4G MOBILE ARCHITECTURE

Rhituparna Paul (0306052)

Nishat Kabir (0306055)

Tahnia Farheen (030680)

A thesis submitted in partial
fulfillment of the requirements for the degree of
BACHELOR OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

December 31st, 2008

Declaration

It is hereby declared that this dissertation is the result of our own work except where explicit reference is made to the work of others, and has not been submitted elsewhere for the award of any degree or diploma.

(Supervisor)
Dr. Md. Abdul Matin
Professor
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology.

(Authors)

Rhituparna Paul (0306052)

Nishat Kabir (0306055)

Tahnia Farheen (0306080)

Acknowledgement

We would like to convey the deepest gratitude and veneration to our honorable thesis supervisor Prof Dr. Abdul Matin, department of Electrical & Electronics Engineering (EEE), Bangladesh University of Engineering And Technology (BUET), Dhaka, Bangladesh, for engaging us in such an important and advanced technology as our "Undergraduate Thesis Program". He is an absolutely perfect guide and his knowledge is so vast in this field that his advice was sufficient for us to perform our thesis. He was always there to lend his abysmal expertise. Moreover, his affection and encouragement helped us greatly to work on such an advanced technology. We are also grateful to the persons who helped us for collection of available information and allowed us to use library facilities.

Contents

List of Figures

Preface

The goal for the next generation of mobile communication system is to seamlessly integrate wide variety of communication services such as high speed data , video and multimedia traffic as well as voice signals. The technology needed to meet theses challenges available is popularly known as Forth Generation (4G) Mobile Systems. There is some different architecture of 4G which meet the demands of new era. In this thesis, different architectures of 4g mobile communication system is described. We here tried to gave a elaborate description of the architectures. We emphasized on new QoS requirement for 4G mobile for real time multimedia service like interactive video conferencing or streaming video and audio in IP network. Finally, we presented not only technical perspective but also economical view to generate sufficient customer demand for 4G networks and possibly beyond by developing applications that are meaningful and attractive to customers and at the same time commercially viable from the standpoint of service providers.

Chapter 1

Introduction to 4G

1.1 What is 4G?

4G is known as beyond 3G,stands as an acronym for Fourth-Generation Communications System. It is used to describe the next step in wireless communications. A 4G system will be able to provide a comprehensive IP solution where voice, data and streamed multimedia can be given to users on an “Anytime, Anywhere” basis, and at higher data rates than previous generations. There is no formal definition for what 4G is; however, there are certain objectives that are projected for 4G.

These objectives include: that 4G will be a fully IP-based integrated system. This will be achieved after wired and wireless technologies converge and will be capable of providing between 100 Mbit/s and 1 Gbit/s speeds both indoors and outdoors, with premium quality and high security. 4G will offer all types of services at an affordable cost.

1.2 Objectives and Approaches

4G is being developed to accommodate the quality of service (QoS) and rate, requirements set by forthcoming applications like wireless broadband access, Multimedia Messaging Service, video chat, mobile TV, High definition TV content, Digital Video Broadcasting (DVB), minimal service like voice and data, and other streaming services for “anytime-anywhere”. The 4G working group has defined the following as objectives of the 4G wireless communication standard:

- A spectrally efficient system (in bits/s/Hz and bits/s/Hz/site)
- High network capacity: more simultaneous users per cell
- A nominal data rate of 100 Mbit/s while the client physically moves at high speeds relative to the station, and 1 Gbit/s while client and station are in relatively fixed positions as defined by the ITU-R.
- A data rate of at least 100 Mbit/s between any two points in the world
- Smooth handoff across heterogeneous networks
- Seamless connectivity and global roaming across multiple networks
- High quality of service for next generation multimedia support (real time audio, high speed data, HDTV video content, mobile TV, etc)
- Interoperability with existing wireless standards, and
- An all IP, packet switched network.

1.3 History of digital communication system

1.3.1 First Generation Cellular System

Almost all of the systems from this generation were analog systems where voice was considered to be the main traffic. These systems could often be listened to by third parties. These systems could often be listened to by third parties. The 1G cellular telephone system divided cities into small cells. This division allowed extensive frequency reuse across a city, allowing millions to use cell phones simultaneously. Numerous incompatible analog systems were placed in service around the world in First Generation.this generation uses FM technology for voice transmission and digital signaling for control information. Other first generation systems include:

- AMPS (Advanced Mobile Phone System)
- NAMPS (Narrowband AMPS)
- TACS (Total Access Cellular System)
- NMT-900 (Nordic Mobile Telephone System).

1.3.2 Second Generation Cellular System

All the standards belonging to this generation are commercial centric and they are digital in form. The 2G (second generation) systems designed were still used mainly for voice applications but were based on digital technology, including digital signal processing techniques. These 2G systems provided circuit-switched data communication services at a low speed. This system operates nationwide or internationally and today's mainstream system, although the data rate for users in this system is very limited. Around 60% of the current market is dominated by European standards. Second generation includes :

- USDC(United States Digital Cellular Standards IS-54 and Is-1 36)
- GSM(Global System For Mobile Communications)
- PDC(Pacific Digital Cellular)
- CDMA-1.

1.3.3 Third Generation Cellular System

3G systems promise faster communications services, including voice, fax and Internet, anytime and anywhere with seamless global roaming. ITU's IMT-2000 global standard for 3G has opened the way to enabling innovative applications and services. The first 3G network was deployed in Japan

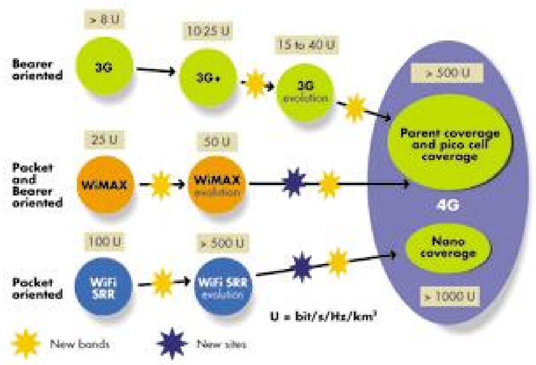


Figure 1.1: 4G Network Coverage

in 2001. 2.5G networks, such as GPRS (Global Packet Radio Service) are already available in some parts of Europe. The systems in this standard are basically a linear enhancement of 2G systems. They are based on two parallel backbone infrastructures, one consisting of circuit switched nodes, and one of packet oriented nodes. The ITU defines a specific set of air interface technologies as third generation, as part of the IMT-2000 initiative. Currently, transition is happening from 2G to 3G systems. For third generation mobile (3G, FOMA) data rates are 384 kbps (download) maximum, typically around 200kbps, and 64kbps upload. 3G networks includes

1.3.4 Forth Generation Cellular System

According to the 4G working groups, the infrastructure and the terminals of 4G will have almost all the standards from 2G to 4G implemented. Even though the legacy systems are in place to be adopted in 4G for the existing legacy users, going forward the infrastructure will however only be packet based, all-IP.

Also, some proposals suggest having an open platform where the new innovations and evolutions can fit. The technologies which are being considered as pre-4G are used in the following standard version: WiMax, WiBro, 3GPP Long Term Evolution and 3GPP2 Ultra Mobile Broadband. Fourth generation (4G) mobile communications will have higher data transmission rates than 3G. 4G mobile data transmission rates are planned to be up to 20 megabits per second.

At present the download speed for mode data is limited to 9.6 kbit/sec which is about 6 times slower than an ISDN (Integrated services digital network) fixed line connection. Recently, with 504i handsets the download

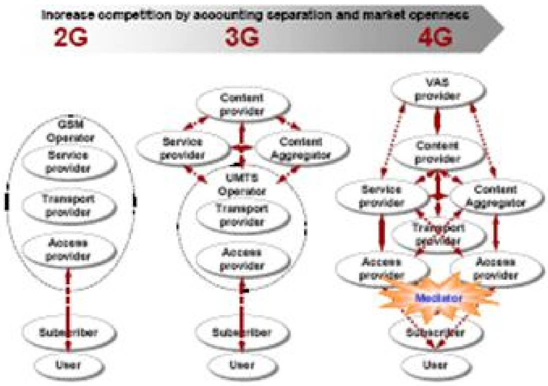


Figure 1.2: Evaluation Of Mobile Value Chain Toward 4G

data rate was increased 3-fold to 28.8kbps. However, in actual use the data rates are usually slower, especially in crowded areas, or when the network is “congested”.

1.4 Reasons To Have 4G

1. Support interactive multimedia services: teleconferencing, wireless Internet, etc.
2. Wider bandwidths, higher bit rates.
3. Global mobility and service portability.
4. Low cost.
5. Scalability of mobile networks.
6. TV, internet, phone, radio, home environment sensors all reachable through one device — the cell phone.

1.5 What’s new in 4G

1.5.1 Business On-The-Move

Even before leaving home to reach the place of a work appointment, the user would like to receive information about train/subway correspondences, door-to-door delays, etc., and more personalized ones, such as knowing how long it takes walking to get on the first correspondence, in order to eventually wait for the next train. According to the user’s decisions, his time-plan must be consequently scheduled in the most efficient way. During his stay on the train, the user would like to download e-mails, listen to the radio, watch the TV, etc. (the environment also enforces the range of applications the user can exploit).

1.5.2 Smart Shopping

The user would like to receive pop-ups informing him of some offer not only when passing by or through a shopping mall, but also anywhere else, where he can start to think about his spare time and maybe plan some fruitful shopping hours. With such service, the targeted advertisements become useful and even precious information for the user. These are not as annoying as massive ones because they become personalized and thus answer a real need. In particular, after giving some hints to the operator about his preferences and hobbies, the user gets, without extra efforts, useful and needed information.

1.5.3 Mobile Tourist Guide

A tourist can use his terminal to get instructions about the way to reach some sightseeing place, but also to interact with the environment in his surroundings warning him in case of some possible worth deviation or giving information about something that is on the way to the final destination that he may most likely miss. Moreover, in a museum instead of buying the brochure or renting some devices that most of the times are defective, all he needs is to download a package in his language for a certain price and enjoy his tour listening to the audio guidance in the earphones of the terminal. For each painting or sculpture, he can automatically listen to the comments and explanations without any effort of browsing through the guide. Also, by buying the ticket via his terminal or by signing up online on the waiting list which sends him back the approximate waiting time, avoiding long queues. The user terminal can also provide information about the culinary specialties of the city/region, where the nearest restaurant for getting a typical meal is situated.

1.5.4 Personalization Transfer

In a music festival or during a concert, the user wants to take pictures and record special moments with his friends and/or the entire event. He has a hand-held device - the most convenient to carry in a concert - that can support such demand. On the way back the pleasure of watching the pictures or videos is not limited on such device, since he can transfer the content to a publicly available larger screen - on the bus, at the train station, at the airport, etc. - and enjoy fully with his friends and the other people that were at the concert. This is very practical when the user is in an outdoor festival for many days because this does not require possessing more than one hand-held device. Then, back home, he can watch the entire recordings on a big screen, even without the support of any computer, but just relaying the video from his terminal to the available monitor. This can easily be also applied to other multimedia content, such as movies, football matches, or TV programs.

1.6 Key Parameters of 4G with respect to 3G

Key Parameters	3G	4G
Major Requirement Driving architecture	Predominantly voice driven - data was always add on	Converged data and voice over IP
Network Architecture	Wide area cell-based	Hybrid - Integration of Wireless LAN (WiFi, Bluetooth) and wide area
Speeds	384 Kbps to 2 Mbps	20 to 100 Mbps in mobile mode
Frequency Band	Dependent on country or continent (1800-2400 MHz)	Higher frequency bands (2-8 GHz)
Bandwidth	5-20 MHz	100 MHz (or more)
Switching Design Basis	Circuit and Packet	All digital with packetized voice
Access Technologies	W-CDMA, 1xRTT, Edge	OFDM and MC-CDMA (Multi Carrier CDMA)
Forward Error Correction	Convolution rate 1/2, 1/3	Concatenated coding scheme
Component Design	Optimized antenna design, multi-band adapters	Smarter Antennas, software multiband and wideband radios
IP	A number of air link protocols, including IP 5.0	All IP (IP6.0)

Chapter 2

About 4G

2.1 What's new in 4G

Unlike previous standards such as 3G (third generation), 4G is based entirely on packet switched networks. In addition, all 4G networks will be digital and will provide higher bandwidths of up to 100 Mbps. 4G is actually a collection of previous standards as oppose to an entirely new standards. Standards such as 3G and Bluetooth will be incorporated in to the 4G standards.

2.2 4G Networks

Clearwire and Intel have teamed up to produce the world's first 4G network that launched on January 6, 2009, in Portland, Oregon. The network called Clear allows consumers and businesses the chance to connect wirelessly anywhere in Portland at true broadband speeds.

Clearwire's 4G network will be available in major metropolitan areas across the U.S. Clearwire's network is currently only available in Baltimore, Maryland, and Portland, Oregon. Baltimore's network will be re-branded from XOHM to Clear in the coming months. Today marked the grand opening events at the Clear retail stores, starting at 9 am to noon.

WiMax is the new technology used on the network, delivers speeds up to 6mbps download using a WiMax-enabled USB modem that plugs directly into their laptops, while mobile users can expect speeds up to 4mbps download. The Motorola modem retails for $49.99 either in store or online, or users can lease the WiMax-enabled USB in home modem that connects to any outlet and a USB port in either their laptop or desktop PC, for as little as $4.99 a month.

Sprint Nextel announced today that it plans to use WiMax technology to power its next-generation high-speed wireless network.

The Sprint Nextel 4G mobility network will use the company's extensive 2.5 GHz spectrum holdings, which cover 85 percent of the households in the top 100 U.S. markets.

In an effort to offer faster speeds, lower cost, and greater convenience and enhanced multimedia quality, Sprint Nextel will work with Intel, Motorola, and Samsung to incorporate WiMaX technology for advanced wireless communications and help make chipsets widely available for new consumer electronics devices.

The new WiMax network would provide download speeds of between 2-4 Mbps, on par with DSL or cable modems. Plans target a launch in trial markets by the end of 2007, with deployment that reaches as many as 100 million people in 2008. Sprint Nextel intends to expand mobile WiMaX network coverage thereafter.

2.3 User Friendliness and User Personalization

In order to encourage people to move towards a new technology, which is a process that usually takes a long time and a great deal of effort from the operators side, a combination of user friendliness and user personalization appears to be the winning concept. User friendliness exemplifies and minimizes the interaction between applications and users thanks to a well-designed transparency that allows the users and the terminals to naturally interact (e.g., the integration of new speech inter-faces is a great step for achieving this goal). For instance,in scenario A, users can get traveling information in the most user-friendly way: text, audio, or video format. User personalization refers to the way users can configure the operational mode of their device and preselect the content of the services chosen according to their preferences. Since every new technology is designed keeping in mind the principal aim to penetrate the mass market and to have a strongly impact on people's lifestyles, the new concepts introduced by 4G are based on the assumption that each user wants to be considered as a distinct, valued customer who demands special treatment for his or her exclusive needs. Therefore, in order to embrace a large spectrum of customers, user personalization must be provided with high granularity, so that the huge amount of information is filtered according to the users' choices. This is illustrated in scenario B, where users canreceive targeted pop-up advertisements. The combination between user personalization and user friendliness provides users with easy management of the overall features of their devices and maximum exploitation of all the possible applications, thus conferring the right value to their expense.

2.4 ACCESS SCHEMES

As the wireless standards evolved, the access techniques used also exhibited increase in efficiency, capacity and scalability. The first generation wireless standards used plain TDMA and FDMA. In the wireless channels, TDMA proved to be less efficient in handling the high data rate channels as it requires large guard periods to alleviate the multipath impact. Similarly, FDMA consumed more bandwidth for guard to avoid inter carrier interference. So in second generation systems, one set of standard used the combination of FDMA and TDMA and the other set introduced a new access scheme called CDMA. Usage of CDMA increased the system capacity and also placed a soft limit on it rather than the hard limit. Date rate is also increased as this access scheme is efficient enough to handle the multipath channel. This enabled the third generation systems to used CDMA as the access scheme IS-2000, UMTS, HSXPA, 1xEV-DO, TD-CDMA and TD-SCDMA. The only issue with the CDMA is that it suffers from poor spectrum flexibility and scalability.

As the wireless standards evolved, the access techniques used also exhibited increase in efficiency, capacity and scalability. The first generation wireless standards used plain TDMA and FDMA. In the wireless channels, TDMA proved to be less efficient in handling the high data rate channels as it requires large guard periods to alleviate the multipath impact. Similarly, FDMA consumed more bandwidth for guard to avoid inter carrier interference. So in second generation systems, one set of standard used the combination of FDMA and TDMA and the other set introduced a new access scheme called CDMA. Usage of CDMA increased the system capacity and also placed a soft limit on it rather than the hard limit. Date rate is also increased as this access scheme is efficient enough to handle the multipath channel. This enabled the third generation systems to used CDMA as the access scheme IS-2000, UMTS, HSXPA, 1xEV-DO, TD-CDMA and TD-SCDMA. The only issue with the CDMA is that it suffers from poor spectrum flexibility and scalability.

The other important advantage of the above mentioned access techniques is that they require less complexity for equalization at the receiver. This is an added advantage especially in the MIMO environments since the spatial multiplexing transmission of MIMO systems inherently requires high complexity equalization at the receiver.

In addition to improvements in these multiplexing systems, improved modulation techniques are being used. Whereas earlier standards largely used Phase-shift keying, more efficient systems such as 64QAM are being proposed for use with the 3GPP Long Term Evolution standards.

2.5 Mobile VOIP

- "Voice Over Internet Protocol"
- Allows only packets (IP) to be transferred, eliminating complexity of 2 protocols over the same circuit
- Allows only packets (IP) to be transferred, eliminating complexity of 2 protocols over the same circuit

  - lower latency data transmission (faster transmission)

- Samples voice between 8,000 & 64,000 times per second and creates stream of bits which is then compressed and put into a packet.

- Increases battery life due to greater data compression.

Figure 2.1: Mobile VOIP

2.6 MIMO

Imagine that you can hear better what you want to listen, and don't hear what bothers you, just by pointing out where message and noise are coming from. Beamforming that is the significant concept of MIMO (Multiple-Input Multiple-Output) allows you do just that using smart antennas system. But that's not all what MIMO has to offer. Spatial multiplexing, achieved by independent simultaneously working antennas, increases bandwidth capacity by modulating and transmitting signal through many paths [15]. Using space-time coding, reliability is improved. MIMO achieves great success thanks to multiple antennas that allows simultaneous directional transmission of two or more unique data streams sharing the same channel. Increasing speed and range, MIMO is already accepted by researchers as one of the main components of projects such as WiBro, WiMAX, WLAN, 802.11n, UMTS R8 LTE, and UMB.

2.7 Quality of Service Challenges

In wireless networks, Quality of Service (QOS) refers to the measure of the performance for a system reflecting its transmission quality and service availability (e.g., 4G is expected to have at least a reliability of 99.99%).

Supporting QOS in 4G networks will be a major challenge. When considering QOS, the major hurdles to overcome in 4G include: varying rate channel characteristics, bandwidth allocations, fault tolerance levels, and handoff support among heterogeneous wireless networks. Fortunately, QOS support can occur at the packet, transaction, circuit, and network levels.

QOS will be able to bet weaked at these different operating levels, making the network more flexible and possibly more tolerant to QOS issues.

Varying rate channel characteristics refers to the fact that 4G applications will have varying bandwidth and transition rate requirements. In order to provide solid network access to support the anticipated 4G applications, the 4G networks must be designed with both flexibility and scalability.

Varying rate channel characteristics must be considered to effectively meet user demand and ensure efficient network management. Spectrum is a finite resource. Incurrent wireless systems, frequency licensing and efficient spectrum management are key issues. In 4G systems, bandwidth allocations may still be a concern.

Another concern is interoperability between the signaling techniques that are planned to be used in 4G (e.g., 3xRTT, WCDMA). In comparison with current 2G and 2.5G networks, 4G will have more fault tolerance capabilities built-in to avoid unnecessary network failure, poor coverage,and dropped calls. 4G technology promises to enhance QOS by the use of better diagnostic techniques and alarms tools. 4G will have better support of roaming and handoffs across heterogeneous networks. Users, even in today's wireless market, demand service transparency and roaming. 4G may support interoperability between disparate network technologies by using techniques such as LAS-CDMA signaling. Other solutions such as software defined radios could also support roaming across disparate network technologies in 4G systems. These major challenges to QOS in 4G networks are currently being studied and solutions are being developed.

Developers believe that QOS in 4G will rival that of any current 2G or 2.5G network. It is anticipated that the QOS in 4G networks will closely approximate the QOS requirements in the wireline environment (99.999% reliability).

2.8 Challenges in the Migration to 4G Mobile Systems

There exist a lot of challenges in design and implementation of 4G systems that will support all the expected features based on current communication systems and standards. One big challenge is how to implement handoffs in IP-based 4G networks with minimum handoff latency and packet loss.Traditionally, handoff management means that system maintains communication connection(s) with a mobile node (MN) when the MN moves from current serving area to a new serving area in a same system. However, in 4G systems, handoff management is more complex to deal with, as it covers not only horizontal handoff but also vertical handoff. Horizontal handoff handles the intra-system handoff when an MN moves between two different cells or access points within the same wireless communication system ,while vertical handoff deals with the inter-system handoff when an MN moves from one wireless communication system to another different wireless system, for example, from GSM cellular network to Wireless LAN network. It is difficult to realize the vertical handoff among different wireless communication systems while meeting the various Quality of Service (QoS) requirements. If handoff latency (i.e., the time spent in handoff) is too long, packets may get lost or disconnections may occur during the handoff, which obviously degrade the QoS in 4G systems. Therefore, fast and seamless handover is a big challenge for 4G heterogeneous networks that are supposed to support real-time highspeed multimedia applications that require small handoff delay and high data-rate transmission.

2.9 Potential applications of 4G

With its high bandwidth and incorporation of several standards, 4G will provide for a vast number of presently non-existent applications for mobile devices. Some believe that it will be possible for users to purchase groceries, watch movies, and open their garages all with one single mobile device. Experts have stated that 4G devices will differ from present day mobile devices in that there will be fewer navigation menus. Instead the devices will interpret the environment of the device and base actions on the users input. In addition, 4G will likely succeed where 3G failed and provide a seamless network for users who travel and require uninterrupted voice/data connections.

Chapter 3

4G QoS Architecture

Fourth generation systems (4G) will provide multimedia group communication sessions to multiple mobile users with distinct requirements. This way, it is expected the control of the quality level, connectivity and ubiquitous access for multi-user multimedia sessions across heterogeneous and mobile networks with seamless capability. This chapter analyses the requirements of a control architecture to provide Quality of Service (QoS), connectivity and seamless handover management for multi-user sessions in 4G systems, and introduces the QoS Architecture for Multi-user Mobile Multimedia (Q3M) proposal. In addition, simulation results present the efficiency of this approach concerning the session setup time and packet losses during handover.

3.1 Introduction

In the past decade, the telecommunications industry has witnessed an ever accelerated growth the usage of mobile communications. As a result, the mobile communications technology has evolved from the so-called second-generation (2G) technologies, GSM in Europe, IS-95 (CDMA) and IS-136 (TDMA) in USA, to the third generation (3G) technologies, UMTS/WCDMA in Europe and CDMA2000 in USA, being standardized by 3GPP and 3GPP2 (respectively), partnership projects between the governmental standards development organizations (SDO) of various countries. Along with the standards development for providing voice service to mobile users, a group of standards to deliver data to the mobile users have evolved from both SDOs and industry. Systems and applications, such as iMode, the mobile Internet access system developed by NTT DoCoMo, and Short Message Service (SMS) for sending and receiving short text messages for mobile phone users, have been built and continue to be developed. The WAP (wireless application protocol) forum and more recently, the Open Mobile Alliance have also been developing applications for wireless networks. In the wireless access field, Bluetooth was developed as a new cable replacement technology, which provides a short-range ($\sim$10m), low bit rate (1Mbps) access in the 2.4GHz spectrums. IEEE also developed a wireless LAN (WLAN) access family of protocol IEEE 802.11 including 802.11b (a 100m, 11Mbps access technology in the 2.4GHz spectrum), 802.11a, and 802.11g, as well as HiperLAN2 developed by ETSI. Nowadays, 802.11 has become one of the most popular and easy ways to provide wireless access for nomadic laptop users; first products of cellular phones that can access IEEE 802.11 base stations have recently been available in the market. The focus is on fourth generation (4G) mobile networks. Even though a universal consensus on what is going to be 4G is not yet reached in the industry or the literature, there is a reasonable understanding of some characteristics of 4G mobile networks. Some of the accepted characteristics are:

- All-IP based network architecture;
- Higher bandwidth;
- Support for different access networks, including WLAN technologies like IEEE 802.11;
- Full integration of"hot spot" and "cellular";
- Support for multimedia applications.

In order to clarify our vision of 4G networks, we could imagine a staff A starts a voice over IP conversation with his boss B (who is at a remote site) on his way to the airport, through an access to the UMTS system. When he arrives at the airport, WLAN access becomes available and the conversation (connectivity) between A and B is expected to be seamlessly continued (upon necessary authentication of A's credential by the network) even with a different access technology and a different operator. Furthermore, data transmission over the wireless link may desire stronger protection and the conversation between A and B may desire certain QoS support from the network. This scenario indicates the integration of different characteristics of data transmission and data protection, and possible different approaches for quality assurance. As 4G is expected to be built on all-IP-based technologies, architectural considerations in IP layer become critical to enable seamless interoperation among these technologies. While the IETF addresses the connectivity problems by its Mobile IP (MIP) protocols for both IPv4 and IPv6 networks, Quality of Service (QoS) and security insufficiencies are apparent: besides a simple IPsec support for MIP registration process , MIP mainly maintains the connectivity between a mobile node (MN) and its corresponding node (CN) while it is moving away from its home networks. It neither supports QoS nor stronger security between the MN and the network. QoS mechanisms, including resource reservation (signaling), admission control and traffic control, allow multimedia applications to get certain quality guarantee e.g., on bandwidth and delay for its packets delivery. Providing QoS guarantees in 4G networks is a non-trivial issue where both QoS signaling across different networks and service differentiation between mobile flows will have to be addressed. On the other hand, before providing network access and allocating resources for an MN, the network needs to authenticate the MN's (or the mobile user's) credential. Furthermore, a security association needs to be established between the MN and the network to ensure data integrity and encryption. Thus, in order to achieve seamless handover, mobility, QoS and security technologies must be integrated. The rest of this chapter presents a new architecture, Seamless Mobility with Security and QoS Support in 4G Networks (SeaSoS), to address these challenges, which integrates QoS signaling and key exchange into the 4G mobile networking infrastructure. We present our views on 4G networks design and analyze underlying fundamental problems in Section 3.2. Section 3.3 describes the Sea SoS architecture and how it addresses these problems. Section 3.4 compares Sea SoS with other approaches and outlines some future work.

3.2 A Basic Model for 4G Networks

QoS, security and mobility can be viewed as three different, indispensable aspects in 4G networks; however all are related to network nodes involving the controlling or the processing of IP packets for end-to-end flows between an MN and the CN. We show in this section how we view the 4G network infrastructure based on which we present the SeaSoS architecture in Section 3.3.

3.2.1 Two Planes: Functional Decomposition

Noting that an IP network element (such as a router) comprises of numerous functional components that cooperate to provide such desired service (such as, mobility, QoS and/or AAA - Authentication, Authorization and Accounting), we identify these components in the SeaSoS architecture into two planes, namely the control plane and the data plane. Figure 3.1 illustrates this method of flexible functional composition in 4G networks. As we are mainly concerned with network elements effectively at the network layer, we do not show a whole end-to-end communication picture through a whole OSI or TCP/IP stack. The control plane performs control related actions such as AAA, MIP registration, QoS signaling, installation/maintenance of traffic selectors and security associations, etc., while the data plane is responsible for data traffic behaviors (such as classification, scheduling and forwarding) for end-to-end traffic flows. Some components located in the control plane interact, through installing and maintaining certain control states for data plane, with data plane components in some network elements, such as access routers (ARs), IntServ nodes or DiffServ edge routers. However, not all control plane components need to exist in all network elements, and also not all network elements (e.g., AAA server) are involved with data plane functionalities. We refer these cases as path-decoupled control and other cases as pathcoupled control.

We argue the separation and coordination of control plane and data plane is critical for seamless mobility with QoS and security support in 4G networks, with the reasons as follows. Per-flow or per-user level actions occur much less frequent than per-packet actions, while per-packet actions are part of critical forwarding behavior, which involves very few control actions (which are typically simply to read and enforce according the install

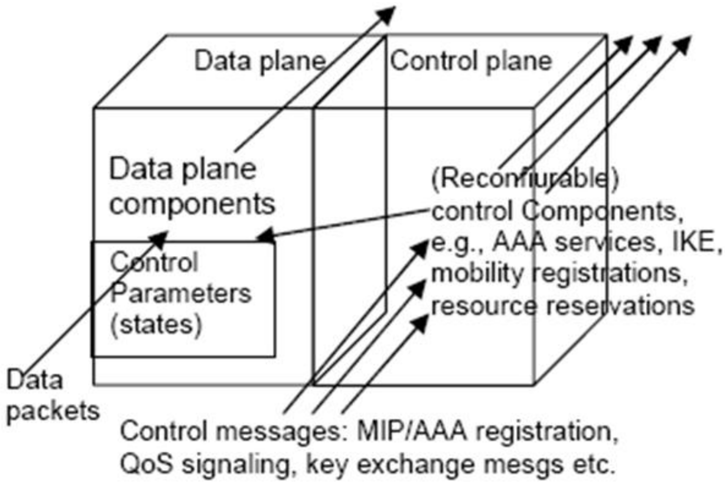


Fig. 1: The decomposition of control plane and data plane functionalities

Figure 3.1: The decomposition of control plane and data plane functionalities.

state during forwarding data). Actually, this separation concept is not new - routing protocols have the similar abstraction together used with the traditional IP packet delivery, this abstraction is recently being investigated in the IETF For CES working group. However, we emphasize the three critical dimensions of future 4G networks: mobility, QoS and security, as well as other new emerging or replacement components might appear, integrated into a unified framework and allowing more extensibility for 4G networks design.

3.2.2 Two Modes of Operation

Besides the functional decomposition, we divide the operations that a 4G network infrastructure used in mobility scenarios into two categories:

1. end-to-end way control, which is related to authentication between the mobile device and the network, and enabling of forwarding end-to-end traffic.

2. the hop-by-hop way, mainly on hop-by-hop trust relationship and resource reservation setup.

3.2.3 Understanding Mobility, QoS and Security Problems in 4G Networks

Mobility involves both control plane and data plane. The control plane is mainly involved with pathdecoupled, end-to-end way of mobility registrations, while data plane concerns mobility-enabled routing for data flows into and from an MN while it moves between different locations. The data plane behavior is achieved by installing/changing certain binding caches upon certain control plane information exchange (e.g., the binding update/acknowledge procedure in MIP). In fact, although MIP does not change the traditional IP routing table, when the MN is away from home and changes its location, associated with its fixed home address information, routing information is added in certain data processing and/or forwarding entities such as mobility agents (e.g., home agent and foreign agent) and systems themselves upon successful MIP registrations. Localized mobility solutions such as fast handover for Mobile IPv6 (FMIPv6) and Hierarchical Mobile IPv6 (HMIPv6) make this a little bit more complicated. Figure 3.2illustrates this issue in various MIPv6 cases.

As shown in Figure 3.2(a), after different combinations of MIP registrations an MN can receive data flows along different paths sending from the CN. For example, after a MIPv6 with route optimization (MIPv6w/RO) registration, data flows traverse along normal IP routing path within bothering mobility agents. However, if a MIPv6 without route optimization (MIPv6w/oRO) registration is enforced, data flows can either traverse through the home agent (HA) towards MN directly (the normal case), or traverse through the HA and the mobility anchor point (MAP) introduced in HMIPv6. Furthermore, when FMIPv6 is applied, the path can be more complicated by the way of further tunneling data packets from the Previous Access Router (PAR) through the New Access Router (NAR) and finally reach the MN. Similarly, Figure 3.2(b) demonstrates the various potential data paths along which flows sent by the MN traverse,including the case after a MIPv6 registration with or without reverse tunneling (MIPv6w/RT or MIPw/oRT), or combined with FMIPv6 and/or HMIPv6. A more detailed description of these scenarios is provided in our prior work [8]. QoS. QoS provisioning also comprises data plane (mainly traffic control e.g., classification and scheduling) and control plane (mainly admission control and QoS signaling) functions. Follow the above exploration of mobility problems, we can identify the fundamental difference of QoS provisioning in all-IP 4G mobile networks from a traditional, wired or wireless IP networks: whereas its resource control mechanisms can be similar to that of traditional networks, changing a location during the lifetime of a data flow introduces changed data path, thus requiring identifying the new path and installing new resource control parameters via path-coupled QoS signaling. Hence, a problem is how to apply any QoS signaling mechanism to achieve end-to end

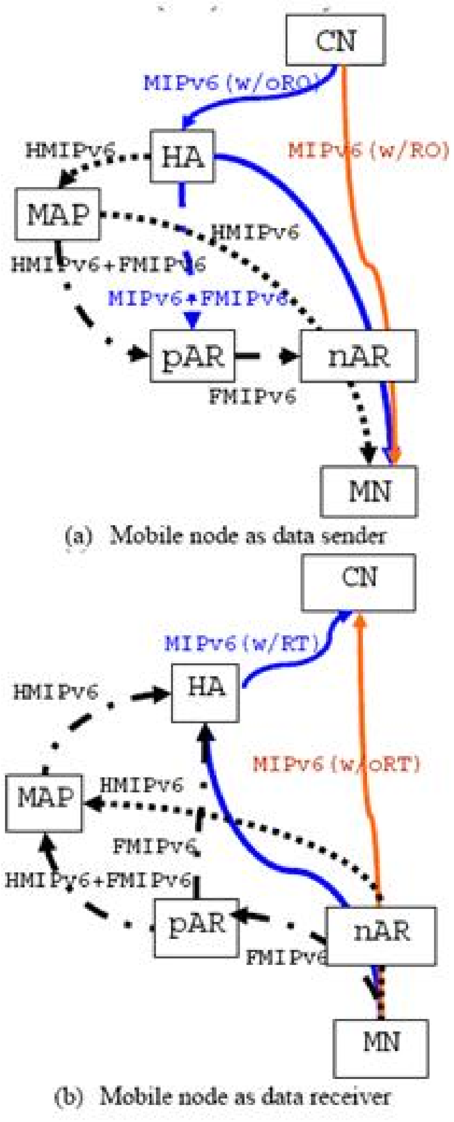


Figure 3.2: A Data plane view of an MN's flow.

resource setup in mobility scenarios. The current QoS signaling protocol, RSVP, exhibits lack of intrinsic architectural flexibility in adapting to mobility requirements. Difficulties arise, for example, because of its inability to adapt to the introduction of mobility routing in the data plane encountered in 4G networks, which results in either too complicated solutions or simply being unable to satisfy the needs. Over the years, research efforts have been made to address this (however it remains still an open issue. Security. Security in 4G networks mainly involves authentication, confidentiality, integrity, and authorization for the access of network connectivity and QoS resources for the MN's flows. Firstly, the MN needs to prove authorization and authenticate itself while roaming to a new provider's network. AAA protocols (such as Radius, COPS or Diameter) provide a framework for such support especially for control plane functions (including key establishment between the MN and AR, authenticating the MN with AAA server(s), and installing security policies in the MN or ARs' data plane such as encryption, decryption, and filtering), but they are not well suited for mobility scenarios. There needs to an efficient, scalable approach to address this. The Extensible Authentication Protocol (EAP), a recently developed IETF protocol, provides a flexible framework for extensible network access authentication and potentially could be useful. Secondly, when QoS is concerned, QoS requests needs to be integrity-protected, and moreover, before allocating QoS resources for an MN's flow, authorization needs to be performed to avoid denial of service attacks. This requires a hop-by-hop way of dynamic key establishment between QoS-aware entities to be signaled on. Finally, most security concerns in this paper lie in network layer functions: although security can also be provided by higher layers above the network layer (for example TLS provides privacy and data integrity between two communicating applications), our study mostly lies on mobility in the sense of network layer information exchange for mobile devices.

3.3 The SeaSoS Architecture

Reification of network architectures for support of QoS provisioning and security in 4G networks calls for new sights of dealing with the complexity of visualizing and architecting networks. Based on the network model described in Section 3.2, we present an architecture for Seamless Mobility with Security and QoS Support (SeaSoS), that integrates mobility schemes with QoS and security measures, and discuss the main issues toward realizing SeaSoS architecture. SeaSoS differs from priori work in two main aspects: 1) it provides a distinct abstraction on functional separation and coordination of various involved network elements, which facilitates the network architects with a systematic exploration of the network design space. 2) SeaSoS allows network operators and end users to modify network attributes using dynamic plug-ins (e.g., replacing a mobility management protocol) or by re-configuring existing network services (e.g., adjusting the configuration parameters of traffic selector in an IPsec architecture). For example, SeaSoS allows MNs to reconfigure their protocol stacks (e.g., from HMIPv6 to standard MIPv6 for mobility support, from Radius to Diameter or COPS for the AAA procedure) in order to dynamically interact with heterogeneous wireless access networks, and choose a certain QoS signaling protocol (such as RSVP or NSIS-QoS) for their end-to-end applications. In one word, SeaSoS identifies the critical infrastructure of future 4G networks, as well as other new emerging or replacement components might appear, integrated into a unified framework and allowing an efficient, scalable and extensible network design for 4G networks. Note a network element may contain zero1, one or many of data plane and control plane components in it. As an example of basic SeaSoS operation, we use MIPv6, RSVP, AAA and EAP together in achieving a seamless handover in 4G networks. As shown in the Message Sequence Chart (MSC) in Figure 3.3, we extend the method proposed in [22], namely apply EAP to perform the mutual authentication between the MN and access network, combined with AAA registration and extended with QoS and mobility support. Note the security association between the MN and the network is not directly transferred over the wireless interface, to avoid malicious nodes to obtain or modify it. As the MIP registration is also an end-to-end way operation, we extend this approach to support efficient MIP registration. In addition, once a mobility registration takes place in the HA, a QoS signaling process can start for the flow destined to the MN. Here we use RSVP Path-Resv two-way signaling but different from traditional RSVP, we use the combination of MN's permanent address (i.e., the home address) and the flow label as the unique identifier, which avoids a double reservation problem as identified in [8]. In order to prevent denial of service of QoS resources, we could apply an RSVP Integrity object to the Path/Resv messages. Before applying this authenticated RSVP signaling procedure one may create a chain of trust relationship (security associations) along the RSVP nodes through the use of ISAKMP with RSVP DOI in the key exchange protocol IKE. To further elaborate SeaSoS concepts, let's assume HMIPv6+MIPv6 is now the replacing mobility scheme. The simplest scenario can be that the MN just moves inside a MAP domain. As shown in Figure 3.4(a), one can integrate the AAA procedure together with the HMIPv6 registration procedure. For example, if we use Diameter as the AAA protocol, we could encapsulate the HMIPv6 Binding Update (HBU) and Acknowledgement (HBA) messages inside the AAA request the MN control plane, similar to the Diameter Mobile IPv4 application.

This is rather simple, and after the MAP control plane receives this message, it forwards only the AAA request part to the local AAA server (AAAL); the latter authenticates the MN and if succeeds, returns a pos-

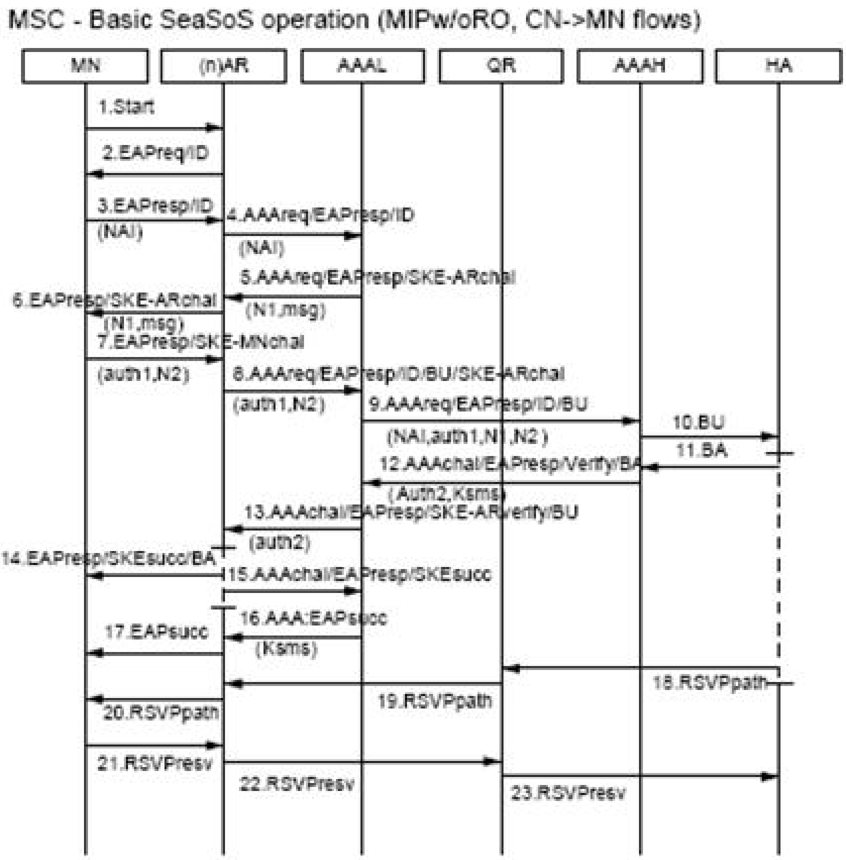


Figure 3.3: Example of SeaSoS basic operation (Control Plane).

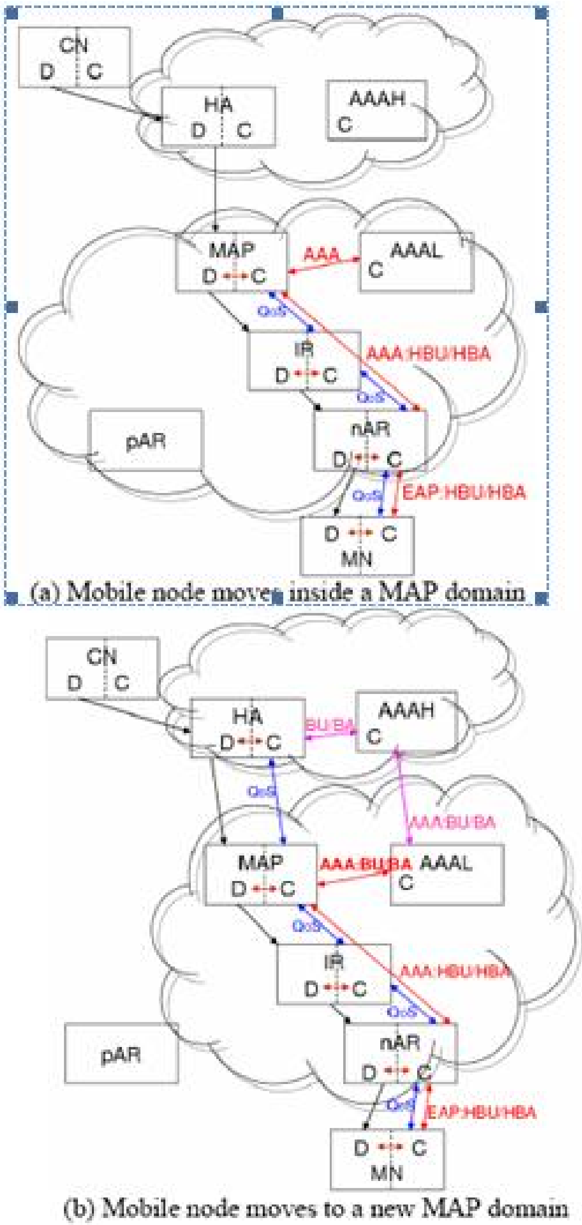
itive AAA response to the MAP. Afterwards, MAP control plane changes its binding cache (i.e., mobility routing information for CN -¿MN traffic) in the data plane. Then the MAP can start two procedures without distinction of subsequence: 1) forward the AAA request with HBA through the nAR towards the MN (while traversing nAR, the control plane in nAR installs a traffic selector in the data plane for CN -¿MN traffic), 2) initialize a QoS signaling process towards the MN. Note 1) and 2) theoretically can be further merged but this increases the complexity of implementation. An inter-domain handover is similar (shown in Figure 3.4(b)).

The difference lies in when the MAP determines it is a request to this domain, it initializes a AAA process combined with a global MIP registration (AAA:BU/BA). When the home AAA server (AAAH) accepts the request, a global binding cache is changed in HA's data plane; the HA can further initialize a QoS signaling process towards the MN. In both cases, we can see the handover process incorporates support of QoS, authentication and authorization for MN's flow in the HMIPv6 registration.

3.4 Summary and Future Work

There have been a few investigations on different aspects on QoS and security in 4G networks, notably MobyDick , SeQoMo, FCAR , and W-SKE. We compare SeaSoS with them in Table 1.

From this table we can see, different from these approaches, which mostly focus on functional aspects and optimization for certain circumstances, SeaSoS introduces the concept of supporting seamless mobility with and QoS mechanisms and security architectural components, which allows dynamic replacing/switching mobility management protocols and re-configuring existing network services in a secure way, and examines how they can be integrated universally to build an environment supporting 4G service requirements. Due to the space limitation we can only sketch the key SeaSoS concepts in this paper. Towards the realization of the SeaSoS architecture, there are a number of issues to be resolved. A key issue is how to tradeoff between efficiency and security, especially when coordinating different control plane components. We are currently developing details of the proposed concepts in the context of seamless inter-domain mobility, and will validate the design through simulations and performance evaluations in a mobile IPv6 environment.



(a) Mobile node moves inside a MAP domain

(b) Mobile node moves to a new MAP domain

Figure 3.4: SeaSoS in HMIPv6 (flows destined to MN).

Table 1: Comparison of SeaSoS with other approaches

Approach	Mobility support	QoS signaling	Security	Key ex-change
MobyDick	MIPv6+ HMIPv6	Implicit session signaling	COPS/ Diameter	*No*
SeQoMo	HMIPv6	QoS Cond.-Handoff	Diameter	*No*
FCAR	MIPv4	Changed RSVP	*No*	*No*
W-SKE	*No*	*No*	EAP+ Radius	Yes
SeaSoS	Any comb-ination of IP mobility	Changed RSVP or NSIS-QoS	EAP+ any AAA	Yes

Chapter 4

IP QoS Architecture

4.1 What is IP QoS?

The term IP means Internet Protocol and QoS refers to Quality of Service.The Internet has recently become an important communications channel. The most demanding application from the service quality point of view was a network virtual terminal as it was an interactive application. The bandwidth required was small and occasional delay variations of order of several seconds could be tolerated .
Recently many interactive or real-time services has been introduced and at the same time the economical importance of the Internet has grown. The IP phones and services based on that technology is threatening the traditional circuit-switched telephone services, specially on long-distance applications.
Transmitting interactive real-time media is the greatest challenge in packet based networks. The end-to-end delay, the delay variations (jitter), and the packet loss must not exceed some limits or usability of the service degrades badly.
Quality of service(QoS) is the ability to provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow. For example, a required bit rate, delay, jitter, packet dropping probability and/or bit error rate may be guaranteed. Quality of service guarantees are important if the network capacity is insufficient, especially for real-time streaming multimedia applications such as voice over IP, online games and IP-TV, since these often require fixed bit rate and are delay sensitive, and in networks where the capacity is a limited resource, for example in cellular data communication. In the absence of network congestion, QoS mechanisms are not required.
Quality of Service (QoS) classifies network traffic and then ensures that some of it receives special handling.It may track each individual dataflow (sender:receiver) separately.It may include attempts to provide better error rates, lower network transit time (latency), and decreased latency variation (jitter).
QoS was not one of the strong points of IP, but it now finds itself in need of it badly. IP is currently a best-efforts service. With the current boon in Internet usage especially due to its proliferation into business, there is a greater and more urgent need for ISPs (Internet Service Providers) to be able to provide QoS. This means providing different treatment to different customers and to be able to guarantee certain levels of service quality according to the users needs.

4.1.1 Factors of QoS

The Quality of Service (QoS) is often quite much abused term. If we look at traditional circuit-switched telecommunication networks, the QoS is formed by several factors, which can be divided into two groups: “human” and “technical” factors as shown in Table 1.

Table: Quality of service factors .

Human factors	Technical factors
stability of service quality	reliability
availability of subscriber lines	expandability
waiting times	effectiveness
fault clearance times	maintainability of the system
subscriber information	congestion waiting
stability of operation of the system	transmission quality
...	...

Figure 4.1:

QoS in Packet Switched Networks In the packet switched networks there are much more factors then in the circuit-switched networks that must be agreed on. The Asynchronous Transfer Mode (ATM) networks have very extensive QoS control as it is intended for real-time traffic. For the IP networks the ITU is developing a recommendation I.380 which defines quite similar metrics for IP packet transfer performance parameters.

IP packet transfer delay (IPTD) This is the delay for a IP datagram (or the delay for the last fragment) between two reference points. Typically a end-to-end delay or a delay within one network.

Mean IP packet transfer delay An arithmetic average of the IP packet transfers delays for packets we are interested about.

IP packet delay variation It is useful that streaming applications know how much the delay varies in network to avoid buffer overflows and underflows (Figure 1). For elastic applications (like TCP) small delay variations are not important but the large ones may cause either unnecessary packet retransmissions or unnecessary long delays before retransmit.

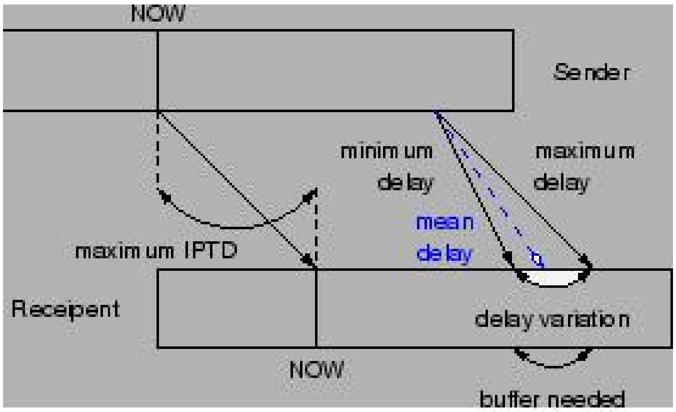


Figure 4.2:

IP packet error ratio (IPER) This is the ratio of eroded packets of all received packets.

$$IPER = \frac{N_{\text{erroneous}}}{N_{\text{successful}} + N_{\text{erroneous}}}$$

Figure 4.3:

IP packet loss ratio (IPLR) The ratio of lost packets from all packets transmitted in population of interest.

$$IPLR = \frac{N_{\text{lost}}}{N_{\text{transmitted}}}$$

Figure 4.4:

The packet loss ratio affects on quality of connection. The applicatios can react on packet loss different ways. Three typical categories are shown on Figure 2. The applications can be divided to similar categories also by required bandwidth and delay.

Fragile If the packet loss exceeds certain threshold, the value of application is lost.

Tolerant The application can tolerate packet loss, but the higher the packet loss the lower is the value of application. There are certain threshold levels which are critical.

Performance The application can tolerate even very high packet loss ratio but its performance can be very low in high packet loss ratio.

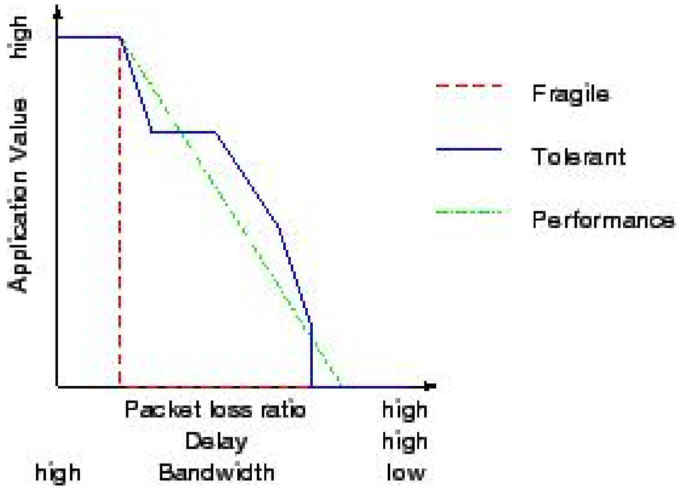


Figure 4.5:

Spurious IP packet rate As it is not expected that this is a number proportional to number of packets transmitted, this is expressed as a rate: number of spurious packets in time interval.

For ATM networks there are also metrics to characterize traffic flow for call admission control (CAC) purposes.

Peak Cell Rate (PCR) The maximum cell rate that connection may have while maintaining jitter less than defined by Cell Delay Variation Tolerance (CDVT).
Sustainable Cell Rate (SCR) The long-term maximum cell rate the connection may have.
Maximum Burst Size (MBS) The number of cells in burst which may exceed SCR but not PCR.
Minimum Cell Rate (MCR) The minimum rate the connection must be able to send at any time.

4.1.2 Grade of Service

The Grade of Service (GoS) has been used in the telecommunications industry to indicate components which contribute to overall quality of service what the user experiences. Many components have both human component and technical component: the technical component can be measured (like bandwidth of voice) and the human component is subjective. There is relation between human and technical components but the exact mapping depends on many factors, for example language used and other culture factors . In the GoS is defined as the following:

It may happen that in a network, or in part of a network, the volume of telephone traffic that arises exceeds the capacity for handling it without limitations, with the result that congestion occurs. These limitations affect the service provided to customers, and the degree of these limitations is expressed by an appropriate GoS parameter (e.g. probability of loss, average delay, failure rate due congestion, etc.). GoS should therefore be regarded as providing information on the traffic aspects of the "quality of service".

In a circuit-switched network the GoS has been divided into two standards :

Loss grade of service This standard has a component internal loss probability: for any call attempt, it is the probability that an overall connection cannot be set up between a given incoming circuit and any suitable free outgoing circuit within the switching network. For international digital telephone exchanges the internal loss probability may not exceed 0.2 % in normal load situation and 1 % in high load. The figures are higher for end-to-end connections: mean 2 and 5 % in local and international connections respectively.

Delay grade of service There are several components in this standard, depending on technology used for signaling information. For the ISDN circuit-switched services the delay components are defined for a pre-selection, a post-selection, a answer signal, and a call release.

4.2 Why QoS needed?

The Internet has worked so far with a best effort traffic model: every packet is treated (forwarded or discarded) equally. This is very simple and efficient model and several arguments has been stated against any need for a more complicated system .
Bandwidth will be infinite The optical fiber has enormous transmission

capacity, tens if not hundreds of terabyte per second in a single 0.5 mm thin (including primary coating) strain of fiber. However, installing new capacity and developing faster equipment will take some time. Also, the networks are generally designed in cost effective manner balancing between over-engineering and over-subscribing.

As wireless systems are more and more common, they limit available bandwidth because capacity on usable radio frequencies is limited. Also the energy conservation in portable equipment may limit available bandwidth.
Corollary of Moore's Law: As you increase the capacity of any system to accommodate user demand, user demand will increase to consume system capacity.
Simple priority is sufficient This is very much true: QoS is all about giving some traffic higher priority over other traffic. The problem is where to assign the priority. The user terminals cannot generally be trusted to give "fair" priorities for different traffic. If there is some billing and policing mechanism, then we do have already some kind of QoS mechanism. In some cases it is useful to give busy signal to user in order to protect the network from being over-subscribed.
There are two approaches to assigning the priority in Internet traffic: hop-by-hop based on reservation (Integrated services) and packet marking at edges of network (Differentiated services).

Applications can adapt While the applications and protocols can adapt to even extreme delays, human user can adapt much less. For example, to maintain dialogue in telephone conversation, end-to-end delays cannot exceed 300 ms to avoid man-on-moon effect .

4.3 Network Quality

4.3.1 Transmission Control Protocol (TCP)

If we want to transport data reliably on top of IP, we need a transport protocol which hides unreliability of IP from application. The most simple solution is that the sender sends one data segment; if the receiver receives it successfully, it acknowledges received packet. As sender receives acknowledgment it may send the second segment. If either segment or acknowledgment is lost or receiver is not online, the sender does not receive acknowledgment in specified time and retransmits segment.

This is not very efficient as the transmission speed is restricted by the round trip delay. A better approach is to use sliding window scheme: the receiver announces how much it is prepared to receive: the sender can send this much without getting acknowledgment.

The Transmission Control Protocol (TCP) provides reliable byte stream using sliding window flow control scheme. The original scheme, however, worked badly in case of congestion as was seen first in 1986 in "congestion collapses" . The scheme was then improved by introducing following methods:

Round-trip-time variance estimation Better estimates to find out when a segment is lost or when it is just late. In case of rising congestion event the delay will increase very much. The retransmit timer is set to value mean plus four times variation.

Exponential retransmit timer back off Limit data rate sent to network to help clear out the congestion.

Slow-start Probe for available bandwidth.
More aggressive receiver act policy Receiver acknowledges data as soon as possible to avoid retransmits.
Dynamic window sizing on congestion Adapts for changed situation on network. **Karn's clamped retransmit backoff** Limit data rate. **Fast retransmit** Fast recovery if only one segment is lost.

In addition to the window used for flow control (e.g.. the sender does not overrun the receiver) the concept of a congestion window was introduced. The congestion window tells the sender how much it can send to network and the sender selects minimum of those two windows.

The improvements made lead to more graceful operation in case of congestion. The problem is that there are difficulties to estimate (specially in case of retransmissions) round trip delay which is vital for maintaining steady flow. This has been addressed with timescale option . Current TCP congestion control algorithms and discussion can be found from.

The TCP is an elastic transport protocol: it adapts its transfer rate to available bandwidth on on network. It does not make any efforts to have minimum rate but only deliveries data in reliable manner.

4.3.2 Guaranteed Quality of Service

The guaranteed service is designed for applications which require certain minimum bandwidth and maximum delay. The service, as it provides firm (mathematically provable) bounds for end-to-end queuing delay, makes possible to provide service that guarantees both the delay and the bandwidth.

The traffic is considered as a fluid model: delivered queuing delays do not exceed the fluid delays by more than the specified error bounds. The maximum delay is

$$d_{\max} = \frac{b-M}{R}\frac{p-R}{p-r} + \frac{M+C_{\text{tot}}}{R+D_{\text{tot}}}, p > R \geq r$$

Figure 4.6:

$$d_{\max} = \frac{M+C_{\text{tot}}}{R+D_{\text{tot}}}, r \leq p \leq R$$

Figure 4.7:

where b is a token bucket depth, r is a bucket rate, p is a token bucket plus peak rate, M is a maximum datagram size, R is a bandwidth allocated to connection, is a end-to-end sum of rate-dependent error terms, and is a end-to-end sum of rate-independent, per-element error terms. When the resource reservation is being made, each node calculates its values for C and D.

As long as the traffic is conforming to traffic specification the network element must transmit the packets conforming to receiver specification .If the traffic exceeds the traffic specification, the non-conforming datagrams must be considered as best-effort datagrams. They should not be given any presence over other best-effort datagrams (to avoid misuse) nor they should be discarded as erroneous packets as a originally conforming traffic may become non-conforming in the network.

4.3.3 Controlled-load Network Element Service

The controlled-load service provides independent the network element load the client data flow with QoS closely approximating the QoS the flow would receive in unloaded network. It uses capacity (admission) control to assure this.

As in the guaranteed service the service is provided for a flow conforming the same TSpec. The applications may assume that only very few if any packets are lost and only very few if any packets greatly exceed minimum transit delay. If a non-conforming packet is received, the network element must ensure that

1. the other controlled-load flows receive expected QoS,
2. the excess traffic does not have an unfair impact on best-effort traffic,
3. the excess traffic is delivered best-effort basis if sufficient resources exists.

4.4 Integrated Services and ATM

The IP and ATM world have one basic difference: the ATM is connection oriented as the IP is connectionless. There are two basic ways to realize IP-over-ATM: using permanent virtual circuits (PVC) which emulate point-to-point links (leased lines) and switched virtual circuits which are set up on-demand. The reservation in ATM is "hard state" (active as long as it is not released) while on RSVP the reservation is "soft state" (active as long there are periodic updates).

The mapping of Integrated services to ATM service classes is quite straightforward and is presented in Table 3. The ATM traffic descriptors (PCR, SCR, MBS) are set to values based on a peak rate, a bucket depth, a RSpec, and a Receiver TSpec.

Table: Integrated Services and ATM QoS mapping

Integrated Service Class	ATM Service Class
Guaranteed Service	CBR or rt-VBR
Controlled Load	nrt-VBR or ABR (with minimum rate MCR)
Best Effort	UBR or ABR

Figure 4.8:

4.5 Bandwidth allocation in subnets

For the point-to-point links the bandwidth allocation and prioritizing is done by router. In multipoint networks there may be some shared media where each host can send as much data as they want. Some mechanism is needed to make sure that the bandwidth allocations do not exceed available bandwidth in network. This is a task for a bandwidth broker or Subnet Bandwidth Manager (SBM).

Each (RVSP) request for a bandwidth goes through SBM and if sufficient capacity in the network exists, it grants the request. It does not, however, policy requests in any way, it is just a book keeper.

4.6 Special Considerations

4.6.1 IP Security

The use of the IP Security protocols [causes some problems for both for the RSVP and differentiated service. If the Encapsulating Security Payload (ESP) is used, the upper protocol layers are encrypted and the network nodes cannot know the port numbers or protocols. The RSVP uses port numbers to make difference between different flows between two hosts (for example one flow for a data communication and one for a audio transfer).

As the transport layer is encrypted the network nodes cannot know the ports and thus cannot differentiate between flows requiring real time handling and bulk transfers. If the datagram is only authenticated the port numbers are visible but they are on a different position which may cause performance problems on routers.

This problem was solved by introducing "virtual destination port" which is actually the IP SEC Security Parameter Index (SPI). We must then make sure that the flows needing different QoS have a different SPI . If the services are classified for differentiated services networks based on port numbers, the encryption hides needed information. In this case the end system should be able to tell the network what kind of service each packet needs.

4.6.2 Tunneling

There are three main uses for tunneling: the first is to build (possibly secure) virtual private networks (VPN), the second one is to provide transport for protocols the network between does not support (as currently IPv6). The third use is tunneling the subscriber traffic from the access server to the Internet service provider (ISP). The access server can be located in local telephone exchanges so Internet connections do not reserve circuit switched capacity from the telephone network.

As the original IP packets are encapsulated to IP packets, the port numbers or DS code points are not visible to routing nodes. This problem has been solved for RSVP by using IP-in-UDP encapsulation. The UDP source ports are used to identify individual (or aggregated) flows. The RSVP reservations are tunneled and corresponding reservations are made for the tunnel also. For the differentiated services the solution is simpler: the DS codepoint is copied to IP datagram which carries the original IP datagram. This way the datagram will receive same service as the original would. The packet reordering may cause problems with some tunneling protocols so packets in same flow should have same DS code.

4.7 Network congestion

Network congestion occurs when a link or node is carrying so much data that its quality of service (QoS) deteriorates. Typical effects include queueing delay, packet loss or the blocking of new connections. A consequence of these latter two is that incremental increases in offered load lead either only to small increases in network throughput, or to an actual reduction in network throughput. Network protocols which use aggressive retransmissions to compensate for packet loss tend to keep systems in a state of network congestion even after the initial load has been reduced to a level which would not normally have induced network congestion. Thus, networks using these protocols can exhibit two stable states under the same level of load. The stable state with low throughput is known as congestive collapse.

Modern networks use congestion control and network congestion avoidance techniques to try to avoid congestion collapse.Congestion is the most important reason for requiring guaranteed QoS. When traffic volumes are low, problems are unlikely to occur even without QoS management. However, networks become unstable as increasing traffic at each terminal causes congestion. Verification of networks with guaranteed QoS and network equipments with QoS management functions requires intentional emulation of high load (congestion) conditions and confirmation that the QoSsystem is functioning correctly.

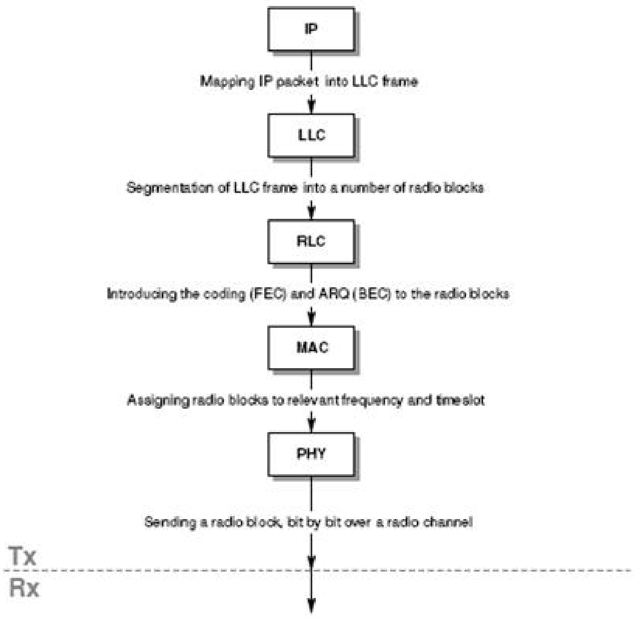


Figure 4.9:

The above setup shows an IP phone system with guaranteed QoS using priority control to assure stable throughput and low delay times. In addition to the IP phone, the network has other traffic, such as email and web browsing, from other PCs. Because the network cannot send packets exceeding the line capacity, overflow packets are discarded as congestion occurs when traffic from each PC increases. In addition, in congested network conditions, transmission delays become large because packets are waiting to be sent and transmission delay variation also increases. However, in a network with guaranteed QoS, traffic is prioritized so the IP phone packets are sent with high priority. A network with QoS management assures that IP phone packets are sent with stable throughout and low delay even under congested conditions. Traffic prioritizing systems use physical port number, MAC address, VLAN(CoS), IP address, ToS (Diffserve), port number, etc.

4.7.1 Network congestion avoidance and control

There is provided a method for congestion control and avoidance in computer networks, which method includes the steps of sensing network congestion (including both sensing and predicting possible future network congestion) and allowing a network node to transmit at least one basic data segment and thereafter to transmit additional data, the quantity of said additional data being a function of the basic data segment, wherein the size of the basic data segment is determined at least in part by the sensed network congestion. Prediction of possible future network congestion is possible, for example, by learning from a history of network load and/or by detecting an increase in the number of users or other indications. When possible future network congestion is predicted, the application of the methods and apparatus of the invention is operative to prevent the development of future congestion altogether or at least to limit the evolving severity level that such future congestion would have otherwise reached. Controlling the transmission rate of network nodes is an important technique to help prevent future congestion altogether and/or to limit the severity of such congestion.

4.8 Delay and jitter

The delay problem is a product of two phenomena. From one side streaming and especially conversational traffic is characterized by a small packet size with high delay performance requirements. When an IP packet reaches the cellular network the following things usually happen. The IP packet is mapped into a Logical Link Control (LLC) frame at the LLC layer. If the size of the packet is too large, then it is fragmented into a number of LLC frames. However this work is primarily aimed at real time traffic, which is usually characterized by small IP packets, hence it is assumed that the packet will not be fragmented and in the rest of the paper an IP packet will be equal to an LLC frame. In the next step the LLC frame is fragmented into a number of radio blocks denoted as Y. This number depends strongly on the radio channel conditions. If the Carrier to Interference ratio (C/I) represents a channel with a high Bit Error Rate (BER) then the RLC has to use strong coding redundancy in order to assure good radio block error resilience and Y reaches its highest value. On the other hand, if the C/I represents a good radio channel condition then the coding redundancy can be smaller and fewer radio blocks are required to transport the same IP packet, yielding a smaller value for Y. The Medium Access Control (MAC) layer then takes the radio blocks and assigns them to the relevant frequency and time slot. The Physical (PHY) layer transmits each radio block over the radio channel. The reverse procedure is applied at the receiver side. The whole process of transporting an IP packet over a cellular network is shown in Figure 4.10.

For the IP packet to be useful, all radio blocks must reach the destination point so that, from the user perspective the important factor for real time traffic is the delay of an IP packet, not the average delay of radio blocks. The goal of the research presented here is to develop a novel method of investigating the delay of IP packets in wireless networks. Additionally there will be a discussion about results coming from the proposed method with its possible implementation.

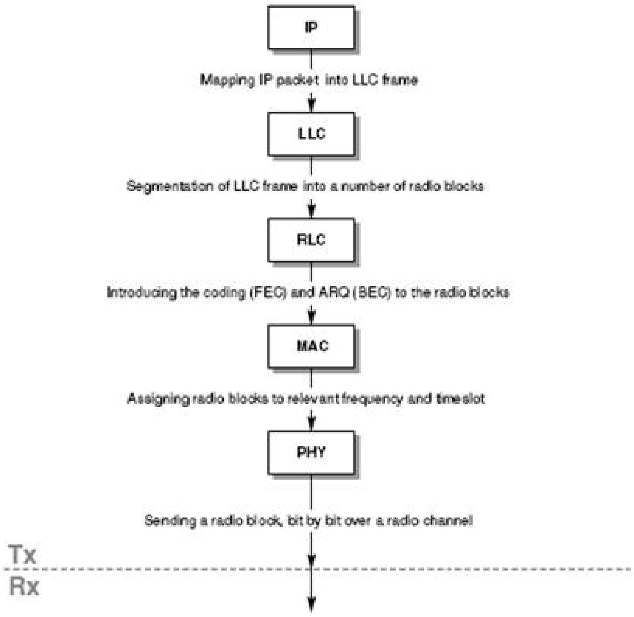


Figure 4.10: Overview of IP packet transport over a cellular network

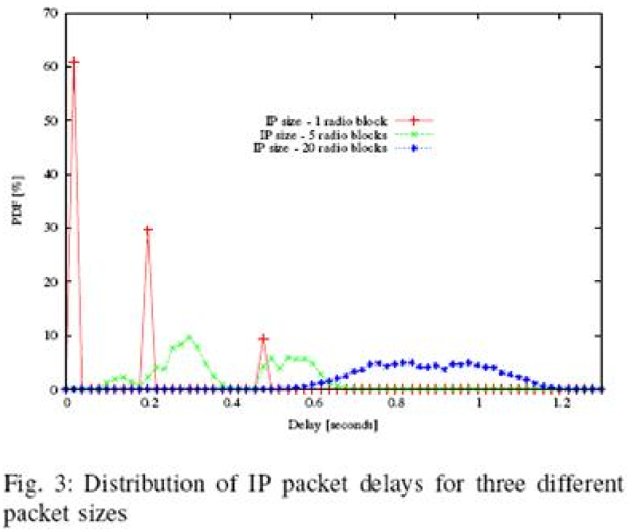


Figure 4.11:

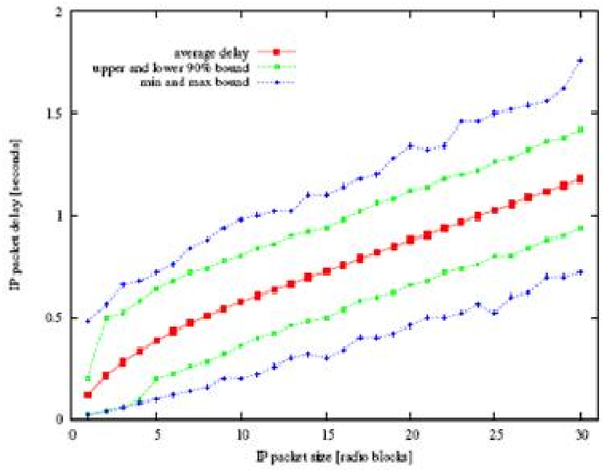


Fig. 4: Average IP packet delay as a function of packet size

Figure 4.12:

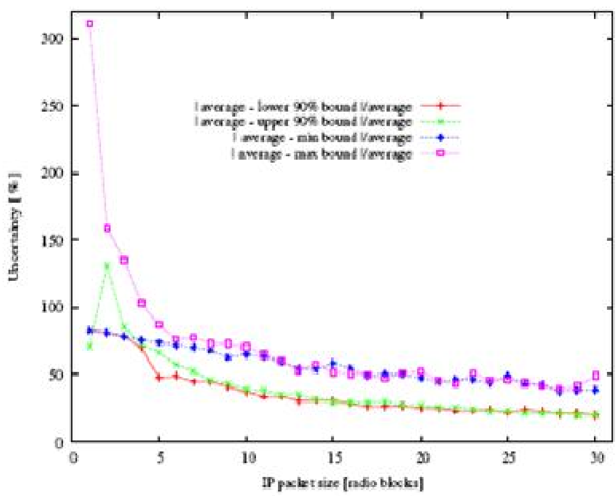


Fig. 5: Uncertainty of IP packet delay

Figure 4.13:

4.9 Conclusions

The well known best effort service is not satisfactory for the real-time applications. The Internet pricing is currently in many cases flat rate: a monthly fee or a fee based on time on-line. The user has no possibility to get a better service even if he is willing to pay more for a premium service.

Currently there are two main efforts to provide control of QoS: the Integrated Services and the Differentiated Services. The first one is based on a resource reservation: the main advantage is possibility to get well defined QoS, the main disadvantage is the need to maintain connection state in all network nodes between end systems. It also needs support from both end systems and from all networks between these two systems to be useful at all.

The Differentiated Services model is based on using per-hop behaviors which are marked as DS code points to IP datagrams. The most important advantage is that the decision at core network node is local both in space and in time which guarantees scalability. It can also deployed step by step and the interoperability is maintained. The disadvantage is that no connection admission control is done which may cause temporary overload situations on low bandwidth connections.

It looks like there are two possible ways to implement QoS: a combination where the diffserv is used in the core networks and RSVP/intserv in access network or diffserv-only network. If the applications do not soon support RSVP, the latter alternative is more probable.

Chapter 5

Open Wireless Architecture

5.1 What is open wireless architecture?

Open Wireless Architecture (OWA) defines the open interfaces in wireless networks and systems, including baseband signal processing parts, RF parts, networking parts, and OS and application parts, so that the system can support different industrial standards and integrate the various wireless networks into an open broadband platform. For comparison, Software Defined Radio (SDR) is only a radio in which the preset operating parameters including inter alia frequency range, modulation type, and/or output power limitations can be reset or altered by software. Therefore, SDR is just one of the implemental modules of the OWA system.

OWA will eventually become the global industry leading solution to integrate various wireless air-interfaces into one wireless open terminal where the same end equipment can flexibly work in the wireless access domain as well as in the mobile cellular networks. As mobile terminal (rather than wireline phone) will become the most important communicator in future, this single equipment with single number and multiple air-interfaces (powered by OWA) will definitely dominate the wireless communication industries.

5.2 Advantages of Open Wireless Architecture

The 4G Mobile communications will be based on the Open Wireless Architecture (OWA) to ensure the single terminal can seamlessly and automatically connect to the local high-speed wireless access systems when the users are in the offices, homes, airports or shopping centers where the wireless access networks (i.e. Wireless LAN, Broadband Wireless Access, Wireless Local Loop, HomeRF, Wireless ATM, etc) are available. When the users move to the mobile zone (i.e. Highway, Beach, Remote area, etc.), the same terminal can automatically switch to the wireless mobile networks (i.e. GPRS, W-CDMA, cdma2000, TD-SCDMA, etc.). This converged wireless communications can provide the following advantages:

- Greatly increase the spectrum efficiency
- Mostly ensure the highest data-rate to the wireless terminal
- Best share the network resources and channel utilization
- Provide the seamless networking among different wireless standards
- Optimally manage the service quality and multimedia applications

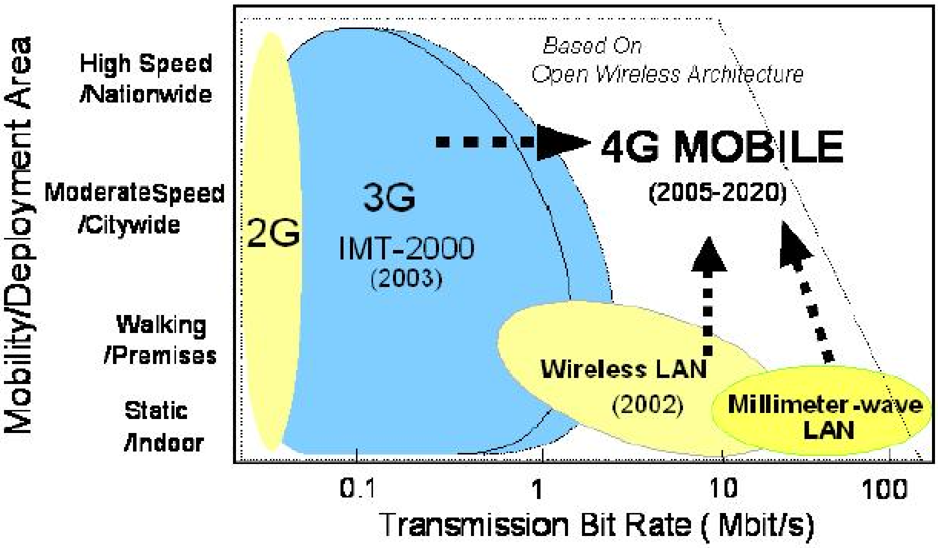


Figure 5.1: Wireless Evolution to 4G Mobile Based on Open Wireless Architecture (OWA).

5.3 Adaptive Modulation and Coding (AMC)

The principle of AMC is to change the modulation and coding format (transport format) in accordance with instantaneous variations in the channel conditions, subject to system restrictions. AMC extends the systems ability to adapt to good channel conditions. Channel conditions should be estimated based on feedback from the receiver. For a system with AMC, users close to the cell site are typically assigned higher order modulation with higher code rates (e.g. 64 QAM with R=3/4 Turbo Codes).

On the other hand, users close to the cell boundary, are assigned lower order modulation with lower code rates (e.g. QPSK with R=1/2 Turbo Codes). AMC allows different data rates to be assigned to different users depending on their channel conditions. Since the channel conditions vary over time, the receiver collects a set of channel statistics which are used both by the transmitter and receiver to optimize system parameters such as modulation and coding, signal bandwidth, signal power,training period, channel estimation filters, automatic gain control, etc.

Figure 5.1 shows the wireless evolution to 4G mobile communications based on OWA platform, where 3G, Wireless LAN and other wireless access technologies will be converged into 4G mobile platform to deliver the best infrastructure of mobile communications with optimal spectrum efficiency and resource management. In fact, this OWA model had already been accepted by many wireless industries, for example, the W-CDMA/W-LAN/WIMAX 3-in-1 terminal is being designed in many companies.

5.4 Adaptive Hybrid ARQ

A successful broadband wireless system must have an efficient co-designed medium access control (MAC) layer for reliable link performance over the lossy wireless channel. The corresponding MAC is designed so that the TCP/IP layers see a high quality link that it expects. This is achieved by an automatic retransmission and fragmentation mechanism (ARQ), wherein the transmitter breaks up packets received from higher layers into smaller sub-packets, which are transmitted sequentially. If a sub-packet is received incorrectly, the transmitter is requested to retransmit it. ARQ can be seen as a mechanism for introducing time-diversity into the system due to its capability to recover from noise, interference, and fades. Adaptive hybrid ARQ shows significant gains over link adaptation alone through e.g. Chase combining. Hybrid ARQ self-optimizes and adjusts automatically to channel conditions without requiring frequent or highly accurate C/I measurements:

1. adds redundancy only when needed;
2. receiver saves failed transmission attempts to help future decoding;
3. every transmission helps to increase the packet success probability.

5.5 Open Broadband Wireless Core

The open wireless platform requires:

- Area and power-efficient broadband signal processing for wideband wireless applications
- Highest industry channel density (MOPS pooling) in flexible new BTS (Base Transceiver System) signal processing architectures
- BTS solutions scalable to higher clock rates and higher network capacity
- Waveform-specific processors provides new architecture for platform reuse in terminals for multiservice capability
- Terminal solutions achieve highest computational efficiency for application with high flexibility
- Powerful layered software architecture using virtual machine programming concept

For example, the key features of open BTS modem include (but not limited to):

- Multi-standard air-interfaces
  - GSM, cdma2000, WCDMA, HDR, TD-SCDMA, WLAN, OFDM, WIMAX
  - proprietary standards
- Highest channel-density
  - 3GPP channels, 3GPP2 channels
  - OFDM channels, WiMax channels
  - ability to support multiple sectors on one chip
  - grow from sectors-on-a-chip to BTS-on-a-chip or System-in-Package
- Scalable data-rates
  - support from 8 kbps to 384 kbps to 2 Mbps to 10 Mbps or higher
- Configurable to mix voice and data
  - programmable allocation of channels
- IP-ready
  - interfaces directly via BTS IP back-haul
- Over-the-network programmable
  - remotely configurable from network operations center

The key features of open wireless terminal include:

- Multi-standard Air Interface
  - GSM, cdma2000, WCDMA, W-LAN, Bluetooth, OFDM, WIMAX, TDSCDMA
- Power Efficient
  - 100 MOPS/mW and more
- Scalable Architecture
  - Breaks the 384 kbps, 2Mbps and 10Mbps plateau
- High-level Modem VMI
  - Simplifies programming for each standard

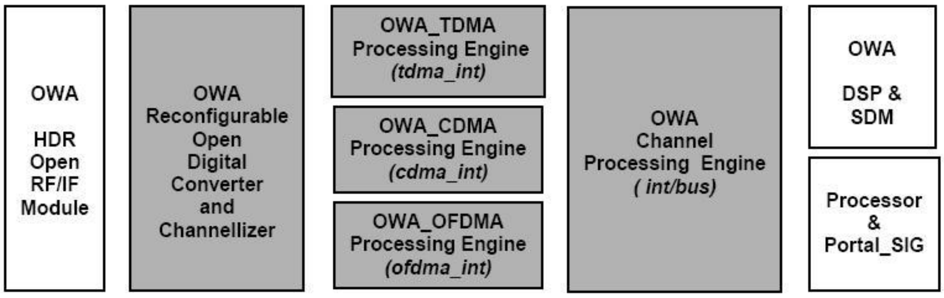


Figure 5.2: Multi-standards BTS Engine for the 4G OWA Platform.

  - Enhances reuse across standards

- Integrates across many platforms
  - No DSP and minimal microprocessor dependent code

SIP Cores (Silicon Intellectual Property)

- Initial engine optimized for B3G/4G applications, for example, TD-SCDMA/WiMax/WLAN 3-in-1 core platform for open service-oriented architecture.

Figure 5.2 shows the Multi-standard BTS Engine for this OWA Platform, where “HDR” means “Hardware Defined Radio”; “SDM” is “Software Defined Module”.

5.6 Open Base Base-band Processing Platform

Same as open computer architecture in the computer system, the OWA shares all the open system resources including hardware and software by mapping different wireless standards to the open interface parameters of baseband, RF and networks. Each OWA system module is an open module, rather than any standard-specific module, and can be easily reconfigured to maximize the system performance, and minimize the power consumption.

To migrate current wireless and mobile systems to such an advanced open wireless system with the features mentioned above, we have to face a number of technical challenges in the open baseband processing of such OWA system:

- Terminal design is much hard than base station design due to its limitation of power consumption, chip area, and processing capability.

- Open architecture requires fully extensible and upgradeable in baseband processing which traditionally can be handled by general-purpose processors and digital signal processors. However, these processors consume more power with less efficiency in system performance.

- Application-specific integrated circuits (ASIC) is a very efficient processor and consumes low power compared to genera-purpose processors and DSPs, but without flexibility in supporting different wireless standards, because ASIC is normally a standard-specific implementation solution.

- Open Wireless Architecture (OWA) demands efficient baseband management system to optimize the open processing modules and system performance.

OWA provides an optimal open baseband processing platform supporting different existing and future defined wireless radio transmission technologies (or air interfaces) including, but not limited to, W-CDMA, TD-SCDMA, GSM, GPRS, OFDMA, WLAN, WPAN and BWA (broadband wireless access system), either in the simultaneous connection mode, or in the selective connection mode of various wireless standards in the user's service geographic region, where different radio standards are mapped into the open interface parameters as inputs to the open processing modules scheduled and administrated by the baseband management system for the optimization of the system performance and resource of the wireless mobile terminal. Figure 5.3 shows an OWA based Baseband Processing System-on-Chip platform for 4G mobile phone where OWA accelerators include baseband open computing machines (OCM) accelerator and open processing kernel accelerator. OWA BIOS is a wireless basic input/output system defined by open interface parameters (OIP) of OWA technology. This OWA Baseband SoC is designed for GSM/ cdma2000/ TD-SCDMA/ WiMax/ WLAN 5-in-1 compact mobile phone terminal.

5.7 Open Network Access Platform

In recent years, access aggregation technologies have been developed that allow a common access and transport network to bear the traffic of subscribers from multiple service providers. Separating access and transport from service accomplishes two points:

- It eliminates the burden of building out an access network, reducing the barrier to entry for new service providers and improving the growth potential for existing service providers.

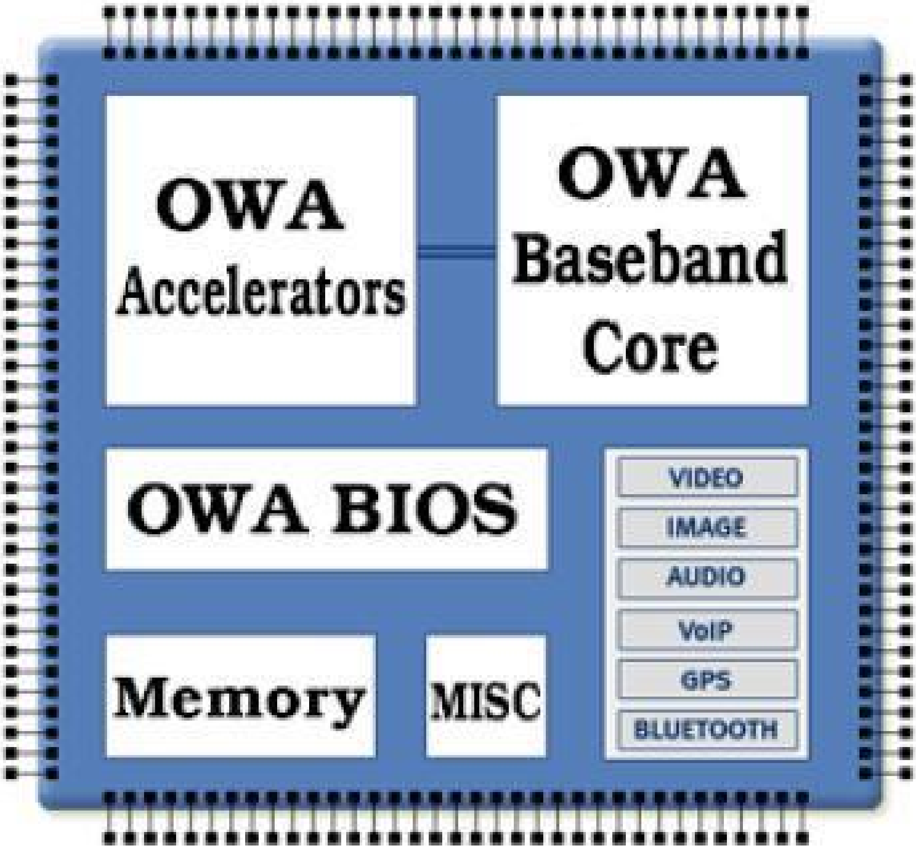


Figure 5.3: OWA Baseband Processing SoC Platform.

- It promotes technical and business efficiencies for access and transport enterprises due to economies of scale and the ability to resell that access infrastructure to multiple service providers.

New systems provide end-to-end direct IP connections for users by extending access aggregation architectures to mobile broadband access. Network and service providers can leverage existing equipment, tool and content bases to support mobile broadband end users, while the end users experience the best of the wireless and wired worlds 3/4 the broadest range of applications and end-user devices, coupled with the freedom to move and high data rates.

5.8 Open Service Service-Oriented Architecture and Open OS Platform

The success of future wireless communications rely mostly on the services provided and the applications the users require, rather than the underlying wireless transmission technologies. The users will dislike different boring names of various wireless standards, such as 802.11, 802.16 or cdma2000, etc, and therefore the service-oriented architecture (SOA) is extremely important for the system design and product development of future wireless communications.

To support this SOA platform, OWA is required to converge various radio transmission technologies onto an open system platform, including baseband processing platform, operating system platform, RF platform and infrastructure platform, to facilitate the future wireless terminal and base-station to handle different communications needs with same open equipment and same number - a truly unique and global personal communication identifier.

Figure 5.4 shows the future open services environment. The OWA layer sits under the SOA layer and the Service Oriented Infrastructure layer, and hence is transparent to the end users. This open environment is designed for the future wireless lifestyle of year 2010 and beyond.

5.9 Open Distributed Wireless Ad Ad-Hoc Networks

Experts doubt that the current path to 3G will succeed. Current 3G migratory paths involve slowly enhancing voice-centric, high-power, hierarchical networks with IPoverlays. Air-interfaces may upgrade from GSM to WCDMA, or CDMA to HSDPA, but such RF adaptations do not address

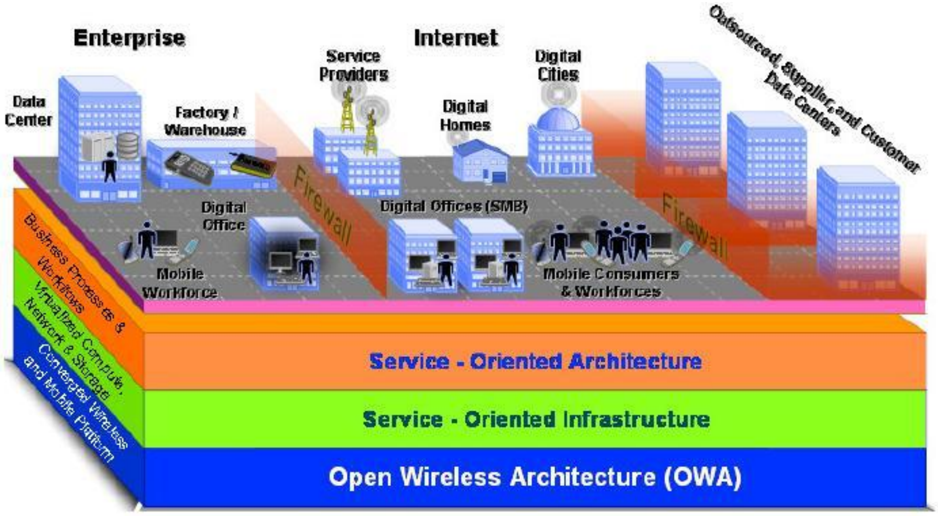


Figure 5.4: The Future Open Services Environment.

the underlying wireless network architecture issue. A concern is whether or not upgrading star networks will work any better in a wireless environment than it did in wireline.

However, low-powered, ad hoc mesh architected networks offer spectrally efficient high performance solutions to this dilemma. In such peer-to-peer networks, end-user wireless handsets act as both end terminals and secure wireless routers that are part of the overall network infrastructure. Upstream and downstream transmission "hop" through subscriber handsets and fixed wireless routers to reach network access points or other end terminals. Routing infrastructure, including handsets, utilize intelligent routing capabilities to determine "best path" for each transmission.

Routing for "best path" must be defined for "least power". That is, network nodes must be able to calculate and update routing tables to send data packets through the paths with minimal power requirements. This is different than network nodes associating with the physically closest available infrastructure. Therefore, subscriber terminals do not "shout" at a centralized base station, but rather whisper to a near-by terminal that routes the transmission to its destination. Therefore subscriber terminals cooperate, instead of compete for spectrum. Spectrum reuse increases dramatically, while overall battery consumption and RF output within a community of subscribers is reduced. Simply put, additional users enhance rather than strain network capacity.

Thus, while the cellular handset can only maintain a 144kbs (for example) link to the base station, the ad hoc mesh device can maintain a multi-megabit link without undue interference.

5.10 Open Standards Will Prevail in Next Decade

With the strong economy growth in East Asia including Korea, China and Japan, and the neighboring countries, the 4G mobile system based on Open Wireless Architecture (OWA) will become the next wave in wireless communications. It is well predicted that Asia-Pacific (AP) will be the major global hub of this 4G mobile in the coming years, and over 70% of world's 4G R&D are based in this region which reflects huge business opportunities and industrial potentials in future wireless communications. Open standards will definitely drive this new storm in the region's information and communication technology industry, especially in China.

5.11 Challenges and Solutions

Some technical challenges are being studied in the USCWC for this emerging 4G-OWA systems, including definitions of open interface parameters (OIP) developed in the OWA core; sharing studies in the common frequent bands between IMT-2000 (3G) and fixed broadband wireless access (BWA like WiMax for example) systems including nomadic applications in the same geographical area; spectrum sharing for IMT-Advanced (the 4G program by ITU) and the principles to prioritize some candidate bands for this OWA converged systems; dynamic bandwidth allocation, radio resource management and adaptive network optimization, etc.

Since an internationally unified standard becomes unfeasible and impossible, there will be many different standards and frequency bands co-existing in the ITU IMT-Advanced era. The OWA platform provides an optimal solution to converge these different radio transmission technologies into a common and shared wireless communications infrastructure, supporting the future service-oriented open architecture. As a practical solution for the initial development of the 4G-OWA system, we focus on the exampled GSM/ TD-SCDMA/ cdma2000/ WiMax/ WLAN 5-in-1 open platform targeting for shortterm strategy towards the year 2010 in some emerging markets including North Americaand East Asia regions. Figure 5.5 shows an implemental OWA radio transceiver architecture supporting multi-bands of 800/900, 1800/1900, 2.4G/2.5G, 3.5G and 5G. This 5-in-1 OWA core platform is being developed by USCWC for the emerging China mobile communication markets.

Figure 5.5: Future All-in-One 4G Terminal by OWA.

Chapter 6

Generic 4G Network Architecture

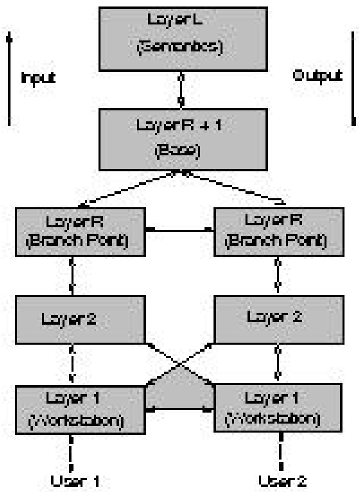


Figure 6.1: Generic System Architecture.

6.1 Introduction to Generic Architecture

The architecture assumes that a user's input/output is processed by a hierarchy of layers. A lower-level layer (that is, a layer closer to the user) manages objects that are interactors of objects in the immediately higher-level layer. We will refer to the latter as abstractions of the former. An interactor of an abstraction creates a presentation of the abstraction, which contains a transformation of the information in the abstraction (e.g. a text field representing an integer, or a bitmap representing a text field) plus some additional information serving as "syntactic sugar" (e.g. a label field or a window scrollbar). Thus, perceptible renderings of abstractions are created by applying the presentation operator successively to their interactors, and the interactors of these interactors, and so on. An abstraction can have a variable number of interactors, which may change dynamically as users create or delete renderings of the abstraction.

The layers communicate with each other using events. Often, this term implies that the communication is sent asynchronously by the sender to the receiver. However, we will use it here in a more general sense and allow the information to be retrieved synchronously from the sender by the receiver. We divide events of a collaboration application into interaction events and collaboration events based on whether they support single-user or collaboration semantics. An interaction event may be an output event or an input event depending on whether it is sent to a lower or upper-level layer.

Abstractions send output events to their interactors and receive input events from the latter. Output events received by objects from their abstractions may be transformed into lower-level events before they are sent to their interactors. Conversely, input events received by objects from their interactors may be transformed into higher-level events before they are sent to their abstractions. Not all input events received by interactors need to be sent to their abstractions - in particular, events that request manipulation of local syntactic sugar. Moreover, not all output events transmitted down by interactors are triggered by output events received from their abstractions. These include not only those events that change local syntactic sugar but also those that generate local echo/feedback in response to requests for changing the higher-level state in the abstraction.

A collaboration event may be a copy or extension of an interaction event or it may be an entirely new kind of event. It may be sent not only to a lower-level and upper-level layer but also a cross layer, that is a layer in an another branch, as shown in the figure.

Some levels in this architecture are shared while others are versioned or replicated. A shared level is associated with a single, shared layer that processes the input/output of multiple users of the application, while a versioned or replicated level is associated with a private layer for each user of the application, which processes the input/output of only that user and collaboration events concerning the user. An object in a private layer is private while an object in a shared layer is shared by multiple users. We refer to the collection of all private objects of a user and the shared objects accessible to the user as the interaction state of that user. All levels below a private level are constrained to be private levels and all levels above a shared level are constrained to be shared levels. Thus, the architecture defines a tree of layers rather than a general graph. We refer to this tree as a protocol tree in analogy with the related networking concept of a protocol stack. We refer to the lowest shared layer as the base, the highest versioned layers as branch points, the base and all layers above it as the stem, and a branch point and all the layers below it as a branch of the architecture. Moreover, we refer to all private layers at a certain level as peers or replicas of each other.

An abstraction may have interactors in zero or more replicated layers. We refer to the different interactors of an abstraction as replicas, peers, or versions. In general, they can create different logical presentations of the abstraction. However, in most current collaboration architectures, they create different physical replicas (for different users) of the same logical presentation. It is for this reason, we have used the term "replica" for a peer interactor and layer, though strictly speaking, the term "version" is more general. In the rest of the discussion, we will use these terms interchangeably.

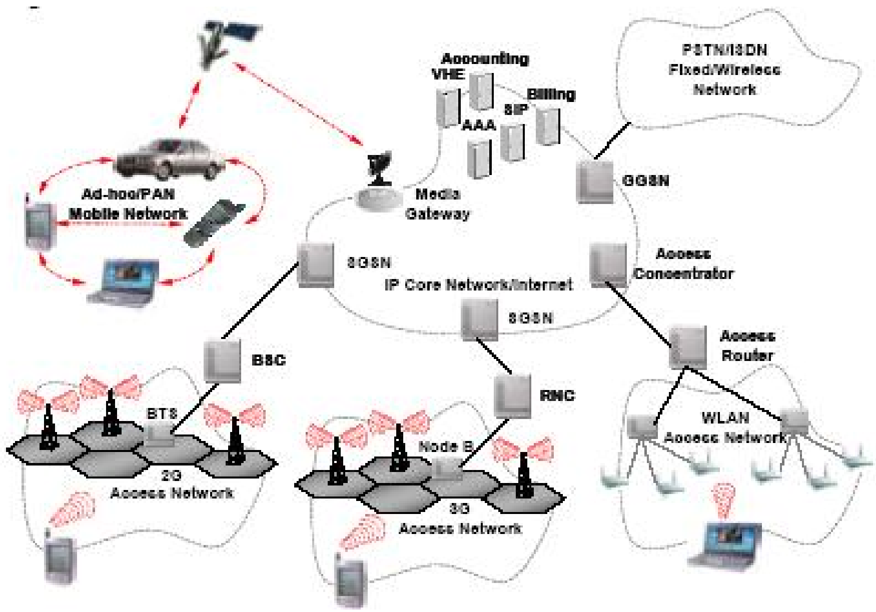


Figure 6.2: Generic 4G Network Architecture.

It is important to note that an interactor in a layer may not have a peer interactor in a peer layer, since not every layer creates an interactor for an abstraction in the layer above. Abstractions and interactors may not only transform interaction events but also control the interaction by checking access rights, consistency, and other constraints.

6.2 Generic 4G Network Architecture

Fourth generation mobile communication system tend to mean different things to different people, for some it is merely a high capacity new radio interface while for others it is internetworking of cellular and wireless LAN technologies that employs a variant of the mobile IPv6 mobility management protocol for intersystem handoff and IETF AAA technologies for seamless roaming. There is no doubt that 4G systems will provide higher data rates. Traffic demand estimates suggest that, to accommodate the foreseen amount of traffic in the 2010-2020 timeframe in an economically viable way, 4G mobile systems must achieve a manifold capacity increase compared to their predecessors. This widely accepted version now sketches a heterogeneous network infrastructure comprising different wireless access systems in a complementary manner where the user supported by his or her personal intelligent agents, enjoys unlettered connectivity and ubiquitous access to applications over the most efficient combinations of available systems. Figure 6.2 illustrates the future 4G mobile network architecture comprising ad-hoc, cellular, hot-spot and satellite radio components.

6.3 Technology Supported by 4G Generic Architecture

6.3.1 Adaptable capability-aware service provision

To provide the mobile user with a consistent list of available applications that are supported by the mobile device he/she is currently employing, it is necessary to discover - and exploit - device capability information. Furthermore, given that wireless access networks differ significantly in terms of coverage area and supported bandwidth, mobile network capabilities (e.g. traffic load) should also be considered along with other important flavors of context information (e.g. user location and mobility patterns) so as to further refine the list of applicable services. In turn, that requires flexible representation formats (e.g., XML) and announcement procedures (e.g., CC/PP) like the ones adopted by the 3GPP MExE specification as well as the employment of sophisticated user profiling mechanisms.

6.3.2 Transparent mobility and universal roaming capability

The variety of wireless access technologies that will coexist in the 4G mobile environment complicates the technical and regulatory aspects of roaming. Seamless user mobility across different wireless access technologies (e.g. WLAN, UMTS) with minimal or zero user intervention must be supported by efficient inter-system mobility management and handover procedures. To reduce signaling load, micro-mobility should be handled by the specific mobility management mechanisms of each wireless access network, while macro-mobility and roaming should built on cross-industry standard protocols and architectures, such as hierarchical Mobile IPv6 and AAA. Considering that handoff to a different system may entail different charges, it may be desirable to include QoS and pricing information as part of mobility management signaling.

6.3.3 Automated protocol configuration mechanisms

In the 4G era, the plethora of available - but disparate - applications combined with the existence of multiple wireless access systems and the need for a user profile-driven decision process, suggests that there may be multiple options capable of accommodating the same set of services. In example, a media stream can be transported to a mobile terminal by means of either a wireless LAN or a UMTS bearer service. However, the decision regarding which particular system to use depends on a number of factors, such as the respective cost of service, availability of network resources, radio link quality and user preferences. Continuing our previous example, we can imagine scenarios where end-to-end application signaling is routed via UMTS, because of its predictable performance and explicit QoS guarantees, while media streams are routed via a nearby WLAN to take advantage of its greater bandwidth capacity and lower cost. Since using different systems will result in accruing different charges, in 4G mobile environments service provision decisions must remain informed of users' pricing preferences.

6.3.4 Policy-based management and information models

Our previous point suggests that the system and protocol configuration procedures should be dynamic and automated to the highest degree possible but also open and flexible enough to facilitate higher-layer control over network bearer services. We find typical network management solutions as too narrow in scope, focusing on management of individual network elements within - rather than across - administrative domains. On the other hand, policy-based management demarcates between enforcer entities and decision entities in the infrastructure, thereby allowing the realization of a flexible management architecture that spans across multiple administrative domains. Furthermore, policy protocols support both outsourcing and provisioning modes of operation, making policy-based management an ideal approach for 4G mobile environments.

6.3.5 Interoperable QoS management across different systems

Currently, the prevalent QoS models for the IP protocol are integrated Services (IntServ) and Differentiated Services (DiffServ). Despite the differences in the scope of their design assumptions, control model and trade-off between accuracy and scalability, IntServ and DiffServ share a common subset of functionality, e.g., the traffic classifications elements. It is possible to treat the common functional components of these QoS architectures as instances of a generic information model for network elements that provide QoS-aware treatment to IP packets. That, in turn, would constitussste part of a generic information model for an entire network infrastructure that provides an aggregate IP forwarding service. Combined with policy-based management, information models can provide a consistent view of network - and mobile terminal - functionality and facilitate its configuration and adaptation regardless of the particular technologies it is built upon. With a flexible open API (e.g. IDL) that exposes the functional features of the network infrastructure in a technologically opaque fashion and allows a (trusted) third party or application to control and coordinate the underlying network mechanisms, the realization of QoS management schemes distributed across multiple administrative domains becomes greatly simplified, especially in the case of heterogeneous infrastructures like 4G.

6.3.6 Flexible pricing and billing mechanisms

The clear demarcation between the network and the service domains that has been architected in existing 3rd generation systems suggests that any future-proof charging and billing architecture should portrait similar - if not greater - flexibility in its design. As a minimum requirement, network-related pricing models must be completely independent from service-related ones, with regard to formulation as well as application matters. In combination with policy-based management, information models support the interoperable specification of pricing models for specific domains (e.g., QoS-based model for the network domain) and the configuration of affected components in the mobile network infrastructure, respectively.

6.3.7 Application and mobile execution environment aspects

To efficiently achieve mass scale deployment over millions of mobile terminals from different manufacturers and with disparate characteristics, application development must adopt the "write once, run anywhere" paradigm. Virtual machine that abstract differences in the operating system and the hardware platform promote a hassle-free application development, while interpreted languages lend themselves nicely to the restricted nature of mobile devices that may lack the resources required for a full compilation of a downloaded application. In addition, independent service providers will be relieved from the burden of developing, supporting and maintaining multiple versions of their applications for each possible client. The execution environment at the mobile terminal should shield applications from mobility-induced events (e.g. change of IP address) in the underlying protocol stacks while enabling the realization of network-aware application behavior. Any API exposed to applications should refrain from using network related information fields (e.g. IP address, port numbers) in its class and method definitions. Information that may change due to network events (e.g., handover) should be handled internally via other libraries that are wrapped by the API that is visible to applications. Ideally, applications will use the transport API to instantiate so-called "flow" objects, allowing them to exchange information with their counterparts. These "flow" objects should be opaque with regard to the protocol details of the underlying connectivity service, thus shielding applications from undesirable network events (e.g. loss of transport socket connection). Such events should be handled by the execution environment that will send the appropriate notification to the application, allowing for the graceful termination of its active communication session, if necessary.

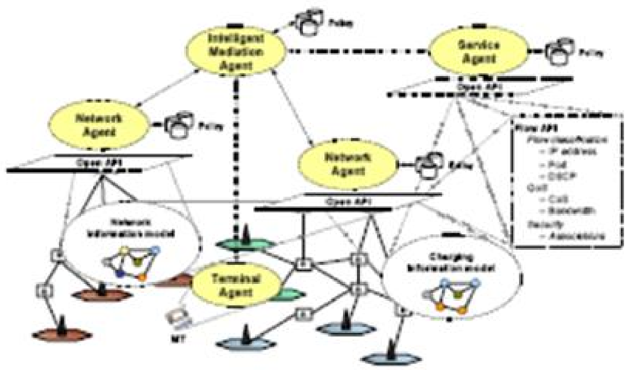


Figure 6.3: Intelligent mediation architecture for 4G.

In addition, the mobile terminal must provide a local QoS Manager to serve application requests for QoS treatment of their traffic flows. QoS signaling must proceed in a two-stage admission control; one at the mobile terminal, as it is a device limited in resources, and one at the intelligent mediation agent that will gather admission control decisions from the network agents of the wireless networks employed to carry the traffic flows of that particular application. Undoubtedly, there will be cases where the mobile terminal rather than the wireless interface is in scarcity of resources (e.g. drained power supply), so this approach can potentially minimize unnecessary signaling over the radio interface and preserve valuable bandwidth. Figure 6.3 below provides an illustration.

6.4 Conclusion

Advances in mobile communication technologies have been rapid and their effects have frequently manifested themselves in ways and places far beyond the ones imagined by their inventors.

Policy-based management and information model concepts, hierarchical Mobile IPv6 and AAA, flexible pricing and billing schemes, capability negotiation processes, and last but not least, open, technology-independent APIs are all important building blocks of 4G mobile systems. Properly combined, the aforementioned technologies can support a 4G system architecture that will be a far cry from its monolithic predecessors.

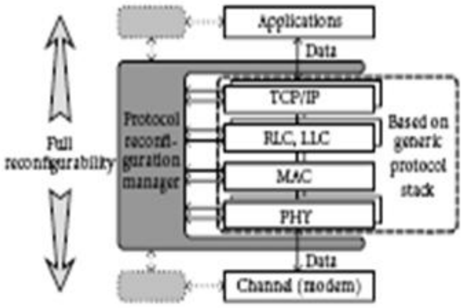


Figure 6.4: Generic Functional Module.

Chapter 7

Distributed Switching Architecture

7.1 What is distributed switching architecture

Distributed Switch Architecture is a protocol for managing hardware switch chips. It consists of a set of MII management registers and commands to configure the switch, and an ethernet header format to signal which of the ports of the switch a packet was received from or is intended to be sent to.

The switches that this driver supports are typically embedded in access points and routers, and a typical setup with a DSA switch looks something like this:

```
+-----------+       +-----------+
|           | RGMII |           |
|           +-------+           +------ 1000baseT MDI ("WAN")
|           |       |  6-port   +------ 1000baseT MDI ("LAN1")
|    CPU    |       | ethernet  +------ 1000baseT MDI ("LAN2")
|           |MIImgmt|  switch   +------ 1000baseT MDI ("LAN3")
|           +-------+ w/5 PHYs  +------ 1000baseT MDI ("LAN4")
|           |       |           |
+-----------+       +-----------+
```

The switch driver presents each port on the switch as a separate network interface to Linux, polls the switch to maintain software link state of those ports, forwards MII management interface accesses to those network interfaces (e.g. as done by ethtool) to the switch, and exposes the switch's hardware statistics counters via the appropriate Linux kernel interfaces.

This initial patch supports the MII management interface register layout of the Marvell 88E6123, 88E6161 and 88E6165 switch chips, and supports the "Ethertype DSA" packet tagging format.

There is no officially registered ethertype for the Ethertype DSA packet format, so we just grab a random one. The ethertype to use is programmed into the switch, and the switch driver uses the value of ETH_P_EDSA for this, so this define can be changed at any time in the future if the one we chose is allocated to another protocol or if Ethertype DSA gets its own officially registered ethertype, and everything will continue to work.

7.2 Use in telephony networks

Distributed switching is often used in telephony networks, though it is often called **host-remote switching**.

In rural areas, population centers tend to be too small for economical deployment of a full-featured dedicated switch, and distances between these centers make transmission costs relatively high. Normal telephone traffic patterns show that most calling is done between people in a community of interest, in this case a geographical one: the population center. Use of distributed switching allows for the majority of calls that are local to that population center to be switched there without needing to be transported to and from the host switch.

The host switch provides connectivity between the remote switches and to the larger network, and the host may also directly handle some rare and complex call types (conference calling, for example) that the remote itself is not equipped to handle. Host switches also perform OAM&P (Operation, Administration, Maintenance, and Provisioning) functions, including billing, for the entire cluster of the host and its remote switches.

A key capability of a remote switch is the ability to act in emergency standalone (ESA) mode, wherein local calls can still be placed even in the event that the connection between that remote and the host has been lost. In this mode, only local calling is available anyway, so the billing capability of the host switch is not required. ESA is increasingly available on digital loop carrier platforms as well as on purpose-built remote switches in order to improve the scope of their utility.

7.2.1 Conference call

A **conference call** is a telephone call in which the calling party wishes to have more than one called party listen in to the audio portion of the call. The conference calls may be designed to allow the called party to participate during the call, or the call may be set up so that the called party merely listens into the call and cannot speak. It is often referred to as an **ATC** (Audio Tele-Conference).

Conference calls can be designed so that the calling party calls the other participants and adds them to the call. In most cases, the participants are able call into the conference call themselves, by dialing into a special telephone number that connects to a "conference bridge" (a specialized type of equipment that links telephone lines).

Usually, most companies use a specialized service provider who maintains the conference bridge, or who provides the phone numbers and PIN codes that participants dial to access the meeting or conference call.

Three-way calling is available (usually at an extra charge) for most customers on their home or office phone line. To three way call, the first person one wishes to talk to is dialed. Then the Hook flash button is pressed and the other person's phone number is dialed. While it is ringing, flash is pressed again. This will put the three people together. This option allows callers to add a second outgoing call to an already connected call.

7.2.2 OAMP

OAMP, traditionally **OAM&P**, stands for **Operations, Administration, Maintenance, and Provisioning**. The term is used to describe the collection of disciplines generally, as well as whatever specific software package(s) a given company uses to track these things.

Though the term, and the concept, originated in the wired telephony world, the discipline (if not the term) has expanded to other spheres in which the same sorts of work are done, including cable television and many aspects of Internet service and network operation.

The Sectors Operations encompass automatic monitoring of environment, detecting and determining faults and alerting admins. Administration typically involves collecting performance stats, accounting data for the purpose of billing, capacity planning using Usage data and maintaining system reliability. It can also involve maintaining the service databases which are used to determine periodic billing. Maintenance involves upgrades, fixes, new feature enablement, backup and restore and monitoring the media health. The major task is Diagnostics and troubleshooting. Provisioning is the setting up of the User accounts, devices and services.

7.3 Distributed switching architecture with respect to 4G

4G is being developed as a successor to 3G networks, not only to overcome the limitations of 3G, but also to make use of the latest developments in the wireless technology domain. Beyond 3G cellular systems (4G) are being developed with two main objectives:

- The first objective is to overcome the shortcomings and limitation of 3G, the chief amongst which is the issue of available bandwidth. 3G networks have a maximum bandwidth of 2Mbps while in realistic scenarios, actual bandwidth will be something around 384Kbps. 2Mbps bandwidth is only possible in restricted cases involving low mobility and Pico cells. Even though 384Kbps is much better than the bandwidth of 2G systems, it is insufficient for multimedia communication. 4G networks are envisioned to offer higher bandwidths of 100Mbps and higher.

Besides this limitation, there are other shortcomings of 3G systems related to the issues of global roaming and network scalability. Originally, the 3G technologies were proposed to provide global roaming. This goal, envisaged by ITU was to have a single radio interface that provided global roaming. However, in actuality, 5 radio interfaces were adopted for 3G networks to cater for competition and migration of the installed base of 2G networks. Hence, the current implementations have failed to achieve the goal of global roaming. Therefore, global roaming is a key requirement for 4G networks.

Further, 3G specifications define three different Core Network (CN) domains, where each domain provides a different set of services. The CS domain provides circuit switched services, the PS domain provides packet switched services and the IMT domain provides IP multimedia services. This architecture is not scalable and suitable for the next generation mobile networks. Hence, for 4G networks, enhanced network architecture is being proposed which is expected to be entirely a packet switched network.

- The second objective is to make use of the achievements in the area of wireless technology. Due to the sluggish pace of 3G network deployment, other wireless technologies have captured a sizeable portion of the market. Wireless LAN (WLAN) and Bluetooth are prominent among these technologies. Satellite based networks like Thuraya are also popular in sparsely populated areas. The 4G architecture is expected to consist of a collection of such wireless technologies/networks.

The vision of 4G is to provide broadband access and global roaming using the most appropriate of modern technologies. The network hierarchy of 4G consists of Satellite Network (highest level), Cellular Network, LAN and PAN (at the lowest level).

The prominent features of such a beyond "3G network" are:

- Higher bandwidths
- Packet switched network
- Stringent network security
- global mobility and network scalability

With today's networks transitioning from switching time division multiplex (TDM) traffic to switching packetized data, 4G multiservice provisioning platforms (MSPPs) designers are faced with the challenge of creating an efficient, cost-effective and flexible system as this transition evolves.

To cater to the QoS and data rate requirements of the forthcoming applications such as High Definition TV and DVB, 3GPP and IEEE (which are along the 3GP working groups), have decided that 4G should be very spectrally efficient, should dynamically share and utilize the network resources, have a high data rate and capacity larger than 3G, have smooth handovers across heterogeneous networks, and should be based on an all-IP packet switched network. Several telecom equipment giants are currently testing 4G communication at 100Mbps while moving and 1Gbps while stationary.

As opposed to the traditional centralized TDM switching architecture, TranSwitch's distributed switching architecture does not require deployment of centralized redundant switching cards. A single component is deployed on the line and tributary cards, providing support for all TDM switching functions, including low-order granularity (Figure 7.1).

The advantage is that switching and performance monitoring at all levels are performed on any path accessible on the line directly without a separate card or device. Loopbacks are supported as a natural part of the switching function. Delay is minimized, as very few elements are used for interconnections. Switching and protection are performed locally at each tributary card. As a result, the speed of protection switching is independent of port count; switching operations will be just as fast in a fully loaded shelf as on a single tributary card. All path data are distributed to every tributary card, simplifying implementation of protection schemes such as SNCP, MS-SPRING, 1+1, 1:N and 1:1, and other functions such as bridge and roll.

In addition to SONET/SDH framing, overhead processing, and high- and low-order pointer processing, each component integrates a strictly non-blocking cross-connect for both high- and low-order path signals. Integrated high-speed serial links (Serdes) are used to transport the traffic from any line card to any tributary card, other line cards or even between tributary cards. Complete performance monitoring for all SONET/SDH traffic with any granularity (VT.15/VC-12 to STS-48/STM4-4c) guarantees full compliance to ITU-T and Telcordia recommendations.

7.4 Circuit-switched network

In telecommunications, a circuit switching network is one that establishes a circuit (or channel) between nodes and terminals before the users may communicate, as if the nodes were physically connected with an electrical circuit.

The bit delay is constant during a connection, as opposed to packet switching, where packet queues may cause varying packet transfer delay.

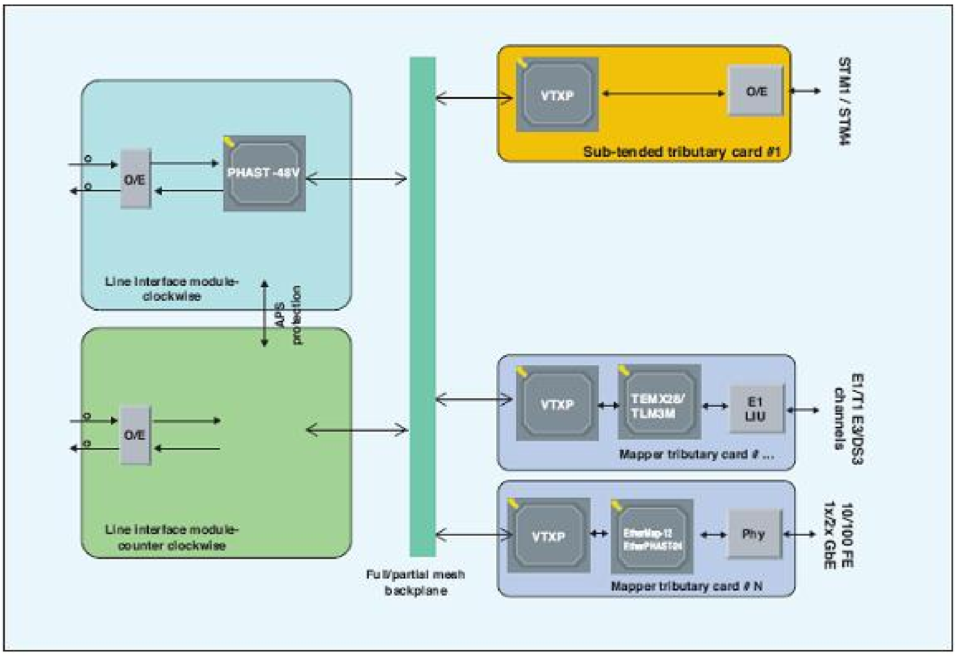


Figure 7.1: A single component is deployed on the line and tributary cards, providing support for TDM switching.

Each circuit cannot be used by other callers until the circuit is released and a new connection is set up. Even if no actual communication is taking place in a dedicated circuit that channel remains unavailable to other users. Channels that are available for new calls to be set up are said to be idle.

Virtual circuit switching is a packet switching technology that may emulate circuit switching, in the sense that the connection is established before any packets are transferred, and that packets are delivered in order.

There is a common misunderstanding that circuit switching is used only for connecting voice circuits (analog or digital). The concept of a dedicated path persisting between two communicating parties or nodes can be extended to signal content other than voice. Its advantage is that it provides for non-stop transfer without requiring packets and without most of the overhead traffic usually needed, making maximal and optimal use of available bandwidth. The disadvantage of inflexibility tends to reserve it for specialized applications, particularly with the overwhelming proliferation of internet-related technology.

Circuit switching is the most familiar technique used to build a communications network.It allows communications equipment and circuits, to

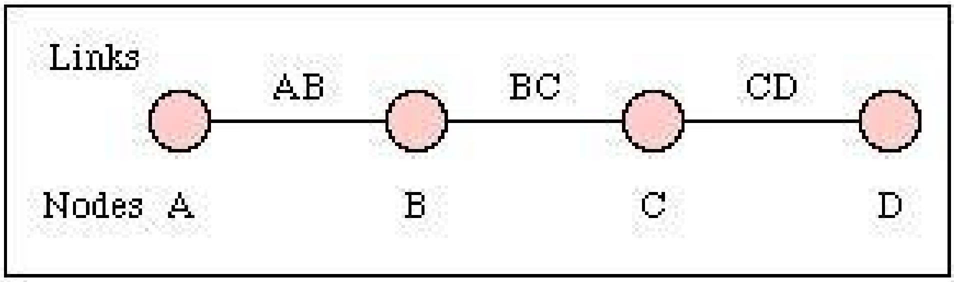


Figure 7.2: A connection between two systems A and D formed from 3 links.

be shared among users. Each user has sole access to a circuit (functionally equivalent to a pair of copper wires) during network use. Consider communication between two points A and D in a network. The connection between A and D is provided using (shared) links between two other pieces of equipment, B and C.

Network use is initiated by a connection phase, during which a circuit is set up between source and destination, and terminated by a disconnect phase. These phases, with associated timings, are illustrated in the figure below.

After a user requests a circuit, the desired destination address must be communicated to the local switching node (B). In a telephony network, this is achieved by dialing the number.

Node B receives the connection request and identifies a path to the destination (D) via an intermediate node (C). This is followed by a circuit connection phase handled by the switching nodes and initiated by allocating a free circuit to C (link BC), followed by transmission of a call request signal from node B to node C. In turn, node C allocates a link (CD) and the request is then passed to node D after a similar delay.

The circuit is then established and may be used. While it is available for use, resources (i.e. in the intermediate equipment at B and C) and capacity on the links between the equipment are dedicated to the use of the circuit.

After completion of the connection, a signal confirming circuit establishment (a connect signal in the diagram) is returned; this flows directly back to node A with no search delays since the circuit has been established. Transfer of the data in the message then begins. After data transfer, the circuit is disconnected; a simple disconnect phase is included after the end

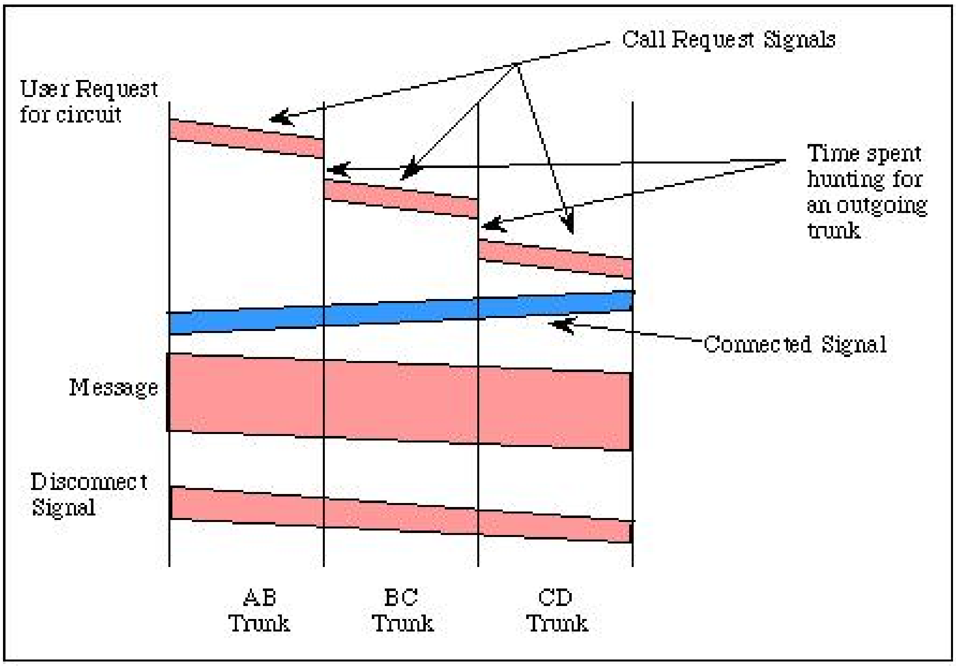


Figure 7.3: A circuit switched connection between A and D. Information flows in two directions. Information sent from the calling end is shown in pink and information returned from the remote end is shown in blue).

of the data transmission.

Delays for setting up a circuit connection can be high, especially if ordinary telephone equipment is used. Call setup time with conventional equipment is typically on the order of 5 to 25 seconds after completion of dialing. New fast circuit switching techniques can reduce delays. Trade-offs between circuit switching and other types of switching depend strongly on switching times.

7.5 Packet switched network

Packet switching is a network communications method that groups all transmitted data, irrespective of content, type, or structure into suitably-sized blocks, called packets. The network over which packets are transmitted is a shared network which routes each packet independently from all others and allocates transmission resources as needed. The principal goals of packet switching are to optimize utilization of available link capacity and to increase the robustness of communication.

Network resources are managed by **statistical multiplexing or dynamic bandwidth allocation** in which a physical communication channel is effectively divided into an arbitrary number of logical variable-bit-rate channels or data streams. Each logical stream consists of a sequence of packets, which normally are forwarded by a network node asynchronously in a first-in, first-out fashion. Alternatively, the packets may be forwarded according to some scheduling discipline for fair queuing or for differentiated or guaranteed quality of service. In case of a shared physical medium, the packets may be delivered according to some packet-mode multiple access scheme. When traversing network nodes, packets are buffered and queued, resulting in variable delay and throughput, depending on the traffic load in the network.

Packet switching contrasts with another principal networking paradigm, circuit switching, a method which sets up a specific circuit with a limited number dedicated connection of constant bit rate and constant delay between nodes for exclusive use during the communication session. Packet mode (or packet-oriented, packet-based) communication may be utilized with or without intermediate forwarding nodes (packet switches).

Packet switching is similar to message switching using short messages. Any message exceeding a network-defined maximum length is broken up into shorter units, known as packets, for transmission; the packets, each with an associated header, are then transmitted individually through the

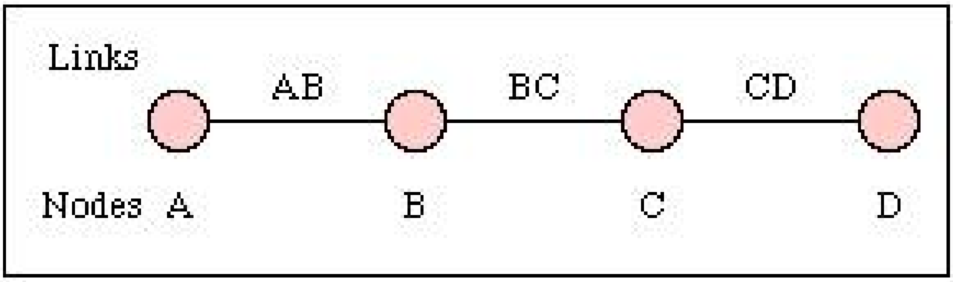


Figure 7.4: Communication between A and D using circuits which are shared using packet switching.

network. The fundamental difference in packet communication is that the data is formed into packets with a pre-defined header format (i.e. PCI), and well-known “idle” patterns which are used to occupy the link when there is no data to be communicated.

A packet network equipment discards the “idle” patterns between packets and processes the entire packet as one piece of data. The equipment examines the packet header information (PCI) and then either removes the header (in an end system) or forwards the packet to another system. If the out-going link is not available, then the packet is placed in a queue until the link becomes free. A packet network is formed by links which connect packet network equipment.

There are two important benefits from packet switching.

1. The first and most important benefit is that since packets are short, the communication links between the nodes are only allocated to transferring a single message for a short period of time while transmitting each packet. Longer messages require a series of packets to be sent, but do not require the link to be dedicated between the transmission of each packet. The implication is that packets belonging to other messages may be sent between the packets of the message being sent from A to D. This provides a much fairer sharing of the resources of each of the links.

2. Another benefit of packet switching is known as “pipelining”. Pipelining is visible in the figure above. At the time packet 1 is sent from B to C, packet 2 is sent from A to B; packet 1 is sent from C to D while packet 2 is sent from B to C, and packet 3 is sent from A to B, and so forth. This simultaneous use of communications links represents a gain in efficiency, the total delay for transmission across a packet

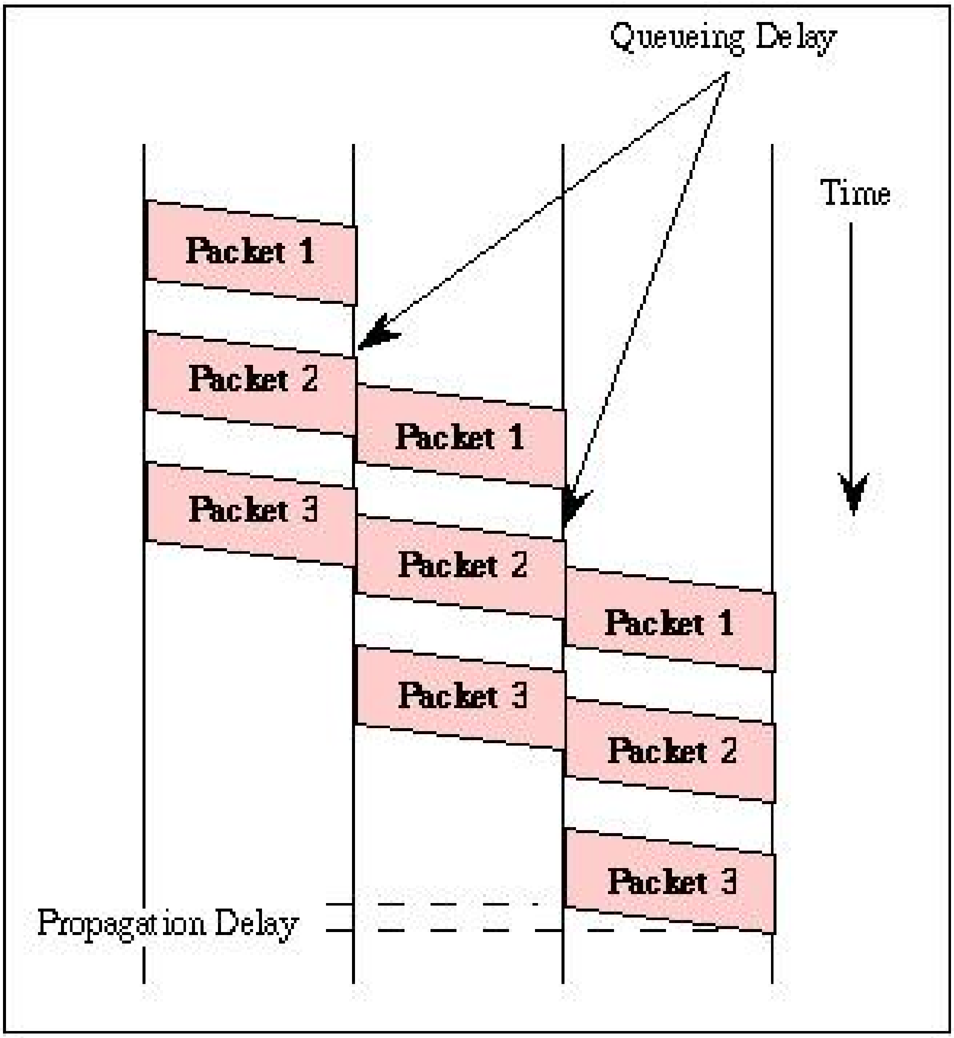


Figure 7.5: Packet-switched communication between systems A and D (The message in this case has been broken into three parts labeled 1-3).

network may be considerably less than for message switching, despite the inclusion of a header in each packet rather than in each message.

Unlike circuit switching, which requires the establishment of a dedicated point-to-point connection, each packet in a packet-switched network contains a destination address. Thus, all packets in a single message do not have to travel the same path. As traffic conditions change, they can be dynamically routed via different paths in the network, and they can even arrive out of order. The destination computer reassembles the packets into their proper sequence. Network protocols such as IP and IPX were designed for packet-based networks.

Two kinds of Packet Switching There are two basic types of Packet Switching.

1. Virtual Circuit Packet Switching Networks

An initial setup phase is used to set up a route between the intermediate nodes for all the packets passed during the session between the two end nodes. In each intermediate node, an entry is registered in a table to indicate the route for the connection that has been set up. The packets passed through this route, have short headers, containing only a virtual circuit identifier (VCI). Each intermediate node passes the packets according to the information that was stored in its table, in the setup phase and according to the packets header content.

In this way, packets arrive at the destination in the correct sequence. This approach is slower than Circuit Switching, since different virtual circuits may compete over the same resources. As in Circuit Switching, if an intermediate node fails, all virtual circuits that pass through it are lost.

The most common forms of Virtual Circuit networks are ATM and Frame Relay, which are commonly used for public data networks (PDN).

2. Datagram Packet Switching Networks

This approach uses a different, more dynamic scheme, to determine the route through the network links. Each packet is treated as an independent entity, and its header contains full information about the destination of the packet. The intermediate nodes examine the header of the packet, and decide the next hop of this packet. In the decision two factors are taken into account:

- The shortest way to pass the packet to its destination - protocols such as RIP/OSPF is used to determine the shortest path to the destination.

- Finding a free node to pass the packet to - in this way, bottle necks are eliminated, since packets can reach the destination in alternate routes. Thus, in this method, the packets don't follow a pre-established route, and the intermediate nodes (the routers) don't have pre-defined knowledge of the routes that the packets should be passed through.

Packets can follow different routes to the destination. Due to the nature of this method, the packets can reach the destination in a different order than they were sent, thus they must be sorted at the destination to form the original message. This approach is time consuming since every router has to decide where to send each packet.

7.5.1 Statistical multiplexing

Statistical multiplexing is a type of communication link sharing, very similar to Dynamic bandwidth allocation (DBA). In statistical multiplexing, a communication channel is divided into an arbitrary number of variable bit-rate digital channels or data streams. The link sharing is adapted to the instantaneous traffic demands of the data streams that are transferred over each channel. This is an alternative to creating a fixed sharing of a link, such as in general time division multiplexing and frequency division multiplexing. When performed correctly, statistical multiplexing can provide a link utilization improvement, called the statistical multiplexing gain.

Statistical multiplexing is facilitated through packet mode or packet oriented communication, which amongst others is utilized in packet switched computer networks. Each stream is divided into packets that normally are delivered asynchronously in a first-come first-serve fashion. Alternatively, the packets may be delivered according to some scheduling discipline for fair queuing or differentiated and/or guaranteed Quality of service.

Statistical multiplexing of an analog channel, for example a wireless channel, is also facilitated through the following schemes:

- Random frequency-hopping orthogonal frequency division multiple access (RFH-OFDMA)
- Code-division multiple access (CDMA), where different amount of spreading codes or spreading factors can be assigned to different users.

Statistical multiplexing normally implies "on-demand" service rather than one that preallocates resources for each data stream. Statistical multiplexing schemes do not control user data transmissions.

7.5.2 Dynamic bandwidth allocation

Dynamic bandwidth allocation is a technique by which traffic bandwidth in a shared telecommunications medium can be allocated on demand and fairly between different users of that bandwidth. This is a form of bandwidth management, and is essentially the same thing as statistical multiplexing. Where the sharing of a link adapts in some way to the instantaneous traffic demands of the nodes connected to the link.

Dynamic bandwidth allocation takes advantage of several attributes of shared networks: (1) all users are typically not connected to the network at one time (2) even when connected, users are not transmitting data (or voice or video) at all times (3) most traffic is “bursty” – there are gaps between packets of information that can be filled with other user traffic.

Different network protocols implement dynamic bandwidth allocation in different ways. These methods are typically defined in standards developed by standards bodies such as the ITU, IEEE, FSAN, or IETF. One example is defined in the ITU G.983 specification for passive optical network (PON).

7.6 Circuit Switching vs. Packet Switching

In principle, circuit switching and packet switching both are used in high-capacity networks. In circuit-switched networks, network resources are static, set in “copper” if you will, from the sender to receiver before the start of the transfer, thus creating a “circuit”. The resources remain dedicated to the circuit during the entire transfer and the entire message follows the same path. In packet-switched networks, the message is broken into packets, each of which can take a different route to the destination where the packets are recompiled into the original message.

In Circuit Switching networks, when establishing a call a set of resources is allocated for this call. These resources are dedicated for this call, and cant be used by any of the other calls. Circuit Switching is ideal when data must be transmitted quickly, must arrive in sequencing order and at a constant arrival rate. There for when transmitting real time data, such as audio and video, Circuit Switching networks will be used.

Packet switching main difference from Circuit Switching is that that the communication lines are not dedicated to passing messages from the source to the destination. In Packet Switching, different messages can use the same network resources within the same time period. Since network resources are not dedicated to a certain session the protocol avoid from waste of resources when no data is transmitted in the session. Packet Switching is more efficient and robust for data that is burst in its nature, and can withstand delays in transmission, such as e-mail messages, and Web pages.

Packet-Switching networks have the property that each of the packets in the message would travel independently to Host B. Each packet is a self-contained unit, containing the address of Host B. Such a packet is referred to as a datagram. It could be the case that all datagrams take different paths. The datagrams might arrive out-of-order at Host B. By contrast, in Circuit-Switching networks, before Host A can send the first packet to Host B, a connection (circuit) must be established between App. A and App. B. This is one single path. Whenever Host A sends signal to Host B it will send it over this path. Packets don't need to carry the address of Host B. Instead they just carry their circuit identier. When App A is done, it tears down connection.

The above is the main difference between Packet-Switching and Circuit-Switching.

Packet-Switching can result in packet loss. Loss occurs when run out of room in router's buer. There are so many packets that want to go on some outgoing link that we don't have room to hold them in the outgoing buer for that link. We'd like to be able to temporarily increase the outgoing bandwidth on the link (so as to get packets out faster) or increase the buer length (so as to have room to store extra packets), but unfortunately these are infinite quantities. Packet-Switching is a free-for-all. There is no way to make reservations of bandwidth or buer space, since the packets move independently of each other. Packets arriving at router from different rows are handled in FCFS order (first-come-first-served). For this reason Packet-Switching often called statistical multiplexing.

circuit-Switching can control loss. Bandwidth and buer space may be reserved for the connection at the time of connection setup. Note, Circuit-Switching doesn't have to mean that a link can only be used by one circuit. A link may be shared by several circuits. These are often called virtual circuits. Each reserves a portion of bandwidth. Each thinks that it is the only circuit. The way in which circuits share a link under Circuit-switching is to use TDM time-division multiplexing. Time Division Multiplexing forces the circuits to take turns. Because one of the big reasons for doing Circuit-Switching is to be able to make bandwidth reservations for the connection, reservations are very often linked to definition of Circuit-Switching.

Brief comparison

- Circuit switching is old and expensive, and it is what PSTN uses. Packet switching is more modern.

- When we are making a PSTN call, we are actually renting the lines, with all it implies. See why international calls are expensive? So if we speak for, say 10 minutes, we pay for ten minutes of dedicated line. we normally speak only when our correspondent is silent, and vice versa. Taking also into consideration the amount of time no one speaks, we finally use much less than half of what we are paying for. With VoIP, we actually can use a network or circuit even if there are other people using it at the same time. There is no circuit dedication. The cost is shared.

- Circuit-switching is more reliable than packet-switching. When we have a circuit dedicated for a session, we are sure to get all information across. When we use a circuit which is open for other services, then there is a big possibility of congestion (which is for a network what a traffic jam is for the road), and hence the delays or even packet loss. This explains the relatively lower quality of VoIP voice compared to PSTN. But we actually have other protocols giving a helping hand in making packet-switching techniques to make connections more reliable. An example is the TCP protocol. Since voice is to some extent tolerant to some packet loss (unless text - since a comma lost can mean a big difference), packet-switching is finally ideal for VoIP service of 4G mobile.

- In circuit switched network for each connection, physical switches are set in the telephone network to create a physical "circuit". That's the job of the switching office. Switches are set up at the beginning of the connection and maintained throughout the connection. In packet Switching, Packetized data is transferred. It is Multiplexed it onto the wire. Packets from different connections share the same link.

- In circuit switched network ,Network resources reserved and dedicated from sender to receiver. Not a very efficient strategy. A connection "holds" a physical line even during "silence" periods (when there is nothing to transmit).In contrast, Packet-Switching networks have the property that each of the packets in the message would travel independently. Each packet is composed by the payload(the data we want to transmit) and a header. The header contains information useful for transmission, such as:
  - Source (sender's) address
  - Destination (recipient's) address
  - Packet size

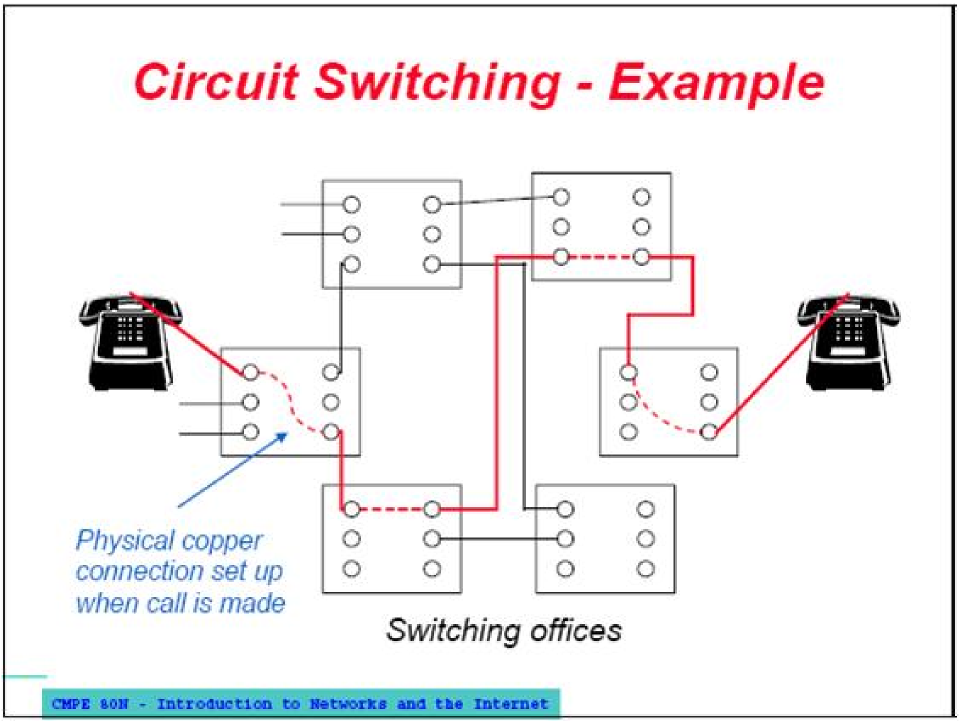


Figure 7.6: A circuit switching example.

- Sequence number
- Error checking information

The header introduces overhead, that is, additional bits to be sent.

Mobile Addresses Each mobile attached to a network is assigned a unique number (called address). A packet contains the address of the mobile that sent it and the address of the mobile to which it is sent. When packets are received, they are put together before the application accesses the data. So, very efficient strategy, as it can transmit data without keeping the line "hold".

7.7 Advantages of distributed switching structure

Compared to the centralized switching architecture, the distributed switching architecture gives the service providers linearly scalable system cost, more service capacity, lower system power, service velocity, and faster and simpler operation, administration, maintenance and provisioning (OAM&P).

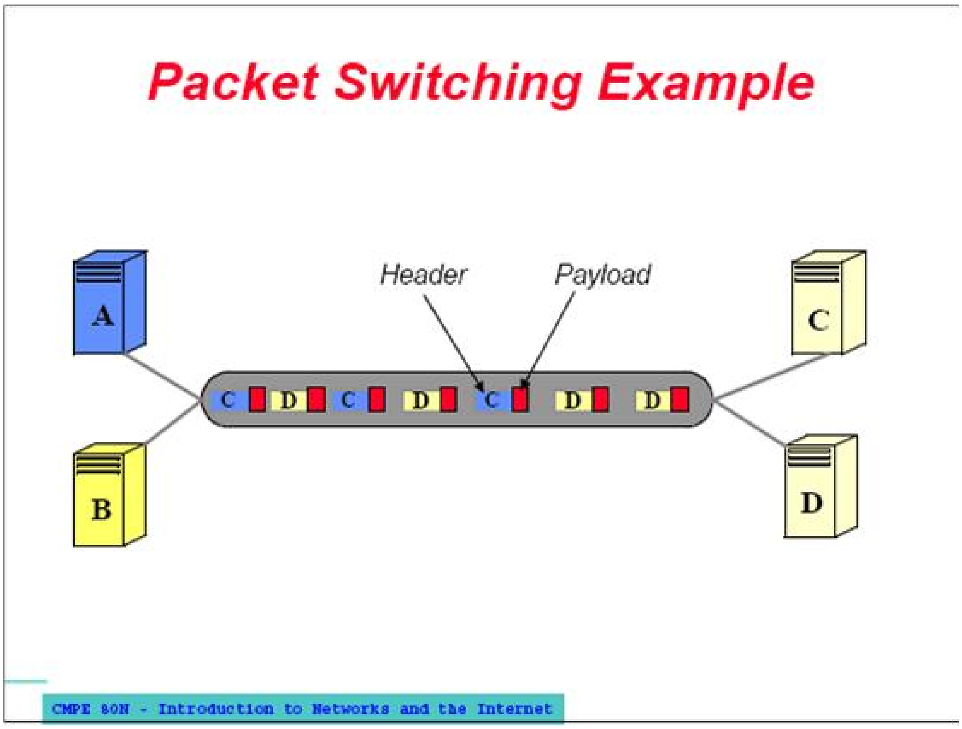


Figure 7.7: A packet switching example.

Carriers' requirement for a "pay-as-they-go" solution is no longer based on financial creativeness or incentives. This architecture does not require initial investment in a fully redundant switch fabric, leading not only to the initial cost of the system being reduced by more than 60 percent, but also supporting a true linearly scalable cost of the system.

Since the distributed architecture eliminates the need for redundant centralized switch fabrics, the result is lower cost even when fully equipped. Furthermore, since two slots are not used for the centralized switch fabric, the distributed architecture accommodates two additional line or tributary cards per shelf, providing a higher capacity for service delivery.

The design effort, system testing and integration NRE are reduced, as fewer system components need to be tested and verified.

Compared to a centralized solution, the power dissipation of the distributed system is directly proportional to the deployed components. The centralized cross-connect components, having been designed for maximum system capacity, dissipate a significant amount of power (typically, 7W-15W per device), even when only a small percentage of their capacity is used (Figure 7.2). Centralized switch fabrics generally pose additional challenges in mechanical design and heat dissipation. The distributed architecture also improves the life-cycle cost to carriers since they only pay for the power and building HVAC actually used by equipped circuits, not for hot standby fault tolerance and capacity not in current use.

By eliminating separate switching cards, the distributed architecture enhances the overall availability of the system. The system is no longer prone to a single point of failure, making it inherently fault-tolerant. This distributed switching architecture eliminates the need for complex time-slot-interchange operations in the centralized switch fabric to implement a recovery from failure. It also permits the simple implementation of N+M protection schemes, since all received line and tributary paths are available after performance monitoring of line and tributary cards.

The ability of the system to implement service changes or react to network configuration changes is known as service velocity. The distributed switching architecture is very simple and fast to configure, only requiring the tributary or line source to be specified for a given connection. When a service provider uses a network-protection mechanism such as UPSR/SNCP, the most important criteria are the time elapsed between the detection of the failure and the restoration of the service. The distributed switching architecture has the benefit of having the protection mechanism implemented where it is needed, such as the UPSR/SNCP sink (e.g. network element traffic egress point). Thus, the recovery of a specific tributary from a failure is not subject to software or capacity delays as in a centralized resource.

OAM&P development and testing effort for any carrier-grade system can easily be two- or threefold the hardware development effort. This software effort is significantly reduced in the distributed switching architecture, which requires neither fault tolerance algorithms nor complex path search algorithms. Manual operations for provisioning or restoration actions also become less complex.

Fourth-generation MSPPs need a different approach to switching to accommodate the evolving traffic mix between TDM and packet data. The introduction of bandwidth-efficient data services like Ethernet over SONET/SDH have placed a new and growing demand on low-order TDM switching and performance monitoring as well as on the need for co-located packet switching-a demand that is not well-satisfied by centralized TDM switching architecture.

It is time for MSPPs to advance beyond STS-1/VC-4 centralized TDM switching architectures. Distributed TDM switching architectures offer a

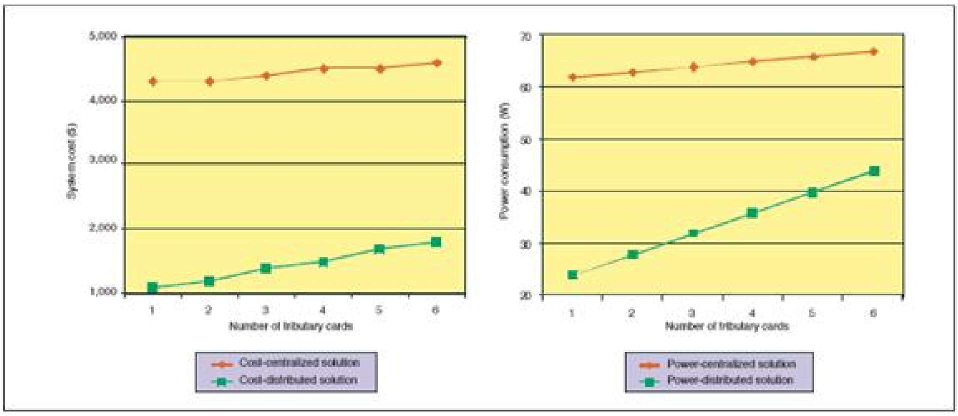


Figure 7.8: The switching architecture does not require initial investment in a fully redudnant switch fabric.

path-size agnostic, inherently fault-tolerant solution that gives the service provider linear cost and power consumption. As MSPPs become more packet-switching-centric, the need for TDM switching will decline. The cost and power consumption burden of a centralized TDM switch should not be an undesired legacy of the system. Distributed TDM switching architecture provides an evolutionary solution to efficiently fulfill current and future TDM traffic requirements.

Chapter 8

4G Mobile WiMAX Chipsets and Power Architecture

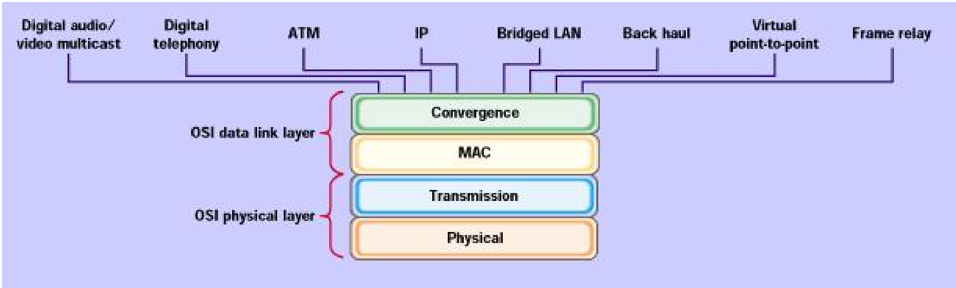


Figure 8.1: IEEE 802.16 Protocol Architecture.

8.1 What is WiMAX?

WiMAX, meaning Worldwide Inter-operability for Microwave Access, is a telecommunications technology that provides wireless transmission of data using a variety of transmission modes, from point-to-multipoint links to portable and fully mobile internet access. The technology provides up to 72 Mbit/s symmetric broadband speed without the need for cables.

8.2 Architecture

8.2.1 IEEE 802.16 Protocol Architecture

IEEE 802.16 Protocol Architecture has 4 layers: Convergence, MAC, Transmission and physical, which can be mapped to two OSI lowest layers: physical and data link.

Physical layer functions

- Encoding/decoding of signals
- Preamble generation/removal
- Bit transmission/reception.

Medium access control layer functions

- On transmission, assemble data into a frame with address and error detection fields
- On reception, disassemble frame, and perform address recognition and error detection
- Govern access to the wireless transmission medium.

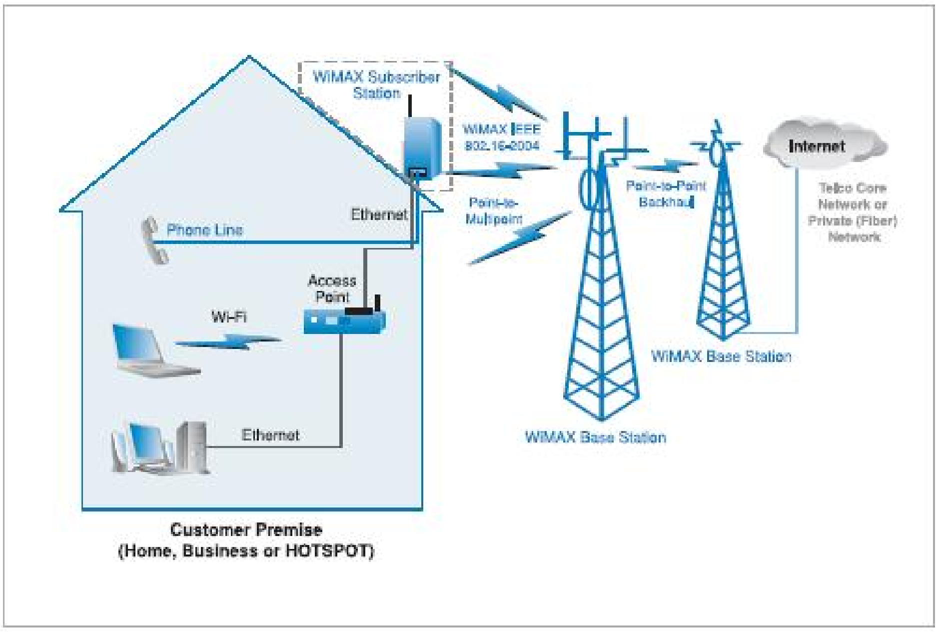


Figure 8.2: P2MP Architecture.

Convergence layer functions

- Encapsulate PDU framing of upper layers into native 802.16 MAC/PHY frames
- Map upper layer's addresses into 802.16 addresses
- Translate upper layer QoS parameters into native 802.16 MAC format
- Adapt time dependencies of upper layer traffic into equivalent MAC service.

8.2.2 P2MP Architecture

- BS connected to Public Networks
- BS serves Subscriber Stations (SS)
- Provides SS with first mile access to Public Networks.

The IEEE only defined the Physical (PHY) and Media Access Control (MAC) layers in 802.16. This approach has worked well for technologies such as Ethernet and WiFi, which rely on other bodies such as the IETF (Internet Engineering Task Force) to set the standards for higher layer protocols such as TCP/IP, SIP, VoIP and IPSec. In the mobile wireless world, standards bodies such as 3GPP and 3GPP2 set standards over a wide range of interfaces and protocols because they require not only airlink interoperability, but also inter-vendor inter-network interoperability for roaming, multi-vendor access networks, and inter-company billing. Vendors and operators have recognized this issue, and have formed additional working groups to develop standard network reference models for open inter-network interfaces. Two of these are the WiMAX Forum's Network Working Group, which is focused on creating higher-level networking specifications for fixed, nomadic, portable and mobile WiMAX systems beyond what is defined in the IEEE 802.16 standard, and Service Provider Working Group which helps write requirements and prioritizes them to help drive the work of Network WG.

The Mobile WiMAX End-to-End Network Architecture is based an All-IP platform, all packet technology with no legacy circuit telephony. It offers the advantage of reduced total cost of ownership during the lifecycle of a WiMAX network deployment. The use of All-IP means that a common network core can be used, without the need to maintain both packet and circuit core networks, with all the overhead that goes with it. A further benefit of All-IP is that it places the network on the performance growth curve of general purpose processors and computing devices, often termed "Moore's Law". Advances in computer processing occurs much faster than advances in telecommunications equipment because general purpose hardware is not limited to telecommunications equipment cycles, which tend to be long and cumbersome. The end result is a network that continually performs at ever higher capital and operational efficiency, and takes advantage of 3rd party developments from the Internet community. This results in lower cost, high scalability, and rapid deployment since the networking functionality is all primarily software-based services.

In order to deploy successful and operational commercial systems, there is need for support beyond 802.16 (PHY/MAC) air interface specifications. Chief among them is the need to support a core set of networking functions as part of the overall End-to-End WiMAX system architecture. Before delving into some of the details of the architecture, we first note a few basic tenets that have guided the WiMAX architecture development.

1. The architecture is based on a packet-switched framework, including native procedures based on the IEEE 802.16 standard and its amendments, appropriate IETF RFCs and Ethernet standards.

2. The architecture permits decoupling of access architecture (and supported topologies) from connectivity IP service. Network elements of the connectivity system are agnostic to the IEEE 802.16 radio specifics.

3. The architecture allows modularity and flexibility to accommodate a broad range of deployment options such as:
    - Small-scale to large-scale (sparse to dense radio coverage and capacity) WiMAX networks
    - Urban, suburban, and rural radio propagation environments
    - Licensed and/or licensed-exempt frequency bands
    - Hierarchical, flat, or mesh topologies, and their variants
    - Co-existence of fixed, nomadic, portable and mobile usage models.

Support for Services and Applications The end-to-end architecture includes the support for: a) Voice, multimedia services and other mandated regulatory services such as emergency services and lawful interception, b) Access to a variety of independent Application Service Provider (ASP) networks in an agnostic manner, c) Mobile telephony communications using VoIP, d) Support interfacing with various interworking and media gateways permitting delivery of incumbent/legacy services translated over IP (for example, SMS over IP, MMS, WAP) to WiMAX access networks and e) Support delivery of IP Broadcast and Multicast services over WiMAX access networks.

Interworking and Roaming is another key strength of the End-to-End Network Architecture with support for a number of deployment scenarios. In particular, there will be support of a) Loosely-coupled interworking with existing wireless networks such as 3GPP and 3GPP2 or existing wireline networks such as DSL and MSO, with the interworking interface(s) based on a standard IETF suite of protocols, b) Global roaming across WiMAX operator networks, including support for credential reuse, consistent use of AAA for accounting and billing, and consolidated/common billing and settlement, c) A variety of user authentication credential formats such as username/password, digital certificates, Subscriber Identify Module (SIM), Universal SIM (USIM), and Removable User Identify Module (RUIM).

WiMAX Forum industry participants have identified a WiMAX Network Reference Model (NRM) that is a logical representation of the network architecture. The NRM identifies functional entities and reference points over which interoperability is achieved between functional entities. The architecture has been developed with the objective of providing unified support of functionality needed in a range of network deployment models and usage scenarios (ranging from fixed - nomadic - portable - simple mobility - to fully mobile subscribers).

The NRM, consisting of the following logical entities: MS, ASN, and CSN and clearly identified reference points for interconnection of the logical entities. The figure depicts the key normative reference points R1-R5. Each of the entities, MS, ASN and CSN represent a grouping of functional entities. Each of these functions may be realized in a single physical device or may be distributed over multiple physical devices. The grouping and distribution of functions into physical devices within a functional entity (such as ASN) is an implementation choice; a manufacturer may choose any physical implementation of functions, either individually or in combination, as long as the implementation meets the functional and interoperability requirements.

The intent of the NRM is to allow multiple implementation options for a given functional entity, and yet achieve interoperability among different realizations of functional entities. Interoperability is based on the definition of communication protocols and data plane treatment between functional entities to achieve an overall end-to-end function, for example, security or mobility management. Thus, the functional entities on either side of a reference point represent a collection of control and bearer plane end-points.

8.3 A WiMAX System Consists Of:

1. A WiMAX tower, similar in concept to a cell-phone tower - A single WiMAX tower can provide coverage to a very large area as big as 3,000 square miles (8,000 square km).
2. A WiMAX receiver - The receiver and antenna could be a small box or Personal Computer Memory card, or they could be built into a laptop the way WiFi access is today.

8.3.1 Tower & Receiver

(a) WiMAX Tower.

(b) WiMAX Receiver.

Figure 8.3: WiMAX Tower & Receiver.

8.4 WiMAX Chips

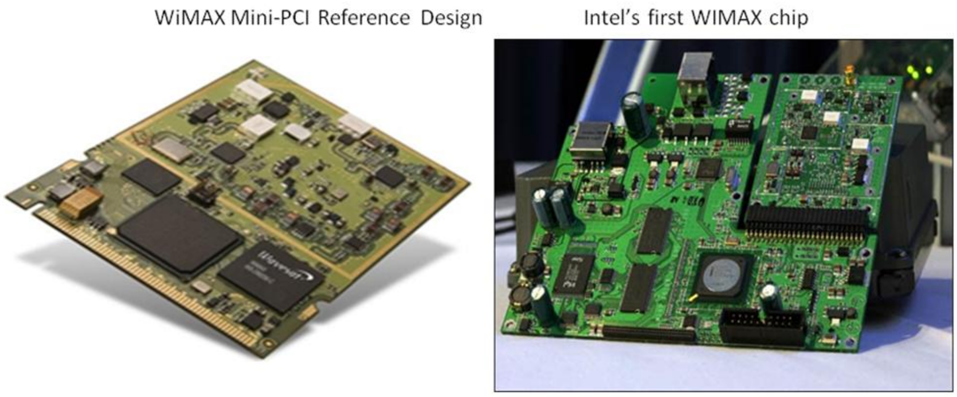


Figure 8.4: WiMAX Chips.

8.5 Frame Structure

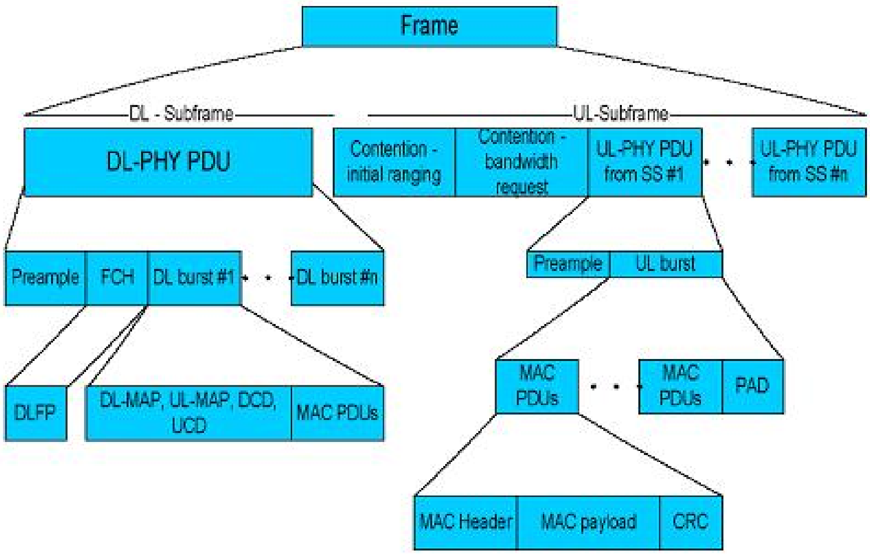


Figure 8.5: Frame Structure.

8.6 Network Entry Process

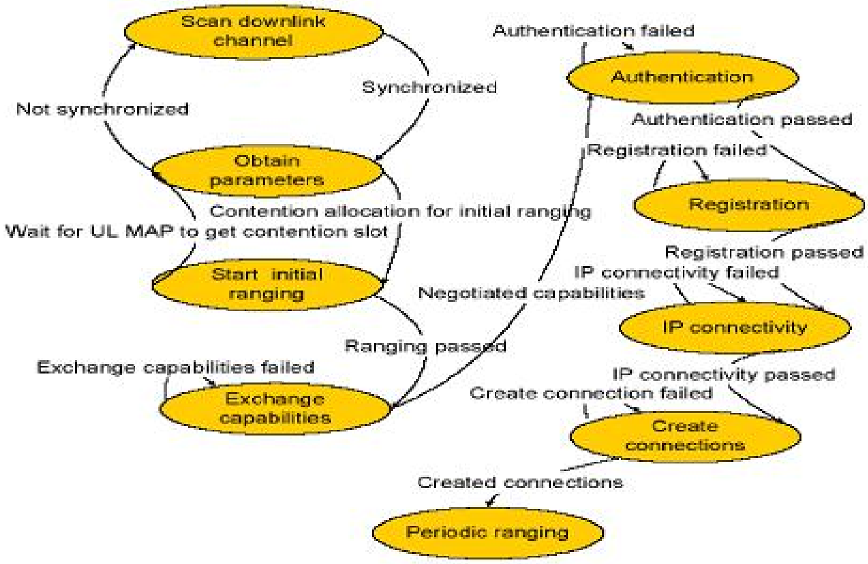


Figure 8.6: Network Entry Process.

8.7 Key Management Messages

Type	Message name	Message description	Connection
0	UCD	Uplink Channel Descriptor	Broadcast
1	DCD	Downlink Channel Descriptor	Broadcast
2	DL-MAP	Downlink Access Definition	Broadcast
3	UL-MAP	Uplink Access Definition	Broadcast
4	RNG-REQ	Ranging Request	Initial Ranging or Basic
5	RNG-RSP	Ranging Response	Initial Ranging or Basic
6	REG-REQ	Registration Request	Primary Management
7	REG-RSP	Registration Response	Primary Management
8		*reserved*	
9	PKM-REQ	Privacy Key Management Request	Primary Management
10	PKM-RSP	Privacy Key Management Response	Primary Management
11	DSA-REQ	Dynamic Service Addition Request	Primary Management
12	DSA-RSP	Dynamic Service Addition Response	Primary Management
13	DSA-ACK	Dynamic Service Addition Acknowledge	Primary Management
14	DSC-REQ	Dynamic Service Change Request	Primary Management
15	DSC-RSP	Dynamic Service Change Response	Primary Management
16	DSC-ACK	Dynamic Service Change Acknowledge	Primary Management
17	DSD-REQ	Dynamic Service Deletion Request	Primary Management
18	DSD-RSP	Dynamic Service Deletion Response	Primary Management

Figure 8.7: A portion of key management message.

Type	Message name	Message description	Connection
26	SBC-REQ	SS Basic Capability Request	Basic
27	SBC-RSP	SS Basic Capability Response	Basic
28	CLK-CMP	SS network clock comparison	Broadcast
29	DREG-CMD	De/Re-register Command	Basic
30	DSX-RVD	DSx Received Message	Primary Management
31	TFTP-CPLT	Config File TFTP Complete Message	Primary Management
32	TFTP-RSP	Config File TFTP Complete Response	Primary Management
33	ARQ-Feedback	Standalone ARQ Feedback	Basic
34	ARQ-Discard	ARQ Discard message	Basic
35	ARQ-Reset	ARQ Reset message	Basic
36	REP-REQ	Channel measurement Report Request	Basic
37	REP-RSP	Channel measurement Report Response	Basic
38	FPC	Fast Power Control	Broadcast

Figure 8.8: Another portion of key management message.

8.8 How WiMAX Works

In practical terms, WiMAX would operate similar to WiFi but at higher speeds, over greater distances and for a greater number of users. WiMAX could potentially erase the suburban and rural blackout areas that currently have no broadband Internet access because phone and cable companies have not yet run the necessary wires to those remote locations.

A WiMAX system consists of two parts:

- A WiMAX tower, similar in concept to a cell-phone tower - A single WiMAX tower can provide coverage to a very large area – as big as 3,000 square miles (~8,000 square km).
- A WiMAX receiver - The receiver and antenna could be a small box or PCMCIA card, or they could be built into a laptop the way WiFi access is today.

A WiMAX tower station can connect directly to the Internet using a high-bandwidth, wired connection (for example, a T3 line). It can also connect to another WiMAX tower using a line-of-sight, microwave link. This connection to a second tower (often referred to as a backhaul), along with the ability of a single tower to cover up to 3,000 square miles, is what allows WiMAX to provide coverage to remote rural areas.

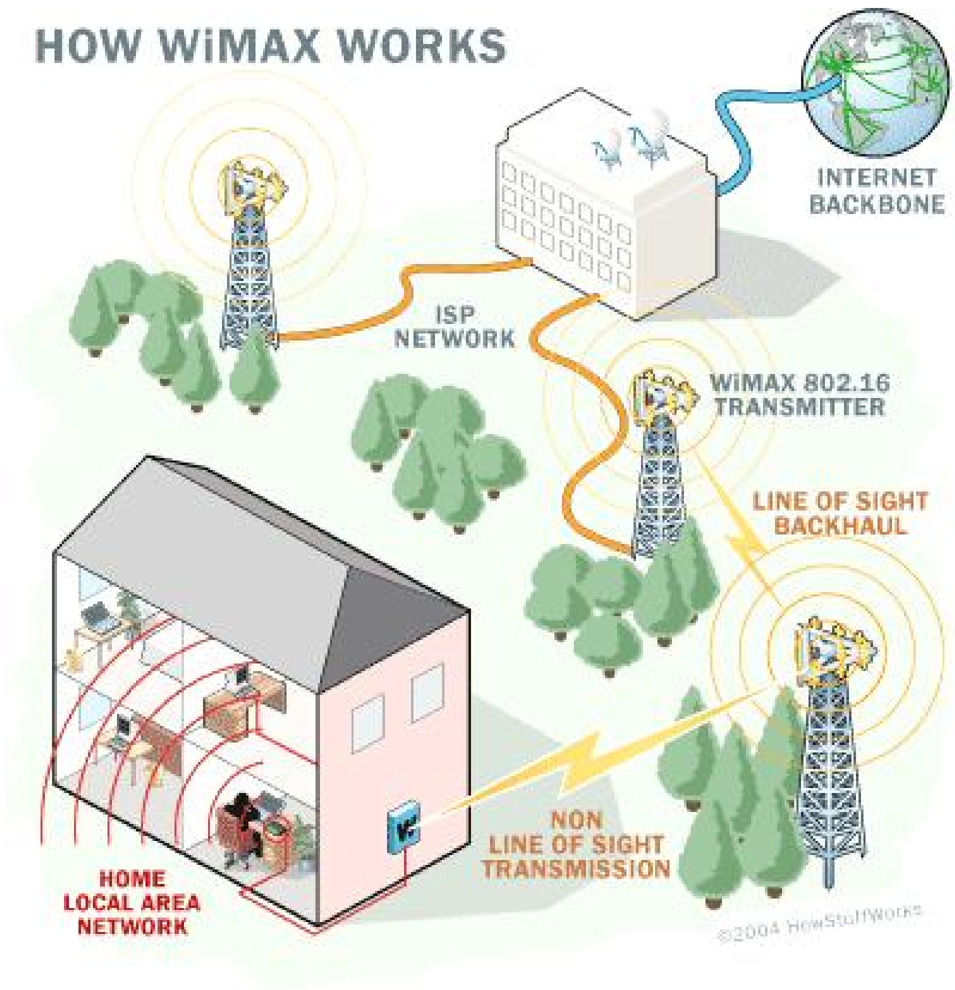


Figure 8.9: How WiMAX Works.

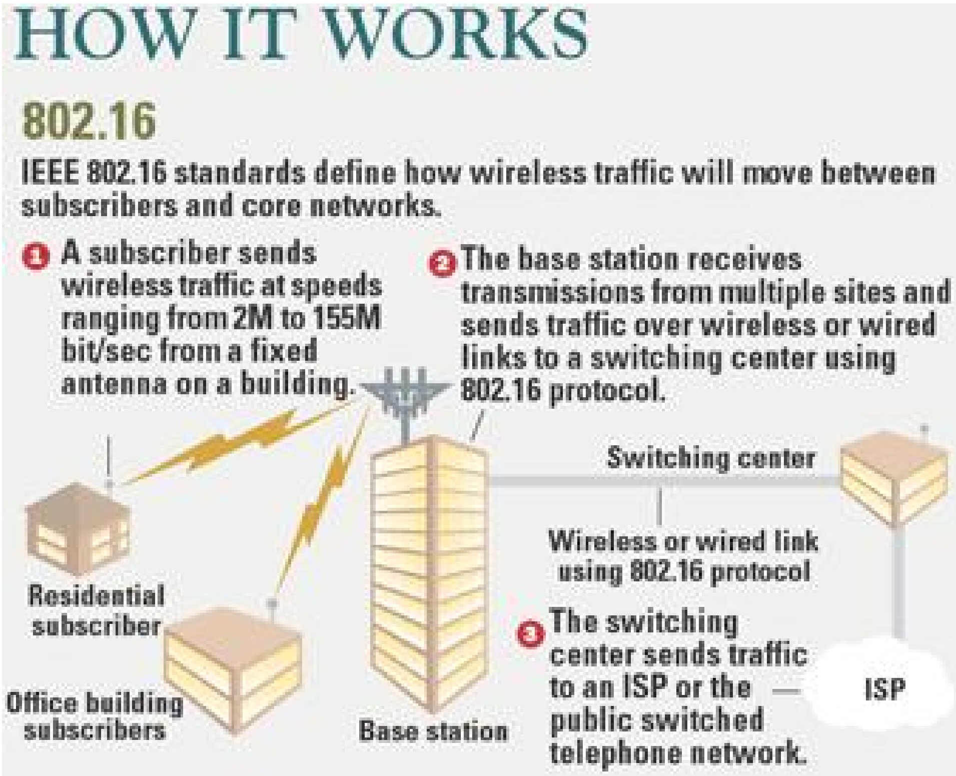


Figure 8.10: How It Works.

What this points out is that WiMAX actually can provide two forms of wireless service: There is the non-line-of-sight, WiFi sort of service, where a small antenna on your computer connects to the tower. In this mode, WiMAX uses a lower frequency range – 2 GHz to 11 GHz (similar to WiFi). Lower-wavelength transmissions are not as easily disrupted by physical obstructions – they are better able to diffract, or bend, around obstacles. There is line-of-sight service, where a fixed dish antenna points straight at the WiMAX tower from a rooftop or pole. The line-of-sight connection is stronger and more stable, so it's able to send a lot of data with fewer errors. Line-of-sight transmissions use higher frequencies, with ranges reaching a possible 66 GHz. At higher frequencies, there is less interference and lots more bandwidth.

WiFi-style access will be limited to a 4-to-6 mile radius (perhaps 25 square miles or 65 square km of coverage, which is similar in range to a cell-phone zone). Through the stronger line-of-sight antennas, the WiMAX transmitting station would send data to WiMAX-enabled computers or routers set up within the transmitter's 30-mile radius (2,800 square miles or 9,300 square km of coverage). This is what allows WiMAX to achieve its maximum range.

8.9 Uses Of WiMAX

WiMAX is designed as a wireless alternative to DSL and cable for last mile broadband access and as way to interconnect Wi-Fi hotspots into a Metropolitan Area Network. Although, the actual uses for WiMAX overlaps those for Wireless Local Area Network up until the mobile Wide Area Network level. Telephone and cable companies are closely probing the potential of WiMAX as a "last mile" connectivity option. This will result to a better-priced service for both home and business customers and not to mention the elimination of the "captive" customer bases for both telephone and cable networks.

In theory, WiMAX can provide connectivity to users within a 31 mile radius even if there is no direct line if sight. However, actual field tests show that the practical limits seem to be just around 3 to 5 miles. According to WiMAX proponents, the technology can provide shared data rates up to 70 Mbit/s. This is enough to connect 60 T1-type connections simultaneously and over a thousand homes running at 1 Mbit/s DSL level connectivity. Practical maximum data rates in actual field tests show can only go between 500 kbit/s to 2 Mbit/s and is quite dependent on the conditions at a given site.

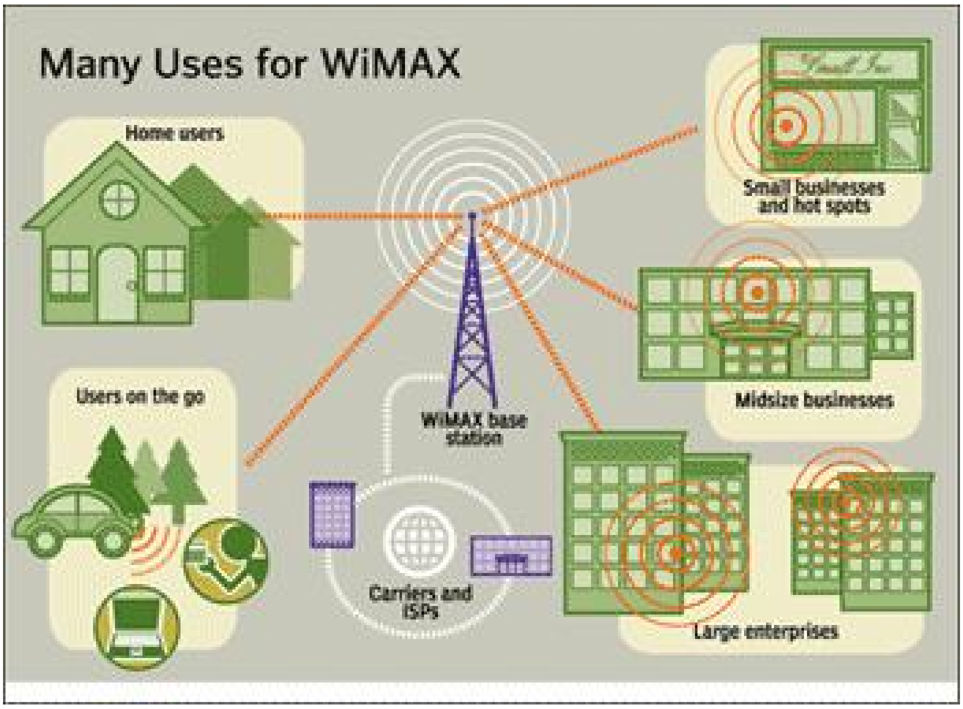


Figure 8.11: Many uses of WiMAX.

Despite the numbers given, there are a lot of ways to improve the speed and range of a WiMAX connection using pre-existing technology. One interesting option for companies with analog cellular network is to let WiMAX "share" a cell tower since it will not interfere with any of the function of the cellular arrays while utilizing the licensed radio frequencies of the analog cellular network to increase its speed and range.

WiMAX antenna can also be directly connected to an Internet backbone using a fiber optic cable. This is one of the means to increase bandwidth for data-intensive applications running across a wireless network or as a back-haul for cellular phone and Internet traffic from a remote area back to a backbone. WiMAX can effectively improve a wireless infrastructure in a decentralized, inexpensive and deployment-friendly manner.

WiMAX is seen as a very good alternative to expensive urban deployments of T1 back-hauls in developing countries with limited wired infrastructure and cruel geography. The cost to install a WiMAX station as a single hub or using an existing cellular tower will be very small compared to a wired solution. WiMAX's 31-mile diametrical range also works well with the low population density and the wide flat areas common to developing countries. Some areas have skipped wired structures due to inhibitive costs and WiMAX can easily fill the gap in-between with its low-cost wireless solution.

There is no global license assigned for WiMAX although it has a very wide RF spectrum under the IEEE 802.16 specifications. The primary band used in the US for WiMAX is around 2.5 GHz although majority of the band is already assigned to Sprint Nextel. In other parts of the world, the bands used are usually around 2.3/2.5 GHz, 3.5 GHz and 5 GHz where the 2.3/2.5 GHz is widely used in Asia.

Chapter 9

Reconfigurable 4G Mobile Environment

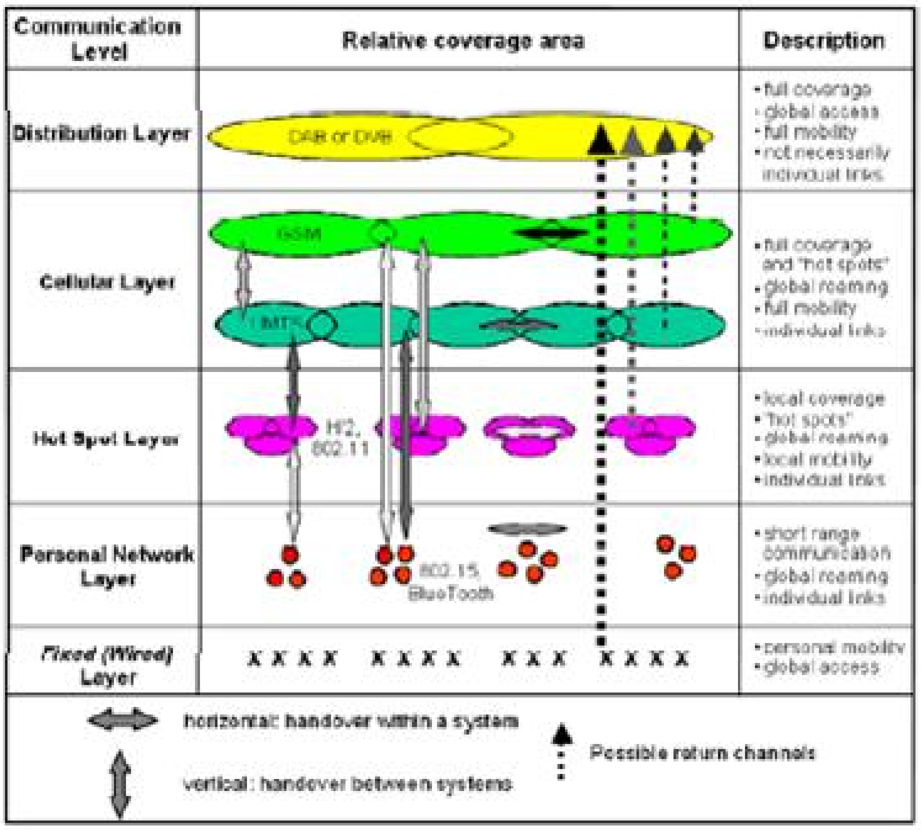


Figure 9.1: Connectivity, mobility & hand off option in 4G.

9.1 Introduction

Over the last decade, the mobile industry has developed into a breeding ground for innovative wireless access technologies. In addition to second (2G) and third (3G) generation mobile communication systems, broadband WLAN type systems as HIPERLAN/2, IEEE 802.11 and broadcast systems as DAB and DVB-T are becoming available and short range connectivity systems like Bluetooth are being developed rapidly. In the fixed access realm, systems such as xDSL and in particular ADSL are increasing the user data rate significantly on the last mile, along with fixed wireless access or wireless local loop systems that are being developed as complementary access solutions to wired access systems.

The combination of these technologies provides a very flexible and powerful platform to support future requirements of services and applications that will emerge in mobile communication systems beyond 3G, broadly termed 4G.

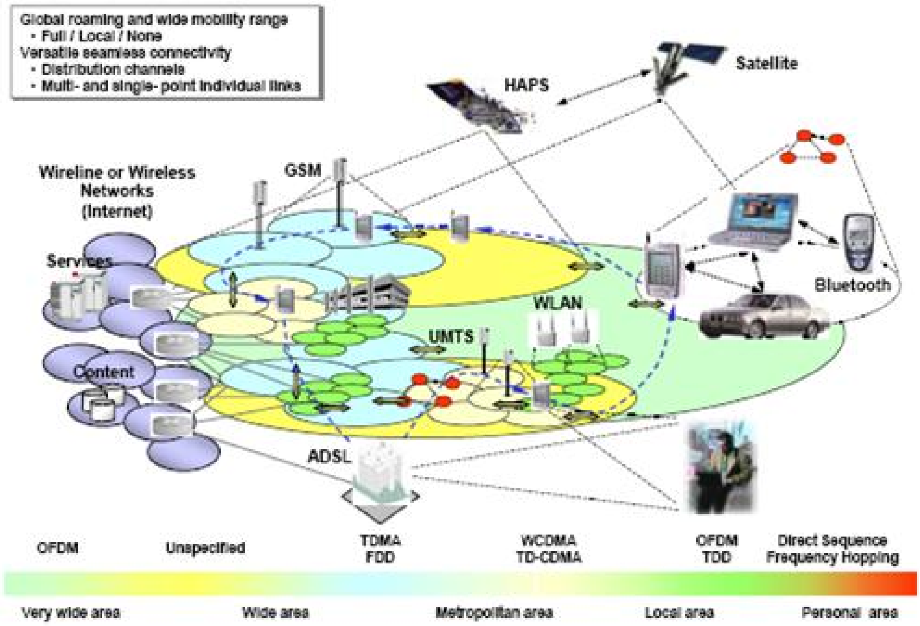


Figure 9.2: The emerging heterogeneous 4G mobile Network.

9.2 User Perspective

The globally accepted vision of 4G entails not only major technical challenges but also upgrades mobile user expectations to a significantly higher level. From the user viewpoint, the major expectation concerning refers to the delivery of a wide variety of multimedia applications and Value-added services, while hiding the technical complexity of the underlying communications schemes. To support a high degree of personalization in 4G mobile service provision, an intelligent personality management mechanism, capable of learning and understanding the preferences of the mobile user and the requirements of the services he/she is currently using, and controlling the behavior of reconfigurable components -that may be dispersed in various locations of the wireless access infrastructure (e.g., mobile device, radio access networks, service gateways, etc) - accordingly.

Ubiquitous mobile access

The end user expects to enjoy ubiquitous, robust connectivity with quality suitable for executing the widest possible range of applications and services.

Open service access

Future mobile users will not tolerate any subscription-based limitations on the access systems they may use at any time. Global roaming will be key requirement and dynamic, instant and personalizable selection of optimal access systems on a per-session basis will rank high among user expectations.

Ease of access to applications and services

Ideally, transparent discovery and switching between services and radio access modes, based on an intelligent establishment and interpretation of user preferences and application requirements will be available in a customizable way.

Task-sensitive QoS adaptation

QoS adaptation and QoS (re)-negotiation must be managed transparently for the end-user, particularly for traffic demanding services (e.g. teleconferencing, broadcast multimedia, voice telephony, etc), yet in a way that considers the nature of each particular application and its importance to the user. The mobile user will demand a sophisticated yet intuitive interface to signal the QoS level he/she prefers and expect the system to automatically select and employ the most appropriate configuration of bearer services, taking into account his/her personal cost-related preferences and without breaking the “transparency” principle.

Relevant service and meaningful billing

This implies intelligent discovery, presentation and selection of service options based on billing schemes; distribution of application processing between network and terminal to economize on terminal resources and billing schemes that hide the inherent complexity (e.g., single bill).

Technology comfort

Given the high degree of complexity in mobile devices and wireless access configurations, the mobile user will expect 24/7 online support for technical problems, making customer relationship management a major concern for any future mobile ISP.

The user perspective has been meeting much attention in the WWRF, a common endeavor between industry, academia and regulatory bodies of the mobile communications sector. The WWRF has also embraced a user-centric viewpoint in the face of the ’multi-sphere’ concept (Figure 9.3). In

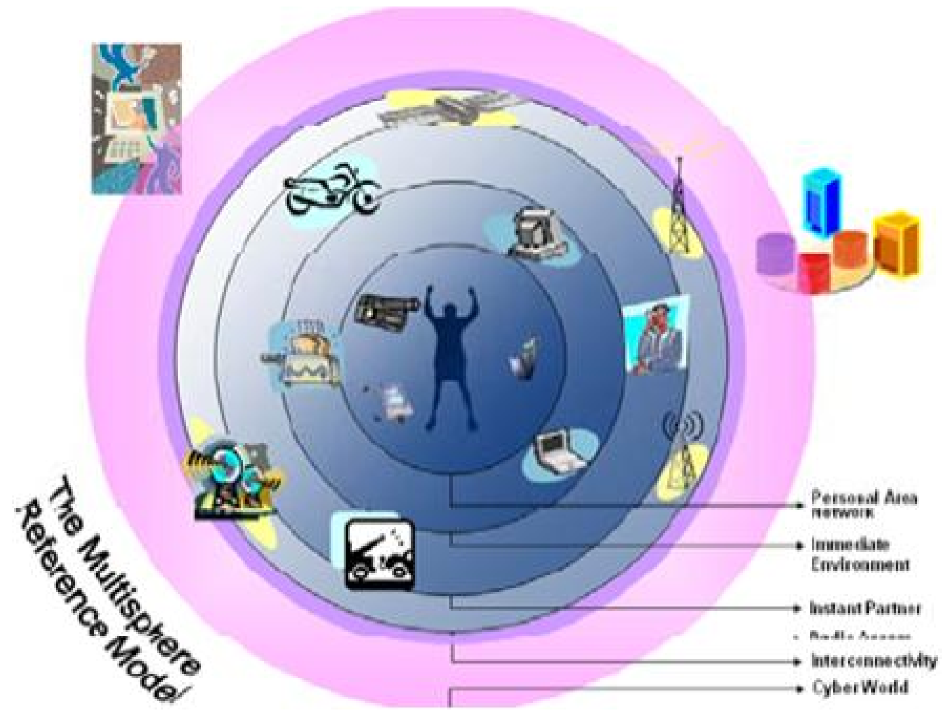


Figure 9.3: The WWRF multi sphere concept.

the future, vertical mobile applications and value-added services will employ a multitude of wireless technologies in an ad-hoc manner. These elements will surround users in a homocentric way, ranging from miniature wireless devices fabricated in wearable items (e.g., user clothes) to in-site environmental (utility) devices (e.g., refrigerator, TV, air conditioner, etc) to outdoor mobile (e.g., car, train, etc) and/or immobile (e.g., WLAN access point, cellular base station, etc) devices, and, ultimately, autonomous intelligent agents in virtual reality fore (i.e., cyberspace).

These requirements as well as the vision, clearly identify the need for a flexible, personalizable and ubiquitous service provision process in 4G.

9.3 FLEXIBLE SERVICE PROVISION

4G will be also characterized by the participation of multiple players in the mobile value chain (e.g., network operators, user application developers, content providers, etc) that will compete alongside the incumbents for a portion of the revenue stream. The foreseen diversified value chain will require more flexible, reconfigurable and co-operating networks, as well an overall intelligence that will dynamically control the behavior of engaged computing and communication resources. The resulting diversity in business models means that a highly configurable service delivery chain will be necessary.

Dynamic user registration

With new access technologies increasing their footprint and competing alongside the incumbent cellular-based access, it becomes apparent that legacy single-technology user registration procedures will fall short of the "Always Best Connected" vision of 4G. On-demand user registration to service discovery facilities that are available under the current radio coverage will be necessary, especially in situations where numerous, a priori unknown, (private) wireless access networks (e.g., WLAN hot spots) exist.

Rapid service deployment

The "walled garden" approach to mobile service provision, with its proprietary instrumentation and long time-to-market will face intense competition in the dynamic mobile value chain, where supporting technologies built on open standards (e.g., IETF protocols, W3C standards, open APIs) will provide an cheap, fast and developer-friendly way of building and deploying value-added services and mobile applications.

9.3.1 Reconfigurability management

Reconfigurability is understood as a major enabler of future mobile systems, while network and terminal reconfigurability management is considered a complicated task that spans multiple layers of the network infrastructure, involves the constant monitoring of network state and requires interaction with underlying equipment that (possibly) resides in multiple administrative domains.

9.3.2 Creation of a VHE-like portal

Mobile users will value a single entry point, through which the discovery and optimal provision of a plethora of services is performed and which can be tailored to the instant service provision context (e.g., terminal capabilities, user location, network characteristics, etc) as well as their personal preferences.

9.3.3 Flexible and user-friendly billing

Ideally, a single bill should be generated per user for all consumed services, including network and value-added ones. That requires billing systems that collect and process data from various network infrastructures (e.g., 2G/3G cellular networks, IP networks, PSTN) and enable the on-the-fly calculation of pricing schemes, the dynamic modification of tariffs and revenue sharing between the involved parties.

9.4 ENABLING TECHNOLOGIES

Based on the user requirements, the ANWIRE task force has identified the enabling technologies for reconfigurability and adaptability in 4G:

9.4.1 Middleware technologies

The consistent provision of 3rd-party applications in a nomadic context as diverse and technologically heterogeneous as the 4G one, is not a trivial task. Typical application development practices become complicated by the need to anticipate all possible networking contexts during application design and application developers become overwhelmed by having to interpolate network-dependent code in the application's core business logic. It is therefore necessary to provide a middleware layer that abstracts the technological details of the underlying mobile network infrastructure and facilitates a focused application development, whilst also providing open interfaces for a number of important tasks (e.g., user profile management, policy management, QoS management, network abstraction, dynamic user and service registration, etc).

9.4.2 Context awareness and adaptation

In 4G, applications will face an enormous range of operational parameters, thereby rendering adaptation as a critical issue. Adaptability will require constant awareness to the evolving service provision context (e.g., user preferences, user location and status, mobile terminal capabilities and resources, etc). A system is context-aware if it explores context to provide relevant information and/or services to the user, where relevance depends on the user's current tasks and preferences. Thus, context-aware adaptability is regarded as the ability to autonomously adapt, either proactively or reactively, to observed or announced changes in the environment (e.g., changes in the user's location or in the mobile terminal's power drain rate). Consequently, middleware must support a flexible context definition scheme and also provide mechanisms for context capture, collection, management and distribution to concerned applications using open interfaces that impose no restrictions on application architecture or implementation. The middleware must hide the technical details of underlying networks while exposing particular contextual aspects (e.g., available QoS classes, resource status, etc) that are important for applications to behave efficiently and to adapt when necessary. An important category of contextual information is user preferences. The middleware must enable the retrieval and handling of user preferences in a generic way so that applications or even the middleware itself - may align their adaptation strategy to the optimal trade-off between network QoS, application presentation fidelity, communication security and overall service cost.

9.4.3 Frameworks for reconfiguration

The advent of frameworks such as OSA/Parlay increases the openness of networking infrastructures (e.g., UTMS). OSA/Parlay capabilities located in specialized servers expose selected network features and allow the dynamic re-configuration of network functionality, thereby facilitating a wide range of previously unimaginable applications.

9.4.4 Reconfigurability control middleware

In 3G and post-3G mobile service provision, the role of middleware is not restricted to handling fundamental (albeit important) tasks like seamless "binding" and communication between distributed application objects; it should also incorporate higher-level functionality to facilitate adaptable service provision. Furthermore, deployment and use of diverse distributed applications may require provisioning - or even reconfiguration - of network equipment in multiple administrative domains, an issue best tackled by service provision platforms that can act as a control point for service provision flexibility and adaptability. Systems that are reconfigurable may change both in terms of structure (e.g. by downloading and instantiating a new radio interface and its associated protocol stack) and behavior (e.g. by reconfiguring the operational parameters of a particular protocol). Reconfigurability management marries the network-level constituents that participate in the service provision with the higher-layer management intelligence required to enforce interoperability and consistency between network services and 3rd-party applications over time and across what may be different implementations of functionality. To this end, the principles of open API frameworks can be exploited to decouple parts of the middleware architecture to achieve abstraction from implementation instruments and to facilitate runtime middleware evolution (Figure 9.1).

9.4.5 Reconfigurable signal-processing algorithms (RSP)

Reconfigurable signal processing tries to bridge the gap between dedicated ASICs and programmable DSPs by offering a better price/performance ratio through the provision of optimal processing of a wide range of algorithms. RSP merges the programmability of the FPGA approach with an architecture that is fine-tuned to signal-processing tasks. This means that the RSP solution is silicon and computationally efficient, striking a high performance.

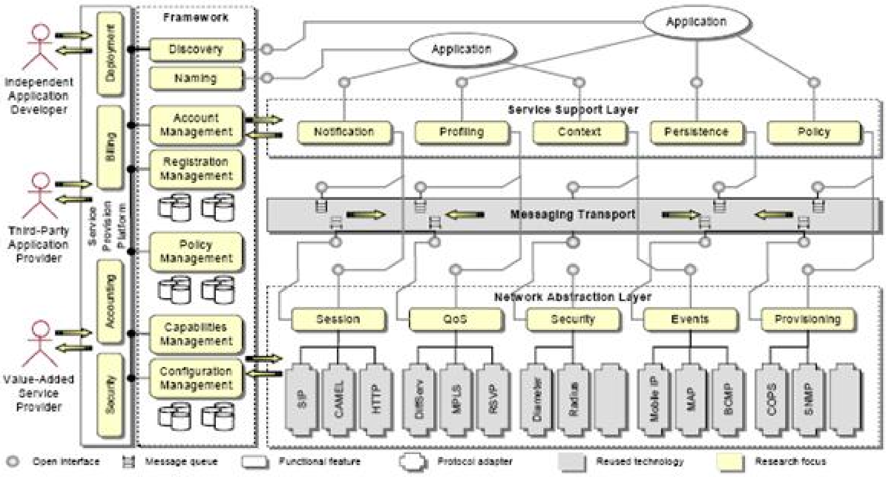


Figure 9.4: The middleware architecture.

Nonetheless, major challenges must be overcome when mobile communication components are designed specifically with runtime reconfiguration in mind. Predicting which parts of the processing are most vulnerable to specification changes and would therefore benefit from being reconfigurable is an extremely important consideration that can be both difficult and time-consuming. This is an ongoing concern due to sheer design complexity, but using RSP to deal with specific functions minimizes the overall design risk. Where reconfiguration is done on the fly in order to share hardware across multiple functions, the scheduling of functionality switching needs to be carefully considered.

9.4.6 H/W and S/W co-design methodologies

According to the IEEE DASC Co-design Study Group, "HW/SW co-design is a design methodology supporting the concurrent development of hardware and software (co-specification, co-development and co verification) in order to achieve shared functionality and performance goals for a combined system". Co-design comprises co-specification, system architecture definition, partitioning, synthesis and co-verification (Figure 9.5).

The purpose of co-specification is to capture the requirements of the complete system. The implementation architecture and style of the system is defined at architecture definition phase. The purpose of partitioning phase is to divide the functionality of the complete system into hardware and software, which are designed in technology dependent synthesis phase. Co verification is needed in each phase to ensure that transformations are done

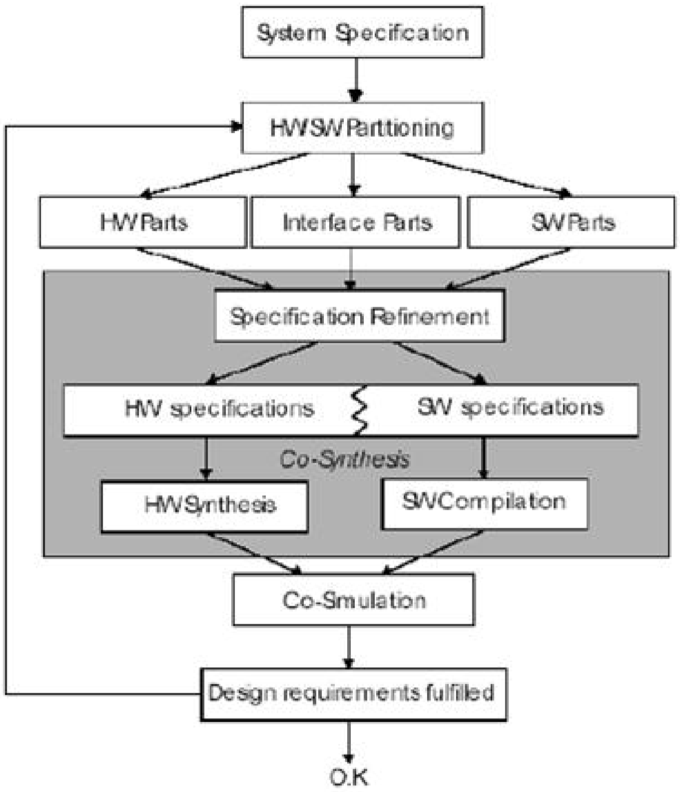


Figure 9.5: The design flow of the general co-design approach.

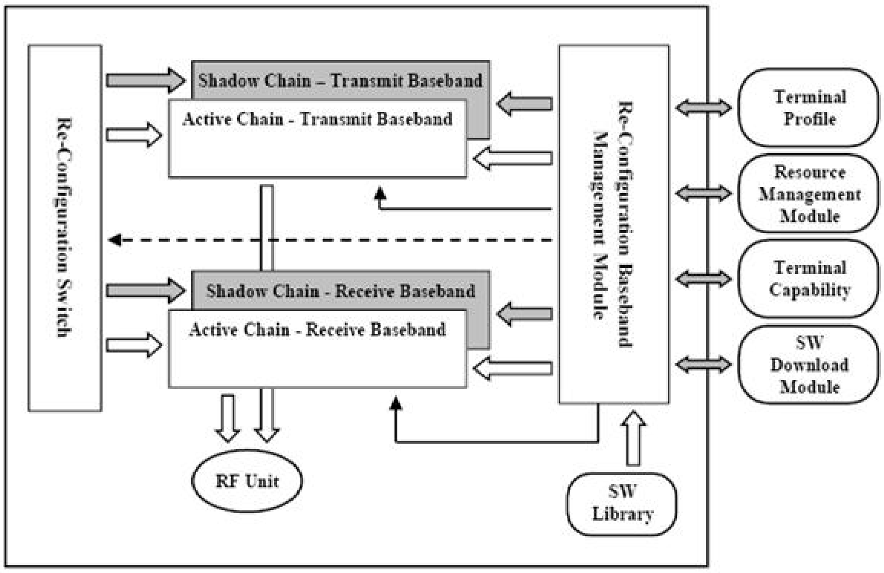


Figure 9.6: Generic software architecture of a reconfigurable baseband.

correctly and that the intended functionality is preserved throughout all the phases.

9.4.7 Reconfigurable baseband architectures

Multi-standard equipment sets very demanding requirements in terms of computational performance and flexibility. Future SDR baseband processing architectures will employ different kinds of processing, interconnection and storage resources with support for partial and total reconfiguration. For seamless switching between radio standards, the baseband architecture must support dynamic creation and binding of its modules. Functionality-implementing classes must be instantiated through dynamic binding, whereby the required functionality (and the respective implementation) of a given class is made available at run-time, while the structure of the class is known a priori. For seamless baseband reconfiguration, a redundant, so-called 'shadow transceiver' is necessary. Designing generalized modules that can switch their operating parameter during runtime can minimize duplication of baseband modules. This approach requires a reconfiguration table that lists all baseband modules and their parameter requirements for all standards supported by the SDR device. The IST project TRUST has designed a software architecture for a reconfigurable baseband that can be embedded into an overall system and network architecture to support the realization of SDR equipment.

9.4.8 Dynamic spectrum allocation (DSA)

The current method of assigning spectrum to different radio systems is fixed spectrum allocation (FSA). With FSA, radio spectrum is allocated to a particular radio standard, in fixed-size blocks that are separated by a guard band to limit interference. The allocated spectrum is available solely for the use of the radio license owner. Considering that almost all services (e.g., telephony, video streaming, web browsing, multicast applications, etc) that are envisaged as future mobile services, have distinct time-varying traffic demands, and the demand for different services on different networks will depend on location, we understand that spectrum usage will vary temporally and spatially. This is the motivation for dynamic spectrum allocation (DSA), which includes temporal DSA and spatial DSA. The DSA can be given temporal and spatial adaptability by partitioning the service area under consideration into DSA areas to give spatial adaptability, and allowing each of these DSA areas to change their spectrum over time independently. Research work indicates that around 30% increases in spectrum efficiency could be gained when DSA is compared to FSA, although the actual value does depend on the traffic patterns used in practice.

9.5 SYSTEM LEVEL ISSUES

Flexibility in service provisioning typically concerns the tasks of service registration, deployment, discovery, adaptation and management. The operation of an advanced service provision framework introduces additional requirements:

9.5.1 User perspective

The system should offer capabilities for dynamic user registration, on-line personalized discovery of available services and selective downloading. The customization of user interface, according to the terminal capabilities and the favorite services, should be supported, to minimize the number of user actions necessary to use a particular service or launch a downloaded application. The architecture should provide users with a concrete and clear means of accessing services from the terminal at hand, without subjecting them to complex installation procedures.

9.5.2 Instrumentation

Use of these supporting functions should be abstracted from specific transport, system and service attributes and implementing technologies through an open API interface. That way, a plethora disparate application and/or value-added services can be deployed - and subsequently discovered and consumed - over a multitude of access networks using a variety of end-user mobile devices. These mechanisms must be interworked to the underlying management architecture and network/system provisioning protocols (e.g., SNMP, COPS, COPS-PR).

9.5.3 Metadata and profiles

In support of the aforementioned requirements, globally accessible repositories are required for maintaining generic user, service and network profiles that can be populated with personalized information and preferences. Multiple profile instances per user must be supported, and processes for adaptation of certain attributes by either the user(s) themselves of their authorized intelligent agent(s) must be made available. Protection of the user profile and access authorization must be supported, as well as the customization of the access control mechanism, to support user anonymity and/or delegation if necessary. Thus, profile flexibility, efficiency and control granularity are strong architectural requirements.

9.5.4 QoS management

Users will want to register on demand with available service providers, discover and download services and media-rich content, so QoS provisioning must be supported by the architecture. Users should be notified about the level of QoS guaranteed by the current service provider(s) in order to decide whether or not they wish to use the service. Moreover, a listing of the various possible options for automatic QoS degradation should be presentable to the user on his/her demand.

9.5.5 Seamless operation

Reconfigurability may manifest itself in various forms, from the dynamic downloading of protocol modules or parts of protocol stacks, the dynamic modification of system parameters through open. Application Programming Interfaces (APIs) to listings that determine the characteristics of switching between different types of operation mode. In all cases, an intelligent orchestration between mobility, resource and reconfiguration management will be required to build end-systems capable of hiding network topology alterations, resource re-allocation and (radio) mode switching in a transparent way.

9.6 Conclusion

There is no doubt that, in 4G, users will expect a high degree of personalization depending on their preferences and (instant) context, with price and

QoS as key decision factors and security and privacy preservation requirements. That means that a whole range of challenging problems must be resolved by network operators and service providers concerning the personalization of services, the provision of QoS support and the respect of user privacy settings. How to share information about user preferences while preserving each user's privacy and providing reliable and transparent security, will be crucial. On the other hand, the integration and interconnection of networks administered by different operators, as well as the business agreements with service providers and application developers, will be essential for the development and operation of the advanced services demanded by the users. The combination of all the technologies required for flexible service provision in the context of future "beyond 3G" systems may result in systems too complex for practical deployment and operation. This point may jeopardize the commercial success of the new systems, consuming a lot of resources during the process of development, deployment and operations. So, it is important that all supporting technologies and solutions remain as simple as possible in order to facilitate efficient interworking and scalable operation. Thus, simplicity and modularity will be key architectural properties of the 4G mobile engineering paradigm. Finally, regulation will be a determinant factor; spectrum allocation (and use) policy must anticipate user demands and provide a fair framework for efficient spectrum usage, be it based on legacy FSA, futuristic DSA or online spectrum trading.

Chapter 10

Features in Future: 4G Visions From a Technical Perspective

Mobile communication is continuously one of the hottest areas that are developing at a booming speed, with advanced techniques emerging in all the fields of mobile and wireless communications. Current times are just the beginning for deploying 3G mobile communication systems, while research on the next generation of mobile communications, 4G wireless and mobile networks begin to pave the way for the future. This paper studies the visions of 4G from a technical perspective. After a brief review on the development history and status of mobile communications and related 4G perspectives, we present an overall 4G feature framework based on the kernel concept of integration, in which two key features (diversity and adaptability) of the three targets (terminals, networks, and applications) are described in detail. The concepts of both external and internal diversity of each target are defined to illustrate the causes and solutions of the adaptability feature. Then, along the entire 4G domain, each feature in the framework is deeply discussed from a technical standpoint, in which promising techniques and possible research issues for sufficient support of adaptability are also proposed. Finally, a short summary on 4G visions is presented as a continuum of features in the development of the mobile communications world.

10.1 Introduction

Mobile communications and wireless networks are developing at an astounding speed, with evidences of significant growth in the areas of mobile subscribers and terminals, mobile and wireless access networks, and mobile services and applications. The present time is just right to start the research of 4G mobile communications because of:

- Possibility, according to the historical indication of a generation revolution once a decade, and now we are near the end of 3G standardization phase and the beginning of 3G deployment.
- Necessity: according to 3G goals, 3G is necessary but not sufficient to the mobile communication strategy, in which many problems are only partly solved and there are still many problems left to be solved in the next generation, i.e. 4G.

There is plenty of related research on the next generation mobile communications [1-4], and the 4G topics are becoming hotter and hotter. However, most of the ongoing research can be classified into two different classes:

1) Many of the related 4G research focuses mainly on one specific technical area, such as distributed computing, mobile agents, multimedia services, or radio air interfaces, etc.

2) Some pieces of research are interested mainly in 4G scenarios from the standpoints of service provider or user, or a market analyst, from a less or non-technical viewpoint.

The difference of this paper to other related pieces of research is that we are going to present overall visions on the features of 4G mobile communications, based on a feature framework and provide detailed proposals to respective support techniques and research topics.

10.2 MOBILE COMMUNICATIONS REVIEW

The history and status of mobile communications are shortly listed in the following, together with the respective evaluations on the chief contributions.

1) Traditionally, wireless systems were considered as an auxiliary approach that was used in regions where it was difficult to build a connection by wireline.

2) 1G was based on analogy technique and deployed in the 1980s. It built the basic structure of mobile communications and solved many fundamental problems, e.g. cellular architecture adopting, multiplexing frequency band, roaming across domain, non-interrupted communication in mobile circumstances, etc. Speech chat was the only service of 1G.

3) 2G was based on digital signal processing techniques and regarded as a revolution from analogy to digital technology, which has gained tremendous success during 1990s with GSM as the representative. The utilization of SIM (Subscriber Identity Module) cards and support capabilities for a large number of users were 2Gs main contributions.

4) 2.5G extended the 2G with data service and packet switching methods, and it was regarded as 3G services for 2G networks. Under the same networks with 2G, 2.5G brought the Internet into mobile personal communications. This was a revolutionary concept leading to hybrid communications.

5) 3G is deploying a new system with new services instead of only providing higher data rate and broader bandwidth. Based on intelligent DSP techniques, various multimedia data communications services are transmitted by convergent 3G networks.

3G still leaves some unsolved problems that it does not concern or concerns only partly. The limitations and difficulties of 3G include:

TABLE 1 Mobile Communication History And Status

Property	1G	2G	2.5G	3G
Starting Time	1985	1992	1995	2002
Driven Technique	Analogue signal processing	Digital signal processing	Packet switching	Intelligent signal processing
Representative Standard	AMPS, TACS, NMT	GSM, TDMA	GPRS, I-Mode, HSCSD, EDGE	IMT-2000 (UMTS, WCDMA, CDMA2000)
Radio Frequency (HZ)	400M-800M	800M-900M, 1800M-1900M		2G
Bandwidth (bps)	2.4K-30K	9.6K-14.4K	171K-384K	2M-5M
Multi-address Technique	FDMA	TDMA, CDMA		CDMA
Cellular Coverage	Large area	Medium area		Small area
Core Networks	Telecom networks	Telecom networks		Telecom networks, Some IP networks
Service Type	Voice Mono-service Person-to-person	Voice, SMS Mono-media Person-to-person	Data service	Voice, Data Some Multimedia Person-to-machine

Figure 10.1:

1) Difficulty in continuously increasing bandwidth and high data rate to meet multimedia services requirements, together with the coexistence of different services needing different QoS and bandwidth.

2) Limitation of spectrum and its allocation.

3) Difficult to roam across distinct service environment in different frequency bands.

4) Lack of end-to-end seamless transport mechanism spanning a mobile sub-network and a fixed one.

The development trends of mobile communications can be summarized by the improvement of three aspects, including network area, e.g. data rate, bandwidth, and network capacity; mobility field, e.g. mobile spatial range, speed, coverage ability; and service property, e.g. services quantity, quality, cost, and category.

Table 1 summarizes the entire development of mobile communications with the properties of each generation including starting time, driven technique, representative standard, radio frequency, bandwidth, multi-address technique, cellular coverage, core networks, and service type. Note that it is a misunderstanding that either radio air interface or bandwidth is the criteria for the identification of different generations. Even they cannot be the representative characteristics for the representative generation.

10.3 4G FEATURES

10.3.1 4G Perspectives Review

Different 4G feature frameworks have been defined from the standpoints of service subscriber, service provider, researcher and engineer. In the following we give some representatives of 4G perspectives.

1) It is easy to say, based on the developing trends of mobile communication, that 4G will have broader bandwidth, higher data rate, smoother and quicker handoff, wider mobile area, more various service, lower cost, etc. Obviously these ideas do not make too much sense as such.

2) Other than the words more, any and/or all are preferred over expressions used by some others, e.g. anyone can communicate with anyone else, anywhere and anytime, or enjoy any service of any network operator, through any network of any network service provider. These sentences are truly attractive from a subscribers viewpoint, and they sound quite like advertisements or word games.

3) DoCoMo introduced the concept of MAGIC for the vision of 4G [5]: Mobile multimedia; Anytime, anywhere, anyone; Global mobility support; Integrated wireless solution; and Customized personal service, which mostly focused on public systems and treat 4G as the extension of 3G cellular service.

4) European Commission (EC) presented a perspective focusing on ensuring seamless service provisioning across a multitude of wireless systems and networks, and providing for optimum delivery via the most efficient network available. Further discussion did continuous promotion around 4G concepts [6-8], e.g. private systems and ad-hoc networks, optimal resource utilization, multiple radio interfaces, WLAN use, standards for interoperability, etc.

5) A broader, all-encompassing perspective of 4G was proposed in [4], according to which 4G will encompass all systems from public to private, operator-driven to ad-hoc, broadband to personal area and ad hoc networks, 2G systems to 3G systems. It focused mainly on personalized services.

It is amusing to see that it is quite easy for anyone to give a prediction on some 4G characteristics, whereas it is more difficult to provide an exhaustive description and sufficient investigations, especially on the support of advanced techniques.

10.3.2 4G Feature Framework

We summarize our proposal of 4G features with one sentence, or even more simply, with one word: integration, i.e. seamless integration of terminals, networks, and applications (together with users). A more detailed analysis and explanation of the definition is as follows.

1) The discussion domain includes three relevant targets, i.e. terminals, networks, and applications. Out of the 4G domain, the user is the only target.

2) The kernel word of the definition is so-called integration, which means the convergence of first the three different targets; second the various modes of each target, which lead to the feature of diversity.

3) The modifier seamless, which means the character and requirement of integration, implies the support of the adaptability feature between the three targets, each one of which is largely miscellaneous.

The 4G vision framework presented by us is illustrated in Fig.1. Note that networks are transparent to the user. In order to clarify the concept, we define two kinds of diversity: external diversity and internal diversity.

1) External diversity is outside the target, which brings along the demand of the adaptability feature to all targets. 2) Internal diversity is inside each of the targets, and it acts as the solution for adaptability requirements.

In short, the need for adaptability is caused by external diversity, and it is solved by internal diversity. Here both the external and internal diversity of users are the cause of all adaptability requirements, which implies that the user is out of the technical domain of 4G visions. The two main features, i.e. diversity and adaptability of the three targets terminal, network, and application are described in detail in the next section.

10.3.3 4G Feature Description

1) User Diversity: The external diversity of users, i.e. people in different situations, includes e.g. culture, educational background, economic capability, physical property, personal preference, etc. The internal diversity of users, i.e. people with different interfaces, include e.g. vision, hearing, speech, touch sense, hands and fingers, body, etc. Note that as for users, both their external and internal diversity are to be adapted by the other two targets: terminal and application. Moreover, for adapting the two kinds of user diversity, both the external and internal diversity of terminals and ap-

plications are the solution.

2) Terminal Diversity and Adaptability: The terminals external diversities are the differences of terminals in both static and mobile attributes. Static attributes include e.g. functionality, weight, size, battery life, human interface, antenna, processing capability, security, style, and cost. Mobile attributes include dynamic attributes of both temporal and spatial features. The former category contains e.g. moving speed and acceleration, along with stationary, pedestrian or vehicular qualities, while the latter is connected to spatial range, e.g. indoors, on-campus, in urban and rural environments, and also direction. The internal diversity of terminals means that one terminal may integrate multiple functions, modes, interfaces, flexibilities, etc. There are three targets for terminal adaptability. For users, it includes the provision of different terminals to satisfy different users and an individual users various requirements. As for applications, we hope that miscellaneous services can be delivered to one single terminal. When networks are concerned, a single terminal can reach a wide range of networks despite of location and mobile rate.

3) Network Diversity and Adaptability: The external diversity of networks is obvious. Internet is assorted by nature, while wireless networks keep the same property. For instance air interfaces can integrate all kinds of standards and work on different frequencies. Moreover, multiple operators deploy networks with multiple standards and protocols. The internal diversity of networks means that one network can interconnect with other different networks and transfer various kinds of loads, e.g. cellular systems with various coverage.

Three targets are related to network adaptability. In reference to terminals, network adaptability aims to make multiform mobile devices with a wide range of moving speeds and mobile areas connectable to wireless networks. For applications, there is a requirement that any type and/or quality of service can be delivered through diverse fixed and mobile networks in the most suitable and efficient way. The target for networks themselves is to make it easy to build a new network or remove an old one, and to make interoperability with ones neighbours seamless despite its heterogeneous nature.

4) Application Diversity and Adaptability: The external diversity of applications will be a reasonable property, and this need not mean that 4G services and applications must be multifarious, in all the aspects of quantity, quality, and type. With internal diversity we mean that one application can be tailored into e.g. multiple levels of quality, various styles, and different kinds of release shape, etc.

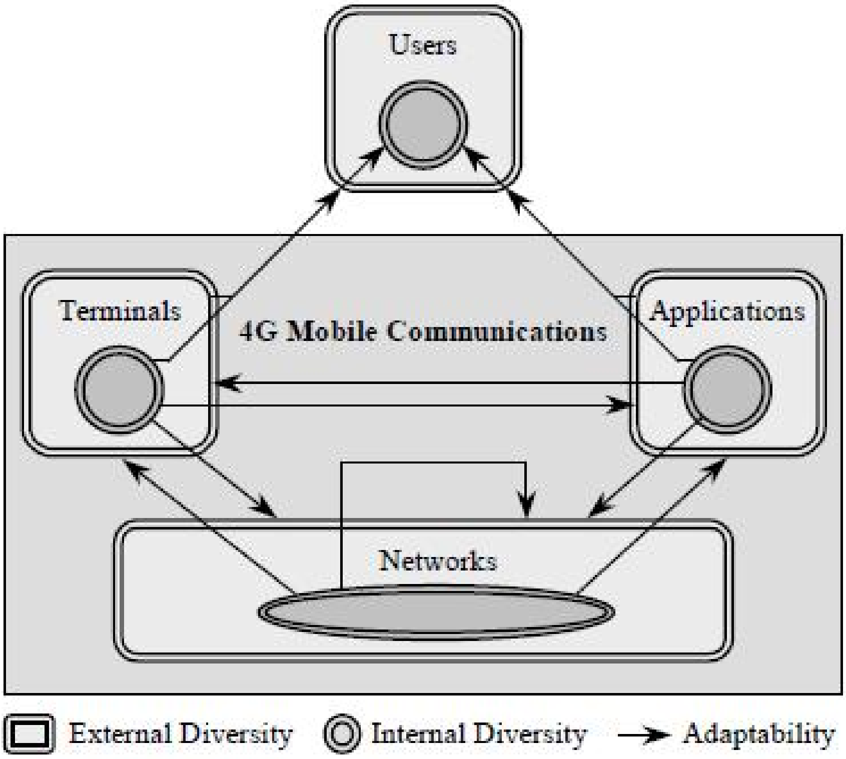


Fig. 1. 4G feature framework

Figure 10.2:

Application adaptability is a main feature of 4G services. To users, this means that services can be delivered automatically according to personal preferences of different users. In view of terminals, we hope that various terminals are able to run one application with different formats, such as e-mail in text message, voice, image, or even video. In connection with networks, applications can be transformed into various forms and levels in order to be transmitted correctly and efficiently.

10.4 4G TECHNICAL PERSPECTIVE

It is obvious that 4G, just like all the previous generations, is driven not only by technology, but also by market requirements. This section mainly discusses, from a more technical perspective, possible topics for research and promising techniques of 4G, and focuses mainly on those techniques that give support to the main feature of adaptability by internal diversity of targets in the 4G domain. Fig. 2 gives an illustration of the discussion domain of 4G.

10.4.1 Terminals

In order to adapt to the diverse applications and networks, together with the various requirements of users, the terminal domain must possess both internal and external diversity. Support techniques of the field may include the following:

1) User interfaces of terminals vary from traditional keyboard, display, and tablet, to new interfaces based on speech, touch, vision, soft buttons, etc. This will be common at a time when one terminal has multiple user interfaces.

2) Adaptive techniques such as smart antennas, software radio, and smart transceivers, enhance interoperability through simultaneous support of several radio interfaces in a single terminal. This makes a terminal roamable across any air interface standard and connectable to any wireless access point by exchanging configuration software. These approaches can also be used on wireless access points as an advanced smart base station.

3) Terminals will be aware of location and context, often based on some wireless low power sensors that are humansensitive and/or environment-sensitive in order to monitor and interact with the physical world to report the human and/or environmental factors. The advances in this area have been used in e.g. wearable computers as a novel terminal type.

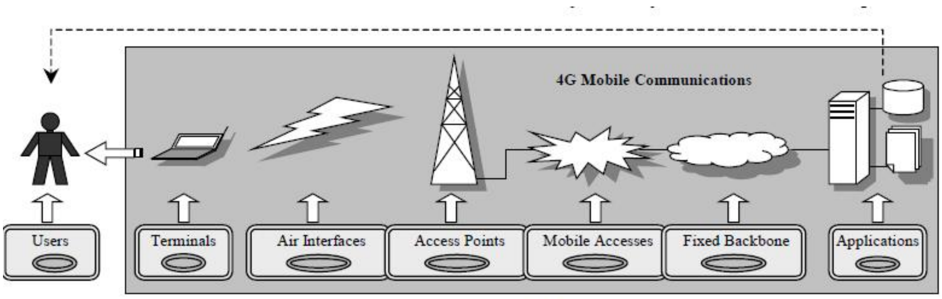


Fig. 2. 4G visions in domains

Figure 10.3:

4) An intelligent terminal is able to dynamically improve its processing capability in order to contain various services. Some function modules can even be downloaded to a terminal when needed.

10.4.2 Networks

More advances in networks are needed to keep pace with the rapidly changing terminals and applications, as follows:

1) Smart antenna, software radio, together with advanced base station are the key techniques to achieve adaptability of wireless access points to diverse terminals, i.e. to make radio systems and air networks re-configurable.

2) Hierarchical and ubiquitous as well as overlay cellular systems, including picocell, microcell, macrocell, and magecell ones, implement seamless network interconnection of both symmetric and asymmetric nature, and seamless terminal handoff of both horizontal and vertical levels respectively.

3) Network layer hierarchical mobility management based on Mobile IPv6 and Cellular IP brings quick and seamless handoff to terminals. The Mobile IPv6 also presents a great contribution to the adaptability of heterogeneous networks.

4) Ad hoc wireless networks are a kind of self-deployed wireless networks to make networks portable and adaptable, and thus dynamically share unlicensed radio spectrum.

5) Network reconfiguration can be obtained by the reconfiguration of protocol stacks and programmability of network nodes. Thus, it can adapt dynamically to the changing channel conditions and low or high data rate

users.

6) Miscellaneous services can be delivered through a mixture of transmission networks including unicast, multicast, and broadcast ones. According to the service types, e.g. real-time attribute, importance, bandwidth demand, or data stream type, multiple levels of QoS can be defined for various services.

7) Network resource can be dynamically allocated to cope with varying traffic load, channel condition, and service environment. Traffic conditions will be dynamically monitored and controlled via techniques such as distributed and decentralized control of network functionalities.

10.5 Applications

Adaptability will be one of the basic requirements to the development and delivery of new mobile services. Promising techniques and possible topics may include:

1) Mobile application should refer to a users profile so that it can be delivered in a way most preferred by the subscriber, such as context-based personalized services. This also brings the applications with adaptability to terminals that are moving in varying locations and speeds. Micro-sensors and GPS receivers are the main driven techniques.

2) Techniques such as adaptive multimedia and unified messaging take the terminal characteristics into account and ensure that the service can be received and run on a terminal with the most suitable form to the host type.

3) Intelligent mobile software agent is a common technique to all of the three targets, which act as a platform for service development, delivery, and auto-configuration.

4) Applications can negotiate with networks so that they can be transferred with the most efficient channel, e.g. indoor networks or WLAN or cellular systems in a wide area. Services will be tailorable in order to fit the different network environments and the varying traffic conditions.

5) Services and applications can also be smoothly delivered across a multiple domain of operators and service providers.

TABLE 2 4G VISIONS SUMMARY

Property	4G
Starting Time	2010-2012
Driven Technique	Intelligent software Auto configuration
Representative Standard	OFDM, UWB
Radio Frequency (HZ)	3G-5G
Bandwidth (bps)	10M-20M
Multi-address Technique	FDMA, TDMA, CDMA
Cellular coverage	Mini area
Core networks	All-IP networks
Service type	Multimedia Machine-to-machine

Figure 10.4:

10.6 CONCLUSION

This chapter presents 4G visions from a technical perspective. After a brief review of the history and status of mobile communications, we propose a 4G feature framework, in which features of 4G mobile communications are defined. The framework is based on the key concept of integration, and it has the following characteristics:

1) Targets in the framework include users, terminals, networks, and applications, which compass the entire technical domain and operating environment of 4G.

2) Core features of 4G are described as diversity and adaptability of the targets, leading to seamless integration.

3) The feature of diversity includes both external and internal diversity, in which adaptability is caused by external diversity and is solved by internal diversity. Technical perspectives are presented for each of the features in the paper, in which also some promising techniques and possible research issues of 4G are introduced. The proposed framework provides a layout view on future communication systems, and challenging research topics are figured for guiding systematic research of 4G.

Chapter 11

Scenarios And Limitations

11.1 Scenarios

11.1.1 Scenario Outline

A key feature of 4G is likely to be the availability of significantly higher data rates than for third-generation (3G) systems. It has been suggested that data rates up to 100 Mbps for high mobility and 1 Gbps for low mobility should be the target value. These data rates suggest higher spectral efficiencies and lower cost per bit will be key requirements for such future systems. Additional important and expected features are likely to be increased flexibility of mobile terminals and networks, multimedia services, and high-speed data connections. Future convergence systems will clearly be another feature. Based on these visions and characteristics of the 4th generation (4G) for future wireless telecommunication, new spectrum allocation issue, and technology feasibility, the advent of 4G service will bring a number of changes of competition environment, regulation and policy as well as service change into future wireless communication. Accordingly, it is very important we expect what kinds of possibility we have for the 4G service to prepare well. Several scenarios are described to display the situations of the wireless communication industry as the 4G. These scenarios are based on different wireless access technologies such as WiMAX, WiBro, 3G LTE, and IEEE802.20.

In the on-going 4G studies in the standardization bodies and relative industries, one of the aims is to establish an integral wireless system that would seamlessly connect the enhanced forms of existing 3G wireless systems such as WCDMA with HSDPA. In this scenario, existing carriers will maintain present customer base and services are integrated 4G. On the other hand, however, it has become possible by technological innovation that non-3G wireless services develop as competitors against 3G services such as WiMAX, or further enhanced IEEE 802 standards. In addition, individuals and organizations have started providing open and free wireless communication services by opening up, through various technologies. Figure 2 shows that these different evolution path toward 4G. For the scenarios we provided in the paper, we assumed that the advent of 4G service will be after 2012. 4G service will be determined whether it can support 4G characteristics technically, and hold the market with service differentiation from competitors. In order to forecast the form of realization of 4G systems, we construct four scenarios in the paper.

11.1.2 Market Trends

In consideration of the market situation, we review market acceptance and adoption of wireless data and deployments of 3GPP and WiMAX networks around the world. 3GPP LTE and WiMAX technologies encompass a huge range of evolving capability, but how well do these technologies actually address market needs. Basically, 3G operators have shown less interest in mobile WiMAX and are more interested in upgrading their own networks that would enable them to compete with WiMAX.

In Korea, launching of WiBro has not been spectacular. What was supposed to become a flagship for Korea's WiMAX technology can still be considered a pilot project. In July 2006, KT launched commercial service in Seoul and surrounding cities where it had provided pilot service. Both KT and SKT offer only the PTC'07 Proceedings Samsung PCMCIA card for the service. However, it is expected that WiBro adoption in Korea will be rather gradual and modest compared to initial estimations.

Currently, UMTS is commercially launched by 107 operators in 50 countries with 79 more networks planned or in deployment. After Cingular Wireless launched the world's first wide scale UMTS/HSDPA network in December 2005, 103 operators have announced their HSDPA deployment plans. It is expected that nearly all UMTS operators will deploy HSDPA, essentially a simple upgrade to the existing system. Furthermore, this upgrade can be continued to 3GPP LTE in a few years.

Currently, UMTS is commercially launched by 107 operators in 50 countries with 79 more networks planned or in deployment. After Cingular Wireless launched the world's first wide scale UMTS/HSDPA network in December 2005, 103 operators have announced their HSDPA deployment plans. It is expected that nearly all UMTS operators will deploy HSDPA, essentially a simple upgrade to the existing system. Furthermore, this upgrade can be continued to 3GPP LTE in a few years.

Although there are still a lot of things to be considered, according to these market changes, to meet needs of consumer demands in a right time, time-to market is also critical factor.

Consumer Demands: New World of Content and Applications

- Deliver Content on My Time and My Location
- Seamless User Experience across all Devices - PCs, portables, iPODs, Smartphones, In-car, At-work
- Customized Content and Services on Demand - Movie Downloads, Live TV, Personalized Communication
- Time/Place Shifting of Content and Applications is demanded and expected

Broadband Enablers Everywhere

- Fixed and WiFI BroadBand at home and at work
- MIMO WiMAX enabling Wide Area, Nationwide Mobility
- Ubiquitous IP and Open Interfaces for Plug and Play
- Ubiquitous IP and Open Interfaces for Plug and Play

11.2 Swot Analysis - 4G

Considering 4G characteristics, expected scenarios and market trends, we can find out strengths, weaknesses, opportunities and threats of 4G with better understandings. The lists and findings follow.

Strengths in 4G

- 4G visions take into account installed base and past investments
- Strong position of telecommunications vendors expected in the marketplace
- Faster data transmission and higher bit rate and bandwidth, allow more business applications and commercialization
- Has advantage for personalized multimedia communication tools

Weakness in 4G

- No large user community for advanced mobile data applications yet
- Growing divergence between telecommunications vendors and operators
- Not possible to offer full internet experience due to limited speed and bandwidth
- Comparatively higher cost to use and deploy infrastructure compared fast mobile generation

Opportunities in 4G

- Evolutionary approach may yield opportunities for the 4G
- Emphasis on heterogeneous networks capitalizes on past investments
- Strategic alliance and coalition opportunities with traditional nontelecommunication industries
- Sophisticated and mature commercialization of 4G technology would encourage more applications of e-commerce and m-commerce

- Worldwide economy recover stimulates consumption and consumer confidence, therefore bring in opportunities for telecommunication sections
- It is expected and predicted that consumers will continue to replace handsets with newer technology at a fast rate
- Desirable higher data capacity rates, the growth opportunity for 4G is very bright and hopeful

Threats in 4G

- Faster rate of growth and developments in other region
- Since 3G mobile is still in the market, it squeezes the market competition in the mobile industry

11.3 Limitations Of 4G

Although the concept of 4G communications shows much promise, there are still limitations that must be addressed.

11.3.1 Operating Area

One major limitation of 4G is operating area.Although 2G networks are becoming more ubiquitous, there are still many areas not served. Rural areas and many buildings in metropolitan areas are not being served well by existing wireless networks. This limitation of today's networks will carry over into future generations of wireless systems. The hype that is being created by 3G networks is giving the general public unrealistic expectations of always on, always available, anywhere, anytime communications. The public must realize that although high-speed data communications will be delivered, it will not be equivalent to the wired Internet - at least not at first. If measures are not taken now to correct perception issues, when 3G and later 4G services are deployed, there may be a great deal of disappointment associated with the deployment of the technology, and perceptions could become negative. If this were to happen, neither 3G nor 4G may realize its full potential.

11.3.2 Cost

Another limitation is cost. The equipment required to implement a next generation network is still very expensive. Carriers and providers have to plan carefully to make sure that expenses are kept realistic. One technique currently being implemented in Asian networks is a Pay-Per-Use model of services. This model will be difficult to implement in the United States, where the public is used to a service-for-free model(e.g., the Internet).

11.3.3 The big makers

Most of the big manufacturers, including Alcatel and Samsung, have also showed interest in 4G technology although not so openly. In August 2004, Samsung organised a 4G Forum, which was attended by 120 mobile communications representatives from 18 countries.

However, 4G has still a long way to go as far as regulations and standardisation are concerned. For example, the radio spectrum that 4G will use will not be decided until 2007.

11.3.4 Technical challenges

Some of the technical challenges posed by 4G are daunting. Apart from all the intrinsic difficulties networks have, they must cope with others: the signal diminishes when the terminal moves more than a kilometre away from the base station; there is great difficult in moving huge amounts of data within a limited area of the spectrum; the mobile prototypes are too big and their energy consumption too high.

All in all, the greatest challenge may be finding customers. Yankee Group says that at present there are no applications that require such high speeds although this may not be a real problem for DoMoCo in Japan where they have cornered the market with their services. This is not so in other countries.

This would explain the major differences foreseen in the commercialisation of 4G: three to five years in Asia and ten to fifteen years in Europe and the United States.

Bibliography

Citations

1. http://www.mediateam.oulu.fi/publications/pdf/86.pdf.
2. http://www.di.uoa.gr/ gazis/data/2003-anwire-workshop-mykonos.pdf
3. a b c Young Kyun, Kim; Prasad, Ramjee. 4G Roadmap and Emerging Communication Technologies.Artech House, pp 12-13. ISBN 1-58053-931-9.
4. 4G - Beyond 2.5G and 3G Wireless Networks. MobileInfo.com. Retrieved on 2007-03-26.
5. Jawad Ibrahim (December 2002). 4G Features. Bechtel Telecommunications Technical Journal. Retrieved on 2007-03-26.
6. Mobility Management Challenges and Issues in 4G Heterogeneous Networks. ACM Proceedings of the first international conference on Integrated internet ad hoc and sensor networks (May 30 - 31, 2006). Retrieved on 2007-03-26.
7. a b c Werner Mohr (2002). Mobile Communications Beyond 3G in the Global Context. Siemens mobile. Retrieved on 2007-03-26.
8. Noah Schmitz (March 2005). The Path To 4G Will Take Many Turns. Wireless Systems Design. Retrieved on 2007-03-26.
9. WINNER - Towards Ubiquitous Wireless Access. WINNER (2007).
10. WINNER II - Public Deliverable. WINNER II (2006-07).
11. DoCoMo Achieves 5 Gbit/s Data Speed. NTT DoCoMo Press (9 February 2007).
12. Press Release: Digiweb Mobile Takes 088
13. RT News article: Ireland gets new mobile phone provider
14. Privatkunden Tarife (de). Deutsche Breitband Dienste. Retrieved on 2007-08-30.
15. WiMAX Day (March 8th, 2007). WiMAX rallies market as Clearwire IPO nets $600 million. WiMAX Spectrum Owners Alliance (WiSOA).
16. WiMAX Has Company. Wireless Week (1 February 2006). Retrieved on 2007-03-26.

Additional Resources

1.Brian Woerner (June 20-22, 2001). "Research Directions for Fourth Generation Wireless". Proceedings of the 10th International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE '01). (118kb) Suk Yu Hui; Kai Hau Yeung (December 2003). "Challenges in the migration to 4G mobile systems". Communications Magazine, IEEE.
2.4G Mobile. Alcatel-Lucent (June 13, 2005).
3.Will Knight (02 September 2005). 4G prototype testing. New Scientist.
4.Caribbean telecoms to invest in 4G wireless networks. Caribbean Net News (June 27 2006).