
- Exposure of All Headers: CGI exposes the application to all header content in an HTTP request through environment variables. This approach can be directly applied to SIP because its methods of encoding messages are similar to those in HTTP.
- Creation of Responses: CGI can control all aspects of a response, including headers, response codes, and reason phrases, as well as content. This flexibility helps in SIP where services are defined largely through response headers.
- Access to Any Resources: The CGI script is an ideal starting point for creating IP telephony services because it is a general-purpose program whose flexible interface can use existing APIs to let the service logic access an unlimited set of network services.
- Component Reuse: Much of CGI components provide easy reading of environment variables and easy parsing and generation of header fields. As SIP reuses the basic syntax of HTTP, these tools are immediately available to SIP CGI.
- Environment Familiarity: Many Web programmers are familiar with CGI.
- Easy Extensibility: Because CGI is an interface rather than a language, it is easy to extend and reapply to other protocols, such as SIP.

17.8.2 Call Processing Language

While SIP CGI is an ideal service creation tool for trusted users, it is too flexible for service creation by untrusted users. Therefore a new scripting language has been developed. This is called the Call Processing Language (CPL), which allows untrusted users to define services. Users can upload CPL scripts to network servers. The logic can be read in and verified and the service instantiated instantly.

17.9 IP MULTIMEDIA SUBSYSTEM (IMS)

IP Multimedia Subsystem (IMS) is an emerging international standard, which looks at total convergence of voice and multimedia. Some literatures even refer IMS as "All IP network". IMS was specified by the Third Generation Partnership Project (3GPP/3GPP2) and is now being embraced by other standards bodies including ETSI. It specifies interoperability and roaming. It provides bearer control, charging and security. It is well integrated with existing voice and data networks. This makes IMS a key enabler for fixed-mobile- multimedia convergence with value-based charging. For a normal user, IMS-based services enable person-to-person and person-to-content communications in a variety of modes that include traditional telephony services and non-telephony services such as instant messaging, unified messaging, push-to-talk, video streaming, multimedia messaging, text, fax, pictures and video, or any combination of these in a personalized and controlled way. For a network operator, IMS takes the concept of layered architecture one step further by defining a horizontal architecture, where service enablers and common functions can be reused for multiple applications. IMS will meet the following requirements:

- Separation of the access and transport layer from the services layer.
- Consistent mechanisms for authenticating and billing end users.
- Consistent mechanisms for sharing user profile information across services.
- Session management across multiple real-time communication services.
- Compatibility with existing Advanced Intelligent Networks (AIN) services (Toll free 800, Premium Service 900, Calling name, Local Number Portability, Customized Applications for Mobile Networks Enhanced Logic (CAMEL), American National Standards Institute-41, etc.).
- Coarse-grained bearer QoS control.
- Transparent interworking with legacy TDM networks, which will support numbering plans, progress tones, etc.
- Convergence of wireline and wireless services.
- Authentication, user management and charging based on existing 2G functions.
- Blending of voice and real-time communications services including Instant Messaging.
- Consistent and blended graphical user interface.
- Open standard interfaces and APIs for new services by service providers and third parties.

IMS will offer following converged services:

- Advanced service components like Presence, Instant Messaging, Push-to-talk over Cellular (PoC), etc.
- Multimedia call (like VoIP, PoC, etc.).
- Multimedia messaging (MMS, etc.).
- Group services (Collaboration, Buddies list, PoC, etc.).
- Infotainment (Audio Visual distribution, Interactive Gaming, etc.).

The IMS services architecture is a unified architecture that supports a wide range of services enabled by the flexibility of SIP. As shown in Figure 17.12, the IMS architecture is a collection of logical horizontal functions, which can be divided into three major layers:

- Communication Layer
- Session Control Layer
- Applications or Service Layer

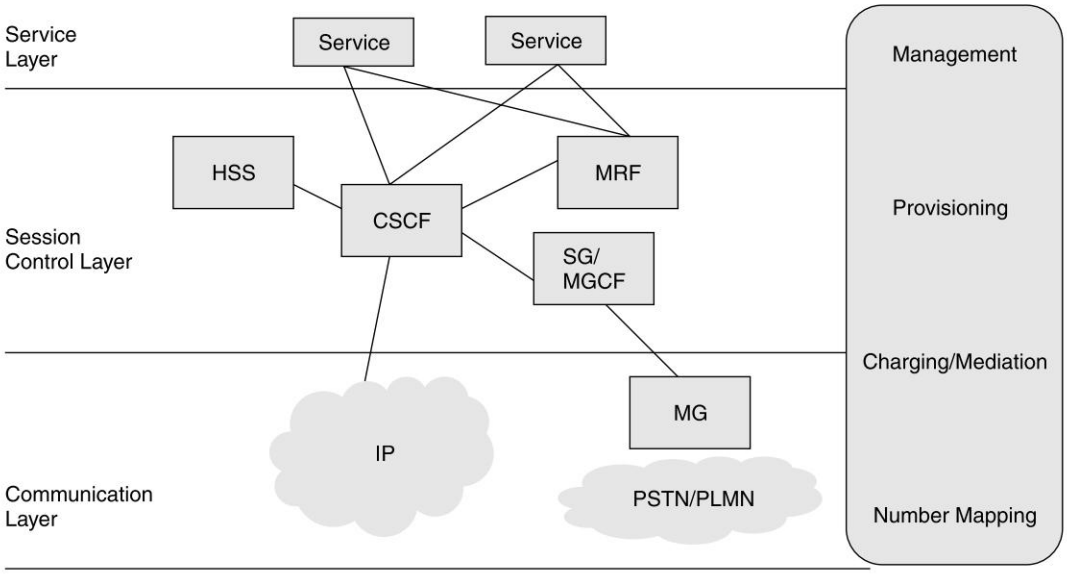


Figure 17.12 Simplified Architecture of IMS

In the IMS architecture, Call Session Control Function (CSCF) is used in the session control layer. The session control layer contains the Call Session Control Function (CSCF), which provides the registration of the endpoints and routing of the SIP signaling messages to the appropriate application server. The session control layer includes the Home Subscriber Server (HSS) database that maintains the unique service profile for each end user. The end user's service profile stores all of the user service information and preferences in a central location. This includes an end user's current registration information (i.e., IP address), roaming information, telephony services (i.e., call forwarding information), instant messaging service information (i.e., buddies list), voice mail box options (i.e., greetings), etc. Media resource function (MRF) includes functions related to conference booking and floor management. Conference booking provides booking information like start time, duration, list of participants, etc. Through floor control, end users (participants, or chairman of the conference) can influence floor and provide information to the MRF Controller on how incoming media streams should be mixed and distributed. Please refer to Chapter 19 for IMS in detail.

17.10 Mobile VoIP

Mobility and wireless issues have not been considered till date in detail within the scope of IP telephony. In particular, while mobility has been considered to some extent within SIP, it has not been addressed comprehensively within H.323 or Megaco either. In a VoIP application, mobility may include terminal mobility, user mobility, and service mobility. Terminal mobility refers to the ability for a terminal to change physical location while the ongoing voice connection is maintained. User mobility is defined as the ability for communications of the mobile user irrespective of the terminal type in use. Service mobility is the ability of a user to access a particular service independent of user and terminal mobility.

In the context of VoIP, roaming refers to the ability that connectivity between endpoints are assured even while one or both endpoints are moving. Such reachability can either be discrete or continuous. Discrete reachability is service portability, implying no online reachability and communications taking place while moving. Continuous reachability is the service mobility allowing seamless communication continuity while roaming. Obviously, mobility encompasses portability, and requires the ongoing connection to be handed off when a mobile terminal is on the move. Upon crossing a region boundary, a handoff must be initiated; otherwise, the connection is broken and the ongoing conversation is interrupted. Mobility management is the key to enabling mobile Internet telephony service over connectionless IP networks. The core operations include registration, call establishment, roaming, and handoff. In H.323 or SIP there is no provision for support for roaming or handoff handling, and called location tracking and location update.

When an H.323 terminal moves across different subnets during a call, it causes the IP address to change. This results in the ongoing connection to be broken. In the intrazone handoff shown in Fig. 17.13 (a), both subnets are under the management of the same GK (Gatekeeper), whereas in the interzone roaming shown in Figure 17.13 (b), they are under the management of different GKs within the same ITG (Internet Telephony Gateway).

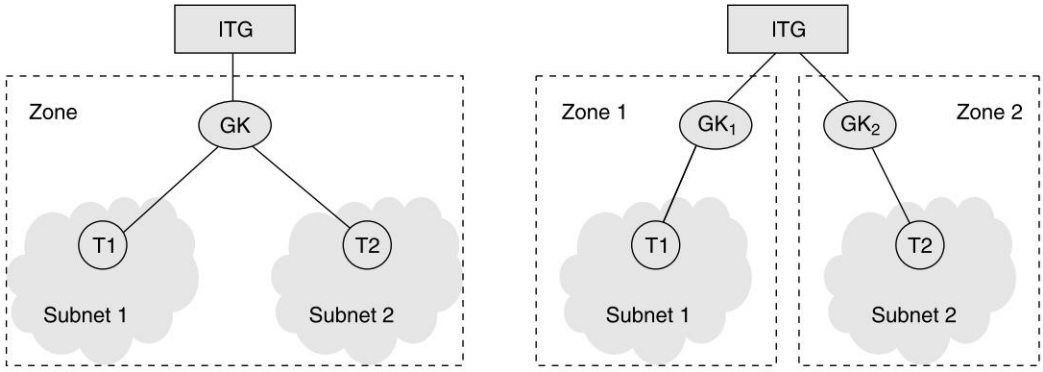


Figure 17.13 Handoff/Roaming Scenario (a) Intra-zone Handoff, (b) Inter-zone

The existing activities in the international standards bodies toward VoIP mobility include efforts made by ETSI TiPHON (Telecommunications and Internet Protocol Harmonization over Networks) Working Group 7, and ITU-T Study Group 16 H.323 Mobile Annex. TiPHON's mandate is to develop specifications (protocol profiles and test suites) to enable end-to-end telephony and multimedia communications services over Next Generation Networks. The standards will cover VoIP mobility including roaming and handoffs.

17.11 Voice over Wireless LAN

Voice over Wireless LAN (VoWLAN) is designed to provide voice over wireless local area network. VoWLAN can be considered as accessing VoIP over wireless device. The principles of VoIP

operation remain the same except with the additional responsibility of handling its wireless nature. Effective QoS provision and robust security features are two of the important challenges to be addressed in case of VoWLAN. The application domains of VoWLAN are enterprise premises, campuses, homes and public hotspots. As many of the just mentioned places already have existing WLAN/LAN infrastructure, the additional voice capability over such networks can be provided with lower costs and in an effective manner. Other terms to denote the concept of VoWLAN are wireless VoIP (wVoIP) and Voice over Wi-Fi (VoFi). Figure 17.14 briefly depicts the VoWLAN architecture.

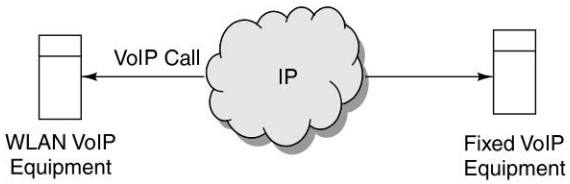


Figure 17.14 Brief Architecture of VoWLAN

The major challenges in VoWLAN deployment can be summarized as under:

1. Handling real-time data can be tedious for the infrastructure which was primarily designed for non real-time data which can tolerate delay latency, re-transmissions, packets drops and jitter in packet stream to some extent.
2. Roaming and handovers need to be addressed keeping in mind the ever increasing mobility of users. Procedures should also address the interworking functions of handling calls in cellular or fixed voice networks domain.
3. The signaling mechanism should be the same for both fixed VoIP and wireless VoIP network infrastructure because VoWLAN equipments would ultimately be integrated with the existing infrastructure of any deployment, be it enterprise, university, home, etc.
4. VoWLAN network planning should be such that there are no dead spaces in the entire intended coverage facility. Commonly ignored areas like staircases, restrooms, elevators, etc., should not be left out for successful VoWLAN deployment.
5. Access control methods and other security mechanisms must be executed in order to provide security for the open air traveling data.

REFERENCES/FURTHER READING

1. "All-IP Core Network Multimedia Domain, IP Multimedia Subsystem–Stage 2", 3GPP2 *X.S0013-002-0 1, Version 1.0 2,* December 2003.
2. "Framework Architecture for Signaling Transport", IETF RFC2719, October 1999, www.ietf.org.
3. Glitho Roch H., Ferhat Khendek and Alessandro De Marco "Creating Value Added Services in Internet Telephony: An Overview and a Case Study on a High-Level Service Creation

Environment", *IEEE Transaction on Systems, Man, and Cybernatics–Part C: Applications and Reviews*, Vol. 33, No. 4, November 2003.
4. Gurbani Vijay K. and Xian-He Sun, "Terminating Telephony Services on the Internet 2004". *IEEE/ACM Transactions on Networking,* Vol. 12, No. 4, p571, August 2004.
5. "Instant Messaging / Presence Protocol Requirements", RFC2779, February 2000, www.ietf.org.
6. Liao Wanjiun, "Mobile Internet Telephony: Mobility Extension to H.323" *IEEE Transaction on Vehicular Technology*, Vol. 50, No. 6, November 2001, p1403.
7. "Megaco IP Phone Media Gateway Application Profile", RFC3054, January 2001, www.ietf.org.
8. *Performance Technologies Learning Center,* http://www.pt.com/learning.html.
9. "Real-Time Control Protocol (RTCP) attribute in Session Description Protocol (SDP)", RFC3605, October 2003, www.ietf.org.
10. "Real-Time Streaming Protocol (RTSP)", RFC 2326, April 1998, www.ietf.org.
11. "Resource ReSerVation Protocol (RSVP)", RFC2205, September 1997, www.ietf.org.
12. Rosenberg Jonathan, Jonathan Lennox and Henning Schulzrinne, "Programming Internet Telephony Services", *IEEE Network*, May/June 1999, p42.
13. "RTP: A Transport Protocol for Real-Time Applications," *IETF RFC1889*, January 1996, www.ietf.org.
14. "Session Initiation Protocol (SIP)-Specific Event Notification", *RFC3265*, June 2002, www.ietf.org.
15. "SIP: Session Initiation Protocol, RFC3261", June 2002, www.ietf.org.
16. *SIP Server Technical Overview; SIP Server version 2.0*, April, 2004, http://www.radvision.com.
17. *Source for Open Source Communication:* www.Vovida.org.
18. "Stream Control Transmission Protocol", IETF RFC 2960, October 2000, www.ietf.org.
19. "The COPS (Common Open Policy Service) Protocol", RFC2748, January 2000, www.ietf.org.
20. *Voice over IP*, www.protocol.com.

REVIEW QUESTIONS

Q1: What is SIP? How does SIP handle call setup and teardown?
Q2: What are the basic functional differences between H.323 and SIP?
Q3: What are the different real-time protocols available for real-time data transmission over IP ? Explain each of them.
Q4: What is convergence between PSTN and IP networks? Discuss its implications.
Q5: What are the different elements in a VoIP architecture? Discuss the role of such elements.
Q6: What is Sigtran? What is it used for?
Q7: Describe Voice over IP applications.

Q8: What is SIP CGI?

Q9: Explain each of the following:

(a) Session Announcement Protocol
(b) Session Description Protocol
(c) H.248
(d) SCTP
(e) SIP to SIP call flow
(f) SIP to PSTN call flow
(g) PTN to IP call flow
(h) Mobile VoIP

CHAPTER 18

Multimedia

18.1 INTRODUCTION

Multimedia (*multi* meaning many and *media* meaning means of storing and/or communication any information) is a type of medium which is created by using more than one conventional medium. It is any blend of text, images, audio, animation and video delivered and manipulated electronically. One example of multimedia could be the collection of text, images and animation in a Computer Based Training (CBT) for lessons in Geography. For a person surfing the World Wide Web, it can be web pages with a diverse collection of text, images, graphics, audio, animation, and embedded video. For a video game kiosk (like a Windows X station), it is a richly simulated environment inclusive of situational audio and video along with textual instructions. We come across multimedia elements in a number of situations. Our lives are so wrapped in digital media that we rarely take cognizance that a particular blend is a combination of a number of individual medium of information transfer.

Information and Communication Technologies (ICTs) have broadened the scope of information delivery and control. Through ICT it is now possible to let people experience simultaneous encounters with more than one kind of medium. When persons are given interactive control to manipulate how such content is delivered, they can be enthralled. When an individual is permitted to control multimedia delivery—when, what, which and how content is presented, it is called interactive multimedia. An example of this can be a video game. When an end user is presented with inter and intra linked multimedia content through which he can navigate, multimedia comes to be known as hypermedia. A piece of multimedia content is called linear if it has a pre-defined beginning, course and end which cannot be altered. Watching a movie on television is an example of linear multimedia. While on the other hand, hypermedia is an example of nonlinear multimedia because the user can navigate through the multimedia content at his own will.

The various types of media which form multimedia are text, images, audio, animation, moving images, and video. Text is one of the oldest and prominent types of information carrier. Text

delivers information which can have an effective meaning. Even to this day, most of the crucial data and information is maintained in textual form only. With the advent of Web, information in the textual format has had the effect of being exploded. The most popular messaging media, SMS (described in Chapter 6) is an example of text.

Sound or audio is another form of medium very commonly included as a part of any multimedia presentation. It is one of the most sensual elements of any multimedia content, the effect of occurrence of which can be perceived over large distances, unlike other media. The greatest advantage of audio is that it does not depend on the literary skill of an individual. People who cannot read or write can use voice to communicate and express their thoughts.

A picture is worth a thousand words. Also, psychologists claim that human being get 80% of information through sight. Therefore, it makes sense to use image or visuals as another media. An image might mean anything from still photographs to graphs to hand drawings and portraits. Images add up to the quality of a multimedia presentation and can communicate more than the text can in one shot. MMS or Multimedia Messaging (described in Chapter 8) is an example of such combination of text and image. Another interesting media is animation; it makes static images come alive. Animation is a phenomenon through which an object can move across the screen. Motion picture or video is one of the most engaging media.

Apart from text, other forms of media can undergo lossy compression and decompression techniques and tolerate loss to some extent during information transfer. Among all forms of media, video is the one which is the most resource-thirsty in terms of storage, transfer and even manipulation. Animation and video are possible due to persistence of vision and a psychological phenomenon called "phi". In the two, a series of images are altered slightly and very quickly, one after another, which seemingly gives a visual illusion of movement.

In Chapter 5 we have seen how voice is digitized in PCM (Pulse Code Modulation)—1 second of voice needs 64K bits. On contrast, a black-and-white image of size 640*480 pixel, with 8 bits (0-255) grayscale, will occupy 307.2KBytes. If the same image is stored in color with 8 bits per basic color, the image will occupy 7372800 bits (921.6KBytes). Now if need to have a moving object of 24 frames/second, a 1 second video will need 176947200 bits (22.1184 MBytes). Therefore, if we want to send 1 second of voice we need a minimum 64K bits/second bandwidth; for image we need 177Mbits/second. In this chapter we will discuss more about these challenges and how to solve them and have multimedia transmission.

18.2 WHY MULTIMEDIA

Multimedia is a new technological phenomenon proliferating on the advancements of faster processing and communication technologies; cheaper and physically shrinking storage technologies; and greater content creation, capability, and awareness. We enumerate the major benefits of using multimedia technologies in the following points:

1. Multiple and interactive media helps in easier grasp of any subject.
2. Digitization helps revolutionizing music, movies, games and virtual reality; most important of all, digitized objects do not need electro-mechanical devices for playing; like we do not need a turntable to play a record or a CD.

3. Digitized objects have other advantages that make it very attractive. It consumes less power, can take advantage of increasing processing power, is easy to compress, easy to store and archive. It can be edited and altered very easily.
4. There remains huge potential for new applications and services based on multimedia. This implies more benefit for markets and avenues for improving our lives (in terms of learning, work and leisure).
5. Smarter devices with lesser memory footprints, easy to use and understand interfaces offer quick network connectivity.
6. Digitization of nearly every form and every chunk of data.
7. Technologies for organizing, browsing, storing and retrieving multimedia content is improving at a very rapid rate. Also, the same storage can be used over and over again.

18.2.1 Major Application Arenas of Multimedia

The application areas for multimedia content are many. Broadly, these can be classified as under:

1. Residential services such as Video on Demand (VoD), Home shopping, Conferencing, etc.
2. Business services including corporate learning, e-business, simulations, conferencing, etc.
3. Education such as distance learning, self learning, digital libraries, etc.
4. Scientific research such as scientific visualization, prototyping, simulations, medicines, etc.
5. Entertainment and leisure activities.
6. Web applications.

Figure 18.1 describes a generic multimedia system. The figure shows how the multimedia objects are acquired, processed, stored, and transmitted. It also shows the presentation of multimedia at the user site. However, the figure just gives an overview of any multimedia system and mechanisms shown are not to be taken as exhaustive.

18.3 Compression and Decompression

We know that to transmit a video of 1 second, we need 177 Mbits/second bandwidth, which is commonly not available. Therefore, the only choice we have, is to compress the digitized media. Following are the major motivations behind compression of multimedia content:

1. Compressed audio and video media consumes lesser amount of storage.
2. Compressed media requires lesser bandwidth for transmission.
3. For error-corrections, redundancy can be intelligently built-in.

For example, let us take a picture taken by a camera phone with 2 Mega pixels. Higher number of pixels means higher resolution of the picture. One pixel is again mapped into a number of bits as, higher number of bits per pixel allows better color. Now, if we say that we represent each colour with 16 bits (2 bytes) then such an image would consume 2000000 * 3 * 16 bits (12 MBytes) for storage. Accordingly, for a 64 Kbits/second transmission link, the time needed to transmit it would be 1500 seconds or 25 minutes. If we are able to compress the given image by a factor of 10, then we can save storage and transmission costs by 90 percent.

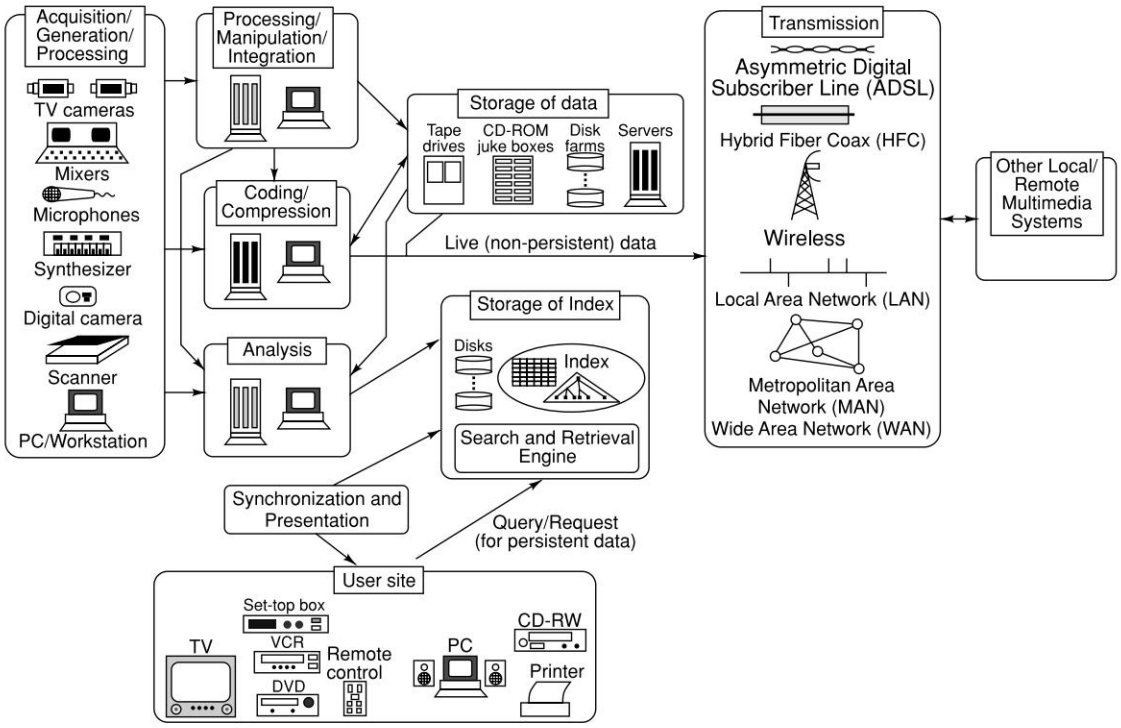


Figure 18.1 General Overview of Multimedia System

The compression process a undergoes a number of steps. In case of audio or voice, the sound is a set of continuously varying analog signals recorded from speech or musical composition. It is then converted into digital signal ready for transmission through a series of steps enumerated below:

1. The samples of audio signal are taken at a pre-defined fixed rate. The sampling rate can be any arbitrary real number. However, according to the Nyquist theorem, this rate must be at least double the highest frequency in the signal.
2. Each of the samples is then encoded into one of the pre-defined finite number of values called quantization values. Each quantization value can be represented by a fixed number of bits.
3. This bit stream then undergoes a process of compression where the total number of bits required to represent the original audio is lesser.
4. The bits of all the samples are concatenated together to be representative of the digital signal of that piece of audio.

For example, if we encode audio analog signals at a rate of 8000 samples per second with each sample represented by 8 bits, post quantization the equivalent digital signal shall have a rate of 64,000 bits per second. The more the sampling rate and number of bits for each quantization value, the better can be the perceived sound quality. Thus, a clear tradeoff is evident between the storage and bandwidth requirement and quality of audio. Let us take an example of video

compression. A video is a sequence of images played out at a constant rate with 24 to 30 images displayed within a second. As we are aware, every pixel can be encoded into a number of bits representing a definite luminance and color. For a five-minute video with a display rate of 30 still images for a display size of 1024 * 768 pixels (a computer screen) per second and 32 bits per color, the storage requirement for 3 colors (red, green, and blue) would be 300 * 30 * 1024 * 768 * 32 * 3 bits or 679477248000 bits (84935GByte).

The process of reducing the space for storing a media is called compression; whereas the reverse process restoring the original media from a compressed media is called decompression. There are two types of compression. These are called lossless compression and lossy compression. Lossless or lossy compression can be either spatial or temporal.

18.3.1 Spatial Compression

There are two types of redundancies in video which are liable to be dropped in case of lossy compression. The first one is spatial redundancy. It is, actually, the redundancy in space within a given image. An image mostly containing one or two colours, e.g., a picture of a sunset. This is an ideal candidate for compression.

18.3.2 Temporal Compression

In a video there is temporal redundancy as well. This refers to the recurrence of repetition in time from one image to subsequent images. For example, if most of the part of second image is same as in the first one then there is hardly any motivation to re-encode the second image fully. The difference between the two images can be encoded instead of the entire second image. For example, in a movie, the actor and actress are moving, but the background scenario like mountains or buildings remain the same from one image to another; therefore, in crude terms, in the second or subsequent images these can be dropped and copied from the first one.

18.3.3 Lossless Compression

In a lossless compression, the original media content is compressed in such a fashion that the output after the decompression process restores the original content without any change. The order of compression that we can achieve in lossless compression is much less compared to its lossy counterpart. Lossless compression can be used for any type of content. However, we use this mainly for text or data; for example, someone doing electronic transfer of 16777216 dollars. Loss of this data may reduce the money to zero. Also, in case of text even a little loss may change the meaning of the text.

18.3.4 Lossy Compression

In lossy compression, the original media content is compressed in such a fashion that part of it is permanently lost. Therefore, when the compressed content is uncompressed, the content appears to be the same, in other words it cannot be perceivably detected by humans. For example, there are many frequencies within the audible frequency spectrum that cannot be perceived by human beings; therefore dropping some of the high frequency components in a voice may not disturb the

intelligibility of the speech. Likewise, we drop some part of video signal to get a higher compression factor. Lossy compression ranges from high quality media with undetectable loss down to massively degraded experiences. The developers of lossy compression technology continually strive to improve the ratio of quality to bit rate or compression.

18.3.5 Lossy Lossless Compression

This is a combination of lossy and lossless compression. This type of compression technique is used in medical imaging. In medical imaging high level of resolution is necessary for correct diagnosis. If the data is to undergo further processing, the repeated application of lossy codecs shall almost certainly degrade the quality of the edited file. Therefore, in medical imaging lossy lossless compression is used. One specific example is ultrasound images; where, the fetus is compressed using lossless techniques; though, the other part of the video other than the fetus is compressed by using lossy compression.

18.4 Coder and Decoder (CODEC)

A codec is a device or program which is capable of performing encoding and decoding on a digital data stream. The word 'codec' is essentially a combination of **co**de and **dec**ode; that in other words signifies Compression and Decompression respectively. In early days of communications, the term codec was primarily used for hardware that encoded and decoded analog signals. But in recent years, codec is applied to a class of software for converting among digital signal formats, and including compander functions.

Codecs (in the software sense) encodes a stream or signal for transmission, storage or encryption and decodes it for viewing or editing. Codecs are often used in video-conferencing and streaming multimedia applications. A video camera's analog-to-digital converter (ADC) converts its analog signals into digital signals, which are then passed through a video compressor for digital transmission or storage. A receiving device then runs the signal through a video de-compressor, then a digital-to-analog converter (DAC) for analog display. The raw encoded form of audio and video data is often called essence, to distinguish it from the metadata information that together make up the information content of the stream and any wrapper data that is then added to aid access to or improve the quality of the stream.

Most of the codecs functions and algorithms we encounter today, use lossy compression technique. There are lossless codecs too, but for most purposes the slight increase in quality is not worth the increase in data size, which is often considerable. Using more than one codec or encoding scheme while creating a finished product can degrade quality significantly. Many codecs are designed to emphasize certain aspects of the media to be encoded. For example, a digital video (using a DV codec) of a sports event, such as baseball or soccer, needs to encode motion well, but not necessarily exact colors, while a video of an art exhibit needs to perform well encoding color and surface texture. There are many codecs ranging from those downloadable for free to ones costing thousands of dollars.

Generally, we will find that codecs can be functionally classified into two main categories: audio and video. However, there are few codecs for text and data too. Again, as mentioned previously, codecs can be proprietary or interoperable.

18.4.1 Audio Codecs

An audio codec is a software or a device that compresses/decompresses digital audio data. The codec is versatile enough to handle pre-defined audio file format or streaming audio format. Normally, codecs are implemented as libraries which interface with one or more multimedia players, such as Winamp or Windows Media Player. However, contextually, audio codec can refer to a hardware implementation with definite inputs and outputs, controlled through software. So, in such a scenario, the phrase audio codec refers to the device encoding an analog audio signal to a digital audio signal and vice-versa. The following is a classified list of popular audio codecs:

Codecs for Non-compression Formats

The following formats are derivatives of Pulse Code Modulation (PCM), except the last one. All of the following formats are non-compression formats. Non-compression format means that the file content is free from having been through the cycle of compression and decompression.

- Audio Interchange File Format (AIFF)
- Resource Interchange File Format (RIFF)
- Microsoft's WAV format
- Linear Pulse Code Modulation (LPCM)
- Pulse-amplitude modulation (PAM)

Codecs for Lossless Data Compression

Here, we enumerate the popular codecs for lossless data compression.

- Apple Lossless Audio Codec (ALAC)
- Audio Lossless Coding (MPEG-4 ALS)
- Direct Stream Transfer (DST)
- DTS-HD Master Audio Dolby TrueHD
- Free Lossless Audio Codec (FLAC)
- Lossless Audio (LA)
- Lossless Predictive Audio Compression (LPAC)
- Lossless Transform Audio Compression (LTAC)
- MPEG-4 Audio Lossless Coding (MPEG-4 ALS)
- OptimFROG (OFR)
- RealAudio Lossless
- RK Audio (RKAU)
- Shorten (SHN)
- True Audio (TTA)
- WavPack (WV)
- Windows Media Audio 9 Lossless

Codecs for Lossy Data Compression

The following are the noted codecs for lossy data compression, in general:

- Adaptive Differential (or Delta) PCM (ADPCM)
- ADX
- Adaptive Rate-Distortion Optimised sound codeR (ARDOR)

- Adaptive Transform Acoustic Coding (ATRAC)
- Dolby Digital (A/52 & AC3)
- DTS Coherent Acoustics
- Impala FORscene audio codec
- MPEG audio
- Musepack
- Perceptual Audio Coding
- QDesign
- TwinVQ
- Vorbis
- Windows Media Audio (WMA)

Following the above list, there is a collection of few more codecs but, in the lower bit rate side, which is optimized for speech/voice. Below is the list.

- Advanced Multi-Band Excitation (AMBE)
- Algebraic Code Excited Linear Prediction (ACELP)
- Code Excited Linear Prediction (CELP)
- Continuously variable slope delta modulation (CVSD)
- Digital Speech Standard (DSS)
- Enhanced Variable Rate Codec (EVRC)
- FS-1015 (LPC-10)
- FS-1016 (CELP)
- ITU standards: G.7xx series
- GSM codecs
- Harmonic Vector Excitation Coding (HVXC)
- Internet Low Bit Rate Codec (iLBC)
- Improved Multi-Band Excitation (IMBE)
- internet Speech Audio Codec (iSAC)
- Mixed Excitation Linear Prediction (MELP)
- QCELP
- Relaxed Code Excited Linear Prediction (RCELP)
- Selectable Mode Vocoder (SMV)
- Speex
- Triple Rate CODER (TRC)
- Vector Sum Excited Linear Prediction (VSELP)

The use of digital samples to represent audio data is subject to some fundamental limitations, irrespective of the class and algorithm of the codec. The bandwidth is limited by the Nyquist–Shannon Sampling Theorem so that the "highest audio frequency that can be reconstructed from the digital data is half the sample frequency". The dynamic range is limited by quantization noise which is half the weight of the least significant bit of each sample. A perfect linear codec (which is usually considered lossless) suffers from these types of signal degradation. Codecs, which are considered lossy, will suffer from the similar types of signal degradation plus some additional lost signal which varies from codec to codec.

18.4.2 Video Codecs

A video codec is a software/device that enables video compression/decompression for digital video. The compression usually employs lossy data compression. In the beginning, video was stored as an analog signal on magnetic tape. When the compact disc entered the market as a digital replacement for analog audio, it also became feasible to begin storing and using video in digital form. However, there exists a complex tradeoff between the video quality, the quantity of the data needed to represent it, the complexity of the encoding and decoding algorithms, robustness to data losses and errors, ease of editing, random access and end-to-end transmission delay.

Video codec designs are often standardized for interoperability, which is necessary in today's Internetworked world. One other reason is because the way films work is different from TV. Also in TV, we use luma (for lumination in gray scale) and chroma (for chromatic) function compared to that in computer display where we use the Red Green Blue (RGB) pattern. The human eye can perceive the difference between the precision of colors and the precision of contours. This brings into light the fact that images should be first converted into a color model separating the luminance from the chromatic (that is color) information before sub-sampling the chromatic planes so as to retain precision of the luminance plane with more information bits. Sub-sampling is a mechanism to reduce an image to smaller size by decreasing the information content in that image. Because of the design of analog video signals (which represent luma and chroma) and color information, a usual step in image compression in codec design is to represent and store the image in a YCbCr color space. YCbCr represents color as brightness and two color difference signals, where Y is the brightness (luma), Cb is blue minus luma (B-Y) and Cr is red minus luma (R-Y). The ITU-R BT.601 international standard defines both YCbCr and RGB color spaces. DVDs, digital TV, Video CDs, and digital camcorders (MiniDV, DV, Digital Betacam, etc.) is coded in YCbCr. The conversion to YCbCr provides two benefits:

1. It improves compressibility by providing de-correlation of the color signals.
2. It separates the luma signal, which is perceptually much more important, from the chroma signal (which can be represented at lower resolution to achieve better data compression ratio).

However, different codecs will use different chroma sub-sampling ratios according to their compression needs. Video compression schemes for Web and Digital Video Disc (DVD) make use of a 4:2:0 color sampling pattern. Various professional video codecs designed to function at much higher bit rates and to record a greater amount of color information for post-production processing, sample in 4:2:2 and 4:4:4 ratios. It is also worth noting that video codecs can operate in RGB space, also. These codecs tend not to sample the red, green, and blue channels in different ratios, since there is no perceptual motivation in doing so.

Figure 18.2 shows Red, Green and Blue (RGB) colors in three dimensions which can be thought of as emanating from either Black or White. It is noteworthy to see the colors Cyan, Magenta and Yellow being formed by varying RGB content.

It is noteworthy to include that some amount of spatial and temporal down-sampling may be used to reduce the raw data rate before the basic encoding process. The most popular transform is the 8 * 8 Discrete Cosine Transform (DCT). Codecs which make use of a wavelet transform, too, are becoming popular, especially in camera workflows which involve dealing with image formatting

in motion sequences. The output of the transform is first quantized, then entropy encoding is applied to the quantized values. In DCT, the coefficients are typically scanned using a zig-zag scan order before the entropy coding typically combines a number of consecutive zero-valued quantized coefficients with the value of the next non-zero quantized coefficient into a single symbol. It also has a feature of indicating when all of the remaining quantized coefficient values equal to zero. The entropy coding method typically uses variable-length encoding tables. It is also possible to compress the video in a multiple step process called n-pass encoding, which performs a slower but potentially better quality compression. The decoding process consists of performing an inversion of each stage of the encoding process. One stage that cannot be exactly inverted is the quantization stage. Therefore, a best-effort approximation of inversion is performed. This process finally represents the video image as a set of macro-blocks.

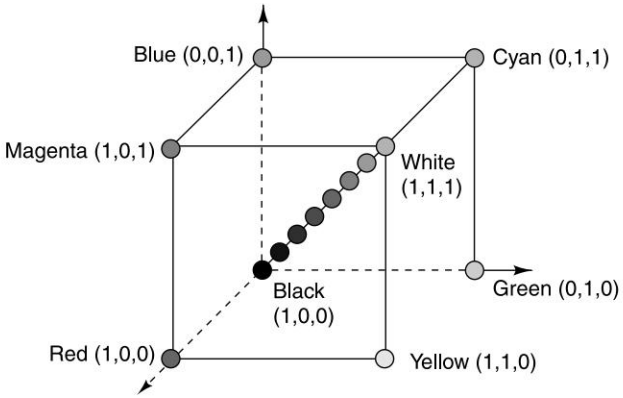


Figure 18.2 Color Dimensions

Codecs for Lossless Video Compression

The following are the popular video codecs for lossless schemes for compression:

- CorePNG
- H.264
- Huffyuv
- Lagarith
- LCL
- MSU Lossless Video Codec
- Tech Smith Screen Capture Codec (TSCC)

Codecs for Lossy Video Compression

The following video codecs are famous when it comes to lossy and more frequently used compression schemes:

- Audio Video Standard (AVS)
- Blackbird FORscene video codec
- Cinepak

- Dirac
- Firebird Original FORscene video codec
- H.261
- H.263
- H.264 (also known as MPEG-4 AVC – Advanced Video Coding)
- Indeo 3/4/5
- KVCD
- MJPEG
- MPEG-1 Video
- MPEG-2 Video
- MPEG-4 ASP (Advanced Simple Profile)
- On2 Technologies VP3, VP6 & VP7
- Pixlet
- RealVideo
- Snow Wavelet Codec
- Tarkin
- Theora
- VC-1 (the primary video codec used by HD DVD and Blu-ray)
- Windows Media Video (WMV)

Nowadays, digital video codecs are found in DVDs, VCDs, emerging satellite and terrestrial broadcast systems, and on the Internet. Online video is encoded in a variety of codecs because codec packs are becoming available. A codec pack is a pre-assembled set of commonly used codecs combined with an installer available as a software package desktop. Encoding media by the public has seen an upsurge with the availability of commonly available and cheap DVD writers.

18.4.3 Open Source Codecs

The following is a list of popular open source codecs. This should not be taken as that no Intellectual Property Rights (IPRs) exist for them. There are regulations existing which control their modification, notification, extension and reproduction.

Video Codecs

- x264 – H.264 (MPEG-4 part 10)
- Xvid – MPEG-4 part 2 codec
- FFmpeg codecs
- Tarkin
- Lagarith
- Theora
- Dirac
- Huffyuv
- OpenAVS

Audio Codecs

- FLAC
- iLBC

- LAME
- Musepack
- Speex
- TTA
- Vorbis
- WavPack
- OpenAVS

There is one popular open source codec for text as well, which is called Writ.

18.5 Popular Compression Techniques

Describing the popular multimedia formats becomes mandatory in the wake of extensive usage of such formats. There are not only such popular formats (or compression/decompression techniques) but their inter-conversion tools, too. Such formats are not only in great usage in the desktop and Internet world, but they are finding increasing acceptance in the wireless world as well.

18.5.1 JPEG

Joint Photographic Experts Group standardizes the JPEG format that is the most commonly used image compression format for photographic images. The Joint Photographic Experts Group committee was formed in 1984 to produce an efficient image compression standard. The group produced its format which was accepted as ISO 10918-1 in 1994. Since then, the group has been producing better compression algorithms while trying to balance the tradeoff between quality and size of image. Most of the images transferred and used on the Internet use this format.

JPEG specifies in detail the whole process of defining how an image is compressed into a stream of bytes and decompressed back into an image as well. Though the mechanism in the JPEG process is lossy, there are variations of JPEG that are lossless. On the application of higher compression ratios, the image quality is severely affected. However, the overall colors and image form might still be identifiable.

The JPEG image compression can be summarized in the following steps:

1. The media will be first converted from Red Green Blue (RGB) space into a color spaceYCbCr. It is possible for lossless JPEG compression schemes not to go through this step while keeping them in RGB space only.
2. Exploiting the fact that human eye can perceptibly differentiate more in brightness rather than the color component, the Cb and Cr components are compressed more than the Y component. This process is called down-sampling. This way the image is reduced to two-thirds or half of its original size.
3. The channels are split into blocks of pixels (like 8 * 8). If the image is not actually divisible into such blocks, then some algorithm is employed to fill up the remaining unfilled pixels. After this, the Y, Cb, and Cr components of each block is converted into a frequency domain representation. One such popular scheme is Discrete Cosine Transform (DCT).

4. The next step is quantization in which most losses in image occur. This exploits the perceptibility of human eye which is better at seeing small differences in brightness over a large area than precisely differentiating the strengths in a high frequency brightness variation.
5. The next step groups similar frequencies together. The well known algorithms for this are Huffman algorithm and Arithmetic coding algorithm. Run Length Encoding (RLE) can also be employed in some cases.

The image decompression is an inverse procedure of the encoding process. Compression ratios like 10:1 is indistinguishable by the human eye. Actually, 100:1 ratio is also possible but then, the image will look distinctly different from the original. The use to which compression-ratio the image will be put is a key factor in determining the compression ratios. In fact, the JPEG encoding does not mandate any particular precision needed for the compressed image. A good example in hand is that in order to support 8-bit precision per pixel output, the de-quantization and inverse frequency transforms are typically processed with at least 14-bit precision in highly optimized decoders. Figure 18.3 demonstrates the perceptible differences between monochrome and grey scale images which are actually the copies of the same image but store just black/white per pixel and varying levels of grey shades per pixel, respectively. Figure 18.4 demonstrates degree of compression and reproduction of copies of the same image with respect to JPEG mechanism.

Monochrome Image

Grey Scale Image

Figure 18.3 Perceptible Difference between Monochrome and Grey Scale Image

When a JPEG image is stored in magnetic media, the most commonly used extensions are .jpeg, .jpg, .jpe, .jfif and .jfi. The JPEG File Interchange Format (JFIF) is an overly simplified version and widely used JPEG format for a wide range of implementations. There is one "Progressive JPEG" format in which data is compressed in multiple passes of progressively higher/lower detail which is useful in an environment of lesser bandwidth. This permits a reasonable preview before all the data has been received.

JPEG 2000 is a wavelet-based image compression standard created by the Joint Photographic Experts Group committee in the year 2000 in order to improve the original DCT based JPEG standard. The extension is .jp2 for ISO/IEC 15444-1 conforming files and .jpx for ISO/IEC 15444-2 ones.

Figure 18.4 Progressive JPEG Compression

18.5.2 GIF

Graphics Interchange Format (GIF) is an 8 bit/pixel image format and supports up to 256 distinct colors from the 24-bit RGB color space. It was introduced by CompuServe in the year 1987. The GIF image compression technique employs lossless compression technique of Lempel Ziv Welch (LZW) algorithm. Hence, it reduces the file size without degrading the perceptible visual quality. GIF supports animations and allows a separate palette of 256 colors for each frame which makes it suitable for images or graphics with solid areas of color. GIF produces animations by storing multiple images in one file, accompanied by control data which produces temporal effects with the stored images.

Nowadays, there are many alternatives existing comparative to GIF format. These are Portable Network Graphics (PNG), Multiple image Network Graphics (MNG) and Scalable Vector Graphics (SVG) formats are the popular replacements. PNG is lossless, bitmapped image format. MNG is open source file format for animated images. SVG is XML specified file format for describing vector graphics. Some of the benefits that helped GIF to be quite popular are due to the following reasons:

1. The interlacing feature in GIF stores image scan lines out of order in such a fashion that even a partially received image is recognizable.
2. GIF format is very suitable for sharp-edged line art with a short range of colors. Hence, it is very useful for small animations and low resolution film clips.
3. GIF89a makes a way around to overcome the limitation of 256 colors.
4. In the early days of GIF, the limitation to 256 colors seemed reasonable because only few people could afford the hardware to display more. Moreover, simple graphics, line drawings, cartoons, and grey-scale photographs normally need fewer than 256 colors.
5. There exist ways to diffuse images by using pixels of two or more different colors to approximate an in-between color, but this occurs with some loss of detail.
6. In Web, GIF is one the most commonly used formats when designing websites.

18.5.3 Windows WAV

Microsoft's .wav is a popular format for audio compression and storage. It can also open and save audio in 8-bit or 16-bit uncompressed PCM format (.wav) from 8 kHz to 48 kHz. The major features of Microsoft's .wav format are:

1. It can record from microphone or headset.
2. For opening .wav files Windows Sound Recorder, the audio codec used by the file must be installed in Audio Compression Manager (ACM).
3. It is inter-compatible with MP3 format.
4. Audio editing is possible like changing bit rate, sampling rate, echo addition, insertion/deletion, etc.

18.5.4 MPEG

Moving Pictures Expert Group (MPEG) is a well known audio and video encoding standard. The group first met in May 1988. Since then the group has been active in evolving many useful and popular audio and video formats. Officially, it bears the designation ISO/IEC JTC1/SC29 WG11.

Until now, MPEG has evolved the following standards:

- MPEG 1
- MPEG 2
- MPEG 3
- MPEG 4
- MPEG 7
- MPEG 21

MPEG 1 defines audio/video encoding and compression standards. MPEG 1 consists of the following parts:

- Part 1: Systems: Synchronization and multiplexing of audio/video.
- Part 2: Video: Compression codec for non-interlaced video signals.
- Part 3: Audio: Compression codec for perceptual coding of audio signals. This has three sub-layers based on the complexity of audio coding:
  - MP1 or MPEG-1 Part 3 Layer 1 (MPEG-1 Audio Layer I),
  - MP2 or MPEG-1 Part 3 Layer 2 (MPEG-1 Audio Layer II), and
  - MP3 or MPEG-1 Part 3 Layer 3 (MPEG-1 Audio Layer III).
- Part 4: Conformance: Procedures for Conformance testing.
- Part 5: Simulation: Reference software.

MPEG 1 video is used by the Video CD format providing quality and performance of a Video Home System (VHS) tape. MPEG 1 video was developed with a goal of achieving acceptable video quality at 1.5 Mbits/second data rates and 29.97 frame per second/25 frame per second resolution. MPEG 1 format is DVD playable. However, MPEG 1 video supports only progressive images.

MPEG 2 is said to be standard for "the generic coding of moving pictures and associated audio information". MPEG 2 provides mechanisms for a combination of lossy audio/video compression that allows storage and transmission of movies using currently available storage media and transmission bandwidth. MPEG 2 consists of the following parts:

- Part 1: Systems : Synchronization and multiplexing of audio/video.
- Part 2: Video: Compression codec for interlaced and non-interlaced video signals.
- Part 3: Audio: Compression codec for perceptual coding of audio signals.
- Part 4: Procedures for compliance testing.
- Part 5: Procedures for Software simulation.
- Part 6: Extensions for Digital Storage Media Command and Control (DSM-CC).
- Part 7: Advanced Audio Coding.
- Part 8: 10 bit video extension (now withdrawn).
- Part 9: Extensions for real-time interfaces.
- Part 10: Conformance extensions for DSM-CC.

MPEG 3 handles HDTV signals in the range of 20 to 40 Mbit/s. The work on MPEG 3 was discontinued when it was discovered that similar results could be obtained through some modifications to MPEG 2.

MPEG 4 handles web, conversation, and broadcast television. In comparison to MPEG 1 and MPEG 2, MPEG 4 had added Virtual Reality Markup Language (VRML) support for 3D rendering, object-oriented composite files and support for various types of interactivity. The following are other important features of MPEG 4:

- MPEG 4 format provides the end users with a wide range of interaction with various animated objects.
- It multiplexes and synchronizes data associated with other media objects so as to let them be transported ahead via network channels.
- It can provide interaction with the audio-visual scene.
- It can be used by network providers for data transparency. The data can be interpreted and transformed into various signals compatible with any available network.

MP3

MPEG 1 Part 3 Layer 3 abbreviates to MP3. MP3 was invented jointly by personnel from Philips, Fraunhofer Society, Centre commun d'études de télévision et telecommunications (CCETT) and Institut für Rundfunktechnik GmbH (IRT). MP3 uses a lossy compression algorithm that is designed to greatly reduce the amount of data required to represent the audio recording. MP3 provisions a representation of PCM encoded audio in much less space by using psychoacoustic models to discard components less audible to human hearing and recording the remaining information in an efficient manner. Normally, the creator of the MP3 file is allowed to set a bit rate (which specifies how many kilobits the file may use per second of audio). Smaller file size will be mean lower bit rate used, that will result in lower audio quality. Similarly, the higher the bit rate used, the higher shall be the quality—and hence, the larger shall be the file size.

An MP3 file is made up of multiple MP3 frames consisting of the MP3 header and the MP3 data. Such a sequence of frames is called an Elementary Stream. Frames are taken as independent entities in the sense that one can cut the frames from a file and an MP3 player would be able to play it. MP3 enjoys widespread popularity and use. The small size of MP3 files has enabled widespread peer to peer file sharing of music. Today, MP3 is the most popular format when it comes to downloading audio files from the Internet and commercial sale.

18.5.5 H.261

H.261 is an ITU-T video encoding standard for transmission over ISDN lines on which data rates are multiples of 64 kbit/s. It is a member of the H.26x family of video encoding standards in the domain of the ITU-T Video Coding Experts Group (VCEG). The encoding algorithm of H.261 was designed to be able to operate at video bit rates between 40 kbit/s and 2 Mbit/s. H.261 supports two video frame sizes: Common Intermediate Format (CIF) - (352x288 luma with 176x144 chroma) and Quarter CIF (QCIF) - (176x144 with 88x72 chroma) using a 4:2:0 sampling scheme. It also has a mechanism for sending still picture graphics with 704x576 luma resolution and 352x288 chroma resolution. It is used by the free VLC media player, MPlayer, ffdshow and FFmpeg decoders projects.

The encoding algorithm uses a hybrid of motion compensated inter-picture prediction and spatial transform coding along with scalar quantization, zig-zag scanning and also incorporates entropy encoding. The basic processing unit of the design, called a macroblock, was used by H.261. Every macroblock consists of a 16x16 array of luma samples and two corresponding 8x8 arrays of chroma samples, using 4:2:0 sampling and a YCbCr color space. The inter-picture prediction reduces temporal redundancy, with motion vectors used to help the codec compensate for motion.

18.6 Networked Multimedia Applications

When we move on to categorize multimedia networking applications, we need to identify criteria of requirements for such applications to predictably run and behave. In a networked world where ICTs are shrinking the boundaries at a rapid rate, applications in isolation virtually have no place. In networks, we encounter two very important challenges: data loss tolerance and timing considerations. In the Internet, multimedia applications can be classified into three major classes: streaming stored audio/video, streaming live audio/video and real-time interactive audio/video. For delay sensitive multimedia applications, timing considerations hold paramount importance. If the packets of any multimedia application encounter a delay of more than a few hundred milliseconds, they are essentially useless. While on the other hand, most of the multimedia applications are loss tolerant to a good extent. Even occasional glitches in playback can be covered up and managed up to un-noticeable loss. Delay sensitive applications, such as online audio/video are to a good extent loss tolerant. Such applications are intrinsically different from elastic applications like the Web, Telnet and file transfer applications. Elastic applications are extremely sensitive to data loss and for them, completeness and integrity of data transferred is invaluably precious.

18.6.1 Streaming Stored Audio and Video

Streaming is a kind of multimedia application, where contents are stored in compressed format on servers. Stored content can mean anything from pre-recorded television documentaries to a professor's lecture. When the clients request such multimedia content, it is streamed. Due to this, an end user can pause, play, rewind and fast forward through the content. However, the time order in which these commands should manifest themselves in response should not be long (not

more than five to 10 seconds). In streaming, part of the content is first downloaded and buffered into the playout system. The player plays the content from the local buffer, and simultaneously keeps downloading the remaining portion of the video in continuous fashion like a stream. Once the playout of such content begins, it should then advance according to the original timing of recording. Content should be received from the server in time before it is queued at the client for playout.

18.6.2 Streaming Live Audio and Video

Imagine viewing a popular soccer match live on the laptop over the Internet. The class of applications of this kind belongs to Streaming Live Audio and Video applications. It is similar to the traditional broadcasting of any live event except that the transmission is over the Internet. The distribution of such live content can be through either IP multicasting or multiple unicast streams. Delays of up to few seconds can be tolerated but constraints are more stringent than those in streaming stored audio and video discussed in the previous section. However, during the live reception of such multimedia content, the media cannot be fast forwarded.

18.6.3 Real-time Interactive Audio and Video

Internet telephony or video conference is the best example of real-time interactive audio and video. As such communication is highly interactive, delays beyond 400 milliseconds render the steam unintelligible.

18.7 Issues in Multimedia Delivery Over the Internet

There are many challenges when it comes to multimedia content delivery over the Internet. We are aware that the Internet provides the best effort service for packet delivery, but delivery of multimedia content over the Internet becomes crucial in wake of the following constraints:

1. As both the transport layer protocols TCP and UDP run over IP, neither of these protocols make any timely delivery guarantees for any content. This is turn results in no guarantees for end-to-end delay amongst the packets of any multimedia stream content. End-to-end delays up to 150 milliseconds are imperceptible by the human listener, between 150 and 400 milliseconds can be tolerated, but beyond 400 milliseconds are beyond intelligent use.
2. Packet jitter imposes constraints on packet delay. Packet jitter is disproportionate between source to destination packet delay within the same packet stream. Jitter introduces glitches in the playback of multimedia stream which can be intolerable, sometimes. If the jitter is high, then the multimedia conversation can soon become unintelligible.
3. Packet loss can occur when the wandering IP datagram gets lost and never reaches the destined machine. TCP retransmission mechanisms are very inefficient for interactive multimedia streams; therefore, UDP is employed. TCP congestion control and three-way connection establishments are another motivating reason for using UDP. Even though, UDP is a connectionless transport layer protocol, it is unable to prevent packet loss, say, if a pathway link is highly congested. Packet loss up to 20 percent can be tolerated, if proper

encoding and compression techniques are in place with respect to the multimedia content in context. But, losses beyond this count are extremely difficult to manage.

There are many schemes suggested to overcome the aforementioned concerns, such as:

1. Employing UDP as transport layer protocol and thus, circumventing TCP's low throughput for inelastic applications like multimedia.
2. Buffer portion of the media at the play-station by delaying playback at the receiver by some hundreds of milliseconds so as to diminish network induced jitter.
3. Pre-fetching the multimedia content when extra bandwidth and client storage are available.
4. Sending redundant information so as to mitigate the network induced packet loss.
5. Addition of more bandwidth, installation of enough caches and increase in switching capacity at various ISP tiers and networks.
6. Use of Content Distribution Network (CDN) and multicast overlay network. Multicast overlay networks have servers scattered throughout the Internet, which amongst themselves form an overlay network. The CDN network multicasts multimedia traffic to thousands of users.

In our discussion ahead, we shall come across the principles and some concrete ways of dealing with limitations of multimedia content transfer over the Internet.

18.8 Multimedia Delivery Over the Internet

Having discussed the issues of delivery of multimedia content over the Internet, we will now look at widely acceptable mechanisms to handle them.

18.8.1 Handling End-to-End Delays

UDP, runs over IP and does not make any delay guarantees, it is still better than TCP due to the following reasons:

1. TCP, being a connection oriented protocol, performs a three-way handshake before the actual data transfer takes place.
2. TCP has time taking mechanisms to control congestion in the links.
3. TCP's retransmissions and acknowledgements are an additional overhead when it comes to tolerant data loss but inelastic interactive multimedia applications.

On the other hand, UDP becomes a choice in protocols because of the following additional features (apart from the absence of the above mentioned TCP's properties):

1. UDP does not maintain any connection state.
2. UDP had just a small packet header overhead of 8 bytes.
3. UDP provides a finer application lever control over the data. This is because as soon as UDP gets the data from the application layer, it immediately encapsulates such data in UDP segments and passes it on to the network layer for delivery. Hence, it better controls what data is sent when.

The above features not only make UDP suitable for multimedia applications, but also for Domain Name Service (DNS), Routing Information Protocol (RIP) and network management.

18.8.2 Handling Packet Jitter

Packet jitter is caused due to random queuing delays in the pathway routers. Jitter can be handled through the following mechanisms:

1. Prefacing each data chunk with a sequence number and timestamp helps in identification of talk spurt to which the packets belong and also in tracking the packets lost, so that suitable correction mechanisms can be employed. Timestamps help in synchronizing various streams in multimedia session as well as in correct identification of right playout time.
2. Delaying the playout at the receiver helps in arrival of packets before their scheduled playout begins. The playout delay can be either of the two types:
   (a) Fixed Playout Delay: In this mechanism, the receiver plays each chunk exactly after q time units (usually milliseconds) after it is generated. So, if a packet bears a timestamp of t, then it is played out at time t+q (presuming that the packet has arrived). If the packets arrive after their scheduled playout time, they are discarded and are of no good value. The tradeoff in this mechanism occurs in judging a good value of q. As a rule of thumb, if delay and variations in delay are small, using a value of q around 150 milliseconds is acceptable.

Adaptive Playout Delay: This mechanism attempts to remove the tradeoff between delay and loss of packet streams. As the name suggests, the delay after which the packet is played out is, adaptive (and not fixed). The mechanism does so by estimating the network delay and variance in network delay and accordingly calculating the playout delay when any talk spurt begins. This makes the playout delay self adjustable with respect to the network delays.

18.8.3 Handling Packet Loss

A packet is considered lost if it either never arrives at the receiver's address or it arrives after its scheduled playout time. Retransmitting a packet is not a viable option. We discuss here few schemes to handle packet loss in interactive multimedia content transfer:

1. Forward Error Correction (FEC): This adds redundant information in the originally sent packet stream. Marginally increasing the transmission rate and using redundant information to reconstruct lost streams (can be either approximations or exact versions) is a good candidate to handle packet loss. FEC can be achieved in a number of ways. One method can be to encode and send a redundant chunk after every n chunks. This redundant chunk would have an XOR-ed representation of n original packets. By keeping the group size n+1 small, a good fraction of lost packets can be recovered. Another method is to send lower resolution multimedia stream as redundant information along with the original stream. If any chunk is lost on the way, then the lower quality stream can be played in place of the lost chunks without making a perceptible difference in the overall quality.
2. Interleaving: Interleaving re-sequences the stream before it is sent. Let us say, that units are of 2 milliseconds and each chunk is of 10 milliseconds. Then, each chunk will have five units. Now, we have five chunks to be sent. For example, in the first chunk we can place the first units of every chunk (that is the units 1, 6, 11, 16 and 21), in the second chunk we can place the second units of every chunk (that is the units 2, 7, 12, 17 and 22) and so on. Even if one chunk is lost, the effect would be mitigated across several chunks.

3. As the multimedia streams depict a good lot of self similarity, the lost packets can be replaced with the packets received just after or just before the lost ones. This works when loss length does not approach the length of any phoneme.
4. Interpolation can be employed in guessing the lost packets. But, the mechanism is computationally more intensive than the packet repetition approach discussed above.

18.9 MULTIMEDIA NETWORKING PROTOCOLS

Witnessing the massive growth in use of multimedia applications and the Internet convergence, it is not surprising that specific standards have been charted out. Amongst such protocols, SIP, RTP (and its companion protocol RTCP), H.323, RTSP and RSVP are the popular ones. We have discussed Session Initiation Protocol (SIP), RTP, RTSP, H.323, in Chapter 17. In this section, we will cover some other aspects of RTSP, H.323 and RSVP related to multimedia.

18.9.1 Real-time Streaming Protocol (RTSP)

To allow the user control over the playback of multimedia content, Real-time Streaming Protocol (RTSP) is used between the media player at the user's end and multimedia streaming server. RTSP provisions the media player (at the user's end) the control the transmission of multimedia content. Through RTSP, the user can control rewind, fast forward, reposition and pause/resume functions. The major features of RTSP are:

1. RTSP is similar to HTTP in the sense that the messages sent between media player and server are in ASCII text and standard predefined messages.
2. RTSP is an out-of-band protocol (again, similar to FTP) that is the control messages between media player and server are sent out of the band of media stream.
3. RTSP runs well over both TCP and UDP.
4. RTSP does maintain state between multimedia player's and server's for the whole session.
5. RTSP can even allow the client to stream towards the server.
6. RTSP is independent of audio and video compression formats used to encode media and their encapsulation in packets.

RTSP enjoys its reputation among multimedia researchers and can do more than discussed above.

18.9.2 H.323

H.323 is another popular umbrella standard for real-time multimedia interactivity on the Internet from the ITU-T (already introduced in Chapter 17). It serves as a viable alternative to SIP. It mandates H.245 as a control protocol, Q.931 as a signaling channel and Remote Access Service (RAS) for registration with the H.323 gatekeeper. Additionally, every H.323 host should support G.711 speech compression. With respect to the video capabilities of such a host, it should minimally support QCIF H.261 standard. An H.323 gatekeeper is analogous to the SIP registrar. This is one of the differences between H.323 and SIP. H.323 is a complete and fully integrated multimedia interactivity protocol suite. It goes on to cover registration, signaling, call admission control to as far as covering codecs. H.323 even mandates RTP for encapsulating audio/video chunks.

The following are the additional features of H.323:

1. It provisions specifications on how H.323 hosts would communicate among each other and with the H.323 gatekeeper.
2. It specifies how H.323 hosts would communicate with circuit-switched telephone networks and their respective gateways.

18.9.3 ReSerVation Protocol (RSVP)

The resource ReSerVation Protocol (RSVP) is an Internet signaling protocol for reserving bandwidth and link buffers for data flows. It is used by a host (representing the applications) to request a specific amount of bandwidth for data flow. Routers, too, use RSVP while requesting resource reservation requests. One feature to make a note is that RSVP only allows hosts to reserve necessary bandwidth. It does not specify how such reservations would be honoured. One additional point to note is that by the word 'resources', we can mean router's buffers, link state variables and bandwidth, unless we are specific. It is wholly left to the network to carry out such an resource provisioning implementation.

The major features of RSVP can be summarized as:

1. The reservation request starts from the receiver up onto the sender of data flow. This means that the RSVP is receiver oriented.
2. RSVP provisions reservation in multicast trees.
3. The routers may merge the reservation requests coming from the downstream hosts and pass it on to their uplink hosts.
4. RSVP also tears down reservations, if no longer needed.
5. An admission test is performed by the router whenever it receives resource reservation requests to cross check if the downlink multicast tree can accommodate the requests. If such a test fails, the router can reject resource reservation requests.

18.10 Content Distribution Networks

When it comes to geographically distributing some popular multimedia content over the Internet at speed in kbps or few mbps, it becomes a challenge. Although web caching partially helps in alleviating this problem, but there are two main hurdles to it. Firstly, if the requesting client is very distant from the server then there are good chances of long delays and packet losses. Secondly, running the packets of the same popular video over the same congested links for different clients does not seem wise. Content Distribution Networks (CDNs) is a solution towards the above mentioned problem of distributing stored and popular multimedia content. A CDN works as follows:

1. A CDN firm installs a number of servers throughout the Internet at various tiers of ISPs. Such servers are generally installed in Data Centres (DCs) owned by other parties. DCs provide a quick access to lower tier ISP access networks.
2. The CDN firm replicates its customer's content in its CDN servers. Such a customer of a CDN firm can be any content provider (such as a news website). The content provider pre-decides which of its content to be provisioned to clients through CDN. The remaining

objects (such as text or low bit rate images) can be distributed by the content provider on their own. Then, the content provider pushes the desired CDN content to CDN host, which transitively replicates this content to its other servers.

3. Through HTTP (precisely URL) and DNS redirections, when a user (or a client) requests a particular content, it is provided by the best delivering CDN server for that particular client. By best delivering CDN server for a client, it might mean a geographically closer CDN, or a CDN from which there exists a better congestion free path based on BGP tables/roundtrip estimates, or some other factor. The CDN firm pre-estimates the best CDN servers for a large number of access ISPs and uses this information to configure DNS.

CDNs are based on the principle to take the (popular) content to clients when they cannot have a best effort path for multimedia streaming from the respective content servers. As we have seen, many independent companies play a role in CDN implementation. The content provider like bbc.com can distribute its content through a CDN firm like Akamai.

CDNs are also employed by large enterprises and universities spanning different centres.

18.11 Principles of Best Effort Delivery

For live multimedia, quality of service (QoS) guarantee is essential. In this section, we will briefly discuss the architectural changes required to handle QoS of a multimedia session. This seems necessary because the traffic from elastic and nonelastic applications, or delay and loss sensitive applications, etc., are going to be treated equally at the routers. This is also necessary to make high fidelity multimedia applications over the Internet popular. And, there are enough business and technological motivations to make this a reality.

We will briefly discuss four basic principles in this respect:

1. Classify packets so as to allow routers (on the source to destination path) to distinguish amongst packets belonging to different classes of traffic. This can be done through packet marking. Then, routers can behave differently with different traffic classes.
2. For many traffic flows passing through a router, there should be some extent of isolation among them. This is necessary because if one traffic flow misbehaves (that is, tries to capture the whole bandwidth at the router for itself), the other should not be adversely affected. This can be achieved by either policing an ongoing traffic flow or through link layer level packet scheduling mechanism for specific bandwidth allocation to each class of traffic flow.
3. Flow isolation amongst various classes of traffic should use the available bandwidth as efficiently as possible. Consider a scenario where a particular traffic flow has decreased its usage of allocated resources. Then, in this case, the unused resources should not be wasted but made allowable for usage of other existing traffic flows.
4. There should be a call admission mechanism in which traffic flows should declare their QoS requirements beforehand. This would facilitate the event in which a flow is admitted into the network only when QoS requirements can be honored (and not when the QoS cannot be facilitated).

18.12 INTSERV AND DIFFSERV

There are two main architectures proposed for QoS provision on the Internet: Integrated services (Intserv) architecture and Differentiated services (Diffserv) architecture.

18.12.1 Intserv

Intserv is a fine grained QoS provision system in which every router and application that expects any kind of guarantees has to make an individual reservation. This is achieved through flow specifications which describe what the reservation is for and RSVP for signaling the reservation request across the network. For a new session requiring QoS guarantees, it should reserve resources in all the routers falling on the path from source to destination. A router should know what resources it has already committed to the ongoing sessions (through it) and whether it can take new reservation requests. Only when all the routers can admit a new session (based on its requirements), is a call set up and the reservation takes their place.

The new session should first declare the type of traffic it would send into the network and QoS required. Rspec (R for reservation) defines specific QoS required by the application and Tspec (T for traffic) characterizes the nature of traffic which would be sent into the network by the application. RSVP is, then, used for signaling resource reservation requests. With these inputs, the source to destination routers can either admit the call or reject it. However, a new session is admitted only when all the source to destination routers reply positive.

Intserv defines two major classes of service: Guaranteed Quality of Service and Controlled Load Network Service. Under the guaranteed service specification, definite bounds of delays that a packet will experience are notified. While on the other hand, under the controlled load network service specification, a session will experience "a quality of service closely approximating the QoS that same flow would receive from an unloaded network element".

One criticism of Intsev is that a number of states must be stored in each router (one for each ongoing session). As a result, IntServ becomes unscalable. One solution is to use a multi-level approach, where per session resource reservation is done in the edge network, while in the core network resources are reserved for aggregate flows only. Other routers (those which lie between these different categories) should adjust the amount of aggregate bandwidth reserved from the core network. This way, the reservation requests for individual flows from the edge network can be better honored.

18.12.2 Diffserv

Diffserv is a coarse grained QoS provision mechanism for classifying and managing network traffic with scalability and flexible service differentiation. That is, it can handle different classes of traffic in different ways. Scalability comes into the picture because there can be millions of simultaneous flows present at backbone Internet router. This is done by placing simple functionalities in the core network and relatively more complex ones in the edge network. Diffserv provides the functional components with which the new services can be built and the existing ones can be modified.

Diffserv consists of two functional components: edge functions and core functions. Edge functions involve packet classification and traffic conditioning. Edge functions mark the incoming traffic

(that is packets) according to the class of traffic, the packet belongs to. This allows the marked packets to receive differentiated service from the core network. Core functions involve forwarding of the marked packets according to per hop behavior associated with the packet class. The per hop behavior influences router's buffers, link state variables and bandwidth amongst various other classes of traffic present. One important feature specific to Diffserv is that per hop behavior depends only on the mark the packet bears.

One clear advantage of DiffServ is that all the policing and classifying is done at the edge functions. This implies that in the core of the Internet, routers can get on with the real job of routing and not care about enforcing agreements.

18.13 Multimedia Service Creation

We will now discuss how to create multimedia service running on Microsoft Windows platform. The application uses Logitech WebCam to take pictures. The picture can be a still picture or a moving video. We will then use Agora Labs' codec for compression and decompression of these video images. We will also use Microsoft Video for Windows (VFW) that provides functions that enable an application to process video data. The video data can be stored in a file or transmitted over some communication channel. The transmission part is left to the choice of the implementer who can use Internet or can open sockets to transmit the compressed image to the remote system.

It is possible to incorporate video capture capabilities into a Windows application by using the AVICap window class. AVICap provides applications with a simple, message-based interface to access video and waveform-audio acquisition hardware and to control the process of streaming video capture to disk. AVICap supports streaming video capture and single-frame capture in real-time. In addition, AVICap provides control of video sources through Media Control Interface (MCI) so that the user can control the start and stop positions of a video source, and augment the capture operation to include step frame capture.

The Windows AVICap class can perform the following tasks:

- Capture audio and video streams to an audio-video interleaved (AVI) file.
- Connect and disconnect video and audio input devices dynamically.
- View a live incoming video signal by using the overlay or preview methods.
- Specify a file to use when capturing and copy the contents of the capture file to another file.
- Set the capture rate.
- Display dialog boxes that control the video source and format.
- Create, save, and load palettes.
- Copy images and palettes to the clipboard.
- Capture and save a single image as a device-independent bitmap (DIB).

18.13.1 Multimedia Video Libraries

We now present example code to develop a multimedia application that uses video codec libraries developed by Agora Labs (http://www.agoralabs.com/). The Agora Labs VX4000S video codec conforms to the ISO/IEC standard MPEG-4. The MPEG-4 standard specifies 10 profiles, each of which is designed for a different application. The VX4000S video codec implements the Simple

Profile of MPEG-4 that supports data partitioning, synchronization, and reversible variable length coding (RVLC). It supports all the picture formats described in the Profile, and non-standard sizes are supported by the encoder by centering the picture within a standard frame, SQCIF (352*288 pixels) or QCIF (176*144 pixels). The encoder generates an error-resilient video bit stream by using packetization, data partitioning, and rate-distortion optimized coding mode selection. The decoder implements an error concealment technique to improve the quality of reconstructed video during transmission in error-prone channels.

Typical encoding latency1 is in the order of 7.0 milliseconds per QCIF frame. Decoding latency1 is approximately half of the encoding time.

Creating a Capture window

```
hWndC = capCreateCaptureWindow (
   (LPSTR) "My Capture Window", // window name if pop-up
   WS_CHILD | WS_VISIBLE, // window style
   0, 0, 160, 120, // window position and dimensions
   (HWND) hwndParent,
   (int) nID /* child ID */);
```

Connecting to a Capture Driver

```
fOK = SendMessage (hWndC, WM_CAP_DRIVER_CONNECT, 0, 0L);
//
// Or, use the macro to connect to the MSVIDEO driver:
// fOK = capDriverConnect(hWndC, 0);
//
// Place code to set up and capture video here.
//
capDriverDisconnect (hWndC);
```

Setting the Video Format

```
bitCount = 24;
m_nCompressionValue = BI_RGB; // BI_RLE8 ; // BI_RGB ;
videoformat.bmiHeader.biSize=sizeof(BITMAPINFOHEADER);
videoformat.bmiHeader.biWidth=176;
videoformat.bmiHeader.biHeight=144;
videoformat.bmiHeader.biPlanes=1;
videoformat.bmiHeader.biBitCount = static_cast<WORD>(bitCount);
videoformat.bmiHeader.biCompression=m_nCompressionValue;
videoformat.bmiHeader.biSizeImage=0;
videoformat.bmiHeader.biXPelsPerMeter=0;
videoformat.bmiHeader.biYPelsPerMeter=0;
videoformat.bmiHeader.biClrUsed=0;
videoformat.bmiHeader.biClrImportant=0;

if (!capSetVideoFormat(hWndC, &videoformat, sizeof(BITMAPINFO)))
   return false;

fOK = capDriverConnect(hWndC, 0);
```

Setting the callback function for Frame Capture

```
capSetCallbackOnFrame(hWndC,FrameCallbackProc);
```

Where callback function could be defined as

```
LRESULT CALLBACK FrameCallbackProc(HWND hWndC, LPVIDEOHDR lpVHdr)
{
.
.
}
```

This function is used by the capture library to inform application that it has captured single video frame and application could use the captured data.

Capturing Frame

The function below captures one frame which results in invocation of the callback the function.

```
capGrabFrame(hWndC);
```

Calling Agora Encoder

As soon as the frame is captured, the next important job is to compress this frame before transmission happens. This involves a series of steps that need to be accomplished; we do these using following functions in the Agora library.

Agora Key Validation

Agora insists to validate the Library key which is shipped along with the DLL before using either an encoder or decoder. This key is an array of 104 characters. Upon the successful authentication of this key the user can proceed with encoding

```
rc = H263PValidateKey(keyBuffer, H263P_KEYLENGTH );
```

First parameter is the Key array which can be initialized or taken at runtime.

Initializing Color

The function **H263PInitColor** initializes the color conversion tables depending on the output color depth **bpp**. This function must be called before allocating memory to the first decoder instance. Subsequent calls to **H263PDecoderOpen** need not be preceded by a call to **H263PInitColor**, unless the arguments passed to it are different.

Video in digital space can be represented using one of two primary color spaces, viz., YCbCr and RGB. YCbCR is used in television whereas RGB is used in computer displays. The difference between YCbCr and RGB is that YCbCr represents color as brightness and two color difference signals, while RGB represents color as red, green and blue. In YCbCr, the Y is the brightness (luma), Cb is blue minus luma (B–Y) and Cr is red minus luma (R–Y). The standard color conversion, with color control added, is:

```
R = mult * (Y + 1.40200 * Cr) + off
G = mult * (Y - 0.34414 * Cb - 0.71414 * Cr) + off
B = mult * (Y + 1.77200 * Cb) + off

H263PInitColor(24, 128, 128, RGB(128, 128, 128));
```

Opening Encoder

The function **H263PEncoderOpen** creates a new instance of the encoder and takes as its argument a void ** pointer to handle the current encoder instance. The function performs the required memory allocation and returns a zero value if successful or a nonzero value if an error occurs.

If the video coder is provided with an external authentication key, the call **H263PEncoderOpen** should be preceded by a call to **H263PValidateKey** to enable that key.

Input Argument(s):

void **hEncode	Pointer to handle to new encoder instance

Output Argument(s):

void **hEncode	Pointer to handle to new encoder instance after memory allocation

```
rc=H263PEncoderOpen(&hEncode);
```

Callback handler for Encoder

```
rc2 = H263PEncoderSetCallback(hEncode, EncoderCallback,0);
```

Allocating memory to Frame Buffer used by encoder and decoder

This allocation depends on size of the frame which has been captured. In the above statement 176*144 specifies the capture size.

```
framebuf = ( BYTE * )malloc((176*144*3));
```

Initializing the Bitmap Structure

```
dibIn.bmiHeader.biSize = sizeof (dibIn.bmiHeader);
      // number of BYTEs in the structure
dibIn.bmiHeader.biWidth = nx;// width of the picture in pixels
dibIn.bmiHeader.biHeight = ny;// height of the picture in pixels
dibIn.bmiHeader.biPlanes = 1;// number of planes (1)
dibIn.bmiHeader.biBitCount = static_cast<WORD> (bitCount);
      // bit count per pixel (8, 16, 24, 32)
dibIn.bmiHeader.biCompression = 0L;// compression format (FOURCC code)
dibIn.bmiHeader.biSizeImage = 38016*2;// size in bytes of the picture
dibIn.bmiHeader.biXPelsPerMeter = 0;
      // horizontal resolution in pixels/meter
dibIn.bmiHeader.biYPelsPerMeter = 0;
   // vertical resolution in pixels/meter
dibIn.bmiHeader.biClrUsed = 0;// number of color indices
dibIn.bmiHeader.biClrImportant = 0;
      // number of important color indices
im_size = (nx*ny*dibIn.bmiHeader.biBitCount/8);
```

Initializing Frame Data

```
SetFrameData.outbisize.cx  = nx;
      // size of picture that must be converted
SetFrameData.outbisize.cy  = ny;
      // but before interpolation filtering
```

```
SetFrameData.offset.cx = SetFrameData.offset.cy = 0;
      // coordinates of offset for pan/tilt
SetFrameData.flags = ENC_PREFILTER;
```

Setting Mode of Encoder

```
encode_mode = ENC_PACKET_CALLBACK | ENC_FAST_CONTROL;
encode_mode |= ENC_IPIC_QUANT_OPT;
```

Convert input color space and Fill Internal Frame Buffer

The functions H263PEncoderSetFrame and H263PEncoderSetFrameEx convert the input color space and fill the internal frame buffer if possible. They currently accept 8, 16, 24 and 32 bit RGB, YVU9, ty2n (IBM ThinkPad), YUV12 (I420, IYUV) , YUV422 (cyuv), YUY2, UYVY and YCrCb 4:1:1 formats. The input picture can be any size. Output picture size is limited to a multiple of 4 pixels in each dimension. The current implementation also restricts the encoded picture to no larger than 720x576.

```
memcpy(framebuf,dataBuf1,(176*144*3));
res = H263PEncoderSetFrameEx(hEncode, // Encoder handle
            &dibIn,  // ptr to DIB header
            framebuf, // ptr to bitmap buffer
            NULL,
            &SetFrameData
            );
```

Set Parameters before Encoding the Frame

```
int rc3 =H263PEncoderSetParams(hEncode, // handle
             1, // encoder mode
             nx,  // horizontal size
             ny,  // vertical size
             0, // reserved
             8, // starting quality
             5, // frame rate fps
             8000, // target bits/frame
             flags, // option flags
             seqFormat, // I-frame/B-frame            8);
```

Finally Encode the Frame

```
int   n = H263PEncode(hEncode,   // handle
      dataBuf,                   // output buffer
      131072,//strlen((const char *)(dataBuf1)),
                    // max size of output buffer in bytes
      frame_index,   // frame index (1/30th seconds)
      &lNextFrame,   // predicted next frame type
      &BPicType);
```

Note: You could send this encoded buffer over socket to remote client or store in a file to be viewed later. This depends on the application requirement. *'dataBuf'* is output buffer which is of importance. It is left to the reader to decode it; something which is any way mandatory is to decode the encoded frame at some stage. So let's have a look at decoding the frames and displaying them.

Opening Decoder

```
H263PDecoderOpen( &hDecode );
```

Set Display window

The function H263PDisplaySetWindow initializes the display window. The function returns a zero value if successful or a nonzero value if an error occurs.

```
int status = H263PDisplaySetWindow(
                    hDecode,
                    hWnd,
                    WS_VISIBLE,
                    &rct,
                    NULL);
```

Initialize Decoder

The function H263PDecoderInit initializes the decoder by re-setting the internal buffers to their initial values, and sets the display mode of the decoder. The function returns a zero value if successful and a nonzero value if an error occurs.

```
rc= H263PDecoderInit( hDecode, 0);
```

Similar to Encoder set Callback for Decoder as well

```
rc = H263PDecoderSetCallback( hDecode, DecoderCallback,0);
```

Decode the encoded Buffer

The function H263PDecodeRTP is the primary function for decompressing and displaying coded bit stream. The function returns several values, as described below. It will also provide several callbacks (if set up by H263PDecoderSetCallback()) depending on the message and type of frame decoded.

```
H263PDecodeRTP( hDecode, dataBuf, n, decode_flags);
```

Note: As soon as the frame is decoded, decoder callback is invoked where the user can trap the different states and errors while decoding the given frame.

Convert the decode picture buffer into Bitmap

The function H263PGetDecodedRGBFrameEx converts an internal decoded picture buffer into the bit map specified by buf. The parameter fmt specifies a H263PSETFRAMEDATAEX structure that specifies information about the decoded picture and dimensions of the target bit map. The color conversion tables used are those built by H263PdecoderInitColor. If H263PdecoderInitColor has not been run for this decoder instance, default color tables are used.

The default action is to convert the entire picture, without regard to the area of the picture that has been updated. If the flags parameter has bit "H263P_UPDATE_ONLY" set, the only part of the picture copied is in a rectangle containing the upper-left corner and all updated blocks. The sizes returned in *hsize and *vsize is the rectangular area that has been copied (either the region that has been updated, or the entire updated frame).

```
H263PGetDecodedRGBFrameEx( hDecode, &hsize, &vsize, framebuf,
&dfmDataEx);   // get IYUV frame
```

Display Bitmap on the window

The buffer received from the above function call can be used to display the frame on the window visible to the user.

```
HWND videoWnd=hWnd;
if (!videoWnd) return ;

HDC winDC = GetDC(videoWnd);
HDC memDC = CreateCompatibleDC(winDC);

HBITMAP bitmap =  CreateBitmap(width, height, 1, 32,
                  (const void *)frameData );

HBITMAP bitmap = CreateBitmap(width, height,1,24, (const void*)
                     frameData);
long lnSize=sizeof(frameData);
HBITMAP oldmap = (HBITMAP) SelectObject(memDC,bitmap);

unsigned int destWidth=176;
unsigned int destHeight=144;

if ( ( destWidth == width ) && ( destHeight == height ) )
{
   BOOL b = BitBlt(winDC,0,0,width,height,memDC,0,0,SRCCOPY);
}
else
{
   BOOL b = StretchBlt(winDC, 0,0,
      // Destination DC, upperleft corner x, upperleft corner y
         destWidth,destHeight,
      // Destination width, height
         memDC,0,0,
      // Source DC, upperleft corner x, upperleft corner y
         width,height, // Source width, height
         SRCCOPY );
      // Raster operation mode (as with BitBlt)
}
```

18.13.2 Multimedia Audio Libraries

For multimedia services, we use both video and audio. We have explained how to use a video library in Section 18.13.1; here we will introduce Microsoft RTC (Real-time Communications) that allows audio and video calls as well as Instant Messaging and collaboration. It integrates the IP

and PSTN network. Applications can be developed to enable the PC to become the endpoint for multimedia communications; it also allows a PC to become the center for multimedia communication. In addition to PC-PC sessions, the user can also create PC to phone calls, phone to phone calls, or text-only IM sessions. RTC also offers presence information that allows users to call contacts through a registrar server that maintains current location information of contacts. The location can be a PC or a telephone and, in the future, a cell phone, pager, or a handheld device. For example, if you dial a contact at her work location and the presence information indicates she is available on the PC at home, your call will automatically be redirected to that location. Users can also maintain privacy by blocking callers from their presence information.

The latest RTC Client API version is 1.3; it requires `Rtcdll.dll` on Windows. It is designed to enable developers to create solutions that include peer-to-peer communication support as well as server-enabled communications. The RTC API can be used to build applications written in C++, C#, and Visual Basic.

REFERENCES/FURTHER READING

1. Agora Laboratory API for VX4000S Dynamic/Static Link Library on Windows Platforms
2. Forouzan B. A. (2005). *TCP/IP Protocol Suite 3/e*, Tata McGraw-Hill, New Delhi.
3. Forouzan B. A. (2006). *Data Communication and Networking 4/e*, Tata McGraw-Hill, New Delhi.
4. http://www.chiariglione.org/mpeg/ - Official MPEG website
5. http://www.iis.fraunhofer.de–Fraunhofer Institute
6. http://www.jpeg.org–Official JPEG website
7. http://www.wikipedia.org–Online encyclopedia
8. http://www.w3.org/graphics/gif/ - GIF website
9. Kurose J. F. (2005). *Computer Networking 3/e*, Pearson Education, New Delhi.
10. Microsoft MSDN (msdn.microsoft.com)
11. Online Free Encyclopedia–Wikipedia (www.wikipedia.com)
12. *Schaum's Outline Computer Networking* (2005). Tata McGraw-Hill, New Delhi.
13. Stallings W. (2003). *Computer networking with Internet Protocols*, Prentice-Hall.
14. Vaughan Tay (2007). *Multimedia: Making It Work 7/e*, Tata McGraw-Hill, New Delhi.

REVIEW QUESTIONS

Q1: What is meant by the term multimedia?
Q2: How is multimedia different from simple media types?
Q3: What are the benefits of using multimedia?
Q4: What are the application areas of multimedia?
Q5: What media types compose multimedia? Explain each of them.

Q6: Explain the role of compression and decompression with respect to multimedia.
Q7: Explain different types of compression with differences between each of them with examples.
Q8: Explain the difference between temporal and spatial compression.
Q9: What is a codec? Explain its role in multimedia.
Q10: List some important codecs each in audio, video and open source categories.
Q11: Elaborate on video codecs.
Q12: Explain the RGB color space.
Q13: Write a brief note on JPEG.
Q14: Write a brief note on MPEG.
Q15: Write a brief note on GIF.
Q16: Write a brief note on MP3.
Q17: Write a brief note on H.261.
Q18: What are various types of networked multimedia applications? Explain the differences between them with examples.
Q19: What are the issues in multimedia delivery over the Internet? How are they tackled?
Q20: Write a brief note on RTSP.
Q21: Write a brief note on H.323.
Q22: Write a brief note on RSVP.
Q23: What is CDN ? Why is it necessary? Also, explain its applications.
Q24: What are the principles of best effort delivery over the Internet?
Q25: Write a brief note on Intserv.
Q26: Write a brief note on Diffserv.
Q27: Explain the difference between Interv and Diffserv.

CHAPTER 19

IP Multimedia Subsystems

19.1 INTRODUCTION

In 1972, when Arpanet and electronic mail were invented, nobody could even dream that there would be a day when people will use one device to talk, write email, send messages, take photographs, engage in video conferencing while at home or office or in motion while travelling in a car or a train. In the last couple of decades, revolution in communication brought about by the convergence of Internet and cellular technologies were never imagined. They both have, undoubtedly, changed the intrinsic and topographical landscape of how various entities are connected, be it human beings, servers, clients devices, phones, business units, or even social interactions. Internet evolved from its early phase of connecting a few research and academic institutions (described in Chapter 2) to it being the information super highway spanning all across the globe. The growth of mobile phones has also been very impressive; with more than 2.5 billion users of cellular telephones (Chapter 5) across the globe, the services provided by these devices evolved from voice to popular services like SMS (Chapter 6) and MMS (Chapter 8). Also, the power of universal roaming combined with universal availability of cellular networks transformed these devices to all-time companions for the masses.

Popularity of cellular phones and the Internet made industry and technologists look at the convergence of circuit-switched network with packet-witched network. In Chapter 3, we have talked about the circuit-switched telephony system and how computer applications can be enabled through telephony. In Chapter 11, we talked about the telephony signaling system. In Chapter 5 through Chapter 8, we discussed the cellular system. In Chapter 9, we discussed the evolution of 3G from 2G and 2.5G.

In this chapter, we discuss the IMS (IP Multimedia Subsystem) that brings about the convergence of services in the 3G networks and beyond. IMS will offer Internet services anywhere and at anytime over cellular network with guarantee on Quality of Service (QoS). Figure 19.1 depicts the conceptual model of convergence of technologies and networks in the 3G/IMS environment.

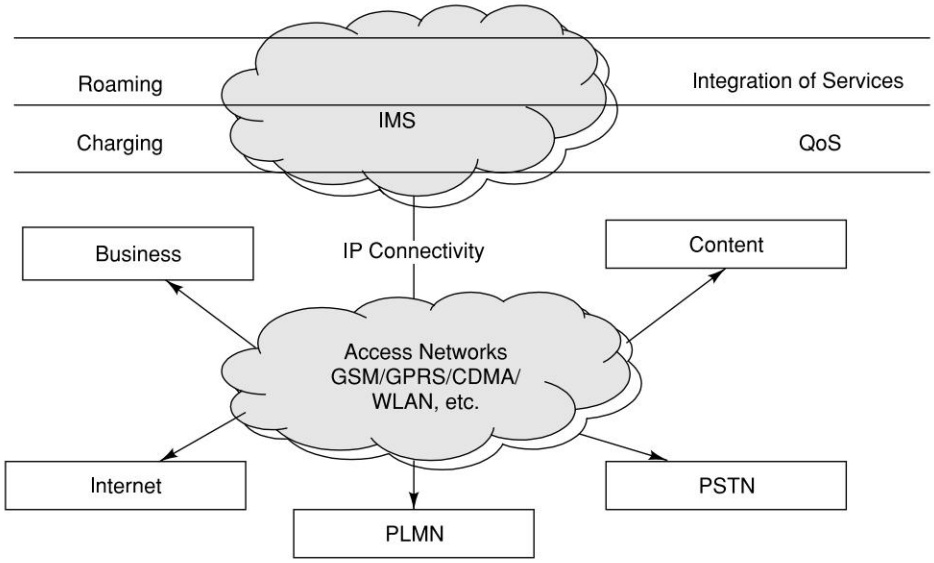


Figure 19.1 Position of IMS in Context of other Access Networks and Businesses

19.2 IMS AND ITS EVOLUTION

IMS is targeted to be platform agnostic convergence at device level, network level, content level, service level, and even at charging level. It can be defined as a service control architecture based on commonly used Internet protocols which enables IP connectivity and rich multimedia services to the end users securely and easily while being independent of underlying access network.

19.2.1 History of IMS

ETSI was the standardization body that defined the Global System for Mobile Communications (GSM). ETSI also defined the General Packet Radio Service (GPRS) network architecture. The last GSM-only standard was produced in 1998, and in the same year the 3GPP was founded by standardization bodies from Europe, Japan, South Korea, the US and China to specify a 3G mobile system comprising Wideband Code Division Multiple Access (WCDMA), Time Division-Code Division Multiple Access (TD-CDMA) radio access, and an evolved GSM core network. The functionality of the release was frozen in December 1999 although some base specifications were frozen afterwards in March 2001. Fast completion was possible because the actual work was divided between two organizations: 3GPP and ETSI. 3GPP developed the services, system architecture, WCDMA and TD-CDMA radio accesses, and the common core network; ETSI SMG (Special Mobile Group) developed the GSM/Enhanced Data Rates for Global Evolution (EDGE) radio access. WCDMA radio access was the most significant enhancement to the GSM-based 3G system in Release 1999. In addition to WCDMA, UMTS Terrestrial Radio Access Network (UTRAN)

introduced the Iu interface as well. Compared with the A and Gb interfaces, there are two significant differences.

Since 1999 and until 2005, there have been six releases. 3GPP has evolved IMS into a finite architecture for SIP-based IP multimedia service machinery. It contains a functionality of logical elements, a description of how elements are connected, selected protocols and procedures. It is important to realize that optimization for the mobile communication environment has been designed in the form of user authentication and authorization based on mobile identities, definite rules at the user network interface for compressing SIP messages and security and policy control mechanisms that allow radio loss and recovery detection. Moreover, important aspects from the operator point of view are addressed while developing the architecture, such as the charging framework and policy, and service control.

19.2.2 Why IMS?

To understand the benefit of IMS, we need to know what exactly are the real advantages in merging Internet and cellular telephony. The answer to this lies in three key areas: Quality of Services (QoS), Charging and Integration of different services.

Internet provides no guarantee about how much time the data transfer would actually take place or the delay incurred by the packets in getting to the destination. Also, due to congestion, packets could be lost or reach the destination out of sequence. Moreover, wireless has its own challenges associated with transmission errors, fading, etc. For real-time multimedia data transfer, therefore, QoS guarantee is absolutely necessary. Consider a user having a video conference. At times, the video quality is good; at times the transfer could just be bearable and at some other times the video might turn incomprehensible. That is why IMS synchronizes session establishment with QoS provision in order to give the user a predictable and un-annoying experience during real-time multimedia sessions.

There used to be a saying in cellular network that “data sells but voice pays”. However, the ARPU (Average Revenue Per User) for voice calls are moving downhill; also, the demand on data is increasing. Therefore, the demand on data services with innovative charging is increasing. In voice networks (PSTN), a user is charged on duration of call; distance the call traverses; and the time when the call is being made. However, in IP, it will have volume-based charging, time-based charging, event-based charging, and content-based charging. Also, operators are offering converged billing where the subscriber receives a single bill for all services like fixedline telephone, broadband over DSL (Digital Subscriber Link), cellular voice, cellular data and services.

Operators do not want to remain just the communications service providers (CSP); they are interested in providing a large array of content and services to the subscribers as well. This will include services created by the operators along with third party service providers. Operators can even offer innovative services by combining services from different service providers. IMS aims at not to standardize the services but to standardize the means to create and integrate such services. IMS also provides the means to deliver these services to the user. This would help the operators to provide a large array of services to users and let its users take advantage of the powerful multivendor service creation industry. Of course, this is in addition to providing the current and future services, available on Internet.

19.2.3 Example of IMS service

Alice decides to send her long-time friend Bob a personalized birthday message while on holiday in the Himalayas. She is having coffee in a roadside shop and enjoying the scenic beauty. She decides to accompany the birthday message for Bob with a video clip of the valley's scenic beauty. She immediately finds a lovely birthday message for Bob on the Internet, adds a line of her own to the greeting message, couples it with the video clip she just took and sends it to Bob. Realizing that Alice is in leisure, Bob calls her up. He thanks Alice for a lovely birthday greeting and invites her to beat him in a car racing game which he is playing on his smartphone. Now, they both get into the game of car racing on Internet through their mobile devices.

19.2.4 IMS Standards

The IMS network has the potential to provide a new telecom business model for both fixed and mobile network operators. There are quite a few standard bodies that are working in areas that relate to convergence and IMS. Major ones are 3GPP, 3GPP2, IETF, ITU-T, ANSI, ATIS, TISPAN, OMA, and GSMA. We have already introduced 3GPP, IETF, ITU-T, ANSI and OMA in Chapter 1. Let us have a look at other standards making bodies like ATIS, TISPAN and GSMA.

The Alliance for Telecommunications Industry Solutions commonly known as ATIS (http://www.atis.org) is a standardization organization for the convergence of information technologies (ITs) and telecommunications technology.

Technologies that are being addressed by ATIS include IMS, Fixed Mobile Convergence (FMC), VoIP, Next Generation Networks (NGN), IPTV, Network interconnection, Network interoperability, Network Security, Telecommunications fraud, etc.

The Telecoms & Internet converged Services & Protocols for Advanced Networks commonly known as TISPAN (http://www.etsi.org/tispan/) is specializing in convergence and Next Generation Networking (NGN) standards. TISPAN is defining a harmonized IMS-centric core for both wireless and wireline networks. Access independent IMS will be a key enabler for fixed/mobile convergence, allowing new services to be rapidly developed and quickly deployed to respond to changing market demands. By making network resources, applications, and user equipment common to all subsystems, TISPAN ensures mobility of users, terminals and services as much as possible, even across administrative boundaries.

GSM Association commonly known as GSMA (http://www.gsmworld.com) is an association representing GSM operators. GSMA plays a pivotal role in the development of the GSM platform and the global wireless industry. GSMA helps its members develop and launch new services, ranging from mobile instant messaging to video sharing to mobile Internet access. The GSM family of technologies consists of today's GSM, GPRS, EDGE and third generation GSM services (3GSM) based on W-CDMA and HSDPA access technologies.

19.3 BENEFITS FROM IMS

As we now know, the IMS aims at delivering a host of useful services to the user that can be developed by operators or any other third party service provider. Operators can develop a new service, integrate the existing ones with a new service from another service provider, or simply change the way a service is being provided to the user. For this end to come true, IMS should help making mobile Internet more and more attractive by combining the latest trends in technology while keeping the doors open to future technologies. IMS should also look into standardizing the service development interfaces (and not the services, except with rare exceptions) so that a common platform exists to develop diverse services. Another key requirement in IMS is that, any service should be able to access any attribute(s) of session(s) in progress and the previous ones. As IMS aims to deliver rich multimedia services to the end users, it should also have support for roaming, operator imposed control over services availed by the user, Internetworking with other networks, etc. These supports from IMS translate into benefits for the user, when in action. IMS supports both voice and non-voice related services over IP that can be summarized as,

1. Person-to-person real-time multimedia services: These include standards telephony, i.e., voice telephony over circuit-switch, packet-switch, as well as video telephony.
2. Customized telephony services: These will allow the service provider or a multimedia application to integrate multiple services around telephony.
3. Multiple synchronized sessions: These allow the user to mix one or multiple person-to-person and person-to-machine sessions.
4. Switching between point-to-point and point-to-multipoint person-to-person sessions: This supports, for example, group conferencing.

To offer these services, IMS must support the following functionalities:

Support for different access networks: Like any other IP network, any network can provide access to IMS be it HFC, WLAN, xDSL, CDMA, WCDMA, PSTN, etc. This meaning that IMS is truly access independent.

Support for roaming: To be able to use mobile devices while on the move has been one of the prime requirements of users. In case of IMS roaming support will go one step forward, which is known as **seamless roaming**. In IMS, a user will be able to roam between various types of networks like WiFi to WiMAX, to UMTS without any interruption in services.

Support for QoS: The ability to negotiate for QoS is a distinguishing feature of IMS. This helps the operators to classify the customers based on QoS awarded to them and thus, maximizing revenue on such classifications.

Support for faster creation of services: IMS targets to achieve faster creation of services by standardizing service creation capabilities rather than services. This means that there would be a common platform from which services would be developed. This will also imply that the services developed would be interoperable and not dependent on any access network or technology.

Support for establishing and maintaining IP based sessions: IMS supports the transfer of multimedia content over packet switched networks and the session control for such transfers. That is, IMS should support peer to peer and peer to content connections over IP networks.

Support for Internetworking: Although, the connection to Internet shall be packet switching based connection, IMS aims to connect to circuit switched networks too. This is because there still exist important circuit switched networks like PSTN networks.

Support for service control: IMS aims to empower operators with effective service control mechanisms. Such mechanisms shall be applicable not only generally (that is, to all users) but also specifically (that is, to a specific user). Such application of service control shall be directly dependent on usage terms of subscribers.

19.4 ARCHITECTURE OF IMS NETWORKS

IMS considers the networking infrastructure by dividing it logically into separate group of functions with standard interfaces between them. Such interfaces are known as "reference points" in the context of IMS networks. Thus, as was mentioned earlier, IMS aims to standardize service developing capabilities than services. There are many interfaces in IMS networks and each defines both the protocol over the interface and the functions between which it operates. However, standards do not dictate what functions should be correlated or how should the functions be grouped into networks nodes. This feature provides enough independence to operators/applications with respect to scalability, requirements and flexibility.

3GPP specifications split the IMS architecture logically into three planes: Application or Service Plane, Control or Signaling Plane, and Transport or User Plane.

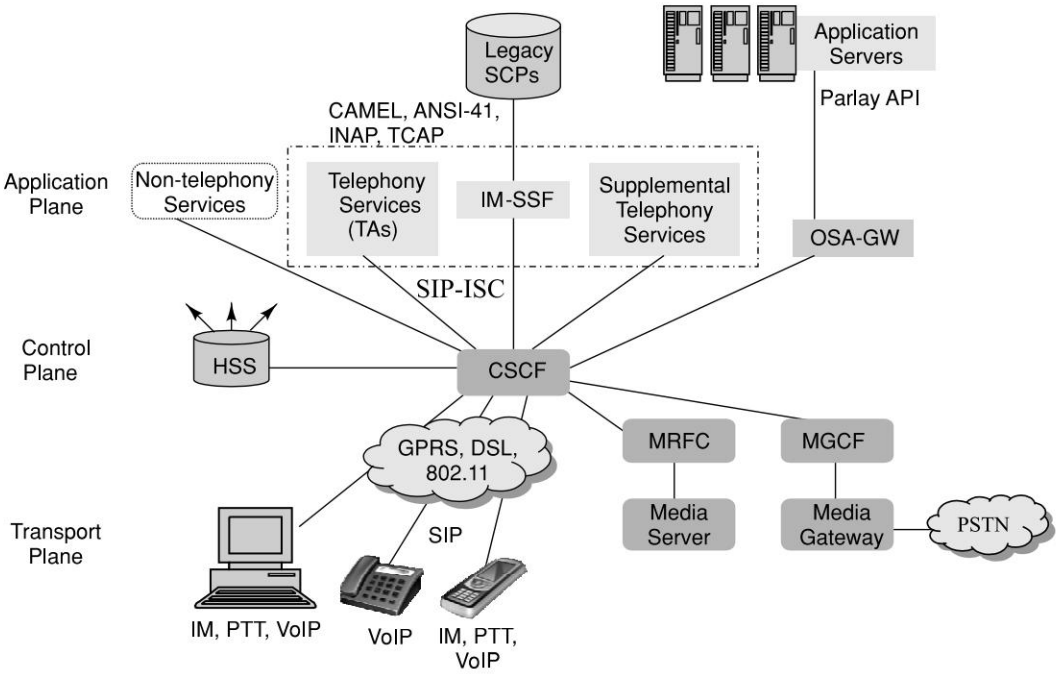


Figure 19.2 IMS Architecture in Relation to Service Planes

The demonstration in Figure 19.2 depicts the IMS architecture in terms of the three service planes. The application plane rests on top of the three planes and is a means of service provision. The services in this plane can be provided through legacy SCPs, Parlay Application Programming Interfaces, etc. As the figure shows, there is a support for non telephony services as well. The IP Multimedia Service Switching Function (IM-SSF) supports the CAMEL Service Environment (CSE) aside from interworking with CAMEL Application Part (CAP). The control function S-CSCF (Call Session Control Function) handles the application servers through SIP as signaling protocol (explained later). The control plane, in between the two service planes, controls and administers the routing and signaling of traffic. The lower service plane-transport plane manages the traffic flow between the user equipment, access and core network.

19.4.1 Application Plane

The Application Plane (or Service Plane), as the name indicates, is responsible for provision of infrastructure and management of services in IMS networks. The major functions of Application Plane include:

1. User identity and status management.
2. Configuration storage.
3. Provision of means of charging.
4. Provision of mechanism of handling control of voice and messaging features (which is handled by the Control Plane).

19.4.2 Control Plane

The Control Plane (also known as Signaling Plane) is the plane responsible for routing the call signals, traffic management at routers with respect to Transport Plane and collecting information for the use of network for billing. It manages the User Plane traffic through Resource and Admission Control Subsystem (RACS). This is actually accomplished by means of Policy Decision Function (PDF), which implements local policy on resource usage. The functions of the Control Plane are mechanized through CSCF. The Control Plane is instrumental in making the IMS networks intrinsically scalable and flexible.

19.4.3 Transport Plane

The Transport Plane (also known as User Plane) is responsible for providing QoS enabled access from User Equipment (UE) over mobile, WiFi and broadband networks. The User Plane provides infrastructure for a wide range of IP multimedia server based and peer to peer services. The User Plane provisions access into the core network through Border Gateways and enforces policies provided by the IMS core. Thus, this plane is able to control traffic flows between the access and the core networks.

19.5 Protocols Used in IMS

While developing the IMS architecture, 3GPP decided to reuse some of the already existing protocols. This was particularly aimed at popular and open protocols. Some of such protocols

were commonly used Internet protocols which are the open ones. The act of producing robust and open protocols was deemed necessary because of the promises IMS made. We have already discussed the benefits and beneficiaries of the open protocols. By robust, it is meant that protocols should be able to run under most of the circumstances and be able to interoperate.

In IMS protocols, there are four major arenas of session control—Authentication, Authorization and Accounting (AAA), real-time data transfer, and policy control. However, there are other protocols also, such as the one for media control. We shall cover the respective protocols one by one.

19.5.1 Session Control Protocol

As we all can guess, the protocol controlling the session in a network is a key protocol. When we come across choosing session control protocol for IMS, we have to bear in mind few basic requirements. First, the protocol should be based on IP. Second, the protocol should work well with multimedia sessions. Session Initiation Protocol (SIP) is chosen to work as session control protocol for IMS. SIP has been described in Section 17.3 in the context of VoIP. SIP is an IETF protocol which establishes and manages multimedia sessions over IP networks. The designers of SIP incorporated the principles of SS7, SMTP, and HTTP while designing SIP. SIP does the following:

1. It manages the ongoing session and can add new multimedia streams while the call is in progress. It can also change the media encoding during the call and invite other users to participate in the ongoing session.
2. It provides mechanisms for call transfer and call hold.
3. It provides establishment of sessions over IP networks while allowing the session participants to notify each other of start of a session, agree on media encodings and end the session.
4. It helps the caller to determine the current IP address of the callee.
5. It always runs on port 5060.

In addition to this, SIP is a text-based protocol. This means that it is easier to debug and extend it. This makes SIP a welcoming protocol to build services upon a key requirement of IMS. In the light of the above description of SIP, it is easy to enumerate the reasons as to why SIP was chosen as session control protocol in IMS.

19.5.2 Authentication, Authorization and Accounting (AAA) Protocol

The role played by Authentication, Authorization and Accounting (AAA) protocol in IMS is evidently clear in terms of handling security and access control. In pre-IMS era, RADIUS (Remote Authentication Dial In User Service) was used for AAA service. RADIUS is defined in RFC 2865. In IMS, Diameter (RFC 3588) is used for handling AAA functions. Diameter is an IETF protocol which was developed to be a successor of RADIUS protocol. RADIUS is extensively used in the Internet to perform AAA. The limitations of RADIUS to support large scale environments and operator defined (centric) accounting features were overcome in Diameter. Diameter is largely an extensible protocol and has its implementation in two broad areas: Diameter base protocol and Diameter applications.

The Diameter base protocol is useful for transfer of Diameter data units, capacity to negotiate and handle errors. The base protocol consists of basic capabilities and must be provided by every

Diameter node. Apart from this, the base protocol is also responsible for providing extensibility. The Diameter application provisions application specific capabilities and related data units.

The key properties of Diameter are:

1. Diameter is a peer-to-peer protocol meaning that any Diameter node can send/receive requests/replies to another Diameter node.
2. From the underlying transport layer protocol, Diameter expects most of the services offered by TCP such as reliability and congestion control.
3. Each session in Diameter signaling can contain several individual requests and replies.
4. Diameter classifies its nodes into three different categories: clients, servers and agents. Client nodes are implemented in the edge devices of a network. Server nodes are responsible for handling AAA requests for a particular domain. Agent nodes are the ones which provide either relay, proxy, redirect or translation services.
5. Diameter is used by IMS in a number of interfaces.
6. Diameter uses a binary header format and transports data units called Attribute Value Pairs (AVPs).

19.5.3 Real-Time Data Transfer Protocol

The facilitation of transfer of real-time data across the nodes of Internet is a key capability of IMS. It is this capability of IMS which would make the sessions interactive and rich of multimedia content. The protocol(s) for real-time data transfer, in effect, not only helps in data transfer but also provides many independent firms to create new and compelling products having the capability to interoperate with each other.

Real-time Transport Protocol (RTP) and RTP Control Protocol (RTCP) are the protocols used by IMS for real-time data transfer. In Section 17.5 we introduced RTP and RTCP in the context of VoIP. Apart from transporting common audio for VoIP, RTP and RTCP can also transport video formats as well. As such, RTP does not provide any mechanism to promise a certain Quality of Service (QoS). This is where the complementary RTCP protocol comes into play. RTP does QoS monitoring using RTCP. RTCP packets are transferred amongst the multimedia session participants as sender/receiver reports announcing statistics which can be useful to the application. Such statistics can include number of packets sent, number of packets lost and inter-arrival Jitter (Jitter is a phenomenon in which packets of the same message experience different delays in reaching a common destination from the same source). The senders and receivers of a multimedia session are free to use such statistics any way they like, and are not bound to use them in any specific manner. Both RTP and RTCP run over UDP. One advantage accrued by both the protocols is that they are complementary to other interactive protocols like SIP and H.323.

19.5.4 Policy Control Protocol

Policy management is a set of rules which determine admission control. It determines what services are availed by which user(s) and enforcement of such rules. This is based on RFC 2753 as "A Framework for Policy-Based Admission Control". Typically, the policies fall into two main categories: general policies and specific policies. General policies are the ones which are applicable to all the subscribers. An example can be a scenario where the operator restricts all its subscribers using a

bandwidth consuming video codec. Specific policies are the ones applicable to either any individual subscriber or a group of subscribers. An example can be permission to a group of subscribers to use both video and audio while availing some service.

The five logical entities in the context of policy management are Access Requestor (AR), Policy Enforcement Point (PEP), Policy Decision Point (PDP), Policy Repository (PR), and the Network Decision Point (NDP). The AR is any endpoint device seeking access to a network resource. PEP is a network element that enforces policy decisions. PDP is a device where a policy decision is made. PR is a data-store which holds policy rules, actions, conditions, and related data. NDP is a network element which interprets network events and sends information to PDP. NDP cannot enforce a policy decision (e.g., an intrusion detection system like Snort), instead it processes events and send them to the PDP for review and enforcement on other devices. We go on to describe policy management with respect to IMS networks in greater detail in Section 19.11.

RFC 2748 describes the "COPS (Common Open Policy Service)" that is used by IMS as a protocol to transfer policies among PDPs and PEPs to ensure QoS. COPS is useful for administration, configuration and enforcement of policies in IMS networks. It does this by employing a client/server model for exchanging policy information. COPS is a binary protocol whose messages contain COPS header followed by typed objects. COPS for policy provisioning (COPS-PR) is a COPS outsourcing and provisioning model. The message format and the use of Policy Information Bases (PIBs) come from the provisioning model while real-time transfer of policy decisions come from the outsourcing model. PIB identifies the policy provisioning data and there can be more than one PIB defined for each area of policy provisioning.

19.5.5 Other Protocols

Session Description Protocol (SDP) is a text-based protocol which describes multimedia sessions in an ongoing IMS session. SDP is defined in RFC 2327; it is, basically, an application layer protocol that helps the caller as well as the callee to express their respective receive capabilities, media exchange formats along with the receive address (or port). SDP messaging transfers session level description, timing description, media type and media format.

Media Gateway Control Protocol (MEGACO), also known as H.248, handles signaling and session management between a media gateway and media gateway controller during a multimedia session. A media gateway controller controls a number of media gateways by acting as a master in a master/slave relationship with media gateways. MEGACO defines the necessary signaling procedures to allow a media gateway controller to control gateways to support multimedia session in an IMS network.

Internet Protocol Security (IPsec) offers security services to upper layers in an IP network (mainly to network layer). IPsec secures communication between security gateways and hosts. IPsec provides access control, confidentiality, data integrity protection, anti-replay protection and data origin authentication. IPsec operates in either tunnel or transport mode. Tunnel mode is used for tunneling traffic between two security gateways while transport mode provisions security services to protocols of upper layer. A security association is set up between nodes interested in exchanging IPsec secured traffic. Every security association is individually identified by address of the participating nodes and security parameter index. IPsec is used by IMS in access as well as network security.

19.6 Building Blocks in IMS Networks

As we have discussed, IMS aims to standardize functions than nodes where such functions are implemented. While going to enumerate such functions, we need to describe the basic entities which form the skeleton of a network where IMS services could be provisioned. It is emphasized that such entities be understood as functions than individual capability nodes.

19.6.1 Databases (HSS and SLF)

In IMS networks, Home Subscriber Server (HSS) and Subscriber Location Function (SLF) form the primary databases. HSS and SLF combined can be considered as the next generation HLR (HLR is described in Chapter 5 in the context of GSM).

HSS stores all the subscriber and service related data of any IMS network. The primary contents include user and service identities, provisioning information, registration data, access parameters, location and security information, service triggering information and S-CSCF (described later) allocated to the user. HSS can be thought of as a 3GPP evolution of HLR of GSM networks.

In case the number of users is large, there can be more than one HSS present. In such scenarios SLF is needed. SLF provides mapping between user profiles to their respective HSSs. This means that if user information is provided to SLF as a query, it will output the corresponding HSS where complete details of that user can be found. It is important to note, however, that the details of any user are never stored across multiple HSSs.

19.6.2 Call Session Control Function (CSCF)

Unlike circuit-switch networks where call flow is handled by SS7 (Chapter 11) signaling network, IMS calls are handled by SIP protocol. For processing SIP signaling in IMS networks, Call/Session Control Function (CSCF) servers are needed. There are three main classes of CSCFs, these are P-CSCF (Proxy-CSCF), I-CSCF (Interrogating-CSCF), and S-CSCF (Serving-CSCF).

Proxy-CSCF (P-CSCF) is the first interface between the IMS terminal (the user device) and the IMS network. This means that the duplex communication which takes place between the user terminal and the IMS network always traverses P-CSCF. The main functions of P-CSCF can be summarized as follows:

1. It verifies the validity of SIP messages sent across through it apart from compressing and decompressing such messages whenever need be.
2. It establishes a number of IPsec security associations towards the IMS terminal. This way it is able to assert the IMS terminal's identity throughout the IMS network and obviate the need of other nodes to re-identify the IMS terminal.
3. It may include a PDP for management of various resources and QoS.
4. It generates charging related data towards a node responsible for charging.

Interrogating-CSCF (I-CSCF) is a SIP contact server within the IMS network for all the communications destined to a user of that network. The primary tasks of I-CSCF are as follows:

1. It obtains the address of the next hop (i.e., SIP application server) from HSS (i.e., providing a routing functionality). It the also provisions an interface between HSS and SLF.

2. It provides Topology Hiding Inter-network Gateway (THIG) functionality. THIG helps in not getting the sensitive information of network disclosed.

Serving-CSCF (S-CSCF) is at the heart of any IMS network. It handles the user registration process, makes routing decisions while maintaining sessions and stores service profiles. Other primary functions of S-CSCF can be summarized as follows:

1. It acts as a PEP and maintains discipline with respect to the services accessed by the user.
2. It interfaces with HSS for various purposes. For example, it might intimate HSS, the S-CSCF allocated to a particular user for duration of registration.

In any IMS network, there can be any number of various CSCFs in order to meet the requirement and scalability needs.

19.6.3 Application Servers

The application servers in IMS networks are responsible for service/application provision. There are three main types of application servers in IMS networks: OSA-SCS (Open Service Access Service Capability Server), SIP-AS (SIP Application Server) and IM-SSF (IP Multimedia Service Switching Function).

Open Service Access-Service Capability Server (OSA-SCS) provides capability to access the IMS network securely from external networks. It provides all the Open Service Access (OSA) capabilities. SIP-Application Server (SIP-AS) can be any SIP server in the IMS network hosting and providing IMS services. IP Multimedia Service Switching Function (IM-SSF) helps the IMS networks use and reuse of intelligent networking functions like CAMEL (discussed in Chapter 11). The application servers might also interface with HSS for various functions, as and when required.

19.6.4 Media and Gateway Related Functions

The Media Resource Function (MRF) provisions media resources in IMS networks. The MRF does all activities with respect to media management. The MRF can be broadly classified into MRFC and MRFP. Media Resource Function Controller (MRFC), as the name suggests, controls the media resources while acting as a SIP user agent and interfacing with S-CSCF. Media Resource Function Processor (MRFP) executes media related functions. The Breakout Gateway Control Function (BGCF) is a SIP server handling routing based on telephone numbers. It comes into play when the IMS terminal initiates a session addressed to circuit switched network.

The Public Switched Telephone Network/Circuit Switched (PSTN/CS) Gateway acts as an interface towards circuit switched networks. The circuit switched network can be PSTN for fixedline, or PLMN for GSM, etc. Its main functions can be decomposed into SGW, MGCF and MGW. Signaling Gateway (SGW) interfaces the signaling and control plane of a circuit switched network by providing means for lower layer protocol conversion. Media Gateway Control Function (MGCF) does call control protocol mapping apart from controlling media resources. Media Gateway (MGW) provides itself as a media plane interface by using inter-protocol mapping function and transcoding (in the case where IMS terminal supports different codecs). These functions are described in Chapter 17 in the context of VoIP.

19.6.5 Identities in IMS

As in other networks, the need to identify a subscriber and services provisioned for this subscriber arises in IMS, too. It is common to identify users in a PSTN network through telephone numbers. Similarly, special services in PSTN networks are identified by numbers starting with 800, 900, etc. In the context of IMS networks, there are three major identities: Public User Identities, Private User Identities and Public Service Identities. It would be analogous to perceive Public User Identities as MSISDN and Private User Identities as IMSI in GSM networks, respectively (Chapter 5).

The identity used for requesting communication with other users is called a Public User Identity. In IMS networks, a single user can have multiple Public User Identities assigned by the operator. This feature is especially useful when there can be a need of personal, public, or business identity. Public User Identities come into use while routing SIP signaling. A Public User Identity in IMS networks can be either a SIP URI or a TEL URL. Examples of a Public User Identity being a SIP URI are sip:mobile.computing@book.com, sip:+91-80-54872187@book.com; (this one provides phone number in SIP URI), etc. An example for a Public User Identity being a TEL URL is tel:+91-80-54872180.

Unlike Public User Identities, Private User Identities are unique for each subscriber and take the format as subscriber@operator.com. Private User Identities are useful only for unambiguous subscriber identification and authentication purposes.

For every service hosted in the application server in IMS networks, there is a unique Public Service Identity allotted to it. However, the Public Service Identity can assume either of the formats used by Public User Identities.

For a subscriber, (multiple) Public User Identities and unique Private User Identity are stored in smart card IP multimedia Services Identity Module (ISIM). Again, ISIM in IMS networks is analogous to SIM in GSM networks. This was the concept until 3GPP release 5. In IMS networks, the HSS (and the S-CSCF) correlate the identity of the subscriber, the Private User Identity and multiple Public User Identites.

19.6.6 IP Multimedia Services Identity Module (ISIM)

As stated earlier, the analogous Universal Integrated Circuit Card (UICC) for IMS networks is ISIM. ISIM primarily contains parameters which are useful for user identification—public and private, user authentication and home network domain URI. A remarkable feature of ISIM is that it can coexist with SIM or other UICCs in the same circuit card.

19.7 Call Flow in IMS Network

We describe a typical IMS registration where a user is taken into the IMS network. The stepwise call flow follows:

1. The dedicated signaling context is established between the user's UE and the Gateway GPRS Support Node (GGSN) in the case of General Packet Radio Service (GPRS) being radio access technology.

2. The UE discovers the address of the Proxy Call Session Control Function (P-CSCF), which it uses as a SIP outbound proxy during registration and for all other SIP signaling while it is registered.
3. The UE sends a REGISTER message to the user's home network to perform SIP registration for the user's public user identity.
4. The Interrogating-CSCF (I-CSCF) selects the Serving-CSCF (S-CSCF) that serves the user while it is registered.
5. The S-CSCF downloads the authentication data of the user from the Home Subscriber Server (HSS).
6. The UE and the P-CSCF agree on a security mechanism.
7. The UE and the network (S-CSCF) authenticate each other.
8. IP Security (IPsec) associations between the UE and the P-CSCF are established.
9. SIP compression starts between the UE and the P-CSCF and UE learns the route to the S-CSCF.
10. The S-CSCF learns the route to the UE while the S-CSCF downloads the user profile of the user from the HSS.
11. The S-CSCF registers the default public user identity of the user.
12. The S-CSCF may implicitly register further public user identities of the user.
13. The UE becomes aware of all public user identities that are assigned to the user and his current registration state and the P-CSCF becomes aware of all public user identities that are assigned to the user and his current registration state.
14. Thus, the user is logged in.

Figure 19.3 describes call flow in a diagrammatic representation when the user accesses the IMS networks from a visited network. It is noteworthy to gauge the role of SGSN and GGSN as the figure conveys that it is GPRS network on which IMS builds up its service control architecture. BG can be Breakout Gateway function which handles session delivery amongst networks.

After the user moves into the IMS network, the subsequent call flows depend on the services availed by the user, the class to which the user belongs (i.e., type of subscriber), type of session, other users, etc.

19.8 IMS CHARGING

The goal of IMS is to allow an operator to charge the users appropriately. When a user transfers data in a real-time multimedia session, loads of bytes are transferred across the network. But during that session, the same user might also be involved in a chat session with another user, surfing the web, reading an electronic mail and the like. Usually, in Internet, a user is charged by the operator on the number of bytes transferred across the user's terminal; the user is also charged on the duration of connection time. The user can also be charged based on the type of service. Operators levy charges on real-time video conferencing sessions on the basis of total number of minutes during which that particular video conferencing session was conducted, regardless of the number of bytes transferred across the user's terminal. The leader of a video conference is also changed based on number of participants on the video conference. Thus, a more logical and intuitive way of charging shall exist. It is important to keep in mind that IMS standards do not provide any particular rating model on how to charge the subscriber nor does it provide any guidelines for it.

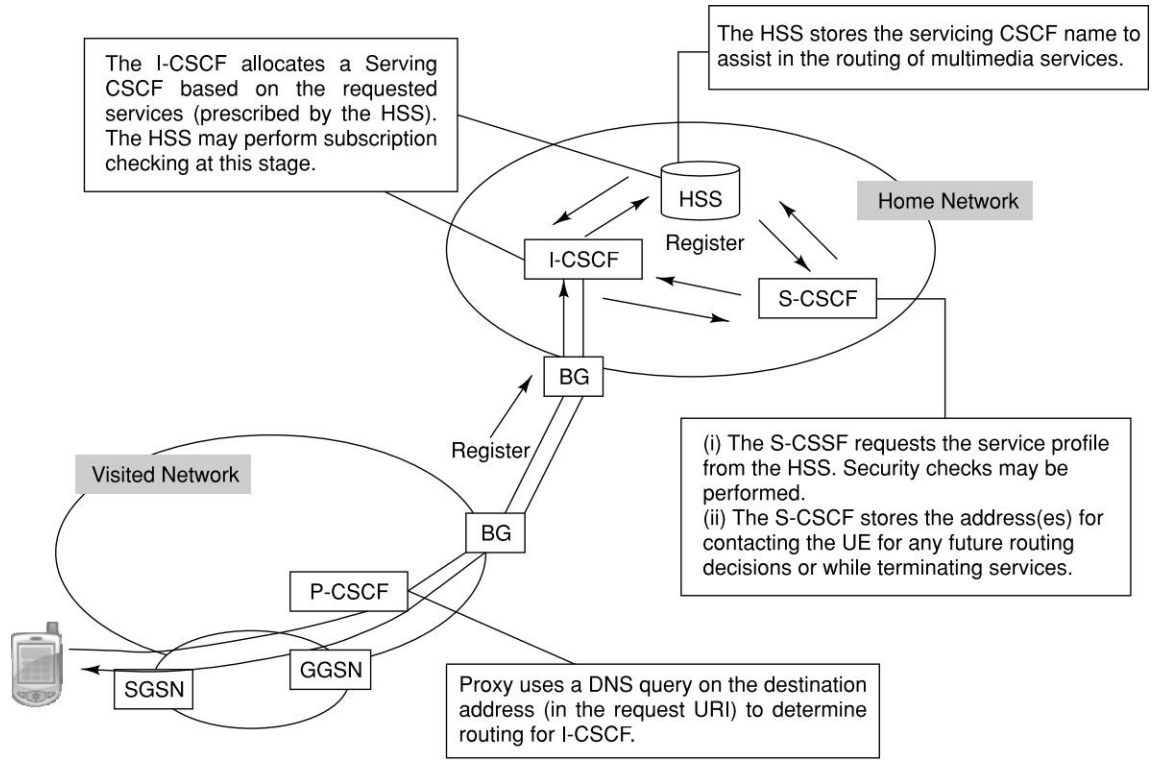


Figure 19.3 Call Flow from Visited Network

IMS charging functions will be any combination of the following criteria:

Volume based charging: In this mode the volume of data transacted will be charged. For example, the charging is done based on MB (Mega Bytes) of data transferred. This is also referred as usage-based charging. This is done through the network flow data. IP Data Record is used to measure and charge a subscriber.

Time based charging: This handles the duration for which the user was logged in into the network. The charging could also be based on time of the day. For example, accessing the network after 11:00 PM at night could be free. Also, there could be different rates during the weekend.

QoS based charging: This type of charging is dependent on quality of service. For example, a user wants to subscriber to some video (Video on Demand) for which the user needs a higher bandwidth for some fixed duration. For this duration the network will guarantee some higher bandwidth and the user needs to pay a different tariff for this activity.

Event based charging: This type of charging is dependent on events. Events could be mail, instant messaging or some other push events like stock quotes when the price of some script moved up or down compared to some user defined threshold.

Service based charging: The operator could charge different rates for different services. For example, a VoIP service may be priced differently compared to Web browsing. Also, browsing some sites could be free. Even browsing web sites with operator injected advertisement could be free.

Content based charging: This type of charging is dependent on the content that is being accessed. For example, a live soccer game could be charged differently compared to other IPTV programs.

19.8.1 IMS Charging Functions

The architecture of IMS allows for various models which can be put in practice. Such models can not only be provisioned to include only the calling party for charging but also the called party. A called party can be charged in case it adds a new media stream to the existing stream or modifies the existing one to a costlier one. It is also possible that the operator is interested in correlating charging information generated in transport and IMS charging levels.

IMS is well supportive of both online and offline charging mechanisms. Online charging directly interacts with session and affects the service rendered in real time. A good example is prepaid service. On the other hand, offline charging does not affect the services consumed in real-time and the operator posts a bill to the user for a period. For supporting the online charging mechanism, IMS network nodes consult the Online Charging System (OCS). OCS maintains a real-time interaction with the user's account and continuously monitors the charges depending on the service usage (Refer to Figure 19.4).

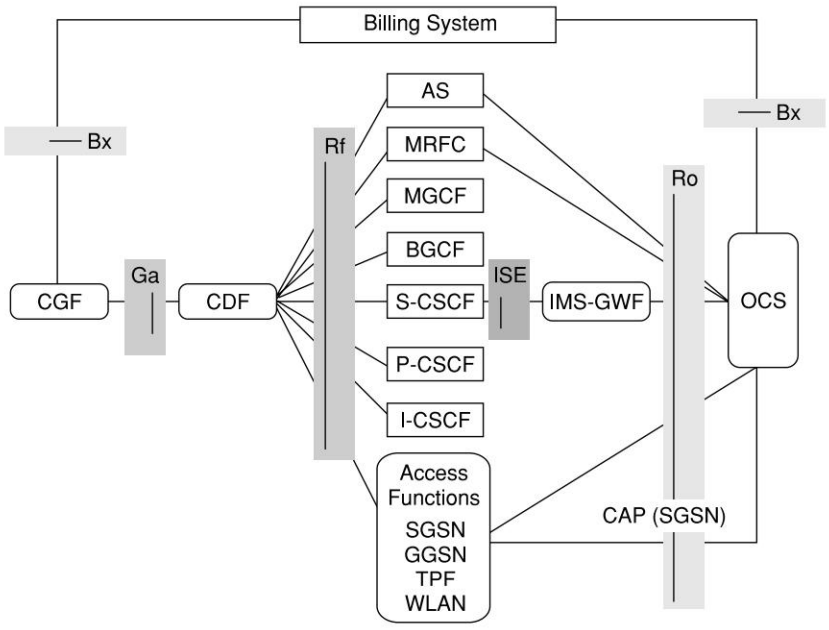


Figure 19.4 Architecture of IMS Charging (Left portion—offline charging and right portion–online charging)

A chargeable trigger condition has to be detected by the IMS network entities to kick start or stop or modify the process of charging. Such a chargeable trigger can be any SIP transaction like session initiation, session modification, session termination, etc. In some cases, it can even be a change in SIP header. After the trigger is detected, the IMS network entities collect the relevant information from SIP messages. This is where the path of charging mechanism has to be identified.

In case of online charging, the IMS network entities continue on processing the SIP transactions and can affect the services being delivered in real time. In online charging, we have witnessed the integral role of three IMS entities—Application Server, Media Resource Function Controller and Serving-Call Session Control Function. It is possible that S-CSCF talks with OCS using IMS Gateway function (IMS GWF) for essential inter protocol conversion. However, it is the IMS Service Control (ISC) reference point which interfaces between S-CSCF and IMS-GWF. OCS recognizes two more reference points: CAMEL Application Part (CAP) for SGSN and Diameter's Ro reference point for other entities. OCS can approve or delegate resource in real-time through credit control handling by creating Charging Data Requests (CDRs).

While in the case of offline charging, these entities send the relevant information to the charging module for creating a Charging Data Request (CDR) for post processing and at the same time permit the SIP request to continue. Such a charging module is called Charging Data Function (CDF). A Diameter-based reference point Rf is used by the IMS entities to interface with CDF. A CDF also collects such information from other network entities as well before creating CDRs which are transitively delivered to Charging Gateway Function (CGF) though Ga reference point. After collecting all the CDRs, the CGF channels them to the billing system through Bx reference point.

19.9 Reference Points in IMS

Reference points form the building blocks of communication between the IMS entities. They form standard interfaces for the entities to 'talk'. The details of some of these important reference points are tabulated below:

S.N.	*Reference points*	*Details*
1	Gm reference point	The Gm reference point connects the UE to the IMS and is used to transport all SIP signaling messages between the UE and the IMS. Its IMS counterpart is P-CSCF. The general procedures in Gm reference point are: • Registration: UE uses the Gm reference point to send a registration request with an indication of supported security mechanisms to the P-CSCF. • Session control: Session control procedures have elaborate mechanisms for both mobile-originated sessions and mobile-terminated sessions. In mobile-originated sessions, the Gm reference point is utilized to forward requests from

(*Contd*)

S.N.	*Reference points*	*Details*
		the UE to the P-CSCF. While, in mobile-terminated sessions, the Gm reference point is used to forward requests from the P-CSCF to the UE. • Transaction: These are helpful in sending stand-alone requests to receive all responses to that request via the Gm reference point. The major difference between transaction procedures and session control procedures is that a dialog is not created in transaction procedures.
2	IMS Service Control (ISC) reference point	This is the reference point for sending and receiving SIP messages between the CSCF and Application Server.
3	Cx reference point	Subscriber and service data are permanently stored in HSS. Such centralized data needs to be utilized by the I-CSCF and the S-CSCF when the subscriber registers or receives sessions. Therefore, there is a need of a reference point between the HSS and the CSCF. This reference point is called the Cx reference point. The procedures can be divided into three main categories: • Location management • User data handling • User authentication
4	Mw reference point	As Gm reference point links the UE to the IMS (through the P-CSCF) so, a SIP-based reference point between different CSCFs is needed. This reference point is called the Mw reference point. Again, procedures in the Mw reference point are divided into three main categories: • Registration • Session control • Transaction
5	Dx reference point	When multiple and separately addressable HSSs are deployed in a network, neither the I-CSCF nor the S-CSCF know which HSS they need to contact. They first need to contact the SLF first. The reference point for this is Dx. The Dx reference point is always used in conjunction with the Cx reference point.
6	Sh reference point	An AS (SIP AS or OSA SCS) may need user data or need to know to which S-CSCF to send a SIP request to. Such contextual information is stored in the HSS. Therefore, another reference point is needed between the HSS and the AS. The Sh reference point works for this. The procedures are divided into two main categories:

(*Contd*)

S.N.	Reference points	Details
		• Data handling • Subscription/notification
7	Dh reference point	When multiple and separately addressable HSSs have been deployed in the network, the AS does not knows which HSS it needs to contact. AS needs to contact the SLF first. This is through the Dh reference point. The Dh reference point is always used in conjunction with the Sh reference point. To get an HSS address, the AS sends Sh request to the SLF, aimed for the HSS. On receipt of the HSS address from the SLF, the AS does its job of sending the Sh request to the HSS.
8	Si reference point	The Si reference point is used to transfer CAMEL subscription information including triggers from the HSS to the IM-SSF.
9	Mm reference point	Mm reference point is helpful in communicating with other multimedia IP networks. It helps the I-CSCF to receive a session request from another SIP server/terminal. Analogously, the S-CSCF uses the Mm reference point to forward IMS UE-originated requests to other multimedia networks.
10	Mg reference point	The Mg reference point links the Circuit Switch (CS) edge function-MGCF to IMS' I-CSCF. Mg allows MGCF to forward incoming session signaling from the CS domain to the I-CSCF.
11	Mi reference point	Mi reference point is needed to forward the session to BGCF. When the S-CSCF discovers that a session needs to be routed to the CS domain, Mi is used for the aforesaid function.
12	Mj reference point	When BGCF receives session signaling via the Mi reference point, it has to select the CS domain in which breakout is to occur. If breakout happens to occur in the same network, then it forwards the session to MGCF via the Mj reference point.
13	Mk reference point	Mk reference point is used for communication between the BGCFs of different networks. When BGCF of one network receives session signaling via the Mi reference point, it selects the CS domain in which breakout is to occur. If the breakout occurs in another network, then it forwards the session to BGCF in the other network through the Mk reference point.
14	Mn reference point	The Mn reference point is the interface between the MGCF and IMS-MGW. Mn interface controls the user plane between IP access and IMS-MGW (Mb reference point). It also controls the user plane between CS access (Nb and TDM interfaces) and IMS-MGS.

(*Contd*)

S.N.	*Reference points*	*Details*
15	Mr reference point	This reference point is used by the S-CSCF to pass SIP signaling to the MRFC, to activate bearer-related services.
16	Mp reference point	This reference point is used by the MRFC when it needs to control media streams.
17	Ut reference point	This is the reference point existing between the UE and the AS. It enables users to securely manage and configure their network services related information hosted on an AS. The users can use the Ut reference point to create Public Service Identities (PSIs).
18	Go reference point	Go helps communication between the IMS (control plane) and the access network (user plane) to ensure that the QoS and source and destination addresses of the intended IMS media traffic matches the negotiated values at the IMS level. Now, charging correlation is an additional functionality.

Figure 19.5 captures the positioning of the above mentioned reference points with respect to IMS architecture. The figure does so in terms of GPRS as radio access technology.

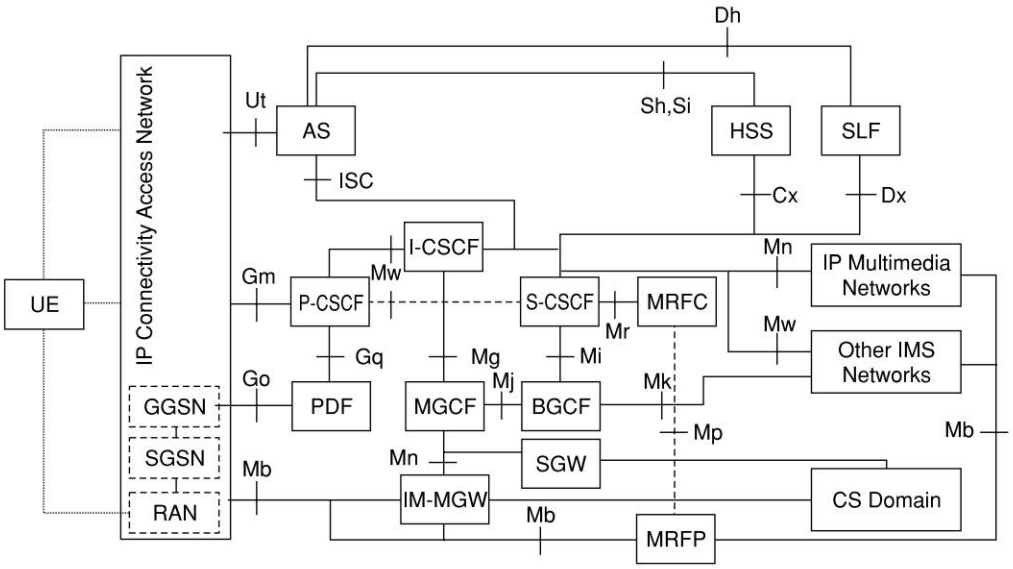


Figure 19.5 Reference Points in IMS Architecture

19.10 Service Creation in IMS

IMS offers standardized person-to-person multimedia services such as Push-to-Talk and services that combine, for example, Instant Messaging or video with voice. IMS also makes a framework for innovation and application creation easily available to the application-developer community through APIs. IMS services can be divided into two major segments, viz., standard services that are defined by various standards bodies; and non-standard services that are developed by third parties and independent software vendors. Like standard services, non-standard services are also deployed in the application server. However, there may not be a standard client application available to access such non-standard services. Therefore, there is a need to develop such non-standard applications to be loaded on the TE (Terminal Equipment). Fortunately, Java addresses this challenge. JSR-281 is intended to be used by application developers who wish to build Java applications for terminal equipments that use the IP Multimedia Subsystem (IMS). Another noteworthy point is that the Java interface in the application servers is a part of routine Java community process. The work is still on to extend the current standardized Java environment, defined in JSR-116, to allow for greater flexibility while building SIP based services. This will facilitate rapid service development.

There can be virtually innumerable non-standard services. These non-standard services are generally called Value Added Services or VAS in short. The following list contains some examples of VAS:

- Implementing the complete architecture with federated solutions including Single Sign On (SSO), privacy management and attribute sharing. Single Sign On is an optimization of the authentication sequence to ease the burden of repeating actions done by the user. It is the ability to use proof of an existing authentication session with provider A to create a new authentication session with provider B.
- Extending and implementing the solution on various access networks like GPRS, CDMA 2000, Wired LAN, etc.
- Provision of user triggered security features.
- Provision of policy control gateways.
- Incorporating new technologies like Automated Trust Negotiation (ATN) systems.

Broadly, we can categorize the emerging services into three types:

1. Real-time user-to-user multimedia telephonic communication (MMTel).
2. Floor-control half-duplex (walkie-talkie) voice communication (Push-to-talk over Cellular).
3. Instant messaging with support for message storing and forwarding (which is in turn called IMS messaging).

Some of these services will be offered as standard service with default client (user agent) available in the TE. While, in other cases, downloadable clients using the standardized Java interface in the terminal can be used. However, such a Java interface can also be used in a terminal-to-terminal session. However, if such new services use the standardized services as a communication service, they can also be used as differentiating services with respect to an operator's own subscribers or services offered by it.

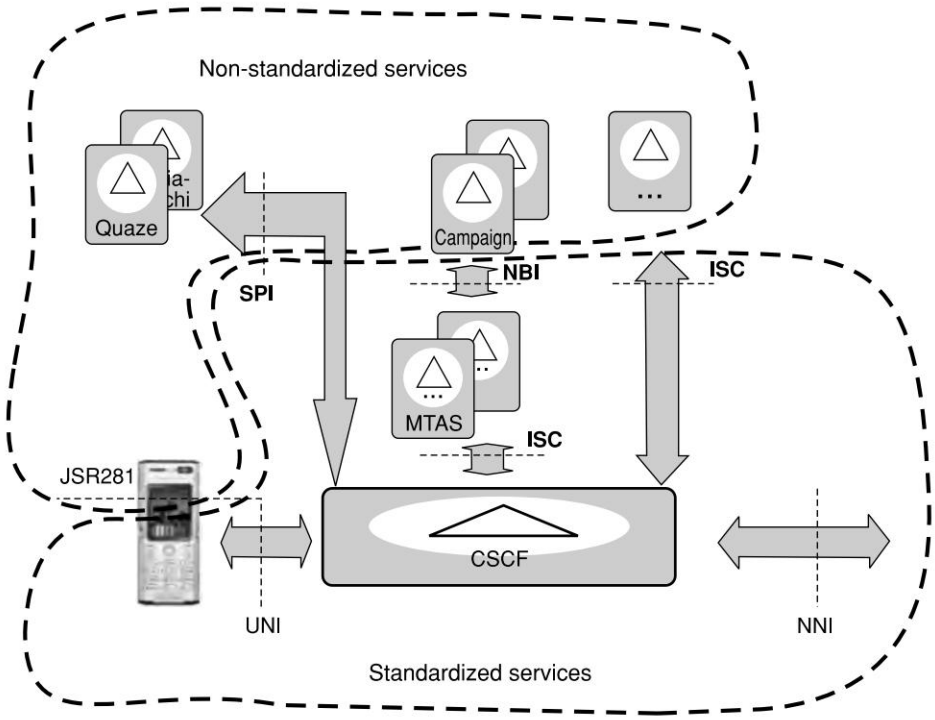


Figure 19.6 The IMS Service Architecture

Figure 19.6 depicts a scenario of IMS application development framework using JSR-281. The figure conceptually brings about the differences between standardized and non-standardized services in IMS. Any operator can use the SPI (Service Provider Interface) as the interface towards third-party service providers. However, this can also be achieved through standardized NNI (Network to Network Interface) and its supportive communication services. In an standard IMS service, the session is initiated by one party from a standard application in the terminal. It accesses the network over the standardized User-to-Network Interface (UNI). The CSCF inspects the signaling and concludes that this particular session is to be managed by the Multimedia Telephony Application Server (MTAS). If the called party is not served by this operator, the session is forwarded over the standardized Network-to-Network Interface (NNI) until after a similar process in the terminating network, the session finally reaches the recipient party over the UNI. The role of North Bound Interface (NBI) on the IMS Application Servers is to extend the standardized services with further application logic plus for being used by other service execution entities. The IMS Service Control (ISC) APIs can also be used to connect application servers to IMS. For example, let us say that a new service may need a non-standard client. Say, it is a prepaid gauge. In a prepaid gauge, the user gets a dash-board for Advice of Charge (AoC) where the user can get details of charges for a call. This application offers a downloadable client that uses the standardized Java interface in the

terminal (JSR281 and its derivatives) to employ its use. Being a downloadable client, this can potentially become a mass-market service.

The telecom community brings out powerful and flexible standards alongwith stable implementations to make sure that the service fulfills market expectations. This will be required from the standardized services from day one. On the other hand, for non-standardized services, it is more important to have shorter time to market to bring innovative service ideas to end users promptly. If a service becomes a success, it will gradually face the same requirements as the standardized services. If it has utilized standardized services as its base, it can easily grow and be offered to all users in the IMS arena.

19.11 Policy Management in IMS

Telecommunication services are real-time and require stringent QoS (Quality of Service) norms; however, Internet Protocol is not sensitive to QoS. Therefore, one of the main goal of IMS is to offer QoS over the IP network. This is achieved through policy management. Policies can be defined as guidelines, directions, rules, etc. To ensure guaranteed level of service, the policy management model in IMS offers need to exercise access control with respect to various classes of subscribers. This is necessary for the following major reasons:

- Performance and Throughput: The system (including user equipment and network) would be better able to finely control and manage the resources like bandwidth and signaling capability, and computational possessions.
- Ability to operate in mobile and highly volatile environments: This is a scenario where services have the ability to run efficiently in incoherent environments.
- Extended control and full management functionality: Unlike the existing methods, the IMS architecture should have full management system functionality enforcing the policies according to combined data that is gathered from the service network.
- Roaming: In a fully mobile environment, a subscriber moves across a number of networks and network operators. But, this may be a source of complications as different providers can have different service agreements with their customers (or subscribers) and therefore offer different service policy management functionalities. Any entity like service policy manager should exist to address these issues by making few assumptions of what the providers offer to the user.

We have already briefly come across the five major entities in any policy management model. Apart from the previously discussed—Access Requestor (AR), Policy Enforcement Point (PEP), Policy Decision Point (PDP), Policy Repository (PR), and the Network Decision Point (NDP); we also define State data and Trigger event. State data contains all the necessary data to describe the state of the managed system, including previously installed policies. To remain up to date, state data must be updated as frequently as possible. Trigger event is any event that causes a new policy decision to be made. For example, it can be a request from any other entity of the system or request from AR to change the codec of the ongoing multimedia streaming.

The relationships between these entities are depicted in Figure 19.7. As we can see, NDPs pass on the state and event information to the PDP. It is re-emphasized that it is the PDP which actually

does the decision making and not NDP. The figure also shows AR which transitively provides inputs to state data and behavior of which triggers decision making in PDP.

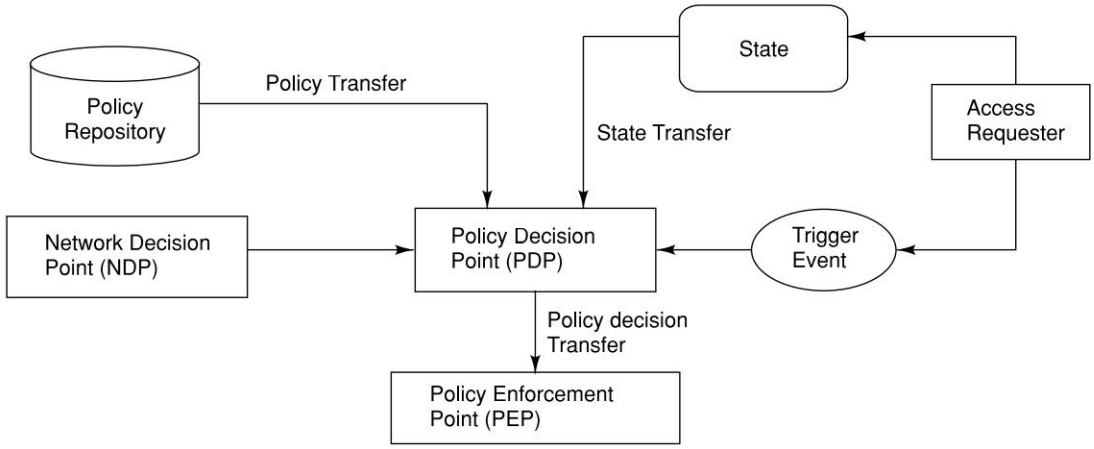


Figure 19.7 Generic Policy Management Model

When a user intends to use a specific service from service provider (SP), some policies need to be followed before the user is allowed to access the service. This could be due to QoS, security or any other policy. Such policies are implemented using PEPs and PDPs. The policies come to matter as an intuitive alternative of choices available to users, SPs and other such entities. When the user tries to access the service at SP, the policy manager sends a query to the PDP. The PDP checks in with the policy repository with the inputs as user credentials and the resource(s) requested. The PDP then replies the PEP whether to allow the resource to that user or not. The PEP can be implemented as part of S-CSCF function, or with S-CSCF and the enforcer plug in at the SP or may even be distributed to other entities such as MRFC, MEGACO, etc. This in reality depends upon the architecture and design of the IMS network.

19.12 Security in IMS

Like any other environment, security considerations in IMS will include functions like secure data transmission, confidentiality, authentication, non-repudiation, integrity, availability, anti-replay, and anti-fraud. IMS uses IP network as its lower layers; therefore, all security vulnerabilities and principles of IP are valid for IMS. However, in IMS a device will be mobile; therefore, we introduce the notion of Network Domain Security (NDS). This helps in provision of IP security between different domains and different nodes within a domain. A security domain can be defined as a network operated by a single administrative authority that intends to maintain a uniform security policy within that domain. However, in most of the cases, a security domain will correspond directly to an operator's core network. Security consideration of IMS therefore needs to address both intra-domain and inter-domain security.

In addition to domain security, IMS needs to address access security and data security. This entity for SIP-based services is a self-sufficient component in itself except that the security parameters for it are derivatives of UMTS Authentication and Key Agreement (AKA) Protocol. In the realm of IMS access security, we deepen our concepts on IMS access security for SIP services and IMS access security for HTTP services. Apart from SIP, user application data also needs to be secured. This is where IMS access security for HTTP services comes into the picture. Securing such an interface involves confidentiality and integrity protection for HTTP data being interchanged. Again, the key establishment and authentications are based on AKA.

3GPP defines following standards for security:

1. Security Architecture and Authentication and Key Agreement (AKA) [3GPP TS 33.102]
2. Network Domain Security (DNS) [3GPP TS 33.310]
3. Access Security for SIP-based Services [3GPP TS 33.203]
4. Generic Authentication Architecture [3GPP TS 33.220]
5. Access Security for HTTP-based Services [3GPP TS 33.222]

The IMS security architecture is depicted in Figure 19.8. Including SIP and the Application Server nodes, there are five different security associations and different needs for security protection for IMS that are numbered 1 through 5 in Figure 19.8.

Security Association 1: This association is for mutual authentication between the UE and the S-CSCF. The HSS collective (comprising the AAA and the associated databases) delegates the performance of subscriber authentication to the S-CSCF. However, the HSS is responsible for generating keys and challenges. The long-term key in the secure memory of the UE and the HSS is associated with the user private identity (IMPI). The subscriber will have one (network internal) user private identity (IMPI) and at least one external user public identity (IMPU). The user's subscription is authenticated by the S-CSCF (home service provider). The security association between the UE and the first access point into the operator's network (P-CSCF) is negotiated based on the protocol defined in RFC3329. The options supported by RFC3329 are: TLS (Transport Layer Security Described in Chapter 20), digest, IPsec-IKE (Internet Key Exchange), IPsec-MAN (Manually keyed IPsec without IKE), and IPsec-3GPP.

Security Association 2: This association provides a secure link and a security association between the UE and a P-CSCF for protection of the Gm reference point (described in Section 19.9). Data origin authentication is provided, i.e., the corroboration that the source of data received is as claimed. Integrity protection shall be applied between the UE and the P-CSCF for protecting the SIP signaling. The UE and the P-CSCF shall agree on security associations, which include the integrity keys that shall be used for the integrity protection. The integrity key shall be the IK, delivered by the S-CSCF to the P-CSCF during the user's IMS authentication process (component of the AKA Authentication Vector), The UE and the P-CSCF shall both verify that the data received originates from a node, which has the agreed integrity key. This verification is also used to detect if the data has been tampered with. Replay attacks and reflection attacks should be mitigated.

Security Association 3: This association provides security within the network domain internally for the Cx-interface.

Security Association 4: This association provides security between different networks for SIP capable nodes. This security association is only applicable when the P-CSCF resides in the VN. If the P-CSCF resides in the Home Network (HN) Security Association 5 applies.

Security Association 5: This association provides security within the network internally within the IMS subsystem between SIP capable nodes.

Note: This security association also applies when the P-CSCF resides in the HN.

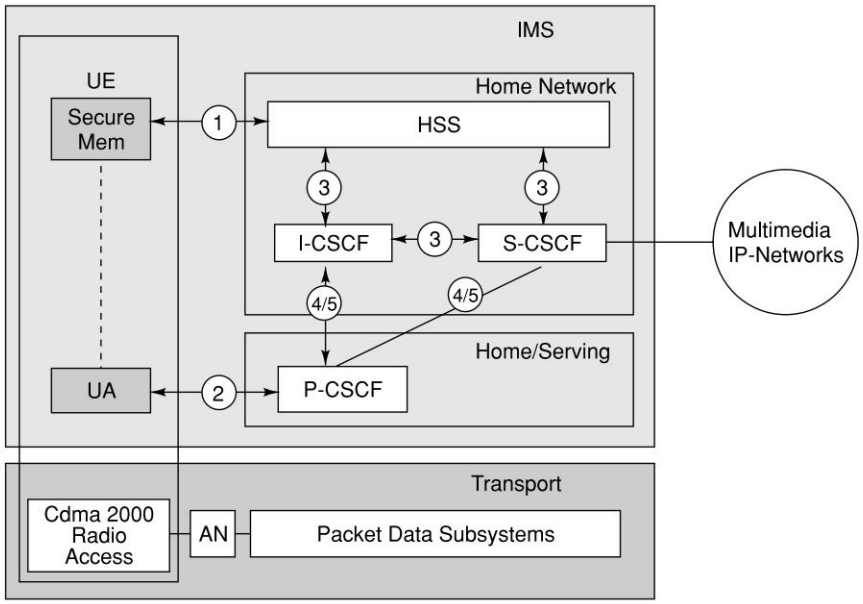


Figure 19.8 IMS Security Architecture

Inter-domain Security

Referring to Figure 19.8, interface 4 provides security between different networks for SIP capable nodes. The involved nodes shall be capable of IPsec. Privacy protection shall be applied with cryptographic strength greater than DES. Integrity protection shall be applied. IPsec may be used in either transport mode or tunnel mode; when used in tunnel mode, one or both of the network security domains may use Security Gateways. Security associations between nodes in different networks shall be negotiated using IPsec/IKE.

Intra-domain Security

The interfaces labeled 3 and 5 in Figure 19.8 are between SIP-capable nodes in the same network security domain. As this interface exists entirely within one network security domain, the administrative authority may choose any mechanism to secure this interface, including physical security where appropriate. Cryptographic methods of security, if applied, shall include both privacy and integrity protection, and be at least equivalent to IPsec using triple-DES and HMAC-MD5.

REFERENCES/FURTHER READING

1. All-IP Core Network Multimedia Domain Overview, 3GPP2 X.S0013-000-0 2 3 4 5 6 Version 1.0 Version Date: December, 2003.
2. Bertrand, *The IP Multimedia Subsystem–An Overview, GET/ENST Bretagne.*
3. Camarillo Gonzalo and Miguel A. Garcia-Martin (2004). The 3G IP Multimedia Subsystem (IMS), John Wiley & Sons, Ltd.
4. Digital cellular telecommunications system (Phase 2+), Universal Mobile Telecommunications System (UMTS); 3G security; Access security for IP-based services (3GPP TS 33.203 version 7.6.0 Release 7).
5. Digital cellular telecommunications system (Phase 2+) Universal Mobile Telecommunications System (UMTS); Generic Authentication Architecture (GAA); Access to network application functions using Hypertext Transfer Protocol over Transport Layer Security (HTTPS) (3GPP TS 33.222 version 7.2.0 Release 7).
6. February 2007.
7. Jeff Fried and Duane Sword, "Making IMS Work: Current Realities, Challenges and Successes", *Business Communications Review*, May 2006.
8. http://www.openimscore.org–Implementation of core elements of IMS/NGN in open source.
9. http://www.openmobilealliance.org–Committed to interoperability in mobile computing.
10. http://www.projectliberty.org–Committed to securing identity management in digital era.
11. http://www.3gpp.org–Collaboration for 3G mobile specifications based on evolution of GSM.
12. http://www.3gpp2.org–3GPP's opposite number in the CDMA2000 realm.
13. http://www.wikipedia.org–Online Encyclopedia.
14. 'IMS Influence on NGN Accounting Solutions', *Telecordia Digest*, December 2005.
15. JSR 116–SIP Servlet API, Java Community Process.
16. JSR 281 – IP Multimedia Subsystem (IMS) Services API for Java Micro Edition, Early Draft version 0.5, Java Community Process.
17. Kessler and Yeung, IMS Client Platform and IMS end to end, Sun Microsystems.
18. Kurapati Krishna (2006). 'Protecting IMS networks from attack', *Business Communications Review*, September.
19. Poikselka, Mayer, Khartabil and Niemi (2006). *IMS Concepts and Services*, John Wiley and Sons Ltd, England.
20. Osterman Sauli, *Master's Thesis, Combining Circuit and Packet Based Services in Converging Networks*, Helsinki University of Technology.
21. Siemens, Security protocols for the use of HTTP at the Mt reference point in the IMS, 3GPP TSG SA WG3 Security, Berlin, Germany , 6–9 May 2003.
22. *Services in the IMS ecosystem*, Ericsson White Paper, 285 23-3109 Uen Rev A.
23. Services in the IMS ecosystem, Ericsson.

24. Universal Mobile Telecommunications System (UMTS) 3G security; Security architecture (3GPP TS 33.102 version 7.1.0 Release 7).
25. Universal Mobile Telecommunications System (UMTS) Network domain security; Authentication framework (NDS/AF) (3GPP TS 33.310 version 7.1.0 Release 7).
26. Xylomenos G., V. Vogkas, and G. Thanos (2007). 'The Multimedia Broadcast/Multicast Service', *Wireless Communications and Mobile Computing.*

REVIEW QUESTIONS

Q1: What is IMS? Explain its evolution.
Q2: How is IMS different from GSM, GPRS, CDMA, etc.?
Q3: Write a short note on the history of IMS.
Q4: Why is IMS necessary in the present world where convergence is happening at a rapid pace?
Q5: What are the motivations of using IMS?
Q6: Describe the IMS architecture and service planes.
Q7: What classes of protocols are used in IMS? What protocols are central to those classes of protocols in IMS?
Q8: Describe the major protocols used in IMS.
Q9: How is Diameter effective as AAA protocol?
Q10: Why is SIP chosen as signaling protocol in IMS?
Q11: Explain how media is important to the concept of IMS.
Q12: Explain policy control protocols in IMS.
Q13: Explain the building blocks in IMS networks.
Q14: Explain how SLF is useful.
Q15: What are CSCFs? Explain their types and then their individual significance.
Q16: Explain how identities are managed in IMS.
Q17: Explain call flow in IMS.
Q18: Write a note on how charging takes place in IMS.
Q19: Explain the difference between Offline and Online Charging. How are they carried out in the context of IMS networks?
Q20: What are reference points in IMS? Explain the commonly used reference points in IMS.
Q21: Write a note on policy management in IMS.
Q22: How is security handled in IMS? What are the issues involved?

CHAPTER 20

Security Issues in Mobile Computing

20.1 INTRODUCTION

Mobile computing is pervading our society and lifestyles very fast. Mobile computing with networked information systems help increase productivity and operational efficiency. This however, comes at a price. Mobile computing with networked information systems increase the risks for sensitive information supporting critical functions in the organization which are open to attacks.

The fundamental premise of mobile computing is that the information will be accessed from outside the organization. As long as the information is within the four walls, the environment will be better known. It may be easier to control this environment and make it secure. When the information or computing environment is outside the controlled environment we do not have much control either from its users or usage patterns. Today, all the computers of the world are interconnected through extranet. Moreover, in a majority of cases, mobile computing uses wireless networks. Wireless media works on the principle of broadcast; information is radiated to everyone within the radio wave range, thus increasing security threats. Unlike a physical attack, cyber attacks can be replicated quite easily. Therefore, unless special care is taken, all systems are open to attack. This chapter discusses different techniques to secure information over the mobile computing environment.

20.2 INFORMATION SECURITY

In any defense system, we need to know our enemy. We also need to determine possible areas—weak points, vulnerabilities—where the enemy may attack. We need to build a defense system around these vulnerabilities. To build an information security system, we need to answer the following questions:

- Who is the enemy?

- What are the vulnerabilities? What are the weak links in the system?
- What could be the possible exploitation of these vulnerabilities by the resulting attacks?
- What needs special protection?
- To protect our assets from attack, we need to build a security system. How much does the security system cost in terms of money, resource and time?
- When the security system is deployed, to what extent will it affect the openness and add to inconvenience?
- Is prevention better that cure? If prevention is expensive or impractical, what is the strategy to recover from the loss following an attack?

There is no absolute security. What may appear to be absolute security in one context may not be absolute security in another. Therefore, while building a security system, we need to arrive at a proper balance between the answers emerging from the above questions. In a mobile environment, the user roams through different networks with heterogeneous security infrastructure. In such an environment where device mobility and network mobility is a necessity, offering homogenous service over heterogeneous devices and networks is the key. In such an environment again, weak security link from a wireless network could become a point of vulnerability for the entire system. Therefore, in a mobile computing environment, it is necessary to have a robust security and trust infrastructure.

20.2.1 Attacks

A security system is a system to defend our assets from attacks. In the physical world, these attacks are carried out at the weak points in the defense system. Likewise in the electronic world, attacks are carried out at the point of vulnerability. When the vulnerability is exploited for some interest or selfish motive, it is an attack on the system. Of course there could be occasions where the vulnerability is exposed by accident as well. Where the vulnerability is exploited, there is a loss. This loss can be either of static information asset (static asset) or an information asset in transit (dynamic asset). If we look at an information system, static assets cover a large portion of the asset base. All the databases, files, documents, etc., in the computers fall in this category. Examples of attacks on static asset are virus deleting files in a computer or jamming a network. An example of an attack on a dynamic asset is the theft of a credit card number while a user is doing a credit card transaction on the web.

Attack on dynamic assets can be of the following types (Fig. 20.1):

- *Interception:* An unauthorized party gaining access to an asset will be part of this attack. This is an attack on confidentiality like unauthorized copying of files or tapping a conversation between parties. Some of the sniffing attacks fall in this category.
- *Modification:* An unauthorized party gaining control of an asset and tampering with it is part of this attack. This is an attack on integrity like changing the content of a message being transmitted through the network. Different types of man-in-the-middle attacks are part of this type of attack.
- *Fabrication:* An unauthorized party inserts counterfeited objects into the system; for example, impersonating someone and inserting a spurious message in a network.
- *Interruption:* An asset is destroyed or made unusable. This is an attack on availability. This attack can be on a static asset or a dynamic asset. An example could be cutting a communication

line or making the router so busy that a user cannot use a server in a network. These are all Denial of service attacks.

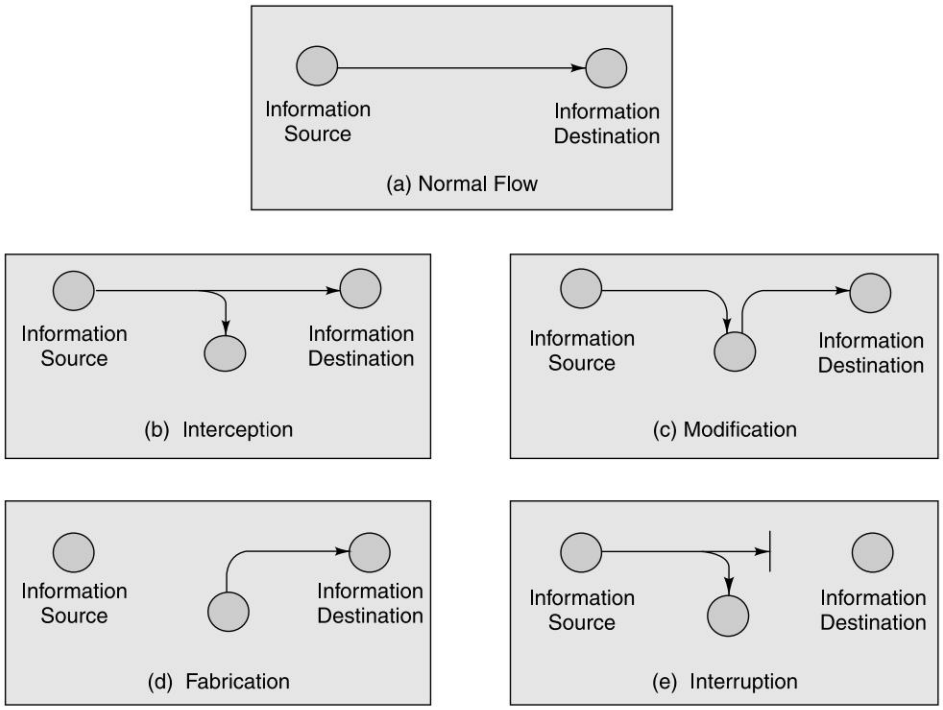


Figure 20.1 Types of Attacks

Attacks on static assets can be of the following types:

- *Virus and Worms:* These are a type of program that replicates and propagates from one system to another. Most of the virus do malicious destructive functions in the system.
- *Denial of Service:* These are attacks on the system to prevent legitimate users from using the service.
- **Intrusion:** These are people or software, which enter into computer systems and perform function without the knowledge of the owner of the asset. These are also called hackers.
- *Replay Attack:* In a replay attack the opponent passively captures the data without trying to analyze the content. At a later time, the same is used in the same sequence to impersonate an event and gain unauthorized access to resource.
- *Buffer Overflow Attacks:* In a buffer overflow attack, the vulnerability of an executable program is exploited to force a stack overflow condition, inducing the program counter of the process to change. The program counter is then manipulated to do the work for the attacker.

- *Trapdoor Attacks:* These are exploitations of some undocumented features of a system. Undocumented functionality are designed to debug, service, support or take control of the system.

A security system needs to be so designed that the system is able to counter and recover from attacks.

20.2.2 Components of Information Security

For centuries, information security was synonymous with secrecy. The art of keeping a message secret was to encrypt the message and thus hide it from others getting to know of it. However, in today's netcentric electronic world, the taxonomy of information security is much beyond encryption. Information security needs to cater to all possible attacks related to confidentiality, integrity, availability, non-repudiation, authorization, trust and accounting (CIANATA). Confidentiality is the property where the information is kept secret so that unauthorized persons cannot get the information. Integrity is the property of keeping the information intact. Availability is the property of a system by which the system will be available to its legitimate users. Non-repudiation is the property by which the identity of both sender and receiver of the message can be identified and verified. Authorization is the property by which the user's properties can be associated to the information access. Trust is the property of expectation, confidence, and belief over time. Accounting is the property of calculating the fee for a service rendered.

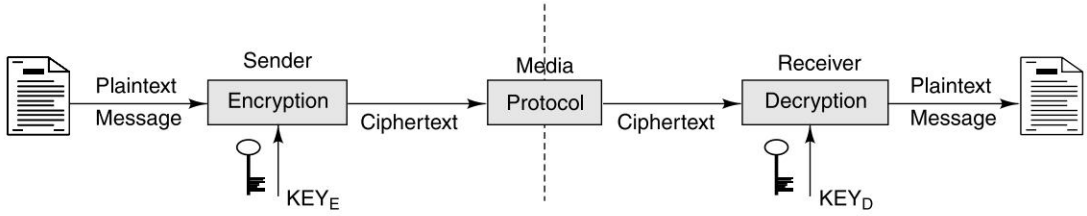


Figure 20.2 Encryption and Decryption with a Key

Confidentiality

Confidentiality is ensured through encryption of the data. To a person a comprehensible message is written in a particular language. The language can be English, Hindi, French or any other language. These messages are called plaintext or cleartext messages. Through encryption (or encipher) we disguise this message in such a fashion that it is no longer understandable by either a person or a machine. An encrypted message is called ciphertext. The process of converting a ciphertext back into plaintext is called decryption (or deciphering). Plaintext need not be a written text. It can even be an audio or video message as well. When leaders of two countries talk, the message is encrypted so that a man eaves dropping cannot make any sense of the conversation. The plaintext message can also be a data file in the computer disk. Figure 20.2 depicts the process of encryption and decryption.

In cryptography there are two components, viz., algorithms and protocols. A cryptographic algorithm is a mathematical function used for encryption and decryption, and protocol relates to the process and procedure of using algorithms. A protocol is the way algorithms are used to ensure

that the security is ensured and the system is less prone to attacks. In a security system the plaintext message is encrypted by using a key KEY_E. The encrypted message is then sent from the sender to the received through a media (wired, wireless, or even postal) using some protocol. The encrypted message is then decrypted using a key KEY_D to extract the original message. A cryptographic key is generally a large number. The range of possible values of a key is called keyspace. The larger the keyspace is, the more difficult it is for an attacker to guess the key and restore the original message. Therefore a larger keyspace makes a ciphertext more secure. This is similar to a lock. A conventional lock of 11 levers is more secure compared to a 7-lever lock.

The art of keeping message secure using the science of encryption and decryption is called cryptography. People who practise cryptography are called cryptographers. There are people who try to break the secrecy of encryptions. These are for many purposes; some are for research purposes to measure the strength of the security and some, for stealing the information. Some are hackers who try to break the security for fun or for a price. These people who try to break the secrecy of the cryptography are called cryptanalysts. The practice of cryptanalyst is called cryptanalysis. There is another science in security engineering. This is called steganography. Steganography is the science of hiding secret message in other messages so that the existence of the secret message is concealed; for example, sending some secret message by changing some bits in a large picture message. By looking at the picture, others will not be able to guess that in reality the picture is carrying a secret message.

Integrity

Integrity is to ensure the integrity of the message. Integrity is achieved by adding additional information into the message. This is done through checksums, message digests or digital signature. In a crypto system, the receiver of the message checks this extra information to verify whether the message has been tampered with. This is similar to a bank cheque. A cheque issued to a customer is honored only when the customer signs it. The cheque number and the signature are verified to ensure integrity. Integrity check is advised for both static asset and asset on transit.

Authorization

Authorization deals with privileges. In any transaction, there is a subject (a person) and an object (data items or file). The subject wants some function to be performed on the object. The privilege to an object is defined through ACL or Access Control List. ACL is used while allowing access to the object. The privilege on an object can be read, write, or execute. Besides objects there need to be privilege-based type of subjects. This is done through authorization.

Authorization is implemented through policy-based resource accessibility. In an organization (or society) where there is a hierarchy, there will be certain functions allowed to certain levels in the hierarchy. A clerk in a corporation may have authorization to approve an expense claim less than a specified threshold, supervisors might have a higher limit, and vice-presidents might have a still higher limit. Similarly, role-based security will be used when an application requires multiple layers of authorization and approvals to complete an action. Privilege management infrastructure together with the role-based authorization allows the administration and enforcement of user privileges and transaction entitlements. In the authorization process, users are checked to see if they have the required rights to access the resource. If they have been granted the required rights, they can access the resource, otherwise they are denied access.

Non-repudiation

Authentication and non-repudiation have some overlapping properties. Authentication is a process by which we validate the identity of the parties involved in a transaction. In non-repudiation we identify the identity of these parties beyond any point of doubt. Non-repudiation can be considered as authentication with formal record. These records will have legal bindings. Like a signature in a cheque, using digital signatures we achieve non-repudiation.

Availability

Media management is not within the scope of security protocols and algorithms. However, media management is part of the larger security framework. Media management is needed to ensure availability of service. For a message a confidentiality may be maintained; also, the integrity is intact but an attacker can manipulate the media to make sure that the message does not reach the destination. This is like there is no theft of power and power quality is good, yet someone blows the transmission line of the power grid.

Attack on availability happens for industrial espionage or from political motivation. During a festive season, one company may target to block the e-commerce site of a competition. In a social framework, someone may try to stop people's voice by using threats or other means of intimidation to compel the author to remove the web page. If these methods prove unsuccessful, various denials of service attacks can be launched against the site to make it impossible to access. In less high-profile cases, people often enjoy far less support for exposing corruption or criticizing employers and particularly litigious organizations. Also, there needs to be some way where terrorist organizations or dictators cannot block the mass opinion. This field of research area is called Censorship-resistant Publishing. Censorship-resistant publishing is achieved through document entanglement.

Trust

Computers rely on user authentication and access control to provide security. Within a network, it may be safe to assume that the keyholder is authentic, and the keyholder is using the key assigned to him or her. However, these strategies are inadequate for mobile computing environments with high level of flexibility. Mobile computing lacks centralized control and its users are not all predetermined. Mobile users expect to access resources and services anywhere and anytime. This leads to serious security risks and access control problems. To handle such dynamic everchanging context, trust-based security management is necessary. Trust involves developing a security policy, assigning credentials to entities, verifying that the credentials fulfill the policy. Also, we need delegation of trust to third parties, and reasoning about users' access rights.

Accounting

For any service, the service provider needs to be paid. The service can be either a content service or a network service. Accounting and billing is a very critical aspect in mobile computing environment. Accounting is the process by which the usage of the service is metered. Based upon the usage, the service provider collects the fee either directly from the customer or through the home network. This will be true even if the user is roaming in a foreign network, and using the services in the foreign network.

RADIUS (Remote Authentication Dial In User Service) protocol (RFC 2865) has been in use for a long time for the AAA (Authentication, Authorization, and Accounting) functions in Internet.

With the demanding service requirement of mobile computing, it is now apparent that RADIUS is incapable of supporting all these complexities. A new protocol called Diameter (RFC 3588) has been released to address the AAA needs for data roaming and mobile computing. Diameter can work in both local and roaming AAA situations.

20.3 Security Techniques and Algorithms

Generally the encryption algorithms are divided into two main groups. These are symmetric key encryption and public key encryption. In a symmetric key encryption, the key used for decryption is the same as the key for encryption. In some cases of symmetric encryption, even the algorithm used for encryption and decryption is the same. In the case of public key algorithms, the key used for decryption is different from the key used for encryption.

20.3.1 Stream Ciphering and Block Ciphering

In stream cipher, a bit or a byte is taken at a time and encrypted. The algorithm looks at the input plaintext as a stream of bits and encrypts them one bit (or byte) at a time as the stream progresses. In this technique, the length of the plaintext and the key size will be same. Wireless LAN (WiFi) uses stream cipher. In this methodology, the key has to be unique for every encryption. If the same key is used for multiple packets, and these packets can be captured, there is vulnerability. The other technique is block cipher. In a block cipher, one block of plaintext is taken as a whole and used to produce a ciphertext block of equal length. Typically a block of 64 bits (8 octets) or 128 bits (16 octets) is used for block cipher. Majority of cryptosystems use block cipher.

20.3.2 Symmetric Key Cryptography

In a symmetric key cryptography, the same key is used for both encryption and decryption. This is like a lock where the same key is used to lock and unlock. In cryptography, symmetric key algorithms are in use for centuries; that is why symmetric key algorithms are called conventional or classical algorithms as well. In this type of encryption, the key is secret and known only to the encrypting (sender) and decrypting (receiver) parties. Therefore, it is also known as a secret key algorithm. Some authors refer to symmetric key cryptography as shared key cryptography as well. This is because the same key is shared between the sender and the receiver of the message. The unique key chosen for use in a particular transaction makes the results of encryption unique. Selection of a different key causes the cipher that is produced for any given set of inputs to be different. The cryptographic security of the data depends on the security of the algorithm used and the key used to encipher the data. The strength of the security depends on the size of the key. Unauthorized recipients of the cipher who know the algorithm but do not have the correct key cannot derive the original data algorithmically. However, anyone who does have the key and the algorithm can easily decipher the cipher and obtain the original data. Symmetric key algorithms are much faster compared to their asymmetric (public key) counterparts.

In a symmetric key cryptography, there are four components. These are plaintext, encryption/decryption algorithm, secret key (key for encryption and decryption), and the ciphertext. In Figure

20.2, if we make $KEY_E = KEY_D$, this becomes a symmetric key algorithm. There are many symmetric key algorithms. The most popular symmetric key algorithms are:

DES: Data Encryption Standard is the most widely used, researched and has had the longest life so far.

3DES: This is a modification of DES. In this algorithm, DES is used three times in succession.

AES: Advances Encryption Standards, this is the current accepted standard for encryption by FIPS (Federal Information Processing Standards) of the US.

Skipjack/FORTEZZA: This is a token-based symmetric algorithm used by defense personnel in the US.

DES (Data Encryption Standard)

In the late 1960s, IBM set up a research project in computer cryptography led by Horst Feistel. In 1971, the project concluded with an outcome of an algorithm named Lucifer. The original algorithm used 64-bits block and 128-bits key. IBM reduced the length of the key to fit the algorithm into a single chip. This algorithm was adopted in 1977 by NIST (National Institute of Standards and Technology) as the data encryption standard (DES). A DES key consists of 64 bits of which 56 bits are randomly generated and used directly by the algorithm. The other 8 bits are used for error detection and not for encryption.

DES employs the principle of scrambling and substitution. These processes are repeated a number of times with keys to ensure that the plaintext is completely transformed into a thoroughly scrambled bit stream. The DES can be divided into the following major functions. These are:

- Permutations of bits in a block. This is the first and last step in DES. In this step the 64-bit plaintext block is rearranged through Initial Permutation **IP**. This is done through a 64-bit register where the bits of the input block are scrambled in a particular fashion. As the last step, the reverse permutation is done through **IP–1**.
- A key dependent computation. This includes multiple rounds (iteration) of transformation through combination of permutation and substitution. This is in the core of the encryption function.
- Swapping of half blocks of data in each round.
- Key schedule; this breaks the 56-bit key into two 28-bit subkeys and use them to compute the bits in data blocks. In each iteration, the bits within the subkey are shifted to generate a new subkey.
- The key-dependent computation is run through 16 rounds. Each round uses the data from the previous round as input.

The beauty of DES algorithm is that the same algorithm is used for both encryption and decryption. DES demonstrates a very high avalanche effect. In an avalanche effect one bit of change in either the input data or the key changes many bits in the output. For example, in DES one bit of change in the input data changes 34 bits, whereas one bit of change in the key affects 35 bits.

3DES (Triple DES): With the increase of processing power available in PCs, 56 bits of key became vulnerable for attack. Therefore, to protect the investment and increase security 3DES (commonly known as Triple DES) was proposed. 3DES uses the same DES algorithm three times

in succession with different keys. This increases the keysize resulting in higher security. Also, as the fundamental algorithm in 3DES is practically the DES, it is easily adaptable without additional investment. There are two different flavors of 3DES. One uses two 56-bit key and the other uses three 56-bit key. By using three 56- bit key, the effective security can be increased to the key size, to 168 bits. Till today 3DES is the most widely used algorithm for symmetric cryptography.

AES (Advanced Encryption Standard)

We have discussed that the strength of security of cryptographic algorithms depends on the size of the key. The larger the size of the key, the longer it takes to decipher the encrypted data through brute force. With GHz of computing power easily available, 56-bit key size is found to be unsafe today. To overcome these challenges, 3DES became popular. However, 3DES was quite slow. Also, scientists found that the 64-bit block which both DES and 3DES use, may not be the best. A higher block size is desirable from efficiency and security point of view.

To overcome these drawbacks, in 1997 NIST (National Institute of Standards and Technology) in the US issued a call for algorithms for advanced encryption standard or AES. According to the call for proposal, the AES standard was to have equal or better security compared to 3DES and more efficient than the 3DES. NIST also specified that AES had to be a symmetric cipher with block size of 128 bits. Also, it has to support keys of size 128-bits, 192-bits, and 256-bits. Many algorithms competed for the AES standard. Following a rigorous evaluation process in November 2001, NIST selected the Rijndael as the AES algorithm. Rijndael is named after two researchers from Belgium who developed the algorithm. They were Joan Daemen and Vincent Rijmen. Rijndael was designed to have the following characteristics:

- Resistance against all known attacks.
- Design simplicity.
- Speed and code compactness on a wide range of platforms.

Like DES, AES also uses permutation and substitution. However, AES does not use Feistel structure. In a Feistel structure, one half of the data block is used to modify the other half of the data and then swapped.

20.3.3 Public Key Cryptography

In symmetric key encryption we use the same key for both encryption and decryption. In public key cryptography we use two different keys, one key for encryption and another for decryption. As there are two different keys used, this is also called asymmetric key cryptography. The development of public key cryptography can be considered as the greatest advance in the history of cryptography. Public key cryptosystem is based on mathematical functions rather than permutation and substitution. However, it is not true that the public key cryptosystem is more secure for general purposes. There is nothing in principle, which makes one algorithm superior to another from the point of view of resisting cryptanalysis. It is computationally infeasible to derive the decryption key given only the encryption key and knowledge of the cryptographic algorithm. The encryption key and the decryption key together form a key pair. One of these keys from the key pair is made public and the other one kept private or secret. That is why this algorithm is called public key cryptosystem.

Whitfield Diffie and Martin Hellman in 1976 came up with the principle of asymmetric key or public key cryptography. Public key cryptography proposed by Diffie and Hellman solved two difficult problems of key distribution and digital signature in cryptography. In public key cryptography, there are six components (Fig. 20.3). These are:

- *Plaintext:* This is the human readable message or data given to the public key algorithm as input for encryption.
- *Ciphertext:* This is the scrambled data produced as output of the encryption algorithm. This is a unique data and depends only on the unique key used for encryption.
- *Encryption Algorithm:* This is the algorithm that does computation and various transformations on the input plaintext. The output of the transformation is too garbled to be decipherable for an intruder.
- *Decryption Algorithm:* This algorithm does the reverse function of the encryption algorithm. This function accepts the ciphertext as input and does some transformation on the data so that the original data is recovered.
- *Public Key:* This is one of the keys from the key pair. This key is made public for anybody to access. This key can be used either for encryption or decryption.

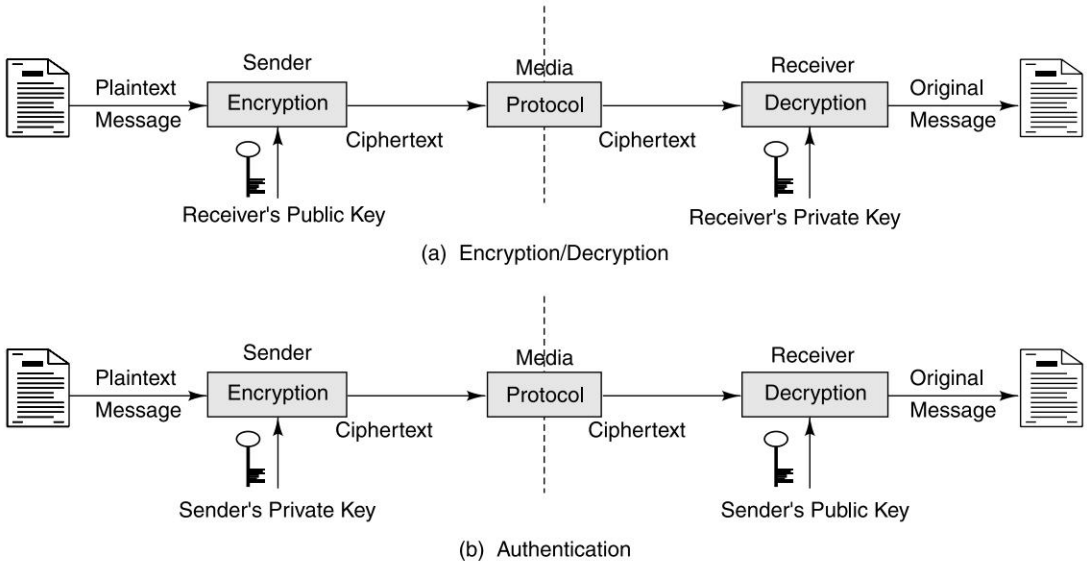


Figure 20.3 Public Key Cryptography

- *Private Key:* This is the other key from the key pair. This key is called the private key, because this is kept secret. This can be used either for encryption or decryption.

There are three public key cryptosystems most widely used today. These are Diffie Hellman, RSA, and Elliptic carve.

The methodology used for encryption of data and the digital signature is different. During the encryption, the sender uses the public key of the receiver. This is because only the receiver should

be able to decrypt the message using his or her own secret private key. If there is a surrogate who is able to intercept the encrypted message, he will not be able to decrypt the message, as the key required to do so is the private key. The receiver's private key is kept secret with the receiver. The methodology used for authentication or digital signature is just the reverse. In case of signing the transaction, the private key of the sender is used by the sender. The receiver uses the public key of the sender to read the signature. This authenticates that the transaction was indeed done by the sender.

20.3.4 Key Exchange Algorithm

Whitfield Diffie and Martin Hellman first introduced the notion of public key cryptography in 1976. In the Diffie Hellman technique, secret keys are never exchanged. However, the technique allows two parties to arrive at a secret key through the usage of public keys. Communicating parties select a pair of private and public keys. Public keys are exchanged. The shared secret key is generated from the private key and the public key of the other party.

Let us assume that there are two parties A and B. A and B choose some prime number p and another number g less than p. These numbers are selected and made available to both A and B in advance. The steps followed in Diffie Hellman algorithms for key generation are as follows:

1. Let these p and g be: $p = 13$ and $g = 3$;
2. A chooses a random number *SA*. This number is kept secret as a private key with A. Let this number be 5.
3. B chooses a random number *SB*. This number is kept secret as a private key with B. Let this number be 7.
4. A takes g and raises it with his secret key *SA* modulo p. This will be $TA = (g \text{ ^ } SA) \bmod p \Rightarrow (3 \text{ ^ } 5) \bmod 13 = (243) \bmod 13 = 9$. This number 9 is A's public key. A already has chosen 5 as his private key.
5. B takes g and raises it with his secret key *SB* modulo p. This will be $TB = (g \text{ ^ } SB) \bmod p \Rightarrow (3 \text{ ^ } 7) \bmod 13 = (2187) \bmod 13 = 3$. This number 3 is B's public key. B has already chosen 7 as his private key.
6. Public keys of A and B are exchanged. This means A sends the public key 9 to B and B sends his public key 3 to A over a public channel like Internet.
7. A takes B's public key and raises it with his own private key mod p. Therefore, we now have $KA = (TB \text{ ^ } SA) \bmod p \Rightarrow (3 \text{ ^ } 5) \bmod 13 = (243) \bmod 13 = 9$.
8. B now takes A's public key and raises it with his own private key mod p in a similar fashion as A. The result will be $KB = (TA \text{ ^ } SB) \bmod p \Rightarrow (9 \text{ ^ } 7) \bmod 13 = (4782969) \bmod 13 = 9$.
9. The value of $(TA \text{ ^ } SB) \bmod p = (TB \text{ ^ } SA) \bmod p = 9$. Though *KA* and *KB* have been calculated by A and B independently; it will always be equal. Therefore, these keys *KA* and *KB* can now be used by A and B as the shared key for payload encryption.

Neither A nor B shared their secret key for use in symmetric encryption, but arrived at that using some properties of modulo arithmetic with prime numbers. The example above may look trivial. However, when these numbers are large, nobody can calculate the key just by knowing p, g and *Sx* in a reasonable period of time. An eavesdropper could not compute discrete logarithm, i.e., figure out *KA* based on seeing *SB*.

RSA

RSA is named after its inventors R.L. Rivest, A. Shamir and L. Adleman. It is a public key algorithm that does encryption/decryption, authentication, and digital signature. The key length is variable and the most commonly used key size is 512 bits. The key length used in India by CCA (Controller of Certifying Authorities) is 2048 bits. Key length can be large for higher security; the key length can be smaller for better efficiency. The plaintext data block is always smaller than the key length. However, the ciphertext block is the same as the key length. RSA is much slower than symmetric key encryption. That is why RSA is generally not used for payload encryption. RSA is used primarily for encrypting a secret key for key exchange.

The RSA algorithm works as follows:

1. Choose two prime numbers p and q.
2. Multiply p and q to generate n. n will be used as the modulus.
3. Calculate $\Phi(n) = (p - 1) \cdot (q - 1)$. $\Phi(n)$ is the Euler's totient function. $\Phi(p)$ is the number of positive integers less than p and relatively prime to p.
4. Choose a number e such that it is relatively prime to $\Phi(n)$.
5. Find d such that it is multiplicative inverse of e; $d = e^{-1} \bmod \Phi(n)$.
6. (e, n) is the public key and (d, n) is the private key
7. To encrypt we use the formula (Ciphertext block) = (Plaintext block)^e mod n.
8. To decrypt we use the formula (Plaintext block) = (Ciphertext block)^d mod n.

Let us take an example where we choose two prime numbers $p = 7$ and $q = 17$.

Calculate $n = p \cdot q = 7 \cdot 17 = 119$

Find the value of $\Phi(n)$ using the formula $\Phi(n) = (p - 1) \cdot (q - 1) = (7 - 1) \cdot (17 - 1) = 6 \cdot 16 = 96$.

Now we need to select an e. e will be relatively prime to $\Phi(n)$ and less than $\Phi(n)$. We can see that 2, 3, 4 have factors with 96, therefore, are not relatively prime. Whereas, 5 is relatively prime to 96. Therefore, we can choose e to be 5.

We know that $d \cdot e = 1 \bmod \Phi(n)$, which in other words $d \cdot e = ((Y \cdot \Phi(n) + 1) \bmod \Phi(n)$. To find the value of d, we use the formula $((Y \cdot \Phi(n)) + 1)/e$. Replace Y with 1 then 2 then 3 and so on until we get an Integer. When we set $Y = 4$, the equation evaluates:

$$d = (4 * 96 + 1)/5 = (384 + 1)/5 = 385/5 = 77$$

Therefore, we get d = 77. We have just generated our key pair. The public key is (5, 119) and private key is (77, 119). We can now use this to encrypt and decrypt values.

To encrypt we use the formula

(Ciphertext block) = (Plaintext block)^e mod n. Assuming that the plaintext block is 8 bits long and the value is 65. Therefore, the ciphertext will be (65 ^ 5) mod 119 => (1160290625) mod 119 = 46. To decrypt, we use the formula (Plaintext block) = (Ciphertext block)^d mod n => (46 ^ 77) mod 119 = (1.07734063167916956809383545838 5e+128) mod 119 = 65.

The example above may look trivial and someone may think that by knowing (5, 119) one can easily find out d. This is almost impossible if the numbers are large, for example 128 bits long. Also, to know the private key, the eavesdropper needs to evaluate p and q from n. The eavesdropper has to factorize the number n to get the two large prime numbers, which is extremely hard even in a huge timeframe. RSA uses the complexity in prime factorization.

Elliptic Curve

A majority of the products and standards that use the public key cryptography use RSA for encryption, authentication, and digital signature. Due to extensive research in cryptanalysis in

RSA and increase in availability of computing power, some vulnerabilities of RSA have been discovered. There are subexponential algorithms available today for breaking RSA and Diffie-Hellman algorithms. To overcome these threats, the size of the RSA key has been increasing over time. This puts a tremendous demand on computing power. Elliptic Curve Cryptography (ECC) has shown a lot of promise for higher security with lesser resource. Elliptic curve cryptography was proposed by Victor Miller and Neal Koblitz in the mid 1980s. Till date there is no subexponential algorithms available to break ECC. An elliptic curve is the set of solutions (x, y) to an equation of the form $y^2 = x^3 + ax + b$, together with an extra point O which is called the point at infinity.

ECC is believed to offer a similar level of security with a much smaller size of key. For example, it is claimed that the level of security that 1024 bits of RSA provide can be achieved by 160 bits of ECC. A 210-bit key of ECC is equivalent to 2048 bits of RSA. This makes ECC very attractive for small footprint devices like cell phones or PDAs.

20.3.5 Hashing Algorithms

Hashing functions are one-way functions used for message digests. Hash function takes an input data of any size and produces an output stream of some fixed size. The outputs are collision free. This means that two different inputs will not produce the same output. It is also not possible to derive the input from a known output. This means that if we have a message digest, it is impossible to derive the original message. The most commonly used hash functions are MD5 and SHA-1.

MD5

MD5 (Message Digest version 5) hashing algorithm is described in RFC 1321. The MD5 algorithm is an extension of the MD4 message-digest algorithm and is slightly slower than MD4. The MD5 algorithm takes a message of arbitrary length as input and produces a 128-bit "message digest" as output. The algorithm processes 512 bits of the input message in blocks. The digest produced by the algorithm can also be considered as a "fingerprint" of the message. It is conjectured that it is computationally infeasible to produce two messages having the same message digest. It is also conjectured that it is computationally infeasible to produce any message having a given message digest. The MD5 algorithm is intended for digital signature applications in a public key cryptosystem.

SHA

The Secure Hash Algorithm (SHA) was developed by the NIST (National Institute of Standards and Technology). SHA was first published in 1993. Later in 1995, a revised version of the algorithm was published as SHA-1. SHA processes input in 512 bits block and produces 160 bits of output. Like MD5, SHA-1 is also based on MD4 algorithm. As both MD5 and SHA-1 are based on MD4, they are quite similar in nature. However, as SHA-1 generates a longer digest of 160 bits compared to 128 bits by MD5, it is considered to be more secure.

MAC

MAC stands for Message Authentication Code. MAC is used to do the integrity check on the message. A secret key is used to generate a small fixed size data block from the message. This is similar to a checksum of the message. Both the sender and the receiver share the same secret key for MAC. When the sender has a message to be sent to the receiver, the message is sent along with the MAC. The receiver receives the message; and calculates the MAC from the message and the

shared key. The receiver checks the MAC received from the sender. If they are the same, the message is considered to be in perfect state. HMAC is another mechanism for message authentication using cryptographic hash functions like MD5, or SHA-1, in combination with a secret shared key. HMAC has been defined in RFC 2104. The cryptographic strength of HMAC depends on the properties of the underlying hash function.

20.3.6 Block Cipher Modes of Operation

Theoretically, an ideal ciphering algorithm must make any plaintext into ciphertext that is completely random in nature - occurrence of every character in the cyphertext will have equal probability. If the ciphertext is random, a hacker cannot even guess the algorithm and get into the original plaintext. Therefore, to have a secure system, the approach always will be to make a known text look random.

In a cryptography algorithm, any key operated on same plaintext will always create the same ciphertext. This characteristic can be exploited by a cryptanalyst or a hacker to locate a repeating pattern in the ciphertext and then launch an attack. To avoid this, and to make the ciphertext appear at random, cipher block chaining (CBC) mode is normally used along with a ciphering algorithm. In the CBC mode, each block of plaintext is XOR-ed with the previous ciphertext block before being encrypted. This technique ensures that the same plaintext will generate a different ciphertext. To make each message unique, an initialization vector is used in the first block. The reverse technique is used to get the plaintext back on the receiving end. The encryption and decryption process in CBC is illustrated in Figure 20.4.

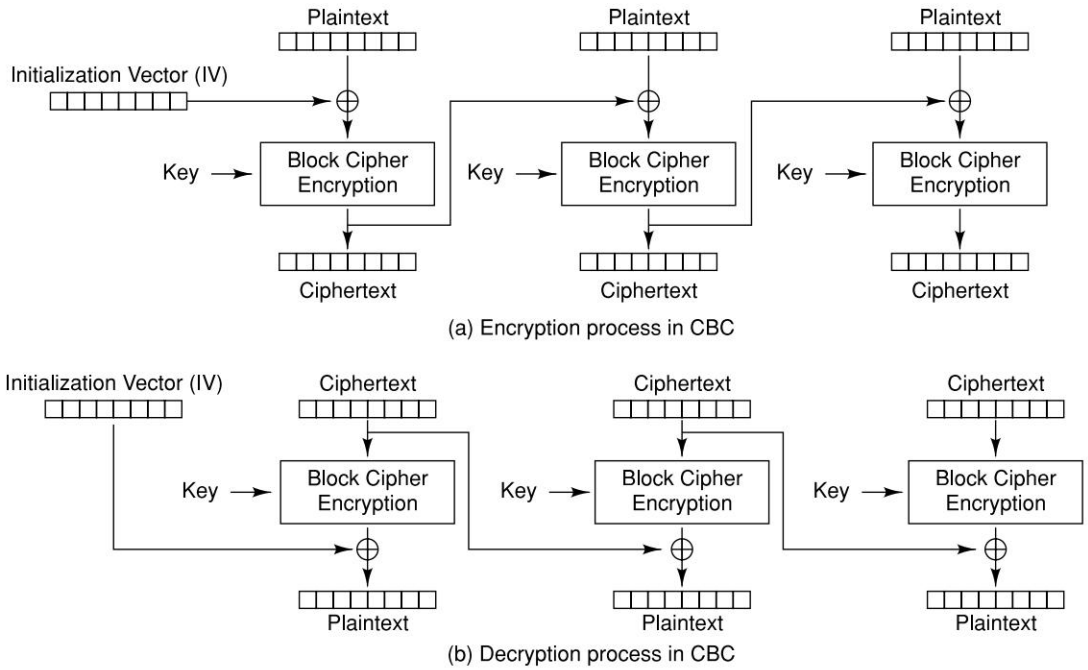


Figure 20.4: Cipher Block Chaining (CBC)

Similar techniques are used to make ciphering look random in few other algorithms with some variations. These are,

- Propagating cipher-block chaining (PCBC)
- Cipher feedback (CFB)
- Output feedback (OFB)
- Counter (CTR)

20.4 SECURITY PROTOCOLS

To provide confidentiality, integrity etc., we need to use different algorithms. However, we need to device protocols that will use these algorithms in such a fashion that vulnerabilities are eliminated and security is ensured. The protocol needs to be so robust that a masquerader is unable to get the message being sent. The protocols need to ensure that if the masquerader is able to modify the message, we can detect it. There are many protocols for secured communication. One such protocol is depicted in Figure 20.5. However, the most popular protocol is SSL (Secured Socket layer–Section 20.4.1). SSL was originally developed by Netscape. The Internet standards for TLS (Transport Layer-Security–Section 20.4.2) and WTLS (Wireless Transport Layer Security–Section 20.4.3) have been derived from the SSL protocol.

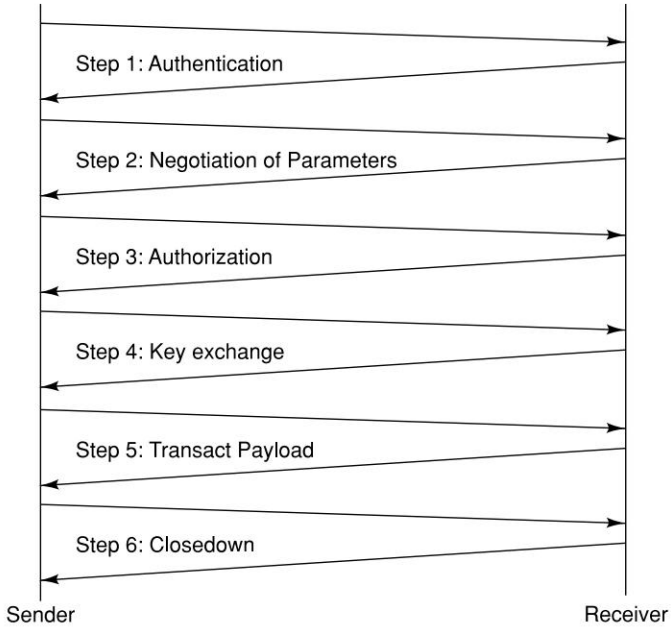
20.4.1 Secured Socket Layer (SSL)

The Secured Socket Layer or SSL protocol is used to provide security of data over public networks like Internet. It runs above the TCP/IP protocol layer and below higher level protocols such as HTTP or IMAP (Fig. 20.5). SSL allows both machines (server and the client) to establish a secured encrypted channel so that all the data transacted between them are confidential and tamper-resistant.

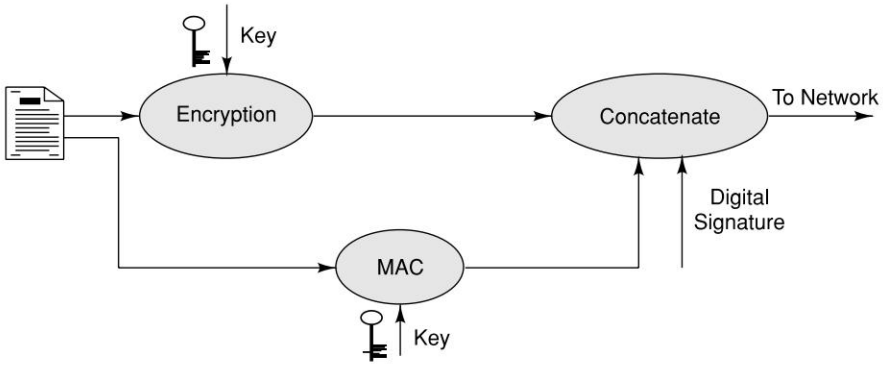
Public-key encryption provides better authentication techniques. On the other hand, symmetric key encryption is much faster than the public key encryption. The SSL protocol uses a combination of both public key and symmetric key encryption. An SSL session begins with SSL handshake. SSL handshake allows the server to authenticate itself to the client using public-key techniques. Optionally, the handshake also allows the client to authenticate itself to the server. It then allows the client and the server to cooperate in the creation of symmetric key. It then uses this shared key for payload encryption, decryption, and tamper detection during the session that follows.

20.4.2 TLS

Transport Layer Security or TLS in short is a security protocol to offer secured communication at the transport layer. TLS protocol is the Internet standard and based on the SSL 3.0 protocol specification. According to RFC 2246 (TLS Protocol Version 1.0), the primary goal of the TLS protocol is to provide privacy and data integrity between two communicating applications. At the lower levels, TLS uses TCP transport protocol. The TLS protocol is composed of two layers: the TLS Handshake Protocol and the TLS Record Protocol.



(a) Protocol for secure communication



(b) Message confidentiality, integrity, and non-repudiation

Figure 20.5 Security Protocols

The TLS Handshake Protocol provides connection security that has three basic properties:

1. Peer's identity can be authenticated using asymmetric or public key cryptography (e.g., Diffie-Hellman, RSA, etc.).

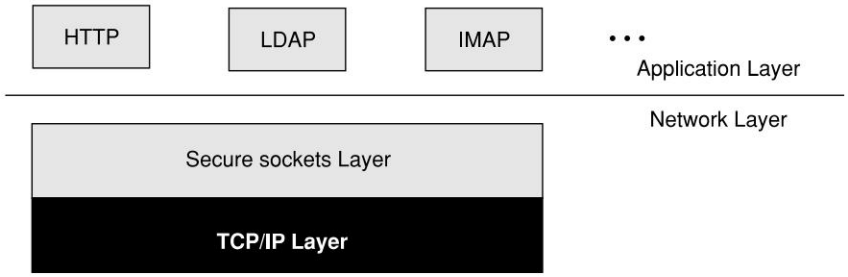


Figure 20.6 SSL Layer

2. The negotiation is reliable: no attacker can modify the negotiation communication without being detected by the parties in the communication.
3. The negotiation of a shared secret is secured: the negotiated secret is unavailable to anybody eavesdropping in the middle of the connection.

TLS Record Protocol provides connection security that has two basic properties:

1. **Privacy:** The confidentiality of the data is maintained through encryption. Symmetric cryptography is used for data encryption (e.g., AES, DES, RC4, etc.). Keys for symmetric encryption are generated uniquely for each connection. These encryption algorithms are negotiated by the TLS Handshake Protocol.
2. **Integrity:** The connection is reliable. Message transport includes a message integrity check using a keyed MAC. Secure hash functions (e.g., SHA, MD5, etc.) are used for MAC computations.

20.4.3 WTLS

The transport layer security protocol in the WAP architecture is called the Wireless Transport Layer Security or WTLS in short. WTLS provides functionality similar to TLS 1.0 and incorporates new features such as datagram support, optimized handshake and dynamic key refreshing. The WTLS layer operates above the transport protocol layer similar to TLS. WTLS provides the upper-level layer of WAP with a secure transport service interface that preserves the transport service interface below it. In addition, WTLS provides an interface for creating and terminating secure connections. The primary goal of the WTLS layer is to provide privacy, data integrity and authentication between two communicating applications. The WTLS protocol is optimized for low bandwidth bearer networks with relatively long latency.

20.4.4 Multifactor Security

In a security system larger key implies higher security. This is simply because, larger key means larger lock. However, it may not be always possible to keep on increasing the size of the key. Therefore, we keep on looking for alternate methods of increasing security. One such method is splitting the key and distributing it. For example, in a bank, a locker cannot be opened with one

key. It requires multiple keys. One key belongs to the customer; the other key is with the bank employee. Both the keys need to be used to open the locker. Take the example of ATM, where an ATM card and the PIN are required to withdraw cash. This technique is called multifactor security. These factors are generally a combination of "what you have", "what you know", and "what you are". Multifactor security can be a combination of any of the following factors.

What You Have

- Magnetic stripe card
- Private key protected by password
- Smart card
- Hardware token
- RF badge
- Physical key

What You Know

- Password
- Pass Phrase
- PIN (Personal Identification Number)
- Answer to some personal questions
- Sequence of numbers
- Predetermined events

Who You Are

- Fingerprint
- Voice Recognition
- Retinal Scan
- Hand Geometry
- Visual Recognition
- Face (picture in passport)
- Other biometric identities

Most of the multifactor security systems in use today are two-factor ones. However, for defense systems and high security establishments three-factor securities are used. In a two-factor security any two of the above factors are used. In a three-factor security, one each from the above factors are used.

20.4.5 Digital Watermark

Watermarks are being used for a long time as a security measure. If we take a 100-rupee currency note of Reserve Bank of India and hold it in front of a light source, we can see Gandhi's face on the white circle. This is called the watermark in the currency note. If we photocopy a currency note using a color photocopier, we will not be able to copy the watermark. The term "digital watermark" refers to a pattern of information inserted in a file. The file can be a digital audio file, digital video file, or a data file that identifies the file's copyright information (author, rights, etc.). The purpose of digital watermark is to provide copyright protection for intellectual property that is in digital format. Unlike printed watermarks which are intended to be visible, digital watermarks are designed to be

completely invisible. In the case of audio clips, the watermarks are inaudible. The information representing the watermark is scattered throughout the file in such a way that it cannot be identified, manipulated or reproduced.

20.4.6 Key Recovery

Encryption is an important tool for protecting the confidentiality of data. This data can be either data on transit over a network or a static data in a file. When suitably strong encryption algorithms are employed and implemented with appropriate assurance, encryption can prevent the disclosure of data to unauthorized parties. However, the unavailability, loss or corruption of the keys may prevent legitimate parties from accessing the data. For law enforcement agencies it may be sometimes necessary to decrypt encrypted data. To facilitate authorized access to encrypted data in the face of such situations, key recovery procedures and standards are needed. This type of system is called Key Recovery System (KRS). KRS will enable authorized persons to recover plaintext from encrypted data when the decryption key is not otherwise available. Key recovery is achieved through different key recovery techniques and key recovery information (KRI). Key recovery information refers to the aggregate of information needed by a key recovery technique to recover a target key. In third party systems like a Certification Authority (CA), the KRI is securely stored with the CA.

20.5 Public Key Infrastructure

Public Key Infrastructure or PKI in short consists of a mechanism to securely distribute public keys. PKI is an infrastructure consisting of certificates, a method of revoking certificates, and a method of evaluating a chain of certificates from a trusted root public key. The framework for PKI is defined in the ITU-T X.509 Recommendation. PKI is also defined through RFC3280. In RFC3280 the goal of PKI is defined as "to meet the needs of deterministic, automated identification, authentication, access control, and authorization functions. Support for these services determines the attributes contained in the certificate as well as the ancillary control information in the certificate such as policy data and certification path constraints."

PKIX is the Internet adaptation for PKI and X.509 recommendation suitable for deploying a certificate-based architecture on the Internet. PKIX also specifies which X.509 options should be supported. RFC2510, RFC2527 and RFC3280 define the PKIX specifications.

20.5.1 Public Key Cryptography Standards

Public-key Cryptography Standards or PKCS in short comprises standards proposed and maintained by RSA lab. These standards are accepted as de-facto standards for public key cryptography helping interoperability between applications using cryptography for security. Most of the crypto libraries available today support PKCS standards. PKCS standards consist of a number of components, which are defined through PKCS #1, #3, #5, #6, #7, #8, #9 #10, #11, #12, #13 and #15.

- *PKCS #1, RSA Encryption Standard:* PKCS #1 describes a method for encrypting data using the RSA public-key cryptosystem. Its intended use is in the construction of digital signatures and digital envelopes as described in PKCS #7. Digital enveloping is a process in which

someone "seals" a plaintext message in such a way that no one other than the intended recipient can open the sealed message. PKCS #1 also describes syntax for RSA public keys and private keys.

- *PKCS #2:* Incorporated as part of PKCS #1.
- *PKCS #3, Diffie-Hellman Key Agreement Standard:* PKCS #3 describes a method for implementing the Diffie-Hellman key agreement whereby two parties, without any prior arrangements, can agree upon a secret key that is known only to them.
- *PKCS #4:* Incorporated as part of PKCS #1.
- *PKCS #5, Password-Based Encryption Standard:* PKCS #5 describes a method for encrypting an octet string with a secret key derived from a password. PKCS #5 is generally used for encrypting private keys when transferring them from one computer system to another, as described in PKCS #8.
- *PKCS #6, Extended-Certificate Syntax Standard:* PKCS #6 describes syntax for extended certificates. An extended certificate consists of an X.509 public-key certificate and a set of attributes, collectively signed by the issuer of the X.509 public-key certificate.
- *PKCS #7, Cryptographic Message Syntax Standard:* PKCS #7 describes a general syntax for data that may have cryptography applied to it, such as digital signatures and digital envelopes.
- *PKCS #8, Private-Key Information Syntax Standard:* PKCS #8 describes a syntax for private-key information. PKCS #8 also describes syntax for encrypted private keys.
- *PKCS #9, Selected Attribute Types:* PKCS #9 defines selected attribute types for use in PKCS #6 extended certificates, PKCS #7 digitally signed messages and PKCS #8 private-key information.
- *PKCS #10, Certification Request Syntax Standard:* PKCS #10 describes a syntax for certification requests. A certification request consists of a distinguished name, a public key and optionally a set of attributes, collectively signed by the entity requesting certification. Certification authorities may also require non-electronic forms of request and may return non-electronic replies.
- *PKCS #11, Cryptographic Token Interface Standard:* This standard specifies an API, called Cryptoki to devices which hold cryptographic information and perform cryptographic functions.
- *PKCS #12, Personal Information Exchange Syntax Standard:* This standard specifies a portable format for storing or transporting a user's private keys, certificates, miscellaneous secrets, etc.
- *PKCS #13, Elliptic Curve Cryptography Standard:* It will address many aspects of elliptic curve cryptography, including parameter and key generation and validation, digital signatures, public-key encryption, and key agreement.
- *PKCS #15, Cryptographic Token Information Format Standard:* PKCS #15 is intended at establishing a standard which ensures that users, in fact, will be able to use cryptographic tokens to identify themselves to multiple, standards-aware applications, regardless of the application's cryptoki provider.

20.5.2 Storing Private Keys

For optimum security, some security information needs to be stored in tamper-resistant storage. This will help in protecting some of the sensitive data like private keys. Also, to provide application level security, part of the security functionality needs to be performed through this tamper-resistant device. The WAP Identity Module (WIM) is designed to address all these needs. WIM will be

used to perform WTLS and application level security functions. In a GSM, GPRS or 3G phones, it can be the SIM (Subscriber Identity Module) or USIM (Universal SIM) card containing additional functionality of the WIM or an external physically separate smart card. Use of generic cryptographic features with standard interfaces like PKCS#15 makes it possible to use the WIM for non-WAP applications like SSL, TLS, S/MIME, etc.

20.6 Trust

A portable computer never connected to a network, a standalone computer, never exposed to any unknown environment, can be assumed to be safe and secure. What happens to the security if we connect the same computer to a small private network? What happens if we connect the same computer to the Internet? What happens if we take this computer out in a football stadium and connect to the Internet over WiFi? The question is, can we trust these environments?

In early days, business was always face-to-face. In those days business used to be carried out between people who knew each other and in close physical proximity. In those days, one handshake literally closed the deal. The problem posed by mobile computing today is very much like that faced by business in the second half of the nineteenth century. During that time, the growth of transportation and communication networks in the form of railroads and telegraphs formed national markets and people were forced to do business with people whom they had never met. Let us take some examples. When a person searches the web for some authentic information on earthquakes, what are the options? The obvious answer is to use an Internet search engine, like Google. There are shops, forums, music groups with the name "earthquake". How do we know out of a few million hits, how much information is authentic? It may be relatively easy for a human being to determine whether or not to trust a particular web page. But is it that easy for software agents in our computers? Like in a database, can we form a SQL like query to extract an authentic technical research paper on earthquakes from the Internet? In another example, let us assume for the moment that you are 55 years old and having chest pain with sweating and vomiting; will you go to Google and give a keyword "chest pain doctor" to look for medical help? The question, therefore, is "Which information sources should my software agent believe?" This is equally important like the question "Which agent software should I believe and allow to access my information source?" If we look at these questions carefully we will find that the first question is about trust and the second question is about security. In mobile computing, we need to address both.

We said the question, "Which agent software should I believe and allow to access my information source?" relates to security. However, there is a catch. Suppose a person by name Anita tries to access my information source. My agent denies access to her. If she produces a certificate that she is a student in my mobile computing course, what action is expected from my agent? Of course, the agent should allow her to access my information source. This is an example of trust. The person who was not trustworthy becomes trustworthy when she produces a certificate. It is interesting to note that this certificate is not the conventional certificate as issued by a CA. Trust is explained in terms of a relationship between a trustor and a trustee. Trustor is a person who trusts a certain entity, whereas, trustee is the trusted entity. Based on the trust in the trustee, a trustor can decide whether the trustee should be allowed to access her resources and what rights should be granted. Therefore, trust plays an important role in deciding both the access rights as well as provenance of information. Trust management involves using the trust information, including recommendations

from other trustees. There are different models of trust. These are direct trust, hierarchical trust and web of trust.

Direct Trust: In a direct trust model, parties know each other. This is like early days where everyone personally knew others in the line of business. A user trusts that a key or certificate is valid because he or she knows where it came from. Every organization today uses this form of trust in some way. Many companies today do business through Internet. However, before they start doing business over Internet, a due diligence and audit is done. Following this they do business over the Internet with proper trust using trusted certificates and known key source.

Hierarchical Trust: In a hierarchical system, there are a number of "root" certificates from which trust extends. This is like the holding company establishing a trust and then member companies use this trust and key (certificate). These root certificates may certify certificates themselves or they may certify certificates that certify still other certificates down the chain. This model of trust is used by conventional CA.

Web of Trust: A web of trust encompasses both of the above models. A certificate might be trusted directly or trusted in some chain going back to a directly trusted root certificate or by some group of introducers. The web of trust uses digital signatures as its form of introduction. When any user signs another's key, he or she becomes an introducer of that key. As this process goes on, it establishes a web of trust. PGP (Pretty Good Privacy) uses this model of trust. PGP does not use the CA in its conventional sense. Any PGP user can validate another PGP user's public key certificate. However, such a certificate is only valid to another user if the relying party recognizes the validator as a trusted introducer.

20.6.1 Certificate

Digital certificate plays a significant role in establishing trust. Through a digital certificate, we can associate a name with a public key. Certificate is a signed instrument vouching that a particular name is associated with a particular public key. It is a mapping between a domain name (like mybank.co.in, for example) and a public key. The structure of certificates is hierarchical, originating from a trusted root certificate. For example, the root certification authority in India is called Controller of Certification Authority (CCA–http://cca.gov.in). CCA is responsible for generating the key pair using SHA-1 and 2048 bit RSA algorithm. CCA issues these certificates to users through different RAs (registration authority). An RA is an organization to which a CA delegates administrative functions of creation, distribution, and book-keeping of the public-private key pair.

Here are the data and signature sections of a certificate in human-readable format taken from an example Cited in the Netscape site:

```
Certificate:
   Data:
      Version: v3 (0x2)
      Serial Number: 3 (0x3)
      Signature Algorithm: PKCS #1 MD5 With RSA Encryption
      Issuer: OU=Ace Certificate Authority, O=Ace Industry, C=US
      Validity:
         Not Before: Fri Oct 17 18:36:25 1997
```

```
      Not After: Sun Oct 17 18:36:25 1999
   Subject: CN=Jane Doe, OU=Finance, O=Ace Industry, C=US
   Subject Public Key Info:
      Algorithm: PKCS #1 RSA Encryption
   Public Key:
      Modulus:
         00:ca:fa:79:98:8f:19:f8:d7:de:e4:49:80:48:e6:2a:2a:86:
            ed:27:40:4d:86:b3:05:c0:01:bb:50:15:c9:de:dc:85:19:22:
            43:7d:45:6d:71:4e:17:3d:f0:36:4b:5b:7f:a8:51:a3:a1:00:
            98:ce:7f:47:50:2c:93:36:7c:01:6e:cb:89:06:41:72:b5:e9:
            73:49:38:76:ef:b6:8f:ac:49:bb:63:0f:9b:ff:16:2a:e3:0e:
            9d:3b:af:ce:9a:3e:48:65:de:96:61:d5:0a:11:2a:a2:80:b0:
            7d:d8:99:cb:0c:99:34:c9:ab:25:06:a8:31:ad:8c:4b:aa:54:
            91:f4:15
         Public Exponent: 65537 (0x10001)
      Extensions:
         Identifier: Certificate Type
            Critical: no
            Certified Usage:
               SSL Client
   Identifier: Authority Key Identifier
      Critical: no
      Key Identifier:
         f2:f2:06:59:90:18:47:51:f5:89:33:5a:31:7a:e6:5c:fb:36:
         26:c9
   Signature:
   Algorithm: PKCS #1 MD5 With RSA Encryption
   Signature:
         6d:23:af:f3:d3:b6:7a:df:90:df:cd:7e:18:6c:01:69:8e:54:
         65:fc:06:
         30:43:34:d1:63:1f:06:7d:c3:40:a8:2a:82:c1:a4:83:2a:fb:
         2e:8f:fb:
         f0:6d:ff:75:a3:78:f7:52:47:46:62:97:1d:d9:c6:11:0a:02:
         a2:e0:cc:
         2a:75:6c:8b:b6:9b:87:00:7d:7c:84:76:79:ba:f8:b4:d2:62:
         58:c3:c5:
         b6:c1:43:ac:63:44:42:fd:af:c8:0f:2f:38:85:6d:d6:59:e8:
         41:42:a5:
         4a:e5:26:38:ff:32:78:a1:38:f1:ed:dc:0d:31:d1:b0:6d:67:
         e9:46:a8:
         dd:c4
```

Here is the same certificate displayed in the 64-byte-encoded form interpreted by software:

```
——BEGIN CERTIFICATE——-
MIICKzCCAZSgAwIBAgIBAzANBgkqhkiG9w0BAQQFADA3MQswCQYDVQQGEwJVUzERMA8G
A1UEChMITmV0c2NhcGUxFTATBgNVBAsTDFN1cHJpeWEncyBDQTAeFw05NzEwMTgwMTM2
MjVaFw05OTEwMTgwMTM2MjVaMEgxCzAJBgNVBAYTAlVTMREwDwYDVQQKEwhOZXRzY2Fw
ZTENMAsGA1UECxMEUHViczEXMBUGA1UEAxMOU3Vwcml5YSBTaGV0dHkwgZ8wDQYJKoZI
hvcNAQEFBQADgY0AMIGJAoGBAMr6eZiPGfjX3uRJgEjmKiqG7SdATYazBcABu1AVyd7c
hRkiQ31FbXFOGD3wNktbf6hRo6EAmM5/R1AskzZ8AW7LiQZBcrXpc0k4du+2Q6xJu2MP
m/8WKuMOnTuvzpo+SGXelmHVChEqooCwfdiZywyZNMmrJgaoMa2MS6pUkfQVAgMBAAGj
NjA0MBEGCWCGSAGG+EIBAQQEAwIAgDAfBgNVHSMEGDAWgBTy8gZZkBhHUfWJM1oxeuZc
+zYmyTANBgkqhkiG9w0BAQQFAAOBgQBtI6/z07Z635DfzX4XbAFpjlRl/AYwQzTSYx8G
fcNAqCqCwaSDKvsuj/vwbf91o3j3UkdGYpcd2cYRCgKi4MwqdWyLtpuHAH18hHZ5uvi0
0mJYw8W2wUOsY0RC/a/IDy84hW3WWehBUqVK5SY4/zJ4oTjx7dwNMdGwbWfpRqjd1A==
——END CERTIFICATE——-
```

20.6.2 Simple PKI

It was thought that digital certificate would address issues related to trust. However, finally, certificates emerged as an instrument for authentication. PKCS also made some attempt to address the need of trust through PKCS#6 and PKCS#9. IETF developed yet another standard called Simple PKI or SPKI (RFC 2692, RFC 2693) in short. SPKI defined a different form of digital certificates whose main purpose is authorization in addition to authentication. Purpose of SPKI is to define a certificate structure and operating procedure for trust management in the Internet.

20.7 Security Models

We have discussed different types of security algorithms. We have also discussed security protocols. These algorithms and protocols are used to protect our assets. These assets can be either static assets in the form of priceless data in a database, files, or documents within a computer or assets in transit. The security and trust model we choose should be able to secure our assets and protect our interests. To protect ourselves from different threats, we need to look at security and trust at system and application levels.

20.7.1 Infrastructure Level Security

Infrastructure level security offers security at the perimeter of the system. This will primarily include networks and the infrastructure. We can call this Network security as well. Infrastructure level security will include protecting the infrastructure or the network so that attacks from worms, viruses, and Trojan horses can be prevented. Prevention from other forms of attacks like intrusion etc., and different firewalls in the network are all part of infrastructure-level security. Virtual private network (VPN) is a part of infrastructure security as well.

Infrastructure level security for a mobile computing environment needs to handle some additional threats compared to a wired network. For mobile computing network, the last mile access network will be wireless in most of the situations. Therefore, at the access level, additional infrastructure security is necessary. An example is encryption in GSM using A5 algorithm. WiFi/wireless LAN

networks use WEP. Some vulnerabilities have been identified in infrastructure security for WiFi; therefore, new security protocols like 802.1x and 802.11i have been proposed to take care of the over-the-air interface in the access network.

20.7.2 System Level Security

In system level security we secure our systems to protect our assets. In the security framework provided by the operating system, shells will be part of system level security. In authentication challenges during Unix login, or login into a mainframe computer through a username, password is the system level security. Access Control list (ACL), File System Security, Memory Security, etc. will also be part of the system level security. It protects the system from worms, viruses and Trojan horses. Prevention from other forms of attacks like buffer overflow attacks, intrusions, etc., can also be part of system level security as also security protocols like SSL and TLS. There is a concept of a capability-based system, where security is policy-driven and managed through capability. Even if a virus enters into such a system, or an intrusion happens, it will not be able to damage any asset in the system. One such operating system is EROS.

Database security is part of system level security. In database security, data in the database is secured by the database software. This can be encrypting a column in a row or some special check based on ACL and capability. Most of the database software today offer security at this layer. This will be over and above the security offered by the operating system.

20.7.3 Policy Based Security

Security systems implemented for wired networks in any organization are primarily policy based. Effective security policies make frequent references to standards and guidelines that exist within an organization. Policy is a set of written down rules about what is allowed or what is not allowed in the network. Policies are usually area-specific, covering a single area. According to the RFC 2196, security policy is defined as "A security policy is a formal statement of the rules by which people who are given access to an organization's technology and information assets must abide." For example, there could be a rule in the organization that nobody will be allowed to have a global IP address. To stop spam and mail bound viruses, there may be another rule that prevents access to external email systems (like hotmail) from the corporate network. To stop the possibility of espionage, there could be a rule that FTP from the Internet is not allowed to any machine within the intranet. A standard is typically a collection of system-specific requirements that must be met by everyone. Standards are necessary when we need interoperability. A guide line is typically collections of system specific procedural specific "suggestions" for best practice. Guidelines are not requirements to be met, but are strongly recommended.

In a wired network where systems are stationary, and the network structure is static, it is possible to define security policies. In such networks, it is possible to enforce these policies or rules. However, things are different when we move to a mobile computing environment. In a mobile computing environment, the user will move from one device to another device, and from one network to another network. These devices or networks may be of similar type or different types. For example, a user moves from a WiFi network to a CDMA2000 network or from a PalmOS to a WindowsCE. In a network with static nodes, it is possible to define a security policy. It may be possible to

enforce such policies as well. However, in case of mobile computing where nodes are roaming from one network to another, it may not be practical to define a security policy and implement it. Therefore, over and above policy based security, for mobile nodes, we need object security. In object security, objects will carry their security signatures and capabilities. This is achieved through the concept of principal. Therefore, when a device moves from network to network, the device carries the security requirement and security signature with it. Principal-based security system is in the process of maturing. OMA DRM is an example of principal-based security.

20.7.4 Application Level Security

Infrastructure and system level security take care of security at the infrastructure and system level respectively. The parameters for these securities are not very flexible; most of the time vendors define them. In a mobile-computing environment we need security at the application level. Application security looks at the security at the content level. This can also be termed as Peer to Peer security. The application at the client device will talk to the application at the server and handle security requirements end-to-end as the content may demand.

In a mobile computing environment, we cannot make any assumption related to the client context or the network context. Therefore, the security needs to be addressed at the content level, using context awareness, J2ME, .NET, WIM or MExE (Mobile Execution Environment) environment. Using cryptographic libraries, we can build security at the application level. This security will be custom-built and can use standard algorithms or new algorithms as agreed by the peer nodes. Of course the system/infrastructure level security, if any, will be over and above the application level security.

20.7.5 Java Security

Security model provided by Java covers both system level and application level security. Java system level security is provided through the "sandbox" model. Sandbox provides a restricted environment for code execution through Java virtual machine. In the sandbox model, local code is trusted to have full access to system resources like file system, memory, etc. However, downloaded code from a remote site as an applet is not trusted. Therefore, applet can access only the limited resources provided inside the sandbox. Java supports digitally signed trusted applet. A digitally signed applet is treated like local code, with full access to resources. Digitally signed applets use public key infrastructure. Prior to transmission, the applet server signs an applet JAR file using its digital certificate. Upon receipt, the client side Java security manager verifies the signature and decides whether the origin and integrity of the application is trusted. Once the authentication is successful, the application code is delivered to the client for execution.

Java offers tools to facilitate various security-related operations. These are:

- *Keytool:* This is a command line tool. Keytool is used to manage keystore, which includes the following functions.
  - ❑ Create public/private key pairs.
  - ❑ Issue certificate requests (which will be sent to the appropriate Certification Authority).
  - ❑ Import certificate replies (obtained from the Certification Authority).
  - ❑ Designate public keys belonging to other parties as trusted keys and certificates are used

to digitally sign applications and applets. A keystore is a protected database that holds keys and certificates for an enterprise. Access to a keystore is guarded by a password. In addition, each private key in a keystore can be guarded by its own password.

- *Jar:* This is a command line tool to create JAR (Java Archive) files. The JAR file format enables users to bundle multiple files into a single archive file. Typically a JAR file will contain the class files and auxiliary resources associated with applets and applications. After importing appropriate keys into the keystore, the jarsigner tool is used to digitally sign the JAR file.
- *Jarsigner:* This is a command line tool to sign JAR files. This is also used to verify signatures on signed JAR files. The jarsigner tool accesses a keystore that is created and managed by key tool, when it needs to find the private key and its associated certificate chain. The jarsigner tool prompts for needed passwords.
- *Policy Tool:* Unlike the other tools, this tool has a graphical user interface. Policy tool is used to create and modify the external policy configuration files that define the installation's security policy.

As a part of application level security, Java framework supports cryptographic library through Java cryptography architecture (JCA). JCA refers to a framework for accessing and developing cryptographic functionality for the Java platform. These cryptographic services are:

- Symmetric key encryption algorithms.
- Public key encryption algorithms.
- Digital signature algorithms.
- Message digest algorithms.
- Message authentication code generation.
- Key generation algorithms.
- Key exchange algorithms.
- Keystore creation and management.
- Algorithm parameter management.
- Algorithm parameter generation.
- Key factory support to convert between different key representations.
- Certificate factory support to generate certificates and certificate revocation lists (CRLs) from their encodings.
- Random-number generation (RNG) algorithm.
- Support of SSL and TLS through http support.

Cryptographic library is available for the entire Java framework. This includes J2EE (Java 2 Enterprise Edition), J2SE (Java 2 Standard Edition), J2ME (Java 2 Micro Edition), and JC (Java Card). However, due to security reasons and resource constraints J2ME and JC functionalities are restrictive. Some of the APIs, which are available in J2EE and J2SE are not supported in Java card.

20.8 Security Frameworks for Mobile Environment

Mobile applications usually span over several networks. One of these networks will be a wireless radio network. Others will be wired networks. At the boundary of any of these networks, there is a need for protocol conversion gateways. These gateways run either at the transport layer or at the

application layers. Moreover, while the user is roaming in foreign networks, there will be multiple wired networks (PLMNs) managed and controlled by different network operators. Multiple gateways and multiple networks make security challenges in mobile environments complex.

In a security system, authentication, and non-repudiation are meaningful only when these are implemented end-to-end between parties that need to authenticate each other. Authorization is a direct function of authentication; therefore, it is also an end-to-end function. Authentication, authorization, and non-repudiation must therefore be implemented at the application layer. Confidentiality and integrity on the other hand can be implemented at any layer or through multiple layers. Confidentiality can be realized by encrypting isolated legs between gateways or even end-to-end. When confidentiality is realized in isolated legs, the gateways or nodes between the legs need to be secured and trusted.

Therefore, to offer secured environment in a mobile environment, security procedures will be a combination of many procedures and functions. Following sections cover some of the vulnerabilities and techniques to offer security in a mobile environment.

20.8.1 3GPP Security

In a mobile computing environment inside a campus the access network is likely to be WiFi. However, outside of the campus, access network will be one of the cellular wide area wireless networks like GPRS, CDMA, or GSM. It could also be WiMax. We have discussed WiFi security in Section 10.8. We have discussed GSM security in Section 5.9. We have discussed the GPRS security in Section 7.3.4. We have also discussed the security issues of CDMA networks in Section 9.3.5.

WiFi security is an extension of LAN security and primarily designed for data and applications. However, security procedures for wireless wide area networks GSM, GPRS, CDMA, are designed primarily keeping the operator in mind. All these security principles mainly try to protect an operator from fraud and network misuse. None of these procedures address the security concerns of user information or the application. Current security procedures in all these wide area wireless networks failed to provide a trusted environment where mobile users felt confident enough to place sensitive information over these networks.

In order to perform an attack in a wireless wide area network, the adversary has to possess one or more of the following capabilities:

- *Eavesdropping:* This is the capability through which the adversary eavesdrops signalling and data traffic associated with a user. Equipment required for such attack is a modified mobile station or a radio receiver.
- *Impersonation of a user:* This is the capability whereby the adversary sends signaling and user data to the network, in an attempt to make the network believe that they originate from a genuine (target of the impersonation) user. Equipment required for such attack is a modified mobile station or a radio transmitter/receiver.
- *Compromising Authentication Vectors in the Network:* The adversary possesses a compromised authentication vector, which may include challenge/response pairs, cipher keys and integrity keys. This data may have been obtained by compromising network nodes or by intercepting signalling messages on network links and then through brute force attack.

- *Impersonation of the Network:* This is the capability whereby the adversary sends signalling and user data to the target user, in an attempt to make the target user believe that the data originate from a genuine network. Equipment required for such attack is a modified base station or a radio transmitter/receiver.
- *Man-in-the-middle:* This is the capability whereby the adversary puts itself in between the target user and a genuine network and has the ability to eavesdrop, modify, delete, re-order, replay, and spoof signalling and user data messages exchanged between the sender and the receiver. The required equipments in such attack are modified base station in conjunction with a modified mobile station.

3GPP looked into these concerns and came up with changes in the security architecture of the current wireless wide area networks. 3GPP proposed a new architecture (Fig. 20.7) through following important changes:

- Changes were made to defeat the false base station attack. The extended security mechanism is now capable of identifying the network.
- Key lengths are increased to allow stronger algorithms for encryption and integrity.
- Mechanisms are included to support security within and between networks.
- Security is based within the switch rather than the base station to ensure that links are protected between the base station and switch.
- The authentication algorithm has not been defined, but guidance on choice will be given.
- Integrity mechanisms for the terminal identity (IMEI) have been included.

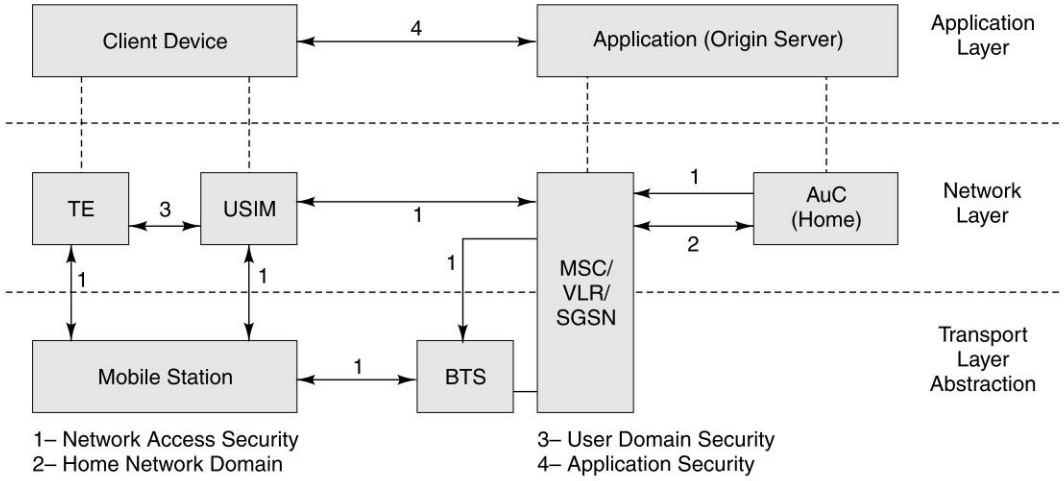


Figure 20.7 3GPP Security Architecture

20.8.2 Mobile Virtual Private Network

Virtual Private Network (VPN) provides an end-to-end security infrastructure. This generally deals with authentication, non-repudiation, integrity, and confidentiality at a layer between the transport

and the application layer. In Section 10.8.8 we discussed Wireless VPN with respect to WiFi. Like wireless VPN, mobile VPN is a private network over a public network (usually the Internet) to connect two endpoints. Instead of using a dedicated physical connection such as leased line, a VPN uses "virtual" connections routed through the Internet from the enterprise's private network to the remote mobile device. VPN implements this through an encrypted private connection between nodes. It generally uses IPsec and other PKI frameworks to offer confidentiality, authentication, non-repudiation (through digital signature), and integrity. With mobile VPN, mobile workers have the freedom to safely use wireless applications on their PDAs, smart phones and other handheld devices in the field as if they are in a private network.

20.8.3 Multifactor Security

In Section 20.4.4, we have discussed multifactor security where factors could be "what you have", "what you know", "what you are". In mobile network, multifactor security can be extended to multiple networks. For example, the security key (session key) in a GPRS network can be split into multiple parts and sent through Internet (TCP/IP) and AMS network. By now we know that SMS uses SS#7 network for its traffic. SS#7 network is closed and protected. SS#7 is physically more secured that the IP network.

20.8.4 Smartcard Security

Smart cards offer data encryption and the ability to store secret information for the purpose of authenticating the cardholder. There are various types of smartcards used in different application scenarios. One such example is the SIM card on a mobile phone. ETSI standard 03.48 specifies procedures for SIM card to be used as a security engine. SIM cards used in mobile phones are processor cards. Processor cards are smart cards with an inbuilt processor and memory. This local processor protects the content in the SIM and makes it tamper resistant. The 03.48 standard specifies the interoperability standards for cryptographic functions. Also, many of the SIM cards have RSA, DES, 3DES, AES algorithms implemented within the card. There are different file systems within a SIM card. These files have very stringent security controls. Files can be protected through passwords known to the user or operator. Private key and many other secret keys can be stored in these files. As these files are protected through password, even if the card is lost, the information is protected. To counter brute force attack, a smartcard processor does not allow more than 10 attempts to read a file data with wrong password. Therefore, a user can make use of the card as a security factory. Using this security factory is quite easy through Java card interfaces. In Java card technology, a Java interface is provided on the SIM card. Java cryptographic architecture (JCA) and Java programs running on the smartcard can make these things quite easy.

20.8.5 Mutual and Spatial Authentication

In Section 20.8.4 we have discussed how a SIM card can be used as a security factory. The SIM card can store secured information like private keys, wireless identity module, and many other private secured information. It also has various algorithms implemented. This can facilitate mutual authentication. In SSL or TLS over Internet client authentication is generally not done. However,

using a SIM card, we can carry out client authentication over wireless wide area networks. This is called mutual authentication; because, using GSM 03.08 procedures, a client can authenticate the server, also, the server can authenticate the client.

SIM cards in a GSM/GPRS network store location information. This includes country, network, and base station information. This information can be obtained and sent from the mobile phone using GSM 03.48 standards specification. Location information can then be used to implement spatial authentication. For example, if the user is in a neighborhood which is insecure, access to some critical applications can be prevented.

20.8.6 RFID Security

Application areas for RFID is increasing. However, it has certain vulnerabilities. For example, using the RF Dump tool (http://www.rfdump. org/), an adversary can detect an RFID-Tag and extract its meta information like Tag ID, Tag Type, manufacturer, etc. It can even be used to rewrite the data stored in some RFID tags using either a Hex or an ASCII editor. All these vulnerabilities pose a serious threat toward RFID based systems starting from merchandise in a store to the e-passports. The US government sometime ago decided to issue passports with RFIDs. It is nicknamed as "epassports". However, the concerns over RFID security delayed this plan. According to the specification of e-passport, there will be 64-bit RFID tags attached in the passport that will contain name, date of birth, place of birth, a digital photograph and a digital face recognition template of the passport holder. This RFID is supposed to work only in a very close proximity. A RFID reader placed beyond the distance of more than 10 centimeters should not be able to read the content of the e-passport. However, in reality it was found that the radio tags' readable distance is as large as 30 feet. This makes the security information in the e-passport available over radio for an adversary to access.

20.8.7 Mobile Agent Security

Mobile agents are processes that can autonomously migrate from one networked computer to another. Mobile agents can be useful for many applications, especially for those in Internet. For example, I give my weekly shopping list to my mobile agent, who visits the web sites of all the stores within a three kilometer radius and tells me which shop has which fish at what price, where tomato is the cheapest, where I can get my favorite pickle, etc. The mobile visits all these stores' web site and prepares a shopping plan for me.

Despite its many practical benefits, mobile agent technology results in significant security threats from both malicious agents and malicious hosts. For example, as the mobile agent traverses multiple hosts that are trusted to different degrees, its state may be changed in a way that can adversely impact the decision making process of the agent.

20.8.8 Mobile Virus

Viruses are common in the PC and desktop environment. However, they were not common in a mobile environment. However, things are changing; as the mobile device becomes more intelligent with more flexibility and higher capabilities, viruses are surfacing—some are already out in the open. In June 2004, as a proof concept a virus called Cabir was developed to exploit Bluetooth

vulnerability. In early November 2004 a mobile virus called "Skull.A" was reported for some models of Nokia phones. A new version named "Skull.B" emerged, which combines Skull.A and Cabir. One more virus identified as Commwarrior.A surfaced in March 2005. This virus uses a combination of Bluetooth and MMS (Multimedia Messaging Service) to propagate. The principles these viruses use are similar in concept as the desktop viruses do.

20.8.9 Mobile Worm

A worm needs to propagate, execute, and reproduce in an automated fashion. To reproduce and then propagate, the worm needs to execute a piece of code (designed by the worm writer) on the target system. Therefore, it is necessary to have an execution environment available to the worm code on the target mobile device. On a mobile equipment today we have various execution environments like:

1. WAP/WML Script (MExE Classmark I)
2. JavaPhone/Personal Java (MExE Classmark II)
3. J2ME (MExE Classmark III)
4. Symbian
5. Windows CE
6. Palm OS
7. Linux

These can access both the TCP/IP and SMS interfaces. Therefore, worms can replicate and propagate through both TCP/IP and SMS interfaces of JavaPhone, PersonalJava or J2ME framework.

Along with Java Phone, technology on the mobile equipment, Java Card facility is also available on the SIM cards. Using all these technologies, it will be possible to develop viruses, worms, and Trojan horses for mobile phones. These viruses and worms will be able to replicate, access the address book, use the network facility and propagate.

REFERENCES/FURTHER READING

1. Calhoun P., J. Loughney, E. Guttman, G. Zorn and J. Arkko (2003), "Diameter Base Protocol", RFC 3588, September.
2. Chen Zhiqun (2000), *Java Card Technology for Smart Cards*, Addison Wesley.
3. Controller for Certification Authority; Government of India: Digital Certificate, http://cca.gov.in.
4. Data Encryption Standard (DES); Federal Information Processing Standard Publication, 1999 October 25, U.S. Department Of Commerce/National Institute of Standards and Technology.
5. Diffie Whitfield and Martin Hellman (1976), *New Directions in Cryptography.*
6. Elliptic Curve Cryptosystem: http://www.certicom.com.
7. Finin Tim and Anupam Joshi. *Agents, Trust and Information Access on the Semantic Web.*
8. Fukuyama Francis. *The Virtual Handshake: E-Commerce and the Challenge of Trust*, http://www.ml.com/woml/forum/ecommerce1.htm.

9. Gindraux S. (2002), "From 2G to 3G: A Guide to Mobile Security", *Proceedings of Third International Conference on 3G Mobile Communications Technologies.*
10. http://www.sans.org/resources/policies/.
11. http://www.tbs-sct.gc.ca/pubs_pol/gospubs/TBM_12A/gsp-psg1_e.asp#poli.
12. Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile: RFC3280.
13. Java Security: http://java.sun.com/products/jdk/1.2/docs/guide/security/CryptoSpec.html.
14. Kaliski Burton S. Jr. *An Overview of the PKCS Standards*, published by RSA Lab http://www.rsasecurity.com/rsalabs/pkcs/.
15. Kaufman Charlie, Radia Perlman and Mike Speciner (2002), *Network Security Private Communication in a Public World*, 2nd Edition; Prentice-Hall of India.
16. Kiran Shashi, Patricia Lareau and Steve Lloyd: PKI Basics-A Technical Perspective, November 2002, http://www.pkiforum.org.
17. Krawczyk H., M. Bellare and R. Canetti (1997), "HMAC: Keyed-Hashing for Message Authentication", February, RFC 2104.
18. Kun Yang, Guo Xin and Liu Dayou. (2000), "Security in Mobile Agent System: Problems and Approaches*", ACM SIGOPS Operating Systems Review.*
19. Levy Henry M. (1984), *Capability-Based Computer Systems*, Digital Press.
20. Waldman Marc and David Mazieres Tangler (2001). *A Censorship-Resistant Publishing System Based On Document Entanglements*, December 8.
21. Messerges, Thomas S. and Ezzat A. Dabbish, (2003), "Digital Rights Management in a 3G Mobile Phone and Beyond*", DRM'03*, October 27, 2003.
22. Report of the Technical Advisory Committee to Develop a Federal Information Processing Standard for the Federal Key Management Infrastructure, http://csrc.nist.gov/keyrecovery.
23. Rigney C., S. Willens, A. Rubens and W. Simpson (2000), "RADIUS" (Remote Authentication Dial In User Service), RFC 2865, June.
24. Rivest, R. (1992), "The MD5 Message-Digest Algorithm", *RFC 1321*, April.
25. Secured Socket Layer (SSL): http://developer.netscape.com/docs/manuals/security/pkin/contents.html.
26. Shapiro Jonathan S. and Norm Hardy (2002), "EROS: A Principle-Driven Operating System from the Ground Up", *IEEE Software Magazine*, January.
27. Shelfer Katherine M. and J. Drew Procaccino, (2002), "Smart Card Evolution", *Communications of the ACM,* July 2002, Vol. 45, No. 7, p83.
28. Stallings William (2000), *Network Security Essentials: Applications and Standards*; Pearson Education.
29. Stallings William (2003), *Cryptography and Network Security Principles and Practices*; Pearson Education, 3rd edition.
30. Talukder Asoke K. and Debabrata Das (2004), "Artificial Hygiene: Non-Proliferation of Virus in Cellular Network*", Journal of Systems and Information Technology*, Volume 8, December 1, pp 10–22.

31. Transport Layer Security Protocol: RFC 2246.
32. Wireless Application Protocol Identity Module Specification, Version 05-Nov-1999, WAP Forum.
33. WPKI, WAP-217-WPKI, Wireless Application Protocol Public Key Infrastructure Definition, Apr-2001.
34. Wireless Transport Layer Security, Version 06-Apr-2001: WAP-261-WTLS-20010406-a; http://www.wapforum.com.
35. 3GPP TS 03.48, Digital cellular telecommunications system (Phase 2+); Security mechanisms for SIM application toolkit; Stage 2, 1999.
36. 3G TR 33.900 V1.2.0 (2000-01) 3rd Generation Partnership Project; Technical Specification Group SA WG3; A Guide to 3rd Generation Security (3G TR 33.900 version 1.2.0), January 2000.

REVIEW QUESTIONS

Q1: What are the different components of information security?
Q2: What is called an attack in terms of network security? Discuss its various types.
Q3: What do you understand by security algorithms and security protocols? What are the differences between them? How are they related?
Q4: Describe symmetric key and public key encryption. If you are required to design a security system, when will you use which one?
Q5: Describe each of the following in brief:
- (a) DES
- (b) AES
- (c) RSA algorithm and its applications
- (d) Elliptic Curve Cryptography

Q6: Explain Diffie – Hellman key exchange algorithm.
Q7: What is a hashing algorithm? Discuss its types.
Q8: Write short notes on:
- (a) SSL and its applications
- (b) TLS
- (c) WTLS
- (d) Multifactor Security and its applications

Q9: What is a digital watermark? Explain its applications.
Q10: Discuss Public Key Cryptography Standards (PKCS).
Q11: Explain each of the following:
- (a) Infrastructure level security

(b) Policy based security
(c) Application level security
(d) Java security
(e) 3GPP security
(f) Smartcard security

Q12: Explain the security framework for mobile computing. How do we ensure security in a mobile environment through Mobile VPN?

Q13: Give examples of RFID security vulnerability.

Q14: Explain the difference between Mobile Virus and Mobile Worm.

The McGraw·Hill Companies

CHAPTER 21

Next Generation Networks

Popularity of Internet helped growth of new research and new business opportunities starting from search engines, wikis, email, social networks, publishing and e-commerce services. Through Internet, enterprises have been able to make their presence global and keep their stores open 24 hours a day and 365 days a year. The demand in Internet forced POTS (Plain Old Telephone Service) to enhance its services into broadband. Various DSL (Digital Subscriber Line) technologies came into existence. The other revolutionary technology of recent times is cellular telephony. Nonetheless, it has changed the landscape of public telephony. The services provided by the cellular operators are no longer limited to speech telephony, they include SMS (Short Message Service) and now MMS (Multimedia Message Service). Nowadays, cellular users are also able to surf the Internet, read electronic mails and avail location-based services on their cellular handsets. While telecommunication networks offered data services, the cable TV industry also realized the need of similar services and offer data services over cable.

All this calls for a judicious blend of Internet, broadband, cellular telephony and TV into a technological mosaic that best combines the advantages of communications, covering their limitations. Next Generation Network or NGN is the converged solution for all these communications services. In next generation networks, all traffic will be packet based and that too, through IP (Internet Protocol) and subsequently, with next generation IP—IPv6. It will offer high bandwidth, be it over the wire or wireless. Also, the NGN will offer seamless roaming not only from one geography to another, but also from one type of network to another type of network within a geography or across geographies. According to ITU (International Telecommunication Union), NGN is defined as "A packet-based network able to provide services including Telecommunication Services and able to make use of multiple broadband, QoS-enabled transport technologies and in which service-related functions are independent from underlying transport related technologies".

NGN is about convergence—all technologies of the past and the future will converge in the even playfield of NGN. It will seamlessly blend the public switching telephone network (PSTN) and the public switched data network (PSDN). Rather than having a large centralized switch,

NGN pushes the central office (CO) functions to the edge of the network with distributed network infrastructure. It can be thought of as object oriented networks where objects are various services. These services are primarily running in various servers and developed using IT (Information Technology). The advantage of IT is that services can be created quickly and commissioned at a low cost. The access methodology will use all kinds of communications and networking technologies.

21.1 All in One—the Converged Scenario

NGN, in essence, is the convergence of IT and CT (Communications Technology). IT and CT are derived from different roots with diverse business objectives; this makes NGN very complex. As NGN will be over IP only, it is also called "All IP Networks" or AIPN. NGN is also referred to in some literature as B3G (Beyond 3G).

21.1.1 Convergence of Voice and Data

Voice is easily comprehensible—it is the sound that we create in the mouth by vibrating our vocal chord or it is the sound produced while human beings speak or shout. Voice is generally perceived through the ear. However, the definition of data is, sometimes, a misnomer. In the context of mathematics, it is used as a basis for reasoning, discussion or calculation. However, in the context of computers or IT, it can be defined as unstructured stream of bytes. When we add a context or put a structure on data, it is converted into information. Majority of information is generally perceived through eyes and visual means. In NGN, voice and data will converge with voice using the VoIP technology and transmitted as data over the IP network.

21.1.2 Convergence of Wireline and Wireless

Wireline technology started with telegraph in 1832; and then, it was telephone in 1871. All these were analog till electronic switches were invented. Also, the analog voice became digital with the invention of PCM. However, all these technologies worked over physical electronic wire, which mainly was copper. Then came fiber optic where digital data is converted into light and transmitted. Fiber allows a higher level of concentration and multiplexing. However, all these—be it analog voice or digitized data, over copper or fiber are transmitted over a physical media—which are termed as fixedline or wireline.

With the discovery of radio in 1901, voice went wireless. Unlike wireline, wireless is bandlimited—there is a fixed band in the electro-magnetic spectrum that is used for radio transmission. In wireline, a transmission can be pointed to a particular destination and can be changed by moving the wire; however, in wireless it functions in broadcast mode transmitting a signal over radio in all directions. Radio started with simplex mode of transmission where there was a radio station with multiple receivers; then it graduated to duplex mode of transmission with walky-talkies. The frequency reuse concept led to the development of cellular telephony. The cellular telephony matured from 1st generation analogue AMPS to 2nd generation GSM. Then it moved to 2.5 generation GPRS to 3rd Generation technologies like UMTS and IMT-2000.

Wireless was also being tried in a variety of applications and services. These were using the free band that is generally known as ISM band. ISM band is unlicensed and free. This motivated

researchers and enterprises to come up with wireless LAN (Local Area Network), that is commonly known as WiFi. The pressure on bandwidth kept on increasing leading to the introduction of broadband wireless commonly known as Worldwide interoperability for Microwave Access (WiMAX).

In NGN wireline and wireless will converge. Both data and voice will be carried over wireline and wireless. However, the backbone transmission will continue to remain wireline. There will not be any functional differentiation between an Internet service provider, mobile service provider, or long-distance carriers.

21.1.3 Convergence of Circuit Switching and Packet Switching

In circuit, we establish an end-to-end channel for communication. This channel is reserved for a period of time irrespective of whether the channel is carrying any traffic. Also, in a circuit, the user has to pay for the period the channel is reserved. Circuits ensure a predictable quality of service (QoS). However, in case of packets a communication channel is shared by many packets that may have different source and destination. Packet switching technology is subject to delay, latency, jitter and loss. Any service that demands for QoS preferably should to be over circuit; on contrast any service that can withstand delay can be over packets. Data transmission is well suited for packet switching. In NGN, circuit switching and packet switching will converge—circuit switched data and packet switched data will all be carried over packet switched networks.

21.1.4 Convergence of IT and CT

The first step towards the convergence between telecommunication and IT happened in 1965 when AT&T used computers to do the switching in Electronic Switching System (ESS). On the other hand, the packet switch network was bringing communication closer to computers. The World Wide Web (WWW), which was started by Tim Berners Lee in 1989 as a text processing software, brought these two faculties of technology together and established Internet as a powerful media. The Internet meets four primary needs of the society: communication, knowledge sharing, commerce and entertainment. This convergence is called Information and Communication Technologies (ICT). Through ICT, we are now moving towards an information based society. ICT will address the need to access data, information and knowledge from anywhere at any time.

The convergence of IT and CT has changed the end user devices as well. Sometime ago, both telephone and computer devices were without any intelligence. These devices were connected to the powerful central switch and central mainframe computer, respectively. Convergence of IT and CT is leading the way to multi-access, multi-use and multi-network devices. We mentioned that NGN can be thought of as an object-oriented network where objects are created using IT. The access methodology by these objects will use all kinds of communications networks. In NGN devices will be a fusion of technologies that will adopt the best features and functions from IT and CT platforms. These devices will not discriminate between different networks, but rather, allow users to move seamlessly between cellular and digital Wi-Fi infrastructures.

The role of device diversity on the NGN will not be limited to handling calls from cell towers to WiMAX networks and back again. A user of the next generation network will expect an ability to connect wherever and however, it is most convenient and (probably) cheapest without being

concerned about which network is being connected to. Today's phone subscribers may be more interested in voice than data services; however, tomorrow priorities may change, with functions like groupware, collaboration, and videoconferencing making an important difference, particularly to road warriors and telecommuters. The incorporation of RFID technology into phones are helping a phone to transform into an electronic wallet that will help payments at fast food drive-ins, retail stores and other venues.

21.1.5 Convergence of OSS and BSS

Traditionally networks are managed through some NMS (Network Management Systems). When these networks are too large like telecommunications networks, where network elements are in hundreds of thousands, it is not sufficient to have simple NMSs. These complex networks need manager of managers that are called OSS (Operations Support Systems) that manage networks through NOC (Networks Operations Centers). OSS systems do the management of the services that a network offers. OSS systems manages operations functions like FCAPS (Fault, Configuration, Accounting, Performance, Security) combined with inventory, and Trouble Ticketing.

While OSS manages the operation, BSS (Business Support System) manages the financial side of the business i.e., the billing, accounting, and revenue of the network operator. Fraud management and churn management also fall within the domain of BSS. BSS also offers the Customer Care interface with products, contracts, and sales.

Traditionally these two systems were asynchronous and used to be run as two independent systems in a network. However, the need of NGN will change all these; OSS and BSS systems will converge where they not only interact real-time, they need to make instant decisions on services related to content.

21.2 Narrowband to Broadband

Multimedia is changing the communications scenario. Users want various contents over unified bearers and devices. This demands higher bandwidth compared to simple data. This is now possible through cable modem, DSL, fixed wireless, satellite or other means. While DSL and Cable Modem do not offer mobility, other wireless technologies offer mobility at slow speed to vehicular state. A single broadband connection can accommodate multiple users simultaneously. More often, it can be left in a stand-by mode as a "Always on" network. Unlike traditional wireline telephony, broadband communications do not require distinct call lengths that can be metered for billing and signaling purposes.

21.2.1 DSL (Digital Subscriber Line) Broadband Networks

DSL is a family of technologies that provide data services over the wires of a local telephone network. DSL can be asynchronous and synchronous. In asynchronous DSL (ADSL) the bandwidth for download is higher that upload. These are quite suitable for Internet access or TV over IP network where the traffic pattern is asymmetric with small number of bytes for upload and a large number of bytes for download. However, for interractive services or VoIP, the traffic load in both upload and download are of similar order. For these types of services synchronous DSL (SDSL)

are more suitable. In a fixedline telephone voice requires 3.4 kHz bandwidth. However, the copper wire that is used for local loop is capable of carrying frequencies well beyond the 3.4 kHz upper limit of POTS. Depending on the length and quality of the loop, the upper limit can be tens of megahertzs. DSL takes advantage of this unused bandwidth of the local loop. As DSL uses frequencies beyond our audible limits, voice and data connection can operate simultaneously without interference. This means that while data exchange is in progress, one can use the telephone to converse with others without interference. Typically, the download speed of consumer DSL services ranges from 256 kilobits per second (kbit/s) to 24,000 kbit/s, depending on DSL technology, line conditions and service level implemented. Other DSL technologies like HDSL (High bit-rate Digital Subscriber line) offer much higher bandwidth.

21.2.2 WiMAX Broadband Wireless Networks

We introduced 802.16 and WiMAX (Worldwide Interoperability for Microwave Access) in Chapter 4 where we categorized it as emerging technology. WiMAX is showing lot of promises and being rolled out by many network providers as wireless broadband. WiMAX can be either fixed or mobile. Fixed WiMAX deployments do not cater for handoff between Base Stations, therefore the service provider cannot offer mobility. Mobile WiMAX on contrast offers a handoff that can be used to deliver both fixed and mobile services. With a line-of-sight environment with a portable Mobile WiMAX device, bandwidth can be expected around 10 Mbit/s over a 10 km radius. However, in urban environments they may not have line-of-sight; therefore, users of WiMAX may only receive 10 Mbit/s over 2 km.

21.2.3 High Speed Broadband Cellular Networks

High Speed Packet Access (HSPA) is a suite of mobile telephony protocols that aim to extend and improve the performance of existing UMTS family of protocols. Better performance is achieved using improved modulation schemes and by fine-tuning the protocols by which cellular handsets and base stations communicate. This helps in efficient utilization of the available bandwidth. Under HSPA, High Speed Uplink Packet Access (HSUPA) and High Speed Downlink Uplink Packet Access (HSDPA) are already existing protocols.

- **High Speed Uplink Packet Access (HSUPA):** Defined by UMTS release 6, the 3GPP definition of HSUPA is "technical purpose of the Enhanced Uplink feature is to improve the performance of uplink dedicated transport channels, i.e., to increase capacity and throughput and reduce delay". HSUPA provides uplink speeds up to 5.76 Mbit/s. HSUPA uses enhanced dedicated channels on which it employs link adaptation methods which help in faster link adaptation and effective retransmissions. There are seven HSUPA categories provisioning speeds from 0.73 Mbits/s to 5.76 Mbits/s.
- **High Speed Downlink Uplink Packet Access (HSDPA):** Defined by UMTS release 5, HSDPA provisions downlink speeds of 1.8, 3.6, 7.2 and 14.4 Mbit/s. Normally, deployments provide up to 7.2 Mbit/s in downlink while supporting uplink speeds to a maximum of 384 kbit/s. HSUPA improves downlink performance using faster scheduling of packets and quick retransmissions (at the base stations) and Adaptive Modulation and Coding (AMC). HSDPA has more than 10 different categories supporting various data rates.

21.2.4 WiBro

WiBro or Wireless Broadband is the South Korean name for IEEE 802.16e (mobile WiMAX) standard. The base stations in WiBro would offer an aggregate data throughput from 30 to 50 Mbit/s while covering a radius of 1–5 km for the use of Internet. It can, actually, provide mobility for moving devices up to 120 km/h (74.5 miles/h) compared to Wireless LAN (WLAN) which provides mobility up to walking speed. WiBro does so by adapting Time Division Duplexing (TDD) and OFDMA while using 8.75 MHz as channel bandwidth.

21.3 All IP and B3G Network

We have mentioned that according to ITU NGN is defined as a packet-based network able to provide various telecommunications and data services. NGN is also called Beyond 3G (B3G) network. The B3G network architectures will evolve to include a much wider range of users, applications and economic deployment. 4G (again, also known as Beyond 3G) is an acronym for Fourth Generation Communications system to identify the next step in wireless communications. There is no industry consensus on what a "Beyond 3G network", will look like; but, as far as the next generation networks are concerned, concepts and ideas include the following:

- Transition towards an "All-IP based network infrastructure".
- Support of heterogeneous technologies (i.e., PSTN, UTRAN, Ad-hoc, WLANs, WiMAX, UMTS, WiFi, etc.).
- Seamless handover across both homogeneous and heterogeneous wireless access technologies.
- Multilayer Mobility Management suitable to support fast mobile users that may access a wide range number of services with diverse characteristics.
- QoS support on the IP layer.
- Use of policy-based mechanisms in order to determine QoS, accounting, and billing mechanisms for multimedia services.
- Network access control of mobile users (i.e., deployment of AAA protocols that allow inter-domain network access control) regardless of heterogeneous wireless access network used.
- Distributed AAA architecture for the dynamic establishment of trust relations in hybrid IPv4/IPv6 networks.
- Secure access to multimedia services across different networking environments.
- Access to multimedia services in hybrid IPv4/IPv6 based networks.
- It will offer RASP that includes reliability, availability, security, and performance.
- More users per cellular cell.
- A highly efficient spectral system.
- Backward compatibility with existing wireless standards.

21.4 OFDM (Orthogonal Frequency Division Multiplexing)

As we move towards convergence and AIPN, the bandwidth requirements for IP increase. To allow higher spectrum utilization OFDM (Orthogonal Frequency Division Multiplexing) is

increasingly being used. Some services that use OFDM today are: DSL, WiMAX, DAB (Digital Audio Broadcast), DVB (Digital Audio Broadcast), 3GPP LTE (3GPP Long Term Evolution), etc.

OFDM transmission scheme is an optimal version of the multi-carrier data transmission scheme. OFDM is based on the principle of spreading the data to be transmitted over a large number of carriers when each of them is being modulated at a low rate. The carriers are made orthogonal to each other by appropriately choosing the frequency spacing between them. Orthogonality helps in elimination of cross-talk between the sub-channels and the use of inter-carrier guard bands. Although the sidebands from each carrier channel overlaps, they can still be received without interference as they are orthogonal in position among each other. This is achieved by putting the carrier spacing equal to the reciprocal of the symbol period. Figure 21.1 depicts an OFDM spectrum.

The data to be transmitted on an OFDM signal is spread across the carriers of the signal wherein each carrier takes part of the payload. This immensely reduces the data rate taken by each carrier. This is effected by adding a guard band time or guard interval into the system. This makes sure that data is sampled only when the signal is stable and no new delayed signals arrive that would alter the timing and phase of the signal.

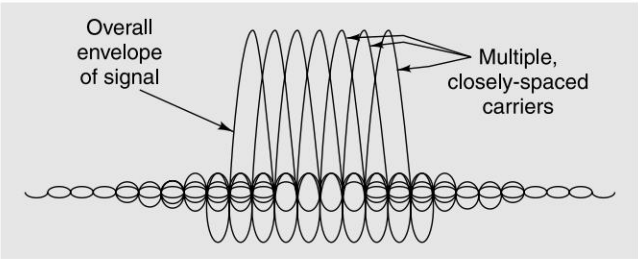


Figure 21.1 OFDM Spectrum

The data distribution across a large number of carriers in the OFDM signal has some significant advantages. Nulls caused by multi-path effects or interference on a given frequency only affect a small number of the carriers, the remaining ones being received correctly. Further, error coding techniques enable many or all of the corrupted data to be reconstructed within the receiver. The receiver at OFDM acts as a bank of demodulators converting each carrier to direct current. The resulting signal is integrated over the symbol period to regenerate the data from that carrier. The same demodulator also demodulates other carriers. As the carrier spacing equal to the reciprocal of the symbol period means that they will have a whole number of cycles in the symbol period and their contribution will sum to zero resulting in no interference. One requirement of the OFDM transceiver systems is that they must be linear as any non-linearity would cause interference between the carriers as a result of inter-modulation distortion. This results in impairing the orthogonality of the transmission.

Although OFDM requires costly circuitry and is sensitive to frequency synchronization issues, it has the following major advantages:

- High spectral efficiency.
- Facilitation of transmitter macro-diversity.
- Robustness against Inter Symbol Interference (ISI) and severe channel circumstances.

21.5 FAMA/DAMA

In order to resolve medium access control issues, the Fixed Assignment Multiple Access (FAMA) protocol was proposed. The applications of FAMA protocol are characterized by the assignment of capacity in a fixed manner amongst multiple stations. Irrespective of the fluctuating demands (of stations), the stations are assigned a fixed channel capacity. However, this results in significant underuse of the overall available capacity.

FAMA protocols can be in of different flavors such as Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA), Code Division Multiple Access (CDMA) or Space Division Multiple Access (SDMA). However, all such FAMA flavored protocols assign a static portion which can be in terms of time, frequency, code or space, of the overall link capacity to different stations. That is the assignment of resources are fixed and do not change according to station traffic patterns. The major benefit of FAMA based protocols is they can provide bounds for delay performance which become of paramount importance in real-time applications. However, following are some potential drawbacks:

- It is difficult to configure FAMA protocols when a new station comes in or moves out of a network system.
- The schemes of tuning FAMA flavored protocols (like TDMA, FDMA and Multi-Frequency TDMA) are labor intensive and unscalable.
- It is difficult to implement FAMA based protocols in a distributed mode.

Improvement over FAMA protocols, Demand Assignment Multiple Access (DAMA) protocols have capacity assignment in a manner that optimally respond to demands from/amongst multiple stations. DAMA protocols assign channels to stations based on the traffic information in the network. The assignment of channels is achieved through reservation or polling techniques so that each station can express their interest in using the channel for transmission based on its own traffic information. Such a process differs for different flavors of DAMA protocols. The reservation process can take place either in the primary communication channel or in a separate signaling channel depending upon the actual protocol. DAMA protocols can be collision free, in which each station will be assigned a fixed reservation slot, or it can be collision based, in which each station needs to compete with other stations when transmitting requests. Polling techniques can be used to decide which station should get the right of transmission over a certain channel during a certain time period as polling is naturally suited to networks with a centralized base station.

21.6 Multi Protocol Label Switching (MPLS)

As a packet in a connectionless network layer protocol like IP travels from one router to the next, each router makes an independent forwarding decision for that packet. A smarter routing is essential for traffic engineering and efficient routing of packets in a converged network where there are payloads from both circuit switched networks and packet switched networks. Multiprotocol Label Switching (MPLS) as described in RFC 3031 (Multiprotocol Label Switching Architecture), does exactly this. MPLS operates at an OSI Model layer that is generally considered to lie between traditional definitions of Layer 2 (data link layer) and Layer 3 (network layer) to perform smarter routing, and thus is often referred to as a "Layer 2.5" protocol.

MPLS is a packet forwarding technology which makes the use of labels for data forwarding decisions. It provides a unified data carriage service for different traffic categories like IP packets, frames of ATM and SONET, etc. Such a model facilitates MPLS to be a generic model for both, circuit switched and packet switched data. The MPLS working group aims to standardize a base technology which can combine the use of label swapping in the forwarding component with network layer routing in the control component and thus, provisioning:

- MPLS to run over any link layer technology while supporting both unicast and multicast traffic flows.
- MPLS to be scalable enough to support Internet growth while being compatible with the IETF Integrated Services Model and its related protocols.
- MPLS to support current IP network operations.

Coming from the family of IETF, MPLS initiates assignment and distribution of label bindings for the establishment of Label Switched Paths (LSPs). LSPs can be created by concatenating one or more label switched hops which, in turn, provision a packet to be forwarded from one Label Switching Router (LSR) to another LSR across the whole MPLS domain. MPLS defines standard based IP signaling and label distribution protocols along with extensions to existing protocols. This helps multi-vendor interoperability. With MPLS, the network layer header analysis is done when the packet enters its domain and this label inspection drives subsequent packet forwarding across the whole of domain. Figure 21.2 depicts an overview of MPLS functioning.

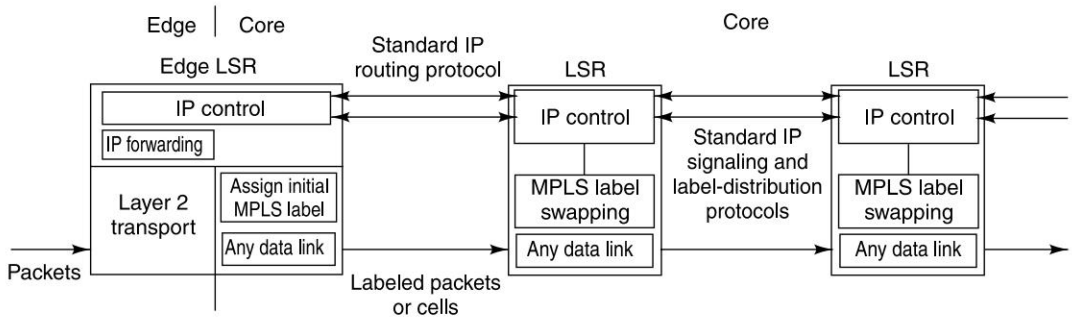


Figure 21.2 Functioning of MPLS

Currently, MPLS has the following major applications:

- *Traffic Engineering*: MPLS helps Traffic Engineering in view of the unprecedented growth in demand for network resources and real-time nature of IP applications. MPLS facilitates Traffic Engineering to allow ISPs to move traffic flows away from the shortest path on to potentially less congested physical paths across the network.
- *Class of Service (CoS)*: MPLS offers great flexibility to the ISPs in terms of different types of services that they can provide to their customers. The precedence bits are used only to classify packets into one of various classes of service. Then onwards, ISPs can determine the specific type of service that is supported by each service classification bits.
- *Virtual Private Networks (VPNs)*: MPLS lets ISPs offer VPN services by providing a flexible and powerful tunneling mechanism.

MPLS offers enhanced routing mechanisms by supporting more than just destination-based forwarding which makes it permit ISPs to deliver new services that can not be easily provided by conventional IP routing techniques.

21.7 WIRELESS ASYNCHRONOUS TRANSFER MODE

Wireless Asynchronous Transfer Mode (WATM) adds up the mobility advantages to the ones provided by ATM networks. The mobility aspect forces a decoupling of the normal mapping of node and switch port. In Wireless ATM, a wireless access point connects the set of wireless nodes while servicing on a single port of ATM switch. As it is known, ATM technology offers speed, scalability, multimedia integration and uniform API features with good cost performance. To develop WATM, the following are needed:

- Protocol extensions to support mobility within an ATM infrastructure.
- Radio access technology necessary for seamless delivery of ATM services to mobile terminals.

According to the WATM working group charter WATM "specification will include both mobile ATM extensions for mobility support within an ATM network as well as radio access layer for ATM-based wireless access. The WATM specifications are intended for use in networks involving terminal mobility and/or radio access, and will be designed for compatibility with ATM equipment adhering to the (then) current ATM Forum specification".

Wireless/Mobility protocols are incorporated into standard ATM stack keeping the following as mandatory:

- ATM cell as basic unit in both wireless and backbone.
- Standard ATM services at transport interface.
- Custom medium access and data link control for wireless segment.
- Mobility extensions to ATM signaling: handoff and location management.

Figure 21.3 depicts the WATM system. It brings out the individual contribution of ATM switches, mobility protocols and WATM radio access. Normally, Wireless ATM systems can be constructed via hardware/software plug-ins to the existing ATM switches.

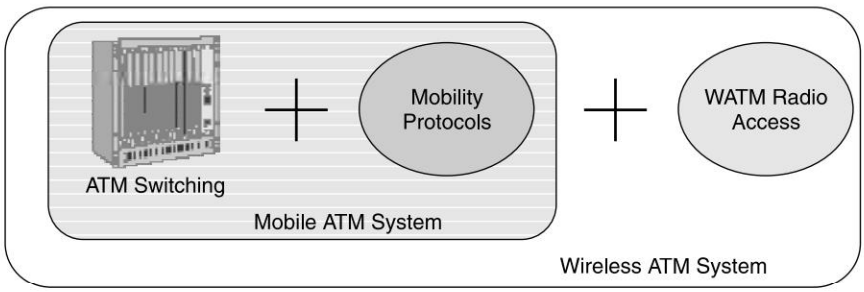


Figure 21.3 Overview of WATM System

The following three are the basic service scenarios for WATM:

1. Providing mobility support in IP
   : ATM connection to appear as a tunnel to IP
   : Mobility of tunnel endpoint oblivious to IP
2. Mobile Telephony
   : ATM base station terminates GSM radio protocol
   : Translates GSM call setup to UNI signaling
   : Bridging GSM air interface and ATM data path
3. Wireless ATM systems
   : End-to-End ATM connection to mobile endpoints
   : Broadband wireless access to the tune of 25 Mbps

As of now, Wireless ATM has firmly established itself as a promising area of research, product development and standardization. The momentum to make progress in each of these three fields is already there due to the efforts of business and regulators to find suitable RF spectrum for local area networks based on the concept of Wireless ATM.

21.8 Multimedia Broadcast Services

Traditionally radio and television have been using analog signals. These have been using AM (Amplitude Modulation) and FM (Frequency Modulation) technique. However, these are changing, radio and TV industry have already embraced digital technology; and now they are moving towards IP too.

21.8.1 Digital Audio Broadcast (DAB)

DAB is a digital technology for radio broadcast. DAB was developed as a research project for the European Union (Eureka project number EU147), which was first tested in Germany in 1988. The MPEG-1 layer-2 audio coding (MP2) technique was created as part of the EU147 project. The protocol was adopted by ETSI in 1997. DAB uses OFDM modulation technique, which is the popular transmission scheme for modern wideband digital communication systems. DAB offers substantially higher spectral efficiency, measured in programmes per MHz and per transmitter site, than analog communication. DAB being digital, can also carry data with the radio signal. DAB therefore carries program guides called *Dynamic Label Segment* from the station giving real-time information such as song titles, music type and news or traffic updates.

Eureka 147 DAB uses a wide-bandwidth broadcast technology and typically spectra has been allocated for it in 174–240 MHz and 1452–1492 MHz, although the technology allows for operation in bands above 30 MKz. The earlier version of DAB that is being used in the UK, Denmark, Norway and Switzerland, uses the MP2. The new DAB+ standard has adopted the HE-AAC (High Efficiency Advance Audio Coding) version 2 audio codec, commonly known as AAC+ that is approximately three times more efficient than MP2. Therefore, broadcasters using DAB+ will be able to provide far higher audio quality or far more stations than they can on DAB, or, as is most likely, a combination of both higher audio quality and more stations will be provided.

21.8.2 Digital Video Broadcast (DVB)

Like radio stations, television has also gone digital and is using the DVB (Digital Video Broadcasting) technology. DVB is a term that is generally used to describe digital television and data broadcasting services that comply with the DVB standard. It delivers compressed images, sound/music or data to the receiver. No restrictions exist as to the kind/format of information in the data containers. The service information in DVB acts like a header to the container ensuring that the receiver knows what it needs to decode.

As such, there is no single DVB standard, but rather a collection of standards, technical recommendations and guidelines. These were developed by the Project on Digital Video Broadcasting. The specifications in DVB concern:

- General aspects of digital broadcasting.
- Channel coding.
- Source coding of audio, data and video signals.
- Transmitting DVB signals over terrestrial and satellite communications paths.
- Scrambling and conditional access.
- Software platforms in user terminals.
- User interfaces supporting access to DVB services.
- The return channel for support purposes.
- Interactive services.

The DVB specifications are interrelated with other recognized specifications. DVB source coding of audio-visual information as well as multiplexing is based on the standards evolved by Moving Picture Experts Group (MPEG). The major objective of the DVB project is to reap the benefits of technical standardization while satisfying the commercial requirements of the project members. Although a large part of the standardization work is now complete, work is still ongoing on issues such as the Multimedia Home Platform. Until now, much of the output of the DVB project has been formalized by ETSI.

At present, the majority of DVB satellite transmissions convey multiple SDTV programmers and associated audio and data. One important feature of DVB is that it is useful for data broadcasting services such as access to World Wide Web.

21.8.3 IPTV

Next generation networks will see convergence of all types of media over IP. This includes voice, radio, TV, and multimedia. Television broadcast over the Internet Protocol is called IPTV. IP/TV was an Mbone (Multicast Backbone) compatible application that moved single and multi-source audio/video traffic, ranging from low to DVD quality, using both unicast and IP multicast RTP/RTCP. Traditionally IP is used for unicast traffic between two points. However, for multimedia services IP needs to be multicast. IGMP (Internet Group Management Protocol) is used for this purpose. The current version of IGMP is version 3 (RFC 3376). A station that wishes to become a receiver sends an IGMP "group join" message to that group's transmitter. Each Layer 3 device that forwards an IGMP join message records the group ID and source interface in its multicast forwarding table. When the transmitter sends IP multicast traffic, the Layer 3 device will then

forward the traffic only to those interfaces from which it has received join messages. Destination IP addresses for multicast traffic fall within the range of 224.0.0.1 through 239.255.255.255 (although some addresses within this range are reserved). Destination Ethernet addresses for multicast traffic begin with 01:00:5E and end with the lower order 23 bits of the destination IP address. Many of the world's major telecommunications providers are exploring IPTV as a new revenue opportunity from their existing markets and as a defensive measure against encroachment from more conventional cable television services. In India BSNL is already offering IPTV through its Internet network. AT&T in the US launched its U-Verse IPTV service in 2006. AT&T offers over 300 channels.

IPTV uses a two-way digital broadcast signal sent through a switched telephone or cable network by way of a broadband connection and a set-top-box (STB) programmed with software that can handle viewer requests to access to many available media sources. The STB is programmed to allow only these channels that the subscriber has paid for. This is managed through close coordination between the CAS (Conditional Access System) at the head-end and a programmed smart-card within the STB.

IPTV covers both live TV as well as stored video through VoD (Video on Demand). The playback of IPTV requires either a PC or a STB connected to a conventional TV. Video content is typically compressed using either a MPEG-2 or a MPEG-4 codec and then sent in an IP Multicast in case of live TV or via IP Unicast in case of Video on Demand.

21.8.4 Internet TV

IPTV deals with TV broadcast over satellite, terrestrial, or mobile network. In IPTV the receiver device is still a HDTV (High Definition TV) or a SDTV (Standard Definition TV). By contrast "Internet TV" generally refers to transport streams sent over IP networks (normally the Internet) with receiver station being a PC or computer. An Internet TV provider has no control over the final delivery and so broadcasts on a "best effort" basis. Elementary streams over IP networks and proprietary variants as used by websites such as YouTube are now rarely considered to be IPTV services. Another main difference between IPTV and Internet TV is that IPTV are paid services that carry both free channels and paid channels; whereas, Internet TV is generally free.

21.9 Multiple Play

Multiple play is a way to describe provision of diverse telecommunication services by vendors that usually offered one or two of such services. Some of the very common such telecommunication services are broadband Internet access, television, telephone (both wired and wireless), etc. Multiple play is a generic term for conglomeration of more than one service into a single bundled product or service. Terms like Triple Play or Quadruple Play describe specific service bundles.

21.9.1 Triple Play

Triple Play means provisioning of two broadband services—high-speed Internet access and television, and one narrowband service—telephone over a single broadband connection. Triple Play is, usually, delivered using a DSL link. In Triple Play, television contents are delivered through

DVB technology using set-top-box. Internet access is usually provisioned through an Ethernet port while voice services can be provided using either existing telephone network or voice over IP (VoIP). There is no standard configurations to offer Triple Play services. There can be different methods as well depending upon the budget at disposal, types of subscribers and geography of subscribers' premises.

The business challenges in offering Triple Play services are associated with ascertaining the right business model, backend processes, customer care support and capital expenditure. There are some technology challenges as well because voice, video and high speed data all have different characteristics and place different burdens on the network that provides access to these services.

21.9.2 Quadruple Play

Triple Play has led to a new term called Quadruple Play where wireless communications is introduced as another medium to deliver Triple Play services. Quadruple Play can be thought as Triple Play plus mobility. In the traditional business model of offering Quadruple Play services, the following are the major drawbacks:

- The networks are separate as the individual services have been offered using their own networks. Usually, wired access has been about voice and broadband access is a data service moving towards entertainment. Wireless networks deliver data and entertainment and, again, IPTV is about video. An operator may be maintaining two or even three different network infrastructures to deliver all these services.
- The user experiences are separate as subscribers use different devices, interfaces and methods to access various services which do not interact with each other.
- The billing systems are separate as subscribers may see all their service charges integrated on one bill but behind the scenes are disparate billing systems that must be maintained separately for each service.

Because of duplication of infrastructure, operating expenses for operators are high. It is difficult and costly to introduce new services. Even if consistent user experience for a service across multiple media can be created, the applications would have to be developed separately on each platform. For subscribers, too, there is not much benefit in moving all their services to one provider, besides the promise of a better price.

However, with advances in Internet Protocol and IP Multimedia Subsystem technologies, there are better options such as:

- IP provides a cost efficient way to converge voice, data and video transport on to a unified network infrastructure.
- IMS provides the next generation network architecture that converges voice, data and IPTV service attributes over multiple access types into a consistent user experience.
- By linking IPTV with IMS, television set-top-boxes can be added to the list of IMS endpoints along with mobile phones, personal computers and other consumer entertainment devices.

Due to advantages garnered by IP and IMS, users can enjoy a consistent user experience across various devices and access networks. Service providers can offer services that help users manage their personal libraries of commercial and private content. Then, such services can be extended to multiple devices and access networks. However, customers are more eager for just price breaks

and a single bill. They are looking for simplicity in managing and using their various services and devices while willing to pay a bit of a premium for it.

21.10 Future Trends

In this section, we discuss some other emerging technologies that look promising in the NGN space. These are 3GPP LTE, and iBurst.

21.10.1 3GPP Long Term Evolution

The 3GPP Long Term Evolution (3GPP LTE) project would result in the release 8 (of 3GPP) having extensions and modifications of the current UMTS system. Sometimes called a fourth generation mobile communications technology, 3GPP LTE will be a wireless broadband Internet system with other services (like voice and data) built over it. Objectively, it has the following goals:

- Backward compatibility with legacy standards.
- Optimal cell size of 5 km, 30 km cell size with reasonable performance and sometimes, more than 30 km (up to a maximum of 100 km) cell size with acceptable performance.
- Download rates of 100 Mbit/s and upload rates of 50 Mbit/s for every 20 MHz of available spectrum with support for 200 active phone calls for every 5 MHz cell.
- Small latencies for small IP packets.
- Increased spectrum flexibility with slices as small as 1.25 MHz supported.

The project is ongoing and general in scope and yet to experience exhaustive trials and demonstrations before touching base.

21.10.2 iBurst

iBurst or High Capacity Spatial Division Multiple Access (HC-SDMA) is a wireless broadband technology which optimally uses the available bandwidth by using smart antennas. HC-SDMA interface provides wide area broadband wireless connectivity for fixed, portable and mobile computing devices. It can be implemented with smart antenna array techniques to substantially improve the radio frequency coverage and performance of the system.

References/Further Reading

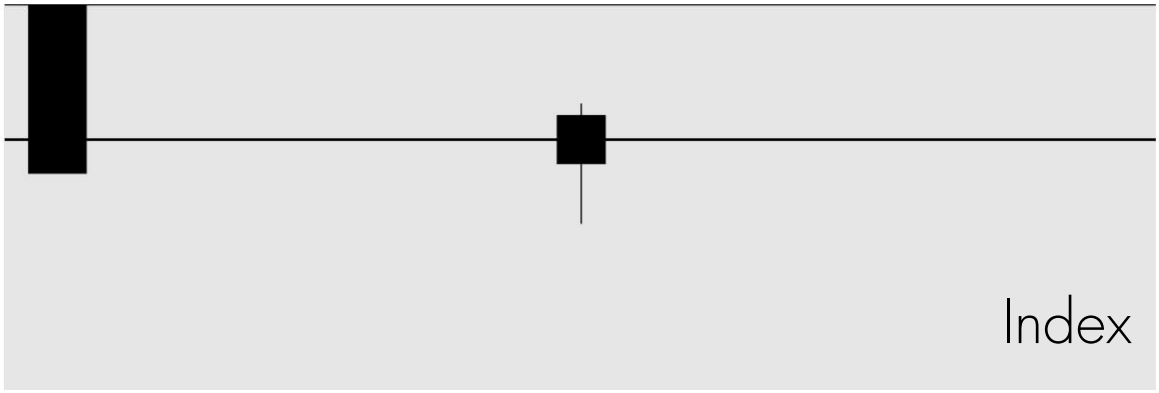
1. Awater GA and Kruys J, *Wireless ATM–An overview*, Lucent Technologies.
2. Armstrong, J. (1999), "Analysis of new and existing methods of intercarrier interference due to carrier frequency offset in OFDM", *IEEE transactions on communications,* Vol 47, No. 3.
3. Bahl V, *Microsoft Research, Future Directions–HIPERLAN, Wireless ATM and FPLMTS.*
4. Baum Kevin L. (1998), "A Synchronous Coherent OFDM Air Interface Concept for High Data Rate Cellular Systems", *IEEE conference proceedings VTC.*
5. Chen Y, *Maximizing data download capabilities for future constellation space missions*, Umea University.

6. Dagiuklas T., C. Politis, S. Grilli, G. Bigini, D. Sisalem, Y. Rebahi, and R. Tafazolli, *Seamless Multimedia Network and Service Access Over All-IP Based Infrastructures: The EVOLUTE Approach.*
7. DW Griffith, Johnson RA, Unkauf MG and Moy A, Multiple Access Raytheon Systems Company, *Modeling Performance improvements in MILSTAR MDR Networks with Demand Assignment.*
8. http://airave.sprint.com
9. http://www.itu.int/ITU-T/ngn/fgngn/index.html
10. http://www.itu.int/ITU-D/imt-2000/documents/Busan/Session3_TTA.pdf
11. Marnik M., *Redefining the Quad Play with IPTV and IMS*, Juniper Networks.
12. Molins M. and Stojanovic, M. *Slotted FAMA: A MAC Protocol for Underwater Accoustic Networks,* Massachussets Institute of Technology.
13. Next Generation Networks, The International Engineering Consortium.
14. Proakis, J., (1995), *Digital Communications*, McGraw-Hill, Inc.
15. Raychaudhuri D. and N.D. Wilson, "ATM based transport Architecture for multiservice wireless personal communications networks", *IEEE J. Select. Areas Communication.*
16. Rickard N., *ABR, Realizing the promise of ATM, Telecommunications*, International edition.
17. St. Arnaud, Bill, Canarie, Inc., *MPLS and Architectural Issues for an Optical Internet Architecture and System Design Session.*
18. Semeria, C., *Multiprotocol Label Switching–Enhancing Routing in the New Public Network*, Juniper Networks.
19. Tariq, S. "MAC Algorithms in Wireless Networks" Masters Thesis, Umea University.
20. Wikipedia–The free online encyclopedia, www.wikipedia.org
21. www.arraycomm.com
22. www.atis.org
23. www.cisco.com
24. www.flarion.com
25. www.iburst.org
26. www.iptp.net
27. www.qualcomm.com
28. www.radio-electronics.com
29. www.rohde-schwarze.com
30. www.stanford.edu
31. www.truphone.com
32. www.tp-alliance.de
33. www.wibro.or.kr
34. www.wimaxforum.org
35. Yan, T.Y., Cheng U, Wang C, Dessouky K and Rafferty W, Jet Propulsion Laboratory, *A FD/DAMA Network Architecture for the First Generation Land Mobile Satellite Services.*

36. Zou, W.Y. and Wu, Y. (1995), "COFDM : An Overview", *IEEE transactions on broadcasting*, Vol. 41, No. 1.
37. Z Sun (2005), *Satellite Networking: Principles and Protocols*, John Wiley and Sons Ltd.
38. www.3gpp.org

REVIEW QUESTIONS

Q1. What is NGN? How is it different from any other generation network? How shall it be popular?
Q2. Explain the following transitions with respect to an NGN network:
  (a) Circuit switching to Packet switching.
  (b) IT and CT into ICT.
  (c) Narrowband to Broadband.
Q3. Explain the role of MPLS in service provisioning. Why is it called 2.5 layer protocol?
Q4. Explain in brief:
  (a) MSP (b) ISP (c) TSP
Q5. Explain OFDM in detail. What is the use of guard period?
Q6. How is orthogonality helpful in OFDM?
Q7. What is an access network? Explain any three of them in details along with their applications.
Q8. What do you mean by All IP Network? Explain its vital characteristics.
Q9. Explain FAMA and DAMA. Bring out the difference between the two of them.
Q10. What is WATM? How does it improve over ATM?
Q11. What are multimedia broadcast services? How is it essential nowadays?
Q12. What is Multiple Play? Explain its types and applications.
Q13. Explain HSPA family of protocols and applications. How would they change the applications of existing networks?
Q14. Write short notes on:
  (a) HSPA (b) HSDPA (c) HSUPA
  (d) HSOPA (e) Evolved HSPA (f) DVB
  (g) DAB (h) IPTV (i) Internet TV
  (j) Triple Play (k) Quadruple Play (l) 3GPP LTE
  (m) iBurst (n) WiBro



Index

Authors' Profiles

Dr. Asoke K. Talukder is Chief Scientific Officer and Director of Geschickten Solutions, Bangalore. He is also Adjunct Faculty, ABV Indian Institute of Information Technology and Management, Gwalior; Adjunct Professor, Department of Computer Science and Engineering, NIT Warangal; and, Adjunct Faculty at Department of Computer Engineering, NITK, Surathkal. He was also the DaimlerChrysler Chair Professor at IIIT, Bangalore and Visiting Professor at VIT University, Vellore.

Dr. Talukder has been with the IT industry for about 30 years. He has held senior positions in different technology companies in India, USA, UK, and Singapore. He was the founder CTO of Cellnext, the pioneering wireless and Mobile Web technology company in India offering technology and solutions in the domains of GSM, GPRS, SMS, MMS, Intelligent Networks, CDMA, and 3G. He has been Corporate Advisor to SaharaNext, Lucknow; Advisor to the Expert Committee, HCL Bangalore; and the CTO and Executive Director––Telecom, Sobha Renaissance Information Technology, Bangalore**.** He has also worked in complex projects for companies like Fujitsu-ICIM, Microsoft, Oracle, Informix, Digital, Hewlett Packard, ICL, Sequoia, Blue Star Infotech, Northern Telecom, NEC, KredietBank, iGate, and many more.

Dr. Talukder did his Ph. D. (Computer Engineering) in Telecom Routing after completing his post graduation in Physics from the University of Calcutta (1976). He set up the first X.25 network in India for the Department of Telecommunications in 1986. Later, he set up the first Java Centre in India in 1998. He was a key engineer for the Oracle Parallel Server for Hewlett-Packard HP-FX fault tolerant computers, as well as the architect for the 64 bit Informix database for DEC Alpha.

He is the recipient of many international awards including All India Radio/Doordarshan Award, ICL Services Trophy, ICL Chief Executive Excellence Award, Atlas Club Excellence Award, ICIM Professional Excellence Award, ICL Excellence Award, IBM Solutions Excellence Award—one of his ubiquitous middleware products was the recipient of this award in 2001, and the Simagine GSM World Congress Award––one of his products on Java Card security was recipient of this award in 2003.

Dr. Talukder has been listed in *Who's Who in the World*, *Who's Who in Science and Engineering*, and *Outstanding Scientists of 21st Century*. He published many peer-reviewed research papers and book chapters. He edited two books and authored two textbooks in the domain of Mobile Computing and Secure Software Engineering.

Dr. Talukder has several patents to his name.

His area of expertise include Product Innovation and Management; Next Generation Internet and Services, Mobile Computing, Secure Software Engineering, Trust and Reputation. Web-based systems and applications include Web 2.0 and Web 3.0, Databases Internals, Compilers, Operating Systems, Cloud Computing and High Performance Computing. His expertise in telecommunications is in the areas of OSS, BSS, Inter-carrier Wholesale Billing, and Intelligent Networks (IN). His current domains of interest are Mobile Computing, Information Security, and Computational Quantitative Biology.

Hasan Ahmed is with Nokia Research and Development, Bangalore. He has experience in the financial domain too. He worked as Information Technology Officer with a large Indian bank prior to his M.Tech from IIIT Bangalore. He is a B. Tech in Information Technology from UPTU, Lucknow. His areas of interest are ubiquitous computing and software engineering.

Roopa R. Yavagal is with Symphony Services, Bangalore. During her stint earlier at Cellnext she, along with her team, developed many mobile applications for various cellular companies in India in the technology domain of GSM, GPRS, and CDMA. Roopa did her Masters from IIIT, Bangalore and her B.E. in Computer Science from Basaveshwar Engineering College, Bagalkot. Her areas of interest include wireless applications, service creation in the wireless domain, and Web 2.0 applications.