WILEY

TIMELY. PRACTICAL. RELIABLE.

3G Handset and Network Design

Geoff Varrall
Roger Belcher

3G Handset and Network Design

3G Handset and Network Design

Geoff Varrall
Roger Belcher

WILEY
Wiley Publishing, Inc.

Publisher: Bob Ipsen
Editor: Carol A. Long
Developmental Editor: Kathryn A. Malm
Managing Editor: Micheline Frederick
Text Design & Composition: Wiley Composition Services

This book is printed on acid-free paper. ♾

Published by Wiley Publishing, Inc., Indianapolis, Indiana
Published simultaneously in Canada

ISBN: 0-471-22936-9

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

I would like to dedicate my contribution to this book to my father Edgar for his gift of curiosity; to my mother Winifred for her gift of confidence; and to my wife Deborah for her gift of our marriage.

Roger Belcher

Contents

Acknowledgments

This book is the product of over 15 years of working with RTT, delivering strategic technology design programs for the cellular design community. This has included programs on AMPS/ETACS handset, base station, and network design in the early to mid-1980s; programs on GSM handset, base station, and network design from the late 1980s to mid-1990s onward; and, more recently, programs on 3G handset, Node B, and network design.

We would like to thank the many thousands of delegates who have attended these programs in Europe, the United States, and Asia and who have pointed out the many misconceptions that invariably creep in to the study of a complex subject.

We would also like to thank our other colleagues in RTT: Dr. Andrew Bateman for keeping us in line on matters of DSP performance and design issues; Miss Tay Siew Luan of Strategic Advancement, Singapore, for providing us with an Asian technology perspective; our valued colleagues from the Shosteck Group, Dr. Herschel Shosteck, Jane Zweig, and Rich Luhr, for providing us with valuable insights on U.S. technology and market positioning; our colleague, Adrian Sheen, for keeping our marketing alive while we were knee-deep in the book; and last but not least, Lorraine Gannon for her heroic work on the typescript.

Also thanks to our families for putting up with several months of undeserved distraction.

Any errors which still reside in the script are entirely our own, so as with all technical books, approach with circumspection.

We hope you enjoy the complexity of the subject, challenge our assumptions, find our mistakes (do tell us about them by emailing geoff@rttonline.com or roger@rttonline.com), and get to the end of the book intrigued by the potential of technology to unlock commercial advantage.

Geoff Varrall and Roger Belcher

Introduction

This book is written for hardware and software engineers presently involved or wanting to be involved in 3G handset or 3G network design. Over the next 20 chapters, we study handset hardware, handset software, network hardware, and network software.

A Brief Overview of the Technology

Each successive generation of cellular technology has been based on a new enabling technology. By *new*, we often mean the availability of an existing technology at low cost, or, for handset designers, the availability of a technology sufficiently power-efficient to be used in a portable device. For example:

First generation (1G). AMPS/ETACS handsets in the 1980s required low-cost microcontrollers to manage the allocation of multiple RF (radio frequency) channels (833 × 30 kHz channels for AMPS, 1000 × 25 kHz channels for ETACS) and low-cost RF components that could provide acceptable performance at 800/900 MHz.

Second generation (2G). GSM, TDMA, and CDMA handsets in the 1990s required low-cost digital signal processors (DSPs) for voice codecs and related baseband processing tasks, and low-cost RF components that could provide acceptable performance at 800/900 MHz, 1800 MHz, and 1900 MHz.

Third generation (3G). W-CDMA and CDMA2000 handsets require—in addition to low-cost microcontrollers and DSPs—low-cost, low power budget CMOS or CCD image sensors; low-cost, low power budget image and video encoders; low-cost, low power budget memory; low-cost RF components that can provide acceptable performance at 1900/2100 MHz; and high-density battery technologies.

Bandwidth Quantity and Quality

Over the next few chapters we analyze bandwidth quantity and quality. We show how application bandwidth quality has to be preserved as we move complex content (rich media) into and through a complex network. We identify how bandwidth quality can be measured, managed, and used as the foundation for quality-based billing methodologies. We show how the dynamic range available to us at the application layer will change over the next 3 to 5 years and how this will influence radio bandwidth and network topology.

We define *bandwidth quality* in terms of application bandwidth, processor bandwidth, memory bandwidth, radio bandwidth, and network bandwidth, and then we identify what we need to do to deliver consistently good end-to-end performance.

Hardware Components

Hardware components are divided into physical hardware and application hardware, as follows:

Physical hardware. The hardware needed to support the radio physical layer—putting 0s and 1s on to a radio carrier, and getting 0s and 1s off a radio carrier

Application hardware. The hardware needed to capture subscriber content (microphones, vocoders, imaging, and video encoders) and to display content (speakers, displays, and display drivers)

A typical 3G handset includes a microphone (audio capture); CMOS imager and MPEG-4 encoder (for image and video encoding); a keyboard (application capture); a smart card for establishing access and policy rights; and, on the receive side, a speaker, display driver, and display. The addition of these hardware components (CMOS imager, MPEG-4 encoder, and high-definition color display) changes what a user can do and what a user expects from the device and from the network to which the device is connected.

Software Components

Software footprint and software functionality is a product of memory bandwidth (code and application storage space), processor bandwidth (the speed at which instructions can be processed), and code bandwidth (number of lines of code). Over the past three generations of cellular phone, memory bandwidth has increased from a few kilobytes to a few Megabytes to a few Gigabytes. Processor bandwidth has increased from 10 MIPS (millions of instructions per second) to 100 MIPS to 1000 MIPS, and code bandwidth has increased from 10,000 to 100,000 to 1,000,000 lines of code (using the StarCore SC140 as a recent example).

The composition of the code in a 3G handset determines how a 3G network is used. Software form factor and functionality determine application form factor and functionality.

Software components can be divided into those that address physical layer functionality and those that address application layer functionality, as follows:

Physical layer software. Manages the Medium Access Control (MAC) layer—the allocation and access to radio and network bandwidth.

Application layer software. Manages the multiple inputs coming from the handset application hardware (microphone, vocoder, encoder) and the media multiplex being delivered on the downlink (network to handset).

Rich Media Properties

It is generally assumed that an application may consist of a number of traffic streams simultaneously encoded onto multiple channel streams. These components are often referred to as *rich media*.

The properties of these rich media components need to be preserved as they move across the radio interface and into and through the core network. By *properties* we mean voice quality (audio fidelity), image and video quality, and data/application integrity.

Properties represent value, and it is the job of a 3G handset and network designer to ensure an end-to-end Quality of Service that preserves this property value.

How This Book Is Organized

The deliberate aim of this book is to combine detail (the small picture) with an overview of how all the many parts of a 3G network fit, or should fit, together (the big picture). In meeting this aim, the content of this book is arranged in four parts of five chapters each, as follows:

Part I: 3G Hardware. We look at the practical nuts and bolts of cellular handset design, how band allocations and regulatory requirements determine RF performance, the processing needed to capture signals from the real world (analog voice and analog image and video), and the processing needed to translate these signals into the digital domain for modulation onto a radio carrier. We discuss the different requirements for RF processing and baseband processing: How we manage and manipulate complex content to deliver a consistent end-to-end user experience. In the following chapters we introduce the various concepts related to bandwidth quality: How we achieve consistent performance over the radio physical layer.

- Chapter 1 reviews some of the design challenges created by the spectral allocation process.
- Chapter 2 shows that making products do something they were not designed to do often leads to a disappointing outcome (as shown in a case study of GPRS/EDGE handset hardware).
- Chapter 3 highlights the hardware requirements of a 3G handset design—how we get a signal from the front end to the back end of the phone and from the back end to the front end of the phone.

- Chapter 4 analyzes how the additional hardware items in a handset—image capture platform, MPEG-4 encoder, color display—influence network offered traffic.
- Chapter 5 reviews some issues of handset hardware configurability.

Part II: 3G Handset Software. We explore how handset software is evolving and the important part handset software plays in shaping offered traffic and building traffic value.

- Chapter 6 case studies application software—what is possible now and what will be possible in the future.
- Chapter 7 analyzes source coding techniques.
- Chapters 8 and 9 begin to explore how we build session value by providing differentiated service quality and differentiated access rights.
- Chapter 10 complements Chapter 5 by looking at software configurability and future handset software trends.

Part III: 3G Network Hardware. We launch into network hardware, returning to the nuts and bolts.

- Chapter 11 reviews some of the design challenges introduced by the spectral allocation process, in particular, the design challenges implicit in delivering efficient, effective base station/Node B hardware.
- Chapter 12 looks at some of the present and future network components—what they do, what they don't do, and what they're supposed to do.
- Chapter 13 covers base station/Node B antennas and other link gain products, including high-performance filters, RF over fiber, and optical transport.
- Chapter 14 talks us through the dimensioning of bursty bandwidth—how we determine the properties of offered traffic in a 3G network.
- Chapter 15 evaluates the particular requirements for broadband fixed access and some of the hardware requirements for media delivery networks.

Part IV: 3G Network Software. We address network software—the implications of managing audio, image, video, and application streaming; the denomination and delivery of differentiated Quality of Service; and related measurement and management issues.

- Chapter 16 analyzes end-user performance expectations, how expectations increase over time, and the impact this has on network software.
- Chapter 17 reviews traffic shaping protocols and the performance issues implicit in using Internet protocols to manage complex time-dependent traffic streams.
- Chapter 18 follows on, hopefully logically, with an explanation of the merits/demerits of Service Level Agreements when applied in a wireless IP network.

- Chapter 19 explores some of the practical consequences of 3G cellular and 3G TV software integration.
- Chapter 20 reviews, as a grand finale, storage bandwidth and storage area network technologies.

The Objective: To Be Objective

We could describe some parts of this book as "on piste," others as "off piste." The on piste parts describe what is—the present status of handset and network hardware and software. Other parts set out to describe what will be. From experience, we know that when authors speculate about the future, the result can be intensely irritating. We argue, however, that you do not need to speculate about the future. We can take an objective view of the future based on a detailed analysis of the present and the past, starting with an analysis of device level evolution.

Predicting Device Level Evolution

Device hardware is becoming more flexible—microcontrollers, DSPs, memory, and RF components are all becoming more adaptable, capable of undertaking a wide range of tasks. As device hardware becomes more flexible, it also becomes more complex. Adding smart antennas to a base station is an example of the evolution of hardware to become more flexible—and, in the process, more complex.

As handset hardware becomes more complex, it becomes more capable in terms of its ability to capture complex content. Our first chapters describe how handset hardware is evolving—for example, with the integration of digital CMOS imaging and MPEG-4 encoding. As handset hardware becomes more complex, the traffic mix shifts, becoming more complex as well. As the offered traffic mix (uplink traffic) becomes more complex, its burstiness increases. As bandwidth becomes burstier, network hardware has to become more complex. This is described in the third part of the book.

As handset and network hardware increases in complexity, software complexity increases. We have to control the output from the CMOS imager and MPEG-4 encoder, and we have to preserve the value of the captured content as the content is moved into and through our complex network. As hardware flexibility increases, software flexibility has to increase.

Fortunately, device development is very easy to predict. We know by looking at process capability what will be possible (and economic) in 3 to 5 years' time. We can very accurately guess what the future architecture of devices such as microcontrollers, DSPs, memory, and RF components will be in 3 to 5 years' time. These devices are the fundamental building blocks of a 3G network.

By studying device footprints, we know what will happen at the system and network level over the next 5 years. We do not need to sit in a room and speculate about the future; the future is already prescribed. That's our justification for including the "what will be" parts in this book. If we offer an opinion, we hope and intend that those opinions are objective rather than subjective.

Bridging the Reality Gap

Too often we fail to learn from lessons of the past. As an industry, we have over 20 years of experience in designing cellular handsets and deploying cellular networks. The past tells us precisely what is and what is not possible in terms of future technology deployment. This allows us to detect when reality gaps occur. Reality gaps are those between technical practicality and wishful thinking. They happen all the time and can be particularly painful when technically complex systems are being deployed.

Almost all technologies start with a reality gap. The technology fails to deliver as well as expected. Some technologies never close the gap and become failed technologies. Some people can make money from failed technologies, but the majority doesn't. Failed technologies ultimately fail because they do not deliver user value.

We also tend to forget that user expectations and customer expectations change over time. A technology has to be capable of sufficient dynamic range to be able to continue to improve as the technology and user expectations mature. Failed technologies often fail because they cannot close the reality gap and cannot catch up with changing user expectations.

Successful technologies are technologies that deliver along the whole industry value chain—device vendors, handset manufacturers, network manufacturers (software and hardware vendors), network operators, and end users.

We aim to show how 3G technology is evolving to become a successful proposition, both technically and commercially. We hope you enjoy and profit from the next 20 chapters.

Before We Start: A Note about Terms

In this book we use the term *handset* to describe a generic, nonspecific portable cellular terminal. When we use the term *mobile*, we are referring to a portable terminal of higher power and capable of traveling at high speed. It is usually vehicle-mounted and may have antenna gain.

In discussing 1G and 2G cellular systems, we use the term *base station* or BTS (base transceiver system). In 3G cellular systems, we refer to this as the Node B. *Node* refers to the assumption that the base station will act as a node supporting Internet protocols. B refers to the fact the node is integrated with a base station. The RNC (radio network controller) is the network subcomponent used in a 3G network for load distribution and access policy control. It replaces the BSC (base station controller) used in 1G and 2G cellular networks.

PART

One

3G Hardware

CHAPTER

1

Spectral Allocations—Impact on Handset Hardware Design

In this first chapter we explain the characteristics of the radio spectrum, how over the past 100 years enabling component technologies have provided us with access to progressively higher frequencies, and how this in turn has increased the amount of RF (radio frequency) bandwidth available. We show how enabling component technologies initially provided us with the ability to deliver increasingly narrow RF channel spacing in parallel with the introduction of digital encoding and digital modulation techniques. We explain the shift, from the 1980s onward, toward wider RF channel spacing through the use of TDMA (Time Division Multiple Access) and CDMA (Code Division Multiple Access) multiplexing techniques and identify benefits in terms of component cost reduction and performance gain, in particular the impact of translating tasks such as selectivity, sensitivity, and stability from RF to baseband.

Setting the Stage

By *baseband*, we mean the original information rate. For analog voice, baseband would be used to refer to the 3 kHz of audio bandwidth. This would then be preprocessed. Preemphasis/de-emphasis would be used to tailor the high-frequency response and reduce high-frequency noise. *Companding* (compression/expansion) would be used to compress the dynamic range of the signal. The signal would then be modulated onto an RF carrier using amplitude or frequency modulation. Usually, an intermediate step between baseband and RF would be used, known as the *IF processing stage* (intermediate frequency). We still use IF processing today and will discuss its merits/demerits in a later section.

In a 2G handset, baseband refers to the information rate of the encoder (for example, 13 kbps) and related digital signaling bandwidth. The data is then channel coded—that is, additional bits are added to provide error protection—and then the data is modulated onto an RF carrier, usually with an IF processing stage. In a 3G handset, baseband refers to the information rate of the vocoder, parallel image and video encoder rates, other data inputs, and related channel coding.

First-generation handsets therefore have a baseband running at a few kilohertz, and second-generation handsets a few tens of kilohertz.

Third-generation handsets have a user data rate that can vary between a few kilohertz and, in the longer term, several megahertz. The user data is channel coded and then spread using a variable spreading code to a constant baseband rate known as the *chip rate*—for example, 1.2288 Mcps (million chips per second; a clock rate of 1.2288 MHz) or 3.84 Mcps (a clock rate of 3.84 MHz). This baseband data, after spreading, has to be modulated onto an RF carrier (producing a 1.25 or 5 MHz bandwidth), sometimes via an IF. The RF will be running at 1900/2100 MHz.

Essentially, the higher the frequency, the more expensive it is to process a signal. The more we can do at baseband, the lower the cost. This is not to downplay the importance of the RF link. The way in which we use the RF bandwidth and RF power available to us has a direct impact on end-to-end quality of service.

Ever since the early experiments of Hughes and Hertz in the 1880s, we have searched for progressively more efficient means of moving information through free space using electromagnetic propagation. By *efficiency* we mean the ability to send and receive a relatively large amount of information across a relatively small amount of radio bandwidth using a relatively small amount of RF power generated by a relatively power-efficient amplifier in a relatively short period of time.

The spark transmitters used to send the first long-distance (trans-Atlantic) radio transmissions in the early 1900s were effective but not efficient either in terms of their use of bandwidth or the efficiency with which the RF power was produced and applied. What was needed was an enabling technology.

Thermionic and triode valves introduced in the early 1900s made possible the application of tuned circuits, the basis for channelized frequencies giving long-distance (and relatively) low-power communication. Tuned circuits reduced the amount of RF power needed in a transceiver and provided the technology needed for portable Morse code transceivers in World War I.

Efficiency in RF communication requires three performance parameters:

Sensitivity. The ability to process a low-level signal in the presence of noise and/or distortion

Selectivity. The ability to recover wanted signals in the presence of unwanted signals

Stability. The ability to stay within defined parameters (for example, frequency and power) under all operating conditions when transmitting and receiving

The higher the frequency, the harder it is to maintain these performance parameters. For example, at higher frequencies it becomes progressively harder to deliver gain—that is, providing a large signal from a small signal—without introducing noise. The gain becomes more expensive in terms of the input power needed for a given output transmission power. It becomes harder to deliver receive sensitivity, because of front-end

noise, and to deliver receive selectivity, due to filter performance. On the other hand, as we move to higher frequencies, we have access to more bandwidth..

For example, we have only 370 kHz of bandwidth available at long wave; we have 270 GHz available in the millimetric band (30 to 300 GHz). Also, as frequency increases, range decreases. (Propagation loss increases with frequency). This is good news and bad news. A good VHF transceiver—for example, at 150 MHz—can transmit to a base station 40 or 50 kilometers away, but this means that very little frequency reuse is available. In a 900 MHz cellular network, frequencies can be used within (relatively) close proximity. In a millimetric network, at 60 GHz, attenuation is 15 dB per kilometer—a very high level of frequency reuse is available.

Another benefit of moving to higher frequencies is that external or received noise (space or galactic noise) reduces above 100 MHz. As you move to 1 GHz and above, external noise more or less disappears as an influence on performance (in a noise rather than interference limited environment) and receiver design—particularly LNA design—becomes the dominant performance constraint.

An additional reason to move to higher frequencies is that smaller, more compact resonant components—for example, antennas, filters, and resonators—can be used. Remember, RF wavelength is a product of the speed of light (300,000,000 meters per second) divided by frequency, as shown in Table 1.1.

During the 1920s, there was a rapid growth in broadcast transmission using long wave and medium wave. The formation of the BBC in 1922 was early recognition of the political and social importance of radio broadcasting. At the same time, radio amateurs such as Gerald Marcuse were developing equipment for long-distance shortwave communication. In 1932, George V addressed the British Empire on the shortwave world service. In practice, there has always been substantial commonality in the processing techniques used for radio and TV broadcasting and two-way and later cellular radio—a convergence that continues today.

Table 1.1 Frequency and Wavelength Relationship

FREQUENCY	SPEED OF LIGHT IN METERS PER SECOND DIVIDED BY FREQUENCY	WAVELENGTH
100 MHz	300,000,000 / 100,000,000	= 3 m
300 MHz	300,000,000 / 300,000,000	= 1 m
900 MHz	300,000,000 / 900,000,000	= 0.33 m
2 GHz	300,000,000 / 2,000,000,000	= 0.15 m

In 1939, Major Edwin Armstrong introduced FM (frequency modulation) into radio broadcasting in the United States. FM had the advantage over AM (amplitude modulation) of the capture effect. Provided sufficient signal strength was available at the receiver, the signal would experience gain through the demodulator, delivering a significant improvement in signal-to-noise ratio. The deeper the modulation depth (that is, the more bandwidth used), the higher the gain. Additionally, the capture effect made FM more resilient to (predominantly AM) interference. Toward the end of World War II, the U.S. Army introduced FM radios working in the VHF band. The combination of the modulation and the frequency (VHF rather than shortwave) made the FM VHF radios less vulnerable to jamming.

Fifty years later, CDMA used wider bandwidth channels to deliver bandwidth gain (rather like wideband FM processor/demodulator gain). Rather like FM, CDMA was, and is, used in military applications because it is harder to intercept.

A shortwave or VHF portable transceiver in 1945 weighed 40 kg. Over the next 50 years, this weight would reduce to the point where today a 100 gm phone is considered overweight.

Parallel developments included a rapid increase in selectivity and stability with a reduction in practical channel spacing from 200 kHz in 1945 to narrowband 12.5, 6.25, or 5 kHz transceivers in the late 1990s, and reductions in power budget, particularly after the introduction of printed circuit boards and transistors in the 1950s and 1960s. The power budget of an early VHF transceiver was over 100 Watts. A typical cell phone today has a power budget of a few hundred milliWatts.

As active and passive device performance has improved and as circuit geometries have decreased, we have been able to access higher parts of the radio spectrum. In doing so, we can provide access to an ever-increasing amount of radio bandwidth at a price affordable to an ever-increasing number of users.

As RF component performance improved, RF selectivity also improved. This resulted in the reduction of RF channel spacing from several hundred kHz to the narrowband channels used today—12.5 kHz, 6.25 kHz, or 5 kHz (used in two-way radio products).

In cellular radio, the achievement of sensitivity and selectivity is increasingly dependent on baseband performance, the objective being to reduce RF component costs, achieve better power efficiency, and deliver an increase in dynamic range. The trend since 1980 has been to relax RF channel spacing from 25 kHz (1G) to 200 kHz (2G GSM; Global System for Mobile Communication) to 5 MHz (3G). In other words, to go wideband rather than narrowband.

Handset design objectives remain essentially the same as they have always been—sensitivity, selectivity, and stability across a wide dynamic range of operational conditions, though the ways in which we achieve these parameters may change. Likewise, we need to find ways of delivering year-on-year decreases in cost, progressive weight and size reduction, and steady improvements in product functionality.

In the introduction, we highlighted microcontrollers, digital signal processors (DSPs), CMOS (complementary metal-oxide semiconductors) image sensors, and displays as key technologies. We should add high-density battery technologies and RF component and packaging technology. RF component specifications are determined by the way radio bandwidth is allocated and controlled—for example, conformance standards on filter bandwidths, transmit power spectral envelopes, co-channel and adjacent channel interference, phase accuracy, and stability.

Historically, there has also been a division between *wide area access* using duplex spaced bands (sometimes referred to as paired bands) in which the transmit frequencies are separated by several MHz or tens of MHz from receive frequencies, and local area access using nonpaired bands in which the same frequency is used for transmit and receive. Some two-way radios, for example, still use single frequency working with a press-to-talk (PTT) key that puts the transceiver into receive or transmit mode. Digital cordless phones use time-division duplexing. One time slot is used for transmit, the next for receive, but both share the same RF carrier.

One reason why cellular phones use RF duplexing and cordless phones do not is because a cellular phone transmits at a higher power. A cordless phone might transmit at 10 mW, a cellular handset transmits at between 100 mW and 1 Watt, a cellular base station might transmit at 5, 10, 20, or 40 Watts. For these higher-power devices, it is particularly important to keep transmit power out of the receiver.

Duplex Spacing for Cellular (Wide Area) Networks

Given that receive signal powers are often less than a picoWatt, it is clear that RF duplex spaced bands tend to deliver better receive sensitivity and therefore tend to be used for wide area coverage systems. Wide area two-way radio networks in the UHF band typically use 8 MHz or 10 MHz duplex spacing, 800/900 MHz cellular networks use 45 MHz duplex spacing, GSM 1800 uses 95 MHz duplex spacing, PCS 1900 uses 80 MHz, and IMT2000 (3G) uses 190 MHz duplex spacing. In the United States, there are also proposals to refarm 30 MHz of TV channel bandwidth in the 700 MHz band for 3G mobile services.

Figure 1.1 shows the duplex spacing implemented at 800/900 MHz for GSM in Europe, CDMA/TDMA in the United States, and PDC (Japan's 2G Personal Digital Cellular standard) in Japan. PDC was implemented with 130 MHz duplex spacing (and 25 kHz channel spacing), thus managing to be different than all other 2G cellular standards.

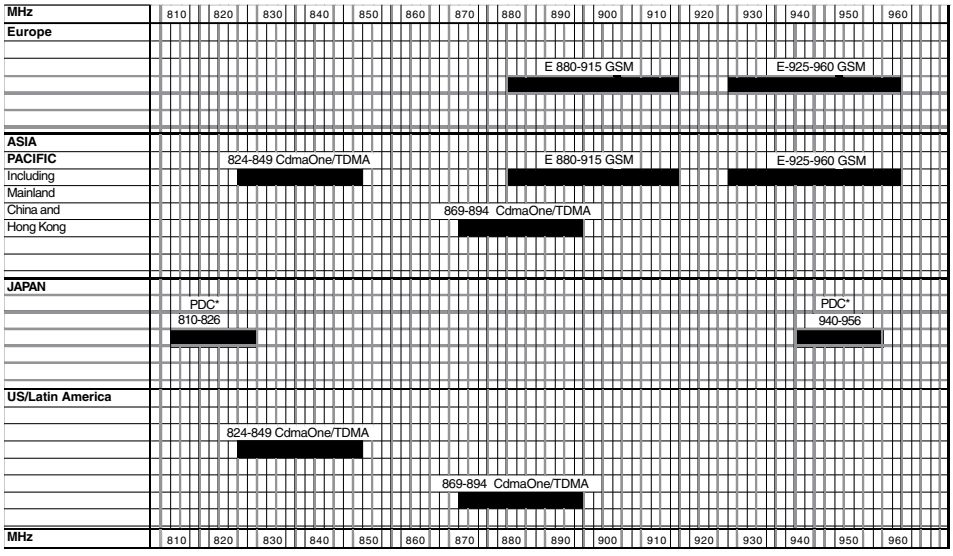


Figure 1.1 Cellular frequency allocations—800/900 MHz with duplex spacing.

In Asia, countries with existing Advanced Mobile Phone System (AMPS), and CDMA/TDMA allocations have a problem in that the upper band of AMPS overlaps the lower band of GSM. As the GSM band is paired, this means the corresponding bands in the upper band of GSM are unusable. The result is that certain countries (Hong Kong being the most obvious example) had a shortage of capacity because of how the spectrum had been allocated. Latin America has the same 800/900 MHz allocation as the United States (also shown in Figure 1.1). In the United States and Latin America, however, the AMPS 2 × 25 MHz allocations are bounded by politically sensitive public safety specialist mobile radio spectrum, preventing any expansion of the US 800 MHz cellular channel bandwidth.

In Europe, the original (1G) TACS allocation was 2 × 25 MHz from 890 to 915 MHz and 935 to 960 MHz (1000 × 25 kHz channels), which was later extended (E-TACS) to 33 MHz (1321 × 25 kHz channels). GSM was deployed in parallel through the early to mid-1990s and now includes 25 MHz (original allocation), plus 10 MHz (E-GSM), plus 4 MHz for use by European railway operators (GSM-R), for a total of 39 MHz or 195 × 200 kHz RF channels

Additional spectrum was allocated for GSM in the early 1990s at 1800 MHz (GSM1800). This gave three bands of 25 MHz each to three operators (75 MHz—that is, 375 × 200 kHz paired channels). As with all duplex spaced bands, handset transmit is the lower band. (Because of the slightly lower free space loss, this is better for a power-limited handset.) Only a fraction of this bandwidth is actually used, rather undercutting operator's claims to be suffering from a shortage of spectrum.

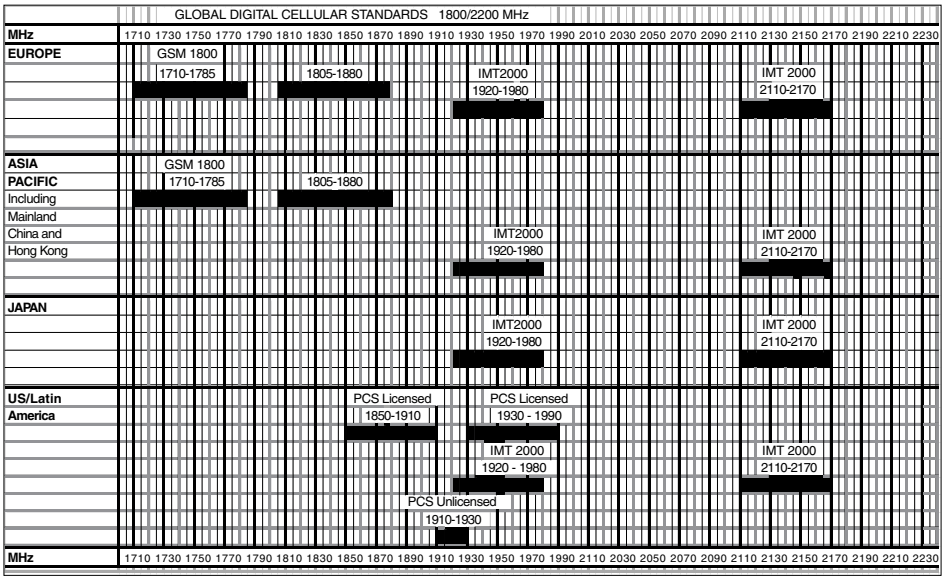


Figure 1.2 Cellular frequency allocations at 1800, 1900, and 2100 MHz.

In the United States and Latin America, 2 × 60 MHz was allocated at 1850 to 1910 and 1930 to 1990 MHz for US TDMA (30 kHz) or CDMA (1.25 MHz) channels or GSM (200 kHz) channels (GSM 1900), as shown in Figure 1.2. Unfortunately, the upper band of PCS 1900 overlaps directly with the lower band of IMT2000, the official ITU allocation for 3G. The intention for the IMT allocation was to make 2 × 60 MHz available, divided into 12 × 5 MHz channels, and this has been the basis for European and Asian allocations to date. In addition, 3 × 5 MHz nonpaired channels were allocated at 2010 to 2025 MHz and 4 × 5 MHz nonpaired channels at 1900 to 1920 MHz. The air interface for the paired bands is known as IMT2000DS, and for the nonpaired bands, it is IMT2000TC. (We discuss air interfaces later in this chapter.)

Figure 1.3 shows the RF bandwidth that needs to be addressed if the brief is to produce an IMT2000 handset that will also work in existing 2G networks (GSM 900, GSM 1800, GSM 1900) co-sharing with US TDMA and CDMA.

Some countries have the 60 MHz IMT2000 allocation divided among five operators. Five licensees sharing a total of 60 MHz would each have 12 MHz of spectrum. As this is not compatible with 5 MHz channel spacing, two operators end up with 3 × 5 MHz paired bands and three operators end up with 2 × 5 MHz paired bands and a nonpaired band (either in TDD1 or TDD2). It will therefore be necessary in some cases to support IMT2000DS and IMT2000TC in a dual-mode handset. The handset configuration would then be IMT2000DS, IMT2000TC, GSM 1900, GSM 1800, and GSM 900. Table 1.2 shows that selectivity and sensitivity are increasingly achieved at baseband, reducing the requirement for RF filters and relaxing the need for frequency stability. The need for backward compatibility, however, makes this benefit harder to realize.

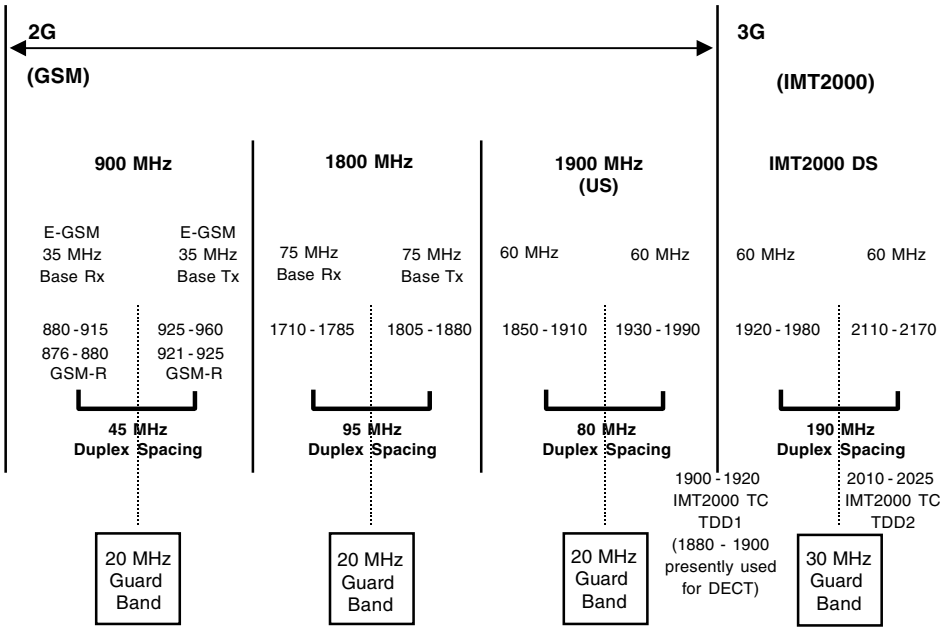


Figure 1.3 Tri-band GSM and IMT2000 allocations.

Table 1.2 Simplified RF Architecture

		SPECTRUM	CHANNEL SPACING	NO. OF RF CHANNELS
1G	E-TACS	33 MHz	25 kHz	1321
	AMPS	25 MHz	30 kHz	833
2G	GSM 900	39 MHz	200 kHz	195
	GSM 1800	75 MHz	200 kHz	375
	GSM 1900	60 MHz	200 kHz	300
3G	IMT2000DS	60 MHz	5 MHz	12
	IMT2000TC	35 MHz	5 MHz	7

First-generation AMPS/ETACS phones were required to access a large number of 25 kHz RF channels. This made synthesizer design (the component used to lock the handset onto a particular transmit and receive frequency pair) quite complex. Also, given the relatively narrowband channel, frequency stability was critical. A 1 ppm (part per million) temperature compensated crystal oscillator was needed in the handset. It also made network planning (working out frequency reuse) quite complex.

In second generation, although relaxing the channel spacing to 200 kHz reduced the number of RF channels, the need for faster channel/slot switching made synthesizer design more difficult. However, adopting 200 kHz channel spacing together with the extra complexity of a frequency and synchronization burst (F burst and S burst) allowed the frequency reference to relax to 2.5 ppm—a reduction in component cost.

In third generation, relaxing the channel spacing to 5 MHz reduces the number of RF channels, relaxes RF filtering, makes synthesizer design easier, and helps relax the frequency reference in the handset (to 3 ppm). Unfortunately, you only realize these cost benefits if you produce a single-mode IMT2000 phone, and, at present, the only country likely to do this—for their local market—is Japan.

Additionally you might choose to integrate a Bluetooth or IEEE 802 wireless LAN into the phone or a GPS (Global Positioning System/satellite receiver). In the longer term, there may also be a need to support a duplex (two-way) mobile satellite link at 1980 to 2010 and 2170 to 2200 MHz. In practice, as we will see in the following chapters, it is not too hard to integrate different air interfaces at baseband. The problem tends to be the RF component overheads.

A GSM 900/1800 dual-mode phone is relatively simple, particularly as the 1800 MHz band is at twice the frequency of the 900 band. It is the add-on frequencies (1.2, 1.5, 1.9, 2.1, 2.4 GHz) that tend to cause design and performance problems, particularly the tendency for transmit power at transmit frequency to mix into receive frequencies either within the phone itself or within the network (handset to handset, handset to base station, base station to handset, and base station to base station interference). And although we stated that it is relatively easy to integrate different air interfaces at baseband, it is also true to say that each air interface has its own unique RF requirements.

Multiplexing Standards: Impact on Handset Design

We have just described how RF channel allocation influences RF performance and handset design. Multiplexing standards are similarly influenced by the way RF channels are allocated. In turn, multiplexing standards influence handset design.

There are three options, or a combination of one or more of these:

- Frequency Division Multiple Access (FDMA)
- Time Division Multiple Access (TDMA)
- Code Division Multiple Access (CDMA)

FDMA

A number of two-way radio networks still just use FDMA to divide users within a given frequency band onto individual narrowband RF channels. Examples are the European ETSI 300/230 digital PMR (Private Mobile Radio) standard in which users have access to an individual digitally modulated 12.5 kHz or 6.25 kHz channel, the French TETRAPOL standard in which users have access to an individual digitally modulated 12.5, 10, or 6.25 kHz channel, and the US APCO 25 standard in which users have access to an individual digitally modulated 12.5 kHz or 6.25 kHz RF channel.

Narrowband RF channels increase the need for RF filtering and an accurate frequency reference (typically better than 1 ppm long-term stability). They do, however, allow for a narrowband IF implementation that helps minimize the noise floor of the receiver. The result is that narrowband two-way radios work well and have good sensitivity and good range in noise-limited environments, including VHF applications where atmospheric noise makes a significant contribution to the noise floor. The only disadvantage, apart from additional RF component costs, is that maximum data rates are constrained by the RF channel bandwidth, typically to 9.6 kbps.

TDMA

The idea of TDMA is to take wider band channels, for example, 25 kHz, 30 kHz, or 200 kHz RF channels and time-multiplex a number of users simultaneously onto the channel. Time slots are organized within a frame structure (frames, multiframes, superframes, hyperframes) to allow multiple users to be multiplexed together in an organized way. The objective is to improve channel utilization but at the same time relax the RF performance requirements (filtering and frequency stability) and reduce RF component costs in the handset and base station.

An example of TDMA used in two-way radio is the European Trans European Trunked Radio Access (TETRA) standard. A 25 kHz channel is split into four time slots each of 14.17 ms, so that up to 4 users can be modulated simultaneously onto the same 25 kHz RF carrier.

TETRA is presently implementing a fairly simple bandwidth-on-demand protocol where a single user can be given one, two, three, or four time slots within a frame. This means that one relatively high rate user per RF channel or four relatively low rate users or any combination in between can be supported. A similar format is used by Motorola in their proprietary iDEN air interface (six slots in a 990 ms frame length).

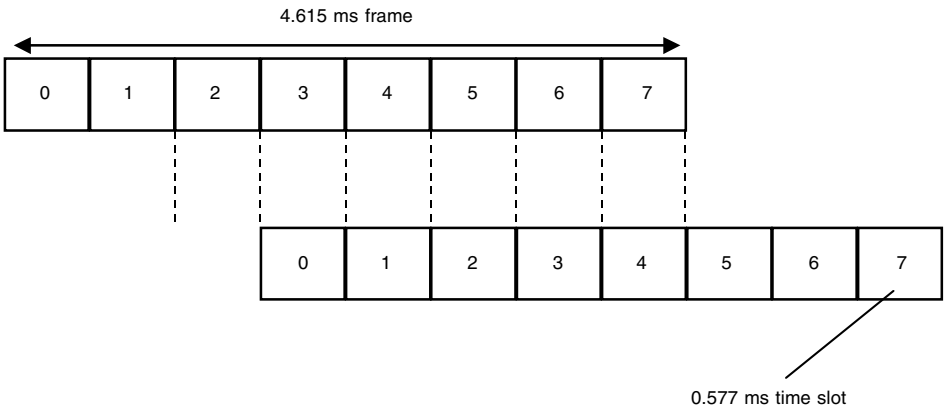


Figure 1.4 GSM slot structure.

In the United States, the AMPS 30 kHz analog channels were subdivided during the 1990s using either TDMA or CDMA. The time-division multiplex uses a three-slot structure (three users per 30 kHz RF channel), which can optionally be implemented as a six-slot structure.

A similar time-division multiplex was implemented in the Japanese Personal Digital Cellular networks but using a 25 kHz rather than 30 kHz RF channel spacing. In Europe, an eight-slot time multiplex was implemented for GSM using a 200 kHz RF channel, as shown in Figure 1.4.

One specific objective of the air interface was to reduce RF component cost by relaxing the RF channel spacing, from 25 kHz to 200 kHz. In common with all other TDMA interfaces, additional duplex separation is achieved by introducing a time offset. In GSM, transmit and receive are both on the same time slot—for example, time slot 2 but with a three-slot frame offset. This helps to keep transmit power (+30 dBm) out of the receiver front end (having to detect signals at –102 dBm or below). The combination of RF and time-division duplexing helps to deliver good sensitivity and provides the option to reduce RF component costs by dispensing with the duplex filter in some GSM phone designs.

Another route to reducing component costs is to use the air interface to provide synchronization and frequency correction as part of the handset registration procedure—an S burst to synchronize, an F burst to provide a frequency fix.

A long, simple burst on the forward control channel aligns the handset, in time, to the downlink time slots. In the frequency domain, the modulation is given a unidirectional $\pi/2$ phase shift for similar successive bits, giving a demodulated output of a sine wave at 1625/24 kHz higher than the center carrier frequency. This means that the F burst aligns the handset, in frequency, to the downlink RF carrier.

CDMA

In the mid-1990s CDMA cellular networks began to be deployed in the United States, Korea, and parts of Southeast Asia. Effectively, CDMA takes many of the traditional RF tasks (the achievement of selectivity, sensitivity, and stability) and moves them to baseband. The objective is to deliver processing gain that can in turn deliver coverage and/capacity advantage over the coverage and/capacity achievable from a TDMA air interface. Endless arguments ensued between the TDMA and CDMA camps as to which technology was better.

In practice, because of political and regulatory reasons and other factors such as timing, vendor, and operator support, GSM became the dominant technology in terms of numbers of subscribers and numbers of base stations deployed, which in turn conferred a cost and market advantage to GSM vendors. However, the technology used in these early CDMA networks has translated forward into 3G handset and network hardware and software. It is easier to qualify some of the design options in 3G handsets if we first cover the related design and performance issues highlighted by CDMA implementation to date.

The original principle of CDMA, which still holds true today, is to take a relatively narrowband modulated signal and spread it to a much wider transmitted bandwidth. The spreading occurs by multiplying the source data with a noise like high-rate pseudorandom code sequence—the pseudorandom number (PN). The PN as a digital number appears to be random but is actually predictable and reproducible having been obtained from a prestored random number generator. The product of the source data and the PN sequence becomes the modulating signal for the RF carrier.

At the receive end, the signal is multiplied by the same prestored PN sequence that was used to spread the signal, thereby recovering the original baseband (source) digital data. Only the signal with the same PN sequence despreads. Effectively, the PN sequences characterize the digital filter, which correlates or captures wanted signal energy, leaving unwanted signal energy down in the noise floor.

Multiple users can exist simultaneously on the same RF channel by ensuring that their individual spreading codes are sufficiently different to be unique. To control access and efficiency on a CDMA network, the spreading code is a composite of several digital codes, each performing a separate task in the link. It is usual to refer to each sequence or code as a *channel*.

IS95 defines the dual-mode AMPS/CDMA technology platform, IS96 the speech coding (currently either 8 kbps or 13 kbps), IS97 and 98 the performance criteria for base stations and handsets, and IS99 data service implementation. What follows is therefore a description of the IS95 air interface, which then served as the basis for CDMA2000.

In IS95, there is one pilot channel, one synchronization channel, and 62 other channels corresponding to 64 Walsh codes. All 62 channels can be used for traffic, but up to 7 of these may be used for paging. The 64 Walsh codes of length 64 bits are used for each of these channels. Walsh Code W0 is used for the pilot, which is used to characterize the radio channel. Walsh Code W32 is used for synchronization. Other Walsh codes are used for the traffic. The Walsh codes identify channels on the downlink, which means they provide channel selectivity.

Walsh codes are a sequence of PN codes that are orthogonal in that, provided they remain synchronized with each other, the codes do not correlate or create co-code or adjacent code interference. Orthogonal codes are codes of equal distance (the number of symbols by which they differ is the same). The cross correlation—that is, code interference—is zero for a perfectly synchronous transmission.

On the uplink, the channel bits are grouped into 6-bit symbols. The 6-bit group (higher-order symbol) generates a 64-chip Walsh code. The orthogonality of the 64 codes gives an increased degree of uniqueness of data on the uplink—that is, it provides selectivity.

The resultant Walsh code is combined with a long code. The composite channel rate is 1.228 Mcps; in other words, the code stream is running at a rate of 1.228 MHz. The long code is a PN sequence truncated to the frame length (20 ms). On the uplink, the long code provides user-to-user selectivity; on the downlink, one long code is used for all base stations but each base station has a unique PN offset (a total of 512 time PN offsets are available). So within a relatively wideband RF channel, individual user channels are identified on the downlink using Walsh codes—with long codes providing cell-to-cell selectivity—individual user channels are identified on the uplink by use of the 6-bit symbols, and long codes are used to provide user-to-user selectivity.

From a handset design point of view, digital filters have replaced the time slots and RF filters used in the TDMA networks. Although RF filtering is still needed to separate multiple 1.25 MHz RF carriers, it is intrinsically a simpler RF channel plan, and it can be implemented as a single-frequency network if traffic loading is relatively light and evenly distributed between cells.

Difference between CDMA and TDMA

An important difference between TDMA and CDMA is that in TDMA, the duty cycle of the RF amplifier is a product of the number of time slots used. A GSM handset using one time slot has a duty cycle of 1/8. Maximum output power of a 900 MHz GSM phone is 2 Watts. Effectively, the average maximum power available across an eight-slot frame is therefore 250 mW.

In CDMA, the handset is continuously transmitting but at a maximum of 250 mW. The total power outputs are therefore similar. In a TDMA phone, the RF burst has to be contained within a power/time template to avoid interference with adjacent time slots.

The RF output power of the TDMA handset is adjusted to respond to changes in channel condition (near/far and fading effects) typically every 500 ms. In an IS95 CDMA phone, power control is done every 1.25 ms, or 800 times a second. This is done to ensure that user codes can be decorrelated under conditions of relatively stable received signal strength (energy per bit over the noise floor). Failure to maintain reasonably equivalent E_b/N_os (energy per bit over the noise floor) between code streams will result in intercode interference.

Traditionally the power control loop in an IS95 CDMA phone requires careful implementation. We discuss power control loop design in a later section in the chapter.

Modulation: Impact on Handset Design

Information can be modulated onto an RF carrier by changing the amplitude of the carrier, the frequency of the carrier, or the phase of the carrier. For example, using Minimum Shift Keying (MSK), the carrier changes phase by +90° or -90° over a bit period (see Figure 1.5).

The example shown in Figure 1.5 is a constant envelope phase modulation scheme. Prior to modulation, the data stream passes through baseband filters. In Gaussian Minimum Shift Keying (GMSK), these are Gaussian filters.

The advantage of GMSK, being constant envelope, is that it can be used with Class C amplifiers, which typically have a power efficiency of between 50 and 55 percent. The disadvantage is that with the GSM implementation of GMSK, because of the filtering, decision points on the modulation trellis are not always obtained, resulting in some residual bit errors. GMSK is a two-level modulation scheme—that is, the two phase states can represent a 0 or a 1.

Higher-level modulation states can be used to carry more bits per symbol. A four-state modulation scheme, for example, QPSK (Quadrature Phase Shift Keying) has 2 bits per symbol (00, 01, 11, 10), an eight-level modulation scheme can carry 3 bits per symbol, a 16-level modulation scheme can carry 4 bits per symbol, a 1024-level modulation scheme (used in fixed point-to-point, for example) can carry 10 bits per symbol. However, as the number of modulation states increase, the distance between phase states reduces and the likelihood of a demodulator error increases. Every time a modulation level is doubled (for example, from two-level to four-level), an additional 3 dB of signal energy is needed to maintain equivalent demodulator bit error rate performance.

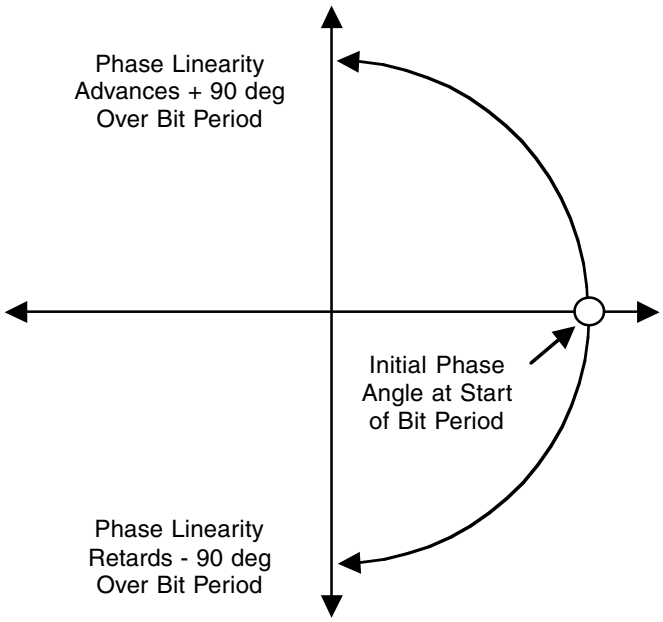


Figure 1.5 Minimum shift keying (MSK).

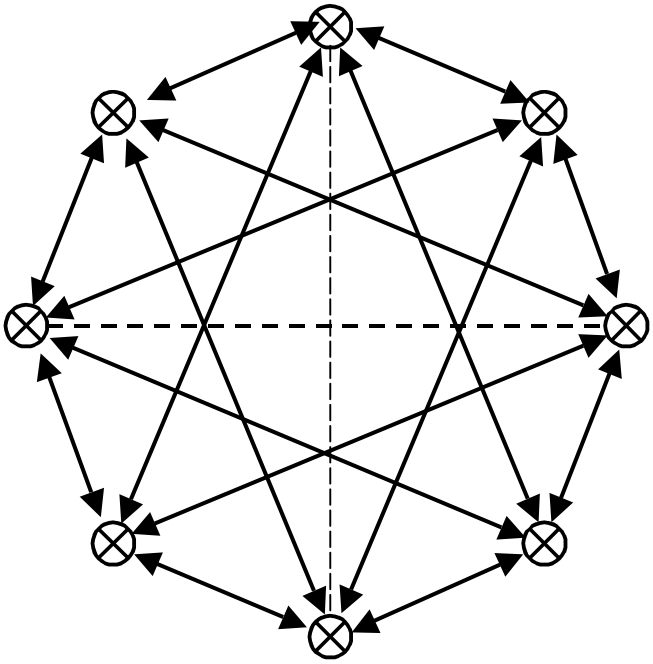


Figure 1.6 IS54 TDMA—modulation vector—I Q diagram for π/4 DQPSK modulation.

Higher-level modulations also tend to contain amplitude components and can therefore not be used with power-efficient Class C amplification. The modulation technique used in IS54 TDMA is an example (see Figure 1.6).

This is a four-level modulation technique known as π/4DQPSK. DQPSK refers to "differential quadrature phase shift keying," the use of four differentially encoded phase states to describe a 00, 01, 01, or 10. The π/4 indicates that the vector is indexed by 45° at every symbol change. This makes it look like an eight-level modulation trellis, which it isn't. It shows that any change from phase state to phase state avoids passing through the center of the trellis, which would imply a 100 percent AM component. Instead the AM component is constrained to 70 percent.

Even so, the modulation requires a higher degree of linear amplification to avoid spectral regrowth during and after amplification. While this is reasonably easily accommodated in low-power handsets, it does result in larger—and hotter—RF amplifiers in IS54 TDMA base stations.

Similarly, CDMA uses QPSK on the downlink and offset QPSK on the uplink (as with π/4DQPSK, OQPSK reduces the AM components and relaxes the linearity requirements of the handset PA). It is, however, difficult to realize efficiencies of more than 7 to 8 percent in base station amplifiers (QPSK), substantially increasing the power and heat dissipation needed, relative to GSM. This is why it has been easier to produce very small picocellular base stations for GSM (1.5 kg) but harder to deliver an equivalent form factor for IS54 TDMA or CDMA products.

Future Modulation Schemes

The choice of modulation has always been a function of hardware implementation and required modulation and bandwidth efficiency. In the 1980s, FM provided—and still provides today—an elegant way of translating an analog waveform onto an (analog) RF carrier.

In the 1990s, GMSK was used for GSM as a relatively simple way to digitally modulate or demodulate an RF carrier without the need for linearity in the RF PA. Note that GSM was developed as a standard in the early 1980s. US TDMA and IS95 CDMA were specified/standardized toward the end of the 1980s, by which time four-level modulation schemes (with AM components) were considered to provide a better efficiency trade-off. Figure 1.7 compares the performance trade-offs of QPSK (1), MSK (2), and GMSK (3). QPSK (1) carries 2 bits per symbol but has relatively abrupt phase changes at the symbol boundaries. MSK (2) has a constant rate of change of phase but still manages to maintain an open eye diagram at the symbol decision points. GMSK (3) has additional filtering (a Gaussian baseband filter that effectively slows the transition from symbol state to symbol state). The filtering ensures the modulation is constant envelope; the disadvantage is that decision points are not always achieved, resulting in a residual demodulated bit error rate.

QPSK is used in IMT2000MC and IMT2000DS on the downlink. HPSK is used on the uplink to reduce linearity requirements. A variant of IMT2000MC known as 1xEV, however, also has the option of using 8 PSK (also used in GSM EDGE implementation) and 16-level QAM. This seems to be a sensible way to increase bandwidth efficiency, given that eight-level modulation can carry 3 bits per symbol and 16 level can carry 4 bits per symbol.

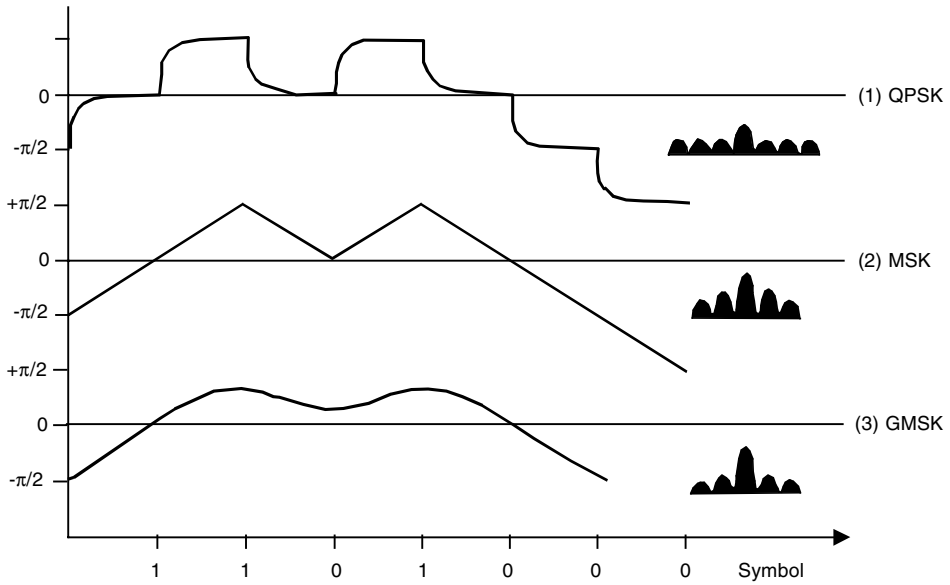


Figure 1.7 QPSK, MSK, and GMSK compared.

It is necessary, however, to qualify the impact of the choice of modulation on the link budget. For every doubling of modulation state, an additional 3 dB of link budget is required to maintain the same demodulation bit error performance. Therefore, 8 PSK needs 3 dB more link budget than QPSK, and 16-level QAM needs 3 dB more link budget than 8 PSK. Provided you are close to the base station, you can take advantage of higher-level modulation, but it will not deliver additional capacity at the edge of a cell. It is also worth verifying the time domain performance of the demodulator.

The usual rule of thumb is that a demodulator can tolerate a quarter symbol shift in terms of timing ambiguity without causing high demodulator error rates

In higher-level modulations, the symbol transition rate stays the same but the number of symbol states increases. The symbol states become closer together in terms of phase and frequency. Given that the vector is rotating, a timing error translates into a phase or frequency error.

Multipath effects cause phase rotation and attenuation. In CDMA, these are partly, though not totally, taken out by the RAKE receiver. Given that none of these adaptive mechanisms are perfect, timing ambiguity translates into demodulator error rate. This effect becomes more severe as bit rate and symbol rate increases. Thus, while higher-level modulation options promise performance gains, these gains are often hard to realize in larger cells, particularly in edge-of-cell conditions where power is limited and severe multipath conditions may be encountered.

An alternative is to use orthogonal frequency-division multiplexing (OFDM). OFDM is sometimes described incorrectly as a modulation technique. It is more correctly described as a multicarrier technique. Present examples of OFDM can be found in wireline ADSL/VDSL, fixed access wireless, wireless LANs, and digital TV.

Standard terrestrial digital TV broadcasting in Europe and Asia uses QPSK. (High-definition TV needs 16- or 64-level QAM and presently lacks the link budget for practical implementation.) The QPSK modulation yields a 10.6 Mbps data rate in an 8 MHz channel. The 8 MHz channel is divided into 8000×1 kHz subcarriers that are orthogonal from each other. The OFDM signal is created using a Fast Fourier Transform. (Fast Fourier Transforms were first described by Cooley and Tukey in 1963 as an efficient method for representing time domain signals in the frequency domain.)

As there are now a total of 8000 subcarriers, the symbol rate per carrier is slow and the symbol period is long compared to any multipath delays encountered on the channel. Continuous pilot bits are spread randomly over each OFDM symbol for synchronization and phase error estimation; scattered pilot bits are spread evenly in time and frequency across all OFDM symbols for channel sounding.

The advantage of OFDM is that it provides a resilient channel for fixed and mobile users. (DVB was always intended to provide support for mobility users.) The disadvantage of OFDM is that it requires a relatively complex FFT to be performed in the encoder and decoder. In digital TV, the power budget overheads associated with the complex transform do not matter, in the context of transmitters producing kiloWatts of RF power and receivers attached to a main supply.

Present implementation of an OFDM transceiver in a 3G cellular handset would, however, not be economic in terms of processor and power budget overhead. OFDM is however, a legitimate longer-term (4G) option providing a bandwidth efficient robust way of multiplexing multiple users across 10, 15, or 20 MHz of contiguous bandwidth.

It also provides the basis for converging the 3G TV and cellular radio network bandwidth proposition. We examine the technology factors influencing 3G TV and cellular network convergence in Chapter 19.

TDMA Evolution

By the end of the 1990s, the mix of deployed technologies included AMPS/TDMA networks (in the United States and parts of Latin America and Asia), using 30-kHz RF channel spacing, GSM networks using 200 kHz RF channel spacing, and CDMA networks using 1.25 MHz channel spacing. The proposed migration route for AMPS/TDMA network operators was to introduce 200 kHz channel rasters and a new 3, 6, 8, 16, 32, or 64 slot frame structure, the idea being to provide more flexible bandwidth-on-demand capability.

Part of the logic here is to take into account the likely dynamic range of the information rate needing to be presented to the channel. For example, a simultaneously encoded voice, image, video, and data stream could result in a composite information rate varying from 15 kbps to 960 kbps, and the rate could change every frame, or every 10 ms. This would be a 64-to-1 ratio (18 dB), hence the choice of a slot structure that can encompass a 64-slot frame where a single user can be allocated anything between 1 slot (a 1/64 duty cycle) to 64 slots (a 64/64 duty cycle) or any value in between. The 16, 32, and 64 slot frames are intended to be used with eight-level PSK, giving a maximum user data rate of 384 kbps.

A second objective is to harmonize the IS54 AMPS/TDMA air interface and GSM. Both air interfaces would have an eight-slot frame in common, both air interfaces would have eight-level PSK in common for higher bit rate applications and the 16, 32, and 64 slot frame structure for high dynamic range applications.

The eight-phase PSK implementation is known as Enhanced Data Rate for GSM Evolution (EDGE) and would be implemented in an AMPS/TDMA network using either 3×200 kHz channels (Compact EDGE) or 12×200 kHz channels (Classic EDGE). A 50 kHz guard band is added on either side to provide protection to and from the IS136 30 kHz channels.

Table 1.3 shows the combined proposal submitted to the ITU and called IMT2000SC (single RF carrier with adaptive time-division multiplexing). The proposal is promoted by the Universal Wireless Communications Consortium (UWCC), now known as 3G Americas.

Table 1.3 2G to 3G Migration—IMT2000SC evolution.

TDMA MIGRATION	IS136	IS136+
Ericsson	Three-slot 30 kHz	Eight-level PSK = 384 kbps
(UWC Proposal)		8, 16, 32, 64 slot
IMT2000SC		

(continues)

Table 1.3 2G to 3G Migration—IMT2000SC evolution. *(Continued)*

GSM MIGRATION	GSM 2G	2.5G
Ericsson	Eight-slot 200 kHz	Eight-level PSK = 384 kbps
		8,16,32,64 slot

The implementation of EDGE into an AMPS/TDMA network requires some care to avoid disturbing existing reuse patterns, using either 1/3 reuse with Compact EDGE or 4/12 reuse with Classic EDGE (see Tables 1.4 and 1.5).

The bandwidth negotiation protocols (multislot allocation and release) would be common to GSM (part of the General Packet Radio Service protocol) and IS54/IS136 TDMA. IMT2000SC is one of the four air interface standards presently being promoted for 3G networks, the others being IMT2000MC, IMT2000DS, and IMT2000TC.

IMT2000MC provides an evolution from the existing IS95 CDMA air interface and is promoted by the CDMA Development Group (CDG) and 3GPP2—the third-generation partnership project standards group dedicated to air interface and network interface standardization and IS95 CDMA backward compatibility (see Figure 1.8).

The original IS95 CDMA networks use a 1.2288 Mcps rate to occupy a 1.25 MHz RF channel. The multichannel (MC) refers to using multiple, that is 3, 6, or 12 × 1.25 MHz carriers to increase per user bit rates. For example 3 × 1.25 MHz carriers will occupy 5 MHz, equivalent to IMT2000DS.

Table 1.4 IMT2000SC—Compact EDGE

IS136	GUARD BAND	COMPACT EDGE	GUARD BAND	IS136
30 kHz	50 kHz	3 × 200 kHz	50 kHz	30 kHz

Table 1.5 IMT2000SC—Classic EDGE

IS136	GUARD BAND	CLASSIC EDGE	GUARD BAND	IS136
30 kHz	50 kHz	12 × 200 kHz	50 kHz	30 kHz

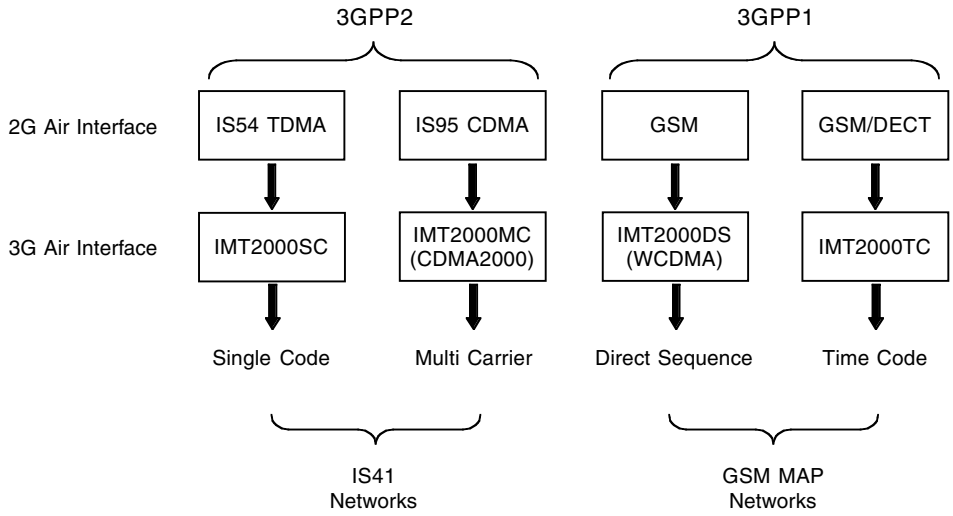


Figure 1.8 2G to 3G air interface evolution.

In practice, there are three ways to increase data rates:

- Allocating PN multiple codes to single users
- Increasing the chip rate, for example, from 1.2288 Mcps to 3.6864 Mcps (or higher multiples)
- Using higher-level modulation schemes such as 8 PSK or 16-level QAM—a version known as 1xEV and promoted by Qualcomm.

At time of writing, 1xEV, rather than the multicarrier IMT2000MCimplementation, is the favored evolution route and is generically known as CDMA2000.

IS54TDMA/IMT2000SC and IS95 CDMA/CDMA2000 are supported by a network standard known as IS41. GSM/IMT2000DS is supported by a network standard known as GSM-MAP (Mobile Application Part).

5 MHz CDMA: IMT2000DS

IMT2000DS is the air interface standard promoted by Ericsson and the 3GPP1, the third-generation partnership project standards group dedicated to promoting interworking with other standards and, specifically, backward compatibility with GSM—an aspect of particular interest to existing GSM network operators. Harmonization with GSM implied using a 13 or 26 MHz clock reference, rather than the 19 MHz clock reference used in IMT2000MC, and a 200 kHz channel raster.

Although RF channel spacing is nominally 5 MHz, it is possible to bring lower-power microcells together with 4.4 MHz RF spacing (but still keeping to the 200 kHz raster). This has the benefit of increasing the guard band between higher-power macrocells and lower-power microcells.

The idea of maintaining a 200 kHz channel raster is to simplify synthesizer design for dual-mode GSM/IMT2000DS phones. Additionally, GSM and IMT2000DS share the same share frame structure from the multiframe upward, the multiframe length being 120 ms (see Figure 1.9). This simplifies the implementation of GSM to IMT2000DS and IMT2000DS to GSM handovers, and could potentially allow for the use of GSM F bursts and S bursts to provide frequency and synchronization for IMT2000.

The IMT2000DS measurement frame and equivalent GSM control channel align to facilitate intersystem handover. Additionally, the code structure was chosen such that frequency accuracy could be transferred from outdoor macrocells or microcells to handsets, relaxing the frequency stability requirements of the handset. In turn, the handsets can transfer the frequency reference to indoor picocells, thereby avoiding the need for a GPS reference to be piped from outside a building to an indoor base station. The code structure is termed *asynchronous*, for reasons we will explain later.

The advantage of the IMT2000MC (CDMA2000) code structure is that it supports very resilient code channel acquisition. When a handset is first turned on, it can acquire the wanted code channel very cleanly. The disadvantage is that timing accuracy within the handset and base station needs to be within a fraction of the chip duration, hence the relatively tight tolerance for IMT2000MC frequency stability.

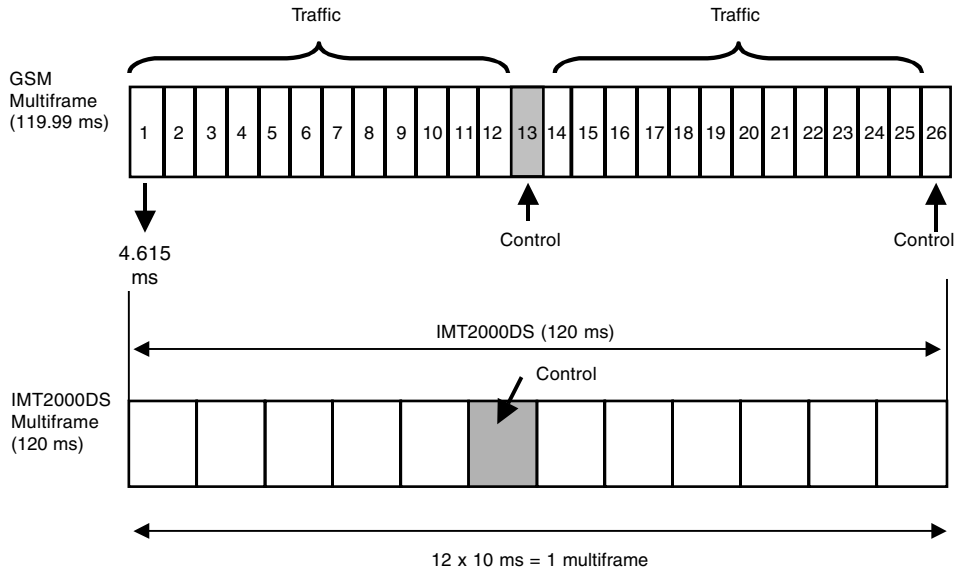


Figure 1.9 Multiframe compatibility between GSM and IMT2000DS.

Short-term stability has also to be tightly controlled (the jitter budget), since this will compromise code correlation in IMT2000MC. This is why higher chip rates in IMT2000MC tend to be avoided, and multiple RF carriers or higher level modulation are used as an alternative method for delivering higher bit rates. Long-term stability is also more critical for IMT2000MC. The relatively relaxed stability requirements for IMT2000DS save component costs in the handset but increase the complexity of code acquisition.

IMT2000TC shares a similar air interface to IMT2000DS—along with the same advantages and disadvantages, but it uses time-division duplexing (similar to a DECT cordless phone). In IMT2000TC the 15 time slots in a frame are used to divide uplink users from downlink users (see Figure 1.10). In effect, this is a CDMA air interface with a time-division multiplex. Each time slot can be additionally subdivided into separate code streams.

As with Digital Enhanced Cordless Telecommunications (DECT), the assumption here is that the air interface will only be used in small cells and that low powers will be used, easing the duplex requirement. The bandwidth can be increased on demand in either direction with between 2 and 13 slots on the uplink and between 1 and 14 slots on the downlink.

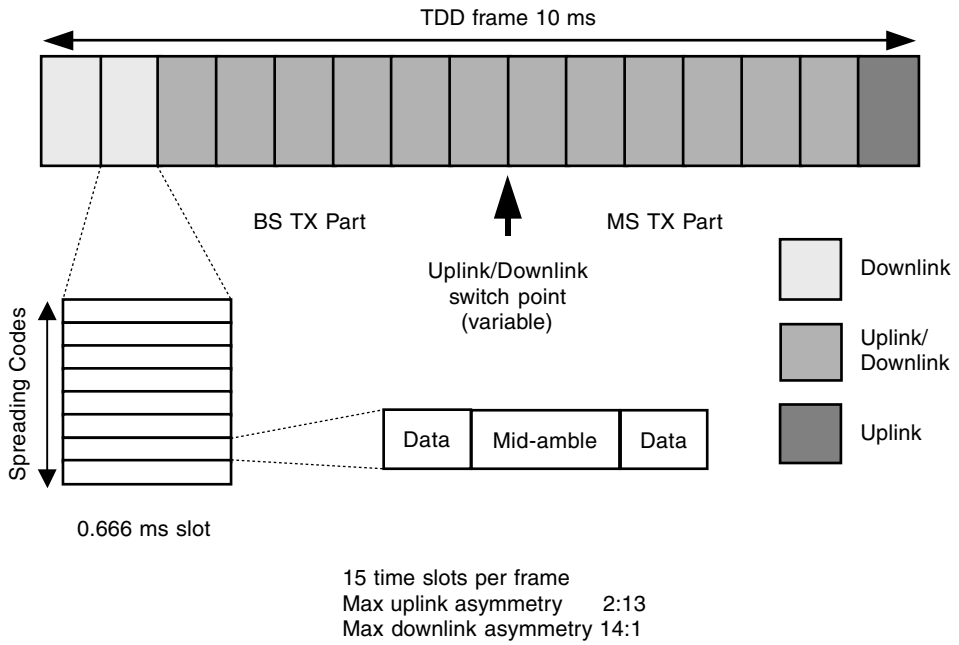


Figure 1.10 IMT2000TC.

Advantages of 5 MHz RF Channel Spacing

We have already highlighted the advantages of wider RF channel spacing in terms of relaxed frequency stability and RF filtering. Additionally, the fading envelope becomes less severe as the bandwidth increases, relative to the operating frequency.

This is known as the *coherence bandwidth*—the bandwidth over which all the signal is affected by multipath fading. As the bandwidth increases, parts of the signal remain unaffected. Fading is a phase cancellation effect and as such is frequency-specific. As the fading depth reduces it becomes progressively easier to follow the multipath fading with a power control loop.

In practice, both slow fading and fast fading (at least for slow mobility users) in a relatively wideband channel can be compensated for by decreasing and increasing the RF power of the handset in sympathy with the fade. In GSM, power control is done every 500 ms (twice a second), in IMT2000MC it is done 800 times a second, and in IMT2000DS it is done 1500 times a second.

This highlights some of the key differences between 3G handsets and GPRS/EDGE handset hardware. 2G air interfaces are designed to support, more or less, constant rate channels—an 8 or 13 kbps codec, for example. The channels themselves tend to be of variable quality; the result of the slow and fast fading experienced by mobile users. 2G can in effect be characterized as being designed to work with constant rate variable quality channels.

Attempts to deliver variable bandwidth (bandwidth on demand) have met with some success in 2G as we will document in the next chapter, but intrinsically the bandwidth limitations of a 200 kHz or 30 kHz channel limit the dynamic range that can be delivered for users with highly variable, bursty data rates. Moving to a wider RF channel makes it easier to deliver variable-rate constant-quality channel bandwidth. As we will see in later chapters, this is more or less essential for the movement and management of time-sensitive multimedia files and is the key performance differentiator between 2G and 3G air interface propositions.

Impact of Increasing Processor Power on Bandwidth Quality

As digital signal processing power increases, bandwidth quality and bandwidth flexibility increases. The objective is to realize these quality improvements and reduce component costs as well as RF and baseband power. There are four areas where quality improvements, cost reduction, and power efficiency benefits can be achieved—multiplexing, source coding, channel coding, and modulation.

Multiplexing

In the 1970s, the consensus emerged that it was going to be easier and cost less to filter in the time domain rather than the frequency domain. This is the reason the wireline world abandoned frequency multiplexing and adopted time-division multiplexing for

wireline backhaul transport. This thinking was taken into the wireless standardization committees. GSM effectively was based, and is still based today, on an ISDN structure and a time-division multiplex on the air interface.

The disadvantage with the time-division multiplex is that RF bursts need to be shaped and modulated onto the RF channel. As we explain in the next chapter, it is proving quite difficult to deliver flexible bandwidth on demand from any of the TDMA options, partly because of the challenge of pulse shaping in a multiple-slot handset.

CDMA moves the process of time domain filtering to baseband and delivers greater flexibility in terms of bandwidth on demand and multiple per-user traffic streams (we study how this is achieved in Chapter 3). Additionally, as described earlier, the CDMA multiplex allows a relaxation of RF channel spacing. CDMA only became possible in the early to mid-1990s, when it became feasible in cost and power budget terms to implement root raised cosine filters and low-cost, low-power budget PN code generators and numerically controlled oscillators (NCOs, studied in detail in Chapter 3).

In fourth-generation cellular, it is likely that CDMA will be combined with OFDM techniques to provide additional channel resilience (using 10, 15, or 20 MHz bandwidths). These hybrid time domain/frequency domain multiplexing schemes are generically described as coded orthogonal frequency-division multiplexing (COFDM). This is only possible when sufficient processing power is available to undertake handset transmit and receive Fast Fourier Transforms, but the benefit will be further improvements in the consistency of bandwidth quality (effectively an increase in coherence bandwidth). For the present, attention is focused on making CDMA work well.

Source Coding

In a first-generation cellular handset, you talk into a microphone and a variable voltage is produced, describing 3 kHz of voice modulated audio bandwidth. The voltage is then FM-modulated onto an RF carrier—an all analog processing chain.

In second-generation handsets, you talk into a microphone and the voice is turned into a digital bit stream using waveform encoding. For example, in GSM, a 104 kbps data stream is produced prior to the vocoder. It is the vocoder's job to reduce this data rate to, for example, 13 kbps or less without noticeable loss of quality. In the wireline world and in digital cordless phones, this is achieved in the time domain by using time domain compression techniques (exploiting sample-to-sample predictability). These are known as *adaptive differential pulse code modulation* codecs. They work well in high background noise conditions but suffer quality loss at low codec rates—for example, 16 kbps or below.

The decision was made that digital cellular handsets should use speech synthesis codecs that coded in the frequency domain (see Figure 1.11). The figure shows a female voice saying "der." Each block represents a 20-ms speech sample. The first block shows the "d," and the second block shows the "er" described in the time domain (y-axis) and frequency domain (x-axis). Each sample is described in terms of frequency coefficients. Compression is achieved by exploiting similarity between samples.

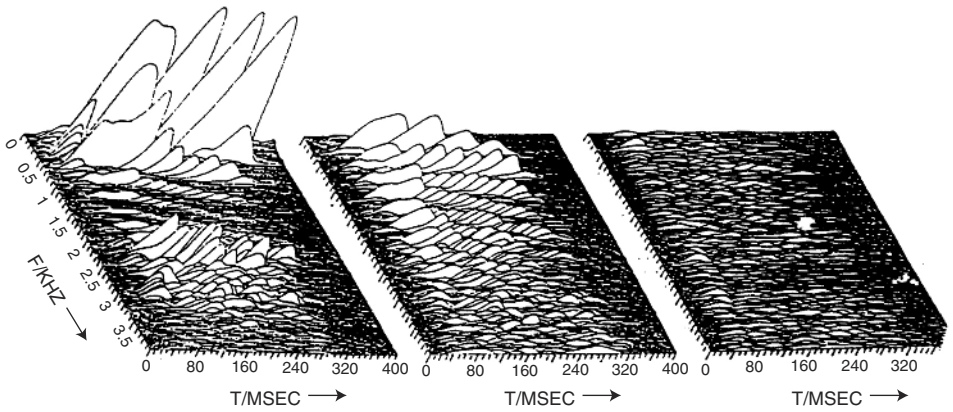


Figure 1.11 Speech coding—voice characteristics.

In the receiver, the frequency coefficients are used to rebuild, or synthesize, the harmonic structure of the original voice sample. The more processing power used in the codec, the better the quality for a given compression ratio.

Alternatively, rather than synthesize waveforms, waveforms can be prestored and fetched and inserted as needed in the decoder. This reduces processor overhead but increases memory bandwidth in the vocoder. These codecs are known as codebook codecs or more precisely *codebook excitation linear prediction* (CELP) codecs. Codecs used in present CDMA handsets and most future handsets are CELP codecs.

Voice codecs are also becoming variable rate, either switchable (for coverage or capacity gain) or adaptive (the codec rate varies according to the dynamic range of the input waveform). The objective of all codecs is to use processor bandwidth to reduce transmission bandwidth. Speech synthesis codecs and codebook codecs can deliver compression ratios of 8:1 or more without significant loss of quality.

3G handsets add in MPEG-4 encoders/decoders to support image and video processing. In common with vocoders, these video codecs use time domain to frequency domain transforms (specifically, a discrete cosine transform) to identify redundancy in the input image waveform. As we will see, video codecs are capable of delivering compression ratios of 40:1 or more with tolerable image quality.

Fourth-generation digital encoders will add in embedded rendering and mesh coding techniques to support motion prediction, motion estimation, and motion compensation.

Channel Coding

Channel coding has been used in digital cellular handsets and base stations for the past 10 years as a mechanism for improving transmission quality in a band-limited, noise-limited Rayleigh faded channel. Channel encoding adds bits to the source coded data calculated from the source coded data. The decoder uses these extra bits to detect and correct errors. Errors are detected when the actual transmitted redundancy value fails to match the redundancy value calculated from the transmitted data.

Two code types are used:

Block codes. Segment the message into blocks adding a parity check number, which is a product of the information bits contained in the block.

Convolutional codes. Also known as tree codes, the encoder has memory and the output code words depend on the current bit value and adjacent bits held within the register.

Block codes are good for detecting bursty errors, and convolutional codes work best with evenly distributed errors. Interleaving is used to help randomize error distribution to ensure convolutional encoders/decoders deliver coding gain. If an error burst lasts longer than the interleaving depth, the convolutional decoder will suffer from error extension, making matters worse. This will hopefully be detected by the block code parity check. The voice, image, or video sample can be discarded and the prior sample reused. Figure 1.12 shows a simple convolutional encoder.

Each time an information bit arrives at the front end of the encoder, a branch code word is produced. As the bit moves through the code register, it influences subsequent branch word outputs. The objective is to increase the distance between 0s and 1s. The memory action enables the decoder to construct and evaluate a multiple decision process on the recovered bits. This weighted analysis provides coding gain. These decoders are commonly described as *maximum likelihood decoders*. Figure 1.13 shows how coding gain is achieved in a GSM vocoder (encoder/decoder).

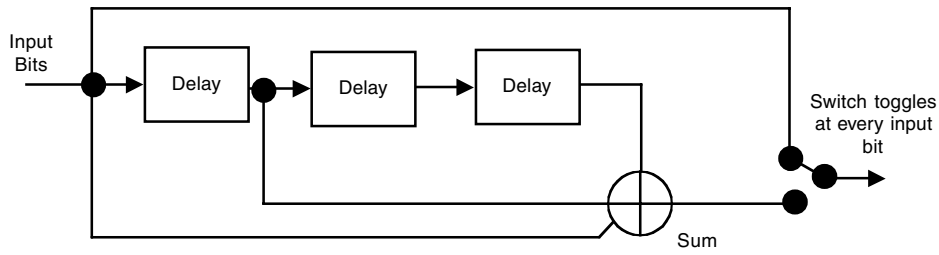


Figure 1.12 Simple convolutional encoder.

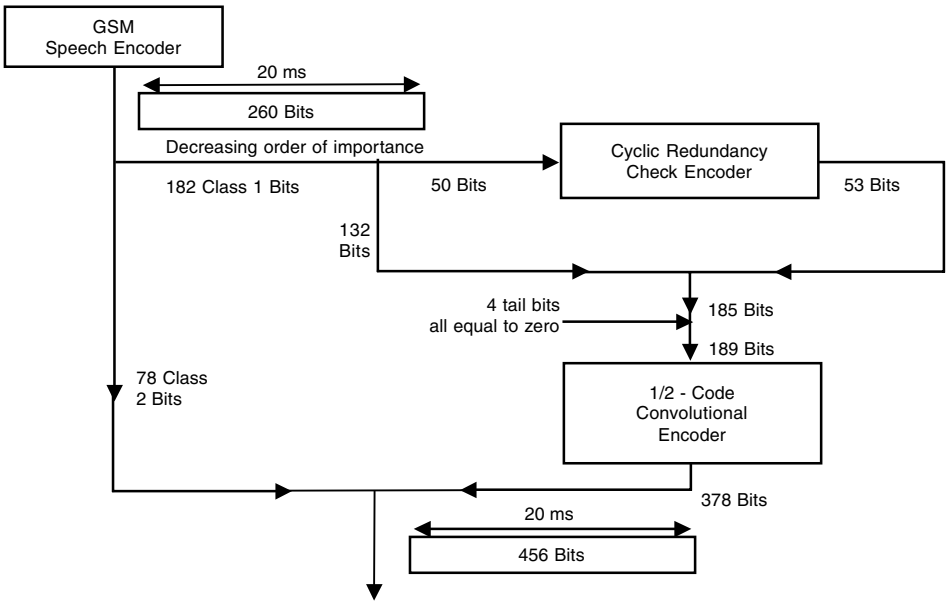


Figure 1.13 Encoding and decoding of a speech burst.

A 20-ms speech sample is described as a 260-bit word, which effectively contains the speech sample frequency coefficients. The 260 bits are split into Class 1 bits, which have parity bits added and are then convolutionally encoded. Note the 1/2 encoder doubles the number of bits—2 bits out for every 1 bit in. Class 2 bits are uncoded. In the decoder, coded bits pass through the convolutional decoder. If the burst errors are longer than the interleaving depth (40 ms in GSM), the block coded parity check detects a parity error, the speech sample is discarded, and the prior sample is reused.

Increasing K, the length of the convolutional encoder, increases resilience against burst errors and delivers additional coding gain ($K = 7$ typically delivers 5.2 dB gain, $K = 9$ delivers 6 dB of gain) but requires an exponential increase in decoder complexity (trading instructions per second against receive sensitivity). This coding gain depends on having sufficient interleaving depth available on the air interface. Interleaving depth in 3GPP1 (IMT2000DS/W-CDMA) is a variable: a minimum of 10 ms, a maximum of 80 ms.

Increasing the interleaving depth from 10 to 80 ms increases coding gain by just under 1 dB for slow mobility users (3 km/h), by just over 1 dB for medium mobility users (20 km/h). However, increasing interleaving depth increases delay. Figure 1.14 shows how interleaving is implemented in GSM. Each 456-bit block is split into 8×57 sub-bit blocks and interleaved over eight time slots and eight frames (approximately 40 ms). This is an irreducible delay. You cannot reduce it by using faster processors, because the delay is a function of the fixed frame rate.

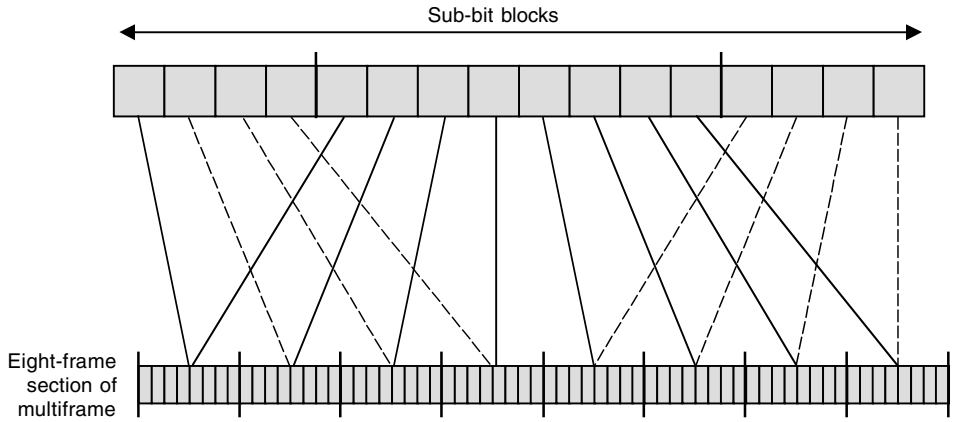


Figure 1.14 GSM channel coding—interleaving.

An alternative is that careful implementation of fast power control, using the 1500 Hz power control loop specified in 3GPP1, makes it possible to follow the fast fading envelope, which was already partly tamed by the coherence bandwidth of the 5 MHz channel. If the fast fading can be counteracted by the power control loop, the Rayleigh channel becomes a Gaussian channel in which burst errors no longer occur. Fast power control in a 5 MHz channel can therefore, theoretically, deliver additional coding gain at least for medium-mobility users (up to 20 km/h) without the need for deep interleaving.

This shows the intricate and rather complex relationship between source rate, convolutional encoding (the choice of 1/2 or 2/3 encoders, for example), interleaving depth and coding gain, which in turn determines uplink and downlink sensitivity. All the preceding parameters can be dynamically tuned to optimize handset performance.

Convolution and Correlation

Convolutional encoding is a key mechanism for delivering coding gain, or sensitivity, in 2G and 3G cellular handsets. A further development of convolutional encoding, called *turbo coding*, is proposed for 3G handsets. Here two or more convolutional encoders are used together with interleaving and puncturing to increase coding distance. Convolutional encoders are effectively implemented as shift registers. They are similar in terms of implementation to the PN code generators used to create long codes for IMT2000DS and IMT2000MC.

In IMT2000DS long codes are used to provide node B-to-node B selectivity on the downlink and user-to-user selectivity on the uplink (covered in detail in Chapter 3).

Both techniques exploit digital domain processes to deliver distance. Convolutional encoders deliver distance between 0s and 1s (sensitivity). PN code generation delivers distance between parallel code streams (selectivity).

There is thus commonality in terms of processing between realizing convolutional encoders/decoders and CDMA code generation (long codes and OVSF codes). Both are designed to be adaptive. You can move in real time from a 1/2 encoder to a 2/3 encoder, and you can move in real time between multiple long codes and variable-length orthogonal variable spreading factor (OVSF) codes, depending on the information to be transmitted and the channel conditions.

Fourth-generation handsets may also use *trellis coding*. Trellis coding is used presently in some satellite systems using higher-level modulation where the pulse and amplitude states are close together—for example, 64-level QAM or higher., in trellis coded modulation schemes where modulation states are close together (A), they are channel coded for maximum distance; where they are far apart (B), they are channel coded for minimum distance. This delivers a significant increase in E_b/N_o performance but at significant cost in terms of DSP processor power and power budget; this was practical in 4G handsets but is not practical today.

We have already described the transition to higher-level modulation methods. As the number of modulation states increases, the requirement for linearity increases. In subsequent chapters we explore the role of the digital signal processor in delivering linearity *and* power efficiency by adapting and predistorting waveforms prior to modulation. DSPs therefore allow us to deliver performance gains both in terms of throughput (higher-level modulation), robustness (channel coding), voice, image and video quality (source coding), and flexibility (the bandwidth on demand and multiple per-user traffic streams available from CDMA).

Summary

Over the past 100 years, a number of key enabling technologies have helped deliver year-on-year performance gains from wireless devices—the development of valve technology and tuned circuits in the first half of the twentieth century, the development of transistors and printed circuit boards from the 1950s onward, the development of microcontrollers in the 1970s (making possible the first generation of frequency-agile analog cellular phones), and more recently, the development of ever more powerful digital signal processors and associated baseband processing devices.

In terms of RF performance, as RF component selectivity, sensitivity, and stability has improved, we have been able to move to higher frequencies, realizing a reduction in the size of resonant components and providing access to an increased amount of bandwidth.

Two-way radio design in the latter half of the twentieth century moved to progressively narrower RF channel spacing. Conversely, cellular networks have moved progressively from 25 or 30 kHz spacing to 1.25 MHz or 5 MHz, with selectivity increasingly being delivered at baseband. This has resulted in simpler RF channelization, though the need to support backward compatibility has resulted in some significant design and implementation challenges, encompassing not only multiple modes (multimode AMPS/TDMA, AMPS/CDMA, GSM/IMT dual-mode processing) but also multiple bands (800, 900, 1800, 1900, and 2100 MHz), both paired and unpaired.

There are various evolutionary migration paths for existing TDMA and CDMA technologies, but at present 5 MHz RF channel spacing is emerging as a reasonably common denominator for 3G handset design. The choice of 5 MHz has made possible the design of handsets delivering variable bit rate—supporting a ratio of 64 to 1 between the highest and lowest bit rates—multiplexed as multiple traffic streams and modulated onto a relatively constant quality radio channel. Variable-rate constant-quality channels provide the basis for preserving information bandwidth value.

Bandwidth quality can be improved by exploiting digital coding and digital processing techniques. This technique can be used to increase throughput, to improve resilience against errors introduced by the radio channel, and to improve bandwidth flexibility.

In the next two chapters, we discuss the RF hardware requirements for a cellular handset and how RF hardware determines bandwidth quality.

A Note about Radio Channel Quality

We also mentioned in passing the Rayleigh fading experienced on the radio channel and the mechanisms we need to adopt to average out these channel impairments. These include interleaving, frequency hopping (a GSM handset must be capable of hopping every frame to one of 64 new frequencies), and equalization. Equalization is needed to correct for the time shifts introduced by the multiple radio paths that may exist between a base station and handset.

Given that radio waves travel at 300,000 km per second, in a millisecond they will have covered 300 km, or looking at it another way, 1 km of flight path equates to 3.33 μs of delay. In TDMA handsets, there needs to be a mechanism for managing multiple paths that may be 4 or 5 km longer than the direct path.

A symbol period in GSM is 3.69 μs. Therefore, a 5 km multipath will create a delayed image of the modulated bit stream 4 bits behind the direct-path component. Multipath is managed in GSM (and US TDMA) by using a training sequence embedded in the bit burst, which effectively models and allows the handset to correct for a 4- or 5-bit time shift. Given that the handset can be up to 35 km away from the base station, the handset needs to adjust for the round-trip timing delay (a round-trip delay of 70 km is equivalent to 63 symbol periods).

The timing advance in GSM is up to 64 symbols (the Tx slot is moved closer to the RX slot in the frame). As we will see in the next chapter, this can be problematic when implementing multislot handsets.

In CDMA the unique and unchanging properties of the pilot signal give the receiver an accurate knowledge of the phase and time delay of the various multipath signals. It is the task of the RAKE receiver to extract the phase and delay of the signals and to use this information to synchronize the correlator in order to align the path signals prior to combining them. This process is detailed in Chapter 3. In CDMA, the pilot channel (IS95 CDMA/IMT2000MC) or pilot symbols (W-CDMA/IMT2000DS) provide the information needed for the receiver to gain knowledge of both the phase and amplitude components of the radio signal received.

The wider the dynamic range of operation required from a handset, the harder it is to deliver channel quality. In GSM, for example, it was decided in the 1980s to support 35 km macrocells (later extended to 70 km for Australia) down to 50-meter picocells. This requires substantial dynamic range. It was also decided to support high-mobility users (up to 250 kmph). This high-mobility requirement makes it necessary to track and correct for Doppler effects in the receiver and requires substantial signaling overhead to manage handovers from cell to cell.

In GSM, 62 percent of the bandwidth available is used for channel coding and signaling overhead; only 38 percent of the allocated bandwidth is actually used to carry user data. Similarly, many decisions in 3G handset design—RAKE receiver implementation, for example—depend on the dynamic range of the operational requirement: the minimum and maximum cell radius and the mobility of the user. Channel quality (and hence bandwidth quality) is dependent on a very large number of variables.

The job of the RF and DSP designer is to make sure handsets can continue to deliver acceptable performance in all operating conditions. As we will see with GPRS, this can be difficult to achieve.

A Note about Radio Bandwidth Quality

Bandwidth quality in a radio system is normally measured in terms of bit error rate. It can also be measured in terms of frame erasure rate—the number of frames so severely errored they have to be discarded.

We have said that one of the key performance parameters we need to achieve is sensitivity. This is generally measured as static sensitivity (a stationary handset) or dynamic sensitivity (a moving handset). The reference sensitivity effectively describes the received signal level needed to achieve a particular bit error rate. For example, the conformance standard for GSM is -102 dBm of signal level to achieve a 1 in 10^3 bit error rate. Similar performance requirements are specified for high-interference conditions and severe multipath conditions. Delay spreads created by multipath will be relatively small in an urban environment (typically 5 μs) and longer in hilly terrain (typically 15 μs). Channel simulations are also established for a variety of channel conditions—for example, a rural area user traveling at 250 kmph would be an RA250 test signal. TU3 would be typically urban, a user moving at 3 kmph. Performance requirements are specified across a wide range of operational conditions.

The idea of static reference sensitivity being specified to deliver a 1 in 10^3 bit error rate is that this equates to the same voice quality achieved by an analog cellular phone assuming a 20 dB SINAD (signal to noise and distortion) ratio—the traditional performance benchmark for an analog handset.

In 3G standards, static sensitivity is specified for 1 in 10^3 and 1 in 10^6 bit error rates for a range of operational conditions. If wireless is to compete directly with wireline connectivity, the bit error rate benchmark will need to improve to 1 in 10^{10}, which is the ADSL gold standard. This will represent a significant challenge. Reducing bit error rates from 1 in 10^3 to 1 in 10^6 requires a 3 dB increase in link budget. More transmit power, more receive sensitivity, or both will be required. Additional power can be made available by increasing network density and improving handset performance.

There is no point in increasing bit rate if bit quality decreases. Bandwidth *quality* is just as important as bandwidth *quantity*.

CHAPTER 2

GPRS/EDGE Handset Hardware

In this chapter we examine the hardware requirements for a GPRS tri-band phone capable of supporting higher-level modulation techniques. We address the design issues introduced by the need to produce the following:

A multislot handset. Capable of supporting GSM (8 slots) and US TDMA (3/6 slots)

A multiband handset. 800, 900, 1800, 1900 MHz

A multimode handset. Capable of processing constant envelope GMSK modulation (GSM) and higher-level modulation with AM components (US TDMA)

We need to combine these design requirements with an overall need to minimize component count and component cost. We also must avoid compromising RF performance.

Design Issues for a Multislot Phone

The idea of a multislot phone is that we can give a user more than one channel. For instance, one slot could be supporting a voice channel, other slots could be supporting separate but simultaneous data channels, and we can give a user a variable-rate channel. This means one 9.6 kbps channel (one slot) could be expanded to eight 9.6 kbps channels (76.8 kbps), or if less coding overhead was applied, one 14.4 kbps channel could be expanded to eight 14.4 kbps channels (115 kbps). Either option is generically described as *bandwidth on demand*.

In practice, the GSM interface was designed to work with a 1/8 duty cycle. Increasing the duty cycle increases the power budget (battery drain) and increases the need for heat dissipation. Additionally, multislotting may reduce the sensitivity of the handset, which effectively reduces the amount of downlink capacity available from the base station, and selectivity—a handset working on an 8-over-8 duty cycle creates more interference than a handset working on a 1-over-8 duty cycle).

The loss of sensitivity is because the time-division duplexing (time offset between transmit and receive) reduces or disappears as a consequence of the handset using multiple transmit slots. For certain types of GPRS phone this requires reinsertion of a duplex filter, with typically a 3 dB insertion loss, to separate transmit signal power at, say, +30 dBm from a receive signal at -102 dBm or below.

Revisiting the time slot arrangement for GSM shows how this happens (see Figure 2.1). The handset is active in one time slot—for example, time slot 2 will be used for transmit and receive. The transmit and receive frames are, however, offset by three time slots, resulting in a two time slot separation between transmit and receive. The time offset disappears when multiple transmit slots are used by the handset. An additional complication is that the handset has to measure the signal strength from its serving base station and up to five adjacent base stations. This is done on a frame-by-frame basis by using the six spare time slots (one per frame) to track round the beacon channels sent out by each of the six base stations. Multislotting results in rules that have hardware implications.

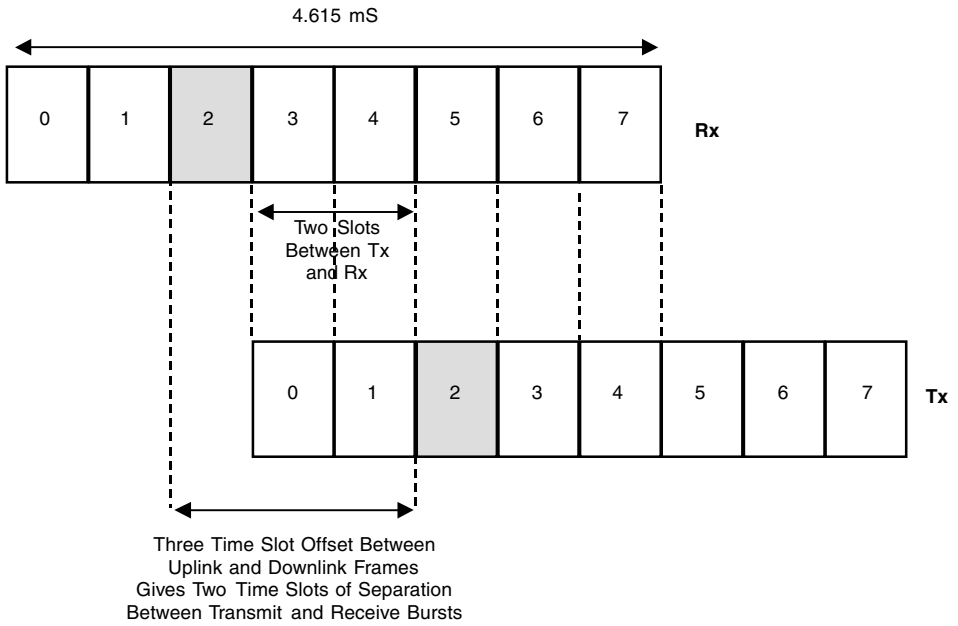


Figure 2.1 Time-division duplexing in GSM.

There are three classes of GPRS handset. Class A supports simultaneous GPRS and circuit-switched services—as well as SMS on the signaling channel—using a minimum of one time slot for each service. Class B does not support simultaneous GPRS and circuit-switched traffic. You can make or receive calls on either of the two services sequentially but not simultaneously. Class C is predefined at manufacture to be either GPRS or circuit-switched. There are then 29 (!) multislot classes.

For the sake of clarity we will use as examples just a few of the GPRS multislot options (see Table 2.1): Class 2 (two receive slots, one transmit), Class 4 (three receive slots, one transmit), Class 8 (four receive slots, one transmit), Class 10 (four receive slots, two transmit), Class 12 (four receive slots, four transmit), and as a possible long-term option, Class 18 (up to eight slots in both directions).

The maximum number of time slots Rx/Tx is fairly self-explanatory. A Class 4 handset can receive three time slots and transmit one. A Class 10 handset can receive up to four time slots and transmit up to two time slots, as long as the sum of uplink and downlink time slots does not exceed five.

The minimum number of time slots, shown in the right-hand column, describes the number of time slots needed by the handset to get ready to transmit after a receive slot or to get ready to receive after a transmit slot. This depends on whether or not the handset needs to do adjacent cell measurements. For example, T_{ta} assumes adjacent cell measurements are needed. For Class 4, it takes a minimum of three time slots to do the measurement and get ready to transmit. If no adjacent cell measurements are needed (T_{tb}), one time slot will do. T_{ra} is the number of time slots needed to do adjacent cell measurements and get ready to receive. For Class 4, this is three time slots. If no adjacent cell measurements are needed (T_{rb}), one slot will do.

The type column refers to the need for a duplex filter; Type 1 does not need a duplex filter, Type 2 does. In a Class 18 handset, you cannot do adjacent cell measurements, since you are transmitting and receiving in all time slots, and you do not have any time separation between transmit and receive slots—(hence, the need for the RF duplexer).

Table 2.1 Multislot Classes

MULTISLOT CLASS	MAX NO. OF SLOTS			MIN. NO. OF TIME SLOTS				TYPE
	RX	TX	SUM	T_{TA}	T_{TB}	T_{RA}	T_{RB}	
2	2	1	3	3	2	3	1	1
4	3	1	4	3	1	3	1	1
8	4	1	5	3	1	2	1	1
10	4	2	5	3	1	2	1	1
12	4	4	5	2	1	2	1	1
18	8	8	N/A	N/A	0	0	0	2

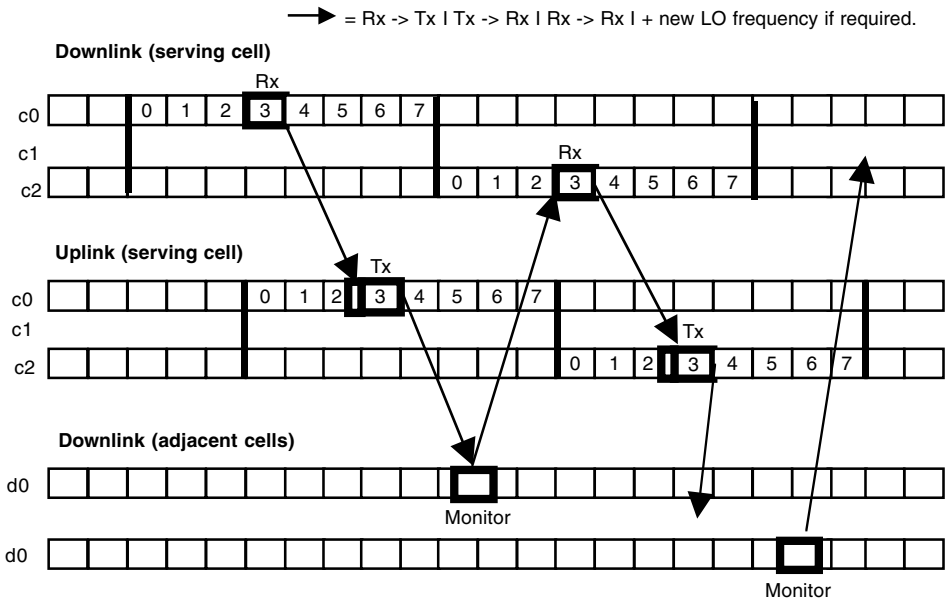


Figure 2.2 Rx/Tx offsets and the monitor channel (Rx/Tx channels shown with frequency hopping).

The uplink and downlink asymmetry can be implemented in various ways. For example, a one-slot or two-slot separation may be maintained between transmit and receive.

Rx and Tx slots may overlap, resulting in loss of sensitivity. In addition, traffic channels may be required to follow a hop sequence, across up to 64 RF 200 kHz channels but more often over 6 or 24 channels.

Figure 2.2 shows the traffic channels hopping from frame to frame and the adjacent cell monitoring being done in one spare time slot per frame. The base station beacon channel frequencies—that is, the monitor channels—do not hop. The transmit slot on the uplink may be moved closer to the Rx slot to take into account round-trip delay.

From a hardware perspective, the design issues of multislotting can therefore be summarized as:

- How to deal with the loss of sensitivity and selectivity introduced by multislotting (that is, improve Tx filtering for all handsets, add duplex filtering for Type 2)
- How to manage the increase in duty cycle (power budget and heat dissipation)
- How to manage the different power levels (slot by slot)

A user may be using different time slots for different services and the base station may require the handset to transmit at different power levels from slot to slot depending on the fading experienced on the channel (see Figure 2.3).

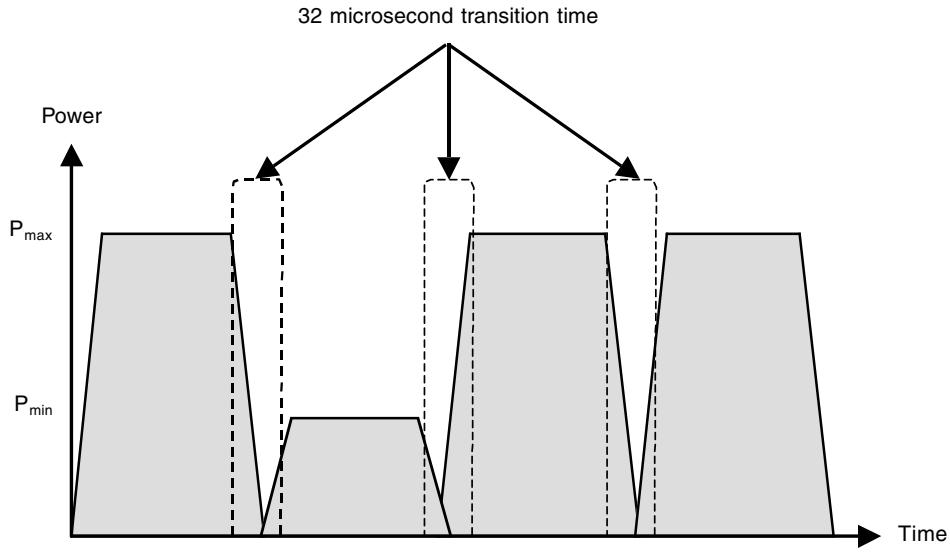


Figure 2.3 GPRS transmitter transition.

The design brief for a multislot phone is therefore:

- To find a way of maintaining or increasing sensitivity and selectivity without increasing cost or component count
- To find a way of improving power amplifier (PA) efficiency (to decrease the multislot power budget, decrease the amount of heat generated by the PA, or improve heat dissipation, or any combination of these)
- To provide a mechanism for increasing and decreasing power levels from slot to slot

Design Issues for a Multiband Phone

In Chapter 1 we described frequency allocations in the 800, 900, 1800, and 1900 MHz bands. For a handset to work transparently in Europe, the United States, and Asia, it is necessary to cover all bands (800 and 1900 MHz for the United States, 900 and 1800 MHz for Europe and Asia).

If we assume for the moment that we will be using the GSM air interface in all bands (we cover multimode in the next section), then we need to implement the frequencies and duplex spacings shown in Table 2.2.

Table 2.2 Frequency Bands and Duplex Spacing

					CHANNELS	
BAND	TOTAL BANDWIDTH	CHANNEL SPACING	TOTAL NO. OF CHANNELS	DUPLEX SPACING	HANDSET TX (MHZ)	HANDSET RX (MHZ)
GSM800	25 MHz	200 kHz	125	45 MHz	824-849	869-894
GSM900	39 MHz	200 kHz	195	45 MHz	876-915	921-960
GSM1800	75 MHz	200 kHz	375	95 MHz	1710-1785	1805-1880
GSM1900	60 MHz	200 kHz	300	80 MHz	1850-1910	1930-1990
Total	**199 MHz**		**995**			

Handset outputs for GSM 800 and 900 are a maximum 2 W. A phone working on a 1/8 duty cycle will be capable of producing an RF burst of up to 2 W, equivalent to a handset continuously transmitting at 250 mW. Multislotting increases the power budget proportionately. Power outputs for GSM 1800 and 1900 are a maximum 1 W. A phone working on a 1/8 duty cycle is capable of producing an RF burst of 1 W, equivalent to a handset continuously transmitting a maximum of 125 mW.

The handsets need to be capable of accessing any one of 995 × 200 kHz Rx or Tx channels at a duplex spacing of 45, 80, or 95 MHz across a frequency range of 824 to 1990 MHz. Present implementations address GSM 900 and 1800 (for Europe and Asia) and GSM 1900 (for the United States). GSM 800, however, needs to be accommodated, if possible, within the design architecture to support GSM GPRS EDGE phones that are GAIT-compliant (GAIT is the standards group working on GSM ANSI Terminals capable of accessing a GSM-MAP network) and ANSI 41 (US TDMA) networks.

The design brief is to produce a tri-band (900/1800/1900) or quad band (800/900/1800/1900) phone delivering good sensitivity and selectivity across all channels in all (three or four) bands while maintaining or reducing RF component cost and component count.

Design Issues for a Multimode Phone

In addition to supporting multiple frequency bands, there is a perceived—and actual—market need to accommodate multiple modulation and multiplexing techniques. In other words, the designer needs to ensure a handset is capable of modulating and demodulating GSM GMSK (a two-level constant envelope modulation technique), GSM EDGE (8 level PSK, a modulation technique with amplitude components), π/4DQPSK (the four-level modulation technique used in US TDMA) and possibly, QPSK, the four-level modulation technique used on US CDMA. GMSK only needs a Class C amplifier, while eight-level PSK, π/4DQPSK, and QPSK all require substantially more linearity. (If AM components pass through a nonlinear amplifier, spectral regrowth occurs; that is, sidebands are generated.)

Implicitly this means using Class A/B amplifiers (20 to 30 percent efficient) rather than Class C amplifiers (50 to 60 percent efficient), which increases the power budget problem and heat dissipation issue. Alternatively, amplifiers need to be run as Class C when processing GMSK, Class A/B when processing nonconstant envelope modulation, or some form of baseband predistortion has to be introduced so that the RF platform becomes modulation-transparent. (RF efficiency is maintained but baseband processor overheads increase.)

The Design Brief for a Multislot, Multiband, Multimode Phone

We can now summarize the design objectives for a GSM multislot GPRS phone capable of working in several (three or four) frequency bands and capable of supporting other modulation techniques, such as non-constant envelope, and multiplexing (TDMA or CDMA) options.

Multiband Design Objectives:

- Design an architecture capable of receiving and generating discrete frequencies across four frequency bands (a total of 995 × 200 kHz channels) at duplex spacings of 45, 80, or 95 MHz while maintaining good frequency and phase stability. To maintain sensitivity, all handsets must be capable of frequency hopping from frame to frame (217 times a second).
- Maintain or improve sensitivity and selectivity without increasing component count or cost.

Multislot Design Objectives:

- Manage the increase in duty cycle and improve heat dissipation.
- Manage power-level differences slot to slot.

Multimode Design Objectives:

- Find some way of delivering power efficiency and linearity to accommodate non-constant envelope modulation.

As always, there is no single optimum solution but a number of options with relative merits and demerits and cost/performance/complexity trade-offs.

Receiver Architectures for Multiband/Multimode

The traditional receiver architecture of choice has been, and in many instances continues to be, the *superheterodyne*, or "superhet." The principle of the superhet, invented by Edwin Armstrong early in the twentieth century, is to take the incoming received signal and to convert it, together with its modulation, down to a lower frequency—the intermediate frequency (IF), where channel selective filtering and most of the gain is performed. This selected, gained up channel is then demodulated to recover the baseband signal.

Because of the limited bandwidth and dynamic range performance of the superhet stages prior to downconversion, it is necessary to limit the receiver front-end bandwidth. Thus, the antenna performance is optimized across the band of choice, preselect filters have a similar bandwidth of design, and the performance and matching efficiencies of the low-noise amplifier (LNA) and mixer are similarly tailored.

For GSM GPRS, the handset preselect filters have a bandwidth of 25 MHz (GSM800), 39 MHz (GSM900), 75 MHz (GSM1800), or 60 MHz (GSM1900). The filter is used to limit the RF energy to only that in the bandwidth of interest in order to minimize the risk of overloading subsequent active stages. This filter may be part of the duplex filter. After amplification by the LNA, the signal is then mixed with the local oscillator (LO) to produce a difference frequency—the IF. The IF will be determined by the image frequency positioning, the selectivity capability and availability of the IF filter, and the required LO frequency and range.

The objective of the superhet is to move the signal to a low-cost, small form factor processing environment. The preselect filter and LNA have sufficient bandwidth to

process all possible channels in the chosen band. This bandwidth is maintained to the input to the mixer, and the output from the mixer will also be wideband. Thus, a filter bandwidth of just one channel is needed to select, or pass, the required channel and reject all adjacent and nearby channels after the mixer stage. This filter is placed in the IF. In designing the superhet, the engineer has chosen the IF and either designed or selected an IF filter from a manufacturer's catalog.

The IF filter has traditionally had a bandwidth equal to the modulation bandwidth (plus practical tolerance margin) of a single channel. Because the output from the mixer is wideband to support multiple channels, it is necessary to position the wanted signal to pass through the selective IF filter. For example, if the IF filter had a center frequency of 150 MHz and the wanted channel was at 922 MHz, the LO would be set to 1072 MHz (1072-922 = 150 MHz) or 772 MHz (922-772 = 150 MHz) to translate the center of the wanted channel to the center of the IF filter. The designer must ensure that the passband of the filter can pass the modulation bandwidth without distortion. Following the selective filtering, the signal passes to the demodulator where the carrier (IF) is removed to leave the original baseband signal as sourced in the transmitter.

The IF filter and often the demodulator have traditionally been realized as electromechanical components utilizing piezoelectric material—ceramic, quartz, and so on. This approach has provided sufficient selectivity and quality of filtering for most lower-level (constant envelope) modulations, such as FM, FSK, and GMSK. However, with the move toward more complex modulation, such as π/4DQPSK, QPSK, and QAM, the performance—particularly the phase accuracy of this filter technology—produces distortion of the signal.

The second problem with this type of filter is that the parameters—center frequency, bandwidth, response shape, group delay, and so on—are fixed. The engineer is designing a receiver suitable for only one standard, for example, AMPS at 25 kHz bandwidth, IS136 at 30 kHz, GSM at 200 kHz. Using this fixed IF to tune the receiver, the LO must be stepped in channel increments to bring the desired channel into the IF.

Given the requirement for multimode phones modes with different modulation bandwidths and types, this fixed single-mode approach cannot be used. The solution is either to use multiple switched filters and demodulators or to adopt an alternative flexible approach.

The multi-filter approach increases the cost and form factor for every additional mode or standard added to the phone and does not overcome the problems of insufficient phase/delay performance in this selective component. A more cost-effective, flexible approach must be adopted.

It is the adoption of increasingly capable digital processing technology at an acceptable cost and power budget that is providing a flexible design solution. To utilize digital processes, it is necessary to convert the signal from the analog domain to the digital domain.

It would be ideal to convert the incoming RF to the digital domain and perform all receive processes in programmable logic. The ultimate approach would be to convert the whole of the cellular RF spectrum (400 MHz to 2500 MHz) in this way and to have all standards/modes/bands available in a common hardware platform—the so-called software radio. The capability to convert signals directly at RF—either narrowband or wideband—to the digital domain does not yet exist. The most advanced analog-to-digital

converters (ADCs) cannot yet come near to this target. To configure a practical cost effective receiver, the ADC is positioned to sample and digitize the IF; that is, the conventional downconverting receiver front end is retained.

The receiver design engineer must decide the IF frequency and the IF bandwidth to be converted. In the superhet architecture, the higher the IF that can be used, the easier the design of the receiver front end. However, the higher the frequency to be converted, the higher the ADC power requirement.

If an IF bandwidth encompassing all channels in the band selected could be digitized, the receiver front end could be a simple non-tuning downconverter with channel selection being a digital baseband function. This is a viable technique for base station receivers where power consumption is less of an issue; however, for handsets, the ADC and DSP power required restricts the approach to digitization of a single-channel bandwidth.

This then returns us to single-channel passband filtering in the analog IF prior to digitization—a less than ideal approach for minimum component multimode handsets. However, minimum performance IF filters could be employed with phase compensation characteristics programmed into the digital baseband filtering to achieve overall suitability of performance.

Another possible approach is to use a single IF selective filter but with a bandwidth suitable for the widest mode/standard to be used. For W-CDMA, this would be 5 MHz. The 5 MHz bandwidth would then be digitized. If it was required to work in GSM mode, the required 200 kHz bandwidth could be produced in a digital filter. This approach needs careful evaluation. If the phone is working predominantly in GSM mode, the sampling/digitizing process is always working at a 5 MHz bandwidth. This will consume considerably more power than a sampling system dimensioned for 200 kHz.

So, in summary, the base station may use a wideband downconverter front end and sampling system with baseband channel tuning, but the handset will use a tunable front end with single-channel sampling and digital demodulation. The required number of converter bits must also be considered.

Again, the power consumption will be a key-limiting parameter, given the issues of input (IF) frequency and conversion bandwidth. The number of bits (resolution) equates directly to the ADC conversion or quantization noise produced, and this must be small compared with the carrier-to-noise ratio (CNR) of the signal to be converted. In a GSM/GPRS receiver, 8 to 10 bits may be necessary. In a W-CDMA receiver, since the IF CNR is considerably worse (because of the wideband noise created signal), 6 or even 4 bits may be sufficient.

In a mobile environment, the received signal strength can vary by at least 100 dB, and if this variability is to be digitized, an ADC of 20 bits plus would be required. Again, at the required sample rates this is impractical—the dynamic range of the signal applied to the ADC must be reduced. This reduction in dynamic range is achieved by the use of a variable-gain amplifier (VGA) before the ADC. Part of the digital processing function is to estimate the received signal strength and to use the result to increase or decrease the gain prior to the ADC.

This process can be applied quite heavily in the handset, since it is required to receive only one signal. However, in the base station, it is required to receive strong and weak signals simultaneously, so dynamic range control is less applicable. In 3G networks, aggressive power control also assists in this process. We consider further issues of the IF sampled superhet in node B design discussions in Chapter 11.

Direct Conversion Receivers

In this section we consider an alternative architecture to the superhet—the direct conversion receiver (DCR). Direct conversion receivers, also referred to as zero IF (ZIF), were first used in amateur radio in the 1950s, then HF receivers in the 1960s and 1970s, VHF pagers in the 1980s, 900 MHz/1800 MHz cordless and (some) cellular phones in the 1990s, and in GPRS and 3G designs today.

The superhet is a well-tried and -tested approach to receiver implementation, having good performance for most applications. However, it does have some disadvantages:

- It requires either additional front-end filters or a complex image reject mixer to prevent it from receiving two frequencies simultaneously—the wanted frequency and an unwanted frequency (the image frequency).
- If multiple bandwidths are to be received, multiple IF filters may be required.
- The digital sampling and conversion is performed at IF and so will require functions to work at these frequencies—this can require considerable current as the design frequency increases.

The DCR is directed at overcoming these problems. The principle is to inject the LO at a frequency equal to the received signal frequency. For example, if a channel at 920 MHz was to be received, the LO would be injected into the mixer at 920 MHz.

The mixer would perform the same function as in the superhet and output the difference of the signal and the LO. The output of the mixer, therefore, is a signal centered on 0 Hz (DC) with a bandwidth equal to the original modulation bandwidth. This brings in the concept of negative frequency (see Figure 2.4).

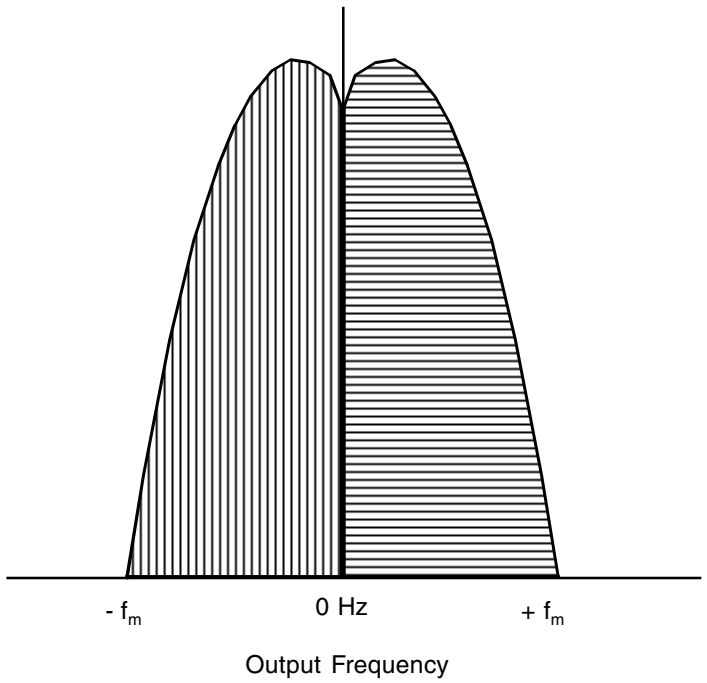


Figure 2.4 The negative frequency.

If the signal is obtained simply as the output of the mixer, it is seen as a conventional positive-only frequency where the lower sideband has been folded onto the upper sideband—the energies of the two sidebands are inseparable. To maintain and recover the total signal content (upper and lower sidebands), the signal must be represented in terms of its phase components.

To represent the signal by its phase components, it is necessary to perform a transform (Hilbert) on the incoming signal. This is achieved by splitting the signal and feeding it to two mixers that are fed with sine and cosine LO signals. In this way an in-phase (I) and quadrature phase (Q) representation of the signal (at baseband) is constructed. The accuracy or quality of the signal representation is dependent on the I and Q arm balance and the linearity of the total front end processing (see Figure 2.5).

Linearity of the receiver and spurious free generation of the LO is important, since intermodulation and distortion products will fall at DC, in the center of the recovered signal, unlike the superhet where such products will fall outside the IF. Second-order distortion will rectify the envelope of an amplitude modulated signal—for example, QPSK, π/4DQPSK, and so on to produce spurious baseband spectral energy centered at DC. This then adds to the desired downconverted signal.

It is particularly serious if the energy is that of a large unwanted signal lying in the receiver passband. The solution is to use balanced circuits in the RF front end, particularly the mixer, although a balanced LNA configuration will also assist (see Figure 2.6).

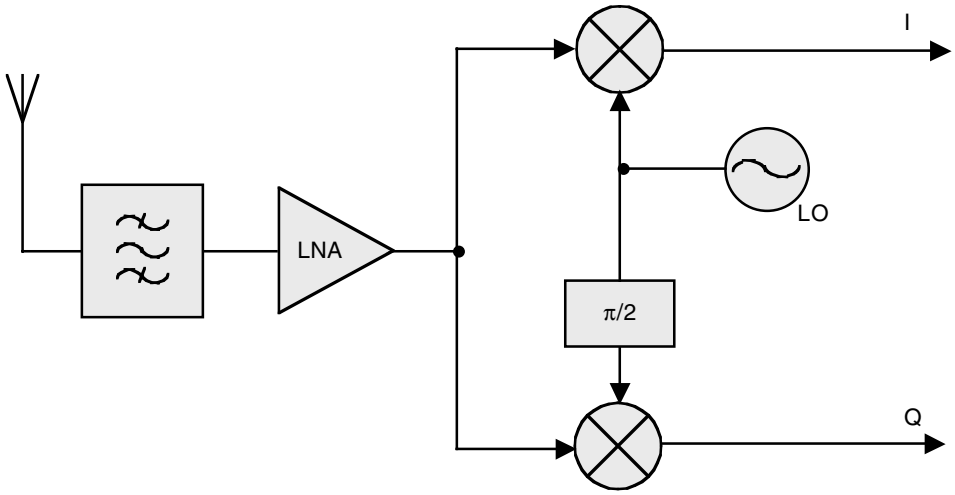


Figure 2.5 I and Q balancing.

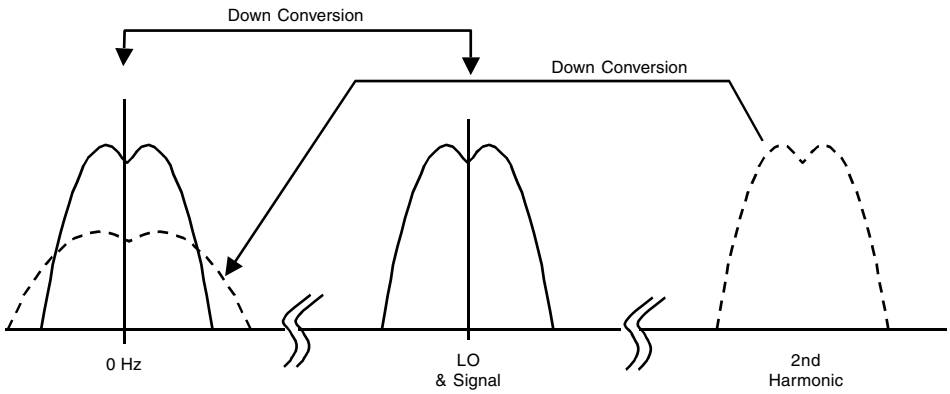


Figure 2.6 Even harmonic distortion.

If the balancing is optimum, even order products will be suppressed and only odd products created. However, even in a balanced circuit, the third harmonic of the desired signal may downconvert the third LO overtone to create spurious DC energy, adding to the fundamental downconverted signal. In the superhet, this downconverted component lies in the stopband of the IF filter.

Although the even and odd order terms may themselves be small, if the same intermodulation performance as the superhet is to be achieved, the linearities must be superior. As circuit balance improves, the most severe problem remaining is that of DC offsets in the stages following the mixer. DC offsets will occur in the middle of the downconverted spectrum, and if the baseband signal contains energy at DC (or near DC) distortions/offsets will degrade the signal quality, and SNR will be unacceptably low.

The problem can have several causes:

- Transistor mismatch in the signal path between the mixer and the I and Q inputs to the detector.
- The LO, passing back through the front end circuits (as it is on-frequency) and radiating from the antenna then reflects from a local object and reenters the receiver. The re-entrant signal then mixes with the LO, and DC terms are produced in the mixer (since $\sin^2$ and $\cos^2$ functions yield DC terms).
- A large incoming signal may leak into the LO port of the mixer and as in the previous condition self convert to DC.

The second and third problems can be particularly challenging, since their magnitude changes with receiver position and orientation.

DCRs were originally applied to pagers using two-tone FSK modulation. In this application DC offsets were not a problem because no energy existed around DC—the tones were + and -4.5 kHz either side of the carrier. The I and Q outputs could be AC coupled to lose the DC offsets without removing significant signal energy.

In the case of GSM/GPRS and QPSK, the problem is much more acute, as signal energy peaks to DC. After downconversion of the received signal to zero IF, these offsets will directly add to the peak of the spectrum. It is no longer possible to null offsets by capacitive coupling of the baseband signal path, because energy will be lost from the spectral peak. In a 200 kHz bandwidth channel with a bit error rate (BER) requirement of 10^{-3}, a 5 Hz notch causes approximately 0.2 dB loss of sensitivity. A 20 Hz notch will stop the receiver working altogether.

It is necessary to measure or estimate the DC offsets and to remove (subtract) them. This can be done as a production test step for the fixed or nonvariable offsets, with compensating levels programmed into the digital baseband processing. Removing the signal-induced variable offsets is more complex. An example approach would be to average the signal level of the digitized baseband signal over a programmable time window. The time averaging is a critical parameter to be controlled in order to differentiate dynamic amplitude changes that result from propagation effects and changes caused by network effects, power control, traffic content, and so on.

Analog (RF) performance depends primarily on circuit linearity usually achieved at device level; however, this is a demanding approach both in power and complexity, and compensation at system level should be attempted. Baseband compensation is generally achieved as part of the digital signal processing and hence more easily achieved. Using a basic DCR configuration, control and compensation options may be considered (see Figure 2.7).

Receiver gain must be set to feed the received signal linearly to the ADC over an 80- to 90-dB range. Saturation in the LPF, as well as other stages, will unacceptably degrade a linear modulation signal—for example, QPSK, QAM, and $\pi/4$DQPSK. To avoid this problem, gain control in the RF and baseband linear front-end stages is employed, including the amplifier, mixer, and baseband amplifiers.

The front-end filter, or preselector, is still used to limit the RF bandwidth energy to the LNA and mixers, although since there is now no image, no other RF filters are required. Selectivity is achieved by use of lowpass filters in the I and Q arms, and these may be analog (prior to digital conversion) or digital (post digital conversion). The principle receive signal gain is now in the IQ arms at baseband, which can create the difficulty of high low-frequency noise, caused by flicker effects or $1/f$.

Another increasingly popular method of addressing the classic DCR problems is to use a low IF or near-zero IF configuration. Instead of injecting the LO on a channel, it is set to a small offset frequency. The offset is design-dependent, but a one- or two-channel offset can be used or even a noninteger offset. The low IF receiver has a frequency offset on the I and Q and so requires the baseband filter to have the positive frequency characteristic different from the negative frequency characteristic (note that conventional filter forms are symmetrical). This is the polyphase filter. As energy is shifted away from 0 Hz, AC coupling may again be used, thus removing or blocking DC offsets and low-frequency flicker noise.

This solution works well if the adjacent channel levels are not too much higher than the wanted signal, since polyphase filter rejection is typically 30 to 40 dB.

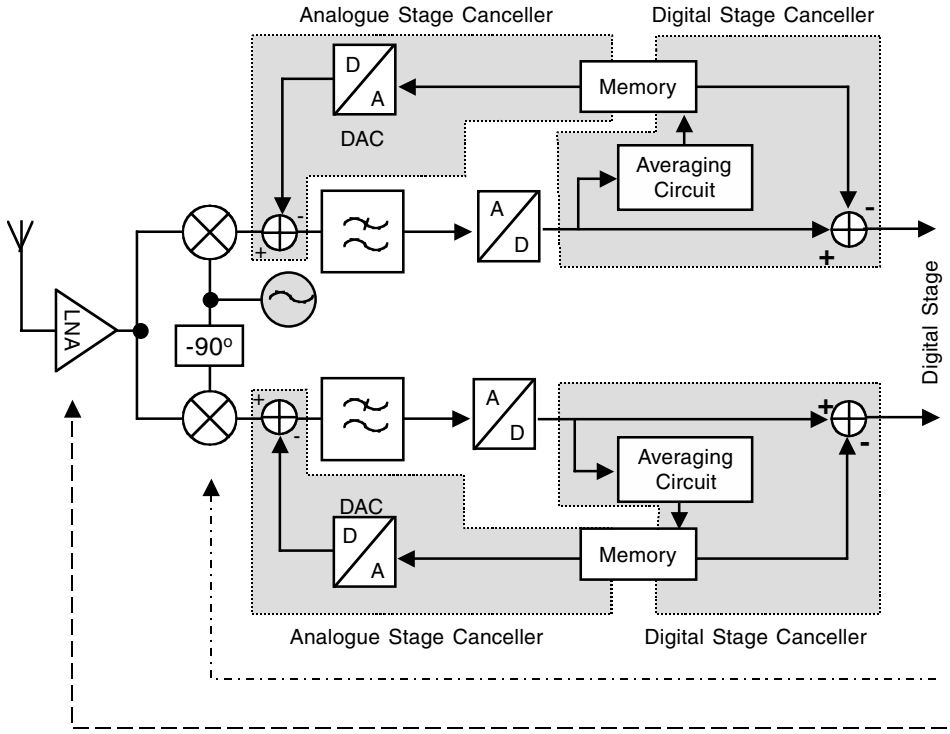


Figure 2.7 Direct conversion receiver control and compensation circuits.

To Sum Up

Direct conversion receivers provide an elegant way of reducing component count, and component cost and increased use of baseband processing have meant that DCR can be applied to GPRS phones, including multiband, multimode GPRS. Careful impedance matching and attention to design detail means that some of the sensitivity and selectivity losses implicit in GPRS multislotting can be offset to deliver acceptable RF performance. Near-zero IF receivers, typically with an IF of 100 kHz (half-channel spacing) allow AC coupling to be used but require a more highly specified ADC.

Transmitter Architectures: Present Options

We identified in Chapter 1 a number of modulation techniques, including GMSK for GSM, π/4DQPSK for IS54 TDMA, 8 PSK for EDGE, and 16-level QAM for CDMA2000 1 × EV. To provide some design standardization, modulation is usually achieved through a vector IQ modulator. This can manage all modulation types, and it separates the modulation process from the phase lock loop synthesizer—the function used to generate specific channel frequencies.

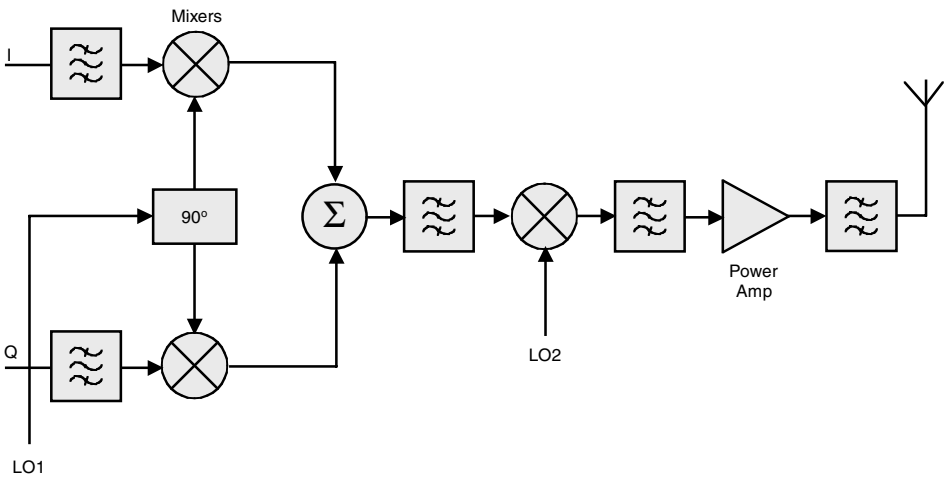


Figure 2.8 Upconverting transmitter.

As in the case of the superhet receiver, the traditional transmitter architecture has consisted of one or more upconversion stages from the frequency generating/modulating stage to the final PA. This approach allows a large part of the signal processing, amplification, filtering, and modulation to be performed at lower, cost-effective, highly integrated stages (see Figure 2.8). The design approach is low risk with a reasonably high performance.

The disadvantage of this approach is that a large number of unwanted frequencies are produced, including images and spurii, necessitating a correspondingly large number of filters. If multiband/multimode is the objective, the number of components can rise rapidly. Image reject mixers can help but will not remove the filters entirely. A typical configuration might use an on-chip IF filter and high local oscillator injection. With careful circuit design, it may be possible, where transmit and receive do not happen simultaneously, to reduce the number of filters by commoning the transmit and receive IFs.

Issues to Resolve

One problem with the architectures considered so far is that the final frequency is generated at the start of the transmitter chain and then gained up through relatively (tens of MHz) wideband stages. The consequence of this is to cause wideband noise to be present at the transmitter output stage—unless filters are inserted to remove it. The noise will be a particular problem out at the receive frequency, a duplex spacing away. (This noise will radiate and desensitize adjacent handsets).

To attenuate the far out noise, filters must be added before and after the PA, and so the duplex filter has been retained. Again, in a multiband design this can increase the number of filters considerably.

To remove the need for these filters, an architecture called the *offset loop* or *translational loop transmitter* has been developed. This relies on using the noise-reducing

bandwidth properties of the PLL. Essentially the PLL function is moved from the start of the transmitter to the output end, where the noise bandwidth becomes directly a function of the PLL bandwidth.

In a well-designed, well-characterized PLL, the wideband noise output is low. If the PLL can be implemented without a large divider ratio (N) in the loop, then the noise output can be reduced further. If such a loop is used directly in the back end of the transmitter, then the filters are not required.

A typical configuration will have a VCO running at the final frequency within a PLL. To translate the output frequency down to a reference frequency, a second PLL is mixed into the primary loop. Tuning is accomplished by tuning the secondary loop, and in this example, modulation is applied to the sampling frequency process. If there are no dividers, modulation transfer is transparent. A number of critical RF components are still needed, however. For example, the tuning and modulation oscillators require resonators, and 1800 MHz channels need to be produced by a doubler or band-switched resonators.

Figure 2.9 shows a similar configuration, but using dividers, for a multiband implementation (single-band, dual-band, or tri-band GSM). Modulation is applied to a PLL with the VCO running at final frequency. Again, this reduces wideband noise sufficiently to allow the duplex filter to be replaced with a switch. Because of the lack of up-conversion, there are no image products, so no output bandpass filters are required.

The advantage of this implementation is that it reduces losses between the transmit power amplifier and the antenna and allows the RF power amplifier to be driven into saturation without signal degradation. The loop attempts to track out the modulation, which is introduced as a phase error and so transfers the modulation onto the final frequency Tx VCO. Channel selection is achieved by tuning the offset oscillator, which doubles as the first local oscillator in receive mode.

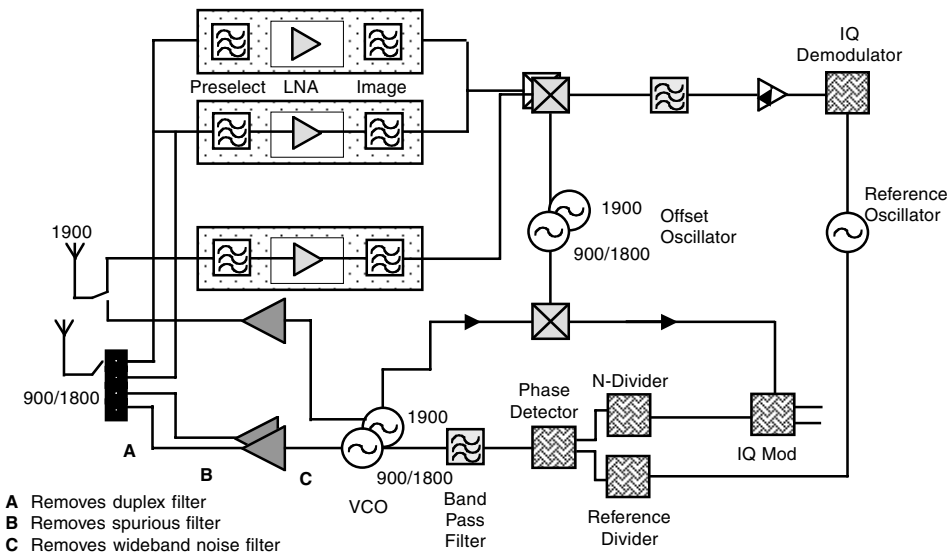


Figure 2.9 Multiband GSM 900/1800/1900 MHz.

There are a number of implementation challenges in an OPLL design:

- The noise transmitted in the receive band and modulation accuracy is determined by the closed-loop performance of the OPLL. If the loop bandwidth is too narrow, modulation accuracy is degraded; if the loop bandwidth is too wide, the receive band noise floor rises.
- Because the OPLL processes phase and can only respond to phase lock functions for modulation, an OPLL design is unable to handle modulation types that have amplitude components. Thus, it has only been applied to constant envelope modulations (for example, FM, FSK, and GMSK).
- Design work is proceeding to apply the benefits of the OPLL to non-constant envelope modulation to make the architecture suitable for EDGE and QPSK. Approaches depend mainly on using the amplitude limiting characteristics of the PLL to remove the amplitude changes but to modulate correctly the phase components and then to remodulate the AM components back onto the PA output.

  With the above options, the advantage of having a simple output duplex switch (usually a GaAs device) is only available when nonsimultaneous transmit/receive is used. When higher-level GPRS classes are used, the duplex filter must be reinstated.

A number of vendors are looking at alternative ways to manage the amplitude and phase components in the signal path. For illustration purposes, we'll look at an example from Tropian (www.tropian.com). The Tropian implementation uses a core modulator in which the amplitude and phase paths are synchronized digitally to control timing alignment and modulation accuracy (see Figure 2.10). The digital phase and amplitude modulator is implemented in CMOS and the RF PA in GaAs MOSFET.

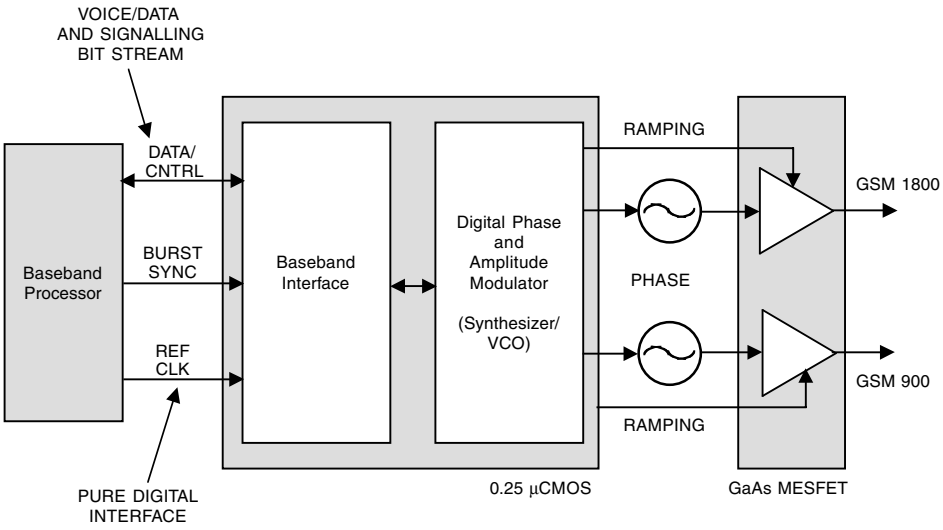


Figure 2.10 Tropian transmitter handset system block diagram.

We revisit linearization and adaptive predistortion techniques again when we study base station hardware implementation in Chapter 11. In the last part of this chapter we focus on the remaining design brief areas for achieving a multiband, multislot, multimode handset.

GPRS RF PA

There are two particular issues with GPRS RF PA design:

- The duty cycle can change from 1/8 to 8/8.
- Power levels can change between time slots.

The need to improve power efficiency focuses a substantial amount of R&D effort on RF materials. Table 2.3 compares present options on the basis of maturity, cost per Watt, whether or not the processes need a negative voltage, and power-added efficiencies at 900/1800 MHz.

Given that power-added efficiencies better than 55 percent are very hard to achieve in practice (even with Class C amplifiers), heat dissipation is a critical issue and has led to the increased use of copper substrates to improve conductivity. Recently, silicon germanium has grown in popularity as a material for use at 1800 MHz/2 GHz, giving efficient gain and noise performance at relatively low cost.

Implementing multislot GPRS with modulation techniques that contain amplitude components will be substantially harder to achieve—for example, the (4DQPSK modulation used in IS54TDMA or the eight-level PSK used in EDGE.

Table 2.3 Device Technology Comparison

DEVICE TECHNOLOGY	PAE1 (%) (4.8V, 900 MHZ)	PAE2 (%) (3V, 1.8 GHZ)	MATURITY	SINGLE BIAS SUPPLY	COST PER WATT
Si BJT	60-70	20-30	Mature	Yes	Low
Si MOSFET	40-60	15-25	Mature	Yes	Lowest
SiGe HBT	60-70	40-50	Early days	Yes	Potentially low
GaAs MESFET	60-70	45-55	Mature	No	Moderate
GaAs P-HEMT	60-70	50-60	Becoming mature	Possibly	High
GaAs HBT	60-70	45-55	Becoming mature	Yes	High

In a practical phone design, linearity is always traded against DC power consumption. Factors that decide the final position on this linearity/efficiency trade-off include the following:

- The semiconductor material, Si, GaAs, SiGe, and so on
- The transistor construction, packaging technique, bond wire inductances, and so on
- The number of components used in and around the power amplifiers
- The expertise and capability of the design engineer

Designers have the option of using discrete devices, an integrated PA, or a module. To operate at a practical power efficiency/linearity level, there will inevitably be a degree of nonlinearity in the PA stage. This in turn makes the characterization of the parameters that influence efficiency and linearity difficult to measure and difficult to model. Most optimization of efficiency is carried out by a number of empirical processes (for example, load line characterization, load pull analysis, and harmonic shorting methods).

Results are not always obvious—for instance, networks matching the PA output to 50 ohms are frequently configured as a lowpass filter that attenuates the nonlinearities generated in the output device. The PA appears to have a good—that is, low—harmonic performance. The nonlinearity becomes evident when a non-constant envelope (modulated) signal is applied to the PA as intermodulation products and spectral spreading are seen.

Until recently the RF PA was the only function in a 2 GHz mobile phone where efficiency and linearity arguments favored GaAs over silicon. The situation is changing. Advances in both silicon (Si) and silicon germanium (SiGe) processes, especially in 3G phone development, make these materials strong contenders in new designs.

Higher performance is only obtained through attention to careful design. Advanced design techniques require advanced modeling/simulation to obtain the potential benefits. Design implementation is still the major cause of disappointing performance. In Chapter 11 we examine GPRS base station and 3G Node B design, including linearization techniques, power matching, and related performance optimization techniques.

Manage Power-Level Difference Slot to Slot

The power levels and power masks are described in GSM 11.10-1. Compliance requires the first time slot be set to maximum power (P_{MAX}) and the second time slot to minimum power, with all subsequent slots set to maximum. P_{MAX} is a variable established by the base station and establishes the maximum power allowed for handset transmit. The handset uses received power measurements to calculate a second value and transmits with the lower one.

This open-loop control requires a close, accurate link between received signal and transmitted power, which in turn requires careful calibration and testing during production. All TDMA transmissions (handset to base, base to handset) require transmit burst shaping and power control to maintain RF energy within the allocated time slot. The simplest form of power control is to use an adjustable gain element in the transmitter amplifying chain. Either an in-line attenuator is used (for example, PIN diode), a variable gain driver amplifier, or power rail control on the final PA.

The principal problem with this open-loop power control is the large number of unknowns that determine the output power—for example, device gains, temperature, variable loading conditions, and variable drive levels. To overcome some of the problems in the open-loop system, a closed-loop feedback may be used.

The power leveling/controlling of RF power amplifiers (transmitter output stages) is performed by tapping off a small amount of the RF output power, feeding it to a diode detector (producing a DC proportional to the RF energy detected), comparing the DC obtained with a reference level (variable if required), and using the comparison output to control the PA and PA driver chain gain.

RF output power control can be implemented using a closed-loop approach. The RF power is sampled at the output using a directional coupler or capacitive divider and is detected in a fast Schottky diode. The resultant signal representing the peak RF output voltage is compared to a reference voltage in an error amplifier. The loop controls the power amplifier gain via a control line to force the measured voltage and the reference voltage to be equal.

Power control is accomplished by changing the reference voltage. Although straightforward as a technique, there are disadvantages:

- The diode temperature variation requires compensation to achieve the required accuracy. The dynamic range is limited to that of the detector diode (approximately 20 dB—without compensation).
- Loop gain can vary significantly over the dynamic range, causing stability problems.
- Switching transients are difficult to control if loop bandwidth is not constant.

An alternative control mechanism can be used with amplifiers employing square law devices (for example, FETs). The supply voltage can be used to control the amplifier's output power. The RF output power from an amplifier is proportional to the square of the supply voltage. Reducing the drain voltage effectively limits the RF voltage swing and, hence, limits the output power. The response time for this technique is very fast, and in the case of a square-law device, this response time is voltage-linear, for a constant load.

The direct diode detection power control system has been satisfactory for analog cellular systems and is just satisfactory for current TDMA cellular systems (for example, GSM, IS54 TDMA, and PDC), although as voltage headrooms come down (4.8 V to 3.3 V to 2.7 V), lossy supply control becomes unacceptable.

CDMA and W-CDMA require the transmitter power to be controlled more accurately and more frequently than previous systems. This has driven R&D to find power control methods that meet the new requirements and are more production-cost-effective.

Analog Devices, for example, have an application specific IC that replaces the traditional simple diode detector with an active logarithmic detector. The feedback includes a variable gain single pole low pass filter with the gain determined by a multiplying digital-to-analog converter (DAC) The ADC is removed. A switched RF attenuator is added between the output coupler and the detector, and a voltage reference source is added. Power control is achieved by selecting/deselecting the RF attenuator and adjusting the gain of the LPF by means of the DAC

The system relies on the use of detecting log amps that work at RF to allow direct measurement of the transmitted signal strength over a wide dynamic range. Detecting

log amps have a considerable application history in wide dynamic range signal measurement—for example, spectrum analyzers. Recently the implementation has improved and higher accuracy now results from improvements in their key parameters: slope and intercept.

Power Amplifier Summary

TDMA systems have always required close control of burst shaping—the rise and fall of the power envelope either side of the slot burst.

In GPRS this process has to be implemented on multiple slots with significant variations in power from burst to burst. Log detector power control implementations improve the level of control available and provide forward compatibility with 3G handset requirements.

Multiband Frequency Generation

Consider that the requirement is to design an architecture capable of generating discrete frequencies across four frequency bands (a total of 995 × 200 kHz channels) at duplex spacings of 45, 80, and 95 MHz while maintaining good frequency and phase stability.

The system block used in cellular handsets (and prior generations of two-way radios) to generate specific frequencies at specific channel spacings is the frequency synthesizer—the process is described as frequency synthesis. PLLs have been the dominant approach to RF frequency synthesis through 1G and 2G systems and will continue to be the technology of choice for RF signal generation in 2.5G and 3G applications.

The synthesizer is required to generate frequencies across the required range, increment in channel sized steps, move rapidly from one channel to another (support frequency hopping and handover), and have a low-noise, distortion-free output.

In 1G systems the network protocols allowed tens of milliseconds to shift between channels—a simple task for the PLL. PLLs were traditionally implemented with relatively simple integer dividers in the feedback loop. This approach requires that the frequency/phase comparison frequency is equal to the minimum frequency step size (channel spacing) required. This in turn primarily dictated the loop filter time constant—that is, the PLL bandwidth and hence the settling time.

GSM has a channel spacing of 200 kHz, and so f_{REF} is 200 kHz. But the GSM network required channel changes in hundreds of microseconds. With a reference of 200 kHz the channel switching rate cannot be met, so various speed-up techniques have been developed to cheat the time constant during frequency changes.

In parallel with this requirement, PLL techniques have been developed to enable RF signal generators and test synthesizers to obtain smaller step increments—without sacrificing other performance parameters. This technique was based on the ability to reconfigure the feedback divider to divide in sub-integer steps—the Fractional-N PLL.

This had the advantage of increasing the reference by a number equal to the fractional division, for example:

Integer-N PLL

- O/P frequency = 932 MHz, f_{REF} = 200 kHz
- N = 932 MHz/200 kHz = 4660

Fractional-N PLL

- If N can divide in 1/8ths, (that is, 0.125/0.250/0.375, etc.)
- O/P frequency = 932 MHz, f_{REF} = 200 kHz × 8 = 1.6 MHz
- N = 932 MHz/1.6 MHz = 582.5

This should have two benefits:

- The noise generated by a PLL is primarily a function of the division ratio, so reducing N should give a cleaner output, for example:

Integer-N PLL

- N = 4660
- Noise = 20log4660
- = 73.4 dB

Fractional-N PLL

- N= 582.5
- Noise = 20log582.5
- = 55 dB

for an 18.4-dB improvement.

- As the reference frequency is now 1.6 MHz, the loop is much faster and so more easily meets the switching speed/settling time requirements. However, the cost is a considerable increase in the amount of digital steering and compensation logic that is required to enable the Fractional-N loop to perform. For this reason, in many implementations the benefit has been marginal (or even negative).

Interestingly, some vendors are again offering Integer-N loops for some of the more advanced applications (for example, GPRS and EDGE) and proposing either two loops—one changing and settling while the second is outputting—or using sophisticated speed-up techniques.

A typical example has three PLLs on chip complete with VCO transistor and varactors—resonators off chip. Two PLLs are for the 900 MHz and 1.8 GHz (PLL1) and 750 MHz to 1.5 GHz (PLL2). The third PLL is for the IF/demodulator function.

Summary

The introduction of GPRS has placed a number of new demands on handset designers. Multislotting has made it hard to maintain the year-on-year performance improvements that were delivered in the early years of GSM (performance improvements that came partly from production volume—the closer control of RF component parameters). This volume-related performance gain produced an average 1 dB per year of additional sensitivity between 1992 and 1997.

Smaller form factor handsets, tri-band phones, and more recently, multislot GPRS phones have together resulted in a decrease in handset sensitivity. In general, network density today is sufficient to ensure that this is not greatly noticed by the subscriber but does indicate that GSM (and related TDMA technologies) are nearing the end of their maturation cycle in terms of technology capability. The room for improvement reduces over time.

Present GPRS handsets typically support three or four time slots on the downlink and one time slot on the uplink, to avoid problem of overheating and RF power budget in the handset.

Good performance can still be achieved either by careful implementation of multiband-friendly direct conversion receiver architectures or superhet designs with digital IF processing. Handsets are typically dual-band (900/1800 MHz) or tri-band (900/1800/1900 MHz).

In the next chapter, we set out to review the hardware evolution needed to deliver third-generation handsets while maintaining backward compatibility with existing GSM, GPRS, and EDGE designs.

CHAPTER

3

3G Handset Hardware

In the two previous chapters we identified that one of the principal design objectives in a cellular phone is to reduce component count, component complexity, and cost, and at the same time improve functionality. By *functionality* we mean dynamic range—that is, the range of operating conditions over which the phone will function—and the ability to support multiple simultaneous channels per user. We showed how GPRS could be implemented to provide a limited amount of bandwidth on demand and how GPRS could be configured to deliver, to a limited extent, a number of parallel channels, such as simultaneous voice and data. However, we also highlighted the additional cost and complexity that bandwidth on demand and multiple slots (multiple per user channel streams) introduced into a GSM or TDMA phone.

Getting Started

The general idea of a 3G air interface—IMT2000DS, TC, or MC—is to move the process of delivering sensitivity, selectivity, and stability from RF to baseband, saving on RF component count, RF component complexity, and cost, and increasing the channel selectivity available to individual users. You could, for example, support multiple channel streams by having multiple RF transceivers in a handset, but this would be expensive and tricky to implement, because too many RF frequencies would be mixing together in too small a space.

Our starting point is to review how the IMT2000DS air interface delivers sensitivity, selectivity, and stability, along with the associated handset hardware requirements.

At radio frequencies, sensitivity is achieved by providing RF separation (duplex spacing) between send and receive, and selectivity is achieved by the spacing between RF channels—for example, 25 kHz (PMR), 30 kHz (AMPS or TDMA), 200 kHz (GSM), or 5 MHz (IMT2000DS).

At baseband, the same results can be achieved by using digital filtering; instead of RF channel spacing, we have *coding distance*, the measure of how separate—that is, how far apart—we can make our 0s and 1s. The greater the distance between a 0 and a 1, the more certain we are that the demodulated digital value is correct. An increase in coding distance equates to an increase in sensitivity.

Likewise, if we take two coded digital bit streams, the number of bit positions in which the two streams differ determines the difference or distance between the two code streams. The greater the distance between code streams, the better the selectivity. The selectivity includes the separation of channels, the separation of users one from another, and the separation of users from any other interfering signal. The distance between the two codes (shown in Figure 3.1) is the number of bits in which the two codes differ (11!).

As code length increases, the opportunity for greater distance (that is, selectivity) increases. An increase in selectivity either requires an increase in RF bandwidth, or a lower bit rate.

Stability between two communicating devices can be achieved by locking two codes together (see Figure 3.2). This is used in TDMA systems to provide synchronization (the S burst in GSM is an example), with the base station providing a time reference to the handset.

In IMT2000DS, the code structure can be used to transfer a time reference from a Node B to a handset. In addition, a handset can obtain a time reference from a macro or micro site and transfer the reference to a simple, low-cost indoor picocell.

Figure 3.1 Coding distance—selectivity.

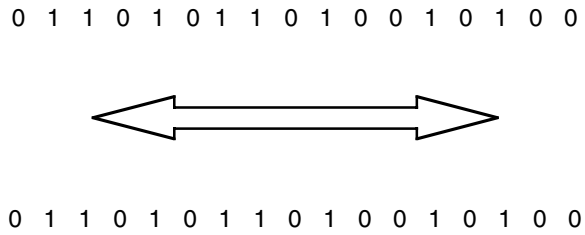


Figure 3.2 Code correlation—stability.

Code Properties

Direct-Sequence Spread Spectrum (DSSS) techniques create a wide RF bandwidth signal by multiplying the user data and control data with digital spreading codes. The wideband characteristics are used in 3G systems to help overcome propagation distortions.

As all users share the same frequency, it is necessary to create individual user discrimination by using unique code sequences. Whether a terminal has a dedicated communication link or is idle in a cell, it will require a number of defined parameters from the base station. For this reason, a number of parallel, or overlaying, codes are used (see Figure 3.3):

- Codes that are run at a higher clock, or chip, rate than the user or control data will expand the bandwidth as a function of the higher rate signal (code). These are *spreading codes*.
- Codes that run at the same rate as the spread signal are *scrambling codes.* They do not spread the bandwidth further.

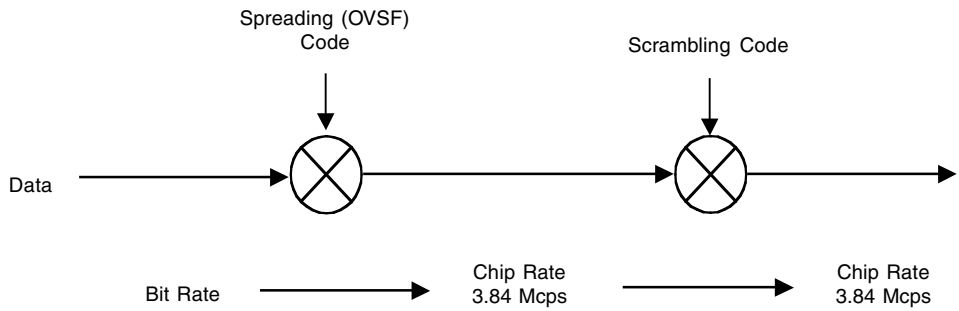


Figure 3.3 Spreading codes and scrambling codes.

The scrambling codes divide into long codes and short codes. Long codes are 38,400 chip length codes truncated to fit a 10-ms frame length. Short codes are 256 chips long and span one symbol period. On the downlink, long codes are used to separate cell energy of interest. Each Node B has a specific long code (one of 512). The handset uses the same long code to decorrelate the wanted signal (that is, the signal from the serving Node B). Scrambling codes are designed to have known and uniform limits to their mutual cross correlation; their distance from one another is known and should remain constant.

Code Properties—Orthogonality and Distance

Spreading codes are designed to be orthogonal. In a perfectly synchronous transmission, multiple codes co-sharing an RF channel will have *no* cross-code correlation; that is, they will exhibit perfect distance. The disadvantage with orthogonal codes is that they are limited in number for a given code length. Also, we need to find a family of codes that will support highly variable data rates while preserving most of their orthogonality. These are known as *Orthogonal Variable Spreading Factor* (OVSF) codes.

The *variable* is the number of symbols (also known as chips) used from the spreading code to cover the input data symbol. For a high bit rate (960 kbps) user, each data symbol will be multiplied with four spreading symbols, a 480 kbps user will have each data symbol multiplied with eight symbols, a 240 kbps user will have each data symbol multiplied with 16 symbols, and so on, ending up with a 15 kbps user having each data symbol multiplied with 256 symbols. In other words, as the input data/symbol rate increases, the chip cover decreases—and as a result, the spreading gain decreases.

The input data (encoded voice, image, video, and channel coding) comes in as a 0 or a 1 and is described digitally as a +1 or as a -1. It is then multiplied with whatever the state of the spreading code is at any moment in time using the rule set shown in Tables 3.1 and 3.2.

Table 3.2 shows the OVSF code tree. We can use any of the codes from SF4 to SF256, though with a number of restrictions, which we will discuss in a moment.

Table 3.1 Exclusive NOR

DATA	SPREADING CODE	OUTPUT
0 (-1)	0 (-1)	1 (+1)
1 (+1)	0 (-1)	0 (-1)
0 (-1)	1 (+1)	0 (-1)
1 (+1)	1 (+1)	1 (+1)

Let's take a 480 kbps user with a chip cover of eight chips per data symbol, giving a spreading factor of 8 (SF 8).

User 1 is allocated Code 8.0

Input Data Symbol

	-1							
Spreading code:	+1	+1	+1	+1	+1	+1	+1	+1
Composite code:	-1	-1	-1	-1	-1	-1	-1	-1
The despreading code will be:	+1	+1	+1	+1	+1	+1	+1	+1
The output code will be:	-1	-1	-1	-1	-1	-1	-1	-1

Output Data Symbols = -1

Effectively, we have qualified whether the data symbol is a +1 or -1 eight times and have hence increased its distance. In terms of voltage, our input data signal at -1 Volts will have become an output signal at -8 Volts when correlated over the eight symbol periods.

User 2 is allocated Code 8.3. The user's input symbol is also a -1.

Input Data Symbol

	-1							
Spreading code:	+1	+1	-1	-1	-1	-1	+1	+1
Composite code:	-1	-1	+1	+1	+1	+1	-1	-1
The despreading code will be:	+1	+1	-1	-1	-1	-1	+1	+1
The output code will be:	-1	-1	-1	-1	-1	-1	-1	-1

That is, -1 is generated for all 8 symbol states.

User 3 is allocated Code 8.5.

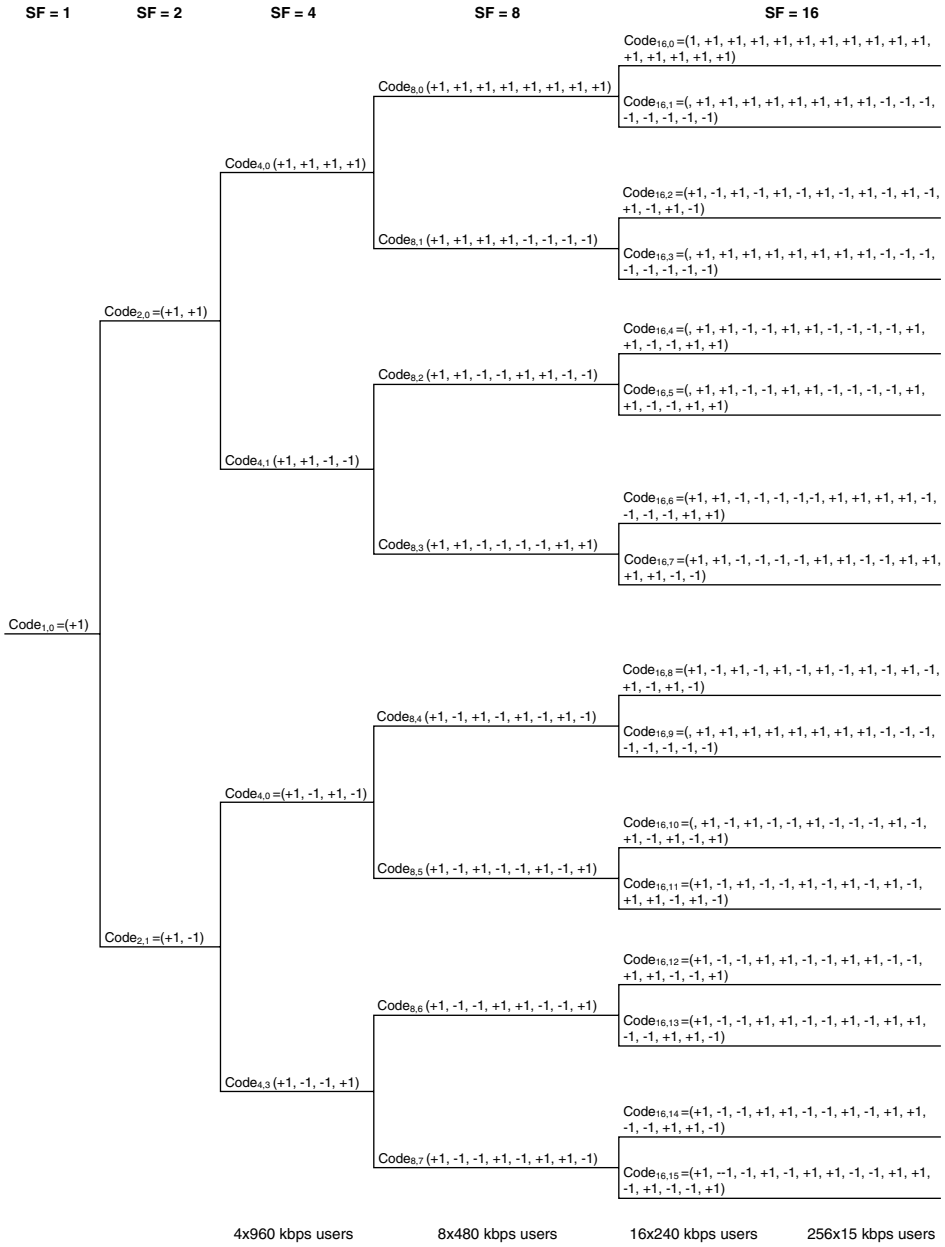
User 2's input data symbol will be exclusive NOR'd by User 3's despreading code as follows.

Input Data Symbol

	-1							
User 2's spreading code:	+1	+1	-1	-1	-1	-1	+1	+1
Composite code:	-1	-1	+1	+1	+1	+1	-1	-1
User 3's spreading code:	+1	-1	+1	-1,	-1	+1	-1	+1
The output code will be:	-1	+1	+1	-1	-1	+1	+1	-1

That is, the output is neither a +1 or a -1 but something in between. In other words, a distance has been created between User 2 and User 3 and the output stays in the noise floor.

Table 3.2 Spreading Codes



Code Capacity - Impact of the Code Tree and Non-Orthogonality

The rule set for the code tree is that if a user is, for example, allocated code 8.0, no users are allowed to occupy any of the codes to the right, since they would not be orthogonal.

A "fat" (480 kbps) user not only occupies Code 8.0 but effectively occupies 16.0 and 16.1, 32.0, 32.1, 32.2, 32.3, and so on down to 256.63. In other words, one "fat" user occupies 12.5% of all the available code bandwidth. You could theoretically have one high-bit-rate user on Code 8.0 and 192 "thin" (15 kbps) users on Code 256.65 through to Code 256.256:

- *1 × high-bit-rate user (Code 8)*—Occupies 12.5% of the total code bandwidth.
- *192 × low-bit-rate users (Code 256*)—Occupy all other codes 256.65 through 256.256.

In practice, the code tree can support rather less than the theoretical maximum, since orthogonality is compromised by other factors (essentially the impact of multipath delay on the code properties). Users can, however, be moved to left and right of the code tree, if necessary every 10 ms, delivering very flexible bandwidth on demand. These are very deterministic codes with a very simple and rigorously predefined structure. The useful property is the orthogonality, along with the ability to support variable data rates; the downside is the limited code bandwidth available.

From a hardware point of view, it is easy to move users left and right on the code tree, since it just involves moving the correlator to sample the spreading code at a faster or slower rate. On the downlink (Node B to handset), OVSF codes support individual users; that is, a single RF channel Node B (1 × 5 MHz) can theoretically support 4 high-bit-rate users (960 kbps), 256 × low-bit-rate (15 kbps) users, or any mix in between.

Alternatively, the Node B can support multiple (up to six) coded channels delivered to a single user—assuming the user's handset can decorrelate multiple code streams. Similarly, on the uplink, a handset can potentially deliver up to six simultaneously encoded code streams, each with a separate OVSF code. In practice, the peak-to-mean variation introduced by using multiple OVSF codes on the uplink is likely to prevent their use at least for the next three to five years, until such time as high degrees of power-efficient linearity are available in the handset PA.

We have said the following about the different code types:

Spreading codes. Run faster than the original input data. The particular class of code used for spreading is the OVSF code. It has very deterministic rules that help to preserve orthogonality in the presence of widely varying data rates.

Scrambling codes. Run at the same rate as the spread signal. They scramble but do not spread; the chip rate remains the same before and after scrambling. Scrambling codes are used to provide a second level of selectivity over and above the channel selectivity provided by the OVSF codes. They provide selectivity between different Node Bs on the downlink and selectivity between different users on the uplink. Scrambling codes, used in IMT2000DS, are Gold codes, a particular class of long code. While there is cross-correlation between long codes, the cross-correlation is uniform and bounded—rather like knowing that an adjacent RF channel creates a certain level of adjacent and co-channel interference. The outputs from the code-generating linear feedback register are generally configured, so that the code will exhibit good randomness to the extent that the code will appear noiselike but will follow a known rule set (needed for decorrelation). The codes are often described as Pseudo-Noise (PN) codes. When they are long, they have good distance properties.

Short codes. Short codes are good for fast correlation—for example, if we want to lock two codes together. We use short codes to help in code acquisition and synchronization.

In an IMT2000DS handset, user data is channel-coded, spread, then scrambled on the Tx side. Incoming data is descrambled then despread. The following section defines the hardware/software processes required to implement a typical W-CDMA receiver transmitter architecture. It is not a full description of the uplink and downlink protocol.

Common Channels

The downlink (Node B to handset) consists of a number of physical channels. One class (or group) of physical channels is the Common Control Physical CHannel (CCPCH). Information carried on the CCPCH is common to all handsets within a cell (or sector) and is used by handsets to synchronize to the network and assess the link characteristic when the mobile is in idle mode—that is, when it is not making a call. In dedicated connection mode—that is, making a call—the handset will still use part of the CCPCH information to assess cell handover and reselection processes, but will switch to using more specific handset information from the Dedicated CHannels (DCH) that are created in call setup.

The CCPCH consists of a Primary CCPCH (P-CCPCH) and a Secondary CCPCH (S-CCPCH). The P-CCPCH is time multiplexed together with the Synchronization CHannel (SCH) and carries the Broadcast CHannel (BCH).

Synchronization

The SCH consists of two channels: the primary SCH and the secondary SCH (see Figure 3.4). These are used to enable the mobile to synchronize to the network in order for the mobile to identify the base station-specific scrambling code.

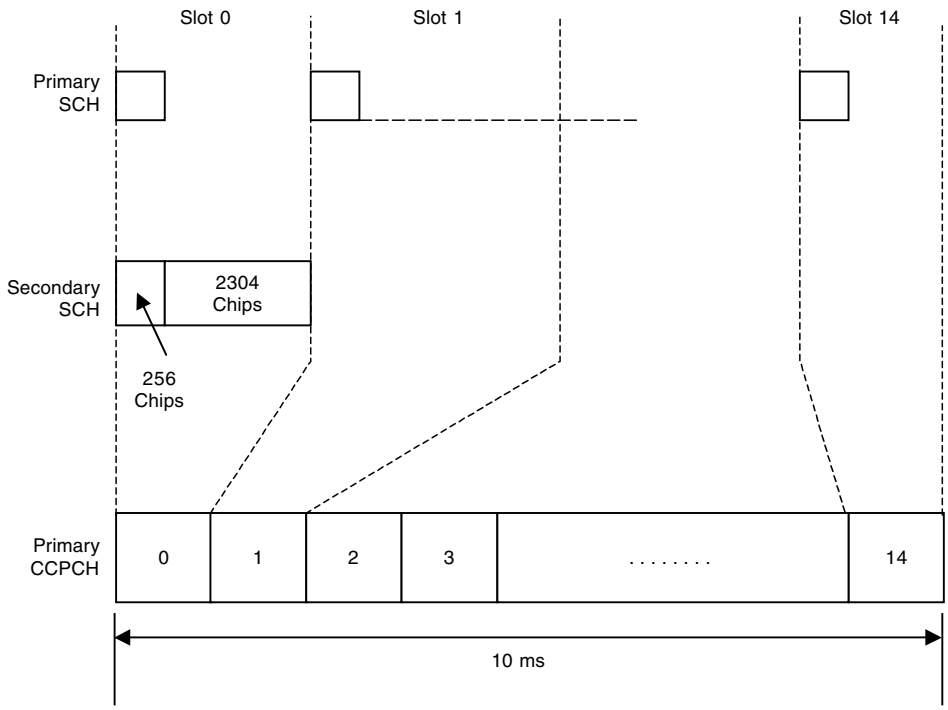


Figure 3.4 Primary and secondary SCH format.

The primary SCH is transmitted once every slot—that is, 15 times in a 10-ms frame. It is a 256-chip unmodulated spreading sequence that is common for the whole network—that is, identical in every cell. It is sent at the front of the 0.625-ms burst and defines the cell start boundary. Its primary function is to provide the handset with a timing reference for the secondary SCH. The secondary SCH, also 256 chips in length, is transmitted in every slot, but the code sequence, which is repeated for each frame, informs the handset of the long code (scrambling) group used by its current Node B.

As the primary SCH is the initial timing reference; in other words, it has no prior time indicator or marker, the receiver must be capable of detecting it at all times. For this reason, a matched filter is usually employed. The IF, produced by mixing the incoming RF with the LO, is applied to the matched filter. This is matching against the 256-bit primary SCH on the CCPCH. When a match is found, a pulse of output energy is produced. This pulse denotes the start of the slot and so is used to synchronize slot-recovery functions.

A 256 tap matched filter at a chip rate of 3.84 Mcps requires a billion calculations per second. However, as the filter coefficients are simply +1 -1 the implementation is reasonably straightforward. The remaining 2304 chips of the P-CCPCH slot form the BCH. As the BCH must be demodulated by all handsets, it is a fixed format. The channel rate is 30 kbps with a spreading ratio of 256, that is, producing a high process gain and consequently a robust signal. As the 256-bit SCH is taken out of the slot, the true bit rate is 27 kbps.

The Common Channels also include the Common PIlot CHannel (CPICH). This is an unmodulated channel that is sent as a continuous loop and is scrambled with the Node B primary scrambling code for the local cell. The CPICH assists the handset to estimate the channel propagation characteristic when it is in idle mode—that is, not in dedicated connection mode (making a call). In dedicated connection mode the handset will still use CPICH information (signal strength) to measure for cell handover and reselection. In connection mode the handset will use the pilot symbols carried in the dedicated channels to assess accurately the signal path characteristics (phase and amplitude) rather than the CPICH. The CPICH uses a spreading factor of 256—that is, high process gain for a robust signal.

Because the mobile only communicates to a Node B and not to any other handset, uplink common physical channels are not necessary. All uplink (handset to Node B) information—that is, data and reporting—is processed through dedicated channels.

Dedicated Channels

The second class of downlink physical channel is the Dedicated CHannel (DCH). The DCH is the mechanism through which specific user (handset) information (control + data) is conveyed. The DCH is used in both the downlink and uplink, although the channel format is different. The differences arise principally through the need to meet specific hardware objectives in the Node B and the handset—for example, conformance with EMC regulations, linearity/power trade-offs in the handset, handset complexity/processing power minimization, and so on.

The DCH is a time multiplex of the Dedicated Physical Control CHannel (DPCCH) and the Dedicated Physical Data CHannel (DPDCH), as shown in Figure 3.5. The DPCH is transmitted in time-multiplex with control information. The spreading factor of the physical channel (Pilot, TPC, and TFCI) may range from 512 to 4.

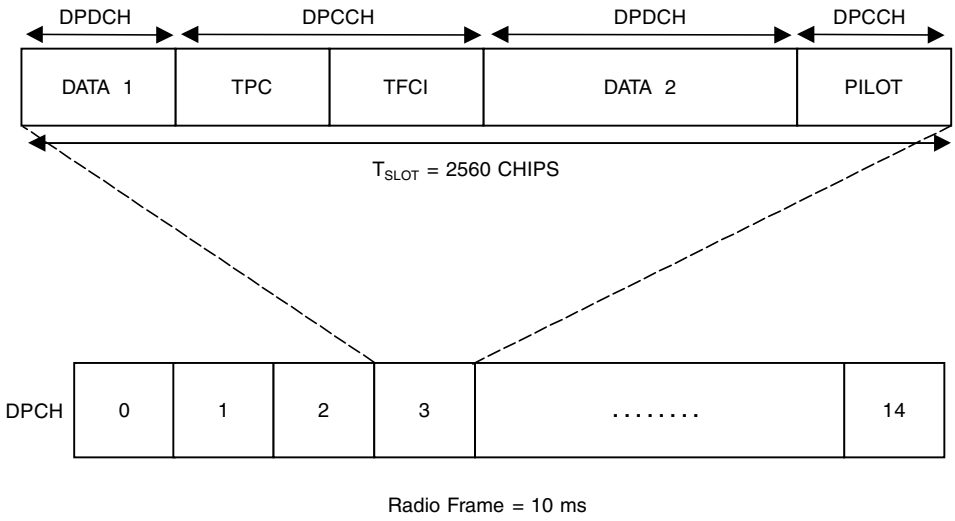


Figure 3.5 Dedicated channel frame structure.

The number of bits in each field may vary, that is:

- Pilot: 2 to 16
- TPC: 2 to 16
- TFCI: 0 to 16
- Data 1: 0 to 248
- Data 2: 2 to 1000

Certain bit/field combinations will require the use of DTX to maintain slot structure/timing.

Table 3.3 shows spreading factors against user data rate. Low-bit-rate users have 24/25 dB of spreading gain, highest-bit-rate users 2/3 dB. From column 5, it is seen that the channel symbol rate can vary from 7.5 to 960 kbps. The dynamic range of the downlink channel is therefore 128:1, that is, 21 dB. The rate can change every 10 ms. In addition to spreading codes, scrambling codes are used on the downlink and uplink to deliver additional selectivity.

In the uplink, user data (DPDCH) is multiplexed together with control information (DPCCH) to form the uplink physical channel (DCH). Multiple DPDCH may be used with a single DPCCH. The DPCCH has a fixed spreading ratio of 256 and the DPDCH is variable (frame-by-frame), from 256 to 4 (see Table 3.4). Each DPCCH can contain four fields: Pilot, Transport Format Combination Indicator (TFCI), Transmission Power Control (TPC), and FeedBack Information (FBI). The FBI may consist of 0, 1, or 2 bits included when closed-loop transmit diversity is used in the downlink. The slot may or may not contain TFCI bits. The Pilot and TPC is always present, but the bit content compensates for the absence or presence of FBI or TFCI bits.

Table 3.3 Downlink Spreading Factors and Bit Rates

MAX USER DATA RATE (KBPS) HALF RATE CODING	SPREADING FACTOR	DPDCH CHANNEL BIT RATE RANGE (KBPS)	NO. OF CODES	CHANNEL SYMBOL RATE (KBPS)
1-3	512	3-6	1	7.5
6-12	256	12-24	1	15
20-24	128	42-51	1	30
45	64	90	1	60
105	32	210	1	120
215	16	432	1	240
456	8	912	1	480
936	4	1872	1	960

Table 3.4 Uplink DPDCH Rates

MAX. USER DATA RATE (KBPS) HALF CODING	SPREADING FACTOR	NO. OF CODE CHANNELS	CHANNEL BIT RATE (KBPS)
7.5	256	1	15
15	128	1	30
30	64	1	60
60	32	1	120
120	16	1	240
240	8	1	480
480	4	1	960

It is the variability of DPDCH (single to multiple channels) that define the dynamic range requirements of the transmitter PA, since multiple codes increase the peak-to-average ratio. From column 4, we see that the channel bit rate can vary from 15 to 960 kbps. The dynamic range of the channel is therefore 64:1—that is, 18 dB. The rate can change every 10 ms.

There are two types of physical channel on the uplink: dedicated physical data channel (DPDC) and dedicated physical control channel (DPCCH). The number of bits per uplink time slot can vary from 10 to 640 bits, corresponding with a user data rate of 15 kbps, to 0.96 Mbps. The user data rate includes channel coding, so the actual user bit rate may be 50 percent or even 25 percent of this rate.

Code Generation

Figure 3.6 shows how the OVSF codes and scrambling codes are applied on the transmit side and then used to decorrelate the signal energy of interest on the receive side, having been processed through a root raised cosine (RRC) filter. Channels are selected in the digital domain using a numerically controlled oscillator and a digital mixer.

Figure 3.7 shows steps in the uplink baseband generation. The DCCH is at a lower bit rate than the DTCH to ensure a robust control channel. Segmentation and matching is used to align the streams to a 10-ms frame structure. The composite signal is applied to the I stream component and the DPCCH carrying the pilot, power control, and TFCI bits with a spreading factor of 256 (providing good processing gain) applied to the Q stream.

The I and Q are coded with the scrambling code, and cross-coupled complex scrambling takes place to generate HPSK.

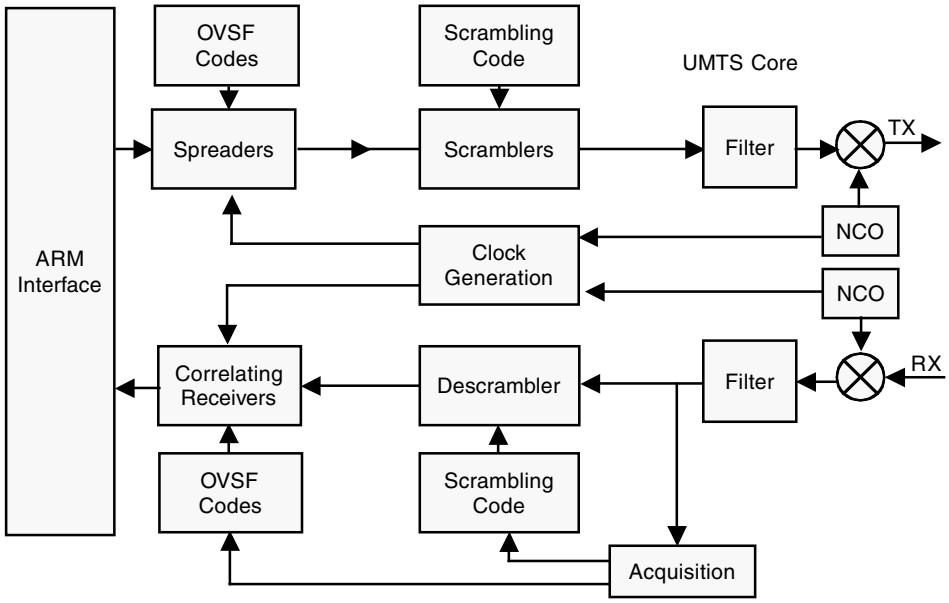


Figure 3.6 UMTS core for multimode 3G phones.

Hybrid phase shift keying, also known as orthogonal complex quadrature phase shift keying, allows handsets to transmit multiple channels at different amplitude levels while still maintaining acceptable peak-to-average power ratios. This process uses Walsh rotation, which effectively continuously rotates the modulation constellation to reduce the peak to average (PAR) of the signal prior to modulation.

Figure 3.7 is taken from Agilent's "Designing and Testing W-CDMA User Equipment" Application Note 1356. To summarize the processing so far, we have performed cyclic redundancy checking, forward error correction (FEC), interleaving, frame construction, rate matching, multiplexing of traffic and control channels, OVSF code generation and spreading, gain adjustment, spreading and multiplexing of the primary control channel, scrambling code generation, and HPSK modulation.

The feedback coefficients needed to implement the codes are specified in the 3GPP1 standards, as follows:

- *Downlink:*
  - 38,400 chips of 2^{18} Gold code
  - 512 different scrambling codes
  - Grouped for efficient cell search
- *Uplink:*
  - Long code: 38,400 chips of 225 Gold code
  - Short code: 256 chips of very large Kasami code

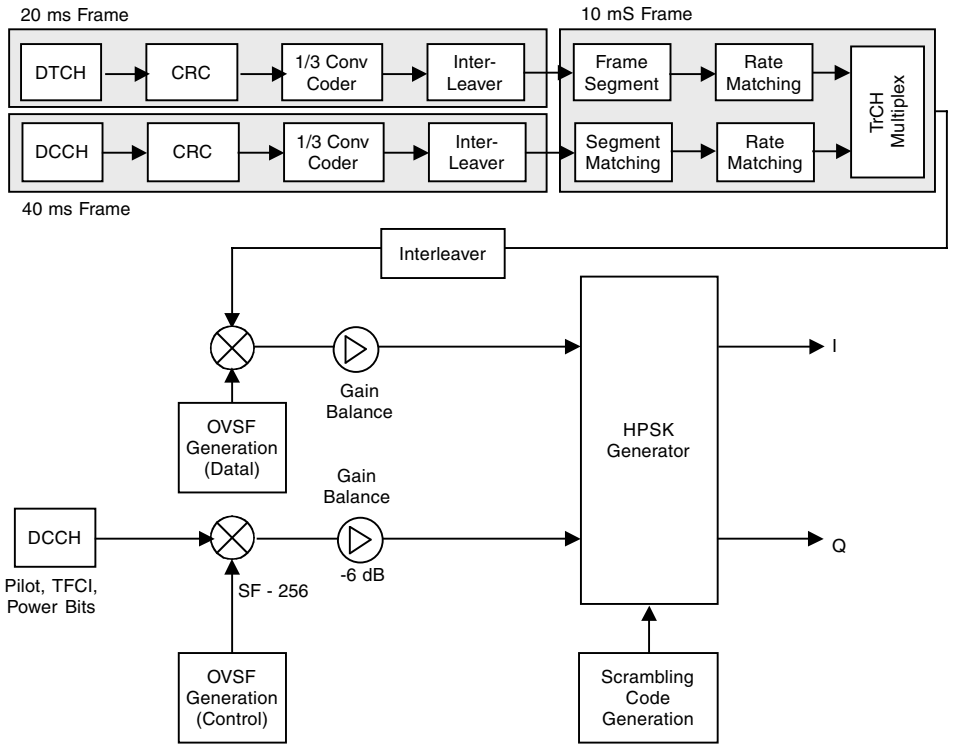


Figure 3.7 Uplink baseband generation.

An example hardware implementation might construct the function on a Xilinx Virtex device—it uses approximately 0.02 percent of the device, compared with the 25 percent required to implement an RRC/interpolator filter function.

Root Raised Cosine Filtering

We have now generated a source coded, 3.84 Mcps, I and Q streamed, HPSK formatted signal. Although the bandwidth occupancy of the signal is directly a function of the 3.84 Mcps spreading code, the signal will contain higher frequency components because of the digital composition of the signal. This may be verified by performing a Fourier analysis of the composite signal. However, we only have a 5 MHz bandwidth channel available to us, so the I and Q signals must be passed through filters to constrain the bandwidth. Although high-frequency components are removed from the signal, it is important that the consequent softening of the waveform has minimum impact on the channel BER. This objective can be met by using a class of filters referred to as Nyquist filters. A particular Nyquist filter is usually chosen, since it is easier to implement than other configurations: the raised cosine filter.

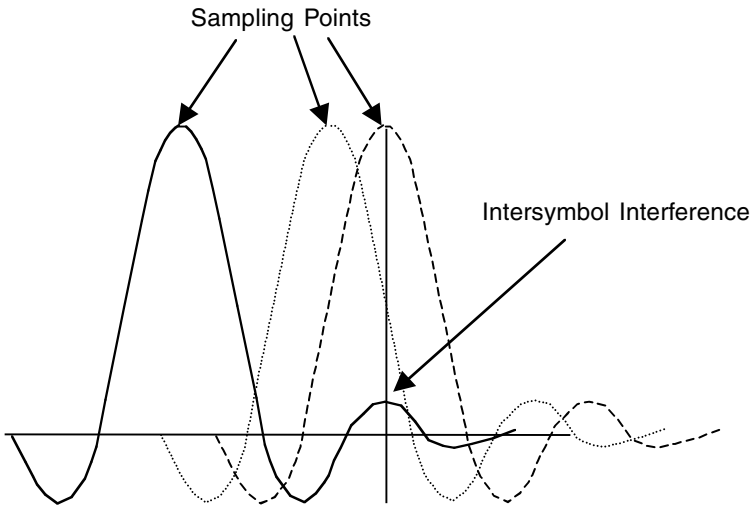


Figure 3.8 Filter pulse train showing ISI.

If an ideal pulse (hence, infinite bandwidth) representing a 1 is examined, it is seen that there is a large time window in which to test the amplitude of the pulse to check for its presence, that is, the total flat top. If the pulse is passed through a filter, it is seen that the optimum test time for the maximum amplitude is reduced to a very small time window. It is therefore important that each pulse (or bit) is able to develop its correct amplitude.

The frequency-limiting response of the filter has the effect of smearing or time-stretching the energy of the pulse. When a train of pulses (bit stream) is passed through the filter, this ringing will cause an amount of energy from one pulse to still exist during the next. This carrying forward of energy is the cause of Inter-Symbol Interference (ISI), as shown in Figure 3.8.

The Nyquist filter has a response such that the ringing energy from one pulse passes through zero at the decision point of the next pulse and so has minimum effect on its level at this critical time (see Figure 3.9).

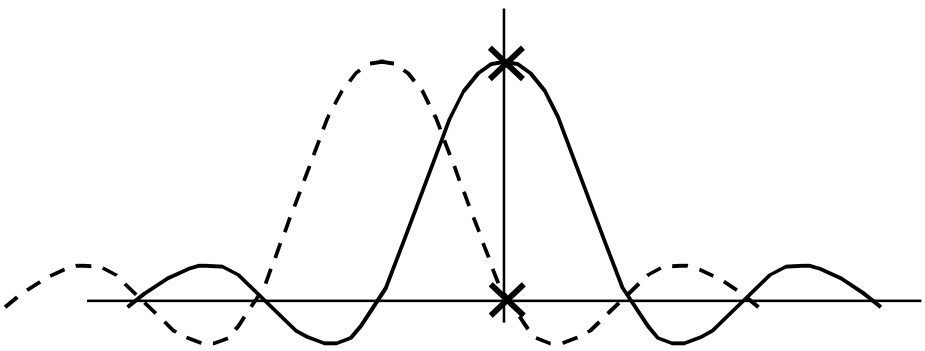


Figure 3.9 The Nyquist filter causes minimum ISI.

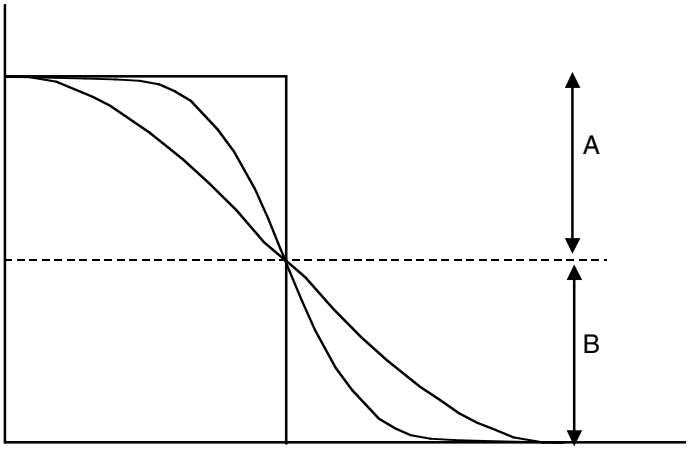


Figure 3.10 Symmetrical transition band.

The Nyquist filter exhibits a symmetrical transition band, as shown in Figure 3.10.

The cosine filter exhibits this characteristic and is referred to as a raised cosine filter, since its response is positioned above the base line.

It is the total communication channel that requires the Nyquist response (that is, the transmitter/receiver combination), and so half of the filter is implemented in the transmitter and the other half in the receiver. To create the correct overall response, a Root Raised Cosine (RRC) filter is used in each location as:

$\sqrt{x} \times \sqrt{x} = x$

Modulation and Upconversion

Because the handset operates in a very power restrictive environment, all stages must be optimized not only for signal performance but also power efficiency. Following the RRC filtering, the signal must be modulated onto an IF and up-converted to the final transmission frequency. It is here the Node B and handset processes differ. The signal could continue to be processed digitally to generate a digitally sampled modulated IF to be converted in a fast DAC for analog up-conversion for final transmission. However, the power (DC) required for these stages prohibits this digital technique in the handset. (We will return to this process in Node B discussions.) Following the RRC filtering, the I and Q streams will be processed by matched DACs and the resulting analog signal applied to an analog vector modulator (see Figure 3.11).

A prime challenge in the design of a W-CDMA handset is to achieve the modulation and power amplification within a defined (low) power budget but with a minimum component count. This objective has been pursued aggressively in the design and implementation of later GSM handsets. Sufficient performance for a single-band (900 MHz) GSM phone was achieved in early-generation designs, but the inclusion of a second and third (and later fourth—800 MHz) band has driven the research toward minimum component architectures—especially filters. Chapter 2 introduced the offset loop transmitter architecture, which is successfully used for low-cost, low component count multiband GSM applications.

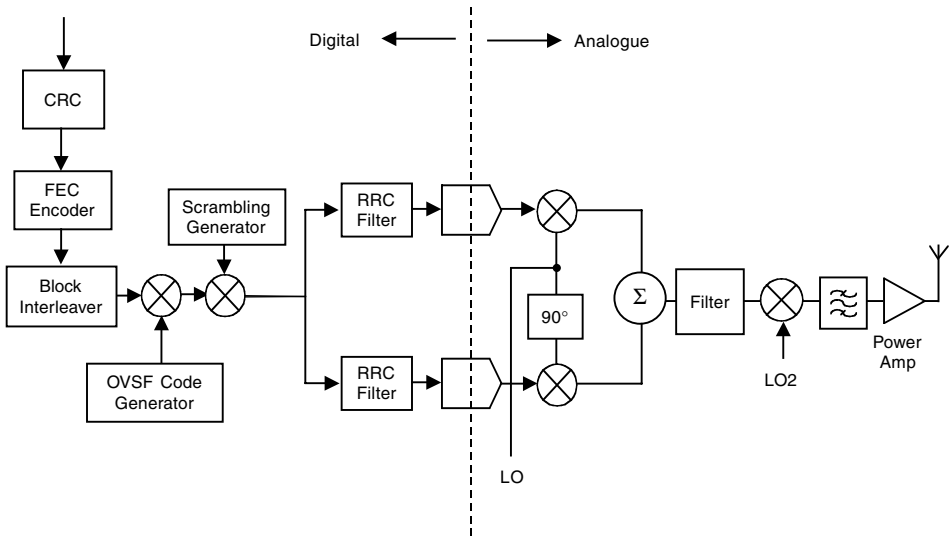


Figure 3.11 Typical digital/analog partitioning in a handset (analog vector modulator).

This architecture is very suitable for the GMSK modulation of GSM, as it has a constant amplitude envelope. The PLL configuration is only required to respond to the phase component of the carrier. The QPSK and HPSK modulation used in W-CDMA is non-constant envelope—that is, the modulated carrier contains both phase and amplitude components. Because the offset loop is unable to reproduce the amplitude components, it is unsuitable in its simple form. However, since it is particularly economic in components, there is considerable research directed toward using this technique for W-CDMA. To use the technique, the offset loop is used, with the amplitude components being removed by the loop function, but an amplitude modulator is used on the PA output to reproduce the amplitude components. This method of processing the carrier separately from its amplitude components is referred to as *Envelope Elimination and Restoration* (EER), as shown in Figure 3.12.

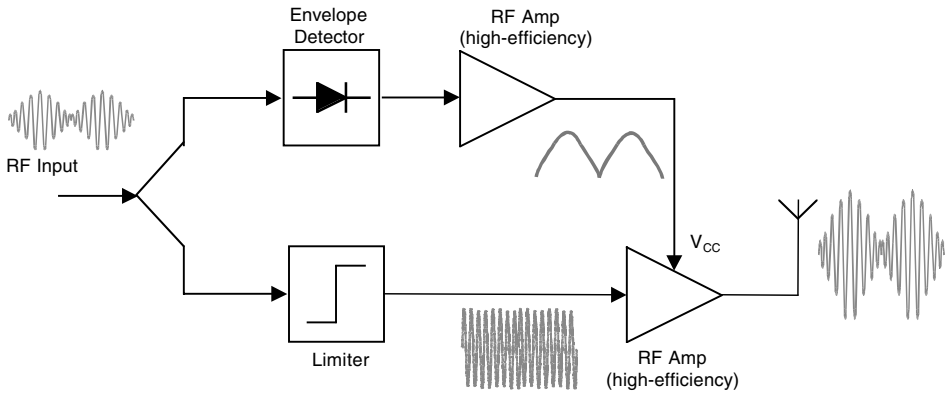


Figure 3.12 RF envelope elimination and restoration.

Other methods of producing sufficient PA linearity include adaptive predistortion and possibly the Cartesian loop technique although the latter is unlikely to stretch across the bandwidth/linearity requirement.

Power Control

All the decorrelation processes rely on coded channel streams being visible at the demodulator at similar received power levels (energy per bit) ideally within 1 dB of each other.

For every slot, the handset has to obtain channel estimates from the pilot bits, estimate the signal to interference ratio, and process the power control command (Transmission Power Control, or TPC), that is, power control takes place every 660 μs, or 1500 times per second. This is fast enough to track fast fading for users moving at up to 20 kmph. Every 10 ms, the handset decodes the Transport Format Combination Indicator (TFCI) which gives it the bit rate and channel decoding parameters for the next 10-ms frame. Data rates can change every 10 ms (dynamic rate matching) or at the establishment or teardown of a channel stream (static matching). The coding can change between 1/3 convolutional coding, 1/3 convolutional coding with additional block coding, /or turbo coding.

Power control hardware implementation will be similar to the detection methods outlined in Chapter 2 but the methods need to comprehend additional dynamic range and faster rates of change. The power control function also needs to be integrated with the linearization process.

The Receiver

As we have outlined in Chapter 2, there is a choice of superhet or zero/near-zero IF receiver architecture. If the superhet approach is chosen, it will certainly employ a sampled/digitized IF approach in order to provide the functional flexibility required. This approach enables us to realize a multimode—for example, GSM and W-CDMA—handset with a common hardware platform but with the modes differentiated by software. The following sections outline the functions that are required after sampling the IF in order to pass the signal to the RAKE receiver processes.

The Digital Receiver

The digital receiver consists of a digital local oscillator, digital mixer, and a decimating lowpass filter (see Figure 3.13). Digital samples from the ADC are split into two paths and applied to a pair of digital mixers. The mixers have digital local oscillator inputs of a quadrature signal—that is, sine and cosine—to enable the sampled IF to be mixed down to a lower frequency, usually positioned around 0 Hz (DC).

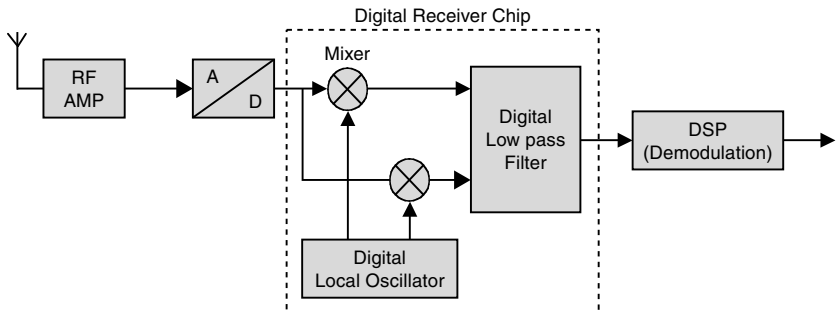


Figure 3.13 The digital receiver.

This process converts the digitized signal from a real to a complex signal, that is, a signal represented by its I and Q phase components. Because the signal is represented by two streams, the I and Q could be decimated by a factor of 2 at this point. However, because the down-shifted signal is usually processed for a single channel selection at this stage, the two decimation factors may be combined.

The LO waveform generation may be achieved by a number of different options—for example, by a Numerically Controlled Oscillator (NCO), also referred to as a Direct Digital Synthesizer (DDS)—if digital-to-analog converters are used on the I and Q outputs. In this process, a digital phase accumulator is used to address a lookup table (LUT), which is preprogrammed with sine/cosine samples. To maintain synchronization, the NCO is clocked by the ADC sampling/conversion clock.

The digital samples (sine/cosine) out of the local oscillator are generated at a sampling rate exactly equal to the ADC sample clock frequency *fs*. The sine frequency is programmable from DC to *fs*/2 and may be 32 bits. By the use of programmable phase advance, the resolution is usually sub-Hertz. The phase accumulator can maintain precise phase control, allowing phase-continuous switching. The mixer consists of two digital multipliers. Digital input samples from the ADC are mathematically multiplied by the digital sine and cosine samples from the LO. Because the data rates from the two mixer input sources match the ADC sampling rate (*fs*), the multipliers also operating at the same rate produce multiplied output product samples at *fs*. The I and Q outputs of the mixers are the frequency downshifted samples of the IF. The sample rate has not been changed; it is still the sample rate that was used to convert the IF.

The precision available in the mixing process allows processing down to DC (0 Hz). When the LO is tuned over its frequency range, any portion of the RF signal can be mixed down to DC; in other words, the wideband signal spectrums can be shifted around 0 Hz, left and right, by changing the LO frequency. The signal is now ready for filtering.

The decimating lowpass filter accepts input samples from the mixer output at the full ADC sampling frequency, *fs*. It uses digital signal processing to implement a finite impulse response (FIR) transfer function. The filter passes all signals from 0 Hz to a programmable cutoff frequency or bandwidth and rejects all signals higher than that

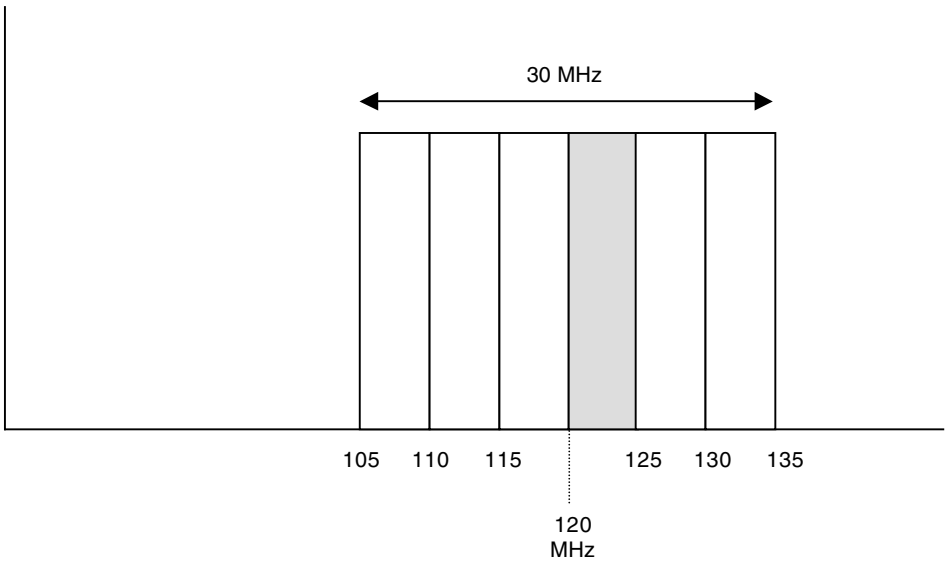


Figure 3.14 A 30 MHz signal digitized at an IF of 120 MHz.

cutoff frequency. The filter is a complex filter that processes both I and Q signals from the mixer. At the output either I or Q (complex) values or real values may be selected.

An example will illustrate the processes involved in the digital receiver function (see Figure 3.14). The bandwidths—that is, number of channels sampled and digitized—in the sample may be outside a handset power budget; a practical design may convert only two or three channels. A 30 MHz (6 by 5 MHz W-CDMA RF channels) bandwidth signal has been sampled and digitized at an IF of 120 MHz.

It is required to process the channel occupying 120 to 125 MHz. When the LO is set to 122.5 MHz, the channel of interest is shifted down to a position around 0 Hz. When the decimating (lowpass) filter is set to cut off at 2.5 MHz, the channel of interest may be extracted (see Figure 3.15).

To set the filter bandwidth, you must set the decimation factor. The decimation factor is a function of both the output bandwidth and output sampling rate. The decimation factor, N, determines both the ratio between input and output sampling rates and the ratio between input and output bandwidths.

In the example in Figure 3.15, the input had a 30 MHz bandwidth input with a ±2.5 MHz bandwidth output. The decimation factor is therefore 30 MHz/2.5 MHz—that is, 12.

Digital receivers are divided into two classes, narrowband and wideband, defined by the range of decimation factors. Narrowband receivers range from 32 to 32,768 for real outputs, wideband receivers 1 to 32. When complex output samples are selected, the sampling rate is halved, as a pair of output samples are output with each sample clock. The downconverted, digitized, tuned around 0 Hz, filtered channel (bandwidth = 5 MHz) now exists as minimum sample rate I and Q bit streams. In this form it is now ready for baseband recovery and processing.

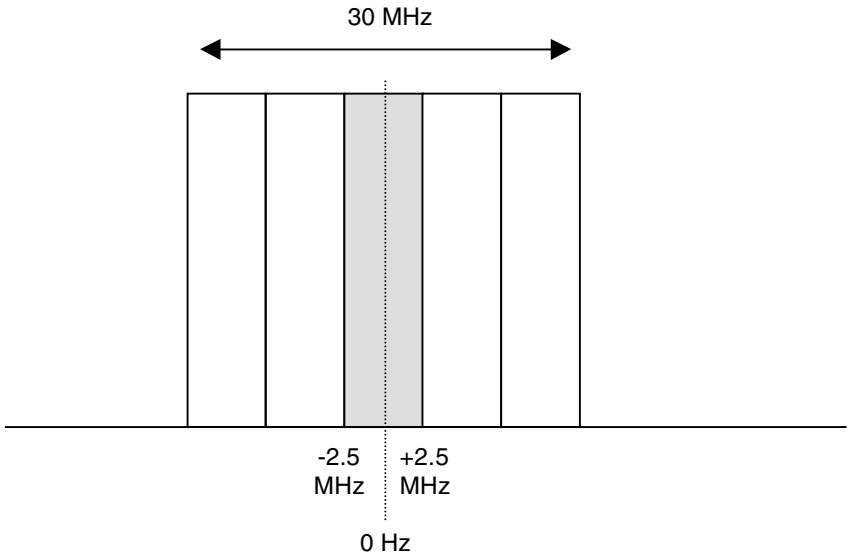


Figure 3.15 Selected channel shifted around 0 Hz.

The RAKE Receive Process

The signal, transmitted by the Node B or handset, will usually travel along several different paths to reach the receiver. This is due to the reflective and refractive surfaces that are encountered by the propagating signal. Because the multiple paths have different lengths, the transmitted signal has different arrival times (phases) at the receive antenna; in other words, the longer the path the greater the delay. 1G and 2G cellular technologies used techniques to select the strongest path for demodulation and processing. Spread spectrum technology, with its carrier time/phase recognition technique, is able to recover the signal energy from these multiple paths and combine it to yield a stronger signal.

Data signal energy is recovered in the spread spectrum process by multiplying synchronously, or despreading, the received RF with an exact copy of the code sequence that was used to spread it in the transmitter. Since there are several time-delayed versions of the received signal, the signal is applied simultaneously to a number of synchronous receivers, and if each receiver can be allocated to a separate multipath signal, there will be separate, despread, time-delayed recovered data streams. The data streams can be equalized in time (phase) and combined to produce a single output. This is the RAKE receiver.

To identify accurately the signal phase, the SCH is used. As already described, the received RF containing the SCH is applied to a 256-chip matched filter. This may be analog or sampled digitized IF. Multiple delayed versions of the same signal will produce multiple energy spikes at the output. Each spike defines the start of each delayed slot. (It is the same slot—the spikes define the multiple delays of the one slot.)

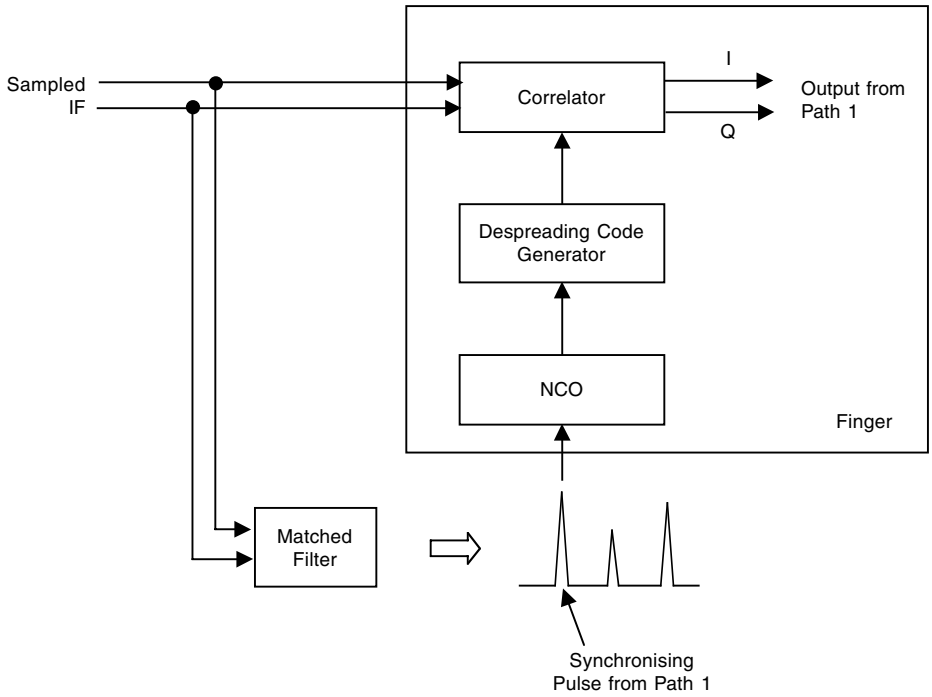


Figure 3.16 Matched filter for synchronizing I and Q.

Each spike is used as a timing reference for each RAKE receive correlator. The despreading code generator can be adjusted in phase by adjusting the phase of its clock. The clock is generated by an NCO—a digital waveform generator—that can be synchronized to a matched filter spike, that is, the SCH phase (see Figure 3.16).

Because the received signal has been processed with both scrambling and spreading codes, the code generators and correctors will generate scrambling codes to descramble (*not* despread) the signal and then OVSF codes to despread the signal. This process is done in parallel by multiple RAKE receivers or fingers. So now, each multipath echo has been despread but each finger correlator output is nonaligned in time. Part of the DPCCH, carried as part of the user-dedicated, or unique, channel is the *pilot code*. The known format of the pilot code bits enables the receiver to estimate the path characteristic—phase and attenuation. The result of this analysis is used to drive a phase rotator (one per RAKE finger) to rotate the phase of the signal of each finger to a common alignment. So, now we have multiple I and Q despread bit streams aligned in phase but at time-delayed intervals.

The last stage within each finger is to equalize the path delays, again using the matched filter information. Once the phase has been aligned, the delay has been aligned, and the various signal amplitudes have been weighted, the recovered energy of interest can be combined (see Figure 3.17).

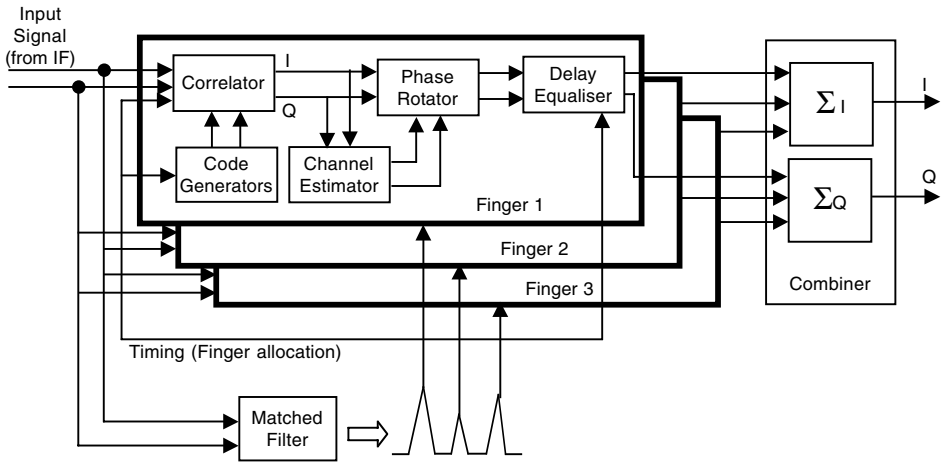


Figure 3.17 Combined signal energy of interest.

Path combining can be implemented in one of two ways. The simpler combining process uses equal gain; that is, the signal energy of each path is taken as received and, after phase and delay correction, is combined without any further weighting. Maximal ratio combining takes the received path signal amplitudes and weights the multipath signals by adding additional energy that is proportional to their recovered SNR. Although more complex, it does produce a consistently better composite signal quality. The complex amplitude estimate must be averaged over a sufficiently long period to obtain a mean value but not so long that the path (channel) characteristic changes over this time, that is, the coherence time.

Correlation

As we have described, optimum receive performance (BER) is dependent on the synchronous application of the despreading code to the received signal. Nonsynchronicity in the RAKE despreading process can be due to the random phase effects in the propagation path, accuracy and stability of the handset reference, and Doppler effects.

The process outlined in the previous section is capable of providing despreading alignment to an accuracy of one chip; however, this is not sufficient for low-BER, best demodulation. An accuracy of 1/8 chip or better is considered necessary for optimum performance.

As DSP/FPGA functions become increasingly power-efficient, greater use will be made of digital techniques (for example, digital filters) to address these requirements of fractional bit synchronization. Currently, methods employing delay lock loop (DLL) configurations are used to track and determine the received signal and despread code phase (see Figure 3.18). Code tracking can be achieved by the DLL tracking of PN signals. The principle of the DLL as an optimal device for tracking the delay difference

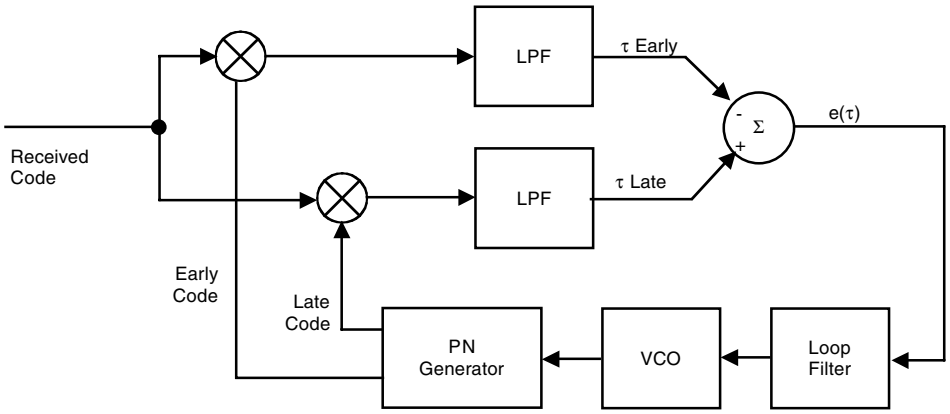


Figure 3.18 The DLL.

between the acquired and the local sequence is very similar to that of the PLL that is used to track the frequency/phase of the carrier. Code tracking loops perform either coherently—that is, they have knowledge of the carrier phase—or noncoherently—that is, they do not have knowledge of the carrier phase.

Two separate correlators are used, each with a separate code generator. One correlator is referred to as the *early correlator* and has a code reference waveform that is advanced in time (phase) by a fraction of a chip; the other correlator is the late correlator and is delayed but some fraction of a chip. The difference, or imbalance, between the correlations is indicative of the difference between the despread code timing and the received signal code.

The output signal e(τ) is the correction signal that is used to drive the PN generator clock (VCO or NCO—if digital). A third PN on time sequence can be generated from this process to be applied to an on-time correlator, or the correction could be applied to adjust the TCXO reference for synchronization.

A practical modification is usually applied to the DLL as described. The problem to be overcome is that of imbalance between the two correlators. Any imbalance will cause the loops to settle in an off-center condition—that is, not in synch. This is overcome by using the *tau-dither early-late tracking loop*.

The tau-dither loop uses one deciding correlator, one code generator, and a single loop, but it has the addition of a phase switch function to switch between an early and late phase—that is, advance and delay for the PN code tuning. In this way imbalance is avoided in the timing/synchronizing process, since all components are common to both early and late phases.

Receiver Link Budget Analysis

Because processing gain reduces as bit rate increases, receiver sensitivity must be determined across all possible data rates and for a required E_b/N_o (briefly, the ratio of

energy per bit to the spectral noise density; we will discuss this further shortly). The calculation needs to comprehend the performance of the demodulator, which, in turn, is dependent on the level of modulation used. Other factors determining receiver sensitivity include the RF front end, mixer, IF stages, analog-to-digital converter, and baseband process (DSP). (See Figure 3.19.)

Let's look at a worked example in which we define receiver sensitivity. For example, let's determine receiver sensitivity at three data rates: 12.2 kbps, 64 kbps, and 1920 kbps at a BER of 1 in 10^6.

The noise power is dimensioned by Boltzman's constant ($k = 1.38 \times 10^{-23}$ J/K) and standardized to a temperature (T) of 290K (17° C). To make the value applicable to any calculation, it is normalized at a 1 Hz bandwidth. The value ($k \times T$) is then multiplied up by the bandwidth (B) used.

The noise power value is then -174 dBm/Hz and is used as the floor reference in sensitivity/noise calculations. The receiver front end (RF + mixer) bandwidth is 60 MHz, in order to encompass IMT2000DS license options. The noise bandwidth of the front end is $10\log_{10}(60\text{MHz}) = 77.8$ dB. The receiver front noise floor reference is therefore -174 dBm +77.8 dB = -96.2 dBm. In the DSP, the CDMA signal is despread from 3.84 Mcps (occupying a 5 MHz bandwidth), to one of the three test data rates—12.2 kbps, 64 kbps, 1920 kbps—and can be further filtered to a bandwidth of approximately:

- Modulation bandwidth = Data rate $\times (1+ \alpha)/\log2(M)$ (where α = pulse-shaping filter roll-off and M = no of symbol states in modulation format)
- For IMT2000, $\alpha = 0.22$ and M=4 (QPSK)

Thus, reduction in receiver noise due to despreading is as follows:

- $= 10\log_{10}$(IF bandwidth/modulation BW)
- $= 10\log_{10}(5\text{ MHz}/7.5\text{ kHz}) = 28.2$ dB for 12.2 kbps
- $= 10\log_{10}(5\text{ MHz}/39\text{ kHz}) = 21.1$ dB for 64 kbps
- $= 10\log_{10}(5\text{ MHz}/1.25\text{ MHz}) = 6.0$ dB for 1920 Mbps

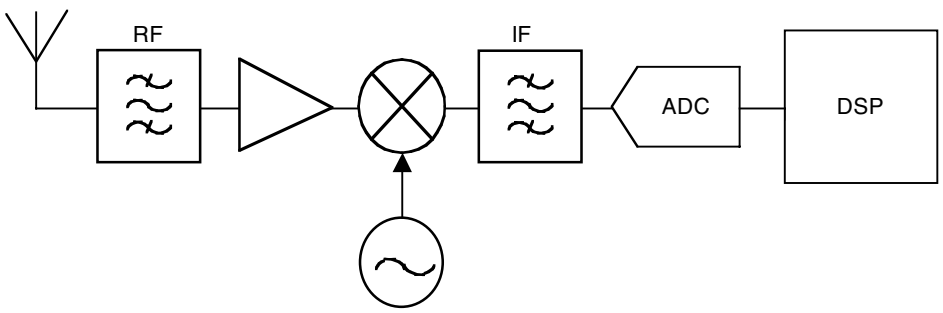


Figure 3.19 The digital IF receiver.

The effective receiver noise at each data detector due to input thermal noise is thus:

SOURCE DATA	RECEIVER NOISE REFERENCE	EFFECTIVE RECEIVER NOISE
12.2 kbps	-107 dBm-28.2 dB =	-135.2 dBm
64 kbps	-107 dBm-21.1 dB =	-128.1 dBm
1920 kbps	-107 dBm-6.00 dB =	-113.0 dBm

The real noise floor for a practical receiver will always be higher because of filter losses, LNA and mixer noise, synthesizer noise, and so on. In a well-designed receiver, 5 dB might be a reasonable figure. The practical effective noise floor of a receiver would then be

12.2 kbps	-135.2 dBm + 5 dB =	-130.2 dBm
64 kbps	-128.1 dBm + 5 dB =	-123.1 dBm
1920 kbps	-113 dBm + 5 dB =	-108.0 dBm

Using these figures as a basis, a calculation may be made of the receiver sensitivity. To determine receiver sensitivity, you must consider the minimum acceptable output quality from the radio. This minimum acceptable output quality (SINAD in analog systems, BER in digital systems) will be produced by a particular RF signal input level at the front end of the receiver. This signal input level defines the sensitivity of the receiver.

To achieve the target output quality (1×10^{-6} in this example), a specified signal (or carrier) quality is required at the input to the data demodulator. The quality of the demodulator signal is defined by its E_b/ N_o value, where E_b is the energy per bit of information and N_o is the noise power density (that is, the thermal noise in 1 Hz of bandwidth). The demodulator output quality is expressed as BER, as shown in Figure 3.20. In the figure, a BER of 1 in 10^6 requires an E_b/ N_o of 10.5 dB.

Because receiver sensitivity is usually specified in terms of the input signal power (in dBm) for a given BER, and since we have determined the equivalent noise power in the data demodulator bandwidth, we need to express our E_b/N_o value as an S/N value. The S/N is obtained by applying both the data rate (R) and modulation bandwidth (B_M) to the signal, as follows:

- $S/N = (E_b/N_o) \times (R/B_M)$
- For QPSK (M=4), $B_M \sim R/2$, thus:
- $S/N = (E_b/No) \times 2 = 14.5dB$ for BER = 1in 10^6

Assuming a coding gain of 8 dB, we can now determine the required signal power (receive sensitivity) at the receiver to ensure we meet the (14.5-8) dB = 6.5 dB S/N target.

DATA RATE	EFFECTIVE NOISE	RECEIVER SENSITIVITY FOR 1 IN 10^6 BER
12.2 kbps	-130.2 dBm	-124.7 dBm
64 kbps	-123.1 dBm	-116.6 dBm
2 Mbps	-108.0 dBm	-101.5 dBm

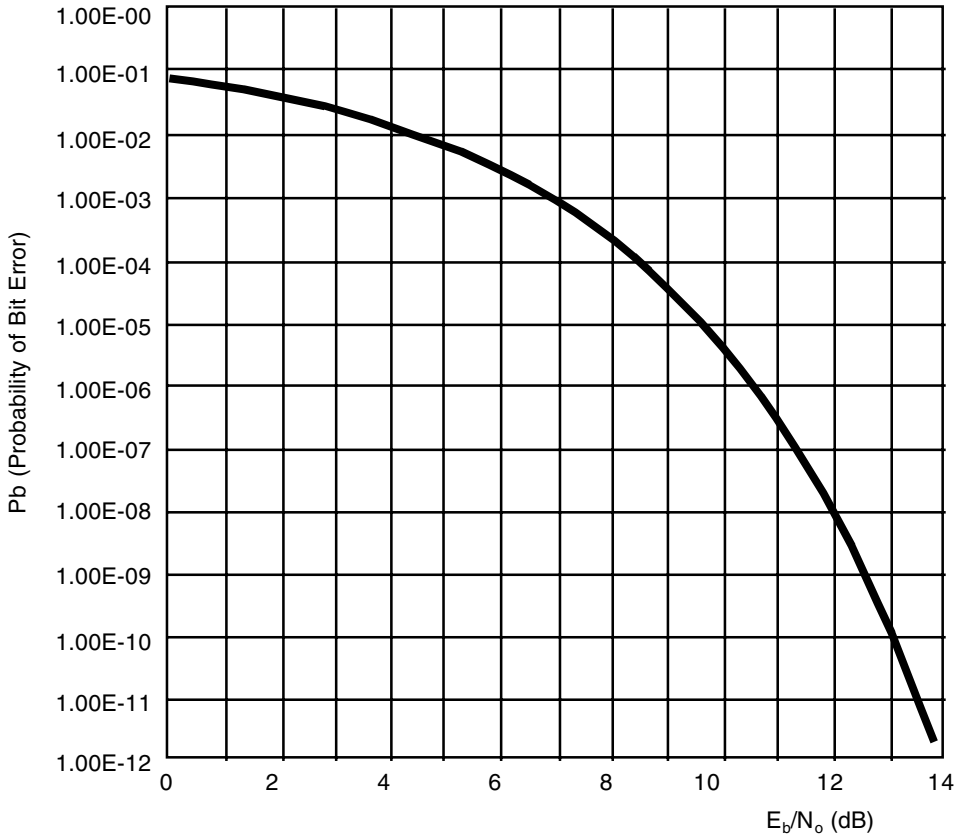


Figure 3.20 Bit error performance of a coherent QPSK system.

There is approximately 22 dB difference in sensitivity between 12.2 kbps speech and 2 Mbps data transfer, which will translate into a range reduction of approximately 50 percent, assuming r^4 propagation, and a reduction in coverage area of some 75 percent!

IMT2000DS Carrier-to-Noise Ratio

In the 2G (GSM) system the quality of the signal through the receiver processing chain is determined primarily by the narrow bandwidth, that is, 200 kHz. This means that the SNR of the recovered baseband signal is determined by the 200 kHz IF filter positioned relatively early in the receive chain; little improvement in quality is available after this filter. Consequently the noise performance resolution and accuracy of the sampling ADC, which converts the CNR, must be sufficient to maintain this final quality SNR. When the W-CDMA process is considered, a different situation is seen. The sampled IF has a 5 MHz bandwidth and is very noisy—intentionally so. Because the

CNR is poor, it does not require a high-resolution ADC at this point; large SNR improvement through the processing gain comes after the ADC. A fundamental product of the spreading/despreading process is the improvement in the CNR that can be obtained prior to demodulation and base band processing—the processing gain.

In direct-sequence spread spectrum the randomized (digital) data to be transmitted is multiplied together with a pseudorandom number (PN) binary sequence. The PN code is at a much higher rate than the modulating data, and so the resultant occupied bandwidth is defined by the PN code rate. The rate is referred to as the chip rate with the PN symbols as chips. The resultant wideband signal is transmitted and hence received by the spread spectrum receiver. The received wideband signal is multiplied by the same PN sequence that was used in the transmitter to spread it.

For the process to recover the original pre-spread signal energy it is necessary that the despreading multiplication be performed synchronously with the incoming signal. A key advantage of this process is the way in which interfering signals are handled. Since the despreading multiplication is synchronous with the transmitted signal, the modulation energy is recovered. However, the despreading multiplication is *not* synchronous with the interference, so spreads it out over the 5 MHz bandwidth. The result is that only a small portion of the interference energy (noise) appears in the recovered bandwidth.

Processing or despreading gain is the ratio of chip rate to the data rate. That is, if a 32-kbps data rate is spread with a chip rate of 3.84 Mcps, the processing gain is as follows:

The power of the processing gain can be seen by referring to the CNR required by the demodulation process. An E_b/N_o of 10.5 dB is required to demodulate a QPSK signal with a BER of 1×10^{-6}. If a data rate of 960 kbps is transmitted with a chip cover of 3.84 Mcps, the processing gain is 6 dB. If a CNR of 10.5 dB is required at the demodulator and an improvement of 6.0 dB can be realized, the receiver will achieve the required performance with a CNR of just 4.5 dB in the RF/IF stages.

It must be considered at what point in the receiver chain this processing gain is obtained. The wideband IF is digitized and the despreading performed as a digital function after the ADC. Therefore, the ADC is working in a low-quality environment—4.5 dB CNR. The number of bits required to maintain compatibility with this signal is 6 or even 4 bits. The process gain is applied to the total spread signal content of the channel.

If the ADC dynamic range is to be restricted to 4 or 6 bits, consideration must be given to the incoming signal mean level dynamic range. Without some form of received signal dynamic range control, a variation of over 100 dB is typical; this would require at least an 18-bit ADC. To restrict the mean level variation within 4 or 6 bits, a system of variable-gain IF amplification (VGA) is used, controlled by the Received Signal Strength Indication (RSSI).

Prior to the change to 3.84 Mcps, the chip rate was at 4.096 Mcps, which when applied to a filter with a roll-off factor $\propto$ of 1.22 gave a bandwidth of 5 MHz. Maintaining the filter at 1.22 will give an improved adjacent channel performance. The process gain is applied to the total spread signal content of the channel. For example, a 9.6-kbps speech signal is channel-coded up to a rate of 32 kbps. The process gain is therefore 10 log(3.84/0.032) = 20.8 dB, *not* 10log(3.84/0.0096) = 26 dB, as may have been anticipated (or hoped for).

Receiver Front-End Processing

In a digitally sampled IF superhet receiver, the front end (filter, LNA, mixer, and IF filter/pre-amplifier) prepares the signal for analog to digital conversion. The parameters specifying the front end need to be evaluated in conjunction with the chosen ADC. An IMT2000DS example will be used to show an approach to this process.

The receiver performance will be noise-limited with no in-band spurs that would otherwise limit performance. This is reasonable because the LO and IF can be chosen such that unwanted products do not fall in-band. Spurs that may be generated within the ADC are usually not a problem, because they can be eliminated by adding dither or by carefully choosing the signal placement.

The superhet receiver will have a front-end bandwidth of 60 MHz, to encompass the total spectrum allocation and a digitizing, demodulation, and processing bandwidth of 5 MHz, as defined for W-CDMA. To meet the stringent power consumption and minimum components requirement, the receiver will be realized as a single conversion superhet (see Figure 3.21).

The first step is a gain and noise budget analysis to X ——- X. The dB figures are converted to ratios:

Filter insertion loss	1.0 dB	= 1.26 (gain = 0.79)
LNA gain	12 dB	= 15.85
LNA noise figure	1.5 dB	= 1.41
Mixer gain	8 dB	= 6.31
Mixer noise figure	12 dB	= 15.85
IF pre-amp gain	0 dB	= 1.00
IF pre-amp noise figure	3 dB	= 1.99
IF filter insertion loss	4 dB	= 2.51

Using the Friis equation, the composite noise factor (linear) can be calculated:

$$F = 1.26 + \frac{(1.41 - 1)}{0.79} + \frac{(15.85 - 1)}{0.79 \times 15.85} + \frac{(1.99 - 1)}{0.79 \times 15.85 \times 6.31} + \cdots\cdots$$
$$= 1.26 + 0.52 + 1.19 + 0.01 + \cdots\cdots$$
$$= 2.98$$

The noise figure is therefore:	4.7 dB.
Gain to X————————X	= -1.0 + 12 + 8 + 0 -4
	= 15 dB.

The front end has a noise figure of 4.7 dB and a conversion gain of 15 dB. An evaluation is made of the noise power reaching X——–X (considered in a 5 MHz bandwidth).

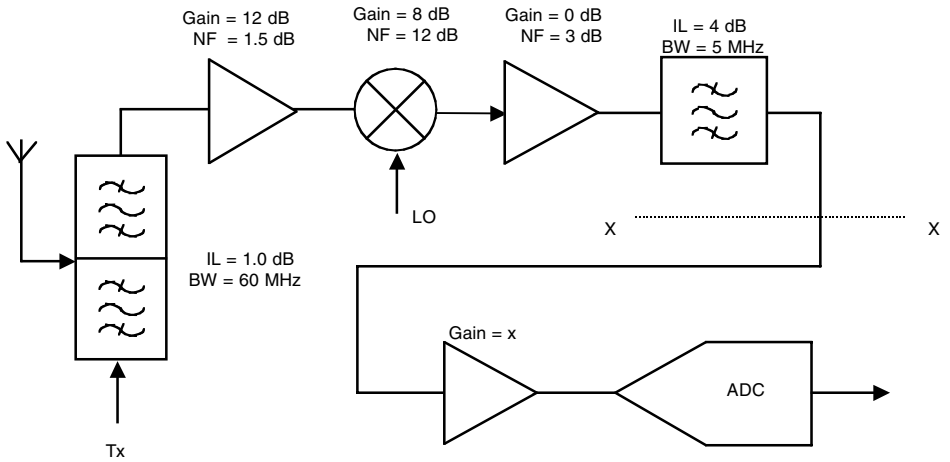


Figure 3.21 W-CDMA superhet receiver.

Noise power in a 5 MHz bandwidth, above the noise floor, is as follows:

= -174 dB/Hz + 10 log 5 MHz

= -174 + 67

= -107 dBm

The RF/IF noise power (in a 5 MHz bandwidth) to be presented to the ADC is as follows:

= -107 + 15 + 4.7

= -87.3 dBm

If the receiver input signal is -114.2 dBm (for a 64 kbps data stream with 10^{-3} BER), then the signal level at X——X is as follows:

= -114.2 dBm + 15

= -99.2 dBm

Therefore, the CNR in the analog IF is as follows:

= -87.3 (-)—92.2

= -11.9 dB.

It is necessary to calculate the minimum quantization level of the ADC. An 8-bit ADC with a full-scale input of 1 V pk-pk will be chosen:

1 bit level = $1 \text{ V}/2^{N}$

where N = the number of bits:

= 3.9 mV pk-pk

= 1.95 mV pk

= $1.95/\sqrt{2}$ mV rms

= 1.38 mV rms

The output of the stage preceding the ADC will be assumed to have a 50-ohm impedance. It is necessary to normalize the ADC minimum quantization level to 50 ohms:

Level in dBm = 20 log ($\sqrt{20} \times 1.38$ mV)

= -44.2 dBm

The noise power threshold presented to the ADC, however, is -87.3 dBm. For the ADC to see the minimum signal, the input must be raised to -44.2 dBm. This can be achieved by a gain stage, or stages, before the ADC, where gain

= -87.3 (-)-44.2

= 43.1 dB

An alternative process is to add out-of-band noise—to dither the ADC across the minimum quantization threshold. This minimizes the impact on the CNR of the signal but may not be necessary given the noisy nature of a W-CDMA signal.

Received Signal Strength

The received signal strength for both mobile handsets and fixed base stations in a cellular network is widely varying and unpredictable, because of the variable nature of the propagation path. Since the handset and Node B receiver front end must be extremely linear to prevent intermodulation occurring, the variation of signal strength persists through the RF stages (filter, LNA, mixer) and into the IF section.

A customary approach to IF and baseband processing in the superhet receiver is to sample and digitize the modulated IF. The IF + modulation must be sampled at a rate and a quality that maintains the integrity of the highest-frequency component—that is, IF + (modulation bandwidth/2). The digitization may be performed at a similar rate (oversampling) or at a lesser rate calculated from the modulation bandwidth (bandwidth or undersampling). Either digitization method must fulfill the Nyquist criteria based on the chosen process (oversampling or bandwidth sampling).

From an analysis of the sampling method, modulation bandwidth, required CNR, and analog-to-digital converter linearity, the necessary number of ADC bits (resolution) can be calculated. The number of bits and linearity of the ADC will give the spurious free dynamic range (SFDR) of the conversion process.

For TDMA handset requirements, the resolution will be in the order of 8 to 10 bits. For IMT2000DS/MC, 4 to 6 bits may be acceptable. Typical received signal strength variation can be in excess of 100 dB. This equates to a digital dynamic range of 16 or 18 bits. The

implementation option is therefore to use a 16 to 18 bit ADC, or to reduce the signal strength variation presented to the ADC to fit within the chosen number of ADC bits.

A typical approach to signal dynamic range reduction is to alter the RF or IF gain, or both, inversely to the signal strength prior to analog-to-digital conversion. The process of gain control is referred to as AGC (automatic gain control) and uses variable-gain amplifiers controlled by the RSSI.

RSSI response time must be fast enough to track the rate of change of mean signal strength—to prevent momentary overload of subsequent circuits—but not so fast that it tracks the modulation envelope variation, removing or reducing modulation depth. The RSSI function may be performed by a detector working directly on the IF or by baseband processing that can average or integrate the signal over a period of time.

Simple diode detectors have been previously used to measure received signal strength. They suffer from limited dynamic range (20/25 dB), poor temperature stability, and inaccuracy. The preferred method is to use a multistage, wide-range logarithmic amplifier. The frequency response can be hundreds of MHz with a dynamic range typically of 80/90 dB.

The variable amplifiers require adequate frequency response, sufficient dynamic range control (typically 60 dB+), and low distortion. Additionally, the speed of response must track rate of mean signal level change. It is a bonus if they can directly drive the ADC input, that is, with a minimum of external buffering.

The concept of dynamic range control has a limited application in base station receivers, as weak and strong signals may be required simultaneously. If the gain was reduced by a strong signal, a weak signal may be depressed below the detection threshold. A wider dynamic range ADC must be used.

IMT2000TC

In Chapter 1 we identified that some operators were being allocated 2 × 10 MHz paired channel allocations for IMT2000 and 1 × 5 MHz nonpaired channel (see Table 3.5). There are two nonpaired bands:

- TDD1 covers 4 × 5 MHz channels between 1900 and 1920 MHz.
- TDD2 covers 3 × 5MHz channels between 2010 and 2025 MHz.

Table 3.5 Band Allocations Including Nonpaired Bands (TDD1 and TDD2)

FREQUENCY (MHZ)	BANDWIDTH ALLOCATION (MHZ)	AIR INTERFACE	NONPAIRED BANDS
1900-1920	20	IMT2000TC	TDD1
1920-1980	60	IMT2000DS	
1980-2010	30	Satellite component (FDD)	

Table 3.5 *(Continued)*

FREQUENCY (MHZ)	BANDWIDTH ALLOCATION (MHZ)	AIR INTERFACE	NONPAIRED BANDS
2010-2025	15	IMT2000TC	TDD2
2110-2170	60	IMT2000DS	
2170-2200	30	Satellite component (FDD)	

Because the channel is not duplex spaced (the same RF channel is used for downlink and uplink), the channel is reciprocal. It is therefore theoretically possible to use the RAKE filter in the handset as a predistortion device. The benefit is that this allows the implementation of a relatively simple (i.e., RAKE-less) picocell base station.

The frame and code structure are slightly different to IMT2000DS. The 15-slot 10-ms frame is retained, but each of slots can support a separate user or channel. Each user or channel slot can then be subdivided into 16 OVSF spreading codes. The spreading factors are from 1 to 16. (Spreading factor 1 does not spread!)

The combination of time-division duplexing, time-division multiplexing, and a code multiplex provides additional flexibility in terms of bandwidth on demand, including the ability to support highly asymmetric channels. The duty cycle can also be actively reduced (a 1/15 duty cycle represents a 12 dB reduction in power). A mid-amble replaces the pilot tone and provides the basis for coherent detection. We revisit IMT2000TC access protocols in Part III of this book, "3G Network Hardware."

GPS

In addition to producing a dual-mode IMT2000DS/IMT2000 TC handset, the designer may be required to integrate positioning capability. There are at least eight technology options for providing location information: cell ID (with accuracy dependent on network density), time difference of arrival, angle of arrival, enhanced observed time difference (handset-based measurement), two satellite options (GPS and GLONASS, or Global Navigation Satellite System), a possible third satellite option (Galileo), and assisted GPS (network measurements plus GPS).

GPS receives a signal from any of the 24 satellites (typically 3 or 4) providing global coverage either at 1.5 GHz or at 1.5 GHz and 1.1 GHz, for higher accuracy. The RF carrier carries a 50 bps navigational message from each satellite giving the current time (every 6 seconds), where the satellite is in its orbit (every 30 seconds) and information on all satellite positions (every 12.5 minutes). The 50 bps data stream is spread with a 1.023 Mcps PN code.

The huge spreading ratio (1,023,000,000 bps divided by 50) means that the GPS receiver can work with a very low received signal level—typically 70 nV into 50 ohms, compared to a handset receiving at 1 μV. In other words, the GPS signal is 143 times smaller. The received signal energy is typically -130 dBm. The noise floor of the receiver is between -112 dBm and -114 dBm (i.e., the received signal is 16 to 18 dB below the noise floor).

Although the GPS signal is at a much lower level, GPS and IMT2000DS do share a similar signal-to-noise ratio, which means that similar receiver processing can be used to recover both signals. The practical problem tends to be the low signal amplitudes and high gains needed with GPS, which can result in the GPS receiver becoming desensitized by the locally generated IMT2000 signal. The solution is to provide very good shielding, to be very careful on where the GPS antenna is placed in relation to the IMT2000 antenna, or to not receive when the handset is transmitting.

If the GPS receiver is only allowed to work when the cellular handset is idle, significant attention has to be paid to reducing acquisition time.

An additional option is to use assisted GPS (A-GPS). In A-GPS, because the network knows where the handset is physically *and* knows the time, it can tell the handset which PN codes to use (which correlate with the satellites known to be visible overhead). This reduces acquisition time to 100 ms or less for the three satellites needed for latitude, longitude, and altitude, or the four satellites needed for longitude, latitude, and altitude.

Bluetooth/IEEE802 Integration

Suppose that, after you've designed an IMT2000DS phone that can also support IMT2000TC and GSM 800, 900, 1800, the marketing team reminds you that you have to include a Bluetooth transceiver. Bluetooth is a low-power transceiver (maximum 100 mW) that uses simple FM modulation/demodulation and frequency hopping at 1600 hops per second over 79 × 1 MHz hop frequency between 2.402 and 2.480 GHz (the Industrial Scientific Medical, or ISM, band). Transmit power can be reduced from 100 mW (+20 dBm) to 0 dBm (1.00 mW) to -30 dBm (1μW) for very local access, such as phone-to-ear, applications.

Early implementations of Bluetooth were typically two-chip, which provided better sensitivity at a higher cost; however, present trends are to integrate RF and baseband into one chip, using CMOS for the integrated device, which is low-cost but noisy. The design challenge is to maintain receive sensitivity both in terms of device noise and interference from other functions within the phone.

Supporting IEEE 802 wireless LAN connectivity is also possible, though not necessarily easy or advisable. The IEEE 802 standard supports frequency hopping and direct-sequence transceivers in the same frequency allocation as Bluetooth. Direct sequence provides more processing gain and coherent demodulation (with 3 dB of sensitivity) compared to the frequency hopping option, but it needs a linear IQ modulator, automatic frequency control for I/Q spin control, and a linear power amplifier.

Infrared

Infrared provides an additional alternative option for local access wireless connectivity. The infrared port on many handsets is used for calibration as the handset moves down the production line, so it has paid for itself before it leaves the factory. Costs are also low, typically less than $1.50.

Infrared standards are evolving. The ETSI/ARIB IRDA AIR standard (area infrared) supports 120° beamwidths, 4 Mbps data rates over 4 meters, and 260 kbps over 8 meters. This compares to a maximum 432.6 kbps of symmetric bandwidth available for Bluetooth.

IEEE 802 also supports an infrared platform in the 850- to 950-nanometer band, giving up to 2 W peak power, 4- or 16-level pulse modulation, and a throughput of 2 Mbps. Higher bit rate RF options are available in the 5 GHz ISM band, but at present these are not included in mainstream 3G cellular handset specifications.

Radio Bandwidth Quality/Frequency Domain Issues

We have just described how code domain processing is used in IMT2000DS to improve radio bandwidth quality. Within the physical layer, we also need to comprehend frequency domain and time domain processing. If we wished to be very specific, we would include source coding gain (using processor bandwidth to improve the quality of the source coded content), coherence bandwidth gain (frequency domain processing), spreading gain (code domain processing), and processing gain (time domain processing, that is, block codes and convolutional encoders/decoders).

Let's first review some of the frequency domain processing issues (see Table 3.6). We said that the IMT2000 spectrum is tidily allocated in two 60 MHz paired bands between 1920-80 and 2110 and 2170 MHz with a 190 MHz duplex spacing. In practice, the allocations are not particularly tidy and vary in minor but significant ways country by country.

Table 3.6 IMT2000 Frequency Plan

TDD1	TDD2				
1900-1920	1920-1980	1980-2010	2010-2025	2110-2170	2170-2200
		SATELLITE			SATELLITE
4 × 5 MHz nonpaired	12 × 5 MHz paired		3 × 5 MHz nonpaired	12 × 5 MHz paired	
IMT2000 TC			IMT2000 TC		

Figure 3.22 shows how spectrum was allocated/auctioned in the United Kingdom. It is not untypical of any country in which the spectrum is divided up between five operators, as follows:

- *License A* (Hutchison) has 14.6 MHz (3 × 5 MHz less a guard band) paired band allocation.
- *License B* (Vodafone) has 14.8 MHz (3 × 5 MHz less a guard band) paired band allocation.
- *License C* (BT3G) has 10 MHz (2 × 5 MHz allocation in the paired band) and 5 MHz at 1910 MHz in the TDD1 nonpaired band.
- *License D* (One2One) has 10 MHz (2 × 5 MHz allocation in the paired band) and 5 MHz at 1900 MHz in the TDD1 nonpaired band.
- *License E* (Orange) has 10 MHz (2 × 5 MHz in the paired band) and 5 MHz at 1905 MHz in the TDD1 nonpaired band.

The German allocation is different in that 10 MHz of paired bandwidth is allocated to six operators (6 × 10 MHz = 60 MHz), then all four of the TDD1 channels are allocated (1900 to 1920 MHz), along with one of the TDD2 channels (see Figure 3.23).

3GPP1 also specifies an optional duplex split of 134.8 and 245.2 MHz to support possible future pairing of the TDD1 and TDD2 bands. Although this is unlikely to be implemented, the flexible duplex is supported in a number of handset and Node B architectures.

The fact that 5 MHz channels are allocated differently in different countries means that operators must be prepared to do code planning and avoid the use of codes that cause adjacent channel interference to either other operators in the same country or other operators in immediately adjacent countries. It is therefore important to explore the interrelationship between particular combinations of spreading codes and adjacent channel performance.

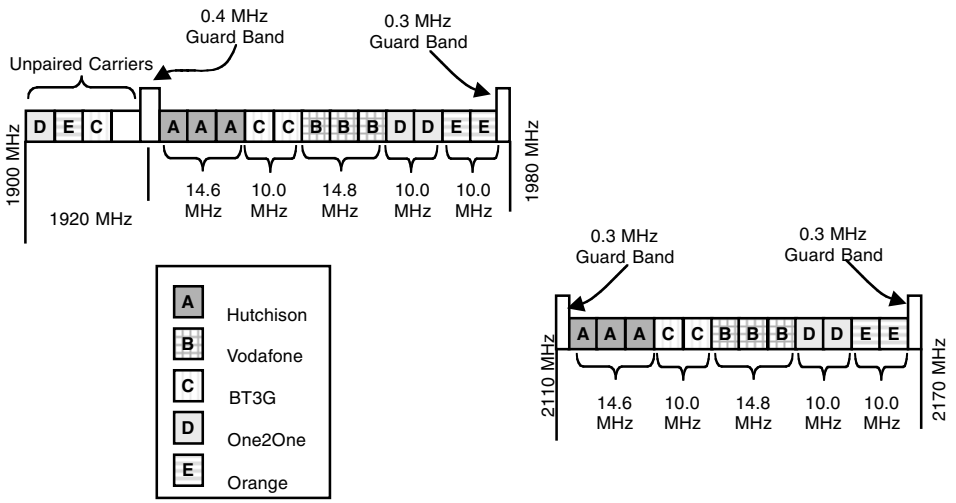


Figure 3.22 Countries with five operators—for example, United Kingdom.

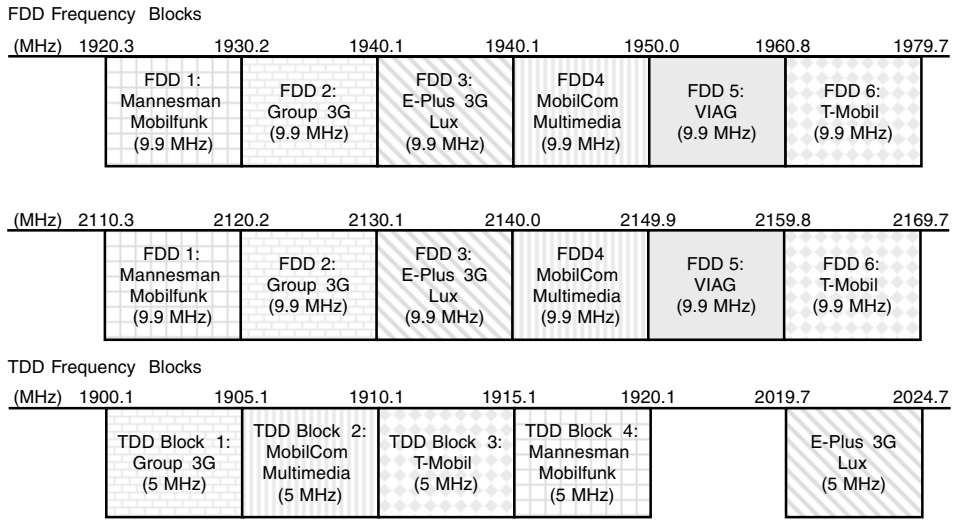


Figure 3.23 Countries with six operators—for example, Germany.

The three measurements used are as follows:

- ACLR (Adjacent Channel Leakage Ratio), formerly Adjacent Channel Power Ratio
- ACS (Adjacent Channel Selectivity)
- ACIR (Adjacent Channel Interference Ratio), formerly Adjacent Channel Protection Ratio.

OVSF code properties also determine peak-to-average ratios (PAR), in effect the AM components produced as a result of the composite code structure. PAR in turn determines RF PA (RF Power Amplifier) linearity requirements, which in turn determine adjacent channel performance. In other words, peak-to-average power ratios are a consequence of the properties of the offered traffic—the instantaneous bit rate and number of codes needed to support the per user multiplex.

We find ourselves in an interactive loop: We can only determine frequency domain performance if we know what the power spectral density of our modulated signal will be, and we only know this if we can identify statistically our likely offered traffic mix.

Out-of-channel performance is qualified using complementary cumulative distribution functions—the peak-to-average level in dB versus the statistical probability that this level or greater is attained. We use CCDF to calculate the required performance of particular system components and, for example, the RF PA.

ACLR is the ratio of transmitted power to the power measured after a receiver filter in the adjacent RF channel. It is used to qualify transmitter performance. ACS is the ratio of receiver filter attenuation on the assigned channel frequency to the receiver filter attenuation on the adjacent channel frequency and is used to qualify receiver performance. When we come to qualify system performance, we use ACIR—the adjacent channel interference ratio. ACIR is derived as follows:

$$ACIR = \frac{1}{\frac{1}{ACLR} + \frac{1}{ACS}}$$

(We review system performance in Chapter 11 on network hardware.)

As a handset designer, relaxing ACLR, in order to improve PA efficiency, would be useful. From a system design perspective, tightening ACLR would be helpful, in order to increase adjacent channel performance. ACLR is in effect a measure of the impact of nonlinearity in the handset and Node B RF PA.

We can establish a conformance specification for ACLR for the handset but need to qualify this by deciding what the PAR (ratio of the peak envelope power to the average envelope power of the signal) will be. We can minimize PAR, for example, by scrambling QPSK on the uplink (HPSK) or avoiding multicodes. Either way, we need to ensure the PA can handle the PAR while still maintaining good ACL performance. We can qualify this design trade-off using the complementary cumulative distribution function.

A typical IMT2000DS handset ACLR specification would be as follows:

- 33 dBc or -50 dBm at 5 MHz offset
- 43 dBc or -40 dBm at 10 MHz offset

Radio Bandwidth Quality/Time Domain Issues

We mentioned channel coding briefly in Chapter 1. 3G cellular handsets and Node Bs use many of the same channel coding techniques as 2G cellular—for example, block coding and convolutional coding. We showed how additional coding gain could be achieved by increasing the constraint length of a convolutional decoder. This was demonstrated to yield typically a 1/2 dB or 1 dB gain, but at the expense of additional decoder complexity, including processor overhead and processor delay.

In GPRS, adaptive coding has been, and is being, implemented to respond to changes in signal strength as a user moves away from a base station. This has a rather unfortunate side effect of increasing a user's file size as he or she moves away from the base station. At time of writing only CS1 and CS2 are implemented.

We also described interleaving in Chapter 1 and pointed out that increasing the interleaving depth increased the coding gain but at the cost of additional fixed delay (between 10 and 80 ms). Interleaving has the benefit of distributing bit errors, which means that convolutional decoders produce cleaner coding gain and do not cause error extension. If interleaving delay is allowable, additional coding gain can be achieved by using turbo coding.

IMT2000 Channel Coding

In IMT2000 the coding can be adaptive depending on the bit error rate required. The coding can be one of the following:

- Rate 1/3 convolutional coding for low-delay services with moderate error rate requirements (1 in 10^3)
- 1/3 convolutional coding and outer Reed-Solomon coding plus interleaving for a 1 in 10^6 bit error rate

In IMT2000 parallel code concatenation, turbo coding is used. Turbo codes have been applied to digital transmission technology since 1993 and show a practical trade-off between performance and complexity. Parallel code concatenation uses two constituent encoders in parallel configuration with a turbo code internal interleaver. The turbo coder has a 1/3 coding rate. The final action is enhancement of E_b/N_o by employing puncturing.

Turbo coders are sometimes known as *maximum a posteriori decoders*. They use prior (*a priori*) and post (*a posteriori*) estimates of a code word to confer distance.

Reed-Solomon, Viterbi, and Turbo Codes in IMT2000

For IMT2000, Reed-Solomon block codes, Viterbi convolutional codes, and turbo codes are employed. The combination of Reed-Solomon and Viterbi coding can give an improvement in S/N for a given BER of 6 to 7 dB. Turbo coding used on the $1 \times 10\text{-}6$ traffic adds a further 1.5- to 3 dB improvement. Total coding gain is ~ 8 dB.

The benefit of coding gain is only obtained above the coding threshold, that is, when a reasonable amount of E_b/N_o is available (rather analogous to wideband FM demodulator gain—the capture effect). Turbo coding needs 300 bits per TTI (Transmission Time Interval) to make turbo coding more effective than convolutional coding. This means turbo coding only works effectively when it is fed with a large block size (anything up to 5114 bits blocks). This adds to the delay budget and means that turbo coding is nonoptimum for delay-sensitive services.

Future Modulation Options

Present modulation schemes used in IMT2000DS are QPSK, along with HPSK on the uplink. 8 PSK EDGE modulation also needs to be supported. In 1xEV, 16-level QAM is also supported.

As the modulation trellis becomes more complex—that is, the phase and amplitude states become closer together and symbol time recovery becomes more critical—-it becomes worthwhile to consider *trellis coding*. In trellis coding, where modulation states are close together, the data is coded for maximal distance; when the data is far apart, they are coded for minimal distance. Trellis coding is used in certain satellite access and fixed access networks.

Characterizing Delay Spread

Delay spread is caused by multipath on the radio channel—that is, different multiple path lengths between transmitter and receiver. Radio waves reflecting off buildings will also change in phase and amplitude. Radio waves travel at 300,000 kmps. This means that in 1 ms they will have traveled 300 km, in 250 μ they will have traveled 75 km, and in 25 μs they will have traveled 7.5 km.

A 1 km flight time is equivalent to an elapsed time of 3.33 2 μ, and a 100 m flight time is equivalent to an elapsed time of 0.33 μs. In other words, delay spread is a function of flight time. Radio waves take approximately 3.33 2 μ to travel 1 km. A 100-m difference between two path lengths is equivalent to a delay difference of 0.33 μs. Chip duration in IMT2000 DS is 0.26 μs. Therefore, multipaths of >70 m are resolvable; multipaths of <70 m are not. If all the energy of each chip in a user's chip sequence falls within one chip period, you do not need a RAKE receiver.

Delay spreads increase as you go from dense urban-to-urban to suburban and are largest, as you would expect, in mountainous areas (termed the "Swiss mountain effect"). Table 3.7 shows typical measured delay spreads. In GSM, the GSM equalizer specification defines that the equalizer should be capable of correcting a 4-bit shift on the channel (that is, a 16 μs delay spread, or a 4.8 km multipath). In practice, most GSM equalizers provide more dynamic range, but early GSM phones quite often suffered overrun in rural/mountainous areas. RAKE receiver dynamic range issues are not dissimilar. Delay spread is independent of frequency; it is a function of flight distance, not frequency.

Practical Time Domain Processing in a 3G Handset

We have established that a third-generation handset needs to process in the frequency domain, the code domain (chip level), and the time domain (bit level and symbol level). These processing mechanisms need to comprehend the ambiguities introduced by the radio path including time ambiguities (delay spread) and phase and frequency offsets.

Table 3.7 Typical Measured Delay Spreads

MEASURED BY	FREQUENCY	ENVIRONMENT	WORST CASE DELAY SPREAD
Rappaport	900 MHz	Washington (urban) Oakland (mountains)	7 to 8 μs 13.5 μs
Parsons	450 MHz	Birmingham (suburban)	2.2 to 3 μs
Turkmani Arowojolu	900 MHz	Liverpool (urban and suburban)	6 μs
Zogg	210 MHz	Switzerland (mountains)	10 to 35 μs

Bandwidth quality is a function of how well those processing tasks are undertaken and how well the processing adapts to changed loading conditions. Let's examine some of the measurements used to qualify adaptive radio bandwidth performance.

Figure 3.24 shows a 12.2 kbps uplink voice channel and 2.5 kbps of embedded signaling. The channels are punctured and rate-matched and multiplexed to give a channel rate of 60 kbps. Spreading factor 64 is applied to provide the 3.84 Mcps rate. Variable gain is applied to take into account the spreading gain. The control channel is 15 kbps and is therefore spread with SF256 and gain-scaled to be -6 dB down from the DPDCH.

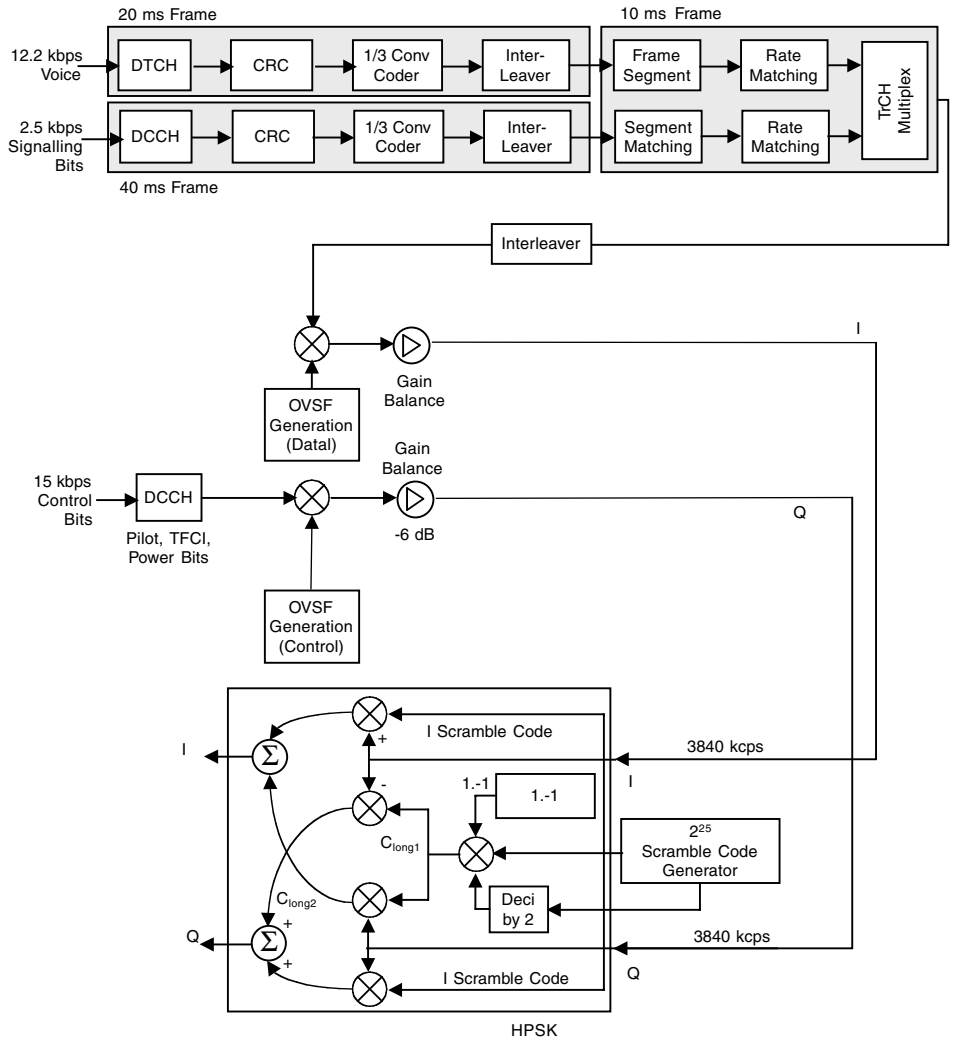


Figure 3.24 Uplink block diagram.

The complex scrambling applied to the uplink is a process known as *Walsh rotation*, which effectively continuously rotates the modulation constellation to reduce the PAR of the signal prior to modulation. It is also known as Hybrid Phase Shift Keying (HPSK) or sometimes orthogonal complex quadrature phase shift keying. HPSK allows handsets to transmit multiple channels at different amplitude levels while still maintaining acceptable peak-to-average power ratios.

Unlike the uplink, where the control bits are modulated onto the Q channel, the downlink multiplexes voice bits, signaling bits, and control bits together across the I and Q channels with slightly different rates resulting; voice and signaling bits are at 42 kbps, and pilot, power control, and TFCI control bits are at 18 kbps to give the 60 kbps channel rate.

Conformance/Performance Tests

You can examine handset and Node B performance at bit level, symbol level, chip level, slot level, and frame level (as shown in Table 3.8), with the Node B exercised by any one of the four reference measurement channels (12.2, 64, 144, and 384 kbps) and the handset exercised by any one of five measurement channels (12.2, 64, 144, 384, and 786 kbps).

The measurement terms are E_b = energy in a user information bit, E_c = energy in every chip, E_b/N_o = ratio of bit energy to noise energy, and I_O = interference + noise density. You will also see the term E_b/N_t used to describe the narrowband thermal noise (for example, in adaptive RAKE design).

Chip level error vector magnitude includes spreading and HPSK scrambling. It cannot be used to identify OVSF or HPSK scrambling errors, but it can be used to detect baseband filtering, modulation, or RF impairments (the analog sections of the transmitter). It could, for example, be used to identify an I/Q quadrature error causing constellation distortion.

QPSK EVM measurement can be used to measure single DPDCH channels, but we are more interested in representing the effect of complex channels. This is done using the composite EVM measurement (3GPP modulation accuracy conformance test), as shown in Figure 3.25.

Table 3.8 Conformance Performance Tests

LEVEL	TEST	SYMBOL LENGTH	RATE
Bits	Bit error rate + EVM	For example, 60 kbps = 16.66 μs	60 kHz
Symbols	Error vector magnitude	For example, 30 ksps = 33.33 μs	30 kHz
Chip	Error vector magnitude	0.26 μs	3.84 MHz
Slot	Power control	666.66 μs	1500 Hz
Frame	Frame erasure	10 ms	100 Hz

A reference signal is synthesized, downconverted (I and Q recovery), and passed through an RRC filter. Active channels are then descrambled, despread, and Binary Phase Shift Key (BPSK) decoded down to bit level. The despread bits are perfectly remodulated to obtain a reference signal to produce an error vector. Composite EVM can be used to identify all active channel spreading and scrambling problems and all baseband, IF, and RF impairments in the Tx chain.

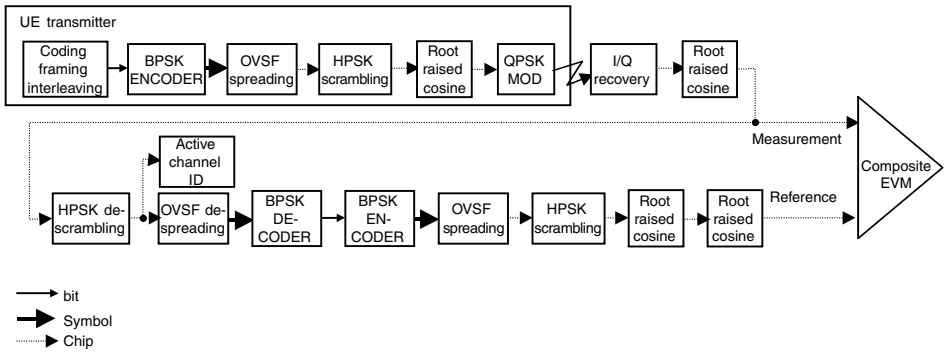


Figure 3.25 Composite EVM (bit-level EVM).

Impact of Technology Maturation on Handset and Network Performance

As devices improve and as design techniques improve handset sensitivity improves. There is also a performance benefit conferred by volume: better control of component tolerance. This could be observed as GSM matured over a 15-year period.

A similar performance evolution can be expected with IMT2000. Effectively the first two generations of cellular have followed a 15-year maturation cycle. It is reasonable to expect 3G technologies to follow a similar cycle.

3GPP2 Evolution

We have spent some time (some U.S. readers may say too much time) on the IMT2000DS/W-CDMA air interface. It is time to review the parallel evolution of CDMA2000/IS95. CDMA2000 is the term used to describe the air interface. IS2000 comprehends the air interface and network interfaces (interfaces to the IS41 network). Table 3.9 shows how IS95A/B has evolved with the adoption of variable length Walsh codes, use of QPSK on the downlink and HPSK on the uplink, more granular power control, supplemental code channels (multiple per-user channels on the downlink and uplink—one fundamental, up to seven supplemental), the option of multiple RF channels within a 5 MHz channel spacing (3xRTT), and the option of higher-level modulation (1xEV).

RC refers to radio configuration and specifies a set of data rates, spreading rates (SR) and coding.

In practice, it seems unlikely that 3xRTT will be implemented, and most deployment is presently focused on 1xEV using the present spreading rate of 1.288 Mcps as the most logical forward-evolution path. It may also be that fixed-length Walsh codes are used rather than variable length—variable data rates can be supported through adaptive modulation. However, variable-length Walsh codes do remain as an option in the standard.

Table 3.9 3GPP2 Evolution

IS95A/IS95B	IS2000 REL 0 TO IS2000A*
cdmaOne	CDMA2000
64 Walsh Codes	128 variable-length Walsh codes
Dual BPSK	QPSK (HPSK uplink) Closed-loop power control (800 Hz) 0.25/0.5/1 dB steps
RC1 9.6 kbps 14.4 kbps	RC2 1.5 to 2.7 to 4.8 to 9.6 kbps voice 19.2 to 38.4 to 76.8 to 153.6 kbps data
1 × RTT	3 × RTT (IMT2000 MC) 1 × EV QPSK/8 PSK/16-level QAM

*IS2000 Rel 0 is backward-compatible with IS95.

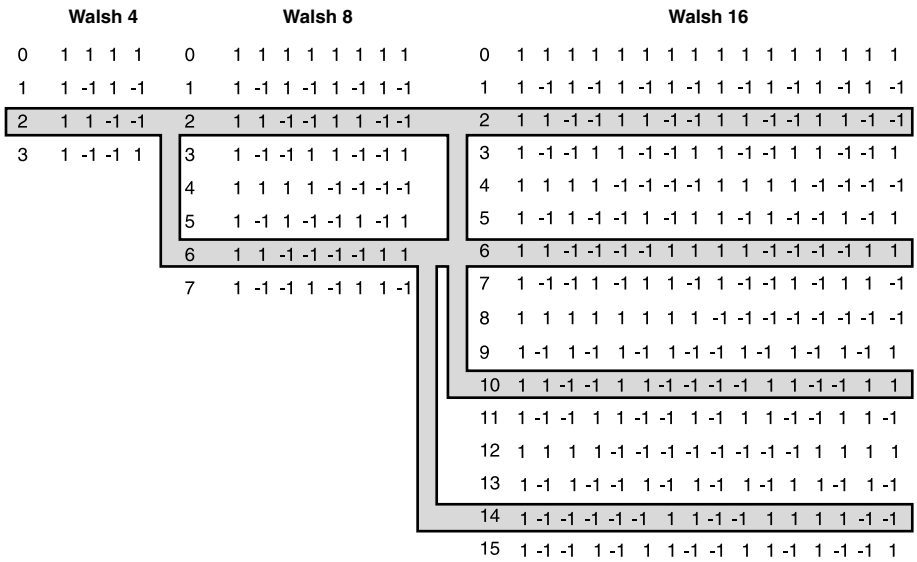


Figure 3.26 Variable length Walsh codes.

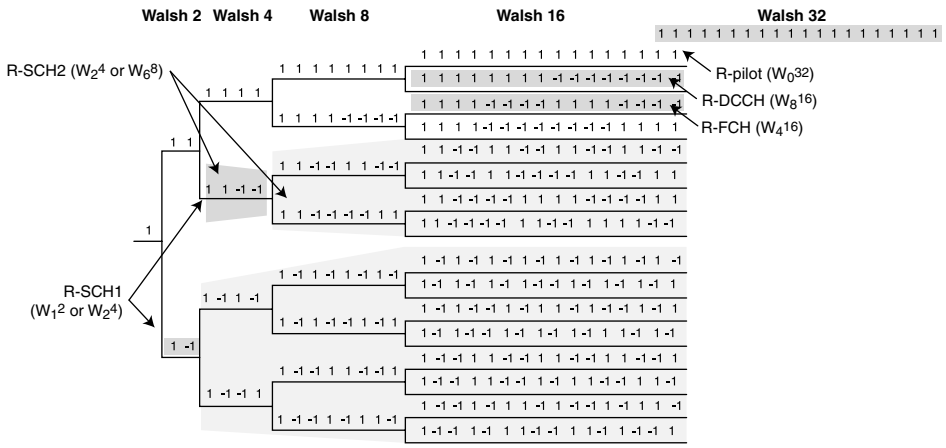


Figure 3.27 Reverse link code structure.

The length of the Walsh code is varied—from 4 to 128 chips—to accommodate different data rates. As the data rate increases, the symbol period gets shorter. The final chip rate stays constant—that is, fewer Walsh code chips are accommodated within the symbol period. If you re-order the code channels so that related code channels are adjacent to each other, you will have reproduced the OVSF code tree used in W-CDMA!

Figure 3.26 shows how, by rearranging the Walsh code tree you end up with an OVSF code tree. The dark portion represents a branch of the OVSF code tree.

As with W-CDMA, the use of HPSK on the uplink restricts which Walsh codes can be used to still keep PAR within acceptable limits. All the light gray codes shown in Figure 3.27. are nonorthogonal to the selected (dark gray) codes.

Table 3.10 gives examples of bit rate versus Walsh code length. Multiplying the code length by the data rate gives you the code rate.

Table 3.10 Bit Rate and Walsh Code Cover

WALSH CODE LENGTH					
128 BITS (WALSH 128)	**64 BITS (WALSH 64)**	**32 BITS (WALSH 32)**	**16 BITS (WALSH 16)**	**8 BITS (WALSH 8)**	**4 BITS (WALSH 4)**
9.6 kbps	19.2 kbps	38.4 kbps	76.8 kbps	153.6 kbps	307.2 kps
$9.6 \times 128 =$ 1.2288 Mcps				$153.6 \times 8 =$ 1.2288 Mcps	

CDMA2000 Downlink and Uplink Comparison

The CDMA2000 downlink has not changed significantly from cdmaOne (IS95). There is one forward fundamental channel (F-FCH) and up to eight forward supplemental code channels (for RC2).

The uplink is significantly different because of the decision that handsets should be capable of transmitting more than one code simultaneously (a multicode uplink). This requires a reverse pilot (R-Pilot + power control), which allows the base station to do synchronous detection, and a reverse fundamental channel (R-FCH) for voice. Other supplemental channels (R-SCH) can be added in as required. The RDCCH (reverse dedicated control channel) is used to send data or signaling information. The channels can be assigned to either the I or the Q path and then complex-scrambled to generate the HPSK signal for modulation, to reduce the peak-to-average ratio.

Implementation Options

The original proposal from 3GPP2 to the ITU was based on the assumed need to fill 5 MHz of RF bandwidth. This could be achieved either by increasing the spreading ratio, as in SR3DS (direct sequence) shown in Figure 3.28, or by putting three CDMA2000 1.25 MHz channels together (SR3MC).

In practice, CDMA2000 is not being implemented into 5 MHz channels, and higher data rates are being achieved by using higher-level modulation schemes (8 PSK and 16-level QAM) in 1.25 MHz RF channels—the variant known as 1xEV.

Linearity and Modulation Quality

If higher levels of modulation are to be supported together with multiple codes per user on the uplink, then significant attention must be paid to PA linearity, both in the handset and the base station, if clipping is to be avoided. As mentioned, each coded channel represents a phase argument. The phase argument can be compromised by a loss of code-to-code orthogonality or AM to PM effects introduced by a PA in compression. It is important also to keep the code power well contained.

On the downlink, the BTS specification specifies that 91.2 percent of the correlated pilot power should be contained in the total transmission power. This means that 8.8 percent of the power produced potentially causes interference to other Walsh channels and embarrassment to the handset receive process.

As with IMT2000DS, modulation quality equates directly to capacity and coverage capability. Modulation quality in CDMA2000 is measured using RHO (equivalent to EVM). Causes of poor RHO could include RF PA compression, amplitude and phase errors in the IQ modulators, carrier feed-through, and spurious Tx signals.

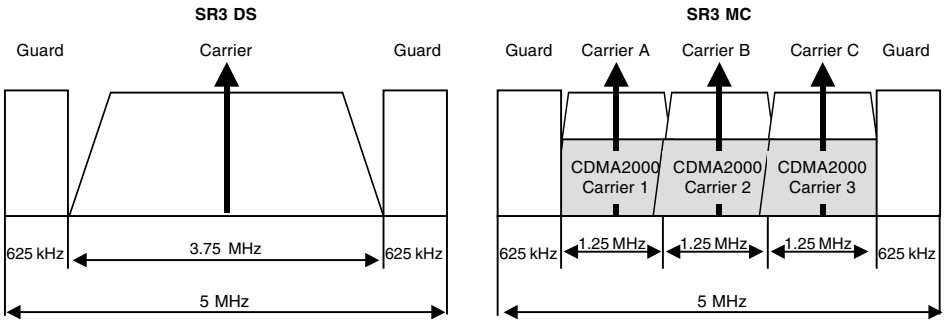


Figure 3.28 Implementation options for a 5 MHz channel.

Frequency Tolerance

Frequency tolerance also needs to be tightly specified. Failure of a GPS receiver will isolate a base station. The base station will still serve local handsets but will drift away from the rest of the network and become invisible—an island cell that takes with it the handsets it is presently supporting. Table 3.11 describes permitted frequency tolerance.

In IMT2000DS, a short code is used to bring the handset onto channel in terms of time synchronization. This helps to relax the frequency reference in the handset (reducing RF component count) but makes the long code acquisition process quite complex.

CDMA2000 is much simpler. All base stations share the same long code, but each base station is offset by 64 chips from the next base station. There are 512 possible offsets. When the handset is turned on, it should lock onto the long code with the shortest PN offset, because this will be, by implication, the nearest base station in terms of flight path.

The only disadvantage to this is that CDMA2000 timing errors need to be carefully managed to maintain acquisition performance and prevent false acquisition. A timing error in the offset higher than 3 μs can cause system performance degradation.

The orthogonality of Walsh codes and OVSF codes disappears if the codes are not time-aligned. Sources of timing errors can be within the application-specific integrated circuit (ASIC); time adjustment parameters), and delay in baseband signal paths or Walsh code intermodulation. The pilot to Walsh channel time tolerance is specified at <50 ns.

Table 3.11 Frequency Tolerance

1900 MHz	± 0.05 ppm	± 99 Hz at 1980 MHz
800 MHz	± 0.05 ppm	± 40 Hz at 800 MHz

Phase errors between the receiver local oscillator and decorrelated Walsh channels create IQ interference and Walsh code intermodulation. The phase tolerance must be less than 2.86 degrees (50 milli-radians). The CDMA2000 handset uses the pilot channel phase as a reference. If the pilot channel phase is not aligned with the traffic channels, the traffic channels will not be demodulated!

Frequency/Power Profile

As with IMT2000DS, CDMA2000 is tightly specified in terms of spurious emissions, measured both for their impact in-band and out-of-channel, as shown in Figure 3.29.

In markets with legacy 30 kHz channel-spaced networks (US TDMA 800 MHz and 1900 MHz), adjacent channel power ratios need to be qualified with respect to adjacent narrowband channels. Similar specifications are required for out-of-band performance. The CDMA2000 specification requires spurious emissions outside the allocated system band (measured in a 30 kHz bandwidth) to be 60 dB below the mean output power in the channel bandwidth or -13 dBm, whichever is smaller.

Frame erasure rate can be used as a measure of receiver performance, provided coding and error correction is applied equally to all bits—that is, there are not classes of bits with different levels of error correction. Frame erasure rate is the ratio of the number of frames of data received that are deleted because of an unacceptable number of errors to the total number of frames transmitted. Frame erasure rate is used as a measure of receiver performance.

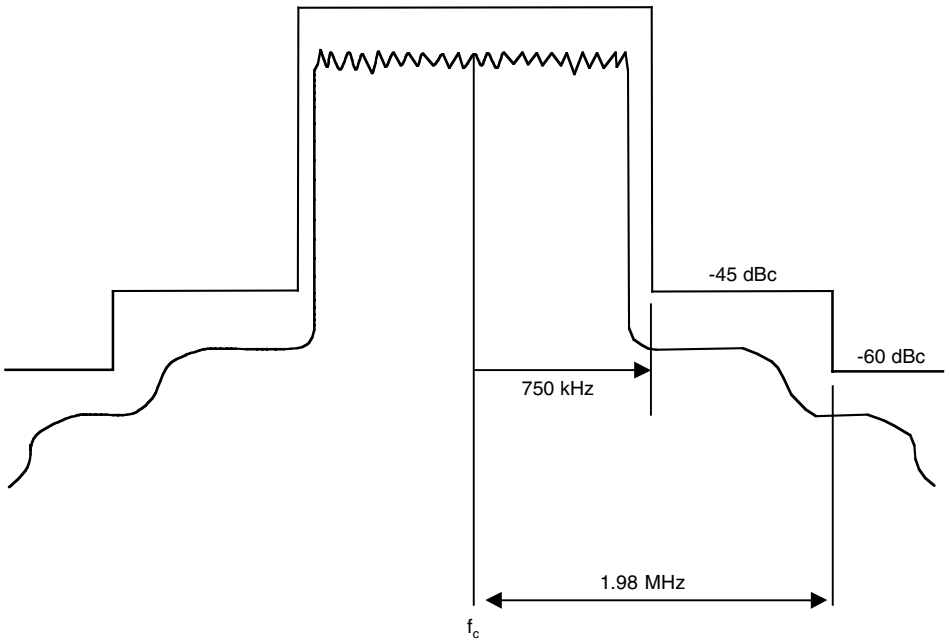


Figure 3.29 In-band/out-of-channel measurements.

We can use frame erasure rate to measure sensitivity and dynamic range, spurious immunity, and performance in AWGN and fading channels. CDMA2000 uses 20 ms frames. Base station receiver performance is expressed in terms of FER versus E_b / N_o. The E_b / N_o required will be a function of data rate and channel requirements.

At system level, the use of a continuous pilot in CDMA2000 provides better channel sounding, compared to IMT2000DS, but it uses more transmit energy. The continuous common pilot channel provides the following:

- More accurate estimation of the fading channel
- Faster detection of weak multipath rays than the per-user pilot approach
- Less overhead per user

Turbo coding is used for higher data rates with K = 9 constraint length.

The forward link coding is adaptive. Interleaving can either be 20 or 5 ms. A 6-bit, 8-bit, 10-bit, 12-bit, or 16-bit CRC is used for frame error checking with 1/2, 1/3, 1/4 rate K=9 convolutional coding. Equivalent rate turbo codes are used on supplemental

channels. Each supplemental channel may use a different encoding scheme. Similarly, downlink coding is adaptive, using a 6-bit, 8-bit, 10-bit, 12-bit, or 16-bit CRC for frame error checking, and 9/16, 1/2, 1/3, 1/4 rate K = 9 convolutional coding. Equivalent rate turbo codes are used on supplemental channels. Each supplemental channel may use a different encoding scheme. Interleaving is again either 5 ms or 20 ms.

Closed-loop power control is carried out at an 800 Hz control rate. The open loop sets Tx power level based on the Rx power received by the mobile and compensates for path loss and slow fading. The closed loop is for medium to fast fading and provides compensation for open-loop power control inaccuracies. The outer loop is implementation-specific and adjusts the closed-loop control threshold in the base station to maintain the desired frame error rate. The step size is adaptive, either 1 dB, 0.5 dB, or 0.25 dB. As with IMT2000DS, power control errors will directly subtract from the link budget.

The power control dynamic range is as follows:

- Open loop ±40 dB
- Closed loop ±24 dB

Power control errors are typically 1.3 dB (low mobility) or 2.7 dB (high mobility).

Dynamic range is similar to other existing networks:

Mobile	79 dB
Base station	52 dB
FDD isolation	(45 MHz, 800 MHz, 80 MHz at 1900 MHz)
Class II mobile	55 dB Tx to Rx
Base	90 dB (higher effective power, 5 dB lower noise floor)

Class IV handsets are equivalent to Power Class 3 handsets in IMT2000DS (250 mW). Class V handsets are equivalent to Power Class 4 handsets in IMT2000DS (125 mW). Both networks also support higher-power mobiles.

Class I:	28 dBm < EIRP < 33 dBm (2 W)
Class II:	23 dBm < EIRP < 30 dBm
Class III:	18 dBm < EIRP < 27 dBm
Class IV:	13 dBm < EIRP < 24 dBm (250 mW)
Class V:	8 dBm < EIRP <21 dBm (125 mW)

CDMA 1xEV has a high data rate option for the downlink, separate 1.25 MHz RF channel, QPSK, 8 PSK, 16-level QAM, and evolution to meet IMT2000MC requirements (3xRTT). 1xEV adds adaptive modulation as a mechanism for increasing data throughput.

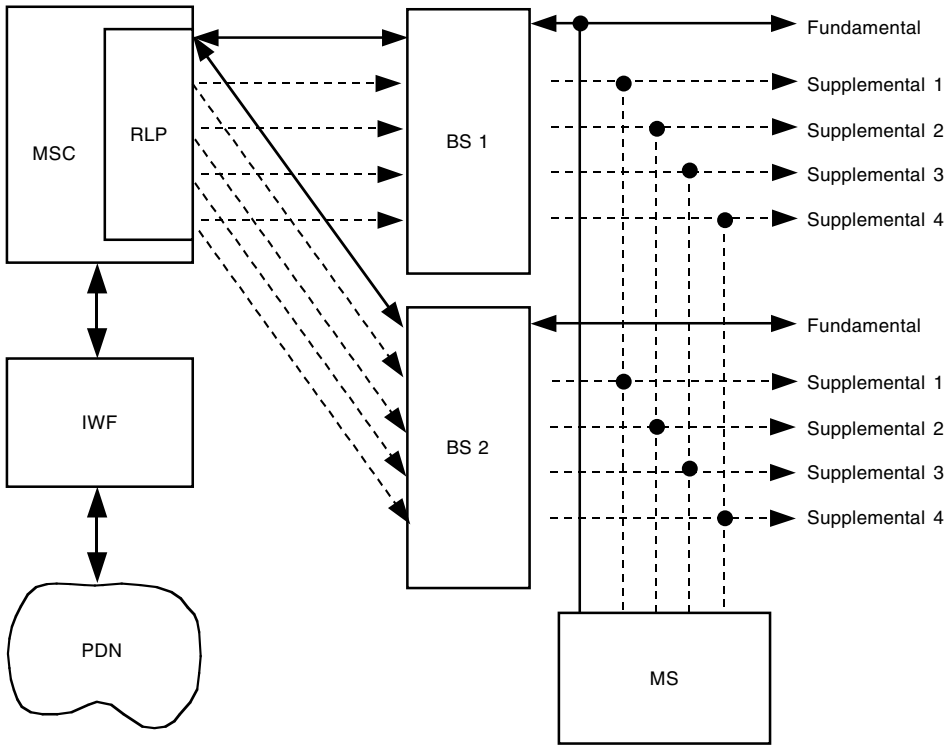


Figure 3.30 CDMA2000 handset in a soft handoff.

The Media Access Control (MAC) layer in IS2000 manages code allocation (the provision of physical layer resources to meet application layer requirements). An active high-rate mobile assigned a fundamental channel on origination negotiates high data rate service parameters. The mobile then sleeps but remains locked to a low-rate channel for synchronization and power control.

The handset signals a high data burst request by indicating to the base station (BS) its data backlog and maximum data rate requested. The handset includes pilot strength information for cells in its neighbor list, which indicates local interference levels. Additionally pilot strength measurements allow the base station to qualify instantaneous downlink capacity.

Supplemental code channels can then be allocated as required. In Figure 3.30, the handset communicates on the fundamental code channel with two base stations (BS1 and BS2). During a burst transmission, one or more supplemental code channels are assigned at BS1, BS2, or both. The MSC performs distribution on the forward link and selection on the reverse link. The Radio Link Protocol (RLP) does an Automatic Repeat Request (ARQ) and the interworking function (IWF) provides access to the packet data network.

When there is backlogged data, the mobile goes into active mode. If backlogged data exceeds a threshold, the mobile requests a supplementary channel (SCRM), sent on the fundamental code channel. The BS/MSC uses pilot strength measurements made by the mobile to decide on burst admission control and allocates supplementary channels. When backlogged data at the IWF exceeds a predetermined threshold, the IWF initiates a request for supplementary channels. The mobile is paged if not already in an active state.

In IS95B, a mobile is either active or dormant, and in CDMA2000, a handset can go into control hold, maintaining a dedicated control channel and power control (burst transmission with no added latency). In suspended state, there are no dedicated channels, although a virtual set of channels are maintained. In dormant state, there are no pre-allocated resources; in other words, the deeper the sleep, the lower the power consumption, but the longer it takes to wake up.

Summary

In this chapter we summarized the main tasks that need to be performed by a 3G handset, and we qualified code domain, frequency domain, and time domain performance issues. Typically, over a 15-year maturation cycle, handset performances improves on a year-by-year basis. and this delivers benefits in terms of network bandwidth quality.

In the following two chapters, we consider 3G handset hardware form factor and functionality and handset hardware evolution.

CHAPTER

4

3G Handset Hardware Form Factor and Functionality

In Chapter 2 we described the physical hardware needed to realize a multislot, multiband, multimode handset. In Chapter 3 we described the physical hardware needed to deliver multiple per-user channel streams. In this chapter we describe the application hardware components needed to realize a multimedia mobile handset and the impact of various hardware items in the handset on the offered traffic mix.

Impact of Application Hardware on Uplink Offered Traffic

First, we review the impact of the microphone and audio vocoder, the CMOS imager (or CCD imager), and the keyboard on uplink offered traffic. We then move on to a brief overview of rich media, followed by a section on the smart card SIM. These components are shown in Figure 4.1.

Voice Encoding/Decoding (The Vocoder)

Let's first look at the microphone and its impact on uplink offered traffic. In Chapter 1 we described briefly the process of talking into a microphone to produce an analog waveform (a varying voltage), which is then digitized in an analog-to-digital converter (ADC). In GSM, this produces a digital bit stream of 104 kbps, which then has to be compressed, typically to 13 kbps or lower, using a transfer from the time to the frequency domain—the basis for all speech synthesis codecs.

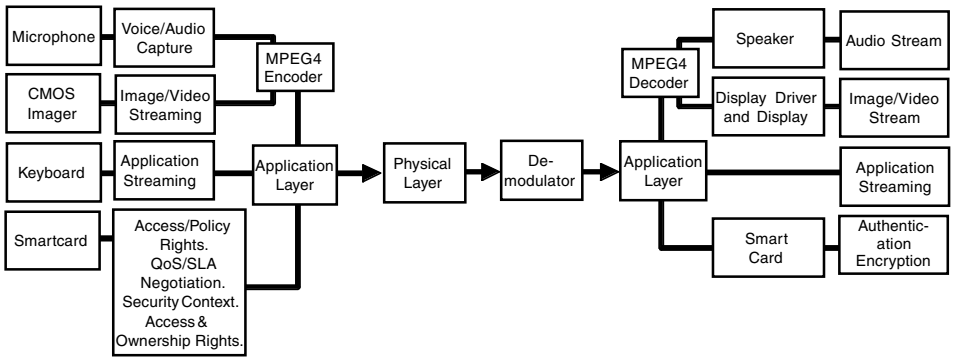


Figure 4.1 Application hardware components in a 3G handset.

Initially, cellular handset codecs were constant rate. The codecs specified for 3G are variable rate. In the case of the adaptive multirate (AMR) codec specified by 3GPP1 (the standards group for IMT2000DS/W-CDMA/UMTS), the rate is switchable between 4.75 and 12.2 kbps. The rate can be chosen to provide capacity gain (lower bit rate) or quality gain (higher bit rate). The codec is an adaptive codebook excitation linear prediction codec, which means speech waveforms are stored in a lookup table in the receiver.

3GPP2 (the standards group responsible for CDMA2000) have specified a variable-rate vocoder described as a selectable mode (SMV) vocoder. It adapts dynamically to the audio input waveform.

Figure 4.2 shows performance comparisons between the SMV and AMR codecs, with the SMV codec providing a better quality/capacity trade-off—at the cost of some additional processing overhead.

Voice quality is measured using a *mean opinion score* (MOS). Mean opinion scoring is essentially an objective method for comparing subjective responses to quality—a group of users listen to the voice quality from a handset and provide a score. A score of 5 is very good (equivalent to a wireline connection); a score of 2.5 would be comprehensible but uncomfortable to listen to, and many of the harmonic qualities of the person's voice will have been lost, to the point where it is difficult to recognize who is speaking. Figure 4.3 shows typical SMV and AMR vocoder performance with the SMV codec, used in 3GPP2, which performs better than the AMR codec, used in 3GPP1, albeit with some additional processing and delay overhead not shown on the graph. The G711 reference is a 16-kbps μ-Law PCM waveform encoder used in wireline voice compression and used in this example as a quality benchmark.

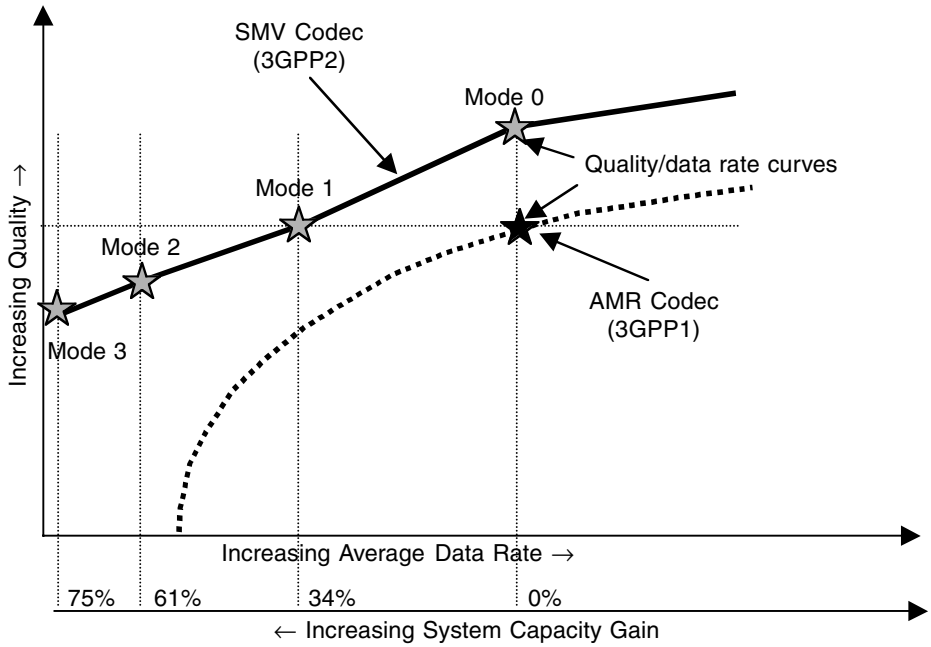


Figure 4.2 Codec performance comparison.

In general, as you would expect, as encoder bit rates increase, voice quality improves. However, more bandwidth will be needed, thereby reducing capacity.

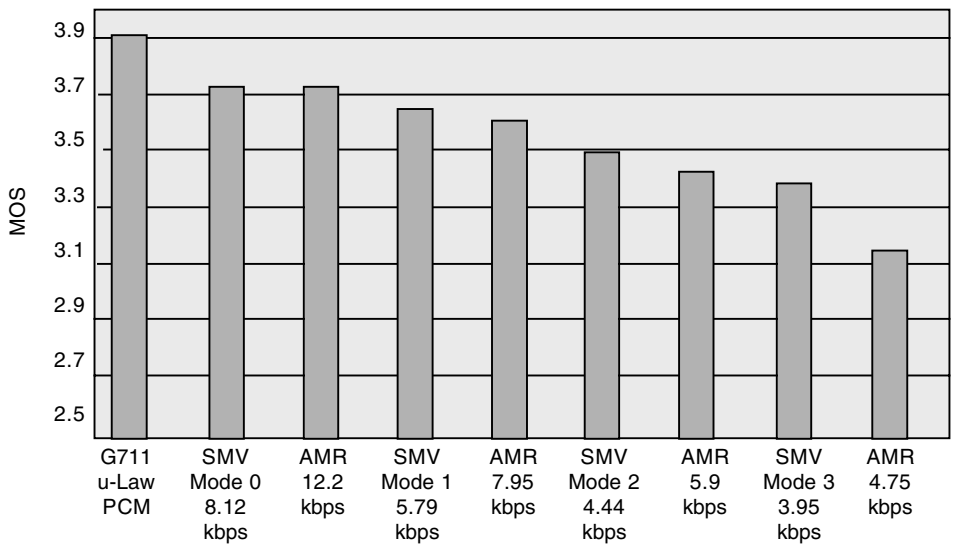


Figure 4.3 SMV and AMR vocoder performance comparison.

Table 4.1 Impact of Audio Processing on Delay Budgets

SEQUENCE	DEVICE	ROUND-TRIP DELAY IN MS
(1)	Handset	
	(a) Encoder delay	33.0
	(b) Encoder processing	10.0
	(c) Channel processing	2.0
(2)	Air Transmission	
	(a) Frame transmission time	20.0
(3)	Base Station	
	(a) Channel processing	2.0
	(b) Viterbi decoding	1.6
	(c) Source decoding	1.0
	Total	69.6

3GPP1 has also specified a wideband version of AMR that encompasses CD-quality audio signals (16 kHz bandwidth versus 3 kHz voice bandwidth). The codec rates are 6.6 kbps, 8.85, 12.65, 15.25; 15.85 kbps, 18.25, 19.85, 23.05, and 23.85 kbps each. This implies an associated need to increase speaker or headset quality in the handset and audio amplifier efficiency.

Parallel work has been undertaken to standardize speech recognition algorithms, with competing candidates from Qualcomm and Motorola/France Telecom/Alcatel. Typical recognition accuracy—that is, user-to-user distance—is >90 percent in a noisy car, five-language test environment. Recognition accuracy is a quality metric.

As we add complexity to audio processing, we increase processing delay, and the delay budget is a not insignificant part of the overall end-to-end delay budget. Table 4.1 details the delay introduced in the send/receive path for each of the audio encoding and encoding processes, including radio transmission framing and channel encoding/decoding. The particular example is a CDMA2000 handset/base station.

CMOS Imaging

Let's now consider image and video quality metrics and the properties of the image bandwidth we are creating by adding CMOS imaging to digital cellular handsets.

Handsets are being designed to integrate digital cameras and high-definition, high-color depth displays. The handset hardware changes the shape and property of the traffic offered to the network, and the RF and baseband performance required is a consequence of the image bandwidth captured by the device.

Table 4.2 CCD vs. CMOS Image Sensors

CCD	CMOS IMAGE SENSORS
More resolution (megapixel images)	Less resolution (100,000 pixels)
Low fixed pattern noise	High fixed pattern noise
Needs multiple voltages	2.8 V or less
Uses more power	Uses less power
Costs more	Costs less

Image bandwidth is determined by the choice of image capture technologies: charge coupled device (CCD) or complementary metal-oxide on silicon (CMOS), as shown in Table 4.2. CCD provides better resolution and more dynamic range (able to work in low-light conditions). Very low fixed pattern noise means a CCD device can take acceptable black-and-white photographs in a completely dark-to-the-human-eye room. The disadvantage of the CCD in digital cellular phones is that it needs multiple voltages, uses more power, and costs more.

CMOS sensors provide less resolution (typically hundreds of thousands of pixels rather than several million). Therefore, megapixel images possible with CCD have relatively high fixed pattern noise and do not work as well in low-light conditions. The advantages of CMOS devices are that they do not need multiple voltages, use less power, and cost less than CCD devices.

To give some present examples, a typical digital camera from Sony using CCD can take 1.6, 2.1, or 3.3 megapixel images and can capture 2500 images on a battery charge. A 12-bit ADC gives wide dynamic range; the display can either be a 4.3 or 3.2 aspect ratio. The camera has a digital zoom, can resize stored images, and has an MPEG-4 movie mode for capturing and e-mailing 60 seconds of moving images (MPEG stands for Moving Pictures Experts Group). An equivalent product from HP provides 2.24 megapixel resolution and 36-bit color depth (to get you to buy that extra-expensive printer!). CMOS image sensors are less ambitiously specified. The most common variants are typically 100,000 pixel devices (352 horizontal and 288 vertical pixels).

An example product from Toshiba can deliver 15 common intermediate format (CIF) frames a second and consumes under 50 mW (five times less than an equivalent CCD product). A DSP is used to double sample the image. Double sampling helps reduce temporal noise and minimize the effects of transistor mismatch. The device uses a 10-bit ADC and runs on a 2.8-V supply. Image lag is reduced by doping the diode transfer switch interface.

Most of the work presently under way on the optimization of CMOS devices is focused on integrating external circuitry to reduce fixed-pattern noise and to improve dynamic range. (A sensor array dynamic range of 68 dB would be a typically achievable figure.) Sensitivity is measured in V/lux/second. A typical achievable sensitivity figure would be 0.52 v/lux/second.

Table 4.3 Color Depth vs. Frame Rate

	FRAME RATES AT 10 BITS VS. 8 BITS PER PIXEL OUTPUT	
NO. OF PIXELS	10 BITS (@ 16 MHZ)	8 BITS (@ 32 MHZ)
1280 × 1024	9.3	18.6
1024 × 768	12.4	24.8
800 × 600	15.9	31.8
640 × 480	19.6	39.2
320 × 240	39.2	78.4

A useful feature of CMOS imaging is the ability to fine-tune resolution, frame rate, and color depth. Table 4.3 shows how this can be resolved into trading-off frame rate against color depth and clock processor speed (power budget). Let's take a 1280 × 1024 pixel image, for example. A fast-moving scene may need an 18 frame per second frame rate. This can be achieved by reducing the color depth from 10 bits to 8 bits and increasing the clock from 16 MHz to 32 MHz. As frame rate increases, our ability to perceive color depth decreases, so effectively, the faster frame rate hides loss of color resolution.

The Keyboard

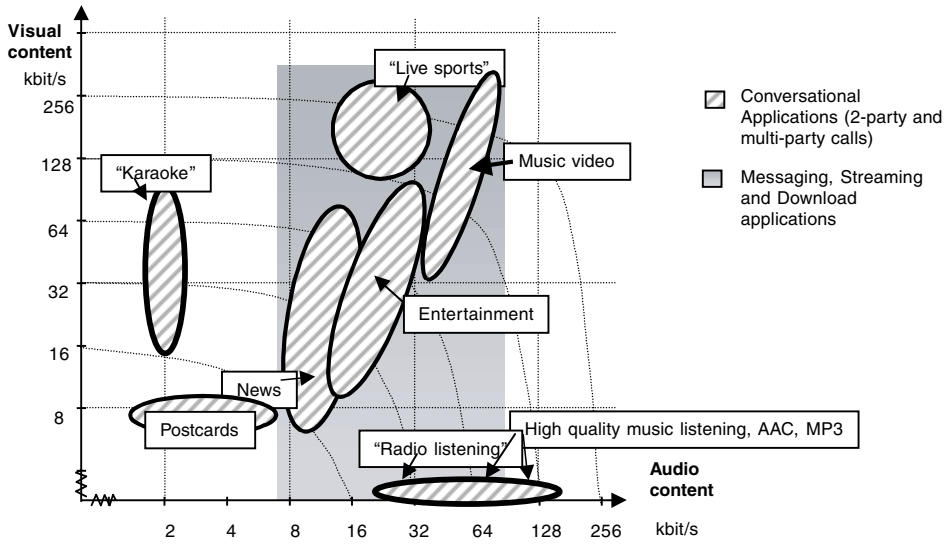
The keyboard is our next component of interest. The choice of application entry is either to use a handset keypad (with or without predictive entry) or use a QWERTY keyboard—either a physical keyboard or virtual keyboard (created on the display). Virtual keyboards require touch-screen displays, which can be quite expensive but are becoming lower-cost over time. Nokia, for example, has a patent on a display that uses capacitive sensing to provide a virtual mouse capability. Conventional keyboards are the most comfortable to use but come with three obvious disadvantages: weight, form factor, and cost.

One option is to make the keyboard a plug-in device—for example, the Ericsson Chatboard—or to have a fold-up device like the Stowaway product for the Palm Pilot. In this instance, weight (224 gm), key pitch (19 mm), and key travel (3 mm) are quality metrics.

A product called Fastap from Digit Wireless increases key density by raising the letter keys and sinking the numbers, making it easier to input Web site addresses.

Rich Media

The microphone, the CMOS imager, and the keyboard are the hardware items used to capture subscriber-generated rich media. *Rich media* is a mix of audio, image, video, and application data. The rich media mix determines the bit rate and bit quality



(Source - Hakan Eriksson presentation at Cannes Wednesday 20th February, 2002)

Figure 4.4 Mobile multimedia services—audio/visual bit rates.

requirements of the offered traffic mix. Figure 4.4, taken from an Ericsson presentation, shows how an image/video bandwidth and audio bandwidth application bandwidth footprint can be realized. Handset hardware determines the offered traffic mix, both on the uplink (dynamic range of the CMOS imager and MPEG-4 encoder) and on the downlink (display and display driver bandwidth).

The Smart Card SIM

The smart card SIM is our next component of interest. As part of the GSM standard in the 1980s, it was decided to incorporate a smart card that would act as a Subscriber Identity Module (SIM)—a mechanism for storing a subscriber's phone number and security information.

The smart card was a French invention and for this reason has seen faster adoption in Europe than the United States . The idea was to take a piece of plastic and put a piece of silicon on it (26 sq mm), on which could be added some memory—an 8-bit microprocessor and a connector. The plastic could either be full IS0 credit-card size (which tended to flex in the early days and later seemed rather large in comparison with handset form factors), a half-size ISO card (which never caught on either), or a plug-in (installed semipermanently), which has become the usual configuration. The market benefit of the SIM was that a subscriber could pick up any handset, add his or her SIM, and be connected to a network.

Smart card SIMs were not initially incorporated into U.S. handsets, although SIMs are now specified by 3GPP2 for use in CDMA2000 (and are known as R-UIM, for Reusable User Identity Modules).

The SIM is now morphing into a new device called a USIM. Depending on whom you talk to and what you read, this stands for a UMTS SIM (Universal Mobile Telephone Standard), a plain and simple Universal SIM, or less often but more appropriately, a User Service Identity Module.

The SIM contains a user-specific encryption key and encryption algorithm, known as the A3/A5/A8 algorithm, which is used to authenticate a user and then to provide encryption using a 58-bit code length across the air interface—that is, over the air. The authentication and encryption algorithms are covered in more detail in Chapter 9, but essentially the A3/A5/A8 algorithm uses a secret key for authentication (k_i) and a secret key for ciphering (k_c). From Chapter 1 you will remember that GSM is based on a frame structure (8 time slots per frame), with the air interface running at 217 frames per second. Above the frame structure sits a multiframe structure, above the multiframe structure sits a superframe structure, and above the superframe structure is a hyperframe that is approximately 3½ hours long. k_c is derived as a product of k_i and the frame number within the 3½ hour cycle that the air interface happens to be at the time the key is established. For all practical purposes this is adequately robust over-the-air encryption.

However, we are now requiring a handset to perform far more functions than just carrying voice. As a result, we need to provide a mechanism for managing access and policy rights, quality of service parameters, service-level entitlements, the particular security context needed for a rich media exchange, and any associated content ownership rights that need to be preserved. If, in addition, the handset is being used to authorize commercial transactions, we need to provide robust, end-to-end authentication and encryption support. *Over the air* means just that—the traffic is secure as far as the network and can then be intercepted by legitimate authorities. *End-to-end encryption* means the traffic remains nontransparent as it moves through the network.

SIM standards have evolved from Phase 1 to Phase 2 to Phase 2+. Table 4.4 shows how the memory requirement has expanded as the standard has evolved.

Typically, available hardware has evolved rather faster than the standard. A typical smart card SIM today has 196 kbytes of ROM, 6 kbytes of RAM, and 68 kbytes of EEPROM, and is now not an 8-bit microcontroller but a 16-bit or even 32-bit controller.

No hardware is totally secure—in the same way that no software is totally secure. Various methods exist to recover RSA keys, including fault injection (subjecting the smart card to ionizing radiation, injecting a single bit error into one of the registers, and comparing the errored and nonerrored outputs) and smart card power analysis (the power drawn by storing a word in a register differs depending on the ratio of 1s and 0s).

Table 4.4 Memory Footprint Evolution

Phase 1 operation	8 kbyte ROM, 251 bytes RAM, 3 kbyte EEPROM, 5 V
Phase 2	16 kbyte ROM, 384 bytes RAM, 8 kbyte EEPROM, 5 V and 3 V operation
Phase 2+	40 kbyte ROM, 1 kbyte RAM, 32 kbyte EEPROM, 5 V and 3 V operation

Figure 4.5 Smart card SIM—example Hitachi.

Figure 4.5 shows a 32-bit smart card EEPROM SIM from Hitachi fabricated in a 0.18 μm process. The device has a 1024-/2112-bit RSA key (rather more robust than the standard GSM A3/A5/A8 58-bit key) to support end-to-end authentication and encryption. The calculations involved in running these keys are nontrivial. This device takes 120 ms to code the 1024-bit key length (which can be a problem for delay-sensitive applications) and is effectively the cost of moving the crypto processor onto the smart card (the benefit is greater security).

In the United States, more mechanically secure hardware packages have been proposed, including *i-buttons*, which are 16-mm computer chips in a steel can. This is an 8-bit microprocessor with 6 kbytes of nonvolatile RAM and a (10 year life) lithium battery. If you try to open the can, all registers are set to zero. The i-button has a 1024-bit key (RSA), which takes just under a second to run, which is fine for non-delay-sensitive applications. (Additional information is available on Dallas Semiconductor's Web site, www.dalsemi.com.)

Other alternatives include fingerprint authentication. A person's fingerprint effectively becomes one of the plates of a capacitor; the other plate is a silicon chip with a sensor grid array. An example product from Veridicom (www.veridicom.com) uses a 300×300 sensor grid array to create a 500 dot per inch image of the ridges and valleys of the fingerprint, which are then processed by an 8-bit ADC to produce a unique digital value. The technology has also been applied in some handsets; for example, a current Sagem dual-band GSM product can recognize up to five fingerprints (www.sagem.com).

Opinions differ as to the long-term security/robustness of fingerprinting as a recognition technique. It is becoming feasible to use modeling techniques to produce artificial fingerprints. Other options exist, such as iris scanning, but most are not particularly practical for present implementation into a digital cellular handset. We are more likely to see a further evolution of the smart card with more memory available. (We cover memory footprints in Part II of this book, which deals with handset software.)

The MPEG-4 Encoder

Now we need to consider the MPEG-4 encoder. MPEG-4 encoders and decoders can be realized in hardware or software. The argument for realizing the encoder/decoder in hardware is to reduce the amount of memory needed, minimize processor delay, and reduce the overall processor power budget. Companies like Emblaze (www.emblaze.com) and Amphion (www.amphion.com) develop specialist ASICs for video processing. The ASICs are optimized to minimize power drain in the handset.

The MPEG-4 encoders use a discrete cosine transform to capture the frequency content of an image that is subdivided into (typically) 16×16 pixel macroblocks. It is the differences from block to block that then get encoded. If two blocks adjacent to each other are both blue—for example, they both show a cloudless blue sky—the description of the second block is effectively the same again. MPEG-4 then adds a frame-to-frame comparison—a process known as differencing or *differential coding* (Chapter 6).

Other Standards

MPEG-4 is not the only standard. Microsoft has its own Windows Media Player. MPEG-4 does have, however, reasonably wide industry support (including support from Microsoft) and builds on earlier work with MPEG-2 and MPEG-3 (audio encoding). One of the problems with these compression standards is that they are optimized to improve storage bandwidth efficiency and are sometimes rather suboptimum when used in variable quality and occasionally discontinuous transmission channels, for example, wireless. MPEG-4 does try to take into account the idiosyncrasies of the radio physical layer. It has also been absorbed into DivX, the PC industry standard for downloadable video, and by Apple in their QuickTime product, so it has some cross-industry adoption.

Figure 4.6 shows the functional diagram for the Amphion MPEG-4 decoder.

The input to the Amphion video decoder core is a compressed MPEG-4 video stream; the data rate is variable from extremely low rates of several kbps up to a maximum defined by the MPEG-4 profile or that possible on the transport stream. For example, the MPEG-4 Simple profile enables up to 384 kbps while the Advanced Simple profile provides up to 8 Mbps (four times the maximum bit rate available from the present W-CDMA physical layer). Higher profiles and bit rates support image scalability, the ability to scale image resolution and frame rate for a given variable channel bandwidth. The MPEG-4 standard supports a wide range of resolutions up to and beyond that of HDTV (high definition television).

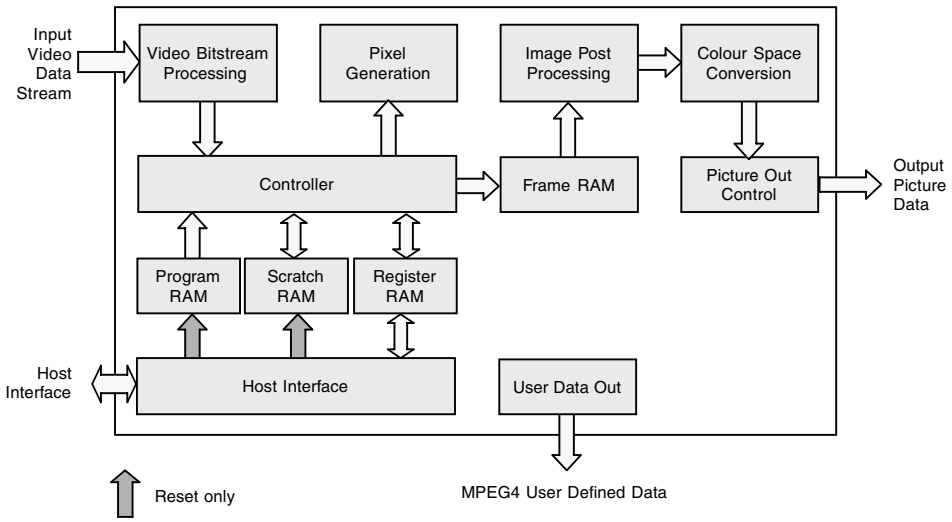


Figure 4.6 MPEG-4 video decoder (Amphion).

AMPHION is a trademark of Amphion Semiconductor Ltd. *http://www.amphion.com.* ARM is a registered trademark of ARM Limited.

The Amphion hybrid architecture of hardware accelerators plus control software on a microcontroller, typically an ARM microprocessor, enables an efficient partition and acceleration of data intensive tasks while maintaining general sequencing and control tasks (such as error resilience) in software. The video bit stream processor extracts variable length symbols from the compressed serial stream, often applying Huffman and run-length decoding for downstream texture decoding and motion compensation. The pixel generation core performs inverse scan, ACDC differential prediction, quantization and discrete cosine transforms on texture coefficients from the video bit-stream processing unit. The image post processor and picture out control does post processing and filtering to take out blockiness and compression artifacts, and then finally color-space conversion and display output timing. Not shown in the functional diagram is the motion compensation accelerator which handles the pixel reference reads, reconstructions and write-backs to frame memory.

Power consumption is much reduced — to below 15 milliWatts — by implementing decompression in a hybrid (i.e., hardware-software) solution because this approach significantly reduces not only the need for processor program and data RAM, but also the overall clock rate required for decoding. Additionally, the main processor can be made available for other tasks such as speech and audio decoding or demux functions. The hardware accelerators can support resolutions and frame rates much higher than any processor-based implementation and thus higher quality video can be supported.

The design challenge for both hardware- and software-based MPEG-4 encoders/decoders is to deliver the functionality needed to support different visual and audio quality metrics: color depth, frame rate, and aspect ratio for imaging, frame rate for video and audio fidelity, which can then be mapped onto a quality-based billing

metric. We discuss quality-based billing in more detail in Chapter 8. Video processing will also support 3D effects (for interactive games), which involves the convergence of MPEG-4 and VRML (the IETF's Virtual Reality Modeling Language). The work groups have taken to describing this as *visual information engineering*. The importance of the MPEG-4 encoder to us is that it effectively defines uplink offered traffic by taking in the imaging and video bandwidth generated by the CMOS or CCD imaging platforms together with other audio and data inputs.

Battery Bandwidth as a Constraint on Uplink Offered Traffic

The other obvious determining factor of uplink offered traffic is battery bandwidth. Battery bandwidth is the ability of a battery to deliver a certain amount of instantaneous RF peak energy, which translates into instantaneous uplink bit rate, and a certain amount of sustained energy, which will determine session length/session persistency. Peak instantaneous RF power is constrained anyway by the regulatory authority and will typically be either a maximum of 250 mW (Class III) or 125 mW (Class IV). We also need to add in the source coding and channel coding. This is not particularly significant for encoding but very significant for decoding, as we will see later in the chapter.

Impact of Hardware Items on Downlink Offered Traffic

Let's move on to look at the hardware aspects of the handset that determine downlink offered traffic. We have already mentioned the MPEG-4 encoder/decoder, but we also need to consider the dynamic range of the speaker, speaker driver, display and display drivers and the way in which these components influence downlink offered traffic properties (and by implication, downlink offered traffic value).

Speaker

First, let's consider the speaker. MPEG-4 encoding (which we cover in a later section) includes higher-rate source coding for enhanced quality audio. Object coding also supports stereo and surround sound. The low-cost speakers used in present handsets may need to be substantially upgraded for future products, or higher-quality headsets, needing higher-quality audio amplifiers, will need to be specified. New loudspeaker technologies provide the basis for additional downlink audio bandwidth. One example is a range of flat speakers from NXT (www.nxt.co.uk). A flat material (cardboard or a translucent material) is actuated across its whole surface to produce a complex audio waveform.

The diaphragm of a conventional loud speaker moves like a rigid piston. NXT's technique is to make a panel vibrate across its whole surface in a complex manner across the entire frequency range of the speaker driver. When applied to a mobile

phone, the technology is called SoundVu, reusing the display so that it can double up as a loudspeaker.

The audio panel is a transparent sheet of an optical polymer material positioned in front of the display screen and separated by a small air gap. The gap varies in size depending on the size of the display (it can be used, for example, for PDA LCD screens, as well as mobile phones). There is some light transmission loss, but the panel can double up to provide electromagnetic interference suppression and an antiglare screen. The sound and image are locked together. They originate from the same point in space and potentially can provide a left-hand channel, center channel, and right-hand channel (surround stereo from your mobile phone).

The power consumption is claimed to be 1/25 of the power consumed by a magnetic speaker, a few milliWatts to support a high-fidelity audio output. The device can also be used to create a touch-screen display. Placing a finger on the screen causes the device to change its vibrational behavior. Using digital signal processing, it is possible to determine the finger's position on the screen within 1 mm. The hardware is already in place (to provide the audio output), so the only additional cost is the processing overhead—the product is known as TouchSound.

At time of writing only proof-of-concept products exist. It does, however, illustrate how changes/developments in handset hardware influence or can influence the offered traffic mix—in this case, using flat panel speaker technology to support wideband audio using the Adaptive Multi Rate Wideband (AMR-W) encoder on the downlink to the handset. The AMR-W codec presently decodes 16 kHz of bandwidth, but consider how user expectations increase over time. We can buy audio products with audio response well above the limits of our hearing (super tweeters with response up to 50 kHz).

Evolution in hardware capability effectively determines the software requirements in the handset. As speaker technology improves, there is a parallel need for increasingly sophisticated audio management. A working group within the Internet Engineering Task Force (IETF) is presently working on Extensible Music Format (XMF) to provide a consistent, standardized way of producing enhanced polyphonic ring tones and game sounds. The software processors are sometimes described as audio engines (www.beatnik.com provides some examples).

Now let's consider downlink image and video bandwidth.

Display Driver and Display

We described earlier how color depth was related to the number of bits used to identify the pels in a pixel (RGB discrimination). The dynamic range of the display driver and display determines what can be shown on the handset—and hence determines the properties of the downlink offered traffic.

Table 4.5 shows the progression from 1-bit to 24-bit color depth. Anything beyond 24-bit color depth is generally not discernible by the human eye, though as with high-range audio products, this doesn't mean people will not buy such products; in fact, some video adapters and image scanners now deliver 32-bit true color. 3GPP has specified a core visual profile that covers from 4 bits (grayscale) to 12 bits, but as we will now see, display capability is rapidly moving toward 16-bit color depth.

Table 4.5 3GPP-3G-324 M

COLOR DEPTH	NUMBER OF POSSIBLE COLORS
1	2 (Black and white)
2	4 (Grayscale)
4	16 (Grayscale)
8	256
16	65,536
24	16,777,216

The most favored candidate for digital cellular handsets to date are conventional but highly optimized LCDs. These come in two flavors: reflective, which work well in bright sunlight, and transmissive, which work well indoors. Products like the Compaq iPAQ use reflective displays to meet the power budget constraint of a PDA that ideally should be capable of running on two AA batteries.

Transmissive LCDs are the standard for laptop PCs. Laptops have a power budget of between 8 and 10 Watts—along with a form factor and rechargeable battery to suit. Digital cellular phones need to be well under a Watt to meet form factor requirements, given existing battery densities. This means they are much closer to PDAs than laptops in terms of power budget constrains. The Nokia 9210 provides a good benchmark for a 2002 product against which future generations of display-enabled handsets can be measured. The device supports 4000 colors.

The more colors a screen can support, the more light filters you need. The more light filters you have, the bigger the backlight. The bigger the backlight, the more power you consume.

The transmission display in the 9210 uses a cold cathode fluorescent lamp. It is positioned right next to the battery and couples light to the display via a wedge-shaped slab waveguide. The wider the prism (that is, the thicker the wedge), the better the coupling efficiency and brightness uniformity of the display and the lower the power consumption.

In this example, the waveguide wedge is 6 mm, which seems to be at present an acceptable thickness/efficiency trade-off. The display can automatically adapt to ambient light conditions. Flat out, the screen emits 100 candelas and consumes 500 mW. At the dimmest setting, it consumes 150 mW. This is, of course, in addition to the existing baseband and RF power budget. Quoted figures from the manufacturer suggest between 4 and 10 hours of use from a fully charged 1300 mAh lithium ion battery.

The resolution achievable is a function not so much dictated by the screen itself but by the driver IC connections. The color screen is 110×35 mm, with a pixel density giving 150 dots per inch (dpi) of resolution at a pixel pitch of 170 μm. The response/refresh cycle of the driver is 50 ms, which is sufficient for a frame rate of 12 frames per second. There is no point in sending such a device a 20 frame per second video stream, as it will be incapable of displaying it. The hardware bandwidth determines offered traffic bandwidth and offered traffic properties.

Table 4.6 Current Hitachi Displays

DISPLAY SIZE	128 × 176 pixels	132 × 176 pixels	128 × 160 pixels
NUMBER OF COLORS	256	4096	65,536
NUMBER OF BITS COLOR DEPTH	8	12	16
DISPLAY RAM CAPACITY (KBPS)	160	278	372
WRITE CYCLE AT 2.4 V	100 ns	100 ns	100 ns
WRITE CYCLE AT 1.8 V	200 ns	200 ns	200 ns

In practice, the hardware in this area is moving rather faster than the software, but it is nice to know that in the future, displays and display driver bandwidth will be capable of supporting increasingly high-resolution, high-color depth displays sent at an increasingly rapid frame rate. Table 4.6 gives some examples of present displays available from Hitachi.

One rather unforeseen consequence of improving display quality and display driver bandwidth is that as display quality improves, compression artifacts become more noticeable; the quality of the display and display driver determines the quality needed in the source coding and physical layer transport. Put another way, if you have a poor-quality display, you do not notice many of the impairments introduced by source coding (compression), channel coding, and the highly variable-quality, occasionally discontinuous radio physical layer.

While color saturation/color depth is reasonably easy to achieve with backlit displays, it is significantly more difficult with reflective (sometimes as described as transflective) LCDs. In an LCD, a single color filter covers each pixel. Transmissive backlit displays use thick filters. Reflective displays use thin filters to allow the light to pass into and back out through the filter. The thinner the filter, the better the reflective properties but the poorer the color saturation. If the thickness of the filter is increased to improve color saturation, the picture becomes too dark.

A reflective LCD from Philips makes one corner of the pixel filter thinner than the rest, which means that light can pass easily, thereby increasing brightness. The rest of the filter is optimized for color saturation.

So here we have another quality metric—brightness—that is directly related to how the display hardware is realized. Additional metrics include uniformity and viewing angle (usually quite narrow with LCDs). One problem with conventional displays is the continued use of glass. Glass is relatively heavy, fragile and does not bend easily. A hybrid approach presently being investigated involves the use of ultra thin glass attached to a flexible sheet. Toshiba has recently shown examples of products that, in the longer term, could provide the basis for foldable lightweight LCDs.

All displays, including flexible, foldable, and conventional displays, require display drivers. The Digital Display Working Group (www.ddwg.org) is presently working to standardize the digital display interface between a computer and its display device, including backward-compatibility with existing analog driver standards. This working group is also producing proposals for micro-displays (50-mm/2-inch diagonal size).

Consider that an SVGA LCD monitor needs to have an address bandwidth/bit rate of 25 megapixels per second (25 million pixels per second). A QXGA cathode-ray tube has an effective bandwidth requirement of 350 megapixels per second.

The refresh rate can be reduced by only refreshing the parts of the display that are changing. This decreases the processor overhead in the driver but increases the memory space needed. Even so, it is not uncommon in PC monitor drivers to encounter driver clock speeds well over 100 MHz. These are power-hungry and potentially noisy devices. Refresh rates in laptop LCDs are now typically 25 ms (the Nokia handset case studied earlier in the chapter had a 50-ms refresh rate). Refresh rate obviously becomes increasingly critical as frame rate increases.

A number of Japanese vendors are sampling display products (with chip-on-glass display drivers) that are supposed to be capable of supporting 30 frames per second. A present example is a Sharp 262,000-color 5-cm reflective display produced on a 0.5-mm substrate, which is claimed to support 30 frames per second at a power consumption of 5 mW per frame—small but efficient.

For backlit (transmissive) displays, performance gains include significant improvements in contrast ratio and parallel reductions in power budget.

Pixel density is moving to more than 200 pixels per inch and contrast ratios are improving from 50:1 to 200:1 or better (see Table 4.7). Power savings are being achieved by using thin film transistors with latch circuits that hold the liquid crystal cell state at the correct potential through the refresh cycle. Fortuitously, investment in LCD-based micro-display technologies can be common both to digital cameras and 3G handsets with digital cameras.

In effect, these are two related but separate product sectors, each of which generate significant market volume. Market volume helps reduce component cost but also tends to improve component performance through better control of component tolerances on the production line. Digital camera performance drives user expectation of how a digital cellular handset with an integrated digital camera will perform. The problem is that the digital cellular handset also has to be able to send and receive pictures and an audio stream over a radio physical layer that will typically consume several hundred milliWatts. There is a balance to be made between memory bandwidth in the handset and how much power to dedicate to sending and receiving image bandwidth, which in turn determines the user experience and user expectations.

Table 4.7 Liquid Crystal Displays—Contrast Ratios

	PRESENT GENERATION	NEXT GENERATION
Contrast Ratio	50:1	200:1
Power	1.2 W	200 mW

There are some other practical issues. Cellular handsets tend to be much more roughly handled than computer products—for example, PDAs. All displays that use glass are inherently fragile and don't take kindly to being dropped onto concrete floors. A very important present design consideration is how to improve the robustness of high-quality displays. Using thin layers of glass bonded to plastic is one option.

How User Quality Expectations Increase Over Time

The Optoelectronics Industry Development Association (www.oida.org) and the Video Electronics Standards Association (www.vesa.org) help to establish standards for pixel density/pixel spacing (resolution), refresh frequency, color depth, brightness, contrast ratio, duty cycle (for example, phosphor degradation in phosphor displays), and power budgets. Quality expectations are influenced by the other display devices that we use each day. Looking back over time, an IBM VGA monitor in 1987 provided a 640 × 480 pixel display with 16 colors. By 1990, XGA monitors were typically 800 × 600 pixel with 16 million colors.

Table 4.8 shows how computer monitor resolution standards are evolving—partly because technology makes the evolution possible (and gives marketing people something new to sell) and partly because high resolution opens up new applications, for example, medical images using Quad Extended Graphics Array (QXGA) resolution. QXGA images are either 3 or 5 megapixels. A 5-megapixel image with a 24-bit color depth produces an image bandwidth of 120 million bits. You would not want to send too many of these to or from a digital cellular handset!

Table 4.9 shows typical LCD screen size and resolution options for laptop PCs. A 21-inch XGA monitor needs 9 million transistors—it is not surprising that LCDs constitute about a third of the component cost of a laptop PC! The smaller the screen, the fewer pixels you need to achieve the same resolution as a bigger screen. Small screens, however, seem bigger if they have higher resolution. It's not just the number of pixels but rather the pixel density that's important.

Table 4.8 Resolution Standards

DESCRIPTION	NUMBER OF PIXELS
VGA	640 × 480
SVGA	800 × 600
XGA	1024 × 768
SXGA	1280 × 1024
UXGA	1600 × 1200
HDTV	1920 × 1080
QXGA	2048 × 1536 = 3 megapixels (Medical imaging > 5 megapixels)

Source: VESA—Video Electronics Standards Association (www.vesa.org)

Table 4.9 LCD Screen Size and Resolution

DESCRIPTION	SCREEN SIZE IN INCHES	NUMBER OF PIXELS
VGA	4*	640 × 480 pixels
SVGA	8.4*	800 × 600 pixels
XGA	10.4*	1024 × 768 pixels
UXGA	15*	1600 × 1200 pixels
XGA	21†	2048 × 1536 pixel

*Toshiba
†IBM

The other factor determining user expectations is digital TV. Present digital TV offerings in the United States, Europe, and Asia do not provide recognizable improvements in terms of image quality (actually because analog TV quality, certainly in Europe and Asia, is already very good). In the longer term, high-definition television will increase user quality expectations.

Digital TV does, however, have an impact on our perception of aspect ratio. A square screen has an aspect ratio of 1:1, standard television sets are 4:3, and widescreen digital TV is 16:9. The 16:9 ratio is chosen to match the typical aspect ratio of human vision, which is supposed to be more comfortable and satisfying to look at.

Aspect ratio has a particular impact on pixel processing overhead—1:1 aspect ratio screen use square pixels. On a 4:3 or 16:9 screen you have to use rectangular pixels (the image pixel is wider than it is taller). The screen pixel information has to be distorted to correct for this.

Alternative Display Technologies

A number of alternatives to the LCD are presently being propositioned and promoted. Polymer LCD displays are one option. Certain polymers will conduct electricity and emit light. These are called Light-Emitting Polymers (LEPs). The example shown in Figure 4.7 was developed initially for use as a backlight in a handset, produced on a glass substrate. The longer-term evolution includes plastic substrates and red/green/blue polymers.

Reverse emulsion electrophoretic displays are a second option. These displays consist of two glass plates. A color reversed emulsion is injected between the two glass plates, which are held together like a sandwich. The emulsion consists of a polar solvent (a liquid with a property like water), a non-polar solvent (a liquid with a property like oil), one or more surfactants (detergents), and a dye, which is soluble in the polar solvent and insoluble in the non-polar solvent.

Figure 4.7 Cambridge Display Technology LEP display.

The result is a lot of colored droplets floating in a clear liquid. The droplets can be electrically charged and made to spread out, which produces color on the display, or compacted, which makes them transparent. The properties of the emulsion change with frequency; in other words, the emulsion is frequency or electrophoretically addressed to provide a dynamic color display—a sort of high-tech lava lamp. (A case study can be found at www.zikon.com.)

Organic electroluminescent (OEL) displays are a third option. OEL displays use thin layers of carbon-based (organic) elements that emit light with current passing through them (electroluminescence). The advantage of OELs is that they do not need a backlight, because they are by nature luminescent. This saves power. They also have a good (160°) viewing angle and can be mechanically compact, because the display driver can be integrated into the substrate. Pixel density is also potentially quite good; Kodak, for instance, claims to be able to get 190,000 pixels individually addressed into a 2.5-inch diagonal space.

Another option might be to use miniaturized—that is, thin—cathode ray tubes, as shown in Figure 4.8. In a thin CRT, an array of microscopic cathodes is deposited on a baseplate using thin film processing. Each cathode array produces electron beamlets that excite opposing phosphor dots (i.e., a cold cathode process producing electrons at room temperature). The process does not need a shadow mask, which means it is relatively power-efficient, and it uses high-voltage phosphors, which give 24-bit color resolution of 16 million colors and high luminance.

The CRT would provide a wide-angle view, which is a major performance limitation with LCDs. It would also produce a 5 ms response time, compared with 25 ms for a typical LCD, and would consume about 3.5 Watts driving a 14-inch display. Whether the technology could scale down to a micro-display is presently unproven, but it has possible future potential, provided the mechanical issues, such as a very high vacuum gap, can be addressed. Additional information is available at www.candescent.com.

Philips has also proposed 3D displays as a future option. These displays use a lenticular lens over each pixel segment, which means specific pixels are only visible from specific angles. Given a reasonably complex driver, a 3D effect can be created.

Although electrophorescent, OEL displays and miniaturized CRTs all offer interesting longer-term options; LCDs (3D or otherwise) are presently preferred, particularly for smaller screen displays, where it is proving relatively easy to deliver good resolution, fast refresh rates, good contrast ratios, an acceptable power budget, and tolerable cost. Displays and display drivers are generally not the limiting factor in delivering end-to-end quality, provided cost targets can be achieved.

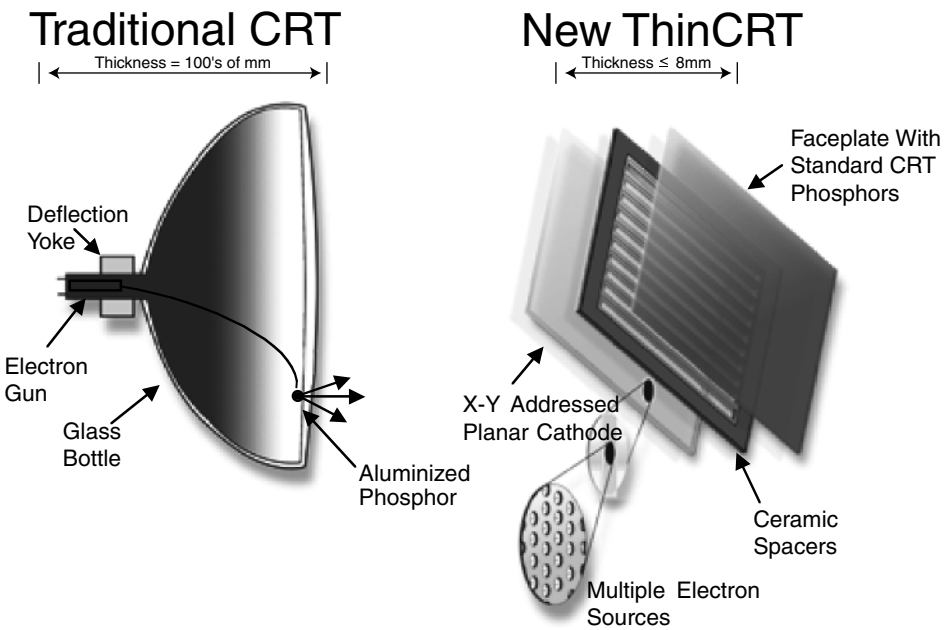


Figure 4.8 Thin CRTs.

MPEG-4 Decoders

We have already covered MPEG-4 briefly earlier in this chapter when we compared hardware and software realization options. Let's revisit the topic, this time focusing specifically on power budgets.

In Chapter 1 we described how, in common with speech codecs, image and video codecs perform a discrete cosine transform to capture the frequency coefficients of the quantized waveform. The DCT/quantizer and the way the information is presented and multiplexed out of the encoder is prescribed by the MPEG-4 standard. (DCT stands for Discrete Cosine Transform.) Other encoding tasks are left to the codec designer, which means they can be optimized without reference to the standard.

Figure 4.9 shows a block diagram of a low-power MPEG-4 encoder/decoder from Toshiba. It features a display interface, camera interface, multiplexer (to manage multiple simultaneous image, video, and data streams prior to multiplexing onto single or multiple code streams on the radio channel), and a video codec hardware block with a Reduced Instruction Set Computing (RISC) processor, Direct Memory Access (DMA) hardware (for optimizing memory fetch routing), and an audio encoder.

The device will simultaneously encode/decode a QCIF (Quarter Common Intermediate Format) video stream at 15 frames per second. It uses Toshiba's variable-threshold CMOS on a 0.25 μm chip with integrated 16-Mbit DRAM and three 16-bit RISC processors. The whole device runs at 60 MHz and consumes 240 mW.

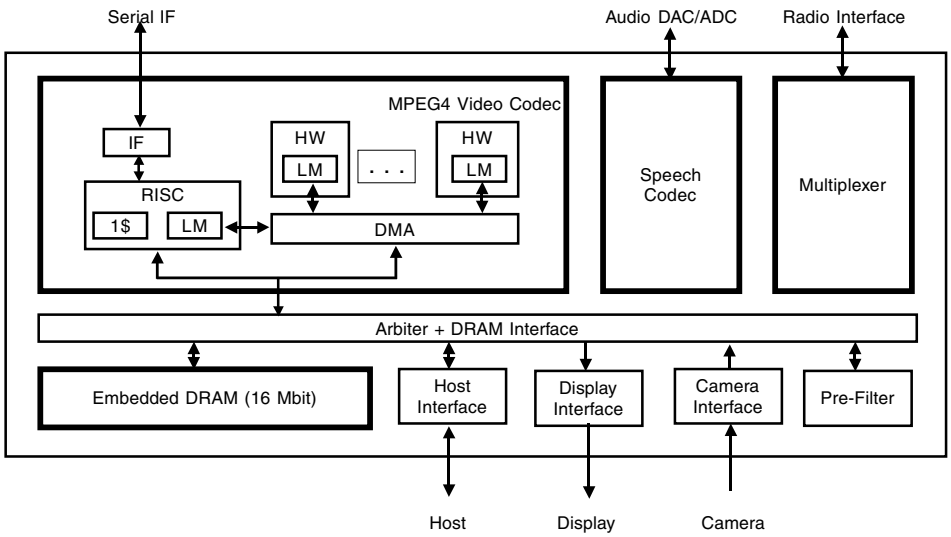


Figure 4.9 Low-power MPEG-4 encoder/decoder.

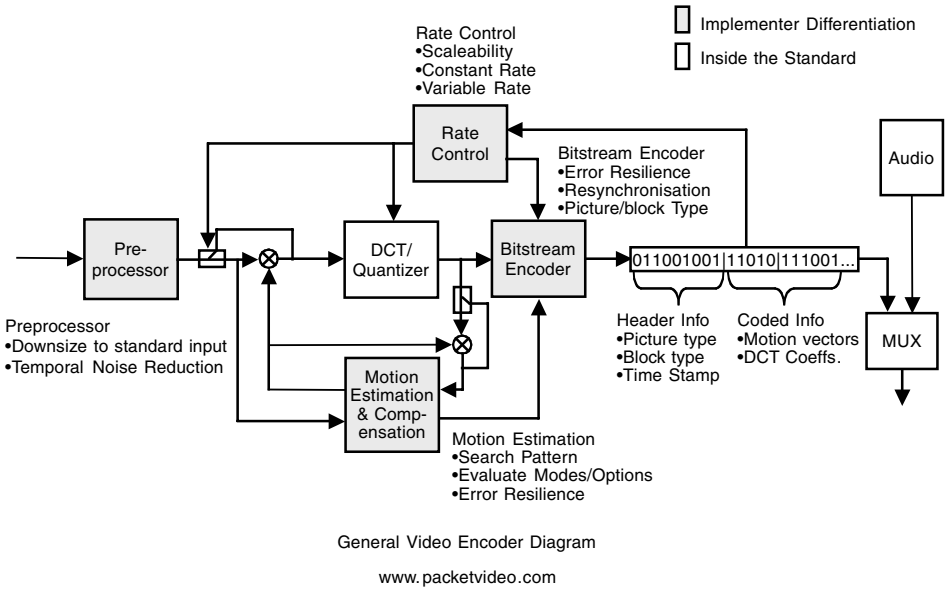


Figure 4.10 Video encoding (MPEG-4).

Figure 4.10 shows an MPEG-4 encoder from PacketVideo (www.packetvideo.com). The blocks in white show the MPEG-4-compliant DCT/quantizer and the way the bit stream is delivered with a packet header, the picture type (for example, QCIF), the type of block coding used, the timestamp, and the data itself (the frequency coefficients and motion vectors). Other white blocks are the multiplexer and audio codec. Blocks in gray show the preprocessor where temporal noise reduction is done, motion estimation and compensation, rate control (fixed rate or variable rate), and the bit stream encoder, where vendor-specific error protection encoding is added.

On the receive path the audio decoder, demultiplexer, depacketizer, motion compensation, and DCT/quantizer are all MPEG-4-compliant (see Figure 4.11). The channel decoding (bit stream decoder) and post processor are vendor-specific. The implications of this are that codec performance will vary between vendors and will depend on how much pre- and post-processing is done, which will determine processor overhead and codec power consumption. It remains to be seen how well codecs from different manufacturers will work together.

At present, the MPEG-4 profile supports 4- to 12-bit color depth, but longer-term profiles will be extended to include 24- or possibly 32-bit color depth, which will need to be accommodated by the encoder. The profiles also describe picture size; CIF (Common Intermediate Format) and the previously mentioned QCIF (Quarter Size Common Intermediate Format) are called "common" because they can be scaled down from National Television Standards Committee (NTSC) and Phase Alternating Line (PAL) images. As screen size reduces, resolution increases. High-resolution small screens often look bigger than lower-resolution bigger screens.

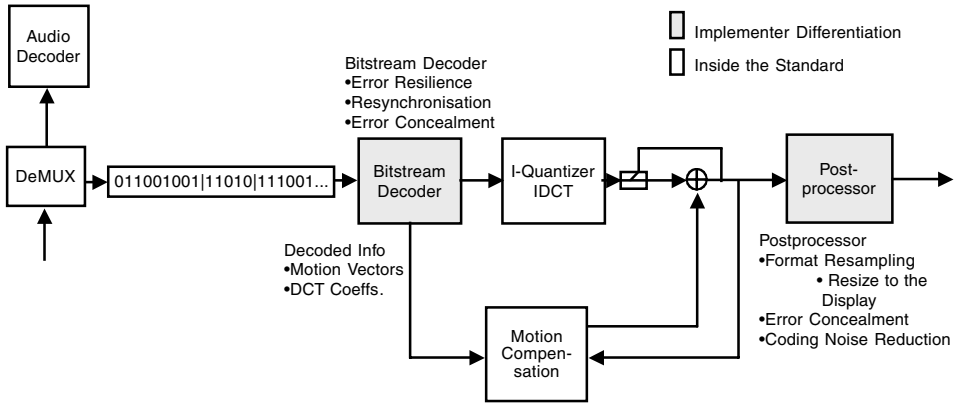


Figure 4.11 Video decoding.

Handset Power Budget

So far in this chapter we have described a number of hardware items that are being added to the handset. These hardware items each consume significant amounts of power. The objective with 3G handset design is to keep the overall power budget equal to, similar to, or, preferably, lower than 2G handset power budgets.

Consider a typical power budget for a GSM phone using the StarCore (Motorola/Lucent) core processor. The example in Table 4.10 is for a traditional superhet with baseband, IF, and RF stages.

In practice, if the baseband can be supported on a 0.9-V supply, the call state power drain can be reduced to less than 40 mW, rather than the 110 mW stated in the table. Also, as network density increases, the RF PA can be run at lower power levels (though sometimes with some decrease in efficiency).

Table 4.10 Typical GSM Power Budget (Motorola/Lucent StarCore)

PARTITIONING	POWER CONSUMED	
Baseband	110 mW in call state (6 mW in idle mode)	
RF/IF	190 mW transmit	134 mW receive
RF PA	400 mW	

Table 4.11 3G Handset Power Budget

DEVICE	POWER BUDGET (MILLIWATTS)
Standard handset	400 mW
Image sensor	+200 mW
MPEG-4 encoder/decoder	+240 mW at 60 MHz
LCD	+200 mW
TOTAL	**1040 mW = 2.4 HOURS**

Assuming the overall power budget can be reduced to 400 mW (RF/IF and baseband receive and transmit on a typical duty cycle), a 700 milliamp hour battery at 3.6 W delivers 2.4 Watt-hours of energy and will support 6.25 hours of use.

Table 4.11 shows how the power budget increases as you add in an image sensor, MPEG-4 encoder/decoder, and LCD. (A transmissive backlit LCD will consume rather more than 200 milliWatts; a passive reflective display rather less.)

As a rule of thumb, you can say that adding multimedia/rich media functionality to a handset easily doubles the power consumption—even before you take into account the additional RF power budget needed to send and receive all the additional image bandwidth created by the device. We have reduced 6.5 hours of use to 2.4 hours if the same capacity battery is used. If we opt for a larger-capacity battery, it may be bigger and weigh more or need to use a more exotic battery chemistry. Either way it will cost more.

In later chapters we also describe how it is the job of handset and network software to increase session complexity and session persistency. Session complexity involves supporting multiple users each with multiple code streams. Because the session has continuous activity, the duty cycle will be 100 percent rather than the 35 percent more typical with existing voice exchanges, though the amplitude (bit rate) of the exchange will be continuously changing as the session progresses.

Both the RF and processor power budgets will need to comprehend this continuous duty cycle. Batteries will also have to be capable of supplying significant peak loads (instantaneous bandwidth) and a high peak to average peak-to-mean ratio. The peak-to-mean ratio will be significantly greater than present GSM handsets. This is a problem for RF PA designers, since it is difficult to get an RF PA to run efficiently when it is lightly loaded. In addition, as session persistency increases, the overall capacity of the battery will need to increase.

Processor Cost and Processor Efficiency

Tables 4.10 and 4.11 show how the processor power budget has increased from 25 percent of the overall power budget to 60 percent as we add image processing and high

bandwidth displays and display drivers. This places substantial focus on processor performance in terms of cost and power efficiency.

The general expectation in the mid to late 1990s was that a 3G phone would need about three times the processor capacity of a 2G phone (300 rather than 100 MIPS). In practice, first-iteration, third-generation handsets are absorbing between 800 and 1000 MIPS; that is, costs and power budgets are rather higher than planned. This is really the consequence of user expectations moving on. The introduction of products like the Palm Personal Digital Assistant (PDA) resulted in people expecting to have handwriting recognition in portable products. Automotive guidance systems have resulted in people expecting to have speech recognition in products. We expect to have simultaneous voice *and* data, not either one or the other, and we expect video quality to be as good as our home DVD system.

In addition, we are trying to offload many of the tasks previously done in the analog domain down to baseband—that is, using DSPs to offload linearity problems and to deliver low-cost filtering and waveform shaping.

In effect we are saying that a multimedia handset incurs substantial physical layer and application layer processor overhead and substantial memory overhead. This gives us a design challenge in terms of cost, product form factor and power budget.

Future Battery Technologies

Adding to the power budget means we need to add in additional battery capacity, in turn adding size, weight, and cost to our 3G handset. We have mentioned battery technologies twice already in this chapter—once in the context of uplink bandwidth and once in the context of needing to meet the peak energy requirements implicit in bursty bandwidth.

Table 4.12 shows a comparison of battery technologies in terms of Wh/kg and Wh/liter. The best-performing batteries at present are based on lithium. Lithium polymer batteries provide a reasonable energy density (70 to 100 Wh/kg) but have relatively high self-discharge rates (they go flat without being used). They also lose about 20 percent of their rated capacity after about 1000 cycles.

Lithium ion batteries using a liquid electrolyte to deliver better energy density (120 Wh/kg) but also have a relatively high self-discharge rate. Lithium metal batteries, using a manganese compound, deliver about 140 Wh/kg and a low self-discharge rate: about 2 percent per month compared to 8 percent per month for lithium ion and 20 percent per month for lithium polymer.

Table 4.12 Battery Density Comparisons

	NI-CADS	NIMH	LITHIUM ION	ZINC AIR	LITHIUM THIN FILM
Wh/kg	60	90	120	210	600
Wh/liter	175	235	280	210	1800

Lithium thin film promises very high energy density by volume (1800 Wh/liter). However, delivering good-through-life performance remains a nontrivial design task. Also, these very high density batteries have high internal resistance; which means they like to hold on to their power. Given that we are trying to design adaptive bandwidth-on-demand handsets that may be delivering 15 kbps in one 10-ms frame and 960 kbps in the next frame, then obviously we need a battery that can support bursty energy needs.

Methanol cells may be a future alternative. These are miniature fuel cells that use methanol and oxygen with (usually) platinum as a catalyst. Fuel cells can potentially deliver better than 100 percent efficiency, since they pull heat from the atmosphere (an answer to global warming!). Even the best diesel engines struggle to get to 30 percent efficiency, so methanol cells with an energy density of 3000 Wh/kg would seem to be a promising way forward.

Motorola has a prototype direct methanol fuel cell (DMFC) that has a membrane electrode assembly, a fuel and air processing and delivery system, a methanol concentration sensor, and a liquid gas separator to manage the release of carbon dioxide. The prototype measures $5 \times 10 \times 1$ cm excluding the control electronics and fuel reservoir. At the moment, the device can produce 100 mW of continuous net power, so there is some way to go before we have a methanol-powered multimedia mobile.

Potentially, however, energy densities of over 900 Watt-hours per kilogram are achievable. A 20-gram battery would be capable of producing 18 Watt-hours of power.

An example is a microfuel cell from Manhattan Scientifics. The device can be produced in kilometer long, thin printed sheets rather like a printed circuit—the main challenge is in the microminiaturization of the air distribution system and the internal plumbing to mix the hydrogen and air sufficiently well to make the device efficient. Manhattan Scientifics claim an energy density of 80 mW/cm^2, equivalent to 940 Wh/kg for the device. NEC has similar products presently in development.

Whether we are talking about conventional batteries or fuel cells, there are essentially two considerations: capacity and the ability to provide power on demand. 3G handsets are either specified for a maximum power output of 250 mW (Class 3) or 125 mW (Class 4), and this determines the instantaneous uplink bandwidth available. The battery has to be capable of meeting this instantaneous demand for power, given the voltage being used in the handset—typically 3 Volts or, in the longer-term, 1 Volt.

Second, the overall capacity of the battery determines overall uplink offered traffic bandwidth from each individual user. A 600 milliamp/hour battery will not be sufficient for uploading video content and also determines downlink processor capacity.

Either lithium or, in the longer term, fuel cell batteries will remain as a key enabling technology in 3G handset and network implementation; battery bandwidth intrinsically determines uplink and downlink network bandwidth and bandwidth value.

Handset Hardware Evolution

3G handset hardware allows us to capture voice (audio bandwidth), image and video bandwidth, and application bandwidth. These are typically multiplexed into multiple traffic streams that may be separately modulated onto multiple OVSF code streams over the physical layer (radio air interface). The choice of image processor (CCD or CMOS) dictates the dynamic range of the image or video stream and other qualities

such as color depth and resolution. In addition, the accurate representation of fast-moving action requires a reasonably fast frame rate. Therefore, the hardware dictates our bandwidth quantity and quality requirements.

3G handset hardware generates bursty bandwidth. Voice, image, and video encoders have historically been *constant rate* encoding devices but are increasingly moving to become variable rate to accommodate the varying amount of entropy and redundancy in the source-coded information. In addition, voice, image, video, and data is being multiplexed together in the encoder. Intuitively you might think this would help to average out some of the burstiness. In practice, peaks of information energy still need to be accommodated. These peaks can either be dealt with by allocating additional bandwidth (bandwidth on demand) or by buffering to smooth out the bit rate. Bursty bandwidth can always be turned into constant rate bandwidth by buffering. The cost is the additional memory needed, delay, and delay variability.

As we will see in later chapters, conversational rich media exchanges are relatively intolerant to delay and delay variability. The best option from an application point of view is to make the delivery bandwidth dynamically responsive to the application bandwidth required, remembering that delivery bandwidth is a summation of radio bandwidth and network bandwidth.

Our handset hardware has captured and described the time domain and frequency domain components of our speech, image, and video waveforms. It is the job of the physical layer to preserve these time domain and frequency domain properties. The physical layer includes the radio link and the network, as well as another radio link to the other side of the network if we are talking about handset-to-handset communication. On the receive side, it is the job of the hardware components to rebuild and reconstruct the original signal (composed of audio, image, and video waveforms), preferably without noticeable loss of quality.

Loss of quality can be caused by a poor, inconsistent radio channel, a badly designed receiver, a poorly designed decoder, or, for image and video, display and display driver constraints. Bandwidth quality in this context is an end-to-end concept that encompasses every hardware component involved in the duplex simultaneous process of send and receive. As a result, the quality of our MPEG-4 encoder/decoder has a direct impact on perceived image and video quality, and the quality of our voice codec (encoder/decoder) has a direct impact on perceived voice quality.

With image and video, there is not much point in transmitting a 24-bit color depth, 30 frame per second video stream if we only have a display capable of supporting 12 bits $\times$ 12 frames per second. Bandwidth quality, therefore, becomes a balancing act. Where do we put our processing power?

Adding MIPS to a voice codec improves quality and reduces the bit rate needed from the radio channel but increases codec processor overhead—and introduces delay. The same principle applies even more so to image and video encoders/decoders. We could transmit a video stream at 12 frames a second and use rendering and interpolation in the decoder to double the frame rate—a perceived quality improvement traded against an increase in processor power.

Bandwidth quality comes with a cost and power budget price tag. As processor costs and processor power budgets improve, quality improves. However, we also need to deliver consistency. A poor-quality channel that is consistent may often be *perceived* as being better quality than a better-quality channel that is inconsistent. (We remember the bad bits.)

The idea of the variable-rate encoder is to deliver constant-quality source coded voice, image, and video (the coding rate changes, not the quality). The idea of having adaptive radio bandwidth that codes out the fast fading on the channel is to deliver constant quality.

Adaptive Radio Bandwidth

At this point, it is worth summarizing what we mean by adaptive bandwidth or, more specifically, *adaptive radio bandwidth*. We cover adaptive network bandwidth later in this book. There are five stages at which we can influence bit rate and bit quality—and hence application quality—are as follows:

- We can change the source coding rate and use processor overhead to pre-process images and video content to make the content more robust and resilient to channel errors. The source coding can be adaptive—responding to the dynamic range of the information stream.
- We can adaptively change the channel coding that we add to the source coded bit stream. For example, we can increase or decrease the interleaving depth, we can choose half rate (2/1) or third rate (3/2) convolutional encoding—two bits out for one bit in, or three bits out for every two bits in—or we can use turbo coding.
- We can change modulation, going from GMSK to 8 PSK (in GSM EDGE) or from QPSK to 8 PSK to 16 level QAM in CDMA2000/1XEV.
- We can provide adaptive bandwidth on demand by using CDMA multiplexing (moving up or down, left or right on the OVSF code tree, or adding or subtracting additional OVSF code streams).
- We can make our RF bandwidth adaptive by varying the power allocation to each user or to each user's channel stream/channel streams.

Even analog (1G) cellular handsets had adaptive bandwidth, in that fairly simple power control was supported together with DTX (discontinuous reception). When you didn't speak, the RF power dropped out. In 2G, DTX is also available and used for voice. For data, variable power is delivered by adding additional slots in addition to the existing power control.

3G effectively brings together adaptive source coding, adaptive channel coding, adaptive modulation, and adaptive multiplexing—in all of which, the RF channel spacing stays constant:

- 25 or 30 kHz for first-generation cellular
- 30 kHz, 200 kHz, or 1.25 MHz for second-generation cellular
- 1.25 or 5 MHz for third-generation cellular

Table 4.13 Coding, Modulation, and Multiplexing in 1G/2G/3G/4G Cellular Networks

GENERATION	SOURCE CODING	CHANNEL CODING	MODULATION	MULTIPLEXING	RF BANDWIDTH
1G cellular	Analog (variable adaptive rate)	None	FM (adaptive)	FDMA	25 kHz
2G cellular	Digital (constant rate)	Block and convolutional coding	GMSK	TDMA	30/ 200 kHz
			QPSK	CDMA	1.25 MHz
3G cellular	Digital (variable adaptive rate)	Adaptive convolutional and turbo coding	GMSK 8 PSK QPSK QAM	CDMA	1.25 MHz 5 MHz
4G cellular	Digital (variable adaptive rate)	Adaptive convolutional, turbo and trellis coding	QPSK and higher-level QAM	CDMA	OFDM (adaptive) 2 MHz/8 MHz 2k or 8k carriers

In 4G cellular, we may also adaptively change the occupied RF bandwidth, as shown in Table 4.13. If we use Orthogonal Frequency-Division Multiplexing (OFDM), for example, we can increase the number of frequency carriers used. In digital TV systems already in place, there is a choice of 2000 or 8000 carriers (2k or 8k systems). It is possible that a similar approach will be taken for fourth-generation cellular. We can thus show the progression over time, that is, how bandwidth has become more adaptive over time.

Analog cellular handsets effectively had adaptive variable-rate encoding and adaptive variable-rate modulation. Some would argue it has taken digital processing 20 years to catch up with analog processing!

Who Will Own Handset Hardware Value?

There are three views of how handset hardware value may be distributed over the next 3 to 5 years . If you are a memory manufacturer, you consider memory as the most important component and add some DSP and microcontroller functionality to your

product proposition. If you are a microcontroller manufacturer, you consider the microcontroller to be the most important component and add some memory and DSP functionality to your product proposition. If you are a DSP manufacturer, you consider the DSP to be the most important component and add some memory and microcontroller functionality to your product proposition.

However, given that we are arguing that much of the future traffic value is uplink-biased (image and video and audio capture), then it could be implied that of all three components, the DSP is probably the most important.

The DSP effectively has a pervasive presence in the cellular handset at RF, IF, and baseband. Although chip-level processing may initially be undertaken by an ASIC, it is likely that, as with GSM, the DSP will creep back in as the most flexible and probably most cost-effective solution. This effectively determines the dominance of the DSP in terms of handset functionality.

In later chapters, we argue that 3G networks will only perform well if there is a common denominator handset hardware and software form factor sending traffic to and receiving information from the network. A DSP vendor is most likely to be in the position to enforce a de facto standard in this area.

Summary

In Chapters 1, 2, and 3, we described how digital processing is used increasingly to deliver RF performance (sensitivity, selectivity, stability). In this chapter, we described how digital processing is used to capture rich media components in the handset (voice, image, and video), to preprocess, compress, and multiplex those components (MPEG-4 encoders) and to recover or reconstruct/synthesize the original component waveforms in the receiver.

This ability to reconstruct/synthesize waveforms in the receiver allows us to deliver significant improvements in perceived bandwidth quality without a parallel increase in radio bandwidth. We have traded off processor bandwidth against radio bandwidth.

In future chapters we will explore the interrelationship of handset hardware, handset software, base station hardware, network hardware, and network software with 3G system planning.

CHAPTER

5

Handset Hardware Evolution

We have just described how bandwidth has become more adaptive over time—adaptive source coding, adaptive channel coding, adaptive modulation, adaptive multiplexing, and, for the future, adaptive RF channel spacing. By implication, this means that it is necessary for hardware to become more adaptive. You can take an all-purpose device like a DSP or a microcontroller and use different parts of it for different purposes. This is fine but arguably a little wasteful of resources. Alternatively, you can dynamically alter—that is, reconfigure—hardware to be reoptimized to a changed application requirement. This is the argument put forward by the makers of field programmable gate arrays (FPGAs). It is certainly true to say that if a number of processes occur sequentially, and provided FPGAs can be reconfigured sufficiently quickly, efficiency gains can be achieved.

A Review of Reconfigurability

Let's review what we mean by reconfigurability. There are three ways to define reconfigurable devices:

- Devices that are reconfigured by the vendor (often prior to shipment)
- Devices that can be reconfigured by a network
- Devices that can reconfigure themselves

Devices that are reconfigured by a vendor do not need to be connected to a network. Devices that are reconfigured by a network or reconfigure themselves include RF: infrared, copper, and optically connected devices. The design decisions to be made are as follows:

- What percentage of a product can remain fixed (that is, supported by conventional logic or an ASIC)?
- How much changeable overhead is required?

Bear in mind that reconfigurability has a price tag attached—the benefits have to outweigh the additional cost. FPGAs can be applied in predelivery reconfigurability. Effectively, FPGAs allow designers to compile C code, to produce algorithms and associated floating-point and fixed-point arithmetic, to decide on hardware and software partitioning, and then to produce hardware and to change their minds when the device doesn't work very well. Alternatively, if the marketing department decides mid-design that the specification needs changing, the hardware can be changed, since FPGAs are much more forgiving when it comes to finalizing gate-level hardware architectures.

FPGAs are also useful if you have a range of products at different prices that are basically the same product but with certain features enabled or disabled—a Bluetooth or GPS add-on, for example. Reconfiguration can be undertaken at any stage of the production cycle or even at point of sale. Handsets with infrared connectors, for example, can be reconfigured on the production line or, if staff are sufficiently well trained and security issues are addressed, in a retail outlet.

The User Service Identity Module (USIM) in a 3G handset is an example. It can be configured at point of sale or reconfigured at any future time, either back at the point of purchase or remotely over the air. A USIM reconfiguration is a change of software in the device. Effectively, in FPGAs, we are applying the same principle to hardware reconfiguration.

We can also define reconfigurability in terms of time scale, years, months, weeks, days, hours, minutes, seconds, millisecond, microseconds. The shorter the time scale, the higher the added value and (sometimes) the higher the performance value.

Static and dynamic rate matching in 3GPP1 is an example of reconfigurability. Here, the implication is that the processing environment, and, possibly, related hardware configuration, can change every 10 ms.

Present vendors of FPGAs are promoting their devices for use in Node B transceivers. The devices are also propositioned for use in handsets, though it will be hard to realize the required cost targets. On the Node B transmit side, FPGAs can be used to realize the linear feedback registers (generating the long and short scrambling codes) and are claimed to provide better silicon area utilization compared to programmable logic devices. Similarly on the receive side, FPGAs can be used to realize the matched filters for extracting the multiple-channel PN codes.

Sometimes, the theoretical benefits of FPGAs are hard to realize because of the need to *simultaneously* process signals in the device. After all, there is not much advantage in reconfigurability if jobs have to be done in parallel.

Figure 5.1, a block diagram from Altera, provides an example of possible partitioning in a Node B design. The chip rate despreading is, at present, a hardware-intensive process, as is the multiuser detection and combining. Essentially, all chip rate processors

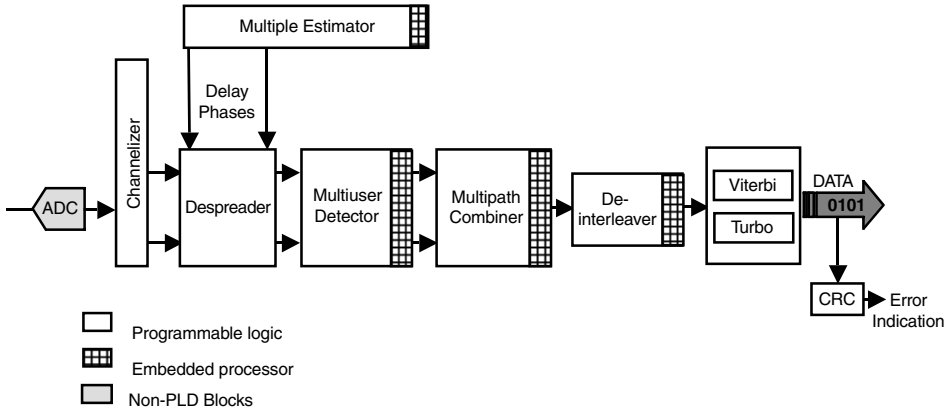


Figure 5.1 Altera 3G platform.

are currently hardware-based. The more complex symbol level processing tasks, such as turbo decoding, are hardware-based. The similar symbol-level tasks, such as de-interleaving, Viterbi convolutional decoding, and block decoding, and bit-level tasks, such as source coding, can be implemented in software on a DSP. Hardware tasks are therefore a candidate for FPGAs. FPGAs are used presently in Node B designs because there is still some fluidity in the standards-making process. Designers still tend to migrate toward ASICs to get maximum hardware performance out of a device, along with lowest per-unit cost. The cost of an ASIC, of course, is the lack of flexibility.

The idea of remote hardware reconfiguration seems attractive in theory but can be quite tricky in practice. An incorrect reconfiguration bit stream could physically damage millions of subscriber handsets, and the damage could be irreversible. Such damage could either be the result of incompetence or malicious intent. To protect against malicious intent, it is necessary to authenticate reconfiguration bit streams and to authenticate devices to which the reconfiguration bit stream is addressed.

The issues are not dissimilar to remote software upgrades—either way, it is always rather nerve-racking to have the potential of physically damaging millions of subscriber products all at once! There are various standards groups working on the security issues of remote reconfiguration, including the Internet Reconfigurable Logic Group supported by Xilinx and Altera.

As we said earlier, FPGAs in Node B designs have a number of well-defined benefits. The use of FPGAs in a handset will be harder to justify.

Figure 5.2 shows how the DSP in a 2G cellular handset does more or less all of the bit-level/symbol-level processing, as well as quite a lot of preprocessing for the RF stages of the device. This was not always the case. In 1992, about the only task the DSP was capable of realizing was the speech encoder/decoder. Between 1992 and 1997, the DSP gradually took over other jobs like convolutional encoding/decoding and channel equalization. The microcontroller looked after higher-level protocol tasks and the man/machine interface. By the end of the 1990s, more or less the whole baseband was being done on (largely TI!) DSPs.

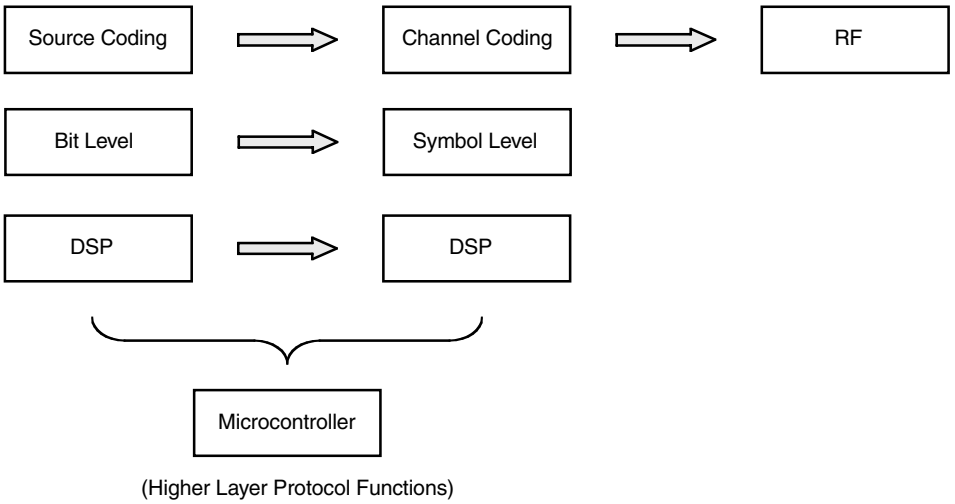


Figure 5.2 Present GSM handset configuration.

Present GPRS handsets are similar in that the DSP is completely dominant in the baseband area, as shown in Figure 5.3. Additionally, tasks such as image processing and MPEG-4 encoding/decoding are typically divided between the DSP and microcontroller. The DSP does repetitive signal processing tasks, such as Fast Fourier Transform (FFT) transitions, and the microcontroller looks after housekeeping tasks, such as organizing memory fetch processes, interrupts, and parallel multitasking. These are all the rather unpredictable tasks for which DSPs are not well suited.

In a 3G handset, early designs are back to the DSP only doing about 10 percent of the baseband processing. In their book *The Application of Programmable DSPs in Mobile Communications* (Wiley, ISBN 0-471-48643-4) Alan Gatherer and Edgar Auslander provide a well-reasoned argument as to how and why the DSP will repeat its 1990s trick of gradually taking over all other baseband processing in the handset, including, in the longer term, chip rate processing (OVSF spreading/despreading).

Presently, however, tricky jobs like the turbo coder/decoder have to be realized using flexible or reconfigurable coprocessors—hardware accelerators running beside the DSP. This adds cost and complexity. The need to manage a number of simultaneous processing tasks, which are very real-time dependent—for example, the processing of time-sensitive, time-interdependent multiple OVSF code streams—has meant that DSPs have to have their own real-time operating system, which, hopefully, will communicate with the microcontroller RTOS.

It is worth noting that the majority of the sources being coded (audio and video) are analog. Also, of course, the RF carrier is analog (a sinusoidal waveform onto which we superimpose analog phase and amplitude signals).

This has led to proposals to produce an analog DSP. An example is from Toumaz (www.toumaz.com) and is a proposed analog implementation of an FFT using bandpass filters.

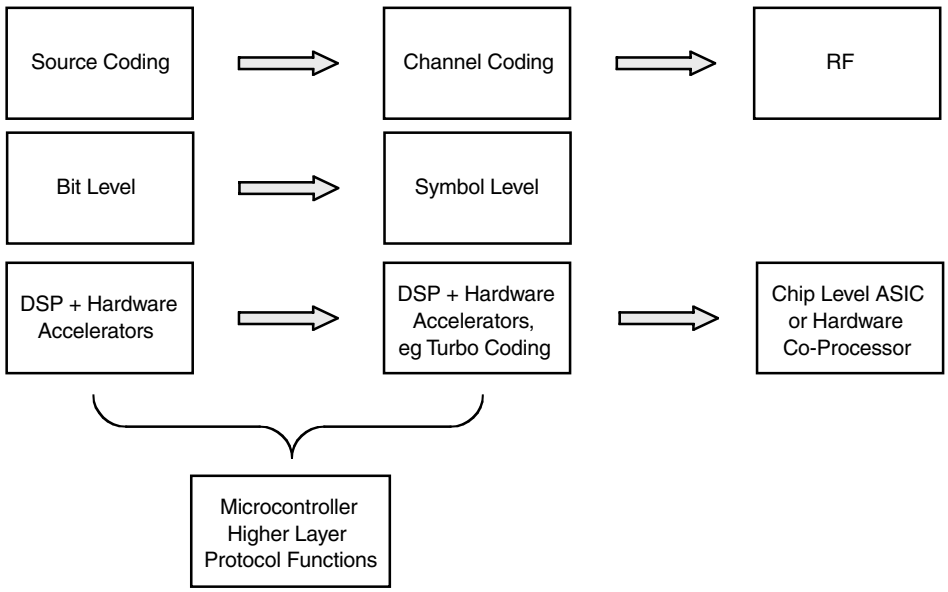


Figure 5.3 3G Handset configuration.

In the meantime, we are left with heavy-lifting DSP solutions for baseband coding/encoding. As coding overheads increase and user data rates increase, processor efficiency becomes increasingly critical. By its very nature 3G is a wide dynamic range, slow-to-fast data rate, multifunction (voice, text, video), multiuser cellular system. This requires optimized low-power, flexible digital-processing capability.

DSPs can provide low-power (relatively), software-adjustable digital processing capability. However, DSP architectures are fixed. There is no capability to modify the interconnects between function units. Under each processing requirement a number of functions will remain unused—a measure of inefficiency. A number of vendors attempt to address this problem with Reconfigurable Communication Processors (RCP).

The vendors, such as Chameleon, Siroyan, Elixent, PicoChip, Prairie, MIPS Technologies, and ARC, are propositioning more efficient DSP core architectures. Approaches include a number of different methods to take advantage of parallelism and pipelining in the algorithm. For example, Chameleon has a three-core architecture where the first core is an embedded processor subsystem, the second core is a 32-bit reconfigurable processor with 108 parallel computation units, and the third core is a programmable I/O (PIO) with a 1.6-Gbps bandwidth. Elixent has a large array of simple arithmetic logic units (ALUs) with distributed memory and local/global interconnects. These are arranged in a chessboard array of tiles. Each tile takes on the responsibility for particular logic functions as demand requires.

Siroyan concentrates on optimizing compiler efficiency and then matching the DSP architecture to the compiler. (Normally you optimize the compiler architecture to the

DSP.) Siroyan's approach is to get the software right first, and then design the hardware around it. In doing so, the hardware ends up being a more scalable distributed DSP optimized for simultaneous parallel processing.

These parallel processing/distributed processing DSP architectures are well suited to image processing: They can be optimized for processing multiple variable-width data streams onto multiple variable-rate physical radio channels (OVSF code channels). Because they move across a complex radio layer into a complex transport layer, they are well suited to preserving the properties of multiple data streams.

Another option is an optical DSP (ODSPE). Optical DSPs promise substantial gains in processor speed and efficiency. An example is Lenslet's EnLight ODSPE product family (www.lenslet.com).

For the moment, electronics will have to do, and we just need to find ways of efficiently using them.

Flexible Bandwidth Needs Flexible Hardware

As bandwidth becomes more bursty, hardware has to become more flexible. The success of the ARM microcontroller has been the combination of the concept of a very long instruction word and the variable-length instruction word—the ability to change register bit width.

In DSPs we have gradually begun to see the use of multiple multiply-accumulate (MAC) units. This allows parallel and flexible processing to be performed, provided the compiler has been designed sufficiently flexibly to take advantage of the multiply-accumulate functions. An example is the StarCore SC140. At 1.5 Volts the processor can handle 1200 million multiply-accumulate functions per second, equivalent to 3000 RISC MIPs, assuming all four MAC blocks are fully utilized. The consumption of the core, excluding peripherals, is 198 mW.

In these devices, typically two-thirds of the overall power requirement is created by the drivers and interfaces. The more this functionality can be brought onto the chip, the better the efficiency of the overall solution. However, this will tend to make the solution less flexible in terms of the application footprint. It is sometimes hard to realize flexibility and power consumption objectives.

Flexibility, along with the ability to parallel process, becomes a hardware quality metric. Distributed DSP provides a good example of how to be flexible and parallel.

The same principle applies to memory, for example, the need for flexible lookup tables when realizing filter structures. These filters are described as distributed arithmetic filters and can deliver significant throughput and efficiency gains.

Summary

As content becomes more complex, channel processing becomes more complex. Complex processing requires complex hardware.

As bandwidth becomes burstier, we have to provide more adaptive delivery bandwidth. This requires adaptive hardware and, as we will see in the next chapter, adaptive software. Adaptive hardware has costs: Designing RF power amplifiers and power supplies to handle highly variable bit rates is quite tricky. Reconfigurable DSPs, variable bit width microcontrollers, and reconfigurable memory add cost and complexity. We need more expensive and exotic battery technologies to support high peak to mean power requirements; bursty bandwidth is expensive bandwidth.

Most present reconfigurable components for example, FPGAs, reconfigurable DSPs, and devices such as optical DSPs are not suitable for implementation in handsets because of power budget and cost constraints. Application knowledge with these devices is gained from their use initially in base stations (Node Bs). As power efficiency improves and costs reduce, these techniques become applicable in handset designs.

3G Handset Software

CHAPTER

6

3G Handset Software Form Factor and Functionality

In Chapter 3 we described the various hardware inputs that we have on a 3G handset—the wideband audio microphone (to capture high-quality audio streaming), the wideband megapixel CMOS imager, the keyboard (application capture), and USIM (access and policy rights control). We now need to consider how handset software is evolving to manage and multiplex these multiple inputs. We will also define how handset software determines session persistency and session quality.

An Overview of Application Layer Software

The raison d'être of the application layer is to take a simple exchange—for example, a voice or messaging exchange—and transform it into a rich media exchange, as follows:

- I talk to a friend.
- The application layer software prompts me to exchange an image file.
- The application software prompts me to send some simultaneous data (information on the image file).
- The application layer prompts me to load a simultaneous video exchange.
- The application software then prompts me to increase the color depth, resolution, or frame rate.
- I end up spending lots of money.

The software has influenced session persistency. What started off as a short, bursty exchange has become a persistent duplex flow of complex content, separately managed on multiple physical layer channel streams.

In our very first chapter, we pointed out how code bandwidth has expanded with each successive cellular generation (see Table 6.1). This has a largely unrecognized but profound effect on network topology. Suppose each handset has 1 million lines of code. For every 1000 subscribers you have 1 trillion lines of code—subscriber-based software code bandwidth. Similarly, if each handset had 10 Gbytes of solid-state or hard disk storage, then every 1000 subscribers represents 10 Tbytes of distributed storage. If each handset is capable of processing 1000 MIPS, then for every 1000 subscribers, there are 1 billion MIPS of distributed processing. As memory and MIPS migrate to the network edge, added value follows.

A traditional AXE switch has 20 million lines of code. In the preceding example, we have said that for every 1000 subscribers we have 1 trillion lines of code. That is 1 trillion lines of code at the edge rather than the center of the network. As code footprint in user equipment increases, the network is increasingly bossed around by the devices accessing the network. The software footprint in the user's device substantially influences offered traffic (uplink loading), which, in turn, influences offered traffic value (uplink value).

We have described how bandwidth burstiness is increasing as we move from constant-rate to variable-rate encoding, and from single to multiple (per user) traffic streams. We can smooth burstiness by buffering, but this absorbs memory bandwidth and introduces delay and delay variability (the latency budget). Application layer software therefore has to be capable of managing these multiple per-user channel streams, which implicitly means the application layer software needs to be good at multitasking.

There are different processes influencing bandwidth burstiness. Each individual content stream is variable rate (or, in the case of video encoding, may be variable rate). In addition, content streams are being added to or subtracted from the application layer and radio physical layer multiplex—in other words, static matching (addition or subtraction of channel streams) and dynamic matching (data rates varied on a 10-ms frame-by-frame basis).

Table 6.1 Software Form Factor and Functionality

CELLULAR PHONE GENERATION	MEMORY BANDWIDTH	PROCESSOR BANDWIDTH (MIPS)	CODE BANDWIDTH (LINES OF CODE)
1G (1980s)	Kilobytes	10	10,000
2G (1990s)	Megabytes	100	100,000
3G (2000-2010)	Gigabytes	1000	1,000,000

There are five candidates for digital cellular handset application software:

Microsoft. This company has traditionally majored on time transparency—each successive generation maintained the look and feel of previous generation products. Usually (up until Windows 2000), the products were more or less fully backward-compatible. This feature, however, came at a cost—additional memory and processor footprint.

Sun/Java. With their J2ME operating system (Java 2 Micro Edition, sometimes also known as Java 2 Mobile Edition) optimized for wireless PDAs, there is a heritage of offering platform transparency—write once, run anywhere (WORA). This is achieved by using byte-level compiling. Effectively the software is abstracted to a higher level to make it easier to write and make Java applets easier to move from platform to platform. This feature, however, came at a cost—additional memory and processor footprint.

Symbian. Here we have a heritage taken from their experience with Psion PDA operating systems. Starting in 1984, Psion produced low power budget PDAs that could run on two AA batteries. To differentiate the hardware product, the operating system and application software were optimized for multitasking—the ability to do several things at once and have several applications open at once, and to be able to move sideways from one task to another (and return to the original task in its original state). This activity provides a good basis for implementing the management and multiplexing of the rich media mix coming from the MPEG-4 encoder in a 3G cellular handset, but multitasking comes at a cost—additional memory and processor footprint.

Palm. With a similar PDA heritage—starting later but with higher (initially U.S.-based) market volume, Palm's differentiation was to provide options for inputting information into the device—for example, handwriting recognition and bar code readers. Given that 3G digital cellular handsets are becoming effectively input appliances, this provides a good basis for the Palm OS to develop into a multi-input management platform, but flexible input platforms come at a cost—additional memory and processor footprint.

Linux. Arguably the wild card of the pack and conceived as a way of reducing Microsoft's dominance in PC software, Linux supports *open code software*, which means the original source code is made available in the public domain to the software design community. Anyone can suggest and help implement improvements in the open source code, resulting in collaborative software development. This helps to avoid disputes over software intellectual property ownership and therefore reduces software component cost.

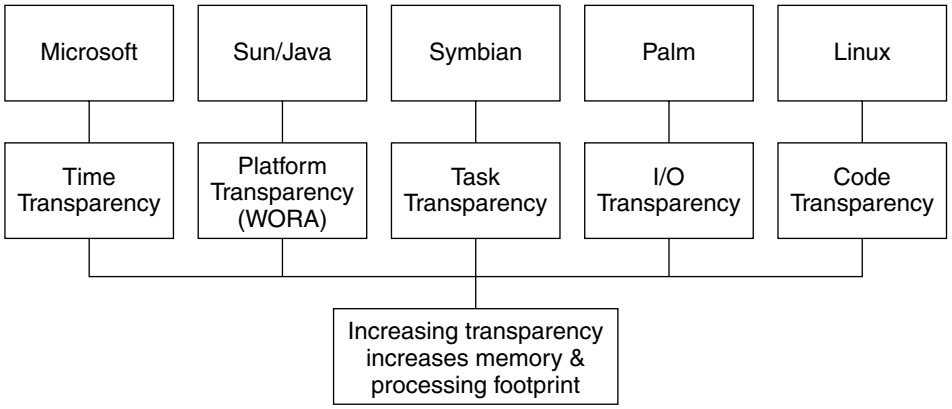


Figure 6.1 Software form factor and functionality.

Higher-Level Abstraction

With any of the preceding application layer operating systems, there is an issue of application transparency, or rather, a lack of transparency across different software and hardware platforms and different radio physical layers. Qualcomm has developed and promoted a higher-level software layer product known as BREW (Binary Run Time Execution for Wireless) whose purpose is to provide this application transparency. Typical services supported include browser functionality, instant messaging, position location, gaming, e-mail management, buddy group management, music file exchange, and information service management. In addition, BREW sets out to standardize the capturing and processing of application-based billing and addresses the need for transparency across different radio physical layers (CDMA2000/W-CDMA) and different networks (GSM-MAP and ANSI 41). These relationships are shown in Figure 6.2.

More information on BREW is available via Qualcomm's Web site, www.qualcomm.com.

The Cost of Transparency

Whatever application software is loaded into a handset, it generally comes at a cost. This is actually a major issue in digital cellular handset economics: Hardware designers are given a target, for example, to reduce GSM hardware component costs (the bill

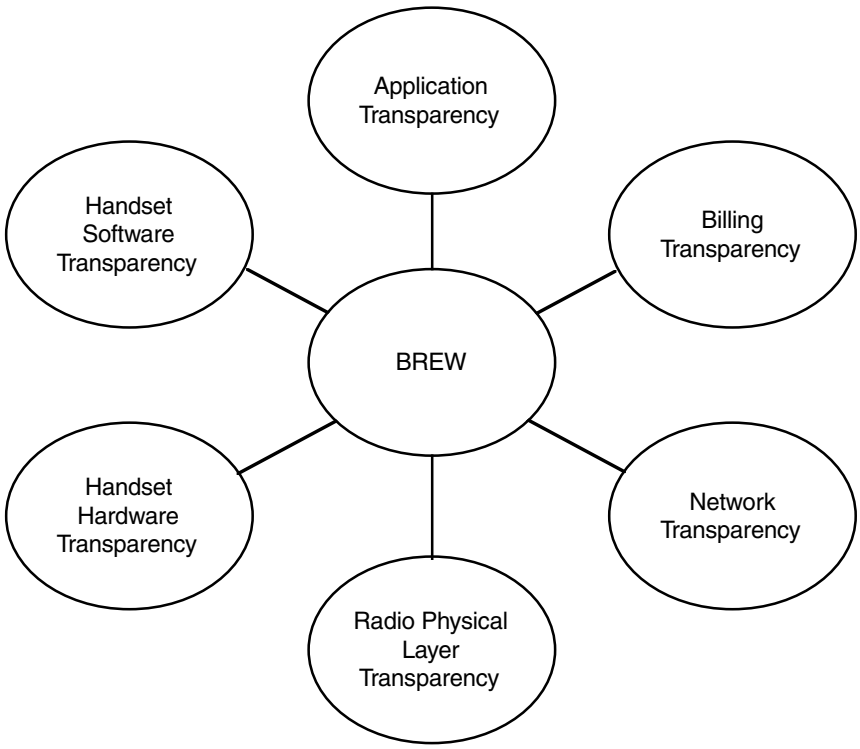


Figure 6.2 Hardware and software application transparency.

of materials) to $40 per handset. Suddenly, the software team announces that their chosen application layer OS will add $10 of licensing cost to each handset. Remember that the hardware component cost includes a material cost, so the hardware margin would be no more than $10. Half the added value of the handset has suddenly moved into software-added value.

The open code software proposed by Linux provides one solution to this. Allowing lots of different design inputs often has the beneficial effect of increasing the application bandwidth. The software can do a wider range of tasks, but this comes at a cost—additional memory and processor footprint.

From an application performance perspective, you would have an operating system that would provide time transparency, platform transparency, task transparency, I/O transparency, and code transparency, but the code and processor overheads would be unsupportable in a portable device.

Typical Performance Trade-Offs

As processor overhead increases, the delay budget increases. As flexibility increases, delay variability increases: The OS allows more interrupts, since more control and choice to the user increases interrupt overheads. The delay and delay variability introduced by the application layer OS becomes, as we will show, a major part of the end-to-end delay and jitter (delay variability) budget. If we are judging quality on the basis of end-to-end delay and end-to-end delay variability, then application layer software response times become a critical performance metric.

We also have to qualify how well the software coexists with the target hardware platform. The more deeply we embed software, the more remote we make the software from the outside world, the more deterministic we can make software performance—that is, the better we can manage delay and delay variability. However the more deeply embedded the software, the less flexible it becomes; the outside world cannot influence the software and therefore has no control over it.

Exploring Memory Access Alternatives

The application software products we've mentioned so far are usually ROM-based products. They do not need to boot up from a hard disk, because the host device does not usually have a hard disk. This makes it hard—actually, impossible—to remotely reconfigure over the air, but it helps to protect the OS from virus infection and means that the software turns on more or less instantaneously.

The ROM-based OS needs to talk to localized and distributed memory within the device. The time taken to go and fetch data from memory and then act on that memory in part determines the delay and delay variability budget. Table 6.2 shows a typical 16-bit microcontroller (from Hitachi) with on-board RAM.

You can reduce the clock speed but only at the cost of increasing the instruction cycle time. Similarly, you could reduce the operating voltage (which is higher than you would want in a digital cellular handset), but this will again slow the cycle time.

Table 6.2 Memory and Microcontroller Specifications

	H8/3062 BF[2]	H8/3064BF	H8/3067F	H8/3068F
Flash size	128 kbyte	256 kbyte	128 kbyte	384 kbyte
RAM size	4 kbyte	8 kbyte	4 kbyte	16 kbyte
Instruction cycle time	80 ns/25 MHz	80 ns/25 MHz	100 ns/20 MHz	80 ns/25 MHz
Operating voltage	5 V			

The problem of memory access is that speed of access is only improving by about 7 percent per year, whereas raw processor speed is rising at about 60 percent per year. It's not the memory or the processor that's the problem; the problem is the interface between the two devices.

The answer is to embed the memory in a system on-chip solution. However, then you need to decide where to put the memory: with the DSP or with the microcontroller or, as is (usually the case, with both, in which case you need to optimize the intercommunication between the microcontroller and DSP.

In terms of organizing memory for maximum performance (minimum delay and delay variability), the general rule of thumb is to have the fast-access storage cells on chip and relatively slow cells on DRAM, and then to work out what should be where and when. This is known in the industry as algorithms of probability and locality. It also becomes important to throw things away when not needed, a bit like good housekeeping. This is referred to in the industry as garbage management.

The problem is that the performance problem that has always existed for off-chip memory access is beginning to reappear for on-chip memory. The solution is to have processors that hide memory access delays by multithreading—that is, handling several tasks at once and switching between them each cycle. Essentially this means that we have a memory real-time operating system that needs to coexist with the microcontroller real-time operating system that, in turn, needs to coexist with the DSP real-time operating system.

Infineon Technologies has tried to bridge the divide that is beginning to open up in terms of design tools and design rules in each of these separate areas. Table 6.3 illustrates an example of a product sampled to the 3G handset design community in the late 1990s that combined a DSP and microcontroller core with Flash, RAM, and ferroelectric random access memory (FRAM). This was a 500-MIPS device. In practice, it has become necessary to have at least 1000 MIPS available. The selling proposition for TriCore is that DSP, memory and microcontroller functions are defined by a common software development environment, which in turn can take advantage of new technologies like FRAM. The table shows the gate density performance benefits realizable from decreasing device geometry from 0.35 micron to 0.18 micron.

Table 6.3 Infineon TriCore Development

TECHNOLOGY	0.35	0.25	0.18
TriCore core	100 MHz	150 MHz	200 MHz
ASIC gates (max)	300 kbit	500 kbit	> 700 kbit
Flash/OTP (max)	4 Mbit (0.35)	16 Mbit	32 Mbit
eDRAM (max)	16 Mbit	32 Mbit	64 Mbit
Flash + eDRAM	N/A	TBD	FRAM
MIPS	130		500
Year	1998	1999	2000

FRAM is a really useful memory product. It is not as dense as DRAM or Flash but is low power, and it will survive about 10 trillion read/write cycles. In addition, these devices are sometimes described as persistent, or nonvolatile, storage devices, which means they do not lose their memory when the handset's battery goes flat, and they have about a 10-year data retention. They also provide fast read, write, and bit-level erase capability. Essentially, you can think of such devices as solid-state hard disks, since both exploit ferroelectric and magnetic effects to provide storage. About 20 times faster than EEPROM, FRAM is beginning to appear both as a standalone product and on smart cards.

Hitachi offer a range of products optimized for the storage and redelivery of multimedia files. These devices come in 16, 32, 64, and 128 Mbyte packages and use interleaving (the simultaneous writing of two or more Flash memories) to deliver write speeds of 2 Mbps and read speeds of 1.7 Mbps. The write time for 500 kbytes of image data from a 3-megapixel digital camera is about 0.25 seconds. This highlights the importance of memory bandwidth performance and, specifically, memory delivery bandwidth performance.

Most of the focus for portable products has been solid-state memory, but it is also worth considering parallel developments in miniature disk device storage. The pervasiveness of laptop PCs has greatly improved the mechanical robustness of hard disk drives. Micro-miniaturization techniques have also made possible miniature disk drives that are both space- and power-efficient—and offer huge amounts of storage bandwidth.

Miniature disk drives (fitting within a Type III 10.5-mm form factor PC card) have been available since 1992 and have increased over the past 10 years from providing a few Mbytes of storage to a few Gbytes. Type III card devices today are capable of storing 15 Gbytes.

In 1999, Type II PC card devices (5 mm thick) became available using magnetic resistance heads with a read density of 8500 tracks per inch and offering about a 10 times reduction in storage cost compared to solid state. The example shown in Figure 6.3 is an IBM 1-Gbyte Microdrive, a hard disk drive in a CompactFlash Type II PC card format. In terms of storage bandwidth, this is sufficient to store 1000 high-resolution photos, 12 music CDs, or 1000 novels. The device delivers a 4.2 Mbps transfer rate, which is over twice as fast as solid-state Flash, and a 1 in 10^{13} bit error rate. It weighs 16 grams, so is not too implausible as an add-in product to a digital cellular handset, which typically weighs 80 grams. (The hamster is not included.)

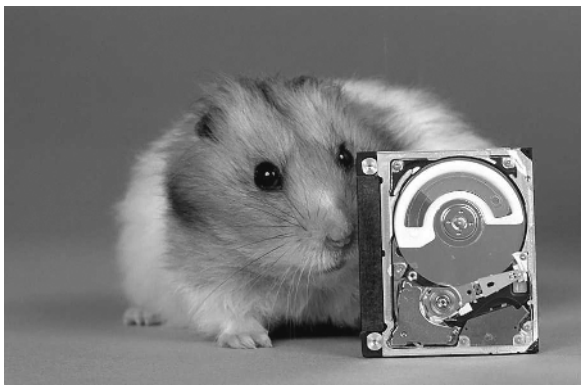


Figure 6.3 IBM 1-Gbyte hard disk drive.

Software/Hardware Commonality with Game Console Platforms

There are a number of parallels between digital cellular handset memory and processor footprints and game console memory/processor footprint. One is that both use what's available and what can be afforded within the product/cost form factor. The performance of game consoles is of interest to us because many of the motion estimation, prediction, and compensation techniques used in today's game products may be used in future digital cellular handsets.

The InTouch product from the wireless technology specialist TTPCom is an example of gaming software integrated into a relatively standard handset (see Figure 6.4).

Table 6.4 shows how the central processor clock speed and memory have increased over the past 8 to 10 years in traditional mains-powered game consoles. The challenge is to realize this type of performance in a portable handheld device.

The PlayStation (PS2) is a Toshiba MIPS device consisting of 13 million transistors in a 0.18 μm process. The device uses a substantial amount of prestored graphics to provide an interactive rich media experience, and it is not unreasonable to expect to see more technology and application convergence with digital cellular devices over the next five years (PS2s also work very well as DVD players).

Microsoft's Xbox uses a hard disk to minimize loading delay (hard disks provide typically twice the access rate of solid-state memory), providing faster manipulation of pixels and polygons.

Figure 6.4 Gaming handset from TTPCom (www.ttpcom.com).

Table 6.4 Processor/Memory Footprints—Games Consoles

	PLAY-STATION 1	NINTENDO 64	DREAM-CAST	PLAY-STATION 2	IBM/GEKKO GAMECUBE	XBOX
Launched	1994	1996	1999	2000	2001	2001
CPU	32 bit	64 bit	128 bit	128 bit	128 bit	Pentium 3
Clock speed	33 MHz	93 MHz	300 MHz	300 MHz	400 MHz	733 MHz
Main memory	2-Mbyte RAM	36-Mbyte DRAM	16-Mbyte RAM	32-Mbyte RAM	28-Mbyte SRAM	64-Mbyte RAM
Video RAM			8-Mbyte RAM	4-Mbyte DRAM		64 Mbyte
Audio RAM			2-Mbyte RAM	2-Mbyte RAM		64 Mbyte
Media	CD	Cartridge	ROM	DVD	Mini-DVD	DVD
Modem			56 kbps			Ethernet
Operating system	PSI		Win CE	PS2		Windows 2000

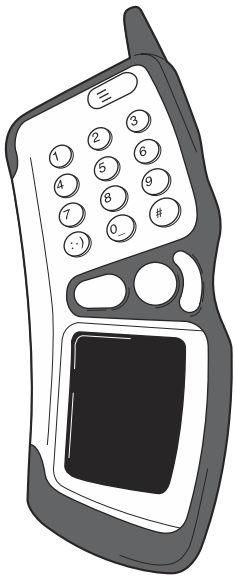


Figure 6.5 Plug-on/add-on fascia from Wildseed.

Add-On/Plug-On Software Functionality

Adding game functionality to a cellular handset is only really useful for people interested in playing games. This means that for everyone else it is an unnecessary overhead (in terms of cost and processor/memory bandwidth overhead) being added to the phone. An alternative is to provide the additional software functionality via a plug-on or add-on or plug-in or add-in component. Plug-on devices are added on top of an existing product and plug-in devices are added in to an existing product.

The product illustrated is a plug-on mobile phone fascia with an Intel StrongARM processor that when added to the handset changes or enhances its functions—for example, ring tones, games, screen savers, and thematic Web links (Web links associated with particular user group interests). The product illustrated in Figure 6.5 is known as Smart Skin. More details can be found on the vendor's Web site, www.wildseed.com.

Add-in/Plug-in Software Functionality: Smart Card SIMS/USIMS

We have already profiled the use of smart card SIMS/USIMS in Chapter 4 on 3G handset hardware form factor and functionality. Essentially, the smart card USIM is just another plug-in memory module with a 16-bit or 32-bit microcontroller.

Full-sized ISO cards are also proposed as additional plug-in memory platforms. The GEMPLUS SUMO card (Secured Unlimited Memory on Card) combines seven flash memory chips on a smart card and can be used in a digital cellular handset (as a 64-Mbyte plug-in SIM) or in a PDA, PC, or set-top box. It can support a total of 224 Mbyte of memory (8 hours of MPEG3 audio, 12 minutes of MPEG video, or 100 e-books) and has a 20 Mbps data link capability.

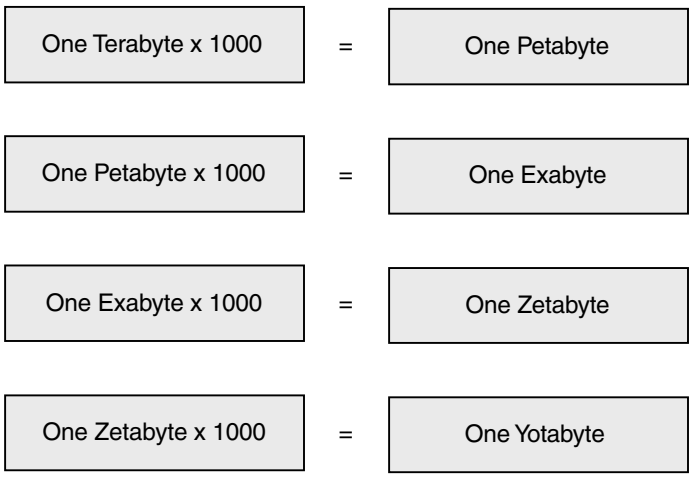


Figure 6.6 Memory bandwidth scalability.

In the introduction, we said that the memory footprint in digital cellular handsets was moving from Megabytes (2G) to Gigabytes (3G). Consider that a typical data warehouse today is about 10 or 20 Terabytes. Vodafone has a 10-Terabyte data warehouse to integrate customer complaints and engineering performance. If you had 10 Gbytes in each subscriber device and 100 million subscribers (to use Vodafone as our example), you would have a 1000-Exabyte data warehouse, equivalent to one Zetabyte (see Figure 6.6).

The Distribution and Management of Memory

As we will see in later chapters, you can build a completely new business model on storage, particularly distributed storage. Distributed storage may even have the benefit of being supplied and paid for by your subscribers. Storage bandwidth does however incur management overheads. We need a memory real-time operating system to go with the DSP real-time operating system (RTOS) and microcontroller RTOS .

Memory has to be managed at a micro and macro level. At the micro level, we have to distribute memory in a handset close to the point of consumption—next to the DSP and next to the microcontroller—and then optimize the partitioning of the memory to match the tasks being undertaken.

At the macro level, we need to decide how much memory we should put in the subscriber product and how much in the network and where. The choice is extended by the provision of Web-based storage. The example in Table 6.5 is from Xdrive Technology's Web site (www.xdrive.com) and shows a number of options for accessing Web-based virtual storage.

Table 6.5 Web-Based Storage

XDRIVE PLUS SERVICE PLANS	TOTAL STORAGE	MONTHLY BY CREDIT CARD
Standard	75 Mbytes	$4.95
Enhanced	150 Mbytes	$9.90
Professional	500 Mbytes	$29.95
Multimedia	1000 Mbytes	$49.95

For wireless devices, this is one option for extending the apparent storage capability of the device but with the proviso that quite a lot of processor power (battery bandwidth) will be needed to recover files *from* storage and quite a lot of RF transmit power (also battery bandwidth) will be needed to send files *to* storage. Remember that over the radio physical layer, we are not particularly short of delivery bandwidth, but we are short of power. This tends to mean that it is better to have storage in the subscriber device rather than in the network (or in the case of Xdrive, in a network the other side of the access network).

We are also introducing delay because of the need to use the radio physical layer to download or upload files. Additional delays may be introduced by the network/networks between the subscriber and the storage. Finally, the remote storage itself will have a latency (delay and delay variability), which will be a consequence of the loading on the server.

Virtual storage solutions define quality of service in terms of access delay, security, and policy management. For example, Amazon identifies customers in terms of their spending power and spending habits. This information can be stored as a cookie in your computer. When you access the Amazon site you get privileged access to the server if you have been identified as a big spender—a process known as *virtual resource management*.

The ability to store complex content either in the subscriber appliance or the network so that either the subscriber or the network can choose when to send or exchange files also helps smooth out some of the peak loading experience in a network.

The ability to shift loading, for instance, from the daytime to nighttime is dependent on the ability to store information that is not delay-sensitive. It is also dependent on having application layer software that is sufficiently intelligent to make and take the decision to store or send, which probably means the ability to negotiate with the network.

We are assuming that the 3G handset will be encouraging subscribers to create their own content. If the local storage in the subscriber device becomes full, the subscriber will want to send the file for remote storage. It would be more efficient for the network if this were done at night. It could also be lower cost for the user. This is only an extension of existing pricing policies, where lower cost calls can be made in the evening.

Examples of subscriber-generated content can be found at the following sites:

- www.my-wedding.com
- www.my-kids.com
- www.my-pets.com
- www.my-party.com

A number of Web-based storage providers have been established to provide remote virtual archiving:

- www.shutterfly.com
- www.ofoto.com
- www.photonet.com
- www.gatherround.com
- www.cartogra.com
- www.photopoint.com

Virtual archiving, however, requires some mechanism for establishing image ownership. We cover authentication and encryption in a later chapter, but for now, we just need to know that we must be able to identify and sign complex content so that we can prove that we own or at least produced the content prior to archiving. Note that we might want to realize value from our stored images. We might also expect image value to appreciate over a number of years. Providing authentication and encrypting files in long-term virtual storage is a tricky proposition. What happens if we lose our authentication key? We cannot access our files, and even if we can, we cannot decrypt them.

In 1986, the BBC spent several million pounds creating a Domesday project file. It was the 900th anniversary of the establishment of the Domesday Book— the systematic recording of taxable assets by William the Conqueror. The BBC decided to create a new Domesday Book that would provide a snapshot of life in 1986 and could be available for study 900 years later (just as the Domesday Book can be read today). Sixteen years later the BBC discovered they had lost the source code used to store and manage the files, and the software engineer involved had long ago left the corporation. The file is at time of writing completely unreadable. This highlights the fact that electronic media storage is far from dependable as a mechanism for long-term storage. (Needless to say, after 900 years, the Domesday records—on parchment—remain in good shape.) The Digital Preservation Coalition has a useful handbook on this topic, produced in association with the British Library and available on www.dpconline.org.

The point about virtual storage and the purpose of digital storage is to earn money from people who wish to store material and *then retrieve material* at some later date. Once we have addressed the issues of long-term key management and long-term digital storage stability and accessibility, the potential exists for realizing value from image redistribution.

Some would argue that the network in the transaction is purely a dumb delivery machine and needs to have no involvement in storage provision. This is the basis for peer-to-peer networking in which files are exchanged directly, between users without

any network interaction—the jargon word used is *disintermediation*. Network operators generally do not want to be disintermediated. Napster was an early example of a company who effectively disintermediated music distributors by setting up peer-to-peer exchange of MPEG-3 audio files. Unfortunately, they used a central server to log exchanges. The records from the central server could be subpoenaed, and Napster (as a free exchange mechanism) was effectively shut down. Napster Mark IIs have appeared, and also Aimster and Mesh, who avoid the use of a centralized server. These companies provide a peer-to-peer exchange product, which at time of writing, appears relatively robust to malign intervention.

Summary

We are putting into people's hands products that can physically capture substantial amounts of simultaneous audio and video bandwidth in four distinct ways, as shown in Table 6.6. We can add additional text and information (via a keyboard) and can digitally sign complex content prior to sending or storing the information. The complex control can either be stored in the device, in the network, or on the other side of the network (for example, a peer-to-peer exchange).

The job of the application software in the user's handset is to manage this complex content. This includes managing the storage requirements of the user or the user's device or both. The software has to have sufficient intelligence to know when subscriber-resident, device-resident storage is about to run out and provide the option to use virtual storage resources. Virtual storage resources, however, present some fairly unique authentication issues.

In a traditional voice phone call, network software (using SS7 signaling) sets up a call, maintains the call, and clears down the call. The call is then billed. In a multimedia exchange, network software (which we describe in Part IV) sets up a session, manages the session, and clears down the session. When virtual archiving is involved, the session might last 1000 years or more! It is uncertain whether any electronic storage media would be sufficiently stable, either in hardware or software terms, to provide this kind of life span.

Table 6.6 The Creative Appliance—the 3G PC

INPUT METHOD	FUNCTION
Microphone	Voice/audio capture
CMOS imaging	Image and video capture
Keyboard	Application capture
Smart cards	Security context, ownership rights (digital signatures), QoS requirement definition

More prosaically, we are reliant on our application software to try and convince our subscriber to spend more money with us. A simple exchange (SMS or voice) needs to be upgraded into a complex exchange of time-sensitive rich media files. This includes bringing multiple participants into a session. It is the job of the software in the handset to increase session persistency, session complexity, and session value.

In 2000/2001, Sony started a global advertising campaign called "Go Create." Essentially, Sony is trying to change the consumer appliance into a creative appliance. (The consumer electronics industry becomes the creative electronics industry.) Consumption is a passive (and ultimately lackluster) pastime. When we create something, such as a media file, our natural inclination is to share our creation with other people—a sort of egocentric rather than network-centric value proposition.

All this suggests the need for an integrated storage management and delivery management real-time operating system. In wireless, this RTOS needs to take into account the qualities (for example, inconsistencies) of the radio channel. In a wireless IP network, the RTOS needs to take into account the qualities (for example, delay and delay variability) of the IP network.

In Chapter 20, "Network Software Evolution," we discuss Sun's Java-based storage network operating system, called Jiro. Such an operating system needs to interact with the OS in the subscriber handset. In other words, we need to integrate radio bandwidth, network bandwidth, and storage bandwidth performance in order to deliver a consistent and predictable end-to-end user experience. This is a subject that we will revisit in substantial detail.

CHAPTER 7

Source Coding

In earlier chapters we discussed the rich media multiplex. How do we capture the properties of wideband audio, image, and video and preserve the properties of the rich media mix as we move across the radio layer—into and through the network? Source coding is arguably the single most important process to address when we look at capturing and preserving complex content value. It effectively dictates how we dimension and prioritize our radio layer and network layer resources. In particular in this chapter we want to review how the evolution of MPEG-4 will influence future handset software functionality.

An Overview of the Coding Process

Let's begin this chapter by reviewing how we separately source-code voice, text, image, and video content. The following sections treat each of these topics in detail.

Voice

We have already discussed the adaptive multirate vocoder and wideband vocoder specified by 3GPP1. This is a speech synthesis codec and, as a result, provides us, conveniently, with the ability to support speech recognition, also specified by 3GPP1. The better the accuracy of the speech recognition (the distance from user to user), the higher the value. Similarly, the better the voice quality (measured on a mean opinion score), the more user value we deliver, but the more it costs to deliver, because of a higher coding rate.

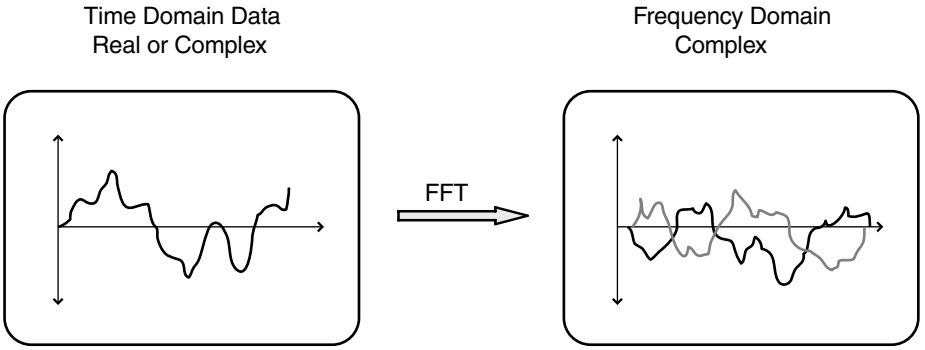


Figure 7.1 Audio codec—time domain to frequency domain transform.

These audio codecs use a time domain to frequency domain transform (discrete cosine transform) to expose redundancy in the input signal (see Figure 7.1). We send filter coefficients that describe the spectral/harmonic (frequency domain) content of the 20-ms speech sample.

MPEG-4 also has an audio coding standard including a very low bit rate harmonic codec (2 to 4 kbps) and a codebook codec (4 to 24 kbps). The codebook codec stores waveform samples in the decoder. When the digital filter coefficients are received, the decoder goes and fetches the closest-match waveform from the decoder—hence, the need for good memory fetch management in these devices. The intention is that the MPEG-4 CELP (codebook excitation linear prediction) codec will be compatible with the AMR-W codec, which has a similar codec rate range.

Text

Having captured our wideband (16 kHz) audio, we now want to add some text. Text source coding has traditionally been realized using ASCII (American Standard for Communications Information Interchange). These are 7-bit words that are used to form a 7-bit alphabet used to describe letters of the alphabet, numbers, full stops, and other text necessities.

ASCII works okay for Latin script (English, etc.) but runs out of address bandwidth if a more complex language has to be described (for example, Japanese, with thousands of characters). Japanese, Chinese, Arabic, or Hebrew SMS can be realized using USC2 (Universal Multiple Octet Coded Character Set), a 16-bit/2-octet character string, or UCS4, a 32-bit/4-octet character string.

ASCII, UCS2, and UCS4 all allow perfectly acceptable representation of text on a grayscale LCD. However, we have said that we are beginning to see an increasing use of high-definition high color depth displays. These displays provide us with the capability to do text rendering by using pixel manipulation.

Pixel elements are made up of pels (picture elements) representing the singular red, green, or blue value of an RGB pixel. Remember that the number of bits used per pixel determines the amount of control you have over the color balance—24 bits gives you high color depth. The size of the image is the product of the number of pixels times the number of bits per pixel.

Text rendering is effectively subpixel manipulation, borrowing subpixels from adjacent whole pixels. The borrowed subpixels are always adjacent to their complementary color pixels, which our eyes mix to form white. We can therefore use subpixel manipulation to clean up jagged edges. Subpixel manipulation also only works on the horizontal resolution of LCDs. Even so, this means we can do the following:

Emboldening (stretching text horizontally)

Ke rning (shifting text horizontally, that is, micro-justification)

Italicizing (slanting type by skewing it horizontally)

Subpixel manipulation only works for LCDs, not CRTs. CRTs are not addressable at subpixel level, but then, as yet, no digital cellular handsets have CRT displays.

This means we can produce book-quality text on our screens, if we so desire. We must be aware, however, that not all LCDs have the same ordering of RGB subpixels. The rendering engine needs to know whether subpixels are arranged in forward or reverse order. Also, text rendering only works for landscape not portrait aspect displays, which means it is not really suitable for e-books, which would be an obvious application. Text rendering is now, however, included in a number of software products (Windows 2000 being one example) and will likely begin to appear further down the portable product food chain at a later date.

Image

Now that we have added beautifully rendered text to our wideband audio, it is time to add image bandwidth. An A4 image scanned at 300 dpi resolution and 24-bit color, however, produces a 24-Mbyte file—potentially a memory and delivery bandwidth embarrassment. As a result, we have a choice of lossless or lossy compression.

In *lossless compression*, all the data in the original image can be completely constructed in the receiver. Lossless compression is typically used in medical imaging, image archiving, or for images where any loss of information compromises application integrity. The problem with lossless compression is that it is hard to achieve compression rates of more than 2:1 or 3:1.

An example of a lossless compression technique used for storage system optimization is a dictionary-based scheme developed by Loughborough University and Actel, a memory product vendor. This compression technique has a learning capability and builds up a dictionary of previously sent data, which it shares with the receiver. If an exact match can be made, only the dictionary reference needs to be sent. If an exact match is not possible, the information is sent literally—that is, with no compression.

In *lossy compression*, we take the decision that a certain amount of information can be thrown away. The impact of discarding the information is either not noticeable or it is acceptable both to the person or device sending or storing the image or to the person or device receiving or storing the image. Compression ratios of 40:1 or higher are relatively easy to achieve with lossy compression. Compression schemes tend to be optimized either to improve storage bandwidth efficiency *or* delivery bandwidth efficiency, but not necessarily both.

Image compression standards are codified by the Joint Picture Experts Group, or JPEG. The Joint Bi-level Image experts Group (JBIG) looks after document compression,

document scanning, and optical character recognition (OCR). *Bi-level* means black and white, but the group also addresses grayscale compression. JPEG 2000 is the unified standard covering lossy and lossless compression and introduces the concept of Q factor.

A JPEG image is built up of a number of 8 x 8 pixel blocks that are transformed (like our audio codec) from the time to the frequency domain. The frequency content of the image is described by a string of digital coefficients. If one pixel block exactly matches the next, effectively, a "same again" message is sent. For example, endless blue sky would produce a whole series of identical pixel blocks. If a cloud appears, this changes the frequency content, and new digital coefficients need to be generated and sent—or perhaps not. We can choose to ignore the cloud, pretend it isn't there, and send a "same again" message, but some important information will have been left out.

A Q factor of 100 means any difference between pixel blocks is coded and sent. A Q of 90 means small block-to-block differences are ignored with some (hardly noticeable) loss of quality. A Q of 70 means larger block-to-block loss of quality, but it still is not very noticeable. In digital cameras, a Q of 90 equates to fine camera mode, and a Q of 70 equates to standard camera mode. We choose 70 when we want to fit more pictures into the memory stick or multimedia card. The choice of Q, however, also determines delivery bandwidth requirements.

As mentioned, the noticeability of quality degradation is also a product of the quality of display being used: A poor-quality display does not deserve a high Q picture; a good quality display is wasted if a poor Q is used.

Say we have a picture taken in fine camera mode (Q = 90), which creates a file size of 172,820 bytes. This will take 41.15 seconds to send over an uncoded 33.6 kbps channel (this is assuming the user data rate is the same as the channel rate with no forward error correction added in). If we took the same picture and had a Q of 5, the file size would reduce to 12,095 bytes and we could send it at the same channel rate in 2.87 seconds. The cost of delivery would be 15 times less for the Q-5 file. The question is, how much would the quality be impaired and how much value would be lost.

This highlights an important issue. Voice-quality metrics are well established. We use a mean opinion score to provide an objective way of comparing subjective quality assessments. For instance, we put 10 people or 100 people in a room and ask them to score a voice for quality, and then produce a mean opinion score (MOS) to describe the perceived quality. JPEG Q gives us an objective measurement of image quality, but we do not presently have a way of setting this against a subjective scorecard. As we will see later, the same problem occurs with video quality.

This is important when we come to negotiate network quality with a customer. In a 2G cellular network, we agree with a network operator to a certain bit error rate (typically 1 in 10^3). This is deemed acceptable and defines the coverage area in which the radio signal will be sufficient to deliver the defined BER or better. We can then show how this BER relates to voice quality and define the MOS achievable across a percentage of the coverage area.

Video

No such established relationship presently exists for image or video quality. We also need to consider that compression ratios increase as processor bandwidth increases.

As a rule of thumb, you can expect video compression ratios to increase by an order of magnitude every 5 years. In 1992, a data rate of 20 Mbps was required for broadcast-quality video. By 1997, this had reduced to 2 Mbps. However, as compression ratios increase, the quality of the source-coded material decreases. Digital TV provides an example, as shown in Table 7.1. A compression ratio of 100:1 yields VHS quality; a compression ratio of 10:1 yields high-definition TV.

Inconveniently, higher compression ratios also mean the data stream becomes more sensitive to errors and error distribution (burst errors) on the channel. These can be coded out by block coding, convolutional coding, and interleaving, but this introduces delay, and, of course, time is money.

If we take the historical trend forward, by 2007 we could have compression ratios of 500 to 1. These will work very well over low BER consistent physical channels—for example, an ADSL line specified at 1 in 10^{10} BER or optical fiber specified at 1 in 10^{12} BER (1 in 10,000,000,000,000 bits errored—effectively an errorless channel). These highly compressed media files will work less well over inconsistent, relatively high BER radio channels.

This brings us to the issue of differential encoding. In JPEG, we compare one pixel block with another and produce a difference figure. In MPEG, we do the same, but in addition, we look for similarities from image to image and express these as a difference coefficient. The problem with differential encoding is that it does not like delivery bandwidth discontinuity—for instance, burst errors on the radio channel or non-isochronous packets in the network.

The problem is partially overcome by using periodic refresh pictures. This is known as *intracoding*. The refresh pictures are only spatially, not temporally, compressed. Even using intracoding, differentially encoded video streams can be very jerky when sent over a wireless network (particularly, as we discuss later, over a wireless IP network). An alternative is to use JPEG for video. Individual still images become moving images by simple virtue of being sent at a suitable frame rate per second. JPEG does not use differencing and therefore avoids the problem, but it does not provide the same level of compression efficiency.

The better answer is to improve radio and network bandwidth quality. Better radio bandwidth quality means avoiding burst errors in the radio channel, better network bandwidth quality means avoiding transmission re-tries and minimizing delay and delay variability. This then allows the efficiency benefits of differential encoding to be realized.

Table 7.1 Compression versus Quality in Digital TV

COMPRESSION RATIO	CHANNEL RATE	RESOLUTION
10:1	20	High definition
20:1	10	Enhanced definition
40:1	5	PAL
100:1	2	VHS

Applying MPEG Standards

Which brings us to the MPEG standards. Existing MPEG codecs are relatively straightforward constant-rate block encoders. An MPEG-2 encoder, for example, takes a 16 x 16 pixel block (macroblock) and codes the motion differences on a block-by-block basis. In HDTV, a 1080-line picture has 1920 pixels per line subdivided down into macroblocks.

It will be a little while before we have high-definition digital TV in a handset. A typical digital TV decoder has nine or more parallel decoders running at 100 MHz producing 20 billion operations per second (BOPS) consuming 18 W of power! We are, however, beginning to see similar techniques being used, albeit on a more modest scale, in digital cellular video compression.

Video encoders today are typically constant rate. This makes them easier to manage over the physical and transport layer, but it means they are less efficient than if they were variable rate, that is, like the adaptive multirate vocoder or SMR vocoder. The SMR vocoder adapts to the dynamic range of the audio waveform. The same principle can apply to video encoders.

Consider that any source-coded content, whether audio or video, consists of entropy, unpredictable or novel material, and redundancy. An ideal compressor would separate out entropy and redundancy perfectly but would be infinitely complex and would have infinite processing delay. Entropy and redundancy ratios are constantly changing. Ideally, the video encoder rate would vary as the amount of entropy increases and decreases. A person jumping up and down will have high entropy (and a low coding rate); a person standing still will have low entropy (and a low coding rate).

A variable-rate video encoder would ideally be matched to a variable-rate radio layer and network layer physical channel. The objective from a user's perspective is to have constant quality**.** Consider as an example DVB/DVD (digital video broadcasting and digital video/versatile disc). In DVB/DVD a complex scene yields a fast encoding rate, a simple scene yields a slow encoding rate (see Figure 7.2).

In 3GPP1 it has generally been considered that variable-rate differential encoding was suboptimal for wireless because of the variability of the radio channel. Constant-rate coding schemes not using differencing, such as H320, were considered to be more suitable. However, as we discussed in Chapter 1, the idea of a 3G 5 MHz channel is to use power control to track out the fast fading—turning our variable quality channel into a constant-quality channel (see Figure 7.3).

We can move from constant-rate variable-quality bandwidth to variable-rate constant-quality bandwidth, but this has to include both radio and network bandwidth consistency. We would argue this points the way toward future MPEG-4 evolution.

The Motion Picture Expert Group (MPEG) was founded in 1993. This makes it young in terms of telecom standards and old in terms of Internet standards. It was originally focused on producing a standard for noninteractive (simplex) video compression but was extended, as MPEG-4 and MPEG-5, to include the manipulation, management, and multiplexing of multimedia content. MPEG proposals tend to be initiated by the broadcast or content producing industry but end up as ISO standards and ITU recommendations. They start in a different place than telecom standards but end up at the same place.

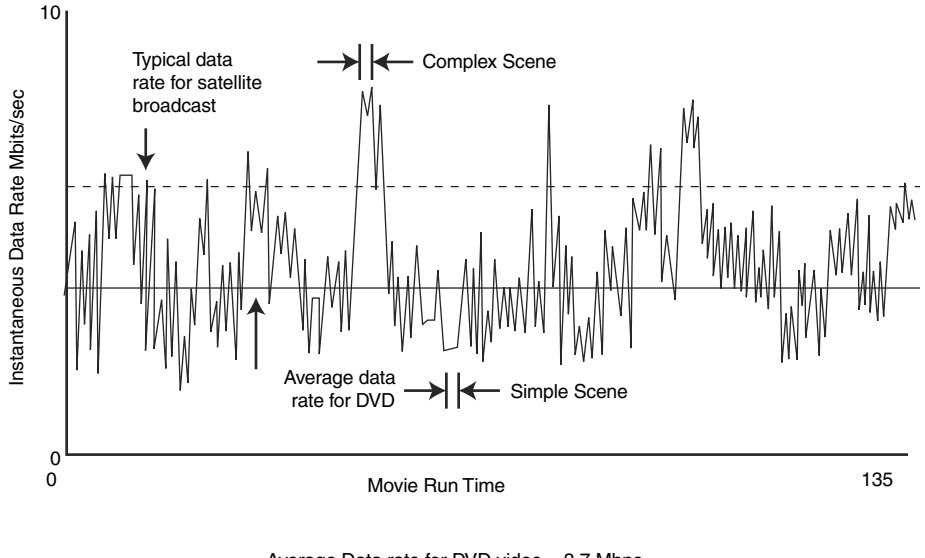


Figure 7.2 Variable-rate encoding.

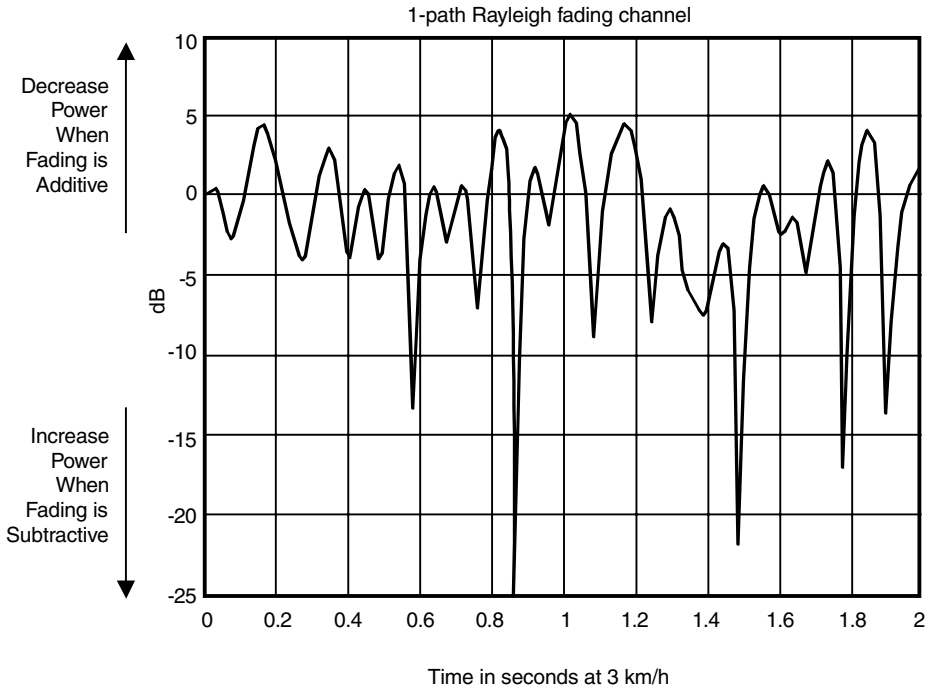


Figure 7.3 Using fast power control to follow the fading channel.

Table 7.2 summarizes the current and in progress MPEG standards. MPEG-1 covers CD-ROM storage, MPEG-2 covers DVB and DVD, MPEG-2—Layer 3 (unofficially but widely known as MPEG-3) covers audio streaming, MPEG-4 adds video streaming (and quite a lot else), MPEG-5 covers multiple viewing angles, MPEG-7 addresses content identification, and MPEG-21 defines—or will define—network quality requirements, content quality, and conditional access rights. MPEG-21 is described as a "multimedia umbrella standard."

The main purpose of MPEG-3 is to improve storage compression efficiency—although, as a consequence, it also reduces delivery bandwidth requirements. An uncompressed 5-minute song creates a 50-Mbyte file that is compressed down to a 5-Mbyte MPEG-3 file. MPEG-3 is a sub-band compression technique; dividing audio bandwidth into 32 sub-bands that are each separately encoded. It helps fit an hour of MPEG-3 music onto a 64-Mbyte memory card or (back to our hard disk!) 150 CDs on an 8-Gbyte hard disk.

MPEG-4 adds video to produce a combined audio/video encoding/decoding standard. In Chapter 4 we describe MPEG-4 as presently implemented—that is, a block coding scheme in which a discrete cosine transform takes time domain information into the frequency domain to exploit macroblock-by-macroblock and image-to-image redundancy. The DCT is precisely prescribed in the standard, as are the multiplexing of the audio and video streams. Other processing tasks are vendor-specific—for example, the preprocessing, motion estimation, compensation and rate control in the encoder, error control and error concealment, and post-processing in the decoder (the implementation of coding noise reduction).

This vendor differentiation is probably not good news for network designers needing to deliver a consistent user experience, as this is going to vary between codecs—particularly when one vendor's codec needs to talk to another vendor's codec. Realistically this will have to be resolved by the vendors. At present, most of the proprietary solutions are constant-rate variable-quality.

Table 7.2 MPEG Standards

MPEG-1	CD-ROM storage compression standard
MPEG-2	DVB and DVD compression standard
MPEG-3	MPEG-2—Layer 3 (MPEG-3) audio streaming standard
MPEG-4	Audio and video streaming and complex media manipulation
MHEG-5	Multimedia hypermedia standard (MPEG-4 for set-top boxes)
MPEG-7	Standard for content identification
MPEG-21	Network quality, content quality, conditional access rights (multimedia umbrella standard)

Object-Based Variable-Rate Encoders/Decoders

The longer-term interest in MPEG-4, however, is the development of object-based variable-rate encoders/decoders. The objective is to deliver variable-rate constant-quality encoding/decoding. What do we mean by object coding? MPEG-4 Version II December 1999 (in parallel with 3GPP Release 99) described a standardized way of moving media objects within a coordinate system. You can have audio objects or video objects. Video images are split into component parts—for instance, a person, a chair, and a table. A table is a primitive object. A completely still person is a primitive object. A person dancing on a table is a complex object and will incur a faster coding rate.

Because MPEG-4 describes the coordinates within which an object moves, we can standardize motion estimation, motion prediction, and motion compensation techniques. An object moving across a background only changes if it deforms, moves into shadow, or rotates. We can predict the axis and direction of movement of an object and reconstruct the movement as a rendering instruction in the decoder. The direction of travel is known as the *optic flow axis*. This means that objects can be manipulated on arrival: We can translate, warp, or zoom objects, we can use transforms (processing algorithms) to change the geometric or acoustical properties of objects, and we can turn audio objects into (three-dimensional) surround sound. (We may not want to do this, but it's nice to know that we can.)

Thus, in the same way that we can render text in the decoder, we can render audio and video objects. The technique is sometimes described as *mesh coding* and borrows memory processing and algorithm prediction technology from the game console software development world. What we are trying to achieve is an increase in the apparent bandwidth available to us in the handset; we can send a small amount of information to and from the handset but turn it into an (apparently) large amount of information by using local processor bandwidth to render and post-process the content.

For example, we might choose to store a generic face in the handset. The encoder has to encode a face, but in practice it only encodes the differences between the face it is seeing (the image stream from the CMOS imaging platform) and the reference face in the encoder (which is the same as the generic reference face in the decoder). The generic face will also be expressionless, so the encoder needs to send difference *and* animation parameters.

The ability to manage objects within a coordinate system also means we can provide motion compensation. Motion compensation can be used to code out camera shake. The problem with camera shake is that it increases entropy. The codec perceives a rapidly shaking image and tries to encode the movement. Motion compensation can cancel out the movement prior to encoding—and therefore reduce the encoder rate *and* improve the quality of the video.

Virtual Reality Modeling Language

MPEG-4 covers some other interesting areas, one of which is the longer-term standardization of meta description using a description syntax known as Virtual Reality Modeling Language (VRML). Meta data is usually described as information about

information. It provides us with a standardized way of describing information such that we can archive it and find it again at some (possibly distant) time in the future (back to our Doomsday project!).

The meta description includes the QoS requirements of the media file; that is, this is *declarative content*—content that defines and describes its radio bandwidth and network bandwidth quality requirements. The quality of service metrics include the following:

- Whether or not the packet stream needs to be isochronous. In an isochronous packet stream, all packets arrive in the same order they were sent. In a non-isochronous packet stream, they do not.
- The buffer and timing requirements—that is, how much buffering will be needed by the complex media file. Table 7.3 shows the buffer size requirements for what are called simple MPEG-4 profiles.

The buffer size expands as the frame size increases (from QCIF to CIF) and as the frame rate increases.

- MPEG-4 also describes how elementary streams from a complex content stream are linked to a complex transport channel. This is very fundamental. In Chapter 3 we described how the OVSF codes are structured on the radio channel downlink and uplink—our complex radio bandwidth transport channel. We need to take these complex composite streams (consisting of up to six elementary streams per user) and preserve their properties, including time interdependencies, as the streams move across the radio layer and into the core network. This is, as we will see in later chapters, absolutely crucial to delivering consistent end-to-end performance in a wireless IP network.
- To help maintain complex-content multiple-stream synchronization, MPEG-4 adds optional timestamping to each elementary stream. This can either allow isochronous packet streams to be reclocked in a receiver or non-isochronous streams to be individually reconstituted, reordered, and reclocked.
- MPEG-4 also supports the defining of buffer size to allow non-real-time data to be sent ahead of a real time exchange—for example, the preloading of a PowerPoint presentation or financial spreadsheet.

Table 7.3 Requirements of MPEG-4 Simple Profile

Level 1	QCIF	15 fps	64 kbps	256 bytes	10240 bytes
Level 2	CIF	15 fps	128 kbps	512 bytes	40960 bytes
Level 3	CIF	30 fps	384 kbps	1024 bytes	40960 bytes

The MPEG-4 encoder is effectively dictating how many per-user channel streams are needed at the beginning of a session, how many per-user channel streams need to be added or removed as the session progresses, and the data rate required on any one of the individual per-user channel streams. This will need to be integrated either with IP session management protocols (such as SIP, which we case study later in the book) or with existing circuit-switched SS7-based session management signaling.

Automated Image Search Engines

We mentioned that MPEG-4 codifies meta descriptions so that complex content can be archived. This work is carried forward into MPEG-7 to provide support for automated image search engines (equivalent to word search engines). Remember that we are performing a discrete cosine transform on each macroblock within an image. We are therefore expressing the spectral content and frequency content (that is, color) of each macroblock in terms of a series of digital filter coefficients.

MPEG-7 exploits this process to produce a standard for automated content description—and hence automated content searching. All images are converted into a common unified format in which image features are identified based on the wavelength of colors making up the scene. The frequency content is described as a 63-bit descriptor.

Now consider this: A medium-sized town has, say, 20 surveillance cameras that are taking pictures every second or so. Those pictures are being stored in a database. The police want to look for someone in a red woolly hat who walked across the right-hand topmost macroblock of camera number 20 at some time during the past 6 months. Previously this would involve several police officers looking through endless video archive footage. The meta descriptor automates this process—just search for red in macroblock x and wait for the results.

MPEG-7 makes wireless-enabled video surveillance far more powerful, because it simplifies the image archive search and retrieval process. The bandwidth uploading from surveillance cameras is generally non-time-sensitive and can occupy the long low-load night hours. Color depth is also important in these applications—the fact that the suspect was wearing a red woolly hat. Contrast ratio is also important.

Digital Watermarking

Which brings us to MPEG-21, our multimedia umbrella standard. There is no point in using a compressed digital image in court if it is not admissible evidence. Digital images can be challenged on the basis that they may have been altered between point of capture (the video surveillance device) and point of presentation (the court).

MPEG-4 started to address the codification of ownership rights and proof of ownership and proof of provenance. The technique is sometimes known as *digital watermarking*—the digital countersigning of an image such that it can be demonstrated that the image was produced by a specific person or device and has not been altered prior to or during storage or delivery, that is, the whole process of authentication. Digital watermarking can be used to provide an audit trail showing the path an image has taken and what has happened to the image.

You must be careful that the compression used does not destroy the digital watermark; it is always a good idea to compress and then watermark a digital image data stream.

The SMS to EMS to MMS Transition

There are over 50 million text messages sent every day in the United Kingdom. An SMS message can be up to 160 characters long. Because it uses the ASCII 7-bit alphabet, this means the maximum message length is $160 \times 7 = 1120$ bits (equivalent to 140 bytes of binary data, $140 \times 8 = 1120$).

SMS is sent in GSM over a traffic channel known as TCH8, running at 80 octets per second—that is, 640bps. A 160-character SMS is 1120 bits long, so it takes 1.75 seconds to send (1120×640). SMS is therefore delivered over a very low rate channel, but because there are not a lot of bits involved, it doesn't take long to send them. It is also a very robust channel (very heavily forward error corrected), so an SMS message will get through when voice will not. So it's good news all around. There are not many bits, so it's easy to send; it doesn't occupy much transmission time or transmission energy.

From a billing perspective, if a user can be charged 10 cents to send or receive an SMS, then the network operator has obtained equivalent revenue (from under 2 seconds of network and radio bandwidth) to a 1-minute phone call. The network margin achievable from SMS, therefore, is considerably higher than the network margin achievable from voice.

SMS has been less pervasive in the United States partly because of internetwork technical compatibility issues. (IS95 CDMA uses a 225-character SMS format, and IS136 TDMA uses a 256-character format.) European and Asian operators, however, are keen to develop the SMS model to include rich media exchange. Some operators have 20 percent of their network margin being generated by SMS traffic. If additional bandwidth can be generated, data would rapidly overtake voice as the main source of network margin.

It's important to recognize the differentiation between average revenue per user (ARPU) and averaging margin per user (AMPU). SMS is economic in terms of radio and network bandwidth utilization but has high-perceived value and, therefore, has relatively higher billing value (whereas voice added value continues to decline). As we also identified earlier, SMS does help to trigger additional voice loading on the network and, just as important, effectively generates a new evening busy hour, increasing network utilization.

Enhanced Messaging Service (EMS) adds picture messaging to the SMS service platform. Picture messages can either be 16 x 16 pixels, 32 x 32 pixels, or 96 x 64 pixels, that is, quite simple low-resolution images. Similarly, animations can be 8 x 8 pixels and

16 x 16 pixels. Multimedia messaging will add the MPEG-4 simple profiles (QCIF and CIF 15 and 30 frames per second) to the SMS platform, at which point the SMS transport layer becomes rather overloaded and rather slow. Since SMS, EMS, and MMS are all intended as store-and-forward services, however, delay (storage bandwidth) can be traded off against delivery bandwidth constraints, but storage bandwidth (store and forward) needs to be factored in as a network cost.

Over a 3 to 5 year period, we argue that SMS, EMS, and MMS will become part of the overall user session mix. In this scenario, the application layer actively manages the physical radio layer multiplex, supporting low-bandwidth data streams (SMS and EMS) on SF256 OVSF code physical channels and MMS and rich media streams on higher user rate (lower spreading factor) code streams. In addition, OVSF code allocation is integrated with RNC-based admission control software platforms (which we cover in a later chapter).

SMS, EMS, and MMS do, however, bring us to a more comprehensive discussion of quality metrics, as perceived by the user.

Quality Metrics

A user is not interested in bit error rates, frame erasure rates, or packet loss. He or she is interested in voice quality and image quality. We can define quality in terms of audible and visible properties, which can be directly experienced and judged by the user. Audio quality metrics are well established and already widely measured, but now we are adding value by *simultaneously* multiplexing images, video, and application data (see Figure 7.4).

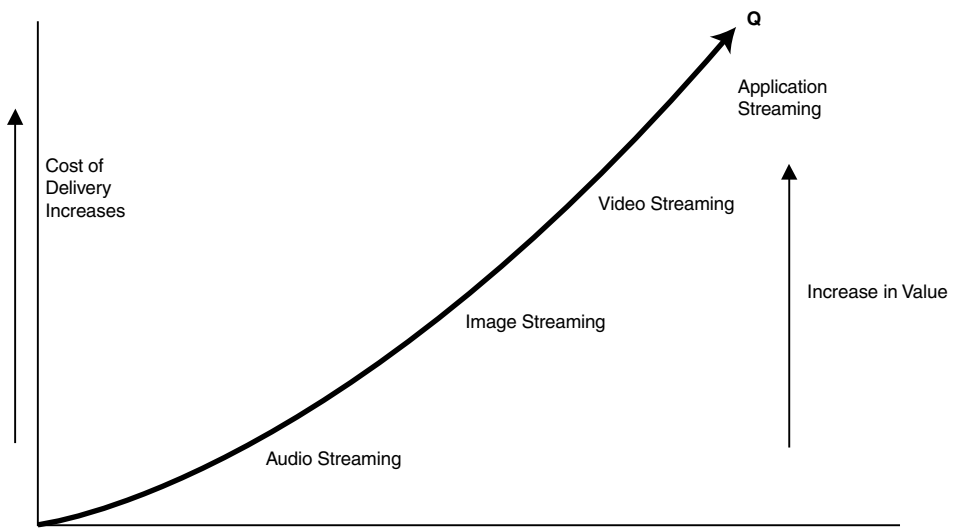


Figure 7.4 Media multiplex (multiplex complexity).

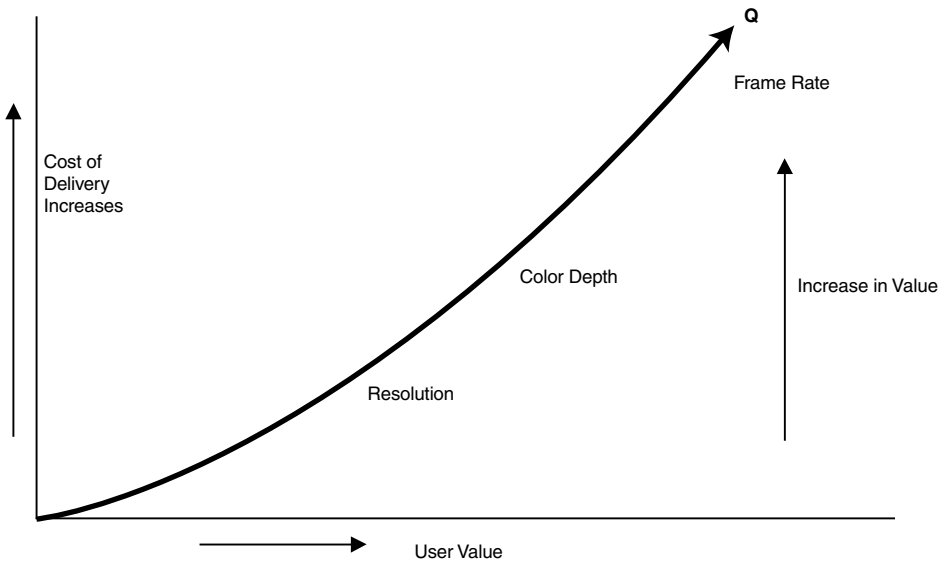


Figure 7.5 Objective quality (resolution, color depth, and frame rate).

Just as we judge audio in terms of fidelity; we can judge image and video streaming in terms of color depth, frame rate, resolution, and contrast ratio; and application quality in terms of application integrity. Figure 7.5 shows the main quality metrics of video. The quality metrics for an image are the same—without the frame rate.

As we increase the complexity of the media multiplex, value increases (we hope), but so does the cost of delivery. Hopefully, the value increases faster than the cost of delivery; otherwise, the whole exercise is rather pointless from a business point of view. Figure 7.6 shows value increasing as delay (and delay variability) increases. We cover this in much more detail in Chapter 11, in a discussion on network bandwidth quality.

Effectively, as we increase delay and delay variability, our cost of delivery reduces and our margin per user should increase, provided the delay and delay variability have not destroyed the value of the user's content, which may or may not be time-sensitive. Remember, we are replacing a user experience (PSTN) where end-to-end delay is typically 35 ms with no end-to-end delay variability. It is always dangerous to assume a user will not notice a reduction in service quality.

Finally, there is the consistency metric (see Figure 7.7). In 1992 when GSM was introduced, the voice quality from the codec was (a) not very good and (b) not very consistent. This was due to a number of factors—codec design, marginal sensitivity in the handset and base station, and insufficient network density (a marginal link budget). It was not until 1995 that voice quality both improved *and* became consistent.

Interestingly, though anecdotally, we are often very forgiving of poor quality provided the quality is consistent. If something is inconsistent, we remember the bad bits. The same applies to video quality in 3G networks. It will take at least 5 years for video quality to be acceptable both in terms of quality (frame rate, color depth, resolution) and consistency. Consistency requires good control of radio bandwidth impairments and irregularities (that is, slow and fast fading) and network bandwidth impairments and irregularities (delay, delay variability, and packet loss).

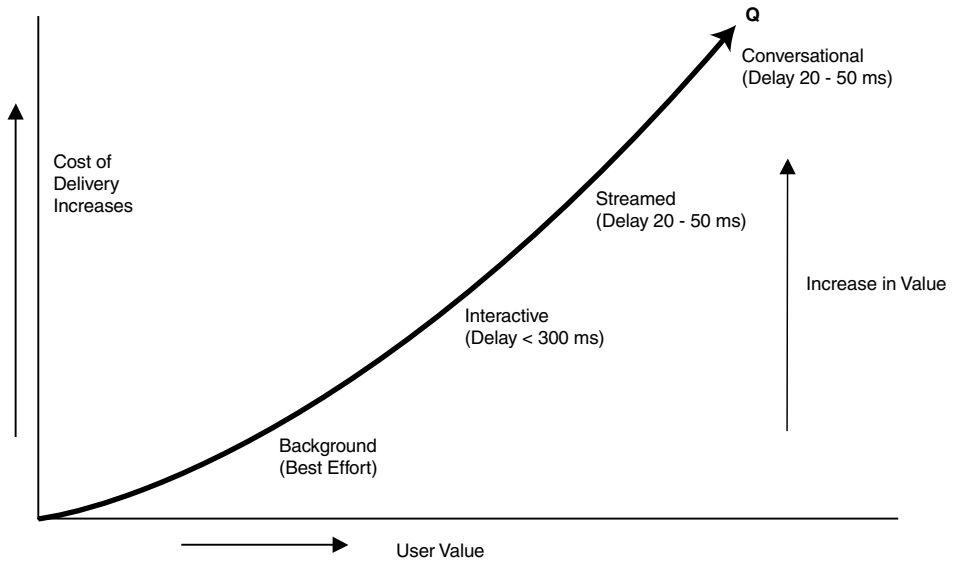


Figure 7.6 Objective quality—the cost of reducing delay and delay variability.

Hopefully, we are beginning to make clear the intimate relationship between radio and bandwidth quality and an acceptable (i.e., billable) user experience.

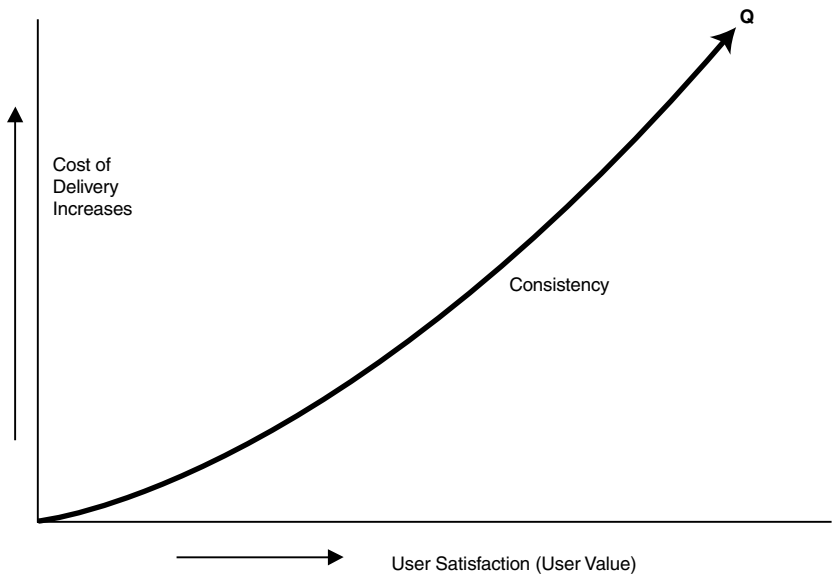


Figure 7.7 Consistency metric.

Summary

We highlighted the transition from constant-rate source coding to variable-rate source coding, both for audio and video capture, and the related significance of MPEG standards evolution, particularly in the longer-term object-based coding technique and rendering engines. We showed how processing in the handset can create the illusion of bandwidth and interactivity, and how preprocessing and post-processing can reduce the amount of radio and network bandwidth needed (including RF power) for an apparently wide-bandwidth application.

The argument was put forward that we should use tangible (easily evident to the user) quality metrics to judge radio and network bandwidth performance and to provide the mechanism for implementing quality-based rather than quantity-based billing. MPEG-4 is generally regarded as a compression standard, but in reality, MPEG also helps us define what network quality requirements are needed to preserve rich media value.

Application layer software needs to evolve within this context. The job of application layer software is to increase session persistency and session complexity, as shown in Figure 7.8.

User value (and user billability) increases as session persistency increases. As session persistency increases, session complexity generally should also increase. A simple data exchange is developed into a data plus voice and video exchange, or a simple voice exchange is developed into a voice and data and video exchange, or a user-to-user exchange is developed into a multiuser-to-multiuser exchange. As session persistency increases, consistency also has to increase. The longer the session, the more obvious it becomes when radio or network bandwidth constraints cause discontinuities in the duplex transfer of real-time rich media information.

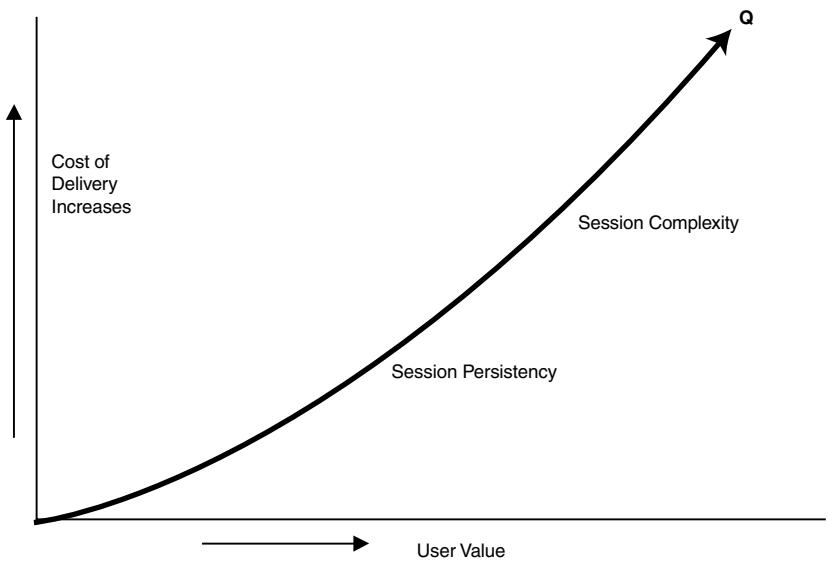


Figure 7.8 Session persistency.

Consistency is often underrated as a quality metric. Consistently poor quality is often perceived as being better than inconsistently good quality. We adjust (and learn to live with) consistent quality, even if the standard is relatively poor. Consistency is a product of protocol performance. Don't send "same again" differentially encoded image and video streams over "send again" channels.

Additionally, software performance is a key element in our overall user quality metric (user Q), which brings us to our next chapter.

CHAPTER

8

MExE-Based QoS

We have identified delay and delay variability as two components that add or subtract from user value. Some of our rich media mix is delay-tolerant but not all of it, and the part of the media mix that is the *most* sensitive to delay and delay variability tends to represent the highest value. Application streaming is part of our rich media mix—Application streaming implies that we have one or more applications running in parallel to our audio and video exchange.

In this chapter, we consider how handset software performance influences the quality of the end-to-end user experience.

An Overview of Software Component Value

Delay and delay variability and the ability to multitask are key elements of software component value. The job of an operating system is to sit between the software resident in the device and the device hardware (see Figure 8.1). Going back in history, an operating system such as Microsoft DOS (Disk Operating System) reads physical memory (hardware) in order to open a file (to be processed by software). In a PDA or 3G wireless handset, the operating systems to date have typically been ROM-based products. This is changing, however, because of the need to support remote reconfigurability and dynamic application downloads.

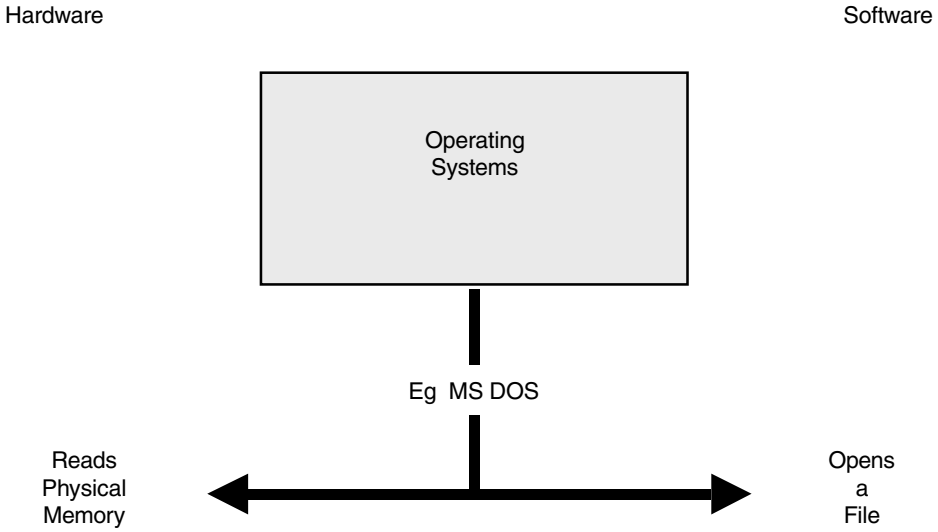


Figure 8.1 Software partitioning.

Applications are sometimes described as *embedded* or even *deeply embedded*. The more deeply embedded the application, the more remote it is from the outside world. In other words, the more deeply embedded the application, the more deterministic it becomes—that is, it performs predictably. As the application is moved closer to the real world, it has to become more flexible, and as a result, it becomes less predictable in terms of its overall behavior.

An example of embedded software is the driver software for multiple hardware elements—memory, printers, and LCDs. There are a number of well-defined repetitive tasks to be performed that can be performed within very closely defined time scales. Moving closer to the user exposes the software to unpredictable events such as keystroke interrupts, sudden unpredictable changes in the traffic mix, and sudden changes in prioritization.

Defining Some Terms

Before we move on, let's define some terms:

Protocols. Protocols are the rules used by different layers in a protocol stack to talk to and negotiate with one another.

Protocol stack. The protocol stack is the list of protocols used in the system. The higher up you are in the protocol stack, the more likely you are to be using software—partly because of the need for flexibility, partly because speed of execution is less critical. Things tend to need to move faster as you move down the protocol stack.

Peers. The machines in each layer are described as peers.

Entities. Peers are entities, self-contained objects that can talk to each other. Entities are active elements that can be hardware or software.

Network architecture. A set of layers and protocols make up a network architecture.

Real-time operating system. We have already, rather loosely, used the term real-time operating system (RTOS). What do we mean by real time? The IEEE definition of a real-time operating system is a system that responds to external asynchronous events in a predictable amount of time. Real time, therefore, does not mean instantaneous real time but predictable real time.

Operating System Performance Metrics

In software processing, systems are subdivided into processes, tasks, or threads. Examples of well-established operating systems used, for example, to control the protocol stack in a cellular phone are the OS9 operating system from Microware (now RadiSys). However, as we discussed in an earlier chapter, we might also have an RTOS for the DSP, a separate RTOS for the microcontroller, an RTOS for memory management, as well as an RTOS for the protocol stack and man-machine interface (MMI) (which may or may not be the same as the microcontroller RTOS). These various standalone processors influence each other and need to talk to each other. They communicate via mailboxes, semaphores, event flags, pipes, and alarms.

Performance metrics in an operating system include the following:

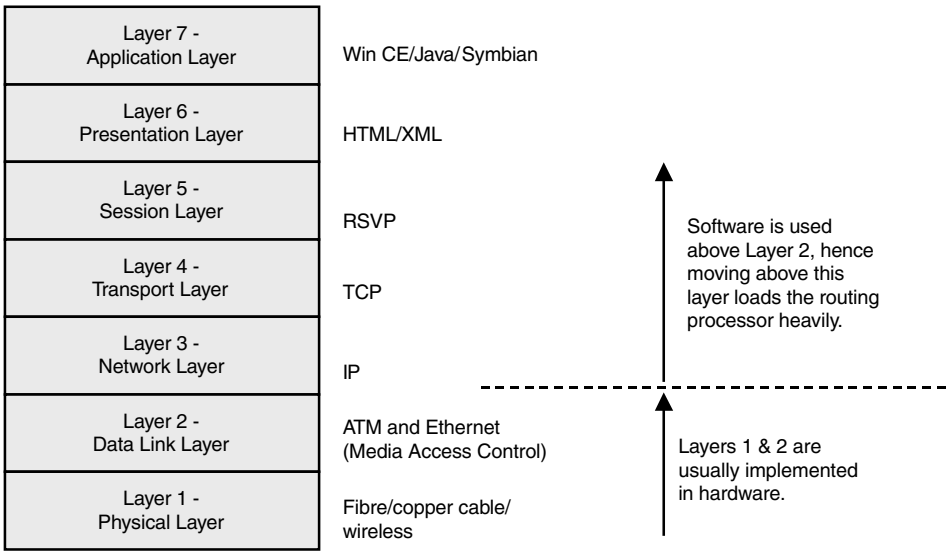
Context switching. The time taken to save the context of one task (registers and stack pointers) and load the context of another task

Interrupt latency. The amount of time between an interrupt being flagged and the first line of the code being produced (in response to the interrupt), including completion of the initial instruction

As a rule of thumb, if an OS is ROM-based, the OS is generally more compact, less vulnerable to virus infection, and more efficient (and probably more predictable). Psion EPOC, the basis for Symbian products, is an example. The cost is flexibility, which means, the better the real-time performance, the less flexible the OS will be. Response times of an RTOS should generally be better than (that is, less than) 50 μs.

The OSI Layer Model

We can qualify response times and flexibility using the OSI layer model, shown in Figure 8.2. The OSI model (Open System Interconnection, also known as Open Standard Interconnection) was developed by ISO (International Standards Organization) in 1984 as a standard for data networking. The growth of distributed computing—that is, computers networked together—has, however, made the OSI model increasingly relevant as a means for assessing software performance.



*Open Systems Interconnection.

Figure 8.2 Software performance—the OSI reference model.

At the Application layer, the software has to be very flexible. This means it must be able to respond to a wide range of unpredictable user requests. Although the execution of particular tasks may be real time (as defined earlier in the chapter), the tasks vary in complexity and may require a number of interactive processes (semaphores and alarms, for example) before the task can be considered as completed (a high degree of interrupt latency).

As we move down the protocol stack, task execution becomes more deterministic. For example, at the Physical layer, things are happening within very precise and predictable time scales. Returning to our reference model, following are descriptions of each protocol layer and the types of tasks performed:

Application layer (7). Windows CE, Symbian, or Java, for example, will look after the user (the MMI and the housekeeping tasks associated with meeting the user's requirement), organizing display drivers, cursors, file transfer, naming conventions, e-mail, diary, and directory management.

Presentation layer (6). This layer does work that is sufficiently repetitive to justify a general solution—page layout, syntax and semantics, and data management. HTML and XML are examples of Presentation layer protocols (for Web page management).

Session layer (5). This layer organizes session conversations: The way our dialogue is set up, maintained, and closed down. Session maintenance includes recovery after a system crash or session failure—for example, determining how much data needs resending. RSVP and SIP (covered later) are examples of Session layer protocols.

Transport layer (4). This layer organizes end-to-end streaming and manages data from the Session layer, including segmentation of packets for the Network layer. The Transport layer helps to set up end-to-end paths through the network. Transmission Control Protocol (TCP) is usually regarded as a Session layer protocol. (SIP is also sometimes regarded as a Session layer protocol.)

Network layer (3). This layer looks after the end-to-end paths requested by the Transport layer, manages congestion control, produces billing data, and resolves addressing conflicts. Internet Protocol (IP) is usually considered as being a Network layer protocol.

Data Link layer (2). This layer takes data (the packet stream) and organizes it into data frames and acknowledgment frames, checks how much buffer space a receiver has, and integrates flow control and error management. ATM, Ethernet, and the GSM MAC (Media Access Control) protocols are all working at the Data Link layer. GSM MAC, for example, looks after resource management—that is, the radio bandwidth requirements needed from the Physical layer.

Physical layer (1). This layer can be wireless, infrared, twisted copper pair, coaxial, or fiber. The Physical layer is the fulfillment layer—transmitting raw bits over a wireless or wireline communication channel.

Any two devices can communicate, provided they have at least the bottom three layers (see Figure 8.3). The more layers a device has, the more sophisticated—and potentially the more valuable—it becomes.

As an example, Cisco started as a hardware company; most of its products (for example, routers and switches) were in Layer 3 or 4. It acquired ArrowPoint, IPmobile, Netiverse, SightPath, and PixStream—moving into higher layers to offer end-to-end solutions (top three layers software, bottom four layers hardware and software). Cisco has now started adding hardware accelerators into routers, however, to achieve acceptable performance.

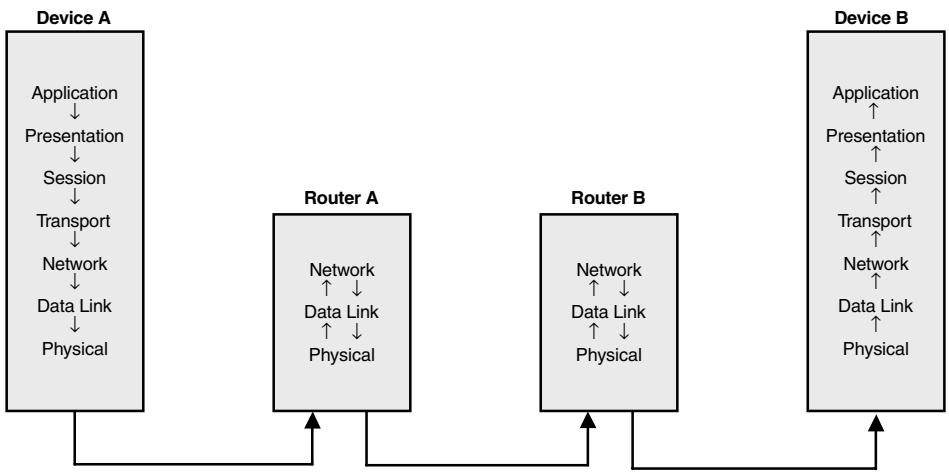


Figure 8.3 Data flow can be vertical and horizontal.

There is, therefore, no hard-and-fast rule as to whether software or hardware is used in individual layers; it's just a general rule of thumb that software is used higher up in the stack and hardware further down. As we will see in our later chapters on network hardware, the performance of Internet protocols, when presented with highly asynchronous time-sensitive multiple traffic streams, is a key issue in network performance optimization. We have to ensure the protocols do not destroy the value (including time-dependent value) generated by the Application layer; the protocols must preserve the properties of the offered traffic.

MExE Quality of Service Standards

So the software at the Application layer in the sender's device has to talk to the software at the Application layer in the receiver's device. To do this, it must use protocols to move through the intermediate layers and must interact with hardware—certainly at Layer 1, very likely at Layer 2, probably at Layer 3, and possibly at Layer 4.

MExE (the Mobile Execution Group) is a standards group within 3GPP1 working on software/hardware standardization. Its purpose is to ensure that the end-to-end communication, described here, actually works, with a reasonable amount of predictability. MExE is supposed to provide a standardized way of describing software and hardware form factor and functionality, how bandwidth on demand is allocated, and how multiple users (each possibly with multiple channel streams) are multiplexed onto either individual traffic channels or shared packet channels.

MExE also sets out to address algorithms for contention resolution (for example, two people each wanting the same bandwidth at the same time with equal priority access rights) and the scheduling and prioritizing of traffic based on negotiated quality of service rights (defined in a service level agreement). The quality of service profile subscriber parameters are held in the network operator's home location register (the register used to support subscribers logged on to the network) with a copy also held in the USIM in the subscriber's handset. The profiles include hardware and software form factor and functionality, as shown in the following list:

Class mark *hardware* description. Covers the vendor and model of handset and the hardware form factor—screen size and screen resolution, display driver capability, color depth, audio inputs, and keyboard inputs.

Class mark *software* description. Covers the operating system or systems, whether or not the handset supports Java-based Web browsers (the ability to upload and download Java applets), and whether the handset has a Java Virtual Machine (to make Java byte code instructions run faster).

Predictably, the result will be thousands of different hardware and software form factors. This is reflected in the address bandwidth needed. The original class mark (Class Mark 1) used in early GSM handsets was a 2-octet (16-bit) descriptor. The class mark presently used in GSM can be anything between 2 and 5 octets (16 to 40 bits). Class Mark 3 as standardized in 3GPP1 is a 14-octet descriptor (112 bits). Class Mark 3 is also known as the Radio Access Network (RAN) class mark. The RAN class mark includes FDD/TDD capability, encryption and authentication, intersystem measurement capability, positioning capability, and whether the device supports UCS2 and

UCS4. This means that it covers both the source coding, including MPEG4 encoding/decoding, and channel coding capability of the handset. Capability includes such factors as how many simultaneous downlink channel streams are supportable, how many uplink channel streams are supportable, maximum uplink and downlink bit rate, and the dynamic range of the handset—minimum and maximum bit rates supportable on a frame-by-frame basis.

MExE is effectively an evolution from existing work done by the Wireless Application Protocol (WAP) standard groups within 3GPP1.

Maintaining Content Value

Our whole premise in this book so far has been to see how we can capture rich media and then preserve the properties of the rich media as the product is moved into and through a network for delivery to another subscriber's device. There should be no need to change the content. If the content is changed, it will be devalued. If we take a 30 frame per second video stream and reduce it to 15 frames a second, it will be devalued. If we take a 24-bit color depth image and reduce it to a 16-bit color depth image, it will be devalued. If we take a CIF image and reduce it to a QCIF image, it will be devalued. If we take wideband audio and reduce it to narrowband audio, it will be devalued.

There are two choices:

- You take content and adapt it (castrate it) so that it can be delivered and displayed on a display-constrained device.
- You take content, leave it completely intact (preserve its value), and adapt the radio and network bandwidth and handset hardware and software to ensure the properties of the content are preserved.

WAP is all about delivering the first choice—putting a large filter between the content and the consumer to try and hide the inadequacies of the radio layer, network, or subscriber product platform. Unsurprisingly, the result is a deeply disappointing experience for the user. Additionally, the idea of having thousands of devices hardware and software form factors is really completely unworkable. The only way two dissimilar devices can communicate is by going through an insupportably complex process of device discovery.

Suppose, for example, that a user walks into a room with a Bluetooth-enabled 3G cellular handset, and the handset decides to use Bluetooth to discover what other compatible devices there are in the room. This involves a lengthy process of interrogation. The Bluetooth-enabled photocopier in the corner is particularly anxious to tell the 3G handset all about its latest hardware and software capability. The other devices in the room don't want to talk at all and refuse to be authenticated. The result is an *inconsistent* user experience and, as we said earlier, an *inconsistent* user experience is invariably perceived as a *poor-quality* user experience.

Pragmatically, the exchange of complex content and the preservation of rich media product properties delivered consistently across a broad range of applications can and will only be achieved when and if there is one completely dominant de facto standard handset with a de facto standard hardware and software footprint. Whichever vendor or vendor group achieves this will dominate next-generation network-added value.

There are two golden rules:

Do not destroy content value. If you are having to resize or reduce content and as a result are reducing the value of the content, then you are destroying *network value*.

Avoid device diversification. Thousands of different device hardware and software form factors just isn't going to work—either the hardware will fail to communicate (different flavors of 3G phones failing to talk to each other) or the software will fail to communicate (a Java/ActiveX conflict for example).

Experience to date reinforces the "don't meddle with content" message.

Network Factors

Let's look at WAP as an example of the demerits of unnecessary and unneeded mediation. Figure 8.4 describes some of the network components in a present GSM network with a wireless LAN access point supporting Dynamic Host Configuration Protocol, or DHCP (the ability to configure and reconfigure IPv4 addresses), a Web server, a router, and a firewall. The radio bearers shown are either existing GSM, high-speed circuit-switched data (HSCSD)—circuit-switched GSM but using multiple time slots per user on the radio physical layer— or GPRS/EDGE.

The WAP gateway is then added. This takes all the rich content from the Web server and strips out all the good bits—color graphics, video clips, or anything remotely difficult to deal with. The castrated content is then sent on for forward delivery via a billing system that makes sure users are billed for having their content destroyed. The content is then moved out to the base station for delivery to the handset.

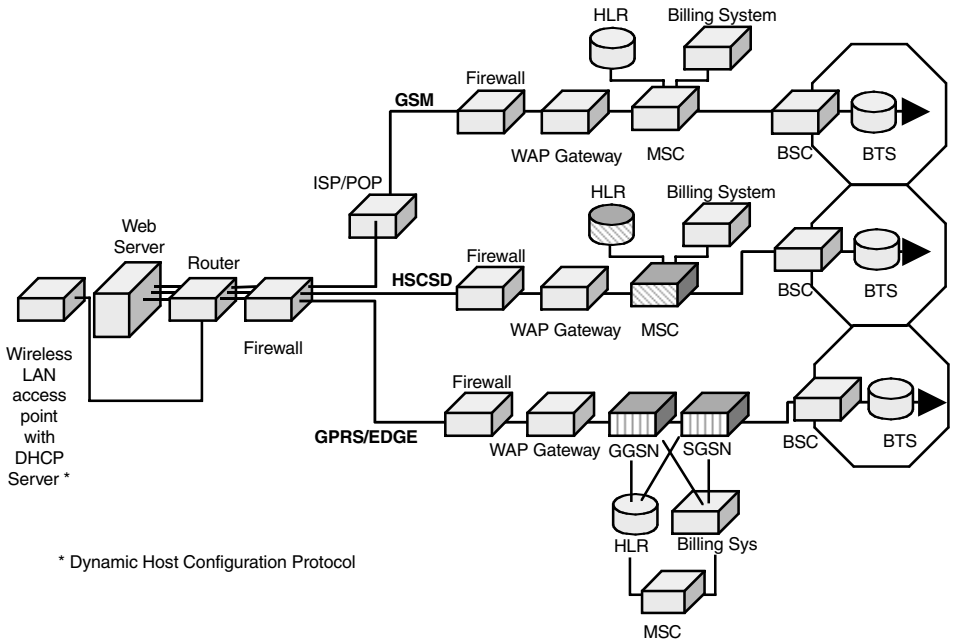


Figure 8.4 GSM Network Components.

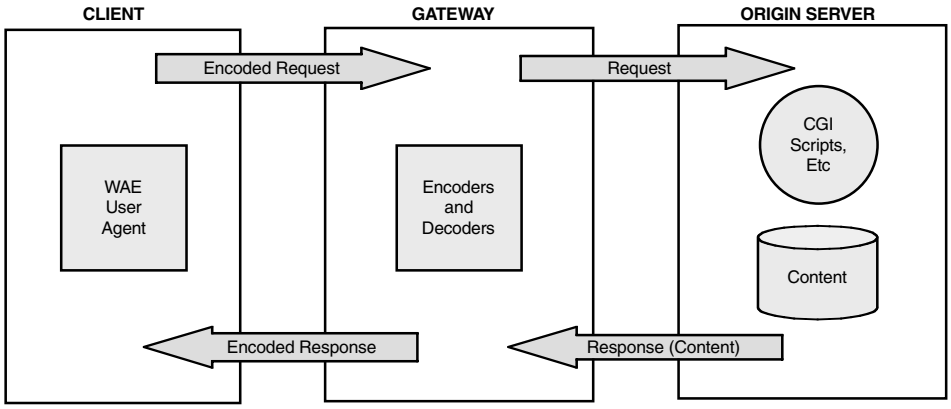


Figure 8.5 WAP gateway.

Figure 8.4 illustrates what is essentially a downlink flow diagram. It assumes that future network value is downlink-biased (a notion with which we disagree). However, the WAP gateway could also compromise subscriber-generated content traveling in the uplink direction.

Figure 8.5 shows the WAP-based client/server relationship and the transcoding gateway, which is a content stripping gateway not a content compression gateway (which would be quite justifiable).

Figure 8.6 shows the WAP structure within the OSI seven-layer model—with the addition of a Transaction and Security layer. One of the objectives of integrating end-to-end authentication and security is to provide support for micro-payments (the ability to pay for relatively low value items via the cellular handset).

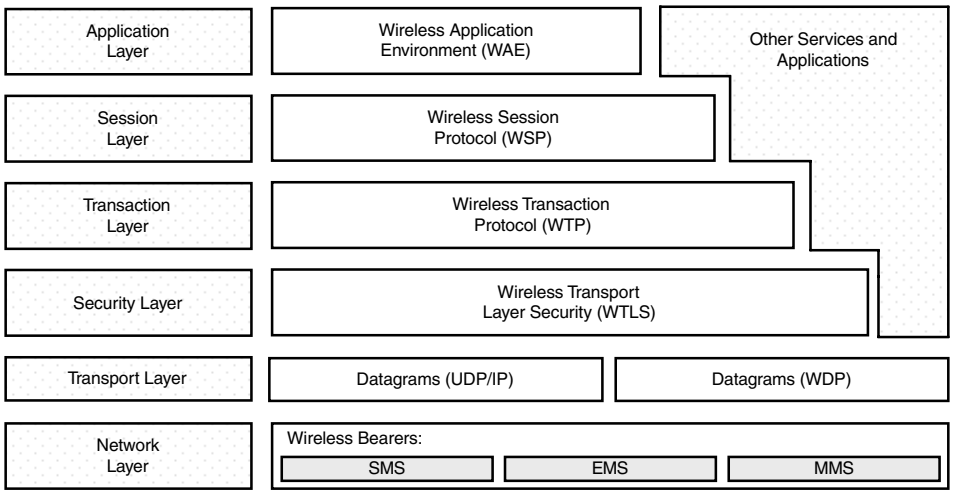


Figure 8.6 WAP layer structure.

The problem with this is verification delay. There is not much point in standing in front of a vending machine and having to wait for 2 minutes while your right to buy is verified and sent to the machine. It is much easier and faster to put in some cash and collect the can.

There is also a specification for a wireless datagram (WDP). Because the radio layer is isochronous (packets arrive in the same order they were sent), you do not need individual packet headers (whose role in life is to manage out-of-order packet delivery). This reduces some of the Physical layer overhead, though whether this most likely marginal gain is worth the additional processing involved is open to debate.

Work items listed for WAP include integration with MExE, including a standardized approach to Java applet management, end-to-end compression encryption and authentication standards, multicasting, and quality of service for multiple parallel bearers. Some of the work items assume that existing IETF protocols are nonoptimum for wireless network deployment and must be modified.

As with content, we would argue it is better to leave well enough alone. Don't change the protocols; sort out the network instead. Sorting out the network means finding an effective way of matching the QoS requirements of the application to network quality of service. This is made more complex because of the need to support multiple per-user QoS streams and security contexts. QoS requirements may also change as a session progresses, and network limitations may change as a session progresses.

As content and applications change then, it can be assumed new software will need to be downloaded into base stations, handsets, and other parts of the network. Some hardware reconfiguration may also be possible. Changing the network in response to changes in the content form factor is infinitely preferable to changing the content in response to network constraints. Reconfiguration does, however, imply the need to do device verification and authentication of bit streams used to download change instructions.

The Software Defined Radio (SDR) Forum (www.sdrforum.org) is one body addressing the security and authentication issues of remote reconfiguration.

Summary

In earlier chapters, we described how the radio physical layer was becoming more flexible—able to adapt to rapid and relatively large changes in data rate. We described also how multiple parallel channel streams can be supported, each with its own quality of service properties. The idea is that the Physical layer can be responsive to the Application layer. One of the jobs of the Application layer is to manage complex content—the simultaneous delivery of wideband audio, image, and video products.

Traditionally, the wireless industry has striven to simplify complex content so it is easier to send, both across a radio air interface and through a radio network. Simplifying complex content reduces content value. It is better, therefore, to provide sufficient adaptability over the radio and network interface to allow the network to adapt to the content, rather than adapt the content to the network.

This means that handset hardware and software also needs to be adaptable and have sufficient dynamic range (for example, display and display driver bandwidth and audio bandwidth) to process wideband content (the rich media mix). In turn this implies that a user or device has a certain right of access to a certain bandwidth quantity and bandwidth quality, which then forms the basis of a quality of service profile that includes access and policy rights.

Given that thousands of subscribers are simultaneously sending and receiving complex content, it becomes necessary to police and regulate access rights to network resources. As we will see in later chapters, network resources are a product of the bandwidth available and the impact of traffic-shaping protocols on traffic flow and traffic prioritization.

Radio resources can be regarded as part of the network resource. Radio resources are allocated by the MAC layer (also known as the data link layer, or Layer 2). The radio resources are provided by Layer 1—the Physical layer. MExE sets out to standardize how the Application layer talks to the Physical layer via the intermediate layers. This includes how hardware talks to hardware and how software talks to software up and down the protocol stack.

The increasing diversity of device (handset) hardware and software form factor and functionality creates a problem of device/application compatibility. Life would be much easier (and more efficient) if a de facto dominant handset hardware and software standard could emerge. This implies a common denominator handset hardware and software platform that can talk via a common denominator network hardware and software platform to other common denominator handset hardware/software platforms.

It is worthwhile to differentiate application compatibility and content compatibility. Applications include content that might consist of audio, image, video, or data. Either the application can state its bandwidth (quantity or quality) requirements or the content can state its requirements (via the Application layer software). This is sometimes described as *declarative content*—content that can declare its QoS needs. When this is tied into an IP-routed network, the network is sometimes described as a *content-driven switched network*.

An example of a content-driven switching standard is MEGACO—the media gateway control standard (produced by the IETF), which addresses the remote control of session-aware or connection aware devices (for instance, an ATM device). MEGACO identifies the properties of streams entering or leaving the media gateway and the properties of the termination device—buffer size, display, and any ephemeral, short-lived attributes of the content that need to be accommodated including particular session-based security contexts. MEGACO shares many of the same objectives as MExE, and as we will see in later chapters, points the way to future content-driven admission control topologies.

Many useful lessons have been learned from deploying protocols developed to accommodate the radio physical layer. If these protocols take away rather than add to content value, they fail in terms of user acceptance. At time of writing, the WAP form is being disbanded and being subsumed into the Open Mobile Alliance (OMA), which aims to build on work done to date on protocol optimization.

CHAPTER 9

Authentication and Encryption

The advent of packet-routed networks and the necessity of sharing transport channels has increased the need for authentication and encryption. The more robust we make the authentication and encryption process, the more value we confer. However, the cost of robust authentication and encryption is an increase in overhead, in terms of processor bandwidth, processing delay, and memory/code footprint. Authentication can be compromised by delay and delay variability, particularly when time-sensitive challenge/response algorithms are used. Network quality and authentication and encryption integrity are therefore intimately related.

This chapter addresses these issues in depth. It also presents several sections of working examples—known dilemmas and possible solutions.

The Interrelated Nature of Authentication and Encryption

Authentication is needed to identify people and devices. It provides people or devices with the authority to access delivery or memory bandwidth—including the right to deposit information in and retrieve information from secure storage. It provides people or devices with the authority to change network parameters—for instance, software upgrades or hardware reconfiguration. It also provides people or devices with the authority to change handset parameters—software upgrades or hardware reconfigurations.

Authentication may be used for:

- Identification and the enforcement of access rights and security policies
- Content distribution
- Application distribution
- Transaction processing
- Virtual data warehousing (storage)

We may need to authenticate device hardware in a network to prevent a security breach. For example, it is technically feasible to replace a router without a network operator's knowledge and then use the router to eavesdrop on traffic or filter out traffic of commercial or political value.

We may also need to authenticate to provide transaction security, for example, if we are using a digital cellular handset to make micro or macro payments.

Authentication can be given for a particular period of time—the length of a session, for example—and then needs to be renewed. Authentication can also be for a long length of time. The right to access storage 900 years from now (recall the Domesday project in Chapter 6) would be an extreme example.

Absolute authentication does not exist. We can never be totally certain that a device is the device that it claims to be or the person is the person he or she claims to be. The more certain we are, however, the more value we confer on the authentication process. Certainty is achieved by *distance*, which is how unique we make the authentication. Distance confers value but also incurs cost. The cost is processor overhead and delay. Usually, authentication requires more information to be sent and therefore also absorbs delivery bandwidth and RF power.

The Virtual Private Network

Once we have authenticated, we can encrypt. A number of techniques are used to provide distance between the user and any (legitimate or nonlegitimate) third parties who wish to read the user's plaintext files. These techniques include the use of nonlinear feedback registers.

The combination of authentication and encryption allows a network operator to provide a virtual private network over a public access network. The virtual private network includes delivery and storage bandwidth. As we will see in later chapters, storage bandwidth can be enhanced by providing archiving, management, search and retrieval systems, and (possibly) higher levels of access security than will be available in a private network.

Key Management

The historic and traditional problem with encryption has been the reliance on a single key to both encode and decode a plaintext message. Ownership of the key gave easy access to the message contents and meant that keys could only be passed to intended recipients by a secure exchange process.

Diffie and Hellman developed the concept of splitting the key into two parts—an encode key and a decode key. Further developments of this concept allowed the exchange of keys through a public, insecure medium and enabled anyone to create an encrypted message but only the trusted recipient would be able to decrypt the message.

This process is achieved through the "lodging" of public keys but the retention of a private key. The actual exchange (and encryption) process relies on the manipulation of very large primes, the product of which is near to impossible to factorize. A worked example is included at the end of this chapter.

As with authentication, there is no such thing as absolute security. Any encryption scheme can be compromised, but the greater the distance—that is, the harder it is to decrypt the traffic—the more value the encryption process confers.

Digital Signatures

The RSA algorithm, developed by Rivest, Shamir, and Adelman, is often used for digital signature verification. This is a large prime number algorithm. An example is included at the end of this chapter.

A key can be established between two consenting devices or two consenting people or between a device and a network or a person and a network or between multiple devices accessing multiple networks. Key administration can therefore become quite tricky. Keys can be organized in such a way that they all become part of a trust hierarchy. Trust in the key is implied by the fact that the key was signed by another trusted key. One key must be a root of the trust hierarchy. This is used in centralized key infrastructures using a Certification Authority and providing the basis for the Public Key Infrastructure (PKI), which we cover later.

The network can in effect provide an additional level of verification value by identifying the user by his international mobile subscriber identity (IMSI), the user's equipment reference (equipment identity number), and a system frame number timestamp. The network then becomes an intermediary in the authentication process.

Senders can also be spenders and may be engaging in micro- or macro-payments (authorizing, for example, large financial transactions). The network can verify the claimed identity of the sender/spender. The sender/spender cannot later repudiate the contents of the message. For example, if the sender/spender has ordered a thousand garden forks, it can be proved that he ordered a thousand garden forks and has to pay for them.

Digital signatures also have the useful ability to replace handwritten signatures but are more flexible. For example, we can sign pictures without making the signature visible to the user.

Network operators also have a legal obligation to make traffic passing through their network available to legitimate eavesdropping authorities—government security agencies, for example. The traffic (voice, image, video, data) has to be available as plaintext.

For this to happen, each user must deposit knowingly or unknowingly his or her secret key with a central authority—a trusted third party from whom the key can be recovered, provided a case for legitimate eavesdropping has been put forward and agreed upon.

Hash Functions and Message Digests

Many signature methods couple authentication and secrecy (encryption/decryption) together. The key used for authentication is also used for encryption. Secrecy, however, comes at a cost: Cryptography involves delay, processor overhead, and memory and delivery bandwidth overhead in the handset. It is therefore often useful to have an authentication process that does not require the whole message to be encrypted. This is sometimes described as a *hash function* or *message digest*.

In Chapter 7, we discussed content ownership and the codification of ownership rights (MPEG-4/MPEG-21). Ownership rights, of an image or video clip, for example, can be protected by computing a message digest consisting of the file countersigned (that is, multiplied by) the user's secret key and possibly also a timestamp—or, across a radio air interface, a system frame number.

The digest, or hash function, has to have three properties:

- Given P (the plaintext), it is easy to compute MD(P).
- Given MD(P), it is effectively impossible to find P.
- No one can generate two messages that have the same message digest. To meet this, the hash should be at least 128 bits long, preferably more.

A number of message digests have been proposed. The most widely used are MD5 and Secure Hash Algorithm (SHA). MD5 is the fifth in a series of hash functions designed by Ron Rivest. It operates by jumbling up bits in a way that every output bit is affected by every input bit. SHA is similar in process but uses 2 bits more in the MD. It is consequently 2^{32} more secure than MD5, but it is slower, since the hash code is not a power of 2.

Public Key Infrastructure

We said that it is a legal requirement for public network operators to provide plaintext access to traffic passing through the network. This is the reason for the Public Key Infrastructure. PKI has the following three functional components (see also Figure 9.1):

Certificate Authority (CA). This is a trusted third party and might be a commercial company such as VeriSign, Entrust, or Baltimore.

Repository Authority (RA). The RA contains the keys, certificates (information about users), and certificate revocation lists (CRLs, which contain information on time expired or compromised certificates).

Management function. This function looks after key recovery, message, or data recovery.

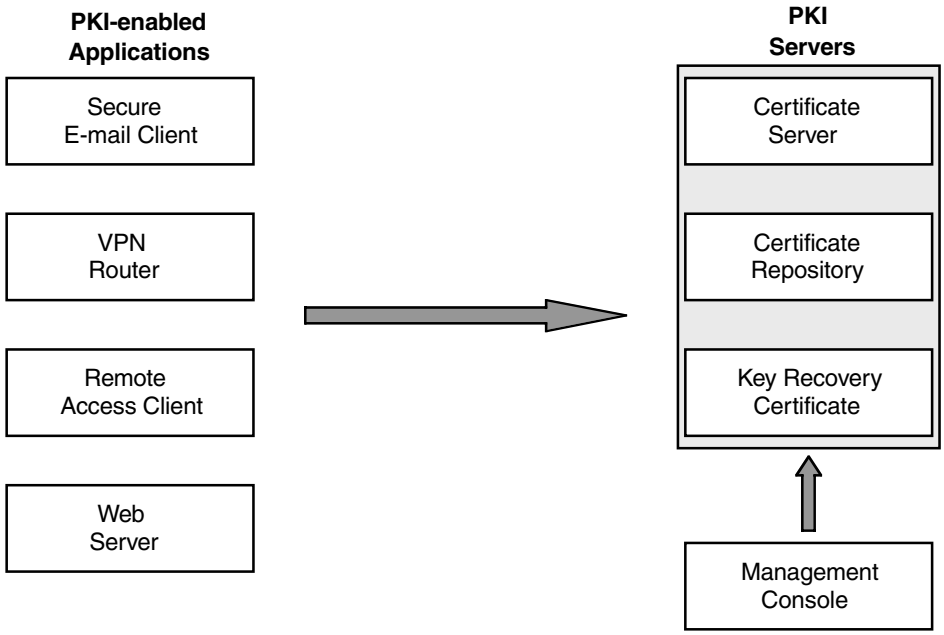


Figure 9.1 PKI server components.

Security Management

The Certificate Authority and Registration Authority functions can be implemented on one or more servers, which may or may not use Lightweight Directory Access Protocol (LDAP). Table 9.1 shows typical functions within a PKI implementation.

Table 9.1 PKI Implementation

FUNCTION	DESCRIPTION	IMPLEMENTATION
Registering users	Collect user information	Function of the CA or a separate RA
Issuing certificates	Create certificates in response to a user or administrator request	Function of the CA
Revoking certificates	Create and publish CRLs	Administrative software associated with the CA

(continues)

Table 9.1 PKI Implementation *(Continued)*

FUNCTION	DESCRIPTION	IMPLEMENTATION
Storing and retrieving certificates and CRLs	Make certificates and CRLs conveniently available to authorized users	The repository for certificates and CRLs. Usually a secure, replicated directory service accessible via LDAP or X500.
Policy-based certificate path validation	Impose policy-based constraints on the certificate chain and validate if all constraints are met	Function of the CA
Timestamping	Put a timestamp on each certificate	Function of the CA or a dedicated time server (TS)
Key life cycle management	Update, archive, and restore keys	Automated in software or performed manually

There are many routine housekeeping functions implicit in PKI administration, for example, multiple key management (users may have several key pairs for authentication, signatures, and encryption), updating, backup (forgotten passwords), a disk crash or virus protection, and archiving (recovering the key used by an ex-employee, for example). Encryption keys have to be archived. Signing keys may also be archived.

PKI forms the basis for providing a virtual private network over a public access network—the more robust the authentication and encryption, the more value the network confers. PKI-based networks don't have to but can use standard IP protocols. Authentication and encryption can convert standard Internet links to provide site-to-site privacy (router to router) or secure remote access (client to server).

Tunneling protocols can be used to wrap/encapsulate one protocol in another protocol. The encapsulated protocol is called Point-to-Point Protocol (PPP); the encapsulating protocol is a standard Internet protocol. The standard for site-to-site tunneling is the IP Security (IPSec) protocol defined by the IETF.

If the network is a wireless network, this could be described as a Wireless Enterprise Service Provision (WESP) platform providing virtual enterprise resources. It could sit side by side with a Wireless Application Service Provision (WASP) platform, which could provide virtual applications (downloading database management software, for example). The WASP could sit side by side with a Wireless Internet Service Provision (WISP) platform providing standard (nonsecure) or secure Internet access.

Downloaded applications need to be verified in terms of their source and integrity, to make sure that they are virus-free. In the PC world, when a new virus appears, it is detected (hopefully) by one of the several virus control specialist companies that now exist (Sophos is one example—www.sophos.com). The virus is then shared amongst each of the specialist antivirus companies who individually work on a counter-virus, which is then sent to their customers. This is an effective pragmatic system, but it does result in the need to store virus signature files on the PC, which can rapidly grow to a memory footprint of many megabytes.

Digital cellular handset software and PDA software has traditionally been ROM based, but the need to remotely reconfigure means that it makes more sense to have the software more accessible (which also means more vulnerable to virus infection). However, it is not a great idea to have to fill up a lightweight portable wireless PDA with megabytes of antiviral signature files, because it wastes memory space in the handset/PDA and it uses up unnecessary transmission bandwidth. The alternative is to use digital signatures to sign any data streams sent out to the handset.

The idea of PKI is to standardize all the housekeeping needed for authentication and encryption when applied across multiple applications carried across multiple private and public access networks (that is, to look after enrolment procedures, certificate formats, digital formats, and challenge/response protocols).

Challenge/response protocols can be quite time-sensitive—particularly to delay and delay variability. The challenge will expect a response within a given number of milliseconds. If a response is received after the timeout period, it will be invalid. This is an important point to bear in mind when qualifying end-to-end delay and delay parameters in a network supporting, for example, mobile commerce (m-commerce) and micro- or macro-payment verification.

The focus for interoperable PKI standards is the PKI working group of the IETF known as the PKI Group (PKI for X509 certificates). X509 certificates are a standardized certificate format for describing user security profiles and access rights. PKI therefore becomes part of the admission protocol that needs to be supported in the handset and the network.

Areas covered by the PKI standard are shown in Figure 9.2 and are as follows:

EDI. Standards for Electronic Data Interchange.

SSL. The Secure Socket Layer protocol used within IETF to provide IP session security.

PPTP. The Point-to-Point Tunneling Protocol.

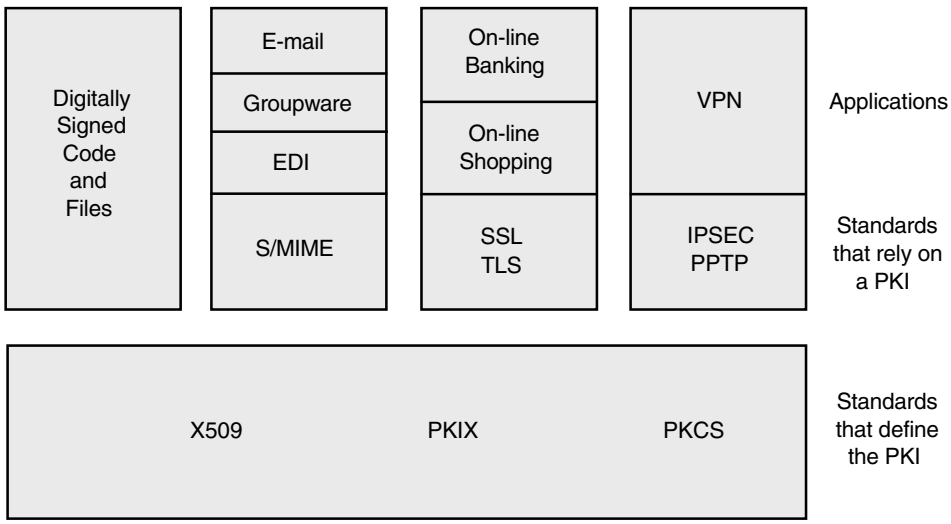


Figure 9.2 PKI standards.

SSL and Transport Layer Security (TLS) are used to provide the basis for secure electronic transactions.

Virtual Smart Cards and Smart Card Readers

One problem with PKI is that it assumes that smart cards are, or will be, a standard component in cellular handsets, workstations, and PCs. Although this is the case with GSM handsets, it has not been the case to date with U.S. cellular handsets. If smart cards and smart card readers are not readily available, some organizations may compromise security by placing private keys on users' hard disks or in temporary cache memory. One answer is to create a virtual smart card and a virtual smart card reader.

The PC, workstation, or smart card-less cellular phone connects to the virtual smart card server and interacts with an emulated smart card as if it were communicating with a hardware smart card connected to a reader. Users activate their virtual smart card with either a memorized static PIN or a dynamic password.

Where to Implement Security

As we are beginning to show, there are a number of standards groups and interest groups implementing authentication, encryption, and security solutions—some at the Application layer (embedded SSL in Windows 2000, for example), some at the IP packet layer, and some (over the air) at the physical layer (see Table 9.2).

The IPSec Standard

IPSec is the standard for protecting traffic at the packet level, using transforms—-that is, changes to the packet structure—to confer security. There are two main transforms used in IPSec: an Authentication Header (AH) transform and an Encapsulating Security Payload (ESP) transform. The transforms are configured in a data structure called a Security Association (SA).

The AH provides authentication (data origin authentication, connectionless integrity, and antireplay protection) to a datagram. It protects all the data in the datagram from tampering as specified in the Security Association, including the fields in the header that do not change in transit. However, it does not provide confidentiality. An AH transform calculates or verifies a Message Authentication Code for the datagram being handled. The resulting MAC code is attached to the datagram.

Before a secure session can begin, the communicating parties need to negotiate the terms for the communication. These terms are those defined in the SA. There needs to be an automated protocol to establish the SAs to make the process feasible for the Internet (that is, a global network). This automated protocol is the Internet Key Exchange (IKE), which is meant for establishing, negotiating, modifying, and deleting SAs. IKE combines the Internet Security Association and Key Management Protocol (ISAKMP) with the Oakley key exchange. Oakley is a working group defining key exchange procedures.

Table 9.2 Security Implementations by Layer

Layer 7 Application layer	Win CE/Java/Symbian	Application layer security
Layer 6 Presentation layer	HTML/XML	
Layer 5 Session layer	RSVP	
Layer 4 Transport layer	TCP	IP packet layer security
Layer 3 Network layer	IP	
Layer 2 Data link layer	ATM and Ethernet Media Access Control	Access control
Layer 1 Physical layer	Fiber/copper cable/ wireless	Over-the-air security

Figure 9.3 shows how IPSec is implemented. The IPSec engine performs AH transforms, ESP transforms, compression transforms, and special transforms (for example, network address translation using IP4). Special transforms also include content- or context-sensitive filtering and automatic fragmentation, if a packet exceeds the maximum transfer unit size. The engine also has to detect denial-of-service attacks.

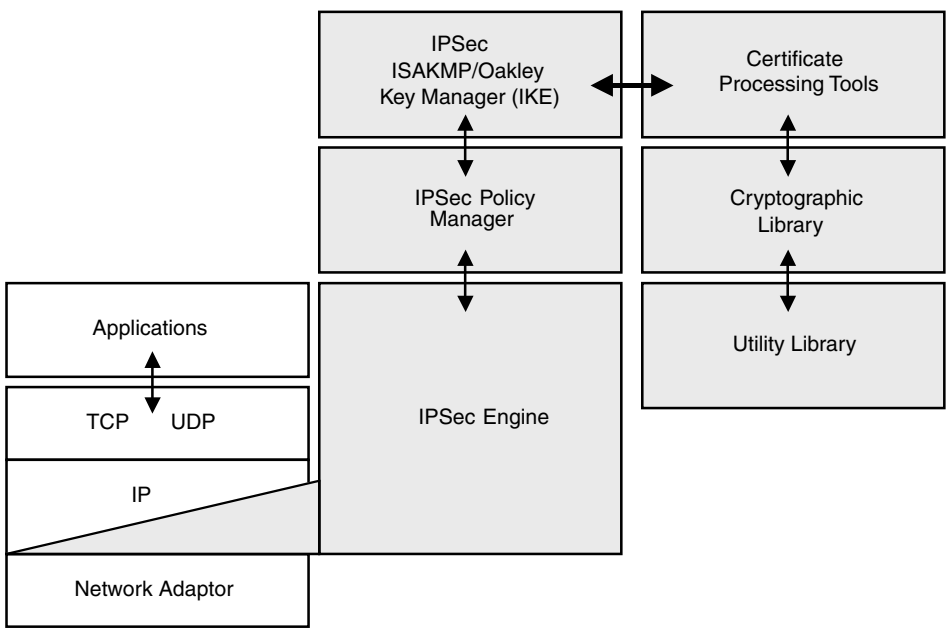


Figure 9.3 IPSec TCP/IP integration.

IPSec can be implemented in the handset, in a Node B, in a radio network controller, and in intermediate routers in the IP network, including firewalls. IPSec, however, can imply significant overheads in terms of delay and delay variability, processor overhead, memory overhead, and additional transmission bandwidth requirements—the cost of security. This is okay if there is a perceived value gain greater than the additional cost. IPSec performance and the performance of processes such as the Diffie-Hellman exchange can be very dependent on good software implementation—assembler optimization, for example. By implication, it becomes a very intimate part of the QoS SLA. A firewall on its own can introduce 150 ms of delay.

The problem becomes more acute if you need to dynamically authenticate a workgroup with users joining and leaving during a session or in multicasting. To quote from an Internet draft (www.ietf.org/internet-drafts/draft-ietf-ipsec-gkmframework-01.txt): "The complexity of these [multicast] cryptography solutions may point to the application layer being the best place for them to be implemented." In other words, because you need flexibility—that is, you cannot predict when users will be joining or leaving the simulcast or multicast—it is better to implement security in the application layer.

The IETF Triple A

We have briefly addressed authentication. We also need to discuss the interrelationship between authentication, authorization, and accounting—or, as described by the IETF, Triple A.

It is not sufficient just to have identity-based authentication. There is also a need to support role-based access control. This has been used for many years in private radio networks to give users specific event-based or role-based access rights. (Motorola calls them storm plans; Ericsson calls them special event plans.) A storm plan might be, for example, a preplanned network response to a terrorist attack. The chief of police, chief of fire, the mayor, or president may acquire a particular set of access rights triggered by the event. Individuals can have particular access rights and groups of users can have access rights. The access rights include the right of access to delivery and memory bandwidth (security data bases, hazardous chemical information, or firefighting information, for example). Similar topologies can be used to qualify spending rights and spending power. IETF Triple A also supports a criticality flag analogous to preemption rights in a storm plan (where the chief of police effectively pulls rank to get channel access). There may be a need to reject legitimate but unwanted users.

In the context of allowing a right of access, level of trust is a relative term. Even if a cryptographically correct certificate is presented, you can never be completely sure a person or device is who they claim to be.

The stability of the access protocol also becomes very critical in these applications. For example, suppose a 747 lands on Downing Street, and 1200 Metropolitan police officers all press their press-to-talk keys on their radio at the same time, expecting instant access and authentication. The access bandwidth is sufficient to support 100 simultaneous users. The authentication bandwidth also has to be sufficient to avoid unacceptable access delay. We thus have another performance metric—protocol performance (also describable as protocol bandwidth). It is relatively easy to become protocol-limited—a frustrating situation where you have access bandwidth available but cannot use it because the protocol cannot respond quickly enough to the immediate/instantaneous bandwidth need.

IETF Triple A also codifies how to deal with protocol security attacks—man-in-the-middle attacks, replay attacks, or bid-down attacks (against which timestamping is generally a useful defense).

Accounting within Triple A includes financial accounting (billing and accountability), session logging, and audit trails to prove a session took place and to protect against repudiation (claiming you didn't order those thousand garden forks). Accounting audit trails can be used commercially *and* to track and search for sessions that may, in retrospect, acquire national security or financial interest (September 11th/ Enron).

Encryption Theory and Methods

When discussing encryption, we need to differentiate *over-the-air* encryption (for example when we use the A3/A5/A8 keys and encryption algorithm on the GSM SIM to provide security over the radio interface) and *end-to-end* encryption (literally from user to user), as shown in Table 9.3. If we require end-to-end encrypted traffic to be read in plaintext by the network, then we need user keys to be stored by a trusted third party and available to be used by the operators on behalf of authorities with a right of access.

Encryption and Compression

Encryption and compression are interrelated. Compression can be used legitimately to improve delivery and storage bandwidth utilization when transmitting or storing voice, text, images, or video content. Compression or steganography (for example, voice files embedded in image files) can be used nonlegitimately as a form of encryption. Highly compressed files have high entropy. It can be hard to distinguish whether files are encrypted or heavily compressed.

We said earlier that compression ratios increase by an order of magnitude every 5 years. This is the result of additional processor and memory bandwidth. Similarly, the gap between encryption and cryptoanalysis increases with time, that is, encryption distance (and, potentially, encryption value) increases with time. This is because increasing computing/processing power confers greater advantage to the cipher user because longer key lengths take polynomially more time to use but exponentially more time to break. Ten years ago, a 56-bit length key would have been considered secure. Today, RSA encryption typically uses 1024- or 2048-bit keys.

Encryption performance can be measured in terms of time to break in MIPS/years (how much computing power in terms of million of instructions per second × how long it would take theoretically to break the code). An RSA 1024-bit key takes theoretically 10^{11} MIPS/years to break. A 2048-bit key takes 10^{20} MIPS/years to break.

Table 9.3 Encryption Differentiation

OVER THE AIR	END TO END
Bit level/symbol level	Bit level/symbol level
Packet level	

So as processor bandwidth increases, it becomes possible to use progressively longer keys. However, as keys get longer, encryption/decryption delay increases, processor overheads and power consumption increases, and code and memory space occupancy increases.

Evolving Encryption Techniques

As a consequence, a number of alternative encryption techniques have been proposed that offer the same distance (security) as RSA but with a shorter key length. Elliptic curve cryptography (ECC) is one option. Table 9.4 shows the ECC key length against an equivalently secure RSA key and demonstrates how the advantage of ECC increases as key length increases.

This delivers significant advantage in terms of processing delay. The examples in Table 9.5 give typical encryption/decryption delays for a system running at a 20 MHz clock rate.

DES to AES

In the United States, data encryption began to be standardized commercially in the 1970s. IBM had a project called Lucifer based on a 56-bit key, and this became the basis for the first generation of cipher standards known as Data Encryption Standard (DES). DES became very outdated and insecure in the 1990s and was replaced with Triple DES, an encrypt/decrypt/encrypt sequence using three different unrelated keys. At present there is a new encryption standard being defined to replace DES known as Advanced Encryption Standard (AES).

AES is able to encrypt streaming audio and video in real time. In addition, it can fit on a small 8-bit CPU (for example, on a smart card) and can be scaled up to work on 32-bit/64-bit CPUs for bulk data encryption. Key lengths are 128, 192, or 256 bit. It is not designed to be particularly secure, but it is computationally expensive to de-encrypt, which means it confers sufficient distance to provide adequate commercial protection without too much delay or processor overhead.

Smart Card SIMS

We covered smart card SIMS briefly in Chapter 4. These are the de facto standard for providing both over-the-air encryption and end-to-end encryption. Hitachi, for example, has a 32-bit microcontroller embedded on a smart card with an onboard cryptographic coprocessor that can do a 1024-bit RSA calculation in less than 120 ms (courtesy of the hardware coprocessor speeding up the calculation). Having the processor hardware on the smart card makes the solution more resilient to hardware attack.

Table 9.4 ECC Key Length/RSA Key Comparison

RSA KEY LENGTH	ECC KEY LENGTH	RATIO
1024	160	7.1
2048	210	10.1
21,000	600	35.1

Table 9.5 Computational Comparisons

Key Length	Encryption Method	DECRYPT	ENCRYPT
1024 bit	RSA	86 ms	24 ms
160 bit	ECC	5 ms	10 ms
2048 bit	RSA	657 ms	94 ms
209 bit	ECC	7 ms	14 ms

Biometric Authentication

We also briefly touched on fingerprint authentication. Fingerprint recognition devices acquire fingerprint images using solid-state capacitance sensing, avoiding the need for passwords by using minutiae data.

An individual's finger acts as one of the plates of a capacitor. The other plate consists of a silicon chip containing a sensor grid array yielding an image. When a finger is placed on the chip's surface, the sensor array creates an 8-bit raster-scanned image of the ridges and valleys of the fingerprint. An analog-to-digital converter digitizes the array output. These devices can be integrated with smart card authentication platforms.

Many body parts can be used for identification (and hence provide the basis for authentication and encryption): fingerprints, eyes, facial features, and so on. These techniques are known as *biometric recognition* techniques.

Some useful Web sites for biometric recognition include the following:

www.authentec.com

www.biocentricsolutions.com

www.cybersign.com

www.eyedentify.com

www.digitalpersona.com

www.identix.com

www.nuance.com (voice recognition)

www.myhandreader.com

www.speechworks.com

www.verivoice.com

www.visionics.com

The more accurate (that is, robust) the recognition metric, the more processor and delay overhead are incurred (but also, the more robust the process, the more value it should have).

Examples of recognition optimization include the following:

- Ridge minutiae recognition algorithms
- Ridge bifurcation and termination mapping (already used in fingerprint search engines)
- Ridge width and pore and sweat duct distribution

Working Examples

This section provides working examples of the authentication and encryption processes detailed in this chapter. The detail of the encryption and security process and maths involved may be illustrated by an attractive analogy taken from Simon Singh's *The Code Book*, published by Fourth Estate (ISBN 1-857-02889-9).

The problem is to be able to transfer messages securely through an unsecure environment. Suppose person A and person B wish to communicate through the postal service. Person A puts his message in a box and attaches a padlock for which only he has the key. He then mails the box to person B. Person B cannot open the box, since he does not have the key. Person B puts another padlock on the box for which only he has the key. Person B then sends the box (with two padlocks on) back to person A. Person A removes his padlock and sends the box back to person B. Person B removes his padlock, opens the box, and removes the message.

If you want the same description the hard way, read through the following sections.

Over-the-Air Encryption

The SIM/USIM encryption works as follows (a GSM/TETRA example):

1. A random challenge is sent from the network of 128 bits.
2. The handset encrypts the challenge using an algorithm known as A3 held on the smart card and the key K: of 128 bits also on the smart card.
3. The handset sends back a signed response (S-RES 32 or 64 bit).
4. S-RES is passed through the A8 algorithm on the smart card to derive the key Kc (54 bits + stuffer bits making up a 64-bit word), which is stored in the non-volatile memory on the SIM.
5. Kc is multiplied with a 22-bit word representing the frame number using the A5 algorithm to produce 114 ciphered bits.
6. The 114 ciphered bits are Exclusive OR'd with 114 coded bits (2×57 coded bits are contained in each bit burst).
7. A5 is embedded in the handset/BTS/Node B hardware.

To provide subscriber identity protection, the IMSI is replaced with a Temporary Mobile Subscriber Identity number (TMSI) when the handset initially talks to the network (before encryption is enabled). The TMSI is a product of the IMSI and the location area identity (LAI).

Public Key Algorithms: The Two-Key System

As stated, early cryptosystems had the weakness of the use of a single cipher key. Ownership of the key broke open the whole system and allowed any key owner to decipher the message. Security therefore related to maintaining the secrecy of the key—if the same degree of protectiveness was applied to the message, encryption would be unnecessary.

This all changed with the invention of the two-key cryptosystem, which uses different encode and decode keys that cannot be derived from one another. A further benefit of this approach is that the keys could be exchanged to relevant parties publicly with security maintained.

This two-key Public Key Algorithm (PKA) is the fundamental process underlying encryption, authentication, and digital signatures—referred to as Public Key Encryption (PKE). If the message to be secured is plaintext P, the keyed encryption algorithm E, and the keyed decryption algorithm D, then the method requires the following logic:

1. D[E(P)] = P
2. It is exceedingly difficult to deduce D from E.
3. E cannot be broken by a chosen plaintext attack.

So:

1. Says that if decryption key D is applied to the encrypted text—that is, E(P)—then plaintext P is recovered.
2. Needs no explanation.
3. Would-be intruders can experiment with the algorithm for an impracticably long time without breaking the system, so the keys can be made public without compromising access security.

In practice, Party A, wishing to receive secure messages, first devises two algorithms, E_A and D_A, meeting the three requirements. The encryption algorithm and key E_A is then made public; hence using public key cryptography. Thus, E_A is public, but D_A is private. Now, the secure communication channel can be operated:

- Party A, who has never had contact with Party B, wishes to send a secure message. Both parties' encryption keys (E_A and E_B) are in a publicly readable file.
- Party A takes the first message to be sent, P, computes $E_B(P)$ and sends it to Party B.
- Party B decrypts it by applying her secret key D_B (that is, they compute $D_B[EB(P)] = P$).

No third party can read the encrypted message, $E_B(P)$, because the encryption system is assumed strong and because it is too difficult to derive D_B from the publicly known E_B. The communication is secure.

So, public key cryptography requires each user to have two keys:

Public key. Used by everyone for sending messages to that user

Private key. Used by the recipient for decrypting messages

Now. let's take a little "back-to-school" refresher course.

Prime Numbers

A natural number (positive integer) is *prime* if it has no factors (numbers whose product is the given number), other than itself and 1. If a number is not prime, it is called *composite* (for example, 17 is prime, but $18 = 2 \times 3^2$ is composite). Finding algorithms for

determining if an integer is prime and the distribution of the prime numbers within a set of natural numbers are still major challenges for mathematicians. There are no computationally efficient algorithms for finding the prime factorization of a given integer. There are, however, computationally efficient ways of testing whether a given integer is prime. This is central to PKE systems.

Although unproven, computation times for factoring an integer grow exponentially with the size of the integer, whereas computation times for integer multiplication grow only polynomially.

Congruency

Two integers a and b are said to be *congruent* (mod n), written $a \equiv b$, if $a = b + nm$ for some integer n. For example, $25 \equiv 17 \equiv 1(\text{mod } 4)$, thus, two integers are congruent if they both have the same remainder when divided by n. We can say $a \equiv b(\text{mod } n)$ if n divides evenly into $(a - b)$. If a and n have no common factors, they are said to be *relatively prime* and then the congruency $ax \equiv 1(\text{mod } n)$ always has a unique solution.

Algorithms have to be found to satisfy all three criteria. One is the RSA algorithm. The method is based on number theory and can be summarized as follows:

1. Choose two large primes, p and q, (typically $> 10^{100}$).
2. Compute $n = p \times q$ and $z = (p\text{-}1) \times (q\text{-}1)$.
3. Choose a number relatively prime to z and call it d.
4. Find e such that $e \times d = 1 \text{ mod } z$.

To apply the algorithm, we divide the plaintext (bit strings) into blocks so that each plaintext message, P, falls in the interval $O \leq P < n$. This can be done by grouping the plaintext into blocks of k bits, where k is the largest integer for which $2^k < n$ is true.

- To encrypt a message, P, compute $C = P^e(\text{mod } n)$.
- To decrypt the message, C, compute $P = C^d(\text{mod } n)$.
- To perform encryption, e and n are needed.
- To perform decryption, d and n are needed.
- Therefore, e, n are the public keys and d is the private key.

The security of the method is based on the factoring of large numbers. If n can be factored, p and q can be found, and from these, z. From z and e, d can be found. However, factoring large numbers is immensely difficult. According to Rivest, Shamir, and Adelman, factoring a 200-digit number requires 4 billion years of computer time. A 500-digit number requires 10^{25} years. In both cases, the best-known algorithm is assumed and a 1 μs instruction time. On this basis, if computers get faster by an order of magnitude per decade, centuries are still required to factor a 500-digit number.

An example will demonstrate the calculations. We will use small numbers:

$p = 3$ and $q = 11$ is chosen.

$n = p \times q$, so, $n = 3 \times 11 = 33$.

$z = (p - 1) \times (q - 1)$, so $z = (3 - 1) \times (11 - 1) = 20$.

A suitable value for d is 7, as 7 and 20 have no common factors, and e can be found by solving:

$de = 1 \pmod{z}$

$\therefore = 1 \pmod{20}$

$\therefore = 1 \pmod{20}/7$

$= 3$

The ciphertext, C, for a plaintext message, P, is given by $C = P^e \pmod{n}$; that is, $C = P^3 \pmod{33}$. The ciphertext is decrypted by the recipient according to the rule $P = C^d \pmod{33}$.

A test case using the word ROGER will demonstrate (see Table 9.6). As the primes chosen for this example are very small, P must be less than 33, so each plaintext block can only be a single character. The result will therefore be a mono-alphabetic substitution cipher—not very impressive. If p and $q \approx 10^{100}$, n would equal 10^{200}, and each block could be up to 664 bits or eighty-three 8-bit characters. The procedure can be followed:

Encrypt R = This is the 18th letter in the alphabet.

This letter will be called plaintext P.

So, $P = 18$.

$P^3 = 5832$.

Ciphertext (C) = $P^3 \pmod{n}$.

$= 5832 \pmod{33}$.

To find $P^e \pmod{33}$:

$5832/33 = 176.7272727$.

So, $C = 33 \times 0.7272727 = 24$.

So, for the letter R, 24 is transmitted and the numeric value 24 is received.

The recipient calculates $C^d = 24^7 = 4586471424$.

and then $C^d \pmod{n} = 4586471424 \pmod{33}$.

to find $C^d \pmod{n}$: $4586471424 = 138983982.545454$.

So, the numeric value of the symbol = $n \times 0.545454 = 18$ (that is, the letter R).

Table 9.6 Test Case

SYM	NUM	P^3	P^3(MOD 33)	C^7	C^7(MOD 33)	SYM
R	18	5832	24	4586471424	18	R
O	15	3375	9	4782969	15	O
G	7	343	13	62748517	7	G
E	5	125	26	8031810176	5	E
R	18	5832	24	4586471424	18	R

RSA is widely used as a public key algorithm but is considered too slow for processing large amounts of data. The Weizmann Institute Key Locating Engine (Twinkle) is an electro-optical computer designed to execute sieve-based factoring algorithms approximately two to three orders of magnitude faster than a conventional fast PC. Designed by Professor Adi Shamir (of RSA), it should crack 512-bit RSA keys in a few days, according to Shamir.

Presently, a rough design, it should run at 10 GHz and uses wafer scale integration (source: *New Electronics*, Sept 99).

Diffie-Hellman Exchange

Public key security relies on the communicating parties sharing a secret key. A major consideration is how both parties can come to share such a key. The obvious exchange possibility is that of a face-to-face meeting, but this may be impractical for many reasons. Therefore, a method of open exchange is required in which keys may be transferred and yet remain secret.

The Diffie-Hellman Key Exchange is an exchange process that meets this requirement of public exchange of private keys. The intention is that a third party should be able to monitor this communication by the first two parties and yet not be able to derive the key.

Parties A and B communicate over a public insecure medium, for example, the Internet. A and B agree on two large prime numbers, n and g, where $(n\text{-}1)\div 2$ is also a prime and certain conditions apply to g. As these numbers (n and g) are public, either A or B may pick them.

- Now A picks a large (512 bit, etc.) number, x, and keeps it secret.
- Similarly, B picks a large secret number, y.
- A initiates key exchange by sending B a message containing n, g, $g^x \bmod n$.
- B responds by sending A a message containing $g^y \bmod n$.
- A takes B's message and raises it to the xth power to get $(g^y \bmod n)^x$.
- B performs a similar operation to get $(g^x \bmod n)^y$.
- Both calculations yield $g^{xy} \bmod n$.
- Both A and B now share a secret key $g^{xy} \bmod n$ (see Figure 9.4).

Vulnerability to Attack

An interested third party (party T) has been monitoring the messages passing backward and forward. T knows g and n from message 1. If T could compute x and y, he would have the secret key. T only has $g^x \bmod n$ and so cannot find x. No practical algorithm for computing discrete logarithm modules from a very large prime number is known.

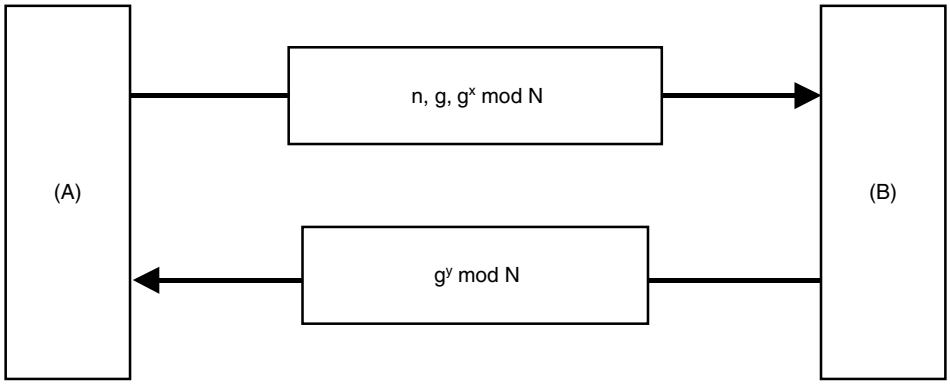


Figure 9.4 Key exchange.

An example (using small numbers for practicality) will show the process:

- The primes chosen by the parties are $n = 47$, $g = 3$.
- Party A picks $x = 8$, and party B picks $y = 10$; both of these are kept secret.
- A's message to B is (47, 3, 28), because 3^8 mod 47 is 28 (that is, $3^8 = 6561$: $6561/47 = 139.5957447$ and $0.5957447 \times 47 = 28$).
- B's message to A is (17).
- A computes 17^8 mod 47, which is 4.
- B computes 28^{10} mod 47, which is 4.
- A and B have both determined that the secret key is 4.
- Party T (the intruder) has to solve the equation (not too difficult for these sized numbers but impossibly long—in time—for long numbers): 3^x mod 47 = 28.

If the intruder T can insert himself in the message channel at key exchange commencement, he can control the communication, as follows (see also Figure 9.5):

- When party B gets the triple, (47, 3, 28), how does he know it is from A and not T? He doesn't!
- T can exploit this fact to fool A and B into thinking that they are communicating with each other.
- While A and B are choosing x and y, respectively, T independently chooses a value z.
- A sends message 1 intended for B. T intercepts it and sends message 2 to B, using the correct g and n (which are public) but with z instead of x.
- T also sends message 3 back to A.
- B sends message 4 to A, which T also intercepts and keeps.

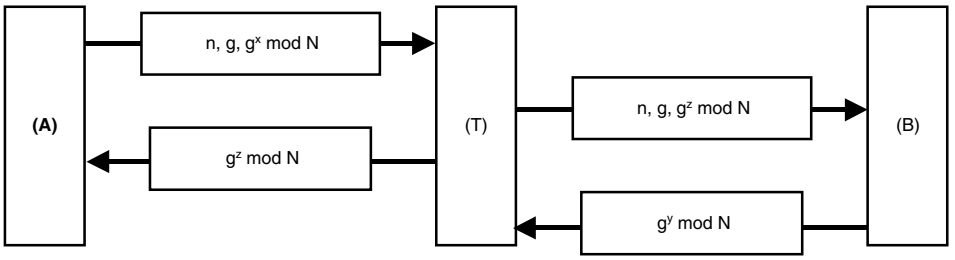


Figure 9.5 Key exchange vulnerability.

All parties now do the modular arithmetic:

- Party A computes the secret key as g^{xz} mod n, and so does T (for messages to A).
- Party B computes g^{yz} mod n, and so does T (for messages to B).
- Party A believes he is communicating with B so establishes a session key (with T). B does the same. Every message that A and B sends is now intercepted by T and can be stored, modified, deleted, and so forth. Party A and B now believe they are communicating securely with each other.

This attack process is known as the bucket brigade attack or the man-in-the-middle attack. Generally, more complex algorithms are used to defeat this attack method.

Authentication: Shared Secret Key

The initial assumption is that party A and party B already share a secret key K_{AB}. The shared secret key will have been established via a secure process (for example, person to person), a Diffie-Hellman exchange, and so on. This protocol is based on the commonly used principle that the first party sends a random number to the second party. The second transforms it in a special way, and then returns the result to the first party.

The authentication protocol type described is called a challenge/response protocol, which is defined as follows (also see Figure 9.6):

- A_i B_i are the identities of the two communicating parties A and B.
- R_is are the challenges, where the subscript identifies the challenger.
- K_i are keys, where i indicates the owner.
- K_s is the session key.

Message 1

- Party A sends her identity (A_i) to B in a way that B understands.
- Party B has no way of knowing whether this message comes from A or an intruder T.

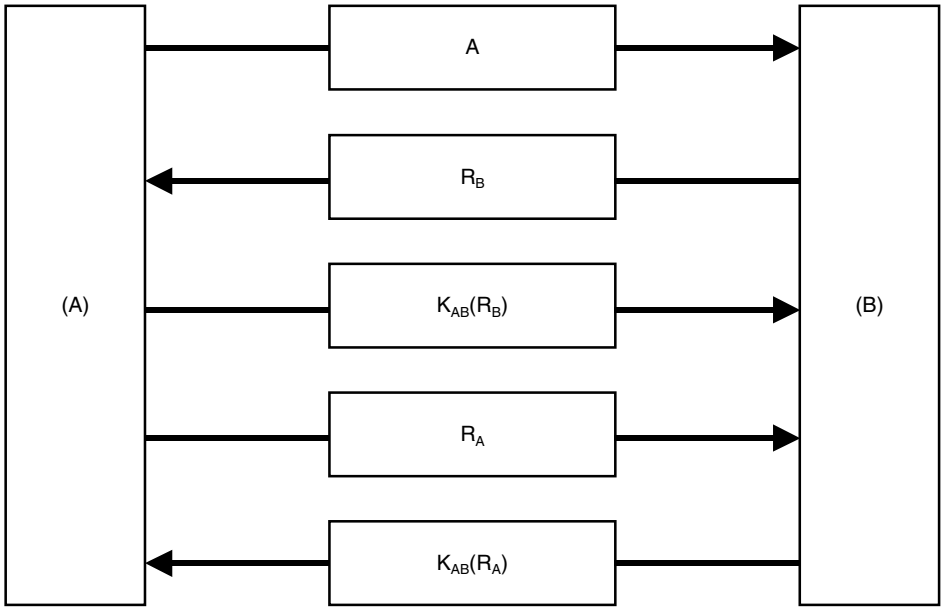


Figure 9.6 Challenge/response protocol—two way authentication.

Message 2

- Party B selects (generates) a large random number R_B and sends it to A in plaintext—the challenge.

Message 3

- Party A encrypts message 2 with the key shared with B and sends the ciphertext, $K_{AB}(R_B)$, back to B.
- When B receives this message, he knows it comes from A, since no one else knows K_{AB} and therefore could not have generated message 3. Also, since R_B was a very large random number (128 bit+), it is unlikely to have been used previously.
- Although B is sure that he is talking to A, A has no assurance that she has been communicating with B.

Message 4

- Party A picks a large random number R_A and sends it to B in plaintext.

Message 5

- Party B responds with $K_{AB}(R_A)$.
- Party A is now sure she is communicating with B.
- If a session key is to be generated, A can pick one, K_S, and send it to B encrypted with K_{AB}.

The protocol described requires a five-message transaction. It is possible to reduce this to a three-message transaction, however, the shortened three-message protocol is vulnerable to a reflection attack if the intruder T can open a multiple session with B.

The following three rules assist in overcoming the three-message vulnerability:

- Have the initiator prove who he or she is before the responder has to. In this case, B gives away valuable information before T has to give any evidence of who he or she is.
- Have the initiator and responder use different keys for proof, even if this means having two shared keys, K_{AB} and K'_{AB}.
- Have the initiator and responder draw their challenges from different sets. For example, the initiator must use even numbers and the responder must use odd numbers.

Digital Signatures

Digital signatures are required as replacements for traditional handwritten signatures. The requirement is for one party to send to another a message signed in such a way that:

- The receiver can verify the claimed identity of the sender.
- The sender cannot later repudiate the contents of the message.
- The receiver cannot possibly have concocted the message him- or herself.

Secret Key Signatures

The secret key signature concept is a system where by a trusted central authority (Certificate Authority, or CA) has possession of and knows all users' secret keys. Thus, each user must generate a personal key and deposit it with the CA in a manner that does not reveal it to any third party, as follows:

- Party A wants to send a signed plaintext message (P) to party B.
- Party A generates $K_A(B, R_A, t, P)$, where t is the timestamp, and sends it to the CA.
- The CA sees that the message is from A, decrypts it, and sends a message to B.
- The message contains the plaintext of A's message and is signed $K_{CA}(A, t, P)$.
- The timestamp is used to guard against the replaying of recent messages reusing R_A.

The shortcoming of the secret key system is that all parties who wish to communicate must trust a common third party—the CA. Not all people wish to do this.

Public Key Cryptography

Public key cryptography (see Figure 9.7) can assist in removing the key deposit process. The assumption is that public key encryption and decryption algorithms have the property that $E[D(P)] = P$, in addition to the usual property that $D[E(P)] = P$ (since RSA has this property, it is not unreasonable).

Assuming the previously mentioned conditions are in effect:

- Party A can send a plaintext message to party B by sending $E_B[D_A(P)]$. Party A can do this, since she knows her own private decryption key, D_A, as well as B's public key, E_B.
- When B receives the *message, he transforms it using his own private key. This yields* D_A(P).
- The text is stored in a safe place and then decrypted using E_A to get the original plaintext.
- If subsequently A denies having sent the message to B, B can produce both P and $D_A(P)$.
- It can be verified that it is a valid message encrypted by D_A by applying E_A to it.

Since B does not have A's private key, the only way B could have acquired the message was if A sent it. If A discloses her secret key, then the message could have come from anyone.

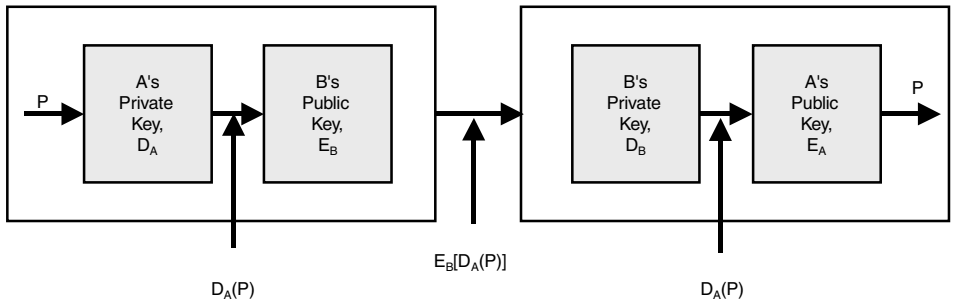


Figure 9.7 Public key signatures.

Summary

The transition to packet-routed networks means that we now share transport channels. This has increased the need for authentication and encryption. The greater the distance we can deliver (the more robust we make the authentication and encryption process), the more value we confer but the greater the overhead in terms of processor bandwidth, processing delay, and memory/code footprint.

Authentication and encryption are part of our overall end-to-end delay budget, but in turn, authentication can be compromised by delay and delay variability, particularly when time-sensitive challenge-response algorithms are used. Firewalls and virus scanning techniques can add many hundreds of milliseconds to our end-to-end delay budget but still have to be taken into account when dimensioning quality of service service level agreements (QoS SLAs).

From the perspective of a digital cellular handset, it makes considerable sense to use the smart card SIM/USIM as the basis both for over-the-air and end-to-end encryption, particularly since hardware coprocessors are now available on the smart card to minimize processing delay. For maximum flexibility, it could be argued that it is better to have authentication and encryption implemented in software at the application layer. Pragmatically, the best option is to integrate SIM/USIM-based admission control with an application layer user interface.

In a packet-routed network, the IP protocol stack may also implement packet-level security. This allows a virtual private network or networks to be deployed within a public IP network. Care must be taken, however, to ensure that network performance does not become protocol-limited. (We revisit IP protocol performance in our later chapter on network software.)

Specialist users can be supported either within private networks or virtual private networks by providing session-specific, location-specific, user group- or implementation-specific keys that can also be given conditional access status (preemption rights). This supports closed user groups and user group reconfiguration.

Key life can be difficult to manage, particularly with multiple user groups where group membership is highly dynamic. Note also that in specialist radio networks, there may be no network—that is, users are talking back-to-back between handsets. In a specialist radio network, a session can be defined as the time during which the press-to-talk key on the radio is depressed. When the PTT is released, the session is completed.

As most specialist users expect virtual instant access to a channel or virtual instant access into a group call, it is imperative that access and authentication protocols work within very strictly defined time limits.

In private mobile radio systems equipped with in-band tone signaling (tone signaling is still sometimes used in taxi radios) the *on to channel rise time*, the time taken to acquire a channel, would typically be 180 ms. Authentication and access protocols therefore have to be close to this in terms of performance and certainly should not introduce more than 250 ms of access delay. Early attempts to produce specialist user group services over GSM resulted in a call set/session setup time of 5 seconds—really not acceptable—an example of protocol performance limitation.

We revisit dynamic user groups in Chapter 17 when discussing mobile IP in ad hoc networks in the context of traffic shaping protocols.

CHAPTER 10

Handset Software Evolution

Our last chapter provided some examples of the trade-offs implicit in software/hardware partitioning when we discussed authentication and encryption. Implementing a function in software provides lots of flexibility, but execution can be relatively slow. Implementing a function in hardware means we lose flexibility, but the execution is much faster. A hardware implementation may also be more secure—less easy to compromise or attack without leaving visible evidence. In hardware, an instruction might typically be expected to execute in four clock cycles. In software, it can take 50 to 60 cycles—the cost of flexibility. This chapter explores the evolution of software in the hardware versus software dilemma.

Java-Based Solutions

Over the past 10 years there has been a move to make software, particularly application layer software (sometimes called platform operating software), easier to write. Java is one example. Originally introduced in 1995 as interoperative middleware for set-top boxes, Java compiles high-level code into Java byte codes.

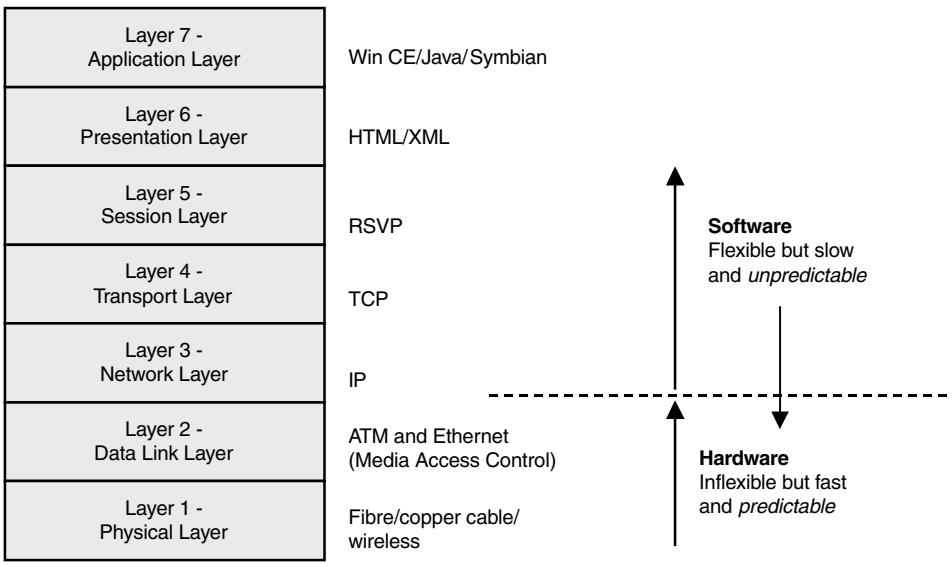
Developers write a version of the Java Virtual Machine in their own hardware's instruction set to run the byte codes. The advantage is that this makes the code fairly portable both between devices and applications. The disadvantage is that you do not have much backward compatibility and you can find yourself using rather more code (memory space and processor bandwidth) than you had originally intended. You would normally run Java on top of an existing real-time operating system.

Over the past 3 to 5 years, microcontroller architectures have been introduced that are specifically designed to support Java byte code execution. However, performance still tends to be hard to achieve within a small processor and memory footprint. ARM (Advanced RISC Machines) has a product known as ARM926EJ-S that supports accelerated Java execution. The byte code processing runs alongside the 32-bit (and reduced 16-bit) instruction set, which means the hardware resources are multiplexed. However, there is insufficient code memory space to directly execute all the Java byte code instructions, so only 134 out of 201 instructions are executed in hardware; the remainder are emulated. The decision then has to be made as to which byte codes should be chosen for direct execution.

ARC, Infineon, and Hitachi have similar Java-friendly microcontroller architectures, but essentially the trade-off is always cost and power budget against speed of execution.

Returning to our seven-layer model (see Figure 10.1), we can say that platform software in the application layer is flexible but unpredictable in terms of execution delay and execution delay variability. As we move down the protocol stack, software performance has to become faster and more predictable, and it becomes increasingly attractive to use hardware to meet this performance requirement.

This continues to be a challenge for Java-based software in embedded applications. The portable byte code instruction set can be implemented in hardware, but this requires processor bandwidth, memory bandwidth, and power—usually rather more than an embedded processor has available. The byte code instruction set can be emulated, but this absorbs memory and is rather inefficient and slow. The answer is to integrate a Java Virtual Machine onto a dedicated processor or coprocessor, but this again absorbs precious space and power resources. Limiting the Java classes installed helps but results in less application flexibility (and a rather inconsistent user experience when less-often-called Java instructions are used).



*Open Systems Interconnection.

Figure 10.1 Meeting hardware/software performance requirements.

Presently efforts are being made to implement a complete Java machine on a smart card supporting J2SE (Java2 Standard Edition) shoehorned into 55 kb on an 8-bit microcontroller. Future 16- and 32-bit smart card microcontroller architectures will make Java-based smart cards easier to implement.

So with Java we are effectively paying a price (cost, processor overhead, and memory/code space) for the privilege of device-to-device and application-to-application portability. We have to convince ourselves that the added value conferred exceeds the added cost. Microsoft offers us similar trade-offs, but here the trade-off tends to be backward compatibility (a privilege that implies additional code and memory space).

Both Java and Microsoft claim to be able to support soft and hard real-time operation, but in practice, exposure to interrupts (a prerequisite for flexibility) destroys determinism, and it is hard to deliver consistent predictable response times because of the wide dynamic range of the applications supported. This is why you often still find a standalone RTOS looking after critical real-time housekeeping jobs like protocol stack execution.

Developing Microcontroller Architectures

Platform Software Solutions (Symbian/Java and Microsoft CE) is developing in parallel with microcontroller architectures. The aim is to deliver better real-time performance by improving memory management (hardware and software optimization) and by supporting more flexible pipelining and multitasking. ARM 7 was a success in the late 1990s because it took GSMs 100,000 lines of source code and made it run about six times faster than the existing 8-bit microcontrollers being used.

A similar change in architecture will be needed, arguably, to support the highly asynchronous multitasking requirements implicit in managing and multiplexing complex multimedia (multiple per-user traffic streams). We are beginning to see similar changes in DSP architecture (for example, the Siroyan example used in Chapter 5, on handset hardware) and in memory architecture.

Software has an insatiable appetite for memory, and access delay is becoming an increasingly important part of the delay budget. If you were to dissect a typical Random Access Memory, you would find almost 60 percent of the die area and 90 percent of the transistor count dedicated to memory access aids. Memory manufacturers are beginning to include (yes, you have guessed already) integrated microprocessors—sometimes known as Intelligent RAM, or IRAM). This means that:

- If you are a microcontroller manufacturer, you add DSP functionality and memory to your product.
- If you are a DSP manufacturer, you add microcontroller and memory functionality to your product.
- If you are a memory manufacturer, you add DSP and a microcontroller to your product.

The software then struggles to adapt to these changing hardware form factors.

Access times are, of course, also a function of operating voltage. Flash memory access times might vary between 70 and 100 ns depending on whether the device is running at 1.8 or 2-5 V. High-speed (very power hungry) S-RAM can reduce access

delay to between 3 and 8 ns. A 5-Volt flash memory might deliver 100 ns of access delay running at 5 V, which could increase to 200 ns if reduced to 3 V.

Hardware Innovations

Application performance, or at least perceived application performance, can be improved by exploiting additional information available from new hardware components in the handset, for example, adding a digital compass to a handset so the handset software knows which direction it is pointing in. Microsoft Research has also proposed adding in a linear accelerometer and proximity sensing, (using the existing infrared devices already embedded in the handset, and user touch detection–capacitive sensing).

The software can therefore turn the device on when it is moved, as follows:

- Moving the phone toward your head enables the phone facilities.
- Putting it flat on the desk enables the keyboard and speakerphone facilities.
- Turning the device 90° when vertical switches the display from portrait to landscape.
- Moving closer to the device zooms the screen font.
- Walking along with the device in your pocket would mean the vibration alert would be turned on.

Microsoft is also proposing a 360° Web cam (or RingCam) for conference meetings. The general principle is that all this then synchronizes seamlessly with your desktop-based applications.

Add-in Modules

Additional hardware/software functionality can either be built into the product or added into an expansion slot. The choice here is to use a memory card slot (for instance, the Sony Memory Stick) or a PC card. PC cards come in three thicknesses:

- Type 1 (3.5 mm)
- Type II (5mm)
- Type III (10.5 mm); incompatible with most PDAs (too thick)

PC cards can either have their own power supply or can be parasitic; that is, they can take power from the host device. PC cards could host a hard disk or a wireless modem or a GPS receiver or a Bluetooth transceiver. Power drain on the host device can be significant.

Hardware does still provide a convenient mechanism for delivering software value. The value of the wireless PC modem is as much in the GSM software protocol stack as in the wireless RF IC. A plug-in CMOS imaging device with embedded MPEG-4 encoder/decoder and memory expansion for image and video storage is an example of add-on plug-in hardware/software value. The combined GPS receiver and MPEG-4

encoder/decoder from Sony goes into the Sony Memory Stick expansion port. The CMOS imager is a 100,000-pixel device. These products are described by Sony as personal entertainment organizers, or PEOs.

Looking to the Future

3G handset software must be capable of managing and multiplexing multiple per-user traffic streams, qualifying the radio bandwidth and network bandwidth requirements by taking into account information provided by, for example, the MPEG-4 encoder. The content itself may be capable of determining its bandwidth requirements (declarative content, or content that can declare its bandwidth quantity and quality needs).

The software then has to be capable of negotiating with the network, which implies an intimate relationship with network-based admission control procedures (we cover these in detail in Part IV of this book on network software). This would involve in the future the qualification of least-cost routing opportunities—but this is unlikely to be very appealing to the network operator.

MPEG-4, MPEG-7, and MPEG-21 provide a relatively stable and well-documented standards platform on which software added value can be built. MPEG-7-based image search engines, as one example, will potentially revolutionize image surveillance as an added value opportunity.

Given that much of the future value generation will be subscriber-based (subscriber-generated added value), handset software becomes progressively more important. The ability to develop session persistency and session complexity is a particularly important prerequisite, as is the ability to manage and multiplex highly asynchronous traffic (bursty bandwidth), including buffer management.

It seems to be generally assumed that there will be a multiplicity of hardware and software form factor in the future. This is *not* a good idea. Hardware needs to talk to hardware, and software needs to talk to software. What is needed is a de facto dominant hardware and software form factor for the handset and a dominant network hardware and software form factor. Ideally (from a technical perspective), this would all be supplied by one vendor, but this might prove rather expensive. To use multiple vendors but avoid device diversity is probably the best technical/commercial compromise.

Authentication and Encryption

Authentication and encryption both confer value to a session-based exchange between two or more people or two or more devices. The more robust the authentication and encryption process, the more value it has but the more expensive it is to implement in terms of processor, memory, and transmission bandwidth.

Authentication can be used to bring together complex user groups who can interact in complex ways. An event can trigger sudden loading on the network. In such circumstances, it is very important that authentication and admission algorithms remain stable. If a number of devices wish to send information at the same time and insufficient instantaneous bandwidth is available, then the software has to be sufficiently intelligent just to buffer the data or slow the source.

Authentication and encryption also enables m-commerce. The ability to bring things via a mobile handset implies the need to barter and negotiate and the need to research price and product/service availability. This would suggest an increasing application space for agent technology.

Agent Technology

Agents are software entities that can access remote databases—software objects that can transport themselves from (electronic) place to (electronic) place.
The following are typical agent capabilities:

- An agent can travel to meet and then interact (for example, negotiate) with another agent.
- Agents can be given a go instruction with a ticket that confers an authority to meet, refer, negotiate, buy, sell, or barter.
- On arrival, the agent presents a petition (for example, the requirement, how long the agent can wait).
- Agents can gather, organize, analyze, and modify information.
- Agents can have limits: time (for example, a 5-minute agent), size (a 1-kbyte agent), or spending (a $1 million agent).
- Because it has authority, an agent can negotiate locally without referral back to its master, which means it is well suited to being disconnected from a network.
- You can also send an agent instructions and messages, such as go to sleep, wake up, or welcome back home.

The problem with agents is that they are analogous to computer viruses—in that, they work in a similar way though without malicious intent. Agents therefore need very good *consistent* authentication. Given that you are effectively allowing software devices to spend money on your behalf, then it is important to have consistent rule management. There is an implicit need to establish and maintain trust between people and machinery. The need to establish trust has to involve mutually suspicious machines (machines that must prove their identity to one another) and mutually suspicious agents (agents that must prove their identity and authority to buy or sell to each other).

It is difficult to establish a consistent and stable trust hierarchy, and to date this has prevented the wide scale deployment of agent technologies. The SIM/USIM-enabled smart card is arguably the pivotal software/hardware component needed to make agent technology realizable on a given basis. Unfortunately, to date, smart card penetration in the United States has been significantly slower than the rest of the world, and this has hindered mass market adoption of agent-based services in digital cellular networks. The gradual integration of the SIM/USIM (a work item in 3GPP1 *and* 3GPP2) will help bridge the gap between the United States and the RoW (rest of the world) agent technology platforms.

Summary

As we will see in our next 10 chapters, it is increasingly important to qualify how handset hardware and software impacts on network hardware and software topology. Specifically, we must qualify how the value generated by handset hardware and software form factor and functionality must be preserved as the product (authenticated and encrypted rich media, parallel application streaming, e-commerce, m-commerce, and so forth) is moved into and through the core network.

PART

Three

3G Network Hardware

CHAPTER

11

Spectral Allocations—Impact on Network Hardware Design

In Parts I and II of the book we covered handset hardware and handset software form factor. We said handset hardware and software has a direct impact on offered traffic, which in turn has an impact on network topology. The SIM/USIM in the subscriber's handset describes a user's right of access to delivery and memory bandwidth, for example, priority over other users. Handset hardware dictates image and audio bandwidth and data rate on the uplink (CMOS imaging, audio encoding). Similarly, handset hardware dictates image and audio bandwidth and data rate on the downlink (speaker/headset quality and display/display driver bandwidth).

In this part of the book we discuss how network hardware has to adapt. We have defined that there is a need to deliver additional bandwidth (bandwidth quantity), but we also have to deliver bandwidth quality. We have defined bandwidth quality as the ability to preserve product value—the product being the rich media components captured from the subscriber (uplink value).

Searching for Quality Metrics in an Asynchronous Universe

Delay and delay variability and packet loss are important quality metrics, particularly if we need to deliver consistent end-to-end application performance. The change in handset hardware and software has increased application bandwidth—the need to simultaneously encode multiple per-user traffic streams, any one of which can be highly variable in terms of data rate and might have particular quality of service

requirements. This chapter demonstrates how offered traffic is becoming increasingly asynchronous—bandwidth is becoming burstier—and how this exercises network hardware.

In earlier chapters we described how multiple OVSF codes created large dynamic range variability (peak-to-average ratios) that can put our RF PAs into compression. This is a symptom of bursty bandwidth. On the receive side of a handset or Node B receiver, front ends and ADCs can be put into compression by bursty bandwidth (and can go nonlinear and produce spurious products in just the same way as an RF PA on the transmit path). As we move into the network, similar symptoms can be seen. Highly asynchronous bursty bandwidth can easily overload routers and cause buffer overflow. Buffer overflow causes packet loss. Packet loss in a TCP protocol-based packet stream triggers "send again" requests, which increase delay and delay variability and decrease network bandwidth efficiency.

We need to consider in detail the impact of this increasingly asynchronous traffic on network architectures and network hardware. In practice, we will see that neither traditional circuit-switched-based architectures nor present IP network architectures are particularly well suited to handling highly asynchronous traffic. We end up needing a halfway house—a circuit-switched core with ATM cell switching in the access network, both optimized to carry IP-addressed packet traffic.

In the first chapter of this Part, we study the RF parts of the network and how the RF subsystems need to be provisioned to accommodate bursty bandwidth. We will find that adding a radio physical layer to a network implicitly increases delay and delay variability. It is therefore particularly important to integrate radio layer and network layer performance in order to deliver a consistent end-to-end user experience.

Typical 3G Network Architecture

Figure 11.1 shows the major components in a 3G network, often described as an IP QoS network—an IP network capable of delivering differentiated quality of service. To achieve this objective, the IP QoS network needs to integrate radio physical layer performance and network layer performance. The IP QoS network consists of the Internet Protocol Radio Access Network (IPRAN) and the IP core replacing the legacy Mobile Radio Switch Center (MSC).

The function of the Node B, which has replaced the base station controller (BTS), is to demodulate however many users it can see, in RF terms, including demodulating multiple per-user traffic streams. Any one of these channel streams can be variable bit rate and have particular, unique QoS requirements. Node Bs have to decide how much traffic they can manage on the uplink, and this is done on the basis of interference measurements—effectively the noise floor of the composite uplink radio channel. A similar decision has to be made as to how downlink RF bandwidth is allocated. The Node B also must arbitrate between users, some of whom may have priority access status. We refer to this process as *IPRAN interference-based admission control*.

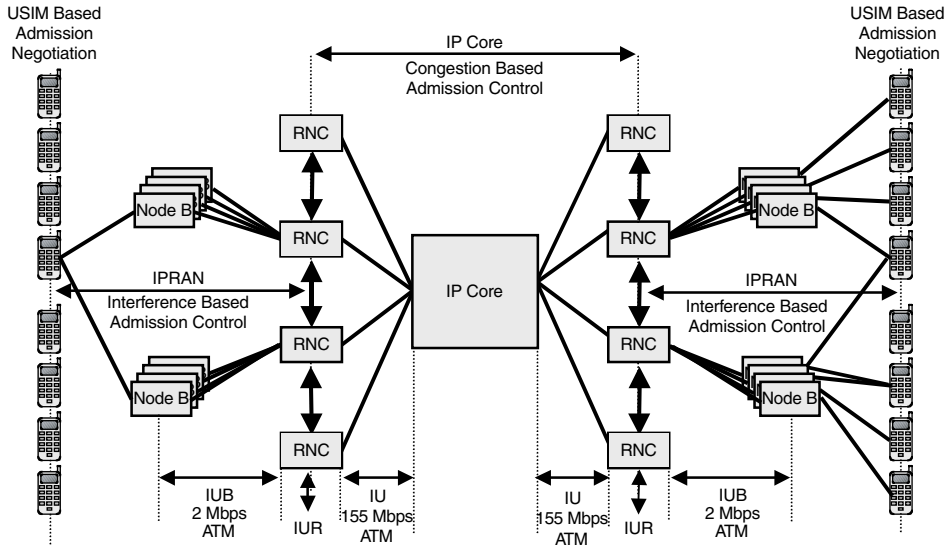


Figure 11.1 IP QoS network.

The RNCs job is to consolidate traffic coming from the Node Bs under its control. The RNC also has to load balance—that is, move traffic away from overloaded Node Bs onto more lightly loaded Node Bs and to manage soft handover—a condition in which more than one Node B is serving an individual user on the radio physical layer. The fourth handset down on the right-hand side of Figure 11.1, for example, is in soft handover between two Node Bs supported by two different RNCs. The handset is, more or less, halfway between two Node Bs. To improve uplink and downlink signal level, the RNC has decided that the handset will be served by two downlinks, one from each Node B. Both nodes will also be receiving an uplink from the handset. Effectively this means there will be two long codes on the downlink and two long codes on the uplink. The two uplinks will be combined together by the serving RNC, but this will require the serving RNC to talk to the other serving RNC (called the drift RNC). The same process takes place on the downlink.

The RNC has to make a large number of very fast decisions (we revisit RNC software in Chapter 17 in our section on network software), and the RNC-to-RNC communication has to be robust and well managed. The RNCs then have to consolidate traffic and move the traffic into the IP core. Admission control at this point is done on the basis of congestion measurements:

- Is transmission bandwidth available?
- If no transmission bandwidth is available, is there any buffer bandwidth available?

If no transmission bandwidth is available and no buffer bandwidth is available then packet loss will occur unless you have predicted the problem and choked back the source of the traffic.

RNC traffic management and inter-RNC communication is probably the most complex part of the IPRAN and is the basis for substantial vendor-specific network performance optimization. This is complex deterministic software executing complex decisions within tightly defined timescales. As with handset design, there is considerable scope for hardware coprocessors and parallel hardware multitasking in the RNC. As with handset design, we will show that network performance is also dependent on the RF performance available from the Node B—the Node B's ability to collect highly bursty traffic from users and deliver highly bursty traffic to users.

The Impact of the Radio Layer on Network Bandwidth Provisioning

Table 11.1 shows how the aggregated bit rate increases as we move into the network. The highly asynchronous traffic loading is supported on a 2 Mbps ATM bearer between the Node B and RNC and a 155 Mbps ATM bearer, or multiple bearers, between the RNC and IP core. The job of the IP core is to process traffic, and (we assume) a fair percentage of the traffic will be packetized. This is a packet-routed or, more accurately, packet-queued network.

The radio physical layer is delivering individual users at user bit rates varying between 15 kbps and 960 kbps. This is aggregated at the Node B onto multiple 2 Mbps ATM wireline transport (copper access), which is aggregated via the RNC onto multiple 155 Mbps ATM (copper). This is aggregated onto 2.5, 10, or 40 Gbps copper and optical fiber in the network core. The IP core may also need to manage highly asynchronous traffic from wireline ADSL/VDSL modems (offering bit rates from 56 kbps to 40 Mbps).

Table 11.1 Access Bandwidth/Network Bandwidth Bit Rates

ACCESS BANDWIDTH	AGGREGATED BIT RATE INCREASES NETWORK BANDWIDTH		
Handsets	Node B	RNC	Core network
RF	Copper		Copper and optical fiber
15 kbps to 960 kbps	25 Mbps to 155 Mbps	155 Mbps to 622 Mbps	2.5 to 10 to 40 Gbps
Wireline			
56 kbps to 8 Mbps to 40 Mbps (VDSL)			

Table 11.2 Time Dependency versus Bit Rate

HANDSETS	AIR INTERFACE	GIGABIT PACKET PROCESSING	TERABIT PACKET PROCESSING
Milliseconds	Microseconds	Nanoseconds	Picoseconds
1 in 10^3	1 in 10^6	1 in 10^9	1 in 10^{12}
For example: 10 ms frame rate	For example: 20 microsecond flight path	For example: OC48 at 2.5 Gbps = 65 nanoseconds 10 Gbps = 16 ns 40 Gbps = 4 ns	

As throughput increases, processing speed—and processor bandwidth—increases (see Table 11.2). Routers must classify packets, and perform framing and traffic management. If we have added packet-level security, the router must perform a deep packet examination on the whole packet header to determine the security context.

The Circuit Switch is Dead—Long Live the Circuit Switch

We could, of course, argue that it is difficult to match the performance or cost efficiency of a circuit switch. Financially, many circuit switches are now fully amortized. They provide definitive end-to-end performance (typically 35 ms of end-to-end delay and virtually no delay variability) and 99.999 percent availability. Circuit switches achieve this grade of service by being overdimensioned, and this is the usual argument used to justify their imminent disposal. However, as we will see, if you want equivalent performance from an IP network, it also needs to be overdimensioned and ends up being no more cost effective than the circuit switch architecture it is replacing.

Circuit switches are also getting smaller and cheaper as hardware costs go down. Ericsson claims to be able to deliver a 60 percent size reduction every 18 months and a 30 percent reduction in power budget.

Consider the merits of a hardware switch. An AXE switch is really very simple in terms of software—a mere 20 million lines of code, equivalent to twenty 3G handsets! Windows 98 in comparison has 34 million lines of code. A hardware switch is deterministic. Traffic goes in one side of the switch and comes out the other side of the switch in a predictable manner. A packet-routed network, in comparison, might have lost the traffic, or misrouted or rerouted it, and will certainly have introduced delay and delay variability.

As session persistency increases (as it will as 3G handset software begins to influence user behavior), a session becomes more like a circuit-switched phone call. In a hardware-switched circuit-switched phone call, a call is set up, maintained, and

cleared down (using SS7 signaling). In a next-generation IP network, a significant percentage of sessions will be set up, maintained, or cleared down (using SIP or equivalent Internet session management protocols).

A halfway house is to use ATM. This is effectively distributed circuit switching, optimized for bursty bandwidth.

Over the next few chapters we qualify IP QoS network performance, including the following factors:

- Its ability to meet present and future user performance expectations
- Its ability to deliver a consistent end-to-end user experience
- Its ability to provide the basis for quality-based rather than quantity-based billing

But let's begin by benchmarking base station and Node B performance.

BTS and Node B Form Factors

Node B is the term used within 3GPP1 to describe what we have always known as the base station. *Node* refers to the assumption that the base station will act as an IP node; *B* refers to base station. The Node B sits at the boundary between the radio physical layer (radio bandwidth) and the RNC, which in turn sits at the boundary between the IP radio access network (RAN) and the IP core network.

We have talked about the need for power budget efficiency in handset design and power budget efficiency in the switch. We also need power budget efficiency in the Node B so that we can get the physical form factor as small as possible. If we can keep power consumption low, we can avoid the need for fan cooling, which gives us more flexibility in where and how we install the Node B.

Typical design targets for a base station or Node B design would be to deliver a cost reduction of at least 25 percent per year per RF channel and a form factor reduction of at least 30 percent per year per channel.

Typical 2G Base Station Product Specifications

Table 11.3 gives some typical sizes and weights for presently installed GSM base stations supplied by Motorola. Although there are 195 × 200 kHz RF carriers available at 900 MHz and 375 × 200 kHz RF carriers available at 1800 MHz, it is unusual to find base stations with more than 24 RF carriers and typically 2 or 6 RF carrier base stations would be the norm. This is usually because 2 or 4 or 6 RF carriers subdivided by 8 to give 16, 32, or 48 channels usually provides adequate voice capacity for a small reasonably loaded cell or a large lightly loaded cell. Having a small number of RF carriers simplifies the RF plumbing in the base station—for example, combiners and isolators, the mechanics of keeping the RF signals apart from one another. Table 11.3 shows that hardware is preloaded with network software prior to shipment.

Table 11.3 Base Station Products: Motorola—GSM 900/1800/1900

RF CARRIERS		M-CELL 6 6	M-CELL 2 2	M CELL MICRO * 2
Dimensions HWD (m)	Indoor	1.76 × 0.71 × 0.47	1.0 × 0.7 × 0.45	0.62 × 0.8 × 0.19
	Outdoor	1.76 × 0.71 × 0.77	1.0 × 0.7 × 0.65	0.62 × 0.8 × 0.22
Weight (kg)	Indoor	234	93	30
	Outdoor	277	135	45
Output power (Watts)	GSM 900	20 W	20 W	2.5 W
	GSM 1800	8 W	8 W	2.0 W
Features include:		6-24 RF Carriers	Optical fiber interconnect	Integrated antenna Low form factor (depth) for wall mounting Hot shipping (Software loaded prior to site delivery)

Products from Nokia have a similar hardware form factor (see Table 11.4). This has the option of a remote RF head, putting the LNA (Low-Noise receive Amplifier) close to the antenna to avoid feeder losses. There is also the choice of weather protection (IP54/IP55; IP here stands for "intrinsic protection").

Table 11.4 Base Station Products—Nokia—GSM 900/1800/1900

	INDOOR SECTOR OMNI	INDOOR MINI	OUTDOOR STREET LEVEL		OUTDOOR ROOF TOP	
No. of carriers (in one cabinet)	3	2	2		2	
Dimensions HWD (m)	2.2 × 0.6 × 0.5	1.4 × 0.6 × 0.5	1.8 × 0.9 × 0.7		1.2 × 0.8 × 0.6 Remote RF head	
Features			IP55		IP54	
			5	5	5	4
			Dust protected	Protected against water jets	Dust protected	Protected against splashing water

Table 11.5 Base Station Products—Ericsson—GSM 900/1800/1900

	MACROCELL RBS 2202 INDOOR	MACROCELL RBS 2102 OUTDOOR	MICROCELL OR LOW CAPACITY RBS 2101 INDOOR OR OUTDOOR
No. of RF carriers	6	6	2
Features			Mast-mounted LNAs
			(to maximize uplink sensitivity)
			Installation database; hardware tracking; Software revision tracking; (common to all Ericsson base station and related modular products)

A Nokia PrimeSite product weighs 25 kg in a volume of 35 liters. This is a single RF carrier base station with two integrated antennas to provide uplink and downlink diversity.

Similar products are available from Ericsson, also including mast-mounted LNAs to improve uplink sensitivity. Table 11.5 gives key specifications, including number of RF carriers, and highlights additional features such as the inclusion of automatic hardware and software revision tracking.

The GSM specification stated that different vendor BTS products should be capable of working with different vendor BSCs. As you would expect, all BTSs have to be compatible with all handsets. Because GSM is a constant envelope modulation technique, it has been possible to deliver good power efficiency (typically >50 percent) from the BTS power amplifiers and hence reduce the hardware form factor. This has been harder to achieve with IS95 CDMA or IS136 TDMA base stations because of the need to provide more linearity (to support the QPSK modulation used in IS95 and the $\pi/4$DQPSK modulation used in IS136 TDMA).

Table 11.6 shows the specification for a Motorola base station capable of supporting AMPS, CDMA, and TDMA. The CDMA modem provides 1.25 MHz of RF channel bandwidth (equivalent to a GSM 6 RF carrier base station) for each RF transceiver with a total of 16 transceivers able to be placed in one very large cabinet to access 20 MHz of RF bandwidth—the big-is-beautiful principle. The linear power amplifier weighs 400 kg! The products are differentiated by their capacity—their ability to support high-density, medium-density, or very localized user populations (microcells).

Table 11.6 Base Station Products—Motorola—AMPS/N-AMPS/CDMA/IS136 TDMA

		HIGH DENSITY	MEDIUM DENSITY	M CELL MICRO
No. of RF carriers	Analog	96	48	1
	Digital	80 channel cards		
	(CDMA)	16 transceivers	160	40
		320 channels		
Dimensions HWD (m)	Indoor			Indoor or outdoor
	Site interface frame	$1.8 \times 0.8 \times 0.6$		
	RF modem	$1.8 \times 0.8 \times 0.6$		$0.7 \times 0.6 \times 0.6$
	Linear power amp	$2.1 \times 0.8 \times 0.6$	$2.1 \times 0.8 \times 0.6$	Note: Depth incompatible with wall mounting
Weight (kg)	Site interface frame	200		
	RF modem	340		
	LPA	400	350	

Table 11.7 shows a parallel product range from Ericsson for AMPS/IS136. These are typically 10 W or 30 W base stations (though a 1.5 W base station is available for the indoor microcell). The size is measured in terms of number of voice paths per square meter of floor space. Again, the product range is specified in terms of its capacity capabilities (ability to support densely populated or less densely populated areas).

Table 11.7 Base Station Products—Ericsson-AMPS/D-AMPS 800 MHz, D-AMPS 1900 MHz

	MACROCELL RBS 884 MACRO-INDOOR	RBS 884 COMPACT (INDOOR/OUTDOOR) ROOF MOUNT/ HIGH CAPACITY DENSELY POPULATED AREAS	MICROCELL RBS 884 MICRO (INDOOR)
No. of carriers	36 (10 W)		
	36 (30 W)	-	-
No. of transceivers	16 (10 W) 8 (30 W)	10 (10 W)	10 (1.5 W)
No. of voice Paths per 1 m^2 of floor space	213		
Max. voice channels		23 (Analog) 71 (Digital)	
features	Autotune combining	Hot repair	Hybrid combining
	Radio frequency loop test	Hybrid combining	
	VSWR alarms		
	RSSI measurement		
Channel spacing	Min. 360 kHz	Min. 120 kHz	Min 120 kHz
(Measure of RF filtering discrimination)			
Receive Sensitivity	Analog -118 dBm for 12 dB SINAD Digital -112 dBm for 3% BER	Same	Same

As with IS95 CDMA, here there is a need to support legacy 30 kHz AMPS channels (833 × 30 kHz channels within a 25 MHz allocation). This implies quite complex combining. The higher the transmitter power, the more channel spacing needed between RF carriers in the combiner. Note also that if any frequency changes are made in the network plan, the combiner needs to be retuned. In this example, the cavity resonators

and combiners can be remotely retuned (mechanically activated devices). If there is a power mismatch with the antenna because of a problem, for example, with the feeder, then this is reflected (literally) in the voltage standing wave ratio (VSWR) reading and an alarm is raised.

The receiver sensitivity is specified both for the analog radio channels and the digital channels. 12-dB SINAD is theoretically equivalent to 3 percent BER. This shows that these are really quite complex hardware platforms with fans (and usually air conditioning in the hut), motors (to drive the autotune combiners), and resonators—RF plumbing overhead. These products absorb what is called in the United States *windshield time*—time taken by engineers to drive out to remote sites to investigate RF performance problems.

In the late 1980s in the United Kingdom, Cellnet used to regularly need to change the frequency plan of the E-TACS cellular network (similar to AMPS) to accommodate additional capacity. This could involve hundreds of engineers making site visits to retune or replace base station RF hardware—the cost of needing to manage lots of narrowband RF channels.

3G Node B Design Objectives

It has been a major design objective in 3G design to simplify the RF hardware platform in order to reduce these costs. As with the handset, it is only necessary to support twelve 5 MHz paired channels and potentially seven 5 MHz nonpaired channels in the IMT2000 allocated band instead of the hundreds of channels in AMPS, TDMA, or GSM.

Unfortunately, of course, many Node Bs will need to continue to support backward compatibility. Broadband linear amplifiers and software configurable radios are probably the best solution for these multimode multiband Node Bs. We look at the RF architecture of a broadband software radio Node B in a case study later in this chapter.

2G Base Stations as a Form Factor and Power Budget Benchmark

In the meantime, customer expectations move on. Over the past 5 years, GSM base stations have become smaller and smaller. Ericsson's pico base station is one example, taking the same amount of power as a lightbulb. Nokia's in-site picocellular reduces form factor further (an A4 footprint).

And even smaller GSM base station products are beginning to appear. The example in Figure 11.2 weighs less than 2 kg and consumes less than 15 W of power.

The continuing reduction of the form factor of 2G base stations represents a challenge for the 3G Node B designer. Vendors need to have a small Node B in the product portfolio for in-building applications. The general consensus between designers is that this should be a single 5 MHz RF carrier Node B weighing less than 30 kg, occupying less than 30 liters. The example shown in Figure 11.3 meets these design requirements. This is a pole-mounted transceiver and as such may be described as having no footprint. It is convection cooled, consuming 500 W for a 10 W RF output and is 500 mm high.

Figure 11.2 Nano base stations (GSM—2G) from ip.access (www.ipaccess.com).

Node B Antenna Configuration

The Node B hardware determines the antenna configuration. The Siemens/NEC Node B shown in Figure 11.3 can either be used on its own supporting one omnidirectional antenna (1 × 360°) or with three units mounted on a pole to support a three-sector site (3 × 120° beamwidth antennas).

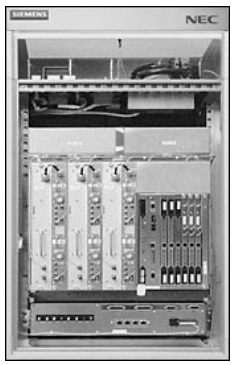
All other Node Bs in this particular vendor's range at time of writing are floor mounted, mainly because they are too heavy to wall mount or pole mount. The example shown in Figure 11.4 weighs 900 kg and occupies a footprint of 600 × 450 mm and is 90 cm high. It can support two carriers across three sectors with up to 30 W per carrier, sufficient to support 384 voice channels.

The same product can be double-stacked with a GSM transceiver to give a 1 + 1 + 1 + 6 configuration, one 5 MHz RF carrier per sector for UMTS and a 6 RF carrier GSM BTS, which would typically be configured with 2 × 200 kHz RF channels per sector. The combined weight of both transceivers is 1800 kg and combined power consumption is over 2 kW.

A final option is to use one of the family of Node Bs illustrated in Figure 11.5. These can be configured to support omnis (360°), four sector (4 × 90° beamwidth antennas), or six sector (6 × 60°), with RF power outputs ranging from 6 W to 60 W per RF carrier.

Figure 11.3 Siemens/NEC Node B (UMTS).

The configuration can also support omni transmit and sectorized receive (OTSR), which has the benefit of providing better receive sensitivity. The physical size is 600 × 450 mm (footprint) by 1400 mm high, and the weight is 1380 kg. Outdoor and indoor versions are available.



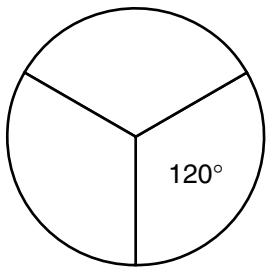


Figure 11.4 Siemens/NEC NB420 Macro Node B.

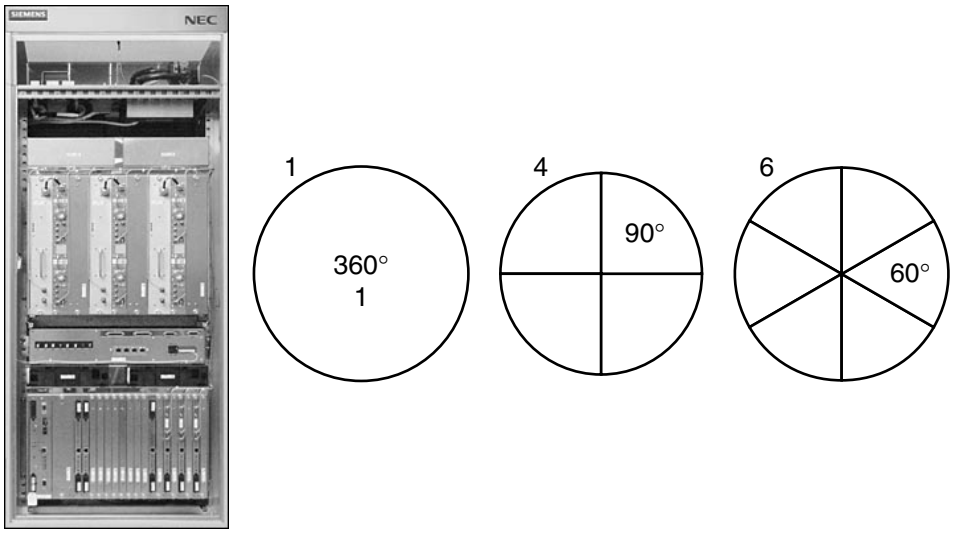


Figure 11.5 Siemens/NEC NB440 Macro Node B—more sectors, more RF carriers, smaller footprint.

The Benefits of Sectorization and Downtilt Antennas

Sectorization helps to provide more capacity and delivers a better downlink RF link budget (more directivity) and better uplink selectivity (which improves sensitivity). Many Node Bs also use electrical downtilt. We discuss smart antennas in Chapter 13; these antennas can adaptively change the coverage footprint of an antenna either to null out unwanted interference or to minimize interference to other users or other adjacent Node Bs. Electrical downtilt has been used in GSM base stations from the mid-1990s (1995 onward). By changing the elevation of the antenna or the electrical phasing, the vertical beam pattern can be raised or lowered. Figure 11.6 shows how the beam pattern can be adjusted to increase or decrease the cell radius. This can be used to reduce or increase the traffic loading on the cell by reducing or increasing the physical footprint available from the Node B. Adaptive downtilt can be used to change coverage as loading shifts through the day—for example, to accommodate morning rush hour traffic flows or evening rush hour loading.

For in-building coverage, an additional option is to have a distributed RF solution in which a Node B is positioned in a building and then the incoming/outgoing RF signals are piped over either copper feeder (rather lossy) or optical fiber to distributed antennas. The optical fiber option is preferable in terms of performance but requires linear lasers to take the (analog) RF signal and modulate it onto the optical fiber and a linear laser to remodulate the optical signal back to RF at the antenna.

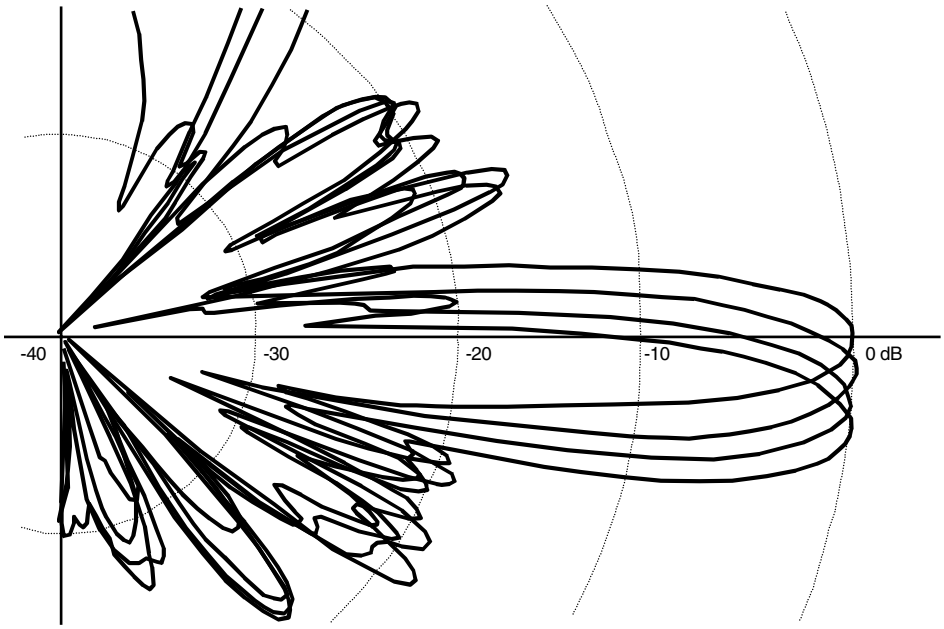


Figure 11.6 Electrical downtilt.

As we shall see in our next section on system planning, the 3G air interface is well suited to a fairly dense low-powered network (less noise rise is produced by adjacent Node Bs and less interference is visible at the Node B receiver). This places a premium on the need to design a small form factor (sub-30 kg) Node B product.

Node B RF Form Factor and RF Performance

Given that operators may be asked to share access hardware and given that operators have been allocated different RF carriers, it may also be necessary to produce small form factor Node Bs capable of processing more than one × 5 MHz RF carrier—ideally 60 MHz, though this is at present unrealistic in terms of digital sampling techniques.

The two major design challenges for Node B products are transmit linearity, including the ability to handle multiple downlink OVSF codes per user, and receive sensitivity, including the ability to handle multiple uplink OVSF codes per user. We have said that receive sensitivity can be improved by using electronic downtilt (reducing the exposure of the Node B to visible interference) and multiuser detection where the Node B uses the short codes embedded in each individual handset's offered traffic stream to cancel out unwanted interference energy. Multiuser detection is a longer-term upgrade (rather like frequency hopping was in GSM in the early 1990s).

Receive sensitivity is also a product of how well the radio planning in the network has been done and how well the Node B sites have been placed in relation to the offered traffic. We address these issues in the next section.

As with handsets, RF power budgets can be reduced by increasing processor overhead. For example, we can implement adaptive smart antennas on a Node B, which will provide significant uplink and downlink gain (potentially 20 or 25 dB). This reduces the amount of RF power needed on the downlink and RF power needed on the uplink. However, if the processor power consumption involved (to support the many MIPS of processing required) is high compared to the RF power saved, then very little overall gain would have been achieved. You will just have spent a lot of money on expensive DSPs.

As with handset design, DSPs can do much of the heavy lifting at bit level and symbol level but run out of steam at chip level. There is also a need for substantial parallel processing to support multiple users, each with multiple uplink and downlink OVSF code streams. These factors presently determine existing Node B form factor and functionality. The design objective has to be to balance good practical RF design with judicious use of DSPs and ASICs to deliver power-efficient processor gain. The requirement, as with GSM, is to keep power consumption for small Node Bs in the region of tens of Watts. This means it is easy to install Node Bs indoors without greatly adding to the landlord/hosting energy bill, and for outdoor applications, it provides the basis for solar-powered or wind-powered base station/Node B implementation.

Simplified Installation

IMT2000DS indoor picocells do not need to have a GPS reference. They can be reclocked by handsets moving into the building. This makes installation substantially easier. An engineer can walk into a building, fix a node B to the wall, plug it into a mains power outlet, plug it into a telephone line (the Node B has its own ADSL modem), turn it on, and walk way. If GPS was needed, the engineer would have to pipe a connection to a window so that the GPS antenna could see the sky. Small Node Bs do not incur the same neighborhood resentment as larger Node Bs (for one reason, they do not look like base stations), and the site is sometimes provided for free by the landlord, which is rarely the case for large outdoor sites.

Radio planning, as we will see later in this chapter, is also partly determined by the product mix of Node Bs available from each vendor. Typically, power outputs will be 40 W, 20 W, 10 W, 5 W, or less. Although some planners would argue the case for smaller numbers of larger, more powerful Node Bs, this goes against existing product and installation trends, which clearly point toward the need to maintain a small form factor (small volume/low weight). This in turn determines the choice of architecture used in the Node B design.

Node B Receiver Transmitter Implementation

In Chapters 2 and 3 we discussed the suitability of the digitally sampled IF superhet and the direct conversion receiver architecture for handset implementation. We concluded that either configuration was capable of meeting the handset specification but

that longer term, the DCR (or near-zero IF) could show a reduction in component count—especially in multistandard environments—although problems of DC offsets required considerable DSP power (baseband compensation).

The 3G Receiver

We also reviewed transmitter implementation and concluded that the architecture (OPLL) developed for cost reduction/multiband requirements in 2G could show similar benefits in 3G if the problem of processing both amplitude and phase components in the modulation (HPSK) could be overcome. We will now consider receiver and transmitter requirements in Node B implementation and assess whether the architectures discussed in the previous chapters are also suitable for Node B designs.

The Digitally Sampled IF Superhet

In analyzing the handset receiver/transmitter options, we recognized that the prime constraint on any decision was that of battery power requirement. To provide for the handset to access any of the 5 MHz channels in the 60 MHz spectrum allocation, a receiver front end tuning with a 12-step synthesizer is necessary to downconvert the selected channel to be passed through a 5 MHz bandwidth IF centered filter to the sampling ADC. The digitized single 5 MHz channel is then processed digitally to retrieve the source-coded baseband signal. This single-channel approach is adopted in the handset in order to comply with the low-power criteria.

If this single-channel approach were adopted in the Node B, where multiple RF channels may simultaneously be required, the requisite number of receivers would have to be installed. As the restriction of Node B power consumption is not as severe, an alternative approach can be considered.

The ideal approach is to implement a wideband front end, to downconvert the 12 5 MHz-wide channels, to pass a number (or all) of the channels through a wideband IF filter, and to sample and digitize this wider bandwidth of channels. The digitized channels would then be passed to a powerful digital processing capability that could simultaneously extract the downconverted baseband signals. The number of channels to be simultaneously processed would again be dependent on the power available both in implementing an RF front end of sufficient dynamic range and an ADC/DSP combination of sufficient processing capability.

Additionally, a greater dynamic range is required by the ADC and DSP, since in the multichannel environment, the channels may be at substantially different signal strengths and so dynamic range control cannot be used. If the IF gain were to be reduced by a strong signal channel, a weak signal channel would disappear into the noise floor.

The Direct Conversion Receiver (DCR)

We have demonstrated that the DCR is a suitable receiver configuration for single-channel operation in the handset. It is similarly suitable for single-channel operation in the Node B.

How will it perform in the multichannel environment?

Consider a wideband approach to receive four simultaneous channels. The receiver front end would still have a bandwidth of 60 MHz—to be able to operate across all 12 W-CDMA channels. The tuning front end would require a local oscillator (LO) having three discrete frequencies in order to downconvert the band in three blocks of four channels each. In the multicarrier receiver, the LO would be placed in the center of the four channels to be downconverted (received), as shown in Figure 11.7.

The output of the I and Q mixers would be the four channel blocks centered around 0 Hz each time the LO was stepped, as shown in Figure 11.8. A typical IC mixer having a Gilbert cell configuration can only achieve at best an IQ balance of some 25 to 30 dB. This means that if the channel 2 to channel 3 amplitude difference is greater than 30 dB, signal energy from channel 3 will transfer into, and hence corrupt, channel 2. Similarly, there will be an interaction between channels 1 and 4.

If we consider the problem of IQ imbalance, we find there are several causes. Typical causes are those of IC manufacturing and process tolerance, variation with supply voltage to the mixers, temperature variation, and other similar effects. This group of causes are predictable, constant effects that can be characterized at the production test stage and compensating factors inserted into the receive processing software.

Compensation can be affected in the digital processing stages by a process of vector rotation, feeding some Q signal into I, or I into Q as required to balance the system. The greater problem is that of IQ imbalance due to dynamic signal variation effects. These are unbalancing effects that cause the operating point of the mixers to shift with signal strength and radiated signal reflection and reentry effects. These effects have been described in Chapter 2 in our discussion of DC offset problems.

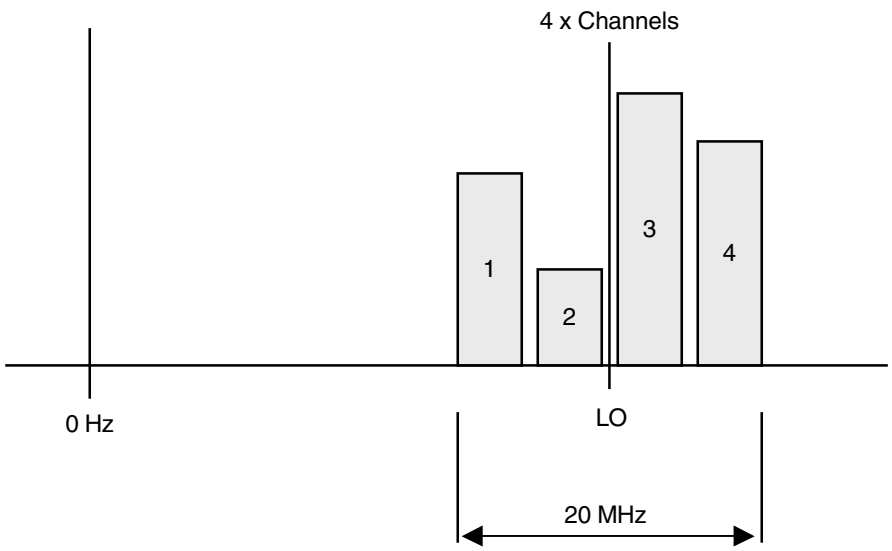


Figure 11.7 Local oscillator positioning for down conversion of channels 1 to 4.

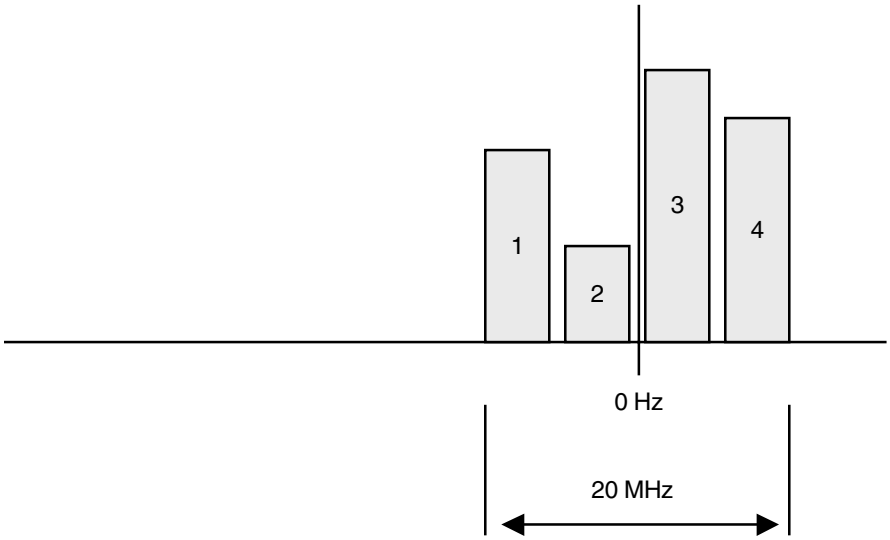


Figure 11.8 Channels 1 to 4 positioned around 0 Hz at mixer outputs.

Although the correction by vector rotation is a relatively simple digital process, the difficulty lies in estimating the instantaneous degree of compensation required given all the variables causing signal amplitude variation. For this reason, the DCR (and near-zero IF) is not chosen for Node B designs. However, the designer should always review the current capability of this technology at the start of any new design, as research is certainly being undertaken to increase the application of the DCR.

The 3G Transmitter

We have considered possible transmitter configurations in Chapter 2, and in Chapter 3, we introduced the need for improved linearity for handset (uplink) modulation. The uplink modulation is HPSK, and this requires considerable linearity to be engineered into the PA. The cost, as always, is increased power consumption. The need for Power-Added Efficiency (PAE) in the handset limits the options that can be used to provide linearity. The additional power available to us in the Node B provides us with a wider number of options.

The RF/IF Section

The application of a linear PA in Node B design should be considered. The downlink signal has QPSK modulation and so has greater amplitude variation than the uplink HPSK. Accordingly, greater linearity is required. There is also the issue of a wideband (multichannel) versus a narrowband (single-channel) approach.

In the Node B we have the option of using 12 separate RF transmitters for the 12 channels and combining their outputs at high power prior to the antenna feed, or, we can create a multichannel signal at baseband (or IF) and pass the composite signal through one high dynamic range, high linearity, high power amplifier. There is a large amount of information, analysis, discussion, and speculation of the benefits of one approach or the other, so we can confine ourselves to a review of linearizing options.

Envelope Elimination and Restoration (EER) and Cartesian approaches were introduced in Chapter 3, so we will consider briefly other alternatives.

Linearization methods fall broadly into two categories:

- Those that use a feedback correction loop operating at the modulation rate.
- Those that use a feedback correction process to update a feed-forward correction process, operating at a slower than modulation rate.

The former method is not particularly well suited to the wide modulation bandwidth of 3G (5 MHz for a single channel and up to 60 MHz for a multicarrier Node B).

Whilst it is relatively simple to extract the envelope from the input RF signal and limit the signal to give a constant envelope drive, there is some advantage to implementing this polar split at the signal generation stage within the DSP.

In particular, it is highly likely that the transfer function through the envelope amplifier and bias/supply modulation process will be nonlinear, and that the drive envelope function will need to be predistorted to compensate for this error. The predistortion factors can be held in a lookup table within the DSP, for example, and updated if necessary by some slow feedback loop from the transmitter output.

The EER approach can yield modest improvements in linearity for quite good efficiency (provided the envelope modulation amplifier efficiency is good). The modulation envelope bandwidth for W-CDMA is, however, quite large (approximately 5×5 MHz to include key harmonics), and the efficiency of switched mode modulation amplifiers falls off quite quickly at high switching rates.

Another method to be considered is RF synthesis. A synthesis engine converts the I and Q (Cartesian) representation of the modulation waveform into two frequency and phase modulated components, the vector sum of which is identical to the source signal (see Figure 11.9). Amplification of these two constant envelope waveforms is performed using Class C or Class F/S switching amplifiers for maximum efficiency, and the outputs are combined to give the composite high-power RF synthesized waveform. One of the key challenges with RF synthesis is the combination of these two high-power FM signals without losing much of the power in the combiner process.

The vector diagram in Figure 11.10 shows how the output signal is synthesized from the summation of the two constant envelope rotating vectors. Full output power occurs when the two vectors are in phase. Minimum output power is synthesized when the two vectors are 180 degrees out of phase. Using a DSP to generate the two constant envelope phase modulated components is quite feasible, since the algorithm is simple. The processing rate, however, must be very high to accommodate the bandwidth expansion of the nonlinear function involved, and the sample rate of the ADCs must also accommodate the bandwidth expansion of the FM modulated outputs. These high sampling rates and the corresponding high power consumption of the DSP and ADC components means that this approach is only feasible for nonportable applications at the present time.

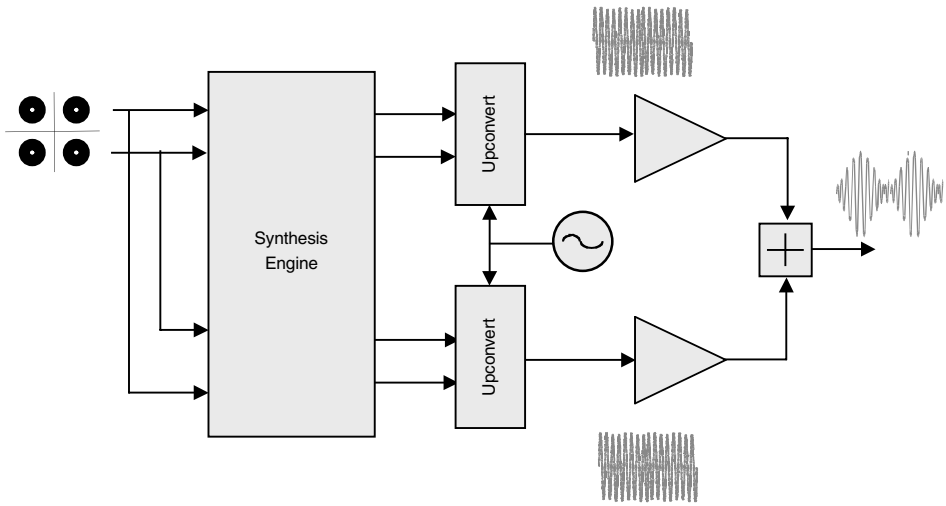


Figure 11.9 Basic RF synthesis operation.

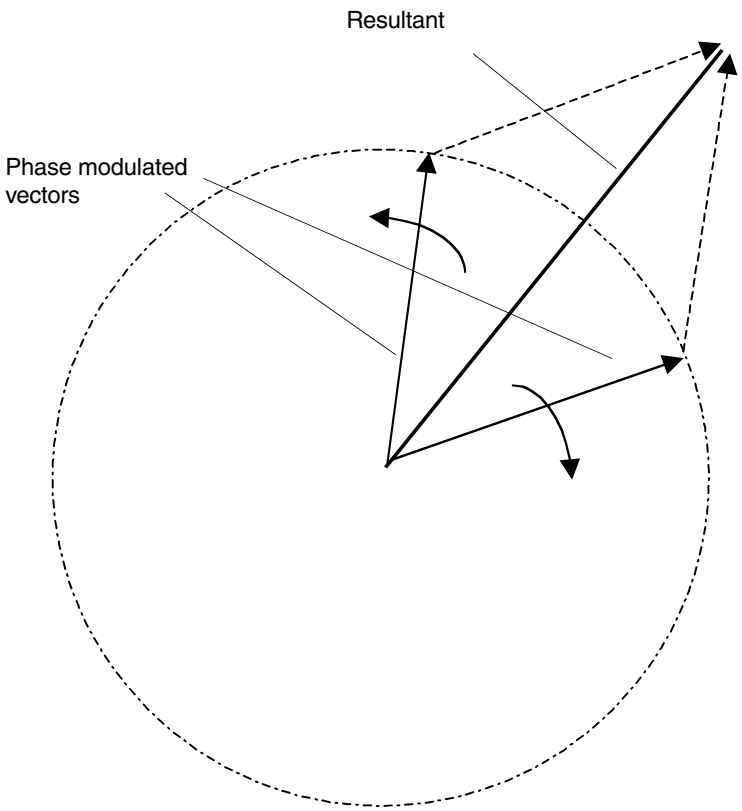


Figure 11.10 RF synthesis vector diagram.

A large number of amplifier linearization solutions are based on predistortion of the signal driving the amplifier in an attempt to match the nonlinear transfer characteristic of the amplifier with an inverse characteristic in the predistortion process. The challenge with predistortion is to be able to realize a predistortion element that is a good match to the inverse of the amplifier distortion—that is, low cost and low power in its implementation—and that can, if necessary, be adapted to track changes in the amplifier response with time, temperature, voltage, device, operating frequency, operating power point, and Voltage Standing Wave Ratio (VSWR).

For complex envelope modulation formats such as multicarrier W-CDMA, the envelope excursions of the composite waveform will cause the amplifier to operate over its full output range. This means that a predistorter element must also match this characteristic over a wide range of input levels if high levels of linearity are to be achieved.

With a typical superhet design of transmitter, there are three locations where predistortion can be implemented. The options are shown in Figure 11.11. An RF solution is attractive, since it is likely to be small and does not require modification of the remainder of the transmit stages. An IF solution is likely to make fabrication of an adaptive predistortion element more practical. A baseband DSP based solution will give ultimate flexibility in implementation, but is likely to take a significant amount of processor cycles and hence consume most power.

One of the simplest RF predistorters to implement is a third-order predistorter. Recognizing that much of the distortion in an amplifier is generated by third-order nonlinear effects, a circuit that creates third-order distortion—for example, a pair of multipliers—can be used to generate this type of distortion but in antiphase. When summed with the drive to the amplifier, significant reduction in the third-order products from the amplifier output can be achieved. Of course, good performance relies on close matching to the gain and phase of the third-order distortion for a particular device, and without some form of feedback control of these parameters, only limited correction is possible over a spread of devices and operating conditions.

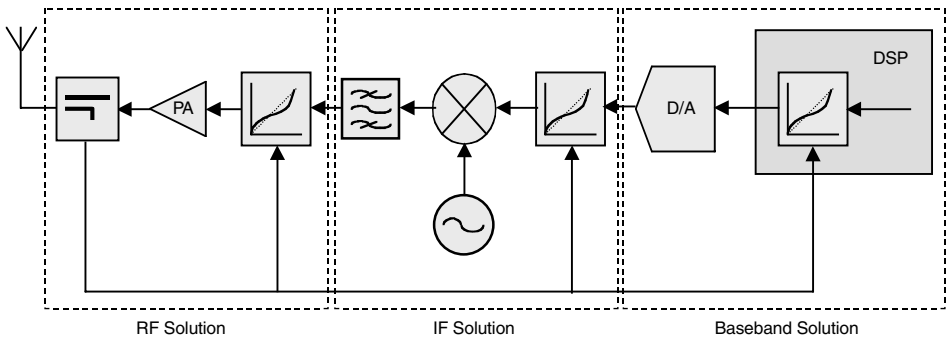


Figure 11.11 Location for predistorter components.

It would be very simple to construct an open-loop DSP-based predistorter using a lookup table; however, for most applications, the characteristics of the transmitter device changes so much with operating point that some form of updating of the predistortion function is needed. As soon as an adaptive control process is introduced, ADC components are needed, additional DSP processing used, reliable and rapid convergence control algorithms must be identified, and the whole process becomes quite complicated. Within a DSP, it is possible to create any predistortion characteristic required and rapidly update the transfer function to follow changes in the amplifier device response.

As the cost and power consumption of DSP engines continues to fall and the processing power increases, the digital baseband predistortion solution becomes more and more attractive—first for Node B use but also for portable use. There are two main options for updating a predistortion lookup table: using power indexing, which involves a one-dimensional lookup table, or Cartesian (I/Q) indexing, giving rise to a two-dimensional lookup table.

Power indexing will result in a smaller overall table size and faster adaptation time, since the number of elements to update is smaller. It does not, however, correct AM-PM distortion, which means that only limited linearization is possible. I/Q indexing will provide correction for both AM-AM and AM-PM distortion and so give optimum results, but the tables are large and adaptation time slow. For wideband multicarrier signals it is necessary for the lookup table to have a frequency-dependant element to accommodate frequency-dependent distortion through the amplifier chain. This can give rise to three-dimensional tables.

In summary, baseband digital predistortion is the most versatile form of predistortion and will become more widely used as the cost and power consumption of DSP falls. Because of the slow adaptation time for a lookup table predistorter, it is not possible to correct for the memory effect in high-power amplifiers, and so this will limit the gain for multicarrier wideband applications. Correct choice of lookup table indexing will give faster adaptation rates and smaller table size; however, frequency dependent effects in the amplifier cannot be ignored.

An alternative to using a lookup table is to synthesize in real time the predistorter function, much like the third-, fifth-, and seventh-order elements suggested for RF predistortion. This shifts the emphasis from lookup table size to processor cycles, which may be advantageous in some cases.

The final linearization method to be considered is the RF feed-forward correction system. This technique is used widely for the current generation of highly linear multicarrier amplifiers designs in use today, and there are many algorithm devices for correcting the parameters in the feed-forward control loops. More recently, combinations of feed forward and predistortion have appeared in an attempt to increase amplifier efficiency by shifting more of the emphasis on pre-correction rather than post-correction of distortion.

A feed-forward amplifier operates by subtracting a low-level undistorted version of the input signal from the output of the main power amplifier (top path) to yield an error signal that predominantly consists of the distortion elements generated within

the amplifier. This distortion signal itself is amplified and then added in antiphase to the main amplifier output in an attempt to cancel out the distortion components.

Very careful alignment of the gain and phase of the signals within a feed-forward linearization system is needed to ensure correct cancellation of the key signals at the input to the error amplifier and the final output of the main amplifier. This alignment involves both pure delay elements to offset delays through the active components, as well as independently controlled gain and phase blocks. The delay elements in particular must be carefully designed, since they introduce loss in the main amplifier path, which directly affects the efficiency of the solution. Adaptation of the gain and phase elements requires a real-time measurement of the amplifier distortion and suitable processing to generate the correct weighting signals. Most feed-forward amplifiers now use DSP for this task. Where very high levels of linearity are needed, it is possible to add further control loops around the main amplifier. Each subsequent control loop attempts to correct for the residual distortion from the previous control loop, with the result that very high levels of linearity are possible but at the expense of power-added efficiency through the amplifier. Feed-forward control requirements are shown in Figure 11.12.

In summary, feed-forward amplifiers can deliver very high levels of linearity over wide operating bandwidth and can operate as RF-in, RF-out devices, making them attractive standalone solutions. Their main drawback is the relatively poor efficiency. Many new multicarrier amplifier solutions are utilizing predistortion correction techniques to try and reduce the load on the feed-forward correction process so that it can operate in a single-loop mode with good main amplifier efficiency. The poor efficiency makes feed forward an unlikely candidate for handset applications; however, since these tend to operate only in single-carrier mode, predistortion techniques alone are likely to give sufficient gain.

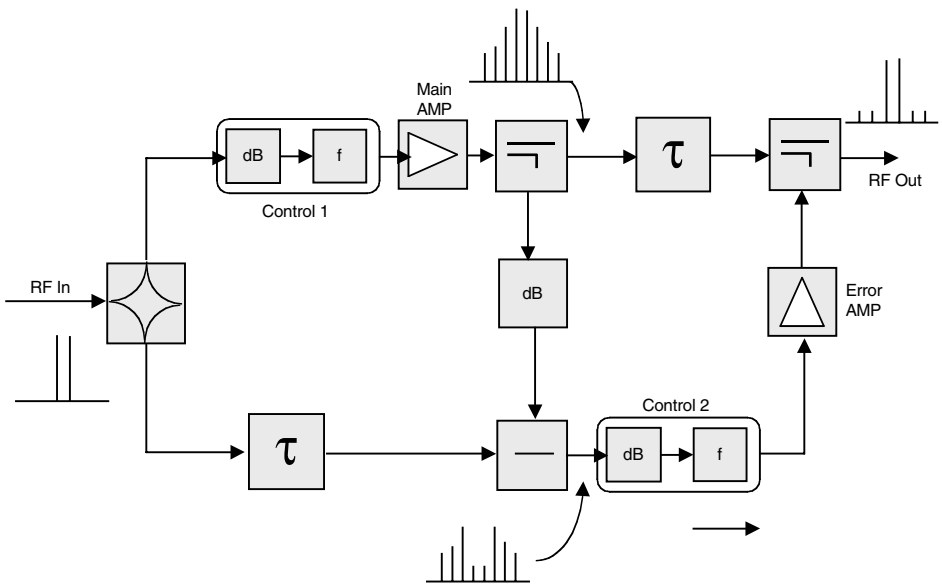


Figure 11.12 Feed-forward control requirements.

The Baseband Section

In Chapter 3 we discussed code generation requirements and root raised cosine filter implementation, and we introduced digital processing methods of producing these functions in the handset. These same functions are required in the Node B transmitter, but given less restriction on power consumption, different trade-offs of software against configured silicon may be made. Also, in the handset it was seen that after the RRC filter implementation, the signals (I and Q) were passed to matched DACs for conversion into the analog domain. An analog-modulated IF was produced that was then upconverted to final transmit frequency. Again, in the Node B, the signal can remain in the digital domain—to produce a modulated IF and only be converted to analog form prior to upconversion. This approach comes nearer to the software radio concept and so provides greater flexibility.

Interpolation

The baseband signal that has been processed up to this stage (RRC filtering) has been constructed at a sample rate that meets the Nyquist criteria for its frequency content. Ultimately, in this example, the signal will be digitally modulated and the IQ streams recombined to yield a real digital intermediate frequency. This will then be applied to a digital-to-analog converter to give a modulated analog IF suitable for upconversion to the final carrier frequency. Because the digital frequency content is increasing (digital upconversion and IQ combining), the sample rate must be increased to re-meet the Nyquist requirement. This is the process of *interpolation*—the insertion of additional samples to represent the increased frequency components of the signal.

QPSK Modulation

Channels are selected in the digital domain using a numerically controlled oscillator (NCO) and digital mixers. Direct digital synthesis gives more precise frequency selection and shorter settling time; it also provides good amplitude and phase balance. The digital filter provides extremely linear phase and a very good shape factor. Figure 11.13 reminds us of the processing blocks.

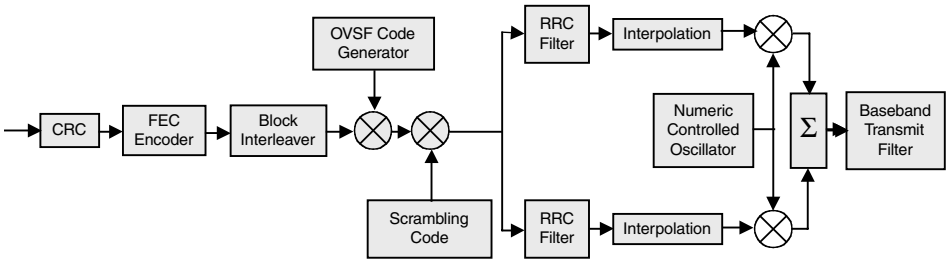


Figure 11.13 Positioning of the NCO.

W-CDMA requirements are as follows:

- Nyquist filter
  - Root raised cosine filter: $\alpha = 0.22$
  - Sampling rate: 3.84 Msps × 4
- NCO
  - 60 MHz bandwidth for channel mapping
  - High spurious free dynamic range (SFDR)

The highest rate processing function in the baseband transmitter is the pulse shaping and vector modulation. These tasks must therefore be designed with care to minimize processing overhead and hence power consumption and chip size. A design example from Xilinx for a Node B unit employs an eight times oversampling approach. (Sample rate is eight times symbol rate.) The output frequency for the channel is set using a digital NCO (lookup table method).

The interpolation process is performed using the RRC filter as the first interpolator filter, with a factor of eight sample rate increase. This is followed by two half-band factor of two interpolators, fully exploiting the zero coefficient property of the half-band filter design.

Because the sample rate at the input to the filters is already very high to accommodate the 3.84 Mcps spread signal, the processing for all three filters is significant. With a total of 3.87 billion MACs used for the I and Q channels, this task alone represents about 38 times the processing load for a second-generation GSM phone.

The use of a digital IF design such as this is clearly not feasible for handset implementation with current technology, since the power consumption of the DSP engines would be too high. Even for Node B use, the approach is using approximately 25 percent of a top-end Virtex 2 FPGA.

Technology Trends

In this and earlier chapters we have seen the need for a dramatic increase in baseband processing capability in the 3G standards. Even a minimum-feature entry-level handset or Node B requires several orders more digital processing than has been seen in 2G products.

It may be argued that it is the practical restrictions of digital processing capability and speed versus power consumption that will be the prime factor in restricting introduction and uptake rate of 3G networks and services. It is not surprising, therefore, that most established and many embryo semiconductor, software, fabless, and advanced technology houses are laying claim to having an ideal/unique offering in this area. Certainly, design engineers facing these digital challenges need to constantly update their knowledge of possible solutions.

Advanced, full-capability handsets and node Bs using smart antennas, multiuser detection and, multiple code, and channel capability will require increasingly innovative technologies. Solutions offered and proposed include optimized semiconductor processes, both in scale and materials (for example, SiGe, InP, and SiC), Micro-Electro-Mechanical Systems (MEMS), cluster DSPs, reconfigurable DSPs, wireless DSP, and even an optical digital signal processing engine (ODSPE) using lenses, mirrors, light

modulators, and detectors. The designer will not only need to weigh the technical capabilities of these products against the target specification but also weigh the commercial viability of the companies offering these solutions.

System Planning

We have considered some of the factors determining Node B RF power and downlink quality (for example, linearity) and Node B receive sensitivity (uplink quality). We now need to consider some of the system-level aspects of system planning in a 3G network.

Many excellent books on system planning are available. Several have been published by Wiley and are referenced in Table 11.8.

Our purpose in this chapter is to put simulation and planning into some kind of historical perspective. Why is it that simulations always seem to suggest that a new technology will work rather better than it actually does in practice? Why is it that initial link budget projections always seem to end up being rather overoptimistic.

Cellular technologies have a 15-year maturation cycles. Analog cellular technologies were introduced in the 1980s and didn't work very well for five years (the pain phase). From the mid-1980s onward, analog cellular phones worked rather well, and by 1992 (when GSM was introduced), the ETACS networks in the United Kingdom and AMPS networks in the United States and Asia were delivering good-quality consistent voice services with quite acceptable coverage. The mid-1980s to early 1990s were the pleasure phase for analog. In the early 1990s there were proposals to upgrade ETACS in the United Kingdom (ETACS 2) with additional signaling bandwidth to improve handover performance. In the United States, narrowband (10 kHz channel spacing) AMPS was introduced to deliver capacity gain. However, the technology started running out of improvement potential, and engineers got bored with working on it. We describe this as the perfection phase.

When GSM was introduced in 1992, it really didn't work very well. Voice quality was, if anything, inferior to the analog phones and coverage was poor. The next five years were the pain phase. GSM did not start to deliver consistent good-quality voice service until certainly 1995 and arguably 1997.

Table 11.8 Further Reading

TITLE	PUBLISHER	LEAD AUTHOR	ISBN
Radio Network Planning for UMTS	Wiley	Laiho	0-471-48653-1
W-CDMA	Artech	Ojanpera	1-58053-180-6
UMTS Networks	Wiley	Kaaranen	0-471-48654
UMTS	Wiley	Muratore	0-471-49829-7
UMTS Networks	Wiley	Castro	0-471-81375-3
W-CDMA For UMTS	Wiley	Holma	0-471-48687-6

The same is happening with 3G. Networks being implemented today (2002 to 2003) will not deliver good, consistent video quality until at least 2005 and probably not until 2007. By that time, 2G technologies (GSM US TDMA) will be fading in terms of their further development potential, and a rapid adoption shift will occur.

Let's look at this process in more detail.

The Performance/Bandwidth Trade-Off in 1G and 2G Cellular Networks

The AMPS/ETACS analog cellular radio networks introduced in the 1980s used very well established baseband and RF processing techniques. The analog voice stream was captured using the variable voltage produced by the microphone, companded and pre-emphasized, and then FM modulated onto a 25 kHz (ETACS) or 30 kHz (AMPS) radio channel.

The old 1200-bit rate FFSK signaling used in trunked radio systems in the 1970s was replaced with 8 kbps PSK (TACS) or 10 kbps PSK for AMPS. (As a reminder, AMPS stands for Advanced Mobile Phone System, TACS for Total Access Communications System, and E-TACS for Extended TACS—33 MHz rather than 25 MHz allocation.)

At the same time (the Scandinavians would claim earlier), a similar system was deployed in the Nordic countries known as Nordic Mobile Telephone System (NMT). This was a narrowband 12½ kHz FM system at 450 MHz.

All three first-generation cellular systems supported automatic handover as a handset moved from base station to base station in a wide area network. The handsets could be instructed to change RF channel and to increase or decrease RF power to compensate for the near/far effect (whether the handset was close or far away from the base station). We have been using the past tense, but in practice, AMPS phones are still in use, as well as some, though now few, NMT phones.

Power control and handover decisions were taken at the MSC on the basis of channel measurements. AMPS/ETACS both used supervisory audio tones. These were three tones at 5970, 6000, and 6030 Hz (above the audio passband). One of the three tones would be superimposed on top of the modulated voice carrier. The tone effectively distinguished which base station was being seen by the mobile. The mobile then retransmitted the same SAT tone back to the base station. The base station measured the signal-to-noise ratio of the SAT tone and either power-controlled the handset or instructed the handset to move to another RF channel or another base station. Instructions were sent to the mobile by blanking out the audio path and sending a burst of 8-kbps PSK signaling.

This still is a very simple and robust system for managing handsets in a mobile environment. However, as network density increased, RF planning became quite complicated (833 channels to manage in AMPS, 1321 channels to manage in ETACS), and there was insufficient distance between the SAT tones to differentiate lots of different base stations being placed relatively close to one another. There were only three SAT tones, so it was very easy for a handset to see the same SAT tone from more than one base station.

Given that the SAT tones were the basis of power control and handover decisions, the network effectively became capacity-limited in terms of its signaling bandwidth.

The TDMA networks (GSM and IS136 TDMA) address this limitation by increasing signaling bandwidth. This has a cost (bandwidth overhead) but delivers tighter power and handover control.

For example: In GSM, 61 percent of the channel bandwidth is used for channel coding and signaling, as follows:

Speech codec	13.0 kbps	39%
Codec error protection	9.8 kbps	29%
SACCH	0.95 kbps	2%
Guard time/ramp time/synchronization	10.1 kbps	30%
TOTAL	33.85 kbps	100%

The SACCH (slow associated control channel) is used every thirteenth frame to provide the basis for a measurement report. This is sent to the BTS and then on to the BSC to provide the information needed for power control and handover. Even so, this is quite a relaxed control loop with a response time of typically 500 ms (twice a second), compared to 1500 times a second in W-CDMA (IMT2000DS) and 800 times a second in CDMA2000.

The gain at system level in GSM over and above analog cellular is therefore a product of a number of factors: *1*. There is some source coding gain in the voice codec. *2*. There is some coherence bandwidth gain by virtue of using a 200 kHz RF channel rather than a 25 kHz channel. *3*. There is some channel coding gain by virtue of the block coding and convolutional coding (achieved at a very high price with a coding overhead of nearly 10 kbps). and *4*. There is a gain in terms of better power control and handover.

In analog TACS or AMPS, neighboring base stations measure the signal transmission from a handset and transfer the measurement information to the local switch for processing to make decisions on power control and handover. The information is then downloaded to the handset via the host base station. In an analog network being used close to capacity, this can result in a high signaling load on the links between the base stations and switches and a high processing load on the switch.

In GSM, the handset uses the six spare time slots in a frame to measure the received signal strength on a broadcast control channel (BCCH) from its own and five surrounding base stations. The handset then preprocesses the measurements by averaging them over a SACCH block and making a measurement report. The report is then retransmitted to the BTS using an idle SACCH frame. The handset needs to identify cochannel interference and therefore has to synchronize and demodulate data on the BCCH to extract the base station identity code, which is then included in the measurement report. The handset performs base station identification during the idle SACCH.

The measurement report includes an estimate of the bit error rate of the traffic channels using information from the training sequence/channel equalizer. The combined information provides the basis for an assessment of link quality degradation due to cochannel and time dispersion and allows the network to make reasonably accurate power control and handover decisions.

Given the preceding information, various simulations were done in the late 1980s to show how capacity could be improved by implementing GSM. The results of base simulations were widely published in the early 1990s. Table 11.9 suggests that additional capacity could be delivered by increasing the reuse ratio (how aggressively frequencies were reused within the network) from 7 to 4 (the same frequency could be reused every fourth cell). The capacity gain could then be expressed in Erlangs/sq km.

In practice, this all depended on what carrier-to-interference ratio was needed in order to deliver good consistent-quality voice. The design criteria for analog cellular was that a C/I of 18 dB was needed to deliver acceptable speed quality. The simulations suggested GSM without frequency hopping would need 11 dB, which would reduce to 9 dB when frequency hopping was used. In practice, these capacity gains initially proved rather illusory partly because, although the analog cellular networks were supposed to be working at an 18 dB C/I, they were often working (really quite adequately) at C/Is close to 5—that is, there was a substantial gap between theory and reality.

The same reality gap happened with coverage predictions. The link budget calculations for GSM were really rather overoptimistic, particularly because the handsets and base station hardly met the basic conformance specification.

Through the 1990s, the sensitivity of handsets improved, over and above the conformance specification, typically by 1 dB per year. Similarly, base station sensitivity increased by about 3 or 4 dB. This effectively delivered coverage gain. Capacity gain was achieved by optimizing power control and handover so that dropped call performance could be kept within acceptable limits even for relatively fast mobility users in relatively dense networks. Capacity gain was also achieved by allocating 75 MHz of additional bandwidth at 1800 MHz. This meant that GSM 900 and 1800 MHz together had 195 + 375 × 200 kHz RF channels available between 4 network operators, 570 RF channels each with 8 time slots = 4640 channels! GSM networks have really never been capacity-limited. The capacity just happens sometimes to be in the wrong place. Cellular networks in general tend to be power-limited rather than bandwidth-limited.

Table 11.9 Capacity Gain Simulations for GSM

	ANALOG FM	GSM	
		PESSIMISTIC	OPTIMISTIC
Bandwidth	25 MHz	25 MHz	
Number of voice channels	833	1000	
Reuse plan	7	4	3
Channels per site	119	250	333
Erlang/km^2	11.9	27.7	40
Capacity gain	1.0	2.3	3.4

Source: Raith and Udderfeldt. *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 2, May 1991.

So it was power, or specifically coverage, rather than capacity that created a problem for GSM 1800 operators. As frequency increases, propagation loss increases. It also gets harder to predict signal strength. This is because as frequency increases, there is more refraction loss—radio waves losing energy as they are reflected from buildings or building edges. GSM 1800 operators needed to take into account at least an extra 6 dB of free space loss over and above the 900 MHz operators and an additional 1 to 2 dB for additional (hard to predict) losses. This effectively meant a network density four to five times greater than the GSM 900 operators needed to deliver equivalent coverage. The good news was that the higher frequency allowed more compact base station antennas, which could also potentially provide higher gain. The higher frequency also allowed more aggressive frequency reuse; though since capacity was not a problem, this was really not a useful benefit.

It gradually dawned on network operators that they were not actually short of spectrum and that actually there was a bit of a spectral glut. Adding 60 + 60 MHz of IMT2000 spectrum to the pot just increased the oversupply. Bandwidth effectively became a liability rather than an asset (and remains so today).

This has at last shifted attention quite rightly away from capacity as the main design objective. The focus today is on how to use the limited amount of RF power we have available on the downlink and uplink to give acceptable channel quality to deliver an acceptably consistent rich media user experience.

TDMA/CDMA System Planning Comparisons

In GSM or US TDMA networks, we have said that the handset produces a measurement report that is then sent to the BSC to provide the basis for power control and handover. The handset does radio power measurements typically every half second (actually, 480 ms) and measures its own serving base station and up to five other base stations in neighboring cells. This is the basis of Mobile-Assisted HandOff (MAHO).

This measurement process has had to be modified as GPRS has been introduced. GPRS uses an adaptive coding scheme—CS1, 2, 3, or 4, depending on how far the handset is away from the base station. The decision on which coding scheme to use is driven by the need to measure link quality. Link quality measurement can only be performed during idle bursts. In voice networks, the measurement has traditionally been done every 480 ms. The fastest possible measurement rate is once every 120 ms (once every multiframe), which is not fast enough to support adaptive coding.

IN E-GPRS (GPRS with EDGE), measurements are taken on each and every burst within the equalizer of the terminal resulting in an estimate of the bit error probability (BEP). The BEP provides a measure of the C/I on a burst-by-burst basis and also provides information on the delay spread introduced by the channel and the velocity (mobility) of the handset—that is, how fast the handset/mobile is traveling through the multipath fading channel. The variation of BEP value over several bursts also provides information on frequency hopping. A mean BEP is calculated per radio block (four bursts), as well as the variation (the standard deviation of the BEP estimation divided by the mean BEP) over the four bursts. The results are then filtered for all the radio blocks sent within the measurement period.

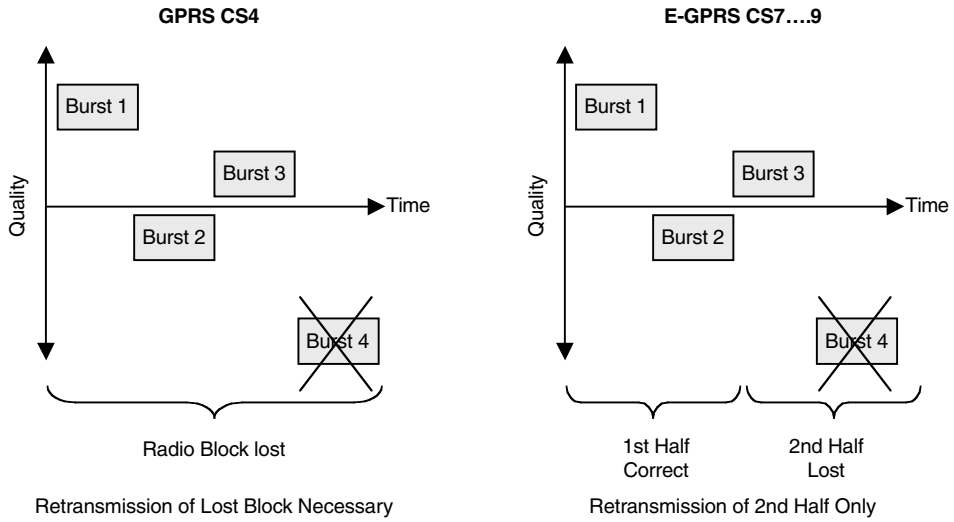


Figure 11.14 Link adaptation.

For the higher coding schemes within GPRS (MCS7 to 9), the interleaving procedure is changed. If frequency hopping is used, the radio channel is changing on a per-burst level. Because a radio block is interleaved and transmitted over four bursts for GPRS, each burst may experience a completely different interference environment. If, for example, one of the four bursts is not properly received, the whole radio block will be wrongly decoded and have to be retransmitted. In E-GPRS, the higher coding schemes MCS7, MCS8, and MCS9 transmit two radio blocks over the four bursts. The interleaving occurs over two bursts rather than four, which reduces the number of bursts that need to be transmitted if errors are detected. This is shown in Figure 11.14. The process is known as *link adaptation*.

These higher coding schemes work better when frequency hopping is used, but you lose the gain delivered from deep interleaving.

Link adaptation uses the radio link quality measurement by the handset on the downlink or base station on the uplink to decide on the channel coding and modulation that should be used. This is in addition to the power control and handover decisions being made. The modulation and coding scheme can be changed every frame (4.615 ms) or four times every 10 ms (the length of a frame in IMT2000).

E-GPRS is effectively adapting to the radio channel four times every 10 ms (at a 400 Hz rate). IMT2000 adapts to the radio channel every time slot, at a 1500 Hz rate. Although the codecs and modulation scheme can theoretically change every four bursts (every radio block), the measurement interval is generally slower.

The coding schemes are grouped into three families: A, B, and C. Depending on the payload size, resegmentation for retransmission is not always possible, thus determining which family of codes are used.

Table 11.10 shows the nine coding schemes used in E-GPRS, including the two modulation schemes (GMSK and 8PSK).

Table 11.10 E-GPRS Channel Coding Schemes

CHANNEL CODING SCHEMES	THROUGHPUT (KBPS)	FAMILY	MODULATION
MCS9	59.2	A	GMSK
MCS8	54.4	A	GMSK
MSC7	44.8	B	GMSK
MCS6	29.6	A	GMSK
MCS5	22.4	B	8PSK
MCS4	17.6	C	8PSK
MCS3	14.8	A	8PSK
MCS2	11.2	B	8PSK
MCS1	8.8	C	8PSK

The performance of a handset sending and receiving packet data is therefore defined in terms of throughput, which is a product of the gross throughput less the retransmissions needed. The retransmission will introduce delay and delay variability, which will require buffering. If the delay variability exceeds the available buffer bandwidth, then the packet stream will become nonisochronous, which will cause problems at the application layer.

We can see that it becomes much harder to nail down performance in a packet-routed network.

As we will see in the next section, link budgets are still established on the basis of providing adequate voice quality across the target coverage area. Adaptive coding schemes if well implemented should mean that when users are closer to a base station, data throughput rates can adaptively increase. When a user is at the cell edge, data throughput will be lower, but in theory, bit error rates and retransmission overheads should remain relatively constant.

Having convinced ourselves that this may actually happen, we can now move on to radio system planning.

Radio Planning

With existing TDMA systems it has been relatively simple to derive base station and handset sensitivity. The interference is effectively steady-state. Coverage and capacity constraints can be described in terms of grade of service and are the product of network density.

In IMT2000, planning has to take into account noise rise within a (shared) 5 MHz channel, an allowance for fast power control headroom at the edge of the cell and soft handover gain. The interference margin is typically 1 to 3 dB for coverage-limited conditions and more for capacity-limited networks. Fast power control headroom is typically between 2 and 5 dB for slow-moving handsets. Soft handover gain—effectively uplink and downlink diversity gain—is typically between 2 and 3 dB. There are four power classes. Class 1 and 2 are for mobiles. Class 3 and 4 are for handsets (mobiles would, for example, be vehicle-mounted). These are shown in Table 11.11. Typical maximum power available at a Node B would be 5, 10, 15, 20, or 40 Watts.

In 3GPP, E_b/N_o targets are set that are intended to equate with the required service level. (As mentioned in earlier chapters, E_b/N_o is the energy per bit over the noise floor. It takes into account the channel coding predetermined by the service to be provided.) The E_b/N_o for 144 kbps real-time data is 1.5 dB. The E_b/N_o for 12.2 kbps voice is 5 dB.

Why does E_b/N_o reduce as bit rate increases? Well, as bit rate increases, the control overhead (a fixed 15 kbps overhead) reduces as a percentage of the overall channel rate. In addition, because more power is allocated to the DPCCH (the physical control channel), the channel estimation improves. However, as the bit rate increases, the spreading gain reduces.

IMT2000 planning is sensitive to both the volume of offered traffic and the required service properties of the traffic—the data rate, the bit error rate, the latency, and service-dependent processing gain (expressed as the required E_b/N_o).

System performance is also dependent on system implementation—how well the RAKE receiver adapts to highly variable delay spreads on the channel, how well fast fading power control is implemented, how well soft/softer handover is configured, and interleaving gain.

Downlink capacity can also be determined by OVSF code limitations (including nonorthogonality) and downlink code power. The power of the transmitter is effectively distributed among users in the code domain. 10 W, for example, gets distributed among a certain amount of code channels—the number of code channels available determines the number of users that can be supported. On the uplink, each user has his or her own PA, so this limitation does not apply.

A Node B will be exposed to intracell and intercell interference. *Intracell interference* is the interference created by the mobiles within the cell and is shown in Figure 11.15.

Table 11.11 Power Classes for Mobiles and Handsets

Power Class 1	+ 33 dBm (+1 to 3 dB)	2 W	Mobiles
Power Class 2	+ 27 dBm	500 mW	
Power Class 3	+24 dBm	250 mW	Handsets
Power Class 4	+21 dBm (±2 dB)	125 mW	

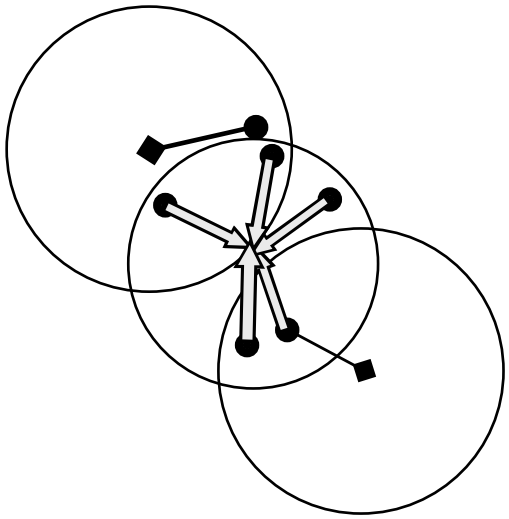


Figure 11.15 Intracell interference.

Intercell interference is the sum of all the received powers of all mobiles in all other cells and is shown in Figure 11.16.

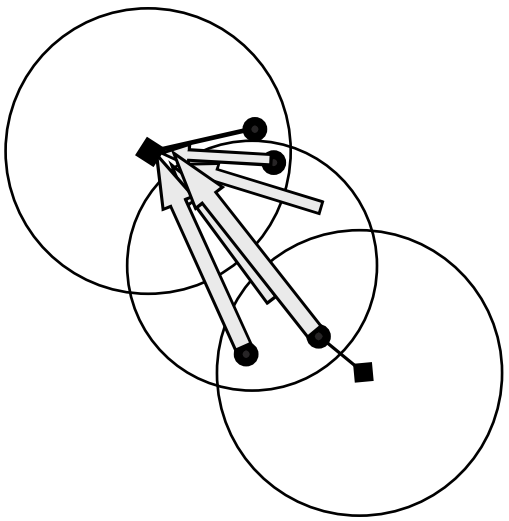


Figure 11.16 Intercell interference.

Interference from adjacent cells initially exhibits a first-order increase. At a certain point, handsets start increasing their power to combat noise rise, which in turn increases noise rise! A first-order effect becomes a second-order effect; the cell has reached pole capacity. The rule of thumb is that a 50 percent cell load (50 percent pole capacity) will result in 3 dB of intracell interference. A 75 percent cell load implies 6 dB of intracell interference.

In other words, say you have a microcell with one transceiver. It will have a higher data rate handling capability than a macrocell with one transceiver because the microcell will not see so much interference as the macrocell.

Rules of Thumb in Planning

Macrocell performance can be improved by using adaptive downtilt to reduce interference visibility, which in turn will reduce noise rise. However, the downtilt also reduces the coverage footprint of the cell site. The other useful rule of thumb is to try and position Node B sites close to the offered traffic to limit uplink and downlink code power consumption. The effect is to increase cell range. Unfortunately, most sites are chosen pragmatically by real estate site acquisition specialists and are not really in the right place for a 3G network to deliver optimum performance.

Figure 11.17 shows how user geometry (how close users are to the Node B) determines cell footprint. As users move closer to the cell center they absorb less downlink code domain power. This means more code domain power is available for newcomers so the cell footprint grows (right-hand circle). If existing users are relatively distant from the Node B, they absorb more of the Node B's code domain power and the cell radius shrinks (left-hand circle).

Interference and noise rise can be reduced by using sectored antennas and arranging receive nulls in a cloverleaf pattern. Typical Node B configuration might therefore include, say, a single RF carrier omnidirectional antenna site for a lightly populated rural area, a three-sector site for a semi-rural area (using 1 × RF carrier per sector), a three-sector site configuration with two RF carriers per sector for urban coverage, or alternatively, an eight-sector configuration with either one or two RF carriers per sector for dense urban applications.

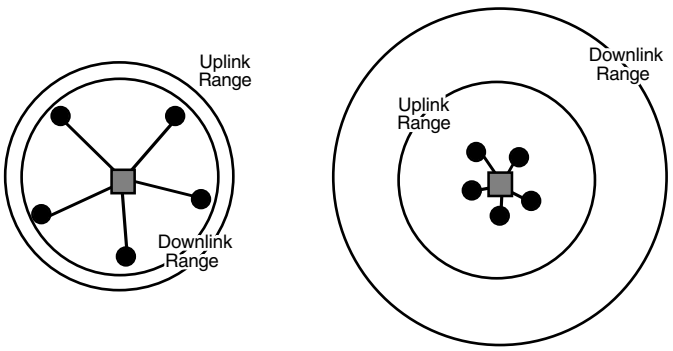


Figure 11.17 Impact of user geometry on cell size.

These configurations are very dependent on whether or not network operators are allowed to share Node B transceivers, in which case four or more 5-MHz RF channels would need to be made available per sector to give one RF channel per operator.

How System Performance Can Be Compromised

System performance can be compromised by loss of orthogonality. OVSF codes, for example, are not particularly robust in dispersive channels (e.g., large macrocells). Degraded orthogonality increases intracell interference and is expressed in radio planning as an orthogonality factor. Orthogonality on the downlink is influenced by how far the users are away from the Node B. The further away they are, the more dispersive the channel and the more delay spread there will be. Loss of orthogonality produces code interference.

We said earlier that we can implement soft handover to improve coverage. Effectively, soft handover gives us uplink and downlink diversity gain; however, soft handover absorbs radio bandwidth and network bandwidth resources, long code energy on the radio physical layer, and Node B to RNC and RNC to RNC transmission bandwidth in the IP RAN. If lots of users are supported in soft handover, range will be optimized, but capacity (radio and network capacity) will be reduced. If very few users are supported on soft handover, range will be reduced, but (radio and network) capacity will increase.

In the radio network subsystem (RNS), the controlling RNC (CRNC) looks after load and congestion control of its own cells, admission control, and code allocation. The drift RNC (DRNC) is any RNC, other than the serving RNC, that controls cells used by the Node B. Node B looks after channel coding and interleaving, rate adaption, spreading, and inner-loop power control.

The RNC looks after combining—the aggregation of multiple uplinks and downlinks. Note that a handset can be simultaneously served by two Node Bs, each of which sends and receives long code energy to and from the handset. In addition, each Node B could be sending and receiving multiple OVSF code streams via two Node Bs to either the CRNC, or if the handset is between RNCs, to both the CRNC and DRNC. The CRNC and DRNC then have to talk to each other, and the CRNC has to decide which long code channel stream to use on a frame-by-frame basis or to combine the two long code channel streams together to maximize combining again. This is a nontrivial decision-making process that will need to be optimized over time. It will take at least 5 years for these soft handover algorithms to be optimized in IMT2000DS networks.

The RNC also has to respond to admission priorities set/predetermined by the admission policy, which is predetermined by individual user or group user service level agreements. The requirements of the traffic (tolerance to delay and delay variability) determine the admission policy and how offered traffic will be distributed between serving cells. There are four service classes in IMT2000. The low-delay data (LDD) services are equivalent to the constant bit rate and variable bit rate services available in ATM and are used to support conversational and streamed video services. The service classes are shown in Table 11.12.

Table 11.12 Classes of 3GPP Service

CLASS	CONVERSATIONAL A		STREAMING B	INTERACTIVE C	BEST EFFORT D
BER	1 in 10^3	1 in 10^6	1 in 10^6	1 in 10^6	1 in 10^8
Bit rate (kbps)	8	144, 384	64, 144, 384, 1920	64, 144, 384, 1920	64, 144, 384, 1920
Delay	Low delay data		Low-delay data	Long constrained delay	Unconstrained delay data
	Constant bit rate		Variable bit rate	Available bit rate	Unspecified bit rate
				Max delay 1 second	No delay limits

LCD and UDD are equivalent to the available bit rate and unspecified bit rate services available in ATM and are used to support interactive and best-effort services. Low bit error rates can be achieved if data is delay-tolerant. A higher-layer protocol, for example, TCP, detects packet-level error rates and requests a packet retransmission. It is possible to reduce bit error rates. The cost is delay and delay variability.

Link budgets (coverage and capacity planning) are therefore dependent on individual user bit rates, user QoS requirements, propagation conditions, power control, and service class. The offered traffic statistics will determine the noise rise in the cell. For example, a small number of high bit rate users will degrade low bit rate user's performance. The failure to meet these predefined service levels can be defined and described as an *outage probability*.

The job of the RNC (controlling RNCs and drift RNCs) is to allocate transmission resources to cells as the offered traffic changes. Loading can be balanced between RNCs using the IUR interface. This is known as *slow dynamic channel allocation*. An RNC balances loading across its own Node Bs over the IUB interface. RF resources (code channels) are allocated by the Node B transceivers. This is described as *fast dynamic channel allocation*, with the RNC allocating network bandwidth and radio bandwidth resources every 10 ms. Both the network bandwidth and radio bandwidth need to be adaptive. They have to be able to respond to significant peaks in offered traffic loading.

Timing Issues on the Radio Air Interface

We said that radio link budgets can be improved by putting handsets into soft handover. Care must however be taken to maintain time alignment between the serving Node B and soft handover target Node B. Path delay will be different between the two serving Node Bs and will change as the user moves. The downlink timing therefore has to be adjusted from the new serving Node B so that the handset RAKE receiver can coherently combine the downlink signal from each Node B. The new Node B adjusts downlink timing in steps of 256 chips ($256 \times 0.26\ \mu s = 66.56\ \mu s$) until a short code lock

is achieved in the handset. If the adjustment is greater than 10 ms, then the downlink has to be decoded to obtain the system frame number to reclock the second (soft handover) path with the appropriate delay.

Use of Measurement Reports

The measurement report in IMT2000DS does the same job as the measurement report in GSM. It provides the information needed for the Node B to decide on power control or channel coding or for the RNC to decide on soft handover.

In IMT2000DS, the measurement report is based on received signal power. This is the received power on one code after despreading defined on the pilot symbols. The decoded pilot symbols provide the basis for the measurement report, which provides the basis for power control or channel coding and soft handover. It also provides the basis for admission control and load balancing. Effectively, it is providing information on the noise floor as perceived by individual users on individual OVSF/long code channels. It provides additional information over and above wideband noise measurements (which can also be used to set admission control policy).

The measurement report includes E_b/N_o (the received signal code power divided by the total received power), signal-to-interference ratio (which is determined partly by cell orthogonality), and block error rate measurements (used for outer-loop power control).

Load estimation can be done either by measuring wideband received power, which will be the sum of intercell, intracell interference, and background receiver noise, or by measuring throughput, which can be measured in terms of bit rate or E_b/N_o. An additional option would be to measure buffer occupancy. Initially, wideband power estimation is probably an adequate way to decide on admission control and load balance at the RNC.

Cell sizes can be increased or decreased physically by increasing or decreasing downlink transmit power or by physically changing antenna patterns (see Chapter 13). It may, for example, be the case that interference patterns change through the day as traffic changes. The loading and interference from users traveling to work by car in the morning may be different from the loading and interference generated from users traveling home at night. The cell site configuration can be adapted to match the offered traffic (and offered noise) as it changes through the day.

In our later chapter on Service Level Agreements, we review how radio and network performance has to be integrated into provable performance platforms, providing proof that a requested grade of service has been delivered.

At radio system level we need to comprehend a number of factors. One important factor is *soft handover gain*. This is the effect of the handset serving two Node Bs. It can be considered as a macro diversity factor, both on the uplink and the downlink. Because the ultimate objective of network control is to maintain an acceptable BER, the additional factor of soft handover gain enables the handset transmission power to be decreased, which in turn reduces both the intercell and intracell interference and so may be expressed as a capacity gain. As the effective path gain is increased—by virtue of the aggregate signal up to two Node Bs—but the uplink transmit power is reduced, the receive power to the network remains the same.

However, the receive sensitivity is no longer a constant design factor of the handset but is also dependent on a number of dynamic factors.

Intercell and intracell interference adds to the noise power. Processing gain influences sensitivity. The E_b/N_o needed will vary (with service, data rate, speed, and multipath channel). In addition a fast fading margin needs to be added to account for the deterioration in E_b/N_o caused by power limiting at the cell edge.

The link budget will change depending on mobility factors as a user moves within the cell, as users move in other cells, and as users move into and out of cells. The link budget will also change on the service factor as users change data rate (the service factor).

We also need to take into account processing gain. Processing gain (see Table 11.13) is based on the ratio of the user data rate to the chip rate (whereas spreading gain is the ratio of the channel data rate to the chip rate—see Chapter 3). These figures need to include the coding gain available from convolutional encoding and interleaving. This will vary depending on what convolutional encoding or interleaving is used, which, in turn, is dependent on the service being supported. This is allowed for by changing the required E_b/N_o for each service.

Table 11.14 shows typical E_b/N_os needed for particular services. The E_b/N_o is lower for the UDD service (unconstrained delay). The delay tolerance effectively makes the data easier to send. Services at a higher mobile speed require a higher E_b/N_o because of power control errors (the inability to follow the fast fading envelope at higher speeds means the fade margin has to be increased). Services in rural areas need a higher E_b/N_o because the delay spread (and nonorthogonality) is greater.

Fast fading allows all handsets operating in a cell to have equally received powers at the Node B receiver. Fast fading power control effectively equalizes the fading character of the radio channel so that the receiver sees a Gaussian-like radio channel. The BER versus E_b/N_o will improve in a Gaussian channel. The improvement in E_b/N_o is called the *fast power control gain*.

Table 11.13 Link Budget Processing Gain

			LINEAR	LOG
8 kbps (voice)	3.84 Mcps	= 3840/8	480	26.8 dB
12.2 kbps (voice)	3.84 Mcps	= 3840/12.2	314	25.0 dB
64 kbps (LCD data)	3.84 Mcps	= 3840/64	60	17.8 dB
144 kbps (LCD data)	3.84 Mcps	= 3840/144	26.7	14.2 dB
384 kbps (LCD data)	3.84 Mcps	= 3840/384	10	10.0 dB
2 Mbps (LCD data)	3.84 Mcps	=3840/2000	1.92	2.8 dB

Table 11.14 Example E_b/N_o (Node B)

	URBAN		SUBURBAN		RURAL	
	MOBILE	PEDES-TRIAN	MOBILE	PEDES-TRIAN	MOBILE	PEDES-TRIAN
8 kbps Voice	4.4	3.3	4.4	3.3	5.0	3.7
LCD 64	2.7	1.1	3.2	1.1	2.9	2.4
UDD 64	2.0	0.7	2.7	1.4	3.0	1.2
LCD 384	2.0	0.7	2.7	1.4	3.0	2.2

Coverage can actually improve for fast mobility users, since they become part of the channel averaging process. Fast power control works well for slow mobility users and provides useful gain particularly (as in the above example) where only minimal multipath diversity is available. Table 11.15 shows how important it is to optimize the power control algorithms in the handset and base station.

The process of power control is as follows:

Open-loop power control. This sets Tx power level based on the Rx power received by the mobile and compensates for path loss and slow fading.

Closed-loop power control. This responds to medium and fast fading and compensates for open-loop power control inaccuracies.

Outer-loop power control. This is implementation-specific, for example, the outer loop adjusts the closed loop control threshold in the base station to maintain the desired frame error rate. Closed-loop implementation at 1500 Hz uses 1/2 dB steps for urban and 1-dB steps for rural areas.

Table 11.15 Mobility Factors and Power Control Uplink E_b/N_o Requirements

CHANNEL	WITHOUT FAST POWER CONTROL	WITH FAST POWER CONTROL	GAIN FROM FAST POWER CONTROL
ITU Pedestrian 3 km/h	11.3 dB	5.5 dB	5.8 dB
ITU Vehicular 3 km/h	8.5 dB	6.7 dB	1.8 dB
ITU Vehicular 50 km/h	6.8 dB	7.3 dB	-0.5 dB

Power control needs to be optimized for certain operational conditions. Power control inaccuracies will substantially reduce capacity and coverage (by adding to noise rise). At higher speeds, fast power control is less effective in compensating channel fading. When carrying out a link budget for 3G, we therefore usually use an E_b/N_o figure for the service, channel type, and data rate assuming fast power control and then subtract a fast fading margin. The fast fading margin is approximately equivalent to the fast fading gain achieved by fast power control when the transmitter had no power limits. We normally expect the fast fading margin to be a few dB.

Uplink Budget Analysis

In previous cellular systems (1G and 2G) the link budget has been calculated on factors that are noninteractive and clearly definable (quantifiable) cell by cell. The differences between theoretical modeling results and practice have been mainly due to the inaccurate characterization of the propagation terrain and clutter factors. These problems still exist in 3G network planning but are added to by the interactive factors that we discussed earlier in this chapter. Link budget calculations will therefore have to take account of these additional factors.

We may consider some example assumptions:

Case 1

A 12.2 kbps voice service traveling at 120 kmph.

Assumptions:

- A 3 dB intracell interference rise (a cell loading of 50 percent).
- No intercell interference rise.
- Being voice, a soft handover is anticipated; hence, there will be some gain, let's assume 3 dB.
- A processing gain of 25 dB (10log3840/12.2).
- An E_b/N_o target of 5 dB.
- A fast fading margin of 0 dB—fast fading is ineffective at 120 kmph.

Case 2

144 kbps real-time data service traveling at 3 kmph.

Assumptions:

- Again, a 3 dB intracell interference rise (50 percent cell loading).
- A high transmit power is available as the mobile is away from the body. Hence, there are no body losses.
- An E_b/N_o target of 1.5 dB—as fast fading power control is effective at 3 kmph.
- Fast Fading Margin of 4 dB (the E_b/N_o target rises by 4 dB as the mobile moves to the cell edge).

It remains to be seen how true-to-life these considerations are (and become!).

Noise rise (as seen by the Node B) will be a function of the offered traffic measured as throughput. The amount, distribution, and burstiness of offered traffic all contribute to the achieved sensitivity in the Node B transceiver.

On the downlink, the capacity constraint may well be OVSF code-limited or orthogonality constraints (interrelated code domain effects). An orthogonality factor of 1 means the code streams are perfectly orthogonal. You will see orthogonality figures quoted in load factor calculations.

The downlink is more load-sensitive than the uplink. For example, one 10 W transmitter is shared amongst all users. On the uplink, each user has his or her own PA. The load limitation effectively limits downlink capacity. Increasing Node B power (downlink power) increases coverage and capacity but only at the cost of reducing the capacity and coverage available from adjacent cells.

Capacity calculations also need to take into account soft handover overheads. For a given number of cells, assumptions have to be made on uplink and downlink throughput and the number of users in soft handover.

Having established some capacity parameters, we need to establish the coverage available to users defined in terms of the service level being delivered to them. The coverage probability will be influenced by the mobility of the user—whether they are walking or driving or riding in a train.

Overall, the lessons learnt from 2G implementation 10 years ago were that early simulations based on (in GSM's case) relatively simple assumptions were overoptimistic. We might expect 3G simulations to be even wider off the mark, given the additional number of variables introduced into the simulation.

In practice, there will be a number of performance limitations in early 3G deployment that will make early link budget simulations hard to achieve.

Handset sensitivity generally improves as a technology matures partly because of device and design optimization and partly because volume product produces performance benefits (better control of component tolerance).

A similar pattern will probably emerge with IMT2000 handsets. Performance degrades if the air interface is required to do something it was not designed to do. This suggests that IMT2000 network performance will begin to settle down and provide good, consistent video and voice quality by 2005 to 2006.

A much more in-depth analysis of 3G system planning can be found in *Radio Network Planning and Optimization for UMTS*, published by Wiley and authored by Jaana Laiho, Achim Wacker, and Tomas Novosad, (ISBN 0-471-48653-1). Our thanks also to Mason Communications for their advice on system planning issues, some of which are referred to in the preceding text. Information on Mason Communications 3G system planning service is available at www.masoncom.com.

Long-Term Objectives in System Planning: Delivering Consistency

Radio bandwidth in a 3G network is more adaptive than radio bandwidth in a 2G network, which in turn is more adaptive than radio bandwidth in a 1G network. The

OVSF code structure provides an elegant mechanism for balancing variable user data rates and allocating (code domain distributed) power. Pilot symbol based measurement reports provide a fast (1500 Hz) mechanism for measuring radio channel quality, which in turn provides the basis for accurate measurement reports, which in turn provides the basis for effective load balancing and admission control. Load balancing and admission control allow resources to be shared across the IP RAN network and radio layer in a more dynamically adaptable way than would be possible in existing legacy access networks (optimized for predominantly constant-rate offered traffic).

At the radio system level, load balancing between Node B transceivers helps to distribute offered noise. This helps improve average Node B sensitivity, which in turn helps improve coverage. Radio system capacity constraints are determined by the amount of RF power available both at the Node B and in the handset. The downlink can relatively quickly become code-limited because of the limited number of OVSF codes available. In larger cells, orthogonality can be a problem because of delay spread on the channel, and this can cause downlink performance degradation.

Coded channel streams (OVSF and long code streams) represent a phase argument that needs to be coherently demodulated, decorrelated, and combined by the receiver. Time, phase, or amplitude ambiguity on the radio channel will potentially impair performance (increase bit error rates).

Network performance is improved over time by optimizing radio layer and network layer performance parameters. There is a need for consistent quality, which in turn requires careful implementation of soft handover algorithms to avoid session discontinuity (or what used to be called high dropped call rates).

In the longer term (3 to 5 years), there is substantial gain potential in IMT2000, which will result in a more consistent radio channel. The challenge will be to deliver equivalent consistency into and through the IP RAN and IP core network to provide measurable and manageable end-to-end performance.

Wireless LAN Planning

In a number of countries, it is now permitted to provide public access services in the unlicensed industrial scientific medical bands. Bands of interest include the ISM band at 900 MHz (between 900 and 930 MHz) used for digital cordless phones in the United States, the ISM band at 2.4 GHz used for IEEE802 wireless LANs and Bluetooth, and the 5 GHz band used for wideband HIPERLANs (high-performance radio local area networks).

The availability of plug-in wireless LAN cards that include IEEE802 and Bluetooth have begun to focus attention on the need to integrate wide area wireless planning with in-building coverage planning.

One of the problems of planning in-building coverage is the unpredictability of in-building propagation, particularly at 2.4 GHz and 5 GHz (propagation unpredictability increases with frequency). Figure 11.18 shows a comparison between free space loss in an empty room and the loss in a room with cubicles partitioned off from one another. Signal levels received in the bare room are typically a few milliVolts. Signal levels received in the room with partitioning quickly attenuate down to a few microVolts. Losses are very dependent on the materials used in the partitioning. Fire-resistant silver foil, for example, will provide a high degree of shielding.

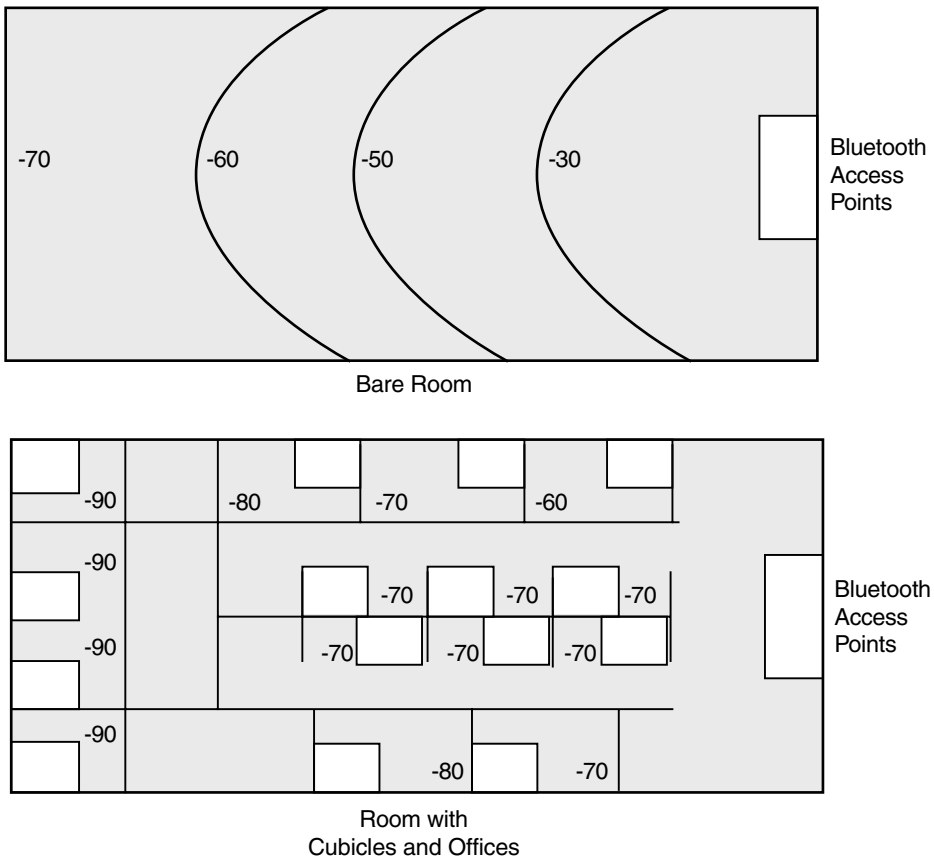


Figure 11.18 In-building received power.

In Figure 11.19 we show the minimum levels of attenuation normally experienced in a building, typically 6 or 7 dB floor to floor and 3 dB through a wall (without fire-resistant foil cladding). The attenuation through the exterior wall of the building will typically be 20 to 30 dB. This is good news and bad news. The good news is we probably do not want signals from inside the building to be visible outside the building and vice versa, both for security reasons and in order to deliver reasonable receive sensitivity.

However, present wireless LAN standards (IEEE802) include handover protocols that allow a user (theoretically) to move within a building (floor to floor) and into and out of buildings and still remain in continuous coverage. In practice, given the very substantial and rapid changes in signal level, it is very easy to drop a call under these conditions. It has also been proposed that wireless LAN to cellular handover should be supported. Again, in practice, this is hard to realize consistently because of the rapid fluctuation in received signal strength in the wireless LAN environment.

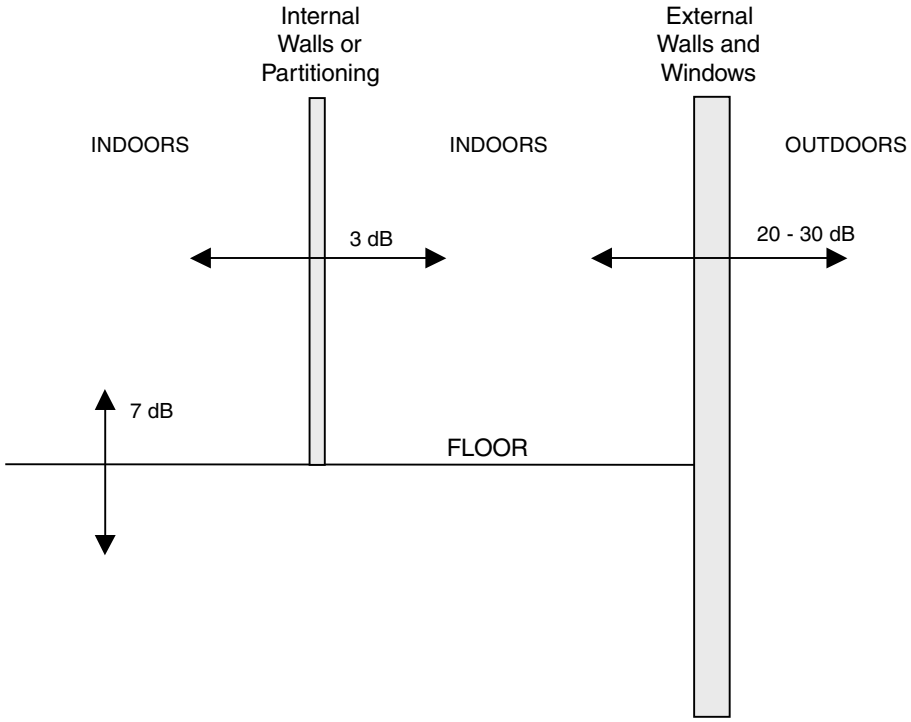


Figure 11.19 Minimum levels of attenuation in a building.

Designing a wireless LAN radio scheme is therefore a reasonably complex process and depends on having knowledge of the building configuration and building materials used. The process is not dissimilar to undertaking heat loss calculations/heat gain calculations from buildings where the sizing of the heating and cooling system is dependent on the building materials used, the size of windows, and whether windows are double or triple glazed. There are also many similarities with lighting design. Radio waves and light waves behave very similarly. In lighting calculations, we have to take into account the polar diagrams (also known as ISO candela diagrams) describing the light distribution available from the luminaire.

Unsurprisingly, this is very similar to looking at an RF antenna specification. Figure 11.20 shows a lighting product from Philips, and Figure 11.21 shows the related ISO candela diagram.

Figure 11.20 Example of a Philips lighting luminaire.

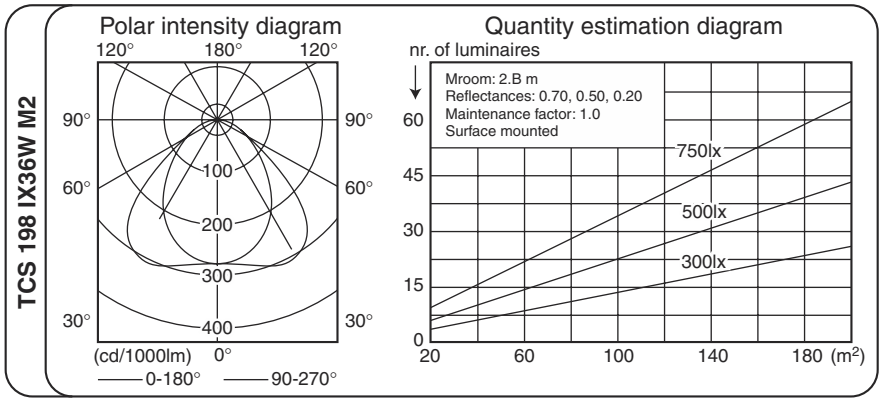


Figure 11.21 Isocandela flux intensity/polar intensity diagram for the luminaire in Figure 11.20.

It makes considerable sense to do a wireless coverage design at the same time as lighting design and heat loss calculations are being done for a building, since many of the inputs needed are common: the configuration and layout of the building and the materials used in the construction of the building. In lighting, we calculate lux intensity at various points in the room—the number of lumens on a desk, number of lumens reflecting off wall surfaces. In RF design, we calculate signal voltages available to RF receivers. The light output available from the luminaires (typically a few tens of Watts) is directly analogous to the RF power available from the wireless LAN transmitter (a few hundred milliWatts). Both lighting and RF design are directly affected by building geometry, user geometry (where people are in the building and what they are doing in the building), and the materials used in the building.

For further information on lighting design and integrated building services design, go to the Chartered Institute of Building Services Engineers Web site (www.cibse.org).

In designing for Bluetooth or IEEE802 RF wireless LAN coverage, we find, because of the wide variability of building geometry and building materials used, that range is not included in the specification, though there are guidelines offered based on the power output and receive sensitivity available. With Bluetooth, the guidelines suggest a range of 10 meters and 100-meter figures based on 0 dB and +20 dBm power output and assuming -70 dBm receiver sensitivity and -5 dBi antenna gain.

Although this may seem to be a reasonable assumption, in practice we have shown that there are typically attenuation effects of several tens of dBs to take into account. It is also in practice very difficult to deliver even moderate antenna efficiency because of space and size constraints and capacitive effects in handheld devices.

Cellular/Wireless LAN Integration

Ensuring consistent wireless LAN in building coverage is far from easy. Most installations are undertaken on the basis of rule of thumb estimates of how many transceivers are needed to provide continuous coverage across a given service area.

In many in-building environments, coverage from the cellular networks (either from outdoor microcells or indoor picocells) may well be more consistent than wireless LAN coverage. This makes handover protocols difficult to implement. Users will continuously be moved from wireless LAN to cellular coverage and back again. This in turn creates more discontinuity rather than less discontinuity in service provision. For these reasons, it is unlikely that wireless LAN/cellular technologies will be successfully integrated, at least for the immediately foreseeable future.

Distributed Antennas for In-Building Coverage

As we move from 2G to 3G technologies, base station form factor (at least temporarily) increases, the need to deliver more linearity increases base station Node B hardware footprint. At the same time, the minimum bandwidth available from a Node B transceiver is 5 MHz, compared to the minimum bandwidth of 200 kHz (an eight-slot, eight-channel single RF carrier mini GSM base station).

For in-building coverage we need small base stations and, often, not a lot of bandwidth. A base station in a small hotel foyer does not need 5 MHz of RF bandwidth. This makes distributed antennas quite attractive, certainly in the early stages of network deployment.

The idea of distributed antennas is to have a donor base station, say, in the basement of a large building. The RF signal is then distributed to a number of antennas mounted throughout the building. The problem with distributed antenna solutions is that losses in copper cable can be quite substantial.

One option is to use RF over fiber. The RF signal is converted to an optical signal using a linear laser and is then delivered down a fiber-optic cable. We cover RF over fiber in Chapter 13 ("Network Hardware Optimization").

Summary

In this chapter we reviewed some of the important system design considerations implicit in implementing a 3G network with a 3G radio physical layer. We have said that the radio physical layer directly influences network performance, and we address this in more detail in future chapters.

We discussed some of the design and performance parameters of the Node B. We said that physical size (form factor) is driven by the ever-decreasing size and volume of 2G base stations and that a particular design challenge is to deliver the additional linearity needed in 3G hardware within a sufficiently compact, lightweight product footprint.

Node B hardware determines how much offered traffic can be supported and how the offered traffic will be accommodated in terms of cell sectorization. We introduced some of the radio layer enhancements that are available, such as downtilt antennas, and highlighted the differences between handset RF design and Node B design and some of the options for implementing Node B hardware (RF/IF and baseband processing). We emphasized that the RF performance of the Node B (code orthogonality on the downlink and receive sensitivity on the uplink) directly influences radio system planning.

In addition, we reviewed some of the lessons learned from system planning in 1G and 2G cellular networks and pointed out that initial coverage and capacity simulations are often overoptimistic. The additional number of variables in CDMA planning make it harder to pin down likely system performance.

We also reviewed some of the present simulations reviewed in the present planning literature and advised some caution in how the present figures should be interpreted. We pointed out that not only Node B RF performance but also handset RF performance is a major component of the RF link budget and that both Node B and handset RF performance increase as the network technology matures (particularly if market volume is achieved—the performance advantage of volume). A 1 dB improvement in Node B on handset sensitivity translates into a 10 percent decrease in network density.

Finally, we reviewed indoor system planning and identified some of the significant attenuation effects introduced by building geometry and partitioning. We said that the rapid changes in signal level typical of in-building coverage presented particular challenges for managing handover in these environments. We suggested that there are commonalities between radio design or in-building coverage and lighting and heat loss/heat gain calculations.

In the longer term, a more integrated approach to radio planning and building design could well be beneficial.

CHAPTER 12

GSM-MAP/ANSI 41 Integration

We have just discussed some of the radio system planning parameters of IMT2000DS—how to deliver adaptive radio bandwidth and how to deliver consistent-quality bandwidth, with sufficient resilience to support persistent rich media sessions between duplex users. We described how radio bandwidth quality is one necessary and important component in the delivery of end-to-end performance guarantees. These guarantees form part of a user's service level agreement, which includes admission rights and policy rights stored in the SIM/USIM.

Approaching a Unified Standard

In a GSM-MAP network, it is the SIM/USIM that dictates or at least describes the quality of service requirements of the user or the user's application. This in turn determines the allocation of radio and network resources. Radio resources are provided either over an IMT2000DS air interface (with backward compatibility to GSM, GPRS and E-GPRS air interfaces) or a CDMA2000 air interface (with backward compatibility to IS95A, B, C).

In addition to having two similar but different air interfaces, we have, worldwide, two similar but different mobility network standards:

ANSI 41 network. Any U.S. TDMA or CDMA2000 air interface, or any AMPS air interface, either in the United States or Asia, will have behind it an ANSI 41 network.

GSM-MAP network. Any GSM or IMT2000DS air interface, either in the United States, Europe, or Asia, will have behind it a GSM-MAP network.

The differences between the two networks are by no means unbridgeable, particularly as both use SS7 signaling to manage network functionality. One practical and important difference historically is that GSM-MAP networks have used the smart card SIM as the basis for controlling radio access to the network, that is, user specific authorization. The user buys a SIM card and can put it into any GSM phone. The SIM card, not the phone, is the device that determines the user's access and priority rights.

In IS41/ANSI 41 networks to date, SIM cards have not been used. Instead, the device is validated for use on the network by virtue of its mobile identity number (MIN) and equipment identity number (EIN). This is now changing, as 3GPP2 (the body working with 3GPP1 on IMT2000DS/CDMA2000 integration) now support the use of the SIM (which in CDMA2000 is actually called an R-UIM—removable user identity module) as an access validation platform.

3GPP1 and 3GPP2 are working together to use the SIM/R-UIM as a basis for bringing together GSM-MAP and ANSI 41. Parallel work is under way to implement GAIT handsets (GSM/ANSI 41 handset interoperability) and the side-by-side compatibility of an ANSI 41 network with the GERAN (GSM/GPRS/EDGE radio access network) and UTRAN (UMTS radio access network), as shown in Figure 12.1. The U-SIM/R-UIM is the mechanism for defining a user's policy/conditional-access rights and is becoming an integral part of the IPQoS proposition.

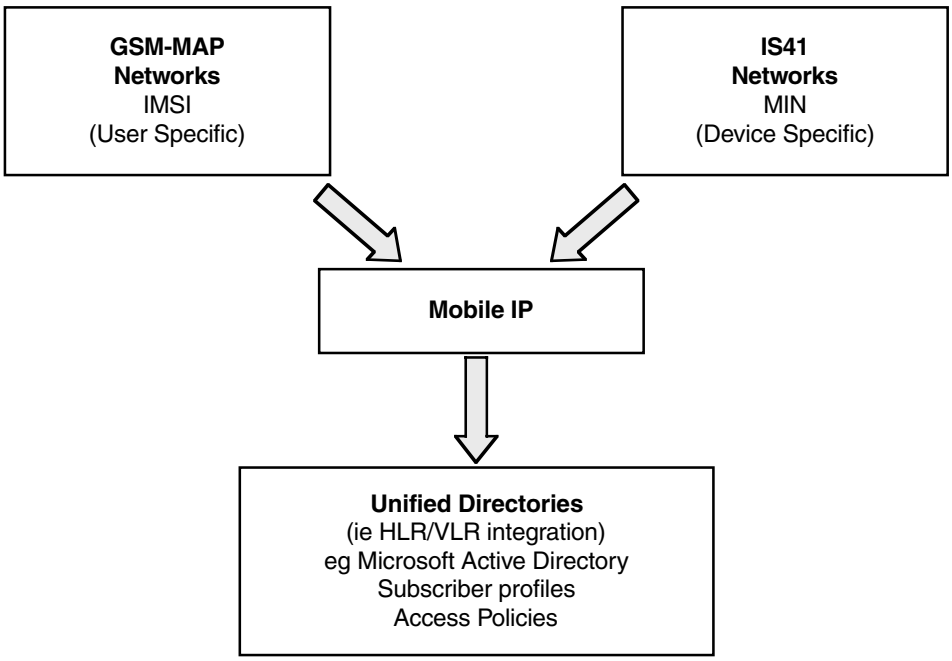


Figure 12.1 GSM-MAP/ANSI 41 integration.

The more radically inclined vendors see IP protocols as an additional mechanism for unification, potentially replacing existing Signaling System 7 (SS7) signaling, which is used to establish, maintain, and clear down telephone calls between users. SS7 provides the signaling control plane for wireless and wireline circuit-switched network topologies. It is a mature and stable standard. In a wireless network, additional functionality is needed to manage the allocation of radio channels (actually a channel pair for the uplink and downlink) and to support mobility management. This is known in GSM as GSM-MAP (Mobile Application Part).

SS7 is often described as an *out of band* signaling system—the signaling is kept functionally and physically separate from the user's voice or data exchange. In a packet-switched network, the routing of calls or sessions relies on a router reading the address on each packet or group of packets transmitted—an in-band signaling system using established Internet protocols (IP). There are standards groups presently working on bringing together IP and SS7 (IP SS7), and significant progress has been made on using both signaling systems to implement always on connectivity in wireline networks (for example, using ADSL).

The additional functionality needed to support wireless connectivity, however, creates a number of implementation problems, which are presently proving difficult to resolve. For example, in a GPRS network, a Packet Common Control Channel (PCCCH) and Packet Broadcast Control Channel (PBCCH) are needed to support always on connectivity. The PCCCH and PBCCH replace the existing Common Control Channel (CCCH) and Broadcast Control Channel (BCCH). There is presently no easy method for ensuring PCCCH- and PBCCH-compliant handsets are backward compatible with CCCH- and BCCH-compliant handsets. This sort of issue can be overcome, but it takes time.

In addition, some network operators question why they should abandon a tried and trusted signaling system that gives good visibility to system hardware performance (including warning of hardware failures) and is (accidentally) well suited to persistent session management. This is an important point. The first two parts in this book argued the case that session persistency would increase over time and become increasingly similar to voice traffic, although ideally with a longer holding time. As session persistency increases, out-of-band signaling becomes increasingly effective, which means session setup, session management, and session clear-down is directly analogous to call setup, call maintenance, and call clear-down.

IP could potentially replace SS7 but would need to emulate the session management and session reporting capabilities of SS7. We revisit this issue when we study traffic shaping and traffic management protocols, the subject of Chapters 16 and 17 in Part IV of this book.

Mobile Network Architectures

The traditional network architecture used in GSM-MAP and ANSI 41 is very hierarchical—a centralized mobile switch controller sits in the center of the network (see Figure 12.2). There may be a number of switches to cover a country. Each switch controls a number of base station controllers, which in turn support the local population of mobile users.

Figure 12.2 Traditional hierarchical network architecture.

Figure 12.3 (see also the following key to the diagram) shows a GSM-MAP network. It is a conventional wireline network based on ISDN, but with a mobility management overlay (Mobile Application Part). This provides the additional functionality needed to move users from cell to cell (power control and handover); to set up, maintain, and clear down mobile calls; and to bill for services provided to mobile users.

Going from left to right, the base stations talk to the BSC over the A-bis interface. This interface takes the 9.6 kbps, 14.4 kbps, or 13 kbps voice traffic from the mobiles (with some embedded signaling) and moves the traffic to and from the BSC over typically multiplexed (120) 16 kbps traffic channels within a 2 Mbps pipe. In this example, voice traffic is then transcoded from the 13 kbps (or 12.2 kbps EFR) codec stream to a

The Visitor Location Register needs to let the user's home network know that the user has moved. If someone now phones the user's home number, the user's call will be forwarded—at some expense—to the user via the visited network. The authentication register looks after SIM/U-SIM based user authentication and the equipment identity register matches the user to the equipment being used. (Stolen equipment can be barred from the network.) These mobility management functions involve, as you would expect, substantial signaling, and this is carried over the SS7 signaling layer on 64 kbps multiplexed land lines.

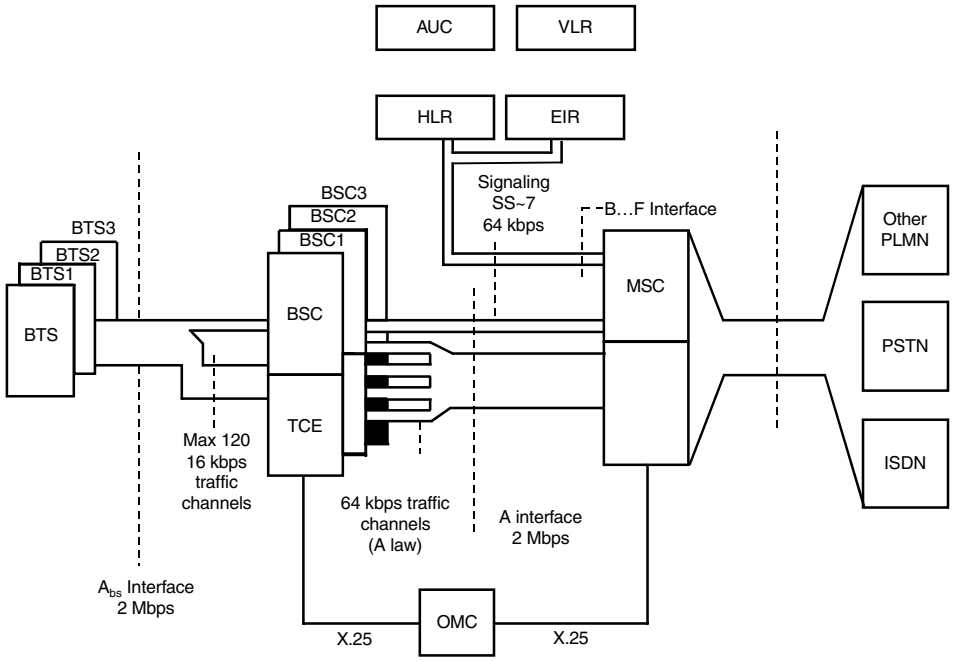


Figure 12.3 GSM network (GSM-MAP).

KEY TO FIGURE 12.3

AUC	Authentication Center
BSC	Base Station Controller
BTS	Base Transceiver Station
DAI	Digital Audio Interface—104 kbps
EIR	Equipment Identity Register
GSM	Global Systems for Mobile Comms
HLR	Home Location Register
ISDN	Integrated Services Digital Network
MS	Mobile Station
OMC	Operation and Maintenance Center
PLMN	Public Land Mobile Network—or Private
PSTN	Public Switched Telephone Network
VLR	Visitor Location Register
TCE	Transcoding Equipment

The mobility management overlay provides the information needed for billing, so it is arguably the most commercially important component in the network.

Traffic to and from mobile users is consolidated in the switch—hardware routed on the basis of the target phone number used in the call setup procedure. If the call is mobile to mobile, for example, the end-to-end link is determined by the sender's IMSI number and the receiver's IMSI. When a call setup request is received at the MSC, the MSC uses Layer 3 (network layer) signaling to allocate access network resources for the call via a BSC and BTS. Layer 3 talks to Layer 2 (the data link layer) to allocate logical channel resources via the BTS to the mobile. Layer 2 talks to Layer 1 to acquire physical channel resources (that is, time slots within an RF channel in GSM/TDMA). Figure 12.4 shows this layer modeling.

This all works fine when the traffic in both directions is more or less constant rate on a per-user basis. Average call length in a cellular network is about 2 minutes. Traffic loading can therefore be very accurately predicted. On the basis of these predictions, decisions can be taken on how much backhaul bandwidth to install (how many 2 Mbps lines to install).

Average call length is actually getting longer year by year as call rates reduce, and, anecdotally, younger people also seem to take more than their parents' share of time on the phone. So call length is increasing as more young people start using mobile phones. This nevertheless still represents quite predictable loading.

Historically, transmission bandwidth in the MSC and copper access network has been overprovisioned to ensure that grade of service is more or less equivalent to fixed access PSTN in terms of availability (so-called five 9s availability). You pick up the phone, and 99.999 percent of the time, you get a line, or put another way, there is a 1 in 10,000 chance of the network being engaged. In practice, the limitation in a mobile network tends to be the radio resource rather than network resources. One of the major rationales of moving to a packet network, however, is to reduce the cost of network transmission.

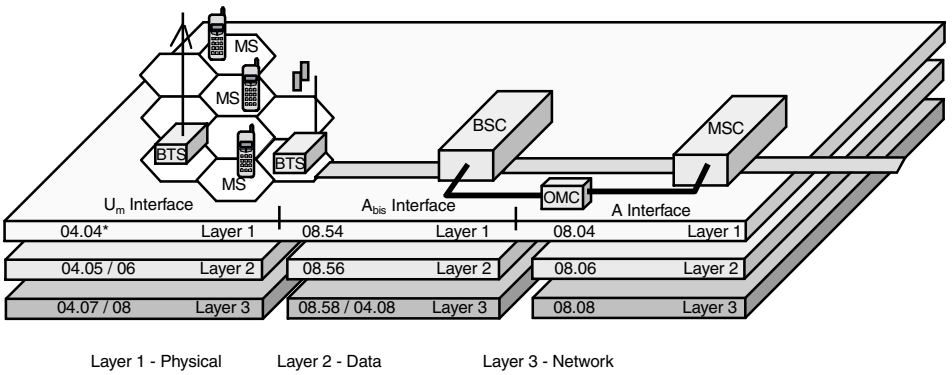


Figure 12.4 Layer modeling.

In a circuit-switched network, a logical channel and physical channel are established end to end for the duration of the call. The logical channel and physical channel exists in two directions simultaneously. Over the radio interface, the channel is a duplex spaced RF channel pair 45 MHz or 190 MHz apart or (in TDD) an uplink and downlink time slot. In a duplex voice conversation, we are only talking for approximately 35 percent of the time; that is, for more than 50 percent of the time we are either listening to the other person or pausing (to draw breath) between words. A pure packet-routed network avoids this wasted bandwidth. Packets are only sent when voice activity is detected.

In defense of circuit-switched networks, it is valid to point out that there is a fundamental difference between logical channel allocation and physical channel allocation. Over the radio air interface, a logical channel pair will have been allocated for the duration of a duplex voice call. However, if the handset and the base station are using discontinuous transmission (RF power is only generated when voice activity is detected), then there is no physical occupancy of the radio layer.

Similarly, because much of the core transmission network has been historically overprovisioned and, in many cases, fully amortized, increasing core network bandwidth utilization is neither necessary nor cost-effective. We do need to take into account, however, the increasingly bursty nature of the traffic being offered to the network; that is, we are justifying the transition to packet networks on the basis of their suitability for preserving the properties of bursty bandwidth—a quality rather than cost-saving justification.

It is as problematic as it is difficult to put a finite value on quality; how much is a 24-bit color depth 15 frame per second video stream worth compared to a 16-bit 12 frame per second video stream. Additionally, we need to factor in the extra costs incurred by deploying packet routing in the network. Figure 12.5 shows the first changes that have to be made—the addition of a GPRS or packet traffic support node.

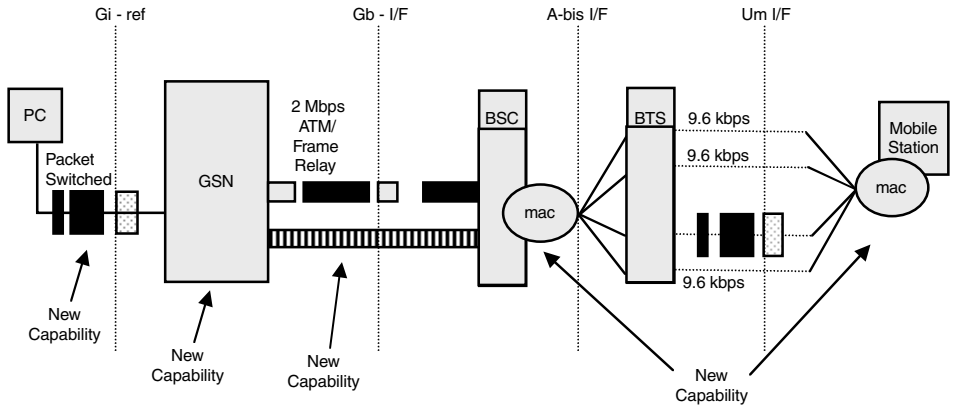


Figure 12.5 Packet-switched data service architecture (GSN-GPRS support node).

The GPRS support node talks to the BSC across a 2 Mbps (2.048 Mbps) ATM transport layer. (We case study ATM in Part IV of the book.) For the moment, all we need to know is that the ATM layer allows us to multiplex bursty traffic and maintain its time domain properties. What you put in at one end of the pipe comes out at the other end of the pipe unaltered—a bit like a filter with a constant group delay characteristic. The traffic experiences some delay because of the multiplexing—and some delay along the transmission path—but the delay is a constant and is equal for all offered traffic. At either end of the ATM pipe, we can, of course, buffer traffic and prioritize access in the ATM pipe. That is, traffic is all treated equally while it is inside the pipe but can be given differential transport priority before it gets into the pipe.

The example shown in Figure 12.5 highlights new MAC (Medium Access Control) functionality (Layer 2 functionality) in the mobile and BSC. This is to support high-speed circuit-switched data from a handset capable of using more than one time slot on the uplink and downlink; that is, variable bit rate can be delivered in increments of additional time slots (or in IS95, additional PN offsets). The A-bis and UM/IF interface, therefore, remains essentially unchanged.

Figure 12.6 shows the addition of a serving GSN that can manage simultaneous circuit-switch and packet-routed traffic talking to the gateway GSN using IPv6 (case studied later). The SGSN talks to the BSC over an ATM transport layer. The BSC talks to the BTS (also over ATM), and the BTS exchanges packets with the mobile using E-GPRS radio blocks to manage packet re-sends (covered earlier in Chapter 2, where we discussed system planning). It is interesting to note the continuing presence of an MSC and an interworking function to manage simultaneous packet-routed and circuit-switched traffic.

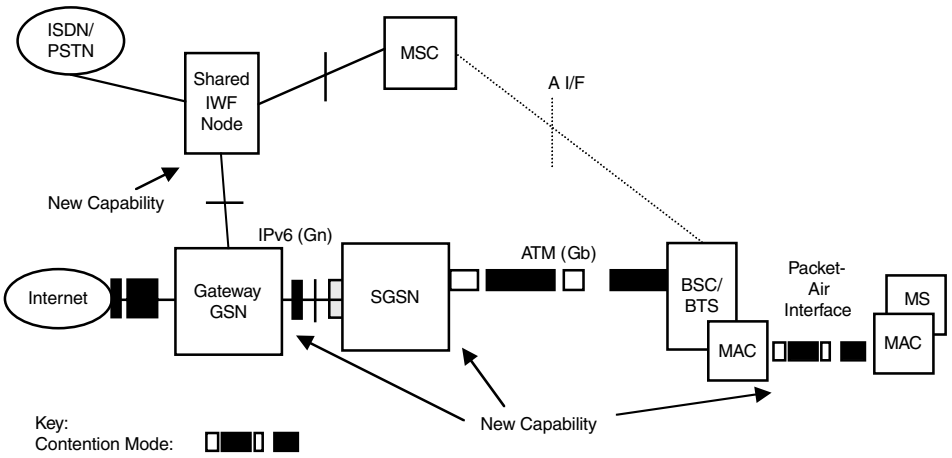


Figure 12.6 GPRS service platform—SGSN—serving GSN.

The forward compatibility selling point is that once an operator has put ATM in between the SGSN and BSC and BTS, it is then relatively easy to implement IMT2000DS with a MAC layer delivering dynamic rate matching on a 10-ms frame resolution, effectively wireless ATM. IPV6 can be used to provide some higher-layer prioritization of packet streams, establishing rights of access and priority/preemption entitlements to network and radio transmission bandwidth.

There is still (particularly in Europe and Asia) a strong circuit-switched feel to the network. ATM is a hardware-based implementation of virtual circuit switching. Remember also that the genesis of the 3GPP specification was to implement wireless ISDN over the radio physical layer.

GSM-MAP Evolution

The original standards documents described the three maximum data rates supportable. Originally the rates were 144 kbps, 384 kbps, and 2048 kbps—equivalent to IDSN 2B + D, ISDN HO, and the lowest entry-level ATM rate. Circuit-switched services would be supported up to 384 kbps, and higher data rates would be packet-switched; 144 kbps would be available in macrocells; 384 kbps would be available in microcells; and 2048 kbps would be available in picocells. The original chip spreading rate in the standard was 4.096 Mcps. This was chosen to support the ATM 2.048 kbps bearer—that is, 2048 kbps equals 1024 kilosymbols; 1024 kilosymbols times a chip cover of 4 equals 4.096 (1024×4).

The chip rate was then reduced. (This was done for political reasons, in the spirit of bringing the chip rate closer to the CDMA2000 chip rate.) However, as a consequence of this, adjacent channel performance improved. The cost was that the 2.048 kbps top user rate was reduced to 1920 kbps equivalent to ISDN H12—that is, 960 kilosymbols ($960 \times 4 = 3.84$ Mcps), as shown in Table 12.1.

There is less focus now on ISDN partly because of the increased need to support very variable user data rates. So, for example, we still use the ISDN rates as a maximum user data throughput but effectively provide an ATM end-to-end wireless and wireline channel for each individual user's packet stream (or multiple per-user traffic streams). Given this shift in emphasis, it was decided to try and improve bandwidth utilization in the ATM copper access transport layer by maintaining the DTX (discontinuous transmission) used over the radio layer as traffic moved into the network core. This is implemented using a protocol known as ATM AAL2.

Table 12.1 Current Maximum Supportable Data Rates

BIT RATES	ISDN NOMENCLATURE
144 kbps	ISDN 2B + D
384 kbps	ISDN H O
1920 kbps	ISDN H 12.

2048 kbps (the lowest ATM rate) was originally included in the 3GPP1 specification.

Our good friend SS7 still stays very much in charge of traffic flow control, but there are proposals to implement broadband SS7 on the basis that the existing 64 kbps-based signaling pipes will become too small. This work is being presently undertaken by 3GPP alongside proposals to implement an IP-based signaling bearer known as SCTP/IP (Signaling Control Transport Plane using IP).

Some network operators and vendors remain unconvinced of the merits of using IP to replace existing (tried, trusted, and effective) signaling protocols, so progress on standardization might be rather slower than expected.

The IUB interface between the Node B and the RNC is 2048 kbps ATM (2.048 Mbps), and the IU interface is 155 Mbps ATM. DTX is implemented using AAL2 across the IUB and In interface (into the core network). Each individual RNC looks after its own family of Node Bs except, referring to Figure 12.7, where a mobile is supported by two Node Bs under separate RNCs. This has to be managed by the IUR interface used by the RNCs to talk to one another. The RNCs also have to manage load balancing, which we covered in the previous chapter.

A typical RNC is configured to support several hundred Node Bs. The RNC is responsible for mobility management, call processing (session setup, session maintenance, session clear-down), radio resource allocation, link maintenance, and handover control. Note the RNC needs to make admission control decisions looking out toward the Node B on the basis of radio layer noise measurements, and admission control decisions looking inward to the core network on the basis of network congestion.

We discuss the software needed for this (reasonably complex) process in Chapter 16.

GPRS Support Nodes

In the core network, we have the GPRS support nodes. These nodes have responsibilities, described in the following sections, which are carried forward into a 3GPP packet-routed 3G network.

The SGSN Location Register

The serving GPRS support node is responsible for delivering data packets to and from mobiles within its service area and looks after packet routing, mobility management, authentication, and charging. The SGSN location register stores the location information (current cell, current VLR) and user profiles (IMSI packet data network addresses) of all GPRS users registered with the SGSN.

The GGSN GPRS Gateway Support Node

The GGSN is the interface between the GPRS backbone and the external packet data networks. It converts GPRS packets from the SGSN into the appropriate packet

data protocol (PDP) format (IP or X25). Incoming data packets have their PDP addresses converted to the GSM address of the destination user, and re-addressed packets are sent to the responsible SGSN. GSN to GSN interconnection is via an IP based GPRS backhaul. Within the backbone, PDN packets are encapsulated and transmitted using GPRS tunneling protocol.

A SGSN may need to route its packets over different GGSNs to reach different packet data networks.

Figure 12.7 shows the intra- and inter-PLMN (Public Land Mobile Network) interconnection. The intra-PLMN backbone connects GSNs of the same PLMN, that is, a private IP-based network specific to the GPRS network provider. Inter-PLMN backbones connect the GSNs of different operators (supported by an appropriate service level agreement).

The Gn and Gp interfaces allow the SGSNs to exchange user profiles when a user moves from one SGSN to another. The HLR stores the user profile, current SGSN address, and PDP for each GPRS user in the PLMN. The Gr interface is used to exchange information between the HLR and the SGSN. The Gs interface interconnects the SGSN and MSC/VLR database. The Gd interface interconnects the SMS gateway with the SGSN.

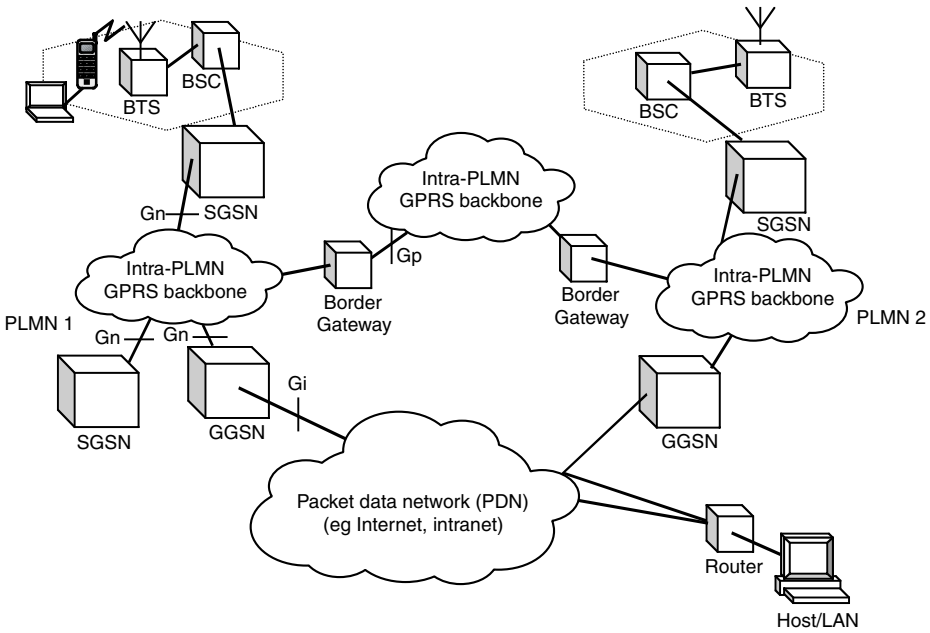


Figure 12.7 GPRS system architecture (routing example).

Table 12.2 GPRS Reliability Levels

RELIABILITY CLASS	LOST PACKET	DUPLICATED PACKET	OUT-OF-SEQUENCE PACKET	CORRUPTED PACKET
1	10^9	10^9	10^9	10^9
2	10^4	10^5	10^5	10^6
3	10^2	10^5	10^5	10^2

GPRS bearer services include point-to-point (PTP) services, which can be connectionless or connection-oriented (for example, X25), or point-to multipoint (PTM) services, for example, supporting multicasting within a geographical area—traditionally referred to in PMR as open channel working or group calling.

GPRS QoS is based on simple service precedence, reliability, delay, and throughput. Service precedence is either high, normal, or low. Table 12.2 shows the three classes of reliability.

GPRS QoS also has four classes of delay, listed in Table 12.3.

Delay is defined as end-to-end transfer time between two communicating handsets or between a handset and the Gi interface to the external PDN. It includes delay for the request and assignment of radio resources and transit delay in the GPRS backbone. Transfer delays outside the GPRS network are not included; as presently specified, GPRS does not support end-to-end guaranteed QoS. Throughput is specified as maximum, peak bit rate, and mean bit rate.

Table 12.3 GPRS Delay Classes

	128 BYTE PACKET		1024 BYTE PACKET	
CLASS	MEAN DELAY	95% DELAY	MEAN DELAY	95% DELAY
1	<0.5s	<1.5s	<2s	<7s
2	<5s	<25s	<15s	<75s
3	<50s	<250s	<75s	<375s
4	Best Effort	Best Effort	Best Effort	Best Effort

Session Management, Mobility Management, and Routing

When you turn on your GPRS handset, it registers with the SGSN of the serving GPRS network. The network authorizes the handset and copies the user profile from the HLR to the SGSN. It assigns a packet temporary mobile subscriber identity (P-TMSI) to the

user (GPRS attach). Detach can be handset or network initiated. After a successful attach, the handset must apply for one or more addresses used in the PDN, for example, an IP address. A PDP context is created for each session, for example:

- PDP type (IPv4/IPv6)
- PDP address
- Requested QoS
- Address of a GGSN serving as the access point to the PDN

The context is stored in the handset, the SGSN, and the GGSN.

Address allocation can either be static or dynamic. In static allocation, the network operator permanently assigns a PDP address to the user. In dynamic allocation, a PDP address is established when a PDP context is established (executed by the GGSN). This might be used, for example, to support prepay packet traffic.

To implement routing, the SGSN encapsulates the IP packets from the handset and examines the PDP context. The packets are routed through the intra-PLMN GPRS backbone to the appropriate GGSN. The GGSN decapsulates the packets and delivers them to the IP network.

Location Management

The network needs to keep track of where a GPRS user is physically to minimize uplink signaling and downlink delivery delay. The network has to rely on the GPRS handset telling it where it is—that is, which base station it is seeing.

The handset is either ready, in idle mode, or in standby, as follows:

- Ready means the handset has informed the SGSN of where it is.
- Idle means the network does not know the location of the handset.
- Standby means the network knows more or less where the handset (that is, within a certain location area subdivided into several routing areas, which will generally consist of several cell sites).

The status of the handset is determined by timeouts.

If an MS moves to a new RA, it produces a routing area update request to the SGSN. The SGSN assigns a new P-TMSI to the user. It does not need to inform the GGSN or HLR, since the routing context has not changed.

Micro and Macro Mobility Management

At a micro level, the SGSN tracks the current routing area or cell in which the handset is operating. At a macro level, the network needs to keep track of the current SGSN. This information is stored in the HLR, VLR, and Gateway GPRS Service Node (GGSN).

The more radical proposals of IP everywhere argue that the HLR, VLR, and GGSN could be replaced with DHCP servers, which could handle dynamic IP4 address allocation, subscriber profiles, and access policies using standard Microsoft Active Directory software. However, the HLR and VLR capability is very well proven within hundreds of GSM networks, so it is unlikely that this change will happen quickly if at all.

Radio Resource Allocation

It is important to differentiate physical and logical channels. A *physical channel* is denoted as a packet data channel (PDCH), taken from a common pool from the cell. Allocation can be driven by traffic load, priority, and multislot class (see Table 12.4).

Logical channels are divided into traffic and signaling (control) channels. One handset can use several PDTCH (data traffic) channels. A packet broadcast channel PBCH supports point-to-multipoint services and carries information on available circuit-switched bearers.

A handset-originated packet transfer can be done in one or two steps. The two-step process involves a resource request and then a channel assignment (logical channel request followed by physical channel assignment) or both steps can be done at once. On the downlink (base to handset), the handset is paged, the mobile requests a physical channel, and the packet is sent.

A physical data channel has a multiframe structure of 52 frames, which is 240 ms long. Note that a 26-frame multiframe (120 ms) is identical to an IMT2000DS multiframe (12×10 ms frames).

As the offered traffic is moved into the network, it is controlled by the data link and transport link layer protocols. GPRS tunneling protocol (GTP) is used to transfer data packets over the transmission plane managed by the GTP tunnel control and management protocol (using the signaling plane) to create, modify, or delete tunnels. UDP is used for access to IP-based packet data networks, which do not expect reliability in the network layer or below. IP is employed in the network layer to route packets through the backbone. All of the packets are carried by the ATM transport layer.

Table 12.4 Radio Resource Allocation Logical Channels

GROUP	CHANNEL	FUNCTION	DIRECTION
Packet data traffic channel	PDTCH	Data traffic	MS (to/from) BSS
Packet broadcast control channel	PBCCH	Broadcast control	MS (from) BSS
Packet common control channel (PCCCH)	PRACH	Random access	MS (to) BSS
	PAGCH	Access grant	MS (from) BSS
	PPCH	Paging	MS (from) BSS
	PNCH	Notification	MS (from) BSS
Packet dedicated control channels	PACCH	Associated control	MS (to/from) BSS
	PTCCH	Timing advance control	MS (to/from) BSS

Operation and Maintenance Center

In Figure 12.3 we showed the OMC as a network component with responsibility for the physical transport links between the RNC and Node Bs, and legacy and 3G service platforms. The RNC monitors the status of the transport links in the network—for example, hardware or software failures. The network could be GPRS, EDGE, or UMTS (UTRAN) or any corporate virtual private network implemented by the operator. As with the RNC, this is a reasonably complex function. In theory, you should be able to use one vendor's OMC with another vendor's RNC or Node B, but in practice there are many vendor-specific variables in terms of implementation.

Summary

We described the major network components in a GPRS network (refer to Figures 12.5 and 12.6), including the SGSN and GSN (providing the gateway to other packet- or circuit-switched networks). We also showed that Internet protocols are present but not pervasive in existing networks, and that practical implementation, particularly in 3GPP1 IMT2000DS/UTRAN networks, is very much based on ATM, both on the copper access and radio access side.

E-GPRS uses higher-level modulation to increase the bit rate over existing GPRS networks. E-GPRS also supports burst error profiles, which helps to make the radio channel more adaptive and helps to reduce retransmission and retransmission delay. GPRS and E-GPRS networks, however, do not, at time of writing, provide robust end-to-end performance guarantees and are unlikely to in the future, as this functionality is not described in the standard.

ATM is increasingly pervasive as a hardware-based distributed switch solution for managing a complex multiplex of time interdependent rich media data streams. The multiplex carries on over the radio layer (which is effectively wireless ATM). 3GPP1 networks are not IP networks but, rather, ATM networks, supporting IP addressing rather than IP-routed traffic streams. We return to this subject in Part IV of this book, which is devoted to network software.

It is also difficult to see how GPRS can ever deliver sufficient dynamic range to support highly burst offered traffic fired into the network from next-generation handsets. 3GPP1 determines a dynamic range excursion of 15 kbps to 960 kbps between two successive 10-ms frames. GPRS, as presently configured, is not able to support this.

Networks still using A-bis interfaces are also constrained on the copper access side. ATM is needed on the IUB and IU interface for managing the incoming and outgoing multimedia multiplex. This implies a significant upgrade to existing copper access connectivity. Copper access quality and copper access bit rate flexibility are two necessary preconditions for preserving rich media value.

Network bandwidth quality is dependent on network hardware quality. Software routing is, generally speaking, insufficiently deterministic and insufficiently fast to process bursty aggregated traffic as it moves into the network core. If IP protocols are used, then substantial use of hardware coprocessing is required to deliver sufficient network performance.

ATM (hardware-based circuit switching) is an alternative now being widely deployed in 3GPP1 networks. It provides generally better measurement capabilities than IP, which, in turn, makes it easier to implement quality-based billing. This suggests that future network evolution may be more about optimizing ATM performance over both the radio and network layer than optimizing IP performance.

On the one hand, we argued that future network value is very dependent on software added value—our million lines of code in every handset. On the other hand, network performance is still very dependent on network hardware, and radio performance is still very dependent on radio hardware, which brings us to our next chapter.

CHAPTER

13

Network Hardware Optimization

In this chapter we review some radio hardware optimization opportunities and their impact on radio bandwidth quality, as well as optical hardware optimization opportunities and their impact on network bandwidth quality). But, first, we need to go back to school and review some basic concepts.

A Primer on Antennas

Figure 13.1 reminds us of how a radio wave travels through free space with an electric field component and a magnetic field component. The distance from trough to trough is the wavelength; the number of waves passing in Hz (cycles per second) is the frequency. Antennas are used to transmit or receive these waves. Antennas are passive components dimensioned to resonate at a particular frequency or band of frequencies.

In Chapter 1 we showed how wavelength decreases with frequency (see Table 1.1). We normally design handset and base station antennas to resonate at fractions of a wavelength. Antennas therefore become more compact as frequency increases, but they also become less efficient and more subject to localized effects such as coupling between antennas on a mast or between antennas and the mast or (in handsets) capacitive coupling effects (the effect of our hand on the outside of the phone).

On handsets, fashion now determines either internal antennas or external stub antennas, which may or may not be 1/4 wave or 1/8 wave. These are inefficient lossy devices. A number of companies have developed proprietary techniques for improving handset performance, for example, by using polarization diversity (capturing both

vertical and horizontal plane energy) or spatial diversity (an antenna either end of the handset). However, fundamental space constraints mean handset antennas are a serious compromise in terms of performance.

As antennas are passive devices, they can only radiate the same amount of energy that is supplied to them. The fundamental reference antenna is the *isotropic radiator*—a theoretical antenna that radiates a total sphere of energy. An antenna is said to have gain if it is dimensioned to focus or concentrate this energy into a specific pattern, direction, or beam. The gain is the ratio of the field strength that would be received at a specific point from the isotropic radiator to the field strength that is received at the same point from the directional antenna. The gain is dimensioned in dBi (dB isotropic).

A practical reference antenna is the *quarter-wave dipole*, and the gain of a directional antenna may also be expressed in dBd (dB referenced to a quarter-wave dipole). Effective isotropic radiated power (EIRP) is the product of the transmitter power and the gain of the transmit antenna. It is expressed in dBW, where 0 dB = 1 W.

We have said there is not much we can do to improve handset antenna performance. Typically, handset antennas show a negative gain—a loss of 1 or 2 dB, or more. Base station antennas, however, give us much more potential for improvement or, rather, producing gain where we need it. Remember: We are also particularly interested in being able to null out unwanted interference that adds unnecessarily to the noise floor of our Node B receiver.

Antenna design principles have not changed much in 70 years. Let's just review some of the basics.

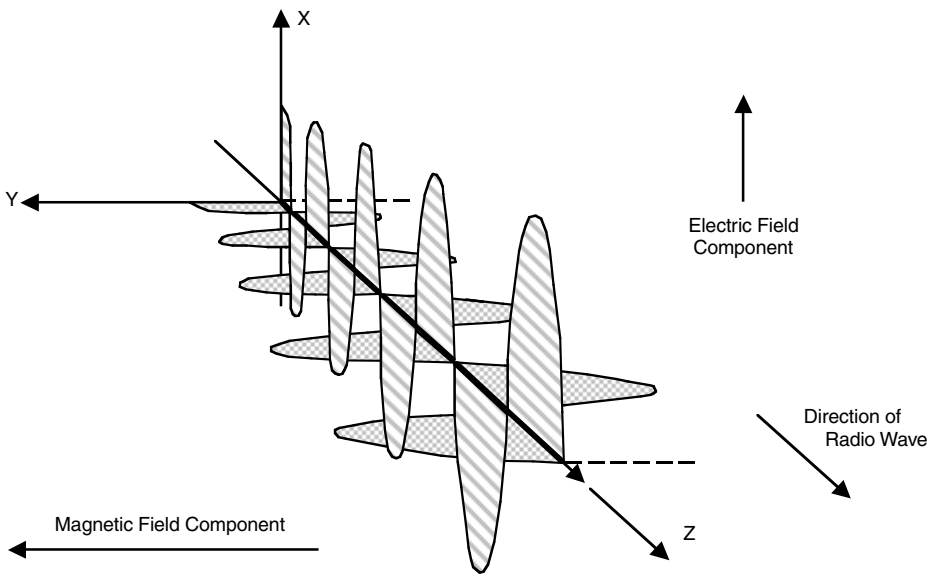
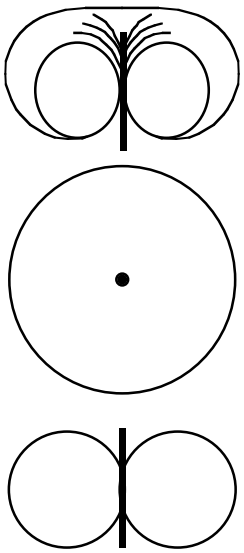


Figure 13.1 Propagation—the wave components.



Simplest Form	Half wave dipole Centre fed End fed
E-Plane	Electric
H-Plane	Magnetic
Vertical Dipole	Vertical Polarisation
Half Wave Dipole	Unity Gain

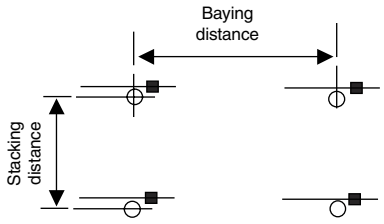
Figure 13.2 Dipole propagation.

Dipole Antennas

The simplest antenna is a dipole, which is either end fed (a pole on a pole) or center fed (a pole off a pole). Figure 13.2 shows the E-plane and the H-plane and the isometric radiation pattern. Imagine if you sat on the doughnut-shaped isometric pattern and squashed it. You would extend the radius looking down on the doughnut from the top, but you would also squash the profile of the doughnut looking at it from the side. That's the theory of antennas in a nutshell (or rather a doughnut).

Directional Antennas

We can create a directional antenna either by putting reflectors behind the driven elements or putting directors in front of the driven elements . A TV aerial is an example of a directional antenna; however, a TV aerial is just receiving, whereas we need to transmit and receive. The more elements we add, the higher the forward gain but the narrower the beamwidth (the bandwidth of the antenna also reduces). Doubling antenna aperture doubles the gain (+3 dB). However, doubling the aperture of the antenna doubles its size, which can create wind loading problems on a mast. A 24-element directional antenna will give a 15 dB gain with a 25° beamwidth but can really only be used at microwave frequencies. Eight element antennas are quite often used at high-band VHF and four-element antennas at low-band VHF.



Stacking and Baying Yagis

Frequency Band	Wavelength Spacing	Distance (m)	Center Frequency (MHz)
Low band	1 λ	3.75	80
Mid band	1.5 λ	3.75	120
High band	1.5 λ	2.81	160
UHF 400 - 470	2.0 λ	1.30	460
790 - 960	2.5 λ	0.88	850
1500	2.5 λ	0.50	1500

Figure 13.3 Stacking and baying.

We can increase the aperture of an antenna by coupling it with other antennas either stacked vertically or bayed horizontally (see Figure 13.3). Every time we double the number of antennas we double the gain. However, when we stack two antennas, we halve the vertical beamwidth; when we bay two antennas, we halve the horizontal beamwidth. At some stage, the coupling losses involved in combining multiple antennas exceeds the gain achieved. We are also adding cost (mast occupancy and wind loading) and complexity.

If we change the wavelength distance between the antennas, we can create nulls on either side of the forward beam (see Figure 13.4). We can use this to reduce interference to other users and to reduce the interference that the base station sees in the receive path.

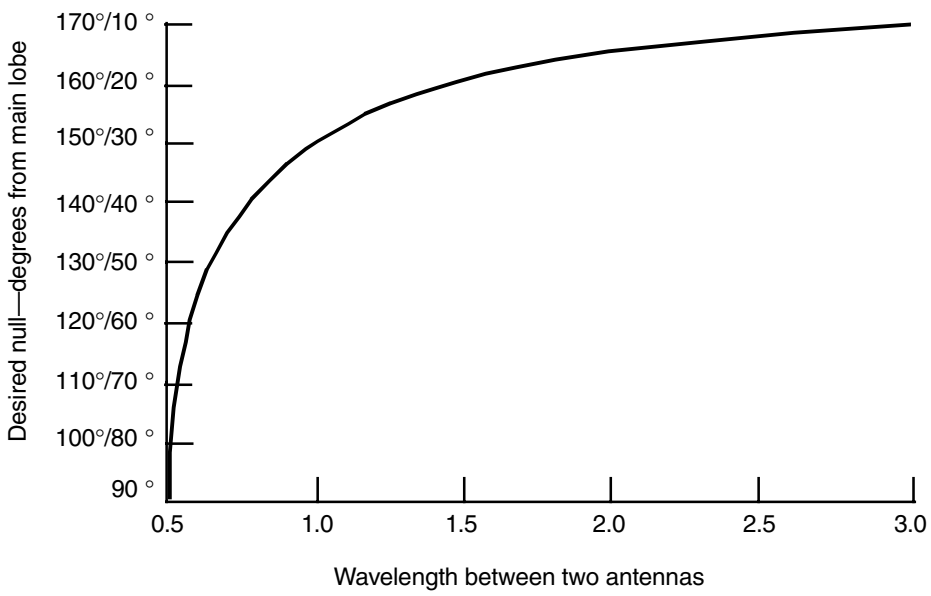


Figure 13.4 Nulling.

Omnidirectional Antennas

Omnidirectional antennas can be end fed and end mounted, or they can be center fed. The example shown in Figure 13.5 is two quarter-wave halves fed in the center, cabled down inside one of the dipole arms. Two end-fed dipoles stacked on top of each other give 3 dB of gain (our squashed doughnut). The beamwidth is narrowed in the vertical plane, but the omnidirectional pattern is maintained in the horizontal plane, and the radius is increased.

Four dipoles in a stack will give 6 dB of gain. In Figure 13.6, the inclusion of VSWR (Voltage Standing Wave Ratio) is a figure of merit. This gives an indication of how well the antennas will match to the transmitter—that is, how much power will be transmitted into free space and how much will be reflected back to the transmitter. An antenna with a VSWR of 1.5:1 will dissipate 95 percent of the power applied to it. In very wideband devices, VSWR can be 2:1 or worse. This example shows how matching deteriorates as you move away from center frequency.

If you change the phase matching between antenna elements, you can uptilt or downtilt the antenna. This is the basis for the adaptive downtilt antennas used on some Node B base stations.

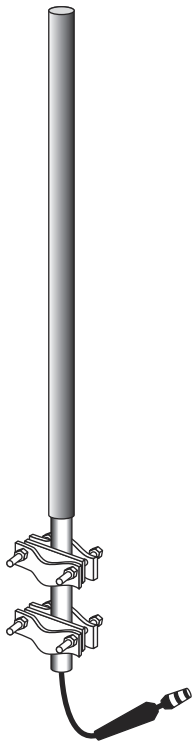


Figure 13.5 Omnidirectional antenna.

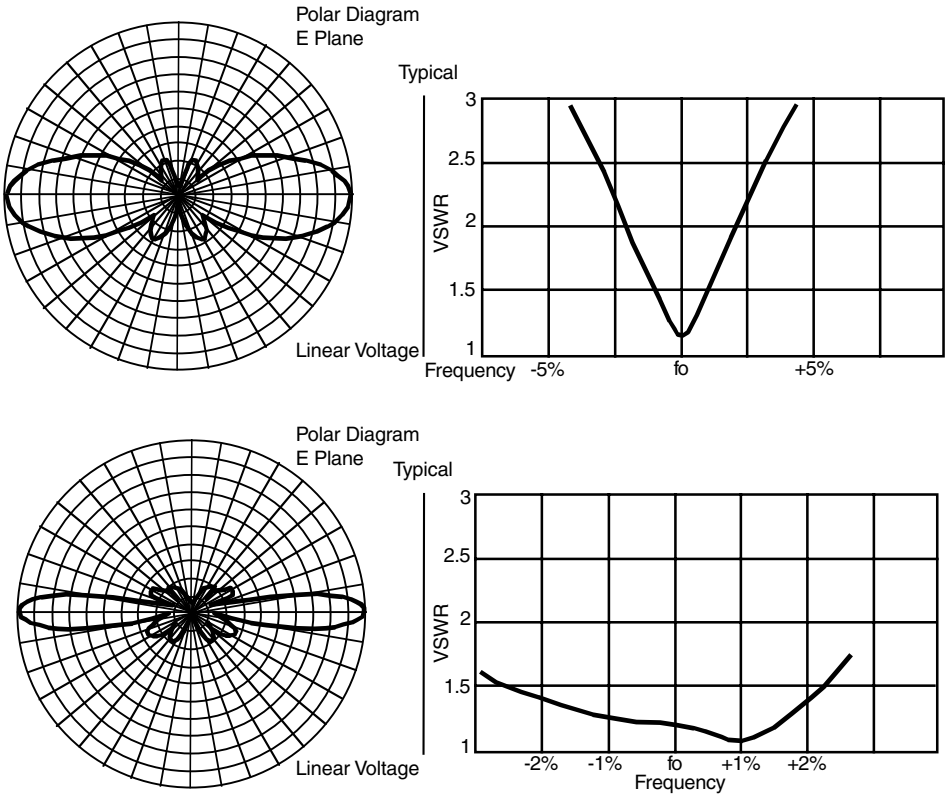


Figure 13.6 VSWR—the matching of antenna power.

If you bay two omnidirectional antennas, you can create any number of different footprints by changing the physical spacing between the two antennas. You can make this adaptive by electrically changing the phase (rather than physically moving the antennas). This is the basis for adaptive antenna design (changing the E-plane and H-plane characteristics of an antenna array).

If you take an omnidirectional colinear and give it a substantial amount of gain, the cell radius will increase, but a hole will appear in the coverage close to the base station. This can be reduced by using electrical downtilt (though it is generally better to have designed the antenna installation correctly in the first place). This is known as the "polo mint" effect.

When you mount antennas on a metal structure (that is, a mast), the antennas should always be at least two wavelengths away from the structure—not particularly a problem at 2 GHz. More of a problem is having multiple antennas close to each other, which then intermodulate with each other. Generally, an isolation of at least 40 to 45 dB is needed between a transmit and receive antenna, and 20 to 25 dB between two transmit antennas.

Dish Antennas

To achieve maximum gain from an antenna, you need to use a dish. The gain of a dish is equal to $(\pi D/\lambda)^2 \times N$, where D is the diameter of the dish, λ is the wavelength, and N is efficiency. Dish antennas are used extensively in cellular networks to move traffic to and from, and between cell sites. More commonly we see dish antennas installed for the reception of satellite TV.

Gain is limited by the tolerance of the reflective surface. Gains of between 30 and 50 dBi are achievable with 1° beamwidths. As frequency increases, particularly above 10 GHz, weather effects become very significant. We study this in Chapter 15, which deals with fixed access wireless.

Installation Considerations

As beamwidth gets narrower, you have to become increasingly accurate when pointing the antenna. Also, if it's windy, the antenna might move, and the link budget will disappear. If the microwave dish is mounted too low on a mast, the horizon (curvature of the Earth) will cause reflection and refraction and will limit the path length.

Dealing with Cable Loss

If an antenna is mounted on a mast remotely from the equipment hut, then feeder loss needs to be taken into account in the link budget. The fatter the feeder, the lower the loss but the more expensive the feeder. Also, the higher the frequency, the higher the loss. In PMR installations at VHF and UHF frequencies, you often see ⅜- or ½-inch feeder used (good imperial measurements here). As frequency increases (for example, 1800 MHz/1900 MHz) either ⅞- or often 1 ½-inch feeder has to be used. Not only is this expensive in material terms, it is also hard to handle because of the bending radius of the copper waveguide inside the corrugated outer plastic jacket (1 ½-inch feeder comes in very large rolls).

Even with fat feeder, a loss of 1 or 2 dB between the antenna and the equipment hut is quite common. This is why there has been an increasing use of mast-mounted low-noise amplifiers in cellular applications. Putting at least the first stage of the Low-Noise Amplifier (LNA) ahead of the feeder loss improves receive sensitivity. Similarly, it is quite common to find High-Power Amplifiers (HPAs) installed at the mast head to maximize transmitted power.

An additional consideration is lightning protection. Earth bonding may be needed to protect people in the hut. Earthing often does not protect equipment, which may well be damaged by a direct strike on the tower. Mast-mounted HPAs and LNAs are intrinsically quite vulnerable to lightning damage.

Smart Antennas

We have shown that you can take a passive antenna and change its coverage footprint by changing the antenna aperture (its size), or by using multiple antennas and physically moving them further apart or closer together. Adaptive antennas effectively take

multiple antennas and move them further apart or closer together electrically—by changing the phase and amplitude relationship between the antenna elements—rather than physically. These are called *smart antennas*, because they can be made smart enough to deliver specific downlink coverage footprints. On the uplink, smart antennas can be used to null out unwanted interference, improving sensitivity.

We said earlier that the difference in free space loss between 900 and 1800 MHz is 6 dB, and the actual loss is typically 1 or 2 dB higher because of increased refraction and reflection at the higher frequency. Increasing our operating frequency to 2 GHz (for IMT2000) adds another dB or so to the loss, which needs to be accommodated in the link budget. By providing additional selective gain, smart antennas can recover some of this lost link budget but at a price.

Smart antennas come in two flavors—switched beam and adaptive. The product from Nortel (see Figure 13.7) is a switched beam smart antenna giving 7 dB gain from a three-sector, four-element array (voting on a slot-by-slot basis between antennas). The HPA and LNA are both mast mounted.

The 7 dB gain effectively means, in theory, that the same coverage is available at 1800 MHz as was available from a 0 dB antenna at 900 MHz. In practice, many 900-MHz cellular installations already use quite substantial gain, for instance, using two physically spaced receive antennas to give diversity gain on the uplink. Typically, about 17 dB of gain is available on the uplink and about 9 dB on the downlink to give a balanced uplink/downlink link budget (because mobiles have less power than base stations). These antenna arrays are typically 65° beamwidth or 90° beamwidth.

The Flexibility Benefit

Smart antennas therefore have to produce some tangible cost/performance benefits over and above conventional passive antenna solutions. One potential benefit is *flexibility*, which is the ability to adapt to changing coverage requirements and a changing interference environment. The flexibility comes from using DSPs to process incoming signals (signal strength, C/I, and angle of arrival) to determine the optimum antenna coverage/visibility footprint.

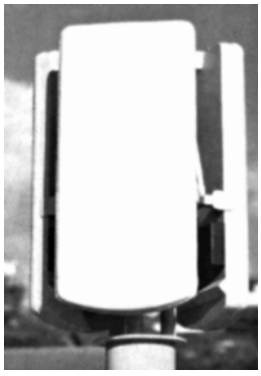


Figure 13.7 Link gain product from Nortel.

The job essentially means doing a lot of parallel processing. (Distributed DSP architectures are good at this.) The algorithms involved are similar to the algorithms used in direction finding, antimissile missile systems (where speed and angle of arrival need to be determined), and medical ultrasound imaging.

Switched Beam Antennas versus Adaptive Antennas

Nonadaptive switched beam antennas select from a range of precalculated beam patterns. Adaptive antennas adjust a weight vector to optimize the beam pattern to the noise and interference environment.

An adaptive antenna maximizes downlink coverage and minimizes uplink interference (by nulling out unwanted signal energy). A nonadaptive switched beam antenna will, by definition, be less efficient than an adaptive antenna but requires considerably less processing overhead.

Adaptive antennas deliver maximum gain in significant interface environments. In low-interference environments (for example, rural areas), switched beams perform almost as well as adaptive.

One practical issue to consider is the impact of these link gain products on other performance metrics—for example, dropped call rates. If you have a switched beam antenna on a site near a highway, say, with $6 \times 60°$ beams facing the highway, a close-in user traveling along the highway will pass through six sectors (separate beams) in a few milliseconds. It is very hard to manage power control and handover, and very easy to lose the user. If the user is three or four miles away, even if he is traveling fast, his rate of progress through the beam patterns will be relatively slow. These issues have to date prevented the wide-scale deployment of smart antennas (in addition to cost and complexity considerations).

Additionally, smart antennas require a very linear transmit/receive path (linear HPA). This means phase accuracy is important, because phase offsets are used to change the beam pattern. If the linearity is needed anyway (for example, in a 3G Node B), then this is not an applicable additional overhead.

Smart antennas can be used to solve internetwork interference problems. In PHS in Japan, where three network operators shared spectrum in a nonduplex nonpaired band allocation, the operators' networks did not clock together. The internetwork, interuser, intersymbol interface created capacity problems. (PHS uses an Ethernet-type MAC access protocol. If a time slot is occupied, the transceiver moves to another time slot). The ISI was detected as occupied bandwidth (that is, the capacity disappeared). The problem was alleviated by using smart antennas to null out interference.

Conventional versus Smart Antennas

Conventional passive antenna design continues to improve—better materials, better matching and coupling techniques, provision of polarization diversity and space diversity to provide uplink gain, and careful sectorization to deliver downlink gain.

Electronic downtilt antennas are semi-smart in that they can be made adaptive to changing interference conditions or can shift loading (if you shrink the cell radius, you support fewer users). Truly adaptive beam forming antennas are, to date, not in widespread deployment; however, the IMT2000 5 MHz allocations, particularly the

allocation of adjacent nonpaired bands, provide plenty of opportunity for co-channel internetwork interference, which may make smart antennas a far more plausible economic proposition.

For the moment, conventional passive antennas answer the majority of deployment requirements. The example shown in Table 13.1 is a typical dual polarized antenna with some downtilt (not dynamically adaptable). The dual polarization is on the receive path. Beamwidth is 90° and the antenna has a 20 dB front-to-back ratio, which means signals from the back of the antenna will be attenuated by 20 dB (relative to the incoming signals from the front) when received by the antenna.

Table 13.1 Dual Polarization Antenna—800 MHz (E-Systems) vs. PCS 1900 (E-Systems)

	800 MHZ	PCS 1900 MHZ
FREQUENCIES		
Receive	A Band: 825 to 847 MHz B Band: 835 to 849 MHz	1850 to 1910 MHz
Transmit	A Band: 870 to 892 MHz B Band: 880 to 894 MHz	1930 to 1990 MHz
GAIN	+12 dBi	+16 dBi (nominal)
POLARIZATION		
Receive	Dual VP and HP	Dual VP and HP
Transmit	VP	VP
BEAMWIDTH	Azimuth: 90 to 92° Elevation: 11.25 to 11.5°	Azimuth: 90 to 105° Elevation: 7 to 10°
PATTERN DOWNTILT	1.5° (other angles available)	1.5° (other angles available)
FRONT TO BACK RATIO	20 dB	25 dB
POWER HANDLING	500 W	250 W

Table 13.1 *(Continued)*

	800 MHZ	PCS 1900 MHZ
VSWR	1.5 to 1	1.5 to 1
INTERMOD		< 152 dB
DIMENSIONS	Height: 6.5 ft Antenna diameter: 6 ins	Height: 64 ins Antenna diameter: 6 ins
INTERNAL STRUCTURES	Aluminum	Aluminum
MAX EXPOSED AREA	3.25 ft	2.67 ft^2
WIND LOAD AT 110 MPH	217 lb point load	217 lb point load
LIGHTNING PROTECTION	All metal parts grounded; Lightning rod included	All metal parts grounded; Lightning rod included
NET WEIGHT	< 50 lb	< 40 lb

In Table 13.1, an increase in gain is achieved at 1900 MHz (+16 dBi rather than +12 dBi) at the cost of a slightly reduced vertical beamwidth (elevation). Front-to-back ratio is also better for the 1900 MHz antenna. Power handling is less (higher losses at 1900 MHz create more heat gain). Wind loading for the 800 and 1900 MHz is quoted as the same, although the 1900 Hz antenna is, as you would expect, smaller and lighter. The intermodulation figure (<152 dB) is also quoted for the 1900 MHz antenna

Companies such as Raytheon and Paratek have gained considerable experience with smart antennas in the satellite space sector (including Iridium and Globalstar). The trick is to bring the technology down to earth at a down-to-earth price. The example shown in Figure 13.8 is from Paratek (www.paratek.com) and is part of its DRWIN (Dynamically Reconfigurable Wireless Network) proposition. Lucent has a similar product proposition.

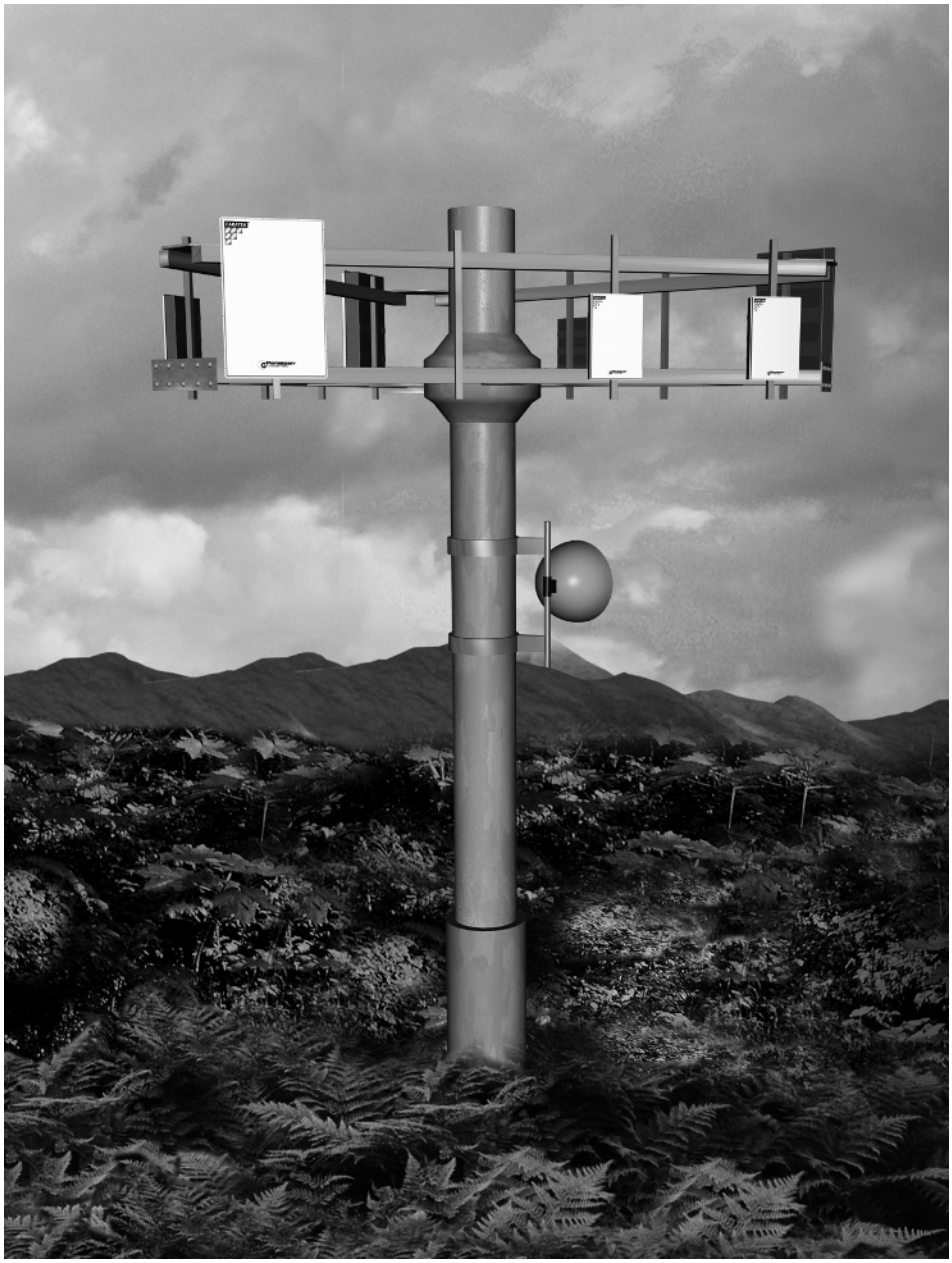


Figure 13.8 Paratek's DRWiN smart antennas.

Distributed Antennas

For in-building coverage we need small base stations and often not a lot of bandwidth. A base station in a small hotel foyer does not need 5 MHz of RF bandwidth. Neither do we want antennas visible within a public building if it can be avoided. This makes distributed antennas quite attractive.

The idea of distributed antennas is to have a donor base station, say, in the basement of a large building. The RF signal is then distributed to a number of antennas mounted throughout the building. The problem with distributed antenna solutions is that losses in copper cable can be quite substantial. One option is to use RF over fiber—a topic dealt with in detail later in this chapter.

A Note about Link Budgets and Power

We have talked about the need to deliver link budget gain by using various techniques (for example, smart antennas). Another way of improving link budget is to move nearer to the base station.

Note that in GSM, the biggest cell radius supported is 35 km. In W-CDMA (3GPP1), the maximum cell radius is 20 km, and the maximum data rate is 144 kbps. A microcell is defined as having a coverage radius of 300 meters, a picocell, a coverage radius of 100 meters.

In either a macrocell, microcell, or picocell, as you move closer to the Node B (base station), you have more power available both on the uplink and downlink to support higher data rates. The dynamic range in a macrocell is typically 70 dB (the difference between being very close to the cell and right at the cell edge). However, you are also trying to follow the fast fading envelope, which means tracking short-term fades of 20 dB or more, so it could be argued the usable dynamic range is typically 50 dB. Nevertheless, this is still a huge difference in link budget. As we move closer to the base station, we can either reduce coding overhead (that is, increase user data rate) or increase modulation complexity—for example, move to 8-level PSK or 16-level QAM in a 3GPP2 air interface (CDMA 2000) to take advantage of the increase in available power.

The problem—which is pretty unavoidable—is that real throughput will vary depending on the user's distance from the base station, which makes it difficult to deliver a consistent user experience. The general principle, then, is just to provide as good a link budget as we can, given the power and cost constraints we have to meet.

In Chapter 4, which dealt with handset hardware, we pointed out that battery bandwidth determines the amount of offered uplink delivery bandwidth available. We have a certain amount of instantaneous RF power available, for instance, either 125 mW for a Class 4 handset and 250 mW for a Class 3 handset (determined by regulatory/standards groups), and a certain amount of peak power deliverable from the battery.

We then have a certain amount of battery capacity (600, 800, 1200, 1800 milliamp/hours for example), which determines how much data we can send from the handset before the battery goes flat.

The closer we are to the base station, the higher the peak data rate available to us and the more data we can send. (This means we can use the power available to send data rather than to overcome the 60 or 70 dB of path loss if we are at the cell edge.) Uplink offered traffic loading is therefore a function of user geometry—how close users are to the base station/Node B, which means user proximity to the Node B determines uplink bandwidth. Another way of looking at this is that revenues increase as network density increases.

Link gain products can be justified on a similar basis. By delivering downlink directivity, more downlink bandwidth can be delivered. Billable bandwidth increases in both directions. Sometimes, investing in link gain products, such as sectored or smart antennas, also gives us parallel side benefits in terms of other services. An example is being able to use sectored antennas or smart antennas to provide location and positioning information.

Positioning and Location

We have just described how you can take an omnidirectional antenna (360°coverage from a colinear antenna consisting of a number of dipoles) and produce gain in the horizontal axis. You can also take a directional antenna and decrease the beamwidth. Typically, directional antennas used in cellular/PCS/3G applications will have 145°, 90°, or 60° beamwidths to support three-, four-, or six-sector cell configurations. As beamwidth reduces, gain (forward directivity gain) increases. A 1 percent beamwidth dish antenna can deliver up to 50 dBi of gain. As beamwidth decreases, the Node B/base station is able to discriminate the position of mobiles (angle of arrival *and* direction of travel) with increasing precision.

For example, a Node B with an omnidirectional antenna knows that a mobile is within its cell radius but does not known where the mobile is within the cell. With a three-, four-, or six- sector cell using directional antennas, the Node B knows which sector the mobile is in. As the mobile moves from sector to sector, direction of travel is known. In addition, TDMA and CDMA cellular networks all have synchronous uplinks. In TDMA networks (GSM and US TDMA) the bit delay introduced by the round-trip between base station and handset is calculated and the mobile is time-advanced (the transmit slot is moved closer to the receive slot), so that all mobiles within the serving cell arrive back at the base station in time with each other. GSM, for example, is able to time-advance by 64-bit symbols. A bit symbol period is equivalent to just under 1 km of flight time; therefore, the number of bit periods of timing advance provide an indication of how far the mobile is away from the base station.

The same time synchronization process is used in IMT2000DS and CDMA2000. Synchronization is achieved by locking the mobile to a short code burst (actually a continuous stream of short code bursts) from the Node B. A chip symbol duration (0.26 μs) is equivalent to 70 meters of flight time, which means potentially very accurate distance information is available (for free) from the air interface. This is the basis for network-assisted positioning and location services, which combine angle of arrival information (from sectored antennas) with distance information.

A mobile can be seen by (and, in IMT2000 and CDMA2000, is supported by) more than one base station/Node B giving additional positional information. Similarly, the handset can see more than one base station. Because the positions (longitude and latitude) of the base stations/Node Bs are known, then either a Node B or handset can work out the handset's position.

Value or quality in positioning/location depends on the accuracy of the fix, reliability (how often the signal is usable), the consistency of the accuracy of the fix, and how long it takes to make the fix (delay and delay variability). These in turn have an impact on the power budget of the handset.

Table 13.2 compares the main options for network-based and handset-based location and positioning systems. Cell ID is the simplest, but accuracy is variable, because it is dependent on network density, as follows:

- Time difference of arrival (TDOA) uses time advance information from three base stations *and* signal strength to give an accuracy of between 300 and 1100 meters. This involves a network upgrade but no handset upgrade.
- TDOA/AOA adds angle of arrival using (smart) phased array antennas to improve accuracy—between 40 and 400 meters, but at the cost of needing to install more complex antenna arrays.
- E-OTD (Enhanced Observed Time Difference) uses the handset to collect information about the time of arrival of signals from the base station and at a number of prespecified location measurement points. This is potentially more accurate than TDOA/AOA and more accurate than TDOA but needs a handset upgrade.

There are then three existing handset-based/satellite-based options: GPS (the U.S. Global Positioning System using 24 satellites)—case studied briefly in Chapter 15), GLONASS (Global Navigation Satellite System, the Russian equivalent), and, possibly longer term, Galileo (the European equivalent to GPS). The advantage of a satellite fix is that, provided at least four satellites can be acquired, satellite fixing gives you longitude, latitude, and altitude.

Hybrid schemes also exist using GPS and network information. This is because GPS works very well in applications where the handset has a clear line of site to three, preferably four, satellites—for example, in rural areas without a lot of nearby high buildings. In an urban environment, satellite line of site is often blocked by buildings, but generally network density is high, so cell ID or sector ID can be used. In addition, the network knows, more or less, where the mobile is (geographic location) and the time, so it can tell the handset which GPS satellites are overhead—that is, which PN sequences to run. This significantly reduces time to fix.

Each GPS satellite uses a 1.023 Mcps PN code onto which is modulated a 50 bps navigation message consisting of the time (repeated every 6 seconds), ephemeris (where the satellite is in orbit, repeated every 30 seconds), and an almanac (where all the satellites are, repeated every 12.5 minutes). Assisted GPS saves the handset from having to store or act on this almanac information. Acquisition times of 100 ms are claimed to be achievable with A-GPS at a power budget of 200 mW per fix (www.globallocate.com).

Table 13.2 Positioning/Location

NETWORK BASED					SATELLITE BASED			HYBRID
OPTIONS	**CELL ID**	**TDOA***	**TDOA / AOA†**	**E-OTD‡**	**GPS**	**GLONASS**	**GALILEO (2008)**	**ASSISTED GPS**
Method		Signal strengths and time synchronization at 3 base stations.	As TDOA but with phased array antennas to provide more accuracy.	Relative time of arrival of BS signals received at handset and at location measurement points.	24 U.S. satellites	21 Russian satellites	30 European satellites	
Merits/ Demerits		Network upgrade but no handset upgrade.	More accurate than TDOA but more complex.	More accurate than TDOA but needs handset upgrade.			Better in Europe?	
Vendors	CellPoint	Cell Loc True Position	SigmaOne / Plextek	Cambridge Positioning Systems	Sirf, Parthus, Trimble, Conexant	Sirf, Parthus, Trimble, Conexant	Sirf, Parthus, Trimble, Conexant	Snap Track (Qualcomm)
Operators	BT, Vodafone			AT&T/ Voice Stream				NTT DoCoMo/ Sprint
Accuracy	Depends on network density	300-1100 meters	40-400 meters E-911	40-400 meters E-911	5-100 meters E-911	100 meters E-911	5-100 meters E-911	10-20 meters

*TDOA = Time Difference of Arrival

† AOA = Angle of Arrival

‡ E-OTD = Enhanced Observed Time Difference

As with all positioning systems, distance is value—in this case, the smaller the distance between the actual position of the user and the calculated position, the higher the value. In GPS, each nanosecond of error represents one foot of measurement error. Certain effects limit the ultimate accuracy of GPS. The density of the atmosphere changes over time. This, and changing gravitational effects, influence the speed at which radio waves travel, so it is impossible to realize absolute accuracy. However, differential GPS schemes, in which signals are calibrated against known locations, can give accuracy down to fractions of a centimeter—sufficiently accurate to detect an earthquake tremor or detect problems with large structures (bridges, skyscrapers, and dams for example).

The practical problem of implementing GPS in a handset is the interference caused by the phone to the GPS receiver. Either the phone needs to stop functioning while measurements are made, which places a premium on acquisition time, or considerable care has to be taken with handset antenna configuration and layout.

In Chapter 10, on handset hardware, we also mentioned the logic of adding a digital compass and infrared distance measurement to the handset, to provide the capability to identify what the handset (and hopefully the handset user) is looking at. The handset can then do a download from the geocoded database in which Web-based information can be searched by geographic location, longitude, latitude, and height. This means the handset displays information on the object being pointed at.

In the United States, the E-911 directive requires network operators to be able to provide positioning information to public safety authorities—emergency rescue services for example. This places an additional premium on positioning and location capability.

Smart Antennas and Positioning

Smart antennas have to be justified primarily on the basis of their link gain budgets and better selectivity and sensitivity. However, they also deliver directivity, and this can be used to acquire positioning information; the more directivity available, the better (more accurate) the positioning. Discussions continue as to whether positioning provided from satellites or positioning provided from the terrestrial network is preferable operationally and economically.

At time of writing, Galileo has just received confirmation of European Union funding. The 30 satellites have an orbit optimized for European coverage (relatively high in the sky in terms of elevation) to provide a good viewing angle (shortest distance through the atmosphere) for users in urban or rural locations. Location accuracy is claimed to be within one meter.

Superconductor Devices

Antennas provide link gain by delivering directivity. A highly directional antenna also delivers selectivity by reducing the amount of visible noise seen by the antenna, which means a high-gain antenna improves radio bandwidth quality. The cost is a larger (more expensive) or more complex antenna (or both), which takes up more physical space on the mast, and in the case of smart antennas, absorbs more processor bandwidth.

Filters can also be regarded as link gain products. By filtering out unwanted signal energy, filters improve the quality of the wanted signal energy. Filters are used on the transmit stages of a base station and handset to prevent interference to other adjacent users, and on the receive side to improve receive sensitivity by delivering selectivity (ratio of wanted to unwanted signal energy).

On the receive side, we are also having to capture and process a very low level signal (a few picoWatts), so we need to find ways of optimizing the process of small signal amplification—that is, optimizing the design and performance of the low-noise amplifier (LNA).

In the following sections, we review how "signal steering" technologies and LNA technologies can together deliver link gain an improvement in radio bandwidth quality.

Filter Basics

First, a reminder about basic filter characteristics. Figure 13.9 shows an ideal lowpass filter. Below cutoff frequency, ω_C, inputs are passed with no attenuation. Above ω_C, inputs are infinitely attenuated. The transition from passband to stopband is infinitely narrow.

An ideal lowpass filter is impossible to obtain. In practice, unwanted products will occur a distance from the filter passband (harmonics, etc.), and so an infinitely narrow transition band is not required. A more practical specification will have a finite width transition band with tolerances and achievable limits.

The Q factor

We have come across Q before when discussing image quality. Q is also used to describe resonance quality.

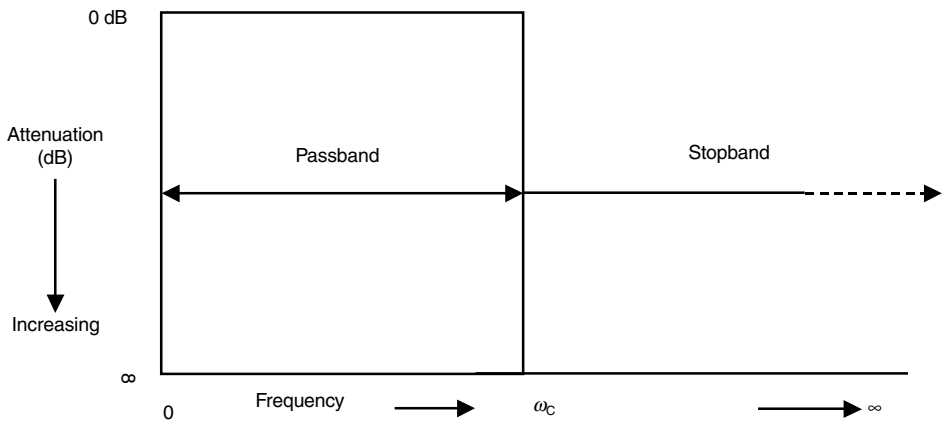


Figure 13.9 Ideal lowpass filter.

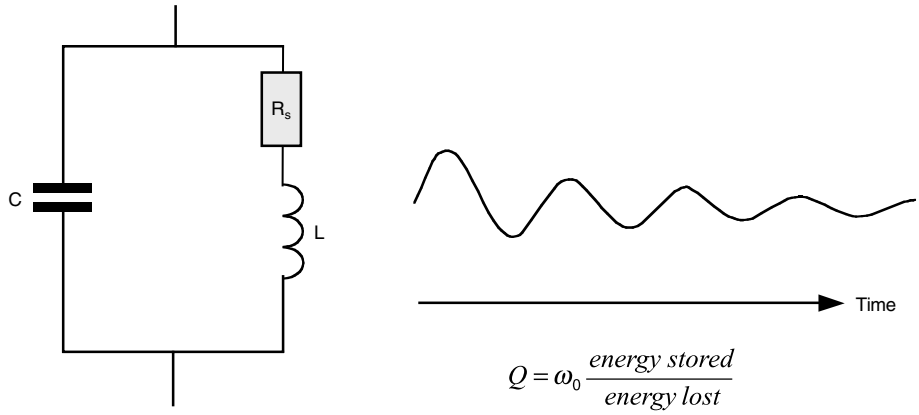


Figure 13.10 The Q of a circuit—loaded Q.

Filters are formed by the cascading of a number of resonant elements. *Resonant elements* are formed by the interaction of inductance, capacitance, and resistance. The figure of merit or quality of the resonant circuit (that is, its sharpness of resonance) is indicated by the Q factor. The higher the ratio of the reactance at resonance to the series resistance, the higher the Q and the sharper the resonance. In a parallel circuit Q = X_L/r_S. In practice a capacitor is less lossy (more ideal) than the inductor, thus it is the resistive component of the inductor that determines the Q.

Another way to envisage Q is a consideration of the energy in a resonant circuit. If the L and C components were ideal, then the energy would circulate ad infinitum.

In practice, energy is lost in the series resistance, and so the energy dies away. Taking energy from the circuit will also produce the same effect. This is referred to as *damping* or *loading* the circuit. The Q of a circuit fed from and loaded by a finite impedance is referred to as the *loaded Q* (see Figure 13.10). Filters may be considered as a number of cascaded resonant sections. The consideration of losses, loading, and Q apply equally to multiple sections as a single section.

The Q of a filter indicates its response performance. It can be expressed as:

$$Q = \frac{f_o}{BW_{3dB}}$$

$$Where\ f_o = centre\ frequency$$

Figures 13.11 and 13.12 show typical practical filter characteristics in terms of Q. Q is important when we come to the RF plumbing needed in a Node B transceiver. Some Node Bs are single RF carrier, but often the transceiver will have multiple 5 MHz RF channels (for IMT2000) and multiple GSM 200 kHz channels. These can be kept apart from each other either at baseband or at RF frequency. Although it is theoretically attractive to do all our selectivity at baseband, in practice we have to provide some additional RF selectivity to meet transceiver sensitivity targets.

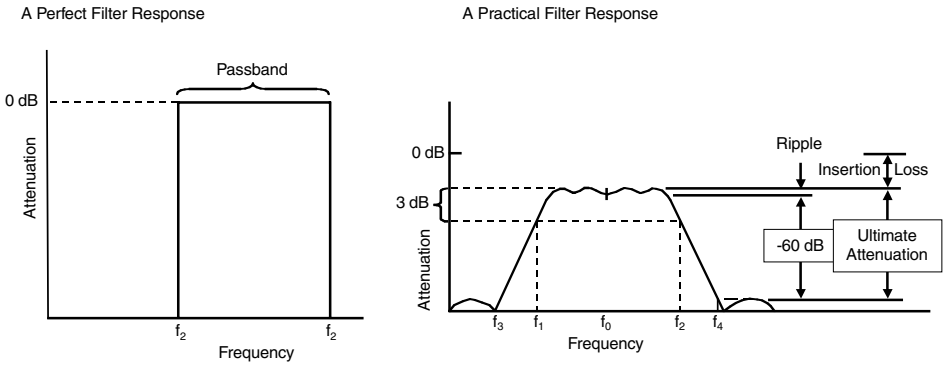


Figure 13.11 Practical filter parameters.

The following sections cover examples of RF system components used particularly in multiuser sites with shared antennas.

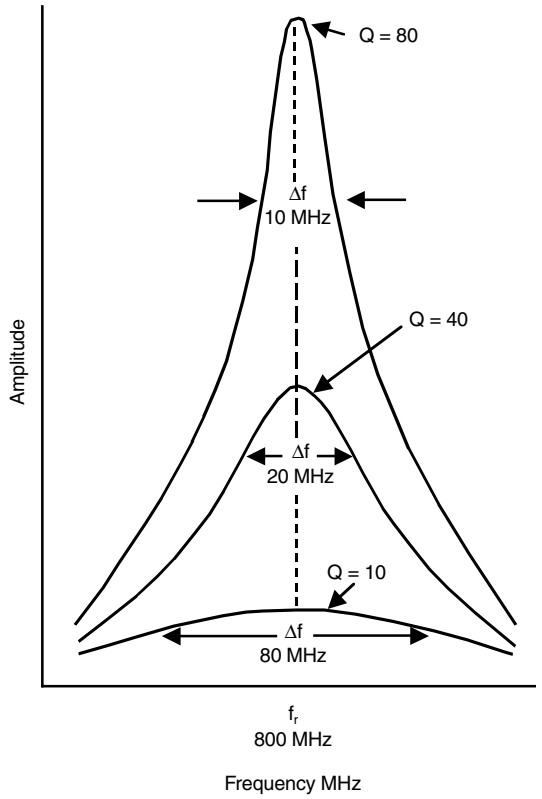


Figure 13.12 Q versus bandwidth.

The Cavity Resonator

The cavity resonator is a bandpass filter circuit with a center resonance frequency related to its physical dimensions. The unloaded Q depends on its physical volume and at VHF can range between 2000 and 10,000. The loaded Q (that is, the damped unloaded Q, produced by the effect of the terminating impedances) is usually designed to be between 500 and 1000.

A system using a cavity resonator gives protection to co-sited receivers by reducing the radiation of wideband noise and spurious products. Cavity resonators may also be connected together to provide additional isolation when multiple transmitters are combined to a single antenna. When used in this application, the cavity resonator is used together with ferrite isolators to ensure unidirectional power flow.

The Cavity Resonator in Multicoupling Applications

If a number of users are multicoupled together on the same base station antenna, several cavity resonators are coupled together with a precisely dimensioned cable harness. The system (multicoupling to a single antenna) uses several cavity resonators coupled together with a precisely dimensioned cable harness.

Insertion losses are typically 2 dB with a relative frequency separation of 1 percent. With very high performance cavities, the separation can be reduced to 0.25 percent with isolations of 20 dB. When used together with a ferrite isolator, isolations of 50 dB are obtainable between adjacent transmitters.

Circulators and Isolators

The ferrite circulator is a practical component used to provide directional isolation at the output of the transmitter. The isolation achieved has to be considered in conjunction with the insertion loss and bandwidth, with the quality of the terminations directly affecting the isolation performance. For very high isolation (>40 dB), a dual circulator version may be fitted.

Circulators are electromagnetic components having three or more connections (ports) in which RF energy circulates in one direction from one port to another, whereas a relatively high attenuation occurs in the opposite direction. Additionally all ports are matched.

The (low) attenuation in the circulation direction is called the *insertion loss* and is in the order of tenths of a dB. The (high) attenuation in the opposite direction is called *isolation* and is usually in excess of 20 dB. Other parameters of importance are VSWR at the ports and the bandwidth of the circulator. The bandwidth is usually limited by construction resonances but is always sufficient for mobile wireless applications. Isolators have the following characteristics:

- They have only two ports.
- The insertion loss is very low in the forward direction.
- The isolation (attenuation) is very high in the reverse direction.

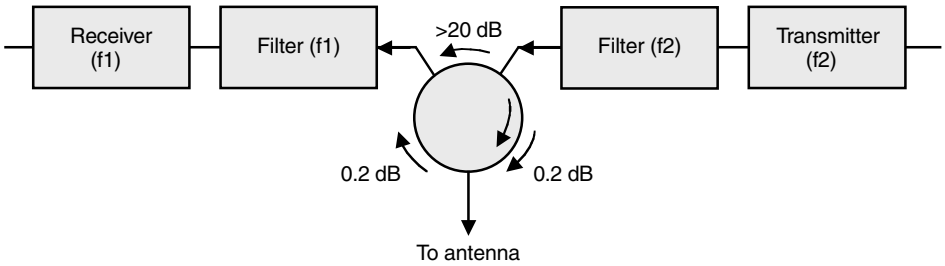


Figure 13.13 Circulators and isolators—filtering to increase isolation.

A magnetically biased ferrite core absorbs the electromagnetic waves in one direction but does not influence them in the other direction. It is the most important part of an isolator or circulator. Let's consider a couple of examples.

Example 1

A circulator can be used as a duplexer, where a receiver and transmitter are connected to a common antenna. The circulator decouples the receiver from the transmitter to the antenna and from the antenna to the receiver. Normally, the transmitter and receiver operate at different frequencies within the bandwidth of the circulator so that additional filters can be used to increase the isolation (see Figure 13.13).

Example 2

A terminated circulator may also be used to decouple two or more transmitters connected to the same antenna. The object is to reduce intermodulation distortion by preventing output energy from one transmitter passing to the second transmitter and creating intermodulation products. The terminating resistor need only be dimensioned for the reflected power. The transmitters can either operate at different frequencies within the bandwidth of the circulator or on the same frequency (where power doubling is required).

Hybrid Directional Couplers

Directional couplers are used to combine or split power from or to transmitters and their loads (see Figure 13.14). These may also be referred to as hybrid *combiners*, *splitters*, or *diplexers*. Applications range from printed circuit functions realized in micro strip, to combining 0.5-MWatt transmitters.

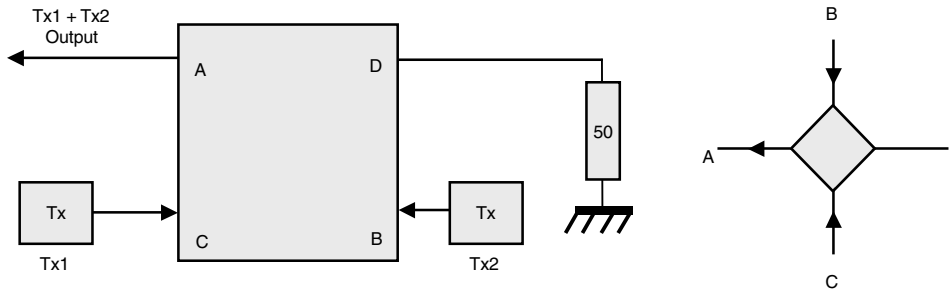


Figure 13.14 Power combining.

In a hybrid network, power fed into any port is split equally between the two adjacent ports, and provided that the loads are perfectly matched, no power reaches the opposite port. Consideration must also be given to the phase difference between the input and the two output voltages.

This gives two classes of hybrid:

Quadrature types. These have two output voltages, differing in phase by 90° and two planes of symmetry.

Sum-and-difference types. These have two output voltages either in phase (0°) or 180° out of phase with each other, depending on which port is used as the input. They have a symmetry of only one plane.

If the paths between ports are labeled according to phase change in proceeding from one port to the next, the properties may be seen. Some types have ports that are balanced, or not directly connected to ground, so voltage phases are ambiguous, and alternative (or relative) phases are given.

If power from a transmitter is split between two loads (for example, antennas) by a simple transformer and T junction, there is a high probability that failure of one load will result in a large change in the amount of power arriving at the remaining load.

A hybrid with a balancing load used to split the power overcomes this problem (as shown in Figure 13.15). Equal amplitude forward waves, with phases appropriate to the hybrid, always reach each termination, and any reflected power is split equally between the transmitter and the load on port D. Any change in voltage applied to one load, brought about by a change in the other, results from a mismatch of the load D or the transmitter. The worst VSWR that can be presented to a transmitter by a single load fault is 50 percent (VSWR 3:1).

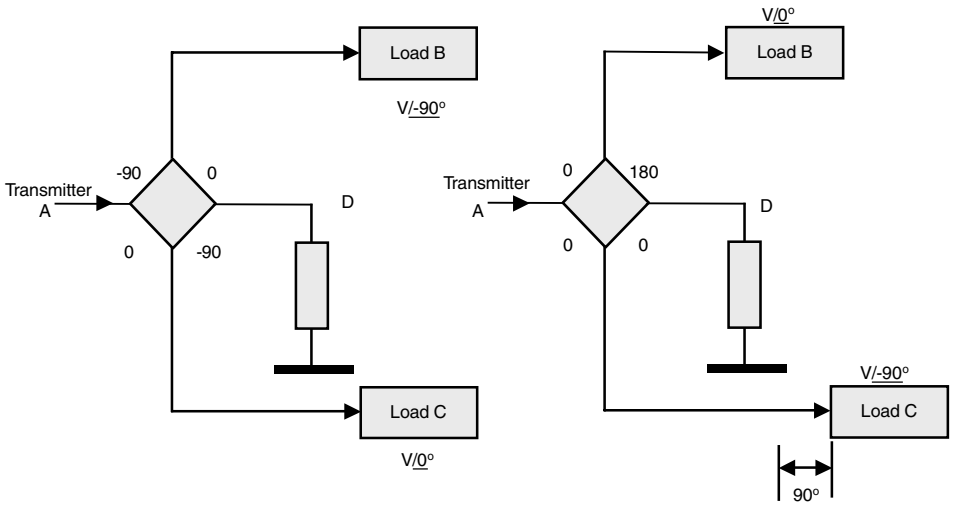


Figure 13.15 Hybrid splitters.

Equal loads, arranged to be fed with equal quadrature currents, will receive their correct relative currents regardless of mismatch. A matched load is always presented to the transmitter, and all reflected power is transferred to load D. By substituting sources for load ports, the power of multiple transmitters may be combined, as shown in Figure 13.16, to one load (for example, antenna). If a transmitter fails, the remaining transmitters will still be correctly terminated and deliver their power to the load.

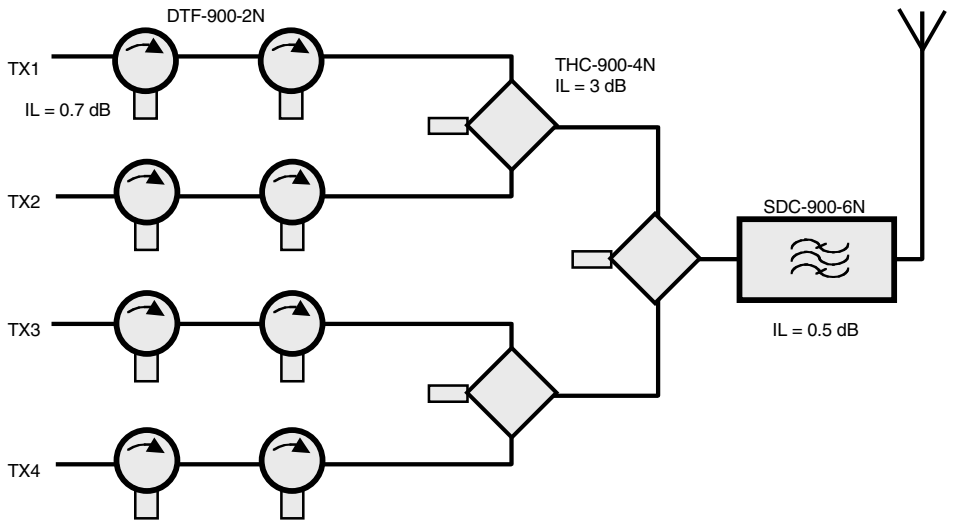


Figure 13.16 Hybrid combiner system.

Multichannel Combining

The devices just described allow us to multiplex different operators using different RF channels onto shared RF hardware (the antenna). In addition, we need to separate RF channels within the base station transceiver. If a single channel is placed through a power amplifier, the amount of linearity required depends on the modulation used. For example, constant envelope GSM can tolerate a nonlinearity of >-20 dBc. Non-constant envelope modulation typically needs a linearity of <-45 dBc.

As sufficient linearity over multichannel system bandwidths have been difficult (read: technical and financial) to achieve, a separate PA has been used for each channel and combining is performed after amplification. Typically two choices of combining are considered—wideband hybrid combining and narrowband cavity filter combining.

The advantage of a wideband combiner is that it needs no retuning for channel additions. The disadvantage is that it has high loss. The disadvantage of a narrowband cavity filter is that it requires retuning with changes in band planning. The advantage is that it has relatively low loss.

If sufficient linearity is available, such that intermodulation products (even at base station powers) are dramatically low (<-75 dBc), then combining may take place at low power, as shown in Figure 13.17 before the multicarrier amplifier. Given sufficient linearity, even multistandard combining may be used—for example, CDMA and AMPS or PCS1900, GSM and CDMA1900 in geographic proximity.

Cost and size constraints are making vendors move relatively quickly to linear amplifier solutions and baseband selectivity. The counter argument is to find a way of improving the Q of the RF filters used and relaxing baseband processor overhead (for example, the DSP overhead involved in PA predistortion and linearization). This is one of the rationales for superconductivity filters.

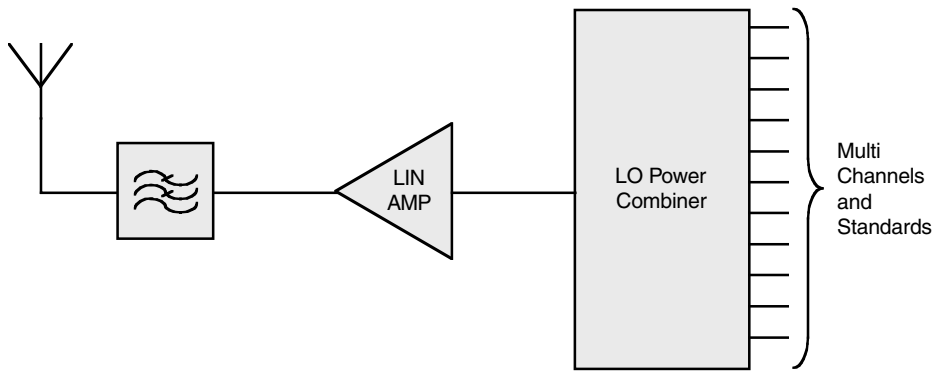


Figure 13.17 The multicarrier/standard amplifier.

Superconductor Filters and LNAs

Superconductivity filters are based on the principle that resistance to current flow reduces with temperature. Superconductivity filters are made from either thin or thick film deposition processing that promises very good conductivity when cooled to temperatures of 90 K or below. Because there is less resistance, the mechanical form factor of the filter can be made smaller for a given selectivity; though you do then need to add a cooler to make the device work.

An example from SuperConductor Technology Inc (www.suptech.com) is a high-temperature superconducting filter and LNA system (dimensions are W: 8.4 inches, L: 23 inches, H: 7 inches) deployed in cellular base stations between the antenna and receiver. Using only 50 W of power, the Stirling cryogenic cooler maintains the filter and LNA at a steady 77 K. Together with the sharpest filters available, the product has a noise figure of < 0.5 dB, as compared to conventional solutions with 2.5 dB.

When used to achieve selectivity, it is claimed that a 16-pole filter can be implemented for 3G applications giving an additional 2.5 MHz of selectivity (-60 dBc) over and above a conventional 11-pole filter, with a 1-dB reduction in the noise floor.

These filters are used mainly in receive applications, since high powers cause heating and hence loss of performance. Although given the low insertion loss, they can handle several Watts. Even on the receive side, a base station can easily be looking at -10 dB/0 dB of received signal energy.

Thick film superconductors are better at handling this incident energy and can deliver a Q of 50,000, giving a close-in, deep rejection figure of better than 100 dB within 3 MHz. However, they are more expensive than thin film superconductors.

Superconductor filters and superconductor LNAs have not as yet been widely deployed in either 2G or 3G wireless networks. Partly this is due to network operators' concerns about maintenance (windshield time), and partly it is because conventional filter technologies continue to improve.

Figures 13.18 and 13.19 show a conventional ceramic filter and preamplifier module. It is a good illustration of how intermod performance can be traded against noise performance. This is a tower-mounted conventional LNA product.

For tower-mounted amplifiers (receive-side amplifiers), network operators generally prefer to use relatively simple devices. The idea of having to climb a mast to service a supercooled LNA is not particularly attractive.

RF over Fiber: Optical Transport

So far in this chapter we have looked at how the characteristics of copper determine the characteristics and quality of the radio signal. We can also take an RF signal and convert it to an optical signal using a linear laser. We need the linearity to preserve the

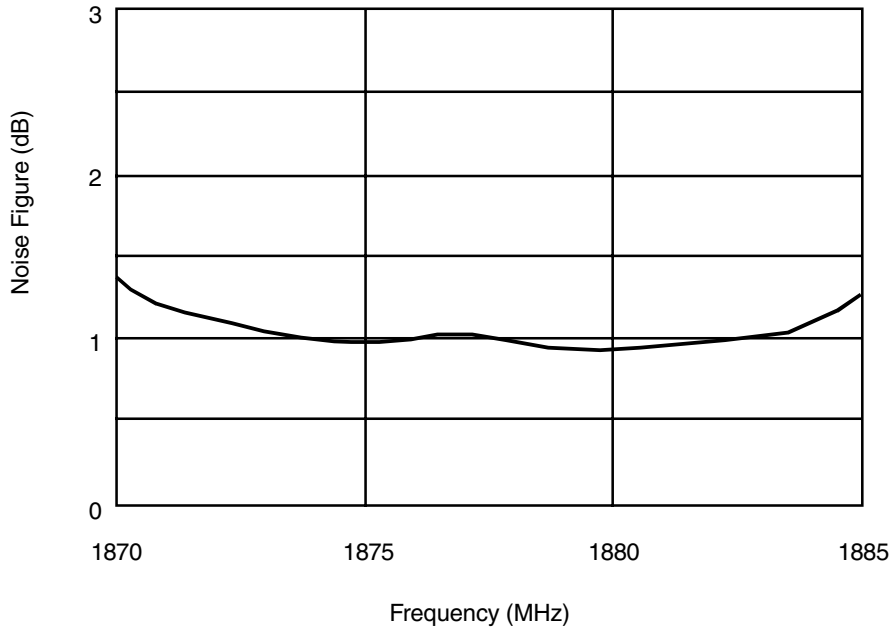


Figure 13.18 Noise figure for PCS tower top LNA (Filtronic Comtek—www.filtct.com).

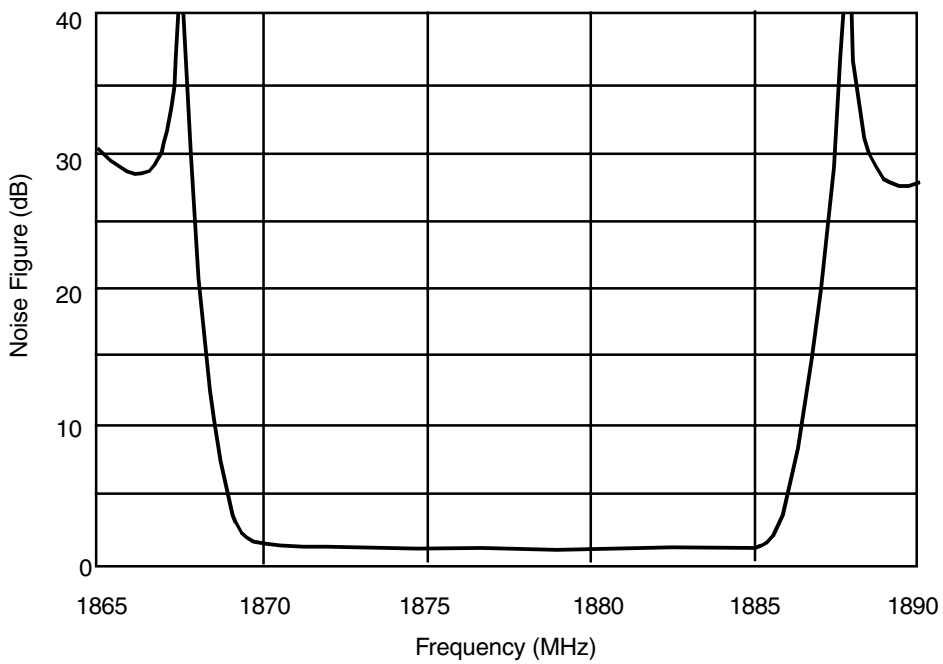


Figure 13.19 Selectivity of ultra low noise tower top module.

phase and amplitude components of our RF signal. The reason we might choose to convert our RF signal to an optical signal is that the loss per kilometer is much lower in an optical fiber when compared to a copper cable.

Typically, the loss in an optical fiber cable is less than 1 dB per kilometer. The loss over a 7/8 copper feeder (a copper waveguide inside a plastic outer sheath) is between 1 and 3 dB per 100 meters, and the loss increases with frequency. Putting an RF signal over optical fiber is not unique to the cellular radio industry but is also widely used in cable TV.

In cellular industry applications, the RF signal at the base station is converted to an optical signal and then sent down the fiber. The signal is delivered to a fiber-optic antenna (which looks rather like a simple detector). A photodiode then converts the optical signal back to RF. The RF signal is then amplified and fed to the antenna. On the uplink, the signal is received at the antenna and then amplified before the laser diode. The laser diode converts the signal to the optical domain for transmission to the laser and is then combined at the base station/hub with other fiber optic signals.

Because the optical link is linear, it effectively becomes transparent to the network; the handset is sending an over-the-air signal, which just happens to be channeled through an optical fiber feed for part of its journey. RF over fiber can be used to solve tricky installation problems. An option for in-building coverage is to install a number of pico base stations.

Often, however, the architect wants to minimize visibility of all external hardware, that is, base stations and antennas. An application example is the Bluewater shopping center in Kent, United Kingdom. This is a small shopping center by U.S. standards but an enormous shopping center by U.K. standards. There are five mobile phone shops in the complex, so coverage has to be good. The problem is that the mall is on two levels, so it is difficult to achieve consistent coverage. The solution was to install 60 GSM transceivers (400 simultaneous phone calls) and to use 64 core and 8 core fiber to connect the base station to small distributed antennas.

Optical Transport in the Core Network

RF over fiber has been made possible by an enabling technology—linear lasers. In the core network, optical transport bandwidth quantity and quality is improving as new enabling technologies became available, particularly linear optical amplifiers, optical filters, and in the longer term, optical memory. In parallel, as we highlighted earlier, optical DSPs have been proposed that support baseband processing at optical speeds. As lasers become more accurate (the ability of a laser to produce a discrete frequency), channel spacing can reduce.

A narrowband channel in the optical domain is 25 GHz. As our ability to synthesize optical frequencies increases, and as the frequency accuracy and stability of new optical devices increase, capacity increases.

Wavelength division multiplexers and multiple-carrier generation frequency management techniques are effectively delivering an order of magnitude increase in optical capacity every decade. Splitting light (for example, in a prism) is potentially easier than carrying a packet in an electronic buffer delivering some interesting multiplexing opportunities.

When we define optical bandwidth quality, many of the performance metrics are not dissimilar to RF bandwidth quality metrics. Dispersion loss over distance increases as optical frequency increases. Higher-frequency, smaller-wavelength optical channels will therefore have less quality than lower-frequency, longer-wavelength channels. Quality of service can also be defined as how physically secure the fiber is—how difficult is it to tap the optical bit stream.

At present, we use electronic switching at the (optical) network edge. In the longer term, we might justify moving optical switching closer to the user/consumer using Multi-Protocol Label Lambda Switching (MPL λS) to provide differentiated quality of service.

Wavelength-division multiplexing provides the ability to configure optical bandwidth to respond to different QoS requirements at an aggregated traffic level and to provide multiple optical routing trajectories for resilience and redundancy, giving good restoration capabilities. At present, the only software-reconfigurable network element is the Optical Layer Cross Connect (OLXC), but in the longer term, we will have tunable lasers and receivers supporting reconfigurable optical add/drop multiplexers.

Multiplexing options include electrical or optical time-division multiplexing to combine input channels into a single wavelength or adaptation grouping in which groups of wavelengths are added or dropped as a group (depending on laser tunability and optical channel spacing).

As bit rates increase (2.5 to 10 to 40 Gbps), power has to increase. Optical networks are power-limited rather than bandwidth-limited, just in the same way that radio networks are power-limited rather than bandwidth-limited. In an optical network, higher power creates problems with impairments and nonlinearities.

Linear impairments are independent of signal power and affect wavelengths individually. Typical impairments include polarization dispersion, chromatic dispersion, and amplifier spontaneous emission. Nonlinear impairments increase as power increases and generate dispersion and cross talk.

As the number of wavelengths increases, the blocking probability of higher priority traffic classes increases. We need to start considering using offset time-based access/priority control to deliver differentiated quality of service. This, however, depends on our ability to provide fast switches. It is hard to get a mechanically tuned grating to switch at less than a millisecond, so this becomes a constraint. (There is no point in having bandwidth available if you cannot provide access to the bandwidth.)

If we can increase the number of optical frequencies (that is, reduce channel spacing), we can reduce the bit rate per optical stream and make switching at either end of the pipe slightly easier. In radio network terms, this is rather like discovering the benefits of using narrowband RF channels. It is also analogous to the way we used OFDM in wireless LANs and digital TV to subdivide the frequency spectrum and slow the bit rate (to improve intersymbol interference, which in turn is analogous to chromatic or polarization dispersion).

Arrayed waveguide gratings are becoming available that can discriminate between 40×100 GHz optical channels and, in the longer term, 80×50 GHz or 160×25 GHz optical channels. If we can switch optically, we can reduce power consumption by about 75 percent compared to electrical switching and deliver a size footprint reduction of 75 percent (glass and air replacing silicon and copper). However, optical switching needs a new generation of enabling technologies.

Presently the options include the following:

- Micro-electromechanical devices (MEMS)
- Liquid crystal devices
- Electro- or thermo-optic devices
- Bubble switching (inkjet technology)

We can, for example, use MEMS to build lots of tiny microcells on a silicon chip. Tilting the mirrors routes the optical data streams between, potentially, several thousand input and output fibers. The trouble is MEMS don't work fast enough for packet switching; we can only use them to reroute around a failed fiber path—that is, for restoration, reconfiguration, or protection (to drop the loading from a compromised light path for example). If we compare optical switching with the existing optical/electrical/optical switching we have today, we can say that an optical switch is fast but stupid, and an optical/electrical/optical switch is smart but slow.

Figures 13.20 and 13.21 show superconductors as a potential halfway house. The example is a 10 Gbps switch proposed by Conductus taking in an optical signal (refer to the right side of Figure 13.20, processing it through a photodetector, performing switching, and then amplifying prior to reconverting to the optical domain via a laser diode. The photo detector, switch, driver, and GaAs pre-amplifier are all supercooled. This adds an intermediate layer between the optical layer and electrical switch layer, that is, a superconductor routing layer (see Figure 13.21).

Alternatively, we might try and do everything in the optical domain, but if we wanted to use IP packet routing, we need the ability to buffer—to give us time to read routing instructions and to smooth bursty traffic—and, at present, we do not have optical RAM. If we are trying to multiplex lots of narrowband optical channels into a single fiber (to relax routing/switching performance), we also begin to lose power in the combining process. Broadband coupling is very lossy (typically 4 dB for a 2-channel multiplex and 13 dB for a 16-channel multiplex), and narrowband filters are bulky and expensive. Demultiplexers are also very hard to design and suffer from high insertion loss and poor sensitivity.

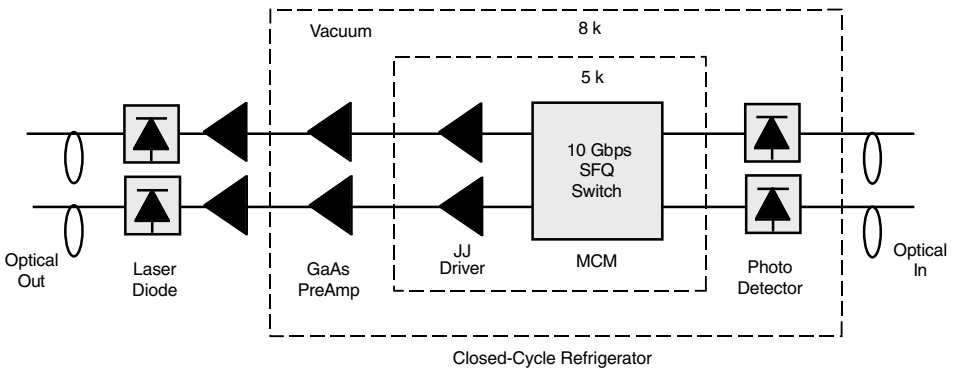


Figure 13.20 Conductus HTS—10 Gbps switch components.

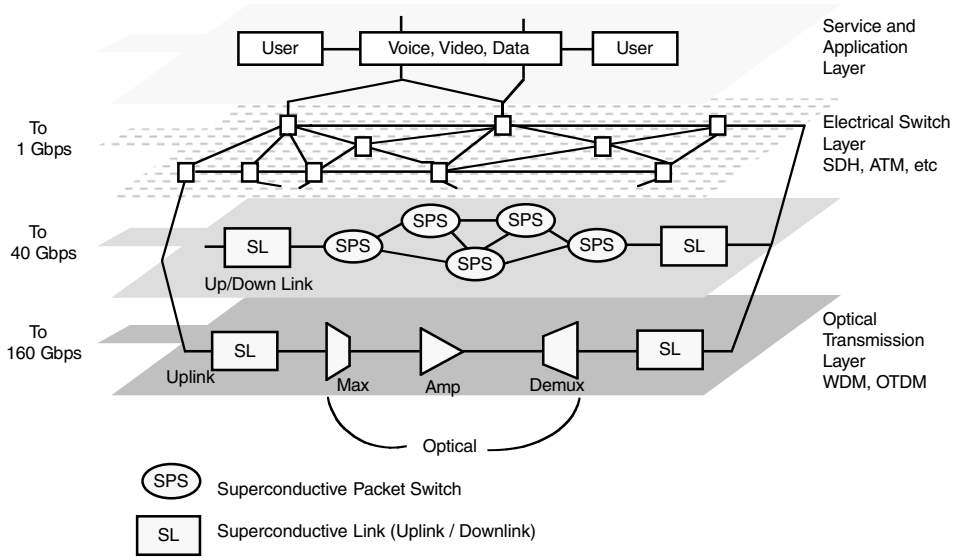


Figure 13.21 Superconducting electrical switch layer.

Optical Selectivity

Just as in the RF domain, we can produce wavelength-selective devices that present a low loss for some wavelengths and a high loss for others by introducing a phase shift in a portion of the light energy. The options are as follows:

Thin film dielectric coatings. These are substrates with alternative layers of high and low refractive index dielectric materials. The transmitted wavelengths are a function of the thickness of the cavity coating. These devices are a very expensive way of providing multiple channel selectivity.

Fiber Bragg gratings. These are devices in which the refraction index of the fiber core can be changed in such a way that a narrow wavelength of light will be reflected and all other wavelengths will be transmitted. The bandwidth is inversely proportional to the length of the grating.

Optical filtering diffraction gratings. Optical filtering diffraction gratings reflect light at an angle proportional to the wavelength, creating constructive and destructive interference. Depending on the wavelength of the incident light, there is an angle for which the individual light waves will be exactly one wavelength out of phase, which means they will add constructively. Because insertion loss is independent of the number of channels, optical filtering diffraction gratings can support lots of channels.

Optical filtering integrated optical devices. These devices are the optical equivalent of an integrated electronic circuit. Cores surrounded by cladding are layered onto a silicon substrate. The cores act as optical waveguides. These devices are good for optical cross connects and demultiplexers.

Optical Transport Performance

Optical transport can be divided into ultra long haul, long haul, and metropolitan.

- Ultra long haul is up to 6000 kilometers, typically using a 40 channel × 10 Gbps multiplex.
- Long haul is up to 600 km, typically 80 channels × 10 Gbps.
- Metropolitan is between 6 and 60 km and may use dense wavelength division multiplexers (40 × 40 Gbps channels).
- Long haul and ultra long haul requires repeaters every 80 kilometers or so. These are usually erbium-doped fiber amplifiers.

At either end of the optical pipe, photodiode detectors exploit the properties of iridium phosphide and iridium gallium phosphide. Iridium phosphide is transparent to photons. Iridium gallium phosphide (conveniently) absorbs photons. We can therefore recover the bit stream from the optical domain, but we still have the problem of how to process packets at optical transport speeds. A 2.5 Gbps packet stream gives us 65 ns to process a packet. A 40 Gbps packet stream gives us 4 ns.

Also in the optical domain, as bit rates increase (and bit duration decreases), the shorter bits become more vulnerable to chromatic and polarization mode dispersion—either in the fiber or in the amplifiers or filters. Iridium phosphide is used for devices because it can deliver nanosecond response times.

In the optical frequency domain, we need to find mechanisms for improving laser wavelength accuracy—for example, the use of phase lock loops to measure the output from the laser and correct for temperature or current drift effects in order to improve the control of optical frequency and phase. This may also include active trimming to correct for device age drift. DCXO age drift characteristics can be determined over time and correction characteristics applied (potentially over many years).

Wavelength Division and Dense Wavelength-Division Multiplexing

Wavelength-division multiplexing (WDM) was introduced in 1994. WDM is defined as using channel spacing of 3 nanometers (375 GHz). Dense wavelength-division multiplexing (DWDM) is defined as 1 nm or less (<125 GHz). The ITU specifies 100 and 50 GHz channel bandwidths and will likely specify 25 GHz spacing as device technologies mature. The two principal allocations are as follows:

- **C band.** 1530 to 1570 nm
- **L band.** 1570 to 1610 nm

This translates as 80 nanometers, 200 × 0.4 nm (50 GHz) channels, and 10,000 GHz.

Table 13.3 defines the ETSI (European Telecommunications Standardization Institute) SDH and ANSI SONET bit rates (SDH is Synchronous Digital Hierarchy; SONET is Synchronous Optical Network). Table 13.4 defines the SONET and SDH Synchronous Transfer Mode (STM) optical carrier (OC) bit rates for dense wavelength-division multiplexers.

Table 13.3 ETSI SDH and ANSI SONET Bit Rates

ETSI SDH	ANSI SONET	
STM 1	OC 3	155 Mbps
STM 4	OC 12	622 Mbps
STM 8	OC 24	1.25 Gbps
STM 16	OC 48	2.5 Gbps
STM 32	OC 96	5 Gbps
STM 64	OC 192	10 Gbps

The chromatic dispersion of four wavelengths at 2.5 Gbps is 16 times less than a single wavelength transmitting at 10 Gbps and, therefore, provides a more robust channel (needs less regenerators). Rather like OFDM, it is better to have a larger number of slower bit rate optical channels.

However, these are synchronous channels and the offered traffic is increasingly asynchronous. Physical layer (Layer 1) traffic patterns are changing in the order of nanoseconds or microseconds, whereas the optical cross-connects (the only existing mechanisms we have to switch in additional optical channels or switch out channels) take a millisecond to switch.

Buffer design either side of the optical domain is therefore very critical, particularly in long-haul fiber.

Table 13.4 DWDM Multiplex

	NO. OF CHANNELS	BANDWIDTH
SONET OC48 / SDH STM 16 (2.5 Gbps)	4	10 Gbps
	8	20 Gbps
	16	40 Gbps
	24	60 Gbps
	32	80 Gbps
OC192/STM 64 (10 Gbits)	32	320 Gbps
	64	640 Gbps
	80	800 Gbps
	128	1.28 Tbps
	160	1.6 Tbps

Note that transmission quality is inversely proportional to the square of the bit rate due to chromatic dispersion.

Summary

This chapter reviewed some radio and optical hardware optimization opportunities and their impact on radio and network bandwidth quality. The following sections summarize the main points made.

Antennas

Most present deployment objectives can be met by using conventional passive antennas with a modest amount of reconfigurability—some electrical downtilt, for example. Smart adaptive antenna schemes may become more economically attractive in the future because of the need to selectively null out interference from in-band users.

Many operators will need antennas that can service the 1900/2100 MHz band, the 1800 band, and the 900 band. The half-wave/quarter-wave relationship between 900 and 1800 MHz makes this feasible although rather suboptimum in terms of antenna performance. As offered traffic moves increasingly toward a more complex rich media product mix, then radio link/radio bandwidth quality will become increasingly important.

Present networks are being designed to support a bit error rate of either 1 in 10^3 or 1 in 10^6. Given that wireline bit error rates are typically 1 in 10^{10}, it could be argued that radio bandwidth quality should be equivalent. Reducing the bit error rate from 1 in 10^3 to 1in 10^6 with the same delay parameters requires a 3 dB increase in the radio link budget. Decreasing from 1 in 10^6 to 1 in 10^9 requires an additional 3 dB.

Smart antennas provide one way in which the link budget can be improved. (An alternative, of course, is just to increase the existing network density using conventional antennas.) Flexible antenna configurations—that is, antennas that can adapt themselves as interference moves in the cell—are a logical way to improve capacity and coverage. The cost economics for widespread deployment, however, are as yet unproven.

Superconductor Devices

Filters are an important ingredient in radio bandwidth quality—the Q of the filter has a direct impact on received C/I and transmit energy purity (keeping transmit energy out of other users' transmit and receive bands). Superconductor filters deliver very substantial performance gains in terms of selectivity with relatively little insertion loss. Practical implementation issues (doubts about the mechanical reliability of cooling engines) have prevented their widespread deployment to date.

Superconductors can also be used to produce very low noise amplifiers. However, in practical networks, thermal considerations may limit their application. In parallel, conventional filter design continues to improve. Better mechanical design and improvements in materials, including the use of silver plating and other passive techniques for improving conductivity, deliver year-on-year performance gains.

Network operators are, by nature, quite conservative and are very sensitive to life-cost liabilities. This prudence tends to prevent aggressive adoption of new technologies,

even when those technologies (on paper) offer substantial performance advantage—a protective technology adoption inertia. Superconductors, therefore, rather like smart antennas, are an educational sell. The customer needs to be convinced of the long-term cost and performance advantage of a new technology, technique, or process. Products that are an educational sell suffer from dwell time. Adoption is slow because operators cannot make up their minds on the technology.

Usually, there are also competing flavors involved—adaptive or switched beam antennas, thin or thick film superconductors—and the customer hears competing and conflicting claims from different vendors. If a customer has too many choices, he or she will often make a positive choice: not to choose any of the choices. When benefits remain unproven, the safest choice is not to make a choice.

Optical Components

Optical components provide access to large quantities of optical bandwidth (several hundred Terahertz). However, bit error rates need to be typically 1 in 10^{12} or better. This places severe demands on component performance and effectively means that our optical transport layer is power-limited (just like the copper access and radio access parts of the network).

As our ability to generate and filter discrete optical frequencies improves, we can use increasingly narrowband channels. This allows us to deliver differential quality of service over the optical transport layer and to provide a measure of adaptive bandwidth (by dropping optical channels in and out of the optical frequency multiplex). This depends, however, on being able to have a fast optical cross connect and presently this presents performance limitations.

One merit of the optical layer, shared with copper access, is that it is a consistent transport medium. It does not suffer from the fading effects encountered on the radio physical layer. Impairments tend to be steady-state. This means they increase with distance. Thus, provided sufficient power can be made available (a sufficient number of repeaters), it is reasonable to assume good consistent quality.

Even so, some wavelengths will provide better quality than others—for instance, impairments increase as optical frequency increases, so lower frequencies will generally deliver better quality. This provides the basis for differentiated quality of service using Internet traffic shaping protocols, such as Multiprotocol Label Switching.

One practical problem of packet routing in the core network is the sheer physical speed at which packets have to be read. The answer tends to be to implement parallel processing. As the number of optical wavelength channels increases, it becomes possible to define different packet routing trajectories, which can be maintained both across the copper transport and optical transport layer. However, this process becomes more complex as the offered traffic becomes increasingly asynchronous over time.

An option is to extend ATM across the optical layer, hardware switching on a 10-ms resolution, to manage and maintain the time domain properties of the rich media products as they move across the copper and optical transport layer. We revisit these protocol issues in Chapter 17, which is devoted to traffic shaping protocols.

CHAPTER

14

Offered Traffic

In our last chapter we looked at the impact of asynchronous traffic on network hardware (antennas, filters, and optical transport). In this chapter, we try and define some of the traffic characteristics and traffic properties and how these characteristics and properties exercise system components in our network.

Characterizing Traffic Flow

In earlier chapters, we described how bursty bandwidth can put RF components (RF power amplifiers) into compression. Bursty bandwidth can also put ADCs into compression. Bursty bandwidth can also cause buffer overflow in routers, resulting in packet loss (buffer bandwidth compression). The properties of the traffic on the network therefore directly impact hardware performance in the network.

To work out how well our network will work, now and in the future, we need to characterize the traffic flowing through the network. The focus today tends to be to try and characterize the content being delivered from the network out to the target device. The alternative approach is to characterize content captured by the subscriber, processed in the subscriber's handset, and then sent to the network for onward transmission—the offered traffic mix.

Six industries, are presently converging all of which either produce or influence content. Computer products, consumer electronics products, and IT products all help to generate traffic. This traffic may come from, or go to, a wireline, wireless, or TV network.

The Internet is used increasingly as part of the delivery process. This means that we need to consider the impact of offered traffic on Internet protocols and the impact of Internet protocols on offered traffic.

The Preservation of Traffic Value (Content Value)

Consider that content is produced by and for a wide range of devices. The value of the content has to be preserved as it moves into and through the network. Quality degradation introduced by compression, bit error rate, packet delay, and packet loss compromises value.

Value can be preserved by bandwidth management—either in band (using, for example, IP protocols) or out of band (SS7). In-band signaling must, however, be sufficiently responsive and granular to respond to rapid changes in the statistical nature and transport needs of the offered traffic, which may be harder to achieve than presently expected.

The Challenge for IP Protocols

IP protocols are proposed as a pervasive solution to access management, traffic management, mobility management, and network management. We can use IP protocols to manage access rights to delivery and memory bandwidth, to determine priority access to delivery and server bandwidth, to determine handover between IP nodes, and to commoditize network management. IP SS7 would replicate the circuit switch capabilities (call setup, call maintenance, call clear-down) and add session management capabilities (session setup, session management, session clear-down)—the "IP over everything" to "everything over IP" transition.

Radio and Network Bandwidth Transition

This transition however, needs to deliver transparent cost and performance benefits, and must take into account future changes in the offered traffic mix and offered traffic properties. These changes include a radio bandwidth transition—a change from constant-rate, variable-quality channels to variable-rate, constant-quality channels, and a network bandwidth transition—a change from synchronous to asynchronous traffic, and a shift from non-isochronous to isochronous traffic. This radio and network bandwidth transition is happening because the offered traffic mix contains an increasingly high percentage of rich media components—audio capture, image capture, video capture, and data capture. This is the basis for bursty bandwidth.

The burstiness of the offered traffic is determined by the application bandwidth and application dynamic range in the subscriber's handset. In 3GPP1, the assumption is that the dynamic range of any single channel stream can vary between 15 kbps and 960 kbps on a frame-by-frame basis, as demonstrated in Table 14.1. This is equivalent to an 18 dB dynamic range or a 64:1 ratio, which is then accommodated over the radio physical layer by the OVSF code structure (the SF4 to SF256 chip cover).

Table 14.1 Dynamic Range of Information Content

	10 MS	10 MS	10 MS
Bit rate	15 kbps	960 kbps	15 kbps
Ratio	1	64	1
		18 dB range	

However, this is the dynamic range excursion limit for a single channel. As we saw in earlier chapters, multiple channel streams produce much larger envelope variations, which are exhibited as large peak-to-average ratios that exercise our RF PA stages in the handset and Node B transceivers. We noted it was hard to deliver the linearity needed for multiple codes, as well as power efficiency. This issue does not disappear as the offered traffic moves into the network; in fact, it gets worse.

Traffic Distribution

You might expect that as traffic streams aggregate together (that is, multiple traffic streams from multiple users), the traffic streams would smooth. However, we have just said this doesn't happen in the handset, so why should it happen when we aggregate traffic together from lots of handsets? Specifically:

- The merging of traffic streams does not necessarily result in traffic smoothing.
- Bursty data streams aggregated together may produce even more bursty data streams.

For some years, we have known that data is different from voice. In August 1996, Jack Scanlon, a well-known and respected Motorola executive, announced at an analyst's briefing: "Network design tools today are voice-related, not data—data is different." He was absolutely right (and he was talking specifically about wireless networks). However, the equivalent lightbulb moment today is the realization and recognition that multimedia is different from data. Networks do not always behave as expected. Behavior is not consistent with traditional queuing theory. The traffic properties are different, and the traffic distribution is different.

Traffic distribution is a function of file size and session length. Traffic properties are a function of error and delay sensitivity (including sensitivity to delay variability). If a user has a big file to send and is going to be online a long time and the traffic being sent is delay- and error-sensitive, then by definition, the user will be bandwidth-hungry and will require a disproportionate allocation of radio and network bandwidth quantity and quality.

For a group of users, you can then assume that a small but significant percentage of users will have a disproportionate amount of RF power and network bandwidth allocated to them. This is known as *heavy tailed traffic distribution*. Traffic distribution and traffic properties directly influence radio bandwidth and network bandwidth provisioning (because bursty bandwidth remains bursty into the network core).

Traffic properties and session properties are interrelated. Session persistency determines traffic distribution; session complexity determines traffic property. Session properties are determined by application software.

The job of application layer software is to increase session length and session complexity. Session length and session complexity together produce heavy tailed traffic distribution, which in turn drives the requirement for radio and network bandwidth quantity and quality. This is an important point. If someone asks you how much radio bandwidth you need, you can only answer the question if you have some idea of what the future traffic mix will be in terms of offered traffic distribution and required traffic properties—ditto with network bandwidth.

Protocol Performance

You also need to be able to qualify protocol performance. We add traffic shaping protocols to manage the allocation of radio and network resources when demand on those resources at times exceeds supply. Traffic shaping protocols, however, absorb bandwidth, which means they can make matters worse. One solution is to overprovision radio and network bandwidth so that peak demands can always be serviced. However, as bandwidth becomes increasingly bursty, this becomes more difficult. The burstier the bandwidth, the higher the peak loads you need to accommodate and the more you will have to overprovision.

So you use traffic shaping protocols. Figure 14.1 shows the protocol stack. The properties of the offered traffic (bandwidth quantity and quality requirements) may be determined at the application layer. For example, the MPEG encoder output may describe what radio layer and network layer QoS is needed, or QoS may be determined at the presentation layer. XML, for instance, defines the QoS requirements needed for transaction processing. It is an information protocol and as such is able to determine and request QoS needs.

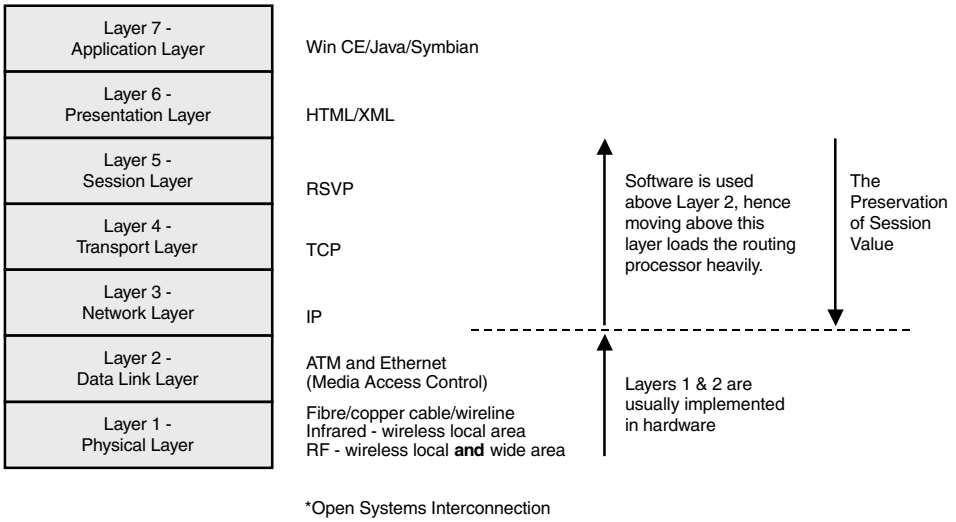


Figure 14.1 OSI seven-layer model (the preservation of session value).

Moving down the protocol stack, session persistency may be defined by RSVP, Diffserv, MPLS, and SIP (all of which we study in Chapter 17). Whether the traffic stream is isochronous or non-isochronous is determined in the transport layer and network layer (TCP/IP). Finally, radio admission is contingent on the MAC layer and allocation of physical radio resources. Admission control is congestion-driven in the network and interference-driven over the radio physical layer.

It is the job of the protocols in the protocol stack to preserve session value—specifically, session consistency in a persistent session and session properties, particularly the time-dependent properties of a rich media exchange.

Admission Control versus Policy Control

At this point we should differentiate admission control and policy control:

- Admission control maintains information about the available resources from a network entity.
- Policy control checks the administrative entitlement to the requested QoS (which may be for an individual user or aggregated services).

We have said that offered traffic distribution and offered traffic properties determine the provisioning of radio and network bandwidth. This includes delivery and memory bandwidth. We are discussing packet-routed networks that are, essentially, queuing networks. It is the process of queuing that delivers bandwidth efficiency by degrading QoS for users who hopefully don't notice or don't care, in order to deliver better service to those who do notice or do care and are willing to pay for the privilege of being given priority access.

Queuing networks do not like bursty bandwidth. Bursty bandwidth fills up buffer bandwidth. When buffer bandwidth overflows, packets are lost and, if using TCP, packets have to be re-sent, absorbing additional delivery bandwidth, as well as RF power. This points to a direct relationship between buffer bandwidth/memory bandwidth and packet loss, which in turn has a direct relationship with end-to-end delay and delay variability.

To quote from an IEEE paper (*IEEE Communications,* January 1999, Zheny, Atiquzzaman, Kouz, Sahinoghu, Tekincy): "Under many conditions, it can be observed that a linear increase in buffer size results in an exponential increase in packet loss." A typical performance for a 2-Mbyte buffer at 60 percent usage might be a 4×10^{-2} packet drop rate. A 64-Mbyte buffer reduces the drop rate to 3×10^{-3}. It is also not only the size of the buffer but the way the buffer memory is partitioned; a router memory that clumps data in batches may actually increase rather than decrease burstiness.

Further, there is a philosophical issue when considering how bursty bandwidth should be managed. You have a choice: You can either make networks application-aware (that is, the network determines the resource requirements of the application), or you make applications network aware (that is, the application can validate, qualify, and quantify the availability of current network resources, and possibly compare availability and cost in other networks).

Needless to say, network operators prefer the first option—put the network in control. Unfortunately, the second option works better, and it potentially confers more benefits, including cost-saving benefits, to the user.

Offered Traffic at an Industry Level

We can also look at offered traffic at an industry level. We said earlier that six industries are converging: computer, consumer electronics, IT, wireless, wireline, and TV. As we will see in later chapters, each of these industries has an interest in using Internet protocols to send information to and information from the World Wide Web.

We have agreed that a significant percentage of next-generation network value will be from subscriber-generated content—image bandwidth capture from computer, consumer electronics, and IT applications. This content or part of this content is delivered to the Web for storage (that is, archiving). Value is generated by the archiving process (content accumulates value over time) and the retrieval process—redelivery of the content back to the originator or the originator's friends or colleagues.

Some of this traffic exchange/traffic movement is time-sensitive, some isn't. The movement and storage of content is a billable moneymaking process, and whatever industry we are in, we like to standardize the mechanics of making money.

Converging Standards

Our six converging industries all have separate standards committees working on hardware and software standardization and content standardization (including, for example, authentication and encryption standards).

The Internet potentially provides a uniform set of protocols that can be used to move content around and make money from it. The Internet standards-making community (the IETF) certainly has an interest in this process. The Web provides the mechanism for storing and retrieving content (and in the process making money from it), so the WWW standards-making community (W3C) certainly has an interest as well. Ideally, all of these industries and their respective standards-making institutions would get together to standardize content management.

Arguably, the IETF is in the best position to do this, since the Internet is the point of intersection between all content generators. All six industries therefore have an interest in the offered traffic mix, offered traffic distribution, offered traffic properties, and offered traffic value. We have also determined that offered traffic value is dependent on the preservation of offered traffic quality (that is, the preservation of the properties of the offered traffic).

The Five Components of Traffic

We have said that traffic consists of five components—voice, image, video, file transfer, and transactions. In the past, we might have built a revenue stream on voice or data, but now we can potentially build five revenue streams, which can become part of a large and complex bill. The question then becomes this: how to build and bill complex value?

Each of the traffic types can be characterized in terms of its rate sensitivity/rate tolerance, error sensitivity/error tolerance, sensitivity to delay, and sensitivity to delay variability. Delay is sometimes described as *latency*. Latency is the amount of delay introduced between a sender and receiver in a simplex or duplex exchange (in a

duplex exchange it will be round-trip delay). Delay variability is sometimes described as *jitter*. Jitter is the amount of variation in latency. As a rule of thumb, jitter should not exceed 10 percent of the latency budget.

Jitter can be divided into macro jitter and micro jitter. *Macro jitter* is congestion-induced jitter caused by a shortage of delivery and memory bandwidth. Micro jitter is the small effects at the physical layer—receiver noise, clock jitter, sampling jitter, time, phase, and frequency effects. *Micro jitter* increases bit error rate. If this triggers "send again" protocols at a higher level in the protocol stack, then the result is macro jitter. This means there is a direct relationship to (in wireless) the radio physical layer (that is, radio bandwidth quality and network bandwidth quality). Radio and network bandwidth quality must be matched to the required service class.

The Four Classes of Traffic

The four classes of traffic in IMT2000 are conversational, streaming, interactive, and background:

- Conversational implies a delay of not more than 80 ms.
- Streaming implies a delay of not more than 500 ms.
- Interactive implies a delay of not more than 1 second.
- For background, delay is not specified.

Conversational traffic is historically constant bit rate. This means most conversational traffic to date has been voice, and voice has used constant-rate encoding. You can also have a conversational exchange of complex content, that is, a conversational rich media exchange. It would be more bandwidth-, power-, and quality-efficient for this to be variable rate.

Streamed content is also likely to be complex content—audio, image, video, data—and is therefore best supported as a variable bit rate service. Interactive uses available bit rate. This is variable bit rate, but with the network deciding on bit rate availability on the basis of congestion measurements from the network and interference measurements from the radio physical layer. Background uses unspecified bit rate, which means it's a best-effort service.

Sources of Delay, Error, and Jitter Sensitivity

We can qualify and quantify sources of end-to-end delay as follows:

- Source encoding introduces delay—typically 20 or 30 ms to source encode an audio or video bit stream. The more complex the encoding process, the greater the delay. Also, as we increase compression ratios, jitter sensitivity increases.
- Encryption may add additional processing delay.
- Channel multiplexing introduces additional delay and is largely determined by interleaving depth (which can vary between 10 and 80 ms in IMT2000DS) and the convolutional or turbo encoder delay.

- Radio path delay in comparison is relatively insignificant—typically a few microseconds. The radio path may, however, be discontinuous. If a user, for example, goes into a tunnel without radio coverage, the session will hang—delay and delay variability will be introduced. Discontinuous radio coverage will cause packet loss and trigger retry requests.

In a receiver the delay budget overhead is a product of demodulation and multiplexing, de-interleaving, channel decoding, and source decoding, including display driver or display delay, which means our conversational delay budget (80-ms end-to-end delay) is already accounted for by the delay introduced by the handset encoder/decoder delay. There is no additional delay budget available for network delay. This is why wireless IP doesn't work as well as wireline IP. Because wireless uses a variable-quality transmission medium (the fading radio channel), we need to add in processing overhead and processing delay to hide the channel variability and provide adequate channel consistency. In addition, we need to manage the gaps in radio transmission, making sure we can restart a session after it has been interrupted.

Table 14.2 characterizes traffic types in terms of their error, delay, and jitter sensitivity. Video, image transfer, file transfer, and transaction processing are all error-sensitive. Voice is error-tolerant. Conversational voice and interactive video are delay-sensitive. Transaction processing can be delay-sensitive if timeout challenge and response authentication algorithms are used. Voice messaging, file transfer, and image transfer are delay-tolerant. Error, delay, and jitter sensitivity are all properties that need to be preserved by the radio and network layer. The problem is that we are adding the delay and delay variability overhead introduced by the radio layer to the delay and delay variability introduced by the network.

Table 14.2 Traffic Characterization

Error-sensitive	Video Image transfer File transfer Transaction processing
Error-tolerant	Voice
Delay-sensitive	Conversational voice Transaction processing Interactive video
Delay-tolerant	Voice messaging File transfer Image transfer
Jitter-sensitive	Conversational voice Compressed video
Jitter-tolerant	Voice messaging File transfer Image transfer

Solutions to Delay and Delay Variability

We can reduce delay and delay variability by overprovisioning, but this reduces bandwidth efficiency and increases cost. We can control delay and delay variability by introducing traffic shaping protocols, which means we can determine that some users will have low delay and delay variability at the expense of others. This is the philosophical basis of an IP network. The challenge is to make a wireless IP network perform as well as a wireline IP network—to deliver wireless/wireline transparency. This means we have to avoid introducing additional delays on the radio layer. We cannot do this because there are a number of irreducible delay mechanisms—source and channel coding/interleaving, for example. We therefore must consider reducing, or at least managing and controlling, network delay/delay variability to make the wireless IP user experience transparent—that is, equal to the wireline IP user experience.

Implicitly this means that we need to provide more control of the end-to-end channel in a wireless IP network. We can achieve this by using IP traffic shaping protocols, effectively in-band packet-level signaling, or by using existing signaling plane control mechanisms (SS7). The SS7 signaling plane sits on top of the network traffic and is designed to manage the process of call setup, call maintenance, and call clear-down—that is, the management and billing involved in a circuit-switched voice call. This is inefficient if users are exchanging short bursts of data. It is therefore more efficient to use IP protocols.

However, we have said that it is the job of the application layer to increase session persistency (and session complexity, because session persistency and session complexity increase session value). A session may start as a nonpersistent session but is manipulated to become more persistent, more complex, and ideally more interactive as the session progresses. In IP terms, this implies that an initial best-effort service may need to be upgraded to an interactive or conversational session as the session progresses. As session persistency increases, we are effectively setting up the session, managing the session, and clearing down the session. Session management becomes directly analogous to call management.

A conversational complex rich media exchange represents the highest added-value service available within our network. We have to be very careful how we preserve the value of this exchange. Continuing to use SS7 for call setup, call maintenance, and call clear-down, and extending SS7 to manage session setup, session management, and session clear-down is one option (and the preferred option presently for most European and Asian, and many U.S. network operators).

If we use IP for session management, using protocols, like Session Initiation Protocol (SIP), then we need to be sure that these protocols perform at least as well and preferably better than the protocols they are replacing.

Managing the Latency Budget

We have to consider the need to deliver end-to-end latency (delay and delay variability guarantees) to meet particular quality of service requirements specified in our users' service level agreement. To do this, we need to take into account all the factors contributing to the latency budget. In a wireless IP network, these factors include the access latency introduced by the radio physical layer (availability of radio resources) and the access latency introduced by the network (availability of network resources). Access

delay and access delay variability may also be caused by poor protocol performance—the inability to allocate available radio or network resources.

The criteria for measuring access latency includes session setup success/failure and session setup delay. If we need to make multiple attempts to establish a session, we introduce delay and delay variability. We also absorb radio and network resources, including signaling bandwidth resource. Once a session is established, we then have the problem of network latency. Network congestion may mean we cannot deliver sufficient network bandwidth to deliver continuous session support, which means the session fails.

In a wireless IP network, session failure can also be caused, as we said earlier, by a discontinuity in the availability of radio resources. The user moves out of radio coverage or the call drops because of a problem with hard or soft handover control. Session continuity is therefore dependent on the continuous consistent availability of radio and network resources.

We also need to consider application latency. If the user-to-user or device-to-device exchange includes the need to access a server, for example, then delay and delay variability may be determined by a lack of server bandwidth or a failure to meet the server's admission permission criteria.

Delivering Quality of Service

From the user's perspective, the critical measurement is application performance. From the network operator's perspective, the criteria required is to be able to demonstrate and prove that pre-agreed application performance metrics have been delivered, which requires some form of proof-of-performance reporting. Delivering a particular quality of service, therefore, requires us to measure and manage access, network, and application latency, and access, network, and application jitter (application delay variability).

Many parameters interact with quality of service and are dependent on our ability to measure and manage radio and network bandwidth quality. These include the following:

Error protection choices. We must decide whether we put error protection at bit level, packet level, or protocol level. A send-again request implemented at protocol level will be more bandwidth-efficient provided it is not used very often. If it is continuously triggered by a discontinuity in the radio or network path, it will absorb radio and network bandwidth and introduce lots of delay and delay variability.

Circuit switching choices. We need to decide whether we circuit-switch, packet-switch, or cell-switch our offered traffic. This depends on whether the traffic is synchronous or asynchronous. If the traffic is asynchronous, how asynchronous is it, which means how bursty is our bandwidth (the frame-to-frame bit rate excursion)? We must decide how much of our traffic needs to be isochronous (packets arrive in the same order they were sent), how much can be non-isochronous. We must know or be able to predict whether our radio and network bandwidth needs to support symmetric or asymmetric traffic. If asymmetric, how asymmetric?

Bandwidth delivery choices. We need to decide whether we deliver transparent or nontransparent bandwidth. A nontransparent channel can give us constant bit error rates but at the cost of variable delay.

Impact on the user's experience. We need to be aware of the consequences of all these decisions on the quality and consistency of the end-to-end user experience.

Source coding issues. We must be aware of the quality issues of high-level source coding, including the impact of error rates and error distribution (over the radio layer), and packet delay, delay variability, and packet loss on highly compressed source-coded rich media channel streams.

Error rate considerations. We must consider that wireless voice networks have traditionally been planned on the basis of delivering 1 in 10^3 bit error rates, whereas wireline networks typically deliver a consistent 1 in 10^{10} bit error rate.

Multimedia requirements. We need to consider the very particular requirements of multimedia, including the need to multiplex multiple per-user channel streams at the application layer and multiple per-user channel streams at the physical layer and the relative requirement to synchronize these multiple streams with each other (or reclock the multiple streams with each other when they arrive in the receiver).

Delivering Wireless/Wireline Transparency

We have said that one of the objectives in 3G network design is to deliver wireless/wireline transparency. The problem is that to deliver performance equivalence, we would need to match wireline throughput, wireline quality, and wireline consistency.

Our expectation of the bit rate/throughput is determined by whether we use ISDN (144 kbps) or ATM (2.048 Mbps, 51, 155, or 622 Mbps, or 2.5 Gbps) or ADSL (8 Mbps down, 640 kbps up) or VDSL (40 Mbps over 2048 frequency bands). ADSL and VDSL are essentially mechanisms for releasing useful bandwidth in the copper access network (last-mile drop). ADSL and G.Lite (splitterless ADSL) occupies copper bandwidth between 20 kHz and 1104 kHz. The G.Lite specification is 1.5 Mbps downstream and 512 kbps upstream, giving an 1800-ft reach over twisted pair at 1 in 10^{10} bit error rate (a distance/quality metric).

Similar throughput gains are being achieved in hybrid fiber/co-ax networks. As with twisted pair, the objective is to user higher frequencies in the co-ax to deliver more bandwidth—potentially up to 900 MHz. Bandwidth provision is predominantly downlink-biased, optimized to deliver content to subscribers rather than capture content from subscribers. As we said earlier, this may not be appropriate, given that uplink loading may tend to increase over time.

Traditional Call Management in a Wireless Network

Figure 14.2 shows a traditional call setup, call maintenance, and call clear-down procedure in a wireless network. The procedure is similar to a wireline network, except the SS7 signaling has to manage interruptions (dropped calls due to lack of coverage or handover problems) introduced by the radio physical layer. Using SS7 signaling, a call is set up, maintained (typically for a couple of minutes or longer), then cleared down. For the duration of the call, the billable event has constant amplitude value.

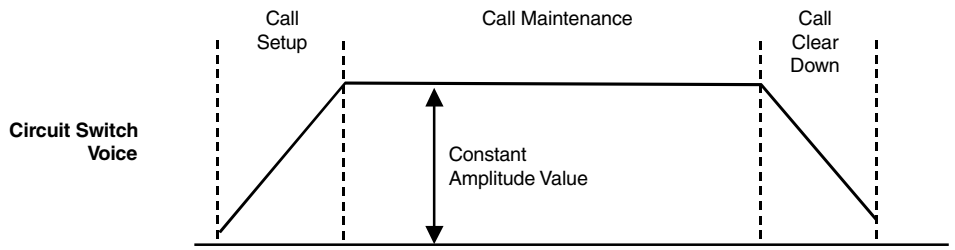


Figure 14.2 A conventional circuit-switched voice call.

Session Management in a 3G Network

Now compare this with a session-switched rich media exchange. As shown in Figure 14.3, in a session-switched exchange, a packet flow is established between two or more users or two or more devices. The job of the application layer software is to increase session persistency (that is, the length/character of the billable event) and session complexity.

Referring to Figure 14.3, note that there is an initial small bandwidth exchange, then at (a) an elementary packet stream is established. At (b) the session changes and dynamic rate matching is used to track the amplitude of the offered traffic. The burstiness is increasing as the session progresses. There could be a number of static matching step functions where new supplemental channel streams are added or taken away. Any one of these channel streams can be variable rate (dynamically matched). At (c), the session is closed down and the whole billable event can be captured by the billing process.

Session persistency and session complexity together determine session value. The complexity axis has variable amplitude value. This is a complex billable event. It is potentially quite complicated to capture and represent the value in such an event—and hence produce a billing rationale. We can simplify the billability by using visible quality metrics—quality metrics that the user can experience directly, such as color depth, frame rate, resolution, and audio quality. The session complexity does not or should not need to be explained as part of the billing process.

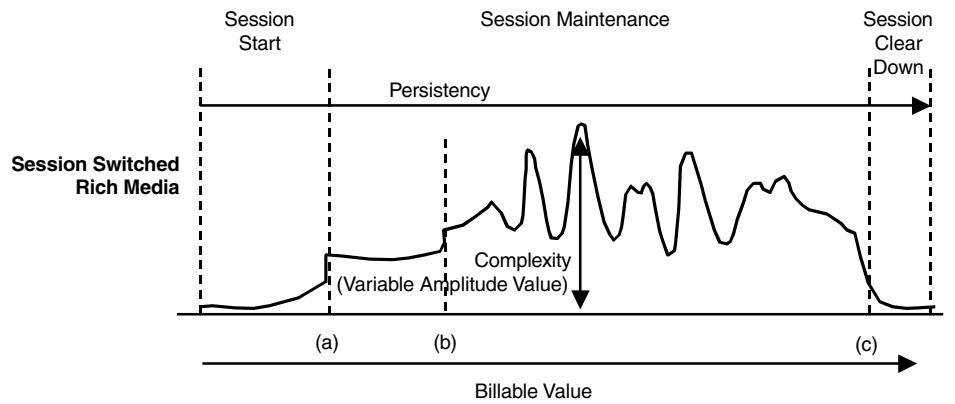


Figure 14.3 A session-switched rich media exchange.

Philosophically this is a connection-oriented exchange—a complex conversation between two people and two devices, or multiple people and multiple devices. It determines the amplitude (burstiness) of our offered traffic, which in turn determines our overall bandwidth requirement.

To support bursty bandwidth, we either need to overdimension the size of the (wireless or wireline) pipe to cope with peak loading or we need to buffer to smooth offered traffic amplitude, which introduces delay and delay variability. This means consistency has a cost: Consistent-quality complex rich media exchange is implicitly bandwidth hungry, which means we then have to qualify whether we have enough bandwidth available.

Table 14.3 compares wireline copper access, RF wireless access, infrared wireless access, and optical access. Optical access gives us potentially 270,000 GHz of bandwidth. Even just C Band and L Band gives us 10,000 GHz.

Bandwidth quantity and quality expectations increase over time and are driven by copper access and optical access performance benchmarks. Application bandwidth also increases over time, as does application dynamic range—the amplitude of the information envelope. The more dynamic range we need to accommodate, the more bandwidth we need.

Table 14.3 Bandwidth Comparisons

	WIRELINE COPPER ACCESS	RF WIRELESS ACCESS	INFRARED	OPTICAL ACCESS (INCLUDING INFRARED)
Frequency	300 Hz-1 GHz	3 MHz-300 GHz	100-118 THz	30-300 THz
Wavelength		300-0.001 meters	850-950 nanometers	240-2400 nanometers
			100 nanometers = 18,000 GHz	C Band: 1530-1570 nanometers L Band: 1570-1610 nanometers 1 nanometer = 125 GHz 80 nanometers = 10,000 GHz
Bandwidth quantity	1 GHz	300 GHz	18,000 GHz	270,000 GHz
Bandwidth quality	1 in 10^{10}	1 in 10^{3}	1 in 10^{6}	1 in 10^{12}
Bandwidth quality	Consistent	Inconsistent	Inconsistent	Consistent

The Challenges of Wireline and Wireless Delivery

We tend to think of copper access as being bandwidth rich, but in practice we become frequency-limited. Attenuation increases rapidly with frequency, which means the distance over which we can travel becomes limited as frequency increases. In practical and economic terms, it is hard to access frequency above 1 GHz. It is, however, a consistent delivery medium and can be very adaptive in terms of bit rate (the whole idea of ADSL).

RF wireless access is not bandwidth-limited. There is plenty of bandwidth available (270 GHz at least), but we are short of RF power. We can increase RF power by increasing network density. Although this has a cost implication, it does provide us with almost infinite bandwidth. However, RF is an inconsistent delivery medium that we need to tame by using measures like adaptive power control. However, there are times when a user will just simply be out of radio range. Depending on network build-out, there will always be coverage black spots both in urban and rural areas. Typically between 10 percent and 20 percent of a developed country with a mature network build-out will still have either marginal or nonexistent coverage. It would be implausibly impractical and expensive to provide the 99.999 percent availability delivered by a wireline network.

Most wireline networks were built in the era of national telecom monopolies where, in return for the right to a monopoly, the telco had a statutory obligation to provide service to all subscribers. Many wireline networks have been amortized over many decades, a luxury not available to wireless network owners. Although wireless theoretically offers coverage cost benefits when compared to new build wireline, these benefits disappear when compared to a legacy wireline network fully amortized many years ago.

In a wireless network, it is still expensive to get enough RF power distributed widely enough to give good consistent coverage. The RF power requirement is dictated by the quality requirement. Wireless network density is normally planned so that even in coverage areas, the typical bit error rate will be 1 in 10^{-3}, rather than the 1 in 10^{10} available over copper or 1 in 10^{12} available in the optical layer.

Infrared works in between 850 and 950 nanometers (100 to 118 Terahertz)—potentially a bandwidth of 18000 GHz, although typical bit rates are a power-constrained 2 to 4 Mbps. Infrared is also, like RF, an inconsistent and implicitly discontinuous delivery medium, working best with a clear line of site between sender and receiver. As an example, the ETSI/ARIB IrDA AIR (Area Infrared) specification delivers up to 4 Mbps over 4 meters or 260 Mbps over 8 meters with a 120° beamwidth.

As mentioned earlier, optical access promises almost infinite bandwidth. Even just taking C Band and L Band (80 nanometers) gives us potentially 10,000 GHz of bandwidth (10 Terahertz), and the optical layer is, of course, a consistent delivery medium.

However, optical fiber is not much use for mobility users. Free space optical transport shares many of the drawbacks of the radio physical layer—high attenuation (including fog and rain dispersion) and the need (in common with higher RF frequencies) for line-of-site communication.

We can never match the consistency available from wireline copper or optical fiber. RF or optical free space transmission is inherently inconsistent. We can, however, manage consistency as a quality metric in the same way we can manage bit error rate, delay, and delay variability.

The Cost of Quality

The advantage of wireless is that we have added mobility to the user experience. After 20 years of using cellular phones, however, we take mobility for granted and now expect to have the same services available to us in a mobile environment as we have from fixed access networks. Not only do we expect the same services, we also expect the same service quality.

Service quality metrics such as consistency can only be delivered from relatively dense wireless networks with robust and stable signaling—the cost of consistency. Low bit error rates and low-latency metrics also require relatively dense networks and robust and stable signaling—the cost of quality.

We have defined content quality in terms of frame rate, resolution, and color depth.

Meeting the Costs of Delivery

If we improve any of these quality metrics—frame rate, resolution, color depth—cost of delivery increases. The issue is whether tariff premiums can increase faster than the cost of delivery (see Figure 14.4). It is also difficult to calculate the real cost of delivery. If we are working with conversational services, we can assume there is no buffering, and there is a direct interrelationship between the application bit rate and radio and network bandwidth occupation.

However, interactive, streamed, and background all use buffering. Do we also cost in buffer occupancy as a part of the delivery cost budget? Consider: A user has a 10-Mbyte file to send. He is given the choice of sending the file in 5 seconds, 5 minutes, or 50 minutes. Now this could mean either the file takes 5 seconds, 5 minutes, or 50 minutes to send, or the network could guarantee that the file would be sent within 5 seconds,

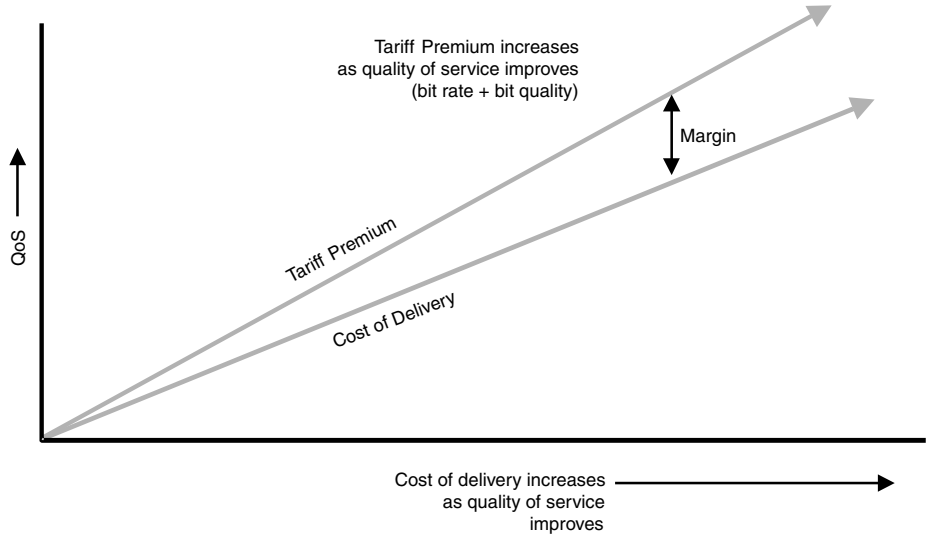


Figure 14.4 Cost of delivery and the tariff premium.

5 minutes, or 50 minutes; that is, the network can make the decision to use a high bandwidth channel for a short period of time or a low bandwidth channel for a long period of time, as long as the overall time taken meets the user's agreed time limit.

The tighter the time constraint, the more priority needs to be given to the traffic stream and, implicitly, the more it will cost to send. However, sending the file more slowly occupies buffer bandwidth. If this buffer bandwidth is memory bandwidth the customer has paid for (that is, buffer bandwidth in the handset), then it might be argued that this bandwidth has a nil cost to the operator (unless he decided to subsidize handsets with extra memory, in which case, it would be a real cost).

You might also argue that sending the file in 5 seconds rather than 5 minutes actually uses less network resource and therefore costs less. From a user's perspective, faster probably represents higher value, but if some users pay for the 5-minute service and get the 5-second service, because the network had instant bandwidth available, then the user paying a premium price for the 5-second service will be upset. Also, the elapsed session time (that is, the promised delivery time and the actual delivery time) needs to be monitored and measured by the network to provide the basis for a time = quality, time = money based billing metric.

Going back to time-based billing might be regarded as a backward step. Billing by the number of bits sent is now a well-established principle. Figure 14.5 shows the way a session might progress. The handset is supported in a steady-state condition of "always on" connectivity. Charges are based on the volume of data transmitted. For much of the time, no data is transmitted, so no charges are incurred. The offered traffic loading from the session is spasmodic and the file size exchanges are usually small—a few kilobytes, for example.

However, as color handsets have become available (typically with 65,000-color displays), and as CCD or CMOS imaging have been introduced, file size has become larger. As file sizes increase, it makes sense to give the user the option of send or receive at a faster or slower rate. The application provides a fast load or slow load option. The user pays a premium for priority.

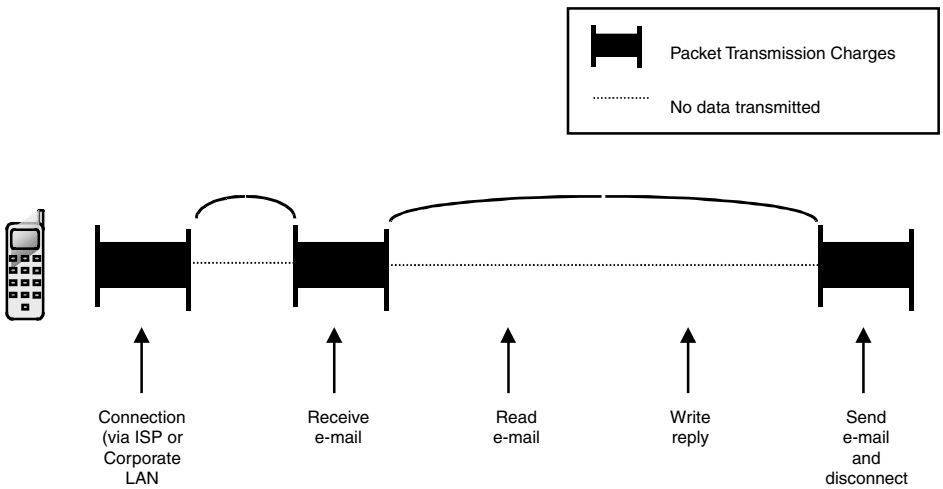


Figure 14.5 Packet transmission.

Example taken from the NTT DoCoMo FOMA service in Japan.

Here we have argued that we can charge more for sending something sooner (not necessarily faster in terms of bit rate, but sooner in terms of elapsed time). The time saving comes from not using buffering. We have also argued that it has cost us less to send! Buffer bandwidth is expensive bandwidth and becomes more expensive as offered traffic becomes more bursty.

We are trying to get away from the "always on, sometimes sending" model to a "sometimes on, always sending model." When a session is in progress, we want to increase the data duty cycle.

The Persistency Metric

In a complex session we want and need to have continuous activity throughout the session. We also want to increase the length of the session—the persistency metric. Figure 14.6 shows an initial channel allocation at (a). Then successive channel additions

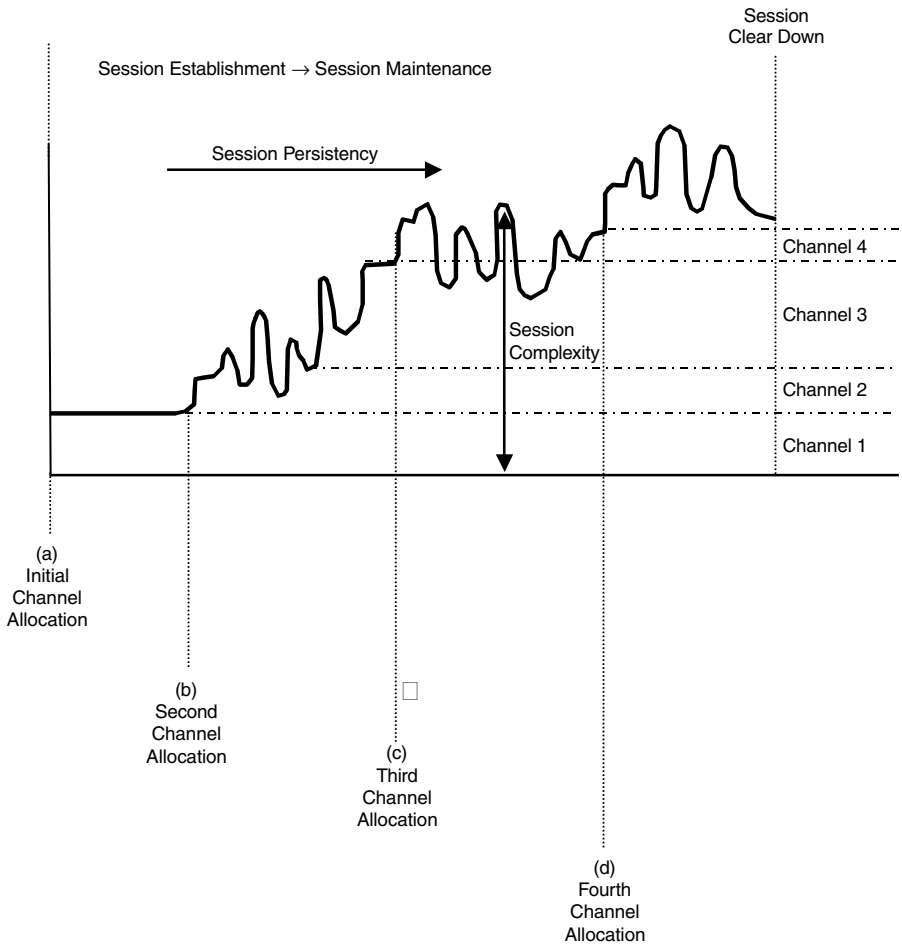


Figure 14.6 Session persistency and session complexity.

(b), (c), and (d) as the session progresses and session complexity/session amplitude increases. Each channel is variable rate. Each channel addition implies a step function increase in billable value.

We can buffer to clip the amplitude peaks in the information rate, but this destroys some of the properties of the offered traffic. If we wish to preserve the properties of the offered traffic, we need to match the information envelope to the physical layer. For example, (b), (c), and (d) in Figure 14.6 represent additional OVSF code allocations, each of which are variable rate (session-based static and dynamic rate matching).

Because we are not using buffering, we can describe this as a *conversational complex content exchange*. This is expensive bandwidth requiring radio and network resources to be matched in near real time (every 10 ms) to the information rate. Because the cost of delivery is high, we would hope that we could obtain a commensurate tariff premium—billing by session length and session amplitude value (session complexity). Consider also the impact of buffering.

- If we buffer the session (that is, downgrade the session to be streamed or interactive), then we are introducing delay and delay variability. We have reduced the value of the session. This may or may not be important to the use or the user's application.
- If we do not buffer the session, we have to overprovision radio and network bandwidth resources, that is, provide sufficient throughput to support the offered traffic at all times throughout the session. The use of dynamic matching, however, minimizes the bandwidth overhead (though it does imply substantial signaling overhead).

Even when we buffer, we will still end up using the same amount of transmission bandwidth but extended over time. Another way of looking at this is to consider that by overprovisioning delivery bandwidth, we can minimize or completely avoid the use of memory bandwidth.

Overprovisioning Delivery Bandwidth

By overprovisioning delivery bandwidth, we can also reduce the need to provide traffic control mechanisms, which in turn absorb signaling bandwidth. As we start to reduce the amount of delivery bandwidth available, we start to have to prioritize bandwidth access—that is, provide good service to some users at the expense of other users. Service quality becomes less consistent.

We revisit this issue when we study traffic shaping protocols in Part IV on network software, but essentially the whole problem boils down to the fact that bursty bandwidth is expensive bandwidth. Supporting bursty bandwidth across the radio physical layer requires static and dynamic rate matching, which in turn absorbs signaling bandwidth, hopefully minimizing the need to overprovision delivery bandwidth. The OVSF code structure provides a variable allocation of code domain power—the amount of RF power is matched to the variable information rate.

In the network, the cost of bursty bandwidth is correlated to the fact that we either have to overprovision transmission resources or overprovision router buffer bandwidth (or both), or we need to deploy a transport layer that can handle highly asynchronous traffic in a deterministic fashion.

In 3GPP1, we manage highly asynchronous traffic by implementing ATM (cell switching) across the access network. In 3GPP2, the problem is managed by overprovisioning. 3GPP1 uses smart thin pipes. 3GPP2 uses fat dumb pipes. Either way, additional delivery cost is incurred.

Traffic in the network has traditionally been managed by a circuit switch. A defined route is established between two points, and the traffic is connected using a hardware connection—a physical change of transistor state in the switch. This is inflexible but fast and consistent. An alternative is to packet-route the traffic. Here a packet header is read by the router, and the router software decides where the packet should be sent. This process introduces delay and delay variability—that is, it is flexible, but slow (relative to circuit switching) and inconsistent.

Session Switching

What we really need is to session switch. Here we set up a session, maintain a session, and clear down a session, with the capability to adapt to changes in session amplitude as the session progresses. This means we need adaptive circuit switching, or more accurately, adaptive session switching.

Packet switching is proposed as a possible solution. In packet switching, some of the software processes involved in packet routing are replaced with hardware implementation functions—that is, hardware coprocessors in the router. The hardware switching functionality is distributed out from the central switch to the routers. Packet-switched routers will then use IPv6, which has a defined header size, defined header fields, and optionally a standardized packet length (which optimizes software and hardware performance in the router). Packet switching effectively emulates ATM.

Preserving and Extracting Traffic Value

Thus we see that the shift in offered traffic (the need to preserve the properties of bursty bandwidth) fundamentally changes the way we have to treat traffic as it moves into and through the network. It is the job of the network to not only preserve this value but extract some value as the traffic passes.

Table 14.4 shows some of the cost/value issues involved in preserving the properties of highly asynchronous offered traffic and capturing some of the value contained in passing traffic. The handset has to have hardware and software capable of capturing and processing the complex multimedia mix—a task requiring substantial multitasking and multiplexing at the application layer. The real-time operating system and man-machine interface (mmi) need to be responsive to these task requirements.

Table 14.4 Cost/Value Distribution

HANDSET		**BASE STATION**		**NETWORK (SWITCH SERVER)**		**TRANSPORT (BACKHAUL)**
APPLICATION LAYER	**PHYSICAL LAYER**		**NETWORK LAYER**		**TRANSPORT LAYER**	
EMBEDDED ADDED VALUE						
RTOS/MMI for example, OS9 Windows CE Java encryption Authentication Compression Multitasking Multiplexing	IS95 CDMA IS136 TDMA GSM TETRA APCO 25 Bluetooth	Servers Routers Edge switching Traffic contract negotiation Distributed billing (Mbyte meters)		Data warehousing Data mining Transaction processing Information distribution Application distribution		WDM Fiber Traffic Management Multi-point to multi-point distribution
ENTERPRISE O/S						
Netscape Navigator Microsoft Explorer CDF/ActiveX/DNA						
TCP/IP/ATM (IPv6)			TCP/IP ATM Frame Relay		TCP/IP ATM Frame Relay, SDH	

We then need to preserve the asynchronous properties of the offered traffic over the physical layer. We need to manage admission control at the base station and RNC, and we need to provision storage bandwidth at various points in the network, including the handset, the base station, and the RNC. We may decide to provide some data warehousing capabilities in the network (adding archiving value to delivery value). If traffic is then moved into and through the core network, we need to preserve offered traffic properties, including offered traffic asynchronicity. Typically at some stage in its journey, highly asynchronous traffic will need to be moved on to a synchronous transport layer SONET, over the wavelength-division multiplexed (WDM) optical layer; that is, we move traffic on through the network in an organized and deterministic way. We may also need to store the traffic, that is, warehouse it for future delivery, a process best described as long-term buffering.

The Cost of Asymmetry and Asynchronicity

We have considered the additional cost implications of asymmetric traffic. A voice network is a symmetric network; the uplink and downlink are balanced. A rich media network is by nature asymmetric, since asynchronous traffic is by its nature asymmetric. The asymmetry is not fixed but constantly varying as the offered traffic rate increases and decreases on a (10-ms) frame-by-frame basis in either direction. We see the symptoms of this in the radio physical layer where the allocation of code domain power is constantly changing on the uplink and downlink to accommodate constantly changing bidirectional user data rates.

Similarly, the allocation of network resources will be constantly changing. It will be statistically rare for the uplink and downlink to be balanced—another argument for using ATM (or equivalent packet-switching capabilities).

Considering the Complexity of Exchange

We also need to consider the complexity of the exchange between users, users and devices, and devices. In a voice network, we have said that the exchange is essentially symmetric and duplex (bidirectional)—a point-to-point exchange.

In a broadcast application, traffic is completely asymmetric. Television would be an example of an asymmetric point-to-multipoint application. However, we need to consider that a lot of broadcast content now contains trigger moments—voting in a game show, for example. A broadcast might trigger a simultaneous response from all or some of the audience generating an instantaneous requirement for uplink bandwidth. Broadcast SMS is an example of a one-to-many exchange that can trigger a many-to-one (multipoint-to-point) response. This defines an important traffic property—the need for instantaneous uplink bandwidth in response to a downlink broadcast.

This loading phenomenon can be observed in the national electricity grid. At the end of a popular TV program or in the commercial break, power demand suddenly surges as everyone goes to put the kettle on to make (in Britain) a pot of tea. The national grid has to be substantially overprovisioned to support these very concentrated and sudden loading peaks. Similarly, with gas distribution, when 20 million

households all cook Christmas dinner at the same time, gas pressure drops. Demand does not go down; ovens have thermostats, so they are smart enough to know they need constant heat.

Anyway, we digress. The point about broadcast streams or trigger streams is that they can deliver substantial uplink offered traffic loading (and hence uplink offered traffic value) but are by nature very peaky in terms of needing to deliver instantaneous access bandwidth. This can either be provided by dramatically overprovisioning uplink bandwidth or by supporting uplink demand with a mix of delivery and buffer bandwidth resource (which implies giving subscribers a variable access delay).

Archiving Captured Content

Content capture is also by nature asymmetric in the uplink direction. Surveillance devices and Web cams are uplink devices. However, value can be generated by archiving captured content. Captured content can then be redelivered to the originating subscribers. This is subscriber-generated content that ends up being consumed by the same subscribers (we enjoy looking at our own content). One objective in content capture is to realize value from content redelivery. Offered traffic loading here is determined by the image capture bandwidth of the device. Image resolution, color depth, and frame rate determine offered traffic volume and offered traffic value.

Redelivery bandwidth is determined by the display bandwidth constraints of the receiving device. Early experience with 3G networks in Japan and Korea confirmed, for example, that 65,000 color displays increased average revenue per user (ARPU). The quality of the experience was better, so users wanted more. Handset hardware and software determines offered traffic loading and offered traffic value, which means handset hardware and software determines network hardware- and software-added value.

Value is greater if we can capture subscriber-generated content, archive this content in the network, and deliver it back to the originating subscriber (and to other subscribers).

The number of participants in this type of exchange, however, can be very fluid and dynamic. If we consider personal subscribers, the offered traffic loading would be determined by the size of the *buddy group*—the number of people participating in a chat room exchange, for example. Chat groups also don't just exchange e-mails, but may exchange images, voice, and video as well (multimedia messaging). People leave and join the chat group as a session progresses (the session may be continuous 24 hours a day).

In a corporate or specialist user application, user groups could be actively configured by the network, and group configuration/reconfiguration can be event based. Motorola calls them *storm plans*, which, as mentioned in an earlier chapter, is the ability to reconfigure a network in response to a particular event, a natural disaster, or terrorist attack. A storm plan will involve preplanning a response to a possible future event. The preplanning will include user group configuration, priority and access rights to delivery bandwidth, and priority and access rights to storage bandwidth—information on hazardous chemicals, for example, or terrorist personality profiles. Sometimes such networks are described as being *situationally aware*.

Flexible user group configuration and reconfiguration can be very complex to implement and involve issues such as authentication and policy control. User group membership lists have to be maintained and updated, including records of individual user profiles. Many of the techniques used in specialist private mobile radio (access

and policy control and user group configuration) are also directly applicable in public access networks. Problems tend to arise when hundreds or thousands of subscribers are participating in a user group with lots of different hardware and software device profiles and lots of different service profiles. Access can also become protocol-limited. Priority protocols are difficult to implement in multipoint-to-multipoint exchanges, particularly when there is a high rate of change in user group configuration (lots of people entering or leaving the user group).

Buddy groups can produce very bursty offered traffic. A comment might provoke a stream of replies; arguments tend to end up with everyone speaking at once. Buddy group interaction is an effective mechanism for increasing session persistency and session complexity (and by default session value). It is the job of the network to sustain and preserve this session value.

Increasing Offered Traffic Loading

What we really need most are applications that increase offered traffic loading in the off-peak hours in a network. Figure 14.7 shows a 24-hour loading of a cell site in Biggleswade, a rather sleepy suburban town in the United Kingdom (with the lowest divorce rate in Britain). The vertical axis is Erlangs (voice traffic loading) and the horizontal axis is time. Between midnight and six in the morning, Biggleswade sleeps (probably why the divorce rate is so low), and so does the network. It is virtually unloaded. At 8:00 A.M. Biggleswade wakes up and goes to work. At 1:00 P.M. Biggleswade stops for lunch. Biggleswade starts going to sleep again as the afternoon progresses. A small evening peak happens just about 8:00 P.M. as Biggleswade arranges the traditional daily visit to the pub.

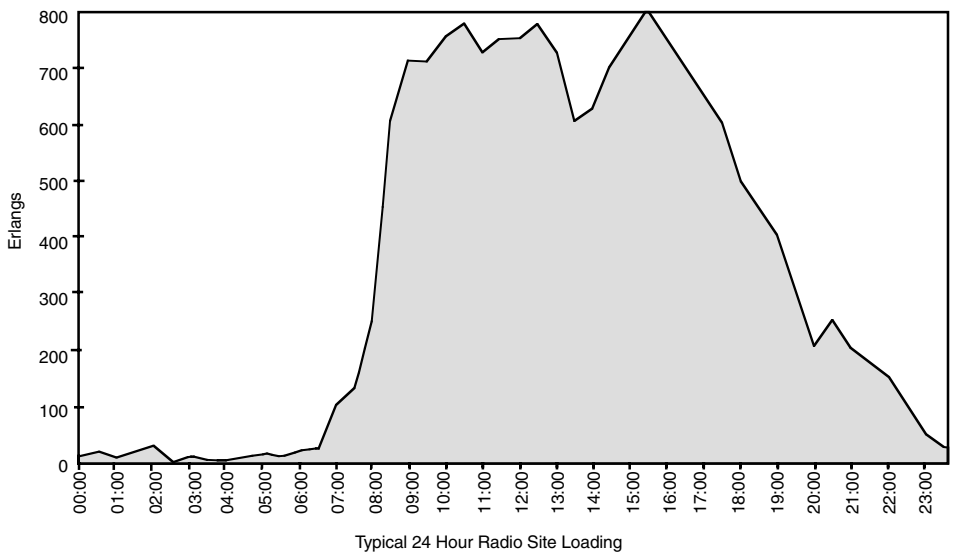


Figure 14.7 Latency bandwidth availability.

Our thanks to our friends at Smith Myers for providing this information to us many years ago.

This is a pre-SMS plot of diurnal loading. The impact of SMS text messaging has introduced additional loading in the evening. One text message may prompt multiple replies (from the SMS buddy list), generating additional network traffic.. This loading includes voice traffic stimulated by the SMS exchange. SMS is very economic in terms of bandwidth used—a few hundred bits and relatively high value in terms of billability. This means it has high perceived value to the user. Even a full-length SMS message only takes 1 ½ seconds to send but can deliver a tariff gain equivalent to a 2-minute phone call.

SMS is store and forward, so not only do we have a low bandwidth requirement, we also have low-cost bandwidth in terms of latency tolerance. This means memory bandwidth allows us to use delivery bandwidth more effectively. In SMS, the traffic delay may be just a few seconds. However, certain types of traffic may be delayed by several hours and not suffer loss of value.

Video surveillance, for example, can be (although not always will be) delay-tolerant. Provided sufficient buffer bandwidth can be made available in the device, then image bandwidth can be delivered in the middle of the night for centralized (network-based) storage. Similarly, if a handset has sufficient memory, it can be updated overnight, that is, trickle fed by the network. The trick is to get the subscriber into the habit of charging his or her phone overnight and leaving it switched on. The network can then download, for example, the subscriber's three most frequently visited Web sites. In the morning, the user has the illusion of instantaneous access to delivery bandwidth; the three sites are immediately available to the user. All the network has to do is difference the site—that is, update any changes that have occurred.

Predicting Offered Traffic Load

The way we use the Internet has a number of interesting implications for offered traffic distribution. The average holding time for an Internet session is 30 times longer than a traditional voice call. Fortuitously, Internet busy hour is not the same as voice busy hour. Session call length peaks at 4:00 A.M., at over 80 minutes. (It is probably best not to speculate what people might be doing on the Internet at four in the morning.) If we can fill the whole of our loading box in Figure 14.7, we have doubled our network bandwidth utilization. We might not have doubled our revenue, but we will definitely have increased our revenue and margin.

We need AMPU (average margin per user) rather than ARPU (average revenue per user). SMS is good because it costs relatively little to send but can be billed at a disproportionately high rate. SMS accounts for 10 percent of revenues but 20 percent of margin in a typical European network.

Delay-tolerant content and delay-tolerant applications are useful for quiet-hour loading. Application or software downloads to handsets and image uploads from surveillance devices are both examples of delay-tolerant content. The loading shift can, however, only be achieved by having buffered bandwidth available both in the device (handset or surveillance device) and the network. It may even be worth subsidizing additional handset-resident/device-resident memory to encourage this load shift effect.

Offered traffic has the habit of being in the wrong place at the wrong time. We can support the offered traffic load by buffering. The buffers can by milliseconds, seconds,

or hours, depending on how delay-tolerant our data is and what we are trying to achieve. It may be that our highest-value offered traffic is very delay-sensitive. We may want to completely avoid buffering—effectively to circuit-switch or virtual circuit-switch the data (or rather, session-switch the data).

We can only dimension the network accurately if we can predict what the future traffic mix will be and what traffic quality will be required. We can predict the likely traffic mix if we can predict the likely mix of devices producing the offered traffic load. Take a million subscribers and determine what the likely hardware and software device mix will be in 3 to 5 years' time. How many subscribers will have digital cameras in their phones? Will the digital cameras be CCD or CMOS? How many subscribers will have 65,000-color screen displays, and display drivers capable of supporting 12, 14, 16, or 24 frames a second? The subscriber product mix will determine uplink loading distribution and downlink load distribution. The subscriber product mix will determine uplink loading traffic properties and downlink traffic properties.

The traffic properties will determine radio and network bandwidth quality requirements, which, in turn, will determine radio and network resource requirements—how much network density is needed, how much backhaul is needed. Defining the subscriber hardware/software product mix also helps us to define how bursty the bandwidth will be, which, in turn, determines how much signaling bandwidth we need and how much additional traffic bandwidth we need.

Handset hardware and software determines offered traffic loading. Offered traffic loading determines future network value and future network cost, and enables us to calculate future network margin.

Summary

Traffic is coming from an ever-increasing number of sources—computer products, consumer electronics products, IT products—and may come from or go to an ever-increasing choice of wireless, wireline, or (digital) TV networks. The traffic is becoming increasingly asynchronous, which means burstier bandwidth. Traffic properties (and traffic value) have to be preserved as traffic is moved into and through wireless, wireline, and TV or radio networks.

An increasing percentage of this traffic may pass over the Internet. This means that we need to qualify the impact of bursty bandwidth on Internet protocols—the impact of Internet protocols on bursty bandwidth. The Internet is a queued network. Queued networks do not like bursty bandwidth. Bursty bandwidth compresses buffer bandwidth. Compressed buffer bandwidth triggers transmission retries, which, in turn, absorb radio and network bandwidth and introduce delay and delay variability (which compromises session value). Cellular networks have conferred the gift of mobility to the user experience. Unfortunately, users now take mobility for granted and expect wireless networks to perform identically to wireline networks.

The performance available from wireline networks, particularly circuit-switched wireline networks, provides a benchmark against which wireless networks are judged. The benchmark is determined by bit rate and bit quality. Bit quality is determined by

metrics such as delay and delay variability, and bit error rate. These quality metrics have a direct impact on session quality. Session quality is determined by the quality of the network (and the radio access layer) and how well the network can accommodate rapid changes in session amplitude. Session amplitude is determined by the dynamic range of each user's offered traffic and the aggregated dynamic range of multiple users when multiplexed together (the per-user and multiuser multiplex).

Inconveniently, bursty bandwidth does not smooth when users are multiplexed together and may become more bursty. This tends to put network components (RF components and router buffers) into compression, which produces packet loss and retries. One answer is to overprovision the radio access network and core network, but this increases delivery cost.

Ultimately, we need to reassure ourselves that the additional delivery overheads implicit in managing the rich media mix can be recovered from higher tariffs. We want to avoid overprovisioning, because this adds unnecessary cost. We want to avoid underprovisioning, because this compromises content value. We can only dimension networks accurately if we can predict the product mix:

- How many handsets or devices there are in our network?
- What is their image capture capability (image bandwidth)?
- What is their video capture capability (video bandwidth)?
- What is their audio capture capability (audio bandwidth)?

This determines uplink offered traffic. Display capabilities (display and display driver, and audio driver bandwidth) determine downlink offered traffic.

A view also has to be taken on the peak-to-mean loading on the network and whether or not we provision (overprovision) for peak loading. These loading issues become particularly complex when we consider trigger moments on the downlink that create a substantial instantaneous peak in demand for uplink bandwidth.

Network planning still tends to be focused on downlink bandwidth provision. In reality, network value is substantially moving toward uplink-generated value. Uplink quality (the ability to handle highly bursty uplink offered traffic) will be a key future requirement.

CHAPTER 15

Network Hardware Evolution

In the previous chapters in this part of the book we described how traffic is becoming increasingly bursty and how this exercises many of our system components. As user bit rate increases, and as the need for bandwidth quality increases, so network density increases.

The Hierarchical Cell Structure

The trend over the past 15 years has been to infill macro sites (up to 35 km radius) with micro sites (500-meter radius) and to infill micro sites with pico sites (100-meter radius). This infilling is called a *hierarchical cell structure*. The handover algorithms are optimized to move traffic from the macro sites to the micro sites to the pico sites, depending on factors such as mobility, signal strength, and traffic loading.

Suppose you are driving down Oxford Street, the main shopping street in London, as you make your journey, you might typically be handed over from several micro sites. Because the traffic is heavy, you will be a slowly moving user and are probably only driving at walking pace. As you turn into Hyde Park (wide open space), you might be handed over to a macro site. If you get out of your car in Oxford Street and walk into a shop, you would be handed over from a micro site to an indoor pico site.

Hierarchical cell structures can provide an almost infinite amount of bandwidth by increasing network density. Cellular networks are therefore very adaptive—able to support a mobile in a 35 km cell, a user in a 100-meter picocell, and handover users from cell

to cell, without any detectable gap or loss of quality. Power control and handover algorithms continue to improve over time. We become expert at optimizing network performance to provide wide area and local area coverage.

There is, however, a cost associated with this flexibility. Power control and handover algorithms absorb signaling bandwidth. As this is part of the overall bandwidth budget, then we will always have significant power control and signaling overheads.

Typically, at least 20 percent of our bandwidth is used for signaling. An additional 40 percent is absorbed by channel coding, much of which is needed to counteract the wide area radio impairments—delay spread from multipath, fast fading, the need to support high mobility (Doppler spread), interleaving to counter burst errors. Not only does this absorb bandwidth (RF power), it also introduces tens of milliseconds of processing delay.

Local Area Connectivity

Now let's consider local area connectivity—for instance, a person in an airport. If a person is using a laptop to download a file, he will be more or less stationary. This is defined as a portable application rather than a mobile application. If the user does move, it will be a relatively slow move (walking from area to area), which will be relatively easy to track. There is less need for a sophisticated and bandwidth-hungry signaling overlay. In addition, because the user is inside and probably close to a base station, there will be little or no multipath to worry about, so much of the channel coding needed for wide area coverage can be discarded.

An optimized air interface can therefore be produced to support portable local area access connectivity without the bandwidth overheads and delay overheads associated with a wide area mobility air interface. This is the thinking behind wireless LANs (WLANs). Table 15.1 summarizes present wireless LAN standards and typical gross and net data rates.

Wireless LAN Standards

The IEEE802 standard has been in existence for over 10 years, and a number of U.S. vendors have been producing wireless LAN products for commercial in-building applications, that is, private access networks. The idea is to extend wireless LANs into the public access network space using either the unlicensed Industrial Scientific Medical (ISM) band at 2.4 GHz or the ISM band at 5 GHz.

Present IEEE 802.11 products use direct-sequence spread spectrum (DSSS). Each bit transmitted is modulated by an 11-bit barker sequence (a pseudorandom sequence) to give just over 10 dB of processing gain. The data is either differentially binary phase shift keyed or differentially quadrature phase shift keyed onto the RF carrier, which uses 25 MHz channel spacing. The 2.4 GHz ISM band allocated in the United States is 2.402 to 2.480 GHz, that is, 78 MHz giving 3×25 MHz channels with some guard band. A frequency-hopping physical layer is also specified and an infrared physical layer. RF products to date have all been DSSS, and this seems likely to remain the case for the foreseeable future.

Table 15.1 Current and Future WLAN Standards

WLAN SYSTEM	CAPACITY PER CHANNEL		MAX RANGE	AIR INTERFACE	CHANNEL BAND-WIDTH	FREQUENCY	NO. OF CHANNELS			QOS
	PHYSICAL LAYER	REAL MAX THROUGH-PUT					U.S.	ASIA	EU	
802.11b	11 Mbps	6 Mbps	100 m	DSSS	25 MHz	2.4 GHz	3	3	4	No
802.11a	54 Mbps	31 Mbps	80 m	OFDM	25 MHz	5 GHz	12	4	0	No
802.11g	54 Mbps	12 Mbps	150 m	OFDM/ DSSS	25 MHz	2.4 GHz	3	3	4	No
HomeRF2	10 Mbps	6 Mbps	50 m	FHSS	5 MHz	2.4 GHz	15	15	0	Yes
HIPERLAN2	54 Mbps	31 Mbps	80 m	OFDM	25 MHz	5 GHz	12	4	15	Yes
5-UP	108 Mbps	72 Mbps	80 m	OFDM	50 MHz	5 GHz	6	2	7	Yes

802.11d—Other RF bands
802.11e—QoS
802.11f—Handover protocols
802.11h—Power control
802.11i—Authentication and encryption
802.11j—802.11/HIPERLAN interworking

The physical layer header determines which RF interface and modulation option are used and defines the packet size and channel coding used (single block code parity check). Some synchronization bits at the beginning of the header provide the basis for locking onto the RF carrier and correlating to the PN-coded channel.

Direct sequence tends to be used because it provides better performance via more processing gain and coherent demodulation. However, the RF PA needs to be linear, and the usual IQ balance issues have to be addressed. The present MAC (Medium Access Control) layer is Ethernet based and doesn't support quality of service, but this is presently being redefined (as 802.11e).

IEEE 802a defines the standard for a 54 Mbps LAN product at 5 GHz using OFDM. This means that a similar multicarrier approach is being adopted by wireless LAN vendors, digital TV (in Europe and Asia), and, as we will see later, a number of fixed-access wireless vendors. This is a 300 MHz bandwidth allocation (12×25 MHz channels). Net throughput per channel is 31 Mbps. 802.11g then backward-engineers this standard to retrofit back into the 2.4 GHz ISM band with the option of using OFDM or DSSS. Net throughput per channel is 12 Mbps.

Other flavors of IEEE 802 include 802.11d to cover application in other RF bands, 802.11f to address handover protocols, 802.11h to cover power control, 802.11I to cover authentication and encryption, and 802.11j to cover 802.11/HIPERLAN interworking.

A number of companies are also promoting OFDM-based solutions with optimized IP protocol backhaul and local area/wide area handover including billing integration. One example is Flarion (www.flarion.com). There are also competing standards from Japan (Home RF2 using frequency hopping) and a proposed future HIPERLAN2 (an ETSI specification).

The addition of authentication and encryption to IEEE802 has made it possible to deliver a public access network proposition, though issues still need to be resolved as to how billing is managed as a user moves from a wireless LAN to cellular coverage. The scenario is that a user arrives at an airport having been on the wide area cellular network. The user's laptop (with an integrated wireless LAN transceiver) locks onto the local wireless LAN base station and downloads/uploads any files waiting to be received/sent. If the user walked back out of the airport, the laptop would camp back on to the cellular network.

Delivering a Consistent User Experience

The challenge will be how to deliver a consistent user experience. Hot spots, such as airports and convention centers, are already saturated with cellular coverage, and the cellular networks often already occupy much of the available soffit space and conduit bandwidth, often using RF over fiber distributed antenna systems.

Wireless LANs will also need to compete with TDD Node B transceivers capable of delivering a (probably sufficient) consistently good-quality 2 Mbps data stream. The ISM band is unlimited, and anyone can use it, provided they do not exceed specified RF power outputs. Machines co-sharing the spectrum could include industrial, scientific, and medical microwave devices and Bluetooth devices (in the 2.4 GHz band). Note also the band allocations at 5 GHz are different for the United States and Asia. Physical layer

connectivity may therefore be rather inconsistent from location to location and certainly inconsistent from country to country. There are also differences in terms of allowable effective isotronic radiated power (EIRP) at 2.4 GHz—1 W for the United States, 100 mW for Europe, and 10 mW per MHz for Japan (26 MHz).

Even if the physical layer connectivity issues are addressed, there still remain substantial application layer connectivity issues. Even if the application layer connectivity issues are resolved, there still remain issues of physical access to hot spot locations already intensively served by existing cellular vendors. And even if the physical access issues are resolved, issues of billing and resolution of revenue capture disputes between competing wireless LAN operators and established cellular vendors still remain.

Sharing the Spectrum with Bluetooth

An added complication is that IEEE 802 shares spectrum with Bluetooth. Bluetooth provides an even more localized connectivity option—for example, to provide a localized RF connection between an earpiece/earbud headset and a cellular phone, or to connect a cellular phone to other peripheral devices, or to connect peripheral devices to other peripheral devices. Table 15.2 shows the ISM band allocation by country at 2.4 GHz.

A Bluetooth transceiver is a low-power device—either 100 mW (+20 dBm), 1 mW (0 dB) or 1 μW (-30 dBm). It uses frequency hopping at 1600 hops per second across 79 hop frequencies at 1 MHz spacings. It uses simple FM modulation to reduce component costs and power consumption.

Table 15.2 ISM Band Allocation by Country at 2.4 GHz

REGIONAL ALLOCATIONS	TOTAL BANDWIDTH	= 83.5 MHZ
Excluding Guard Bands:		
North America	2.402 to 2.480 GHz	= 78 MHz
		= 79 hop frequencies
Europe frequencies	2.402 to 2.480 GHz	= 20-79 hop
Except:		
Spain	2.447 to 2.473 GHz	20-27 hop frequencies
France	2.448 to 2.482 GHz	20-35 hop frequencies
Japan	2.473 to 2.495 GHz	23 hop frequencies

Present devices are either two-chip or single-chip implementations. For cost reasons, the processes use CMOS. This tends to compromise receive sensitivity (high noise floor), particularly in single-chip devices. Range is not included in the specification, but vendors have derived figures of 100 meters for the 100 mW device and 10 meters for the 1 mW device, assuming -70 dBm receiver sensitivity and -5 dBi antenna gain.

Working in a Real Office Environment

In practice, a real office environment typically requires an antenna gain of 10 to 20 dB because of the absorption and reflection from walls and furniture.

In practice, performance across the radio physical layer is very variable if distances of a few feet are needed. If the link is just between a headset and a handset, then the link budget is okay, but then arguably a wired connector could do the same job at a faster, more robust, and more consistent data rate, and it avoids the problem of needing a battery in the headset device.

Several voice codecs are supported, including continuously variable slope delta modulation (CVSD)—a voice codec that can work in a high bit error rate channel (4 percent bit error rate). Connections can either be synchronous for voice or asynchronous connectionless for packet data. There are two types of block encoding, 1/3 rate and 2/3 rate, or simple ARQ (automatic repeat request).

Gross data rate achievable is 1 Mbps typically divided into 2×432.6 kbps symmetric channels or a 721 kbps unidirectional channel with a 57.6 kbps return channel. Each packet is transmitted at a different hop frequency. A single Bluetooth transceiver can support seven simultaneous links, that is, up to eight connected devices that then share the available bandwidth. Devices can be organized in a scatternet—a collection of multiple and nonsymmetrical pico nets. Alternatively, one of the devices can be elected as the master unit (usually the first device to be turned on). The master unit clock and hop sequence is used to synchronize all other devices in the pico net.

Joining the Scatternet Club

It must be said that to date, most applications using Bluetooth have been one to one—a host and slave device (handset and earbud, for example). In practice, it is very hard to develop consistent rules for joining and leaving a scatternet—for instance, authentication procedures. You could have 100 Bluetooth-enabled handsets all in the same room, but you might not want them all to talk to each other. If the device is in discovery mode, it would spend all its time interrogating other devices and inviting them to join its scatternet club. All the other devices could potentially be doing the same. Who decides which club to join?

Because there are potentially so many different hardware and software form factors, it is difficult to define a common ontology that can be used for disparate devices to communicate with one another. There is not a huge amount of point in getting a Bluetooth earpiece headset to talk to a Bluetooth-enabled printer. It would not have much to say. Just because something is possible doesn't mean you should do it!

So as with wireless LANs there are physical layer performance issues. Bluetooth and IEEE802 wireless LANs used together in the 2.4 GHz band add mutually to each

other's noise floor. The Bluetooth device suffers from poor sensitivity anyway, so the quality of the connection will be far from constant.

The Bluetooth Price Point

In addition, Bluetooth needs to be delivered at a very low price point. The choice of FM helps reduce costs. There is no need for linearity in the PA, and you can use a simple FM modulator and FM discriminator. The use of CMOS helps to keep costs low, as does a single chip RF/baseband execution, but both these factors decrease receive sensitivity.

The Bluetooth add-on cost has to be equivalent to the cost of adding infrared to a handset—about $1.50. However, the infrared port is sometimes used by manufacturers to calibrate a handset as it moves along a production line. It has more than paid for itself before the phone has been built. You could not use Bluetooth to do this (it's a bad idea to use an RF device to calibrate an RF device).

Dealing with Infrared

Bluetooth also needs to perform against an infrared port that is evolving over time. The ETSI/ARIB IrDA AIR (Area InfraRed) specification defines a point-to-multipoint capability in which multiple devices can be supported from a host device provided they are within a 120° beamwidth. The data rate is 4 Mbps over 4 meters and 260 kbps over 8 meters.

The disadvantage with infrared is that it doesn't work so well in strong sunlight. As with RF, free space optical transport is an inconsistent delivery medium subject to blocking, refraction, and reflection. IrDA also suffered, still suffers some would say, from incomprehensibly difficult driver software, which has to be installed on a target device—for example, to get a laptop with IrDA to talk to a cellular handset with IrDA so the laptop can access the Internet. This problem has been carried forward into Bluetooth—devices that are deaf to each other because of software incompatibility or user incomprehension.

Plug-in Modules

If Bluetooth or wireless LAN cards are added as plug-in modules, care must be taken if the Bluetooth or wireless LAN card relies on the host device for power. This would seriously compromise a PDA running on two AA batteries.

Plug-in cards come in three thicknesses—Type I cards are 3.5 mm deep, Type II are 5 mm deep, and Type III are 10.5 mm deep—usually too deep to be compatible with most PDAs.

Present products being sampled include integrated wireless LAN/Bluetooth, wireless LAN cards from Intersil, and integrated GPRS and Bluetooth wireless cards—the example shown in Figure 15.1 is by Plextek Limited in the United Kingdom.

Local area connectivity can, of course, include wireless headsets. A common application is where a user has a handset connected to his or her belt with Bluetooth connectivity to an earpiece.

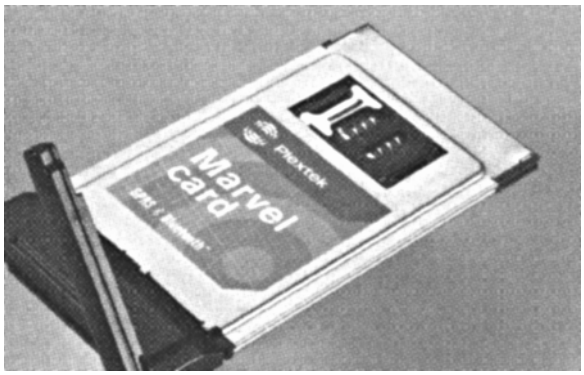


Figure 15.1 GPRS and Bluetooth plug-in card from Plextek Limited.

Going cordless has its advantages and disadvantages. The advantages are fairly obvious: no cord to get tangled. The disadvantages include the need to worry about two sets of batteries (one in the handset, one in the headset) and RF and application layer compatibility between different vendors. The jury is still out as to whether consistent headset-to-handset RF and application layer connectivity can be achieved.

Headsets are, however, an example of personal networks in which various devices are carried on the body—or sewn into clothing—with each device having the capability to communicate or fulfill a particular function—for example, heart-rate monitoring. Heart-rate information may have medical value (astronauts in space), competitive value (knowing how calm a racing driver is), or entertainment value (the pulse rate of a football or baseball player after they have scored).

A Network within a Network within a Network

In a perfect world, devices would roam seamlessly between Bluetooth, IEEE, picocells, microcells, and macrocells (see Figure 15.2). Picocells, microcells, and macrocells work together very well but use substantial signaling bandwidth to achieve the happy state of effective communication. Getting cellular handsets coordinated to work with wireless LANs and Bluetooth is technically difficult and probably impossible in terms of standards integration, because too many different vendor interests are involved. This is the concept of *data clouds*—the ability to log on seamlessly to a public access or private access wireless LAN and upload or download or exchange high-quality audio and video, either using IEEE802 or Bluetooth.

In Europe and Asia, Bluetooth devices are being added to cellular handsets, but, with the exception of earbud applications and PDA synchronization, they are not widely used. Public access wireless LANs have not achieved widespread market adoption to date in European and Asian markets because of technical constraints (power limitations and problems with co-existence with Bluetooth) and regulatory reasons. Cellular operators are only marginally interested, since they already have saturated coverage and adequate capability in hot spot areas and will greatly add to that capacity with the introduction of Node B 5 MHz transceivers.

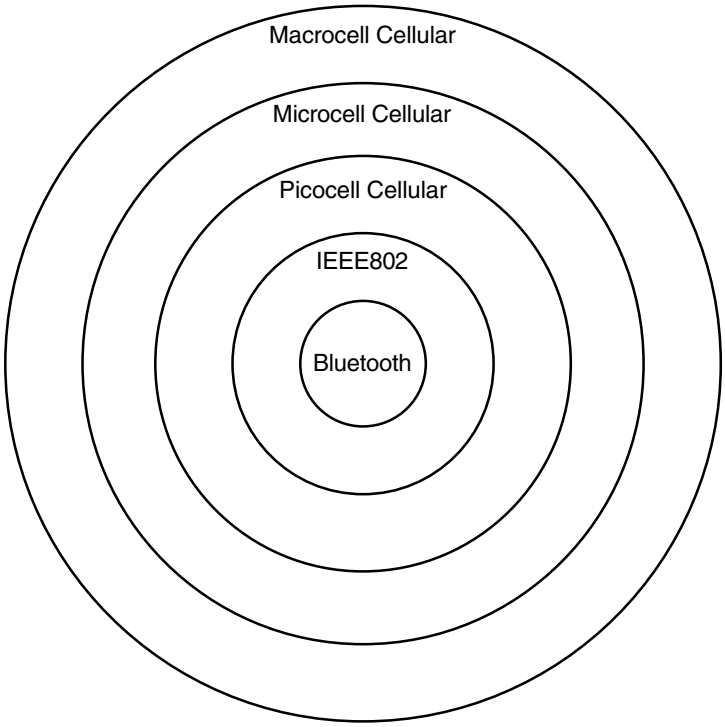


Figure 15.2 A network within a network within a network.

In the United States, public access wireless LAN networks may achieve a measure more success, but it is difficult to imagine wireless LANs becoming pervasive on a worldwide basis for technical, commercial, and political reasons.

Low-Power Radio and Telemetry Products

Although there is presently considerable market focus on 2.4 GHz ISM-based networks, there are many radio channel allocations at lower frequencies that support radio telemetry and telecommand applications. The RF channel spacing may be either 12.5 kHz, 20 kHz, or 25 kHz, and the allocations can be anywhere between 100 MHz and 860 MHz.

Products tend to be divided into very low power (1 to 10 mW) SAW filter or crystal-based devices, quite low power (500 mW), or reasonably high power (2 W). Very low power devices are used for applications where very low power consumption is needed—devices installed on pipes underground, for example, which need to be remotely interrogated occasionally from the surface (sleeper meters).

Typically the products can be configured to work at VHF (125 to 180 MHz), UHF (400 to 500 MHz), and in the band allocated within Europe between 868 and 870 MHz.

Range and data throughput are dependent on power output and receive sensitivity. Table 15.3 shows a typical product specification.

Table 15.3 Typical Telemetry Product Specification

Modulation type	GMSK
Frequency range	VHF 125–180 MHz or UHF 400–500 MHz and 868–870 MHz
Channel spacing	12.5 kHz/20 kHz/25 kHz available
RS232 Baud input rate	300, 600, 1200, 2400, 4800, 9600,19,200 and 38,400 (configurable)
RF power output	500 mW (high); 5 mW (low)
Type approval	ETS 300 113, MPT1329 and ETS 300 339 (EMC)

Source: www.woodanddouglas.co.uk

One of the problems with the telemetry and telecommand market is that spectrum allocations differ from region to region (the United States, as always, is different), and this makes it difficult to achieve economy of scale in terms of RF hardware. In terms of software, some commonality is emerging in that IP protocols are being used increasingly. The IETF Simple Network Management Protocol (SNMP) helps improved application transparency. Software code for a utility meter reading application in the United States, for example, does not need to be completely rewritten for a utility meter reading application in Europe.

VHF and UHF telemetry is used very successfully in motor racing to monitor and manage engine performance and driver performance (our previous heart-rate example). This is a very high mobility (over 200 mph) application and also can be quite broadband, because there is a lot of engine information to capture. It is actually easier to deliver broadband to a fixed rather than moving object, as we will see in our next section.

Broadband Fixed-Access Network Hardware Evolution

Over the past 5 years, spectrum has been allocated on a country-by-country basis at 3.5 GHz, 10 GHz, 26 GHz, 28 GHz, 38 GHz, 39 GHz, and 40 GHz for broadband fixed wireless access. Available bandwidth increases as frequency increases, but propagation becomes more unpredictable. Over 10 GHz, communication effectively has to be line of site.

The link budget also becomes very dependent on climatic conditions. As you move into the millimeter bands (30 GHz to 300 GHz), the wavelength is the same length as a raindrop (or a hydrometeor to use the correct description). Raindrops in effect scatter the radio waves. This is known as *nonresonant absorption*.

Weather Attenuation Peaks

There are also some attenuation peaks caused by water vapor and oxygen resonance effects. This is called *resonant absorption*. Attenuation peaks for water occur at 22 and 183 GHz. Attenuation peaks for oxygen occur at 60 and 119 GHz. Figure 15.3 shows these attenuation characteristics. Note that the 60-GHz oxygen line exhibits a propagation loss of 15 dB per kilometer. This is good news and bad news. You can get very high reuse ratios at 60 GHz, but you have to accommodate the propagation loss in the link budget.

Broadband fixed-access networks avoid these resonant absorption lines, but they do all suffer from weather effects. The wetter the climate, the more fading margin needs to be taken into account in the link budget.

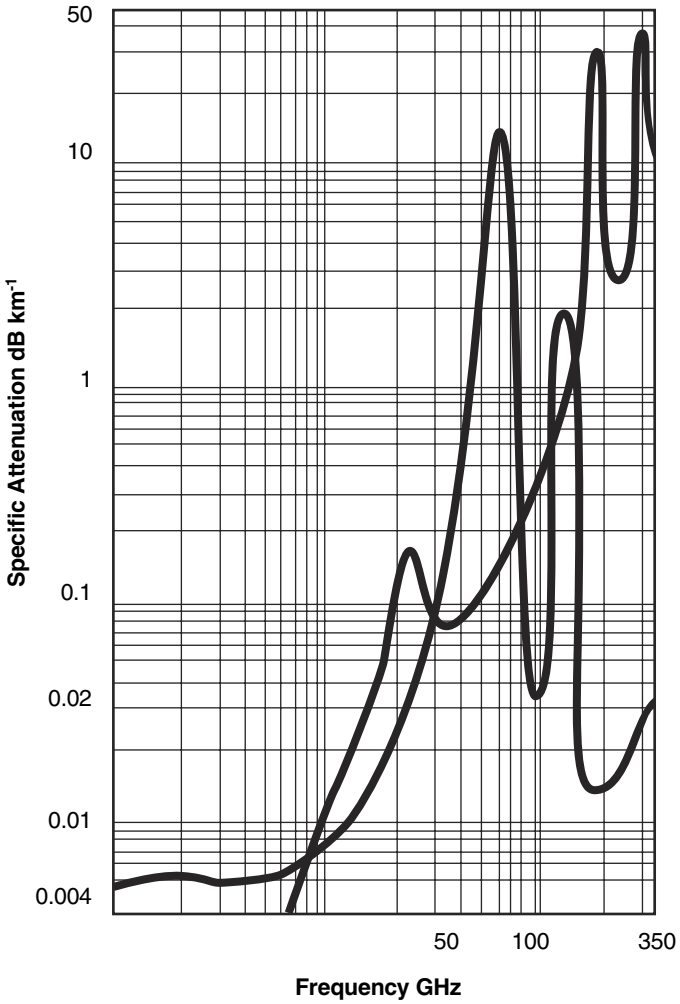


Figure 15.3 Millimetric attenuation characteristics.
With thanks to Rutherford Appleton Laboratory, Didcot, Oxfordshire, England.

Table 15.4 Lucent On-Demand Wireless Access System Specifications for 10 GHz (ITU Version)

RANGE AVAILABILITY (TYPICAL)	7 MHZ CHANNEL	14 MHZ CHANNEL
CCIR climatic Zone K, 99.99% availability	13.1 km	10.3 km
CCIR climatic Zone N, 99.99% availability	7.7 km	6.5 km

The impact of weather can be observed on wideband fixed-access performance specifications, as demonstrated in Table 15.4. Zone K is rather dry (Arizona); Zone N is rather wet. Note also in the table, which is based on a Lucent product implemented at 10 GHz, how the range reduces with the wider band channel (a wider noise floor). Availability is specified as 99.99 percent (four nines availability) rather than the five nines availability of wireline copper access. To deliver five nines availability would require a higher link budget. Range would reduce further.

As frequency increases, weather effects become more pronounced. In Table 15.5—a Lucent product implemented at 26 GHz—some additional gain is provided by a separately mounted (nonintegral) antenna.

Table 15.5 Lucent On-Demand 26 GHz—ITU

RANGE AVAILABILITY (TYPICAL)	7 MHZ CHANNEL	14 MHZ CHANNEL
CCIR climatic Zone E, 99.9% availability		
Integral antenna	3.8 km	3.3 km
390 mm nonintegral antenna	5.4 km	4.6 km
CCIR climatic Zone K, 99.99% availability		
Integral antenna	2.6 km	2.3 km
390 mm nonintegral antenna	3.4 km	3.0 km
CCIR climatic Zone N, 99.99% availability		
Integral antenna	1.5 km	1.3 km
390 mm nonintegral antenna	1.9 km	1.7 km

Table 15.6 Lucent On-Demand 38 GHz FCC

Modulation format	Hub to Customer Terminal	4 QAM, 16 QAM, FDM/TDM
	Customer Terminal to Hub	4 QAM, 16 QAM, FDM/TDMA
Receive sensitivity—Typical at BER = 1×10^6 after FEC	4 QAM, 12.5 MHz	-84 dBm
	16 QAM, 12.5 MHz	-76 dBm
System capacity net throughput per radio channel	4 QAM, 12.5 MHz	13 Mbps
	16 QAM, 12.5 MHz	26 Mbps

Range is also dependent on the modulation used. (Higher-level modulation techniques absorb more power for the same demodulator bit error performance.) Tables 15.6 and 15.7—a Lucent product implemented at 38 GHz—shows how moving from 4-level QAM to 16-level QAM requires an additional 8 dB of link budget, a doubling of the power and some implementation loss. The benefit here is taken as capacity gain rather than range gain.

As frequency increases, coverage becomes increasingly line of site dependent. The benefits of a nonintegral antenna (more directivity) become more significant. Note also the antennas become more compact as frequency increases.

A number of vendors are also promoting OFDM/OFDM solutions for broadband fixed access. Cisco, Broadcom, and Pace have a product that uses OFDM within a 6 MHz channel to give a duplex 20 Mbps traffic stream or 40 Mbps in a 12 MHz channel. Note the commonality with digital TV multicarrier OFDM implementation in Europe and Asia.

Table 15.7 Lucent On-Demand 38 GHz FCC

RANGE AVAILABILITY (TYPICAL)	7 MHZ CHANNEL	14 MHZ CHANNEL
CCIR climatic Zone E, 99.99% availability		
Integral	2.4 km	1.4 km
300 mm nonintegral	3.2 km	2.1 km
CCIR climatic Zone K, 99.99% availability		
Integral	1.6 km	1.0 km
300 mm nonintegral	2.1 km	1.4 km

Mesh Networks

Because deployment is very line site sensitive, it is worthwhile to consider adaptive antennas that can search for the strongest signal and adapt to new signal paths as network density increases. These are known as *mesh networks*. Some meshed networks are claimed to be able to support 1 in 10^{12} bit error rates and 99.999 (five nines) availability.

Meshed networks (see Figure 15.4) depend on users acting as repeaters This results in a distributed communications network.

When a transceiver is installed, the antenna searches for the strongest signal within a 360° radius and then locks to that signal. If a heavy rainstorm disrupts the signal, the antenna can look in another direction to reestablish an alternative path—adaptive fixed-access bandwidth. Figure 15.5 is an adaptive antenna array designed for Radiant Networks. The antenna array is motor driven and physically rotates to establish links with other users.

Fixed-Access Wireless Access Systems

In common with wide area mobility networks, fixed-access wireless networks need to be able to handle bursty bandwidth. This means the same issues of protocol performance apply, albeit without the added complication of a mobility management overlay. Users in a fixed-access network, by definition, stay in the same place. Fixed-access networks can either be deployed as a number of dumb fat pipes or a larger number of smart thin pipes—for example, using ATM.

Fixed point-to-point hardware is already widely deployed in existing terrestrial cellular networks, predominantly 38 GHz point-to-point links between cell sites or between cell sites and BSC or RNC or MSC switch nodes. Similar hardware is used to provide links in digital TV transmission and TV distribution networks. Theoretically this should deliver some economics of scale and common deployment experience.

Figure 15.4 Example of mesh deployment.
Picture courtesy of Radiant Networks (www.radiantnetworks.com).

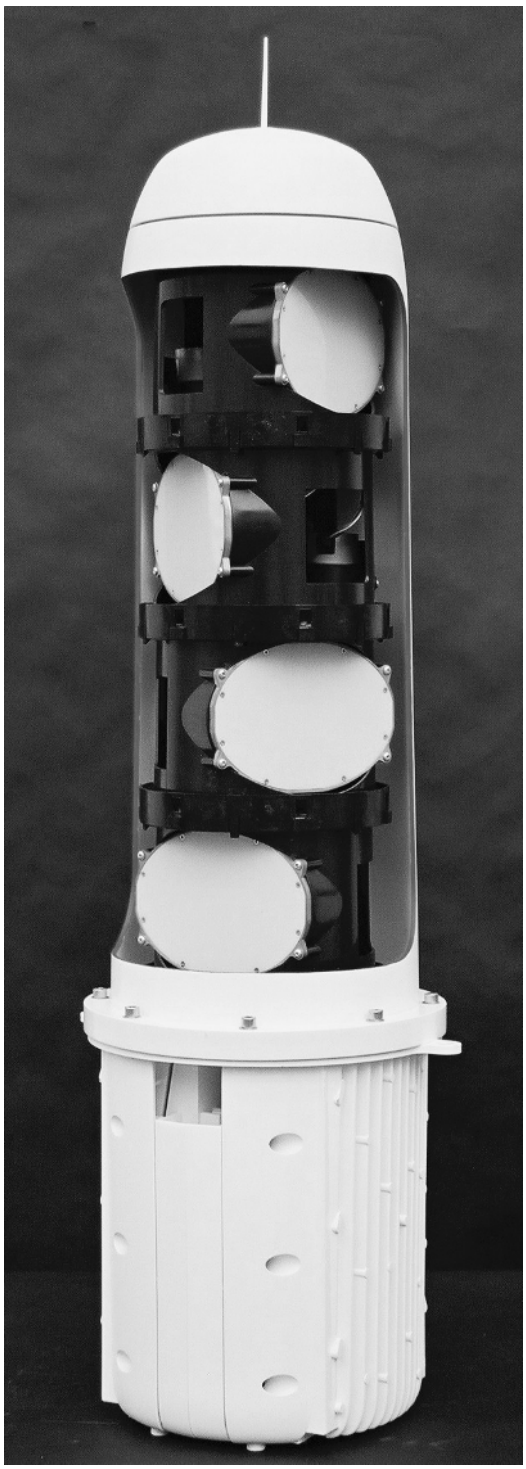


Figure 15.5 Adaptive antennas for fixed access.
Picture courtesy of Radiant Networks (www.radiantnetworks.com).

In practice, there are so many different flavors of different hardware at different frequencies that this has tended to prevent widespread deployment of fixed-access radio as a substitute for wireline access. If a vendor has too many products to choose from, he or she often makes a choice not to choose any of them. The problem has been compounded by a confused and disparate fixed-access wireless standards-making process.

The fragmentation of the market in terms of technology and the fact that so many different frequencies have been allocated in so many different countries make it hard to realize economy of sale when manufacturing radio transceivers. RF components are still quite expensive above 10 GHz, and it is difficult to achieve consistent RF performance between units leaving a production line.

In parallel, wireline access performance is improving (with techniques like VDSL), and wireline access costs are reducing. Wireline networks have often already been fully amortized over many years.

The fact that fixed wireless access systems are still, even if only to a small degree, weather dependent also militates against their widespread adoption.

Alternative Fixed-Access and Mobility Access Wireless Delivery Platforms

As we move up in frequency, particularly above 10 GHz, radio waves behave in a very similar way to light. We showed in Chapter 11 that it makes a lot of sense to do indoor RF calculations, using similar design techniques to those used in lighting design. It also potentially makes sense to use outdoor lighting platforms—in other words, street lamps—as an alternative method for delivering RF bandwidth. Figure 15.6, taken from a Philips Lighting catalogue, shows the photometric data from a street lamp, giving the distribution of light intensity in isocandelas. You will notice this is very similar to antenna coverage patterns used in RF system planning.

A number of attempts have been made to integrate base stations and antennas into street furniture—some more successful than others. One company specializing in this area is Stealth Network Technologies (www.stealthsite.com). Base station and network hardware is concealed in flag poles, silos, water towers, or billboards, typically using RF-transparent materials to hide the hardware.

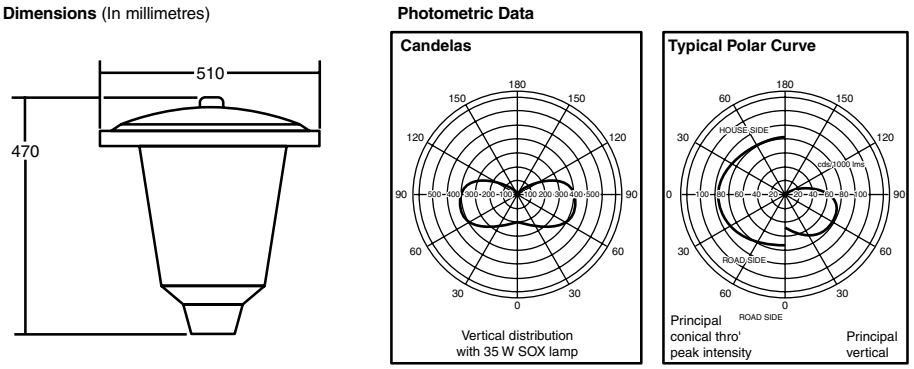


Figure 15.6 Photometric data from a Philips street lamp.

The NIMBY Factor

One reason for hiding base stations and antennas is that if people don't know that the hardware is there, they cannot complain about it—the "not in my back yard," or NIMBY factor. Although it is well beyond the scope of this book to discuss exposure limits, it is an area of increasing—and possibly legitimate—public concern.

There are a number of software packages now available that calculate emission levels at particular elevations and distances from base stations, given site variables such as type, height, bearing, antenna tilt, power output, feeder and system losses, and number of RF carriers. The software calculates surrounding RF emission levels from the base station and produces an elevation plot at a number of user-specific points—giving the field strength, for example, in nearby high-rise buildings (displayed in polar or rectangular format). An example product can be found on Link Microtek's Web site, www.linkmicrotek.com.

Setting the Stage for Satellite

An alternative method of hiding RF and switch hardware, and avoiding the NIMBY factor, is to put the hardware up in the sky or into space. There have been a number of proposals and demonstration projects showing the possibility of mounting transceivers in weather balloons or unpiloted high-flying aircraft (typically at 100,000 feet or so), although as yet, none of these schemes have been taken through to full-scale deployment. Deployment of systems in space is, however, well established.

Satellite Networks

In the 1980s, first-generation cellular systems generally provided quite poor geographic coverage. It varied substantially from country to country and from operator to operator, but typically you might experience demographic coverage of 70 percent to 80 percent of the population, which would equate to 50 percent to 60 percent geographic coverage. The difference between demographic and geographic coverage is that, conveniently, people tend to live in concentrated urban areas. Only a few people live in deserts or on remote mountaintops.

If people wanted coverage in remote rural areas, then they would need to use a mobile rather than a handset. The mobile had more transmit power available. Also, if installed in a car, with a gain antenna, the mobile would have better transmit and receive performance.

However, what people really wanted was ubiquitous coverage, even if using standard handsets. One way to provide ubiquitous coverage, including coverage of mountains and largely uninhabited rural areas was to use satellites.

Early Efforts

Motorola proposed what turned out to be a rather overambitious project known as Iridium. Other vendor consortia, which included companies like Hughes and Raytheon with space sector experience, proposed similar projects. After a number of years, spectrum was allocated, and after a number of years, systems were built.

In the meantime, terrestrial networks—first and second generation—had extended their geographic coverage in many countries to 80 percent, 90 percent, or in some countries, over 90 percent, and by implication, greatly reduced the geographic addressable market for satellite-based services.

In urban areas, satellite-based systems did not work very well because of building blocking. You needed to go outside to make your phone work. The solution was to produce dual-mode phones giving satellite access and cellular access, but this made the phones rather large and expensive.

These were the main technical reasons why the satellite-based systems (initially Iridium and Globalstar) failed commercially. There were also many marketing and business reasons for their failure, but these are outside the scope of this book.

However, many valuable lessons have been learned from the deployment experience to date. It is quite rational to expect that satellite-based systems will play a significant role in fourth-generation cellular service provision, so it is well worth reviewing present and likely future technology options.

Present and Future Options

The United States sent up its first rocket in 1926. In 1945, Arthur C Clarke wrote about a wireless network of extraterrestrial relays. In 1957, the Russians launched Sputnik, and the space race began. In 1962, Telstar, the first commercial telecommunications satellite was launched followed by Intelsat in 1965, hailed by President Johnson as "a milestone in the history of communications between people and nations." Then came the Apollo missions, the continuation of substantial military spending on the space sector in the 1980s and 1990s, and the development of reusable launch capability (the shuttle).

By the 1990s, it was technically (if not, as a subsequently proved, commercially) feasible to implement a satellite-based system providing an uplink and downlink with sufficient link budget to support users with portable handsets.

The choice of orbits includes the following:

VLEOs. Very low Earth orbits, typically at 350 km (similar to the space shuttle/space station). The problem with very low Earth orbits is that there is some residual atmospheric drag that shortens the life of the satellite. Also there is quite a lot of space debris floating around.

LEOs. Low Earth orbits, typically at 700 km. LEOs were used (are used, as the satellites are still in space) for Iridium.

MEOs. Medium Earth orbits at 10,000 km.

HEOs. High Earth orbits at 30,000 km.

Suborbital solutions. Available at about 100,000 feet (18 miles), either using weather balloons or high altitude pilotless planes.

The higher the orbit, the fewer satellites you need to provide coverage but the bigger the satellites need to be. Bigger because they need to have more RF power for the downlink, since they are covering a larger geographic footprint. They also need more processing power for the uplink and downlink, because they are, hopefully, supporting more users.

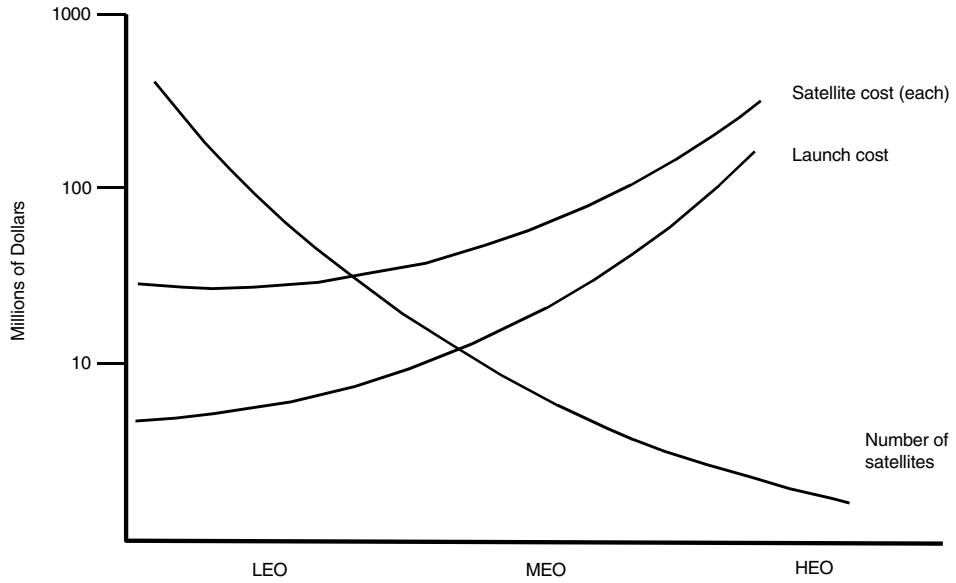


Figure 15.7 Satellites for mobile—cost trade-offs.

Figure 15.7 shows a typical cost trade-off. The higher the orbit, the fewer satellites you need, but the satellites get bigger and more expensive. This graphic was produced by Globalstar, which uses an orbit at 1410 km (an orbit known as the Walker orbit). The graphic shows the optimum cost trade-off is somewhere between a LEO and a MEO once launch costs are taken into account. Conveniently, this is where the Globalstar satellites are positioned.

Note that orbits are a mathematical construct. Attempts are being made in the United States and Russia to obtain patent rights for existing and possible future orbits, including some of the highly elliptical options like Molyna and Tundra. Apart from needing fewer satellites, higher orbits are cleaner, since there is less space debris above 2000 km.

Some of the elliptical orbits pass through the Van Allen belt, which means that devices need to be radiation hardened, further adding to their cost. Higher orbits also introduce additional transmission delay—20 ms for a LEO, 133 ms for a MEO, and 500 ms for a HEO.

Iridium

The Iridium system uses 66 satellites, and each satellite has 48 beams. The satellites take 100 minutes to go around the Earth and are in view to an individual user for about 10 minutes. There are six equally spaced orbits going north to south, with 11 operational satellites in each orbit. Originally there were seven orbits making up 77 satellites, hence the name Iridium (the 77th element).

Actually, there are 74 active satellites in orbit (8 in orbit spares). Each satellite provides a 48-beam cell coverage on Earth with each cell having a nominal 700 km diameter. The

satellite's beam pattern is fixed, and users are handed from beam to beam within the satellite, and then between satellites as the satellite passes over.

Each satellite weighs 700 kg and supports a 530 W payload and 250 W service link. The service links are at 1610 to 1626 MHz; the gateway links are at 19 GHz (downlink) and 29 GHz (uplink). The satellites can also communicate directly with each other at 23 GHz. The satellites were designed to be launched from the United States (Delta), Russia (Proton), or China (Long March) launch vehicles.

Not everything went quite to plan with Iridium. Additional costs included a launch failure when a Delta 2 rocket exploded at Cape Canaveral. Also, the handsets needed to be subsidized, because they were quite expensive to manufacture (and rather big to hold).

Globalstar

Next up is Globalstar. Being a high LEO (or low MEO), Globalstar only needs 48 satellites. The Walker orbit results in eight orbit planes with six satellites per plane. Each satellite has 16 spot beams and weighs 350 kgs, supporting a 550 W payload. Globalstar uses a CDMA air interface (Iridium is TDMA). The uplink is at 1.6 GHz; the downlink is at 2.4 GHz. Coverage is good up to about a latitude of 70°. It does not provide polar coverage.

Also proposed, but not yet deployed, is a MEO system known as an ICO/MEO system (ICO stands for Intermediate Circular Orbit). The ICO/MEO would be at 10,400 kilometers.

ORBCOMM

ORBCOMM has offered limited commercial services in the United States since 1996 and services in Europe from 1999. It is a VHF 137–138/148–149.9 MHz system for sending and receiving short data messages (229 characters). Most of the user terminals include a GPS capability.

Inmarsat

Inmarsat was established in 1979 by international treaty but is now a privatized entity. There are four second-generation and five third-generation satellites all placed in a geostationary orbit. The Inmarsat4 program known as Broadband Global Area Network (B-GAN) is due to be launched in 2004. There will be two Inmarsat4 satellites offering a 72 kbps transmit and 216 kbps receive facility for pocket-size palmtop terminals, and 144 kbps transmit and 432 kbps receive service for laptop-sized terminals. Inmarsat and Iridium are both widely used by the military, particularly for non-mission-critical applications (for example, logistics).

Calculating the Costs

The lack of commercial success of satellite-based systems for mobility access to date does not mean the technology does not have a future role to play. Indeed, now that Iridium's launch costs and development costs have been written off, it will probably make a profit (the wonders of fiscal engineering).

Table 15.8 IMT2000 Frequency Plan

TDD1 1900-1920	1920-1980	1980-2010 SATELLITE	TDD2 2010-2025	2110-2170	2170-2200 SATELLITE
4 × 5 MHz Nonpaired IMT2000 TDD1	12 × 5 MHz IMT2000DS	Mobile (Handset) Uplink	3 × 5 MHz Non-paired IMT2000 TDD2	12 × 5 MHz Paired IMT2000DS	Mobile (Handset) Downlink

30 MHz of uplink bandwidth and 50 MHz of downlink spectrum has been allocated in the IMT2000 frequency plan for mobile satellite systems (MSSs), as shown in Table 15.8.

As processor costs reduce, it becomes possible to deliver substantial coding gain, both on the uplink and downlink. This makes it easier to achieve an economic uplink and downlink link budget, which in turn makes the economies of mobile satellite systems more favorable in the longer term.

Provided the same air interface is used, then subscriber handset costs do not need to be substantially higher. The spectrum is proximate to the IMT2000DS paired bands, so the RF implementation of the handsets can be relatively straightforward.

Next-generation shuttle launch technologies promise to reduce launch costs (and the satellites are becoming smaller and lighter anyway). Note how important it is to deliver linearity and power efficiency on the satellite, as this dictates its traffic-carrying and hence revenue-producing capacity. Improvements in smart antenna technologies also promise significant link budget gains (sensitivity and selectivity gain), which will help to improve uplink performance, reducing power budget drain in the handset.

The experience gained by Iridium in intersatellite switching also promises substantial longer-term network performance benefits. Delay budgets for LEO systems are really quite acceptable, and very predictable and controllable. We rest our case for an MSS technology renaissance at some time in the next 3 to 5 years.

Satellites for Fixed Access

In parallel to mobile access provision there have been a number of proposals to implement fixed-access service including broadband service provision. The commercial argument is that it is expensive to install fiber to remote wireline subscribers and that satellite could provide a viable commercial and technical alternative. This will depend on how much fiber costs reduce over the next 3 to 5 years and the success (or possible lack of success) of other terrestrially based wireless access options.

An example of a fiber bypass is the Teledesic broadband LEO. This proposal uses 288 satellites at 1300 kilometers arranged in 12 planes (24 satellites per plane). The uplink is at 28.6 to 29.1 GHz and the downlink at 18.3 to 19.3 GHz. Using large numbers of satellites means that users can view a satellite at a relatively high elevation angle. This avoids the problem of low-elevation angles where the radio path through the atmosphere is relatively long and subject to rain attenuation.

The Teledesic proposal is based on an ATM switch fabric. Because it is designed for fixed access, the user is expected to use a dish receive and transmit antenna. The larger the dish, the higher the data rate. Proposed services include 16 kbps voice, 64 kbps video, and 2 Mbps multimedia with a bit error rate of 1 in 10^{10}. Transmission delay introduced by the space segment should be less than 20 ms.

Downlink delivery is, of course, already well established technically and commercially through the provision of satellite TV. It is the uplink that is tricky to deliver. However, we have argued that much of the added value in next-generation networks is subscriber-generated uplink value. Digital TV lacks uplink bandwidth.

The technical opportunity may therefore be in the deployment of OFDM-based satellite systems that can provide broadband downlink and uplink service to fixed-access and mobility users. The commercial opportunity will be the capturing of subscriber-generated content, the archiving of subscriber-generated content, and the redelivery of that content to the originating subscriber and other interested parties.

Bandwidth quality is dependent on the delivery of a robust link budget. This in turn is dependent on the maturation of enabling technologies such as smart antennas and efficient linear RF PAs (to increase downlink capacity) and good low-noise amplifier performance (to improve uplink capacity). The use of the 2 GHz IMT2000 bands will limit the impact of weather on the link and should result in a reasonably consistent user experience, even for mobility users (particularly for mobility users in rural areas where building blocking will not generally be a problem).

The ability to provide coverage even in very remote areas will confer an advantage—witness the value of present satellite telephones in recent international conflicts. Almost by default, we will expect to be able to upload and download complex content wherever we are, and satellites will play a part in the delivery and storage of rich media products over the next 3 to 5 years.

Summary

In this chapter we set out to differentiate wide area access connectivity and local area connectivity, showing how local access can provide better bandwidth utilization, provided that capabilities such as handover and power control are not required. We drew attention to some of the device hardware compatibility issues when deploying local area scatternets. This, in turn, can deliver a rather inconsistent user access experience. In other words, the problems of device compatibility are directly related to network performance.

We reviewed a rather overlooked part of the industry—telemetry and telecommand, using low-power radio—to remind us that VHF and UHF frequency allocations still exist and can deliver good functionality.

We then looked at the wireless technology options for broadband fixed access. These technologies provide an alternative to cable systems and can provide the basis for the delivery of TV content and other broadband content. However, particularly at frequencies over 10 GHz, care must be taken to provide line of sight coverage, and we need to take weather effects into account in the link budget.

We argued the merits/demerits of using satellites to provide mobility access and fixed access, and we pointed out that success here was dependent not only on traditional technology engineering but also required some imaginative fiscal engineering. There is a clear argument in favor of integrating satellite-based mobility service provision with terrestrial radio service provision at some time in the future.

Over the five chapters in this part of the book, we described the role of network hardware in determining service quality—the user experience. In the next five chapters, we go on to describe the role of network software in delivering service quality and service value.

PART

Four

3G Network Software

CHAPTER 16

The Traffic Mix Shift

In earlier chapters we showed how handset hardware is changing the traffic mix in a wireless network. The bandwidth of the CMOS imager, audio codec, and MPEG-4 encoder determines the uplink offered traffic mix; the bandwidth of the display, display driver, and speaker determine the downlink offered traffic mix. In earlier chapters we showed how handset software is changing the traffic mix in a wireless network—how session persistency and session complexity increase over time, and how session complexity increases as session persistency increases. This means the longer the session, the more complex you can make it.

The Job of Software

In essence, the job of the software in the handset is to prompt the user to use the handset. For example, if the user has taken pictures or a short video with the in-built camera, then the software prompts the user to send the pictures either to other people or other places—for example, to a virtual storage site. The software provides the option of choosing 6-bit, 8-bit, or 24-bit color depth, providing a choice of resolution. For video, a choice of frame rate is provided and, for audio, a choice of narrowband or wideband (high-fidelity) sound. If the user has a buddy list, this provides the option to send the video and audio clip to multiple destinations. If a number of these recipients reply, there is an opportunity to build a complex real-time exchange where a number of people are supported in a conference session (not a conference *call*).

The user group could be commentating on video capture being provided by one of the members of the group or several members of the group. The longer the session can be made to last, the more complex it becomes. Session value increases as session length (session persistency) increases. Complex exchanges involving complex content require complex admission control. Users may join, leave, or rejoin a user group as a session progresses. Users may have equal access rights or users may be given priority. In addition, users may need to be authenticated and user traffic may need to be encrypted.

At the start of a session, someone has to decide on the traffic properties (the traffic class), as follows:

- Is the exchange conversational? In this case no buffer bandwidth is allowed or required.
- Is the session streamed? In this case some buffering will be needed.
- Is the session interactive? In this case uplink and downlink buffering will need to be provided and controlled.

Critical Performance Metrics

For conversational traffic, end-to-end delay is a critical performance metric. For streamed or interactive traffic, delay and delay variability are critical performance metrics. Delay variability is a product of buffering and protocol-induced delay—that is, the triggering of send-again protocols when packet loss occurs (usually as a result of buffer overflow).

Conversational, streamed, and interactive traffic needs careful session management. The only part of the traffic mix that is truly fire-and-forget is background traffic, which, by definition, has no specified minimum delay requirement. However, even background traffic value can be destroyed by excessive delay—a misrouted message that takes several days to arrive. A message that never arrives has had all its value destroyed.

Radio Bandwidth Quality

We have discussed how radio bandwidth quality is related to service quality and the preservation of content and session value. A poor-quality radio channel can induce packet loss or produce discontinuous traffic, which can compromise application and content integrity.

Radio bandwidth quality is determined by how well the handset works in RF terms and how well the base station hardware performs. We can also improve radio bandwidth quality by increasing network density. Either way, we add cost. We could keep the same number of base stations and improve their performance by adding smart antennas, low-noise RF components, and high Q filters, or we could keep the base stations as they are but have more of them (which improves the link budget). Denser networks, however, absorb more signaling overhead and are therefore less efficient.

In the next chapter we review how network bandwidth quality is related to service quality. We have described in previous chapters how we can become code-limited—that is, not have enough OVSF codes to support our multiple-channel per-user multiplex. Similarly, we find we can become protocol-limited, which means we may have network bandwidth available but are unable to allocate and access the bandwidth when we need it.

The Performance of Protocols

The performance of protocols when presented with highly asynchronous traffic is particularly important. The radio physical layer can support a change of data rate every 10 milliseconds. Imagine an application in a handset. The application can request a dynamic rate matched channel, which can vary between 15 kbps and 960 kbps every frame (every 10 ms). The uplink can request additional supplemental channels. (The theoretical limit set by the 3GPP1 standard is six simultaneous uplink channels, any of which could be changing data rate on a frame-by-frame basis.) This determines the shape or property of the uplink offered traffic.

In practice, the baseband processing and RF hardware limitations of the handset substantially limit the ability to utilize this potential uplink code bandwidth. You would also quickly become code-limited at the base station (which only has a limited number of OVSF codes available). Nevertheless, it would not be unreasonable to expect offered traffic rates on a per-user basis to be varying between 9.6 kbps and 384 kbps spread across one or more OVSF code channels. It would also not be unreasonable to expect a similar per-user loading on the downlink.

Network Resource Allocation

The per-user multiplex at the physical layer is therefore complex and dynamic. Note also that we may have multiple users all engaged in the simultaneous exchange of complex content, which requires the dynamic allocation of radio bandwidth and network bandwidth resources. Network resource allocation has to be sufficiently fast and flexible to adapt to rapidly changing user application requirements. If the network response is too slow, by the time resources are allocated, the application requirement will have changed. Buffering can be used to reduce the immediacy needed, but this introduces delay and delay variability, which may compromise application value.

Consider some of the decisions that need to be made:

- A request comes from the handset application layer for more bandwidth. The decision must be made as to whether enough bandwidth is available:
  - Radio bandwidth allocation is determined on the basis of interference measurements from the radio physical layer.
  - Network bandwidth allocation is determined on the basis of congestion measurements within the network.
- The decision may also need to be tempered by knowledge of particular hardware and software constraints in the receiving device—how to keep fast senders from swamping slow receivers.

Service Parameters

The service requirements may be described as *service primitives*. A primitive defines an action to be taken in response to a request or an indicator requiring a response and confirmation. Primitives may and usually do have parameters:

- Upper and lower bounds
- Minimum and maximum bit rates
- Number of channels per user
- Maximum session length

These parameters must be negotiated and then confirmed or unconfirmed. The decision may need to include the choice of connectionless or connection-oriented delivery bandwidth. In a connectionless exchange, each message will carry the overhead of a full destination address and will or may be independently routed through the network. When two messages are sent, the first one sent is usually the first to arrive, but not always. In a connection-oriented exchange; a connection is established; the connection is used; and the connection is cleared down (that is, a circuit-switched transaction or circuit-switched exchange). Note that both connectionless and connection-oriented exchanges can be characterized by quality of service.

A decision may or may not involve a process of negotiation. The decision can be taken theoretically by the sender, the receiver, or the network in the middle. Network operators always prefer to make the decision in the network. The information needed for the decision must be captured and delivered to the point where the decision is to be made. The decision then has to be distributed to the interested and relevant parties.

Power Control and Handover

Power control and handover is an example of the previously described process in action. A measurement report is compiled by the handset and includes, for example, received signal strength, bit error rate, and frame erasure (how many frames exceed a specified bit error rate). This information is used by the BSC, or in a third-generation network, the RNC, to decide whether to increase or decrease the power output of the handset. If the handset is working at or close to its maximum power, the BSC or RNC needs to decide whether to move the user to another base station and which of the target handover base stations to choose.

Note that in a first-generation network, these decisions were typically taken in the Mobile Switch Center (MSC). As networks became denser with more base stations, it became harder to manage the handover process and it became difficult to control the dropped call rate. Hence, in second-generation networks, the decision typically was distributed out to the BSC, which would look after its particular subgroup of base stations. The MSC in turn looked after the BSCs, which means some of the decision making had been decentralized.

The Evolution of Network Signaling

Note how signaling bandwidth needed to increase from first generation to second generation. In a first-generation network, signaling over the radio layer was accomplished using short bursts of 8 kbps or 10 kbps data transfers (blank and burst signaling). In second generation, each multiframe has a frame dedicated to collecting and sending radio channel measurements—the slow associated control channel. The network takes this information and passes back the decision to the handset using a fast associated control channel—generally a traffic channel reused as a signaling channel. In the network, control channel information is moved across the A-bis interface (3 kbps within each 16 kbps channel) and then aggregated into multiple 64 kbps channels using ISDN.

Second-Generation Signaling

Let's just review second-generation network signaling using a GSM-MAP network as an example. There are three signaling protocols:

- LAPD-M looks after the signaling between the mobile (handset) and base station.
- LAP-D looks after signaling between the base station and base station controller (and is adapted from ISDN).
- MTP looks after communication between the base station controller and the mobile switch center and uses SS7 (Signaling System 7).

Figure 16.1 shows how the signaling planes are divided into user-facing functions, operator-facing functions, and external network functions. The operating subsystem looks toward the operator and is responsible for managing traffic, along with the hardware and software needed to manage traffic flowing into and out of the network.

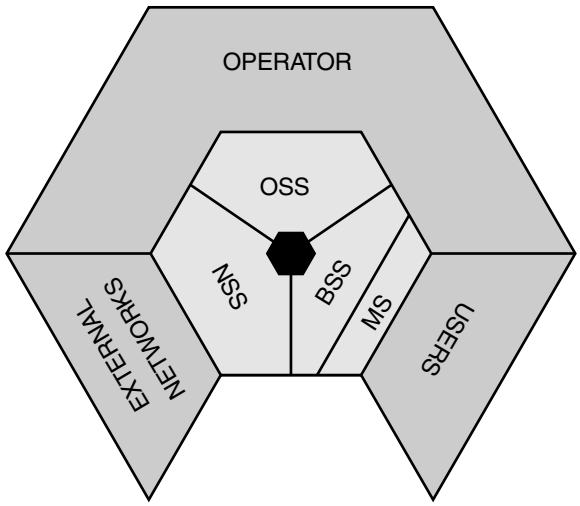


Figure 16.1 Subsystem organization.

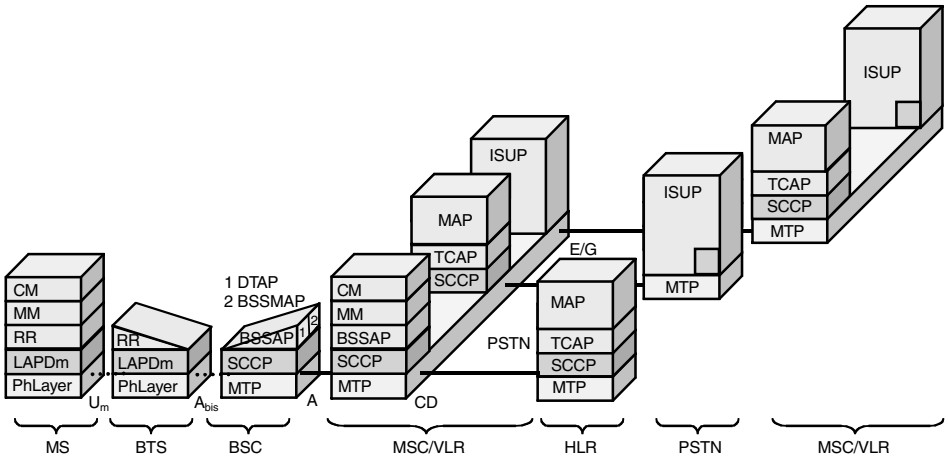


Figure 16.2 GSM protocol stacks.

The network and switching subsystem bridges the link between the operator's network and other external networks. (Remember: We have to manage roaming, so we need a mechanism for supporting users as they move out of the home network to other networks—that is, macro mobility management.)

The network and switching subsystem must be able to communicate with the home location register and visitor location register, that is, globally manage mobile users. The base station subsystem and the mobile subsystem look toward the users and manage physical radio resource allocation.

Figure 16.2 shows the protocol interfaces, which are exercised as traffic moves into the network. Call maintenance (CM) looks after communication management—call setup, call maintenance, call clear-down. This is really quite simple in a second-generation wireless network. A call is set up, maintained, and cleared down. The only complication introduced by the radio physical layer is that the call might be dropped because of some problem with the radio channel (the user disappears into a tunnel with no RF coverage) and the call may have inconsistent quality. Other than that, CM is the same whether it's a wireless or wireline connection.

Third-Generation Signaling

Consider, however, how much more complex CM becomes in a third-generation network. We set up a session. The session may need just one physical channel, but we still have to decide on the parameters available for that channel—maximum bit rate, minimum bit rate, required bit error rate, and a non-isochronous or isochronous bit stream. The application layer then requests a second channel. We need to decide on the parameters available for that channel. We now have two channels active for the single user. Note also: If we are using dynamic rate matching, either of these channels can change rate every 10 ms.

Theoretically we may need to support up to six channels per user, and any one channel could be continuously changing its bandwidth quantity and quality metrics. Call maintenance (session maintenance) becomes a far more complex process. For instance, we may also need to be supporting multiple users all co-sharing the same user group channel, all individually needing more or less bandwidth as the session progresses. We then need to clear down the session and find some way of describing the session activity (the session persistency and complexity metrics), so that the network can bill the participants.

Protocol Stack Arrangement

Figure 16.3 shows how the protocol stack is arranged in an existing (2G) handset. Consider that in a 3G network, CM effectively is replaced with a session management (SM) protocol that establishes session properties at the start of a session and adapts to any changes in session property needed as the session progresses.

In a 2G network, the mobility management protocol takes the user information from the SIM (international mobile subscriber identity number) and uses that information to negotiate access to the network, which may be the host network or a visited network—that is, user authentication (AUC) via the Home Location Register (HLR) or Visitor Location Register (VLR).

In a 3G network, the U-SIM holds much more information on the user and the user's right of access:

- Is the user entitled to use multiple channels?
- Does the user have a right of access to virtual storage?
- Does the user have right of access to particular user groups?

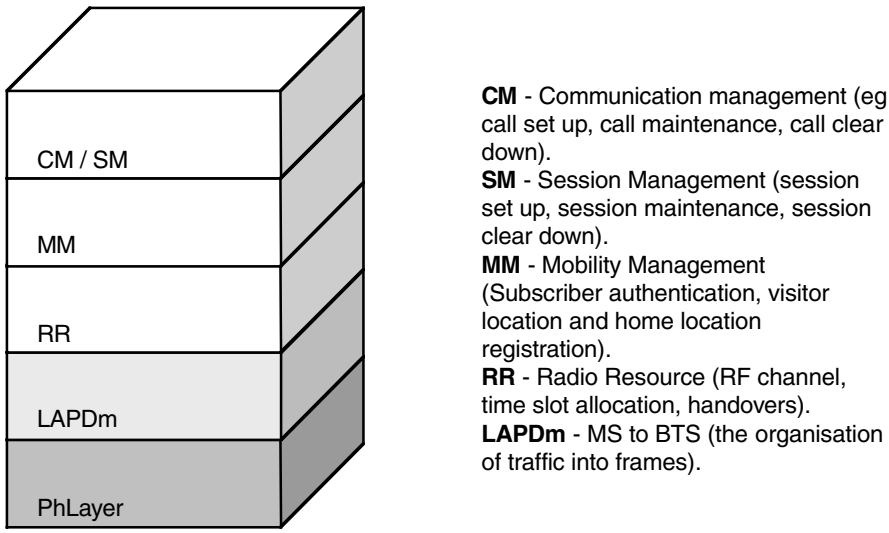


Figure 16.3 Mobile station protocol stack.

In addition, session properties will be dependent on the user's hardware and software form factor and functionality:

- Does the user have a high color depth high-resolution display?
- Does the user have a wideband audio coder, CMOS imager, and MPEG4 encoder?
- Does the user have the right to use these capabilities?
- Does the user have end-to-end encryption?
- Does the user have the right to use end-to-end encryption?
- Has the user deposited a key with the trusted third party to provide for lawful interception?
- Does the user have the right to make micro-payments (small purchases)?
- Does the user have the right to make macro-payments (large payments)?

Load Distribution

In a 2G network, radio resource (RR) allocation looks after time slot allocation and RF channel allocation. When a handset accesses a network, it will be authenticated and told which base stations to monitor (a serving base station and up to five handover candidates). If the mobile requests a call, a time slot will be allocated within an RF channel (actually a duplexed spaced RF channel pair for the uplink and downlink). If the network needs to move the handset to another time slot, or to another RF channel, or to another base station, this will be achieved by using the fast associated control channel.

In a 3G network, the RNC will also be looking after load distribution to manage the highly dynamic changes in offered traffic. Load distribution includes the need to move handsets into and out of soft handover involving, as described in earlier chapters, substantial communication between serving and drift RNCs. Radio resource allocation becomes much more complex and much more dynamic and may change continuously as a session progresses.

For example, a change in radio resource allocation in a 2G network occurs because of a handset moving from one cell to another while a call is in progress. This will also happen in a 3G network, but in addition, as a session progresses, OVSF code channel allocation may change every 10 ms on existing allocated channels, and new channels may be added or subtracted from the user's physical layer multimedia multiplex.

The allocation decision may be driven by the application requirement, but whether or not resources are allocated will be driven by whether the network has radio, transmission, and storage resources available, and whether, even if these resources are available, the network wants to make those resources available. It may be that the decision is made to reserve resources in case more important users need them, which means the decision-making process becomes highly dynamic and adaptive, and becomes dependent on multiple technical, operational, and commercial decision criteria.

Just because a user or the user's application has requested a certain allocation of bandwidth and just because a user has a right to a certain allocation of bandwidth does not ensure that the user or application will receive that allocation, either at the beginning of the session or as the session progresses. For instance:

- What happens if the network delivers the application requirement for 90 percent of the session but fails to deliver the application requirement for 10 percent of the session, does the user get a 10 percent rebate?
- How does the user or the application know that the bandwidth allocated has only matched the bandwidth requested for a certain percentage of the session?
- How does the user or the application prove that the performance delivered was not the same as the performance requested; does the network bill for the performance requested or the performance delivered? (Can you guess which one is more likely?)

Somehow this complex performance capture process has to be undertaken (preferably in the user's device), and the failure to perform has to be communicated back to the network to provide the basis for rebate-based billing. We are adding technical and commercial complexity, which, in turn, results in additional signaling complexity.

3G Frame Structure

Referring to Figure 16.3, the last part of the protocol stack in the handset is the LAP Dm protocol that looks after framing between the mobile and the base station—or in GSM, 8 slots in a frame, 26 or 52 frames in a multiframe or multiple multiframes concatenated into superframes, and hyperframes. We described the 3G frame structure in earlier chapters (that is, commonality with GSM at multiframe level). The frame structure in the 3G air interface is actually simpler. The complexity comes from the need to support GSM/IMT2000 intersystem internetwork handover, which requires the handset to be able to compile management reports from both networks and then send those reports back (probably) to the MSC for an internetwork handover decision to be made.

2G Versus 3G Session Management

Figure 16.4 shows the protocol stack in a 2G base station. The base station is in essence a dumb modem that just moves traffic into the network and out from the network to its local flock of handsets. Power control and handover decisions are made by the BSC.

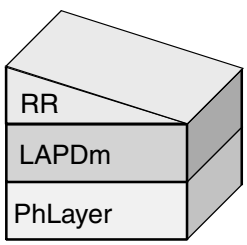


Figure 16.4 Base station protocol stack.

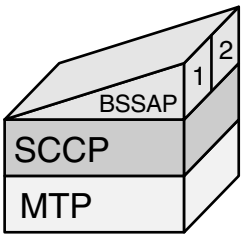
In a 3G or Node B, dynamic OVSF code management needs to be supported in addition to power control and handover management (radio resource allocation). LAPD-M will need to support GSM-to-IMT2000 handovers.

Potentially, the Node B could be the point where packet stream prioritization is undertaken, although present thinking suggests that prioritization and load balancing will be done at the RNC. This is an important philosophical point. In present cellular networks, call/session status information—power levels, serving base stations, bit error rate, and so on—is moved across the same physical channels as our traffic (for example, voice), but it is kept logically separate in its own time slot or its own frame (or both).

Similarly, in a 3G network, status information and measurements share the same physical RF channel but are kept logically separate. As we hit the network, this logical separation between traffic and signaling is maintained. We move the status and management information to the place in the network where a decision has to be communicated back to the device that originated the status and measurement information and to other interested devices (neighboring Node Bs, for example).

If the handset had sufficiently intimate and up-to-date knowledge of the network resources available, it could decide for itself on the most economic/fastest route to be taken through the network and could use that knowledge to describe the required routing trajectory in a packet header. The job of the Node B would be to read the packet header and move the packet stream on into the network on the basis of the packet header instruction. All the signaling effectively stays in band, which means it physically stays with the traffic, even though it may be kept logically separate.

This is not what happens in present wireless networks. Status and management information from the radio layer is extracted by the base station controller and sent on the SS7 signaling path to the MSC. Figure 16.5 shows how signaling is presently arranged in a base station controller.



1 DTAP
2 BSSMAP

Figure 16.5 BSC protocol stack.

The parts are as follows:

- DTAP (Data Transfer Application Part) looks outward to the user and the user's device (the handset).
- BSSMAP (Base Station Subsystem Mobile Application Part) looks after the management of the base station serving the user and other handover candidate neighbor base stations.
- SCCP (Signaling Connection Control Part) looks after the management of traffic channels and routers, that is, the physical routing of offered traffic.
- MTP (Mobile Transaction Part) looks after the commercial needs of the offered traffic, determining the following:
  - Is the user authenticated?
  - Does the user have particular rights of access?
  - Has the user roamed from another network?

There is a lot of network control involved here. This implies substantial signaling load, which costs money and introduces delay. Network operators do, however, like to feel in control of their traffic, and this may be regarded as a necessary unavoidable overhead.

Figure 16.6 shows how all this signaling finally ends up in the mobile switch center. The parts are as follows:

- TCAP (Transaction Capability Application Part) looks after the commercial properties of the traffic—the right of access, the right to be billed for a certain service/quality of service.
- MAP (Mobile Application Part) looks after the housekeeping needed to keep track of a mobile user—the mobility management. Note this is going on the whole time your phone is turned on, whether or not you are making a call. The phone is continuously measuring its serving base station and making the measurement report, which is fed back to the BSC and MSC, so that the MSC knows where the mobile is physically (so traffic can be delivered from and to the mobile) and so that it knows the quality of the radio link presently maintained.
- The ISUP (ISDN User Part) looks after the call setup, call maintenance, call clear-down—that is, ISDN-based call management.

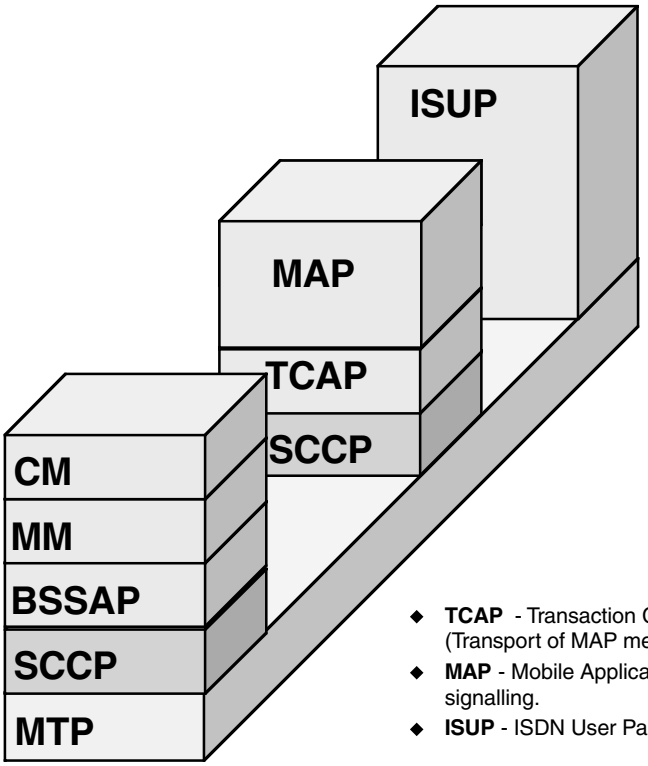


Figure 16.6 MSC/VLR.

In a 3G network, all these MSC-based functions become substantially more complex. In a 2G network, TCAP just has to decide whether a user has a right of access to a call—a voice circuit. In a 3G network, TCAP has to decide on the traffic properties and hence the required session properties needed for the user:

- Does the network have sufficient resources available?
- Can these resources continue to be made available as the session progresses?
- Are additional participants permitted to join the session, and what are their required session properties?

The MAP part in a 3G network needs to comprehend traffic shedding and load distribution in the radio access network, as well as the mechanics of managing mobile users in a complex network. Because the network is a mixture of GSM and IMT2000, MAP must be able to capture and act on intersystem radio bearer measurements.

ISUP has to look after session-related signaling (as opposed to call-related signaling). ISUP in a wireless network has always been a bit more complicated than a

wireline network, because the addition of a radio channel in the communications link means that a call may be terminated unexpectedly or may become unacceptably noisy.

Similarly in session management, the session may terminate or become unacceptably noisy. What has changed, however, is that session properties are now far more dynamic. A call is a call. A voice circuit is established, maintained and cleared down, and billed.

A session is a chameleon, changing its color as the session progresses—becoming more complex, less complex. The application bandwidth and required quality metrics are constantly changing. ISUP has to respond to that constantly changing bandwidth quality requirement, provision the bandwidth, and then qualify whether the bandwidth requirement was or was not fulfilled for the duration of the session. Note also that the signaling has to communicate with other interested or involved networks.

Communications between Networks

Figure 16.7 shows the ISUP link to the Public Switched Telephone Network (PSTN). Consider that there are now between three and six operators in most countries and roughly 100 countries (or separate political entities with their own national communication networks), so there are over 500 cellular networks that have to be capable of talking to each other and to their respective wireline networks. In a 2G network, the information in the HLR and VLR is reasonably simple—a description of the user (SIM identity) and the user's hardware (stored in the Equipment Identity Register).

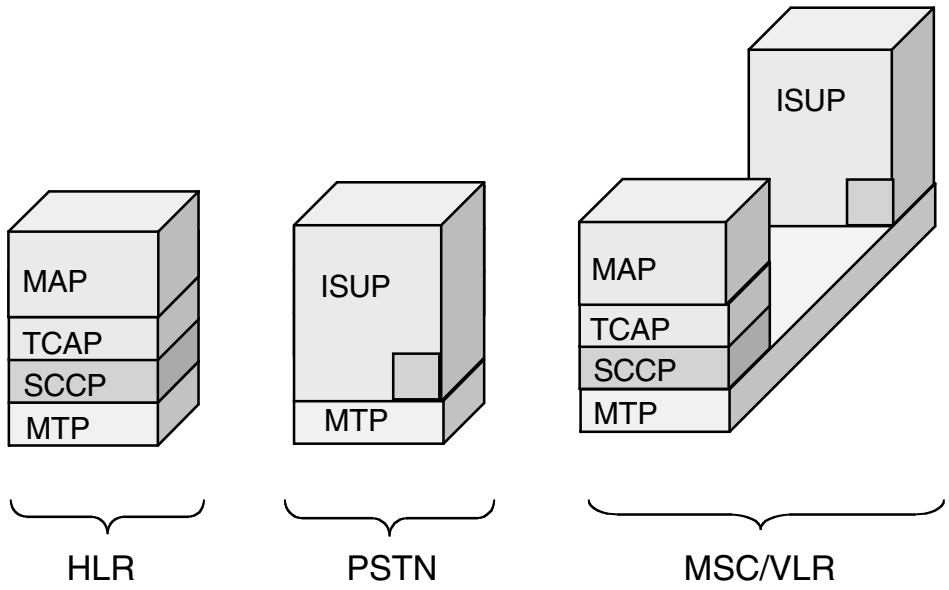


Figure 16.7 HLR/PSTN/MSC/VLR.

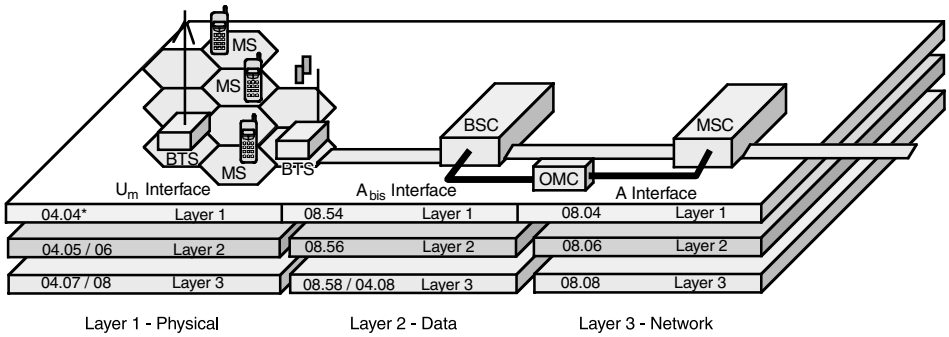


Figure 16.8 Layer modeling.

In a 3G network, the information is far more complex—the user's service profile, the user's device hardware and software form factor and functionality, the user's access rights to delivery and storage bandwidth, associated buddy groups and user group configuration, and on and on. Figure 16.8 summarizes the present signaling structure in a wireless network.

The components are as follows:

- The UM interface looks after the link between mobiles and the base station.
- The A-bis interface looks after the link between the base station and base station controller.
- The A interface looks after the link between the base station controller and the MSC.
- The OMC (Operation and Maintenance Center) tracks any hardware failures occurring in the network (and from other related networks) and looks after software-based service platforms.

Why We Need Signaling

Let's summarize why we need signaling:

- Signaling allows network software to make intelligent decisions on the basis of information received from various parts of the network (or other related networks).
- Signaling has a very direct impact on radio and network bandwidth quality. If signaling is poorly implemented, either the wrong decision will be made or the decision will be made too slowly.
- Signaling overhead increases as networks become more complex. Networks become more complex as network density increases. If base stations are closer together and the mobility of users stays the same, then the network will have to manage more handovers.

- If we try to get more performance out of our radio physical layer, we often find our signaling load increases. In a 3G network, we employ soft handover to give us uplink and downlink diversity gain. This substantially increases signaling bandwidth at and between the serving and drift RNC.
- As session complexity increases, signaling complexity increases. If our traffic properties are constantly changing, our signaling has to respond.
- Signaling introduces delay and delay variability. (Remember: We are trying to support static rate matching where channels are added and subtracted on a per-user basis as a session progresses, and we are trying to support dynamic rate matching where channels can change their offered traffic properties—bit rate and channel coding—every 10 ms.

Moving Beyond the Switch

Historically, much of the decision making in a cellular network has been centralized in the switch—20 million lines of software code providing centralized control. This has historically worked well, because the offered traffic has been very deterministic—that is, highly predictable. A request for a call is received with a destination address; we set up the call, maintain the call, clear down the call, and bill the call. Call setup, call maintenance, call clear-down, and call billing is easy. Session setup, session maintenance, session clear-down, and session billing is not so simple.

A mobile switch center behaves very predictably when loaded with predictable traffic, particularly if the traffic is constant rate and constant quality. However, we have said that our per-user traffic mix is now neither constant rate nor constant quality. The quality requirements of our traffic will change dynamically as our traffic mix changes and our session progresses. The object of the whole exercise, however, is to give the user the feeling that he or she is getting a constant-quality user experience, which means at any given moment, we have to provide sufficient radio and network bandwidth quantity and quality such that we avoid compromising the user's application quality (and by implication, application value).

As we discussed in earlier chapters, this may mean adding in extra channel coding, changing modulation state, increasing power on the radio channel, or, in the network, providing a particular routing trajectory to deliver a defined end-to-end delay and delay variability metric. Our network software, therefore, has to be capable of taking highly dynamic decisions on the basis of highly dynamic information often received from multiple destinations. These are decisions that rely on feedback control.

Feedback processes work best. They are fastest in decentralized distributed networks. There is no point in taking very time-dependent decisions on, for example, radio resource allocation, if the signaling path introduces more delay than the decision window can accommodate.

Letting the Handset Make the Decisions

The ultimate logic is to move the decision making into the handset. The handset intuitively knows the properties needed from the radio physical layer and network in

order to preserve the application properties being generated by the handset. It is, or should be, possible for the handset to discover available radio and network resources and then to negotiate access to those resources.

It should also be possible to express this bandwidth quality requirement in a packet header. The base station/Node B reads the packet header and obeys the routing instruction contained in the header. The routing decision has already been pre-agreed, which means decision making has been distributed out to the user's device.

In theory, this makes complete sense. In practice, most network operators would be very unwilling to live with the idea of allowing a subscriber or subscriber's application to determine routing options or right of access to delivery or storage bandwidth.

Dealing with SS7 and Existing Switching Architectures

In addition, existing signaling protocols (SS7) and switching architectures (circuit switching) have been tested and refined over many years. They provide very certain end-to-end delivery conditions. In a circuit-switched network, we know what the end-to-end delay will be, and we know what the delay variability will be—very little. This makes it much easier to send a bill to someone and avoids the problem of that user complaining that he or she has not received what he or she was expecting.

The SS7 signaling provides us with a transparent view of a call as it progresses. This is partly because the signaling plane is logically separate from the traffic. Effectively, the signaling is looking down at the traffic below. If we were to embed the signaling at packet header level, we would arguably lose this transparency.

However, if SS7 signaling is to be used to support session management, and if session properties will be constantly changing, this implies very substantial signaling load on the network. This means considerable signaling bandwidth will be absorbed, resulting in occupied bandwidth, which is expensive—the cost of signaling overhead.

Additionally, if the signaling path is physically long, there will be too much hysteresis. The signaling responses will be too slow to adapt to the changing loading conditions—the network software implications of managing audio, image, video, and application streaming—which means the signaling path will introduce delay. If the decisions being taken are complex, the signaling will also introduce delay variability.

Signaling is therefore an important ingredient in network bandwidth quality and session consistency.

Making a Choice

We then have a choice: We can throw away SS7 signaling and circuit switching or adapt SS7 signaling and circuit switching to accommodate complex sessions. As session persistency increases, it becomes increasingly attractive to keep SS7. The longer the session, the more economic out-of-band signaling becomes—and the less attractive in-band packet header-based signaling becomes.

The problem is more the issue of whether to retain the circuit-switching capabilities of existing networks. We have said that circuit switching provides very dependable, very predictable performance, which, when combined with SS7 signaling, provides a very robust basis for billing. It is, however, relatively inflexible and nonoptimum for dealing with highly asynchronous and constantly varying traffic. A circuit switch is

essentially a hardware platform with a relatively small amount of software (20 million lines of code). It is fast and efficient, but it is not particularly flexible and not particularly adaptive.

An option is to distribute the hardware functionality closer to the edge of the network. This reduces the signaling load on the network, reduces the signaling delay, increases the speed of decision making, and makes the hardware architecture more adaptive to local loading conditions. This is the thought process behind 3GPP1's inclusion of ATM cell switching as a mechanism for managing asynchronous traffic loading in the radio access network.

Note the difference: In a pure play packet routed network, the handset would define routing priorities needed to support the locally generated application. This would be expressed in the packet header. The software in the router in the Node B and RNC would interpret the packet header and route the packet stream as required and described by the packet header.

In an ATM implementation, routing is pre-agreed, probably for the whole session. Any decisions to be taken while a session progresses will be, typically, whether to add or subtract channels for an individual user or change the properties of these channels. The routing stays the same, and the switching is hardware- rather than software-enabled. This means that this is a distributed hardware switched architecture rather than a distributed software routed architecture—an ATM cell switched network rather than an IP routed network.

It will be less flexible than an IP routed network (which relies on software-based decision making to make decisions based on information contained in individual packet headers), but it will be more predictable in terms of the way it behaves when presented with complex content. This means it is a better solution for complex session management. ATM is however not particularly well suited to supporting complex multiuser-to-multiuser exchanges, particularly if feedback is used to manage access control (available bit rate).

We could, of course, use IP protocols for session management and mobility management, as well as network management, but we need to be sure we can replicate all of the functionality presently available to us from existing signaling solutions. We need to find a precise way of qualifying and quantifying protocol performance, particularly traffic shaping protocol performance.

Summary

As offered traffic becomes more asynchronous, it places more demands on the signaling in the network. The signaling needs to become more responsive. This means we have to qualify the hysteresis implicit in the feedback decision-making processes involved, including the time taken to execute a decision. This is very analogous to some of the design considerations we have at chip level where we have to qualify memory fetch routines and housekeeping routines in a device's microcontroller and memory architecture. At network level, similar decisions are being taken—for example, on bandwidth allocation and access prioritization—but the distances are much greater (sometimes many miles).

It is therefore essential to qualify protocol performance at network level—the subject of our next chapter.

CHAPTER

17

Traffic Shaping Protocols

In the previous chapter, we discussed the fact that circuit switching was designed to move essentially constant rate traffic from users at one side of the network to users at the other side of the network (or networks). In this chapter we look at how session value is developed on the basis of session persistency and how this, in turn, has an impact on traffic shaping protocols. We examine the hardware and software performance constraints in a typical router—how router hardware and software introduces delay and delay variability when presented with highly asynchronous traffic. We also identify the important issue of policy complexity. How policy complexity is increasing over time and why this presents a challenge to Internet Protocol performance.

An Overview of Circuit Switching

Circuit switching works on a number of well-defined and relatively constant rule sets. A call request comes in with a destination (a telephone number), and a path is established between sender and receiver (actually two paths: an uplink and downlink); the path is then maintained for the duration of the call and then cleared down and billed. There is no buffering involved, so the end-to-end performance is deterministic and consistent over time. The queuing behavior of voice traffic is well understood, so circuit-switched capacity can be provisioned to reduce blocking to an effectively unnoticeable level—that is, five nines availability, meaning it is available 99.999 percent of the time.

The argument against circuit switching is that it is inefficient. In a voice call, we are talking for only about 35 percent of the time. The rest of the time we are pausing between words or are listening to what is being said in reply. It is therefore wasteful to dedicate a two-way channel pair for the duration of the call.

However, we need to consider two points:

- In a wireless network, we have to differentiate between logical and physical channels. Over the radio layer, if we use discontinuous transmission, we only apply RF power to the channel if we detect voice activity. Thus, although we are consuming logical channel bandwidth, we are not consuming physical channel bandwidth.
- We should consider how the offered traffic mix is changing.

The assumption is that a high percentage of traffic will be discontinuous—short little bursts of data with long periods of inactivity in between. Clearly, circuit switching is not particularly efficient as a mechanism for moving this traffic. However, we have said that the whole purpose of our application layer, both in the handset and the network, is to build session persistency and session complexity.

In other words, we are trying to move from a discontinuous duty cycle (in voice or data exchanges) to a continuous duty cycle.

Moving Toward a Continuous Duty Cycle

Revisiting the loading model introduced in an earlier chapter, we see that once a session has been established, we have continuous activity throughout the session through to session clear-down. We describe this as session switched. In reality it is a circuit-switched session or at least a hardware-switched session with the capability of supporting variable bit rate—a combination of circuit-switched and ATM-switched traffic.

Remember that we are trying to realize subscriber asset value through the delivery of deterministic services. Deterministic services need to be based on end-to-end latency guarantees. End-to-end latency guarantees are dependent on radio and network bandwidth quality. What we want to be able to do is provide quality of service guarantees against which we can bill. These guarantees have to address end-to-end performance. End-to-end performance can be compromised by access delay, network delay, and application delay. We need to make sure that our overall system can either describe those delay components or preferably manage and minimize the delay and delay variability, working as a mechanism for delivering end-to-end bandwidth quality and, hence, bandwidth value.

Deterministic Response to Asynchronous Traffic

In Chapter 8 we said that the IEEE definition of a real-time operating system was "a system that responds to external asynchronous events in a predictable amount of time." The equivalent definition for a wireless network is "a network that can provide throughput to asynchronous traffic in a predictable amount of time." In other words, it

Table 17.1 Circuit Switch to Packet Routing

Circuit switch	15% bandwidth utilization
ATM	50% bandwidth utilization
IP	85% bandwidth utilization

has a deterministic response to asynchronous offered traffic. We need to be able to determine the performance requested and we need to be able to determine performance delivered in order to support quality-based billing.

The general assumption has been that there will be more routing (software based) in next-generation networks and less switching (hardware based). Software-based routers give us lots of flexibility, but the cost of flexibility is delay and delay variability. Hardware-based switching gives us less flexibility but delivers better, more predictable—or deterministic—performance.

If the offered traffic is discontinuous, then IP-based packet routing can be demonstrated to show major efficiency gains over ATM or circuit switching (see Table 17.1). However, as we try and improve quality of service in a packet-routed network, bandwidth utilization reduces.

If we try and deliver exactly the same deterministic performance and availability that we get from circuit switching but using IP protocols, we will find that IP protocols are no more efficient than circuit switching.

As session persistency and session complexity increases, the need for deterministic performance increases. It is oversimplistic to say that Internet protocols deliver efficiency benefits. Sometimes they do; sometimes they don't.

Dealing with Delay

Efficiency depends on the performance criteria required. A PSTN end-to-end circuit switch introduces typically a 35-ms delay. This delay will be consistent and constant. Typical Internet delay can be 200 to 400 ms—sometimes less, sometimes more—which means the delay is variable. Delay and delay variability are intrinsic to the Internet and are the consequential cost of flexibility and resilience.

Delay in an IP-routed network is the composite delay introduced by packet capture, buffering, routing, and queuing. The components of this delay are as follows:

Packet capture. The time taken to process an entire packet before forwarding to the router.

Buffering. The delay caused by the need to smooth bursty offered traffic.

Routing. The time taken to check the header and routing table.

Queuing. The time spent by packets waiting in router buffers while routers deal with other packets—typically 20 to 30 ms.

We can reduce queuing delay down to 5 or 10 ms by introducing packet shaping protocols such as Diffserv and MPLS (which we case study later in this chapter). However, packet shaping/packet management protocols reduce bandwidth efficiency because they increase protocol overhead.

Deep Packet Examination

Consider that our software-based routers in a packet-routed network must deal with packets delivered from thousands of different devices. Each one of these devices could be sending multiple packet streams that require separate QoS treatment. Each device and each device packet stream may require authentication and may have particular access rights that have to be considered, since this is a policy-based network. A policy-based network depends on what is referred to in the Internet world as "deep packet examination." If we express the security context and access rights in the packet header, we have to read that part of the packet header, check the information against locally stored or remotely stored lookup tables, decide what to do with the packet and then, finally, send the packet on to its destination (or discard it or send it back). The role of a software router is not simple.

The growth of devices and the growth of destination addresses has already created ever-expanding lookup tables. Adding in security fields and priority fields makes things even more difficult. In particular it causes problems with packet lookup rates and packet lookup delay.

The general idea is that the router receives a packet, examines the destination address, determines the address of the next-hop router, and sends the packet on to the next hop, as shown in Figure 17.1. Performance is determined by the packet forwarding rate and the routing table size.

If we add in data security, it may be necessary to examine the entire packet header and do a multiple-table lookup, which introduces delay and delay variability—that is, classification delay.

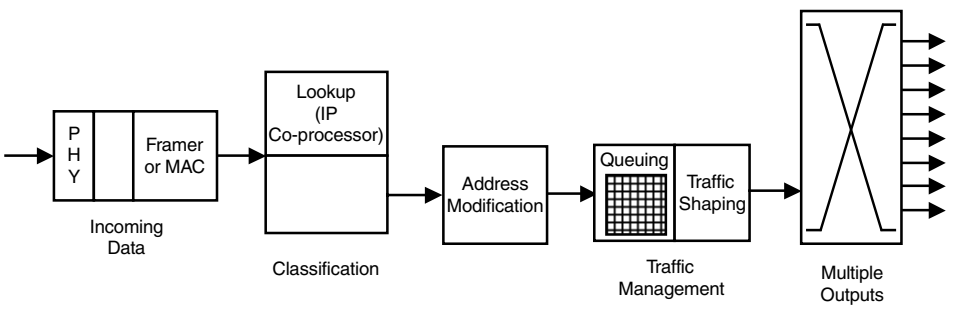


Figure 17.1 Router packet processing.

To try and minimize delay, the router divides packet processing into several tasks. The data comes in from the physical layer and is demultiplexed in accordance with the MAC (Medium Access Control) layer rule set—packets within frames within multiframes. The packet is then sent for classification. A hardware coprocessor may be used to improve lookup performance, which is defined by the number of searches per second, the number of entries in the table, and whether or not multiprotocol tables are used. Multiprotocol tables are needed if differentiated classes of service are supported.

A software-based standard processor-based solution can take several hundred instruction cycles to classify a packet with QoS or security attributes (or both). A coprocessor can perform the task in a single clock cycle but lacks flexibility. It can only be used when the decisions to be taken are largely predefined and repetitive, which describes hardware-based switching rather than software-based routing. Policy complexity is increasing over time as session complexity increases, and the need for packet-by-packet header examination decreases as session persistency increases, which helps to relax the loading on the processor.

Note, however, that as you move into the network, these routers need to cope with aggregate bit rates of 40 Gpbs or more (the bit rates typically coming down from the optical layer). Typically, coprocessors have to achieve anything up to 100 million table lookups per second. This can only be achieved by simplifying the rule sets determining traffic prioritization.

Address Modification and Queuing

Once the packets are classified, they are moved on for address modification and queuing. The user/policy tagged packet is then passed to a network processor that acts on the tagged packet instructions (that is, access and delivery priority).

The network processor may be a general-purpose RISC processor or an ASIC or multiple RISC processors used in parallel (although managing lots of parallel processors becomes quite tricky at higher bit rates, particularly if the traffic loading keeps changing). The traffic management function needs to sort out priority packets from best-effort packet streaming. If individual user quality of service guarantees are supported, the traffic management engine has to implement individual queues for each subscriber—a separate queue for each traffic flow.

In the core network, this implies that thousands of queues are required, which would be very hard to support on a software platform. The multiple queues then need multiple outputs, because the traffic may be going in different directions. The routing fabric needs to be integrated with the queuing control and traffic shaping to avoid buffer overflow and packet loss.

The device by now consists of several ICs—an IP coprocessor, an IP traffic management IC, and a specialized network processor. This is a hardware-based IP switch not a software-based IP router. As such, it can arguably match the performance of an ATM

or circuit-switched solution, but don't expect it to be either more efficient or more flexible, or necessarily lower cost, than existing options. The underlying challenge is getting predictable performance from an IP routed network.

Packet Loss and Latency Peaks

If we fail to match the routing fabric (egress bandwidth) with the offered traffic (ingress bandwidth), then we will suffer buffer overflow. This, in turn, causes packet loss. Figure 17.2 is from NetTest (www.nettest.com), a test equipment and network validation vendor. It shows some real-world measurements made in a GPRS network (not identified for obvious reasons) in Q4 2001. The figure shows how repetitive packet loss is triggered probably because of buffer overflow. Note the periodic nature of the effect.

Similar periodic impairments show up when latency is measured (see Figure 17.3). Although average latency is about 200 ms, the latency peak is over 3 seconds and occurs every 40 seconds or so.

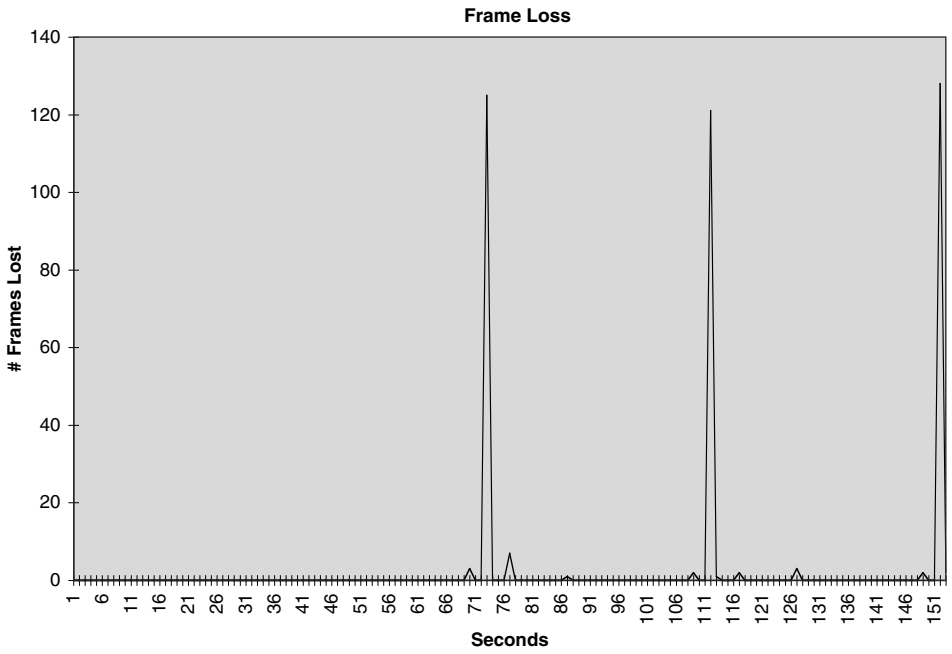


Figure 17.2 Packet loss results.

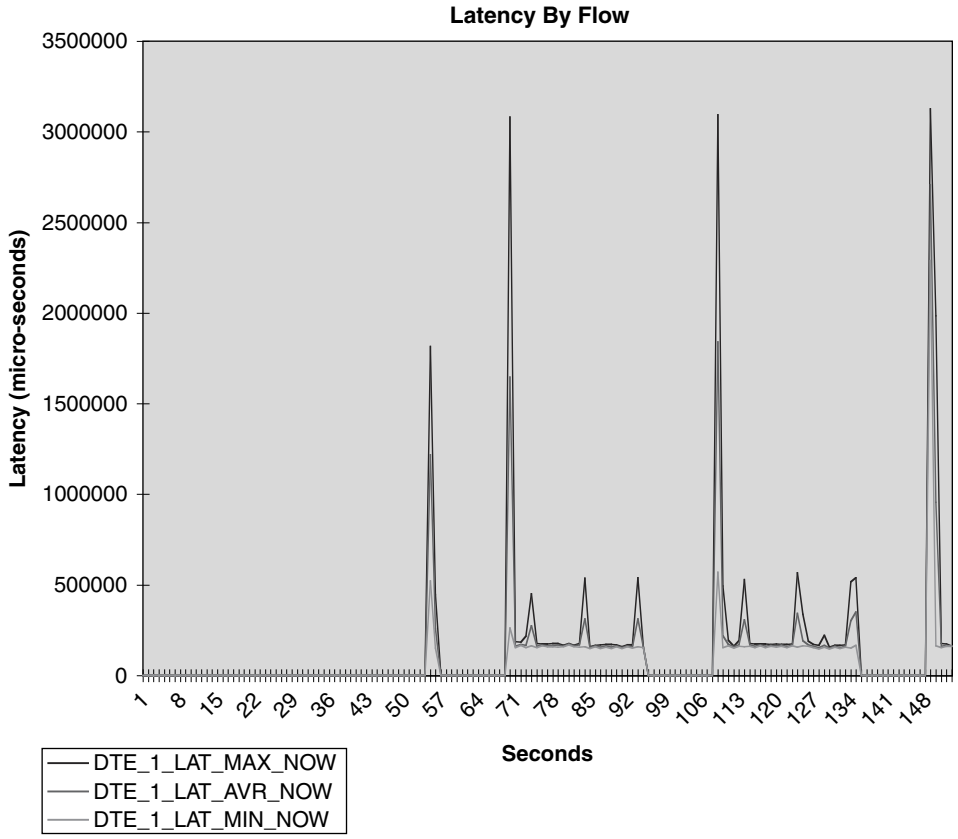


Figure 17.3 Latency results.

The reason for the packet loss and latency peaks becomes clear when you look at the network loading. Network loading, as illustrated in Figure 17.4, is peaking at the same time interval as the packet loss and latency peaks (every 40 seconds or so). The network is effectively being put into compression by the offered traffic. The symptom of compression is packet loss and packet delay.

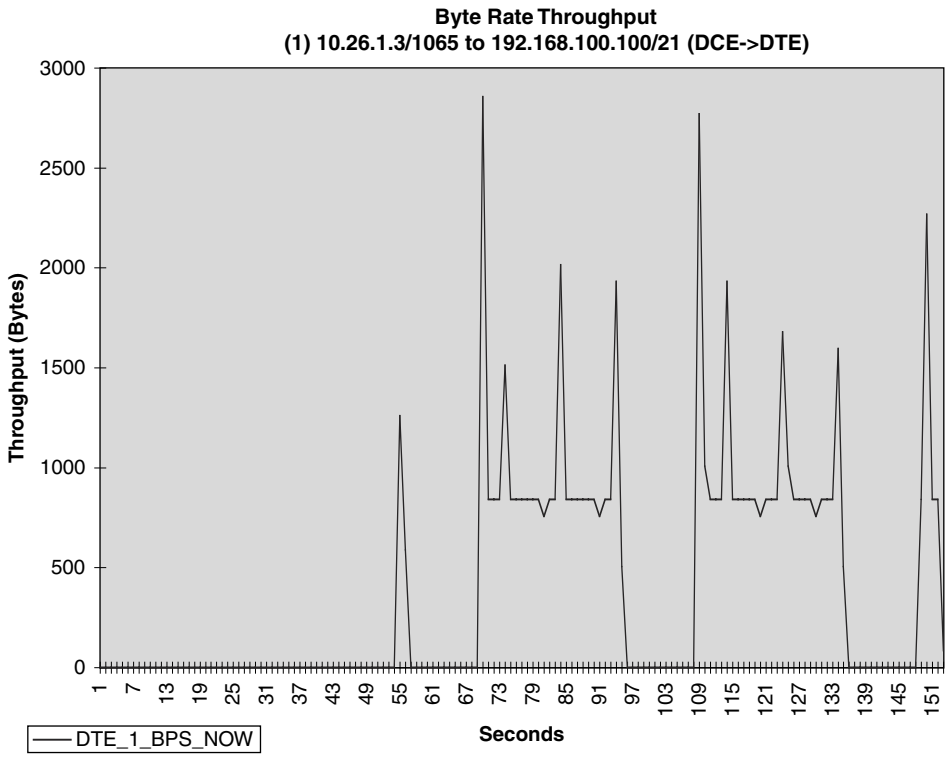


Figure 17.4 Network loading.

In the trace featured in Figure 17.4, there is a periodic drop in throughput, where no data seems to pass through the BSS (or the BSC). During these zero throughput times, packets queue up in the BSC until there is overflow (leading to the losses shown in the loss graph). The zero throughput times last for many seconds. Once packets can get through again, the queue is flushed out. Queues in the BSC are first in, first out queues of fixed size, and to accommodate time sensitive data, it appears that once the queue fills in the BSC, the packet that has been in there longest is discarded to make room for a new one. Now that the throughput has been reestablished, the queue flushes out, and you see a long stream of packets with decreasing latency (first in, first out queue). Since the queue is quite large, this causes the jitter shown in Figure 17.5.

We could, of course, make the performance of this network more consistent by band-limiting the offered traffic—slowing down the source of the traffic. We could also increase the amount of buffer bandwidth—which would add to the delay but reduce packet loss and hopefully reduce delay variability. Another alternative would be to overprovision radio and network bandwidth resources. Any of these options, however, either adds to our delivery cost or reduces traffic value.

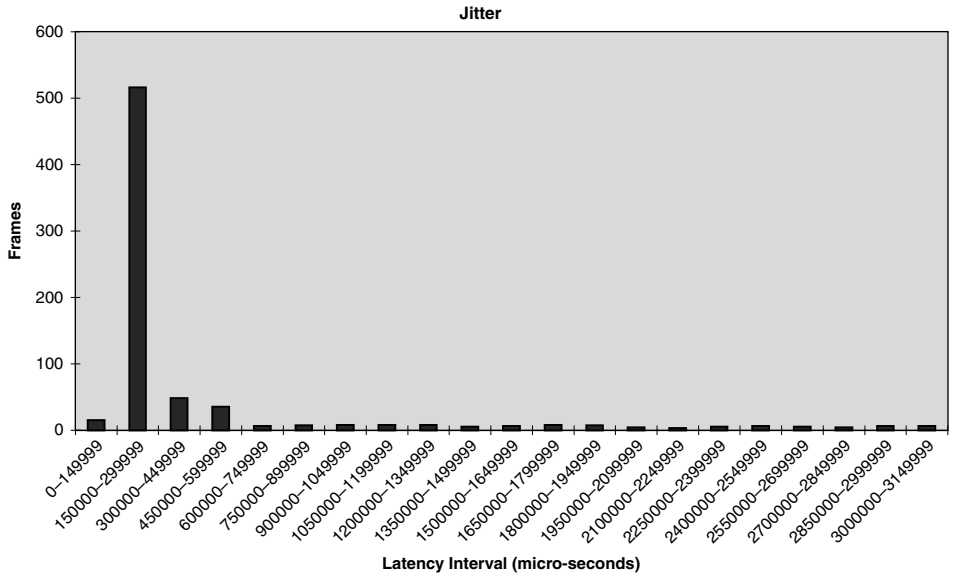


Figure 17.5 Jitter in a real-world network.

Buffering Bandwidth

It might be argued that we would be much better off moving the traffic through a bufferless transport channel. By buffering the traffic, we are increasing the amount of time the traffic is occupying network resources (absorption of memory bandwidth). If we trigger buffer overflow, we end up using more transmission bandwidth than we would have needed for the bufferless channel (unnecessary additional absorption of transmission bandwidth).

In effect, a software-routed network is a bit of a no-win proposition. Larger buffer sizes reduce buffer overflow and the risk of packet loss but increase delay; shorter packets create less processing delay at router buffers but introduce more overhead. Traffic shaping protocols can deliver better quality of service to some users but only at the cost of worse quality of service to other users, packet-level security. Packet-level policy engines (enforcement of access and policy rights) increase processor overhead and processor delay. Finally, firewalls can introduce additional delay (150 ms or more is not untypical).

To add insult to injury, if the software in a router is not well implemented (search and fetch routines, for example), then burstiness can increase as the packet stream comes out of the router buffer. We may need to increase buffer bandwidth just to avoid overloading router table lookup routines. As the traffic mix changes (traffic becomes more bursty), buffer bandwidth has to increase. As buffer bandwidth increases it becomes harder to deliver predictable performance. Deterministic performance from

the network becomes harder to achieve. The harder it is to deliver predictable performance, the harder it is to deliver a consistent user experience—and the harder it is to bill for that experience.

Buffer bandwidth is, in effect, a hidden cost implicit in packet routing. Protocol overhead is a hidden cost in packet routing. It can also be argued that there is an additional routing overhead cost.

Multiple Routing Options

The Internet was originally conceived as an inherently resilient network. The resilience was the product of multiple routing—the ability to send packets either on the shortest routing path or on alternative routing paths, which statistically will, of course, be longer. As a network becomes loaded it becomes progressively more likely that alternative routing paths will be used. If user-specific quality of service is implemented, priority users may be given shortest path hop sequences, while best-effort users will be sent on extended multi-hop routes, which are less preferred routes.

Thus, though multiple routing options are generally considered to increase network utilization, counterintuitively they can also increase network occupancy. For example, in the United States an average circuit switch voice call travels 100 km to 800 km. An average data packet travels 2700 km. As network loading increases, nonoptimal routing increases and, hence, source to destination traveled distance increases.

Let's revisit our OSI seven-layer model. Our IP protocols (TCP/IP) are at Layer 3 or 4, our traffic shaping protocols (RSVP/MPLS/SIP, Diffserv) are at Layer 5. All these protocols tend to be software based, although as we discussed earlier in the chapter, we are seeing an increasing need to introduce hardware accelerators to boost performance by reducing delay and delay variability. ATM and circuit switching are hardware based and can be considered as being implemented at the MAC layer and physical layer—switching physical bits through physical hardware.

As we move up the protocol stack, we get more flexibility but less performance. We are more exposed to changing user requirements, and it is harder to deliver predictable deterministic performance. As we move down the protocol stack, we move further away from the user. We lose flexibility but gain in performance.

IP Switching

Wouldn't it be nice if we could combine flexibility and performance? This is the promise of IP switching.

We have said that session persistency is increasing. We are trying to create long-lived packet flows from user to user. These flows are connection-oriented rather than connectionless. The idea of IP switching is to take a long-lived flow—a multimedia stream—and switch it using hardware. The definition of a *flow* is a sequence of packets treated identically as they move across a potentially complex routing function.

The Transition from IPv4 to IPv6

The transition to IP switching is facilitated by, though not dependent on, the move to IPv6 from the existing IPv4 address protocol. IPv6 updates IPv4 by including priority, labeling, QoS differentiation, and flow control. It also increases address bandwidth from the 32-bit header used in IPv4 to 128 bits.

A 32-bit address header supports 4 billion unique addresses. This might seem enough until you consider that 6 billion people might want IP addresses at some time in the future, 2 billion televisions might need IP addresses, 1 billion cellular phones might need IP addresses, and 14 billion microcontrollers might need an IP address. In IPv4, temporary addresses can be issued by a DHCP server and used locally. The same addresses can be reused elsewhere in the Internet. Network address translators are used to make sure locally issued addresses do not escape into the outside world. This works but requires administration.

Using 128 bits as an address bandwidth gives us 340 billion billion billion billion unique addresses, equivalent to 1500 addresses for every square meter on Earth. The cost is that we are adding extra bits to every message that we send (128 bits rather than 32 bits).

The IPv4 to IPv6 transition, shown in Table 17.2, does, however, have other benefits. IPv4 uses a variable-length header consisting of 14 fields. IPv6 has a fixed 40-byte header with 8 fields. In an IPv6 router, the router knows what it is looking for and looking at—a 40-byte header with 8 fields. In contrast, in an IPv4 router, the router has to discover the header length, determine the information contained within the 14 fields, and then act on that information. The IPv6 router can therefore provide more consistent performance because of the simpler consistent structure of the header.

Table 17.2 How IPv4 and IPv6 Headers Differ

IPv4			
Version	IHL	Type of Service	Total Length
Identification	Flags	Fragment Offset	
Time to Live	Protocol	Header Checksum	
Source Address			
Destination Address			
Options	Padding		
IPv6			
Version	Traffic Class	Flow Label	
Payload Length	Next Header	Hop Limit	
Source Address			
Destination			

To summarize the differences between IPv4 and IPv6:

- IPv6 header is a fixed 40 bytes long compared to IPv4's variable length.
- The simplified fields in IPv6 allow for more efficient processing through the router. The eight fields in the IPv6 header have the following functions :
  - **Version.** 4 bits long—identifies the protocol version number
  - **Traffic Class.** 8-bit field—similar to type of service in IPv4
  - **Flow Label.** 20-bit field—special handling requests
  - **Payload Length.** 16-bit field—indicates payload (excluding header) of up to 64 kb (or jumbograms by exception)
  - **Next Header.** 8-bit field—authentication, encryption, and fragmentation
  - **Hop Limit.** 8-bit field—prevents perpetual forwarding
  - **Source Address.** 128 bits
  - **Destination Address.** 128 bits

Delivering Router Performance in a Network

Consider the performance that an IPv4 or IPv6 router has to deliver as we move into the network. Table 17.3 shows how aggregated bit rate increases from the edge to the core of the network. Individual user devices produce between a few kilobits and 2 Mb per wireless device or up to a few Mbits for wireline copper access. At the Node B (in a 3G wireless network), user bit rates aggregate to between 25 and 155 Mbps. At the RNC, user bits aggregate to between 155 and 622 Mbps. In the core network, the bit rate aggregates to between 2.5 and 40 Gbps.

Table 17.3 Access Bandwidth/Network Bandwidth Bit Rates

ACCESS BANDWIDTH	NETWORK BANDWIDTH		
HANDSETS	**BASE STATION (NODE B)**	**RNC**	**CORE NETWORK**
RF (WIRELESS)	***COPPER***		***COPPER AND OPTICAL FIBER***
16 kbps to 2 Mbps	25 Mbps to 155 Mbps	155 Mbps to 622 Mbps	2.5 to 10 to 40 Gbps
WIRELINE			
56 kbps to 8 Mbps to 40 Mbps (VDSL)			

Table 17.4 Time Dependency versus Bit Rate

HANDSETS	AIR INTERFACE	GIGABIT PACKET PROCESSING	TERABIT PACKET PROCESSING
Milliseconds	Microseconds	Nanoseconds	Picoseconds
1 in 10^3	1 in 10^6	1 in 10^9	1 in 10^{12}
For example, 10-ms frame rate	For example, 20 μs flight path	For example, OC48 at 2.5 Gbps = 65 ns 10 Gbps = 16 ns 40 Gbps = 4 ns	

Table 17.4 shows the impact this has on router performance. A 3G handset is organizing its packet stream on a frame-by-frame basis (bit rate and channel coding can change frame by frame, which means every 10 ms).

Over the radio layer, we encounter time domain impairments (delay spread—a few microseconds). These are reasonably easily accommodated within our baseband processing. As we move into the access and core network and as the aggregated bit rate increases, so time dependency increases. In Gigabit packet processing, we need to process in nanoseconds; in Terabit packet read processing, we need to process in picoseconds.

If we can define an IP flow such that we read the first header of the flow (flow setup), any interim header—when and if the flow characteristics have to change (flow maintenance)—and the last header of the flow (flow clear-down), then we have made life much easier for the router, which means we have less flexibility but more performance. We have effectively re-created the characteristics and properties of a circuit switch. However, what we really need is session switching where the characteristics of the session can change every 10 ms (a dynamically rate-matched session).

This would make flow maintenance in an IP switch really quite complex—lots of interim headers that need to be processed each time the flow characteristics change. This in turn can only really be achieved by using fixed-length packets, which will generally be less efficient—unless the predefined packet size can be accurately matched to the payload. Adaptive bandwidth is expensive bandwidth.

IPv6 tries to address some of these performance and efficiency issues. IPv4 checksums are taken out in IPv6 and done at a higher protocol level. The IPv4 type of service becomes IPv6 flow control and priority, and the IPv4 options field is replaced by IPv6 extension headers. The IPv6 extension headers support fragmentation, defining the rule set for packet size—in addition to security (authentication and encryption), source routing (where the source requests its preference routing trajectory), and network management (hop-by-hop reconfiguration).

Improving Router Efficiency

The idea is to improve router efficiency, to be able to differentiate and prioritize traffic flows, and to adapt to traffic flow requirements as they change as a session progresses. The source routing header, for example, can be used to describe and request a requirement to receive a certain amount of bandwidth for a certain amount of time using the RSVP protocol (which we will cover shortly).

Having launched the packet stream into the network, IPv6 does not allow intermediate fragmentation, which means the defined rule set is maintained for the whole routing trajectory. Packet size can be up to 64 kbps. Larger packets are subject to additional negotiation and are subject to exception routing (a bit like trying to send an elephant through the mail; in fact, extra large packets are called jumbograms, which we discuss at length at the end of the chapter). If the bandwidth requirements keep changing, it is very convenient to keep the packet length the same. It is easier to multiplex packets together when they are all the same length.

IPv6 also includes the digital signature for authentication and encryption in the header, which helps support complex multiuser exchanges when users are continuously joining and leaving the session and adds a 4-bit flow label to describe 16 priority levels (from least important to vital).

Effectively, IPv6 provides more specific advice to the router in a format that is easier and faster to read, allowing the router to adapt to changes in the user group configuration or changing bandwidth requirements (flow property reconfiguration).

Traffic Shaping Protocols: Function and Performance

Let's consider some of the traffic shaping protocols, their intended function, and how they perform. Internet protocols are defined by the IETF. The workgroups include vendor representatives and other interested parties. Anyone can contribute to IETF standards making. It is a very democratic process constrained only by time, competence, motivation, and interest.

The open process works well—and relatively quickly when compared to existing telecom standards making practice—but the openness has a cost. The protocols often end up doing something other than the job they were originally designed to do. Given that these standards do not have a fundamental impact on hardware configuration and are generally realized in software, this mission flexibility is not really a problem.

Resource Pre-Reservation Protocol

RSVP is the short description for the Resource Pre-Reservation Protocol. It was originally intended as a way of pre-allocating an IP session resource for a predetermined period—that is, virtual circuit switching. The problem is that the protocol could only handle just over 2000 simultaneous flows and was very nonoptimum for multipoint-to-multipoint multiuser-to-multiuser session control.

Partly because of Microsoft's involvement (embedding RSVP into Windows 2000), RSVP morphed into being a per-conversation per-session protocol defining the QoS requirement from the user's device.

RSVP defines four levels of service:

High quality—application driven. This is for applications that are able to quantify their resource requirements—for example, an application using MPEG-4 encoding that can describe the bandwidth properties needed to preserve application integrity, which means declarative applications, or applications able to declare their bandwidth quality needs.

Medium quality—network driven. Here the application has an approximate rather than accurate idea of what it needs. For instance, it knows whether or not it needs an isochronous packet flow (all packets arriving in the same order they were sent). The application trusts the network to provide appropriate prioritization to preserve application value.

Low quality—network driven. Here the application may have defined and negotiated some basic latency bounds and minimum bandwidth guarantees that the network will endeavor to deliver.

Best effort—network driven. You get what you get—the "perhaps it will get there sometime sometimes" option.

So RSVP is used to define an application's bandwidth quality requirement, not to preallocate session resources as originally intended. It, however, remains as a session layer protocol. It describes session quality requirements.

Multiprotocol Label Switching

The session response allocation job has been taken over by Multiprotocol Label Switching (MPLS). Promoted by a number of vendors including Ipsilon (Nokia) and Cisco, MPLS breaks down packets into fixed-length cells and groups all packets within an IP session into a single flow. The session packets are tagged so that each router treats each packet in the flow identically (the definition of a flow—a sequence of packets treated identically by a possibly complex routing function).

MPLS provides a bridge between the Network layer (Layer 3) and the Data Link layer (Layer 2) and assumes ATM is used at the data link layer. It swaps Layer 3 Network layer labels for Layer 2 transport layer labels. The fixed-length packets mean that the packet stream can be prefragmented to fit into the ATM cell structure (a 48-byte payload). Thus:

- RSVP describes the session quality requirement.
- MPLS delivers consistent routing and consistent packet fragmentation for the duration of the IP flow.
- A subset of the MPLS standard, known as MP Lambda (MPλS), addresses tag switching over the optical layer (also known as "Lambda switching").

Diffserv

RSVP, MPLS and MPλS are intended to be used in parallel with Diffserv. Diffserv was originally conceived as a mechanism for defining four levels of service at the network edge—a job now done by RSVP. There were four standards—Platinum, Gold, Silver, and Bronze. Diffserv then morphed into a mechanism for grouping traffic flows together that share similar QoS characteristics—QoS flow aggregation—and is used as the basis for negotiating service level agreements (SLAs) between carriers.

This isn't one single network but a series of back-to-back agreements between a number of network operators, all of whom supply bandwidth to one another. To provide an end-to-end service level guarantee, each operator needs to know what will happen to the traffic (or more accurately the users' traffic) as it moves across other networks not directly or even indirectly under the operator's control.

Diffserv provides a rather rough-and-ready way of defining network bandwidth quality and provides a basis for internetwork SLA peering (that is, back-to-back network bandwidth quality agreements). Diffserv defines the performance expected—not necessarily the performance delivered. The performance delivered depends on how well each of the protocols performs individually and how well the protocols perform together.

Figure 17.6 shows Diffserv spanning four protocol layers and MPLS spanning two protocol layers. In practice, MPLS can also be regarded as a session layer protocol. It is session-specific in that it establishes a flow for the duration of a session.

Session Initiation Protocol

We also need to add Session Initiation Protocol (SIP) to the session layer. SIP has gained more traction in the IETF standards process as people have gradually started to recognize how and why session persistency is increasing. It is included in Microsoft Windows XP and is supported by 3GPP.

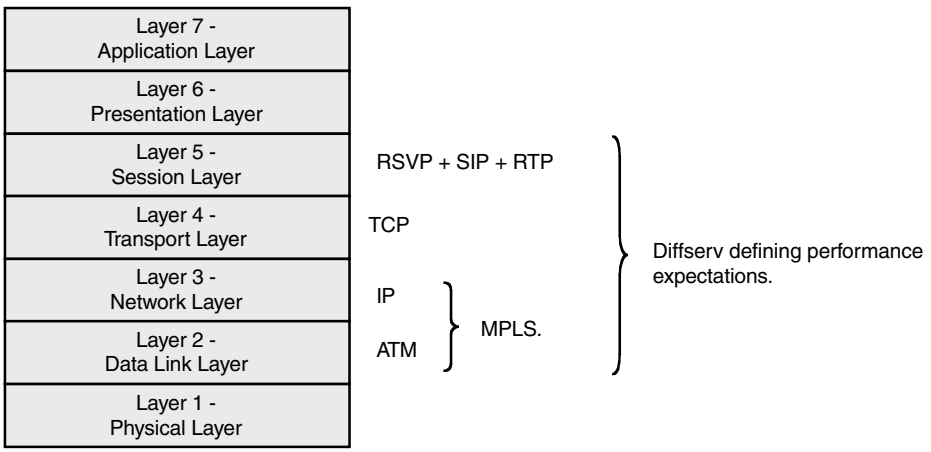


Figure 17.6 Diffserv and MPLS protocol spans.

SIP has evolved from the H323 standard for IP telephony. In H323 the protocol establishes a session, negotiates the features and capabilities of the session, and then clears down the session. In SIP, the process is the same, with all users identified by their Universal Resource Locator (URL) integrated with the user's phone number or hostname. However, whereas H323 is predominantly a protocol for managing IP telephony, SIP adds in media types and formats that the caller or the caller's application wants to use.

The problem with H323 and SIP is that call setup can take several seconds longer than the PSTN. It is a protocol that re-creates circuit-switched performance but with added delay.

It does, however, apparently work well when used to build up multiple user sessions. This involves the delivery of *presence* information—that is, whether someone is available or able to participate in a session. This is used in existing products like ICQ or AOL Instant Messenger to manage buddy lists.

The SIP Forum (www.sipforum.org) describes SIP as "a control protocol for creating, modifying, and terminating sessions with one or more participants." To work in a telecom environment, SIP also has to integrate signaling and service management. This is arguably SIP's Achilles' heel, since it is hard to re-create the functionality and stability of SS7. Why replace something that already works well? This will tend to make adoption slow in the traditional telco sector.

Real-Time Protocol

Last, but not least, we can add in the IETF Real-Time Protocol (RTP). The job of RTP is to help manage multiple per-user traffic streams coming down from the application layer on their way to being coded to separate OVSF code streams at the physical layer. RTP assumes that the individual streams are complementary and need to maintain their time interdependence. This time interdependence can be compromised in a number of ways. Individual streams can be sent via different routes. Some of the streams can be badly errored and subjected to transmission retries, and some may be sent on a non-isochronous channel (packets not arriving in the same order they were sent).

RTP adds timestamps to each of the packet streams so that they can be reclocked in the application layer of the receiver (which, of course, means the media is no longer strictly real time, since buffering is needed, thereby introducing delay and delay variability).

Measuring Protocol Performance

This brings us to a general summary of the quality issues related to packet-routed networks. Packet-routed networks use buffering to delay offered traffic in order to improve transmission bandwidth utilization. Traffic shaping protocols—RSVP, MPLS, Diffserv—address issues of traffic flow prioritization, improving quality of service for some users at the expense of other users. Traffic shaping protocols do not address the problem of packet loss.

In TCP (Transmission Control Protocol implemented in Layer 4, the transport layer), if packet loss is detected, TCP stops traffic until all the lost packets have been re-sent. An alternative is to use User Datagram Protocol (UDP). UDP allows packets to drop. If packets are dropped in TCP, "send again" protocols introduce variable delay. If packets are dropped in UDP, they are lost.

UDP is used in conversational sessions—for example, IP telephony. It assumes there is sufficient forward error correction on the channel to keep packet loss at an acceptable threshold. UDP is also often used in streaming applications where transmission retries would be seriously disruptive. Packet loss, however, directly degrades application quality.

The information in Table 17.5 is from the GPRS QoS specification determining packet loss probability, including the probability of a lost packet, the probability of a duplicated packet, the probability of an out-of-sequence packet, and the probability of a corrupted packet.

Levels of Reliability and Service Precedence

There are three levels of reliability (1, 2, and 3) and three levels of service precedence (high, normal, and low). This expresses UDP performance. Table 17.6 describes delay performance using TCP—assuming delay variability introduced by transmission retries.

Delay is defined as the end-to-end transfer time between two communicating mobiles or between a mobile and a Gi interface to the external Packet Data Network (PDN). It includes delay for request and assignment of radio resources and transit delay in the GPRS backbone. Transfer delays outside the GPRS network are not included. It is therefore not useful to us if we are trying to implement quality-based billing based on end-to-end performance guarantees.

Table 17.5 GPRS QoS

	PROBABILITY FOR			
RELIABILITY CLASS	LOST PACKET	DUPLICATED PACKET	OUT-OF-SEQUENCE PACKET	CORRUPTED PACKET
1	10^{9}	10^{9}	10^{9}	10^{9}
2	10^{4}	10^{5}	10^{5}	10^{6}
3	10^{2}	10^{5}	10^{5}	10^{2}

Table 17.6 Delay Parameters

	128-BYTE PACKET		1024-BYTE PACKET	
CLASS	MEAN DELAY	95% DELAY	MEAN DELAY	95% DELAY
1	<0.5s	<1.5s	<2s	<7s
2	<5s	<25s	<15s	<75s
3	<50s	<250s	<75s	<375s
4	Best effort	Best effort	Best effort	Best effort

Classes of Traffic in GPRS and UMTS

Traffic classes in GPRS and UMTS are described in terms of bearer quality of service rather than end-to-end service quality. The four traffic classes are as follows:

Conversational class. Minimum fixed delay, no buffering, symmetric traffic, guaranteed bit rate.

Streaming class. Minimum variable delay, buffering allowed, asymmetric traffic, guaranteed bit rate.

Interactive class. Moderate variable delay, buffering allowed, asymmetric traffic, no guaranteed bit rate.

Background class. Big variable delay, buffering allowed, asymmetric traffic, no guaranteed bit rate.

This is a rather oversimplistic classification and probably underestimates the need to support variable-rate conversational rich media exchanges, potentially the highest value part of the traffic mix. A complex conversational exchange implies minimum fixed delay, no buffering (no variable delay), dynamically changing bit rates and dynamically changing uplink and downlink asymmetry (a minimum/maximum uplink/downlink bandwidth guarantee).

Switching and Routing Alternatives

We can circuit-switch this traffic, cell-switch the traffic (asynchronous transfer mode), or packet-route the traffic. No one option is mutually exclusive of the other: We can packet-route and circuit-switch; we can map IP packet streams onto ATM.

Comparing circuit switching, ATM, and packet routing in terms of performance and bandwidth utilization:

- Circuit switching provides the most deterministic performance.
- IP provides the best bandwidth utilization.
- ATM provides flexible bandwidth. In particular, ATM has the ability to deliver flexible bandwidth in a controlled and predictable (and by implication billable) way.

ATM: A Case Study

ATM was developed in the 1980s as a mechanism for managing the movement of multimedia in local area networks—from one place to another, from one user to another. It is optimized to support the multiplexing of multiple media streams onto a common physical channel, maintaining the time interdependency of the media streams (including multiple per-user streams). Rather like a circuit switch, you load some traffic onto an ATM link, and you know precisely when and how the traffic will reappear on the other side. There will inevitably be some transport delay but no delay variability. It's rather like a diode or a capacitor: You know what will happen; it is a totally deterministic process. ATM achieves this predictability by using fixed-length cells of 53 bytes, of which 5 bytes are used as a header.

In early Ethernet implementations, ATM was generally considered as a very expensive protocol. A typical Ethernet payload might be 1500 bytes long. It was ridiculous to segment this payload into 48-byte cells and have to add 5 bytes of address and control overhead for each cell. However, fortuitously, as content has become more complex, the rich media mix, typical payload packet lengths have reduced. Consider audio or video sampling. Typically, a 10- or 20-ms audio or video sample will be described using, say, 160 bits per sample. Once this is channel coded, you have 320 bits or about 40 bytes. Unsurprisingly, packet lengths in a multimedia bit stream are typically 40 bytes long. They fit very neatly into an ATM cell.

The sales pitch for ATM is that it allows us to create virtual paths and virtual channels through a switched network. Because we get visibility and predictability from the fixed-length cell structure, we can deliver a defined quality of service with control of all the deliverable variables—peak cell rate, sustained cell rate, the maximum burst size, and cell delay variation. We can minimize cell delay variation for voice and video, and we can minimize cell loss or cell misinsertion for data. We have the capability of aggregating multiple per-user traffic streams and maintain stream properties as the traffic is moved over the physical layer, and we have in-built monitoring that allows us to have proof-of-performance reporting.

A user can be supported on a number of virtual channels. The virtual channels can be grouped together and sent on a virtual path that will follow the same route through

the network. The virtual channels and virtual paths are defined in the first 4 bytes of the header, with the fifth byte used as a check scan, which also provides synchronization for the receiver. We can emulate circuit switch performance but with more multiple channel and variable bit rate flexibility.

In the early days of ATM (the 1980s) it really didn't work very well. If the network switch became congested, cells would be dropped and the whole frame would have to be retransmitted. If a source was overwhelming the network, there was no way of choking back the source. The retry rate would increase and absorb yet more network bandwidth, and the network would go into meltdown.

The solution was to implement the Available Bit Rate protocol (ABR).

Available Bit Rate Protocol

In the ABR protocol, fields in a resource management cell are used to specify the cell rate required over the virtual channel (see Figure 17.7). There is a forward resource management (FRM) cell and a backward resource management cell (BRM). FRM and BRM are directly analogous to the Transport Format Combination Indicator (TFCI) used in the IMT2000DS physical layer, which establishes the rate requirement for the next frame.

The forward resource management cells and backward resource management cells in ATM ABR distribute flow and congestion information through the system. Note that network congestion can be measured and predicted either from the level of buffer occupancy or from rate information. If congestion is detected, the source can be told to either stop sending or reduce the send rate. It is a feedback-based control mechanism based on congestion measurement. You can think of it as the mirror image of the IMT2000 radio physical layer, which is a feedback control-based mechanism based on interference measurement. Both ATM and the radio physical layer work on a 10-ms frame duration.

The problem with ABR is that it works best for short round-trip delays and is nonoptimum for multipoint-to-point or multipoint-to-multipoint exchanges, where it is difficult to consolidate feedback from multiple sources.

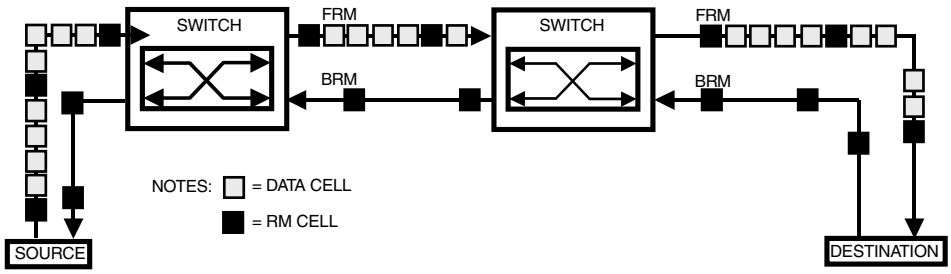


Figure 17.7 ABR protocol.

The Four Options of ATM

ABR is one of four protocols in ATM. The others are Constant Bit Rate (CBR), Unspecified Bit Rate (UBR), and Variable Bit Rate (VBR). Following is a brief description of each:

CBR. Peak cell rate is negotiated during connection setup and guaranteed for the duration of the connection. If you have a variable-rate source-encoded data stream, it is a bit wasteful to carry it on a constant-rate virtual channel. An alternative is to use constant-rate video streaming, which delivers varying quality as the entropy increases and decreases (not particularly good for the user).

VBR. Peak cell rate and sustainable cell rate are generated for the duration of the connection. Variable bit rate is rather like a blank check. What the user asks for the user gets. ABR is similar to VBR but adds congestion measurement as a feedback control mechanism. What the user asks for the user gets, provided the network has the bandwidth available or the network wants to make the bandwidth available.

UBR. This is whatever the network decides to make available—equivalent to best effort. Rich media is deliverable over CBR, VBR, or ABR. It could, of course, also be delivered using UBR provided the network was sufficiently well dimensioned (over dimensioned) to provide an adequate on-demand data rate, but over-dimensioning would add to the cost of delivery.

ABR. Of all four options, ABR provides a good compromise and gives the network operator some control over service quality. It also makes it reasonably easy to integrate the radio physical layer and network physical layer in terms of admission control (a topic we return to later in the chapter). However for multiuser or multipoint-to-multipoint exchangers, we are better off using VBR, CBR, or UBR, since the feedback control in ABR tends to fall over when faced with multiple feedbacks.

Efficient Network Loading

We have said that we use circuit switching when we have a bufferless end-to-end channel. We use packet routing when we have buffering. The buffering means we can even out the loading on the network and load the network more efficiently, because we don't have to dimension the network to accommodate instantaneous peak loading. We are, in practice, band-limiting the offered traffic, which, of course, implicitly degrades offered traffic value. Buffer bandwidth therefore has a performance cost: delay and delay variability, including the delay variability caused by retries when buffer overflow occurs. Buffer bandwidth has a capital cost. We have to pay for it.

We also use memory in ATM ABR. If we need to slow the source, we need to buffer. We also may need to buffer to smooth the multiplexing of multiple traffic streams, each of which may be variable rate. TCP/IP slows data flow if data is lost and speeds up when not. The rate flow control response is ½ to 1 second. IP routers therefore have to have sufficient memory to buffer data from all inputs for 1 second. In ATM ABR, the flow control response is 5 to 10 ms. ATM routers only have to have sufficient memory to buffer data for 10 ms.

Put another way, if the source can be stopped 100 times faster (comparing ATM to TCP/IP), localized memory can be utilized more efficiently. Efficient memory utilization minimizes cell loss (equivalent to packet loss in TCP/IP). You are less likely to suffer from data loss in ATM because you are using your memory more efficiently.

In addition, because the radio physical layer and ATM data link layer share a 10-ms response granularity, it is easier to match radio bandwidth performance to network bandwidth performance—for example, optimizing source coding to the common 10-ms frame length and optimizing packet length to fit within the common radio and network physical layer frame length structure.

ATM, TCP/IP Comparison

To summarize some of the pros and cons of ATM and TCP/IP:

ATM is fast and efficient. This is provided packet lengths can be organized to match the cell size. ATM has substantial address overhead (5 bytes in a 53-byte cell), but then so does TCP/IP by the time you have added IPv6 and various traffic shaping protocol overheads.

IP is flexible and slow. We can improve throughput for some users by using traffic management/traffic shaping protocols, but this degrades service for other users. We can also improve performance physically by using hardware coprocessors, but then we lose flexibility. It is hard to reconfigure coprocessors; it is easy to reconfigure software. We can improve software-based routing performance by doing tasks in parallel, but this makes task management more tricky and implies additional hardware overhead.

IP routers analyze the destination IP address in a packet and match the address to a stored list of subnet addresses. The software has to search, match, and execute the packet transfer. It may also have to decide on multiple routing options. It is hard to pin down this process in terms of the time taken to decide and execute the decision. It is implicitly an indeterministic process. Software-based decision making implies delay and delay variability.

ATM looks up an output port in the connection table and switches in hardware. It is inflexible but fast and predictable. The rigid rule set (fixed-length cells) is an advantage if a complex multiplex has to be supported. A complex multiplex would be multiple users, with each user having multiple streams, with each stream having multiple and continuously variable data rates (changing every 10-ms frame), and with the multiple per-user streams having a time interdependency that needs to be preserved both across the radio physical layer and through the network.

The convergence of parallel per-user traffic streams is done in the ATM adaption layer, also known as the adaptation layer. This handles parallel traffic streams requiring different error handling or latency characteristics that may or may not change as a user-to-user or multiuser-to-multiuser session progresses. ATM allows determinism to be maintained as multiple per-user, multiple-user bit streams aggregate into the network. The segmentation can also support multiple low bit rate users co-sharing an ATM cell.

We have said that ATM uses memory/buffer bandwidth more efficiently, and this in turn reduces the packet loss rate. A 2-Mbyte buffer at 60 percent utilization typically triggers a 4 in 10^2 packet drop rate. Increasing the buffer size to 64 Mbytes reduces the drop rate to 3 in 10^8. Effectively, we are reducing buffer utilization to a few percent, which means we can relatively easily accommodate loading peaks. As bandwidth becomes burstier, buffer utilization becomes an increasingly important network performance differentiator. Bursty bandwidth effectively puts our network buffer bandwidth into compression.

ATM makes our network bandwidth more adaptive—that is, better able to manage and respond to rapid changes in offered traffic loading. That's why 3GPP1 has specified ATM in the IUB and IU interface

The Node B base station takes the 10-ms frame packet stream from the radio physical layer and maps it into a 10-ms frame packet stream using 2 Mbps ATM to deliver traffic to the radio network controller (RNC). The RNC aggregates the packet streams from multiple Node Bs, each supporting multiple users who each have multiple traffic streams that require individual quality of service differentiation. These packet streams (virtual circuits) are grouped into virtual paths and sent to the core network using a 155 Mbps ATM physical layer.

Figure 17.8 shows how admission control is implemented in an IP network. The application layer generates a complex session. This may include Web pages and graphical material that will be formatted by the presentation layer using HTML/XML. The session layer uses RSVP/MPLS or Diffserv (or both) to describe and request required session properties, which are delivered using TCP/IP. The radio physical layer then takes the offered traffic and accepts it or refuses it on the basis of interference measurements. As and when the traffic is offered to the core network, admission control is managed on the basis of congestion measurements.

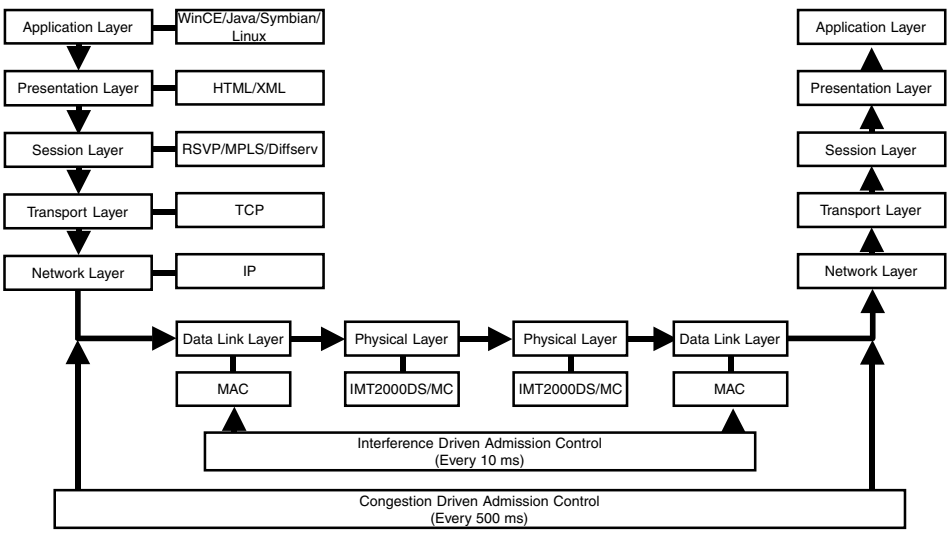


Figure 17.8 Admission control.

The congestion-based TCP/IP admission control works on a 500-ms to 1-second feedback loop; the radio physical layer admission control works on a 10-ms feedback loop (dynamic rate matching). The argument is that the congestion-based admission control using TCP/IP is not responsive or adaptive enough to support conversational rich media or the rapidly changing properties of the offered traffic mix.

It can be made to work by using ATM to discipline IUB and IU interface performance, which brings us back to our IP QoS network model, first introduced in Chapter 11 and covered in further detail in the next section.

The IP QoS Network

The challenge for an IP QoS network, is to adaptively manage admission control in the RNC including load balancing and traffic prioritization—that is, responding to User Service Identity Module (USIM) based admission requests and responding to congestion-based interference measurements in the IP core and interference-based measurements in the IP RAN (radio layer).

The RNC is one of the trickiest—and hence highest-value—components in the IP RAN, given that it has to arbitrate admission control requests coming from the USIM with interference measurements from the radio layer and congestion measurements from the network. Note the RNC also has to shed and share load sideways with other RNCs. This is not an IP network. It is an IP over ATM network.

The Future of ATM: An All-IP Replacement

It may be that ATM will be eradicated from future networks, but if this happens, the all-IP replacement will be no more or no less efficient, and no less or no more expensive. An all-IP implementation will have to replicate ATM functionality. Implicitly this means it will have similar overheads and a similar level of complexity.

Given that ATM has a more established track record, particularly in terms of its ability to deliver consistent predictable and measurable performance and given that consistent predictable and measurable performance is a precondition for quality-based rather than quantity-based billing, then it is our bet that ATM will remain with us for a very long time. The use of ATM in digital TV transmission is another reason for its increasing prevalence in future multimedia/rich media delivery networks (including content capture networks.)

The alternative is to keep circuit-switched architectures in place and overprovision transmission and switching bandwidth to accommodate peak loading (the bursty bandwidth overhead). Ericsson claims that it is still executing a form factor reduction for circuit switches of at least 60 percent every 18 months and a 30 percent reduction in the power budget. Switches are getting smaller, more efficient, and cost less as well. If you have to overprovision packet-routed or ATM networks to provide deterministic services, you may as well retain your (probably already overprovisioned) circuit switch.

IP Wireless: A Summary

Wireless IP is used generically to describe the application of TCP/IP to wireless networks. Specifically, wireless IP encompasses IP-based traffic management, IP-based network management, IP-based mobility management, and IP-based access management.

The IPv4-to-IPv6 Transition

Moving from IPv4 to IPv6 increases IP address bandwidth from 32 to 128 bits. This allows a 4-bit flow label to be used that supports 16 access priority levels—the basis for traffic discrimination. The first 8 bytes of the 128-bit IP header tell routers how to direct a packet steam (routing trajectory management) and where and how the user is attached to the network (mobility management). IPv6 also includes digital signatures for authentication and encryption—the basis for access management.

IP Traffic Management

In 3G networks, we are assuming a proportionate increase in video and image streaming with or without simultaneous audio—the traffic mix shift. MPEG-4 variable-rate differentially encoded information is presented to the physical layer. This traffic is by its nature highly asynchronous. In general, this asynchronous traffic will be multiplexed together, usually into an ATM cell switch fabric, and then delivered on to SONET or frame relay transport. The packet stream can be either isochronous or nonisochronous.

IP traffic management overlays a number of routing protocols, each of which provides a measure of traffic shaping functionality—for example, RSVP. RSVP was originally intended as a way of pre-allocating an IP session resource for a predetermined period—a kind of virtual circuit switching. The problem was that the protocol could handle only 2300 simultaneous flows. The protocol didn't scale well and was less than optimal for multipoint-to-multipoint streaming. RSVP is now evolving to become a per-conversation/per-session protocol defining QoS requirement from the device—for example, RSVP embedded in Windows 2000.

Most network designers assume that a network should be application- and device-aware. There will, however, be thousands of different application form factors with thousands of different QoS requirements and thousands of different device form factors with widely differing image capture, image processing, and display capabilities. In practice, a high-quality QoS can only be delivered when QoS requirements are application driven from the device. The application becomes network-aware rather than the network being application-aware. In RSVP, the highest level of service is application driven. The application is readily able to quantify its resource requirements; the lower levels of service (medium quality, low quality, or best effort) are network driven.

Having used RSVP to define QoS, MPLS is then used as a flow switching protocol. MPLS breaks packets into fixed-length cells—an ATM look-alike structure optimized for a multimedia multiplex. The packets within an IP session are then grouped into a single flow and tagged for expediting through the router hops (usually being mapped into an ATM or frame relay circuit). Finally, Diffserv is used to provide additional traffic shaping. Diffserv was originally proposed as a protocol for defining four levels

of service at the network edge. It is now being proposed (and promoted in various IETF workgroups) as a way of grouping traffic flows together that share similar QoS characteristics—QoS flow aggregation. It is then used as the basis for negotiating service level agreements between carriers (internetwork SLA peering).

IETF Real-Time Protocol (IETF RTP) may be used to synchronize complementary traffic streams being moved into and through the network—for example, timestamping parallel per-user bit streams carrying audio and video streaming. The IETF Synchronized Multimedia Integration Language (SMIL) protocol may be used to perform a similar function at the application/presentation layer.

IP-Based Network Management

All of the preceding configurations require new measurements to be made to provide an effective audit of network performance—for example, end-to-end delay, end-to-end delay variation, access latency, network latency, and application latency (server bandwidth constraints). These become embodied into IP SS7—an IP-based upgrade of SS7 and or an evolution of Simple Network Management Protocol (SNMP).

IP-Based Mobility Management

As vendors begin to embed IP addresses into base stations and routers, IP-based mobility management becomes an option; potentially Home Location and Visitor Location Registries could be replaced by DHCP servers tied into a directory-enabled network—that is, an IT solution superimposed onto a telecom network structure.

IP-Based Access Management

By access management, we need to clarify that we mean access to delivery bandwidth and to storage bandwidth. Note how encryption is used both to secure delivery bandwidth privacy and storage bandwidth privacy.

Similarly with authentication: We authenticate to arbitrate on the right of access both to delivery and storage bandwidth, which means access to servers or virtual (network-resident) storage—the storage area network proposition. IP Sec can be used to provide both delivery and storage bandwidth security including integration with X500 directory standards (X509 certificates) and Public Key Infrastructure (PKI) products. Note also how Triple A (the authentication, authorization, and accounting proposals from the IETF) also addresses the perceived need for unified billing.

IP protocols based on IPv6 have the potential to impact most, if not all, areas of 3G network topology. However, they have to demonstrate that they deliver an economic benefit and equivalent or better performance than presently achievable in circuit-switched network topologies.

The economic benefit of IP protocols is based on buffering. Buffering allows us to smooth our offered traffic loading, which means we can increase our network bandwidth utilization by spreading the offered load in the time domain. However, buffering implies a performance cost: delay and delay variability. Delay and delay variability degrades application value. In addition, we might argue that buffer occupancy implies

additional cost. The longer the information takes to pass through the network, the more it costs us.

Using IP protocols to route traffic—that is, using software to interpret packet header information—introduces uncertainty in terms of delay and delay variability, which makes it hard for us to guarantee end-to-end performance. IP routing is not by nature deterministic. It is difficult to accurately predict performance, particularly when the network is presented with highly bursty traffic.

TCP/IP has a 500-ms to 1-second feedback delay. As the traffic mix shifts toward an increasingly time-sensitive mix of rich media products and content, this delay will become increasingly expensive in terms of buffer utilization. As buffer utilization increases, packet drop rates increase. As packet drop rates increase, transmission bandwidth utilization reduces. As bandwidth gets burstier, the performance and efficiency of IP routing reduces.

The options are as follows:

- We can use IP as the basis for network management, but we have very well tested solutions already in place. As bandwidth gets burstier, the need for signaling increases. Narrowband SS7 will, at some stage, need to be replaced with broadband SS7. It does not necessarily need to be broadband IPSS7.
- We can use IP for access management—packet-level authentication and encryption, but we have well tested SIM smart card-based solutions already in place.
- We can use IP to support mobility management, but we have very well tested solutions (HLR/VLRs) already in place.

Thus, although IP protocols have the potential to impact most if not all areas of 3G network topology, in practice, their presence will likely be less pervasive than originally expected.

Network operators like to feel in control. Existing centralized network architectures provide that control. Control includes knowledge of end-to-end performance. Circuit switching provides absolute knowledge of end-to-end performance. Circuit switching was, however, never really designed to manage highly bursty bandwidth. Essentially, ATM is distributed hardware switching optimized for the transmission of highly bursty bandwidth.

Our definition of bursty bandwidth is not discontinuous bandwidth—short little bursts of low-bandwidth data. It is, rather, a persistent session in which activity is always present, but where the amplitude of that activity is constantly changing.

The IMT2000DS air interface is a wireless ATM air interface. The air interface is optimized to deliver bursty bandwidth, managing admission control every 10 ms on the basis of interference measurements. It is logical to match this radio air interface to an ATM-based access network optimized to deliver bursty bandwidth, managing admission control every 10 ms on the basis of congestion measurements. You can have IP-addressed packets and move them across the ATM switch fabric, but you are using ATM rather than IP to shape and manage and control the offered traffic.

The general assumption has been that an IP transition would occur. The first stage of the transition would be "IP over everything," and in practice, we are beginning to see this happening. The second part of this transition is going to be much slower than expected. "Everything over IP" implies certain performance compromises, and the

jury is still out as to whether "everything over IP" can deliver sufficient control of end-to-end bandwidth quality for conversational rich media exchanges, which represent the highest value part of the offered traffic mix. In addition, everything over IP implies having sufficient dynamic range available to adapt to the rapid changes in the bit rate and bit quality requirements of the offered traffic mix.

We could, of course, ignore all this and argue that we don't really need a network at all, which brings us to the topic of mobile ad hoc networks.

Mobile Ad Hoc Networks

Ad hoc networks are networks that do not have fixed assets. There are no base stations, Node Bs, RNC, or central switches, which means the users are the network and communicate with each other directly.

Figure 17.9 shows a number of users, A to P. All these users are in radio range and can talk to each other. The exchange can be user to user or user to multiuser or multiuser to multiuser.

The second user group, A to D, is out of range. However, if the two clouds of users are brought physically closer together, such that E of the first user group becomes visible to one of the users in the second group, then E can act as a repeater and users from both groups can talk to one another.

The disadvantage is that E's battery will very quickly go flat. For this reason, it is difficult to conceive of ad hoc networks being used to provide public access service, since we would never really know who is using our battery bandwidth. Similarly, all users could share memory and processor bandwidth. One user could store information on another user's device and use the other user's processing bandwidth to run applications that could be user- or group-specific. Again, in a public access network, it might be unsettling to have other users sharing your storage and processor bandwidth.

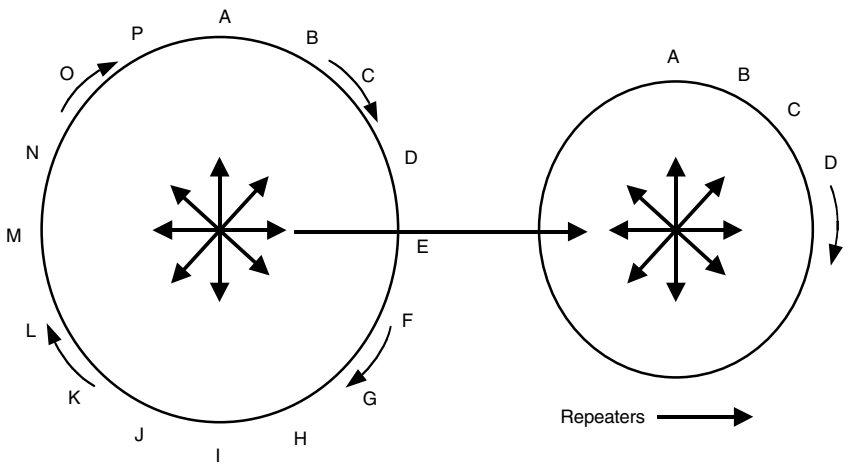


Figure 17.9 Ad hoc networks.

However, ad hoc networks do work well for private user groups. They are not new and have been used in private mobile radio for many years. In a PMR network, working directly between handsets is often called *back-to-back working* or *direct mode*. In an underground fire, for example, the first equipment failures are often the base stations or feeders to and from the base station. Firefighters still have to communicate. Back-to-back working allows them to talk to one another on an open channel. This functionality is available in many trunked radio networks, including, for example, TETRA networks in Europe and Asia. Typically, back-to-back working uses either proprietary protocols to manage access or protocols specific to whichever radio technology is being used.

The Internet Protocol Alternative

An alternative is to use Internet protocols using either IPv4 or IPv6. In IPv4, handsets are treated as nodes. A node sends a request to a DHCP server to lease an IP address for a period of time. In IPv6 handsets are also treated as nodes, but the handset would more likely be given its own permanent IP address.

Using IPv6, the 128-bit address header is split into two. The first 8 bytes tell a router how to direct the packet and where and how the user is attached to the group; the second 8 bytes is the end user's globally unique address. The router will typically be another handset. IPv6 also supports authentication, so access rights to the group can be managed.

Routing protocols are either proactive or reactive. In a proactive protocol, all possible routes within the network are continuously evaluated. When a packet needs to be forwarded, the optimal route is already known. This minimizes routing delay but absorbs network bandwidth. In a reactive protocol, route determination is invoked on demand. The access delay variability implicit in a reactive protocol means it cannot be used for real-time communication.

Zone and Interzone Routing

Figure 17.10 shows a simple example of zone routing where the protocols provide a rule set by which on-demand routing trajectories can be established. Zones are defined for each node on the basis of the number of nodes whose minimum distance, in hops from X, is at most some predefined number.

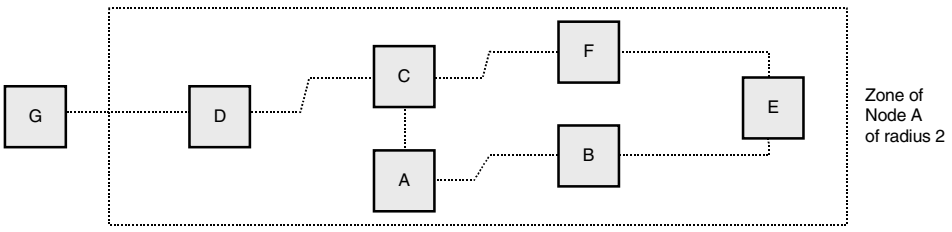


Figure 17.10 Zone routing.

In Figure 17.10, the zone radius is equal to 2, which means Nodes B through F are within the routing zone of A. Node G is outside A's routing zone. E can be reached by two paths from A: one with length 2 hops and one with length 3. Since the minimum is less or equal then 2, E is within A's routing zone.

Peripheral nodes are nodes whose minimum distance to the node in question is equal exactly to the zone radius—that is, nodes D, F, and E are A's peripheral nodes.

Figure 17.11 is an example of interzone routing. Node A has a datagram to send to node L. Assume a uniform routing zone radius of 2 hops. Since L is not in A's routing zone (which includes B, C, D, E, F, and G), A bordercasts a routing query to its peripheral nodes: D, F, E, and G. Each one of these peripheral nodes checks whether L exists in their routing zones. Since L is not found in any routing zones of these nodes, the nodes bordercast the request to their peripheral nodes. G bordercasts to K, which realizes that L is in its routing zone and returns the requested route (L-K-G-A) to the query source, A. This is sometimes described as a *broadcasting protocol*.

In Figure 17.11, the protocol relies on the source node and the intermediate nodes learning the position and hop distance of each of their neighbors. The header is made up of a destination address—in this example, 32 bits: the next-hop address (32 bits), the hop address after that, and so on. The length of the route to the destination has to be measured and recorded (in a 4-bit shop count field). This is known as a *neighbor discovery/neighbor-maintained protocol*.

The choice of protocols depends on whether a proactive or reactive response is needed. If a proactive response is needed, all the intermediate hops need to maintain routing information in order to route the packets they receive. These routing tables can quickly grow to great proportions.

If a reactive response is needed, you use a neighbor discovery protocol to learn the ad hoc network topology to allow the packets to be sent to allow route discovery to be done. The time taken to do this gets longer as the network grows in size.

Either way, route discovery and route maintenance protocols have to be carefully designed.

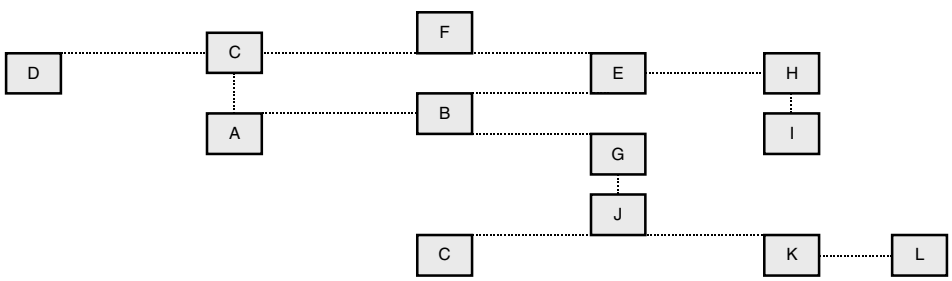


Figure 17.11 Interzone routing.

Route Discovery and Route Maintenance Protocols

Consider how route discovery is typically achieved:

- The source node broadcasts a route request packet. Each node hearing the request passes on the request and adds its own address to the header. The forwarding of the route request propagates out until the target of the request is found. The target then replies.
- As a node overhears other users it can use, data packets or control packets to do route discovery or route maintenance, inserting routes into its route cache. This is called *promiscuous snooping*.
- If a forward error occurs, a route error message is sent back to the originator saying that the link is broken. All nodes overhearing or forwarding the route error packet can update their routing information.

The protocol has to stay stable and avoid route reply storms. One way to achieve this is to use cluster-based routing protocols. These protocols divide the nodes of an ad hoc network into a number of overlapping or disjointed clusters in a distributed manner. A cluster head is elected for each cluster to maintain cluster membership information.

This gives rise to the term clouds and clusters. *Clouds* are groups of users or devices who intercommunicate with each other. *Clusters* address group membership issues within a number of clouds. Clusters are established to try and avoid the problem of counting to infinity where discovery protocols propagate out indefinitely.

IP Terminology Used in Ad Hoc Network Design

Let's summarize some of the IP terms used in ad hoc network design. Some of these terms will be less familiar to people with a telecom background.

Nodes. As in Node B in a 3G network, are simply devices that implement IP.

Routers. These are nodes that forward packets not directly addressed to them. In an ad hoc network, nodes can also be routers.

Host. This is an access node that is not a router.

Link. This is any physical medium—copper, wireless, fiber—over which nodes can communicate.

Packet. This is an IP header with a payload.

Neighbor association. A process whereby nodes copy IP addresses within their visible subnet and (from other subnets available) into a temporary access direction to form an instant virtual community. This is the basis of the protocols used in Bluetooth for local device discovery.

Promiscuous transmit mode. Only available when IPv6 is used, this allows a node to multicast to other nodes. If no response is received from one or more of the addressed nodes, these are deleted from the directory.

Promiscuous receive mode. Where a mobile node monitors router advertisements. It can also set itself to receive all packets on the link including those not addressed to it.

Dynamic discovery. Where mobile nodes detect their own movement by learning the presence of new routers as the mobile moves into wireless transmission range and by learning that previous routers are no longer available.

Multicast mesh. Link connections between forwarding group members.

Join request. Packets sent establishing/updating group membership and routing.

Join table. A table sent round to describe the join request.

Member table. A table sent round to describe the join table.

Note that ad hoc networks can be democratic. All users have equal access and equal management rights—that is, all users mutually agree on access and policy rights. Alternatively, one use can be a master—the first mobile to be turned on, for example—and all subsequent new members are slaves. The master handset may take responsibility for managing the user group, so, in effect, the handset acts like a base station. The group can either have equal access rights, or some users may be given priority. If some users are given priority, a system of fairness has to be established.

Beacon. A control message sent by a master handset node or a base station node or a master handset acting as a base station node informing all other nodes in the neighborhood of its continuous presence. Beacon channels are widely used in wireless air interfaces. GSM and TDMA have a beacon channel.

Cluster. A group of nodes within close physical proximity.

Cluster head. A node often elected in the cluster information process that has complete knowledge about group membership.

Convergence. The process of approaching a state of equilibrium in which all nodes of the network agree on a consistent collection of state about the topology of the network and in which no further control messages are needed to establish the consistency of the network topology—a steady-state condition.

Convergence time. The time required for a network to reach convergence after an event (typically the movement of a mobile node) has changed the network topology.

Distance vector. How many hops to store in a routing header.

Fairness. Not the same as equality, it may be that some users have access priority. In this case, all users are not equal but the access policy may still be regarded as fair.

Flooding. Delivering data or control messages to all nodes within the network.

Goodput. The total bandwidth used less the protocol overhead.

Laydown. The relative physical location of nodes within the network.

Mobility factor. The relative frequency of node movement compared to the convergence time of the routing protocols.

Payload. The actual data within a packet.

Scenario. The characteristics of network laydown, path loss, and mobility factor properties.

Security parameter index. The security context between defined router pairs.

Spatial reuse. Reuse of channels spatially far enough apart to avoid interference.

Administering Ad Hoc User Groups

All of the terms defined in the previous section can be used to set up, manage, and maintain ad hoc user groups. In principle it is very similar to setting up a session, maintaining a session, and closing down a session, except that the routing topology may be changing continuously as the session progresses and the users in the user group may be changing continuously as the session progresses.

A Sample Application

Suppose 100 soldiers land in a remote mountainous region of a remote mountainous country to fight a remote mountainous enemy. The first radio to be turned on becomes the master. As each soldier turns on his or her handset, it joins the user group (having been authenticated and having had over-the-air encryption enabled). Certain radios in the group might have predefined priority access rights, and there may be some predetermined reconfiguration capabilities that could be event-dependent (the master radio gets captured, for example). Provided all radios are within coverage of one another, all the users can talk to each other either on a one-to-one basis or in virtual open channel, and the users can hear all voice exchanges on the channel. The user group can also exchange images, video capture, and battlefield data—position coordinates, for example.

If one of the user group loses contact or if a number of users lose contact, it may be necessary to use one or more handsets as repeaters, which means choosing a volunteer to go and stand on the nearest hill. Group one can now talk to group 2, but you will now need to invoke interzone routing protocols to maintain communication.

Achieving Protocol Stability

This could be defined as a wide area network in that, theoretically (assuming sufficient density of users), the geographic coverage could be infinite. The issue becomes one of protocol scalability:

- How long does it take to admit a new member who may be an old member who has physically moved from one cluster to another?
- How many users have to be informed of that change?

The protocols need scalability and stability. If users are joining and leaving the group continuously, the signaling load can be substantial. Additionally, authentication ideally has to be achieved within milliseconds to provide an acceptable channel delay.

This is the obvious point to make about ad hoc networks. We have had ad hoc networks in the private mobile radio industry for at least 30 years—well before IP protocols existed. Typically, vendor-specific or technology-specific protocols were deployed (and are still deployed today—for example, in Motorola ASTRO networks, Nokia or Ericsson EDACS networks, or in TETRA networks).

We can replace these existing protocols with IP protocols, provided we get equivalent or better performance and other benefits (of which one benefit might be a more universal address and management standard). IPv6 specifically allows us to introduce more functionality into these ad hoc networks—more precise and predictable admission control (packet-level authentication), more precise and predictable prioritization,

more precise and predictable quality of service differentiation. For these reasons, it is reasonable to expect greater use of IP protocols in future ad hoc networks.

Macro Mobility in Public Access Networks

If the protocols could be made truly scalable, then of course we could envisage their application in a wider context—used for macro mobility management in public access networks, rather than micro mobility management in private access networks.

Let's assume every cellular handset has an IPv6 address. Let's also assume that every base station (Node B), every RNC, and all other subcomponents in the radio access network have an IP address—the IP RAN. When a base station (Node B) is installed and switched on, it becomes a member of the Node B community and is automatically configured within the IP RAN. When a handset is turned on, it logs on using the SIM-based International Mobile Subscriber Identity Number (IMSI) or equivalent identity number (EIN). It could also log on using its IPv6 address. In theory, you would not need to use the IMSI or the EIN, which means the device could have an IP address as well as the user.

This is one way of bringing together GSM-MAP networks and IS41 networks. The GSM-MAP networks provide access control on the basis of the IMSI; IS41 networks provide access control on the basis of the EIN. Both could be replaced by a composite of two IP addresses—the user address and the device address. This is known as *mobile IP* and is the basis of work presently being done by 3GPP1 and 3GPP2 to produce GAIT (GSM/ANSI Interoperable Terminal) handsets.

Mobile IP

To be at all useful, mobile IP has to be able, obviously, to cope with mobility. Handsets have a home address. An IP address is assigned to a mobile handset when it is logged on to its home link. If the handset moves, it becomes a mobile node. If it changes its point of attachment, it still needs to be reachable via its home address.

This means that IP addresses either need to replace or be added to the existing mobility management functions that manage migration of users from base station to base station, from RNC to RNC, from network to network, and from country to country. It has to be said that the mobility management presently deployed in GSM-MAP works remarkably well. Thousands of expensive and generally clever people have worked for years to establish and test international roaming procedures (including billing and settlement).

It may be that adding IP addresses to Home Location Registers and Visitor Location Registers can be justified in terms of additional functionality, but it is hard to justify on the basis of any obvious present shortcomings in the existing mobility management procedures (the "if it's not broken, don't fix it" principle). In addition, the situation is presently complicated by the uncertainties being introduced by the IPv4-to-IPv6 transition.

IP addresses are allocated by the regional Internet registries reporting to the Internet Assigned Numbers Authority (IANA). People can have IP addresses (personal access networks), and devices can have IP addresses (device access networks). Usually, access protocols need to be able to qualify the user and end user's device, so generally you

will need an IP address for the user and his or her device. This means lots of addresses and quite a lot of address overhead (2 × 32 bit addresses using IPv4, 2 × 128 bit addresses using IPv6).

Macro Mobility Management

3GPP1 has been keen on IPv6 because it is really the only way to provide adequate performance in terms of bandwidth quality differentiation and bandwidth quality control. Japan, for example, has decreed that all routers in Japan will have to be IPv6-compatible from 2005 onward. U.S. operators are less eager because of the added overheads introduced by IPv6.

What happens to a device IPv6 address when a device is discarded? Does the address return to the public domain? How does it return to the public domain? What happens when a user churns from one network to another? Does the personal IP address move with the user? If not, how does the IP address return to the public domain?

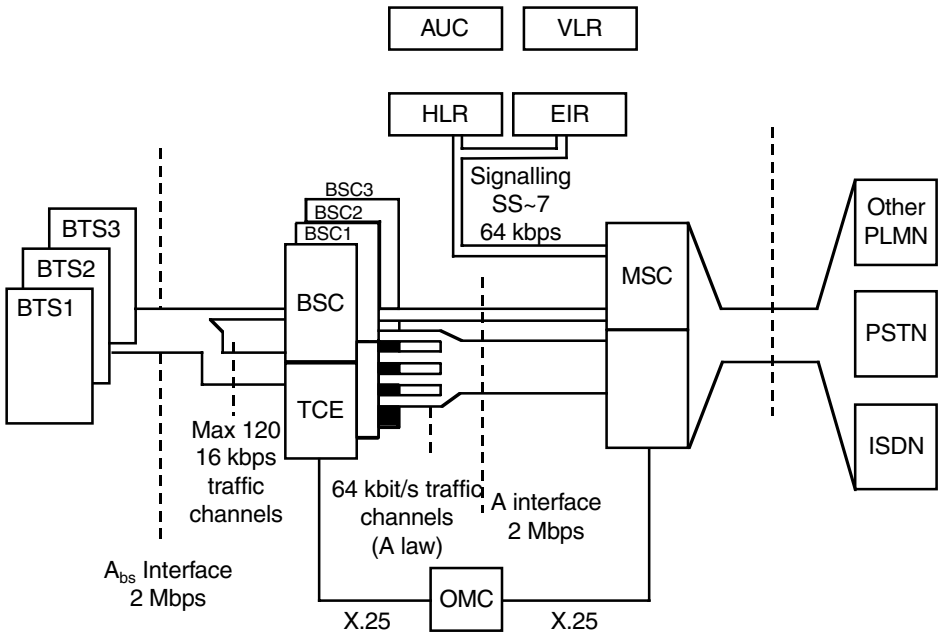
One way to avoid these issues is to use dynamic allocation: An IP address is allocated to a user for a session. This could either be an IPv4 address or an IPv6 address. However, this makes macro mobility management quite tricky to control. Without clear billable benefits, most network operators at this point ask the "why bother" question. For these reasons, the adoption of IPv4 or IPv6 for macro mobility management within 3G cellular networks is likely to be much slower than initially expected.

If public wireless LANs became pervasive, then arguably this would greatly increase the use of IP for macro mobility management. However, it may be that wireless LAN deployment in public areas will be slower than expected and may only become relatively ubiquitous as a subset of 4G network deployment.

In 3G public access networks the probable outcome is that while IP addressing may be used on a relatively wide scale, it will be used for addressing rather than for traffic or mobility management.

Use of IP in Network Management

If we were to give every network subcomponent an IP address, we could use IP protocols in the Operation and Maintenance Center (OMC). Figure 17.12 shows the position of the OMC in a traditional GSM-MAP network. The OMC provides a link between the base station controller and the mobile switch center.



AUC	Authentication Centre
BSC	Base Station Controller
BTS	Base Transceiver Station
DAI	Digital Audio Interface (104 kbit/s)
EIR	Equipment Identity Register
GSM	Global Systems for Mobile Comms
HLR	Home Location Register
ISDN	Integrated Services Digital Network
MS	Mobile Station
OMC	Operating and Maintenance Centre
PLMN	Public Land Mobile Network (Or Private)
PSTN	Public Switched Telephone Network
VLR	Visitor Location Register
TCE	Transcoding Equipment

Figure 17.12 GSM network (GSM-MAP).

Figure 17.13 shows the substitution of the RNC for the BSC, the substitution of the Node B for the BTS, and the creation of a more or less self-contained radio network subsystem consisting of a family of Node Bs and their host RNC.

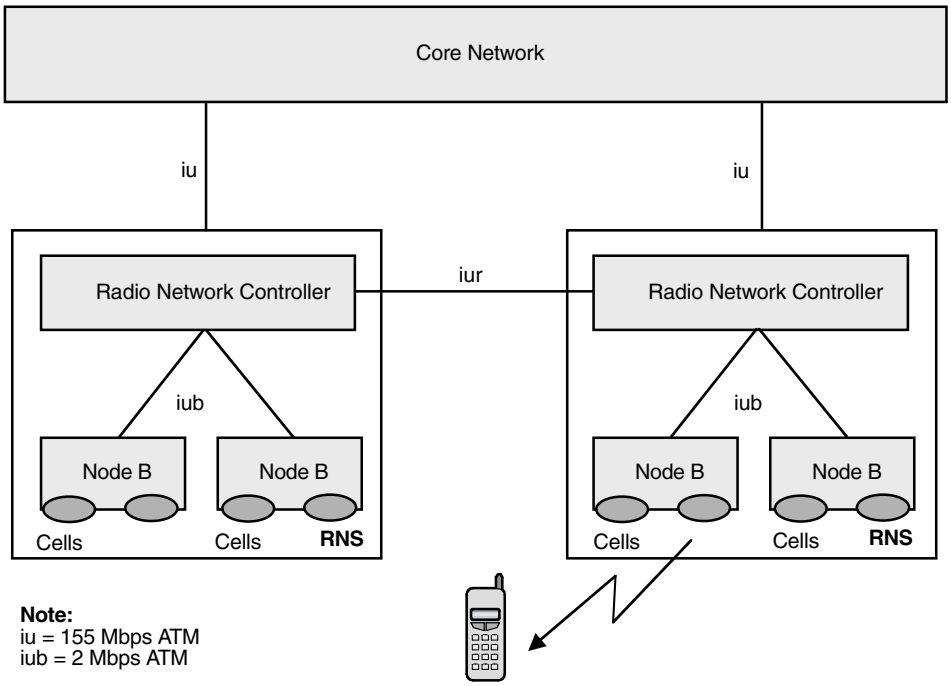


Figure 17.13 3G network topology.

Again, this could be a logical progression and could result in better interoperability between different vendors' OMCs. This may, of course, not be in the vendors' best interest. A vendor-specific OMC helps capture and retain a network operator as a customer and prevents other vendors supplying parts of the network subsystem.

Note that one of the ostensible purposes of telecom standards making is to prevent this from happening. One vendor's RNC should be capable of talking to another vendor's Node B and a third vendor's core network solution. The OMC is the glue that holds these different subsystems together, tracking hardware and software malfunctions. The recipe for the glue tends to vary from vendor to vendor and will probably remain that way.

The Impact of Distributed Hardware and Distributed Software in a 3G Network

We have said that a typical 2G network MSC has 20 to 30 million lines of code. In a 3G network, we are moving some of the decision making involved in traffic and mobility management to the RNC. The RNC arguably will be the network subcomponent with the biggest influence on radio access performance. How well does the RNC perform load distribution? How well does the RNC perform traffic prioritization?

As bandwidth becomes burstier, decision making has to be moved out to the network edge. Intrinsically this means that software value moves out to the network edge. Network software performance can be measured in terms of decision-making performance. Decisions are better made close to the point of execution, as this reduces signaling delay.

Signaling delay can seriously compromise network performance (network bandwidth quality). If signaling load exceeds signaling bandwidth, signaling delay will be increased—and, hence, signaling delay variability will increase, meaning the signaling will become less deterministic. Signaling performance is therefore an integral part of network software performance. We need to discover:

- How efficiently we capture the information needed to make a decision?
- How efficiently we move that information to the place where the decision is made?
- How efficiently/predictably we execute the decision?
- Does the decision making stay stable when presented with highly dynamic loading conditions?
- Does the decision making stay stable when presented with highly complex and rapidly changing user bandwidth quantity and quality needs?
- Can the network software adequately control and measure network performance (to which it itself is a contributor) such that a subscriber can be billed for a proven rather than promised session-based quality of service?

IP over Everything

IP protocols have been promoted as a panacea—a magic pill that promises performance and efficiency benefits. It is difficult to have both. As you improve IP performance, its efficiency goes down.

The network operator needs a high level of visibility to real-time network performance, including real-time measurements of the network's response to a highly variable offered traffic load. This visibility is probably easier to achieve by optimizing existing signaling protocols (the evolution of SS7 to broadband SS7), and existing network software platforms.

"IP over everything" looks achievable, as well as rational and justifiable; "everything over IP" less so.

A Note about Jumbograms: How Large Is that Packet in Your Pocket?

We have waxed lyrical on the merits of ATM as a mechanism for managing multimedia in a real-time environment—delivery with known, preferably minimal, and nonvarying delivery delay. The more complex the per-user media multiplex, the more sense it makes to have defined packet sizes—the 48-byte payload in ATM, for example.

However, not all traffic is delay-sensitive. A very large percentage of traffic is delay-tolerant. A very large percentage of traffic does not need to be conversational, interactive, or streamed, which means it can be best-effort.

Similarly, in an IP packet stream, we do not have to have defined packet sizes. Consider IP over ATM. If we decide to use a defined packet size of 40 bytes to fit inside a 48-byte payload, we have the IP header overhead on the 40 bytes, the ATM cell overhead on the 48 bytes, and 8 bytes spare that nobody knows what to do with. We have achieved determinism, which in this case we don't need, at the cost of efficiency.

Efficiency increases in an IP packet as packet length increases. The header becomes a proportionately smaller overhead. Big packets might need to be exception-routed to avoid causing possible indigestion at node choke points, but provided we can manage this, we can optimize the throughput efficiency of our data file. This is the rationale behind jumbograms.

A jumbogram is an IPv6 packet containing a payload larger than 65,535 octets and up to 4,294,967,295 octets. Jumbograms can also be classified as urgent. If you want to bring an IP network to its knees, send a series of urgent jumbograms. Provided traffic is best-effort, IP can deliver substantial efficiency benefits, both in terms of delivery bandwidth utilization and minimal address overhead.

There is little or no prenegotiation to be done before the packet is sent. There is little or no renegotiation needed while the packet is being sent. There is therefore very little need for signaling bandwidth. It is a very simple session with simple, time-tolerant, low-cost bandwidth requirements.

As sessions become more complex and we move toward conversational rich media exchanges, the increasing need for close control of end-to-end delivery delay and delay variability and the need to provide highly adaptive bandwidth makes IP less optimum both in terms of access control and traffic management. We have to harden up network software decision making.

We also need to consider the complexity involved in supporting multiple physical channel streams per user, each with user specific or application specific quality of service requirements. The arbitration of this complex multiplex demands a high degree of determinism difficult to deliver from an all IP network. If we want this performance, then we need to have IP over ATM, and we need to bear in mind the costs implied in this choice. Complex content and complex applications require complex control.

Software-Defined Networks

In Chapters 5, 10, and 11 we profiled the merits/demerits of software and defined radio hardware and software configurability—the ability of a radio to reconfigure itself at the physical layer to allow interoperability with different air interfaces operating at different frequencies. Physical layer flexibility includes the ability to change the modulation used (GMSK, π/4DQPSK, QPSK, 16-level QAM) and the ability to change digital filters. In practice, it will always be hard to deliver operational transparency across widely spaced frequency bands. You will always need frequency-conscious components, RF hardware that is specific to the operating frequency, such as antennas and front-end filtering.

The same principles apply to base station and Node B configuration. It is possible to have broadband linear RF PAs, and it is possible to have relatively broadband antennas, but there is generally a performance cost associated with the additional flexibility.

Similarly, we can say that smart antennas give us a certain amount of reconfigurability—the ability to adapt to geographic changes in offered traffic distribution. For instance, the loading in the morning rush hour is different from the loading in the evening rush hour, since people are traveling the opposite direction. Smart antennas coupled with smart antenna software capable of recognizing and responding to changing traffic conditions can be considered to be providing us with a partially software-defined network.

Software-defined radios have been promoted in military applications as a way of enabling incompatible radio technology to intercommunicate. The army, navy, and air force have traditionally specified their own unique radio systems. As warfare has become more complex, it has become increasingly necessary to find some way of getting these radio systems to talk to one another. This includes physical layer capabilities and compatibility in the higher layers. The radio has to be able to talk to other radios and other networks.

As we move up the OSI protocol stack, it becomes increasingly easy to reconfigure both radio and network functionality. Note, however, that reconfigurability does not necessarily equate to compatibility. In fact, the more flexible we make a process, the harder it is to ensure the compatibility of that process with other (equally flexible) processes. We have said that we have predominantly hardware solutions at Layer 1, predominantly software solutions at Layer 7, and a mix of hardware and software in between.

It is reasonably easy to test hardware compatibility: Does it fit or doesn't it; does it work or doesn't it? It is much harder to test software compatibility. Does it work given a certain set of inputs and operational loading parameters? What if X happens, will Y result? How is Y affected as user requirements change?

Software behavior is far less predictable, which means it is less deterministic than hardware behavior, less easy to model, and less easy to test. Because it is harder to model and harder to test, it is harder to ensure compatibility.

The Argument for Firmware

The answer according to present thinking is to use firmware rather than software. Firmware is consistent software. It trades flexibility for consistency. Firmware cannot be changed dynamically in the way that software can be changed, but it does have much more predictable behavior, and it is therefore easier to ensure compatibility.

As applications and content become more complex, software and hardware becomes more complex. We have said that it is harder to test software compatibility than it is to test hardware compatibility. This suggests that compatibility problems are going to increase rather than decrease. Firmware is partly a solution, but only at the cost of flexibility.

In practice we are beginning to see this in present network deployment. After considerable effort, we get Bluetooth devices to talk to one another at the physical layer; we ensure compatible signaling is deployed and common traffic management and access protocols are used only to find that the application layers fail to intercommunicate.

3G Network Considerations

In a 3G network, it is going to be very difficult to get RNC software from one vendor to intercommunicate with RNC software from another vendor. There are just too many lines of code involved. The job of application layer software is to increase session complexity, which in turn increases session value. However, as session complexity increases, handset and network software has to work harder, continuously adapting to changing user and user application needs.

We need software-defined radios and software-defined networks to help provide the flexibility needed to adapt to dynamically changing user requirements as a session progresses and dynamically changing loading as a session progresses. Software-defined radios and software-defined networks have a cost: they increase the likelihood of incompatibility. There may also be a performance cost—for example, the use of software in routers rather than hardware switching.

Software-defined radios and software-defined networks do not implicitly reduce costs and do not implicitly improve compatibility. Actually, they tend to add cost and make compatibility harder to achieve. Software-defined radios and software-defined networks are, however, a necessary component in the delivery of the flexible adaptive bandwidth needed to support flexible adaptive applications.

Summary

In this chapter we reviewed the merits/demerits of IP-based traffic shaping, the need to deliver IP flow management, and some of the hardware/software performance issues involved in router design. We said that not only is traffic becoming more asynchronous, but session persistency is also increasing. This means Internet protocols have to replicate the functionality of circuit switching (tight control of end-to-end delay and no delay variability) and the flexibility of ATM (along with its ability to multiplex highly asynchronous traffic). We reviewed SIP as an IP-based session management protocol, but made the general point that protocol overheads increase as traffic/session management complexity increases.

We reviewed ad hoc network topologies, ad hoc network management, and access management protocols, showing how zone routing can be implemented. We also pointed out that mobile IP can potentially be used to provide both micro and macro mobility management.

Finally, we looked briefly at software-defined networks and suggested that flexibility and adaptability do not necessarily deliver compatibility.

Traffic shaping protocols have a very direct impact on the user experience—the quality of service as seen by the user. Quality of service has to be expressed in some kind of comprehensible and consistent manner. This is achieved, or should be achieved, by implementing a service level agreement—the subject of our next chapter.

CHAPTER

18

Service Level Agreements

Service Level Agreements (SLAs) are used in private and public networks to define an agreed set of performance metrics being delivered from one party to another or between multiple parties. This chapter addresses all the variables, the specialized needs, quality billing models, and consequent challenges implicit in current and future SLA implementations.

Managing the Variables

If we are to implement a quality-based billing model, an escalating complex of variables must be managed, within prescribed limits, so we can prove we have delivered the quality of service requested and required.

Wireline circuit-switched networks. These SLAs are really simple, specifying how much bandwidth is available, the bit error rate of the bandwidth (typically 1 in 10^{10}), and network availability (typically 99.999 percent).

Wireless circuit-switched networks. These SLAs are reasonably simple but need to comprehend physical/geographic coverage and dropped call rates. Note that physical/geographic coverage can vary on a day-to-day basis either because of changing propagation conditions or changes in offered traffic loading (creating noise rise at the base station receiver). Wireless circuit-switched performance is therefore less predictable than wireline circuit-switched performance because of the unpredictability introduced by the radio physical layer.

Wireline packet-routed networks. These SLAs need to comprehend a number of additional—and largely unpredictable—variables, which include packet loss, packet delay, and packet delay variability.

Wireless packet-routed networks. These SLAs need to reflect all these variables plus the additional variability introduced by the radio physical layer. Because the radio physical layer has a higher bit error rate than copper or optical access, it will intrinsically introduce a bigger variation in performance.

Defining and Monitoring Performance

SLAs can be internal or external. External SLAs define and monitor required performance levels against predicted performance levels—for example, the performance variability introduced by IP traffic shaping protocols. Internal SLAs set and manage user or customer expectations. In addition, SLAs need to measure network availability and performance (network response), and application availability and performance (application response).

SLAs are simple to implement when network performance is consistent and predictable. They are hard to implement when network performance is inconsistent and unpredictable. The radio layer is inconsistent and unpredictable. Packet routing is inconsistent and unpredictable. Add the two together in a wireless IP network and you get very inconsistent, very unpredictable, performance. This is the challenge for wireless IP radio network design. Consistent end-to-end delay and delay variability and packet loss become part of the Quality of Service guarantee, which, in turn, becomes part of the SLA.

Determining Internet Service Latency

An Internet service provider latency agreement might cover backbone latency or end-to-end latency. Backbone latency might be between 100 and 150 ms. End-to-end latency (assuming a wireline termination) might be 150 ms.

We can test these latency figures by sending a packet on a round-trip and seeing what happens to it. If it never returns, we have detected packet loss; if it returns after a certain time, we know the round-trip delay. This is sometimes known as "ping testing"—that is, channel sounding, rather like echo sounding in submarines.

So the typical Internet latency as expressed in the SLAs is between 100 and 150 ms, however, typical Internet latency is about 200 to 400 ms. The difference is that the SLA latency is average latency—for example, over a one-month period—whereas Internet latency reflects peak latency. If latency is average latency, delay at peak times can be substantial. This is probably the time of day when we also most need the performance.

Addressing Packet Loss Issues

A Service Level Agreement might also be defined on the basis of packet loss. A low packet loss might be 2 percent to 5 percent of packets sent. A high packet loss might be defined as 20 percent to 30 percent.

In practice, if we are using TCP/IP, a packet loss of 30 percent will trigger so many retransmission requests that the network becomes unusable. If we are using UDP with a packet loss of 30 percent, the application becomes unusable.

Network Latency and Application Latency

We have also said that we have to differentiate between network latency and application latency. An application SLA could say, for example, "95 percent of transactions will have less than a 2-second response time, and 5 percent may have a response time of 2 to 5 seconds." This might be used in an enterprise resource planning (ERP) application.

A bandwidth SLA might say something like this: "Information streams up to 100 kbps will be delivered with an end-to-end delay not exceeding 100 ms with a known bound to delay variability of not more than 10 percent. Bit rates above 100 kbps will be handled on a best-effort basis."

In a wireless network, this might be combined with a grade of service figure such as "for 95 percent of all locations, a user will receive." This is a *grade of service SLA* (GoS SLA).

QoS and Available Time

We can also describe quality of service in terms of available time, with available time defined as error-free seconds. In a wireline SLA, this could be something like the following (see also Figure 18.1):

Error-free seconds (EFS). Available seconds in which no bit errors occurred

Errored seconds (ERR SEC). Available seconds in which at least one bit error occurred

Severely errored seconds (SES). Available seconds in which the BER was worse than 1 in 10^3

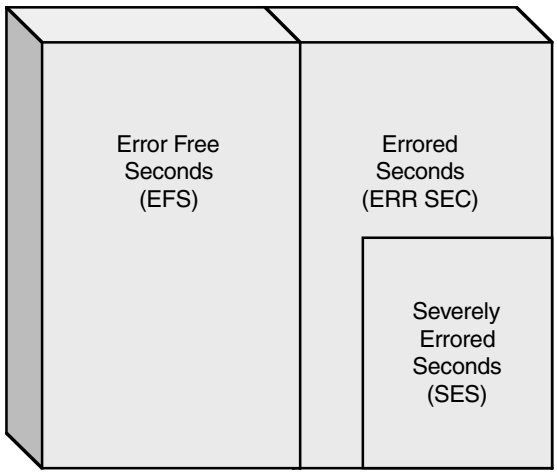


Figure 18.1 QoS description in a wireline SLA.

Only error bursts with a BER worse than 10^{-3} (severely errored seconds) are counted as unavailable time, which means the granularity of how you describe bit error rates averaged over time becomes an integral part of the billing process. For example, we could count degraded minutes—the number of minutes during which an average bit error rate of, say, 1 in 10^6 bit errors or worse recurred—degraded hours, days, or months. The longer the time constant, the easier it is to deliver the SLA.

However, we need to remember that the radio physical layer introduces more bit errors and more bit error rate variability, so not only has the SLA become more complex (harder to express), it has also become harder to deliver.

Billing and Proof-of-Performance Reporting

Unless you are an expert IT manager, you will now be thoroughly confused, thoroughly bored, or both. A potentially simple transaction has turned into a complex legal document expressing highly variable performance metrics, which are difficult to prove. This is not a good basis for billing.

If a network fails to deliver against the SLA, we also need to agree on a refund policy:

- Do we just say "sorry"; do we say we will try harder next time; do we give some money back or provide some free bandwidth?
- Who is responsible for proving a certain level of service was provided, the network operator or the end user? If it's the network operator, how can we be sure that what the operator is telling us is right?

The more variables we have, the harder it is to manage and measure the variables, and the harder it becomes to bill for bandwidth quality. Adding in a radio physical layer makes it even harder.

Real-Time or Historical Analysis

Proof-of-performance reporting can either be done on the basis of real-time analysis or historical trend analysis—packet delay or packet loss over a particular period. If, as a network operator, I have failed to deliver adequate performance over an extended period as expressed in an SLA, I may be faced with economic harm litigation. An e-commerce site or m-commerce site, for example, might be sensitive to end-to-end delay and delay variability and certainly sensitive to packet loss.

My customers may well be able to prove that revenue loss is directly linked to transaction delay introduced by my network. Litigation liability may be a hidden cost implicit in a packet-routed network.

Measuring Performance Metrics

Figure 18.2 shows the results of a survey commissioned by Lucent asking a number of IT managers the question: "What parameter do you test when assessing network performance?"

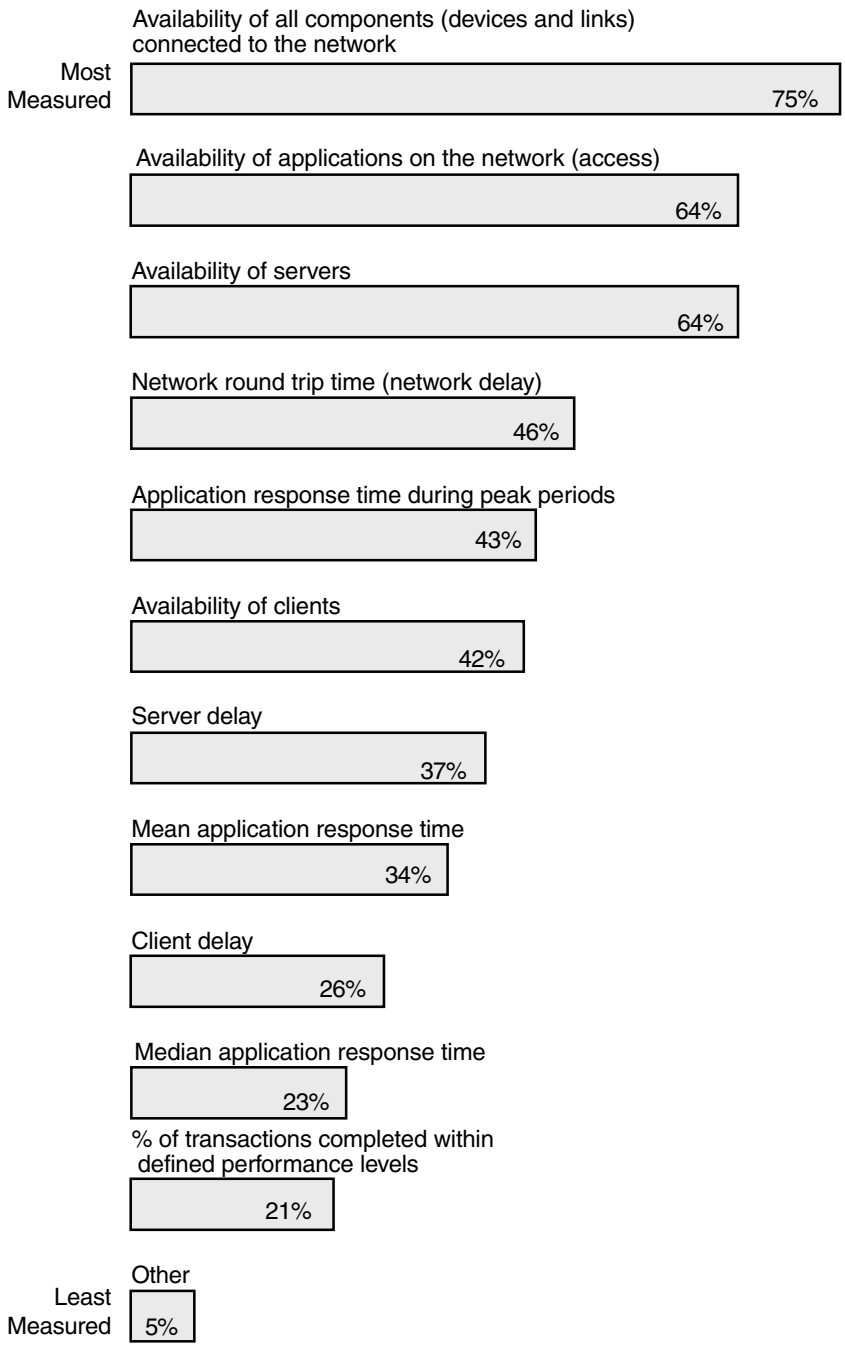


Figure 18.2 Network performance testing—a Lucent survey.

The most tested parameter was network component availability—that is, availability of network hardware. This is because it is the easiest metric to measure and also the least useful. The least measured metric was the percentage of transactions completed

within defined performance levels. This is the hardest metric to measure and arguably the most useful. The table needs to be completely reversed, with the presently least measured metrics becoming the most important and the presently most measured metrics becoming the least important. It is not the hardware you want to measure; it's the performance of the hardware that counts.

Of course, the cost of measuring all these performance metrics might exceed the income from the service provided (the cost of measuring packets may be greater than their value). We have made our transmission bandwidth more efficient, but we have hugely added to the cost of our application bandwidth (the cost of proof-of-performance reporting). The only way we can recover the cost is to send the user a completely incomprehensible bill and hope he or she just pays without working out that we have failed to deliver what he or she has paid for. We have experienced this problem with GPRS.

GPRS Billing

GPRS billing is based on GPRS call records. GPRS call records are generated in the GPRS service nodes. Packet counts are passed to a *charging gateway* that generates *call detail records* that are sent to the billing system. However, charging records may be generated in more than Serving GPRS Support/Service Node (SGSN) and in one or more Gateway GPRS Support/Service Node (GGSN). For any one Packet Data Protocol (PDP) context, you therefore need charging records from the SGSN and the GGSN.

These then become part of the Support Node Call Detail Records (S-CDR). The S-CDR may contain triggers—a set of containers that are added when specific trigger conditions are met—for example, uplink and downlink volume count. This provides the basis for quantity-based rather than quality-based billing.

A typical bill might be $15 per month for 5 Megabytes, $30 per month for 20 Megabytes, $50 per month for 50 Megabytes, and $1 per Megabyte thereafter, with the bill being incremental and additional to existing voice tariffs.

In addition to the S-CDR, there is also a Mobility Management Call Detail Record (M-CDR). If a change of routing area occurs because of the movement of the subscriber, a change of location container is added to the M-CDR.

In addition to the M-CDR, there is also a Gateway GPRS Support Node CDR (G-CDR). The job of the G-CDR is to handle handover information as the serving node changes. The triggers are the same as the S-CDR triggers.

Components in the container include data volumes in octets separately identified for uplink and downlink traffic streams, initial and subsequent changes of quality of service requested, and initial and subsequent changes of quality of service received, as shown in Table 18.1. These two are not necessarily the same, which means there are three components to consider:

- Quality of service requested
- Quality of service negotiated
- Quality of service delivered

The GPRS Tunneling Protocol (GTP) is used to deliver GPRS CDRs from the network element or functional entities generating charging records—for example, from the SGSN and GGSN. The CDRs are then delivered from the charging gateway to the billing system.

Table 18.1 Container Record

QoS requested = QoS1 QoS negotiated = QoS1	QoS negotiated = QoS2	
Data volume uplink = 1 Data volume uplink = 2	Data volume uplink = 5 Data volume downlink = 6	Data volume uplink = 3 Data volume downlink = 4
Change condition = QoS change Timestamp = TIME1	Change condition = Tariff change Timestamp = TIME2	Change condition = Record closed Timestamp = TIME3
Session Progression ⇒		

To date, Quality of Service billing has only been implemented between backbone operators and large corporate VPNs. The vast majority of billing is done on the basis of packets sent, sometimes combined with components from existing voice tariffing, such as time of day, duration of the session, and distance.

The practical problem with GPRS has been that the billing data is not captured from a single point but from many points. In a voice call, circuit-switched calls are charged in the anchor MSC. Billing capture is guaranteed by routing all signaling through the anchor MSC, even if the traffic channel of a call is routed through another MSC because of handover. The process in GPRS is far less consistent. There is really no satisfactory way of describing session complexity. As we cannot describe complexity, we cannot bill for it. In addition, there is really no satisfactory way presently to provide proof-of-performance reporting:

- What were the end-to-end delay and delay variability parameters of the session (including packet loss and packet retry statistics)?
- Did the network meet the maximum peak loading requirements of the session?
- What was the percentage of errored seconds as the session progressed?

Even if we had a way of capturing these performance metrics, we would need to find a relevant way of describing them to the user. In practice, these problems are being resolved by the development of *service detail records* (SDRs) that become an integral part of the call detail record, which may, just to be confusing, evolve into a *session detail record* (also SDR).

Session-Based Billing

As session persistency increases, it makes sense to move back closer to the traditional session-based billing used in circuit-switched voice networks, using the anchor MSC to track session setup, session maintenance, and session clear-down in the same way that the anchor MSC presently tracks call setup, call maintenance, and call clear-down.

However, we also have to capture session complexity:

- How many multiple channels per user are supported?
- What are the QoS attributes of each of these channels?

- Were these QoS attributes preserved or compromised by the network?
- Did this have an impact on application value (was it noticeable to the user or the user's application?
- Were the user or the user's application bothered about the quality degradation—that is, did they know or did they care, and if not, can you bill them anyway?

There may also be sources of subsidy to take into account—advertising revenues set against session cost.

We have said that the job of handset hardware, handset software, network hardware, and network software is to increase session value. There is no point in increasing session value if we either destroy that value or cannot find a way of describing that value and billing for it.

Toward Simplified Service Level Agreements

In practice, Service Level Agreements very easily become overly complex—too complex to present to a user in a convincing way. This implies a return to more easily tangible quality factors.

Qualifying Quality

In a voice call, quality can be qualified by users in terms of audio fidelity (and dropped call rate). In a complex rich media exchange, quality can be quantified in terms of resolution, frame rate, color depth, and contrast ratio. Audio and video quality can also be defined in terms of consistency (consistent quality delivered consistently from the beginning to the end of the session—that is, consistently consistent quality). Additional value parameters are conferred by other factors, such as security (the robustness of the authentication and encryption used in the session) and any storage resources used as the session progresses.

This results in a much simpler SLA based on audible and visible (directly experienceable) quality metrics.

Bandwidth Quality versus Bandwidth Cost

From a network operator's perspective, as we increase delivered bandwidth quality, we increase delivered bandwidth cost. Provided the tariff premium achievable exceeds the cost premium we have delivered an improved average margin per user (AMPU), much more valuable to us than the usual measure used—average revenue per user (ARPU).

To derive AMPU we, however, need to know how much different services have cost to deliver. We know, for example, that an SMS message takes 1 second to send and consists of a few ASCII 7-bit character strings, which means the message will be a few hundred bits long. We can approximate the cost of delivery.

The costing process becomes more complex as content complexity increases. For example, there is presently no consistent method for costing the overheads associated

with aggressive SLAs. There is no consistent way of costing the hidden costs of rebating against failure to deliver SLAs, and there is no consistent way of costing buffer bandwidth—the cost of buffering packet flows as they move through the network.

Additionally, we presently have no consistent way of expressing long-term storage costs and storage revenues (or the fact that archived content may increase in value over time). These metrics need to be defined in storage SLAs—access delay, consisting of access delay over time, access priority, download delay (download bandwidth), upload delay (upload bandwidth), and storage security over time.

Both for delivery bandwidth SLAs and storage bandwidth SLAs, it really all comes down to the fact that you cannot afford not to fulfill an SLA. You also need to consider the hidden costs implicit in rebate provision—customer dissatisfaction, churn rates (where customers move to other networks), the precedent set by rebates, the disputes implicit in rebates, and the litigation risks created by a failure to deliver an agreed Quality of Service.

Although hard to prove, it's probably a better value to overinvest in the network in order to ensure that the quality of service provided meets the quality of service requested for most of the time (in which most is defined as four nines or five nines QoS availability). This results in a clean, dispute-free billing environment, satisfied customers, and low churn rates. As we will see in the next chapter, such billing schemes already exist and are successfully applied in the digital TV/digital broadcasting industry.

Personal and Corporate SLA Convergence

Unlike specialist SLAs, personal and corporate SLAs are becoming more similar over time. Traditionally, corporate SLAs have included metrics like transaction response time and access to corporate data (storage SLAs). However, these metrics are now included, or can be included, in personal subscriber SLAs—reflecting the fact that many people now work at home and need work and leisure access to remote data bases and the facility to store and retrieve data.

Specialist SLAs

Specialist users, however, will have a number of additional metrics. They may, for instance, expect the network to be situationally aware. The ability of the network to respond to a particular situation becomes part of the SLA. This would include an ability to reconfigure network topology or to change admission control procedures and access policy rights (and maintain network resilience).

Range and Coverage

Specialist user Service Level Agreements are, as you would expect, more specialized. Range (coverage) is often more important than capacity. Many networks still use tone modems in a narrowband analog channel, so high peak bit rates are not a huge issue.

Many transactions—searching a police database to check a car registration, for example—require only small amounts of information to be sent. Many users are therefore still very well served by narrowband legacy networks, which may be analog or digital.

Narrowband digital radio schemes may be voice-centric or data-centric or both. Mobitex networks are still widely deployed, providing 8 kbps data channels using GMSK. These networks are very adequate for many utility applications where one of the main functions is simple job status updates.

Here, coverage is the single most important quality metric. The network will be designed to provide a certain geographic coverage which will be the basis for the GoS agreement.

Onto Channel Time

In voice networks or safety critical data networks, onto channel time may be tightly specified. For many years the police, fire, and ambulance have used tone signaling systems with a typical rise time of 180 ms. Public access networks have needed to meet or improve this figure in order to be considered as a serious and credible alternative to existing private networks.

User Group Configurations

User group configuration flexibility is also important. GPRS, for example, was proposed as a way of providing flexible user group access. In practice it has been hard to match the performance of simpler but task-optimized trunking networks (providing open channel working).

Where access delay is not a critical performance parameter, a GSM, TDMA, CDMA, or 3G network can provide an interesting alternative to a private network. An example would be the use of GSM modules to provide simple connectivity for utility meters, vending machines, industrial machinery, parking meters, security, and surveillance devices.

Content Capture Applications

In turn, this supports specialist content capture applications (for example, surveillance applications). Here the storage SLA might comprehend the storage, archiving, and retrieval techniques used to support the application. For example, with the use of MPEG-7-based image search and image retrieval techniques, the SLA would include the typical time taken to fulfill an image search and retrieval request, or the time taken to do a fingerprint match (an application response SLA).

Specialist Handsets

Specialist handsets have been proposed that retain press-to-talk and band selection facilities. Examples would include GPRS-enabled handsets that meet the GSM ASCI (Advanced Speech Call Item) requirement to support broadcast and group calls, priority and preemption, and (reasonably by GSM standards) fast onto channel access times.

Specialist applications include oil refineries and chemical plants. Here the quality metric is that the handsets and the base stations have to be intrinsically safe to avoid the risk of explosion. Because overhead water hydrants are often used to limit the impact of gas leaks, the handsets often have to be waterproof. They also must be capable of working in very high noise environments and may use noise canceling microphones and noise canceling headsets, which means part of the SLA will be the ability of handsets to work satisfactorily in extreme conditions.

Site-Specific Software Issues

In these applications, part of the Service Level Agreement may need to include identification of site-specific interference issues—the use of specialist machinery at 2.4 GHz, for example. The network may need to store and to be able to disseminate hazardous chemical data (to be situationally aware, event aware, or both).

Network software may need to be optimized for specialist user applications. Motorola has a product called PoliceWorks—a Windows-based applications suite for managing traffic offenses, accidents, incidents, arrests, booking, and towing—the day-to-day needs of the police officer on the beat. A similar product called WaveSoft has been developed by Motorola for firefighters. Both products use Motorola's Magic Pipe software operating system.

The response time of the software to particular application demands or situational requirements may be part of the SLA.

Mandatory Interoperability

Service Level Agreements may also define mandatory interoperability between previously noncompatible radio networks. The U.S. Navy Space and Naval Warfare System Command (SPAWAR) buys software-defined radios from Motorola to provide interoperability between the U.S. Navy, the U.S. Coast Guard, and NATO.

Here, the software-defined radio has to provide a bridge between incompatible radio physical layers and incompatible application layers (for example, the use of different incompatible authentication and encryption standards). This has become a particular issue in the U.S. post September 11 to ensure public safety agencies can intercommunicate by radio.

Hardware Physical Test Requirements

User products often have to be extremely well ruggedized to survive combat conditions. This includes ruggedized information capture devices including surveillance devices.

Here, the SLA might typically include physical tests that user hardware has to pass, such as drop tests or splash tests. Battery performance, including through-life duty cycle, might also be defined. A typical police radio, for example, might be used for three 8-hour shifts every day. Its performance might be specified over 1500 duty cycles and would include performance in extreme conditions, such as subzero or very hot temperatures. Other aspects of handset hardware, for example, display visibility in high ambient light conditions, might be specified.

Specialized Network Solutions

Specialized users need specialist network solutions. These can often be hard to accommodate within public access network topologies or hard to describe within a public access network SLA. The following list gives an overview of some typical specialist user requirements:

Wide area coverage. Good geographic grade of service.

All informed user capability. Where the whole user group can overhear real-time traffic or, if sharing video, oversee real-time traffic.

Virtual open channels and instant access. The ability to have instant access to channels—less than 250-ms onto channel access times.

Multigroup announcements. The ability for some or all users or the dispatcher to make a multigroup announcement.

Wide area broadcast messages. The ability to wide area broadcast.

Priority levels (dynamically changeable). The ability to provide dynamically changeable access priority levels.

Security (encryption) and radio inhibit sleeper phones. The ability to stun (disable) stolen radios. For example, with a stolen police radio, it is useful to be able to re-enable the speech return path. The device then becomes (unbeknownst to its new owner) a surveillance device, and the police have at least some chance to recover the device and to capture the thief.

Voice clarity. Important to police, other public safety officers, or military users, who may need to issue safety-critical instructions. Note that instructions may be issued in a high noise environment which further reduces voice clarity. To illustrate this, British armed forces were given the example of an instruction sent over a field telephone "Send reinforcements we are going to advance" being interpreted as "Send three and four pence, we are going to a dance."

Specialist audio processing. May be needed to block out background information—for example, noise suppression and hands-free operation. Whisper phones may be needed, so you can talk very quietly in a covert surveillance operation.

Press to talk (PTT). PTT capability or a keypad may be necessary.

Talk groups. For example, geographic and/or functional.

Interworking or interoperability. Network reconfigurability might also be required.

Storm plans. For example, Motorola's storm plans or Ericsson's special event plans.

Covert surveillance (miniature handsets). Very very small handsets might be needed for surveillance applications.

These are all feature sets based on hardware and software handset and network functionality, which might typically be included in a specialist private radio network performance specification and would need to be reflected in a specialist user SLA.

In addition, specialist user groups often have to perform against operational SLAs—how long it takes for an ambulance to respond to a call out, for example. Some of these operational response criteria are described on the Web site supported by the Forum of Incident Response and Security (www.first.org).

Often it will be the information technology manager's responsibility to make sure radio coverage is available across the whole area over which incident response SLAs need to be delivered. A failure to communicate can, of course, result in loss of life or possible litigation, so radio coverage and coverage consistency can be a very sensitive issue.

Good radio coverage is provided by having lots of RF power available. An APCO Project 25 specialist radio at 800 MHz is either 5 Watts (portable) or 110 Watts (mobile), and the base stations are 500 Watts. (TETRA's 25-Watt base stations look pretty puny in comparison.)

Specialist users can also often get access to highly dominant base station sites providing very wide area coverage—for example, on mountaintops or on top of very high buildings.

The Evolution of Planning in Specialist Mobile Networks

As with personal subscribers and corporate subscribers, specialist user behavior is determined by the user's handset. This means the handset's hardware and software determines how the user accesses the network and determines the properties of the traffic offered to the network.

Twenty or thirty years ago, the dominant consideration in two-way radio planning was coverage (that is, range), and this was the key metric described in private mobile radio/specialist mobile radio GoS agreements. These GoS agreements effectively implied noise-limited networks.

As traffic increased, capacity became more of an issue, particularly where a relatively large number of users needed to be accommodated on a relatively small number (two, three, or four) of relatively narrowband (12 kHz) channels. Much more recently, we have had to consider content—the impact of police, fire brigade, and public safety officers capturing image and video content and needing to move that content into and across a radio network.

Consequently, we now have to combine content (peak loading introduced by a more complex mix of multimedia content) and context to define peak loading. Context can be determined by time of the day, day of the month, month of the year, or year or decade. Some events are intrinsically predictable (centennial celebrations, jubilees), and some are less predictable (riots and revolutions, hurricanes, earthquakes). These events create peak loads on the network. It is arguably unrealistic to dimension radio networks for any possible future event, and we may find it necessary to undertake risk analysis—the task of determining the actuarial risk of an event occurring or reoccurring. The actuarial risk is then described in the Service Level Agreement. This relationship is shown in Figure 18.3.

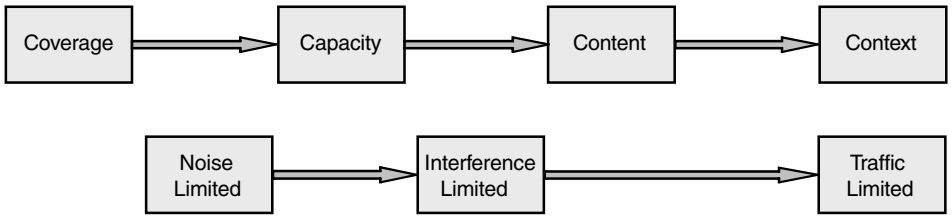


Figure 18.3 Content and context in Service Level Agreements.

Summary

Service Level Agreements have been used for many years in private radio networks to cover performance requirements such as coverage, voice quality, system availability, onto channel times, blocking, resilience, and redundancy, in addition to specialist user requirements.

SLAs have been introduced into public access networks over the past 3 to 5 years. They are the consequence of moving from circuit switching to packet routing.

In a packet-routed network, we are trying to improve bandwidth utilization by loading our network more intensively. However, some of our users require a level of service that is equivalent to circuit switching. We can provide this level of service but only at the expense of other users. Implicitly, some users will have better service than others. The differentiated Quality of Service has to be described in the SLA. An SLA might be session based, which means established just for the duration of a session between two users or multiple users. The SLA might also be for a defined period—a year, 2 years, or 10 years. Either way, we have to define the service level requested, the service level negotiated (which may or may not be the same thing), and the service level delivered (which again, may or may not be the same).

If the service level delivered does not match the service level promised, then we may have to provide some form of rebate or recompense. The hidden costs of rebating may be substantial.

As content becomes more complex, Service Level Agreements become more complex. As Service Level Agreements become more complex, they become harder to deliver, harder to measure and manage, and by implication, harder to bill.

The solution may be to revisit session-based, time-based billing with quality defined in terms of tangible experienceable metrics and audio fidelity and image/video quality and session consistency. Note that session consistency becomes increasingly important as session persistency increases.

An increase in session persistency and an increase in session complexity together create significant challenges for the future evolution of Service Level Agreements based on traditional QoS metrics (delay, delay variability, and packet loss).

As bandwidth becomes burstier, it becomes harder to describe in an SLA. If we cannot describe something accurately, we cannot bill for it accurately. If we cannot bill something accurately, our billing will be subject to dispute. Disputes destroy customer/subscriber asset value.

Our prime objective is to build and sustain session value and describe session value in terms that are tangible to the end user—quality-based billing based on tangible quality metrics.

Without transparency to the cost of delivery, we cannot produce a cost-based billing metric. Substantial work still needs to be done in this area. Present work is focusing on adding more information to the container data records (CDRs) or the IP equivalent (IPDRs), but no real consensus yet exists as to how to produce a consistent measurement of delivery and storage (access and network) bandwidth cost against differentiated Quality of Service delivery metrics.

We also need transparency, however, in terms of the cost of delivery against specific service level criteria. If we cannot define and quantify cost of delivery, we cannot define average margin per user.

CHAPTER

19

3G Cellular/3G TV Software Integration

In this chapter we review the potential convergence between 3G TV networks and 3G cellular networks. We compare U.S. digital TV, and European and Asian DAB (Digital Audio Broadcasting) and DVB (Digital Video Broadcasting) standards. The DAB and DVB standards provide an air interface capable of supporting mobility users, providing obvious opportunities for delivering commonality between cellular and TV applications. We point out that TV has plenty of downlink power but lacks uplink bandwidth (which cellular is able to provide). There are also commonalities at network level, given that both 3G cellular and 3G TV networks use ATM to move traffic to the broadcast transmitters (3G TV), and to and from the Node Bs, RNCs, and the core network in a 3G cellular network.

The DAB/DVB radio physical layer is based on a relatively complex frequency transform, which is hard to realize at present in a cellular handset transmitter (from a power efficiency and processor overhead perspective). We are therefore unlikely to see 3G cellular phones supporting a DVB radio physical layer on the uplink, but the phones could use the existing (W-CDMA or CDMA2000) air interface to deliver uplink bandwidth (subscriber-generated content). We will also look at parallel technologies (Web TV).

The Evolution of TV Technology

To date there have been three generations of TV technology (see Table 19.1):

Table 19.1 The Three Generations of TV

1930	1950/1960/1970/1980	2000/2002
Introduction of black-and-white 405-line TV	Introduction of UHF NTSC color TV in the United States and PAL in Europe and Asia. Switching off 405-line VHF transmission in the 1980s.	Introduction of digital TV (similar UHF frequencies as 2G TV)

- The start of black-and-white TV in the 1930s (fading out or being switched off in the 1980s)
- Second-generation color introduced in the 1950s in the United States (NTSC), and 1960s in Europe and Asia (PAL)
- Third-generation digital TV, presently being deployed at varying rates and with various standards in Europe, Asia, and the United States

Technology maturation in television works in 50-year cycles. Black-and-white 405-line TV survived from the 1930s to the 1980s. Analog color (introduced in the 50s and 60s) is still very much with us today—though it will be switched off in Europe by mandate by 2010.

Getting 2 billion people to change their TV sets takes a while. In practice, there has been an evolution of other new TV technologies, including Web TV and IPTV.

The Evolution of Web-Based Media

In 3G cellular networks and wireline telecom networks we have defined the changes taking place in the traffic mix—from predominantly voice-based traffic to a complex mix of voice, text, image, video, and file exchange.

A similar shift is occurring on the World Wide Web—initially a media dominated by text and graphics but now increasingly a rich media mix of text/graphics, audio, and video streaming.

As the percentage of audio, image, and video streaming increases on the Web, user quality expectations increase. These expectations include the following:

- Higher resolution (text and graphics)
- Higher fidelity (audio)
- Resolution and color depth (imaging)
- Resolution, color depth, and frame rate (video)

These expectations increase over time (months and years). They may also increase as a session progresses. Remember: It is the job of our application software to increase

session complexity—and by implication, session value. We want a user downloading from a Web site to choose:

- High-quality audio
- High-value, high-resolution, high color depth imaging
- High-value, high-resolution, high color depth, high frame rate video

Maximizing session complexity maximizes session value. As we make a session more complex, it should also become longer (more persistent). Also, as we increase session persistency, we should be able to increase session complexity, as shown in Figure 19.1.

We are trying to ensure that session delivery value increases faster than session delivery cost in order to deliver session delivery margin. Over the longer term (months/years), we can track how user quality expectations increase. As with cellular handsets, this is primarily driven by hardware evolution. As computer monitor displays improve, as refresh rates on LCDs get faster, as resolution and contrast ratio improves, as display drivers become capable of handling 24-bit and 32-bit color depth, we demand more from our application.

We don't always realize the quality that may be available to us. For example, we may have a browser with default settings, and we may never discover that these settings are changeable (arguably a failure of the browser server software to realize user value).

The browser describes the display characteristics preset, set by the user, or set by the user's device. This information is available in the log file that is available to the Internet service provider or other interested parties. This information can be analyzed to show how, for example, resolution settings are changing over time.

A present example would be the shift of users from 800 x 600 pixel resolution to 1280×1024 pixel resolution and 1152×864 pixel resolution.

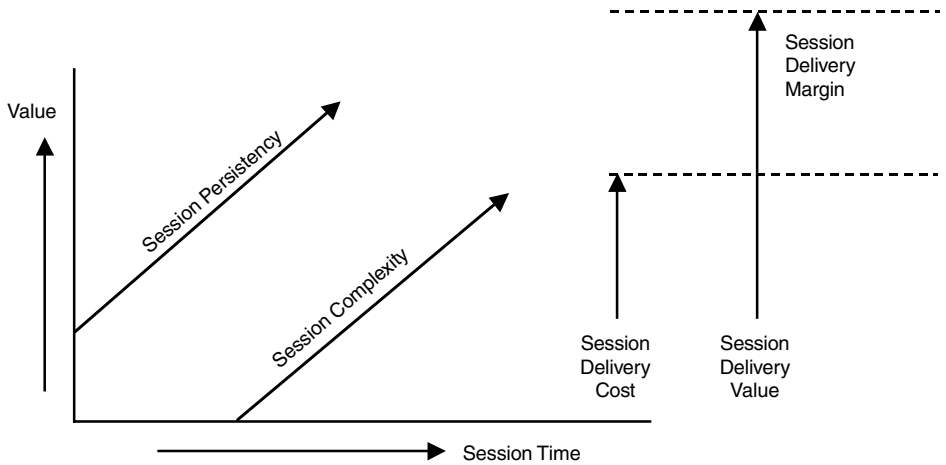


Figure 19.1 Session persistency and complexity.

Resolving Multiple Standards

As the Web moves toward being a rich media experience, it moves closer to television as an entertainment and information medium. This poses some challenges as to who should make the standards needed to provide interoperability and content transparency (the ability to deliver content to different devices without destroying content value). Figure 19.2 shows the different standards-making groups presently involved. They are as follows:

- The W3C (World Wide Web Consortium) has workgroups focusing on presentation layer standards (HTML).
- The IETF has workgroups presently trying to standardize the delivery protocols needed at the session layer, transport layer, and network layer to preserve the value of Web-based complex content as the content is moved into and through a complex network, including control of the real-time components of the rich media mix.
- The ATVEF (Advanced Television Enhancement Forum) is working on a standard for HTML-based enhanced television (XHTML).
- The SMPTE (Society of Motion Picture and Television Engineers) is working on a standard for declarative content.
- The digital TV industry, predominantly in Europe and Asia, is working on future digital TV standards including MPEG-4/MPEG-7/MPEG-21 evolution.

Ideally, the best work of each of these standards groups would be combined in a common standard. Although this may not happen, it is, however, well worth reviewing some of the common trends emerging.

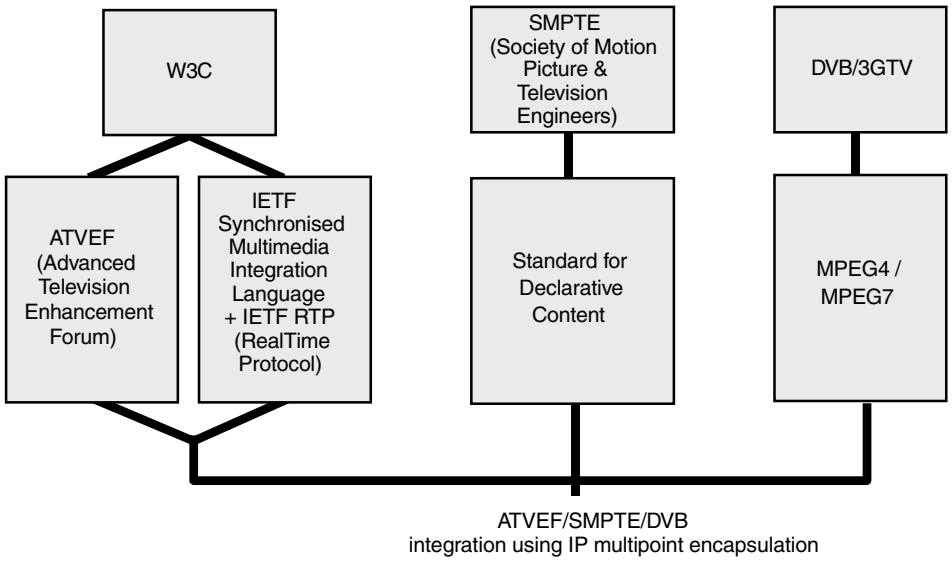


Figure 19.2 Multiple standards.

Working in an Interactive Medium

In common with TV, the Web has been to date largely a one-way delivery medium.

The Web, however, provides interactivity and an almost infinite choice of download material. Increasingly, the Web contains subscriber-generated content (an almost infinite choice of content of almost infinitely variable quality). This is due to the availability of low-cost hardware—Web cams and low-cost recording devices and the software associated with that hardware.

As mentioned, TV has traditionally been a one-way medium. Journalists produce content, which is then delivered to us. Increasingly, though, television is becoming a two-way medium based on subscriber-generated content or subscriber-generated entertainment. For example:

- Talk shows
- Phone-ins
- Fly-on-the-wall documentaries (or "reality shows")
- Game shows, where we vote for the winners and losers
- National IQ competitions

Both the Web and traditional TV can provide real-time or virtual real-time coverage of events—elections, sports fixtures, game shows, or competitions. These events have trigger moments when the subscribers are asked to vote. Trigger moments can generate a huge demand for instantaneous uplink bandwidth.

We may often charge a premium rate to subscribers for the right to vote, the right to contribute or comment. Uplink bandwidth is becoming a principal source of income. Unfortunately, however, we have designed our networks (Web delivery and TV networks) to be asymmetric in the downlink direction, whereas our value is increasingly generated in the uplink direction.

Delivering Quality of Service on the Uplink

This presents us with a major uplink Quality of Service issue. Broadcast triggers generate large numbers of uplink requests. We all want to be first to vote or bid or buy (in an online auction, for example). We all feel we have the right to be first to vote or bid or buy, but it may take several hours to deliver our vote (or our bid or buy request) because of the lack of instantaneous uplink bandwidth. The only satisfactory resolution to this problem is to dimension our uplink bandwidth to accommodate these peak loading points. This will, however, result in gross overprovisioning for the vast majority of the time (and add to uplink bandwidth cost).

We can ease the uplink problem a little by providing the illusion of interactivity. In Europe and Asia Teletext is an integral part of the TV experience. Teletext involves sending additional program information (information about programs and additional information for programs) on a subcarrier simultaneous with the mainstream programming. An example would be local weather. When watching the weather broadcast, viewers are told they can go to Teletext to look up local weather. This information has been pre-sent to and is stored in the receiver; it is then updated on a rolling basis.

Viewers can also look at information such as flight availability or flight status (incoming flight arrival times). This experience provides the illusion of interaction without the need for any uplink bandwidth. It is only when they choose to act on the information (book a flight, for example), that they need an uplink—typically a telephone link to the service provider.

The ATVEF Web TV Standard

The ATVEF is working on a standard very similar to Teletext intended for Web TV. Web TV receivers are required to have a kilobyte of memory available for session cookies (storage of short-lived, session-long attributes needed to support an interactive session). Receivers also need to be able to support 1 Mbyte of cached simultaneous content. The ATVEF is also trying to address the problem of large numbers of uplink requests responding to broadcast triggers. The only solution (assuming the service provider does not want to overprovision uplink access capacity) is to provide local buffering and a fair conditional access policy (priority rights).

Topics addressed by the ATVEF include how to place a Web page on TV, how to place TV on a Web page, and how to realize value from a complex resource stream. A complex resource stream is a stream of content that can capture subscriber spending power, which means the content has spending triggers—the ability to respond to an advertisement by making an instant purchase, respond to a charity appeal by making an instant donation, or respond to a request to vote by making a (premium charge) telephone call. The ATVEF has a broadcast content structure that consists of services, events, components, and fragments:

Service. A concatenation of programs from a service provider (analogous to a TV channel)

Event. A single TV program

Component. Represents the constituent parts of an event—an embedded Web page, the subtitles, the voice over, the video stream

Fragment. A subpart of a component—a video clip, for example

Integrating SMIL and RTP

In parallel, the IETF is working on a Synchronized Multimedia Integration Language (SMIL). SMIL supports the description of the temporal behavior of a multimedia presentation, associates hyperlinks with media objects and describes the layout of the screen presentation, and expresses the timing relationships among media elements such as audio and video (animation management). Note how important end-to-end delay and delay variability becomes when managing the delivery of this complex content, particularly when the content is being used as the basis for a complex resource stream whose sole purpose is to trigger a real-time uplink response, which means we are describing a conversational rich media exchange.

The intention is to try and integrate SMIL with the IETF Real-Time Protocol (RTP). The purpose of RTP is to try and determine the most efficient multiple paths through a network in order to deliver (potentially) a number of multiple per-user information

streams to multiple users. RTP allows but controls buffering and re-tries to attempt to preserve end-to-end performance (minimize delay and delay variability) and preserve real-time content value. The IETF describes this as IP TV.

The Implications for Cellular Network Service

So perhaps we should now consider what this means for cellular network service provision. The W3C is working on standards for Web TV. The IETF is working on standards for IP TV. Hopefully, these standards will work together over time. Web TV and IP TV are both based on the idea of multiple per-user channel streams, including simultaneous wideband and narrowband channel streams. This is analogous to SMS-coded data channels in IMT2000. We know that the IMT2000 radio physical layer is capable of supporting up to six simultaneous channel streams per user. We also know that the hardware (handset hardware) is, or at least will be, capable of supporting potentially up to 30 frames per second high color depth high-resolution video (for the moment we are rather limited by our portable processor power budget).

Media delivery networks depend on server architectures that are capable of undertaking massive parallel processing—the ability to support lots of users simultaneously. Note that each user could be downloading multiple channel streams, which means each user could be downloading different content in different ways. This is reasonably complex. It is even more complex when you consider that each user may then want to send back information to the server. This information will be equally complex and equally diverse.

To put this into a broadcasting context, in the United Kingdom, 8 million Londoners are served by one TV broadcast transmitter (Crystal Palace). This one transmitter produces hundreds of kiloWatts of downlink power to provide coverage. In the new media delivery network model, 8 million subscribers now want to send information back to that transmitter. The transmitter has no control over what or when those subscribers want to send.

The network behind the transmitter has to be capable of capturing that complex content (produced by 8 million subscribers), archiving that content, and realizing revenue from that content by redelivering it back to the same subscribers and their friends. This is now an uplink-biased asymmetric relationship based on uplink-generated value, realized by archiving and redelivering that content on the downlink.

Alternatively, it may be that subscribers might not want to use the network to store content and might prefer to share content directly with other subscribers (the peer-to-peer network model).

Either way, the model only works if the content is reasonably standardized, and this only happens if the hardware (and the software needed to manage the hardware) is reasonably standardized. Note this is harder to achieve if the hardware is flexible (reconfigurable), as this makes the software more complex, which makes consistent interoperability harder to achieve.

Television hardware is reasonably consistent, though it is less consistent than it used to be. For instance, we can now have 16:4 wide-screen TVs or 4:3 aspect ratio TVs, which complicates the content delivery process. But even taking this into account, cellular phones are far more diverse in terms of their hardware footprint—display size, aspect ratio, color resolution, processor bandwidth, and so on.

Device-Aware Content

This has resulted in the need to make content device-aware—using transcoders and filters to castrate the content. As we have argued, it is actually better to make the device content-aware and to make sure that user devices are reasonably consistent in the way in which they deliver content to the network. In other words, devices need to be able to reconfigure themselves to receive content from the network and need to be adaptive in the way in which they deliver content to the network.

This is made much easier if the devices share a common hardware and software platform. This either emerges through the standards process (unlikely) or via a de facto vendor dominance of the market. Note this might not be a handset vendor but will more likely be a component vendor. Intel's dominance of the computer hardware space is an example of de facto vendor-driven device commonalty. Similarly, it is the device (hardware) commonality that has helped Microsoft establish a de facto vendor-driven software/operating system.

Given the dominance of the DSP in present cellular phones and the likely dominance of the DSP in future cellular phones, it is most likely to be a DSP vendor who imposes this necessary hardware commonality.

The Future of Digital Audio and Video Broadcasting

In the meantime, we need to consider the future of digital TV (3G TV). Specifically, we need to consider the future of Digital Audio Broadcasting and Digital Video Broadcasting. Digital Audio Broadcasting is based on a 1.536 MHz frequency allocation at 221.296 to 222.832 MHz. This is shown in Table 19.2.

The physical layer uses QPSK modulation and then an Orthogonal Frequency-Division Multiplex (OFDM)—that is, a time domain to frequency domain transform, with the bit rate spread over either 1536 frequency subcarriers, or 768, 384, or 192 frequency subcarriers. The more subcarriers used, the longer the modulation interval and therefore the more robust the channel will be to intersymbol interference.

However, as we increase the number of frequency subcarriers, we increase the complexity of the transform and therefore increase processor overhead. This does not matter when we have one or several or a small number of mains-powered transmitters transmitting just a downlink. It would be computationally expensive and therefore power hungry to make the present receive-only receivers into receiver/transmitters. Apart from the processor overhead, the peak-to-average ratios implicit in OFDM would make power-efficient small handsets difficult to implement. This is why you do not have OFDM in any 3G cellular air interface.

Table 19.2 European Digital Audio Broadcasting Modulation Options

NO OF CARRIERS	1536	768	384	192
Modulation interval	1.246 ms	0.623 ms	0.312 ms	0.156 ms
Guard interval	0.246 ms	0.123 ms	0.062 ms	0.031 ms
Modulation	QPSK	QPSK	QPSK	QPSK

Table 19.3 TV Bands Across Europe

TV BAND	FREQUENCY	CHANNELS
I	47-60 MHz	2-4
III	74-233 MHz	5-11
IV	470-790 MHz	21-60
V	790-8672 MHz	40-69

OFDM does deliver a very robust physical channel. It trades processor overhead against channel quality and is therefore a strong contender for 4G cellular air interfaces (by which time processor overhead and linearity will be less of an issue). The DAB channel delivers a gross transmission rate of 2.304 Mbps and a net transmission rate of 1.2 Mbps that supports up to 64 audio programs and data services or up to 6×192 kbps stereo programs with a 24 kbps data service delivered to mobile users traveling at up to 130 kph.

Digital TV uses a similar multiplex to deliver good bandwidth utilization and good radio performance, including the ability to support mobility users. Four TV bands were allocated across Europe by CEPT in 1961 for terrestrial TV, as shown in Table 19.3.

In the 1990s proposals were made to refarm bands 4 and 5, dividing 392 MHz of channel bandwidth into 8 MHz or 2 MHz channels. This was to be implemented in parallel with a mandated shutdown of all analog transmissions by 2006 in some countries (for example, Finland and Sweden) and by 2010 in other countries (for example, the United Kingdom). In addition, some hardware commonality was proposed between digital terrestrial TV (DVB-T), satellite (DVB-S), and cable (DVB-C), including shared clock and carrier recovery techniques.

Partly because of its addiction to football and other related sports (Sky TV, the dominant DVB-S provider, bought out most of the live television rights to most of the big sporting fixtures), the United Kingdom now has one of the highest levels of digital TV adoption in the world (over 40 percent of all households by 2002).

DVB-S, DVB-C, and DVB-T use different modulation techniques. DVB-S uses single-carrier QPSK, DVB-C uses single-carrier 64-level QAM, and DVB-T uses QPSK or QAM and OFDM. This is due to the physical layer differences. The satellite downlink is line of sight and does not suffer from multipath; the cable link by definition does not suffer from multipath. Terrestrial suffers from multipath and also needs to support mobile users, which would be difficult with satellite (budgeted on the basis of a line-of-site link) and impossible for cable.

In DVB-T, transmitter output power assumes that many paths are not line of site. For DVB-S, downlink power is limited because of size and weight constraints on the satellite. Terrestrial transmitters do not have this limitation (though they are limited ultimately by health and safety considerations). As in DAB, the carrier spacing of the OFDM system in DVB-T is inversely proportional to the symbol duration. The requirements for a long guard interval determines the number of carriers: A guard interval of $\approx$250 μs can be achieved with an OFDM system with a symbol time of $\approx$ 1 ms and hence a carrier distance of $\approx$ 1 kHz resulting in $\approx$ 8000 carriers in an 8-MHz-wide channel. The

OFDM signal is implemented using an inverse Fast Fourier Transform, and the receiver uses an FFT in the demodulation process. The FFT size 2^N, where N = 11 or N = 13 are the values used for DVB-T. The FFT size will then be 2048 or 8192, which determines the maximum number of carriers. In practice, a number of carriers at the bottom and top end of the OFDM spectrum are not used in order to allow for separation between channels (guard band).

Systems using an FFT size of 2048 are referred to as 2k OFDM, while systems using an FFT size of 8192 are referred to as 8k OFDM. The more subcarriers used, the longer the guard interval duration, and the longer the guard interval duration, the more resistant the path will be to multipath delivery. There is some lack of orthogonality between the frequency subcarriers that is coded out—hence, the term COFDM: coded orthogonal frequency-division multiplexing.

Planning the Network

Although this may seem rather complicated, it makes network planning quite easy. In many cases, you can implement a single-frequency network using existing transmitter sites (assuming the 8k carrier is used). More frequency planning is needed for the 2k carrier implementation. The guard interval can be used to compensate for multipath between the transmitter and a receiver and the multipath created from multiple transmitters transmitting at the same frequency. Presently, Europe is a mixture of 8k and 2k networks, the choice depending on the coverage area needed, existing site locations, and channel allocation policy.

We said that the modulation used in DVB-T is either QPSK or QAM. Actually, there are a number of modulation, code rate, and guard interval options. These are shown in Table 19.4.

Going from QPSK to 16-level QAM to 64-level QAM increases the bit rate but reduces the robustness of the channel. Remember, the link budget also has to be based on only some users being line of site (satellite tends to win out in the bit rate stakes). This relationship is shown in Table 19.5. The link budget needs to be increased in a 2k network when higher-level modulation is used and depending on what coding scheme is used. In (1) we have a heavily coded (2 bits out for every 1 bit in) 16 QAM channel. Reducing the coding overhead, in (2), but keeping the 16 QAM channel requires an extra 4 dB of link budget, though our user data throughput rate will have increased. (3) shows the effect of moving to 64-level QAM and using a 2/3 code rate (3 bits out for every 2 bits in). The user data rate increases but an extra 4 dB of link budget is needed.

Table 19.4 DVB-Modulation Options

DVB-T		
QPSK	10.6 Mbps	More robust
16 QAM	21 Mbps	↑
↓		
64 QAM	30 Mbps	Less robust

Table 19.5 2K Carrier Comparison (NTL)

	MODULATION	CODE RATE	FAILURE POINT (C/I NEEDED)	DATA RATE
(1)	16 QAM	1/2	12 dB	12 Mbps
(2)	64 QAM	3/4	16.5 dB	18 Mbps
(3)	16 QAM	2/3	20 dB	24 Mbps

Because DVB-T needs to cope with moving objects in the signal path and user mobility, additional coding and interleaving is needed, together with dynamic channel sounding and phase error estimation, which means the use of pilot channels consisting of scattered pilots—spread evenly in time and frequency across OFDM symbols for channel sounding—and continuous pilots—spread randomly over each OFDM symbol for synchronization and phase error estimation.

The idea of the adaptive modulation and adaptive coding schemes is to provide differentiated Quality of Service with five service classes:

LDTV. Low definition (video quality)—multiple channel streams

SDTV. Standard definition—multiple channel streams

EDTV. Enhanced definition—multiple channel streams

HDTV. High definition—single channel stream 1080 lines; 1920 pixels per line

MMBD. Multimedia data broadcasting (including software distribution)

Conditional access billing will then be based on the service class requested or the service class delivered (or both). We use the future tense because, presently, the link budget—determined by the existing transmitter density and present maximum power limits—prevents the use of the higher-level modulation and coding schemes. This has, to date, prevented the deployment of EDTV or HDTV—the cost of quality bandwidth. However, the building blocks exist for quality based billing based on objective quality metrics (pricing by the pixel).

In the transport layer, digital TV is delivered in MPEG-2 data containers—fixed-length 188-byte containers. ATM is then used to carry the MPEG-2 channel streams between transmitters. The MPEG-2 standard allows for individual packet streams to be captured and decoded in order to capture program specific information. This forms the basis for an enhanced program guide, which can include network information, context and service description, event information, and an associated table showing links with other content streams—a content identification procedure. We can therefore see a measure of commonality developing both at the physical layer and at the application layer between 3G cellular and 3G TV. Physical layer convergence is unlikely to happen until next-generation (4G) cellular, since the COFDM multiplex is presently too processor-intensive and requires too much linearity to be economically implemented in portable handheld devices.

There is, however, convergence taking place or likely to take place at the application layer and transport layer. At the application layer, digital TV is establishing the

foundations for quality-based billing, which will be potentially extremely useful for cellular service providers. At the transport layer, both 3G cellular and 3G TV use ATM to multiplex multiple channel streams for delivery to multiple users. The IU interface in 3G cellular and 3G TV uses 155 Mbps ATM to move complex content streams into and through the network, from the RNC to and from the core network in 3G cellular and between transmitters in 3G TV.

3G cellular and 3G TV both use MPEG standards to manage the capture and processing of complex content—the rich media mix. 3G TV has lots of downlink bandwidth (392 MHz) and lots of downlink RF power (hundreds of kiloWatts). 3G cellular has plenty of bandwidth but not a lot of power (10 or 20 W on the downlink and 250 mW on the uplink). 3G cellular, however, does have a lot of uplink bandwidth (60 MHz), which digital TV (3G TV) needs. This relationship is shown in Figure 19.3.

Digital TV needs uplink bandwidth because uplink bandwidth is the mechanism needed to capture future subscriber value. Digital TV needs a return channel, preferably a return channel that it owns or has control of (that does not need to share uplink revenue with other transport owners).

Many cellular operators already share tower space with TV transmission providers. Most of the dominant tower sites were set up by the TV broadcasting companies many years ago. Some companies, the BBC, for example, have sold off these towers for use by TV and cellular providers. There are already well-established common interests between the two industries. Content convergence will help to consolidate this common interest.

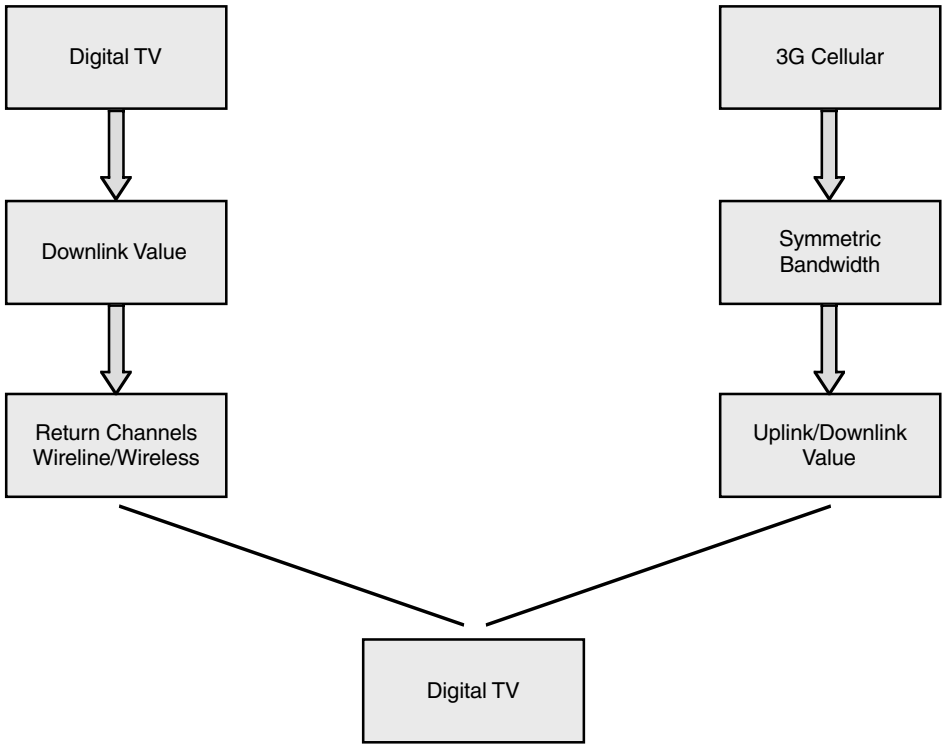


Figure 19.3 Downlink/uplink value.

In the United States, digital TV is experiencing slower adoption. U.S. digital TV standards are the responsibility of the Advanced Television Systems Committee (ATSC). The standard is very different to and in competition with the European and Asian DVB standard. The physical layer uses a modulation known as 8 VSB (vestigial side band) rather like single side band AM with a 6 MHz channel to deliver a gross channel rate of 19 Mbps. It is not designed to support mobile reception. The typical decoder power budget is 18 Watts, so, as with DVB, inclusion of a digital TV in a digital cellular handset is presently overambitious. It is difficult to see how the economics of digital TV using the ATSC standard will be sustained in the United States without access to the rest of world markets presently deploying DVB-T, DVB-S, or DVB-C technologies.

From 2010 onward there would appear to be very clear technology advantages to be gained by deploying integrated 3G TV and 3G cellular networks and moving toward a common receiver hardware and software platform. European and Asian standards making (including the MPEG standards community) is beginning to work toward this goal. The proximity of spectral allocation in the 700- and 800-MHz bands will provide a future rationale for 3G TV/3G cellular technology integration.

The Difference Between Web TV, IPTV, and Digital TV

Traditionally the Web has been described as a lean-forward experience, which means you are close to a monitor and keyboard, searching for a particular topic to research, using a search engine or going from Web site to Web site using hyperlinks. The time spent on any one Web site is typically quite short. TV is often described as a lean-back experience—more passive. You sit back and watch a program at a defined time, for a defined time. The program is broadcast on a predetermined predefined RF channel (DVB-T or DVB-S broadcast). Session length can typically be several hours.

IP TV gives you access to either Web-based media or traditional TV content. Expatriates often use IP TV to access favorite radio and TV programs from their home country that are not available locally. It provides access via the Internet to a variety of media from a variety of sources. However, these distinctions are becoming less distinct over time.

The job of a Web designer is to increase the amount of time spent by visitors on a site. Web casting particular events at a particular time (a Madonna concert, for example) is becoming more popular. The mechanics of multicasting—delivering multiple packet streams to multiple users—are being refined.

In parallel, digital TV is becoming more interactive. Access to the Internet is available through cable TV or via a standard telephone connection, and TV programs have parallel Web sites providing background information on actors, how the program was made, and past and future plotlines.

Probably the most visible change will be display technology. High-resolution widescreen LCD or plasma screens for home cinema requires ED TV or HDTV resolution (and high-fidelity surround sound stereo). Equally significant will be the evolution of storage technology (the availability of low-cost recordable/rewritable DVDs) and the further development of immersive experience products (interactive games). Note the intention is to use Multimedia Data Broadcasting (MMDB) for software distribution, which will include distribution of game infotainment products.

So it all depends on how attached we become to our set-top box and whether we want to take it with us when we go out of the house. We can already buy portable analog TVs (and some cell phones—for example, from Samsung—integrate an analog TV receiver chip). These products do not presently work very well. As digital TV receiver chip set costs reduce and as TV receiver chip set power budgets reduce, it will become increasingly logical to integrate digital TV into the handset—and it will work reasonably well because DVB-T has been designed to support mobility users.

Additionally, digital TV and Web TV can benefit from capturing content from digital cameras embedded into subscriber handsets—two-way TV (uplink added value).

We are unlikely to see any short-term commonality of the radio physical layer because of the processor overheads presently implicit in OFDM, including the requirements for transmit linearity. Use of OFDM in fourth-generation cellular would, however, seem to be sensible.

Co-operative Networks

A number of standards groups are presently involved in the definition of co-operative networks. Co-operative networks are networks that combine broadcast networks with 3G cellular or wireless LAN network technologies.

There are a number of European research projects also presently under way, including DRIVE (Dynamic Radio for IP Services in Vehicular Environments) and CISMUNDUS (Convergence of IP-Based Services for Mobile Users and Networks in DVB-T and UMTS Systems). The research work focuses on the use of a 3G cellular channel as the return channel with the downlink being provided by a mixture of DAB and DVB transmission, collectively known as DXB transmission.

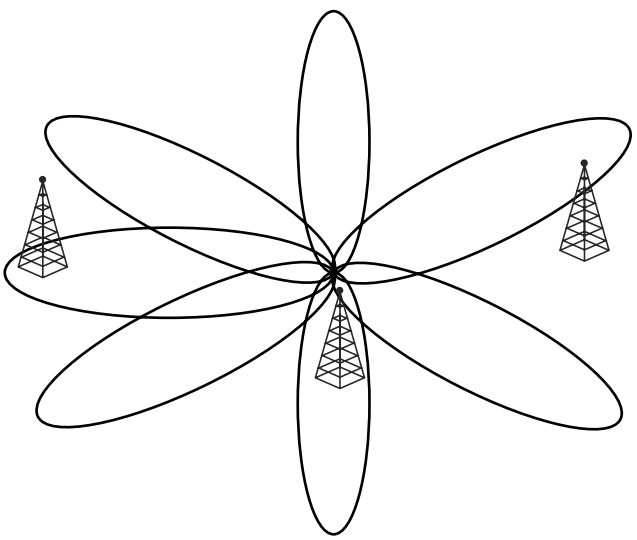


Figure 19.4 Data-cast lobes.

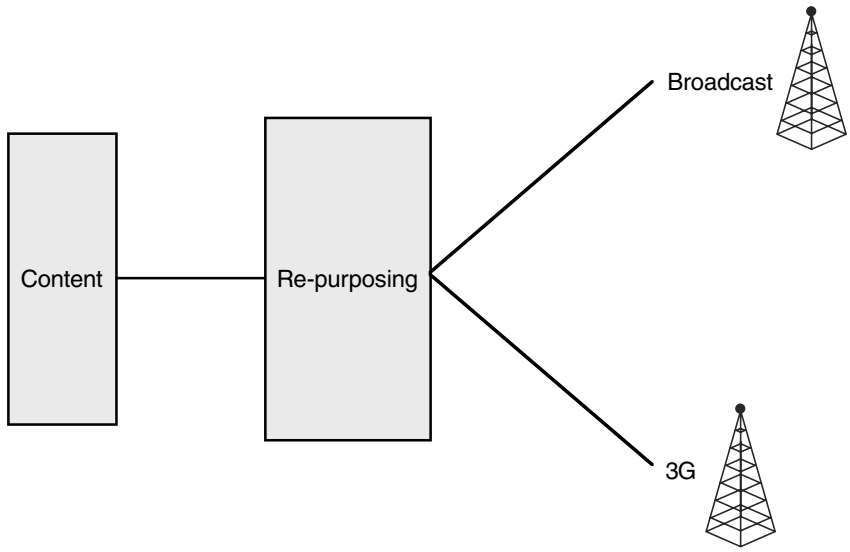


Figure 19.5 Content re-purposing.

A European forum known as the IP Data Casting Forum has proposed that DVB and DAB be used to deliver IP data using data cast lobes—reengineering the broadcast transmitters to provide some directivity to support localization of deliverable content. Data-cast lobes are shown in Figure 19.4. As with cellular networks, this would also allow for a more intensive use of the UHF digital TV spectrum.

The dynamic range of display technologies involved (high-quality CRT, high-quality LCD, low-quality LCD), including a range of aspect ratios (16×9, 4×3, 1×1), requires a measure of re-purposing of content, as shown in Figure 19.5.

This is similar in concept to the transcoding used in the Wireless Application Protocol. Care will need to be taken to maintain consistency of delivered content quality. Co-operative networks will need to be consistently co-operative.

Summary

Over the next 10 years there are increasing opportunities to develop physical layer commonality between 3G TV and 3 or (more likely) 4G cellular networks. There are also commonalities in the way that traffic is handled and transported across digital TV and 3G cellular networks using ATM cell switching. In addition, digital TV is increasingly dependent on uplink-generated value, which is available, as one option, on the cellular uplink. Trigger moments in TV content can, however, create a need for large amounts of instantaneous uplink bandwidth, and care needs to be exercised to ensure adequate uplink bandwidth and uplink power is available to accommodate this.

Highly asynchronous and at times highly asymmetric loading on the uplink places particular demands on the design of network software and its future evolution—the subject of our next and final chapter.

CHAPTER

20

Network Software Evolution

In the previous chapter we talked briefly about TV network hardware and TV network software, and the parallel development of Web TV and IP TV. The products delivered (TV programs or Web-based content) can either be real time (broadcast when an event happens), near real time (broadcast within a defined time delay—milliseconds, seconds, minutes, or hours after the event), or non-real time. Non-real time includes content that is not event based but, rather, information that is collated and stored to be downloaded on demand at some indeterminate time in the future. Near real time requires buffering. Non-real time requires buffering and long-term storage.

Generically, these networks can be described as media delivery networks. We made the general point that these networks are also able to capture content from subscribers, store that content, and redeliver it back to the originating subscribers and other interested parties—two-way TV.

In this final chapter, we look at *service delivery networks*—networks that can support a variety of one-way and two-way services. We also want to highlight storage bandwidth as a mechanism for realizing subscriber asset value.

A Look at Converging Industries and Services

We argue that there are six industries presently converging:

- The computer industry
- The consumer electronics industry

- The IT industry
- The wireless industry
- The wireline industry
- The TV industry

The Internet is a point of intersection between these industries. All six industries use storage as part of their added value proposition.

Services provided include application service provision, Internet service provision, and enterprise service provision. The computer industry tends to focus on application service provision; the consumer electronics industry tends to focus on Internet service provision; the IT industry tends to focus on enterprise service provision.

If we deliver these services over a wireless network we call them:

WASPs. Wireless application service providers

WISPs. Wireless Internet service providers

WESPs. Wireless enterprise service providers

Managing Storage

Storage can be either network based or subscriber based—for example, in the subscriber's handset. Storage bandwidth needs to be managed using storage software. Access to storage may be defined by a storage SLA (covered briefly in Chapter 18). The job of storage software is to extract value from storage bandwidth. Search engines are an example of how to realize storage bandwidth value (although in practice, with one or two exceptions, to date it has proven hard to turn user value into network dollar value).

Managing Content

We can produce content or get other people to produce content on our behalf—including our subscribers—and produce copies of the content that we can distribute toward the point of final consumption—setting up mirror Web sites that, for example, replicate content on a country-by-country basis. The replicated content may be localized, and we may add local language content. The management of content includes filtering, access priority, and delivery priority—storage service attributes. Storage information can be organized using meta data tags, labeling, cataloging, and describing content so it can be retrieved consistently (known within W3C as the "resource description framework").

Information can be geocoded so that it can be scanned against geographic location—by longitude, latitude, or height (geospatial coding). Consider that a 3G handset using Assisted GPS (A-GPS) can get a positional fix within 10 to 20 meters within 100 ms at a power budget cost of about 200 mW per fix, which means highly accurate, fast, and usable positioning. Accuracy with the European Galileo satellite system (in service by 2008) will be ± one meter. This information can be used as the basis for locality service platforms.

Using Client/Server Agent Software

Content delivery is all about not throwing content value away. Agent software acts as an intermediary between both parties. The agent software can alternatively be embedded directly on the server or client platform or become part of a peer-to-peer relationship. Consider how agent software can help us add value:

- We can add value to content by using knowledge management techniques—the process of gathering, organizing, refining, and disseminating information.
- We can pull information from subscribers (gathering).
- We can catalog, filter, link, and index information (organizing).
- We can contextualize, project, or compact information (refine).
- We can push or share information or enable collaboration (dissemination).
- We can discover and realize value from patterns and trends in stored information by using data mining techniques.
- We can group information sources together in terms of their attributes.
- We can classify.
- We can codify information.
- We can determine subscriber behavior patterns and optimize the storage and distribution of information on the basis of the knowledge that we can acquire about a user.
- We can use this information to store relevant information physically close to the user.
- We can invest in storage rather than delivery bandwidth. However, there is not much point storing information if we cannot serve it to people, so we have to invest in server bandwidth.

The following list shows the different jobs we may want our server to perform depending on whether we are a WASP, WISP, or WESP. In effect, we need a server to deliver services—the service delivery network proposition. The types of server needed will include:

- File server
- Web server
- Database server
- Application server
- Groupware server
- Print server
- Transaction server
- Object server
- Consolidated server (a combination of all the above)

Delivering Server and Application Transparency

The problem is how to deliver service transparency and application transparency. This is the issue that Sun addresses with Jiro. Jiro is doing what Java has done for software—the Write Once, Run Anywhere philosophy (WORA), becoming the store once, share anywhere philosophy (SOSA).

Jiro proposes the use of agents to resolve application incompatibility and application priority issues. It is known as a Federated Management Architecture (FMA) and uses SNMP (Simple Network Management Protocol) or WBEM (Web-Based Management Protocol) as a basis for managing resources. To put this into the context of our OSI seven-layer model, Jiro adds a Management Logic layer, an Agent layer, and a managed resource layer into the model, thereby becoming a nine-layer model, as follows:

- Application layer
- Presentation layer
- Management logic layer
- Agent layer
- Managed Application Resource layer
- Transport layer
- Network layer
- Data Link layer
- Physical layer

The Management Logic layer allows, or should allow, management services to be distributed across a network in a reasonably consistent way and allows storage services to be managed automatically. This includes the management of storage access policy.

Storage Area Networks

In addition to being described as a service delivery network, the network may be described as a *Storage Area Network* (SAN), sometimes also known as a Storage Access Network. Key performance metrics in a SAN could be order processing times or transaction times—the measure of SAN bandwidth quality.

Processing delay and transaction delay are generally the result of server bandwidth limitations. Additional performance metrics include reliability and availability of the server and application interoperability. (There's not much point in an application running quickly if it won't talk to other applications.)

Application interoperability also requires a consistent implementation of authentication and encryption—the management of security permissions. This includes front-door checks (controlling user access) and backdoor checks (controlling access by the service developer). Essentially, the code has a digital signature, the author of the code has a digital signature, and the user of the code has a digital signature—the basis of a security policy.

Application Persistency

The policy is dependent on the persistency of the application. Application persistency is often similar to and may be the same as session persistency. Application persistency is influenced by storage bandwidth, server bandwidth, and bandwidth availability. For example, an application may crash because insufficient storage capacity or server capacity is available.

The job of Jiro would be to detect that this has happened or predict that it will happen and try to solve the problem. In more general terms, Jiro is designed to provide automated policy-based management of any storage device or data service running on any operating system on any network, or, to quote the Jiro mantra, "platform independence for storage."

Figure 20.1 shows how the Jiro management logic area sits on top of the managed resource layer using what's known as the Common Information Model (CIM) and WBEM (Web-Based Enterprise Management). Because of their position between higher-layer software and lower-layer execution protocols, these products are sometimes known as *middleware*. The middleware in particular is responsible for application performance monitoring and intelligent backup—that is, system recovery after a system crash. Additional case study information is available either on the Jiro site (www.Jiro.com) or from the Storage Networks Industry Association (www.snia.org) or the Distributed Management Task Force (www.dmtf.org).

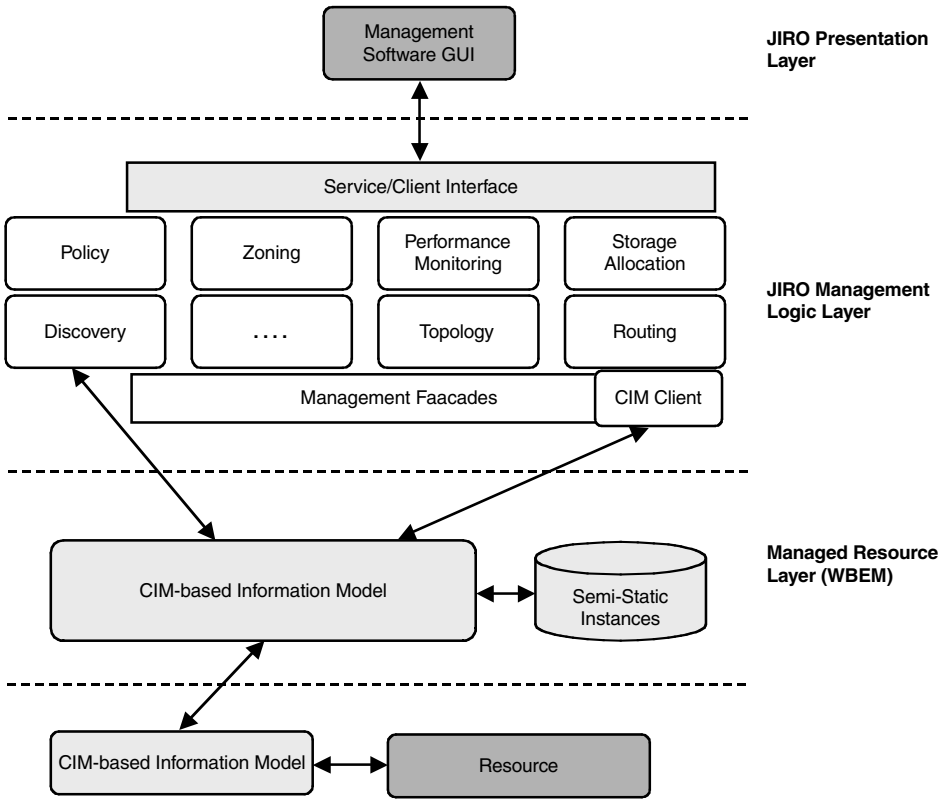


Figure 20.1 The common information model including Jiro.

Interoperability and Compatibility

We have used Sun's Jiro as an example of IT management. We could equally well have used Microsoft, which has similar products, including ActiveX and Microsoft Active Directory. Whereas Sun offers interoperability between Java and Jiro products and Unix-based operating systems, Microsoft offers interoperability between Microsoft products, including Active X, Active Directory, and Windows NT. Neither Sun nor Microsoft necessarily guarantees interoperability between each other's products. Given that most networks support a mixture of Windows NT and Unix-based operating systems, software interoperability remains and will continue to remain an issue.

Software has the benefit of being relatively easy to reconfigure. Configurability does not confer compatibility—rather the opposite. Some of the components that require compatibility in a cellular network are shown in Figure 20.2.

The Relationship of Flexibility and Complexity

Software flexibility has had to increase as hardware flexibility has increased. Software complexity has increased as hardware complexity has increased. Hardware complexity and hardware flexibility allows us to support highly bursty application bandwidth, which in turn increases application complexity and application flexibility. This relationship is shown in Figure 20.3.

A complex application is an application that draws on a variety of sources—audio, image, video, and data—which in turn demand either an instantaneous or noninstantaneous allocation of network resources. Network resources in a wireless network include radio bandwidth allocation, network delivery bandwidth allocation, and, optionally, allocation of network storage resources (managed by network storage software).

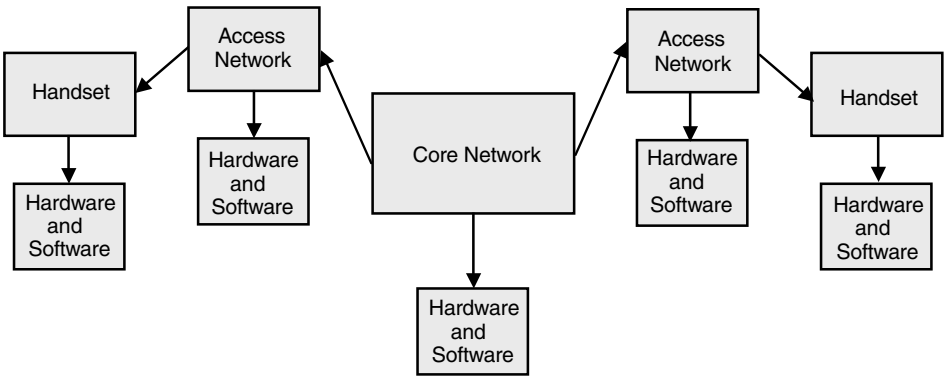


Figure 20.2 Component compatibility in a cellular network.

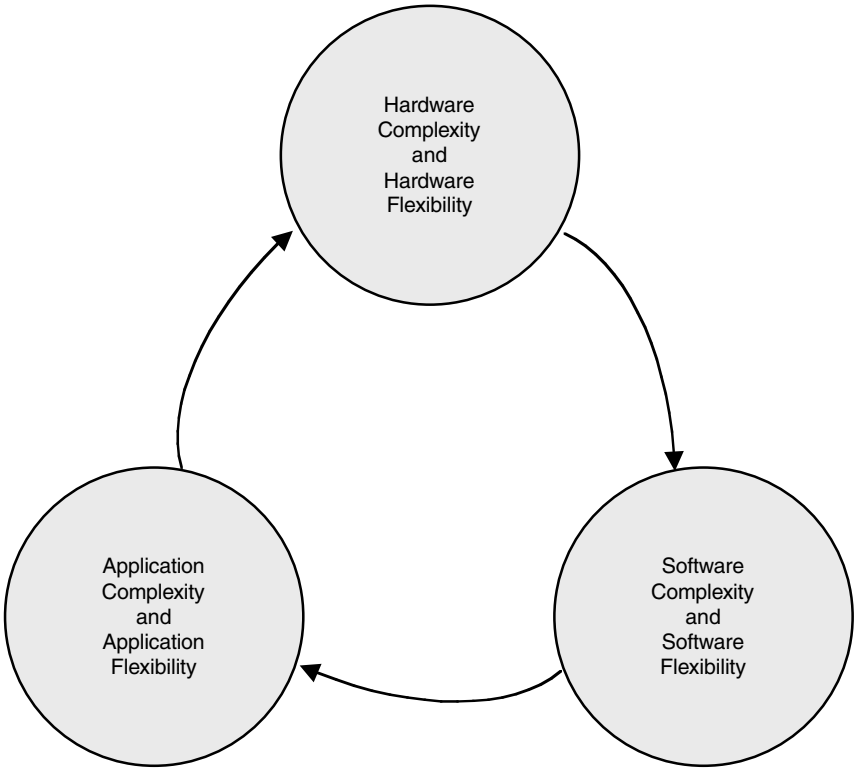


Figure 20.3 Complexity relationships.

Wireless and wireline networks already make money out of storage:

- Voice messaging uses storage resources.
- A store-and-forward service (SMS) uses storage resources.
- Downloadable ring tones use storage resources.
- Backing up subscriber telephone address books in the network uses storage resources.

Remember also, stored information can gain value over time (storage asset value appreciation), provided we can maintain storage stability (memory volatility) and access capability. The increasing diversity of storage media, both solid-state and disk-based, has created a number of interoperability issues partly, though not totally, addressed by existing vendors. Storage compatibility and application compatibility will continue to be an important network quality metric with a direct and indirect impact on delivered service quality—particularly service consistency.

Hardware compatibility issues tend to be compounded rather than resolved by software compatibility issues. In addition, as we have said before, software compatibility is harder to test and to prove.

Network Software Security

We have argued the case for configurability and have pointed out that configurability has a cost in terms of compatibility. Configurability also has a cost in terms of security: the need to consider the impact of computer viruses on network software integrity. In an IT network, antivirus products can be positioned at a number of points, shown in Figure 20.4.

For example, we could expect to see antivirus products at the gateway, at the server, and at the desktop. Because PDAs have traditionally had their operating systems embedded in read-only memory, it has never been considered necessary to provide protection in remote access (wireless) devices. However, adding over-the-air configurability to these devices makes them vulnerable to infection. Antivirus software, however, introduces additional overheads including access delay and access delay variability, memory (storage of virus signatures) and processor overhead. For additional background on network software virus control, try www.symantec.com or www.sophos.com.

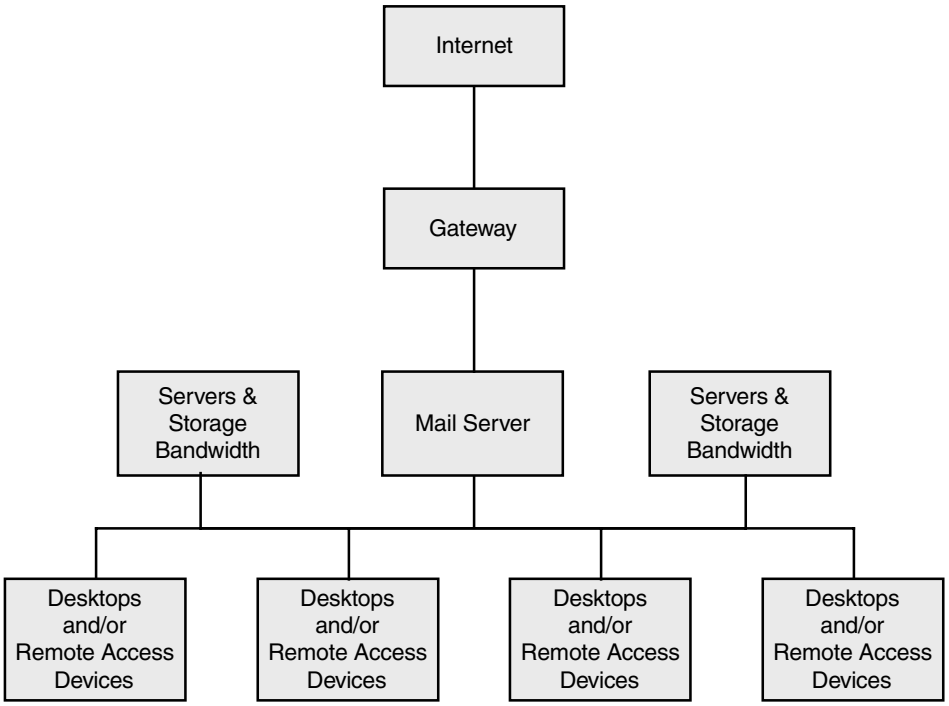


Figure 20.4 The positioning of security products.

Model-Driven Architectures

We have said that we are trying to deliver bandwidth flexibility, configurability, and compatibility. Software helps us deliver flexibility but is intrinsically difficult to test and can, in particular, create interoperability issues. The more lines of code, the harder it is to get software components to communicate consistently under all possible operating conditions.

The Object Management Group (www.omg.org) is a group of vendors presently trying to address the software interoperability issue by introducing *Model-Driven Architectures*. This is an IT industry initiative developed from CORBA (Common Object Request Broker Architecture) and OMA (Object Management Architecture)—two earlier attempts to improve intervendor software interoperability. A side objective is to decrease the cost of software development by reducing the duplicated effort that tends to be part of the process of software component production.

The quite reasonable assumption is that although applications may seem on the surface to be very different (a police control room would seem to be very different from an air traffic control system, which would seem to be very different from a sewage and waste management system, which would seem to be very different from a telecom network), they do in practice have a number of similarities as far as underlying software processes are concerned and can be modeled using a Unified Modeling Language. The Unified Modeling Language has to be used in parallel with the Meta Object Facility (MOF) and Common Warehouse Metamodel (CWM)—that is, the interoperability of software controlled memory and delivery bandwidth.

The standard describes *pervasive services*. These are services that can be found generically in almost all application domains—for example, directory services, event handling, persistence, transactions, and security with standardized domain models for specific vertical industries. Potentially, this ought to mean that a telecom operator could aggregate a number of vertical market corporate, personal subscriber, and specialist user software solutions from a number of different vendors with a reasonable chance that the applications might work with each other. For example, the police service solution might be expected, quite reasonably, to work with the fire service solutions platform even with different vendors involved.

There is still a fairly big mountain to climb: computer systems have not historically been as consistently reliable as telecom networks, and as Mac and PC users the world over will testify, have never been particularly interoperable. Telecom networks, however, increasingly use store-and-forward services to move content into and through the network and, as such, have needed to take on the job of managing network storage—and, by implication, storage performance metrics such as access delay and access delay variability, access control, and access security.

Testing Network Performance

Even with well-designed software, we still need to be able to test and measure network performance. Compatibility and consistent hardware and software performance depends on functional testing and interoperability testing. Specifically:

- We need to test how individual hardware and software components perform.
- We need to test how well individual hardware and software components work with each other.
- We need to ensure handset hardware works with handset software and that network hardware works with network software.
- We also need to ensure handset and network hardware compatibility, and handset and network software compatibility.

These relationships are shown in Figure 20.5.

We need intervendor compatibility for handset hardware and software. Ideally, we also benefit from intervendor compatibility for network hardware and software. This has been hard to achieve consistently in 2G and 2.5G networks and will be harder to achieve in 3G networks because of the increase in network software component count and complexity.

The Challenge of Software Testing

In general, software is harder to test than hardware. Hardware testing is relatively deterministic. It either works or it doesn't under a predetermined set of operational conditions, such as temperature and humidity. We can drop test, heat test, and splash test handset and network hardware. We can test individual hardware components—passive and active components, DSPs, memory, and microcontrollers against precise pre-agreed performance specifications.

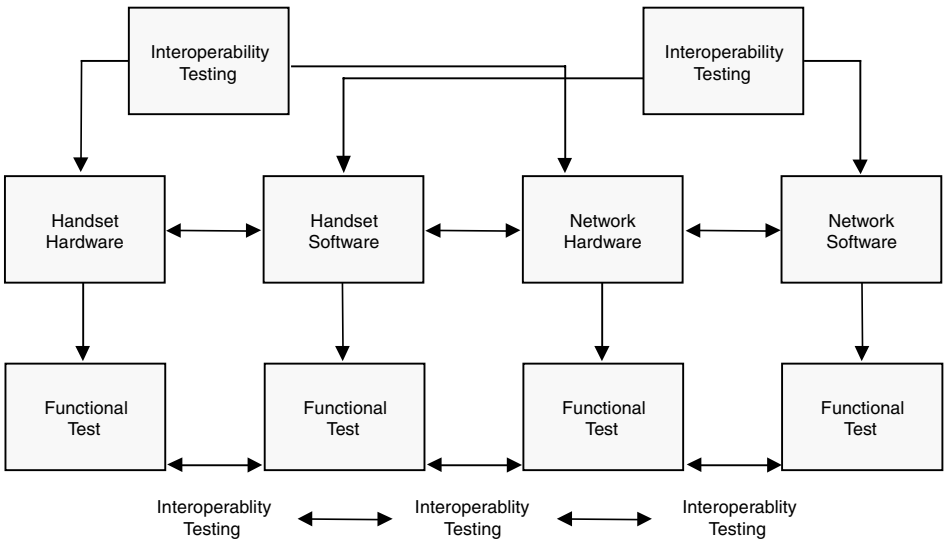


Figure 20.5 Test and measurement relationships.

Software by its nature is less deterministic. We cannot always simulate or predict software behavior under all operational conditions. The way in which code has been written and compiled by individual vendors may determine the behavior of the code. Even with 2G handsets, the majority of product recalls and in-service problems tend to be related to software implementation. Given that the 2G to 3G transition involves an order of magnitude increase in handset code footprint (from 100,000 to 1 million lines of code), it is not unreasonable to expect an order of magnitude increase in software-related compatibility and performance issues.

Software compatibility and software performance issues remain with us as we move into the network. Consider, for example, the functions that need to be performed by the software in the RNC. The RNC accepts and manages traffic to and from the Node Bs (base station transceivers) and to and from the core network. In addition, the RNC has to supervise hard and soft handover and internode and inter-RNC load distribution.

The RNC is responsible for preserving the properties of the offered traffic going to and coming from each of the users served by each of the Node Bs under its control. This includes arbitrating access and policy rights and managing session quality for individual users including the allocation of radio and network resources to meet user application needs that are, potentially, constantly changing. The RNC has to decide on resource allocation and admission on the basis of interference measurements received from the radio physical layer and congestion measurements received from the network core. This is a dynamic, decision-intensive, software-intensive process.

Because it is difficult to define all possible loading and operational conditions, it is difficult to define how RNC software will behave. This makes it hard to agree on consistent intervendor rule sets against which RNC performance can be measured. It also makes it hard to define specific performance metrics. If we cannot define specific performance metrics for the RNC, we cannot test the RNC.

Given that the RNC substantially defines network performance—including the Quality of Service metrics that we are using as the basis for billing—we have by implication an intrinsically untestable network. Untestable or hard-to-test networks, or a lack of agreed upon performance metrics against which network performance can be tested, tends to result in performance and payment disputes, which are undesirable for all parties involved.

As with handset software, the 2G-to-3G transition (from 2G BSC to 3G RNC) implies an order of magnitude increase in RNC software component count and complexity. It is not unreasonable to expect an order of magnitude increase in software-related network compatibility and network performance issues.

Test Languages

Software testability extends to the need to agree on a consistent language to describe test requirements. Within telecom testing, a standard language known as TTCN has been used to describe tests for GSM and ATM, in addition to IPv6 and SIP. TTCN originally stood for Tree and Tabular Combined Notation but now stands for Testing and

Test Control Notation. It looks and feels like a programming language and can be used to specify interoperability testing, robustness testing, performance testing, regression analysis, system testing, and integration testing. However, TTCN can slow down the test process, and quite often test vendors start by supplying products that test what is known as a prose specification, or in other words, plain English.

Although TTCN is derived from the prose specification, it adds in low-level settings, which help provide a more detailed and consistent test definition. It helps to test the tests.

Measuring and Managing Consistency

In previous chapters we identified consistency as a key performance indice (KPI) in a 3G wireless network. We need, in parallel, to develop consistent ways of measuring and managing consistency.

Why Is Consistency Important?

The importance of consistency has already been clearly established in previous generations of cellular network deployment. Consistency in a 2G network is a product of voice quality and dropped call performance. Consistent voice quality and the ability to complete a call without interruption are still the best way to maintain an acceptable level of user/customer satisfaction. Note that even if voice quality is relatively poor, if it is consistent, we perceive the quality to be better than it actually is. Conversely, even if voice quality is good, if it is inconsistent, we perceive the overall quality to be poor.

3G Consistency Metrics

We have said that session consistency and session persistency provided the basis for building session value together with session immediacy and session complexity. To review:

- As session complexity increases, it becomes harder to maintain session consistency.
- As session persistency increases (as a session gets longer), it becomes harder to deliver consistency.
- As session immediacy increases, it becomes harder to deliver consistency.

We have maintained, and still maintain, that delay and delay variability degrade session value. The highest-value component in our offered traffic mix is a conversational complex content (rich media) exchange. The complex content exchange consists of time-sensitive, time-interdependent simultaneous audio, image, video, and data streaming. Conversational complex content cannot be buffered, since buffering introduces delay and delay variability. Conversational complex content can therefore only be delivered over a circuit-switched or closely managed ATM cell-switched transport layer or an IP session with equivalent control of end-to-end delay and effectively no end-to-end delay variability. Consistency, when considered in the context of a time-sensitive complex content exchange, is also critically dependent on the close control of admission and access policy.

The RNC has to manage complex content and complex admission control on the basis of interference measurements from the radio physical layer and congestion measurements from the core network. Inconsistent interpretation of this feedback information will deliver inconsistent network performance, which will translate into an inconsistent user experience.

Attempts to improve bandwidth utilization/bandwidth efficiency (to decrease delivery cost) generally have a performance cost in terms of loss of consistency. We can improve bandwidth utilization in an IP network by queuing. Queuing implies user access prioritization. It becomes difficult to maintain a consistent interpretation of access and policy rights, particularly when individual user requirements are constantly changing.

We can only deliver consistency if we have a high degree of control over network performance. It is difficult to have a high degree of control over IP network performance, particularly when we add the inherent inconsistency introduced by the radio physical layer (dropped calls and varying transmission quality due to the fading channel).

Consistency implies a transparent view of network bandwidth, which can only be securely delivered by using out-of-band signaling. This means the use of SS7 in a signaling plane physically separated from the traffic flow; the cost of consistency is significant signaling overhead.

We can compensate for a loss of consistency by rebating customers/users to whom we fail to deliver a pre-agreed level of service. Compensation, particularly the administration of rebates, however, incurs hidden costs, including the need to define the cost of managing customer complaints. As always, it remains good practice to provide consistent levels of service that are better than customers expect. If we provide inconsistent service, we need to ensure the level/degree of inconsistency is either unnoticeable or below the level of customer indifference.

Consistency is arguably the most underrated metric in present IP network performance planning. The addition of the radio physical layer adds to the problem. Consistency will also become increasingly important as content becomes more complex over time.

Summary

Over the past 20 chapters we have shown how the changing hardware and software form factor of the user's handset is changing the offered traffic mix and offered traffic value (subscriber-generated uplink added value). We have suggested that a shift is occurring from consumer electronics devices to creative electronics devices—devices that fire into the network. We have asserted that network-generated content does not have intrinsic value, that in order to increase future revenues, we need to create addiction and dependency, and that addiction and dependency are based on our instinctive need to create and share our creativity with other people. This in turn is influenced by the Internet shift toward a more egocentric, subscriber-centric value model where value is created by the consumer—for instance, phone-ins, chat groups, and Web cams.

We can realize subscriber-generated content value by preserving the properties of the content—the qualities of the rich media multiplex. We have suggested the need to

choose visible and tangible performance metrics—billing by quality not quantity. We have identified the components needed to deliver radio and network bandwidth quantity and quality and explored how bandwidth quantity and quality can be used to realize revenue return (subscriber asset value appreciation). We have also identified that delivery cost increases as quality increases.

We highlighted how difficult it is to combine flexibility, efficiency, and performance: how important it is to qualify delay metrics in a network having to capture, store, and deliver rich media products. We identified storage bandwidth as a mechanism for capturing and exploiting subscriber-generated content and the radio bandwidth and network bandwidth implications of the media mix shift. This included the need to provide sufficiently flexible bandwidth on demand—that is, variable-rate, constant-quality channels.

We suggested a shift from centralized to distributed value (as MIPS and memory migrate to the network edge, added value follows). We showed how added value is becoming centered on the subscriber device, with value moving from the core network to the RNC and Node B. We extended the concept of bandwidth management to include delivery and memory bandwidth quality metrics. We suggested that complex content needs a complex network and that complex networks need close control in order to provide quality-based billing opportunities. This explains why some vendors are placing an apparently disproportionate amount of effort on ATM deployment both in terms of network bandwidth and radio bandwidth deployment.

We summarized the interrelationship between the subscriber appliance and the network proposition and showed how six industries are effectively converging: computer, consumer electronics, IT, wireless, wireline, and TV. We talked about technology time lines—how the TV industry works in 50-year cycles, and how and why the cellular industry works on 15-year cycles (it is physically easier to change a cellular handset than to change a television).

The Phases of Cellular Technologies

We said that all cellular technologies go through the following three stages:

- The first 5 years are the pain phase (nothing works very well).
- The second 5 years are the pleasure phase (most things work quite well).
- The last 5 years are the perfection phase (everything works, but we expect the network to do things it wasn't designed to do).

Every time we introduce a new network technology (the pain phase), it has to compete with an existing technology introduced 10 years before (entering its perfection phase). Dates of introduction vary from region to region, but the 15-year rule still generally applies.

This problem is not exclusive to cellular. In fact, every new technology suffers from a reality gap. In the pain phase, technologies fail to live up to marketing promises or, more importantly, user expectations. In the pleasure phase, successful technologies start to deliver over and above user expectations. In the perfection phase, the introduction of new user requirements may overstretch the technology. Note some technologies never reach the pleasure phase. They never catch up with user expectations. We define these as failed technologies.

Successful technologies can also be defined as technologies that deliver value down the whole industry value chain—the device component vendor, appliance manufacturer, network manufacturer, network operator, and end user. Failure to deliver value at any stage of the value chain will mean the technology will fail. The wireless value chain relationship can be illustrated as follows:

- Silicon
- Appliance manufacturer
- Network manufacturer
- Network operator
- End user

Provided technologies get to the pleasure phase and achieve market volume, they will also benefit from volume-based cost and performance advantage. In RF products particularly, volume brings specific performance benefits (sensitivity, stability, selectivity) because of the ability to keep tighter control of production line component tolerance spread. You also tend to have more engineers working on device and design refinements and performance optimization.

Provided market volume is achieved, we can expect to see at least a dB a year of performance improvements from 3G handsets over the next 5 years. Performance will then stabilize or be compromised as we start to try and make the handsets do things they were not designed to do.

On the basis of the present maturation cycle, we can expect to be introducing fourth-generation cellular product in 2012. By this time, there will be substantial commonality between digital TV and cellular physical layer and application layer connectivity and network topology.

Software compatibility rather than hardware compatibility will emerge as one of the most important issues for network operators and their vendors. Flexibility and configurability do not imply compatibility. Flexibility and configurability do not imply interoperability. It will be hard to deliver deterministic performance. Software is not by nature particularly predictable or consistent. We need predictability and consistency in order to justify quality-based added value.

We will see a transition from circuit switching to session switching, with session value determined in terms of session persistency and session complexity. Session complexity implies multiple channels per user, each channel being variable bit rate—bursty bandwidth.

Bursty bandwidth is expensive bandwidth. It exercises all of the system components in our network—the RF amplifier in the handset, the Node B receiver, the ATM multiplex on the IUB and IU interface, and router bandwidth in the network.

We can only achieve deterministic performance in the network if we have a high level of control over end-to-end performance. This is difficult to deliver in a packet-routed network.

We need deterministic network performance to preserve the properties of conversational time-sensitive rich media. It may be that conversational rich media is a small part of our overall traffic mix, but it will potentially represent the highest value part of the traffic mix.

It is important to consider the relative cost of delivery. If the cost of delivery exceeds the tariff premium achievable, we may as well all go home.

Optimistically, we should be able to translate bandwidth quality into bandwidth value. The problem is how to put a price on quality. How much more will we pay for a 24-bit color depth image rather than a 16-bit color depth image, how much more will we pay for a high-resolution image, how much more will we pay for a 25-frame-per-second video stream compared to a 12-frame-per-second video stream.

In practice, we are prepared to pay a premium price for perfection. We buy hi-fi systems for $80,000 (which are probably only infinitesimally better than an $800 system); we buy cars that can travel at twice the legal speed limit. Vanity purchasing is high added value purchasing.

If we have taken a high-resolution, high-color-depth picture or a fast-frame-rate, high-color-depth, high-resolution video clip with high bandwidth audio and we are given the option of either maintaining image/video/audio quality or degrading image/video/audio quality, then most of us will, hopefully, decide to keep the quality even though it costs us more to send and store or store and send.

Given that most of us are also by nature impatient, we will often also want to send and share information in a conversational exchange. After all, this is the basic functionality that we have come to expect from the plain old telephone system (POTS) over the past 100 years.

Session immediacy is therefore one of the four components of session value. These are shown in Figure 20.6.

We need immediacy and consistency to build session value, particularly when multiple users are involved. We need immediacy, consistency, and complexity to build session value, which means session value is effectively built on an increase in the data duty cycle. Handset hardware, handset software, network hardware, and network software have to work together to build session value in order for us to be able to bill session value.

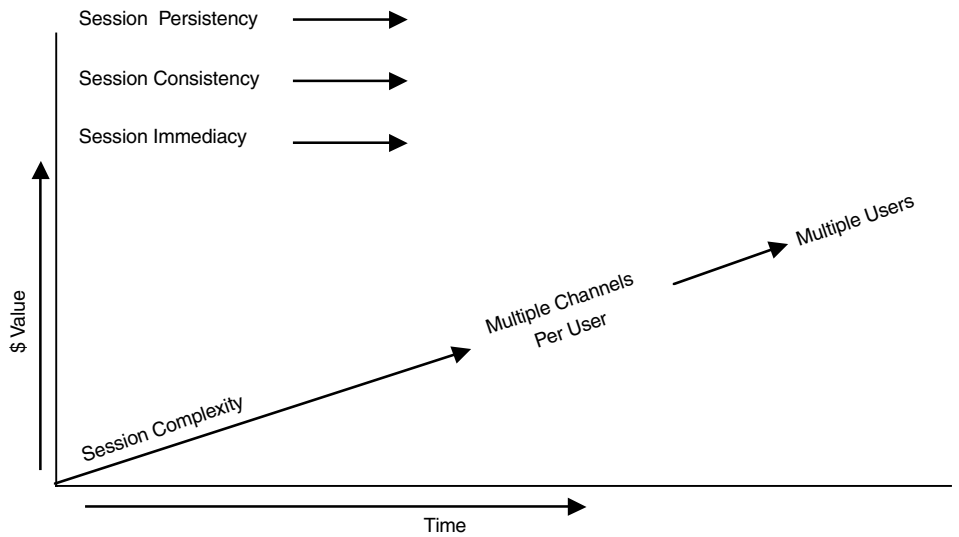


Figure 20.6 The four components of session value.

As voice quality improved in 2G cellular networks and as service consistency improved (lower blocked call rates and dropped call rates), we used our phones more. We made more phone calls. We made longer phone calls. The network operators (some at least) made money. The same principle applies to 3G network deployment: We will only realize session value when we can deliver a consistent-quality, rich media conversational exchange.

Preserving Bursty Bandwidth Quality

This brings us back to the starting point of the book. Traffic is becoming more asynchronous over time. Bandwidth is becoming burstier. The dynamic range of the bit rate coming from and going to individual users is increasing. Bandwidth quality and session quality are intimately interrelated. Figure 20.7 shows the many different aspects of bandwidth that we need to consider.

Delivering bandwidth quality and building and preserving session value is dependent on the successful integration of design skills in all these areas. In the past 20 chapters we have tried to highlight these design issues in the context of handset hardware and software design, and network hardware and software design. It is difficult, probably impossible, to be a specialist in each and every one of these topics, but it is increasingly important to be aware of the design issues involved in areas parallel to our own particular sphere of interest.

We have tried to reflect this in the Resources section at the back of this book, which is intended as a guide for future research.

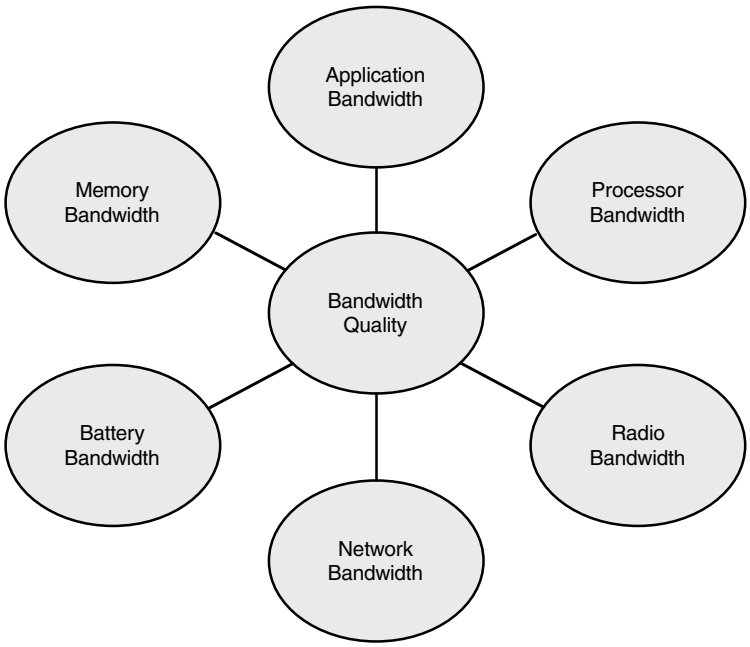


Figure 20.7 The components of bandwidth quality.

APPENDIX

Resources

In this section we have listed the standards bodies involved in 3G standardization and related areas. Most of these sites have hot links to member companies and organizations.

If you require additional detail then please visit our web site www.rttonline.com and go to the HOT LINKS Links section where we have resources listed by industry sector:

Also on our Web site you will find our HOT TOPICS. These are posted monthly on the site and are a digest of current RTT research. You can receive these 4 to 5 weeks prior to posting by joining our auto send PUSH LIST.

For more detail. Although we have covered a lot of topics in this book, there is always scope for studying any particular subject in more detail. You may, for example, have an interest in RF PA or oscillator design, or have a particular interest in handset or network hardware design. If you would like more detail on any particular topic, you will find our full range of design programs listed on the site including 3G handset hardware and software design and network hardware and software design.

Standards and Industry Organizations

COMPANY NAME	NOTES	WEB SITE ADDRESS
3rd Generation Partnership Project	Vendor group	www.3gpp.org
ACTS (Advanced Computer and Telecommunication Services)	EEC funded research	www.actsline.org
Advanced Television Systems Committee	ATSC vs. TV standards authority	www.atsc.org
American Telecoms Standardization		www.atis.org
ANSI (American National Standards Institute)	LANs and WANs (standards body)	www.ansi.org
Application Service Provider Consortium		www.aspinindustry.org
Association for Local Telecommunications Services		www.alts.org
Association of Public Communication Officers		www.apcointl.org
ATM Forum	ATM vendor consortium	www.atmforum.com
Biometrics Association	Biometrics	www.bioapi.org
Bluetooth Vendor Consortium		www.bluetooth.com
Broadband Content Delivery Forum	Nortel-sponsored forum	www.bcdforum.org
Broadband Wireless Association		www.broadband-wireless.org
CDMA Development Group		www.cdg.org
CDPD Forum		www.cdpd.org
Cellular Telecommunications Industry Association		www.wow-com.com
Common Switch Interface Consortium	Soft switching consortium	www.csix.org
Compact Flash	Vendor consortium	www.compactflash.org
Computer Measurement Group	IP performance measurement vendors	www.cmg.org

Standards and Industry Organizations *(Continued)*

COMPANY NAME	NOTES	WEB SITE ADDRESS
Consumer Electronics Manufacturers Association		www.eia.org
CSSA	The U.K. association for the software IT services and information industries	www.cssa.co.uk
DECT Forum		www.dect.ch
DECT Vendor Group		www.dectweb.com
Digital Display Group	Digital display working group	www.ddwg.com
Directory Enabled Networks		www.dutf.org
DMTF (Desktop Management Task Force)	Common information model (vendor consortium)	www.dmtf.org
Electronics Industry Association		www.eia.org
Enterprise Computing Telephony Forum	CT1, PSTN, and IP integration	www.ectf.org
ETSI (European Telecommunications Standards Institute)	Telecommunications standards body	www.etsi.org
European Broadcasting Union	Broadcasting association	www.ebu.ch
European Radiocommunications Office		www.ero.dk
Federal Communications Commission		www.fcc.gov
Federation of Communications Services (FCS)	U.K. association	www.fcs.org.uk
Frame Relay Forum	Frame relay (vendor consortium)	www.frforum.com
GEA (Gigabit Ethernet Alliance)	Gigabit Ethernet (vendor consortium)	www.gigabit-ethernet.org

(continues)

Standards and Industry Organizations *(Continued)*

COMPANY NAME	NOTES	WEB SITE ADDRESS
Geospatial Information and Technology Association	Geographic information systems vendor group	www.gita.org
Gigabit Ethernet		www.gigabit-ethernet.org
GSM Association	GSM MOU site	www.gsmworld.com—links into GSM operator web sites.
GSM Data Group	Vendor group promoting GSM-based data solutions—for example, Intel, Cellnet	www.gsmdata.com
GSM Suppliers Association		www.gsassociation.org/
HIPERLAN2	HIPERLAN workgroup	www.hiperland2.com
Home RF/Swap	Shared Wireless Access Protocol	www.homerf.org
ICAP Forum	Internet Content Adaptation Protocol	www.i-cap.org
IEEE (Institute of Electrical and Electronic Engineers)	Standards body	www.ieee.org
IETF (Internet Engineering Task Force)	Internet and related technologies (standards body)	www.ietf.org
IMTC	International Multimedia Teleconferencing Consortium	www.imtc.org
International Wireless Telecommunications Association	Trunked radio association	www.iwta.org also www.imta.org
Internet 2	Next-generation Internet	www.internet2.edu
Internet Content Adaption Protocol Forum		www.i-cap.org
Internet Telephony Consortium		www.itel.mit.edu
IPv6 Forum	Forum for IPv6 adoption	www.ipv6forum.com

Standards and Industry Organizations *(Continued)*

COMPANY NAME	NOTES	WEB SITE ADDRESS
ISO (International Organization for Standardization)	Information technology standards body	www.iso.ch
ITU	ITU regulatory Web site	www.regulate.org
ITU-T (International Telecommunications Union-Telecommunications Standardization Sector)	Telecommunications standards body	www.itu.ch and www.itu.int
Java/Sun	Java card spec	www.java.sun.com. commerce
Java/Sun	Java Telecom Object Network	www.jtone.com
J-PEG	Joint Photographic Experts Group	www.jpeg.org
Low Power Radio Association		www.lpra.org
Mobile Data Association		www.mobiledata.org
MPEG	Motion Picture Experts Group	www.mpeg.org
Multilayer Switching	Anritus portal	www.multilayerswitch .com
Multimedia Telecommunications Association	Broadband cable standards association	www.mmta.org
National Association of Broadcasting	US TV association	www.nab.org
National Telecommunications Information Administration		www.ntia.doc.gov
Netcentric	Metered Service Information Exchange (MSIX)	www.netcentric.com
Network Management Forum	Net management and vendor consortium	www.nmf.org

(continues)

Standards and Industry Organizations *(Continued)*

COMPANY NAME	NOTES	WEB SITE ADDRESS
OMG (Object Management Group)	Object-oriented software vendor consortium	www.omg.org
Open Service Gateway	Internet Toasters!	www.osgi.org
Optical Electronics Industry Development Association		www.oida.org
Optical Interworking Forum	Optical switching	www.oiforum.com
OURS (Open User Recommended Solutions)	Information technology (vendor consortium)	www.ours.org
PCS Data	GSM vendor group (U.S. focus)	www.pcsdata.com
Peer to Peer Working Group	Vendor group	www.p2pwg.org
Personal Communications Industry Association		www.pcia.com
PHS International		www.phsi.com
Portable Computing Association		www.pcca.org
Protocol Portal	Access to protocol Web sites	www.protocols.com
QoS Forum		www.qosforum.com
Radiocommunications Agency		www.open.gov.uk/radiocom/
SIP Forum	SIP vendor interest group	www.sipforum.org
Smart Card Industry Association		www.scia.org
Society of Motion Picture and Television Engineers		www.smpte.org
Software and information Industry Association	Vendor consortium	www.siia.net
Software Defined Radio Forum	Vendor forum	www.sdrforum.org
Spectrum Auctions	Auction process review	www.spectrum auctions.gov

Standards and Industry Organizations *(Continued)*

COMPANY NAME	NOTES	WEB SITE ADDRESS
SS7 Signaling	Nortel/Agilent site	www.access7.com *also* www.ips7.net
Storage Network Industry Association		www.snia.org
Telecommunications Industry Association		www.tiaonline.org
TETRA MoU		www.tetramou.com
Tetrapol Standards Group		www.tetrapol.com
The IEEE Computer Society		www.btg.com/IEEE
The Open Group	Open systems (vendor consortium)	www.opengroup.org
The Patent Office		www.patent.gov.uk
Third Generation Partnership Project	Pressure group formed to support WCDMA	www.3gpp.org
TI	U.S. Telco Interconnect standards	www.t1.org
UMTS		www.umts-forum.org
UTC		www.2dgsys.com
UWCC	IS136 vendor group	www.uwcc.org
V-card Forum	Pdi-info@imc.org	
Video Electronics Standards Association	Vendor group	www.vesa.org
Virtual Socket Interface Alliance		www.vsi.org
W3C (World Wide Web Consortium)	World Wide Web (vendor consortium)	www.w3c.org
Wi-Fi Group	802 interoperability testing	www.wi-fi.org
Wireless Application Protocol (study group)		www.xwap.com and www.wapforum.org
Wireless Data		www.wirelessdata.org
Wireless Ethernet Site		www.wireless ethernet.org
Wireless LAN Alliance		www.wlana.com
XML Vendor Group		www.xmg.org

Index

A

C

F

H

I

M

N

Q

R

S

T

U

V

W

X

Z