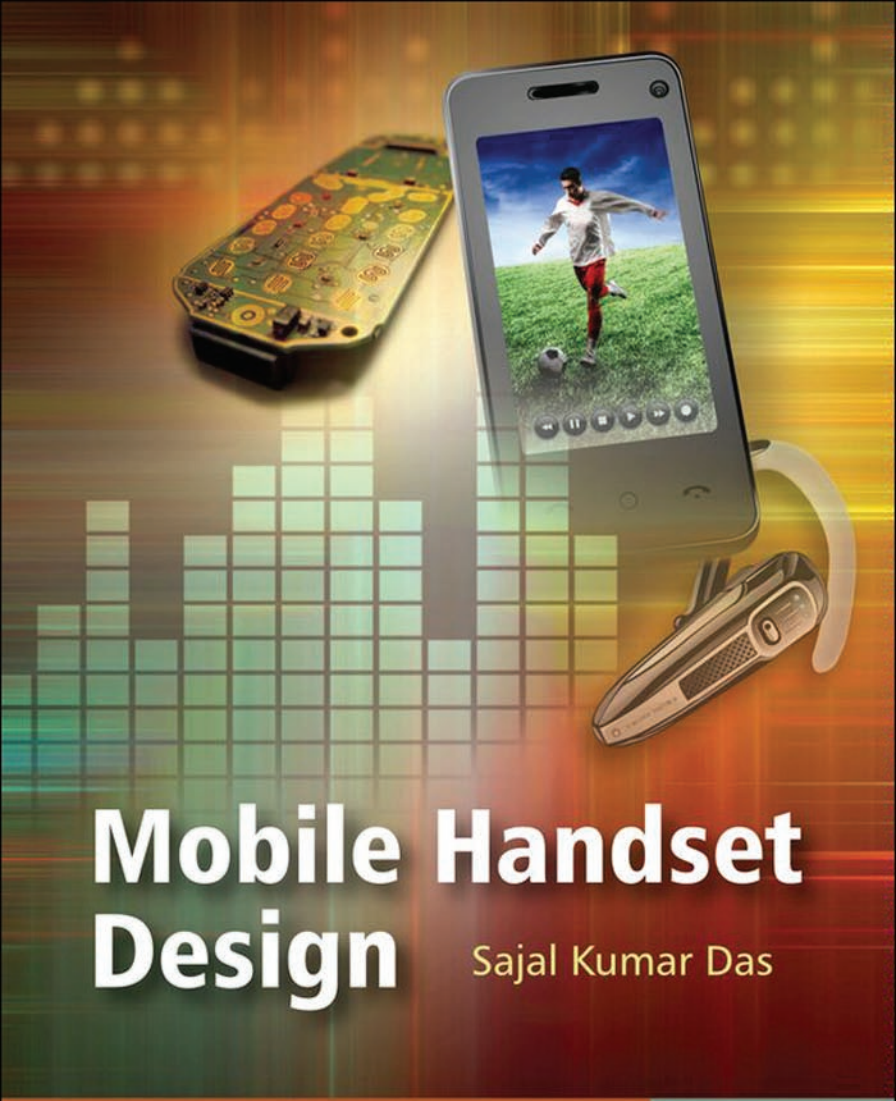


Mobile Handset Design

Sajal Kumar Das

WILEY

MOBILE HANDSET DESIGN

MOBILE HANDSET DESIGN

Sajal Kumar Das
Nokia R&D Center, India

John Wiley & Sons (Asia) Pte Ltd

Other Wiley Editorial Offices

John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstrasse 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons Canada Ltd, 5353 Dundas Street West, Suite 400, Toronto, ONT, M9B 6H8, Canada

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Das, Sajal K.
Mobile handset design / Sajal Kumar Das.
p. cm.
ISBN 978-0-470-82467-2 (cloth)
1. Mobile communication systems. 2. Wireless communication systems. 3. Cellular telephones. I. Title.
TK6570.M6D35 2009
621.3845′6–dc22

2009018214

ISBN 978-0-470-82467-2 (HB)

Typeset in 9/11pt Times by Thomson Digital, Noida, India.
Printed and bound in Singapore by Markono Print Media Pte Ltd, Singapore.
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Dedicated to Everyone who likes Mobile Phones

Contents

Preface

Over the last decade, there has been a considerable resurgence in interest in mobile communication. Because of the ever increasing demand for higher data rate, support of more complex applications, and seamless hand-over between the various networks, the mobile system has evolved over several generations from first generation to fourth generation. As a result of advancements in the technology, not only have new wireless standards been developed but also the challenges in the design of the mobile handset have been varied, which has thrown up more challenges for the design of unique handsets that can offer solutions providing low power consumption, low cost, smaller size, high performance, and tremendous flexibility. This challenge has created a significant opportunity for mobile phone design houses and manufacturers to develop innovative products.

The purpose of this textbook is to initiate students, working practitioners, and mobile communication readers to the technologies of the mobile handset, which is one of the fastest growing fields in today's engineering world. The inspiration for writing this text came into my mind when I was working for various companies in the mobile communication field. As there is no such book presently available on the market that covers the mobile communication principle and handset design aspects, I felt such a book might be useful for students, as well as for mobile communication practicing engineers, and project managers. This textbook is based upon my working experience as a design engineer in the field of wireless and mobile communications and is modeled on an academic course developed for electronics communication engineering students.

This book covers all the aspects of mobile handset technology, starting from the very basic elements of wireless communication systems, which will help everyone to master mobile handset design. It covers the mobile communication working principle, different mobile communication standards and the anatomy of mobile phones over the last four generations of mobile communication systems. This book is written after consulting the literature, technical papers, books, and valuable notes in the mobile communication field. I acknowledge with due courtesy the sources consulted in preparation of this book. I express my sincere thanks and deep sense of gratitude to all colleagues (especially Srinath, Mohan, Karthik, Sundar, Avinash, Mukesh, Arnab and Arvind), senior colleagues, friends, and family members for their valuable suggestions. The author would greatly appreciate any positive criticisms and suggestions from the readers for improving the quality of the book and to serve the wireless community in a more useful way.

Introduction

This book is written to cover all aspects of mobile phones, beginning with the understanding of the working principles, to practical engineering, and to making the final product. There are some basic principles on which mobile communication systems working today are based. At present, there are various technologies available to select from when making a mobile system. Thus, we need to choose from what is available according to our requirements (for example, data speed, mobile speed) and collect these in a rule book to produce a standard so that everyone can do business together, as one can not develop all the components of the entire system individually. As the technology evolves, so many new standards are being developed. In this book, the reader is first introduced to the working principle of a mobile phone. Next, the different problems and the corresponding design solutions (including the various latest research) into the RF and baseband design are described. These are the basic principles that any designer has to comprehend in order to analyze and understand the design of any mobile device. Then the details of the design of mobile phones through the different generations are discussed. As a case study, the anatomy and internal details of 2G, 3G, and 4G mobile phones are described. Finally, some of the hardware and software design challenges for mobile phone are covered.

Chapter 1 "Introduction to Mobile Handsets"

In this chapter, the reader is first introduced to the different terminologies and parameters associated with mobile communication. It describes the origins and the need for telecommunication, then the quest for wireless communication and gradual movement towards cellular communication through the long range cordless phone concept. The mobile handset is introduced and there is a brief discussion (by way of an introduction to the components) of the working principles of the various essential components inside it. Finally, the chapter scans through the evolution of the mobile handset covering the generations of mobile communication (from 1G to 5G).

Chapter 2 "Problem Analysis in Mobile Communication System"

A wireless channel is more complex than a traditional wired transmission channel. A range of problems such as multi-path, fading, interference, environment noise, burst noise, time dispersion or delay spread are involved in signal transmission through a wireless channel. In this chapter, how the signal transmission–reception process becomes gradually complex from a *point to point – wire-line* scenario to a *multi-user – wireless – mobile* scenario is discussed. A detailed description is provided of the various problems and issues associated with a wirelss mobile channel, which need to be mitigated in the mobile receiver design and will be helpful for channel and receiver simulation.

Chapter 3 "Design Solutions Analysis for Mobile Handsets"

Various wirelsss channel problems have a strong negative impact on the bit error rate and receiver performance of any modulation technique. The main reason why detection in a fading channel has poor

performance compared with an AWGN (additive white Gaussian noise) channel is because of the randomness of the channel gain due to the effect of deep fade. Thus, in order to combat these effects, we need to establish the corresponding radio receiver design solutions. It is mainly diversity, equalization, and channel coding techniques that are used to improve the quality of the received signal and to reduce the bit error rate. Also, different modulation techniques are chosen to meet the data rate and spectrum efficiency. In this chapter, we discuss the various design solutions for combating the wireless channel issues, which were discussed in Chapter 2. The designer can select a range of design solutions based on the system requirements, cost, size, and performance trade-off.

Chapter 4 "Mobile RF Transmitter and Receiver Design Solutions"

An RF module is the analog front-end module that is responsible for signal reception from the air, down-conversion to a baseband signal on the receiver side and up-conversion of the baseband signal to an RF signal, and then transmission in the air on the transmitter side. There are various design solutions available for mobile front-end RF modules. In this chapter, the RF front-end architecture and the design aspects for a mobile transmitter and receiver are discussed in detail.

Chapter 5 "Wireless Channel Multiple Access Techniques for Mobile Phones"

In the case of wireless communication, we use air or free-space as the medium and EM waves as the information carrier. Air, being a public channel, needs to be multiplexed among various simultaneous users. In this chapter, different multiple access techniques are discussed in order to share the same wireless air channel between numerous users and wireless mobile standards. These are mainly: Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA), Code Division Multiple Access (CDMA), and Orthogonal Frequency Division Multiple Access (OFDMA).

Through the above chapters, the working principles and the different design solutions for any mobile handset should be well understood. Next, as a case study, the mobile phones for 2G, 2.5G, 3G, and 4G systems are discussed with respect to the system architecture, software protocol design, operations, procedures, and hardware anatomy.

Chapter 6 "GSM System (2G) Overview"

GSM (Global System for Mobile Communications) is world's first cellular system to specify digital modulation, network level architectures, and services. Today, it is the world's most popular second generation (2G) technology, with more than one billion subscribers worldwide. In this chapter, the reader is given a brief introduction to the GSM standard, network architecture, and various interfaces of the GSM system.

Chapter 7 "GSM Radio Modem Design: From Speech to Radio Wave"

A user speaks in front of phone's microphone and the analog speech signal is converted into a digital signal and source coded. Then error correction (FEC) coding is applied to the speech data in order to protect it from the channel disturbances. After this the coded data are interleaved to protect them from burst error and then ciphered for security purpose. Next these are assembled into bursts along with the training sequence and tail and guard bits. It is then ready for transmission over the radio interface. Digital modulation is performed and then up-converted to the GSM RF channel frequency band and transmitted

via the air. The reverse process is performed at the receiver side. In this chapter, the radio design aspect of a GSM system is discussed in detail, which includes various logical and physical channel concepts, various physical layer processing, baseband design parameters, and the different blocks in a digital transceiver.

Chapter 8 "GSM Mobile Phone Software Design"

As discussed in the earlier chapters, a GSM mobile handset system consists of several essential hardware blocks, the necessary software for driving these hardware blocks, a protocol stack for governing communication, and application software for running the various applications. Typically, the software part consists of several modules such as a boot loader, initialization code, protocol stack, device drivers, and RTOS. Apart from this, there may be audio and video related software, a Bluetooth stack, and also some other application software such as gaming, a calculator, and so on. In this chapter, the software system of a GSM mobile phone is discussed in detail.

Chapter 9 "GSM Mobile Phone Operations and Procedures"

When a mobile phone is switched ON, its first task is to find a suitable BTS (base transceiver substation) through which it can gain access to the network. All BTSs broadcast their allocated BCCH (broadcast control channel) carrier (different neighbor BTSs are allocated different radio beacon frequencies to transmit according to frequency planning and each BTS has a single broadcast (BCCH) frequency) in the respective cell. Generally upper layer protocol stack commands layer-1 to program the RF module to measure the RSSI (received signal strength indication – which is usually measured as the square root of I^2 and Q^2) for different carrier frequencies and then once it is performed, layer-1 indicates the result to the upper layer. On the basis of this result, the upper layer decides on which carrier frequency it should search for FB and SB bursts (generally it selects the carrier frequency on the basis of the highest RSSI value). This process is called cell selection. In this chapter different operations of mobile phones, starting from the switch ON, to camp on, to call setup, to handover, to call release are discussed in detail with a message flow diagram.

Chapter 10 "Anatomy of a GSM Mobile Handset"

The GSM mobile handset has evolved over a period of time and improved its efficiency with respect to size, weight, complexity, application support, performance, and battery life. In Chapter 1, we have already briefly discussed the internal components of any mobile phone. The basic phone architecture and the associated peripherals such as display (LCD), keypad, speaker, and microphone remain almost the same with respect to the air interface technology or mobile standard used. However, based on the mobile standard chosen, the front-end RF unit, the baseband processing unit as well as the protocol stack used (especially up to layer-3) will be different. This chapter discusses in detail the internal structure of a GSM mobile phone and some of its internal components and functional blocks.

Chapter 11 "Introduction to GPRS and EDGE (2.5G) Supported Mobile Phones"

In order to address the inefficiencies of circuit switched radio transmission, two cellular packet data technologies have been developed: Cellular Digital Packet Data (CDPD) (for AMPS, IS-95, and IS-136) and the General Packet Radio Service (GPRS). Basically, GPRS is based on the packet radio principle and to support this some modifications have been done on both the network and the mobile sides. In this chapter, GPRS and EDGE (known as 2.5G) systems and the corresponding design changes from GSM mobile phones to support these technologies are discussed.

Chapter 12 "UMTS System (3G) Overview"

Second generation (2G) mobile communication systems have several limitations. To satisfy the increasing demand for higher data rate, tighter data security, larger network capacity, and support of various multimedia applications, the International Telecommunication Union (ITU) has defined a set of requirements, which specify what is needed from the next generation (3G) mobile systems. In this chapter, UMTS (Universal Mobile Telecommunication System) network architecture and its evolution from GSM (2G) is briefly discussed.

Chapter 13 "UMTS Radio Modem Design: From Speech to Radio Wave"

This chapter presents a WCDMA (Wideband Code Division Multiple Access) air interface, also referred to as UMTS terrestrial radio access (UTRA), which has been developed through the Third-Generation Partnership Project (3GPP) and radio modem design aspect of a UMTS mobile handset. This chapter discusses the various logical, transport, and physical channels and the corresponding physical layer modem design blocks.

Chapter 14 "UMTS Mobile Phone Software and Operations"

As with a GSM phone, a UMTS mobile phone also contains modem software module and several other applications processing software modules, along with the OS and associated software modules. UTRAN interface consists of a set of horizontal and vertical layers of protocol architecture. The WCDMA protocol has a layered structure designed to give the system a great deal of flexibility. The WCDMA structure is divided vertically into an "access stratum" and a "non-access stratum," and horizontally into a "control plane" and a "user plane." Protocol layers-1 and -2 are in the access stratum. Protocol layer-3 is divided between the access and non-access strata. In layers-2 and -3, control plane and user plane information is carried on separate channels. Within layer-1 some channels carry only control plane information, while others carry both user and control plane data. This chapter discusses the software and protocol design aspects of a UMTS mobile handset.

Chapter 15 "Anatomy of a UMTS Mobile Handset"

This chapter provides more information on the various internal blocks, and hardware–software components of a UMTS mobile phone.

Chapter 16 "Next Generation Mobile Phones"

Over the last few years, there has been a considerable resurgence in interest in wireless communication. Owing to the ever increasing demand for higher data rate, support of more complex applications, and seamless handover between the various networks, the wireless system has evolved over several generations. 4G is an initiative to move beyond the limitations and problems of 3G, which is having trouble being deployed and meeting its promised performance and throughput. 4G is intended to provide high speed, high capacity, low cost per bit, IP based services. The 4G mobile communication systems are projected to solve the still-remaining problems of 3G systems and to provide a wide variety of new services, from high-quality voice, to high-definition video, to high-data-rate wireless channels. One term used to describe 4G is $4G = C.A^5$, where, C is **C**ommunication, A is **A**nytime, **A**nywhere, with **A**nyone, on **A**ny device, through **A**ny network. In this chapter, the latest trends towards different mobile phone system

design complexities and a range of alternatives are reviewed. As a case study, a WiMAX based mobile device is discussed. Finally, the next generation wireless mobile radio systems (4G and above) are discussed, including cognitive radios.

Chapter 17 "Competitive Edge in Mobile Phone System Design"

Today's mobile handset system is not just a piece of hardware or collection of software; rather it is a combination of both hardware and software. In the present competitive market the key factors to success are designing a system that can work with minimum resources (such as memory size and MIPS), which offers high performance in terms of execution speed, low power consumption, and high robustness. It is not always difficult to write a piece of software to work on a system that is logically right, but it really is a big job to write a piece of software that will work in an environment of limited resources (such as memory, MIPS) with greater speed of operation and that is logically correct. This chapter examines various factors that contribute towards the development of a competitive mobile phone hardware and software protocol stack. Both technical and non-technical aspects are considered. The key issues addressed include protocol architecture, system performance in terms of memory, CPU, operating system (OS), electrical power consumption, processing power (MIPS), cost, optimum hardware/software partitioning and productization.

1

Introduction to Mobile Handsets

1.1 Introduction to Telecommunication

The word telecommunication was adapted from the French word *télécommunication*, where the Greek prefix *tele-* (τηλε-) means- "far off," and the Latin word *communicare* means "to share." Hence, the term telecommunication signifies communication over a long distance. In ancient times, people used smoke signals, drum beats or semaphore for telecommunication purposes. Today, it is mainly electrical signals that are used for this purpose. Optical signals produced by laser sources are recent additions to this field. Owing to the evolution of major technological advances, today telecommunication is widespread through devices such as the television, radio, telephone, mobile phone and so on. Telecommunication networks carry information signals from one user to another user, who are separated geographically and this entity may be a computer, human being, teleprinter, data terminal, facsimiles machine and so on. The basic purpose of telecommunication is to transfer information from one user to another distant user via a medium. God has given us two beautiful organs: one is an eye to visualize things and other is an ear to listen. So, to the end users, in general information transfer is either by voice or as a real world image. Thus, we need to exchange information through voices, images and also computer data or digital information.

1.1.1 Basic Elements of Telecommunication

The basic elements of a telecommunication system are shown in the Figure 1.1. In telephonic conversation, the party who initiates the call is known as the calling subscriber and the party who is called is known as the called subscriber. They are also known as source and destination, respectively. The user information, such as sound, an image and so on, is first converted into an electrical signal using a transducer, such as a microphone (which converts sound waves into an electrical signal), a video camera (which converts an image into an electrical signal) and so on, which is then transmitted to the distant user via a medium using a transmitter. The distant user receives the signal through the use of a receiver and this is then fed to an appropriate transducer to convert the electrical signal back into the respective information (for example, a speaker is used as a transducer on the receiver side to convert the electrical signal into a sound wave or the LCD display is used to convert the electrical signal into an image).

Before getting into an in-depth discussion, we will first familiarize ourselves with some of the most commonly used terms and mathematical tools for mobile telecommunication system design and analysis. We will learn about some of these as we progress through the various chapters.

Mobile Handset Design Sajal Kumar Das

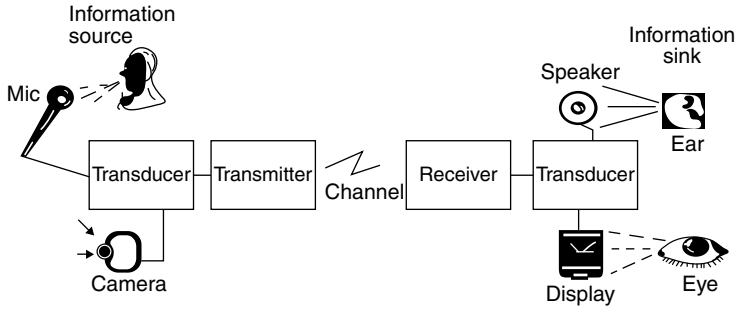


Figure 1.1 Basic elements of telecommunication

1.1.1.1 Signal

The amplitude of a time varying event is described as a signal. Using Fourier's theory a signal can be decomposed into a combination of pure tones, called sine or cosine waves, at different frequencies. Different sine waves that compose a signal can be plotted as a function of frequency to produce a graph called a frequency spectrum of a signal. The notion of a sinusoid with exponentially varying amplitude can be generalized by using a complex exponential. Based on the nature of the repetition of the signal amplitude with respect to time, the signal can be classified as periodic (repeats with every period) and aperiodic (not a periodic waveform). Also, signals can be either continuous or discrete in nature with respect to time.

Analog Signal

Analog signals are continuous with time as shown in Figure 1.2. For example, a voice signal is an analog signal. The intensity of the voice causes electric current variations. At the receiving end, the signal is reproduced in the same proportions.

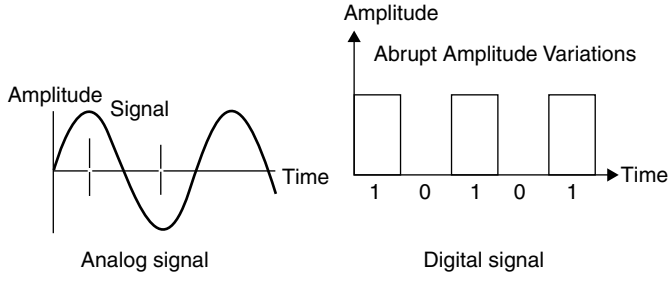


Figure 1.2 Analog and digital signal

As shown in the Figure 1.3, the signal can be represented in complex cartesian or polar format. Here, T is period of a periodic signal ($T = 1$/frequency) and A is the maximum amplitude. ω_o is the angular frequency ($=2\pi f$) and Φ is the phase at any given instant of time (here at $t = 0$).

Signals having the same frequency follow the same path (repeat on every period T), but different points on the wave are differentiated by phase (leading or lagging). From the figure it is obvious that

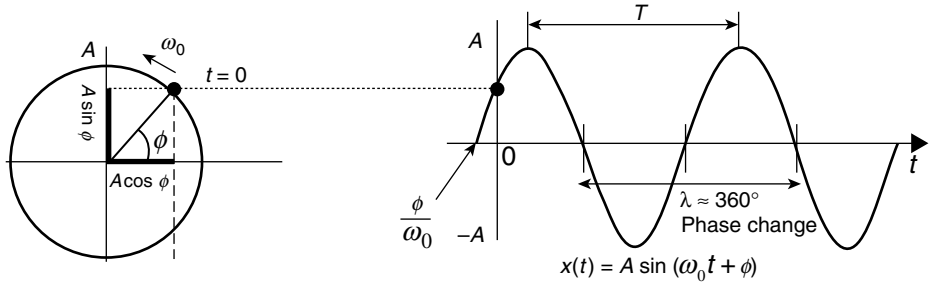


Figure 1.3 Signal representation in polar and cartesian format

one period (T) is 360° of a phase (a complete rotation). Harmonics ($2f$, $3f$, ...) are waves having frequencies that are integer multiples of the fundamental frequency. These are harmonically related exponential functions.

Digital Signal

Digital signals are non-continuous with time, for example, discrete. They consist of pulses or digits with discrete levels or values. The value of each pulse width and amplitude level is constant for two distinct types, "1" and "0", of digital values. Digital signals have two amplitude levels, called nodes and the value of which is specified as one of two possibilities, such as 1 or 0, HIGH or LOW, TRUE or FALSE and so on. In reality, the values are anywhere within specific ranges and we define the values within a given range. A system which uses a digital signal for processing the information is known as a digital system. A digital system has certain advantages over an analog system, as mentioned below.

Advantages – (1) Digital systems are less affected by any noise signal compared with analog signals. Unless the noise exceeds a certain threshold, the information contained in digital signals will remain intact. (2) In an analog system, aging and wear and tear will degrade the information that is stored, but in a digital system, as long as the wear and tear is below a certain level, the information can be recovered perfectly. Thus, it is easier to store and retrieve the data without degradation in a digital system. (3) It provides an easier interface to a computer or other digital devices. Apart from this ease of multiplexing, ease of signaling has made it more popular.

Disadvantages – From their origins, voice and video signals are analog in nature. Hence, we need to convert these analog signals into the digital domain for processing, and after processing again we need to convert them back into the original form to reproduce. This leads to processing overheads and information loss due to conversions.

Digital Signaling Formats

The digital signals are represented in many formats, such as non-return to zero, return to zero and so on, as shown in Figure 1.4. In telecommunication, a non-return-to-zero (NRZ) line code is a binary code in which "1s" are represented by one significant condition and "0s" are represented by the other significant condition, with no other neutral or rest condition. Return-to-zero (RZ) describes a line code used in telecommunications signals in which the signal drops (returns) to zero between each pulse. The NRZ pulses have more energy than an RZ code, but they do not have a rest state, which means a synchronization signal must also be sent alongside the code.

Unipolar Non-Return-to-Zero (NRZ) – Here, symbol "1" is represented by transmitting a pulse of constant amplitude for the entire duration of the bit interval, and symbol "0" is represented by no pulse. This allows for long series without change, which makes synchronization difficult. Unipolar also contains a strong dc component, which causes several problems in the receiver circuits, such as dc offset.

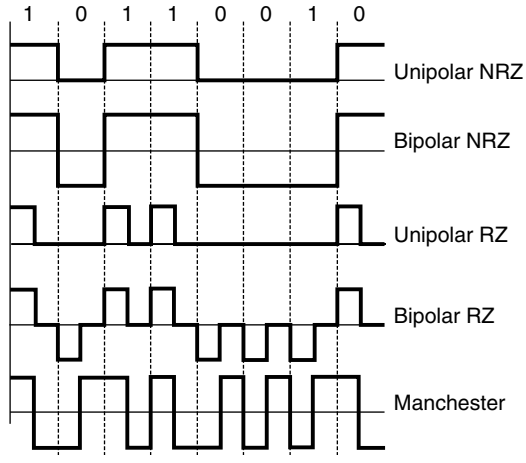


Figure 1.4 Digital signal representations

Bipolar Non-Return-to-Zero – Here, pulses of equal positive and negative amplitudes represent symbols "1" and "0." (for example, $\pm A$ volts). This is relatively easy to generate. Because of the positive and negative levels, the average voltage will tend towards zero. So, this helps to reduce the dc component, but causes difficulties for synchronization.

Unipolar Return-to-Zero – Symbol "1" is represented by a positive pulse of amplitude A and half symbol width and symbol "0" is represented by transmitting no pulse.

Bipolar Return-to-Zero – Positive and negative pulses of equal amplitude are used alternatively for symbol "1," with each pulse having a half-symbol width; no pulse is used for symbol "0." The "zero" between each bit is a neutral or rest condition. One advantage of this is that the power spectrum of the transmitted signal has no dc components.

Manchester Coding – In the Manchester coding technique, symbol "1" is represented by a positive pulse followed by a negative pulse, with each pulse being of equal amplitude and a duration of half a pulse. The polarities of these pulses are reversed for symbol "0." An advantage of this coding is that it is easy to recover the original data clock and relatively less dc components are present. However, the problem is it requires more bandwidth. For a given data signaling rate, the NRZ code requires only half the bandwidth required by the Manchester code.

1.1.1.2 Analog to Digital Conversion

To convert an analog signal into a digital signal an electronic circuit is used, which is known as an analog-to-digital converter (ADC). Similarly, to convert a digital signal into an analog signal, a digital to analog converter (DAC) is used. The concept is depicted in Figure 1.5. Most of the ADCs are linear ADC types, where the range of the input values map to each output value following a linear relationship. Here, the levels are equally spaced throughout the range, whereas in the case of non-linear ADC, all the levels are not equally spaced. So, in this case, the space where the information is most important is sampled with a lesser gap and space, and where information is important, it is sampled at a higher rate. Normally, a compander (compressors and expanders) is used for this purpose. An 8-bit A-law or the μ-law logarithmic ADC covers the wide dynamic range. ADCs are of different types; below a few of the commonly used ADCs are discussed.

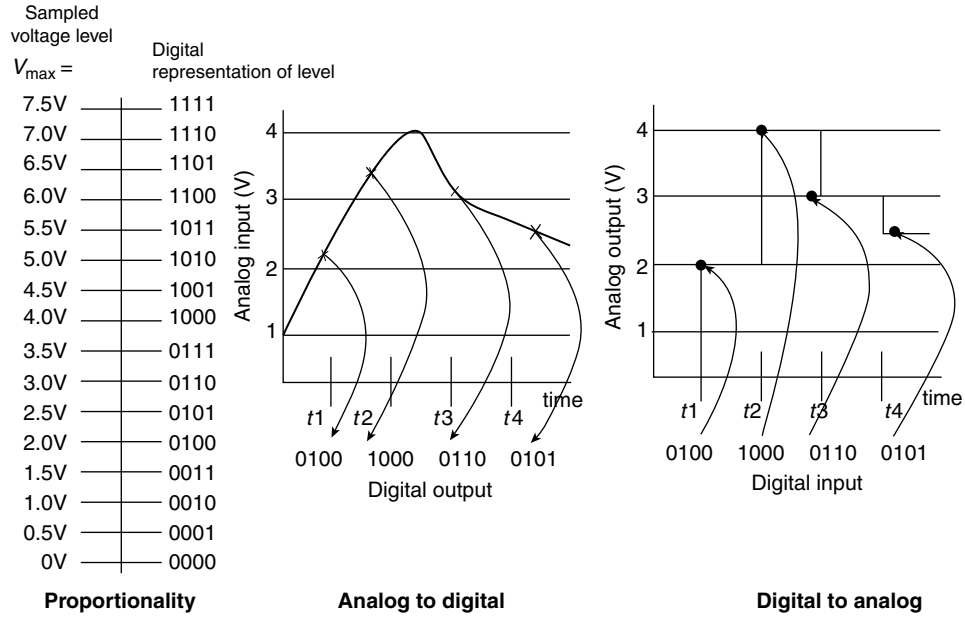


Figure 1.5 Analog to digital and digital to analog conversion

A. **Direct conversion ADC (flash ADC)** – As shown in Figure 1.6, this consists of a bank of comparators; each one outputs their decoded voltage range, which is then fed to a logic circuit. This generates a code for each voltage range. This type of ADC is very fast, but usually it has only 8 bits of resolution (with 256 comparators). For higher resolution, it requires a large number of

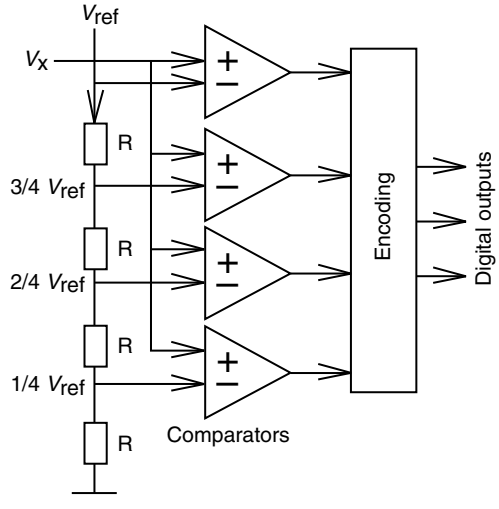


Figure 1.6 Flash analog to digital converter

comparators, this leads to larger die size and a high input capacitance, which makes it expensive and prone to produce glitches at the output. This is often used for wideband communications, video or other fast signals conversion.

B. **Sigma–delta ADC** – This works using two principles: over sampling and noise shaping. The main advantage of this filter is its notch response. It offers low cost, high resolution, and high integration. This is why this is very popular in the present day's mobile receivers. This is discussed in detail in Chapter 10.

1.1.1.3 Sampling

The process of converting a continuous analog signal into a numeric sequence or discrete signal (or digital signal) is known as sampling. A sample refers to a value or set of values, at a point in time and/or space. The discrete sample instance may be spaced either at regular or irregular intervals.

Sampling Interval – A continuous signal varies with time (or space) and the sampling is done simply by measuring the value of the continuous signal at every T units of time (or space), and T is called the sampling interval. Thus the sampling frequency $(fs) = 1/T$.

Sampling Rate and Nyquist Theorem – For a single frequency periodic signal, if we sample at least two points over the signal period, then it will be possible to reproduce a signal using these two points. Hence the minimum sampling rate is $f_s = 2/T = 2f$. A continuous function, defined over a finite interval can be represented by a Fourier series with an infinite number of terms as given below:

$$f(x) = a_0/2 + \sum_{k=1}^{\alpha} a_k \cdot \cos kx + b_k \cdot \sin kx \tag{1.1}$$

However, for some functions all the Fourier coefficients become zero for all frequencies greater than some frequency value $-f_H$. Now, let us assume that the signal $f(x)$ is a band-limited signal with a one-sided baseband bandwidth f_H, which means that if $f(x) = 0$ for all $|f| > f_H$, then the condition for exact reconstruction of the signal from the samples at a uniform sampling rate f_s is, $f_s > 2f_H$. Here, $2f_H$ is called the Nyquist rate. This is a property of the sampling system (see Figure 1.7). The samples of $f(x)$ are denoted by: $x[n] = x(nT)$, $n \in N$ (integers). The Nyquist sampling theorem leads to a procedure for reconstructing the original signal $x(t)$ from its samples $x[n]$ and states the conditions that are sufficient for faithful reconstruction. Nyquist, a communication engineer in Bell Telephone Laboratory in 1930, first discovered this principle. In this principle, it is stated as "Exact reconstruction of a continuous-time baseband signal from its samples is possible, if the signal is band-limited and the sampling frequency is greater than twice the signal bandwidth."

On the other side, the signal is being recovered by a sample and hold circuit that produces a staircase approximation to the sampled waveform, which is then passed through the reconstructive filter. The power level of the signal coming out of the reconstructive filter is nearly same as the level of the original sampled input signal. This is shown in the Figure 1.8.

Aliasing – If the sampling condition is not satisfied, then the original signal cannot be recovered without distortion and the frequencies will overlap. So, frequencies above half the sampling rate will be reconstructed as, and appear as, frequencies below half the sampling rate. As it is a duplicate of the input spectrum "folded" back on top of the desired spectrum that causes the distortion, this is why this type of sampling impairment is known as "foldover distortion" and the resulting distortion is also called aliasing. For a sinusoidal component of exactly half the sampling frequency, the component will in general alias to another sinusoid of the same frequency, but with a different phase and amplitude. The "eye pattern" is defined as the synchronized superposition of all possible realizations of the signal of interest viewed within a particular signaling interval. It is called such because the pattern resembles the human eye for binary waves. The width of the eye opening defines the time interval over which the received signal can be sampled without error from inter-symbol interference.

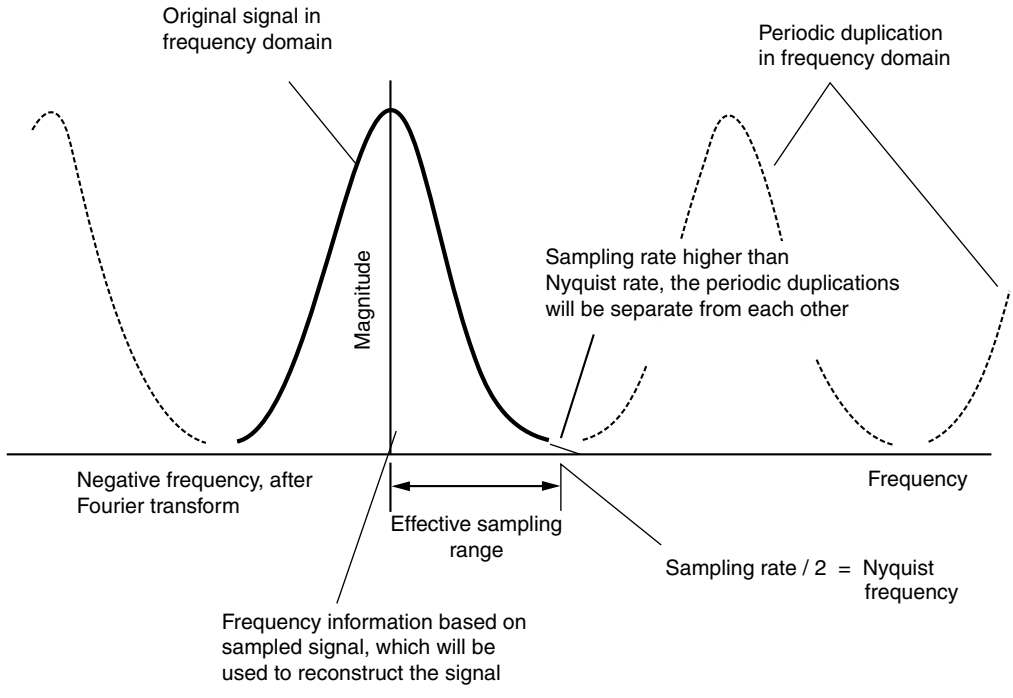


Figure 1.7 Sampling frequency

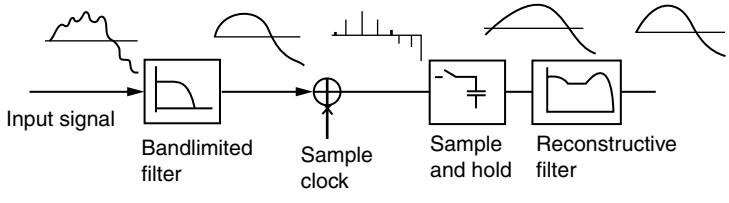


Figure 1.8 Sampling and reconstruction of signal

As shown in Figure 1.9, if we sample $x(t)$ too slowly, then it will overlap between the repeated spectra resulting in aliasing for example, we cannot recover the original signal, so aliasing has to be avoided. To prevent or reduce aliasing, two things need to be taken into consideration. (1) Increase the sampling rate more than or equal to twice the maximum signal frequency (whatever the maximum signal frequency present in that band is). (2) Introduce an anti-aliasing filter or make the anti-aliasing filter more stringent.

Although we want the signal to be band-limited, in practice, however, the signal is not band-limited, so the reconstruction formula cannot be precisely implemented. The reconstruction process that involves scaled and delayed sinc functions can be described as an ideal process. However, it cannot be realized in practice, as it implies that each sample contributes to the reconstructed signal at almost all time points, which requires summing an infinite number of terms. So, some type of approximation of the sinc functions

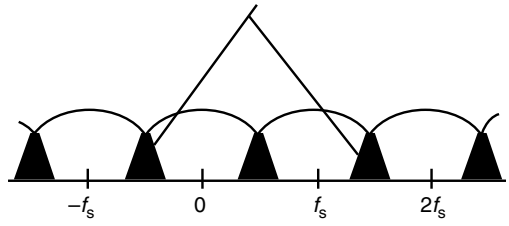


Figure 1.9 Signal distortion (energy overlap) due to low sampling rate

(finite in length) has to be used. The error that corresponds to the sinc-function approximation is referred to as the interpolation error. In practice digital-to-analog converters produce a sequence of scaled and delayed rectangular pulses, instead of scaled, delayed sinc functions or ideal impulses. This practical piecewise-constant output can be modeled as a zero-order hold filter driven by the sequence of scaled and delayed Dirac impulses referred to the mathematical basis. Sometimes a shaping filter is used after the DAC with zero-order hold to give a better overall approximation.

When the analog signal is feed to the ADC based on the maximum and minimum value of the analog signal amplitude, the range of the analog signal is defined and the resolution of the converter indicates the number of discrete values that it can produce over the range of analog values. The values are usually stored in binary form, hence it is expressed in the number of bits. The available range is first divided into several spaced levels, and then each level is encoded into n number of bits. For example, an ADC with a resolution of 8 bits can encode an analog input to one in 256 different levels, as $2^8 = 256$. The values can represent the ranges from 0 to 255 (that is, unsigned integer) or −128–127 (that is, signed integer), depending on the application.

1.1.1.4 Accuracy and Quantization

The process of quantization is depicted in Figure 1.10. The signal is limited to a range from VH (15) to VL (0), and this range is divided into M (=16) equal steps. The step size is given by, $S = (VH - VL)/M$.

The quantized signal Vq takes on any one of the quantized level values. A signal V is quantized to its nearest quantized level. It is obvious that the quantized signal is an approximation to the original analog signal and an error is introduced in the signal due to this approximation. The instantaneous error $e = (V - Vq)$ is randomly distributed within the range (S/2) and is called the quantization error or noise. The average quantization noise output power is given by the variance $\sigma^2 = S^2/12$.

The toll quality speech is band limited to 300–3400 Hz (speech signal normally contains signals with a frequency range in between 300 and 3400 Hz). To digitize this waveform the minimum sampling frequency required is $2 \times 3400\,\text{Hz} = 6.8\,\text{KHz}$ in order to avoid aliasing effects. The filter used for band limiting the input speech waveform may not be particularly ideal with a sharp cut off, thus a guard band is provided and it is sampled at a rate of 8 KHz.

Dithering – The performance of ADC can be improved by using dither, where a very small amount of random noise (white noise) is added to the input signal before the analog to digital conversion. The amplitude of this is set to be about half of the least significant bit. Its effect is to cause the state of the LSB to oscillate randomly between 0 and 1 in the presence of very low levels of input, instead of sticking at a fixed value. So, effectively the quantization error is diffused across a series of noise values. The result is an accurate representation of the signal over time.

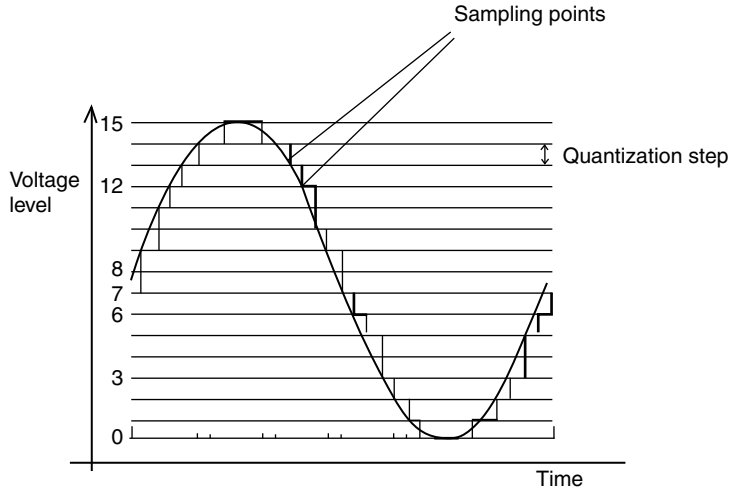


Figure 1.10 Quantization process

Sample Rate Converter – Sampling rate changes come in two types: decreasing the sampling rate is known as decimation and when the sampling rate is being increased, the process is known as interpolation. In the case of multimode mobile devices, the sample rate requirement is different for different modes. In some instances, where one clock rate is a simple integer multiple of another clock rate, resampling can be accomplished using interpolating and decimating FIR filters. However, in most situations the interpolation and decimation factors are so high that this approach is impractical. Farrow resamplers offer an efficient way to resample a data stream at a different sample rate. The underlying principle is that the phase difference between the current input and wanted output is determined on a sample by sample basis. This phase difference is then used to combine the phases of a polyphase filter in such a way that a sample for the wanted output phase is generated. Compared with single stage, a multi-stage sampling rate conversion system offers less computation and more flexibility in filter design.

1.1.1.5 Fourier Transforms

At present, almost every real world signal is converted into electrical signals by means of transducers, such as antennas in electromagnetics, and microphones in communication engineering. The analysis of real world signals is a fundamental problem. The traditional way of observing and analyzing signals is to view them in the time domain. More than a century ago, Baron Jean Baptiste Fourier showed that any waveform that exists in the real world can be represented (and generated) by adding up the sine waves. Since then, we have been able to build (or break down) our real world time signal in terms of these sine waves. It has been shown that the combination of sine waves is unique and any real world signal can be represented by a combination of sine waves, also there may be some dc values (constant term) present in this. The Fourier transform (FT) has been widely used in circuit analysis and synthesis, from antenna theory to radiowave propagation modeling, from filter design to signal processing, image reconstruction, stochastic modeling to non-destructive measurements.

The Fourier transform allows us to relate events in the time domain to events in the frequency domain. We know that various types of signals exist, such as periodic, aperiodic, continuous and discrete. There are several versions of the Fourier transform, and are applied based on the nature of the signal. Generally, Fourier transform is used for converting a continuous aperiodic signal from the time to frequency domain and a Fourier series is used for transforming a periodic signal. For aperiodic–discrete signal (digital), discrete time Fourier transform is used and for periodic–discrete signals that repeat themselves in a periodic fashion from negative to positive infinity, the discrete Fourier series (most often called the discrete Fourier transform) is used.

The transformation from the time domain to the frequency domain is based on the Fourier transform. This is defined as:

$$S(\omega) = \int_{-\infty}^{\infty} s(t)[\mathrm{e}^{-j2\pi \cdot f \cdot t}] \cdot \mathrm{d}t \tag{1.2}$$

Similarly, the conversion from frequency domain to time domain is called inverse Fourier transform, which is defined as:

$$s(t) = \int_{-\infty}^{\infty} S(\omega)[\mathrm{e}^{j2\pi \cdot f \cdot t}] \cdot \mathrm{d}f \tag{1.3}$$

Here $s(t)$, $S(\omega)$, and f are the time signal, the frequency signal, and the frequency, respectively, and $j = \sqrt{-1}$, angular frequency $\omega = 2\pi f$. The FT is valid for real or complex signals, and in general, it is a complex function of ω (or f). Some commonly used functions and their FT are listed in Table 1.1.

Table 1.1 Some commonly used functions and their Fourier transforms

Time domain	Frequency domain
Rectangular window	Sinc function
Sinc function	Rectangular window
Constant function	Dirac Delta function
Dirac Delta function	Constant function
Dirac comb (Dirac train)	Dirac comb (Dirac train)
Cosine function	Two, real, even Delta function
Sine function	Two, imaginary, odd Delta function
Exp function –{ $j$exp($j\omega t$)}	One, positive, real Delta function
Gaussian function	Gaussian function

1.1.1.6 System

A system is a process for which cause (input) and effect (output) relations exist and can be characterized by an input–output (I/O) relationship. A linear system is a system that possesses the superposition property, for example, $y(t) = 2x(t)$, and an example of non-linear system is $y(t) = x^2(t) + 2x(t)$.

Time-invariant system: a time shift in the input signal causes an identical time shift in the output signal. Memory-less (instantaneous) system: if present output value depends only on the present input value.

Otherwise, the system is called memory (dynamic) system.

Causal system (physically realizable system): if the output at any time t_0 depends only on the values of input for $t \leq t_0$. For example, if $x_1(t) = x_2(t)$ for $t \leq t_0$, then $y_1(t) = y_2(t)$ for $t \leq t_0$.

Stable system: a signal $x(t)$ is said to be bounded if $|x(t)| < B < \propto$ for any t. A system is stable, if the output remains bounded for any bounded inputs, this is called bounded-input bounded-output (BIBO) stable system.

Practical system: a non-linear, time-varying, distributed and non-invertible.

1.1.1.7 Statistical Methods

When a signal is transmitted through the channel, two types of imperfections can cause the received signal to be different from the transmitted signal. Of these, one is deterministic in nature, such as linear and non-linear distortion, inter symbol interference and so on, but the other one is nondeterministic, such as noise addition, fading and so on, and we model them as random processes.

The totality of the possible outcomes of a random experiment is called the sample space of the experiment and it is denoted by *S*. An event is simply a collection of certain sample points that is subset of the sample space. We define probability *P* as a set function assigning non-negative values to all events *E*, such that the following conditions are satisfied:

a. $0 \leq P(E) \leq 1$ for all events
b. $P(S) = 1$
c. For disjoint events $E1$, $E2$, $E3$, ..., we have $P\left(U_{i=1}^{\infty} Ei\right) = \sum_{i=1}^{\infty} P(Ei)$

A random variable is a mapping from the sample space to the set of real numbers. A random variable X is a function that associates a unique numerical value $X(\lambda_i)$ with every outcome λ_i of an event that produces random results. The value of a random variable will vary from event to event, and, based on the nature of the event, it will be either continuous or discrete. Two important functions of a random variable are cumulative distribution function (CDF) and probability density function (PDF).

The CDF, $F(X)$ of a random variable X is given by

$$F(X) = P[X(\lambda) \leq x] \tag{1.4}$$

where $P[X(\lambda) \leq x]$ is the probability that the value $X(\lambda)$ taken by the random variable X is less than or equal to the quantity x.

The PDF, $f(x)$ of a random variable X is the derivative of $F(X)$ and thus is given by

$$f(x) = \mathrm{d}F(X)/\mathrm{d}x \tag{1.5}$$

From the above equations, we can write

$$F(x) = \int_{-\infty}^{x} f(z)\mathrm{d}z \tag{1.6}$$

The average value or mean (m) of a random variable X, also called the expectation of X, is denoted by $E(X)$. For a discrete random variable (X_d), where n is the total number of possible outcomes of values x_1, x_2, ..., x_n, and where the probabilities of the outcomes are $P(x_1)$, $P(x_2)$, ..., $P(x_n)$, it can be shown that

$$m = E(X_\mathrm{d}) = \sum_{i=1}^{n} xi \cdot P(xi) \tag{1.7}$$

For a continuous random variable X_c, with PDF $f_c(x)$, it can be represented as

$$m = E(X_c) = \int_{-\infty}^{\infty} x \cdot f(x)\mathrm{d}x \tag{1.8}$$

The mean square value can be represented as

$$E(X_c^2) = \int_{-\infty}^{\infty} x^2 f(x)\mathrm{d}x \tag{1.9}$$

A useful number to help in evaluating a continuous random variable is one that gives a measure of how widely its values are spread around its mean. Such a number is the root mean square value of $(X - m)$ and is called the standard deviation σ of X.

The square of the standard deviation, σ^2, is called the variance of X and is given by

$$\sigma^2 = E[(X-m)^2] = \int_{-\infty}^{\infty} (x-m)^2 f(x)\mathrm{d}x \tag{1.10}$$

The Gaussian (or normal PDF) is very important in wireless transmission. The Gaussian probability density function $f(x)$ is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(x-m)^2/2\sigma^2} \tag{1.11}$$

when $m = 0$ and $\sigma = 1$ the normalized Gaussian probability density function is achieved.

1.1.1.8 Basic Information Theory

Information theory was developed to find the fundamental limits on data compression and reliable data communication. It is based on probability theory and statistics. A key measure of information in the theory is known as information entropy, which is usually expressed by the average number of bits needed for storage or communication. The entropy is a measure of the average information content per source symbol. The entropy, H, of a discrete random variable X is a measure of the amount of *uncertainty* associated with the value of X. If X is the set of all messages x that X could be, and $p(x)$ is the probability of X given x, and then the entropy of X is defined as

$$H(x) = E_x[I(x)] = \sum_{x \in \chi} p(x) I(x) = \sum_{x=0}^{\chi-1} p(x) \cdot \log_2[1/p(x)] \tag{1.12}$$

In the above equation, $I(x)$ is the self-information, which is the entropy contribution of an individual message. The special case of information entropy for a random variable with two outcomes is the binary entropy function:

$$H_b(p) = -p \log_2 p - (1-p)\log_2(1-p) \tag{1.13}$$

We assume that the source is memory-less, so the successive symbols emitted by the source are statistically independent. The entropy of such a source is

$$H_b(p) = -p_0 \log_2 p_0 - p_1 \log_2 p_1 = -p_0 \log_2 p_0 - (1-p_0)\log_2(1-p_0) \tag{1.14}$$

From the above equation, we can observe

1. When $p_0 = 0$, the entropy $= 0$, when $p_0 = 1$, the entropy $= 0$.
2. The entropy $H_b(p)$ attains its maximum value $H_{max} = 1$ bit, when $p_1 = p_0 = 1/2$ for example, symbol 0 and 1 are equally probable. This is shown in Figure 1.11.

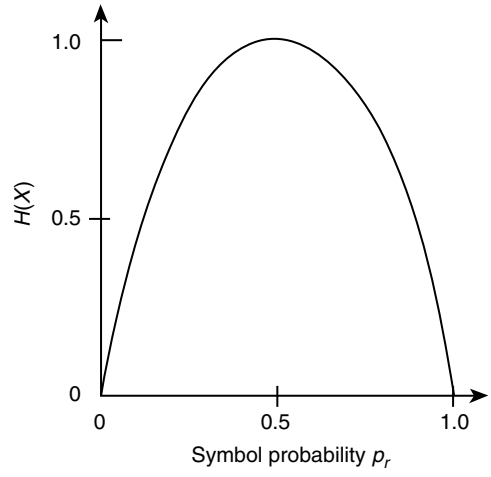


Figure 1.11 Entropy function $H(p)$

A binary symmetric channel (BSC) with crossover probability p is a binary input, binary output channel that flips the input bit with probability p. The BSC has a capacity of $1 - H_b(p)$ bits per channel use, where H_b is the binary entropy function as shown in the Figure 1.12.

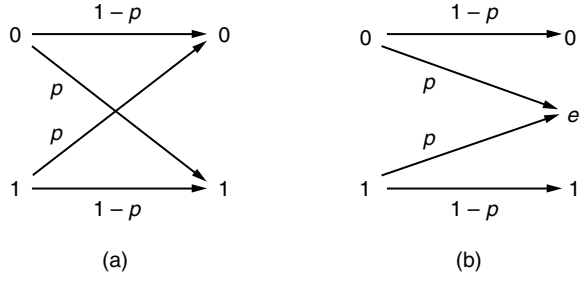


Figure 1.12 (a) A binary symmetric channel with crossover probability p is a binary input. (b) A binary erasure channel (BEC) with erasure probability p is a binary input

A binary erasure channel (BEC) with erasure probability p is a binary input, ternary output channel. The possible channel outputs are 0, 1, and a third symbol "e" called an erasure. The erasure represents complete loss of information about an input bit. The capacity of the BEC is $1 - p$ bits per channel use.

Let $p(y|x)$ be the conditional probability distribution function of Y for a given X. Consider the communications process over a discrete channel, for example, X represents the space of messages transmitted, and Y the space of messages received during a unit time over our channel. The appropriate measure to maximize the rate of information is the mutual information, and this maximum mutual information is called the channel capacity and is represented by:

$$C = \max_{f} I(X;Y) \tag{1.15}$$

1.1.1.9 Power and Energy of a Signal

The energy (and power) of a signal represent the energy (or power) delivered by the signal when it is interpreted as voltage or current source feeding a 1 ohm resistor. The energy content of a signal $x(t)$ is defined as the total work done and is represented as:

$$E(x) = \int_{-\infty}^{\infty} |x(t)|^2 \mathrm{d}t \tag{1.16}$$

The power content of a signal is defined as work done over time and is represented as:

$$P(x) = \lim_{T \to \infty} \int_{-T/2}^{T/2} |x(t)|^2 \mathrm{d}t \tag{1.17}$$

Conventionally, power is defined as energy divided by time. A signal with finite energy is called an energy-type signal and a signal with positive and finite power is a power-type signal. A signal is energy type if $E(x) < \infty$ and is power type if $0 < P < \infty$.

1.1.1.10 Bandwidth (BW)

The signal occupies a range of frequencies. This range of frequencies is called the bandwidth of the signal. In general, the bandwidth is expressed in terms of the difference between the highest and the lowest frequency components in the signal. A baseband signal or low pass signal bandwidth is a specification of only the highest frequency limit of a signal. A non-baseband bandwidth is the difference between the highest and lowest frequencies.

As the frequency of a signal is measured in Hz (Hertz), so, the bandwidth is also expressed in Hz. Also, we can say that the bandwidth of a signal is the frequency interval, where the main part of the power of the signal is located. The bandwidth is defined as the range of frequencies where the Fourier transform of the signal has a power above a certain amplitude threshold, commonly half the maximum value (half power ~ -3 dB, as $10 \log_{10}(P/P_{\text{half}}) = 10 \log_{10}(1/2) = -3$. Power is halved ($P/2$) at the 3 dB points for example, $P = P/2 = (V_0/\sqrt{2})(I_0/\sqrt{2})$, where V_0 and I_0 are the peak amplitude of voltage and current, respectively; refer to Figure 1.13.

However, in digital communication the meaning of "bandwidth" has been clouded by its metaphorical use. Technicians sometimes use it as slang for baud, which is the rate at which symbols may be transmitted through the system. It is also used more colloquially to describe channel capacity, the rate at which bits may be transmitted through the system.

Bit Rate – This is the rate at which information bits (1 or 0) are transmitted. Normally digital system require greater BW than analog systems.

Baud – The baud (or signaling) rate defines the number of symbols transmitted per second. One symbol consists of one or several bits together, based on the modulation technique used. Generally, each symbol represents n bits, and has M signal states, where $M = 2^n$. This is called M-ary signaling.

1.1.1.11 Channel Capacity

The maximum rate of communication via a channel without error is known as the capacity of the channel. In a channel where noise is present, there is an absolute maximum limit for the bit rate of transmission.

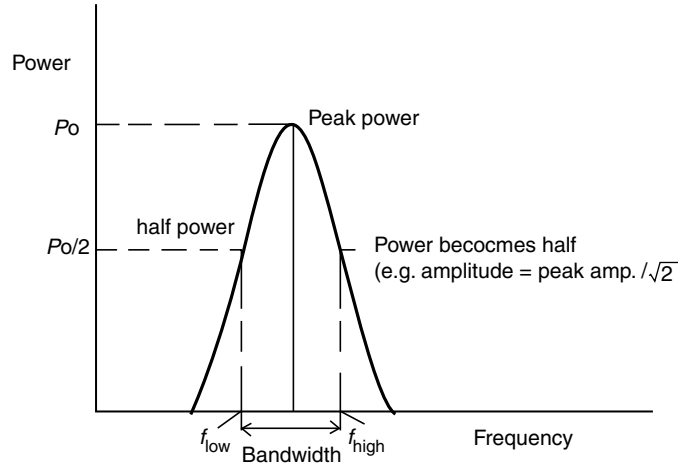


Figure 1.13 Bandwidth of a signal

This limit arises when the number of different signal levels is increased, as in such a case the difference between two adjacent sampled signal levels becomes comparable to the noise level. Claude Shannon extended Nyquist's work to a noisy channel.

Applying the classic sphere scenario, we can obtain an estimate of the maximum number of code words that can be packed in for a given power constant P, within a sphere of radius $\sqrt{(NP)}$. The noise sphere has a volume of $\sqrt{(N\sigma^2)}$. Thus, as shown in the Figure 1.14, the maximum number of code words that can be packed in with non-overlapping noise spheres is the ratio of the volume of the sphere of radius $\sqrt{\{(N\sigma^2) + NP)\}}$, to the volume of the noise sphere: $[\sqrt{\{(N\sigma^2) + NP)\}}]^N/[\sqrt{(N\sigma^2)}]^N$, where, N is the signal space dimension and σ^2 is the variance of the real Gaussian random variable

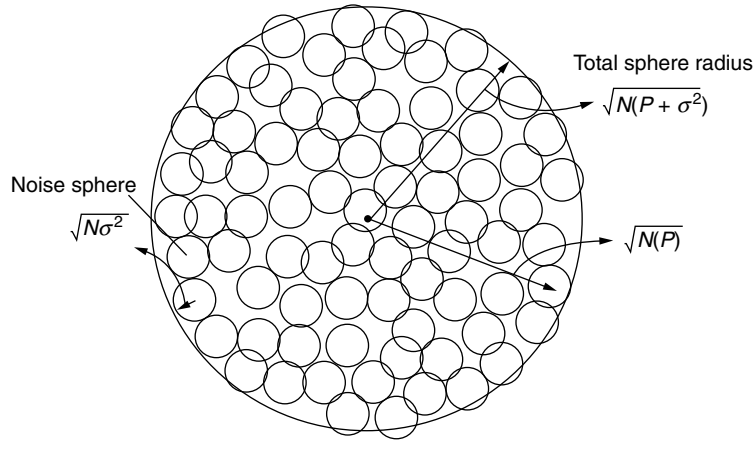


Figure 1.14 Number of noise spheres that can be packed into N dimensional signal space

with mean μ. This equation implies that the maximum number of bits per symbol that can be reliably communicated is

$$(1/N)\log\left[\sqrt{\{(N\sigma^2)+NP)\}}\right]^N/\left[\sqrt{(N\sigma^2)}\right]^N = (1/2)\log(1+P/\sigma^2) \tag{1.18}$$

This is indeed the capacity of the AWGN (additive white Gaussian noise) channel. In later chapters, we will see that for complex channels, the noise in I and Q components is independent. So it can be thought of as two independent uses of a real AWGN channel. The power constraint and noise per real symbol are represented as $P_{av}/2B$ and $N_0/2$, respectively. Hence the capacity of the channel will be

$$C = (1/2)\log(1+P_{av}/N_0 \cdot B) \text{ bits per real dimension} = \log(1+P_{av}/N_0 \cdot B) \text{ bits per complex dimension} \tag{1.19}$$

This is the capacity in bits per complex dimension or degrees of freedom. As there are B complex samples per second, the capacity of the continuous time AWGN channel is

$$C_{awgn}(P_{av}, B) = B\log(1+P_{av}/N_0 \cdot B)\text{bits/s} \tag{1.20}$$

Now, the signal to noise ratio (SNR) $= (P_{av}/N_0 \cdot B)$ – which is the SNR per degree of freedom. So, the above equation reduces to

$$C_{awgn} = \log(1+\text{SNR})\text{ bits/s/Hz} \tag{1.21}$$

This equation measures the maximum achievable spectral efficiency through the AWGN channel as a function of the SNR.

1.2 Introduction to Wireless Telecommunication Systems

The medium for telecommunication can be copper wire, optical fiber lines, twin wire, co-axial cable, air or free space (vacuum). To exchange information over these mediums, basically the energy is transferred from one place to the other. We know that there are two ways by which energy can be transferred from one place to another: (1) through the bulk motion of matter or (2) without the bulk motion of matter, which means via waves. With waves the energy/disturbance progresses in the form of alternate crests and troughs. Again, the waves are of two types: (1) mechanical waves and (2) electromagnetic waves. A *mechanical wave* can be produced and propagated only in material mediums that posses elasticity and inertia. These waves are also known as elastic waves; a sound wave is an example of an elastic wave. An *electromagnetic wave* does not require any such material medium for its propagation, it can travel via free space; light is an example of an electromagnetic wave. This is why we see the light from the sun, but do not hear any sound of bombardments from the sun.

In a conventional wire-line telephone system, the wire acts as the medium through which the information is carried. However, with a wireless system, as the name WIRELESS indicates, there is no wire, so removes the requirement for a wire between the two users. This means that the information has to be carried via air or free space. Thus the air or free space medium will be used for transmitting and receiving the information. But, *what will help to carry the information through this free space or air?* The answer was given in the previous paragraph – electromagnetic waves. As electromagnetic waves can travel through free space and air, we can use electromagnetic waves to send–receive information. Thus we conclude that for wireless communication, air or free space will be used as the channel/medium and electromagnetic waves will be used as the carrier. Then next problem to arise is how to generate this electromagnetic (EM) wave?

1.2.1 Generation of Electromagnetic Carrier Waves for Wireless Communication

In 1864 James Clark Maxwell theoretically predicted the existence of EM waves from an accelerated charge. According to Maxwell, an accelerated charge creates a magnetic field in its neighborhood, which in turn creates an electric field in this same area. A moving magnetic field produces an electric field and vice versa. These two fields vary with time, so they act as sources of each other. Thus, an oscillating charge having non-zero acceleration will emit an EM wave and the frequency of the wave will be same as the oscillation of the charge.

Twenty years later, in the period 1879–1886, after a series of experiments, Heinrich Hertz come to the conclusion that an oscillatory electrical charge $q = q_o \sin \omega t$ radiates EM waves and these waves carry energy. Hertz was also able to produce EM waves of frequency 3×10^{10} Hertz. The experimental setup is shown in the Figure 1.15. To detect EM waves, he also used a loop S, which is slightly separated as shown in the figure. This is the basis for the theory of antenna.

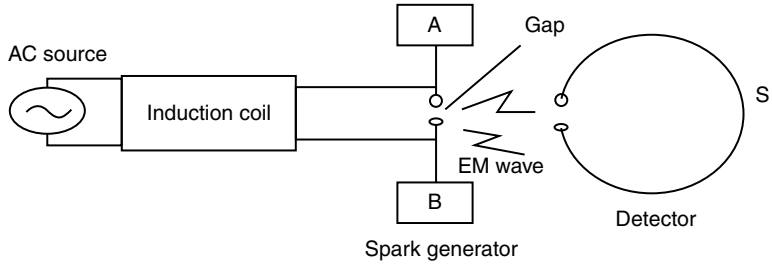


Figure 1.15 Hertz experiment, generation of EM wave

In 1895–1897, Jagdish Bose also succeeded in generating EM waves of very short wavelength (~25 mm). Marconi, in Italy in 1896, discovered that if one of the spark gap terminals is connected to an antenna whilst the other terminal is earthed, then under these conditions the EM waves can travel up to several kilometers. This experiment launched a new era in the field of wireless communication.

So, now we know that an antenna is the device that will help to transmit and receive the EM waves though the air or free space medium. Next, we will see how the antenna actually does this.

1.2.2 Concept of the Antenna

An antenna is a transducer that converts electrical energy into an EM wave or vice versa. Thus, an antenna acts as a bridge between the air/free-space medium (where the carrier is the EM wave) and the communication radio device (where the energy is carried in the form of low frequency electrical signals). From network theory, we know that if the terminated impedance is matched with a port, then only the maximum power will be transmitted to the load, otherwise a significant fraction of it will be reflected back to the source port. Basically, the antenna is connected to a communication device port, so the impedance should be matched to transfer maximum power; similarly on the other side, the antenna is also connected to the air/free-space, so the impedance should be matched on that side to transfer maximum power.

Physically, an antenna is a metallic conductor, it may be a small wire, a slot in a conductor or piece of metal, or some other type of device.

1.2.2.1 Action of an Antenna

When an ac signal is applied to the input of an antenna, the current and voltage waveform established in the antenna is as shown in the Figure 1.16. This is shown for a length of a wire of $\lambda/2$.

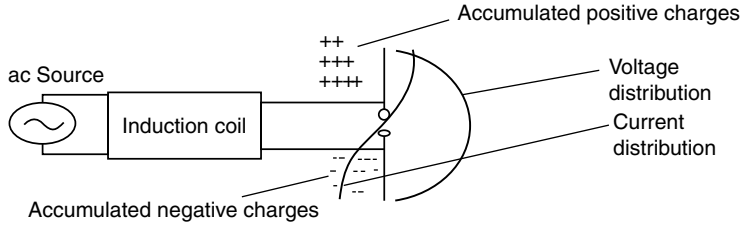


Figure 1.16 Charge, current and voltage distribution

If the length changes then the distribution of charge on the wire will also change and the current–voltage waveform will also change. In Figure 1.17, various waveforms for different antenna lengths are shown. Wherever the voltage is a maximum, then the current is a minimum as they are 90° out of phase.

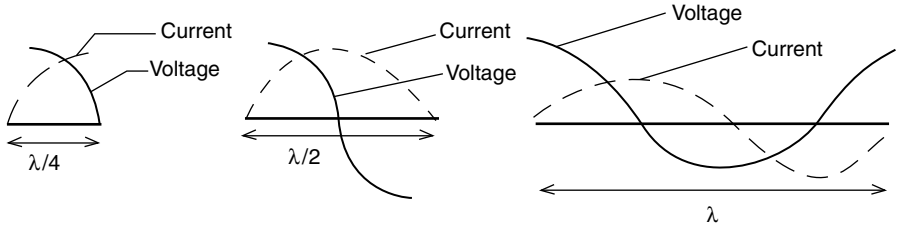


Figure 1.17 Voltage and current distribution across the antenna conductor for different lengths

It is evident that the minimum length of an antenna required to transmit or receive a wave effectively will be at least $\lambda/4$. This is because the property of a periodic wave resides mainly in any section of length "$\lambda/4$" and then it repeats over an interval of $\lambda/4$. This means that by knowing only $\lambda/4$ of a wave, we can reconstruct the wave. Basically, the distance between the maxima and minima of a signal waveform is $\lambda/4$, so all the characteristics of the wave will remain there.

How Does the EM Wave Carry Energy Over a Long Distance? – As shown in Figure 1.18, when an RF (radio frequency) signal is applied to the antenna, at one instant of time (when the signal is becoming positive- (a)), conductor A (the dipole) of the antenna will be positive due to a lack of electrons, and at that instant B will be negatively charged due to the accumulation of electrons. As the end points of A and B is an open circuit, so charge will accumulate over there. From the Maxwell equations, we know that electrical flux lines always start from the $+$ve point and end at the $-$ve point, and form a closed loop. A current flowing through a conductor produces a magnetic field and voltage produces an electric field. So, lines of electric and magnetic fields will be established across the antenna conductors as shown in Figure 1.18a.

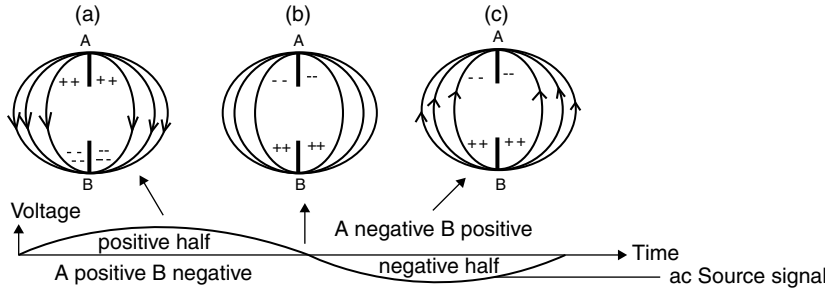


Figure 1.18 Creation of flux lines with the RF signal variation

In the next instance of time (b), when the applied input RF signal decreases to zero, at this time there is no energy to sustain the closed flux loops of the magnetic and electric field lines, which were created in the previous instance of time. So, these will detach from the antenna and remain self sustained.

In the next instance of time (c), when A becomes negative and B positive, the electric and magnetic flux lines will be created again, but this time the direction will be reversed, as the $+$ve and $-$ve points are interchanged. Now, when the RF signal goes to zero, these flux lines again becomes free and self sustained.

The flux lines that were created in the first instance when A was positive and B was negative will have the opposite direction to the flux lines that were created when A is $-$ve and B is $+$ ve. Thus, these flux lines will repel each other, and will move way from the antenna as shown in Figure 1.19.

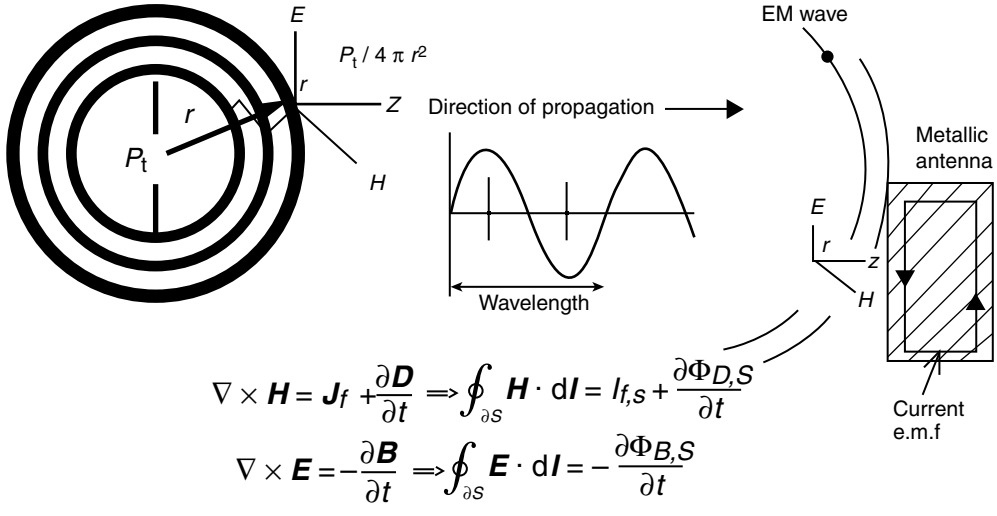


Figure 1.19 Signal reception by the antenna

This phenomenon is repeated again and again according to the variation with time of the input radio frequency (RF) signal (for example, according to its frequency) and a series of detached electric and magnetic fields move onwards from the transmitting antenna.

In a 3D space (free space) these flux lines will actually be spherical in shape. If the radius of the sphere is r, and the power of the RF signal (V^*I) during the time when these flux lines were generated is P_t, then the

power P_t is spread over the surface of a sphere whose radius is r, so the power density will be $P_t/4\pi r^2$. Taking this energy, the flux lines will move away from the transmitting antenna. Now, as they move away from the antenna, the size of the sphereincreases, r increases but the same power (P_t) is contained within it. Thus the power density $P_t/4\pi \cdot r^2$ decreases as it travels far from the transmitting antenna (for example, increase in r), and the problem of transmission of electrical energy via air/free-space is solved.

How is Energy Received on the Other Side? – Again, the antenna helps to solve this problem too. It transforms the received EM wave into an electrical signal. When the transmitted wave arrives at the receiving end, it tries to penetrate via another the metallic antenna. We know that the EM wave consists of an electric field and a magnetic field and that these are perpendicular to each other, and also that they are perpendicular to the direction of propagation. Thus when the EM wave touches the metallic antenna (from Maxwell's third equation) the magnetic field (H) will generate a surface current on the metallic antenna, which will try to penetrate via the metal (as it is a good conductor). However, it will die down after traveling a thickness of the skin depth, and, the EM wave will generate an electrical current in the metal body of the antenna. Similarly (from Maxwell's fourth equation), the electric field will generate an electric voltage in the antenna, as shown in Figure 1.19.

This phenomenon can be experienced by placing a radio inside a closed metallic chamber and finding that it does not play, as the EM wave can not penetrate via the thick metallic wall. However, it can penetrate through a concrete wall. For the same reason, a mobile telephone call disconnects inside an enclosed metallic lift due to the degradation of the signal strength.

Thus we have converted the transmitted energy (which was transmitted using the carrier of the EM waves) back into the electrical signal through the help of another antenna. So the antenna helped in transmitting and receiving the information through the air. As the user wants to send and receive the information, ideally the user should have both transmitting and receiving antennas. However, in general, in a mobile device, the same antenna is used for transmission as well as receiving purposes (refer to Chapter 4).

Thus we now know how to transmit and receive the information via the air (for example, via a wireless medium) using antenna. However, the problem at this stage is whether the baseband signal is transmitted directly, as its frequency is low (~KHz), and so it can not be sent directly via the air due to the following problems:

1. A larger antenna length ($\sim\lambda/4$) is required.
2. Much less BW is available at the lower frequency region.

The solution to this is to up-convert the baseband signal to a high frequency RF signal at the transmitter and then similarly down-convert at the receiver for example, which requires RF conversion techniques. How is the baseband signal up-converted/down-converted? The solution for up-conversion is the use of analog or digital modulation and mixing techniques (on a transmitter block) and the solution for down-conversion is the use of demodulation mixing techniques (on a receiver block). These will be discussed in more detail in Chapter 4. Next we will establish what else, apart from the antenna, is required inside the transmitter and receiver to transmit or receive the information.

1.2.3 Basic Building Blocks of a Wireless Transmitter and Receiver

We know that a digital system is more immune to noise, and that in addition to this there are many such advantages of digital systems over the old analog systems. So, from the second generation onwards wireless systems have been designed with digital technology. However, there is a problem, in that voice and video signals are inherently analog in nature. So how can these signals be interfaced with a digital system? These signals have to be brought into the digital domain for processing using an analog-to-digital converter (ADC) and then again reverted back into an analog signal using a digital-to-analog converter (DAC) and then sent via an antenna. A typical transmitter block diagram of a wireless system is shown in the Figure 1.20.

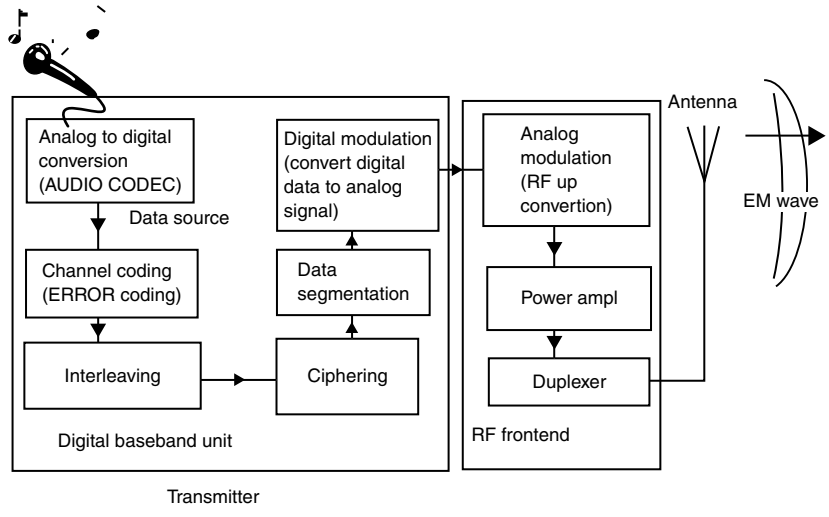


Figure 1.20 Transmitter system block diagram

As shown in the Figure 1.20, on the transmitter side, when the user speaks in front of a microphone, it generates an electrical signal. This signal is sampled and converted into digital data and then fed to the source codec, for example, a speech codec unit (this is discussed in more details in Chapter 8), which removes the redundant data and generates the information bits. These data are then fed into the channel coder unit. When the signal travels via the medium, during this time it can be affected by signal noise, so we need some type of protection against this. The channel coder unit inserts some extra redundant data bits using an algorithm, which helps the receiver to detect and correct the received data (this is discussed in more detail in Chapter 3). Next, it is fed to an interleaving block. When data pass through the channel, this time there may be some type of burst noise in the channel, which can actually corrupt the entire data during this burst period. Although the burst lasts for only a short duration, its amplitude is very high, so it corrupts the data entirely for that duration. In order to protect the data from burst error, we need to randomize the data signal (separate consecutive bits) over the entire data frame, so that data can be recovered, although some part will be corrupted completely. An interleaving block helps in this respect (this is discussed in detail in Chapters 3 and 8). Next, it is passed to a ciphering block, where the data are ciphered using a specific algorithm. This is basically done for data security purposes, so that unauthorized bodies cannot decode the information (ciphering is discussed in Chapters 7 and 9). Then the data are put together in a block and segmented according to the data frame length.

The data processing is now over, and next we have to pass it for transmission. This data signal can not be sent directly using an antenna, because, it will be completely distorted. Also, the frequency and amplitude of the data signal is less, as we know the length of the transmitting antenna should be a minimum of the order of $\lambda/4$. So, the required size of the antenna will have to be very large, which is not feasible. This is why we need to convert it into a high frequency analog signal using modulation techniques. The digital modulator block transfers the digital signal into a low amplitude analog signal. As the frequency of this analog signal may be less, we therefore may need to convert it into a high frequency RF signal, where the wavelength is small and the required antenna length will also be small. The analog modulator block helps to up-convert the analog signal frequency to a high RF carrier frequency (the modulation technique is discussed in Chapter 5). It is then fed into a power amplifier to increase the power of the signal, and after that it is passed to duplexer unit. We know that our telephone acts as a transmitter as well as a receiver, so it

should have both a transmission and a reception block inside. As we want to use the same antenna for transmission and for reception purposes, so we connect a duplexer unit to separate out the transmitting and receiving signals. The transmitted signal goes to an antenna and this antenna radiates the signal through air/free space.

As shown in Figure 1.21, the reverse sequence happens in the receiver, once it receives the signal. The antenna actually receives many such EM waves, which are of different frequencies. Now, of these signals, the receiver should receive only the desired frequency band signal that is transmitted by the transmitter. This is done by passing the signal from the duplexer via a band-pass filter. This filter will allow only the desired frequency band signal to pass through it and the remiander will be blocked. After this, it passes via the low noise amplifier to increase the power of the feeble signal that is received. Then it is RF down-converted (analog demodulated), digital demodulated, de-interleaved, de-ciphered, decoded and the information blocks are recovered. Next, it is passed to the source decoder and ultimately converted back into an analog voice signal by the DAC and passed to the speaker to create the speech signal.

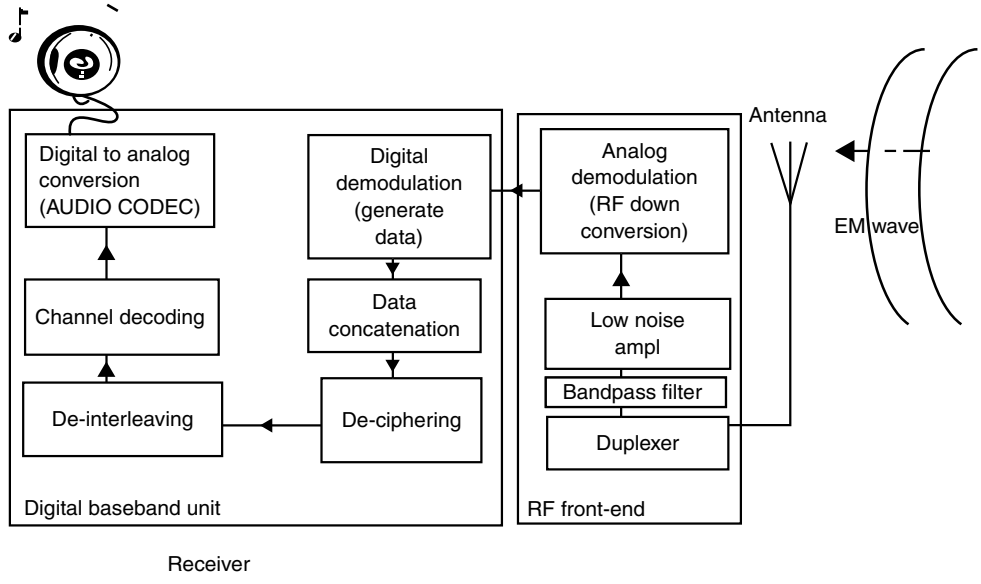


Figure 1.21 Receiver system block diagram

In this figure the front blocks deals with the analog signals. This front-end part that deals with the analog signal is called the RF front-end or RF transceiver. Sometimes the digital modulation/de-modulation unit is also put together with the RF transceiver unit (the design of the RF transmitter and receiver are discussed in detail in Chapter 4). The back-end part, where the baseband digital signal (baseband signal) is processed and the signaling and protocol aspects are dealt with is known as the baseband module.

1.2.4 The Need for a Communication Protocol

We have seen how the sender and receiver communicate via a wireless channel using a transmitter and a receiver. To set up, maintain and release any communication between the users, we need to follow certain

protocols, which will govern the communication between the various entities in the system. The Open System Interconnection (OSI) reference model was specified by ITU, in cooperation with ISO (International Organization for Standardization) and IEC (International Electrochemical Commission). The OSI reference model breaks down or separates the communication process into seven independent layers, as shown in Figure 1.22. Generally, in wireless communication systems, similar to the ISO-OSI model, a layered protocol architecture is used. However, in most instances, only the lower three layers (physical, data link and network) are modified according to the needs of the various wireless systems. A layer is composed of subsystems of the same rank of all the interconnected systems. The functions in a layer are performed by hardware or software subsystems, and are known as entities. The entities in the peer layers (sender side and receiver side) communicate using a defined protocol. Peer entities communicate using peer protocols. Data exchange between peer entities is in the form of protocol data units (PDUs). All messages exchanged between layer N and layer $(N-1)$ are called primitives. All message exchange on the same level (or layer) between two network elements, A and B, is determined by what is known as peer-to-peer protocol.

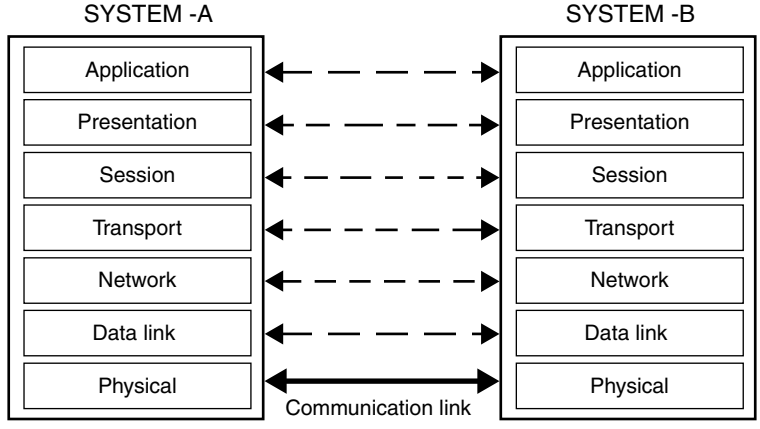


Figure 1.22 Peer to peer protocol layers

Various wireless standards, such as GSM (Global Systems for Mobile Communication) and UMTS (Universal Mobile Telecommunications System) have been developed following different sets of protocols as required for the communications. This is discussed later in more detail in the appropriate chapters.

1.3 Evolution of Wireless Communication Systems

Since the invention of the radio, wireless communication systems have been evolving in various ways to cater to the need for communications in a variety of segments. These are broadly divided into two categories:

1. **Broadcast communication systems** – This is typically simplex in nature, which means that one side transmits and other side receives the information, and there is no feedback from the receiver side. Typically radio and TV transmissions are of this nature. This is generally point-to-multi-point in nature. The transmitter broadcasts the information and all the intended receivers receive that information by tuning the receiver.

2. **Point-to-point communication systems** – This is typically duplex in nature, for example, both sides can transmit as well as receive the information. In this instance, there is always a transmitter–receiver pair, for example, each transmitter sends to a particular receiver and vice versa over the channel. Here, one thing we need to remember is that in the case of a telecommunication system, every user would like to be able to connect with every other user. So if there are n users in a system, then a total of $n\,(n-1)/2$ links are required to connect them. This is very ineffective and expensive. To overcome this issue a central switching office or exchange is placed in between, as shown in Figure 1.23. With the introduction of the switching systems (the exchange), the subscribers are now not directly connected to one another; instead they are connected to the local exchange.

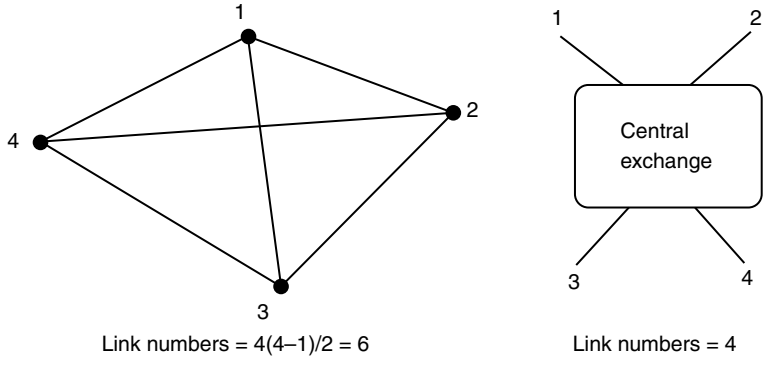


Figure 1.23 A network of four users with point-to-point links without and with central exchange

Examples of point-to-point wireless communication systems are walkie-talkies, cordless phones, mobile phone and so on. These are again classified into two categories: (a) fixed wireless systems and (b) mobile wireless systems. In the case of walkie-talkies and cordless phones, the transmitter and receiver pairs are fixed, whereas for mobile phones the transmitter and receiver pairs are not fixed, they can vary from time to time according to the user call to a particular called party, and these are called multi-user systems.

The basic differences between these two are given in the Table 1.2.

Table 1.2 Differences between the broadcast and point-to-point systems

Broadcast	Point-to-point
Broadcast systems are typically unidirectional	These systems are bi-directional
Range is very large	Range is less
High transmitted power	Moderate transmitted power
No feed-back from receiver	Generally have feedback mechanism to control power
Less costly receiver	Generally receivers are expensive
Typically one transmitter and many receivers	Generally works with transmitter and receiver pairs

So far we have learnt the basic techniques for sending and receiving information via a wireless channel. This is the basic concept used for all types of wireless communications. If the sender and the receiver are fixed at their respective locations, then whatever we have discussed so far is sufficient, but this is not the real situation. The user wants to move around while connected. Thus we have to support user's mobility. How is that done?

1.3.1 Introduction of Low Mobility Supported Wireless Phones

Cordless phone systems provide the user with a limited range and mobility, for example, the user always has to be close to the fixed base unit, in order to be connected with the telecommunication network. Cordless telephones are the most basic type of wireless telephone. This type of telephone system consists of two separate units, the base unit and the handset unit. Generally the base unit and handset unit are linked via a wireless channel using a low power RF transmitter–receiver. The base unit is connected to the tip and ring line of the PSTN (public switched telephone network) line using an RJ11 socket.

The cordless telephone systems are full-duplex communication systems. The handset unit sends and receives information (voice) to and from the base unit via a wireless link and as the base unit is connected to the local telephone exchange via a wired link, the call from the handset is routed to the exchange and finally from the exchange to the called party. In this instance, the range and mobility available to the user are very limited.

1.3.2 Introduction to Cellular Mobile Communication

The limitations of cordless phones are the restricted range and mobility. However, the users want more mobility to be able to roam around over a large area, and at the same time they want to be connected with the other users. So, how can this problem be addressed?

One solution could be to transmit the information with a huge transmitting power, but this has several problems: first of all, if the frequencies of the transmitters of the various users are not properly organized then they will interfere with each other, and secondly this will only cover a particular zone. Thus this will also be restricted to local communication only. The ideal solution for this problem was discovered at the Bell Labs, by introducing the cell concept, where a region is geographically divided into several cells as shown in the Figure 1.24. Although in theory, the cells are hexagonal, in practice they are less regular in shape. Ideally the power transmitted by the transmitter covers a spherical area (isotropic radiation for example, transmits equally in all directions), so naturally for omnidirectional radiation the cell shapes are circular.

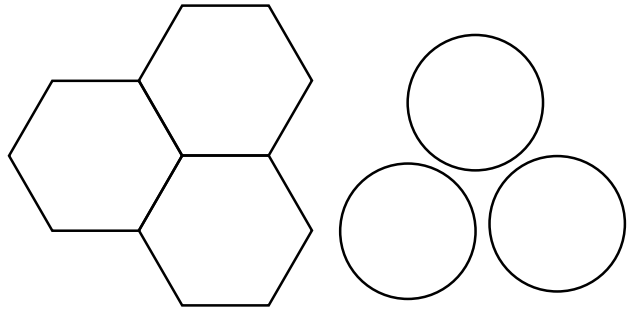


Figure 1.24 Shape of cells in a cellular network

Users in a cell are given full mobility to roam around using a mobile receiver. Typically, a cell site (for any cellular provider) should contain a radio transceiver and controller (similar to the base unit in a cordless system), which will manage, send and receive information to/from the mobile receivers (similar to the cordless handset). These radio transmitters are again connected with the radio transmitters in the other cell and with the other PSTN or mobile or data networks. This is just like a star-type inter-connection of topology but in a wireless medium. This means that now users in any cell are connected to all the other users in different cells or different networks for example, they are globally connected. So, all users get the mobility they wanted. As the user's receivers are by nature mobile, this is why it is called a mobile handset/receiver and as communication takes place in a cellular structure, so it is called cellular communication.

One of the main advantages of this cellular concept is frequency (or more specifically channel) reuse, which greatly increases the number of customers that can be supported. In January 1969, Bell Systems employed frequency reuse in a commercial service for the first time. Low-powered mobiles and radio equipment at each cell site permitted the same radio frequencies to be reused in different distant cells, multiplying the calling capacity without creating interference. This spectrum efficient method contrasts sharply with earlier mobile systems that used a high powered, centrally located transmitter to communicate with high powered car mounted mobiles on a small number of frequencies, channels which were then monopolized and could not be re-used over a wide area. A user in one cell may wish to talk to a user in another cell or to a land phone user, so the call needs to be routed appropriately. This is why another subsystem is added with a base station for switching and routing purposes, as shown in Figure 1.25. This is discussed in more detail in Chapter 6 onwards, and the architecture of this network system is different for different mobile communication standards.

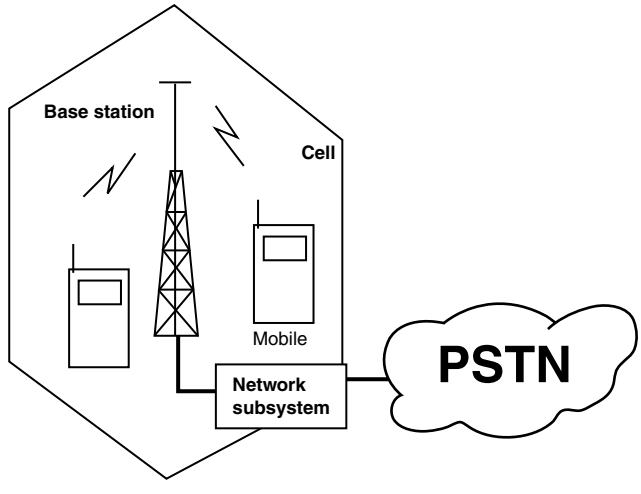


Figure 1.25 Architecture of a typical cellular system

As the frequency is treated as a major resource, so it should be utilized in such a way to enhance the capacity of the subscribers in a cell. We know there will be many mobiles and one base station (for one operator) in a cell, so all the mobiles in a cell have to always be listening to the base station in order to receive any information related to the system, synchronization, power, location and so on. Every base station uses a broadcast channel to send this information. Similarly, it uses a different channel to convey any incoming information to the mobile. To start any conversation, the mobile requests a

channel from the base station, then the base station assigns a free channel to that mobile for communication and the mobile then releases it after it has been used. These procedures are discussed in depth in the respective chapters.

1.3.2.1 Concept of Handover

Handover is a key concept in providing this mobility. It helps a user to travel from one cell to another, while maintaining a seamless connection. A handover is generally performed when the quality of the radio link between the base station and the moving mobile terminal degrades. The term "handover" refers to the whole process of tearing down an existing connection and replacing it by a new connection to the target cell into which the user is handed over, because there is a higher signal strength there. The network controller usually decides whether a handover to another cell is needed or not based on the information about the quality of the radio link contained in measurement reports from the mobile. Knowledge about radio resource allocation in the target cell and the correct release of channels after the handover is completed are vital for a successful handover. The inability to establish a new connection in the target cell for several reasons is referred to as a "handover failure." As expanding markets demand increasing capacity, there is a trend towards reducing the size of the cells in mobile communications systems. A higher number of small sized cells lead to more frequent handovers and this situation results in the need for a reliable handover mechanism for efficient operation of any future cellular mobile network.

1.3.3 Introduction to Mobile Handsets

A mobile handset is the equipment required by the user to send–receive information (voice, data) to/from another party via base station and network subsystems. This is like a wide area cordless telephone system with a long range of communication. In this equipment high-powered transmitters and elevated antennas are used to provide wireless telephony typically over a cell of radius of 20–30 miles.

Evolution of the Generations of Mobile Phones In 1887, Guglielmo Marconi was the first to show a wireless transmission using Morse code to communicate from ship to shore. Later, he commercialized his technology by installing wireless systems in transatlantic ocean vessels. Since then, Marconi wireless systems have been used to send distress calls to other nearby boats or shoreline stations, including the famous ship the *Titanic*. This first wireless system used a spark-gap transmitter, which could be wired to send simple Morse code sequences. Although the transmission of the signal was easy, the reception was not quite so simple. For this, Marconi used a *coherer*, a device that could only detect the presence or absence of strong radio waves. Since then things have come a long way, and wireless telephone systems have become fairly mature, with fixed wireless communication devices, cordless phones, short range wireless phone gradually being introduced.

In December 1947, Douglas H. Ring and W. Rae Young, Bell Labs engineers, proposed hexagonal shaped cells for mobile phones. Many years later, in December 1971, AT&T submitted a proposal for cellular service to the Federal Communications Commission (FCC), which was approved in 1982 for Advanced Mobile Phone Service (AMPS) and allocated frequencies in the 824–894 MHz band. In 1956, the first fully automatic mobile phone system, called the MTA (Mobile Telephone system A), was developed by Ericsson and commercially released in Sweden. This was the first system that did not require any type of manual control, but had the disadvantage that the phone weighed 40 kg. One of the first truly successful public commercial mobile phone networks was the ARP network in Finland, launched in 1971; this is also known as 0th generation cellular network.

In 1973, Motorola introduced Dyna-Tac, the world's first cell phone with a size of a house brick at $9 \times 5 \times 1.75$ inches, and weighing in at a strapping 2.5 lbs. This contained 30 circuit boards, and the

only supported features were to dial numbers, talk, and listen, with a talking time of only 35 minutes. In 1949, Al Gross (the inventor of the walkie-talkie) introduced the first mobile pager, for use by hospital doctors. In 1979, the first commercial cellular phone market opened up in Tokyo using a type of analog FM (frequency modulation) to send voice signals to users. Similar systems in North America and Europe followed. By the late 1980s, analog cellular communications were a commercial success, and companies were pressing government regulatory agencies to open up new radio spectra for more voice services. Nokia introduced the world's first handheld phone, the Mobira Cityman 900. On July 1, 1991, Nokia manufactured and demonstrated the first GSM (Global System for Mobile Communication) call.

First Generation (1G)

In the 1960s Bell Systems developed the Improved Mobile Telephone Service system (IMTS), which was to form the basis of the first generation mobile communications systems. 1G (first generation) is the name given to the first generation of mobile telephone networks. Analog circuit-switched technology is used for this system, with FDMA (Frequency Division Multiple Access), as an air channel multiple access technique, and worked mainly in the 800–900 MHz frequency bands. The networks had a low traffic capacity, unreliable handover, poor voice quality, and poor security. The examples of such 1G system are Analog Mobile Phone Systems (AMPS) – the analog systems implemented in North America, Total Access Communication Systems (TACS) – the system used in Europe and other parts of the world.

Second Generation (2G)

The problems and limitations of analog circuit based 1G mobile networks were soon realized, so the digital technology based 2G (second generation) mobile telephone networks were introduced. Many of the principles involved in a 1G system also apply to a 2G system, and both use a similar cell structure. However, there are differences in signal handling, and a 1G system is not capable of providing some of the more advanced features of a 2G system.

The most popular 2G system is GSM. In GSM 900, the band at 890–915 MHz is dedicated to uplink communications from the mobile station to the base station, and the band at 935–960 MHz is used for the downlink communications from the base station to the mobile station. Each band is divided into 124 carrier frequencies, spaced 200 kHz apart, in a similar fashion to the FDMA method used in 1G systems. Then each carrier frequency is further divided using TDMA into eight 577 μs long "time slots," each one of which represents one communication channel – the total number of possible channels available is therefore 124×8, producing a theoretical maximum of 992 simultaneous conversations. In the USA, a different form of TDMA is used in the system known as IS-136 D-AMPS, and there is another US system called IS-95 (CDMAone), which is based on the Code Division Multiple Access (CDMA) technique. 2G systems are designed as a voice centric communications network with limited data capabilities such as fax, Short Message Service (SMS), as well as Wireless Application Protocol (WAP) services.

Second Generation Plus

Owing to the rapid growth of the Internet, the demand for advanced wireless data communication services is increasing rapidly. As the data rates for 2G circuit-switched based wireless systems are too slow, the mobile systems providers have developed 2G + technology, which is packet-based and have a higher data speed of communication. Examples of 2G + systems are: High Speed Circuit-Switched Data (HSCSD), General Packet Radio Service (GPRS), and Enhanced Data Rates for GSM Evolution (EDGE). HSCSD is a circuit-switched based technology that improves the data rates up to 57.6 kbps by introducing 14.4 kbps data coding and by aggregating four radio channels timeslots of 14.4 kbps. The GPRS standard is packet-based and built on top of the GSM system. GPRS offers a theoretical maximum 171.2 kbps bit rate, when all eight timeslots are utilized at once. In Enhanced GPRS (EGPRS), the data rate per timeslot will be

tripled and the peak throughput, including all eight timeslots in the radio interface, will exceed 384 kbps. Enhanced Data Rates for GSM Evolution (EDGE) is a standard that has been specified to enhance the throughput per timeslot for both HSCSD and GPRS. The basic principle behind EDGE is that the modulation scheme used on the GSM radio interface should be chosen on the basis of the quality of the radio link. A higher modulation scheme is preferred when the link quality is good, and when the link quality is bad it uses modulation scheme with lower data rate support.

Third Generation (3G)

2G mobile communication systems have some limitations and disadvantages, such as lower system capacity, lower data rate, mostly voice centric, and so on. Hence the demand for a newer generation of telecommunication systems, which are known as third generation (3G) systems. Third generation systems support higher data transmission rates and higher capacity, which makes them suitable for high-speed data applications as well as for the traditional voice calls. The network architecture is changed by adding several entities into the infrastructure. Compared with earlier generations, a 3G mobile handset provides many new features, and the possibilities for new services are almost limitless, including many popular applications such as multimedia, TV streaming, videoconferencing, Web browsing, e-mail, paging, fax, and navigational maps.

Japan was the first country to introduce a 3G system due to the vast demand for digital mobile phones and subscriber density. WCDMA (Wideband Code Division Multiple Access) systems make more efficient use of the available spectrum, because the CDMA (Code Division Multiple Access) technique enables all base stations to use the same frequency with a frequency reuse factor of one. One example of a 3G system is Universal Mobile Telecommunication Systems (UMTS). UMTS are designed to provide different types of data rates, based on the circumstances, up to 144 kbps for moving vehicles, up to 384 kbps for pedestrians and up to 2 Mbps for indoor or stationary users. This is discussed from Chapter 12 onwards.

Fourth Generation (4G)

As 3G system also have some limitations, so researchers are trying to make new generations of mobile communication, which are known as the fourth generation (4G). 4G will be a fully IP-based (International Protocol) integrated system. This will be achieved after wired and wireless technologies converge and will be capable of providing 100 Mbit/s and 1 Gbit/s speeds both indoors and outdoors, with premium quality and high security. 4G will offer all types of services at an affordable cost, and will support all forthcoming applications, for example wireless broadband access, a multimedia messaging service, video chat, mobile TV, high definition TV content, DVB, minimal service such as voice and data, and other streaming services for "anytime-anywhere." The 4G technology will be able to support interactive services such as video conferencing (with more than two sites simultaneously), wireless internet and so on. The bandwidth would be much wider (100 MHz) and data would be transferred at much higher rates. The cost of the data transfer would be comparatively much less and global mobility would be possible. The networks will all be IP networks based on IPv6. The antennas will be much smarter and improved access technologies such as Orthogonal Frequency Division Multiplexing (OFDM) and MC-CDMA (Multi Carrier CDMA) will be used. All switches would be digital. Higher bandwidths would be available, which would make cheap data transfer possible. This is discussed in Chapter 16.

Today, the 4G system is evolving mainly through 3G LTE (Long Term Evolution) and WiMAX systems. People who are working with the WiMax technology are trying to push WiMax as the 4G wireless technology. At present there is no consensus on whether to refer to this as the 4G wireless technology. WiMax can deliver up to 70 Mbps over a 50 km radius. As mentioned above, with 4G wireless technology people would like to achieve up to 1 Gbps (indoors). WiMax does not satisfy the criteria completely. To overcome the mobility problem, 802.16e or Mobile WiMax is being standardized. The important thing to remember here is that all the research on 4G technology is based around OFDM.

Fifth Generation (5G)

5G (fifth generation) should make an important difference and add more services and benefits to the world over 4G. For example, an artificial intelligence robot with a wireless communication capability would be a candidate, as building up a practical artificial intelligence system is well beyond all current technologies. Also, apart from voice and video, the smell of an object could be transmitted to the distant user!

The evolution of mobile communication standards is shown in Figure 1.26 and added features are discussed in Table 1.3.

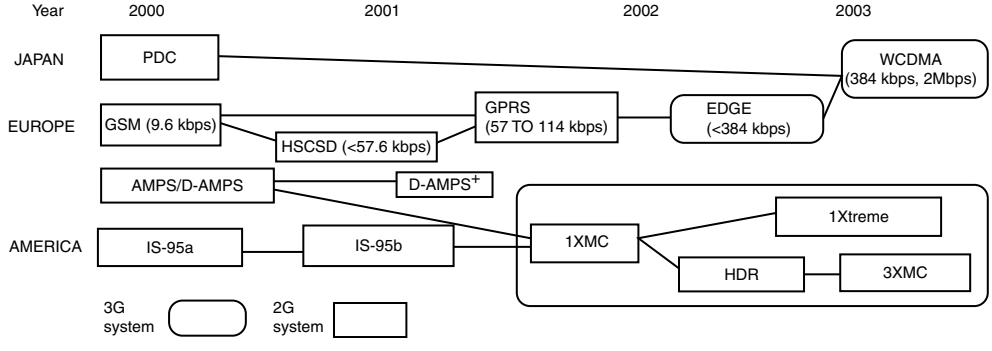


Figure 1.26 Evolution of standards and technologies

1.3.3.1 Basic Building Blocks of a Mobile Phone

The mobile phone is a complex embedded device, consisting of various modules to support its operations. These can be broadly divided into three categories: (1) modem – which takes care of transmission and reception of information over the channel and consists of a protocol processing unit, modulation/demodulation unit, and RF unit; (2) application – this is used to support various user applications, such as for a voice support microphone, speaker, speech coder, A/D–D/A converters and so on and if for example a video support camera or LCD are used; and (3) power – this module takes care of providing battery power to the different subsystems. A typical internal block diagram of a cellular mobile handset is shown in Figure 1.27. Based on the supported cellular mobile system standards (such as GSM or UMTS), the design of the RF, modulation, baseband, and protocol processing units section varies. However, the basic operational blocks remain more or less the same.

The basic blocks are briefly mentioned below, but these are described in detail later, in the appropriate chapters.

1. **Antenna** – Converts the transmitted RF signal into an EM wave and the received EM waves into an RF signal. The same antenna is used for transmission and reception, so there is a duplexer switch (or processor controllable switch) to multiplex the same antenna.
2. **RF Block** – In the receive path, the signal is first passed through the band-pass filter to extract the signal of the desired band, and then passed through the low noise amplifier to amplify the signal. Next, the input RF signal is down-converted into a baseband signal (using heterodyne or homodyne receiver architecture), so that sampling can be performed at a much lower rate. Sampling at the RF signal

Table 1.3 Generations of mobile phone features

Mobile system generation		Technology	Bandwidth (kbit/s)	Features
First generation mobile	AMPS/NMT	Advanced Mobile Phone System Nordic Mobile Telephony	9.6	Analog voice service No data capabilities
Second generation mobile	GSM	Global System for Mobile Communication	9.6 → 14.4	Digital voice service Advanced messaging Global roaming Circuit-switched data Telephone e-mail SMS Digital Text Delivery
	HSCSD	High-Speed Circuit Switched Data	9.6 → 57.6	Extension of GSM Higher data speeds
	GPRS	General Packet Radio Service	9.6 → 115	Extension of GSM Always-on connectivity Packet-switched data
	EDGE	Enhanced Data Rate for GSM Evolution	64 → 384	Extension of GSM Always-on connectivity Faster than GPRS Mobile banking Voicemail, web Mobile audio player Digital newspaper publishing Digital audio delivery Mobile radio, karaoke Push marketing/targeted programs Location-based services Mobile coupons

(*continued*)

Table 1.3 (*Continued*)

Mobile system generation		Technology	Bandwidth (kbit/s)	Features
Third Generation Mobile	IMT-2000/UMTS	International Mobile Telecommunications 2000/Universal Mobile Telecommunications System	64 → 2048	Always-on connectivity
				Global roaming
				IP-enabled
				Mobile video conferencing
				Video phone/mail
				Remote medical diagnosis and education
				Mobile TV/video player
				Advanced car navigation/city guides
				Digital catalog shopping
				Digital audio/video delivery

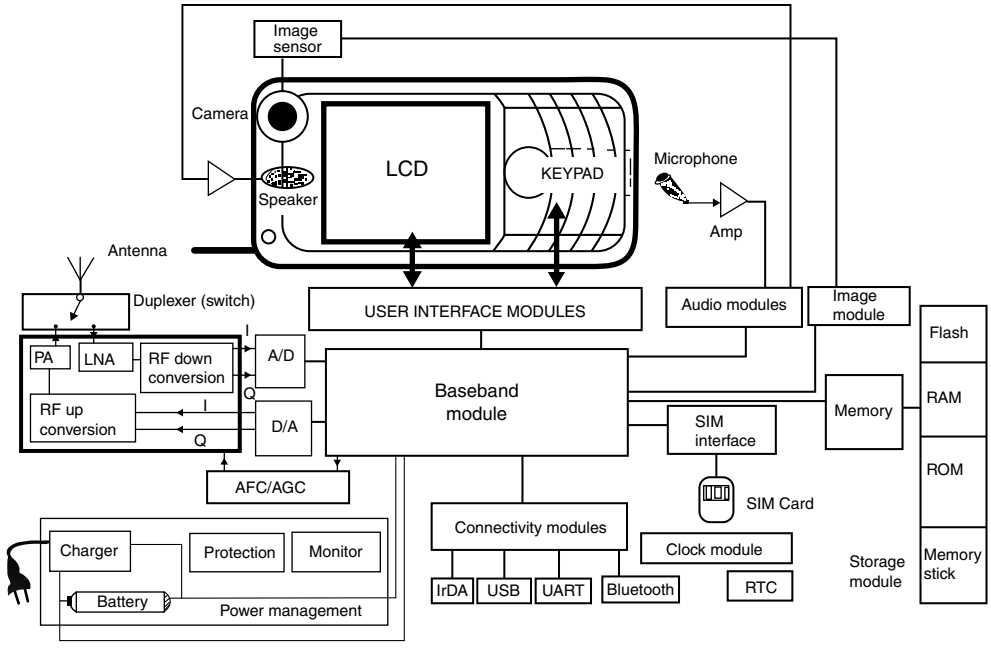


Figure 1.27 Internal block diagram of a typical mobile handset

level is not feasible, as according to Nyquist's theorem, the sampling frequency requirement is $f_s = 2 \times f_{min}$, which is too high and will generate huge volume of sampled data per second, which is difficult to process using the presently available DSP (digital signal processor). Similarly, in the transmit path the input signal is up-converted to an RF frequency and amplified, bandpassed and transmitted via the antenna.

3. **Analog to Digital and Digital to Analog Converter** – The incoming I and Q samples are sampled by an ADC unit and fed to the digital baseband block. Similarly, the transmitted baseband data are converted into I and Q analog signals by a digital modulator (or DAC) unit.
4. **Baseband Module** – This is the heart of the handset module. It controls all the devices and processes the digital information. Generally, all the physical layer modem processing (such as channel coding, interleaving, channel estimation, decoding and so on) is performed by this unit. Apart from this, it also processes the protocol for communication and interfaces. One or two processors are usually used for the baseband module implementation.
5. **AGC/AFC Module** – The automatic gain control (AGC) unit controls the gain of the mobile handset receiver and the automatic frequency control (AFC) unit controls and corrects the frequency error during operation. The baseband unit provides the information to the RF unit to do this. AFC is used to lock the transmitter clock to the desired frequency from the base station and tracks the frequency continuously, so that the phone receives a stable frequency and is synchronized with the base station clock. The AGC amplifier is used to maintain a constant output level of power at the receiver. To maintain a constant level of received signal, the AGC should be set before each received burst, this is called pre-monitoring. The baseband unit measures the received signal level and adjusts the AGC gain.

6. **Microphone** – This is a transducer, which converts the speech energy into an electrical signal.
7. **Speaker** – This is a transducer, which converts the electrical signal into sound waves (voice).
8. **Audio Modules** – The audio module controls the speaker, microphone unit, and ADC–DAC unit, which converts the electrical signal from the microphone into digital data and similarly baseband input digital data into an electrical signal, and then passes it to the speaker. It also consists of different speech and audio codec (such as MP3, AAC).
9. **Camera** – This is a sensor, which converts the light energy into electrical energy.
10. **Image Module** – This manages the camera modules and sends or receives the signal from the camera sensor and interfaces to the base--band unit.
11. **Keypad** – This is a set of buttons arranged in a block which usually bears digits and the alphabet. it is used to dial the number and make a call, accept a call or for typing SMSs and so on.
12. **Display Unit (LCD)** – This is used to display the digits or images. Generally a liquid crystal diode (LCD) is used for display. It has low energy requirements and is easy to read. LCD screens are made by sealing a liquid compound between two pieces of glass and/or a filter. The screen has many dots, and as these are changed they reflect lights (or absorb light) and display the digits. Some LCD screens have an electroluminescence panel behind them and are termed as back lit, which helps to provide background light.
13. **User Interface Module** – The user interface module controls and interfaces with the LCD and keyboard.
14. **SIM Interface Module** – This is the interface to the SIM card. It reads and writes the data to the SIM card.
15. **Memory** – There are several types of memory used in a cell phone, for example, Flash memory is used for storing the program, SRAM are used for internal memory for processing storage requirement, ROM for storing tables, fixed data, and memory sticks, and Flash card for image or data storage.
16. **Connectivity Module** – To connect the mobile with other local external devices for data transfer or any other purposes,several communication devices exist on the mobile handset, such as a serial port (UART), USB, IrDA or Bluetooth.
17. **Battery** – This is the source of energy to the handset circuits. Generally, a lithium ion (Li-ion) type of battery is used. It is lighter in weight and has a long life cycle. The other types of batteries are nickel–cadmium (NiCd), nickel–metal hydride (NiMH) and so on.
18. **Power Module** – This consist of a battery, battery charging, monitoring, and power management units.
19. **Clock Module** – It distributes the clock signal to the system.
20. **Real Time Clock (RTC)** – In a mobile phone the real time clock (RTC) is used as a base clock to provide clocks for hours, minutes, seconds and so on. It is also used for the calendar, and timer with an alarm function, for the power ON/OFF program and so on. A battery back up is provided to the RTC, so that the clock can keep running even when the phone battery is disconnected. For this a rechargeable polyacene battery, which can keep the clock running for about 30 min is used. When the main battery is disconnected for more than an hour or so, the RTC loses its content and when the main battery is connected again, the RTC clock restarts. The RTC restarts only after the 32 KHz is settled, the count keeps the baseband processor in reset mode during this time.
21. **Vibra Alert Device** – In the mobile phone a vibra alert device is used to give a silent alert signal to the user about the incoming call. This device is not placed on the phone board; rather it is placed inside the special vibra battery packs. Generally, the vibration is made using a specially designed motor and controlled with a pulse width modulation (PWM) signal via the battery terminal.

Apart from these hardware modules, there will be one or more microprocessor/controller modules to control the whole system and to run the protocol software and various applications such as voice call, data call, audio player, game, and so on. These will be discussed in more details in the next chapters.

Further Reading

Das, S.K. (2000) *Microwave Signals and Systems Engineering*, Khanna Publishers, Delhi.
Haykin, S. (2005) *Communication Systems*, John Wiley & Sons Inc., Hoboke
Proakis, J.G. and Salehi, M. (2005) *Fundamentals of Communication Systems*, Pearson Prentice Hall, Englewood Cliffs, NJ.
Tse, D. and Viswanath, P. (2005) *Fundamentals of Wireless Communication*, Cambridge University Press, Cambridge.

2

Problem Analysis in Mobile Communication System

2.1 Introduction to Wireless Channels

Apart from the transmitter and receiver circuit design issues, several other critical factors appear in the design of a mobile communication system, primarily due to the wireless channel and user mobility. A wireless channel is more complex than a traditional wired transmission channel, with lots of issues being involved in signal transmission through a wireless channel. Before discussing these issues in detail, first let us examine the basic working principles of a typical communication system. The information flow between two parties communicating is depicted in Figure 2.1. The information from the source is digitized, source coded (redundant bits are removed), passed through the protocol layers (where extra header bits are inserted), processed in the physical layer (channel coding, puncturing, interleaving, burst formation, etc.), modulated (converted into an analog signal), amplified, and sent via the channel. The signal propagates via the channel and finally reaches the receiver. The receiver receives the transmitted signal and reverses the operations to recover the source information. In this communication process, the channel plays a significant role, as its characteristics severely affect the signals that are propagated through it.

First let us try to understand how the signal transmission–reception process gradually becomes complex from a *point-to-point – wire-line* scenario to a *multi-user – wireless – mobile* scenario. In the case of a wired channel, it is typically copper wire that is used to connect the users with the local exchange. The problems associated with this are as follows. (1) *Noise* – this signal is unwanted to all receivers. It is mainly thermal noise that is the big issue here, which is generated due to temperature variations of the line and this noise gets added into the user's desired signal. The thermal noise distribution function is Gaussian in nature, so channel characteristics can be represented by a *Gaussian channel*. (2) *Attenuation* – this is because of the finite resistance of the copper wire. As discussed in Chapter 1, from *Shannon's theory* the capacity of the channel (C) can be written as $C = B \log (1 + S/N)$, where S is the signal strength and N is the noise at the receiver end. We design a receiver by defining the required link budget (the minimum S/N requirement) and set the transmitted power accordingly to meet this requirement. Defining this is easy and a one time job, as here the channel characteristics do not vary with respect to time.

The situation becomes little bit complex for the multi-user scenarios in a frequency division multiplexing system. In this scenario, the interference (the signal is wanted by at least one user and unwanted by the rest) from the other users comes into the picture. There are two types of interferences that occur in the frequency domain, and these are known as co-channel interference (CCI) and adjacent

Mobile Handset Design Sajal Kumar Das

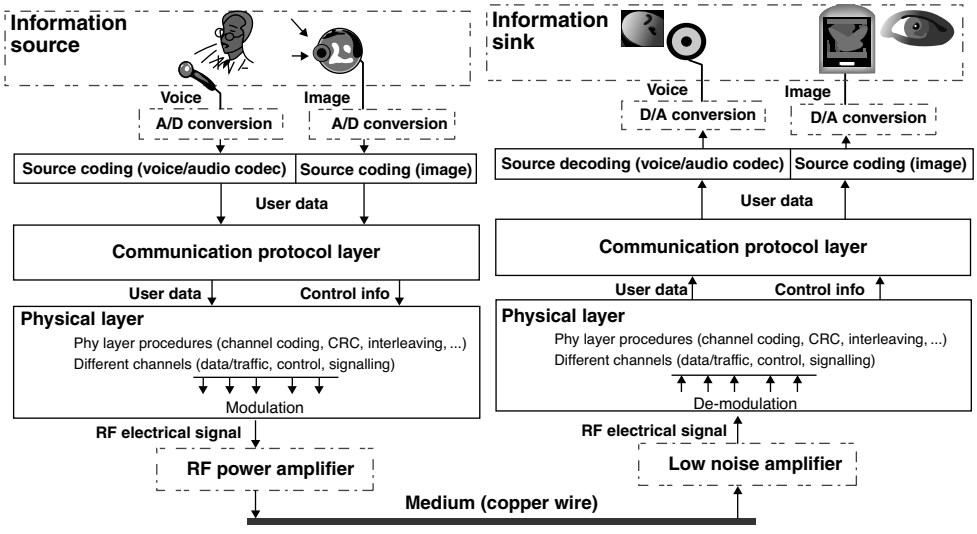


Figure 2.1 Information flow between two communicating parties

channel interference (ACI). In addition to this, in the time domain, there is inter-symbol interference (ISI), and this is generated, when the previous data signal overlaps with the next one, because of the delay spread and higher data rate, for example, a lesser symbol period.

The situation becomes more complex in the case of a long range ($r \sim 10$ km for example, cellular system) wireless scenario. Here, apart from the path-loss and attenuation, there are several other phenomenon such as the *multi-path effect* (the same signal arrives at the receiver via different paths after multiple reflections and add up if they are in phase or cancel out if they are out of phase), shadow fading, interference, environment noise, burst noise, time dispersion or delay spread, and so on, seriously degrade the quality of the signal reception.

The situation become even worse in the mobile environment, as the the position of the receiver moves with respect to time, which causes the spread of the frequency bands due to the *Doppler effect*. There are two terms normally used to characterize the channel: (a) coherent time (T_c), which is the time interval over which the channel impulse response in essentially invariant, for example, over this period the channel does not change much; and (b) coherent bandwidth (B_c), which is the BW over which the channel transfer function remains virtually constant, for example, over this BW, all frequency bands will be equally affected by channel impairment.

As a result of all the phenomena discussed above, the received signal strength in a mobile wireless environment fluctuates from maximum to minimum. This is called signal fading and it makes the wireless channel extremely unpredictable. Shannon's capacity equation can be modified for a mobile wireless channel and can be represented as $C = B\,log\,(1 + |h|^2.\,S/N)$, where, $|h|^2$ is the channel gain and its value depends on the channel fading statistics. The major differences between a wireless and wire-line channel are summarized in Table 2.1.

The quality of a wireless link between the transmitter and a receiver depends on the radio propagation parameters, mobile environment and air channel's characteristics. Different problems arise due to various reasons at the transmitter, channel, and receiver locations:

- **At Transmitter** Variation of transmitter characteristics with respect to transmitted power, modulation, non-linearity of amplification, data rate, signal bandwidth, operating frequency and so on.

Table 2.1 Differences between wire-line and wireless channels

Wireless communication	Wire-line communication
Air channel is used as the medium, which is a public channel	Generally, copper wire or fiber optic cables are used as medium and these are private channels (belonging to operator)
Signal is transmitted as electromagnetic waves	Signal is transmitted as electrical signal
The characteristic of air channel varies frequently with respect to time and frequency; the channel is unpredictable most of the time	The characteristic of wire channel does not vary much with respect to time (for example, over a long period of time, it is almost constant)
Channel loss is more	Channel loss is less
Generally receiver is mobile	Receiver is stationary
Generally the channel is characterized by Rayleigh, Rician distribution	Generally channel is characterized by AWGN
Security is a major threat	As the lines belong to specific operators, so although security is a major concern, it is less severe than wireless channel, where anyone can eavesdrop into the public air channel

- **At Channel** Variation of signal propagation in the unpredicted air channel due to attenuation, path loss, fading, multi-path, Doppler spread and so on.
- **At Receiver** Mobility of the receiver with respect to time and location, interference, noise, synchronization and so on.

All these factors are related to variability introduced by the mobile user because of signal reflection, diffraction, scattering, and a wide range of environmental variations that affect the signal propagation characteristics. These problems play a significant role in network, cell size, and receiver architecture design. In this chapter, we will first analyze various problems associated with communication over the wireless channel and then in Chapter 3, we will discuss various techniques or design solutions introduced to overcome these difficulties in wireless receivers across the generations of the wireless system.

2.2 Impact of Signal Propagation on Radio Channel

There are generally three basic phenomenon, reflection, diffraction, and scattering, which impact the signal propagation in a mobile wireless environment.

2.2.1 Reflection

Reflection occurs when a propagating electromagnetic wave impinges on a smooth surface of very large dimensions (>wavelength), as shown in Figure 2.2.

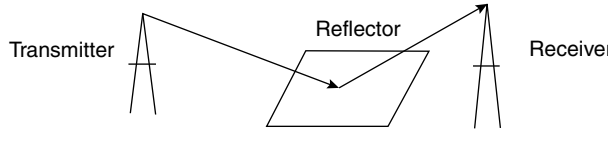


Figure 2.2 Reflection of a wave

2.2.2 Diffraction

Diffraction occurs when the radio path between the transmitter and the receiver is obstructed by a dense object with a sharp edge, causing deflection of the secondary waves in various directions away from the sharp edge, as shown in Figure 2.3. This forms secondary waves behind the obstructing body. Diffraction is a phenomenon that accounts for RF energy traveling from the transmitter to receiver without a line-of-sight path between the two. It is often termed *shadowing* because the diffracted field can reach the receiver even when shadowed by an impenetrable obstruction (Figure 2.3).

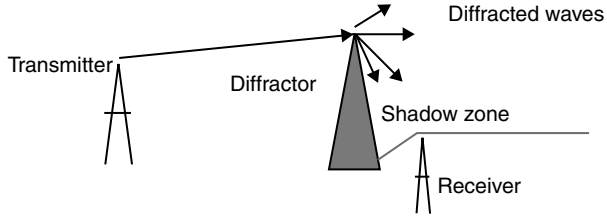


Figure 2.3 Diffraction of a wave

2.2.3 Scattering

Scattering occurs when a radio wave impinges on either a large rough surface or any surface whose dimensions are of the order of λ or less, causing the reflected energy to spread out (scatter) in all directions. This is shown in Figure 2.4. In an urban environment, typical signal obstructions that yield scattering are lampposts, street signs, and foliage (Figure 2.4).

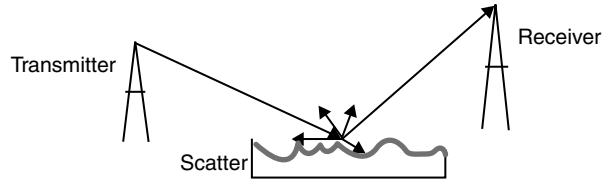


Figure 2.4 Reflection of a wave

When the surface roughness increases a little, it creates two components: a specular reflection and a scattering component. The component of specular reflection is called the coherent component, while that of scattering is called diffuse or the incoherent component. However, when the surface roughness increases, only diffuse components will remain without any specular reflection component. Such surface scattering depends on the relationship between the wavelength of the electromagnetic radiation and the surface roughness, which is defined by the Rayleigh or Fraunhofer criteria. According to the Rayleigh criterion: if $\Delta h < \lambda/8 \cos\theta$, the surface is smooth. However, in the Fraunhofer criterion: if $\Delta h < \lambda/32 \cos\theta$, then the surface is smooth, where Δh is the standard deviation of surface roughness, λ is wavelength, and θ is the angle of incidence. Generally, the scattering coefficient, which is scattering per unit area, is a function of incident angle and scattering angle. Flat surface reflection coefficient (r_s) is multiplied by a scattering loss factor:

$$r_s = \exp\left[-8\left(\frac{\sigma_h}{\Delta h_0}\right)^2\right] I_0\left[8\left(\frac{\sigma_h}{\Delta h_0}\right)^2\right] \tag{2.1}$$

where $\Delta h_0 = \lambda/(\pi \cos \theta)$, σ_h is the standard deviation of the surface height, I_0 is the modified Bessel function of first kind and zero order. The electric field, for minimum to maximum protuberance $(h) >$ critical height (h_c) can be solved for a rough surfaces using the modified reflection co-efficient: $\Gamma' = \Gamma \cdot r_s$

2.3 Signal Attenuation and Path Loss

As discussed in Chapter 1, a transmitting antenna radiates a fraction of the given amount of power into free space. Considering the transmitter at the center of a radiating sphere (Omni directional radiation), the total power, which is found by integrating the radiated power over the surface of the sphere, must be constant regardless of the sphere's radius. After propagating a distance through the wireless channel from the transmitting antenna, some part of the signal impinges on the receiver antenna. For most antenna-based wireless systems, the signal also diminishes as the receiver moves away from the transmitter, because the concentration of radiated power changes with distance from the transmitting antenna.

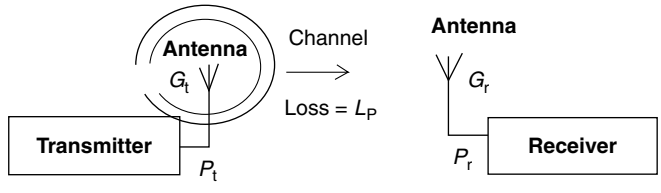


Figure 2.5 Free space signal transmission

As shown in Figure 2.5, let us consider that the power P_t is fed to the transmitting antenna and it has a gain of G_t and also assume that the antenna is transmitting equally in all directions. Then the effective transmitter power (equivalent isotropic radiated power, EIRP) will be $P_t^* G_t$. This power will be radiated over a sphere. At a distance R from the transmitter the sphere radius will be R and as the energy retention by the sphere is constant so the total energy over the sphere will always be same, for example, $P_t \cdot G_t$. The energy density over the sphere will be $P_t G_t/4\,\pi\,R^2$ and this indicates the received power at a distance R. Thus it is evident that when the receiver moves far away from the transmitter, R increases and received signal power $P_r = P_t G_t/4\,\pi\,R^2$ decreases as a proportion of R^2. Also, if A is the effective area of the receiver antenna (where this transmitted wave will impinge), then the total received power at the receiving antenna will be:

$$Pr = A \cdot (P_t\, G_t/4\pi R^2) \tag{2.2}$$

Again, let us consider G_r is the gain of the received antenna. Thus:

$$G_r = 4\,\pi A/\lambda^2 \tag{2.3}$$

$$P_r = A \cdot (P_t\, G_t/4\,\pi\,R^2) = (G_r \cdot \lambda^2/4\,\pi) \cdot (P_t\, G_t/4\,\pi\,R^2) = P_t \cdot G_t \cdot G_r \cdot (\lambda/4\,\pi\,R)^2 = P_t \cdot G_t \cdot G_r/L_p \tag{2.4}$$

$$\text{Propagation loss} - L_p = (4\pi R/\lambda)^2 \tag{2.5}$$

If other losses are also present, then we can rewrite the above equation:

$$P_r/P_t = G_t \cdot G_r/L_p \cdot L_0 \tag{2.6}$$

where L_0 is other losses expressed as a relative attenuation factor, and L_p is the free space path loss.

2.3.1 Empirical Model for Path Loss

Several empirical models exist for path loss computation, such as Okumura–Hata, COST 231 (Walfisch and Ikegami), and Awkward (graph) model. Amongst these, the Okumura–Hata model is the most widely used in radio frequency propagation for predicting the behavior of cellular transmissions in the outskirts of cities and other rural areas. The model calculates attenuation taking into account the percentage of buildings in the path, as well as any natural terrain features. This model incorporates the graphical information from the Okumura model and develops it further to suite the need better.

2.3.1.1 Okumura–Hata Model

Okumura analyzed path loss characteristics based on several experimental data collection around Tokyo, Japan. This model calculates the attenuation taking into account the parentage of buildings in the path, as well as the natural terrain features. Hata's equations are classified into three models as described below:

1. **Typical Urban**

$$L_{50} = 69.55 + 26.16 \log f_c + (44.9 - 6.55 \log h_b) \log d - 13.82 \log h_b - a(h_m)\text{dB} \tag{2.7}$$

where $a(h_m)$ is the correction factor for mobile antenna height and is given by
For large cities:

$$a(h_m) = 8.29[\log (1.54h_m)]^2 - 1.1 \qquad f_c \leq 200\,\text{MHz} \tag{2.8}$$

$$a(h_m) = 3.2\,[\log (11.75h_m)]^2 - 4.97 \qquad f_c \geq 400\,\text{MHz} \tag{2.9}$$

For small and medium cities:

$$a(h_m) = [1.1 \log (f_c) - 0.7]\, h_m - [1.56 \log (f_c) - 0.8] \tag{2.10}$$

2. **Typical Suburban**

$$L_{50} = L_{50}(\text{urban}) - 2[(\log (f_c/28)2) - 5.4]\,\text{dB} \tag{2.11}$$

3. **Rural**

$$L_{50} = L_{50}(\text{urban}) - 4.78(\log f_c)^2 + 18.33 \log f_c - 40.94\,\text{dB} \tag{2.12}$$

where, f_c is the carrier frequency, d is the distance between the base station and the mobile handset (in km), h_b is the base station antenna height, and h_m is the mobile antenna height (in m).

2.3.1.2 COST 231 Model

The COST 231 model, also called the Hata model PCS extension, is a radio propagation model that extends the Hata and Okumura models to cover a more extensive range of frequencies. This model is applicable to open, suburban, and urban areas.

The model is formulated as:

$$L = 46.3 + 33.9 \log f - 13.82 \log h_B - C_H + [44.9 - 6.55 \log h_B] \log d + C \quad (2.13)$$

where $C = 0$ dB for medium cities and suburban areas and 3 dB for metropolitan areas, $L =$ median path loss (in dB), $f =$ frequency of transmission (in MHz), $h_B =$ base station antenna height (in m), $d =$ link distance (in km), and $C_H =$ mobile station antenna height correction factor.

2.4 Link Budget Analysis

Link budget is the budget of signal energy at the receiver on a given link accounting for all the risks in the link. A link budget relates TX power, RX power, path loss, RX noise and additional losses, and merges them into a single equation. A link budget tells us the maximum allowable path loss on each link, and helps to determine which link is the limiting factor. This maximum allowable path loss will help us to determine the maximum cell size. So, instead of solving for propagation loss in the prediction equations, we can take the maximum allowable loss from the link budget and calculate the cell radius R, from the propagation model. The link budget is simply a balance sheet of all the gains and losses on a transmission path, and usually includes a number of product gains/losses and "margins." For a line of sight radio system, a link budget equation can be written as:

$$P_{RX} = P_{TX} + G_{TX} - L_{TX} - L_{FS} - L_M + G_{RX} - L_{RX} \quad (2.14)$$

where $P_{RX} =$ received power (dBm), $P_{TX} =$ transmitter output power (dBm), $G_{TX} =$ transmitter antenna gain (dBi), $L_{TX} =$ transmitter losses (coax, connectors, . . .) (dB), $L_{FS} =$ free space loss or path loss (dB), $L_M =$ miscellaneous losses (fading margin, polarization mismatch, body loss, other losses, . . .) (dB), $G_{RX} =$ receiver antenna gain (dBi), and $L_{RX} =$ receiver losses (coax, connectors, . . .) (dB).

The receiver systems exhibit a threshold effect, when the signal to noise ratio (SNR) drops below a certain value (threshold value), the system either does not work at all, or operates with unacceptable quality. For acceptable performance, the necessary condition is: SNR (at receiver) $\geq$ threshold SNR, which indicates that $P_r \geq P_{r0}$. P_r is limited by Equation 2.4 to provide satisfactory performance and P_{r0} is receiver sensitivity.

As an example the parameters are analyzed here for a particular receiver:

1. **Transmit Power** – (>30–45 dBm for base stations and approximately 0–30 dBm for mobiles) this is simply the EIRP of the transmitter.
2. **Antenna Gain** – (>18 dBi for base stations) this is a measure of the antenna's ability to increase the signal.
3. **Diversity Gain** – (>3–5 dB) by utilizing various frequencies, time, or space, the system can extract signal information from other replicas and this translates into a gain.
4. **Receiver Sensitivity** – (> −102 to −110 dBm) the lowest signal that a receiver can receive and still be able to demodulate with acceptable quality.
5. **Duplexer Loss** – (>1 dB) the loss from using a duplexer unit, which duplexes the uplink and downlink.
6. **Combiner Loss** – (>3 dB) the loss from using a combiner unit, which combines multiple frequencies onto one antenna system.
7. **Filter Loss** – (>2–3 dB) the loss occurred due to the use of filters in the circuit.
8. **Feeder Loss** – (>3 dB) the loss from the cables connecting the base station with the antenna system.
9. **Fade Margin** – (>4–10 dB) this accounts for fading dips, especially for slow moving mobiles, because for fast moving mobiles they tend to move out of a dip faster than the channel changes. Some special curves (Jake's curves) are used to compute this parameter based on a certain reliability of coverage (percentage $\gg$ 75–95%).

10. **Interference Margin** – (>1 dB) this accounts for high interference from the other users.
11. **Vehicle Penetration** – (>6 dB) accounts for the attenuation of the signal by the chassis of a car.
12. **Building Penetration** – (>5–20 dB) accounts for the penetration of building material for indoor coverage. This depends on the type of building and the desired quality at the center of the interior.
13. **User Body Loss** – (>3 dB) accounts for the signal blockage created by a mobile user's head (sometimes called head loss).

Using tools such as Planet (MSI) or Wizard (Agilent) the coverage can be analyzed and we can also estimate co-channel and adjacent channel interference. The tools can be used for automatic frequency and code planning.

2.5 Multipath Effect

As many obstacles and reflectors (tall buildings, metallic objects, water surfaces, etc.) are present in a wireless propagation channel, so the transmitted signal is reflected by such reflectors and arrives at the receiver from various directions over multiple paths, as shown in Figure 2.6. Such a phenomenon is called a multipath. It is an unpredictable set of reflections and/or direct waves, each with its own degree of attenuation and delay. Multipath is usually characterized by two types of paths:

- **Line-of-Sight (LOS)** the straight line path of the wave from the transmitter (TX) directly to the receiver (RX).
- **Non-Line-of-Sight (NLOS)** the path of a wave arriving at the receiver after reflection from various reflectors.

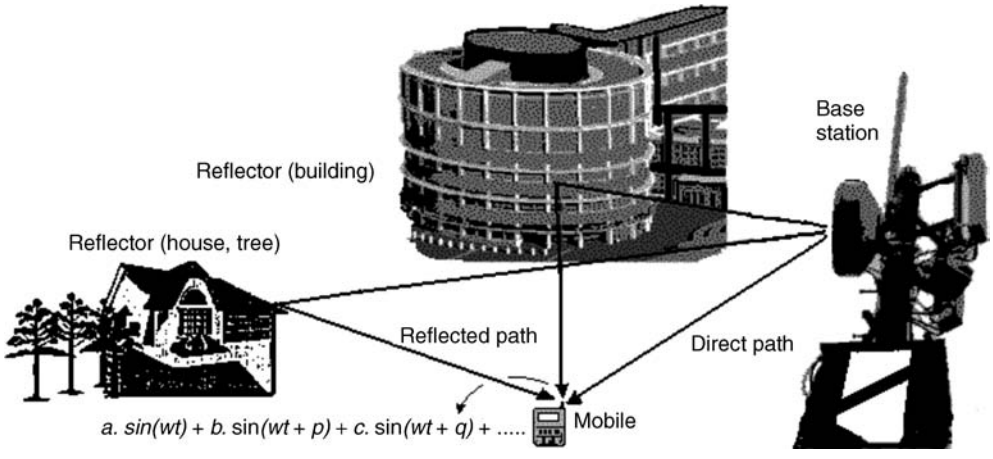


Figure 2.6 The multipath effect in a wireless channel

Multipath will cause fading (amplitude and phase fluctuations), and time delay in the received signals. When multipath signals are out of phase with the direct path signal, reduction of the signal strength at the receiver occurs; similarly when they are in phase, reinforcement of the signal strength occurs. This results in random signal level fluctuations, as the multipath reflections destructively (and constructively) superimpose on each other, which effectively cancels part of the signal energy for brief periods of time. The degree of cancellation, or fading, will depend on the delay spread of the reflected signals, as embodied

by their relative phases, and their relative power. One such type of multipath fading is known as "Rayleigh fading" or "fast fading."

2.5.1 Two Ray Ground Reflection Model

We will first consider two signals coming from the transmitter and one of them is reflected by a reflecting surface, whereas other one arrives directly at the receiver following the path d_D, as shown in Figure 2.7. The same principle can be extended to compute the resultant effect when considering multiple paths, because many such direct and reflected waves will arrive at the receiver.

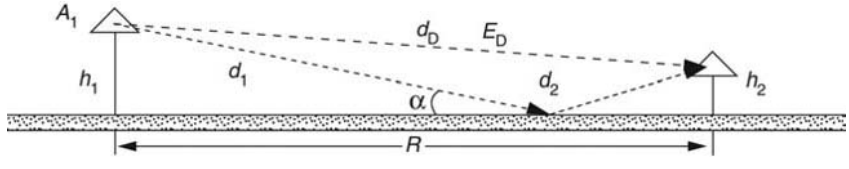


Figure 2.7 The two ray reflection model in a wireless channel

Total received field at the receiver is:

$$E\,(\text{total}) = E\,(\text{direct}) + E\,(\text{reflected}) = E_D + E_R e^{j\Delta\phi}$$

where

$$E_D = \frac{A}{d_D};\; E_R = \frac{A.\Gamma}{d_1 + d_2};\; |E_t| = \frac{A}{d_D}|1 + \Gamma\frac{d_D}{d_1 + d_2}e^{j\Delta\phi}| \tag{2.15}$$

where E_D = direct LOS component, E_R = reflected component, $\Delta\varphi$ = phase difference, and Γ = complex reflection coefficient. The reflection introduces amplitude and phase fluctuation, for example, fading. The phase difference is:

$$\Delta\phi = \frac{2\pi}{\lambda}(d_1 + d_2 - d_D) = \frac{2\pi}{\lambda}\Delta d \tag{2.16}$$

As the mobile operating frequency (for example in GSM $\sim$ 900 MHz) is very high. for example, the wavelength is smaller than the distance, so, we can assume:

$$d_1 + d_2, d_D \gg \lambda; d_1, d_2, d_D \gg h_1, h_2; \frac{d_D}{d_1 + d_2} \approx 1$$

where h_1 and h_2 are the height of the transmitter and receiver antenna. Thus the path difference $((d_1 + d_2) - d_D) = \sqrt{((h_t + h_r)^2 + R^2)} - \sqrt{((h_t - h_r)^2 + R^2)}$. When $R \gg (h_t + h_r)$, then $\sqrt{((h_t + h_r)^2 + R^2)} - \sqrt{((h_t - h_r)^2 + R^2)} = \sqrt{[R^2\{((h_t + h_r)/R)^2 + 1\}} - \sqrt{[R^2\{((h_t - h_r)/R)^2 + 1\}} = R\cdot[\sqrt{\{((h_t + h_r)/R)^2 + 1\}} - \sqrt{\{((h_t + h_r)/R)^2 + 1\}}]$.

$$= R.\left[1 + \tfrac{1}{2}\cdot((h_t + h_r)/R)^2 - 1 - \tfrac{1}{2}\cdot((h_t + h_r)/R)^2\right] = 2\cdot h_t\cdot h_r/R \tag{2.17}$$

For small α ($\alpha \ll 1$), this indicates that $\Gamma \approx -1$.

With this approximation, the total received field becomes:

$$|E_t| \approx \frac{A}{d_D}|1 - e^{j\Delta\phi}| \approx \frac{4\pi h_1\cdot h_2 A}{\lambda R^2}, R > \frac{20 h_1\cdot h_2}{\lambda}$$

It is important to observe that, as discussed earlier, in free space, as the distance R increases, the electric field decreases as $1/R$ and thus the power decreases as $1/R^2$. However here, considering the multipath effect, the electric field reduces at a rate of $1/R^2$. Hence the power reduces as $1/R^4$.

$$P_r \approx |E_t|^2 \approx \frac{1}{R^4} \tag{2.18}$$

The path loss is

$$L_P = \frac{R^4}{h_t^2 h_r^2} \tag{2.19}$$

2.6 Delay Spread

To understand this, we should consider the effect of channel impulse response. In an ideal case, a *Dirac pulse* $\delta(t)$ is defined as $\delta(t) = 1$, only at $t = 0$ and elsewhere $\delta(t) = 0$. (However, in practice, the pulses are of finite width, as shown in Figure 2.8) If we assume a very short pulse of extremely high amplitude (delta pulse) is sent by the transmitting antenna at time $t = 0$ and considering a practical situation, this pulse will arrive at the receiving antenna via direct and numerous reflected paths with different delays τ_i and with different amplitudes (because of the different path distance and path loss). Thus, various parts of the pulse will arrive at the receiver at different instances in time based on the pulse width and the channel conditions. The impulse response of the radio channel is the sum of all the received pulses. Now, because of the mobility of the receiver (the receiver and some of the reflecting objects are also moving), the channel impulse response is a function of time and the delays (τ_i). The channel impulse response can be represented as:

$$h(t, \tau) = \sum_N a_i \delta(t - \tau_i) \tag{2.20}$$

The channel may vary with respect to time, which indicates that delta pulses sent at different instances (t_i) will cause different reactions in the radio channel. The *delay spread* can be calculated from the channel impulse response. As shown in Figure 2.9, for any practical channel the inevitable filtering effect will cause a spreading (or smearing out) of individual data symbols passing through a channel.

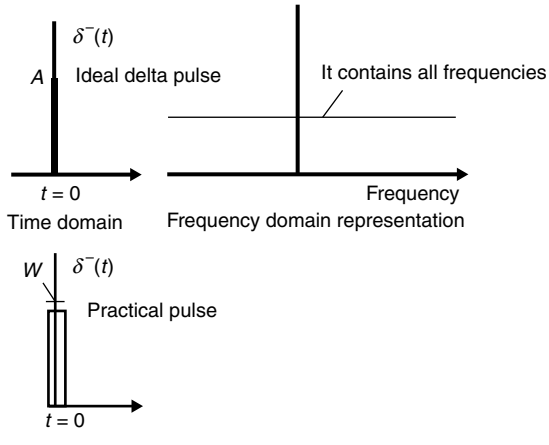


Figure 2.8 Nature of a delta pulse

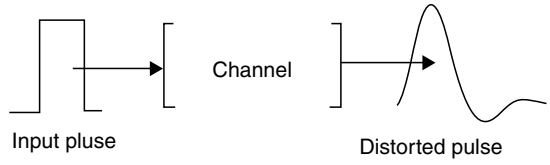


Figure 2.9 Channel delay spread

Two commonly used terms for delay spread calculations are the average and RMS delay spread and these are defined as below:

$$\text{Average delay spread } \mu_\tau = \frac{\int_0^\infty \tau\phi(\tau)\mathrm{d}\tau}{\int_0^\infty \phi(\tau)\mathrm{d}\tau} \tag{2.21}$$

$$\text{RMS delay spread } \sigma_\tau = \sqrt{\frac{\int_0^\infty (\tau-\mu_\tau)^2\phi_C(\tau)\mathrm{d}\tau}{\int_0^\infty \phi_C(\tau)\mathrm{d}\tau}} \tag{2.22}$$

If power density is discrete as shown in Figure 2.10a, then the average and RMS delay spread for a multipath profile can be written as:

$$\overline{\tau} = \frac{\sum_k \tau_k p(\tau_k)}{\sum_k p(\tau_k)};\ \sigma_\tau = \sqrt{\overline{\tau^2} - (\overline{\tau})^2} \tag{2.23}$$

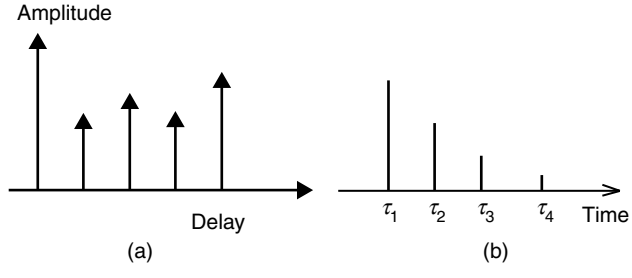


Figure 2.10 (a) Discrete power density, and (b) arrival of various multipath signals at different time scales

where

$$\overline{\tau^2} = \frac{\sum_k \tau_k^2 p(\tau_k)}{\sum_k p(\tau_k)}$$

Thus delay spread is a type of distortion that is caused when identical signals (with different amplitudes) arrive at different times at the destination as shown in Figure 2.11 as. The signal usually arrives via multiple paths and with different angles of arrival. The time difference between the arrival moment of the first multipath component and the last one is called delay spread. This leads to time dispersion, resulting in inter symbol interference. This causes significant bit rate error, especially in the case of a TDMA based system.

$$\text{Delay spread: } T_{\text{d}} = \max_i(\tau_{\text{i}}) - \min_i(\tau_i) \tag{2.24}$$

Delay spread increases with frequency. The RMS delay spread is inversely proportional to the coherence bandwidth.

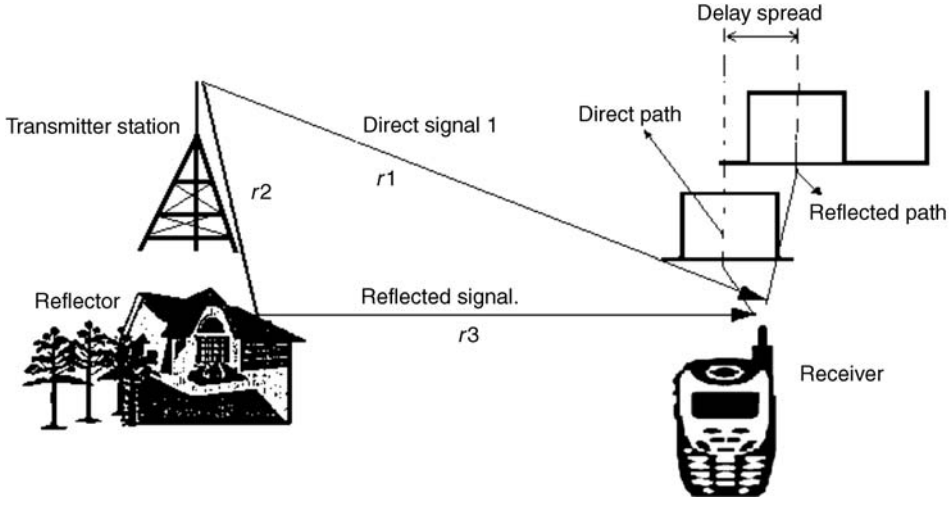


Figure 2.11 Delay spread

2.6.1 Coherent BW (Bc)

Coherence bandwidth is the bandwidth over which the channel transfer function remains virtually constant. This is a statistical measure of the range of frequencies over which the channel can be considered as "flat" (for example, a channel that passes all spectral components with approximately equal gain and linear phase). Equivalently, coherence bandwidth is the range of frequencies over which two frequency components have a strong potential for amplitude correlation. Formally the coherence bandwidth is the bandwidth for which the auto co-variance of the signal amplitudes at two extreme frequencies reduces

from 1 to 0.5. For a Rayleigh fading WSSUS channel with an exponential delay profile, we can write, $B_c = 1/(2\ \pi\ \sigma_\tau)$, where σ_τ is the RMS delay spread. This result follows on from the derivation of the correlation of the fading at two different frequencies. It is important to note that an exact relationship between coherence bandwidth and RMS delay spread does not exist. In general, spectral analysis techniques and simulation are required to determine the exact impact of a time varying multipath on a particular transmitted signal. Coherence BW characterizes the channel responses – frequency flat or frequency selective fading. The signal spectral components in the range of coherence bandwidth are affected by the channel in a similar manner.

2.7 Doppler Spread

If the mobile receiver moves away from or near to the transmitter with some velocity, then the approach velocity of EM wave changes based on its direction of movement (Figure 2.12). This leads to a change in frequency. This phenomenon is known as the Doppler effect. This means, in the case of a mobile receiver, due to the motion of the receiver and some reflecting objects in the medium, the receive frequency shifts as a result of the Doppler effect.

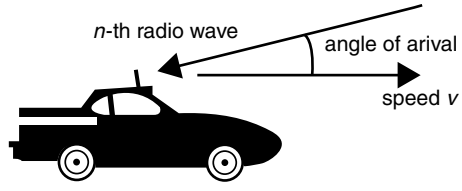


Figure 2.12 Mobile receiver and the Doppler effect

For single-path reception, this shift is calculated as follows:

$$f_d = \frac{v}{c} f_c \cos \alpha \tag{2.25}$$

where v = speed of the vehicle, c = speed of light, f_c = carrier frequency, α = angle between v and the line connecting the transmitter and receiver.

The Doppler shift = $(v \cos \alpha)/\lambda$, which is positive (resulting in an increase in frequency) when the radio waves arrive from ahead of the mobile unit, and is negative when the radio waves arrives from behind the mobile unit. The maximum Doppler shift = $v/\lambda = f_m$. Hence the above equation can be written as, $f_d = f_m \cos \alpha$. Doppler shift leads to (time varying) phase shifts of individual reflected waves. If it is for a single wave, then this minor shift does not bother radio system designers very much, as a receiver oscillator can easily compensate it. However, the fact is that many such waves arrive with different shifts. Thus, their relative phases change all the time, and so it affects the amplitude of the resulting received composite signal. Hence the Doppler effects determine the rate at which the amplitude of the resulting composite signal changes.

With multipath reception, the signals on the individual paths arrive at the receiving antenna with different Doppler shifts because of the different angles α_i, and the received spectrum is spread (spread in the frequency domain). The models behind Rayleigh or Rician fading assume that many waves arrive with their own random angle of arrival (thus with their own Doppler shift), which is uniformly distributed within $[0,2\pi]$, each independent of the other waves. This allows us to compute a probability density function of the frequency of the incoming waves. Moreover, we can obtain the Doppler spectrum of the received signal.

If the arrival angle α can be viewed as being uniformly distributed, then the Doppler frequency $f_d = f_m \cos\alpha$ is cosine distributed. Received power in $d\alpha$ around α is proportional to $d\alpha/2\pi$. Using

$$\frac{d\cos^{-1}x}{dx} = -\frac{1}{\sqrt{(1-x^2)}}$$

The Doppler power spectral density can be written as:

$$S(f_D) \propto \frac{d\theta}{2\pi df_D} = \frac{d[\cos^{-1}(f_D/f_m)]}{2\pi df_D} = C/\sqrt{1-(f_D/f_m)^2}$$

This implies that the Doppler shift causes frequency dispersion.

- Single frequency f_c broadened to a spectrum of $(f_c - f_m, f_c + f_m)$.
- Signal with bandwidth 2B center at f_c broadened to a bandwidth of approximately $2B + 2f_m$.

Doppler spread B_D is defined as the "bandwidth" of the Doppler spectrum. It is a measure of spectral broadening caused by the time varying nature of the channel. The reciprocal of the Doppler spread is called the coherent time of the channel. Coherence time ($T_c \propto 1/B_D$), is used to characterize the time varying nature of the frequency dispersion of the channel in the time domain.

The effect of fading due to Doppler spread is determined by the speed of the mobile and the signal bandwidth. Let the baseband signal bandwidth be B_S and symbol period T_S, then

- "**Slow fading" channel**: $T_S \ll T_c$ or $B_S \gg B_D$, signal bandwidth is much greater than Doppler spread, and the effects of Doppler spread are negligible.
- "**Fast fading" channel**: $T_S > T_c$ or $B_S < B_D$, channel changes rapidly in one symbol period T_S.

Of course, other Doppler spectra are possible in addition to the pure Doppler shift; for example, spectra with a Gaussian distribution using one or several maxima. A Doppler spread can be calculated from the Doppler spectrum.

2.7.1 Coherence Time (T_c)

Coherence time is the time duration over which the channel impulse response is essentially invariant (Figure 2.13). If the symbol period of the baseband signal (reciprocal of the baseband signal bandwidth) is greater than the coherence time, then the signal will distort, as the channel will change during the transmission of a symbol of the signal.

We can say that this is the time interval within which the phase of the signal is (on average) predictable. Coherence time, T_c, is calculated by dividing the coherence length by the phase velocity of light in a medium; given approximately by $T_c = \lambda^2/(c\Delta\lambda)$, where λ is the central wavelength of the source, $\Delta\lambda$ is the spectral width of the source, and c is the speed of light in a vacuum.

Coherence time characterizes the channel responses – slow or fast fading. It is affected by Doppler spread.

2.8 Fading

The fluctuation of signal strength (from maximum to minimum for example, deep) at the receiver due to the channel condition variation is known as fading. The effect of the channel on the various parameters of a signal is shown in Figure 2.14.

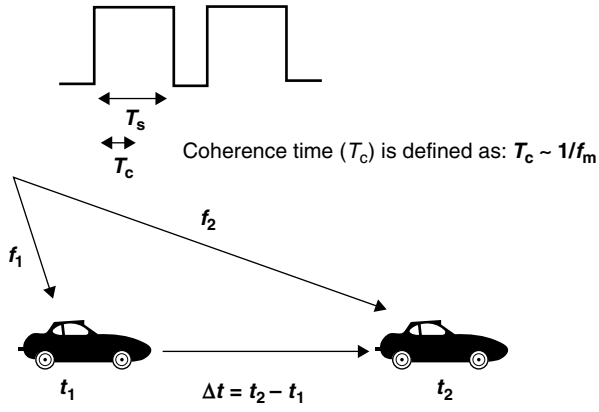


Figure 2.13 Coherence time (T_c)

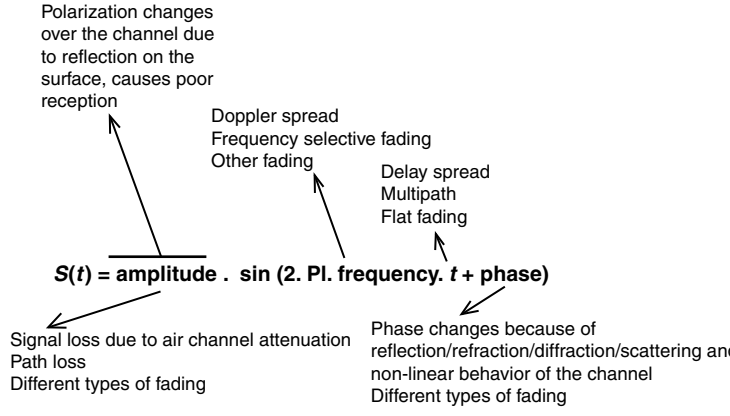


Figure 2.14 Channel effect on the various parameters of a signal

Based on the signal fluctuation over distance scale, the fading can be classified into two categories:

2.8.1 Large-Scale Fading

Large-scale fading represents the average signal power attenuation or path loss. Free space attenuation, shadowing, causes large-scale fading. This is mostly dependent on prominent terrain contours (hills, forests, billboards, clumps of buildings, etc.) between the transmitter and receiver. The receiver is often represented as being "shadowed" by such prominences. This occurs, when the mobile moves through a distance of the order of the cell size. This is also known as "large-scale path loss," "log-normal fading," or "shadowing." This is typically frequency independent.

2.8.2 Small-Scale Fading

Small-scale fading refers to fluctuations in the signal amplitude and phase that can be experienced as a result of small changes in distance between transmitter and receiver, due to constructive and destructive interference of multiple signal paths. This happens in a spatial scale of the order of the carrier wavelength. This is also known as "multipath fading," "Rayleigh fading," or simply as "fading." Multipath propagation (delay spread), speed of the mobile (Doppler spread), speed of surrounding objects, and transmission bandwidth are the main factors that influence small-scale fading. This is typically frequency dependent.

A received signal, $r(t)$, is generally described in terms of a transmitted signal $s(t)$ convolved with the impulse response of the channel $h_c(t)$. Neglecting the signal degradation due to noise, we can write $r(t) = s(t) * h_c(t)$. In the case of mobile radios, $r(t)$ can be partitioned in terms of two component random variables, as follows:

$$r(t) = m(t) \times r_0(t)$$

where $m(t)$ is known as the large-scale fading component, and $r_0(t)$ is the small-scale fading component (see Figure 2.15a). $m(t)$ is sometimes referred to as the local mean or log-normal fading because the magnitude of $m(t)$ is described by a log-normal pdf (or, equivalently, the magnitude measured in decibels has a Gaussian pdf). $r_0(t)$ is sometimes referred to as multipath or Rayleigh fading.

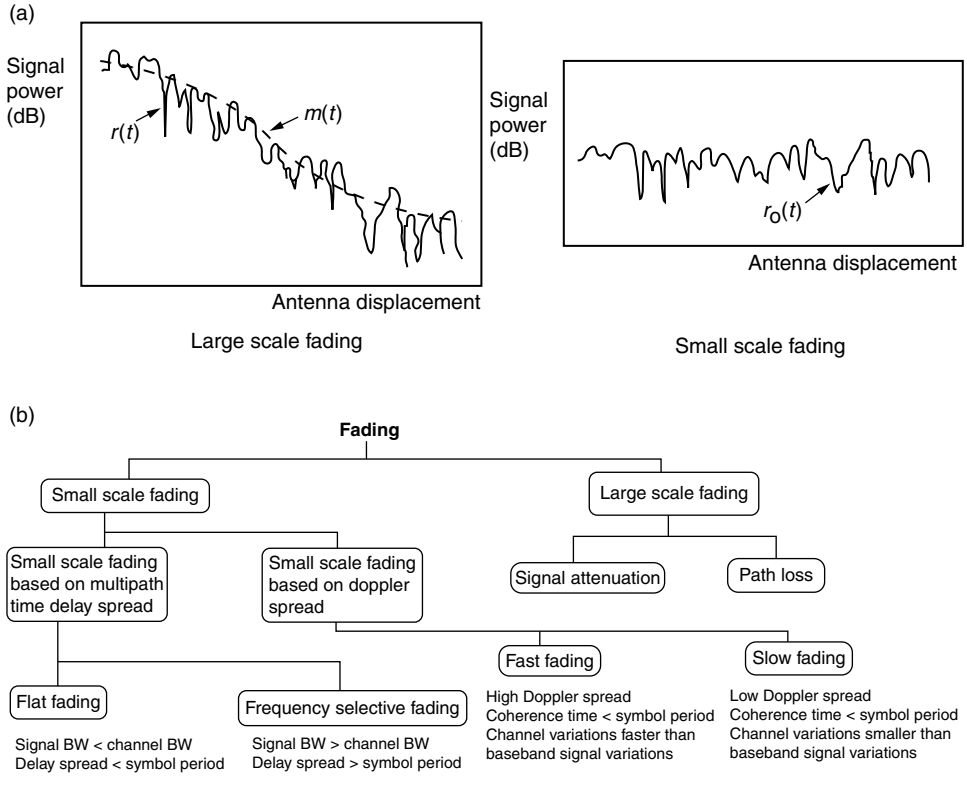


Figure 2.15 (a) Large- and small-scale fading. (b) Classification of fading

As shown in Figure 2.15b, there are four main types of small-scale fading based on the following causes:

- **Doppler Spread Causes:**
  - **Fast Fading** – High speed mobile environment that means high Doppler spread, coherence time $<$ symbol period.
  - **Slow Fading** – Low speed that means low Doppler spread, coherence time $>$ symbol period.
- **Multipath Delay Spread Causes:**
  - **Flat Fading** – BW of signal $<$ coherence BW, delay spread $<$ symbol period.
  - **Frequency Selective Fading** – BW of signal $>$ coherence BW, delay spread $>$ symbol period.

We will discuss these below in more detail.

2.8.3 Flat Fading

If the mobile radio channel has a constant gain and linear phase response over a bandwidth that is greater than the bandwidth of the transmitted signal, which means $B_s \ll B_c$ or $T_s \gg \sigma_T$, then under these conditions flat fading occurs (Figure 2.16). Flat-fading channels are also known as amplitude varying channels and are sometimes referred to as narrowband channels, as the BW of the applied signal is narrow when compared with the channel flat-fading BW.

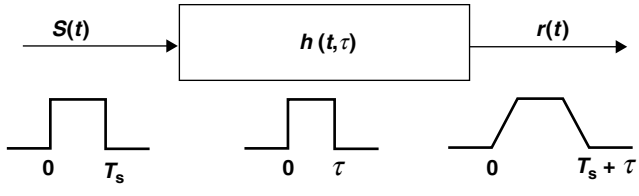


Figure 2.16 Flat fading

Flat fading occurs when (a) signal BW $\ll$ coherence BW, and (b) $T_s \gg \sigma_T$.

2.8.4 Frequency-Selective Fading

If the channel possesses a constant-gain and linear-phase response over a BW that is smaller than the BW of the transmitted signal, then the channel creates frequency-selective fading. Here, the bandwidth of the signal $s(t)$ is wider than the channel impulse response (Figure 2.17).

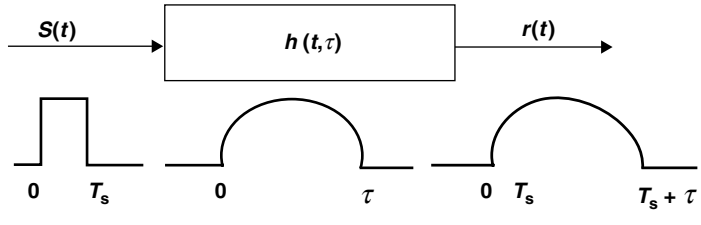


Figure 2.17 Frequency selective fading

This occurs when (a) $B_s > B_c$ and (b) $T_s < \sigma_T$. This causes distortion of the received baseband signal and inter-symbol interference (ISI).

2.8.5 Fast Fading

Fast fading occurs if the channel impulse response changes rapidly within the symbol duration. This means the coherence time of the channel is smaller than the symbol period of the transmitted signal. This causes frequency dispersion or time-selective fading due to Doppler spreading. The rate of change of the channel characteristics is larger than the Rate of change of the transmitted signal and the channel changes during a symbol period. In the frequency domain, the signal distortion due to fading increases with increasing Doppler spread relative to the bandwidth of the transmitted signal. Thus the condition for fast fading is:

$$Bs < Bc \text{ and } Ts > Tc$$

The channel also changes because of receiver motion.

2.8.6 Slow Fading

Slow fading is the result of shadowing by buildings, mountains, hills, and other objects. In a slow-fading channel, the channel impulse response changes at a rate much slower than the transmitted baseband signal $S(t)$. In the frequency domain, this implies that the Doppler spread of the channel is much less than the bandwidth of the baseband signals. Thus in this instance the channel characteristic changes slowly compared with the rate of the transmitted signal.

So, a signal undergoes slow fading if the following condition is satisfied"

$$T_s \ll T_c \text{or} B_s \gg B_D$$

where B_s = bandwidth of the signal, B_D = Doppler spread, T_s = symbol period, and T_c = coherence bandwidth.

The information is summarized in a Table 2.2, which will be very useful when designing any new wireless system.

Table 2.2 Different types of fading

Physical parameters:-
v = velocity of mobile, c = velocity of light, B_s = transmitted signal bandwidth, B_c = coherence bandwidth, T_c = coherence time, D_s = Doppler shift, T_d = delay spread

Physical parameters	Relations with other parameters
Doppler shift for a path (D_s)	$D_s = f_c \cdot v/c$
Coherence time (T_c)	$T_c \sim 1/(4 \cdot D_s)$
Coherence bandwidth (W_c)	$B_c = 1/2 \cdot T_d$
Types of fading and defining characteristics	
Fast fading	$T_c \ll$ delay requirement
Slow fading	$T_c \gg$ delay requirement
Flat fading	$B_s \ll B_c$
Frequency selective fading	$B_s \gg B_c$
Under spread channel	$T_d \ll T_c$

2.9 Signal Fading Statistics

For better receiver design, we want a statistical characterization to explain how quickly the channel changes, how much it varies, and so on. Such characterization requires a probabilistic model. Although the probabilistic models show poor performance for wireless channels, they are very useful for providing insight into wireless systems.

The fading distribution, describes how the received signal amplitude changes with time. It is a statistical characterization of the multipath fading. The received signal is a combination of multiple signals arriving from different directions, phases, and amplitudes. By received signal, we mean the baseband signal, namely the envelope of the received signal [$r(t)$]. Generally, three types of statistical distributions are used to describe fading characteristics of a mobile radio:

- Rayleigh distribution;
- Rician distribution;
- Log-normal distribution.

The rapid variations in signal power caused by a local multipath are represented by a Rayleigh distribution, whereas with an LOS propagation path, often the Rician distribution is used (Figure 2.18). Rician distributions describe the received signal envelope distribution for channels where one of the multipath components is a line-of-sight (LOS) component, that is, there is at least one LOS component present. A Rayleigh distribution describes the received signal envelope distribution for channels where all the components are non-LOS, that is, there is no LOS component and all are defused components. The long-term variations in the mean level are denoted by a log-normal distribution.

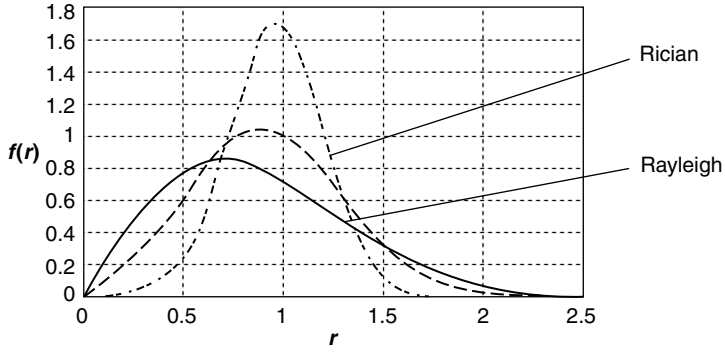


Figure 2.18 Rician density function and Rayleigh density function

2.9.1 Rician Distribution

When there is a dominant stationary (non-fading) LOS signal component present, then the small-scale fading envelope distribution is Rician. The Rician distribution degenerates to a Rayleigh distribution when the dominant component fades away.

In this case the probability distribution function (pdf) is given by:

$$p(r) = (r/\sigma^2) \cdot \mathrm{e}^{-(r2+A2)/(2\sigma 2)} \cdot I_0(Ar/\sigma^2) \cdot \text{For} A \geq 0,\ r \geq 0$$

where A is the peak amplitude of the dominant signal, and I_0 is the modified Bessel function of the first kind,and zero order, $r^2/2 =$ instantaneous power, and σ is the standard deviation of the local power.

The Rician distribution is often described in terms of parameter K, which is known as the Rician factor and expressed as $K = 10.\log\ (A^2/2\ \sigma^2)$ dB.

As $A \geq 0$ and $K \geq \alpha$ dB, because the dominant path decreases in amplitude, this generates a Rayleigh distribution.

2.9.2 Rayleigh Distribution

Typically this distribution is used to describe the statistical time varying nature of a received envelope of a flat-fading signal, or an envelope of individual multipath components.

Rayleigh distribution has the probability density function (pdf) given by:

$$p(r) = \frac{r}{\sigma^2} e^{\left(-\frac{r^2}{2\sigma^2}\right)} \text{ for } (0 \leq r \leq \infty) \text{ and for } (r < 0)$$

where σ^2 is the average power of the received signal before envelope detection, and σ is the RMS value of the received voltage signal before envelope detection.

The Rayleigh-fading signal can be described in terms of the distribution function of its received normalized power, for example, the instantaneous received power divided by the mean received power $= \Phi = r^2/2/\sigma^2$.

$$d\Phi = (r/\sigma^2) \cdot dr$$

As $p(r)\ dr$ must be equal to $p(\Phi)\ d\Phi$, hence

$$P(\Phi) = [p(r) \cdot dr]/[(r/\sigma 2) \cdot dr] = (r/\sigma 2) \cdot e - (r2/2\sigma 2)/(r/\sigma 2) = e - \Phi \cdot\ 0 = \Phi = \alpha$$

This represents a simple exponential density function, which indicates that a flat-fading signal is exponentially fading in power. The average power is

$$P_{av} \propto V_{rms}^2$$

Observations:

1. When $A/\sigma = 0$, the Rician distribution reduces to a Rayleigh distribution. As the dominant path decreases in amplitude, the Rician distribution degenerates to a Rayleigh distribution.
2. The Rician distribution is approximately Gaussian in the vicinity of r/σ, when A/σ is large.

2.9.3 Log-Normal Distribution

The log-normal distribution describes the random shadowing effects that occur over a large number of measurement locations which have the same transmitter and receiver separation, but have a different level of clutter on the propagation path. Typically, the signal $s(t)$ follows the Rayleigh distribution but its mean square value or its local mean power is log-normal.

In many cases Nakagami-m distribution can be used in tractable analysis of fading channel performance. This is more general than Rayleigh and Rician.

Based on the channel fading type and speed of the mobile receiver, different types of channels are defined and used for channel simulation. For example, static channel (no fading, zero speed), RA100 (rural area, with a speed of 100 kmph), HT100 (hilly terrain, with a speed of 100 kmph), TU50 (typical urban, with a speed of 50 kmph).

2.10 Interference

In a multi-user scenario, different users communicate using the same medium. Under these conditions, every user's signal is considered as unwanted to the rest of the users (except to the intended receiver of that user's signal). We know that any unwanted signal is treated as noise. However, here in the true sense these signals are not noise, as this signal may be a wanted/desired signal for someone, for example, every signal has one intended receiver for which it was sent and the remainder may be unintended receivers of that signal. In a wireless system, one user's desired signal can interfere with another user's desired signal and corrupt it; this phenomenon of one user's desired signal interfering with another user's desired signal is known as interference. This means that one user is interfering with the others. Interferences arise due to various reasons, and although they corrupt the signals they are sometimes unavoidable.

2.10.1 Inter-Symbol Interference

The digital data are represented as a square pulse in the time domain. Figure 2.19 represents the digital signal of amplitude A and pulse width τ. Owing to the delay spread as described earlier, the transmitted symbol will spread in the time domain. For consecutive symbols this spreading causes part of the symbol energy to overlap with the neighboring symbols, which leads to inter-symbol interference (ISI).

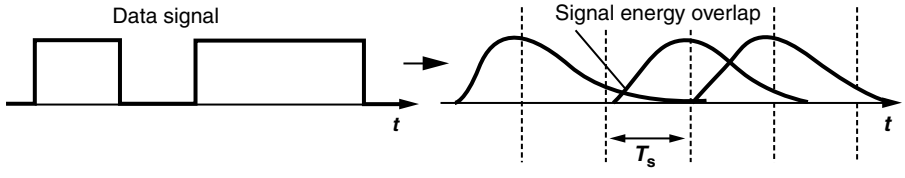


Figure 2.19 Inter-symbol interference

Ideally, a square wave in the time domain is a sync pulse in the frequency domain with infinite bandwidth. As all the practical communication systems are band-limited systems, there is always a certain amount of energy leak from the neighboring symbol and this amount is dependent on the bandwidth being used (refer to Chapter 4).

2.10.2 Co-Channel Interference

As the frequency is a very precious resource in wireless communication, in order to support many users, the same frequency is reused in some other distant cell. This reuse of frequency may cause an interference in another cell's mobile, where the same frequency is being used again. The co-channel interference occurs when two or more independent signals are transmitted simultaneously using the same frequency band.

In GSM, mobile radio operators have generally adopted a cellular network structure, allowing frequency reuse. The primary driving force behind this is the need to operate and support many users with the limited allocated spectrum. Cellular radio can be described as a honeycomb network set up over the required operating region, where frequencies and power levels are assigned in such a way that the same frequencies can be reused in cells, which are separated by some distance. This leads to co-channel interference problems. A co-channel signal that is at the same frequency band can easily bypass the RF filters and affects the energy level of the true signal. There will be more discussion on this in the Chapters 3 and 7.

2.10.3 Adjacent Channel Interference

Signals from nearby frequency channel (adjacent channel) leak into the desired channel and causes adjacent channel interference (Figure 2.20). This can be caused by bad RF filtering, modulation, and non-linearity within the electronic components. In such cases the transmitted signal is not really band limited, due to that the radiated power overlaps into the adjacent channels. To avoid this a guard band is generally inserted in between two consecutive frequency bands.

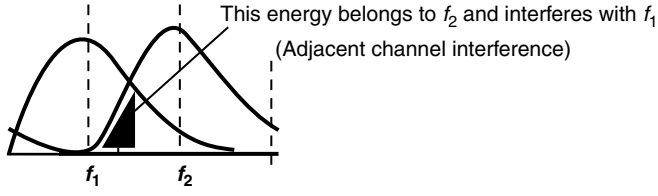


Figure 2.20 Adjacent channel interference

2.11 Noise

Any undesired signal is considered as a noise signal. The sources of noise may be internal or external (atmospheric noise, man made noise) to the system. In a receiver electrical circuit, this occurs as some electrons move in a random fashion, causing voltage/current fluctuations. In the case of a wireless channel, where the signal is not an electrical signal but rather it is an EM wave, hence the signal fluctuation, obstruction, environment RF interference, and superposition/mixing with other waves; all these disturbances (as discussed earlier) are also considered as noise. Here instead of signal to noise ratio (S/N), the signal/(interference + noise) ratio, and carrier/(interference + noise) ratio equations are used. As noise is random in nature it can only be predicted by statistical means, and usually shows a Gaussian probability density function as shown in Figure 2.21 is used for this.

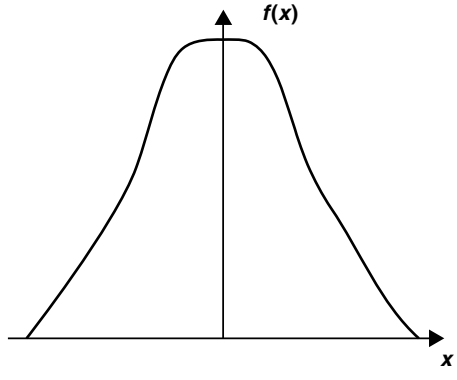


Figure 2.21 Gaussian distribution function

The random motion of electrons causes votage and current fluctuation. As, the noise is random, so the mean value will be zero. Hence, we use the mean square value, which is a measure of the dissipated power.

The effective noise power of a source is measured in root mean square (rms) values. Noise spectral density is defined as the noise content in a bandwidth of 1 Hz.

2.11.1 Noise in a Two-Port Circuit

To reduce noise added by a receiver system, the underlying causes of the noise must be evaluated. Broadband noise is generated and subsequently categorized by several mechanisms, including thermal noise and shot noise. Other causes include recombination of hole/electron pairs (G–R noise), division of emitter current between the base and the collector in transistors (partition noise), and noise associated with avalanche diodes. Noise analysis is based on the available power concepts.

Available Power This is defined as the power that a source would deliver to a conjugate matched load (maximum power transferred). Half the power is dissipated in the source and half the power is dissipated (transmitted) into the load under these conditions. As shown in Figure 2.22, a complex source is connected to a complex conjugate load. In this case the amount of power delivered can be easily computed. So if

$$Z_s = Z_L^*$$

this means that

$$Z_s = R_s + jX_s \text{ and } Z_L = R_L - jX_L$$

then

$$V_L = V_s/2 \text{ and } P_L = V_L^2/R_L = V_s^2/4R_L = V_s^2/4R_s = P_{\text{available}}$$

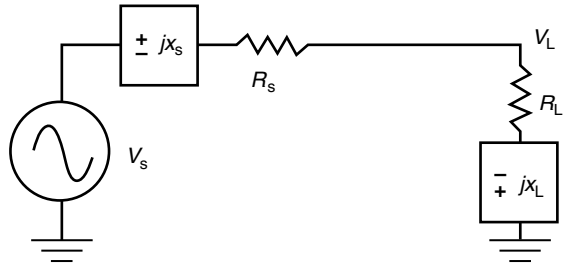


Figure 2.22 Complex source connected to complex conjugate load

2.11.1.1 Power Measurement Units

Power gain in units of dB (decibel) $= 10\log_{10}(P_r/P_t)$, the log-ratio of the power levels of the two signals. This is named after Alexander Graham Bell and can also be expressed in terms of voltages, $20\log_{10}(V_r/V_t)$, as $P = (V^2/R)$, where watts $= 10^{\text{dB mW}/10} \times 10^{-3}$.

$$\text{dBm(dB milliwatt)} - \text{relative to 1 mW, i.e. 0 dBm is 1 mW(milliwatt)}$$
$$\text{dBm} = 10\log_{10}(P \text{ in watt}/1\text{ mW})$$
$$\text{dB } \mu\text{W} = 10\log_{10}(P \text{ in watt}/1\ \mu\text{W})$$

2.11.2 Thermal Noise

When the temperature of a body increases, then the electrons inside it start flowing in a more zigzag fashion. This random motion of electrons in a conductor prevents the usual flow of current through the device and this type of unwanted signal generated as a result of temperature is known as thermal noise (Figure 2.23). This is proportional to the ambient temperature. Power is defined as the rate of energy removed or dissipated. The available power is the maximum rate at which energy is removed from a source and is expressed in joules/second (watt). The thermal noise power that is available is computed by taking the product of Boltzmann's constant, absolute ambient temperature, and the bandwidth (B) of the transmission path. Boltzmann's constant k is 1.3802×10^{-23} J/K. With an increase in the ambient temperature, the electron vibration also increases, thus causing an increase in the available noise power. Absolute temperature (T) is expressed in degrees Kelvin. The thermal noise is expressed as $P_W = kTB$.

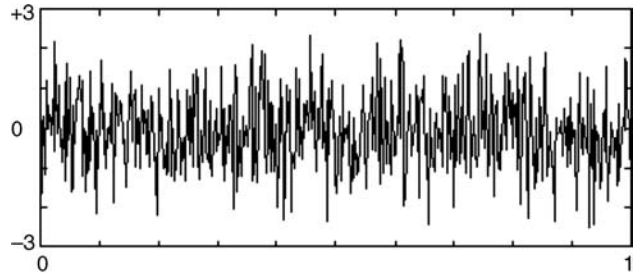


Figure 2.23 Thermal noise

2.11.3 White Noise

Noise where the power spectrum is constant with respect to frequency is called the white noise (Figure 2.24). Thermal noise can be modeled as white noise. An infinite-bandwidth white noise signal is purely a theoretical construction, as by having power at all frequencies, the total power of such a signal is infinite and therefore impossible to generate. In practice, however, a signal can be "white" with a flat spectrum over a defined frequency band ($\pm 10^{14}$ Hz).

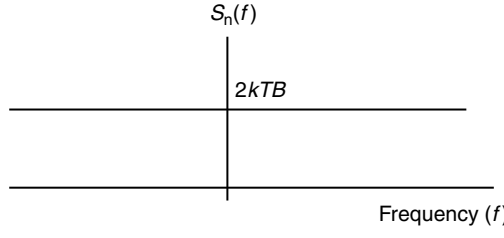


Figure 2.24 Power spectrum of white noise

2.11.4 Flicker Noise

Flicker noise is associated with crystal surface defects in semiconductors and is also found in vacuum tubes. The noise power is proportional to the bias current and unlike thermal and shot noise, flicker noise

decreases with frequency, for example, when the frequency is less, the flicker noise is more (Figure 2.25). An exact mathematical model does not exist for flicker noise, because it is so device-specific. However, the inverse proportionality with frequency is almost exactly $1/f$ for low frequencies, whereas for frequencies above a few kilohertz, the noise power is weak but essentially flat. Flicker noise is essentially random, but because its frequency spectrum is not flat, it is not a white noise. It is often referred to as pink noise because most of the power is concentrated at the lower end of the frequency spectrum. Because of this, metal film resistors are a better choice for low-frequency, low-noise applications, as carbon resistors are prone to flicker noise.

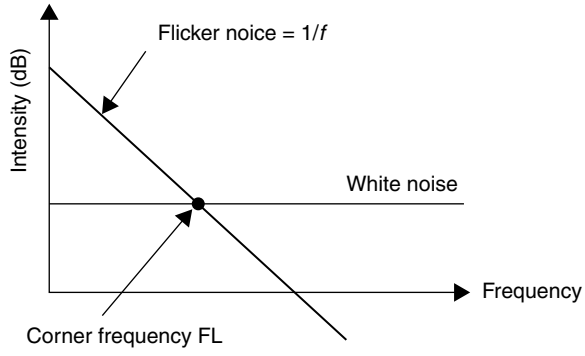


Figure 2.25 Flicker noise defined by corner frequency

2.11.5 Phase Noise

Phase noise describes the short-term random frequency fluctuations of an oscillator. Phase noise is the frequency domain representation of rapid, short-term, random fluctuations in the phase of a wave (in RF engineer's terms), caused by time domain instabilities ("jitter" in a digital engineer's terms). An ideal oscillator would generate pure sine waves but all real oscillators have phase modulated noise components in them. The phase noise components spread the power of a signal to adjacent frequencies, resulting in sidebands. Typically this is specified in terms of dBc/Hz (amplitude referenced to a 1-Hz bandwidth relative to the carrier) at a given offset frequency from the carrier frequency.

Phase noise is injected into the mixer LO (local oscillator) port by the oscillator signal as shown in Figure 2.26. If perfect sinusoidal signals are input to the RF port, the LO signal and its phase noise mixes with the input RF signals and produces IF (intermediate frequency) signals containing phase noise. If a small desired signal f_d and a large undesired signal f_u are input to the RF port, the phase noise on the larger conversion signal may mask the smaller desired signal and this would hinder reception of the desired signal. Thus, low phase noise is crucial for oscillators in receiver systems. During the detection of digitally modulated signals, the phase noise also adds to the RMS phase error.

2.11.6 Burst Noise

Burst noise or popcorn noise is another low frequency noise that seems to be associated with heavy metal ion contamination. Measurements show a sudden shift in the bias current level that lasts for a short duration before suddenly returning to the initial state. Such a randomly occurring discrete level burst would have a popping sound if amplified in an audio system. Like flicker noise, popcorn noise is very

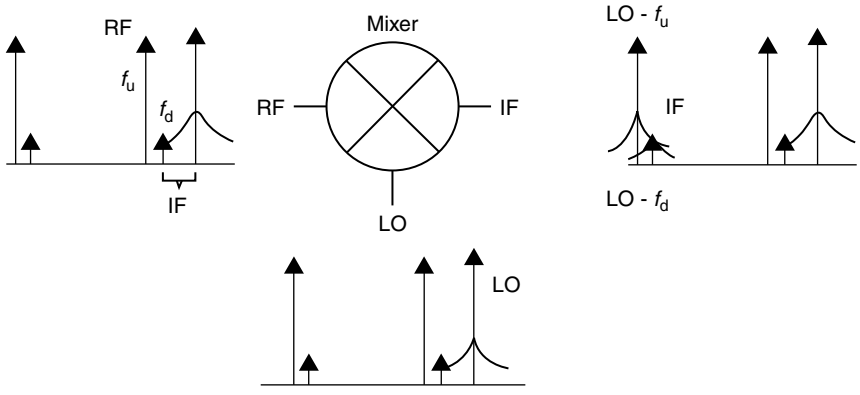


Figure 2.26 Frequency conversion limitation due to phase noise

device specific, so a mathematical model is not very useful. However, this noise increases with bias current level and is inversely proportional to the square of the frequency $1/f^2$.

2.11.7 Shot Noise

This noise is generated by current flowing across a P–N junction and is a function of the bias current and electron charge. The impulse of charge q depicted as a single shot event in the time domain can be Fourier transformed into frequency domain as a wideband noise.

$$\overline{i^2(f)} = 2eI_0BA^2/\text{Hz}$$

where I_0 is the dc current and $e =$ electron charge $= 1.6 \times 10^{-19}$ coulomb.

The power produced by shot noise is directly proportional to the bias current. Like thermal noise, shot noise is purely random and its power spectrum is flat with respect to frequency (Figure 2.27). Measurements confirm that the mean-square value of shot noise is given by

$$I_n^2 = 2q\,I_{dc}B;\; I_n = \sqrt{(2qI_{dc}B)}$$

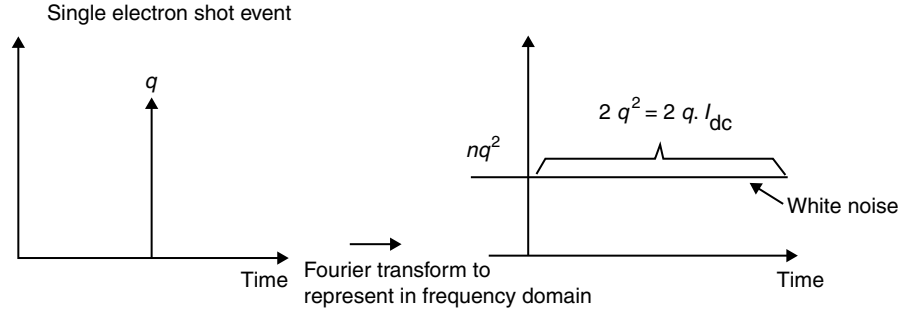


Figure 2.27 Shot noise

where $I_n =$ rms average noise current in amperes, $q = 1.6 \times 10^{-19}$ coulombs (C), the average/electron, $I_{dc} =$ dc bias current in the device in amperes, $B =$ bandwidth in which measurement t is made, in Hz.

2.11.8 Avalanche Noise

Avalanche noise occurs in zener diodes, because in reverse biased P–N junctions the breakdown happens after a certain voltage. This noise is considerably larger than shot noise, so if zeners have to be used as part of a bias circuit then they need to be RF decoupled.

2.11.9 Noise Figure (NF)

This is the ratio of the signal to noise at the input to the signal to noise at the output of a device. This indicates how much extra noise has been added to the signal by the device. The noise figure is represented as:

$$F = \frac{SNR(\text{at input port})}{SNR(\text{at output port})} = (S/N)_{\text{in}}/(S/N)_{\text{out}}$$

Noise figure (NF) is a measure of the degradation of the signal to noise ratio (SNR), caused by the components in the device circuit. Noise figure is always greater than 1, and the lower the noise figure the better the device. When T is the room temperature represented by T_o (290 K), and noise temperature is T_e, then the factor $(1 + T_e/T_o)$ is called the noise figure of an amplifier.

The noise figure is the decibel equivalent of noise factor: $F = 10^{NF/10}$, where $NF = 10 \log (F)$. If several devices are cascaded, the total noise factor can be found using the Friis formula:

$$F = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + \frac{F_4 - 1}{G_1 G_2 G_3} + \cdots + \frac{F_n - 1}{G_1 G_2 G_3 \cdots G_{n-1}}$$

where F_n is the noise factor for the nth device and G_n is the power gain (numerical, not in dB) of the nth device. This indicates that the NF of the first block should be as minimal as possible to keep the noise under control. This is why one low noise amplifier is placed in the receiver circuit at the first stage to boost the signal strength without increasing the overall noise level.

Further Reading

Goldsmith, A. (2005) *Wireless Communications*, Cambridge University Press, Cambridge.

Rappaport, T.S. (1996) *Wireless Communications: Principles and Practice*, Prentice Hall, Englewood, NJ, ISBN 9780133755367.

Tse, D. and Viswanath, P. (2005) *Fundamentals of Wireless Communication*, Cambridge University Press, Cambridge.

3

Design Solutions Analysis for Mobile Handsets

3.1 Introduction

In the previous chapter, we discussed about various problems (multi-path, scattering, Doppler spread, etc.) associated with wireless fading channels. These effects have a strong negative impact on the bit error rate of any modulation technique. The main reason why detection in a fading channel shows poor performance compared with an AWGN channel is because of the randomness of the channel gain due to the deep fade event. Thus, in order to combat these effects, we have to establish the corresponding radio receiver design solutions. In this chapter, we will mainly focus on baseband design solutions and in the next chapter we will consider the RF design solutions.

To improve the received signal quality and to reduce the bit error rate it is mainly diversity, equalization, and channel coding techniques that will be used. These are discussed below.

3.2 Diversity

We learnt about these following points in Chapter 2. (1) The channel conditions (channel transfer function or channel gain) vary with respect to time, this variation may be very fast or slow (depending on the coherent time of the channel). (2) The channel reacts (showing different attenuation, delay, etc.) differently to signals with various frequencies passing through the channel, and signals coming via a variety of paths may experience different types of attenuation due to temperature, moisture content, reflector, and obstacle variation. (3) The signal strength degrades as the received mobile signal is obstructed by a large building or hills, and so on. (4) The channel parameters vary when the mobile moves at speed.

From these above mentioned points, we can explore some solutions, for example:

1. As the channel characteristics change with respect to time, we can send the same information at different time instances, for example, repeat the same information again and again, so that at one instance of time the channel may be in bad condition (deep fade), but at another it may be in good condition. This indicates that we may receive the information properly during the second instance. Thus we can think of sending the same information over a number time instances, which leads to time diversity.

Mobile Handset Design Sajal Kumar Das

2. As the channel behaves differently to different frequencies (frequency selective fading channel), we can send the same information via signals of different frequencies. This means, if one frequency is in deep fade then another one will help to carry the information correctly to the receiver. Thus we can think of sending the same information over various frequencies, which leads to frequency diversity.
3. Because of the presence of a variety of obstacles in different places, the signal strength varies over with the range of paths the signals travel. Thus, placing numerous antennas at various locations may help a better signal to be received in one place when the signal conditions at the other places degrade. Therefore, we can consider placing several receive antennas at different places (separated by a distance of at least one wavelength), which leads to space diversity.

These solutions to improve the performance of the receiver over a fading channel are called diversity techniques. The basic idea is to send signals that carry the same information via a range of means and multiple independently faded replicas of the data symbols are obtained at the receiver end, from which more reliable detection can be obtained by choosing the correct one or by combining them all. The cost for this diversity reception is additional receiver complexity, because of the path tracking and the additional signal processing overheads.

There are many ways to obtain diversity, such as: (1) time diversity, (2) frequency diversity, (3) space diversity, and (4) polarization diversity.

The design decision when choosing any diversity technique depends on various factors, such as channel coherent bandwidth, coherent time, and the presence of a number of independent signal paths at any one instant.

3.2.1 Time Diversity

Diversity over time can be achieved if the same information is repeated over different instances in time. As with a fading channel, the channel condition varies over time, so if the same information is sent many times then there is a higher probability that in at least one instance of time the channel will be out of deep fade. So, to receive multiple repetitions of the signal with independent fading conditions, time diversity repeatedly transmits information with time spacing that exceeds the coherence time of the channel. Generally, the diversity over time is obtained by using repetition coding, interleaving, and some other techniques, as explained below.

3.2.1.1 Exploitation of Natural Phenomena for Obtaining Time Diversity

Because of multipath effects, the same signal information comes via a range of paths and as they travel different distances with the same speed (the speed of light), they arrive at the receiver antenna at different times. We can consider this phenomenon to be as if the transmitter is sending the same information but with various time delays. This effect can be treated as time diversity. Thus there is no need to resend by the transmitter, rather, nature will resend it for us. This is good!!

Now we have to exploit this effect to improve the signal quality at the receiver. We know that the λ path difference of a wave is equivalent to a 2π phase difference (see Figure 1.3). If the signals are traveling at the velocity of light c ($=3 \times 10^{10}$ cm/s) then at time t, it will travel a distance of $c.t$. Hence a path difference of $c.t$ will be equivalent to the phase difference $(2\pi/\lambda)\cdot c\cdot t = \theta$. Now, for the various paths the time "t" will be different, so "θ" will also be different. As shown in Figure 3.1, if we receive several such multipath signals (for example, $a\cdot\sin(\omega t + \theta_1)$, $b\cdot\sin(\omega t + \theta_2)$, $c\cdot\sin(\omega t + \theta_3)$, ...), and as these are not phase aligned, we can not simply add these (because if we add, we can not see the combined effect as they are not phase aligned). Strictly speaking, the phase is the condition of the wave at a specific time, and phase and time are

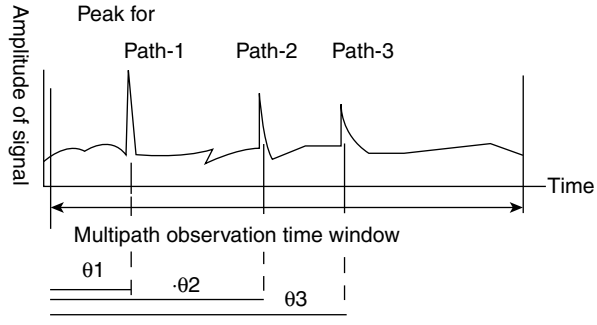


Figure 3.1 Multipath signal peaks at receiver

interrelated. All these multipath signals are the same in nature as they originate from the same signal but some might have advanced in phase, as they have arrived early compared with others. So first we need to align them to the same phase. Now, if we have several multipath searchers and each of them is tuned to one multipath signal (look for the peak in the signal amplitude for that path) and track them individually, then the peak for all of these has to be detected over the period of a time window. The peaks for the various multipaths will be placed at different positions across the time window.

These are then rotated by the same angle as they are delayed (for example, $\theta 1, \theta 2, \theta 3, \ldots$) respectively, which will align them in the same phase [$a{\cdot}\sin(\omega t)$, $b{\cdot}\sin(\omega t)$, $c{\cdot}\sin(\omega t)$, ...]. Now, they can be added together (as their angles are the same) and this will produce a large signal strength. This process is called multipath combined. This could very well result in higher SNR in a multipath environment than in a "clean" environment. This works very well when the time gaps between the peaks of the multipaths are well separated and numerous multipath peaks are present over a reasonable width of the time window.

This special type of receiver architecture, where several receivers are working in parallel to receive signals from multiple paths, is called a Rake receiver. The Rake receiver is so named because of the analogy with the function of a garden rake, with each branch collecting bit or symbol energy, which is similar to how the tines on a rake collects leaves. Rake receivers are common in a wide variety of radio devices, including mobile phones (CDMA based) and wireless LAN equipment.

3.2.1.2 Rake Receiver

The basic idea of a Rake receiver was first proposed by Price and Green and patented in 1956. As discussed, the Rake receiver uses a multipath diversity principle – it rakes the energy from the multipath propagated signal components and combines it.

M-ray multi-path models can be used for this purpose as shown in Figure 3.2. Each of the M paths has an independent delay τ, and an independent complex time variant gain G. Here $r(t)$ is the transmitted signal received at the receiver front-end and $r_p(t)$ is received signal after processing.

As shown in the Figure 3.3, a Rake receiver utilizes multiple correlators to separately detect M strongest multipath components, which experienced different delays during their travel over the channel. Each correlator detects a time-shifted version of the original transmission, and each finger correlates with a portion of the signal, which is delayed by at least one chip in time from the other fingers. The outputs of each correlator are weighted to provide a better estimate of the transmitted signal than is provided by a single component. Outputs of the M correlators are denoted as Z_1, Z_2, ..., and Z_M and the outputs are

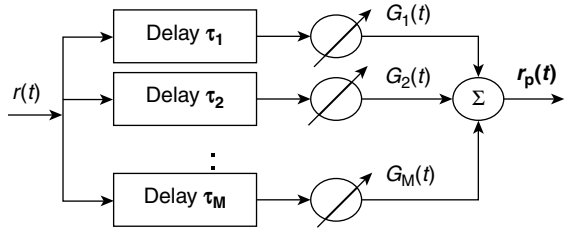


Figure 3.2 Rake receiver basic elements

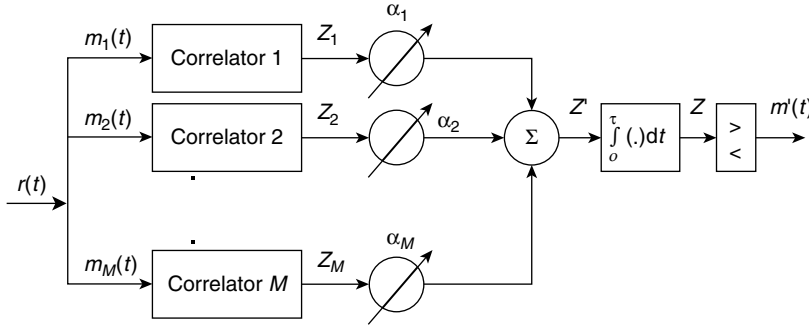


Figure 3.3 Rake with correlated and weighted sum decision

weighted by $\alpha_1, \alpha_1, \alpha_2, \ldots, \alpha_M$, respectively. Once these are correlated, weighted and summed up, then this is passed for demodulation and bit decisions.

The weighting coefficients are based on the power or the signal to noise ratio from each correlator output. If the power or SNR from a correlator is small, then a small weighting factor α will be assigned accordingly. Now if maximal-ratio combining is used then the following equation can be written for the summed output:

$$Z' = \sum_{m=1}^{M} \alpha_m Z_m$$

The weighting coefficients, α_M, are normalized to the output signal power of the correlator

$$\alpha_m = \frac{Z_m^2}{\sum_{m=1}^{M} Z_m^2}$$

Choosing weighting coefficients based on the actual outputs of the correlator leads to better Rake receiving performance.

Rake Receiver Architecture in a Digital Receiver (Using I-Q Modulation)

The typical architecture of a Rake module is shown in the Figure 3.4. The tasks of each module are:

1. **Matched Filter** – Impulse response measurement, assigning the largest peaks to Rake fingers, timing to delay equalizer, tracks and monitors peaks with a measurement rate depending on speeds of mobile station and on propagation environment.

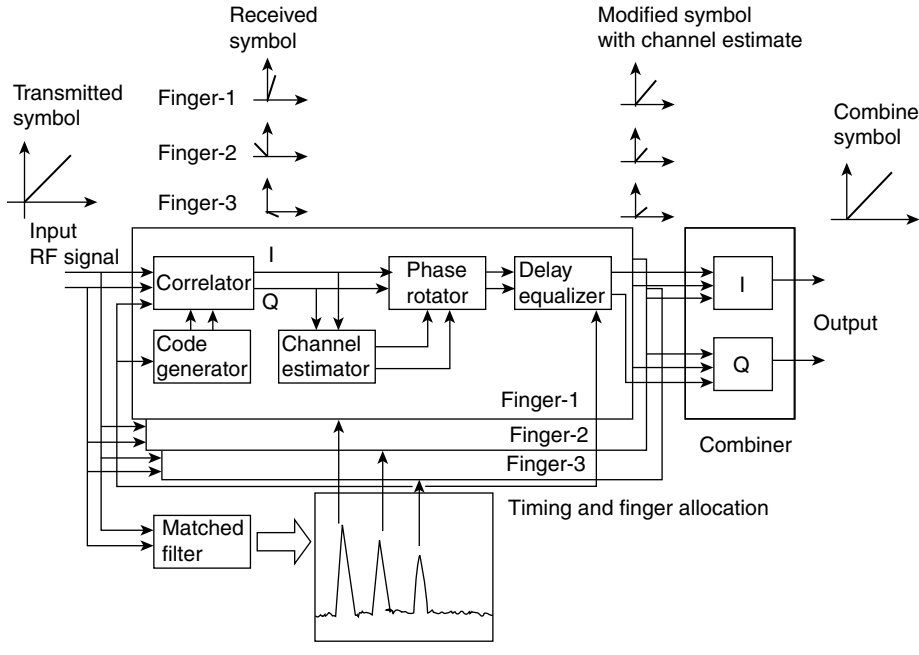


Figure 3.4 Rake internal blocks

2. **Code Generators** – Generates PN codes for the user or channel.
3. **Correlator** – Despreading and integration of user data symbols.
4. **Channel Estimator** – Estimates the channel state, channel effect corrections.
5. **Phase Rotator** – Phase correction.
6. **Delay Equalizer** – Compensates delay for the difference in the arrival times of the symbols in each finger.
7. **Combiner** – Adding of the channel compensated symbols, multipath diversity against fading.

Applications of Rake Receiver

Apart from in a CDMA/WCDMA receiver, the Rake receiver can also be used in a non-CDMA based receiver. Basically, a Rake receiver "rakes" in the symbol energies from various multipath delays, and combines them using a maximal-ratio combining technique. An adaptive FIR equalizer will do the same job, basically equalizing the impulse response of the channel; this is equivalent to a matched filter that is matched to the different channel delays. There is nothing about a Rake receiver that makes it applicable only for CDMA type signals. Things that need to be considered are as follows. (1) The total length of the time window to collect the multipaths: if the time window length is more then the possibility of providing many multipaths will be more, but this leads to more complexity of the receiver and if the time window is shorter then we may not get enough multipaths in that time window. Based on cell size and reflector objects in the cell this may vary. (2) Chip rate or symbol rate: if the chip rate is too low (for example, wider time gap) and the cell size is small (smaller time to travel from transmitter to receiver over any path), then in this case multipath signals will be separated by a very small time gap and the chip/symbol from different multipaths may overlap with each other at the receiver. Thus, resolving multipaths will be hard.

3.2.2 Frequency Diversity

The likelihood is that the signals at different frequencies will not suffer the same level of attenuation. Therefore different frequencies are used to transmit the same information and at the receiver the energies from these are combined or the strongest one is used for decoding. This technique of using many frequencies to transmit the same information through the channel to the receiver is known as frequency diversity.

3.2.3 Space Diversity

This is a method of transmission or reception, or both, in which the effects of fading are minimized by the simultaneous use of two or more physically separated antennas, ideally separated by one or more wavelengths. Space diversity is also known as antenna diversity.

3.3 Channel Estimation and Equalization

The transmitted signal experiences various problems when it travels through a wireless medium. On the receiver side, we first need to know how the channel behaved to any transmitted burst/packet during that time. This process of evaluating or estimating the channel impulse response during transmission of the burst is known as channel estimation. Now, once the channel characteristics are known using the channel estimation process, we will try to compensate for the effect by using an equalization process as shown in Figure 3.5. Thus, channel estimation will help us to know how the different parts of the transmitted signal are affected by the channel and then knowing that the receiver will do the reverse effect to compensate for this channel effect.

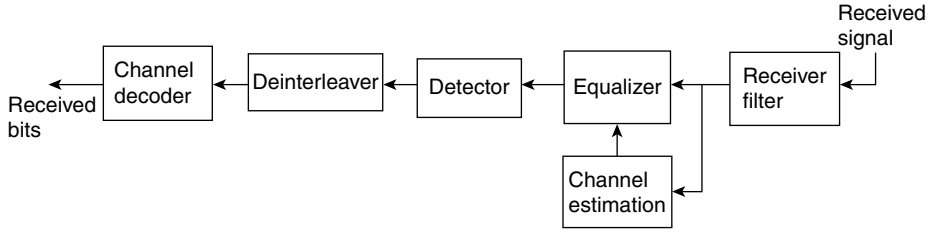


Figure 3.5 Basic channel estimation blocks in a receiver

3.3.1 Study of Channel Characteristics – Channel Estimation

In general a channel can be studied using various techniques, and based on the channel knowledge at the transmitter and receiver, we can broadly categorize these as below:

- **Receiver knows the channel characteristics** – Here the transmitter sends a training sequence or a pilot channel or pilot bits to study the channel periodically. A training sequence or pilot bits are bits that are known to the receiver. Generally these are sent along with the data bits. The receiver receives the data bits along with the training sequence bits and finds out how the known bits (training sequence bits) have been affected by the channel by using correlation techniques. Knowing this, it tries to equalize the data

bits by assuming that data bits have also been affected in a similar manner during that time by the channel. Some systems use a separate channel (pilot channel) for this purpose.

- **Transmitter and receiver know the channel characteristics** – Transmitter sends a training sequence or a pilot channel or pilot bits to study the channel and then the receiver feeds back the measurement reports to the transmitter intimating the channel conditions.
- **Transmitter does not have any knowledge of channel realization except for the statistical characterization (Rayleigh, Rician, etc.)** – Here the transmitter just statistically characterizes the channel and based on this it makes the transmission decision.

As shown in the Figure 3.6, when we receive the signal there are various energy components that contribute towards the total received energy over that symbol period, such as energy due to inter symbol interference, co-channel interference, adjacent channel interference, and noise. These energies are not the transmitted signal's energy; they are the energy from some other signal, which has leaked into this symbol period and is increasing or decreasing the total energy of the desired signal. These need to be estimated and should be removed or suppressed before passing to the decoder. Similarly, if the same signal's energy is residing in some other path or with some delay path then we need to tap these energies

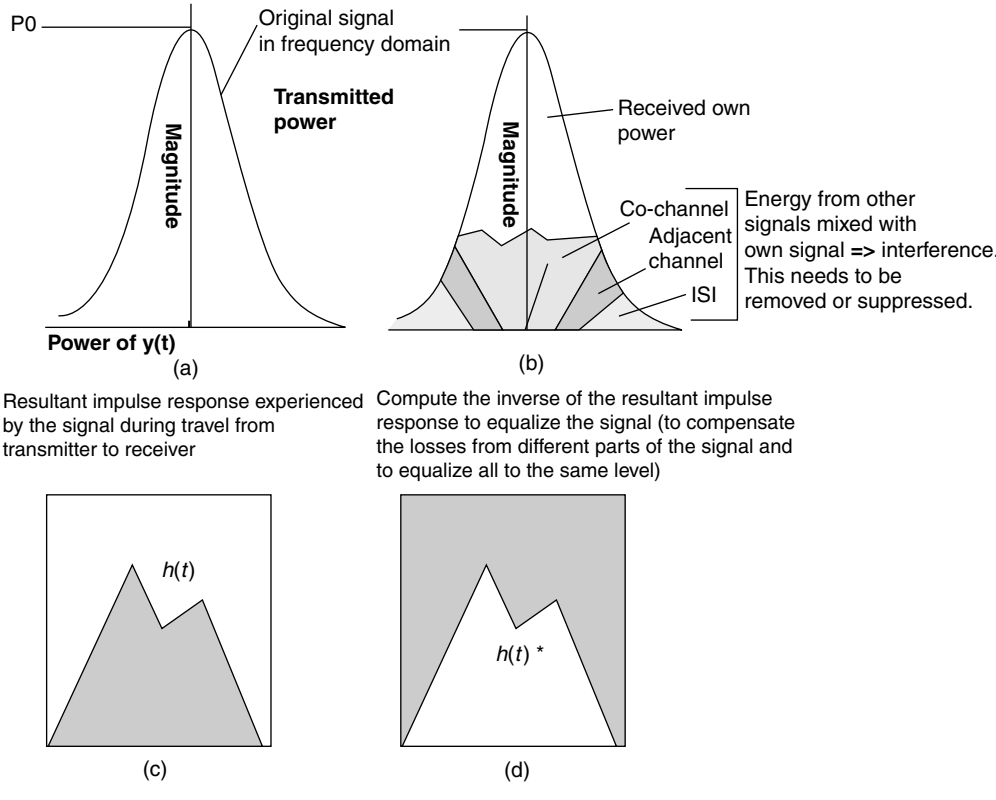


Figure 3.6 (a) Transmitted signal, (b) received signal with interference signal, (c) transfer function of channel, and (d) inverse transfer function at the receiver

and combine them to increase their own energy, which is known as combining, as when using different diversity techniques.

To remove ISI from the signal, many types of equalizers can be used. Detection algorithms based on trellis search (for example, MLSE or MAP) offer a good receiver performance, and often without too much computation. Therefore, these algorithms are currently fairly popular. However, these detectors require knowledge of the channel impulse response (CIR), which can be provided by a separate channel estimator. Usually the channel estimation is based on the known sequence of bits, which is unique for a certain transmitter and is repeated in every transmission burst. Thus, the channel estimator is able to estimate CIR for each burst separately by exploiting the known transmitted bits and the corresponding received samples. Usually CIR is estimated based on the known training sequence, which is transmitted in every transmission burst.

A channel estimator for a single signal will now be described, which is basically a least-squares (LS) channel estimation technique. Consider first a communication system that is only corrupted by noise as depicted in Figure 3.7. Digital signal "x" is transmitted over a fading multipath channel h_L, where h_L is the channel transfer functions of length L, for example, we are assuming that the energy of the digital signal x is spread over L number of delay paths as shown in Figure 7.12. We need to estimate these tap gains $(h_1, \ldots, h_L)$ for signals delayed by different amounts (as if for different multipaths we are trying to estimate the gain factor and combining them together as these energies over different paths belong to the same symbol, for example, the concept is similar to previously discussed Rake receiver). Thermal noise is generated at the receiver and it is modeled by additive white Gaussian noise n, which is sampled at the symbol rate. The demodulation problem here is to detect the transmitted bits x from the received signal y. Simply, we can write the received signal as $y = x^*h + n$, then we need to compute h from this equation and provide the same to the equalizer, which is the main task of the channel estimator.

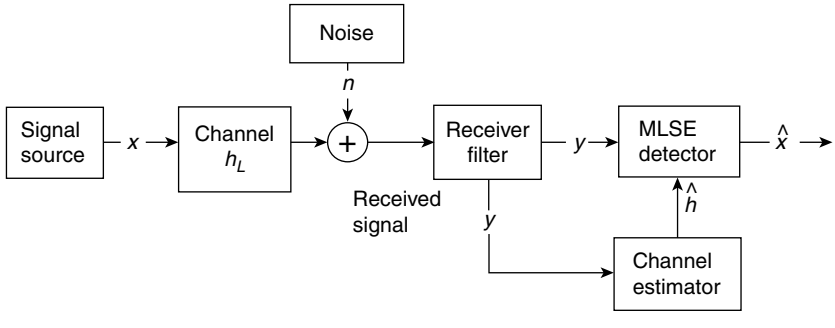


Figure 3.7 LS channel estimation technique

As shown in the Figure 3.7, the received signal first passes through the received filter and the output from the filter is represented by y, which can be expressed as $y = Mh + n$, where n denotes the noise samples, and the complex channel impulse response h of the required signal is expressed as $h = [h_0\, h_1 \ldots h_L]^T$. The transmitter sends a unique training sequence in each transmission burst, which is divided into a reference length of P and guard period of L bits, and this is denoted by $m = [m_0\, m_1, \ldots, m_{P+L-1}]^T$, where $m_i \in \{-1, +1\}$. Now, to obtain the equation $y = Mh + n$, the circulant training sequence matrix M is formed as below:

$$M = \begin{bmatrix} m_L \ldots m_1 m_0 \\ m_{L+1} \ldots m_2 m_1 \\ \ldots \\ m_{L+P-1} \ldots m_P m_{P-1} \end{bmatrix}$$

The LS channel estimates are found by minimizing the following squared error quantity:

$$\hat{h} = \arg_h \min ||y - Mh||y - Mh||^2$$

Assuming the channel to be white Gaussian, the solution can be expressed as:

$$\hat{h}_{\mathrm{LS}} = (M^{\mathrm{H}} M)^{-1} M^{\mathrm{H}} y$$

where $()^{\mathrm{H}}$ denotes the Hermitian and $()^{-1}$ the inverse matrices. The periodic autocorrelation function of the training sequence is ideal with the small delays from 1 to L, as the correlation matrix $M^{\mathrm{H}}M$ becomes diagonal. With this assumption, the given solution can be further simplified to

$$\hat{h} = \frac{1}{P} M^{\mathrm{H}} y$$

This holds good for GSM training sequences, where the reference length 16 is chosen for a normal burst training sequence (total training sequence bits are 26, where the front and back 5 bits are copied from the back and front part of the actual 16 bits sequence, respectively).

Now, the channel can change very fast, thus we need to compute the channel taps gain (h) adaptively and pass it to the equalizer. This can be done for burst or symbol based on the complexity and system requirement. If I–Q modulation is used, then the requirements are the same for computing the I and Q channels, for example, h will be complex [for example, ($h1_{\mathrm{I}} + j\, h1_{\mathrm{Q}}$), …, ($hn_{\mathrm{I}} + j\, hn_{\mathrm{Q}}$) value].

3.3.2 Equalization

Generally equalization means equalizing all parts of the received signal (with respect to time delay and frequency) to the same attenuation level (as the various components of the signals suffer different levels of attenuation or fading when travelling through the medium). This helps to mitigate the inter symbol interference (ISI), fading effect and so on. The mobile fading channel is random and varies with time, so equalizers must track the time varying characteristics of the mobile channel, and thus are called adaptive equalizers. An equalizer is usually implemented at the baseband or at the IF section in a receiver. In a typical wireless system the RF communication occurs in a passband [$f_{\mathrm{c}} - B/2, f_{\mathrm{c}} + B/2$] of bandwidth B with a center frequency f_{c}. Most of the processing such as coding/decoding, modulation/demodulation, and synchronization are performed in the baseband.

As shown in Figure 3.8, if $x_{\mathrm{b}}(t)$ is the original baseband information signal and $h(t)$ is the combined complex baseband impulse response of the transmitter, channel, and RF/IF section of the receiver, then the signal at the receiver will be $y_{\mathrm{r}}(t) = x_{\mathrm{b}}(t) \otimes h(t) + N_{\mathrm{d}}(t) + I_{\mathrm{d}}(t)$, where, $h(t)$ is the combined channel gain (its value is highly dependent on the channel fading). $N_{\mathrm{d}}(t)$ and $I_{\mathrm{d}}(t)$ are the AWGN noise and interference components, which are added with the signal during the transmission–reception process. Now, our task is to find $h(t)$ in the above equation, then compensate for the same amount. The signal received by the equalizer may be expressed as $y_{\mathrm{b}}(t) = x_{\mathrm{b}}(t) \otimes h^*(t) + n_{\mathrm{d}}(t)$, where $h^*(t)$ denotes the complex conjugate of $h(t)$ and $\otimes$ denotes the convolution, and $n_{\mathrm{d}}(t) = N_{\mathrm{d}}(t) + I_{\mathrm{d}}(t)$. If the impulse response of the equalizer is $h_{\mathrm{e}}(t)$, then $C(t) = x_{\mathrm{b}}(t) \otimes h^*(t) \otimes h_{\mathrm{e}}(t) + n_{\mathrm{d}}(t) \otimes \mathrm{h_e(t)}$, and $g(t) = h^*(t) \otimes h_{\mathrm{e}}(t) = \delta(t)$. The combination of transmitter channel and receiver appear to be an all-pass channel. In the frequency domain this equation can be written as $H_{\mathrm{eq}}(f)\, F^*(-f) = 1$, where $H_{\mathrm{eq}}(f)$ and $F(f)$ are the Fourier transform of $h_{\mathrm{eq}}(t)$ and $g(t)$. Hence, an equalizer is actually an inverse *filter* of the channel. If the channel is time varying, then it

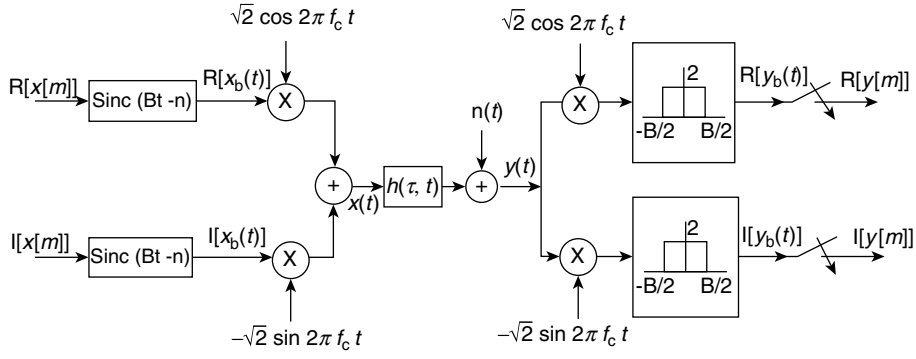


Figure 3.8 Transmitted signal $x_b(t)$ over the channel and received signal $y_b(t)$

compensates the delay, or if it is frequency selective it emphasizes the frequency components, which are highly attenuated over the channel.

The goal of channel equalization is to remove the effects of the channel on the transmitted symbol sequence $[x_k]$, that is, inter symbol interference (ISI). This can be done either by inverse filtering [for example, linear equalization (LE) or decision-feedback equalization (DFE)] or by applying sequential detection (for example, a Viterbi algorithm).

As shown in Figure 3.9, let us consider $h_T(t)$, $h_C(t)$, and $h_R(t)$ as being the transfer function of the transmitter, channel, and the receiver, respectively. During transmission from transmitter Tx to receiver Rx, the signal is distorted by the overall impulse response $h(t)$. It is composed of the impulse responses of the transmitter $[h_T(t)]$, the channel $[h_C(t)]$ and the receiver $[h_R(t)]$. When the channel characteristic is unknown and $h_T(t)^* h_C(t)^* h_R(t)$ has a large ISI, we need to use an equalizer to reduce the ISI. The equalizer can be implemented as a linear FIR or IIR or as a non-linear filter (decision feedback).

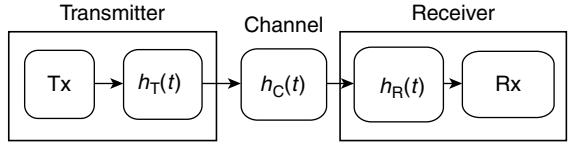


Figure 3.9 Overall channel model

The channel dynamics may not be known at the start-up and later the channel may vary with time, so an adaptive implementation of the equalizer is necessary. The following different modes of adaptation can be distinguished:

1. **Adaptation using a training signal (periodic training sequence used in GSM):** It is assumed that the channel remains unchanged over a coherent time. So a predefined known bit pattern (training sequence) is sent to study the channel, and derive the channel impulse function, which will be applied for the data demodulation, or the training sequence is sent along with the data and used to tune the filter coefficient.

2. **Decision directed adaptation:** An error signal is generated by comparing in- and output of the decision device.
3. **Blind adaptation:** Exploiting signal properties instead of using an error signal for adaptation. Example, CMA and subspace method.
4. **Rake receiver**

3.3.3 Equalizer Implementation

As shown in Figure 3.10, an EQ filter typically allows the user to adjust one or more parameters that determine the overall shape of the filter's transfer function.

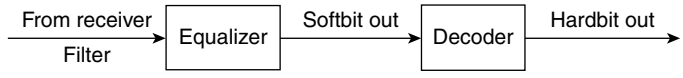


Figure 3.10 Typical implementation of conventional equalization

Many filter structures are used to implement equalizers, and in each structure there are numerous algorithms used to adapt the equalizer. The adaptive equalizer tracks the channel variations and adjusts the filter coefficients accordingly.

The basic structure of an adaptive equalizer is shown in Figure 3.11. Subscript k is used to denote discrete time index. There is a single input $y_k(t)$ passed into the equalizer block at any time instant. The value of $y_k(t)$ depends on the noise and instantaneous radio channel. The transversal filter based adaptive equalizer is shown in Figure 3.11, which has K delay elements, $K + 1$ taps, and $K + 1$ tunable multipliers, which are called weights. The weights are continuously updated by the algorithm either on a sample by sample basis or on a block by block basis.

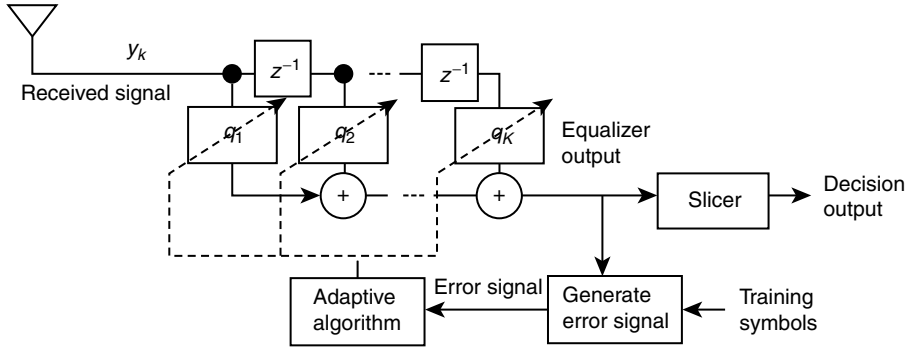


Figure 3.11 Adaptive equalizer

The error signal (e_k) is derived by comparing the output of the equalizer $y_{\text{eq}}(t)$ with some signal $y_s(t)$, which is either an exact scaled replica of the transmitted signal $x_k(t)$ or which represents a known property of the transmitted signal. Generally an adaptive algorithm is controlled by the error signal e_k. The mean

square error (MSE) between the desired and the output signal is denoted by $E\,[e(k)\,e^*(k)]$ is the most common cost function.

The least mean squares (LMS) algorithm searches for the optimum or near-optimum filter weights by performing an iterative operation. Upon reaching the convergence the adaptive algorithm sets the filter weights until the error signal exceeds an acceptable level or until a new training sequence is sent.

Methods of computing equalizer coefficients are: (1) finite-length versus infinite-length equalizers, (2) batch (ZF, MMSE, MSINR, etc.) versus recursive (LMS, RLS, etc.) equalizers, (3) symbol-spaced versus fractionally-spaced equalizers, and (4) training-based versus blind equalizers.

Two types of equalization based data detection are: (1) interference suppression – detects a signal from one user at a time while treating other users as interference; (2) multi-user detection (MUD) – detects signals from all users simultaneously and then subtracts them one by one, except for the desired one, for decoding.

3.3.4 Signal Model

Using the signal model as shown in Figure 3.12, various mathematical models are used for transfer function computation as mentioned below:

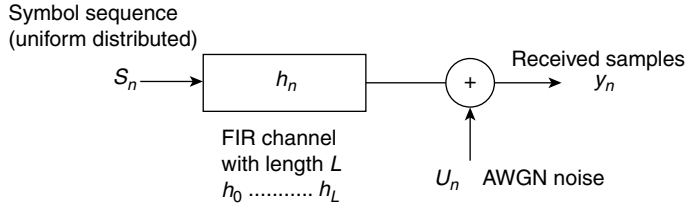


Figure 3.12 Signal model

a. **Convolution model:**

$$y_n = \sum_{l=0}^{L} h_l S_{n-l} + U_n$$

b. **Vector model:**

$$y_n = [h_0 \ldots h_L] \begin{bmatrix} S_1 \\ \cdots \\ \cdots \\ S_{n-L} \end{bmatrix} + U_n$$

c. **Matrix model:**

$$\begin{bmatrix} y_n \\ \cdots \\ \cdots \\ y_{n-N} \end{bmatrix} = \begin{bmatrix} h_0 \ldots h_L \\ \cdots \\ \cdots \\ \ldots h_0 \ldots h_L \end{bmatrix}_n \begin{bmatrix} S_1 \\ \cdots \\ \cdots \\ S_{n-L} \end{bmatrix} + \begin{bmatrix} U_n \\ \cdots \\ \cdots \\ U_{n-N} \end{bmatrix}.$$

3.3.5 Types of Equalizers

Possible equalizer types are: linear equalizer, decision feedback equalizer (DFE), maximum *a posteriori* probability (MAP) equalizer, soft-output Viterbi (MLSE) equalizer, and the possible decoder types for these types of equalizers are maximum *a posteriori* probability (MAP) decoder, and Viterbi (MLSE) decoder.

Equalization techniques can be broadly divided into two categories: linear and non-linear (see Figure 3.13). If an output decision maker block is not used in the feedback path to adapt the equalizer then it is called linear, otherwise it is called a non-linear equalizer. Linear equalizers are simple, but perform poorly under certain channel conditions. Non-linear equalizers are more complicated but significantly outperform linear equalizer.

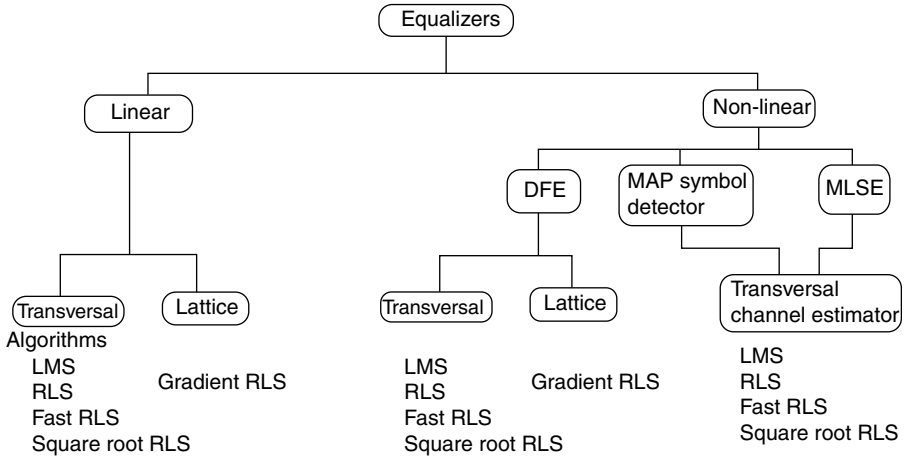


Figure 3.13 Types of equalizers

3.3.5.1 Linear Equalizer (LE)

The basic working principle of a linear equalizer is shown in the Figure 3.14. The received sequence r_k is obtained by filtering of the symbol sequence a_k by the linear channel $H(z)$ and adding noise n_k. Then the linear equalization filter $C(z)$ is applied to the input. The error signal e_k is defined as the difference between the output of the equalizer and the output of the decision device (slicer). Typically, a slicer is employed to

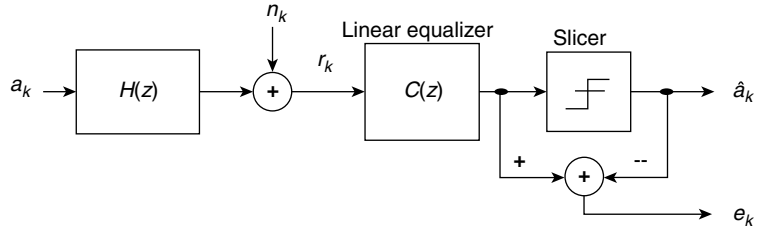


Figure 3.14 Linear equalizer blocks

obtain tentative decisions for the transmitted data symbols and to determine the noise associated with the slicer decision. The slicer simply quantizes the input to the nearest alphabet symbol.

In a zero-forcing (ZF) equalizer, if the channel response for a particular channel is *H(s)*, then the input signal is multiplied by the reciprocal of this. The zero-forcing equalizer tries to removes all ISI, and is ideal when the channel is noiseless. However, when the channel is noisy, the zero-forcing equalizer will boost the noise greatly at frequencies *f* where the channel response $H(j2\pi f)$ has a small magnitude in an attempt to invert the channel completely.

The minimum mean square error (MMSE) does not eliminate ISI completely, but rather minimizes the total power of the noise and ISI components in the output. In statistics and signal processing, a minimum mean square error (MMSE) estimator describes the approach that minimizes the mean square error (MSE). Let X be an unknown and Y be a known random variable, then an estimator $\hat{X}(y)$ is any function of the measurement Y, and its MSE is $E\{(\hat{X}-X)^2$, where the expectation is taken over both X and Y.

Linear Transversal Equalizer

A linear transversal equalization is the simplest equalizer, which is shown in Figure 3.15. The tap coefficients are adapted to suit the current channel conditions. Normally this adaptation is performed on a training sequence. In the presence of severe amplitude and phase distortion, the required inverse filter tends to result in an unacceptable degree of noise amplification.

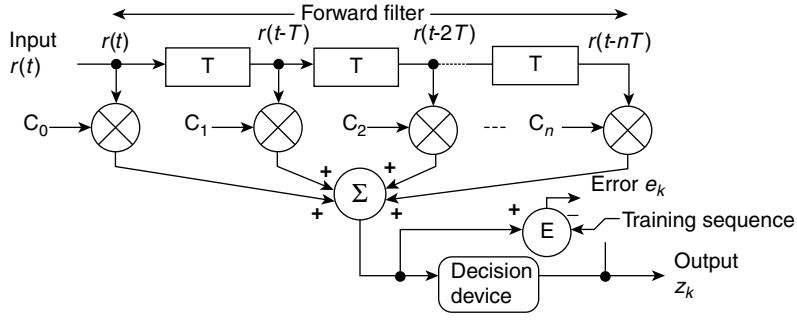


Figure 3.15 Linear transversal equalizer

Matched Filter

A matched filter is obtained by correlating a known signal, or template, with an unknown signal to detect the presence of the template in the unknown signal. The matched filter is the optimal linear filter for maximizing the signal to noise ratio (SNR) in the presence of additive stochastic noise. A filter that is matched to the signal waveform $s(t)$, where $0 \le t \le T_b$, has an impulse response $h(t) = s\,(T_b - t)$. The impulse of the optimum filter is a time reserved and delayed version of the input signal, for example, it is matched to the input signal. A linear time invariant filter defined in this way is called a matched filter. If $y(t)$ is the output of the matched filter when the input signal is $s(t)$, then

$$y(t) = \int_0^t s(\tau)\, h(t-\tau)\, \mathrm{d}\tau = \int_0^t s(\tau)\, s(T_b - t + \tau)\mathrm{d}\tau$$

Now if we sample $y(t)$ at $t = T_b$, we will get:

$$y(T_b) = \int_0^{T_b} s^2(t)\,dt = E$$

where E is the energy of the signal $s(t)$. Thus the matched filter output at the sampling instance $t = T_b$ is identical with the output of the signal correlator.

We can use a set of matched filters to build a detector, which is shown in Figure 3.16.

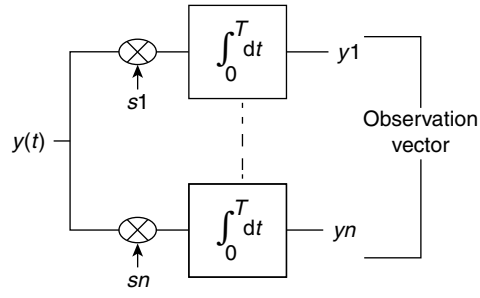


Figure 3.16 Demodulator part of matched filter receiver

For a digital signal, we can also say that the matched filter is an ordinary FIR with the conjugate channel as taps. The match filtering is performed by filtering with the time reversed, conjugate channel estimate:

$$y_{out}[t] = \sum_{k=0}^{L-1} h^*[k] y_{in}[t+k]$$

In an AWGN non-ISI channel, the received signal is mainly corrupted by white Gaussian noise and in this case matched filter can provide the optimal BER performance.

3.3.5.2 Non-Linear Equalizer

When channel distortion is too severe, then non-linear equalizers are used. Linear equalizers do not perform well on channels that have deep spectral nulls in the pass band. The most commonly used non-linear equalizers are: decision feedback equalization (DFE), maximum likelihood symbol detection, and maximum likelihood sequence estimation (MLSE).

Decision Feedback Equalization (DFE)

The basic limitation of a linear equalizer, such as the transversal filter, is the poor performance for channels having spectral nulls. A decision feedback equalizer (DFE) is a non-linear equalizer that uses previous detector decisions to eliminate the ISI on pulses that are currently being demodulated. In other words, the distortion on a current pulse that was caused by previous pulses is subtracted.

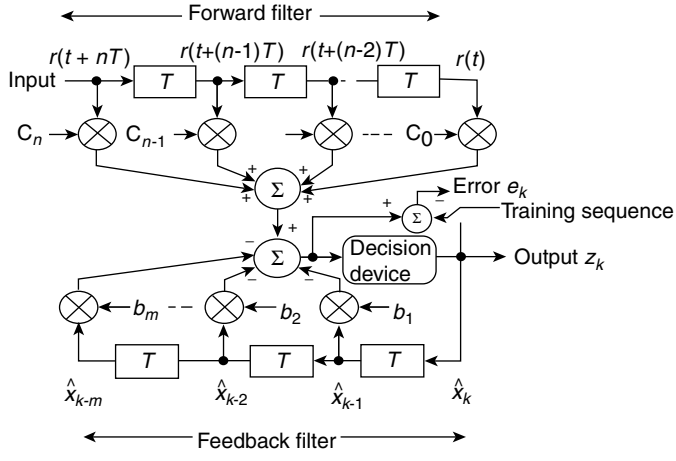


Figure 3.17 Block diagram of DFE

Figure 3.17 shows a simplified block diagram of a DFE where the forward filter and the feedback filter can each be a linear filter, such as a transversal filter. The non-linearity of the DFE stems from the non-linear characteristic of the detector that provides an input to the feedback filter. The basic idea of a DFE is that if the values of the symbols previously detected are known, then ISI contributed by these symbols can be cancelled out exactly at the output of the forward filter by subtracting past symbol values with appropriate weighting. The forward and feedback tap weights can be adjusted simultaneously to fulfil a criterion such as minimizing the MSE.

Maximum Likelihood Sequence Estimation (MLSE)

The MLSE tests all possible data sequences using a channel impulse response simulator within the algorithm and then chooses the data sequence with the maximum probability as the output. This has a large computational requirement, especially when the delay spread is large. First Forney proposed an MLSE estimator structure and implemented it with the Viterbi algorithm, which is described later. A block diagram of an MLSE receiver based on DFE is shown in Figure 3.18. The MLSE requires knowledge of

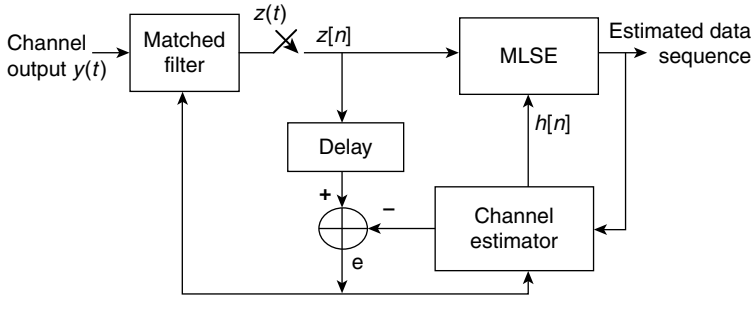


Figure 3.18 Block diagram of MLSE based receiver

channel characteristics in order to compute the metrics for making decisions. The statistical distribution of noise corrupting the signal is also necessary. Thus the probability distribution of the noise determines the form of metric for optimum demodulation of the received signal. Notice that a matched filter operates on a continuous time signal, whereas MLSE and a channel estimator rely on discretized samples.

Adaptive Equalization Using LMS Algorithm

An LMS algorithm adapts to changing channel characteristics by recursively adjusting the tap weight coefficients (C_i) to reduce the average mean square error. This is shown in Figure 3.19.

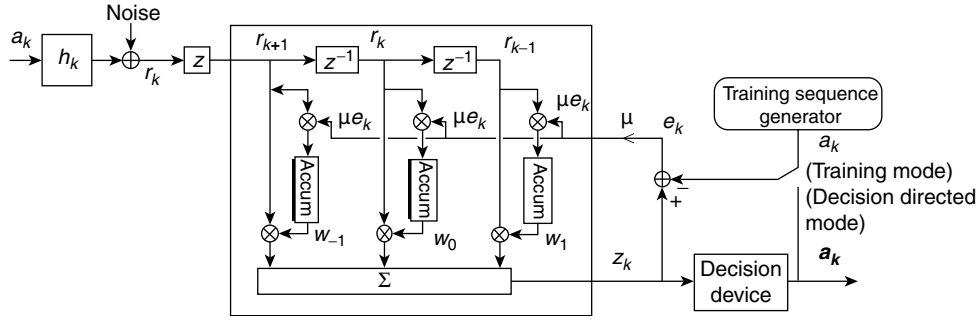


Figure 3.19 Adaptive equalizer using LMS algorithm

$$\text{LMS algorithm: } w(k+1) = w(k) - \mu e_k r_k^*$$

$$\text{Error: } e_k = z_k - a_k.$$

$$\text{Equalizer output : } v_k = w_k^* r_k = w_{-1}\, r_{k+1} + w_0 r_k + w_1 r_k = [w_{-1} w_0 w_1] \begin{bmatrix} r_{k+1} \\ r_k \\ r_{k-1} \end{bmatrix} = w^T r_k$$

Least Minimum Mean Square Error (LMMSE) Receiver

A least mean square equalizer is a robust equalizer, where the minimization of the mean square error between the desired and actual equalizer output is used as the criteria. The LMS algorithm seeks to minimize the mean square error. Least minimum mean square error (LMMSE) is a linear data detector and in the presence of ISI, it provides better BER performance than a matched filter. It can be viewed as a Wiener filter, which minimizes the mean squared estimation error. Here the mean squared estimation error serves as a cost function. As discussed in the previous section, error is given by: $e_k = z_k - a_k$. Then the mean square error $|e_k|^2$ at a time instance k, can be computed using $E[e_k^* e_k]$. If $\hat{S}_{\omega,n}$ denotes the Wiener filter output, then the estimation error $e_{w,k}$ can be expressed as: $e_{w,k} = S_n - \hat{S}_{\omega,k}$. The mean squared error can thus be denoted as: $E[|e_{w,k}|^2]_k = E[|S_k - \hat{S}_{\omega,k}|^2]_k$, where [.] denotes the average with respect to time. Filter coefficients of an LMMSE receiver are obtained by minimizing the mean square error: $w_o = \arg_{\min} E[|S_k - \hat{S}_{\omega,k}|^2]_k$. If the transmitted symbols are complex (for example, in an I,Q form), then an IQ LMMSE equalizer is used. The complexity of the IQ LMMSE receiver is at least twice that of the LMMSE receiver.

For GMSK modulated signals, BER performance can be improved by *IQ* splitting. However, this is a non-linear process and approximately doubles the complexity compared with LMMSE. BER in an AWGN ISI channel can be further improved by another type of non-linear signal processing, which models the signal coming from a first-order hidden Markov source.

3.4 Different Techniques for Interference Mitigation

Different methods can be employed for reducing the overall interference level in a system. Some of these methods are discussed below.

3.4.1 Frequency Hopping

Frequency hopping is a technique for dynamically changing the transmitter frequency with respect to time in a defined manner, instead of a statistically allocated frequency. It involves a periodic change of transmission frequency and the set of possible carrier frequencies is called the hopset. In a hopset each carrier frequency is known as a channel and has a defined bandwidth. The bandwidth of a channel used in the hopset is called an instantaneous bandwidth and the bandwidth of the spectrum over which the hopping occurs is called the total hopping bandwidth. On the other hand, the receiver knows the hopping pattern, so, it tunes the receiver accordingly to receive the channel information. Thus, ideally in a wireless system as two users are not simultaneously utilizing the frequency band, the overall interference level and probability of error decreases. However, if two users transmit simultaneously in the same channel then collision may occur. This is known as a hit and it depends on the pseudo random hopping pattern used.

This technique helps to reduce the over all interference level in the system, and is discussed further in Chapters 5 and 9.

3.4.2 Discontinuous Transmission (DTX)

Discontinuous transmission (DT) is a feature in mobile systems, where the transmitter is muted, when there is no information to be sent, such as during periods of silence. This feature prolongs battery life in portable mobile phones and reduces interference in wireless systems, as the mobile does not inject any transmission energy into the system during that period of silence.

The possibility of invoking DTX functions has extended the original speech codec specifications to include two additional functional modules: voice activity detection (VAD) and comfortable noise generator (CNG). VAD classifies the input signal into active speech, inactive speech or background noise. Based on the VAD decisions, DTX inserts silence insertion descriptor (SID) frames during the silence intervals. Throughout the silence, SIDs are periodically sent to the CNG module, which generates ambient/comfort noise during periods of inactive speech on the receiver side to indicate to the user that the call is still active.

3.4.3 Cell Sectorization

Dividing the cell into a number of sectors helps the use of a directed transmitted beam and also to reduce the transmitted power towards an unintended direction. This assists in reducing the overall interference level of the system.

3.4.4 Use of Adaptive Multi-Rate (AMR) Codec

Use of multi-rate codec helps to reduce the interference level in the system. This is discussed in detail in Chapter 8.

3.4.5 MIMO

Any system with multiple inputs into the receiver and multiple outputs from the transmitter is a MIMO system. Jack Winters at Bell Laboratories filed a patent on wireless communications using multiple antennas in 1984. He also published a paper on MIMO in 1985, based on Winters' research. Winters and many others published articles on MIMO in the period from 1986–1995. MIMO terminology focuses on the system interface with antennas rather than the air interface. The basic working principle of MIMO is: multiple antennas receive more signal and transmit more signal, which is shown in Figure 3.20.

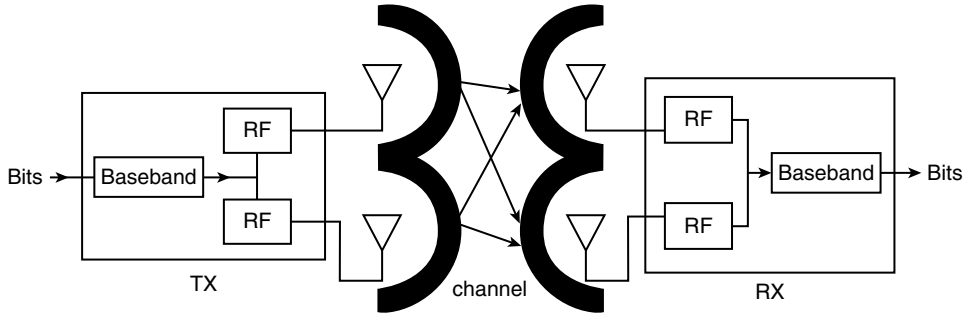


Figure 3.20 MIMO principle

Advantages of MIMO – (1) Array gain provides increased coverage and QoS (quality of service), (2) diversity gain provides increased coverage and QoS, (3) multiplexing gain provides increased spectral efficiency, and (4) co-channel interference reduction provides increased cellular capacity.

Disadvantages of MIMO – Requires multiple antennas at the transmitter and receiver side; placing multiple antennas in a mobile handset is a real issue as space is a constraint in a small mobile phone.

3.4.5.1 Single Antenna Interference Cancellation (SAIC)

In the previous chapter, we studied what interference and noise are. Basically, noise is totally unwanted signal and is not injected into the system by any transmitter, it is created naturally and unintentionally due to channel variation or by receiver components. Of these, in a single user scenario, the noise is the only concern in receiver design. However, the situation is different in case of a multi-user scenario, where many transmitters are injecting signals into the system. The signal from one transmitter looks like an undesired signal to all other un-intended receivers, except for the intended one. This signal is know as an interference signal. As discussed previously, several types of interference exist, such as, in the frequency scale adjacent channel interference (ACI) and co-channel interference (CCI), and in the time scale inter symbol interference (ISI).

For reduction of ISI, we design a system to have various equalization techniques. For reduction of ACI, we design the filters and RF components to reject or stop the signal coming from the adjacent channels. Also there is a limit put in the standard for the maximum allowed adjacent channel interference. Co-channel interference (CCI) refers to the power coming to the receiver within the band of interest due to a certain frequency reuse in distance cells in the cellular network. Frequency reuse as employed in GSM means that frequencies are repeated according to a certain reuse pattern. The larger the number of frequencies used in the reuse pattern, the lower the network capacity (measured in users/MHz/cell). Cell layout, with frequency reuse factors of 7 and 3 is shown in Figure 7.14 of Chapter 7. So smaller cell size means less transmitted power is required, but as cells are closely spaced so more frequency reuse factor means more co-channel interference.

During the initial rollout phase of GSM networks the main concern was to ensure sufficient coverage at a reasonable cost. Although today coverage is still an important factor, due to huge demand many networks are now limited by the number of users they can serve simultaneously with a good quality. Thus in order to accommodate more users many networks have introduced smaller cells and a tighter frequency reuse to increase the number of physical channels over a geographical area. This approach helps to increase the user support capacity but leads to higher co-channel interference and today the capacity of many networks is in fact limited by interference. Hence the scenario has been changed from capacity limited to interference limited.

It is known that maximum GSM/EDGE spectral efficiency is obtained in an interference limited operation, and interference therefore constitutes a major limiting factor for the performance. Hence, introduction of techniques to mitigate the effects of interference in GSM/GPRS/EDGE networks is of high importance. The interference sources typically depend on the implementation of different network elements, locations of the interfered and interfering sites with respect to each other as well as type and size of the cells/sectors.

An alternative to multiple antennas or antenna arrays (MIMO) technique is the single antenna interference cancellation (SAIC) technique, where a single antenna is used along with interference suppression or a cancellation algorithm. SAIC techniques can considerably improve the receiver performance with minimum software upgrade in a communications device. SAIC was introduced by 3GPP in Release 6. SAIC is introduced mainly to boost the capacity of GSM networks. To improve the receiver quality there are several features recommended in the downlink, out of these one is the use of SAIC. Thus SAIC capable mobiles are also often referred to as DARP mobiles (downlink advanced receiver performance).

There are several advantages of using the SAIC technique:

1. Requires one antenna only, so, easier to fit into a mobile.
2. With a given amount of system resources, a network is able to support more SAIC mobile terminals than conventional non-SAIC terminals.
3. For a given number of mobile terminals in a network, SAIC mobile terminals experience more user satisfaction in terms of frame error rate than conventional mobile terminals.
4. Owing to better receiver performance, base stations serving SAIC terminals can transmit at lower power levels. This reduces the overall level of interference in the network, which enables *all* terminals to transmit at lower power levels, reducing the interference in the network even further. Because of this effect one would also expect to see an improvement in the performance of non-SAIC terminals.

By their nature, SAIC terminals improve the bit-error-rate performance, and the mobile reported bit-error-rate governs the power output by the base station. Thus, a lower bit-error-rate reported will in turn trigger less power from the base station, and hence less interference into the system. Consequently, SAIC terminals will be able to run with lower power from the base station even if the network does not

know the SAIC capability of the terminal. This has a positive effect even on conventional terminals, as they will experience lower interference levels caused by the reduced power when transmitting to SAIC terminals. Thus, the benefit from SAIC terminals is fairly evenly shared within the terminal population, without any change of the standard or any modification in the network. The result shows that there is 40–60% increased capacity due to the use of SAIC and link level gains of 6–7 dB for GMSK modulation. The supporting companies for SAIC include Cingular, Philips, Nortel Networks, Nokia, Motorola, Ericsson, AT&T Wireless, and Intel.

Working Principle

GSM uses GMSK modulation, which has *I* and *Q* channels and carries the same information in both channels. This can be considered as two separate channels carrying the same information and experiences channel impairments independently. The real (*I* channel) and imaginary (*Q* channel) data are considered as if they are coming from two separate antennas and then use standard combining algorithms to suppress interference. The oversampling helps further, so, a received signal is over sampled at 2× and treat the real and imaginary parts for the on-time and delayed samples as four separate antennas. It makes use of four virtual channels:- space (2 channels) – *I* and *Q* channels, and time (2× Over sampled), which implies on time and delayed samples.

It then estimates the interference in the midamble (training sequence) part and applies the inverse of this correlation matrix to suppress the interference in the data part.

SAIC Implementation Algorithms

As mentioned before, SAIC refers to a class of processing algorithms (see Figure 3.21) that attempt to cancel or suppress interference using a single receive antenna.

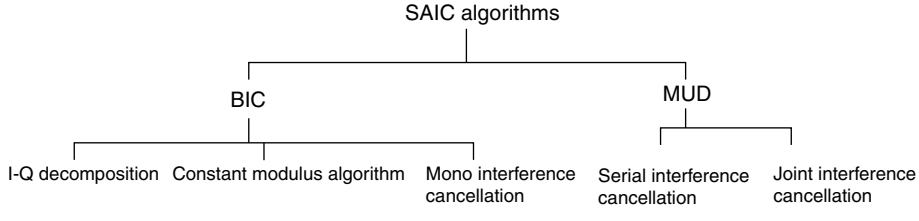


Figure 3.21 SAIC implementation algorithm

The two most prominent families of SAIC algorithms are: (1) joint demodulation (JD), also called joint detection and (2) blind interference cancellation (BIC). The basic difference between JD and BIC is that the JD receivers attempt to jointly process (demodulate) both the desired signal and one or more of the interferers, while BIC receivers only process (demodulate) the desired signal while canceling or suppressing the interference. As the BIC algorithm only specifically deals with the desired signal, it is said to be "blind" to the interferer. BIC algorithms may not be applicable to all wireless systems, but they can be applied to GMSK signals used in GSM due to the nature of the modulation.

A receiver including BIC is estimated to be between 1.5 and 3 times more complex than a conventional receiver and some JD-based algorithms with reduced complexity have claimed similar complexity. Classical JD algorithms are even more complex due to the requirement to jointly perform simultaneous demodulation, and to detect the modulation type of the interfering signal. Nevertheless, JD SAIC algorithms can sometimes offer better performance in certain interference scenarios, and thus, there is a classic trade-off in performance versus complexity between the two approaches. Most SAIC results

presented to date have been for GMSK modulation that is used for basic voice services and GPRS. Conceptually, the JD technique can also be applied to 8PSK. However, the current complexity of this approach for 8PSK makes it impractical for mass volume handsets.

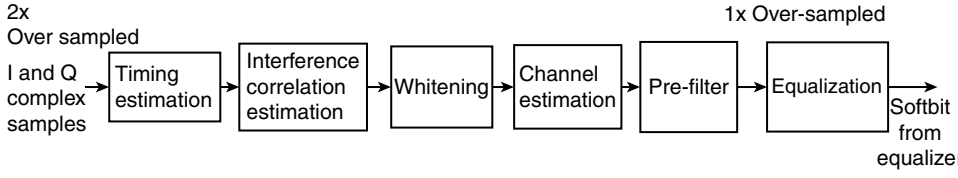


Figure 3.22 Blocks for SAIC receiver

As shown in Figure 3.22, in an SAIC receiver the timing estimation block corrects the burst timing with respect to burst reception. The basic idea of I–Q whitening is to perform joint space–time whitening of the received signal based on an estimate of the interference signal. Generally, an auto-regressive (AR) filter model is used for better interferer power estimation. The effect of removing I–Q correlation (ideally I and Q correlation is zero), that is, decorrelating the signal real and imaginary parts, is what we understand by I–Q whitening. Conventional equalizers are based on the assumption that the residual signal is white and Gaussian. So, using I–Q whitening, we transfer the color noise to white noise by using an adaptive decorrelating filter. To perform whitening we need to estimate the residual signal correlation matrix. The I–Q correlation is only partially captured by the correlation matrix of the complex interferer. We unfold the complex signal into a real-valued vector signal with twice as many samples by multiplexing the real and imaginary parts. The correlation matrix of this unfolded signal fully captures the I–Q correlation. Based on the I–Q correlation, the signal is then whitened. Next, the channel is re-estimated from the whitened signal, the pre-filter down-samples to $1\times$ oversampling and converts the channel estimate into its minimum phase equivalent, thus moving the energy towards the first channel taps. Then the equalizer detects the received softbits.

3.5 Channel Coding

As discussed in Chapter 2, owing to the inevitable presence of noise and fading in the wireless channel, the transmitted data sequences are corrupted. This increases the bit error probability at the receiver and this level of reliability is unacceptable for many applications. The design goal of channel coding is to detect and correct the bit errors in the received data sequence by adding some extra redundant bits into the transmitted data sequence.

Apparently it appears as if the source encoder (discussed in a later chapter) and the channel encoder work towards opposite goals. The source encoder extracts redundancy from the information source module and removes uncontrolled redundancy, so that the generated source data volume is less and can be easily transmitted via the channel satisfying the channel data rate, especially for any bandwidth limited system. However, the channel encoder adds redundant bits in order to provide protection against wireless channel impairment or a degree of error control for the received signal. There are two different approaches for error control coding. (1) We can add a small number of redundant bits to the original data sequence to allow the receiver only to detect the errors (see Figure 3.23). Now if the error is detected, the receiver can ignore the data or it may request a retransmission. This is known as automatic retransmission request (ARQ). This is mostly implemented by the link layer protocol (L2). (2) We can insert a large number of redundant bits to the data sequence to allow the receiver both to detect and then correct the error bits in the data sequence. This strategy is known as forward error correction (FEC). This is mostly taken care of by the physical layer (L1).

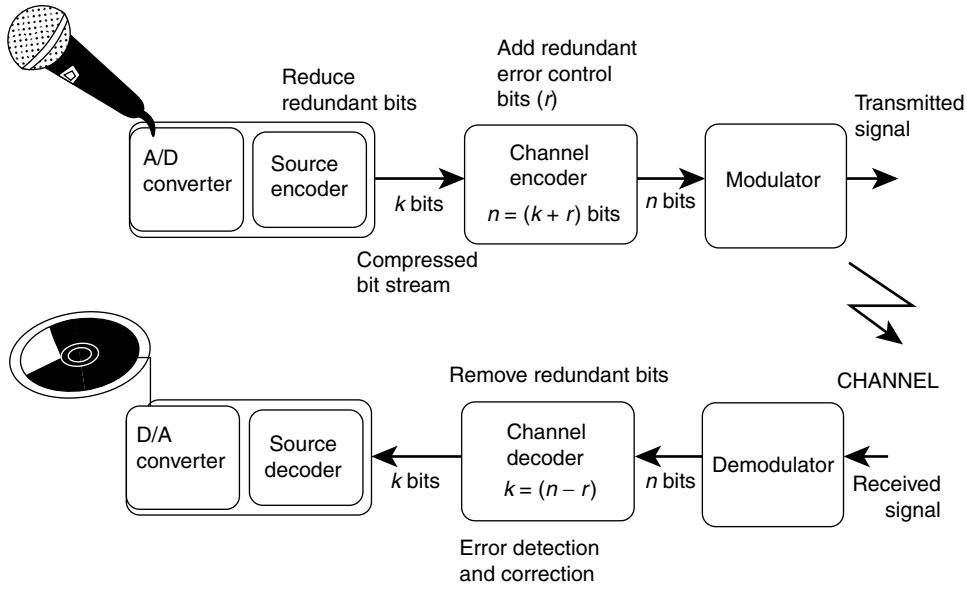


Figure 3.23 Channel encoder and decoder blocks

Generally, the error correction and detection codes can be broadly classified into three categories: block codes, convolution codes, and turbo codes.

3.5.1 Block Codes

A block code operates on fixed-length input blocks of information bits, which are known as message blocks. Here the general principle is to segment the information into blocks and add a parity check number, which is normally the product of information bits contained in the block. Block codes are FEC codes that help to detect and correct a limited number of errors. In a block encoder, k is the input information bits to the encoder unit and the encoder then uses different generator polynomials and adds r numbers of extra bits to k bits. So the total number of output bits from the encoder will be $n = r + k$. The block code is referred to as an (n, k) code, and the throughput or rate of the code is defined as $R_c = k/n$. With a k input bit sequence, 2^k distinct codewords can be transmitted. For each codeword, there is a specific mapping between the k message bits and the r check bits. The code is systematic because a part of the sequence in the codeword coincides with the k message bits. As a result, it is possible to make a clear distinction in the codeword between the message bits and the parity bits. The code is binary as these are constructed from bits and linear as each codeword can be created by a linear modulo-2 addition of two or more codewords.

Besides code rate, there are some other parameters, which are most commonly used for performance measurement, and these are:

Code Distance – The distance between two codewords is known as code distance and it is the number of elements in which two codewords C_i and C_j differ.

$$D(C_i, C_j) = \sum_{l=1}^{N} C_{i,1} \oplus C_{j,1} \,(\text{modulo } q)$$

In the above equation, D is the distance between two codewords and q is total number of possible values of C_i and C_j. The length of each codeword is N elements or characters. If the code is binary code, then the distance is known as the Hamming distance. Generally, we define the *Hamming distance* (or simply the distance) between two codewords of a binary code, as the number of digits in which they differ. For example $d(0,1) = 1$, $d(001,011) = 1$, $d(000,111) = 3$, $d(111,111) = 0$.

The *minimum distance* of a code is the minimum of all distances between distinct codewords. In symbols, the minimum distance of $C = \min \{d(x,y) \mid x,y \varepsilon C\}$.

The Euclidean distance between any two signals is: $d_{ij}^2 = |s_i - s_j|^2$ (when i is not equal to j). The relation between the Euclidean distance and the Hamming distance is very simple. If the sequences s_i and s_j differ in d_{ij}^{H} locations, then their Euclidean distance d_{ij}^{E} and Hamming distance d_{ij}^{H} are related by $(d_{ij}^{E})^2 = 4\, d_{ij}^{H}\, E$, where E is the energy related to each bit.

Weight of Code – The weight of a codeword of length N is given by the number of non-zero elements in the codewords. For a binary code the weight is basically the numbers of 1s in the codewords and is given by:

$$W(C_i) =^{N} \sum_{l=1}^{N} C_{i,1}$$

The encoder for a block code is memory-less, which means that the n digits in each codeword depend only on each other and are independent of any information contained in previous codewords.

3.5.1.1 Examples of Block Codes

There are many types of block codes available, several of these are discussed below.

Hamming Codes

In 1950, Richard Hamming introduced a linear error-correcting code known as Hamming code. Hamming codes can detect and correct single-bit errors. In other words, the Hamming distance between the transmitted and received code words must be zero or one for reliable communication. Alternatively, it can detect (but not correct) up to two simultaneous bit errors, whereas, simple parity code cannot correct errors and cannot be used to detect more than one error. In mathematical representation, Hamming codes are a class of binary linear codes. For each integer $m > 1$ there is a code with parameters: $[2^m - 1, 2^m - m - 1, 3]$. The parity-check matrix of a Hamming code is constructed by listing all columns of length m that are pair-wise independent. For example (7, 4) Hamming code, encodes 4 data bits into 7 bits by adding 3 parity bits. This can detect and correct single-bit errors and can detect double-bit errors. The code generator matrix and the parity-check matrix are:

$$G: = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}; \; H: = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

For example, 1011 is encoded into 0110011.

Golay Codes

Marcel J.E. Golay generalized Hamming's idea for perfect single-error correcting codes based on any prime number. An x-error-correcting code must have a minimum distance of at least $2^x + 1$. For the code to be perfect, the number of vertices of the unit n-cube inside a packing sphere of radius x must be a power of r, where r is the radix of the code. In the binary case: $\sum_{i=0}^{x} \binom{n}{i} = 2^k$ for some integer k. This is the sum of the first $x + 1$ entries of the nth row of the Pascal triangle. Golay found two such numbers, $\sum_{i=0}^{2} \binom{90}{i} = 2^{12}$ and $\sum_{i=0}^{3} \binom{23}{i} = 2^{11}$. For $n = 23$, Golay found a 3-error correcting (23,12) code and provided a matrix for it. Golay codes are linear binary (23,12) codes with a minimum distance of 7 and an error correction capability of 3 bits. This is the only nontrivial example of perfect code.

Cyclic Codes

A code is defined as cyclic, if any cyclic shift of any codeword is also a codeword. For example, if 101101 is a codeword, then after shifting the last bit to the first bit position and then all other bits to the right by one position, the resultant codeword is 110110. This subcategory of codes is simple to implement using hardware by linear shift registers with feedback. Two important types of cyclic block codes used in practical systems are Bose–Chaudhiri–Hocquenghem (BCH) and the Reed–Solomon (RS) codes.

BCH Code

BCH codes are a class of linear cyclic block codes discovered by R.C. Bose and D.K. Ray-Chaudhuri and independently by A. Hocquenghem. This allows multiple error correction and provides a large choice of block lengths and code rates. The most commonly used BCH codes are binary with following parameters:

Codeword length: $n = 2^m - 1$ bits, where $m = 3, 4, \ldots$ Number of check bits: $n - k <= m\,t$, where t is the number of correctable errors per codeword.

Reed–Solomon Code

Reed–Solomon code is non-binary code, which is capable of correcting errors that appear in bursts and are commonly used in concatenated coding systems. In non-binary block codes, the input bit stream is converted into symbols of m bits long. These symbols are segregated into message blocks, each of which are k symbols in length. The encoder then adds r check symbols of m bits long and creates a codeword of length n symbols. RS (n,k) codes exist for all n and k for which: $0 < k < n < 2^m + 2$.

However, for the most commonly used RS (n, k) code, $(n, k) = (2^m - 1, 2^m - 1 - 2t)$, where, t is the number of symbols that are correctable per codeword. The number of parity symbols, $n - k$, equals $2t$. For RS codes, the code minimum distance $d_{\min} = n - k + 1$.

3.5.2 Convolution Codes

Historically, convolution codes were first introduced by Elias in 1954 and become popular after the discovery of an efficient decoding algorithm by A.J. Viterbi in 1967. As for block codes, the convolution codes divide the bit stream from the source into k-bit blocks. Each k-bit block is then encoded into an n-bit block, but unlike block codes, the values of n bits depend not only on the values of the k bits in the corresponding source block but also on the values of the bits in the previous k-bit source blocks. Although covolution code is more complex, these are more powerful than block codes as they exploit past history.

The idea is to make every codeword symbol to be the weighted sum of the various input message symbols. This is like the convolution used in LTI systems to find the output of a system, when you know the

input and impulse response. It is also known as tree codes, the encoder has memory and the output codeword depends on the current bit values as well as on the previous bit values held within the registers. A binary convolution encoder is structured as a mechanism of shift registers and modulo-2 adders, where the output bits are modulo-2 additions of selective shift registers content and present input bits.

The convolution codes are commonly specified by three parameters: k – the number of input bits that form the block input to the encoder, so k is the number of input bits being shifted into the encoder at any time tick; n – the number of coded output bits; and m – the number of input registers. These represent the number of previous k-bit blocks that influence the encoding of the present block.

Often manufacturers of convolution code chips specify the code parameters as (n, k, L), where the quantity L is called the constraint length of the code and is defined by $L = k\ (m - 1)$. It represents the number of bits in the encoder memory that affect the generation of the output bits. The L parameter is known as the constraint length, which is equal to the number of k-tuple stages in the encoder shift register. Each tuple stage of the shift register contains k registers to hold the k number of input bits coming at a time.

The encoder shift register of $L.k$ bit stages and n output generators are shown in Figure 3.24. From the figure it is evident that the first register is unnecessary (as it stores the last input bit), and therefore the required number of shift register stages is ($L.k - 1$), instead of $L.k$ stages.

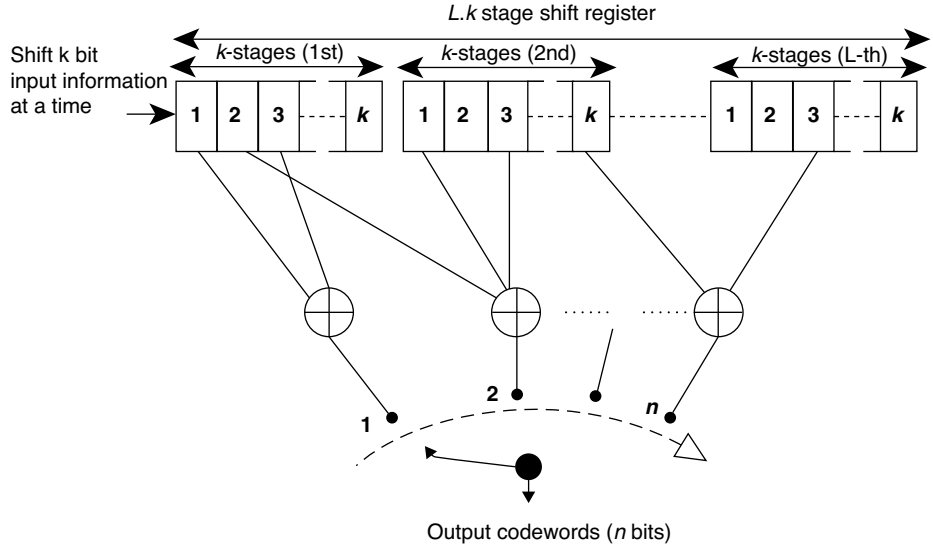


Figure 3.24 Convolution encoder of rate k/n with constraint length L

It consists of n modulo-2 adders. At each shift instant, k bits are shifted into the first k stage of the register and all other bits, which were already in the register, are shifted by k bits to the right. The outputs of n adders are then sequentially sampled by the commutator to generate n output code bits at a time.

The output bits are generated according to the bits present in the registers and the generator polynomials used to connect different registers to the adders. A convolution code is usually defined in terms of sequences of the convolution code, denoted by $g_1, g_2, g_3, \ldots, g_n$. For example, the generator sequence

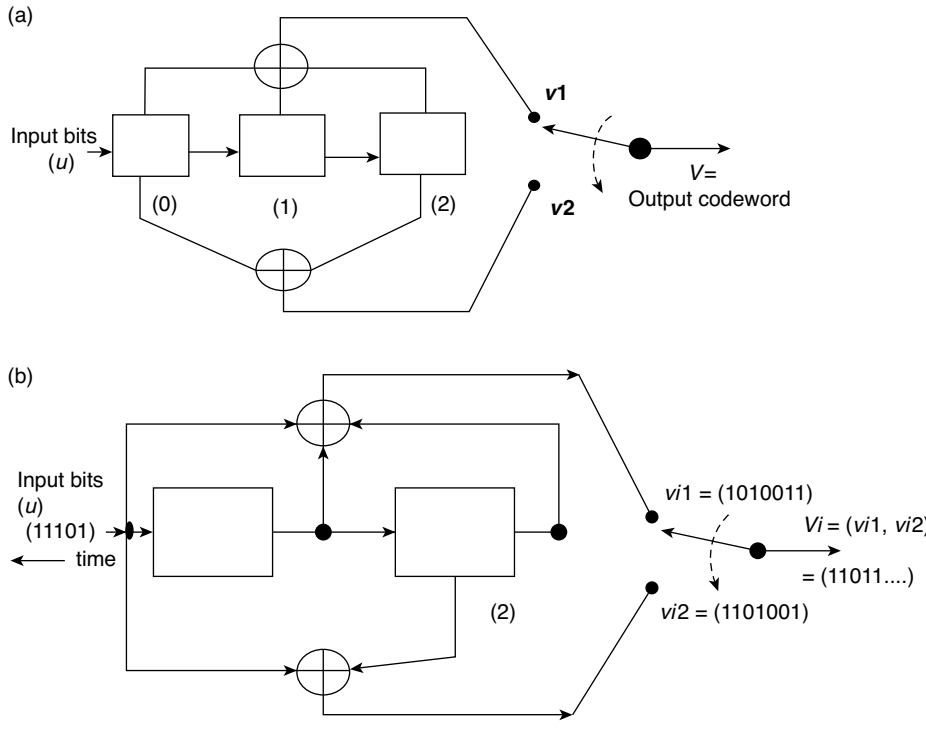


Figure 3.25 (a) Half-rate convolution encoders with constant length $L = 2$. (b) Equivalent representation of (a) with input and output data stream

can be represented by: $g_1 = [00101001]$, $g_2 = [000000001]$, and $g_3 = [100000001]$. Here, 0 indicates that the register number is not connected to the adder, and 1 indicates that the register is connected to the adder. We can also define the generator matrix of the convolution code:

$$G: = \begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

An encoder for the binary (2,1,2) convolution code is shown in Figure 3.25b, where u is the information sequence and v is the corresponding code sequence (codeword).

The input bit sequence $u = 11101$, and generator sequence $g_1 = [111]$, $g_2 = [101]$.

Here 1 input at a time is inserted. When the first bit "1" is inserted, $v1 = 1 \oplus 0 \oplus 0 = 1$, $v2 = 1 \oplus 0 = 1$. V1 = (1, 1). (Register bit sequence: 1 0 0.) When the second bit "0" is inserted, $v1 = 0 \oplus 1\ (+)\ 0 = 1$, $v2 = 0 \oplus 0 = 0$. $V1 = (1, 0)$. (Register bit sequence: 0 1 0.) When the third bit "1" is inserted, $v1 = 1 \oplus 0 \oplus 1 = 0$, $v2 = 1 \oplus 1 = 0$. $V1 = (0, 0)$. (Register bit sequence: 1 0 1.)

Similarly, the other sequences will be generated: $V = (11\ 01\ 10\ 01\ 00\ 10\ 11)$, where the encoded sequence:

$$
\begin{aligned}
v &= u.G \\
&= (1,1,1,0,1) \cdot \begin{pmatrix} 11 & 10 & 11 & & & & \\ & 11 & 10 & 11 & & & \\ & & 11 & 10 & 11 & & \\ & & & 11 & 10 & 11 & \\ & & & & 11 & 10 & 11 \end{pmatrix} \\
&= (11, 01, 10, 01, 00, 10, 11)
\end{aligned}
$$

It is assumed the shift registers that generate convolution code is loaded with "0" before the first information bit enters. Thus the information bit sequence is padded with $(L-1).k$ number of "0"s to bring back the convolutional encoder to an all zero state. We assume that the length of the information bit sequence is a multiple of k. If the length of the input sequence is not a multiple of k, we pad it with 0s such that the resulting length is a multiple of k.

The encoder for the convolutional code uses a look-up table to do the encoding. The look-up table consists of four items: (1) input bit, (2) the state of the encoder, which is one of the eight possible states, (3) the output bits, and (4) the output state that will be the input state for the next bit.

3.5.2.1 Representation of Convolution Codes

The encoder can be represented in several different but equivalent ways. These are described below.

Generator Polynomial Representation

We can represent a convolution encoder by a set of n generator polynomials, one for each of the n modulo-2 adders. We simply write a polynomial $G_j(x) = g_{j1} + g_{j2}x + \ldots + g_{jk}x^{k-1}$, based on the K-tupple encoder connection vector $G_j = (g_{j1}, g_{j2}, \ldots, g_{jk})$. Generator representation shows the hardware connection of the shift register taps to the modulo-2 adders. A generator vector represents the position of the taps for an output. A "1" represents a connection and a "0" represents no connection. This has already been shown in Figure 3.25.

State Diagram Representation

Here 1/01 represents that the input binary digit to the encoder was 1 and the corresponding codeword output is 01, for example, input bits/output bits (see Figure 3.27). The encoder states are just a sequence of bits. The (2,1,2) code has a constraint length of 2 (see Figure 3.26). The shaded registers hold these

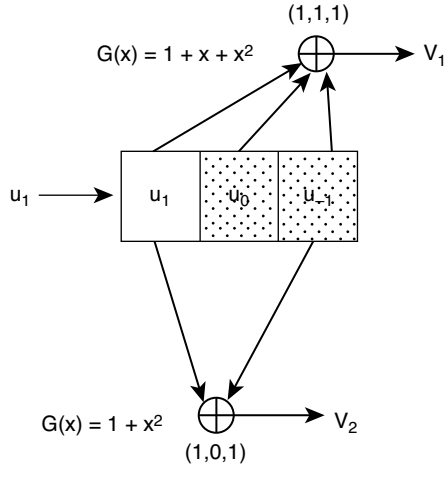


Figure 3.26 Binary (2, 1, 2) convolutional code

bits and the unshaded register holds the incoming bit. This means that 2 bits or 4 different combinations of these bits can be presented in the memory registers. These 4 different combinations will determine what output we will get for v_1 and v_2, the coded sequence. The number of combinations of bits in the shaded registers are called the states of the code and are defined by: number of states $= 2^L$, where L is the constraint length of the code and it is equal to $k\cdot(m-1)$.

The corresponding state diagram representation of the encoder (which is shown in Figure 3.25) is shown in Figure 3.27, where $L = k$. $(m-1) = 1$. $(3-1) = 2$. So the total number of possible states $= 2^L = 2^2 = 4$ and these are 00, 01, 10, 11.

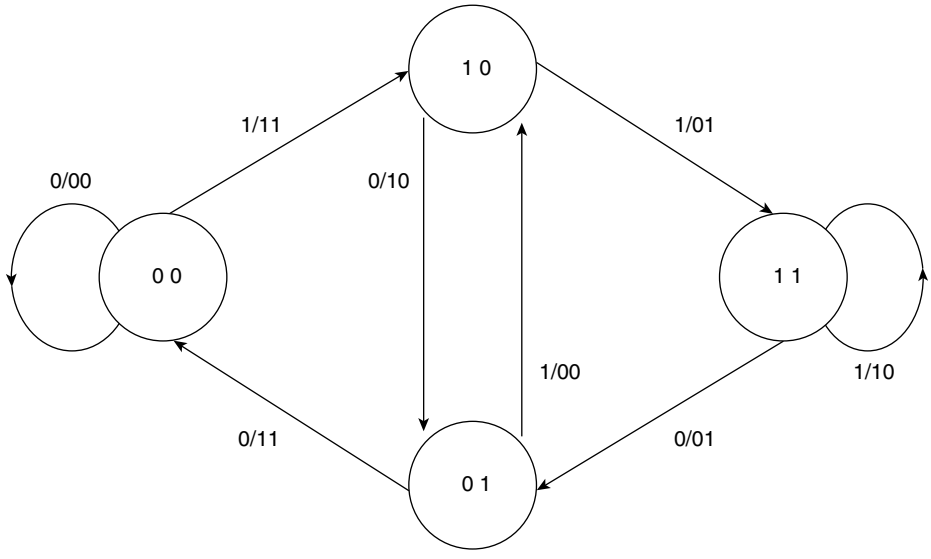


Figure 3.27 State diagram representation of the encoder with code (2,1,4)

In the state diagram, the state information of the encoder is shown inside the circle. Each input information bit causes a transition from one state to another. In the figure, the path information between the states in denoted as x/c, where x represents the input information bit, and c the output encoded bits. It is customary to begin convolution encoding from the all zero state. For example, let us consider that the initial state was "00", now if "0" comes as input the state will change to the "00" state again and the output codeword will be $(0 \oplus 0 \oplus 0, 0 \oplus 0) = (0,0)$. Thus the path is represented by 0/00 with self transition. However, if 1 comes as input, the state will change to (10)0, for example, 10, as 0 (the right most bit) will be pushed out. Now the output code will be $(1 \oplus 0 \oplus 0, 1 \oplus 0) = (1,1)$. The path will be represented as 1/11. Similarly, from state 10, if 0 arrives as input, the next state transition will be shifted to (01)0, for example, 0 will be pushed out, and the output codeword will be $(0 \oplus 1 \oplus 0, 0 \oplus 0) = (1, 0)$. The path will be represented as 0/10. If 1 arrives as input, the state will switch to the 11 state and the output bit sequence will be $(1 \oplus 1 \oplus 0, 1 \oplus 0) = (0,1)$. The path will be represented as 1/01. Now, at state (11) if 0 arrives as input the state will change to (01) and the output bit seqence will be $(0 \oplus 1 \oplus 1, 0 \oplus 1) = (0,1)$. Then if 1 arrives as an input, the state will switch to (11) and the output will be $(1 \oplus 1 \oplus 1, 1 \oplus 1) = (1,0)$. From state (01) if 0 arrives as input, the next state transition will be to state (00) and output state will be $(0 \oplus 0 \oplus 1, 0 \oplus 1) = (11)$. If 1 arrives as the input bit, the next state will be (10) and the output sequence will be $(1 \oplus 0 \oplus 1, 1 \oplus 1) = (0,0)$.

Tree Diagram Representation

The tree diagram attempts to show the passage of time as we go deeper into the tree branches. Here instead of jumping from one state to another, we go down the branches of the tree depending on whether a 1 or 0 is received. The starting state is assumed to be 000. If a 0 is received, we go up and if a 1 is received, then we go downwards. It is somewhat better than a state diagram, but still not the preferred approach for representing convolutional codes.

Trellis Diagram Representation

A close observation of tree representation shows that the structure repeats itself once the number of stages is greater than the constraint length. Two nodes having the same transition label can be merged and by doing this throughout the tree diagram, we can obtain another diagram called a trellis diagram, which is a more compact representation. A trellis diagram is a state transition diagram plotted versus time. All possible states are shown on the y-axis and the discrete time along x-axis. A code of (n,k,L) will have a trellis diagram with $2^{(L-1)k}$ states, and each state is represented by dots arranged across a vertical column and this column is repeated horizontally across the x-axis with the number of received bit sequence. We move horizontally through the trellis with the passage of time and each transition means new bits have been received. The transitions between the states are indicated by branches connecting these dots. On each branch connecting the states of the present and the next column, a transition label is marked and "n" binary symbols indicate the encoder output corresponding to that transition. Also note that we always start from the all-0 state (for example, 000 in the case of $L = 3$), move through the trellis following the branches corresponding to the given input sequence, and return to the all-0 state. So, the codewords of a convolutional code correspond to paths through the corresponding trellis, starting at the all-0 state and returning to the all-0 state. Although the trellis diagrams are messy they are generally preferred over both the tree and the state diagrams because they represent a linear time sequencing of events. As with state or tree diagrams, the trellis diagram is unique to each code. We can draw the trellis for as many periods as we want. Each period repeats the possible transitions. We always begin at an all-0 state, such as 000. Starting from here, the trellis expands and in L bits becomes fully populated such that all transitions are possible. The transitions then repeat from this point on.

As example of trellis diagram representation is shown in the Figure 3.28, where for example the total number of states is 8, generator polynomials $g_1 = (1,1,1,1)$, $g_2 = (1,1,0,1)$ and each vertical column of nodes corresponds to the possible states at a single time step. As shown, the trellis indicates the assumption that the decoder is always initialized to a 000 state at time 0. Based on the value of first input bit (0 or 1), the encoder can remain in state 000 or transition to 100 for time 1. Remaining in state 000 is indicated by the upper branch (solid line) leaving state 000 time 0 and entering state 000 time 1. This transition branch is labeled 0(00), indicating that 0 is the input bit and (00) is the output bit from the coder.

To understand this more clearly, the Figure 3.29 shows the trellis diagram of code (2,1,4) with an input bit sequence 1011000 and the corresponding generated output bit sequence 11011111010111. The incoming bits are shown on the top. Here the total number of states $= 2^{(4-1)^*1} = 8$, for example, from 000 to 111, as shown by the dots in a column (nodes). We will start at an all-0 state at the top left most corner dot, where $k = 1$, so each time 1 input bit arrives and $n = 2$ for example, two output bits. The input bit sequence is 1011000. In the branch line the 1 input bit is indicated first and then the corresponding 2 output bits are indicated inside the brace. Thus first input bit 1 and then output bit (11), so the label represented is 1 (11). The previous state was 000, when 1 pushes inside the register, so the new sequence will be 100 (and the right most 0 goes out). Hence it goes to the 100 state. It continues like this depending on the input bit.

3.5.2.2 Decoding

Several methods exist for decoding of convolution codes. These are broadly classified into two basic categories: (a) sequential decoding, for example the Fano algorithm and (b) maximum likelihood decoding, for example Viterbi decoding.

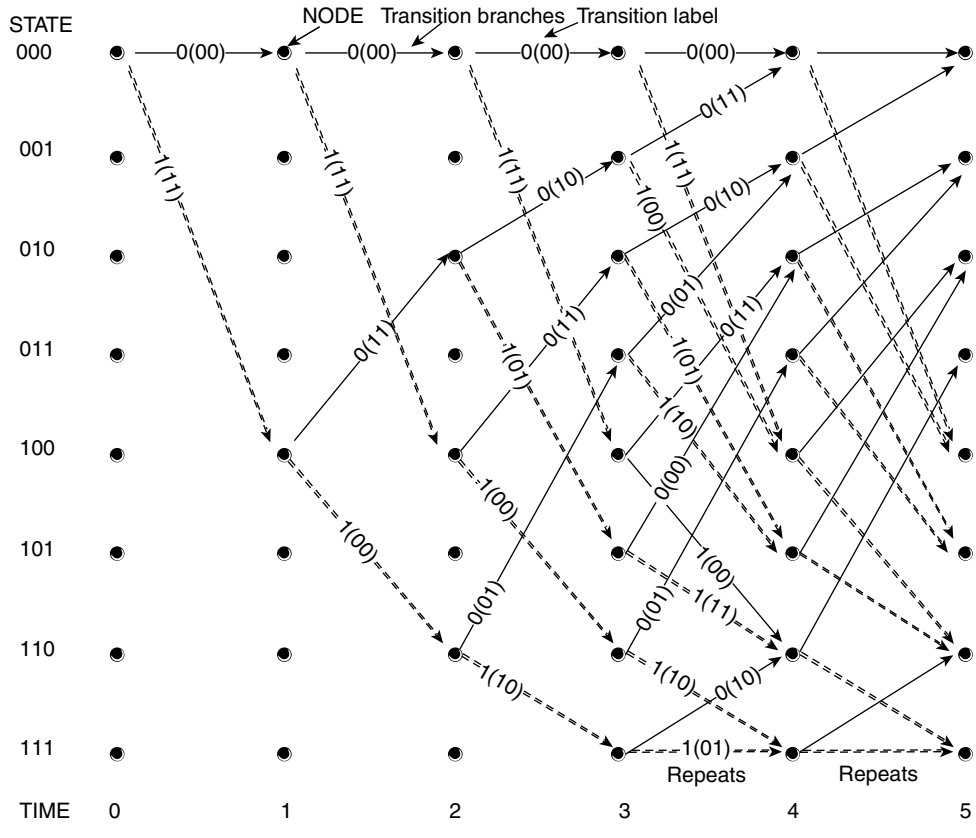


Figure 3.28 Trellis diagram of (2,1,4) code

Other decoding algorithms also exists, such as the stack algorithm, feedback decoding, and although the basic principle of decoding is same, the former two decoding methods represent two different approaches. If a data string of length n bits is sent after passing through a $^1/_2$ rate convolution coder, then we receive 2^*n number of bits. Now, these 2^*n received bits may or may not have errors. We know from the encoding process that these bits map uniquely. So an n bit sequence will have a unique 2^*n output bits. However, owing to errors, we can receive any and all possible combinations of the 2^*n bits. We can then apply one of the following to make a decision:

a. We can compare this received sequence with all possible sequences and select the one with the smallest Hamming distance (or bit disagreement) – this is the idea behind hard decision decoding.
b. We can do a correlation and pick the sequences with the best correlation – this is the idea behind soft-decision decoding. Soft-decision decoding is superior by about 2–3 dB.

If a data stream of length $2n$ bits is received, then the possible number of codewords is 2^{2n}. As the number of bits increases, the number of calculations required to do decoding in this brute force manner increases such that it is no longer practical to do decoding this way. We need to find a more efficient method that does not examine all options and has a way of resolving the ambiguity.

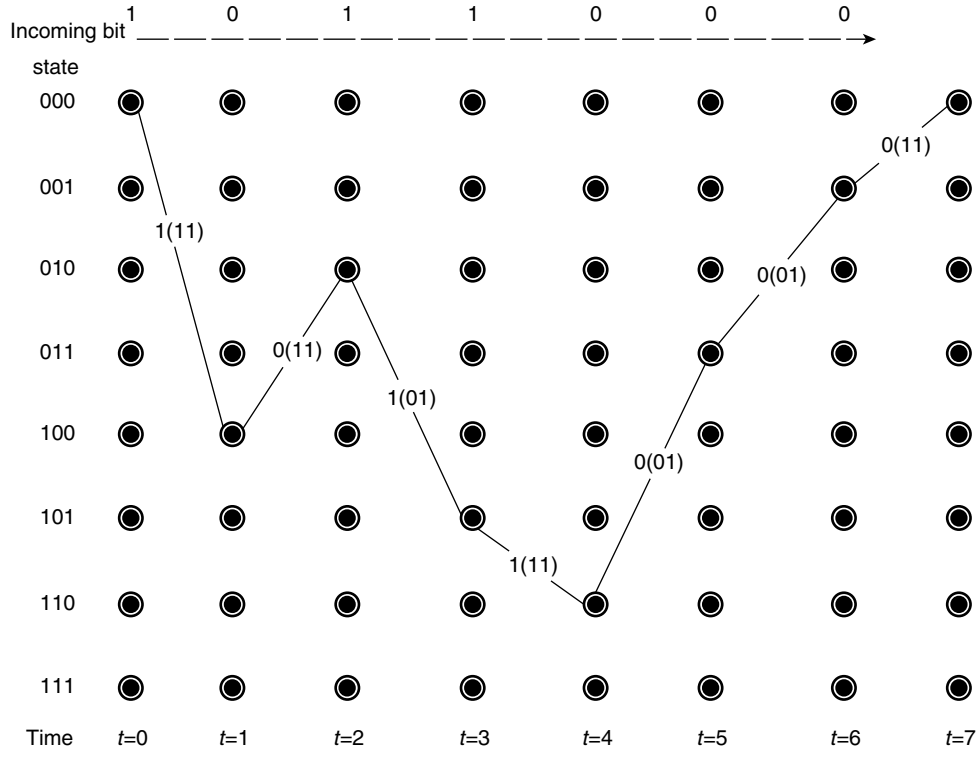


Figure 3.29 Encoded sequence of input bits 1011000, output bit 11011111010111 for (2,1, 4) code

Sequential Decoding

Sequential decoding is one of the first methods proposed for decoding a convolutionally coded bit stream. This was first proposed by Wozencraft and later a better version of it was proposed by Fano. This decoding mechanism is best described by analogy. For example, you are given some directions about how to get to a place, which consists of particular landmarks. However, the given direction was not correct, which is why occasionally you do not recognize a landmark and you end up on the wrong path. When you end up in a wrong path the landmarks do not match, so you realize that you may be on a wrong path. Thus you backtrack to a point where you do recognize a landmark and take an alternate path until you see the next landmark. In this way finally you arrive at the destination. In this process, you may back track several times based on how good the directions were. Similarly, in sequential decoding we are dealing with just one path at a time. This allows both forwards and backwards movement through the trellis, and the decoder keeps track of its decisions, each time it makes an ambiguous decision, it tallies it. If the tally increases faster than some threshold value, then the decoder gives up that path and retraces the path back to the last fork where the tally was below the threshold. The memory requirement of sequential decoding is manageable and so this method is used with long constraint length codes where the S/N is also low.

3.5.2.3 Maximum Likelihood and Viterbi Decoding

In 1967 A.J. Viterbi came up with an algorithm, which is known as the Viterbi algorithm and this is most widely used decoding method for convolution codes. Viterbi decoding is the best known implementation of the maximum likelihood decoding. In such a decoder, the received sequence is compared with each possible transmitted code vector and the one closest to the received sequence is selected as the correct transmitted code sequence. The term closest is used in the sense of the minimum number of differing binary symbols (Hamming distance) between the code vectors under investigation. We narrow the options systematically at each time tick. These concepts are used to reduce the choices, the errors occur infrequently and the probability of error is small. The probability of two errors in a row is much smaller than a single error, which indicates that the errors are randomly distributed.

Given a convolution code and a received bit sequence, the Viterbi algorithm can be summarized as below.

1. Parse the received sequence into m subsequences each of length n. k is the information bits in each block. Draw a trellis of depth m for the code under study. As a code tree is equivalent to a trellis, we may equally limit our choice to possible paths in the trellis representation of the code. The reason for preferring a trellis over a tree is that the number of nodes at any level of the trellis do not continue to grow as the number of incoming message bits increases; rather it remains constant at 2^{L-1}, where L is the constraint length of the convolution code.
2. For the last $L-1$ stages of the trellis, draw only paths corresponding to the all-0 sequences. This is done as we know that the input sequence has been padded with $k(L-1)$ 0s. Set $l=1$ and set the metric of all-0 state equal to 0. Find the distance of the l-th subsequence of the received sequence to all branches connecting the l-th stage to the $(l+1)$-th stage states of the trellis.

The Viterbi decoder examines an entire received sequence of a given length. The decoder computes a metric for each path and makes a decision based on this metric. All paths are followed until two paths converge on one node. Then the path with the higher metric is kept and the one with lower metric is discarded. The paths selected are called the survivors. For an N-bit sequence, the total number of possible received sequences is 2^N. Of these only 2^{kL} are valid. The Viterbi algorithm applies the maximum-likelihood principles to limit the comparison to 2 to the power of kL surviving paths instead of checking all paths. The most common metric used is the Hamming distance metric. This is just the dot product between the received codeword and the allowable codeword. These metrics are cumulative. If the hard decision decoding is used, this algorithm finds the path that is at the minimum Hamming distance from the received sequence, and if the soft-decision decoding is employed, the Viterbi algorithm finds the path that is at the maximum Euclidean distance from the received sequence.

Example – Let us consider a simple convolution encoder with code (2,1,2) as shown in Figure 3.30.

Say the input bit sequence to encoder is *100010100*, where the 2-bit tail of zeros is added at the end to flush out the encoder (so, total $7+2=9$ input information bit). Output bit sequence $y=(g_0, g_1)=((u_1 \oplus u_2 \oplus u_3), (u_1 \oplus u_3))$.

The initial state was 000, when input bit 1 is inserted, the state transition will be 000 → 100, output will be $((1\oplus 0\oplus 0)\ (1\oplus 0))=(11)$. When the second input bit 0 is inserted, the state transition will be 100 → 010, output will be $((0\oplus 1\oplus 0)\ (0\oplus 0))=(10)$. When the third input bit 0 is inserted, the state transition will be 010 → 001, output will be $((0\oplus 0\oplus 1)\ (0\oplus 1))=(11)$. When the fourth input bit 0 is inserted, the state transition will be 001 → 000, output will be $((0\oplus 0\oplus 0)\ (0\oplus 0))=(00)$. When the fifth input bit 1 is inserted, the state transition will be 000 → 100, output will be $((1\oplus 0\oplus 0)\ (1\oplus 0))=(11)$. When the sixth input bit 0 is inserted, the state transition will be 100 → 010, output will be $((0\oplus 1\oplus 1)\ (0\oplus 0))=(10)$. When the seventh input bit 1 is inserted, the state transition will be

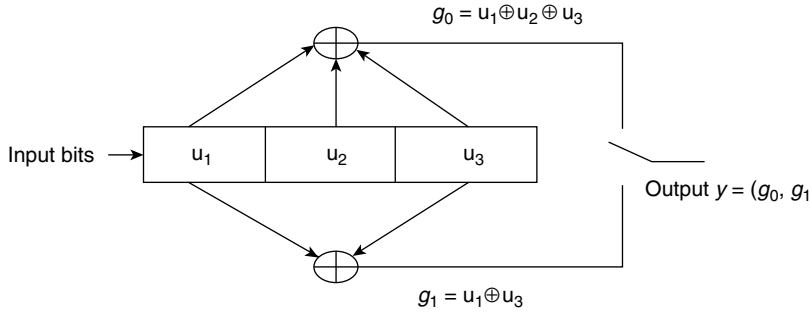


Figure 3.30 Simple convolution encoder (2,1,2)

$010 \rightarrow 101$, output will be $((1 \oplus 0 \oplus 1)\,(1 \oplus 1)) = (00)$. When the eighth input bit 0 is inserted, the state transition will be $101 \rightarrow 010$, output will be $((0 \oplus 1 \oplus 0)\,(0 \oplus 0)) = (10)$. When the ninth input bit 0 is inserted, the state transition will be $010 \rightarrow 001$, output will be $((0 \oplus 0 \oplus 1)\,(0 \oplus 1)) = (11)$.

Thus the $9 \times 2 = 18$ bit output generated by the encoder is – 11 10 11 00 11 10 00 10 11.

This is transmitted through the channel and suppose that two bit errors occurred, then the received code sequence is 10 10 11 00 11 11 00 10 11.

The length of the received sequence is 9. We have to draw a trellis of depth 9. Also, note that as the input information sequence is padded with $k.(L-1) = 2$ number of 0s, for all final two stages of the trellis, we will draw only the branches corresponding to all-0 inputs. This also means that the actual length of the input sequence is 7, which after padding with two 0s, has increased to 9. The total number of states in the trellis will be $2^{(L-1)k} = 4$, so 4 numbers of nodes are drawn in a column (at a given time click) (see Figure 3.31).

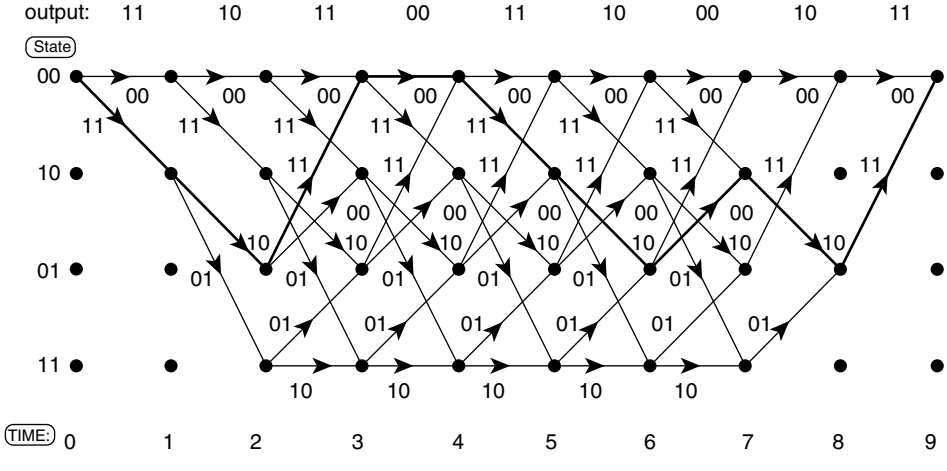


Figure 3.31 Encoder trellis diagram showing the path traversed to encode input sequence 100010100

The receiver knows:

a. The generator polynomials $y = (g_0, g_1) = ((u_1 \oplus u_2 \oplus u_3), (u_1 \oplus u_3)$ and for each two output bits there was one input bit inserted into the encoder in the transmitter side.
b. It also knows that the encoder started at the 00 state.
c. Received code sequence is 10 10 11 00 11 11 00 10 11 and there might be several errors in this received sequence when it passed through the channel.

Now from this, the decoder has to decode the correct transmitted output bit sequence, which is 11 10 11 00 11 10 00 10 11. Then from this it can easily find out the input bit sequence *100010100* using a mapping table or generator polynomials or paths. As shown in the Figure 3.32, the starting point will be state 00 at time 0 (left most corner node in the trellis). Therefore, after receiving the first two output bits 10 (which corresponds to one input information bit – either 0 or 1), the receiver knows that one of two possibilities might have occurred:

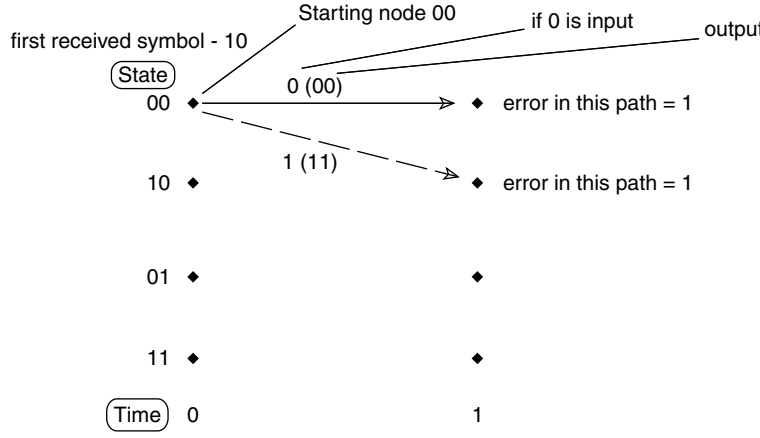


Figure 3.32 Partial encoder trellis at time tick 1

1. If the first input information bit was 0, then the state transition from time 0 to time 1 will be 00 → 00. The output will be $y = (g_0, g_1) = ((0 \oplus 0 \oplus 0), (0 \oplus 0) = (00)$, which is marked as transition label.
2. If the first input information bit was 1, then the state transition from time 0 to time 1 will be 00 → 10. The output will be $y = (g_0, g_1) = ((1 \oplus 0 \oplus 0), (1 \oplus 0) = (11)$.

The received sequence is (10), so the error if option (1) is chosen will be simply the bit difference between (10) and (00) = 1. Similarly, the error if option (2) is chosen will be simply the bit difference between (10) and (11) = 1. The two possibilities are equally likely, as each implies that one bit error was introduced by the channel. The result is summarized in Figure 3.32.

Upon receiving the second symbol (10), the receiver knows that one of four possibilities must have occurred. (1) It was at state 00 and the input bit was 0. Then the state transition from time 1 to time 2 will be 00 → 00. The output will be $y = (g_0, g_1) = ((0 \oplus 0 \oplus 0), (0 \oplus 0) = (00)$. (2) It was at state 00 and the input bit was 1. Then the state transition from time 1 to time 2 will be 00 → 10. The output will be $y = (g_0, g_1) = ((1 \oplus 0 \oplus 0), (1$ XOR $0) = (11)$. (3) It was at state 10 and the input bit was 0. Then the state transition from time 1 to time 2 will be 10 → 01. The output will be $y = (g_0, g_1) = ((0 \oplus 1 \oplus 0), (0 \oplus 0) = (10)$. (4) It

was at state 10 and the input bit was 1. Then the state transition from time 1 to time 2 will be 10 → 11. The output will be $y = (g_0,g_1) = ((1 \oplus 1 \oplus 0), (1 \oplus 0) = (01)$.

When the received sequence is (10), so the error if option (1) is chosen will be the bit difference between (10) and (00) = 1. Similarly, the error if option (2) is chosen will be the bit difference between (10) and (11) = 1. The error if option (3) is chosen will be the bit difference between (10) and (10) = 0. The error if option (4) is chosen will be the bit difference between (10) and (01) = 2. The total error in path connected (time 0, node 00) to (time 2, node 00) is 1 + 1 = 2, and (time 0, node 00) to (time 2, node 10) is 1 + 1 = 2, and (time 0, node 00) to (time 2, node 01) is 1 + 0 = 1, and (time 0, node 00) to (time 2, node 11) is 1 + 2 = 3. The result is summarized in Figure 3.33.

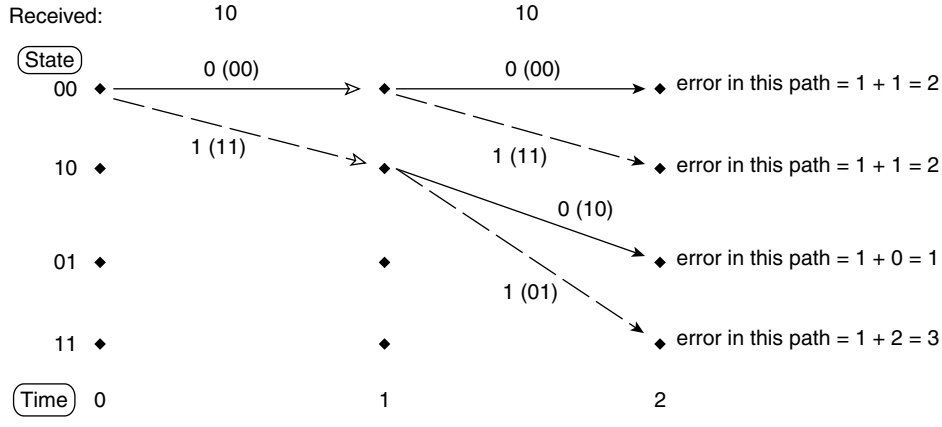


Figure 3.33 Partial encoder trellis at time tick 2

Upon receiving the third symbol (11), the receiver knows one of the eight possibilities must have occurred:

1. It was at state 00 and the input bit was 0. Then the state transition from time 2 to time 3 will be 00 → 00. The output will be $y = (g_0,g_1) = ((0 \text{ XOR } 0 \text{ XOR } 0), (0 \text{ XOR } 0) = (00)$.
2. It was at state 00 and the input bit was 1. Then the state transition from time 2 to time 3 will be 00 → 10. The output will be $y = (g_0,g_1) = ((1 \text{ XOR } 0 \text{ XOR } 0), (1 \text{ XOR } 0) = (11)$.
3. It was at state 10 and the input bit was 0. Then the state transition from time 2 to time 3 will be 10 → 01. The output will be $y = (g_0,g_1) = ((0 \text{ XOR } 1 \text{ XOR } 0), (0 \text{ XOR } 0) = (10)$.
4. It was at state 10 and the input bit was 1. Then the state transition from time 2 to time 3 will be 10 → 11. The output will be $y = (g_0,g_1) = ((1 \text{ XOR } 1 \text{ XOR } 0), (1 \text{ XOR } 0) = (01)$.
5. It was at state 01 and the input bit was 0. Then the state transition from time 2 to time 3 will be 01 → 00. The output will be $y = (g_0,g_1) = ((0 \text{ XOR } 0 \text{ XOR } 1), (0 \text{ XOR } 1) = (11)$.
6. It was at state 01 and the input bit was 1. Then the state transition from time 2 to time 3 will be 01 → 10. The output will be $y = (g_0,g_1) = ((1 \text{ XOR } 0 \text{ XOR } 1), (1 \text{ XOR } 1) = (00)$.
7. It was at state 11 and the input bit was 0. Then the state transition from time 2 to time 3 will be 11 → 01. The output will be $y = (g_0,g_1) = ((0 \text{ XOR } 1 \text{ XOR } 1), (0 \text{ XOR } 1) = (01)$.
8. It was at state 11 and the input bit was 1. Then the state transition from time 2 to time 3 will be 11 → 11. The output will be $y = (g_0,g_1) = ((1 \text{ XOR } 1 \text{ XOR } 1), (1 \text{ XOR } 1) = (10)$.

When the received sequence is (11), so the error if option (1) is chosen will be the bit difference between (11) and (00) = 1. Similarly, the error if option (2) is chosen will be the bit difference between (11) and (11) = 0. The error for option (3) = dif (11), (10) = 1, error for option (4) = dif (11), (01) = 1, error for option (5) = dif (11), (11) = 0, error for option (6) = dif (11), (00) = 2, error for option (7) = dif (11), (01) = 1, error for option (8) = dif (11), (10) = 1. The total error in path connected (time 0, node 00) to (time 3, node 00) via (time 2, node 00) is 1 + 1 + 2 = 4, similarly total errors for different paths are shown in the Figure 3.34. Of the two paths arriving at each node for time 3, the less likely one (more errors) can be pruned away as shown in Figures 3.34, 3.35 and 3.36.

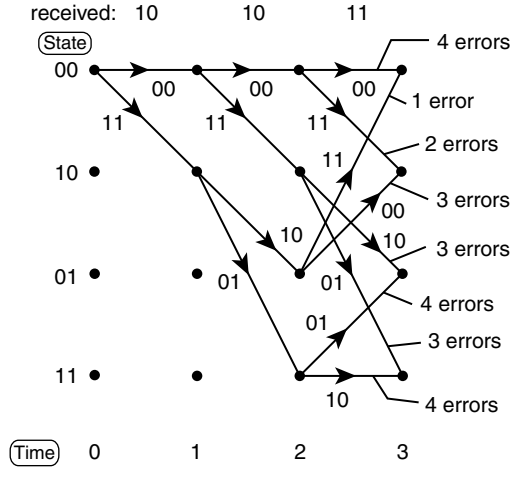


Figure 3.34 Partial encoder trellis at time tick 3

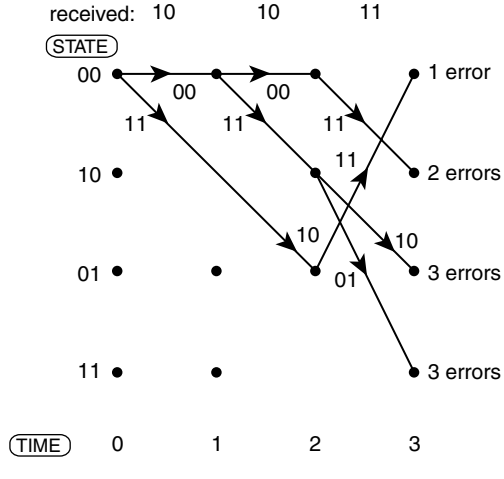


Figure 3.35 Partial encoder trellis at time tick 3 after removing the non-survivor paths

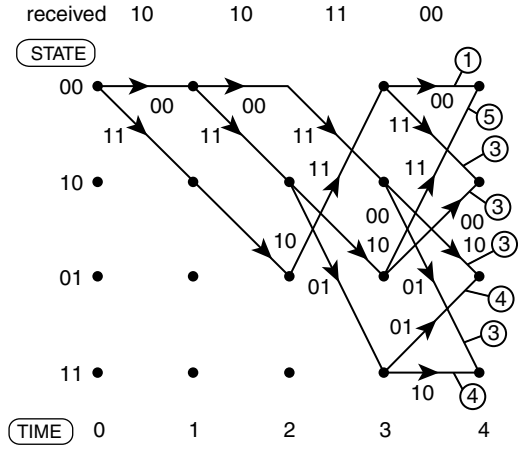


Figure 3.36 Partial encoder trellis at time tick 4

In general, the trellis for time k shows all eight possible transitions from states at time $k-1$. Then, in the trellis for time $k+1$, non-surviving paths from $k-1$ to time k are pruned away. In cases of a tie, the surviving path can be selected arbitrarily. The soft decision metrics greatly reduce the indices of a tie. This way, the process continues and is finally reached at time 9. The partial encoder trellis for time 9 is shown in Figure 3.37. From this we can select the best path and the transition on the best path provides the output bits, and so the corresponding input bits are chosen. Thus the correct input information bits are found.

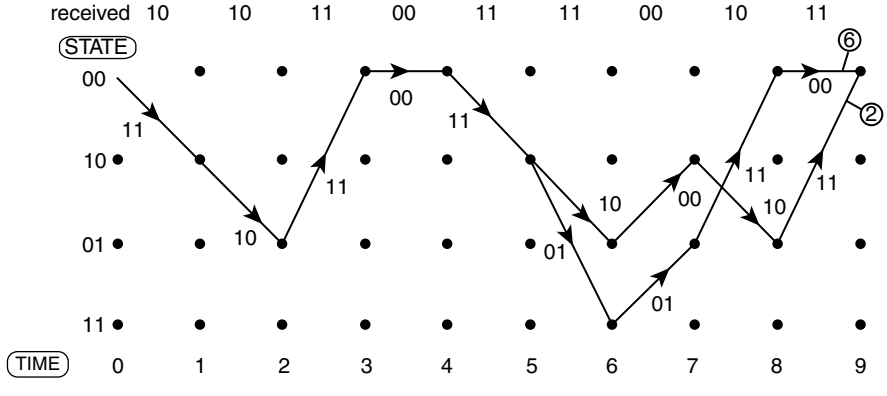


Figure 3.37 Partial encoder trellis at time 9

3.5.2.4 Viterbi Decoding with Soft Decision

In the previous example, we considered when the Viterbi decoders involved only hard decisions and the input to the decoder were 00, 01, 10, and 11. However, instead of hard input bits, if we input to the decoder the soft decision values from the demodulator, then in this situation the Viterbi decoder performs better.

The soft decisions convey an indication of received signal quality and how confident the receiver is regarding the decisions that have been made (indicates with confidence value).

Generally, in a detector circuit we try to estimate the energy value over a bit period by integrating the signal energy over the bit period as shown in Figure 3.38, and then take a decision of 1 or 0 comparing with a threshold value. This is basically making a hard decision by deciding whether it is strictly 1 or 0. Assume a demodulator output voltage of $+1$ V corresponds to a bit value of 1 and an output voltage of -1 V corresponds to a bit value of 0. Making a hard decision means that a simple decision threshold that is usually between the two signals is chosen. In very simple terms, this also means the maximum-likelihood decoding.

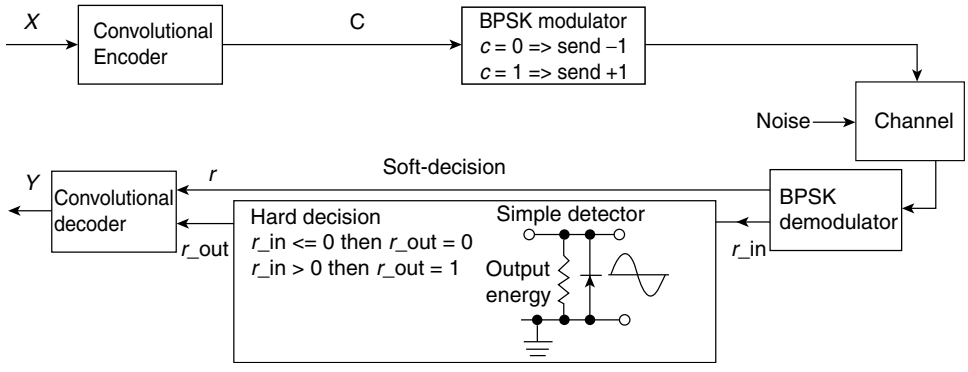


Figure 3.38 Hard decision and soft decision decoding

However, when the received signal is corrupted by the noise, then determining the 0 or 1 bit from the received signal energy spectrum becomes difficult. The demodulator output for binary 1 or 0 will have a probability density function (PDF) as shown in Figure 3.39b. The signal spreads out and the energy from one signal leaks into the other, which is a function of S/N. If the added noise power is small, that is, noise

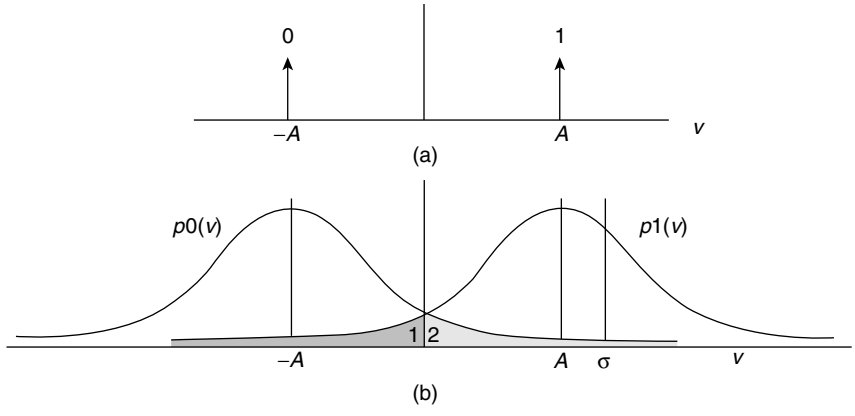


Figure 3.39 (a) Two signals representing a 1 and 0 bit (hard decision representation), and (b) noise of S/N = 2 spreads out to spill energy from one decision region into another (soft decision representation)

variance (σ) is small, then the spread will be less. Intuitively we can say that decoding errors will be less likely if the S/N is high or noise variance is small.

We can quantify the error that is made with this decision method. The probability that "1" was sent but decoded as "0," is a function of the two shaded areas as shown in the Figure 3.39b. Here the area "1" represents the energy that belongs to the signal corresponding to "1" that has leaked into the opposite decision region and hence erroneously leads to the decoding of the bit as "0", and area 2, which is the energy from the bit 0 that has leaked into the region of "1". This affects the value of the sampled received voltage, hence potentially causes an error in the decision. Given that a 1 was sent, the probability that it will be decoded as 0 will be

$$Pe1 = \frac{1}{2} erfc\left(\frac{A - v_t}{2\sigma}\right)$$

where, v_t = decision threshold voltage, which is considered as 0 for this case, and σ = variance of noise or its power. We can also rewrite the above equation as a function of S/N:

$$Pe1 = \frac{1}{2} erfc\left(\sqrt{\frac{S}{N}}\right)$$

The above equation is the well known bit error rate equation. Here we are making a hard-decision, but instead of having just two regions of decision, we divided the area into 4 or more regions as shown in the Figure 3.40.

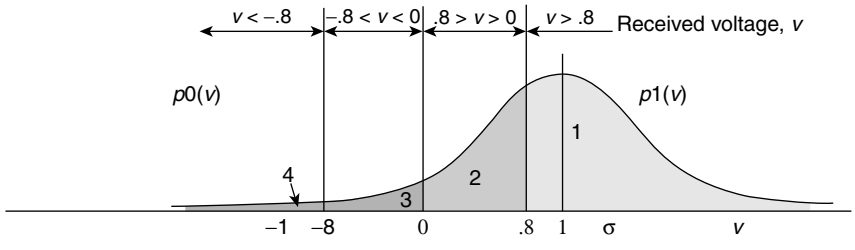


Figure 3.40 Creating four regions for making decision

As shown in the Figure 3.40, the probability that the decision is correct can be computed from the area under the Gaussian curve. We chose four regions as follows:

Region 1 = if (received voltage >0.8 V), region 2 = if (0 < received voltage < 0.8 V), region 3 = if (−0.8 V < received voltage < 0 V), region 4 = if (received voltage < −0.8 V).

Now say the received voltage falls in region 3, then what is the probability of error that a 1 was sent? If a hard decision method is used then the answer is easy as it can be calculated using the above equation. However, if soft decision method is used, then to calculate the similar probabilities for a multi-region space, we use the Q function. The Q function gives us the area under the tail defined by the distance from the mean to any other value. Thus $Q(2)$ for a signal, the mean of which is 2, would give us the probability of a value that is 4 or greater.

In this example as shown in Figure 3.41, we have assumed a v_t of 0.8 but it can be any number. Pe1 (probability that a 1 was sent if the received voltage is in region 1) = $1 - Q(A - v_t/\sigma)$. Pe4 (probability that a 1 was sent if the received voltage is in region 4) = $Q(2(A + v_t)/\sigma)$. Pe2 (probability that a 1 was sent if the

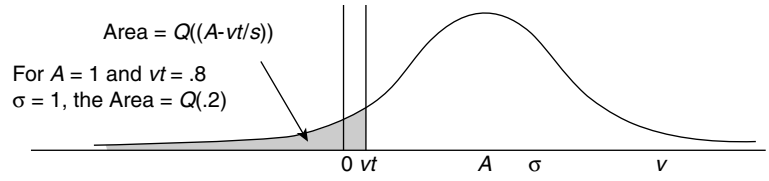


Figure 3.41 The Q function to determine the probabilities of a normally distributed variable

received voltage is in region 2) $= 1 - \text{Pe1} - Q(A/\sigma)$. Pe3 (probability that a 1 was sent if the received voltage is in region 3) $= 1 - \text{Pe1} - \text{Pe2} - \text{Pe4}$.

We have computed these for S/N $= 1$ assuming that v_t is $0.8A$ and $A = 1.0$ and also that both bits 0 and 1 are equi-probable. This is also called the a priori probabilities for bits 0 and 1.

So the area of region 4 or Pe4 is equal to $Q(1.8)$, which we can look up from the tables. The area of region 1 is then equal to $1 - Q(.2)$. The others can be quickly calculated in this manner.

This process of subdividing the decision space into regions greater than two is called soft decision. These probabilities are also called the transition probabilities. Here the soft bit will indicate the sign value (which is shown as 1 or 0) and is followed by the confidence value (for example, how confident the detector is about this decision of whether it is 1 or 0). Soft decision improves the sensitivity of the decoding metrics and improves the performance by as much as 3 dB in the case of an 8-level soft decision.

In the decoding section, we calculated a Hamming metric by multiplying the received bits with the codewords. We do the same thing now, except, instead of receiving 0 and 1, we get one of these voltages. The decoder looks up the metric for that voltage in its memory and makes the following calculation. Assume voltage pair ($v3$, $v2$) are received. The allowed codewords are 01, 10.

$$\text{Metric for } 01 = p(0|v3) + p(1|v2) = -4 + -4 = -8$$
$$\text{Metric for } 10 = p(1|v3) + p(0|v2) = -1 + -1 = -2$$

We can say, just by looking at these two, that 01 is much more likely than 10. When these metrics add, they exaggerate the differences and help the decoding results.

The decoder trellis for the same example in the previous section using soft decision is shown in the Figure 3.42.

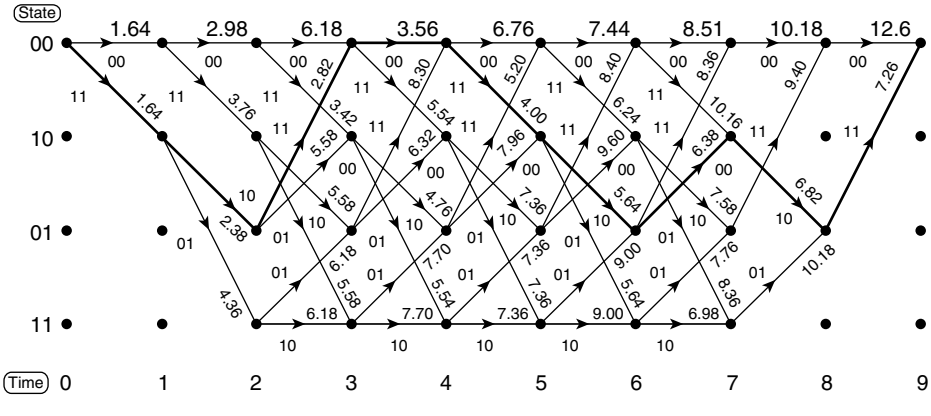


Figure 3.42 Decoder trellis using soft decision method

The decoding delay and amount of memory required for decoding a long information sequence are unacceptable. The decoding can not be started until the whole sequence is received and also the total surviving paths have to be stored.

q-Format Representation

The q-format is specified by a value representing the number of bits on the right side of the decimal separator. In a soft decision of an x-bit value, the sign tells the polarity of the decision and the amplitude tells its certainty. The magnitude of decision represents the confidence in that decision; for example, for 0.3, its $q - x$ representation will be 0.3×2^x and a $q15$ representation will be 0.3×2^{15}.

Generally, all fixed-point formats are specified using 2s complement notation and are defined through the two types of Q formats (signed and unsigned) as given in Table 3.1.

Table 3.1 Q format representation

Attributes	uQy_x	Qy_x
Total number of bits	x	x
Number of fractional bits	y	y
Number of sign bits	0	1
Number of integer bits	$x - y$	$x - y - 1$
Range	$[0; 2^{x-y} - 2^{-y}]$	$[-2^{x-y-1}; 2^{x-y-1} - 2^{-y}]$
Resolution	2^{-y}	2^{-y}

Log-Likelihood Ratio

There are other ways of improving decoding performance by playing around with the metrics. One important metric to know about is called the log-likelihood metric. This metric takes into account the channel error probability and is defined by

$$\text{Metric for agreement} = \frac{\log_{10} 2(1-p)}{\log_{10} 2}$$

$$\text{Metric for disagreement} = \frac{\log_{10} 2(p)}{\log_{10} 2}$$

For $p = 0.1$, the metric for agreement is $= -20$ and the metric for disagreement $= -1$. Thus, if we have received bits 01 and the codeword is 00, the total metric would be $-20 + -1 = -21$ and the metric for complete agreement would be -40. A Fano algorithm used for sequential decoding uses a slightly different metric and the purpose of all these is to improve the sensitivity. Viterbi decoding is fairly important as it also applies to decoding of block codes. This form of trellis decoding is also used for trellis-coded modulation (TCM).

3.5.3 Turbo Codes

Turbo codes are a class of advanced error correcting codes, which offer robust error correction capabilities for a wide variety of channels. Using these codes, it is possible to get as close as a fraction of a dB to a Shannon limit at low SNR. Turbo codes are a special class of concatenated codes, where an interleaver is placed between two parallel or serial encoders (constituent codes). The basic structure of a Turbo encoder and decoder are shown in Figures 3.43 and 3.44. The constituent codes

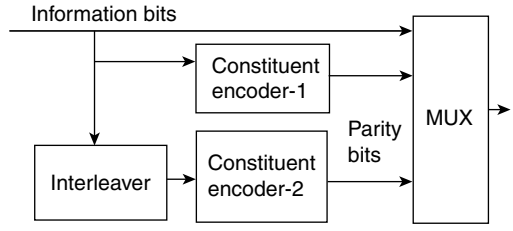


Figure 3.43 Turbo-encoding

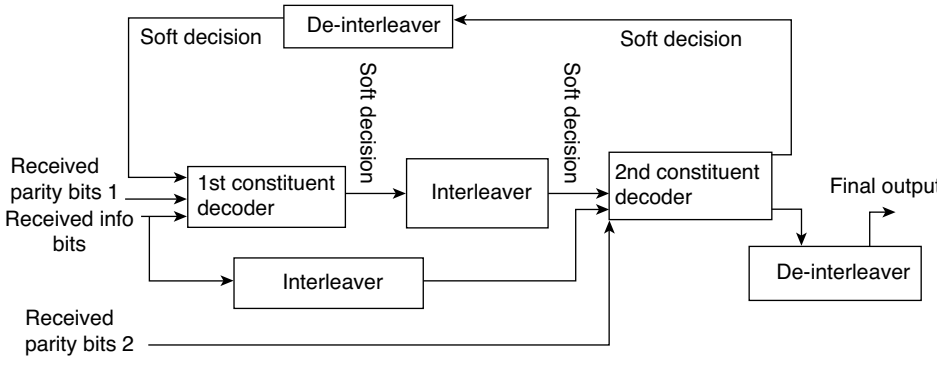


Figure 3.44 Turbo-decoding

are usually recursive systematic convolutional codes (RSCC) of rate 1/2 and usually the same code is used in encoder 1 and 2. Recursive convolutional codes (inside the encoder) have a feedback path from the delay elements back to the input. As the conventional convolutional code does not have the feedback path, so it behaves as an FIR digital filter, whereas the RSCC are like IIR filters. Here, the N information bits enter the first encoder, and pass through the interleaver and then through the second encoder. As the encoders are systematic, so each one generates N parity bits. After the encoding a total N information bits and $2N$ parity bits, that is a total of $3N$ bits are transmitted. On the decoder side an iterative algorithm is used. Generally, the maximum a posteriori probability (MAP) decoding method or soft-output Viterbi algorithm (SOVA) is used. In the first decoder, using this method the likelihood of different bits are computed and passed to the second decoder. The second decoder computes the likelihood ratios and passes them to the first decoder. This process is repeated until the likelihoods suggest a high probability of correct decoding for each bit. Then the final decision is made.

3.5.3.1 Coding Gain

Coding gain indicates how much better the user's decoded message performs as compared with the raw bit error performance of the coded transmission within the channel. Each error control code has a particular coding gain, which depends on the code used, decoder implementation, and channel

BER probability (P_{ch}). A good approximation for the decoded message error probability (P_B) is given by:

$$P_B \sim= (1/n) \cdot \Sigma_{i=k+1}{}^{n} i \binom{n}{i} P_{ch}^{i} (1-P_{ch})^{n-i}$$

where k denotes the number of errors that can be corrected in an (n,k) block code.

3.6 Automatic Repeat Request (ARQ) and Incremental Redundancy

So far, we have discussed different physical layer procedures for error free transmission, but in spite of this, errors can still happen and then it is up to the higher layer to decide what to do in this scenario, for example, whether to request for repetition or discard it (or automatically repeat the same information many times or add many error protection bits).

Automatic repeat request (ARQ) is an error detection mechanism used in the link layer (L2). Basically, the receiver informs the transmitter that a block has been incorrectly received and the transmitter resends it. This is not very efficient as it involves delay and throughput loss.

This can be done with a stop and wait (SAW) procedure, where the transmitter sends a block and waits for the receiver response before sending a new block or resending the incorrect one. This is not very efficient, as the transmitter is inactive until it gets a response. The solution used for HSDPA is N-channel SAW, which is a generalized version of the dual channel and can be used by multiple users.

Hybrid-ARQ-I is a combination of ARQ and forward error correction (FEC). In Hybrid-ARQ-II, the erroneous blocks are kept and are used for a combined detection with the retransmissions. This is of three types.

a. **Code Combining** – a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets, code selection, packet format, and reliability weighting of packets, and so on.
b. **Incremental Redundancy (IR)** – successive retransmissions of an erroneous block are sent with additional redundancy that is increased with each retransmission.
c. **Chase Combining** – with chase combining, the retransmitted data frames are identical with the original but when combined for detection, they are weighted with their SNR.

3.7 Interleaving

In the mobile environment the error in the transmitted bits tend to occur in bursts, as the signal enters into and goes out of deep fades and also because of the environmental impulse noise. The burst error is more severe, as it completely corrupts a part of the transmitted message. The convolutional error correcting codes are most effective, when the errors are randomly distributed throughout the bit stream. However, when errors are clustered in a specific area, then the convolution encoder will not be very effective. For this reason the coded data are interleaved before they are transmitted over the radio interface. This helps to convert the burst noise effect into a random noise effect by separating the consecutive message bits across the time scale. Hence the two consecutive message bits will not experience a similar type of channel fading effect. This implies that after the time separation the probability of both the original consecutive bits being corrupted will be less.

To understand the interleaving concept, we can take a simple example. Suppose we want to transmit a message- "MIKKI IS NOT WELL," and in this message some information such as "NOT WELL" is very important and if this part is corrupted during the transmission then it will have a major impact compared with other parts. However, if the deep fade condition occurs at the time when "NOT WELL" is passing through the channel then there is a high probability of corrupting this part of message completely and it will be very difficult to interpret the meaning of the message at the receiver end. One solution to this can be to rearrange the entire message in such a way that each different letter of the alphabet of the various words are separated by some distance, for example, arranged as MINW ISOE KTL KI. Then because of the impulse/burst noise, some portion such as ISOE may be corrupted, but as some letters of the different words arrive at the receiver without the effect of burst error, so the message recovery will be feasible through error correction. The channel characteristics generally change over a period of coherent time, so the separation of consecutive message bits in time should be more than the coherent time of the channel to achieve the maximum gain from of it.

An interleaver can be of two forms: a block interleaver or a convolutional interleaver. A bock interleaver formats the encoded data into a rectangular array of m rows and n columns, and interleaves nm bits at a time, for example, fills data row wise and sends the data column wise. The reverse operation happens on the receiver side. Convolutional interleavers can be used in place of block interleavers in the same way. These are ideally suited for use along with convolution codes.

3.8 Modulation

The original user information generated from a microphone or camera sensor is analog, whereas information from a computer or any digital equipment is a digital signal. These original signals, which are generated by the source (or the input transducer) or digital machine, are known as baseband signals.

A. **Baseband Communication** – Baseband signals can be sent without any carrier. In this instance, it is known as baseband communication. In baseband communication, the baseband signals have sizable power at low frequencies, so that they can be transmitted as they are, for example, without any carrier and without shifting the frequency range. They can be transmitted via pairs of cables, coaxial cables, but they can not be transmitted directly via radio links. Local telephone communication, pulse code modulation (PCM) is an example of baseband communication.
B. **Carrier Communication** – Transmitting and receiving the baseband signal without any carrier via a channel causes lot of problems. In the case of wireless communication, the analog and digital baseband signal (lower frequency, for example, larger wavelength signals) can not be sent directly, as they will require very large sized antenna (required antenna size $\propto$ wavelength) for signal transmission and reception. The second problem is that as the baseband signal from the various users lies in the same frequency band, so they will interfere with each other and distort the signals heavily. Thus there is a problem of channel multiplexing. In wire-line communication, the digital signal looks like a square wave pulse and cannot be sent as it is via the channel, as the data shape will be distorted because of the channels characteristics (capacitive, conductive, resistive transmission line load) and the bandwidth requirement will be too high. Therefore, we need to convert it into the analog domain for long distance transmission.

The purpose of digital modulation is to convert an information-bearing discrete-time symbol sequence into a continuous-time waveform (analog). This means transferring the digital domain information into a real world analog domain signal. A modem (modulator and demodulator) is used for this digital–-analog (or vice versa) conversion purpose.

In analog modulation, the low frequency analog signal is shifted from the low frequency domain to the high frequency domain by modulating it, using a carrier signal. The baseband signal (modulating signal) and the carrier signal forms a new signal pattern, and the resultant signal is called the modulated signal. Based on the nature of the baseband signal the modulation techniques may be of different types. If the baseband signal is analog then the analog modulation techniques are used and if the baseband signal is digital then digital modulation techniques are used.

3.8.1 Analog Modulation

In the case of analog modulation, the baseband signal or the modulating signal is a low frequency information signal, which is modulated by another high frequency carrier signal and the resultant signal is called a modulated signal. Analog modulation schemes are employed in the first generation of mobile radio systems. We will first discuss some of the analog and then the digital modulation techniques. There are three components of an RF carrier that can be changed during the modulation: the amplitude, the frequency, or the phase, and maybe a combination of amplitude, and frequency or phase. On this basis, the modulation type varies.

3.8.1.1 Amplitude Modulation (AM)

In amplitude modulation (AM), the carrier signal $A\cos(\omega t)$ has its amplitude "A" modulated in proportion to the baseband information signal $m(t)$ and provides a modulated output signal as $A[1 + m(t)]\cos\omega_c t$ as shown in Figure 3.45. The magnitude of $m(t)$ is chosen to be less than or equal to 1. The modulation index is then defined to be $\beta = \max_t m(t)$.

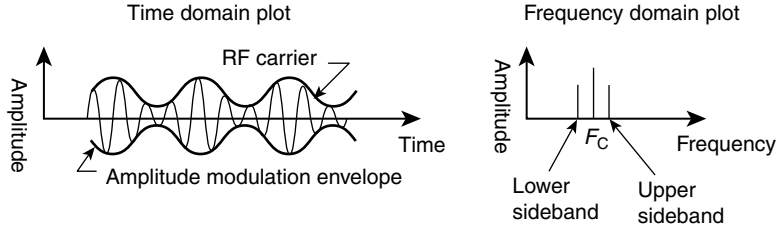


Figure 3.45 AM modulation with modulation index 0.2

The modulation index is often expressed as a percentage. The AM signal can be expressed as $A\ (1 + \beta\ \sin\omega_m t)\ \cos\ \omega_c t = A\ \cos\ \omega_c t + (A\ \beta/2)\ \{\cos\ [(\omega_c + \omega_m]\ t\} + \cos\ [(\omega_c - \omega_m)\ t)]$. The frequency components are the carrier (ω_c) and two sidebands ($\omega_c + \omega_m$ and $\omega_c - \omega_m$). As both the sidebands of an AM signal carry the same information, so one of these can be removed without losing any information. It is more efficient to transmit only one of the sidebands (the so-called single sideband AM, known as the USBAM (upper sideband AM) and the LSBAM (lower sideband AM), respectively). Two common techniques used for generating an SSB signal are the filter method and the balanced modulator method.

3.8.1.2 Frequency Modulation (FM)

In frequency modulation the amplitude is kept constant and the carrier signal's frequency is modulated by the amplitude of the modulating signal. The frequency of the carrier signal is varied linearly with the baseband message signal $m(t)$. If a_c is the amplitude of the carrier, f_c is the carrier frequency, and k_f is the frequency deviation constant, the modulating signal is a sinusoid of amplitude A_m and frequency f_m, then the FM signal can be represented as: $v = a_c \sin [\omega_c t + (k_f.A_m/f_m) \sin \omega_m t]$. This can be rewritten as a sum of components of constant frequency using the properties of the Bessel functions.

$$v = a_c\{J_o(m) \sin(\omega_c t) + J_1(m)[\sin(\omega_c + \omega_m)t - \sin(\omega_c - \omega_m)t] + J_2(\mathrm{m})[\sin(\omega_c + 2\omega_m)t + \sin(\omega_c - 2\omega_m)t] + J_3(m)[\sin(\omega_c + 3\omega_m)t - \sin(\omega_c - 3\omega_m)t] + \ldots$$

This expression implies that the FM spectrum consists of a component at ω_c and an infinite number of lines at $\omega_c \pm n\omega_m$ and that the amplitudes of the components are given by the Bessel functions. If we replace the carrier frequency term with a time-varying frequency and with Δf as the peak frequency deviation, then the carrier frequency term $f_c + (\Delta f/V_{mo}).V_m(t)$ now varies between the limits of $f_c - \Delta f$ and $f_c + \Delta f$. The modulation index for FM is defined as $\beta = \Delta f/f_m$, where f_m is the maximum modulating frequency used.

Performance of FM system: (1) Bandwidth – FM has a significantly larger bandwidth than AM. The bandwidth of an FM signal is: BW $\approx 2(\beta + 1)f_m$. (2) Efficiency – The efficiency of a signal is the power in the sidebands as a fraction of the total. In FM signals, because of the considerable sidebands produced, the efficiency is generally high. (3) Noise – FM systems are inherently immune to random noise and far better at rejecting noise than AM systems. The noise is generally spread uniformly across the spectrum. In the AM system the change in amplitude can actually modulate the signal and be picked up. As a result, AM systems are very sensitive to random noise.

3.8.2 Digital Modulation

As digital systems are becoming more common, so the use of digital modulation is also becoming very popular. The baseband or modulating signal is digital baseband data and modulated signal is an analog signal. The digital signals can be of various types, as discussed in Chapter 1.

3.8.2.1 Amplitude Shift Keying (ASK)

In ASK the amplitude of the carrier is changed according to the baseband data signal (1 or 0). Generally, if 0 needs to be transmitted then there is no signal (or zero signal amplitude) and a signal with a particular amplitude is sent when 1 needs to be transmitted (see Figure 3.46).

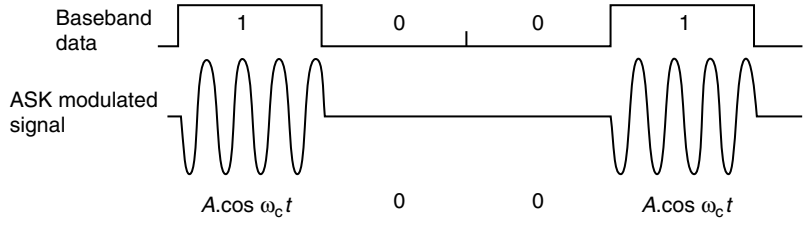


Figure 3.46 ASK signal

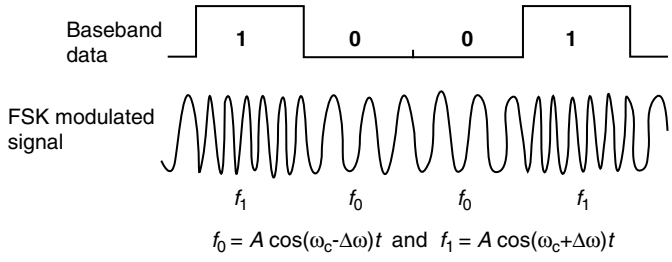


Figure 3.47 FSK signal

3.8.2.2 Frequency Shift Keying (FSK)

In frequency shift keying (FSK), the frequency of the carrier is changed according to the baseband data. One particular frequency is used for 1 and another for 0 as shown in the Figure 3.47.

$$FSK(t) = \begin{cases} \sin 2\pi f_1 t & \text{for bit 1} \\ \sin 2\pi f_2 t & \text{for bit 0} \end{cases}$$

Typically a binary FSK signal is implemented using a simple modulator and demodulator circuit (see Figure 3.48). The modulator is a voltage controlled oscillator (VCO) which is biased to produce the center frequency, when no modulation is applied. The demodulator is implemented as a phase locked loop, for example, a VCO, a phase detector, and a loop filter. The phase detector measures the difference in phase between the FSK signal and the VCO output. The loop filter necessarily slows the demodulator response to minimize the effects of noise.

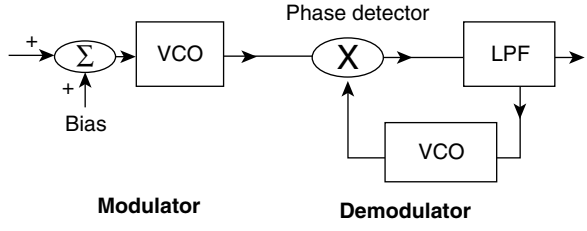


Figure 3.48 FSK modulator and demodulator

3.8.2.3 I/Q Format

Although it is common to describe the signal vector by its rectangular coordinates of I (in-phase) and Q (quadrature), polar notations are also the basis of many representations used in digital communications. In digital communications, modulation is often expressed in terms of I and Q. This is a rectangular representation of the polar diagram (see Figure 3.49). On a polar diagram, the I axis lies on the zero degree phase reference, and the Q axis is rotated by 90°. The signal vector's projection onto the I axis is its "I" component (in-phase) and the projection onto the Q axis is its "Q" component (quadrature phase).

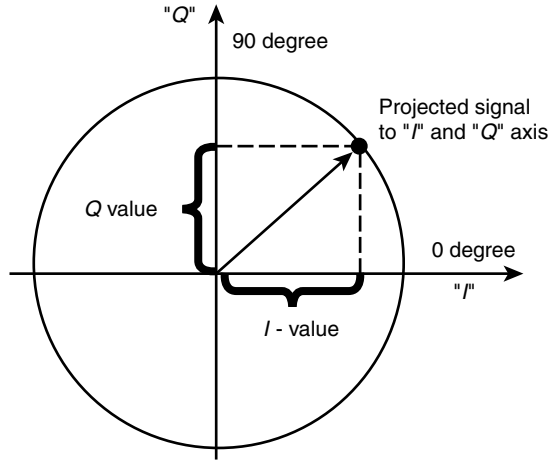


Figure 3.49 Polar to rectangular conversion

In polar coordinates:

$$A\,e^{j\omega t} = A(\cos\omega t + j\sin\omega t) = A(\cos\omega t + j\sin(\omega t + 90^\circ) = I + j\,Q$$

Digital modulation is easy to accomplish using the *I/Q* modulation technique. Most of the digital modulation techniques map the input data to a number of discrete points on the *I/Q* plane. These are known as constellation points. As the signal moves from one point to another, simultaneous amplitude and phase modulation usually results.

I and Q Modulation in a Radio Transmitter – In the transmitter, *I* and *Q* signals are mixed with the same local oscillator (LO) signal and a 90° phase shifter is placed in one of the LO paths. This creates two signals separated by 90° and are orthogonal to each other or in quadrature. As the signals are in quadrature, they do not interfere with each other when these signals are combined to a composite output signal (as the sin and cos signals are orthogonal). This simplifies the design of digital radios. The main advantage of *I/Q* modulation is the symmetric ease of combining independent signal components into a single composite signal and later splitting such a composite signal into its independent component parts.

I and Q Modulation in a Radio Receiver – The composite signal with magnitude and phase (or *I* and *Q*) information arrives at the receiver input (see Figure 3.50). This composite input signal (in terms of magnitude and phase) is thus broken into an in-phase (*I)* and a quadrature (*Q)* component by mixing with the local oscillator signal at the carrier frequency.

3.8.2.4 Phase Shift Keying (PSK)

In phase shift keying (PSK) the phase of the signal is varied when the input data bit changes. In Figure 3.51, the input data stream is 1001 and the signal is $a_0 \sin(\omega.t + \theta_0)$. When a bit changes from 1 to 0, then a phase value of θ is added to the signal phase, for example, the signal will become $a_0 \sin(\omega\, t + \theta_0 + \theta)$. Similarly, when the data changes from 0 to 1 a phase value of θ is subtracted, for example, the signal will become $a_0 \sin(\omega t + \theta_0 - \theta)$. Thus there will be an abrupt phase change at that instant. Based on the variation of the phase θ, the modulation level can be changed. The maximum value of θ is 360°. Hence, we can divide this maximum value of θ into 2, 4, 8, ..., sectors and change the modulation level.

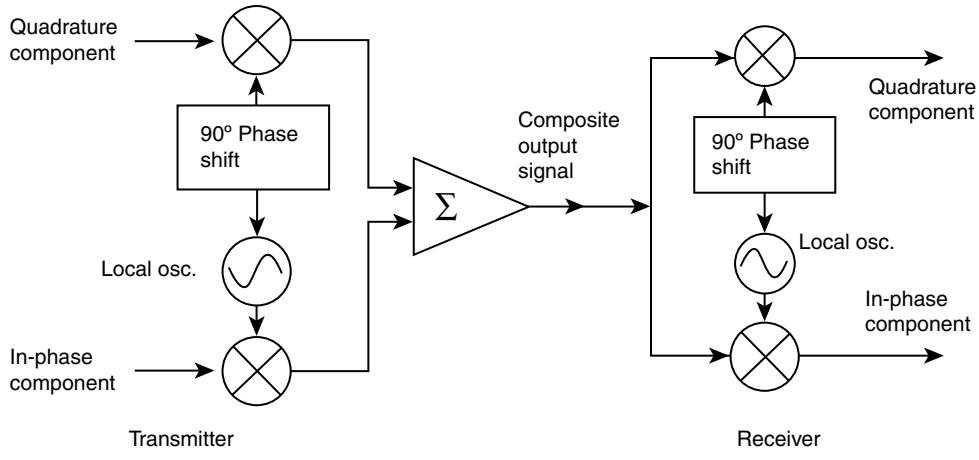


Figure 3.50 *IQ* signal modulation circuit and demodulation circuit

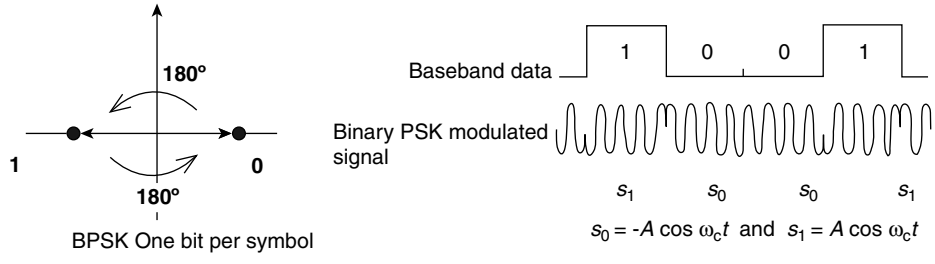


Figure 3.51 BPSK modulation

Binary Phase Shift Keying (BPSK)

If we divide max θ by 2, then $\theta = 360°/2 = 180°$. Now, if we plot these in a Cartesian graph as shown in Figure 3.51, there will be two points, which are separated by 180°. Using two points, only 1 bit value can be represented, either 1 or 0. So one point will be 0 and the other 1. Depending on the input data stream the phase will vary. For example, if the input data is 100011 then the phase variation will be 1 to 1 no change, and: $\theta = 0$, 1 to 0 $\theta = -180°$, 0 to 0 $\theta = 0$, 0 to 0 $\theta = 0$, 0 to 1 $\theta = 180°$.

The probability of bit error of a BPSK (binary phase shift keying) signal can be represented by the Q function: $Q\ [\sqrt{(2E_b/N_o)}]$, where E_b is the energy per bit.

Quadrature Phase Shift Keying (QPSK)

If we divide max θ by 4, then $\theta = 360°/4 = 90°$. Now, if we plot these in a Cartesian graph as shown in Figure 3.52, there will be four points, which are separated by 90°. Using four points, only 2 bit values can be represented: 11, 10, 01, and 00. Thus the points will be 11, 10, 01, 00. Here two bits will be taken at a time as input. These two bits together will form a symbol. Now break into even and odd symbols and pass the even one to the *I* and odd one to *Q* path. Depending on the input symbol stream the phase of the *I* and *Q*

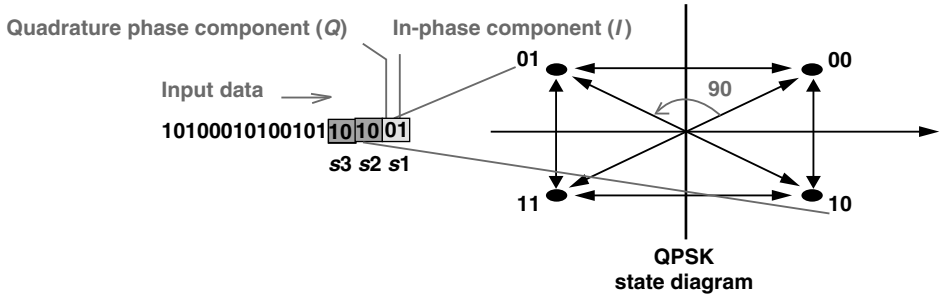


Figure 3.52 QPSK modulation

signals will vary. For example, the input binary bit stream $\{dk\}$, $dk = d0, d1, d2, \ldots = 1\mathbf{1000111}\ldots$ arrives at the modulator input at a rate $1/T$ bits/s, which is then separated into two data streams $d_I(t)$ and $d_Q(t)$ containing even and odd bits, respectively, for example, $d_I(t) = d0, d2, d4, \ldots = 1001$ and $d_Q(t) = d1, d3, d5, \ldots = \mathbf{1011}$, as if two separate pulse streams are modulated simultaneously (see Figure 3.53). The pulse stream $d_I(t)$ modulates the cosine function with an amplitude of ± 1, which is equivalent to shifting the phase of the cosine function by 0 or π; consequently this produces a BPSK waveform. Similarly, the pulse stream $d_Q(t)$ modulates the sine function, yielding a BPSK waveform orthogonal to the cosine function. The summation of these two orthogonal waveforms is the QPSK (quadrature phase shift keying) waveform. This is possible because the two signals I and Q are orthogonal to each other and can be transmitted without interfering with each other. This resultant modulated signal can be achieved by amplitude modulating two square data streams (odd and even in this instance) by following mathematical expression:

$$\begin{aligned} s(t) &= (1/\sqrt{2})\, dI(t) \cos(2\pi ft + \pi/4) + (1/\sqrt{2})\, dQ(t) \sin(2\pi ft + \pi/4) \\ &= A \cos[2\pi ft + \pi/4 + \theta(t)] \end{aligned}$$

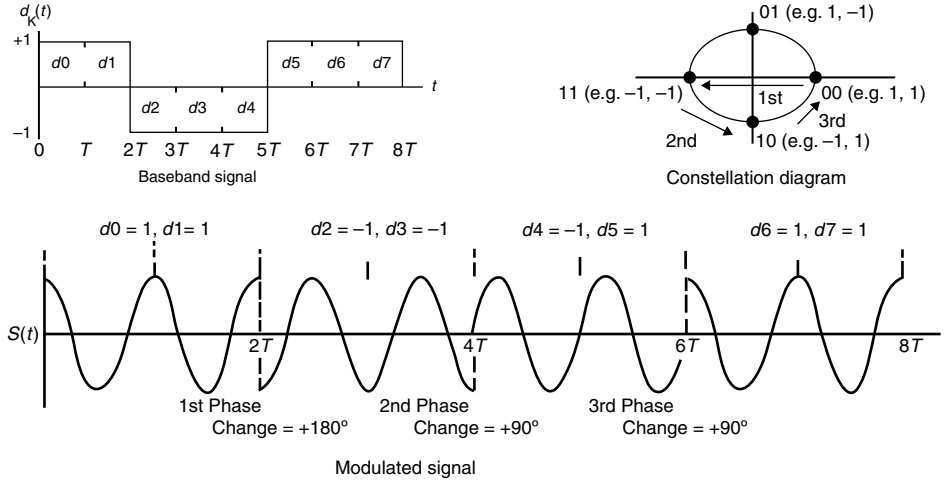


Figure 3.53 Baseband signal and QPSK modulated signal

The values of $\theta(t) = 0, -(\pi/2), \pi/2, \pi$ represent the four possible combinations of $a_I(t)$ and $a_Q(t)$. Each of the four possible phases of carriers represents two bits of data. Thus there are two bits per symbol. As the symbol rate for QPSK is half the bit rate, twice as much data can be carried in the same amount of channel bandwidth as compared with BPSK. In QPSK the carrier phase can change only once every $2T$ s. If from one T interval to the next one, neither bit stream changes sign, the carrier phase remains unchanged. If one component $a_I(t)$ or $a_Q(t)$ changes sign, a phase change of $\pi/2$ occurs. However, if both components change sign then a phase shift of π occurs.

The probability of bit error for a QPSK signal can be represented by the Q function as: $Q[\sqrt{(2E_b/N_o)}]$, where E_b is the energy per bit.

3.8.2.5 Minimum Shift Keying (MSK)

Minimum shift keying (MSK) modulation is derived from offsetting QPSK, where, instead of a rectangular pulse wave, we use a half-cycle sinusoidal pulse as shown in the Figure 3.54.

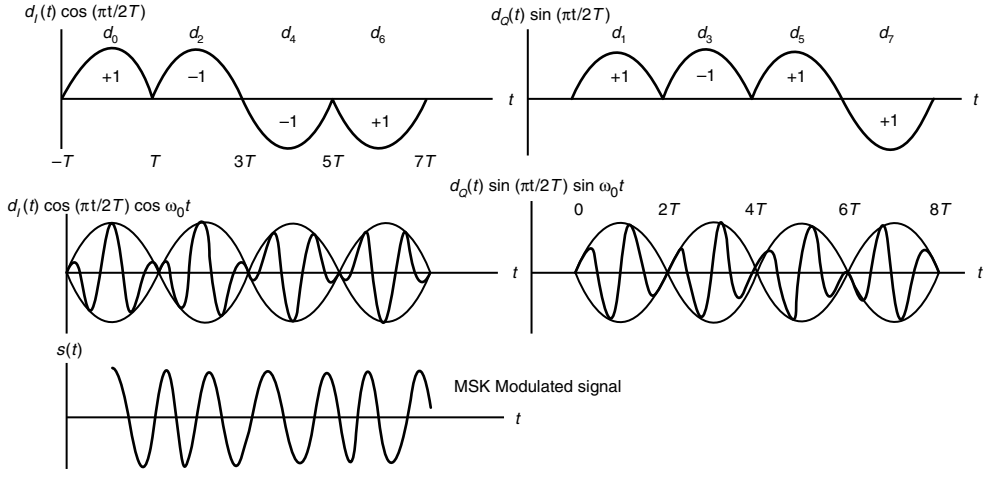


Figure 3.54 Replacing a rectangular pulse with half-cycle sinusoidal pulse for even sequences, replacing a rectangular pulse with half-cycle sinusoidal pulse for odd sequences, and MSK modulated signal

The MSK signal is defined as: $S(t) = d(t)\cos(\pi t/2T)\cos 2\pi ft + d(t)\sin(\pi t/2T)\sin 2\pi ft$. The phase modulation is improved by minimizing the amplitude fluctuations (number of types of phase). MSK is a continuous phase modulation scheme, where the modulated carrier contains no phase discontinuities and frequency changes occur at the carrier zero crossings. MSK is unique due to the relationship between the frequency of a logical zero and one: the difference between the frequency of a logical zero and a logical one is always equal to half the data rate. In other words, the modulation index is 0.5 for MSK, and is defined as:

$$m = \Delta f \cdot T$$

where

$$\Delta f = |f_{\text{logic 1}} - f_{\text{logic 0}}|$$
$$T = 1/\text{bit rate}$$

MSK has several advantages over other modulation schemes but the disadvantage of MSK is that its power spectrum density does not fall fast enough to completely reduce interference between adjacent signals. Therefore the spectrum is not compact enough to realize data rates approaching the RF channel BW. For wireless data transmission systems, which require more efficient use of the RF channel BW, it is necessary to reduce the energy of the MSK upper side lobes. A straightforward means of reducing this energy is low-pass filtering the data stream prior to presenting it to the modulator (pre-modulation filtering). The pre-modulation low pass filter must have a narrow BW with a sharp cut-off frequency and very little overshoot in its impulse response. This is where the Gaussian filter characteristic comes in. It has an impulse response characterized by a classical Gaussian distribution. A Gaussian-shaped impulse response filter generates a signal with low side lobes and a narrower main lobe than the rectangular pulse. This type of filtering has a delayed and shaped impulse response that has a Gaussian-like shape. This modulation is called Gaussian minimum shift keying (GMSK).

3.8.2.6 Gaussian Minimum Shift Keying (GMSK)

Figure 3.55 depicts the impulse response of a Gaussian filter for BT = 0.3 and 0.5. BT is related to the filter's −3 dB BW and data rate by BT = $f_{-3\ \text{dB}}$/bit rate.

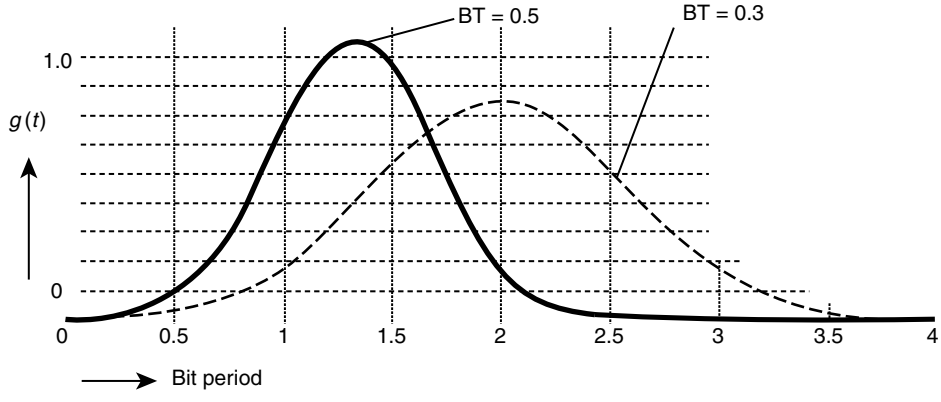


Figure 3.55 Gausssian filter impulse response for BT = 0.3 and BT = 0.5

The reduced side lobe energy for GMSK and a bit is spread over approximately 3 bit periods for BT = 0.3 (and two bit periods for BT = 0.5), lead to inter symbol interference (ISI), which is inherently introduced here in order to make the spectrum more compact. Thus the channel spacing may be tighter for GMSK compared with MSK for the same adjacent channel interference.

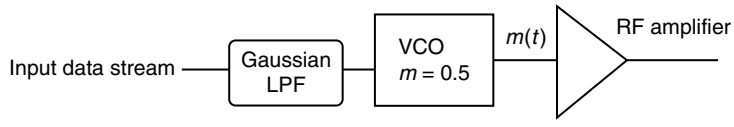


Figure 3.56 GMSK VCO-modulator

GMSK Modulation Implementation

There are two commonly used methods to generate GMSK, one is frequency shift keyed modulation and the other is quadrature phase shift keyed modulation. The first method as shown in Figure 3.56, based on GMSK VCO-modulator architecture, but this is not suitable for coherent demodulation due to component tolerance problems A block diagram of the modulator based on the second methodis shown in Figure 3.57. The steps followed in the modulator are as: (a) create the NRZ (−1,1) sequence from the binary (0,1) input sequence; (b) create N samples per symbols; (c) integrate the NRZ sequence; and (d) convolute with a Gaussian function then compute the corresponding I and Q components (at this stage, we have the quadrature components of the baseband GMSK equivalent signal); (e) multiply the I and Q components by the corresponding cos$n\omega t$ and sin $n\omega t$ carriers, and (f) then add the two signals.

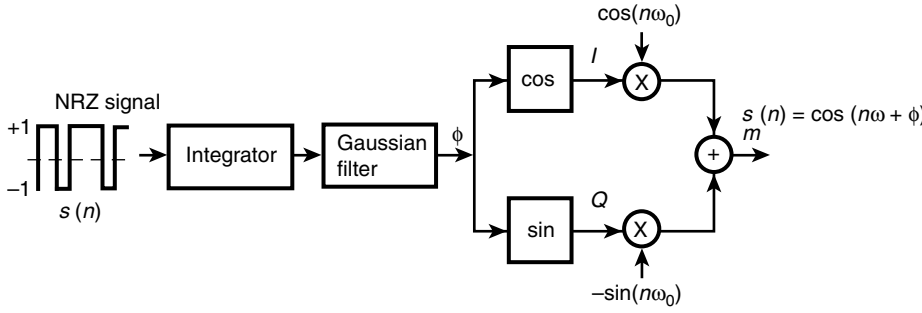


Figure 3.57 GMSK modulation block diagram

GSM uses GMSK with a modulation index $h = 0.5$, BT equal to 0.3 and a modulation rate of 271 (270 5/6) kbauds. This means B = 81.3 kHz when T is about 3.7 μs. The GMSK modulation is chosen as a compromise between spectrum efficiency and a reasonable demodulation complexity (see Figure 3.58). The constant envelope allows the use of simple power amplifiers and the low out-of-band radiation minimizes the effect of adjacent channel interference.

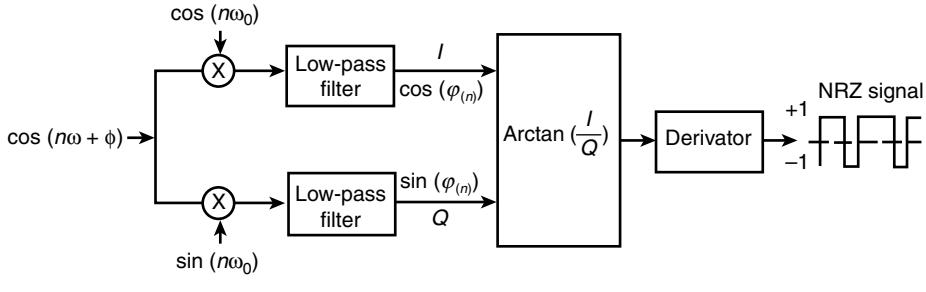


Figure 3.58 GMSK demodulation blocks

In GSM, differential encoding combined with 90° rotation converts MSK into BPSK type of modulation. This rotation helps to avoid zero crossing, during the switch from one constellation point to the other, which in turn helps to reduce the distortion in RF amplification. Thus, for GMSK $\pi/2$ and in 8-PSK $3\pi/8$, rotation is defined in GSM and EGDE.

3.9 Bit Rate, Baud Rate, and Symbol Rate

To understand and compare different modulation format efficiencies, it is important to first understand the difference between bit rate and symbol rate. The signal bandwidth for the communications channel required depends on the symbol rate, not on the bit rate.

- **Bit rate** is the rate at which information is passed.
- **Baud rate** (or signaling rate) defines the number of symbols per second.
- **Symbol rate** = bit rate/the number of bits transmitted with each symbol.

As the symbol rate is the bit rate divided by the number of bits that can be transmitted with each symbol, so if one bit is transmitted per symbol, as with BPSK, then the symbol rate would be the same as the bit rate. If two bits are transmitted per symbol, as in QPSK, then the symbol rate would be half of the bit rate. The symbol rate is also sometimes called the baud rate.

The baud (or signaling) rate defines the number of symbols per second. Each symbol represents n bits, and has M signal states, where $M = 2^n$. This is called M-ary signaling.

Different methods of modulation provide different data rates as shown in Table 3.2.

Table 3.2 Theoretical bandwidth efficiency limits for different modulations

Modulation	Theoretical bandwidth efficiency limits
MSK	1 bit/s/Hz
GMSK	1 bit/s/Hz
BPSK	1 bit/s/Hz
QPSK	2 bits/s/Hz
8 PSK	3 bits/s/Hz
16 QAM	4 bits/s/Hz
32 QAM	5 bits/s/Hz
64 QAM	6 bits/s/Hz
256 QAM	8 bits/s/Hz

Several parameters are used to compare the efficiency of the various modulation techniques. These are described below.

1. Power efficiency is a measure of how much the signal power should be increased to achieve a particular BER for a given modulation scheme. Signal energy per bit/noise power spectral density = E_b/N_0.
2. Bandwidth efficiency is measured as data rate per hertz (rate/bandwidth – bits/s/Hz). This is also known as spectral efficiency.
3. Channel capacity per bandwidth.

Table 3.3 Modulation techniques for different applications

Modulation	Applications
MSK, GMSK	GSM, CDPD
BPSK	Cable modems, deep space telemetry
QPSK, PI/4 DQPSK	CDMA, NADC, TETRA, PHS, PDC, LMDS, DVB-S, cable modems
OQPSK	DECT, paging, AMPS, land mobile, CT2
VSB	ATV, broadcast, cable
8PSK	Satellite, aircraft, broadcast video systems
16 QAM	Microwave digital radio, modems, DVB-C, DVB-T
64 QAM	DVB-C, modems, broadband set top boxes
256 QAM	Modems, DVB-C, digital video

In the case of coherent detection, the received signal is processed with a local carrier of the same frequency and phase. However, for non-coherent detection it requires no reference values. For example, PSK is a coherent modulation whereas ASK is non-coherent. A variety of types of modulation are used for different wireless systems based on the system requirements, data rate requirement, and cost. These are shown in Table 3.3.

3.10 Inband Signaling

An inband signaling modem communicates digital data over a voice channel of a wireless telecommunications network. As an input it receives the digital data. An encoder converts the digital data into audio tones, which synthesize the frequency characteristics of human speech. The digital data is also encoded to prevent voice encoding circuitry in the telecommunications network from corrupting the synthesized audio tones representing the digital data. It then outputs the synthesized audio tones to a voice channel of a digital wireless telecommunications network.

Further Reading

Doelz, M.L. and Heald, E.H. (1961) Collins Radio Co., "Minimum-shift data communication system," US Patent 2 977 417, Mar. 28.

Forney, G.D. Jr. (1973) The Viterbi Algorithm. *Proceedings of the IEEE*, **61** (3), 268–278.

Haykin, S. (1996) *Adaptive Filter Theory*, Prentice-Hall, Englewood Cliffs, NJ.

Haykin, S. (2001) *Communication Systems*, John Wiley & Sons, Hoboken.

Jeruchim, M.C., Balaban, P., and Shanmugan, K.S. (1992) *Simulation of Communication Systems*, Plenum Press, New York, p. 731.

Kay, S.M. (1998) *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Englewood, NJ, p. 595.

Laster, J.D. (March 1997) *Robust GMSK Demodulation using Demodulator Diversity and BER Estimation*, PhD thesis. Blacksburg, VA.

Murota, K. and Hirade, K. (1981) GMSK Modulation for Digital Mobile Radio Telephony. *IEEE Transactions on Communications*, **com-29** (7), 1044–1050.

Nefedov, N. and Pukkila, M. (2000) Turbo equalization and iterative (Turbo) estimation techniques for packet data transmission. Second International Symposium on Turbo Codes, Brest, France, 4–7 September, 2000.

Price, R. and Green, P.E. (1958) A communication technique for multipath channel. Paper presented at the Proceedings of the IRE, 555–570.

Proakis, J.G. (1995) *Digital Communications*, 3rd edn, McGraw-Hill, New York, p. 929.
Proakis, J.G. and Salehi, M. (2004) *Fundamentals of Communication Systems*, Pearson Education, Upper Saddle River, NJ.
Pukkila, M. (2000) Channel Estimation Modeling, Master Thesis, HUT, Nokia Research Center.
Qureshi, S. (1985) Adaptive Equalization. *Proceedings of the IEEE*, **73**, 1349–1387.
Rappaport, T.S. (2001) *Wireless Communications: Principles and Practice*, Prentice-Hall, Englewood Cliffs, NJ.

4

Mobile RF Transmitter and Receiver Design Solutions

4.1 Introduction to RF Transceiver

In Section 1.3.3.1 of Chapter 1, we discussed the basic building blocks of a mobile phone. For a digital mobile phone, the modem part can be broadly divided into two main blocks: (a) the RF module and (b) the baseband module. The RF module is the analog front-end module. On the receiver side it is responsible for signal reception from the air and down conversion onto the baseband signal, and on the transmitter side it is responsible for up conversion of the baseband signal to the RF signal, then transmission in the air. The baseband module deals with digital processing of the baseband signal and protocols. An ADC and DAC are placed in between the analog RF and digital baseband units. In this chapter, we will discuss the different design solutions for analog RF front-end modules. The basic building blocks of a typical radio RF front-end section of any mobile device (using digital baseband) is shown in the Figure 4.1.

1. **Antenna** – This acts as a transducer. It converts the incoming EM wave into an electrical signal on the receiver side and the outgoing electrical signal into an EM wave on the transmitter side. The basic principle of antenna action has already been discussed in Chapter 1, and different types of mobile phone antennas are discussed in detail in Chapter 10.
2. **Tx–Rx Path Separation Block (Duplexer or Tx–Rx Switch or Diplexer)** – Generally the same antenna is used for transmission as well reception purposes. So we have to have a mechanism to multiplex the same antenna between the transmit and receive path. There are several techniques available:
   a. **Tx–Rx Switch** – Here the same antenna is time switched between the Rx and Tx paths (Figure 4.2). Systems where the Tx and Rx path signals are not present simultaneously (for example, a half duplex system GSM), then a Tx–Rx switch can be used. Diodes can be used as switching elements and switching is controlled by the processor to connect the Tx or Rx path with the antenna. Single frequency can also be used for Tx and Rx, which is suitable for a time division based system.
   b. **Diplexer** – The Tx and Rx frequency bands are different and they are first separated by filters and then connected to the Rx or Tx path (Figure 4.3). It only works for systems with separated Rx and Tx frequency bands (FDD).

Mobile Handset Design Sajal Kumar Das

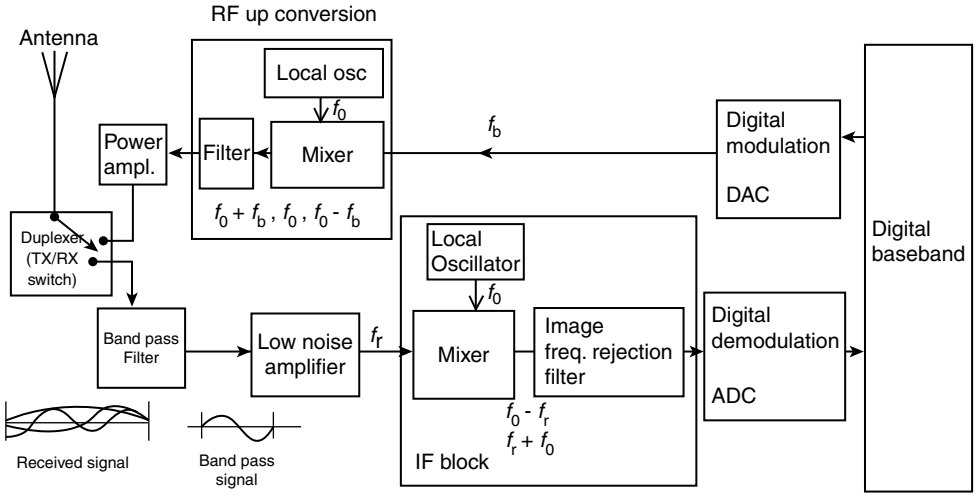


Figure 4.1 Various blocks of an RF transceiver (analog front end)

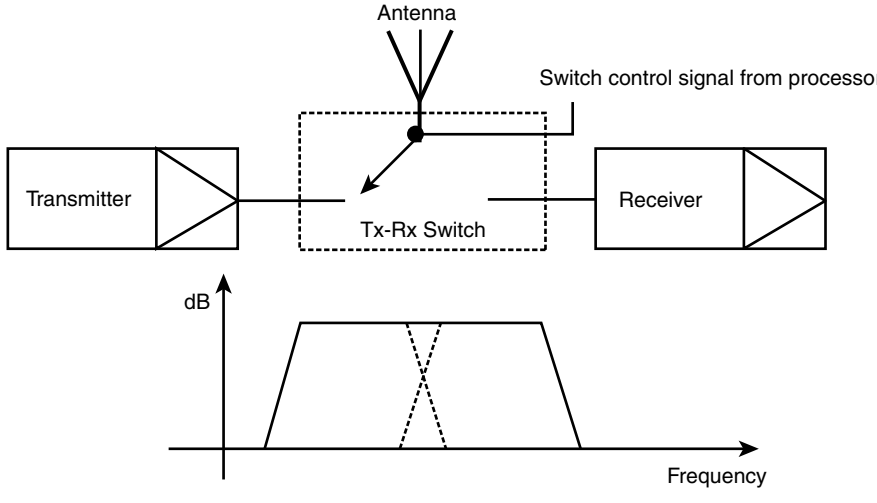


Figure 4.2 Antenna multiplexing by using a Tx–Rx switch

c. **Duplexer** – The Tx circuit, Rx circuit, and antenna are connected to the three port duplexer (Figure 4.4). Tx and Rx are separated by a path difference of $\lambda/2 =$ phase difference of "π", for example, they are opposite in phase ($-$ve) and will cancel each other out. Thus the Tx port is isolated from the Rx port, but the signal from Tx and Rx will arrive at the antenna at same phase ($\pi/2$) as the path difference between Tx or Rx with an antenna port is $\lambda/4$.

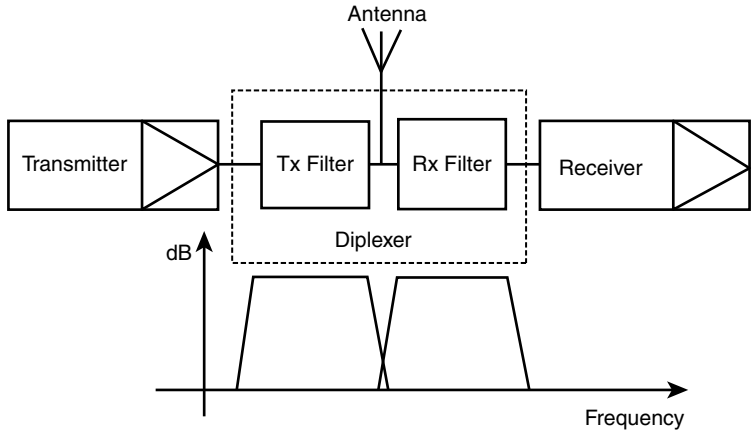


Figure 4.3 Diplexer

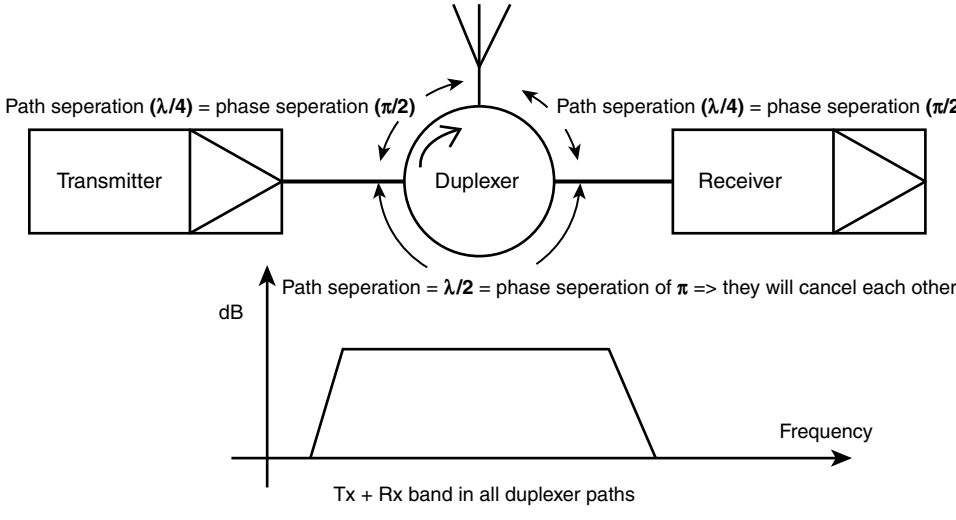


Figure 4.4 Duplexer

A comparison of these three techniques for antenna multiplexing between Tx and Rx path is given in the Table 4.1.

Electrically a duplexer is a device that uses sharply tuned resonate circuits to isolate the transmitter circuit from the receiver circuit. This allows both of them to use the same antenna at the same time without the transmitter RF frying the receiver circuit. The separation or isolation between the transmit and receive signal paths is mandatory in order to avoid any destruction of the receiver when the Tx signal is injected, or at least to avoid any degradation of the receiver sensitivity due to the frequency proximity of the high power signal from the transmitter block.

Table 4.1 Comparison between different antenna multiplexing techniques

Duplexer	Diplexer	Tx-Rx switch
This is a passive device, so no power supply is needed	This is a passive device, so no power supply is needed	Power supply may be required
Power handling capability is less	Power handling capability medium	Power handling capability is good but it is switch dependent
High isolation, low insertion loss is critical for it's operation (typical isolation: 20 dB)	Isolation is dependent on the filter performance	Isolation and insertion loss is not very critical
Size can be smaller than diplexer	Filters are normally bulky	This is space saving design
Permanent VSWR matching	Lowest third inter-modulation products	Less problem
As isolation is dependent on wavelength, so it offers narrower frequency bands	Narrower frequency band (as filters are designed accordingly)	Wider frequency bands
Tx and Rx can be of same frequency, and they can be transmitted simultaneously, can be used for both TDD and FDD system; good for multi-mode receiver	Tx and Rx should be different frequency, can not be used for FDD system	Tx and Rx should not be operational at the same instant; can be used for TDD system

In the cellular band, the duplexer's jobs are: (1) to isolate the transmitted signal from the received signal in the receive band to avoid any degradation of the receiver sensitivity; (2) attenuating the power amplifier (PA) output signal to avoid driving the low-noise amplifier (LNA) into compression; (3) attenuating the receiver's spurious responses (first image and others); (4) attenuating first local oscillator (LO) feed-through using the first mixer LO-RF ports; and (5) attenuating transmitter output harmonics and other undesired spurious products.

3. **Band-Pass Filter** – This is used to extract the desired band of signal from the entire band of the received signal. Whatever EM waves impinge on the antenna, based on their length (∼wavelength) the antenna will convert those waves into RF electrical signals. In the reception path there will be many such RF signals with different frequencies, which will be mixed up and appear in the receiver circuit. Out of these, we need to take only the desired frequency band by using an appropriate band-pass filter.
4. **Low-Noise Amplifier** – The amplitude of the received signal will be much less. So we need to boost this received feeble signal without adding any extra noise signal into it, for example, this stage of amplification should have a very low level of noise. This is why we use a low-noise amplifier (LNA).
5. **Mixer** – Mixers are key components in receivers as well as in transmitters. Mixers translate the signals from one frequency band into another, by mixing frequencies. The output of the mixer consists of multiple images of the input signals to the mixer, where each image is shifted up or down by multiplication of the local oscillator (LO) frequency. The mixer's output signals are usually the signals translated up and down by one LO frequency. Generally mixers (sometimes known as frequency converters), modulators, and balanced modulator circuits work on the same basic principle.

   This module contains local oscillators to generate RF signal locally and mixers to mix the incoming and locally generated LO signal in order to down convert the incoming RF signal.

Suppose the received RF signal is $A_r \sin\omega_r t$ and the local oscillator signal is $A_o \sin\omega_o t$. If these are mixed (that is, multiplied) the resultant signal will be $A_r \sin\omega_r t$. $A_o \sin\omega_o t = (A_r A_o/2) \cos 2\pi (f_o - f_r) t + \cos 2\pi (f_o + f_r) t$. That means it generates two frequencies $(f_o - f_r)$ and $(f_o + f_r)$. This signal is passed through a channel select filter, which will filter out the frequency $(f_o + f_r)$, so that only $(f_o - f_r)$, known as the intermediate frequency (IF) will be passed forward. In some receivers, two or more such IF stages are used. Thus the RF signal is down converted in several steps and finally it arrives at the last stage, where the $(f_o - f_r)$ becomes equal to the baseband signal frequency $f_{baseband}$. This is then sampled at a minimum rate of $2 * f_{baseband}$ = Nyquist rate, to recover the baseband data signal and then digitally demodulated. Similarly, the mixer is used for frequency up conversion in the Tx path.

In an ideal situation, the mixer output would be an exact replica of the input signal. However, the reality is that the mixer output is distorted due to non-linearity in the mixer. In addition, the mixer components and a non-ideal LO signal introduce more noise in the output. Bad design might also cause leakage effects, complicating the design of the complete system. The mixer design is discussed in more detail in Section 4.2.

6. **Digital Demodulation** – For the receiving system, which uses digital technology, the received analog signal is digitally demodulated. Various techniques are used for digital modulation–demodulation, such as ASK, FSK, PSK, QPSK, MSK, GMSK, and QAM. Modulation/demodulation has already been discussed further in Chapter 3.
7. **A/D Converter** – The received signal is sampled and converted into a digital signal. There are different types of ADC used for this purpose. Among these, sigma delta is very popular, which is discussed in Chapter 10.
8. **Digital Modulation** – The transmitted digital data from the baseband is digitally modulated using different digital modulation techniques.
9. **D/A Converter** – This converts the baseband digital signal into a baseband analog signal.
10. **Power Amplifier** – The feeble RF signal is amplified to high power by a power amplifier,sent to a duplexer unit, and then to an antenna.

4.2 Mixer Implementations

The mixer is implemented using various properties, and below some commonly used mixer implementations are discussed.

Ideal Mixer – The ideal mixer will perform the mathematical multiplication of the two input signals, creating components positioned at frequencies equal to the sum and difference of the input signals with no additional components. This is why the mixing (multiplying) device must be perfectly linear and there must be no leakage of the input signals to the output port.

Single Balanced Mixer and Double Balanced Mixer – The term balanced mixer is used to imply that neither of the input terms will appear at the mixer output. However, in practice, suppression of these input components is never perfect in an analog mixer circuit. A balanced mixer can be implemented using a transformer coupled diode arrangement, or by using an active transistor based design. Both types of mixer produce signals at odd harmonics of the carrier frequency, particularly the diode ring mixer. In most instances, these can be easily filtered out.

The single balanced mixer (SBM) performs multiplicative mixing because its RF and LO signals are applied to different ports. This is more commonly seen in two diode mixer configurations, a balanced transformer drives the diodes out of phase for the LO and in phase for signals present at the RF port. Adding two more diodes and another transformer to the singly balanced mixer results in a double balanced mixer (DBM), as shown in Figure 4.5. Its frequency response is largely determined by the frequency response of its transformers. As second-order harmonics are the most difficult to suppress, so double balanced mixers are the favored solution.

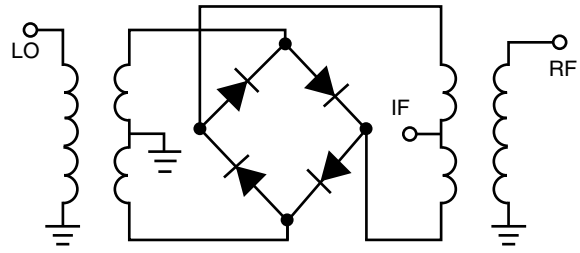


Figure 4.5 Double-balanced mixer

4.2.1 Design Parameters

The mixer performance is dependent on parameters such as conversion loss, isolation, dynamic range, dc offset, dc polarity, two-tone third-order inter-modulation distortion, and intercept point. Although a mixer works by means of amplitude non-linear behavior, we generally want it to act as a linear frequency shifter. The degree to which the frequency-shifted signal is attenuated or amplified is an important parameter in mixer design. A mixer also contributes noise to the output frequency shifted signals. The degree to which a mixer's noise degrades the SNR of the signals is evaluated in terms of noise factor and noise figure. The load presented by a mixer's ports to the outside world can be of critical importance to a designer for VSWR matching. Isolation between ports plays a major role in reducing dc offset in a mixer. The dynamic range of any RF/wireless system can be defined as the difference between the 1 dB compression point and the minimum discernible signal.

4.2.1.1 Inter-Modulation Distortion and Intercept Points

If there is a large interfering signal presents within the bandwidth of the RF input filter, then mixer distortion heavily limits the sensitivity of a receiver. There are two aspects of distortion that are of concern: (1) compression, and (2) inter-modulation distortion. The 1 dB compression point (CP1) is the point where the output power of the fundamental crosses the line that represents the output power extrapolated from small-signal conditions minus 1 dB. The third-order intercept point (IP3) is the point where the third-order term, as extrapolated from small-signal conditions, crosses the extrapolated power of the fundamental. Both CP1 and IP3 illustrated in Figure 4.6.

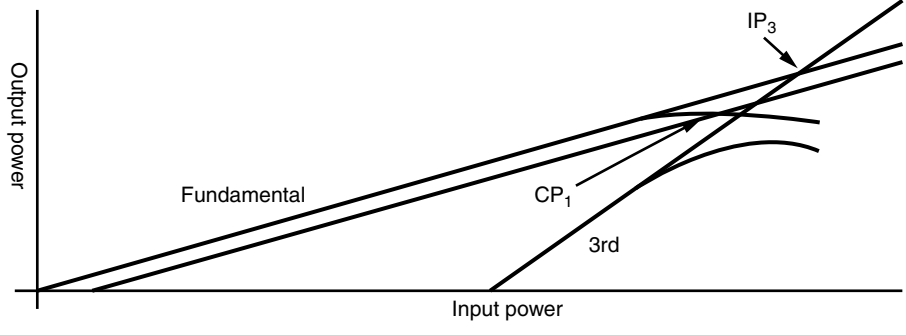


Figure 4.6 Intercept points

Distortion of the output signal occurs, because several of the odd-order inter-modulation tones fall within the bandwidth of the circuit. Inter-modulation distortion is typically measured in the form of an intercept point. As shown in Figure 4.6, the third-order intercept point (IP3) is determined by plotting the power of the fundamental and the third-order inter-modulation product versus the input power.

4.2.1.2 Basic Difference Between Mixer and an Amplitude Modulator

The mixer and AM modulator work in the same fashion and generate three output signal frequencies. The only difference is that in a mixer the two different signals are multiplied ($A_r \sin\omega_r t\, A_o \sin\omega_o t$), whereas in an AM modulator the amplitude of the RF carrier signal is varied according to the input signal: $v(t) = A(1 + m \sin\omega_m t) \sin\omega_c t$.

4.3 Receiver Front-End Architecture

In the previous section, we saw the various blocks of a radio receiver. The receiver architecture mainly varies based on the RF front end.

4.3.1 Different Types of RF Down Conversion Techniques

Typically, radio communication systems operate with carrier frequencies at many hundreds of MHz to several GHz. If we want to sample the received signal at the antenna itself (this will help to bypass RF hardware blocks and will help to process everything in the digital domain, which is most desirable for software defined radio), then the minimum sampling frequency requirement will be 2^*fs ($\sim 2^*1$ GHz, for example, 2 giga samples will be generated per second and then each sampled level will be converted into several bit streams by the ADC), and the sampled data will be too huge to handle by any of the present day's DSP (digital signal processor). Also, directly converting the antenna signals into digital form in an integrated ADC would require prohibitively large sensitivity, selectivity, linearity, and very high conversion speed. As of today, such analog-to-digital converters that could offer this services do not exist. As a result, the received RF signals need to be converted into lower frequencies (baseband frequency), for the signal processing steps, such as channel selections, amplification, and detection. This conversion is accomplished by a mixing process, producing a down-converted (in the receiver block) and an up-converted (used in the transmitter block) component.

Now, based on the mixing of the local oscillator (LO) frequency with the desired incoming RF frequency, several down-conversion techniques exist. We can mainly classify these into two broad categories: heterodyne and homodyne receivers. (1) In a heterodyne receiver the LO frequency and desired RF frequency is set to be different –examples of such architecture are super heterodyne, low IF, and wide IF receivers. (2) Whereas in the case of a homodyne receiver (same mixing) the LO frequency and the desired RF frequency are set to be the same, so the IF (intermediate frequency) is zero. Prior to the selection of optimum receiver architecture, different types of RF down-conversion receiver architectures will be reviewed and compared.

4.3.1.1 Heterodyne Receiver

Conventional radio receivers utilize the so called heterodyne architecture (hetero = different, dyne = mix). This architecture, translates the desired RF frequency into one or more intermediate frequencies, before demodulation.

What is Heterodyning? – "Heterodyne" means mixing two different frequencies together (one incoming signal frequency from the antenna and other locally generated from the local oscillator), to produce a beat frequency, namely the difference between the two and the sum of the two. Example: $A_r \sin \omega_r t\ A_o \sin \omega_o t = (A_r A_o/2) \cos 2\pi (f_o - f_r)\ t + \cos 2\pi (f_o + f_r)\ t$.

What is Superheterodyning? – When we use only the lower sideband (the difference between the two frequencies), we are superheterodyning. Strictly speaking, the term superheterodyne refers to creating a beat frequency that is lower than the original signal. Edwin Armstrong came up with the idea of converting all incoming frequencies into a common frequency. The superheterodyne receiver, invented in 1917, has enjoyed a long run of popularity.

We have discussed that superheterodyning is simply reducing the incoming signal frequency by mixing. In a radio application (Figure 4.7), we are reducing the incoming AM or FM signal frequency (which is transmitted on the carrier frequency) to some intermediate frequency, called the IF (intermediate frequency) $= f_o - f_r$.

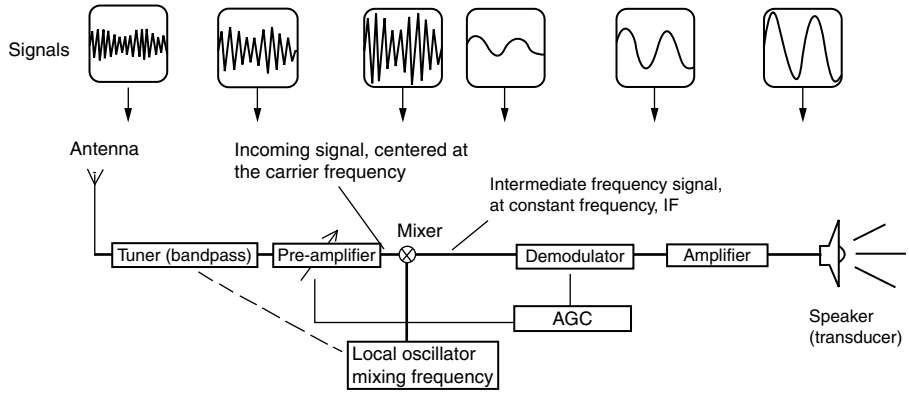


Figure 4.7 Basic block diagram of a super-heterodyne receiver (analog receiver)

This is essentially the conventional receiver with the addition of a mixer and a local oscillator. The local oscillator is linked to the tuner, because they must both vary with the carrier frequency. Let us look at a specific example. An FM radio is tuned to a station operating at 89.9 MHz. This signal is mixed with an LO signal at a frequency of 100.6 MHz. The difference frequency at the output of the mixer is 10.7 MHz. This is the IF signal. If the FM radio is tuned to a different station at 107.9 MHz, the LO frequency is also re-tuned to 118.6 MHz. The mixer once again produces an IF signal of 10.7 MHz. In fact, as the FM radio is tuned across the band from 87.9 to 107.9 MHz, the local oscillator (LO) is tuned from 98.6 10 118.6 MHz. No matter what frequency the radio is tuned to (in its operating range), the mixer's output will be 10.7 MHz.

The superheterodyne overcomes the variable sensitivity and selectivity problems of the RF receiver module by doing most of the amplification at the intermediate frequency, where the gain and selectivity can be controlled carefully (Figure 4.8). However, the superheterodyne introduces some new challenges. Firstly, the LO signal must always differ from the input signal by exactly the IF frequency, regardless of whatever input frequency is selected. This is known as "tracking". Secondly, there are two different frequencies that can mix with the LO signal to produce the IF signal. One of these frequencies is our input signal frequency; the other is known as the "image frequency." The image, input, IF, and LO frequencies are related as follows.

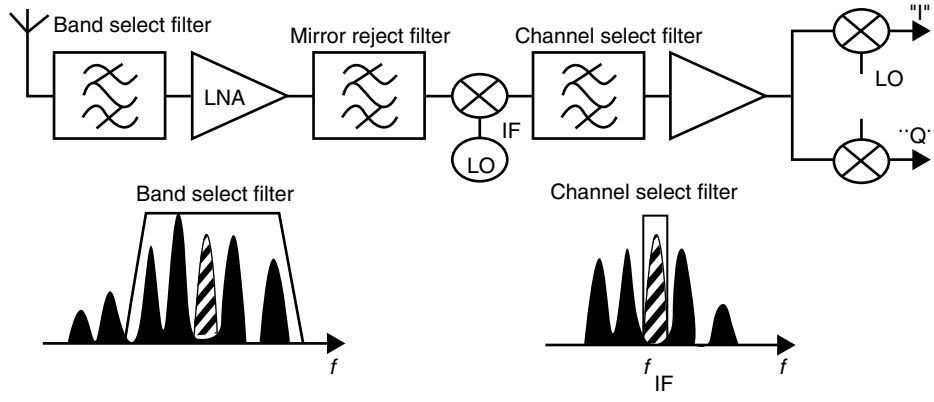


Figure 4.8 Super-heterodyne receiver architecture (digital receiver: I–Q modulated)

The image is another incoming frequency that is mistakenly treated as the desired input signal.

$$\text{IF} = \text{LO} - \text{input}: \ \text{IF} = \text{image} - \text{LO}; \ \text{for example, image} = \text{IF} + \text{LO} = \text{IF} + \text{IF} + \text{input}.$$

$$\text{Thus, image} = \text{input} + 2^{*}\text{IF}$$

Let us take an example that references our earlier discussion about the FM radio. When the receiver is tuned to 89.9 MHz, the 89.9 MHz signal can mix with the LO signal of 100.6 MHz to create a 10.7 MHz IF signal. However, a signal (or noise) at 111.3 MHz can also mix with the LO signal to create a 10.7 MHz IF. Therefore, any incoming noise or interference signals at this frequency have to be rejected. One filter is used to stop ($f_0 + f_i$) and pass ($f_0 - f_i$) to the demodulator unit. $f_{if} = f_0 - f_i$, which is allowed to pass, but there may be another frequency that is ($f_0 + f_{if}$) when this is input to the mixer, then the resultant will be ($f_0 + f_{if}$) $- f_0 = f_{if}$. Thus this will also pass via the filter. However here, IF is not because of the desired input signal (f_i) rather it is another input signal ($f_0 + f_{if}$). This is known as image frequency as discussed earlier ($f_0 + f_{if}$) $= f_{image}$. This also needs to be rejected as it causes the following issues.

Image Frequency Problem

The desired RF input signal frequency $= f_{if} = f_0 - f_i$ (which is allowed to pass via the filter). However, there is another RF input signal frequency $f_0 + f_{if} = f_{image}$, and when this is fed to the mixer, the resultant frequency $= f_0 - f_{image} = f_0 - f_0 + f_{if} = f_{if}$. As this resultant frequency is same as the IF frequency, so this will also pass via the filter (Figure 4.9). This is derived from unwanted RF signal, and is known as the

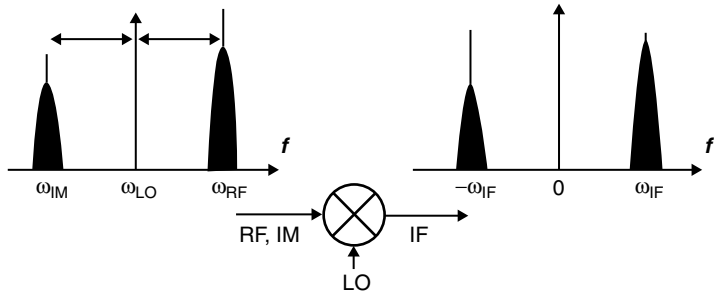


Figure 4.9 Image frequency

image frequency, which needs to be stopped, otherwise this will corrupt the original signal and appear in the detector.

To reject the mirror/image frequency signal, an additional filter is often applied as front of the mixer, which is known as an image rejection filter (IR). This is bulkier and makes the on-chip integration difficult.

On-Chip Superheterodyne Receiver Architecture

The on-chip architecture of the above discussed superheterodyne receiver is shown in Figure 4.10. A passive band-pass filter limits the input spectrum provided by the antenna. Noise is introduced into the mixer, so the signal is first amplified by a low-noise amplifier (LNA) before mixing. Mixers translate the RF signal into IF frequencies. The LO signal, which is tuned at a particular spacing above or below the RF signal, is injected into the mixer circuits. Hence, first these bands have to be removed by an image reject filter. For this, the signal goes off-chip into an image rejection (IR) filter using passives with high-quality factors. Then, on mixing with a tuneable LO signal, the selected input channel frequency is down converted to an IF. This LO1 output needs to be variable in small frequency steps for narrow band selection. To alleviate the aforementioned sensitivity–selectivity trade-off in image rejection, an off-chip, high-Q band-pass filter performs partial channel filtering at a relatively high intermediate frequency. A second down-conversion mixing step translates the signal down to baseband frequency, which can be treated in the digital domain. This reduces the requirement for the final, integrated analog channel selection filter, as now it can be done digitally.

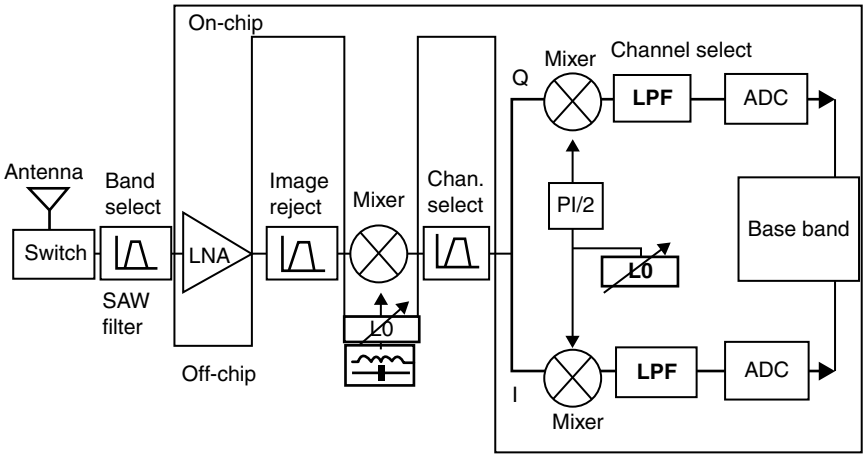


Figure 4.10 Heterodyne RF converter (with on-chip and off-chip components)

Digital modulation schemes, which use both in-phase (I) and quadrature (Q) elements of a signal, and thus both components can be generated in the second mixing stage, as shown in Figure 4.10. As the channel of interest has already been selected by the first mixer, the frequency of the second LO is fixed.

Off-chip passive components provide filters with a high Q-factor and result in good performance for both sensitivity and selectivity, which makes the heterodyne architecture a common choice. Furthermore, noise introduced by the local oscillator is less problematic, as it is filtered by the subsequent channel selection. Image rejection and adjacent channel selectivity is better in this type of architecture. The filters can be manufactured using different technologies such as SAW (surface acoustic wave external filter), bipolar, and CMOS. However, off-chip filtering comes at the expense of extra signal buffering (driving typically 50 ohm loads), increased complexity, higher power consumption, and larger size. Narrow-bandwidth passive IF filtering is typically accomplished using crystal, ceramic, or SAW filters (these are

passive filters). These filters offer better protection than the receivers of zero-IF gyrator filters (active filter) against signals close to the desired signal because passive filters are not degraded by the compression of a signal resulting from large signals. The active gyrator circuit does not provide such protection.

However, undesired signals that cause a response at the IF frequency in addition to the desired signal are known as spurious responses. In the case of the heterodyne receiver, spurious responses must be filtered out before reaching the mixer stages. One spurious response is known as an image frequency. An RF filter (known as a pre-selector filter) is required for protection against the image unless an image-reject mixer is used. Additional crystal-stabilized oscillators are required for the heterodyne receiver.

Generally, superheterodyne (superhet) receivers cost more than zero-IF receivers due to the additional oscillators and passive filters. These items also require extra receiver housing space, for example, increased size. However, the superior selectivity of a superheterodyne receiver may justify the greater cost and size in many applications.

The benefits of the superhet architecture are enormous: Most of the filtering and gain takes place at one fixed frequency, rather than requiring tunable high-Q band-pass filters or stabilized wideband gain stages. In some systems, multiple IFs are used to distribute the gain and selectivity for better linearity.

Advantages and Disadvantages of Superheterodyne Receiver

Advantages: (1) High selectivity and sensitivity as a result of the use of high-Q filters and the double RF down conversion scheme. (2) Good image rejection capability.

Disadvantages: (1) Image frequency problem; to reject the mirror frequency signal (image frequency) an additional filter (IR) is often applied in front of the mixer. (2) Poor integration; port integration becomes difficult as it uses high-Q devices, a double conversion scheme, IR, and channel filter (passive). Hence, it cannot be integrated in small packages such as inside a chip or set of chips. (3) Larger in size and weight. (4) It consumes more direct current (dc) power. (5) Only a fixed data bandwidth. (6) Higher cost for higher number of components, improved protection and larger physical size, requires extra printed circuit board (PCB) real estate.

Applications: The superhet receiver is typically used in radio receivers and satellite receivers.

4.3.2 Homodyne Receiver

4.3.2.1 Zero-IF Receiver (DCR)

The homodyne (homo = same, dyne = mix) architecture uses a single frequency translation step to convert the incoming RF channel directly into baseband, without any operations at intermediate frequencies. It is therefore also called as zero-IF or direct conversion receiver (DCR) architecture. Here, the IF frequency is chosen as 0 Hz (dc frequency) by selecting the local oscillator frequency to be the same as the desired RF input signal frequency. So after mixing at both I and Q channels, the generated frequency components will be $(f_0 - f_i) = 0$, and $(f_0 + f_i) = 2f_i$ as $f_0 = f_i$. This is shown in Figure 4.11. The portion of the channel translated into the negative frequency half-axis becomes the image to the other half of the same channel translated into the positive frequency half-axis. After the down conversion the input signal has a bandwidth of B Hz around the center 0. Figure 4.12 shows this architecture in the case of quadrature down conversion (I–Q de-modulation receiver) receiver. As with the heterodyne receiver, in this architecture an off-chip RF filter first performs band limitation, before the received signal is amplified by an integrated LNA. Channel selection is done by tuning the RF frequency of the LO to the center of the desired channel, making the image equal to the desired channel. Thus in this situation the problem of images is not present and the off-chip IR filter can be omitted. A subsequent channel selection low-pass filter (LPF) then removes nearby channels or interferers prior to A/D conversion.

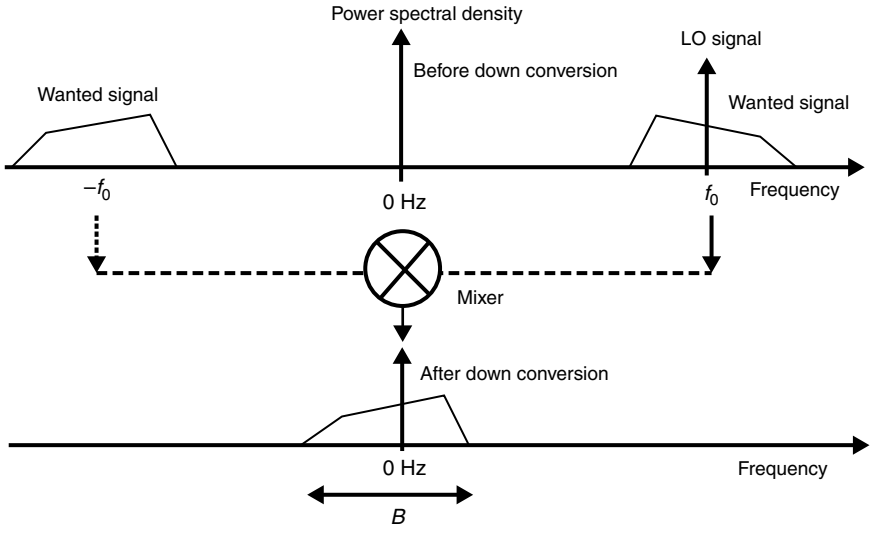


Figure 4.11 Direct conversion technique

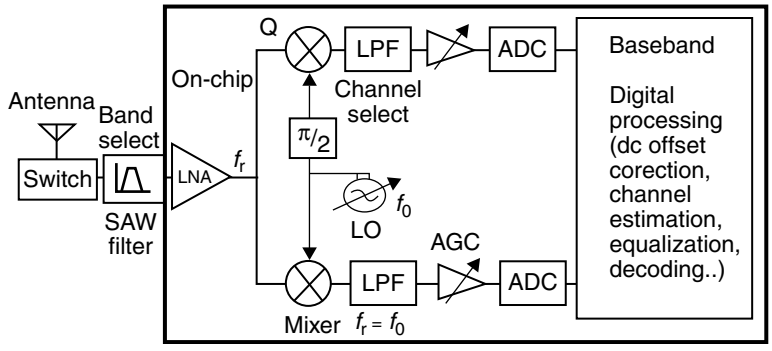


Figure 4.12 On-chip zero-IF direct conversion RF converter

Owing to the direct conversion to DC, homodyne receivers are more susceptible to disturbances arising from I/Q phase mismatches, non-linearities, and flicker noise, than heterodyne designs. To control the performance loss, additional circuitry and design efforts are required. However, there is no need for image rejection or other off-chip filters, which helps to save power and reduces the total receiver size.

Advantages of this Architecture

1. No image frequency problem – an advantage of the zero-IF receiver is that no image exists and an image-reject filter (or image-reject mixer) is not required.
2. LPF can be integrated – channel filtering is now possible entirely on-chip as after the down conversion the frequency is now at the baseband frequency. The zero-IF receiver can provide narrow baseband filtering with integrated low-pass (LP) filters. Often the filters are active op-amp-based filters known as gyrators. The gyrators provide protection from most undesired signals. The gyrator filters eliminate

the need for expensive crystal and ceramic IF filters, which take up more space on a printed circuit board.
3. Eliminate passive IF and image-reject filter – the IF SAW filter, IR filter, and subsequent stages are replaced with low-pass filters (LPF's) and baseband amplifiers that are amenable to monolithic integration. The LNA need not drive a 50 ohm load because no image rejection filter is required.
4. Increased ADC dynamic range because of limited filtering.
5. Good SSB digital modulation.
6. Reduced component numbers.
7. Reduced power consumption – the filtering and gain can now take place at dc, where gain is easier to achieve with low power, and filtering can be accomplished with on-chip resistors and capacitors instead of the expensive and bulky SAW filters.
8. High level of integration – the zero-IF topology offers the only fully integrated receiver currently possible. This fully integrated receiver solution minimizes the required board real estate, the number of required parts, receiver complexity, and cost. Most zero-IF receiver architectures also do not require image-reject filters, thus reducing cost, size, and weight.
9. Good multi-standard ability – the placement of the filter at the baseband (usually split between the analog and digital domains) permits multiple filter bandwidths to be included at no penalty to the board area, as the filtering is accomplished on-chip. Thus, direct conversion is the key to multimode receivers for the future.

Problems of this Architecture and Possible Alternative Design Solutions

Several well known issues that have historically plagued direct-conversion receivers are: self-detection due to LO-RF leakage, dc offset, and AM detection.

1. **Local Oscillator Leakage** – One of the most well know problems in direct-conversion receiver architecture is the spurious LO leakage. This arises because the LO in a direct-conversion receiver is tuned exactly to the desired input signal frequency, which is the center of the LNA and antenna pass-band. Owing to imcorrect isolation, a small fraction of this LO signal leaks through the mixer and travels towards the input signal side, passes through the LNA and arrives at the antenna (Figure 4.13). Then it radiates out through the antenna. This becomes an in-band interferer to other nearby receivers tuned to the same band, and for some of them, it may even be stronger than the desired signal. Regulatory bodies, such as the FCC strictly limit the magnitude of this type of spurious LO emission. Each wireless standard and the regulations of the Federal Communications Commission (FCC) impose upper boundariess on the amount of in-band LO radiation, typically between 50 and 80 dBm. The issue is less severe in heterodyne and image-reject mixers because their LO frequency usually falls out of the reception band. Also, it leaks to other side of the entire receiver signal chain, which appears as a dc

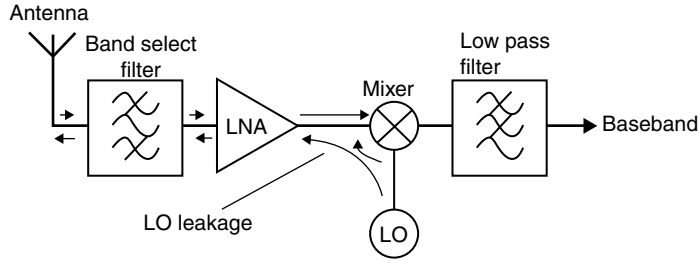


Figure 4.13 LO leakage

offset. The problem of LO leakage becomes severe as more sections of the RF transceivers are fabricated on the same chip.

Design Option – With differential local oscillators, the net coupling to the antenna can approach acceptably low levels.

2. **Self-reception** – Because the local oscillator is tuned to the RF frequency, self-reception may also be an issue (Figure 4.14).

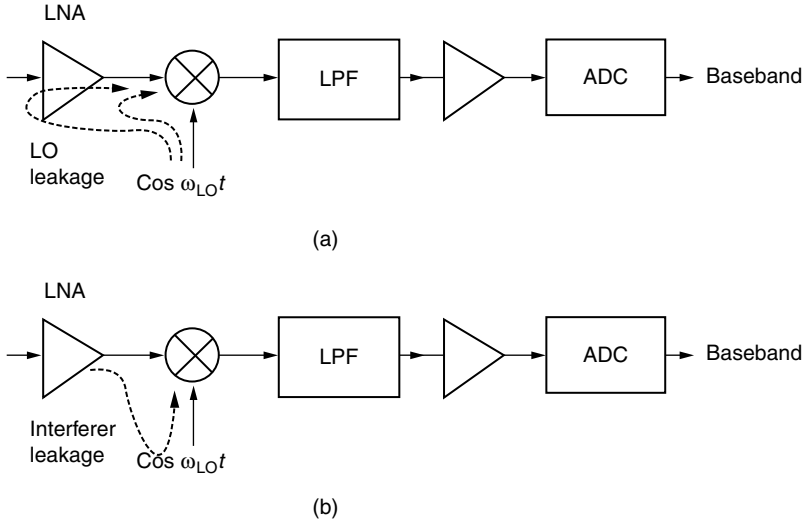


Figure 4.14 (a) Self-mixing LO and (b) interferers mixing

Design Option – Self-reception can be reduced by running the LO at twice the RF frequency and then dividing by two before injecting into the mixer. Because the zero-IF local oscillator is tuned to RF frequencies, the receiver LO may also interfere with other nearby receivers tuned to the same frequency. However, the RF amplifier reverse isolation prevents most LO leakage to the receiver antenna.

3. **Dc Offset Problem** – The basic operation of a direct-conversion receiver can be described as mixing an input signal frequency of ($f_C + f_m$), where f_m is the bandwidth of the modulation, with a local oscillator at f_{LO}, yielding an output at $f_{MIXOUT} = (f_C + f_m - f_{LO})$ and $(f_C + f_m + f_{LO})$. The second term is at a frequency twice that of the carrier frequency and can be filtered out very easily by the channel select filter. However, the first term is much more interesting, as $f_{LO} = f_C$, and substitution yields f_{MIXOUT} $f_{LO} + f_m - f_{LO} = f_m$. This means that the modulation has been converted into a band from dc to the modulation bandwidth, where gain, filtering, and A/D conversion are readily accomplished. The dc-offset problem occurs when some of the on-channel LO (at f_C) leaks to the mixer RF port, creating the effect that $f_{LO} - f_{LO} = 0$ (or dc). This can corrupt desired information that has been mixed down around zero Hz.

So, when the leaked LO signal appears at the input of the mixer, then leaked LO signal and LO signal are mixed, this will result in a zero frequency output or dc output as they are at same frequency. Also, remember, in DCR the desired down converted signal is centered around zero frequency. Thus, self-mixing caused by leakage from the local oscillator to the LNA (or vice versa) will corrupt the baseband signal at dc and saturate subsequent processing blocks. This leads to a narrower dynamic

range of the electronics, because the active components become saturated easier than in the case of a zero offset.

Dc offsets is a severe problem in homodyne receivers. If the receiver moves spatially, it receives reflected LO signals at the antenna, which generates time varying offsets. Causes of dc offset are either drift in the baseband components (for example, op amps, filters, A/D converters), or dc from the mixer output caused by the LO mixing with itself or with the mixers acting as square law detectors for strong input signals. Dc offsets from various sources lie directly in the signal band, and in the worst case they can saturate the back-end of the receiver at high gain values.

Design Options – From the above discussion, we infer that DCRs require some means of offset removal or cancellation.

a. **Ac Coupling:** A possible approach to remove the offset is to use ac coupling, that is, high-pass filtering, in the down-converted signal path. However, as the spectrum of random binary (or M-ary) data exhibits a peak at dc, such a signal may be corrupted if it is filtered with a high corner frequency.

   One technique is to disregard a small part of the signal band close to dc and employ a high-pass filter with a very sharp cutoff profile at low corner frequencies. This requires large time constants, and hence, large capacitors, that is, area. It is only practical for wideband applications (WCDMA), where the loss of a few tens of hertz bandwidth at dc does not degrade the receiver performance significantly. The system can either be ac coupled or incorporate some form of dc notch filtering after the mixer. However, for narrow band applications (GSM), this would cause large performance losses.

   A low corner frequency in the HPF may also lead to temporary loss of data in the presence of the wrong initial conditions. If no data are received for a relatively long time, the output dc voltage of the HPF drops to zero. Now if data are applied, the time constant of the filter causes the first several bits to be greatly offset with respect to the detector threshold, thereby introducing errors.

   A possible solution to the above problems is to minimize the signal energy near dc by choosing "dc-free" modulation schemes. A simple example is the type of binary frequency shift keying (BFSK) used in pager applications.

b. **Offset Cancellation:** Wireless standards, which incorporate time-division multiple access (TDMA) schemes, where each mobile station periodically enters an idle mode, so as to allow other mobiles to communicate with the base station. Thus, in the idle time slot, the offset voltage in the receive path can be stored on a capacitor and this is subtracted from the signal during actual reception. Figure 4.15 shows a simple example, where the capacitor stores the offset between consecutive TDMA bursts while introducing a virtually zero corner frequency during the reception of data. However, the major issues are the thermal noise and difficulty with offset cancellation in a receiver as interferers may be stored along with offsets. This occurs because reflections of the LO signal from

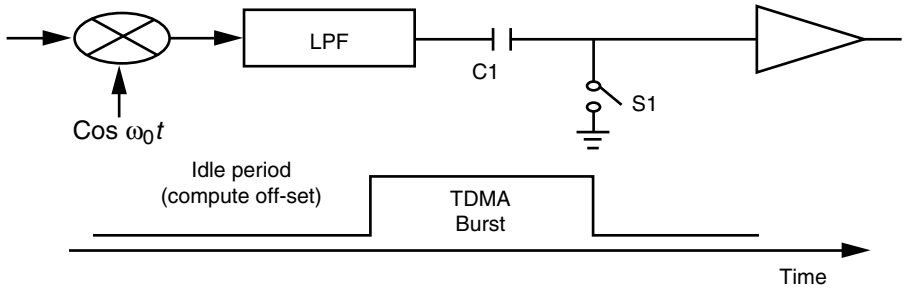


Figure 4.15 Off-set cancellation

nearby objects must be included in offset cancellation and hence the antenna cannot be disconnected (or "shorted") during this period. While the timing of the actual signal (the TDMA burst) is well defined, interferers can appear any time. A possible approach for alleviating this issue is to sample the offset (and the interferer) several times and average the results.

c. **Shielding and Other Layout Techniques:** These are often used to reduce this effect. Another approach is to convert an off-channel (or even out-of-band) LO signal to an on-channel LO inside the chip, reducing leakage paths. Also, operating the LO at half (or twice) the necessary injection frequency is a good solution for single-band applications; a regenerative divider simplifies multi-band designs. Once the dc offset due to LO-RF leakage has been reduced, a second problem arises: inherent dc offset in the baseband amplifier stages and its drift with temperature. Here, the best solution is to employ extreme care in the design of the gain stages and to make sure that enough gain, but not too much, is provided. Excessive gain in the baseband section may cause offsets that can be corrected momentarily but that may drift excessively and require additional temperature compensation. There are three possible methods by which offsets may be handled in the receiver: continuous feedback, track-and-hold, and open-loop. The continuous-feedback scheme (in software or hardware) attempts to null dc error at the mixer output. This generally requires tight coupling between the baseband processor and software and makes it difficult to mate an RF IC from one vendor with a baseband controller and software from another vendor. In the "track and hold" method, the dc offset is estimated just prior to the active burst (track mode) and then stored (hold mode) during the active burst. Such schemes are generally completely integrated with the radio IC and can be made transparent to the user by locally generating all the necessary timing signals. Practical issues with the scheme include dealing with multi-slot data (GPRS) where the baseband gain may be changing on a slot-by-slot basis (without adequate time to recalibrate) and also ensuring that the dc estimate obtained during the track mode is accurate. Such schemes can be implemented in either digital or the analog domains.

The latest-generation radios using the open-loop approach have substantially lowered dc offsets and can operate with lower-performance A/D converters (typically 60–65 dB of available dynamic range), without any special software requirements.

4. **Need for High-Q VCO** – As neither image rejection filter nor channel select filtering is done prior to mixing, all adjacent channel energy is untreated. This requires the LPF and ADC to have a sharp cutoff profile and high linearity, respectively. With respect to low-Q values of integrated components this implies tougher design challenges.
5. **Even-Order Distortion** – Even-order distortion, especially second-order non-linearity, can degrade the direct-conversion receiver's performance significantly, because any signal containing amplitude modulation generates a low-frequency beat at the baseband.

   Design Options – Because of the inherent cancellation of even-order products, differential LNAs and double-balanced mixers are less susceptible to distortion. However, the phenomenon is also critical for balanced topologies, due to unavoidable asymmetry between the differential signal paths. However, the problem is, if the LNA is designed as a differential circuit, it requires higher power dissipation than the single-ended counterpart to achieve a comparable noise figure.
6. **Flicker Noise (1/f Noise)** – As the down-converted spectrum is located around zero frequency, the noise of devices has a profound effect on the signal, a severe problem in MOS implementations.

   Design Options – The effect of flicker noise can be reduced by a combination of techniques. As the stages following the mixer operate at relatively low frequencies, they can incorporate very large devices (several thousand microns wide) to minimize the magnitude of the flicker noise. Moreover, periodic offset cancellation also suppresses low-frequency noise components through correlated double sampling. A bipolar transistor front-end may be superior in this respect to an FET circuit, but it is also possible to use auto-zero or double-correlated sampling to suppress flicker noise in MOS op-amp-based circuits.

7. **I/Q Mismatch** – As discussed earlier, in I–Q modulation, to achieve maximum information, we should take both parts of the signal. This is done by a method known as quadrature down-conversion. The principle of this method is that the signal is at first divided into two channels and then down-converted by an LO signal, which has a phase shift of 90° in one channel with respect to another. The vector of the resulting signal is described as: $|Signal| = \sqrt{I^2 + Q^2}$, $\arg(Signal) = \varphi = arctg\frac{Q}{I}$.

   The problem of the homodyne receiver, or, more concretely, of the I–Q (in-phase–quadrature) mixer, is mismatches in its branches. Assuming a mismatch of ε for the amplitude and θ for the phase, we can estimate the error caused by these mismatches. In this way we get:

$$E_{IQ} = \left| \frac{S_{ideal} - S_{miss}}{S_{ideal}} \right| \approx \left(\frac{1}{2} + \varepsilon \right) \theta^2 \tag{4.1}$$

   For typical values of $\varepsilon = 0.3$ and $\theta = 3^0$ this gives an error of $1.5 \cdot 10^{-3}$.

   I–Q modulation requires an exact 90° phase shift between the RF and LO signal or vice versa. In either case, the error results a in 90° phase shift and mismatches between the amplitudes. Signals corrupt the down-converted signal constellation, thereby increasing the bit error rate. All sections of the circuit and paths contribute to gain and phase error. This will show the resulting signal constellation with finite error. This effect can be better seen by examining the down-converted signals in the time domain. Gain error simply appears as a non-unity scale factor in the amplitude. Phase imbalance, on the other hand, corrupts one channel with a fraction of the data pulses in the other channel, in essence degrading the signal-to-noise ratio if data streams are uncorrelated. However, the mismatch is much less troublesome in DCR than in image-reject architectures.
8. **Need for AGC and AFC** – Sensitivity and rejection of some undesired signals, such as inter-modulation distortion, can be difficult to achieve in DCR in order to enhance the performance. The active gyrator filters compress with some large undesired signals. Once the gyrator is compressed, filter rejection is reduced, thus limiting the protection. Zero-IF receivers typically require an automatic gain control (AGC) circuit to protect against large signal interference that compresses the gyrator filters.

   Zero-IF receiver limitations require tighter frequency centering of the LO and RF frequencies. Significant offsets in the RF or LO frequencies degrade the bit error rate (BER). One solution for zero-IF designs is to add automatic frequency control (AFC). AFC prevents the centering problem by adjusting the frequency of the LO automatically.

Applications of this Architecture

Different modulation schemes exhibit varying susceptibilities to the problems in DCR. Quadrature-phase shift keying (QPSK) modulated spread spectrum schemes such as CDMA and WCDMA have almost no signal energy near dc and are more immune to dc offsets. This architecture is particularly suitable for the DS-SS (direct sequence spread spectrum) standard because of the wide channel bandwidth, and the removal of some small amounts of energy near zero for dc offset compensation will not have much impact on the over all received energy. Conversely, Gaussian minimum shift keying (GMSK) modulated GSM signals do have a dc component in the data and are under time constraints placed by the TDMA system. For this reason, the GSM signal cannot simply be ac-coupled at the baseband, nor can the dc offsets be easily filtered, because either of these methods would simultaneously remove wanted and unwanted signals. This is why, zero-IF DCR are not very useful for a GSM receiver. As discussed earlier, recent work using this architecture suggests that the effects of various imperfections can be alleviated by means of circuit design techniques.

The direct-conversion receiver architecture was successfully used in pagers (where ac coupling is allowed) and satellite receivers.

Direct conversion based transceiver solutions currently do not benefit from the most cost effective CMOS technologies due to the susceptibility to 1/*f* noise. This is because the 1/*f* noise in the mixer and baseband filtering stages appears directly on top of the down-converted signal in a direct conversion radio. This effectively increases the receiver noise figure (NF), especially in narrowband applications such as GSM. In practice, bipolar transistors will prove more appropriate for the LNA and mixer design, with MOS transistors allocated to the subsequent baseband stages. BiCMOS designs will be forced into more expensive and larger feature-size processes, hindering radio integration roadmaps aimed at cost-reduction.

Again careful design can minimize this problem, but still with the present scenarios, due some limitations, this could be the reason why direct conversion will not work for every application. Its monolithic integration capabilities make the homodyne architecture an attractive alternative for wireless receivers. If the RF signal is down-converted in a single step to a low (but not to dc) frequency, then limitations at dc have less impact on the receiver performance. This approach is followed in low-IF architectures, discussed next.

4.3.3 Low-IF Receiver

The digital low-IF receiver leverages the performance advantages of the superheterodyne with the economic and integration advantages of the direct conversion approach. This is accomplished by band selecting and down-converting the desired RF signal into a frequency very close to the baseband (for example 100 kHz) instead of zero, as illustrated in Figure 4.16. Next, the low-IF signal is filtered with an LPF and amplified before conversion into the digital domain by the analog-to-digital converter (ADC). The final stage of down conversion to baseband and fine gain control is then performed digitally.

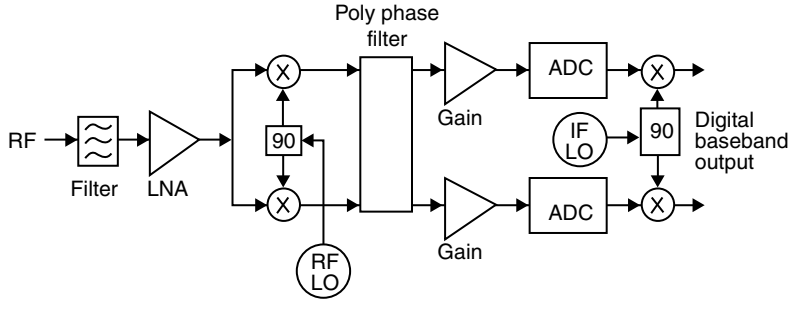


Figure 4.16 Low IF receiver

High-resolution, over-sampling, delta-sigma converters allow the channel filtering to be implemented with DSP techniques rather than with bulky analog filters. The signal can then directly interface to a digital BBIC input or a digital-to-analog converter (DAC) can be used to output analog I and Q signals to a conventional baseband integrated circuit (BBIC).

Similar to the DCR, the digital low-IF receiver is able to eliminate the off-chip IF SAWs necessitated by the superheterodyne approach. While the digital low-IF approach does encounter an image frequency at the adjacent channel, an appropriate level of image rejection can still readily be achieved with a well designed quadrature down-converter and integrated I and Q signal paths. This nullifies the need for external image reject filters. At the low-IF frequency, the ratio of the analog channel filter center frequency

to the channel bandwidth (moderate Q) enables the on-chip integration of this filter. Followed by amplification, the signal is then converted into the digital domain with an ADC. This ADC requires a higher level of performance than the equivalent DCR implementation because the signal is not at baseband. A digital mixer operating at 100 kHz can then be used for the final down-conversion to baseband where digital channel filtering is performed (Figure 4.17).

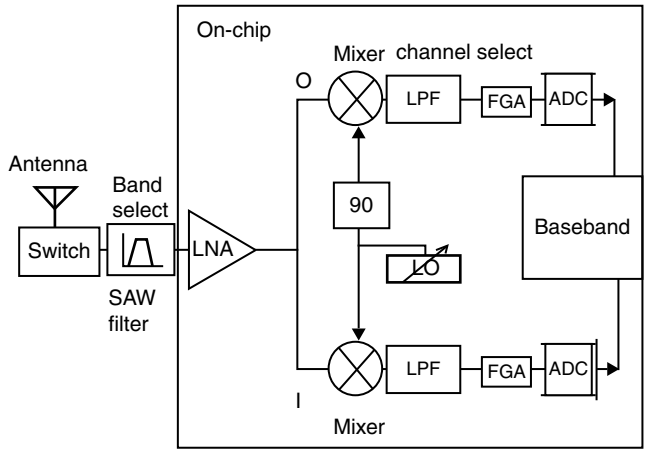


Figure 4.17 Integrated low-IF RF converter

The migration of these traditionally analog functions into the digital domain offers significant advantages. Fundamentally, digital logic is immune to operating condition variations that would corrupt sensitive analog circuits. Using digital signal processing improves design flexibility and leverages the high integration potential, scalability, and low cost structure of CMOS process technologies. While the digital low-IF receiver does add a down-conversion stage (mixer and filter), because the extra stage is digital, it is possible to implement this functionality in an area smaller than that occupied by the analog baseband filter of the DCR architecture. Digital low-IF receivers will also find it easy to comply with the developing DigRF BBIC interface standard for the next generation transceiver applications.

The digital low-IF architecture described curtails issues associated with dc offsets. As the desired signal is 100 kHz above the baseband after the first analog down-conversion, any dc offsets and low frequency noise due to second-order distortion of blockers, LO self-mixing, and $1/f$ noise can easily be eliminated by filtering. Once in the digital domain and after the down-conversion to baseband, dc offsets are of negligible concern. The desired signal is no longer as small and vulnerable and digital filtering is successful in removing any potential issues.

With dc offset issues avoided at the system level, digital low-IF receivers will greatly relax IP2 linearity requirements and will still meet the critical AM suppression specification with relative ease.

Manufacturers adamantly demand the most reliable, easy to implement, and low-cost components and ICs for each handset function. The immunity of digital low-IF receiver to dc offsets has the benefit of expanding part selection and improving manufacturing. At the front end, the common-mode balance requirements on the input SAW filters are relaxed, and the PCB design is simplified. At the radio's opposite end, the BBIC is one of the handset's largest BOM contributors. It is not uncommon for a DCR solution to be compatible only with its own BBIC in order to address the complex dc offset issues. Fortunately, digital low-IF based transceiver solutions can empower the system designer with multiple choices when considering BBIC offerings. This is because there is no requirement for BBIC support of complex dc offset calibration techniques.

In addition to flexibility, digital low-IF based transceivers may be able to capture notable sensitivity improvement from the BBIC. Many BBICs for GSM systems employ dc filtering as a default to compensate for large dc drifts that may occur when they are coupled with a DCR based design. When these same BBICs are paired with low-IF transceivers, such filtering is not needed. The handset designer is then in a position to work with the BBIC vendor to disable the unwanted filtering in the software. This has the benefit of regaining the valuable signal bit energy around baseband frequencies that had been thrown away by the filtering. The handset designer can then enjoy a potential sensitivity enhancement of 0.2–0.5 dB at little expense!

4.3.3.1 Advantages

The gain and filtering are done at a lower frequency than the conventional high-IF superhet. This reduces the power and opens up the possibility of integrating the filter components on-chip, thus reducing the total number of components. If the gain stage is ac coupled, any issues relating to dc offsets should be eliminated. The main advantages are: (a) no image frequency problems, (b) LPF can be integrated in the IC/digital module, (c) eliminates the passive IF and image reject filter, (d) high level of integration, (e) reduced number of components compared with a superheterodyne receiver, (f) reduced dc offset problem, and (g) less $1/f$ noise compared with a zero-IF receiver.

4.3.3.2 Disadvantages

(1) More baseband processing power requirement (MIPS). (2) ADC requires a higher level of performance than the equivalent DCR implementation because the signal is not at baseband. (3) Receiver's polyphase filter requires more components than an equivalent low-pass filter used in a DCR. We know that in true mathematics $\cos(\omega t) = \cos(-\omega t)$, so the negative frequency can not be identified. As mentioned before, we want to discriminate between positive and negative frequencies in order to realize on chip selectivity. This is not possible with real signals but is possible with two-dimensional signals or complex signals. We can imagine positive and negative frequencies as being phasors rotating in the complex plan in opposite directions. The complex signals used in a receiver are called polyphase signals, which consist of a number of real signals with different phases. A quadrature signal consists of two real signals with $\pi/2$ phase shift. The polyphase bandpass filter ensures the rejection of the mirror frequency and provides the antialiasing necessary in the digital signal processor (DSP) which does the final down-conversion to baseband and demodulation of the signal. The wanted signal is multiplied with a single positive frequency at f_{LO}. The mirror signal will be mixed down from f_{mirror} to $-f_{IF}$ and the wanted signal at f_{IF}. With a polyphase filter it is possible to discriminate between the negative and positive frequencies and therefore, the mirror frequency will be filtered out. (4) Image cancellation is dependent on the LO quadrature accuracy. (5) In hybrid implementations, where the image-reject function is divided into analog and digital phase-shift stages, the A/D conversion process occurs at the IF frequency. This generally requires higher power than baseband converters and more stringent control of the sampling clock, as clock jitter will degrade the conversion of an IF signal.

4.3.3.3 Applications

It is best suited to GSM (GMSK) receivers, but can also be used in multi-mode receivers.

4.3.4 Wideband-IF Receiver

An alternative to the low-IF design is the wideband-IF architecture shown in Figure 4.18, this receiver system takes all of the potential channels and frequencies translates them from RF into IF using a mixer

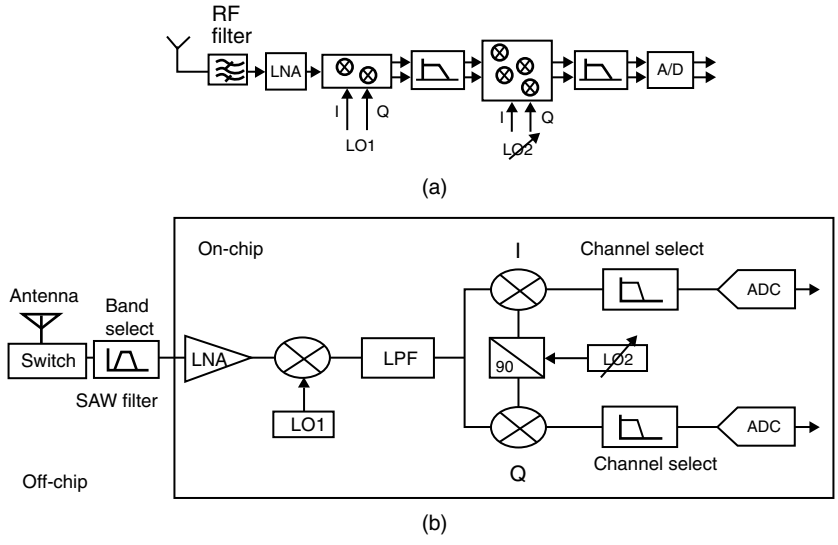


Figure 4.18 (a) Wide-IF RF converter and (b) on-chip implementation of wide-IF receiver

with a fixed frequency local oscillator (LO1). A simple low-pass filter is used at IF to remove any up-converted frequency components, allowing all channels to pass to the second stage of the mixers. All of the channels at IF are then frequency translated directly to baseband using a tunable, channel-select frequency synthesizer (LO2). Alternate channel energy is then removed with a baseband filtering network where variable gain may be provided.

This approach is similar to a superheterodyne receiver architecture in that the frequency translation is accomplished in multiple steps. However, unlike a conventional superheterodyne receiver, the first local oscillator frequency translates all of the received channels, maintaining a large bandwidth signal at IF. The channel selection is then realized with the lower frequency tunable second LO. As in the case of direct conversion, channel filtering can be performed at baseband, where digitally programmable filter implementations can potentially enable more multi-standard capable receiver features.

In contrast to the previous architectures, the first local oscillator frequency is fixed. All available channels are converted into an intermediate frequency, resulting in a wide bandwidth at IF. Up-converted frequency components are removed by a simple low-pass filter. Channel selection and filtering are done at IF. Owing to the lower operation frequency, the requirements for the tunable LO and low-pass filter in the second down-conversion stage are relaxed. Hence, a narrow channel can be selected and filtered without off-chip components. Furthermore, filtering can be performed partly in the digital domain, which adds to multi-standard operation capabilities of this architecture. This flexibility comes at the expense of higher linearity requirements of the ADC.

The wideband IF architecture offers two potential advantages with respect to integrating the frequency synthesizer as compared with a direct conversion approach. The foremost advantage is the fact that the channel tuning is performed using the second lower frequency, or IF, local oscillator and not the first, or RF, synthesizer. Consequently, the RF local oscillator can be implemented as a fixed-frequency crystal-controlled oscillator, and can be realized by several techniques that allow the realization of low-phase noise in the local oscillator output with low-Q on-chip components. One such approach is the use of wide phase-locked loop (PLL) bandwidth in the synthesizer to suppress the VCO contribution to phase noise near the carrier. Note that the VCO phase noise transfer function has a high-pass transfer function close to the carrier and the bandwidth of suppression is related to the PLL loop bandwidth. In addition, as channel

tuning is performed by the IF local oscillator, operating at a lower frequency results in a reduction in the required divider ratio of the phase-locked loop necessary to perform channel selection. The noise generated by the reference oscillator, phase detector, and divider circuits of a PLL all contribute to the phase noise performance of a frequency synthesizer. With a lower divider ratio, the contribution to the frequency synthesizer output phase noise from the reference oscillator, phase detector, and divider circuits can be significantly reduced. Moreover, a lower divider ratio implies a reduction in spurious tones generated by the PLL. An additional advantage associated with the wideband IF architecture is that there are no local oscillators operating at the same frequency as the incoming RF carrier. This eliminates the potential for the LO re-radiation problem that results in *time-varying* dc offsets. Although the second local oscillator is at the same frequency as the IF desired carrier in the wideband IF system, the offset that results at baseband from self mixing is relatively constant and is easily cancelled.

As the first local oscillator output is fixed and is different from the channel frequencies, the problem of dc offset is alleviated in the wideband-IF architecture. The still existing self-mixing in LO1 or LO2 results in constant dc offsets that can be removed either in the analog or digital domain. Isolation from the channel selection oscillator (LO2) to the antenna is much larger than in the heterodyne case. This greatly reduces problems associated with time varying offsets. Using a fixed frequency at LO1 allows for phase noise optimization for this oscillator. Frequency conversion into IF introduces images again. These can be removed using a Weaver architecture, but mismatches between the I and Q paths limit the image suppression.

Also, additional components from the second conversion stage inevitably result in larger power consumption. These problems are balanced by good monolithic integration capabilities and improved multi-standard prospects due to programmable filtering in the DSP.

4.3.4.1 Advantages

(1) Allows for high level of integration, (2) relaxed RF PLL specification, VCO could be made on chip, (3) channel selection performed by IF PLL lower the required divider ratio, (4) good multi-standard ability, and (5) alleviated dc offset problem.

4.3.4.2 Disadvantages

(1) Increase of 1 dB compression point for the second set of the mixer, and (2) increased ADC dynamic range requirement because of limited filtering in comparison with the heterodyne receiver.

Applications – Feasibility has not been proven for GSM, but can be used in satellite radio receivers.

4.4 Receiver Performance Evaluation Parameters

Optimizing the design of a communications receiver is inherently a process of compromise. There are several factors that govern the performance of a radio receiver:

1. **Selectivity and Sensitivity:** The most important characteristics of a receiver are its sensitivity and selectivity. Sensitivity expresses the level of the smallest possible input signal that can still be detected correctly (that is, within a given BER). Selectivity, on the other hand, describes the receiver's ability to detect a weak desired signal in the presence of strong adjacent channels, so called interferers. Thus, sensitivity is the lowest signal power level that the receiver can sense and selectivity is the selection of the desired signal from the many (which are received by the antenna). For a good receiver the selectivity and sensitivity should be higher than the reference level.

2. **Noise Figure (NF):** Indicates how much the SNR degrades as a signal passes through the system. Low NF enables successful reception, even for low levels of received signal power.
   The equation which relates the noise figure and sensitivity is as below:

$$S = F.B.k.T_0\left(\frac{S_0}{N_0}\right) \tag{4.2}$$

   where
   S = sensitivity in watts
   F = numeric system noise figure
   B = receiver bandwidth in Hz
   k = Boltzman's constant = $1.38 * 10^{-23}$ J/K
   T_0 = 290 Kand
   S_0/N_0 = receiver's output SNR (numeric)
3. **Image Rejection (IR):** Measures the ratio of the desired signal to the undesired image. The higher the IR the better the receiver.
4. **Phase Noise:** Phase noise describes an oscillator's short-term random frequency fluctuations. Noise sidebands appear on both sides of the carrier frequency. Typically, one sideband is considered when specifying phase noise, thus giving single sideband performance. Thus, low phase noise is crucial for oscillators in receiver systems.
5. **Receiver Non-linear Performance:** Amplifiers usually operate as a linear device under small signal conditions and become more non-linear and distorting with increasing drive level. The amplifier efficiency also increases with increasing output power, thus, there is a system level trade-off between the power efficiency or battery life and the resulting distortion. It is desired that the receiver's non-linear performance should be good.
6. **Processing Power to Drive Different Applications:** A higher MIPS is always desirable to drive the different complex processings but there should be a trade-off between cost and power consumption.
7. **Cost and Size:** This is the driving factor that is most required for design.
8. **Complexity:** The implementation of receiver architecture should be simple.

4.4.1 Receiver Architecture Comparison

The parameters for various receiver architectures are compared in the Table 4.2.

4.5 Transmitter Front-End Architecture

As discussed in the first section, the RF transmitter module mainly consists of a root raised cosine filter, modulator, power amplifier, RF filter, duplexer, and antenna.

4.5.1 Power-Limited and Bandwidth Limited Digital Communication System Design Issues

Communication system design involves trade-offs between performance and cost. Performance parameters include transmission speed, accuracy, and reliability, whilst cost parameters include hardware complexity, computational power, channel bandwidth, and required power to transmit the signal. Generally, a communication system is designed based on the criteria:

1. bandwidth limited system
2. power limited system.

Table 4.2 Receiver architecture comparison

Parameter	Super heterodyne	DCR	Low-IF
Transceiver IC process Technology	Bipolar BiCMOS, GaAS	BiCMOS, CMOS (rarely chosen)	CMOS
Integration	Low	High	High
Off-chip IR filter	Required	Not Required	Not required
IF-filtering	Requires IF SAW	On-chip LPF (may need external capacitors)	On-chip
Noise-figure	10.7 dB	Same	Same
Image rejection	−11 dB	−25 dB	−28 dB
Second intercept point (IP2)	N/A	43 dBm	18 dBm
Factory IP2 calibration	N/A	43 dBm	18 dBm
Third intercept point (IP3)	−19 dBm	Same	Same
Dc off-sets	Not there, easily filtered	Yes there and inherently susceptible	Not there and easily filtered
Flicker noise (1/*f*)	No	Yes	No
LO-self mixing	No	Yes	No
Interferer leakage	No	Yes	No
RCVR dc off-set calibration	No	Yes	No
BBIC dc off-set calibration support	No	May be required	No
Option to disable dc filtering	N/A	No	Yes – potential sensitivity improvement
Die size	Large	Small	Moderate to small
Power consumption	Moderate	Small	Small
Component selection	Difficult	Moderate	Easy
PCB layout	Moderate	Difficult	Easy
Cost reduction roadmap	Difficult	Moderate	Easy
Cost	High	Low	Moderate
Solution risk	Low	Moderate	Low

In bandwidth limited systems, spectrally efficient modulation techniques can be used to save the bandwidth at the expense of power, whereas in power limited systems, power efficient modulation techniques can be used to save power at the expense of the bandwidth. In a system that is designed to be both bandwidth and power limited, error correction coding (channel coding) can be used to save power or improve error performance at the expense of bandwidth. Trellis coded modulation (TCM) schemes can also be used to improve the error performance of bandwidth limited channels, without an increase in bandwidth.

Generally, there are two main causes of error in a communication system: *noise* and *distortion*. We will first just consider distortion as noise was discussed in Chapter 2. The distortion occurs in an ideal baseband channel, only if the bandwidth of the transmitted signal exceeds the bandwidth of the channel and also at the transmitter's power amplifier module.

In order to come up with a proper design trade-off between accuracy, transmitted power, and transmission speed or bandwidth requirement, we need to examine how the energy of the baseband data pulse is distributed throughout the frequency band. The time and frequency band representations of single rectangular pulses of height A and width τ, is shown in the Figure 4.19.

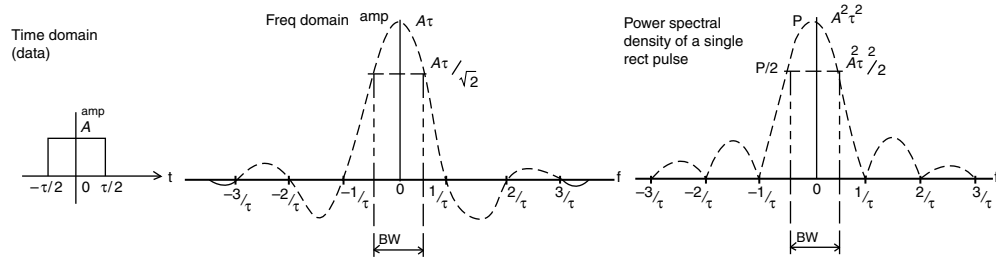


Figure 4.19 Rectangular pulse, its frequency domain representation and power spectral density

From this, we can calculate and plot an average normalized power spectral density for a series of n number of such pulses as shown in Figure 4.20 and its value will be:

$$G(f) = \left\{ nA^2\tau^2 \sin c^2 \frac{(\pi.f.\tau)}{n\tau} \right\} = A^2\tau \sin c^2(\pi.f\tau) \tag{4.3}$$

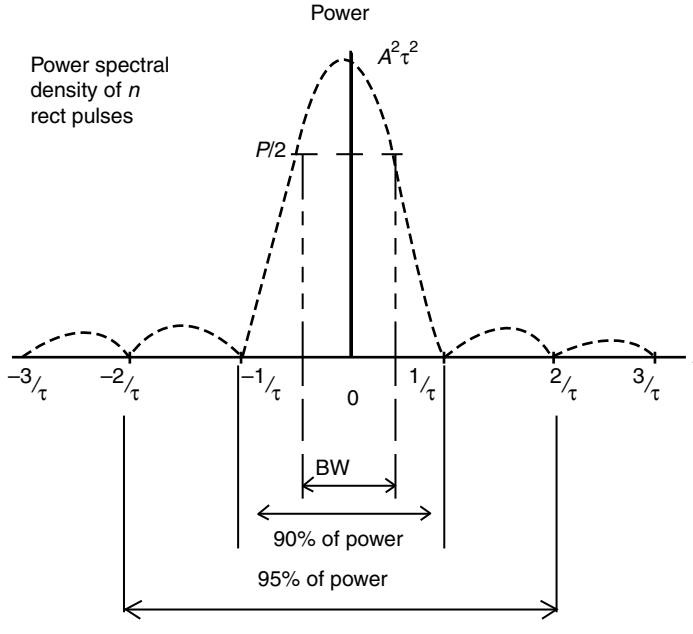


Figure 4.20 Power spectral density of n rectangular pulses

From Figure 4.20, it is obvious that most the power lies in the main lobe and inside the calculated BW. We can quantify the accuracy of the received signal for an ideal baseband channel by stating the percentage of the transmitted signal's power that lies within the frequency band passed by the channel. From Equation 4.2, it is evident that the accuracy (which is dependent on the transmitted power spectral density $G(f)$ at a pre-defined label), bandwidth or speed of transmission (τ) and the amplitude of the data pulse are interdependent. This indicates that if the transmission rate is more, which requires more bandwidth, then to keep the accuracy at the same label (for example, 95%), we need to increase the amplitude or power level of the data pulse and vice versa. More BW means more average transmitted power will lie inside the frequency band, so the amplitude of the signal A (for example, transmitted power level) can be reduced. Thus more available BW requires less transmitted power for the same level of accuracy (desired power in the selected band).

If we increase the amplitude of the pulse, this will lead to more power consumption, more interference, and more non-linear distortion in the amplifier. Hence we need to investigate how we can increase transmission speed without reducing accuracy or increasing the bandwidth and amplitude of the transmitted pulse. Special shaped pulses, which require less bandwidth than rectangular pulse need to be considered. We know that bandwidth is inversely proportional to pulse width τ and we want the pulses to be as wide as possible to reduce bandwidth, but we do not want the pulses to overlap. This is accomplished by selecting a pulse width of $\tau = T$, making each pulse as wide as its corresponding bit period. Thus we can relate the optimum pulse width and transmission speed by: $\tau_{opt} = 1/r_b$.

Our requirement should be: (1) a spectrally compact and smooth shaped pulse, as this will contain lower frequency components; and (2) the pulse transmitted to represent a particular bit should not interfere at the receiver with the pulse transmitted previously, for example, there should not be any inter symbol interference (ISI).

Keeping these two points in mind, the sync-shaped pulse in the time domain satisfies both the requirements. It is important to observe that there will be no ISI at exactly the center of each bit period. So at the receiver we need to sample the received signal exactly in the center of each bit period to avoid ISI. However, if the receiver is not completely synchronized with the transmitter, then this will cause timing jitter. Now the question is how we can reduce the timing jitter or ISI. This indicates we need to use a pulse that is smooth, like a sync pulse, but has a narrower main lobe and flatter tails. Consider the waveform as shown in Figure 4.21, which is a sync-shaped pulse multiplied by a damping factor, and is known as raised cosine pulse shape. The larger the damping factor β, the narrow the main lobe and the flatter the tail of the pulse. Thus a larger value of β means less effects of ISI and less susceptibility to timing jitter, but, greater the value of β, the greater is the bandwidth. The roll-off factor $\alpha = \beta/(r_b/2)$ allows us to express the

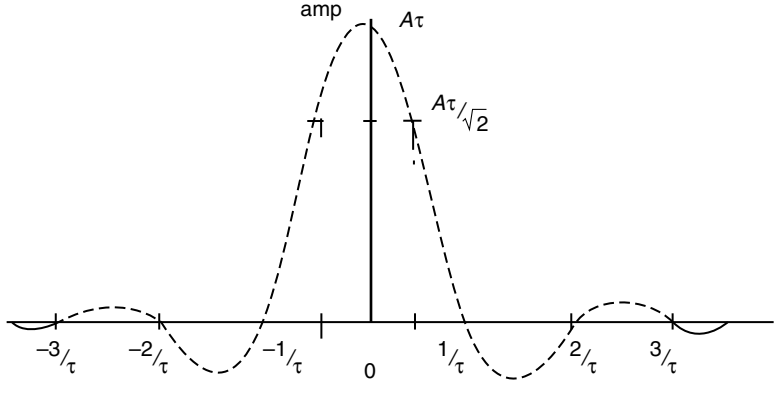


Figure 4.21 Damped sync pulse

trade-off of additional bandwidth for less susceptibility to jitter in a manner that is independent of transmission speed.

$$P(t) = \frac{A \sin c(\pi.t.r_b)}{[\cos(2.\pi.\beta.t/1-(4\beta t)^2]} \tag{4.4}$$

The trade-offs for selecting the pulse shapes for binary PAM is shown Table 4.3.

Table 4.3 Advantages and disadvantages of using different pulse shapes

Pulse shape	Bandwidth	Advantage	Disadvantage
Rectangular $\tau = n/r_b$	$2.r_b$ (95% in-band power)	No ISI, minimum susceptibility to jitter	High bandwidth requirement
Sinc $\tau = 2/r_b$	$0.5.r_b$ (100% in-band power)	Low bandwidth, no ISI only if receiver is perfectly synchronized	Susceptible to timing jitter
Raised cosine (freq. domain) $\tau = n/r_b$	$r_{b/2}.(1 + \alpha)$ (100% in-band power)	No ISI, less susceptibility to jitter than sync pulse	Requires more bandwidth than sync pulse but less than rectangular pulse

Obtaining tighter synchronization requires more complex equipment within the receiver, thus transmitting raised cosine pulses will reduce the receiver's complexity relative to transmitting sync pulses. A filter, composed of discrete components, is designed to produce an impulse response resembling a time delayed version of the raised cosine pulse. Then a series of narrow pulses are input to the filter, one pulse per bit period, with a positive narrow pulse representing each "1" whereas a negative narrow pulse represents each "0." The drawback of the analog method is it requires large numbers of discrete components.

Generating raised-cosine shaped pulses using digital circuitry is much easier than using analog circuits. As we have observed, these design parameters such as accuracy, transmitted power, BW, data rate, and complexity are all inter-related and hence the proper trade-offs depend on system's application.

4.5.2 *Investigation of the Trade-offs between Modulation and Amplifier Non-Linearity*

The choice of modulation has always been a function of hardware implementation, and required modulation and BW efficiency. Amplifiers usually operate as a linear device under small signal conditions and become more non-linear and distorting with increasing drive level. The amplifier efficiency also increases with increasing output power, but this leads to the problem of increased non-linear distortion and reduced battery life. For most commercial systems, this trade-off is constrained by interference with adjacent users, power efficiency, battery life, and resulting signal distortion. Thus, in many cases the amplifier signal levels are reduced or "backed off" from the peak efficiency operating point. Hence, we need to investigate the amplifier–modulation combination to minimize the energy required to communicate information.

Linear transmitter power amplifiers, such as class-A or class-B type of amplifiers, offer good quality, low output signal distortion but with significant penalties in heat dissipation, size, and efficiency. Alternatively, non-linear amplifiers such as class-C amplifiers offer very good efficiency and low heat

dissipation but introduce more distortion into the output signal. As the class-C and -AB type of amplifiers offer good efficiency so, for better power usage purposes, this type of amplifier is generally used as the RF transmitter power amplifier.

A higher level of modulation states is used to carry more information bits per symbol. However, every time a modulation level is doubled an additional 3 dB of signal energy are needed to maintain the equivalent demodulator bit error rate performance.

For GSM, the modulation used is GMSK, where the filtering ensures the modulation is a constant envelope, the disadvantage is that decision points are not always achieved, resulting in residual demodulator bit error rate. 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. As OFDM shows how highly sensitive it is to non-linear effects, so it requires more linear amplification than other modulation schemes. A multi-carrier modulated signal has a very large peak power, so the influence of the non-linear amplifier becomes large. Increase in peak power leads to input signal saturation, which leads to non-linear amplitude distortion and this leads to out of band radiation and degradation of BER.

The modulation schemes used for WLAN, UMTS, and GSM can be broadly divided into two categories.

4.5.2.1 Constant Envelope (Non-Linear) Modulation

In this case the signal envelope is fixed. It employs only phase information to carry the user data, along with a constant carrier amplitude. This allows the use of non-linear amplifier stages, which can be operated in class-AB or -C, so good power efficiency can be achieved. The most common standard employing non-linear modulation is GSM (Global Standard for Mobile Communication), which uses Gaussian minimum shift keying (GMSK) with a **BT** factor of **0.3** and raw data rate of 270.833 kbps.

4.5.2.2 Non-Constant Envelope (Linear) Modulation

In this case the signal envelope varies with time. The user information is conveyed in both the phase and amplitude of the carrier. Thus the transmitters should not distort the waveform at all, and hence amplifier stages must be operated in a linear class-A fashion. QPSK and BPSK are the modulation types used for the UMTS and WLAN OFDM systems. This means, WLAN and UMTS use non-constant envelope (linear) modulation whereas GSM uses a non-linear or constant envelope modulation scheme.

In the case of a multi-mode system, the modulation schemes for different modes are already defined. Therefore we have no scope to change the modulation scheme. However, we can come up with a best suitable transmitter architecture and power amplifier for this multi-mode terminal solution.

4.6 Transmitter Architecture Design

Transmitter design has unfortunately not resulted in a single preferred architecture suitable for all applications, due to the differing requirements for linear and non-linear modulation schemes.

4.6.1 Non-Linear Transmitter

The favored architecture for a constant-envelope transmitter is the offset phase-locked loop. This utilizes an on-frequency VCO, which is modulated within a phase-locked loop. A block diagram of this architecture is shown in Figure 4.22.

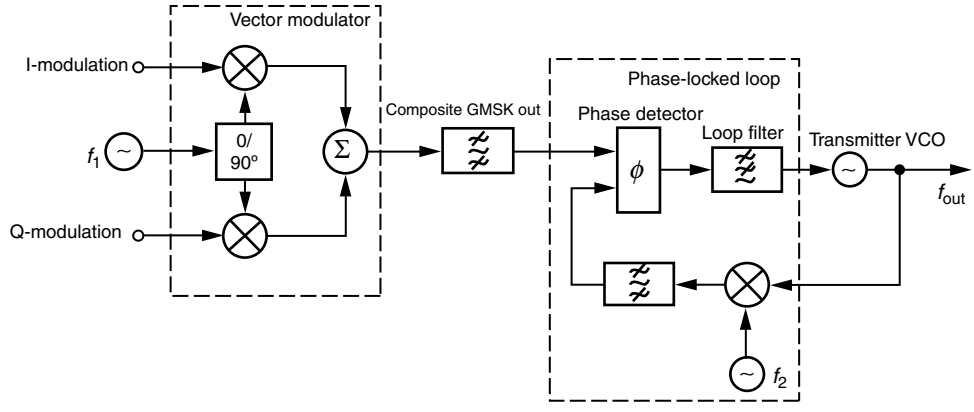


Figure 4.22 Transmitter architecture for non-linear modulation schemes

A modulated carrier is generated at an IF of f_1 using an I–Q vector modulator. The modulated carrier is applied to the phase-locked loop, which modulates the VCO phase in order to track the phase of the feedback part of the phase comparator. The output signal f_{out} is converted back down to f_1 using a mixer with an LO at f_2 such that $f_1 = f_{out} - f_2$, for comparison in the phase detector.

This architecture is used in most of the current GSM phones, as it provides optimum power efficiency, cost, and performance by minimizing the amount of filtering required at the output. It also readily extends to future GPRS enhancements. The offset approach eliminates the problems of LO leakage, image rejection, and spurious sideband conversion associated with heterodyne architectures, reducing the filtering requirements at the PA output. The PLL has a low-pass response to the vector modulator, which invariably has a high wideband noise floor. The noise is therefore rejected by the loop before the signal reaches the PA. The output is also protected from the high noise figure of the offset mixer, which is not the case in heterodyne architectures. As the signal is of constant amplitude, it is possible to apply power control within the power amplifier stages, which follow. This allows the main transmitter to be optimized for power consumption.

4.6.2 Linear Transmitter

Linear modulation transmitters are required to preserve both the phase and amplitude of the signal. The consequence of this is that the offset phase-locked loop transmitter cannot be used as it only transmits phase information. It would be possible to apply the amplitude modulation component at the VCO output, but there are technical difficulties associated with this technique, in particular AM-to-PM conversion in the power amplifier, which have yet to be solved to give a viable solution. Instead a conventional heterodyne architecture is usually employed, comprising an IF modulator and an up-conversion mixer.

The power control requirements of the standards, usually call for power control to be distributed through the transmitter module. This is because the required dynamic range requires more than one variable gain stage. For a cellular system the final transmit carrier frequency can be up to 2 GHz. Variable gain amplifiers at the final transmit frequency are difficult to implement with large gain control ranges. Thus it is necessary to perform some of the gain control earlier in the transmitter chain.

The transmitter architecture for a linear scheme is shown in Figure 4.23. With the offset PLL architecture, the carrier modulation is achieved at an IF f_1. Here the baseband I and Q signals contain information in both phase and amplitude. The signal is band-pass filtered to remove unwanted products and wideband noise from the vector modulator output, and a variable gain stage enables some power control (subject to carrier leakage limitations). This signal is then up-converted in a mixer using an LO

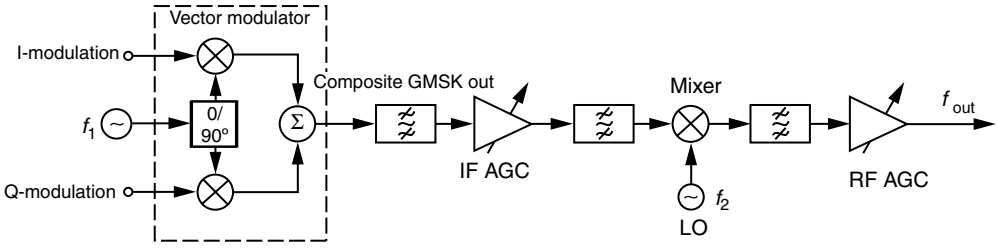


Figure 4.23 Transmitter architecture for linear modulation schemes

at frequency f_2. The output is filtered for image rejection, and so on, and a further variable gain stage is used to give the total required dynamic range. The distribution of the power control needs to be carefully planned to maintain the SNR along the chain. In particular, the vector modulator and up-convert mixer generate wideband noise levels, which need to be considered in the transmitter level plan. The subsequent power amplifier is required to be linear for the entire dynamic range of the transmitter, which can lead to power inefficiency. However, some transmitters do switch the PA bias from high to low power, to help this situation.

4.6.3 Common Architecture for Non-Linear and Linear Transmitter

The future of wireless communication can be considered as a plethora of heterogeneous systems operating together with current and legacy technologies. A mobile handset designed to operate on multiple RF bands is called multi-band phone and when it is designed to operate for different cellular standards it is called multi-mode phone. One viewpoint of the next generation wireless communication is the seamless integration of a wide variety of systems (cellular – GSM, GPRS, EDGE, UMTS, and WLAN) for efficient service provision. WLAN handsets offer a significant improved data rate over cellular handsets. However, they have a very limited range and access nodes can only be found in high use areas. On the other hand, a cellular handset, which has a larger range at the expense of data rate, can solve this problem of insufficient range. For cellular handsets, several standards have been introduced, which will co-exist. Also, various frequency bands have been introduced at different geographical locations. So one major challenge will be to find a robust solution to incorporating both WLAN and different cellular modes (GSM-900/1800 and UMTS) into single user terminal equipment (UE), which should have the capability to select the appropriate mode in any given coverage area available, ideally with a seamless interface between the different modes of operation.

For each standard, the handset must be able to transmit and receive with conformity to the ANSI/IEEE or 3GPP standard. The Table 4.4 tabulates different parameters used for different modes of operation.

Table 4.4 Use of various modulation techniques for different systems

Mode	Duplex	Multiple access	Transmit band	Receive band	Modulation
GSM-900	FDD	TDMA/FDMA	890–915	935–960	Constant envelope (GMSK)
GSM-1800	FDD	TDMA/FDMA	1710–1785	1805–1880	Constant envelope (GMSK)
UMTS	FDD	W-CDMA	1920–1980	2110–2170	Non-constant envelope (BPSK/QPSK)
WLAN	TDD	OFDMA/CSMA	2400–2483.5	2400–2483.5	Non-constant envelope (QPSK)

A number of issues arise when so many functionalities for different standards co-exist in small single user equipment, but we will mainly focus on the transmitter design challenges.

Unfortunately, conventional GSM/GPRS transmitter architectures are designed to deliver constant-envelope half-duplex signals and have little in common with UMTS architectures, which generate envelope-varying full-duplex signals. As a result, combining these transmitters within one multimode handset can be an expensive proposition. However, by changing the transmitter's modulation system from quadrature (I–Q) to polar, the architectures can be designed to deliver constant-envelope and envelope-varying signals. Polar transmitter-based architectures have no requirement for linear radio-frequency (RF) circuitry, which means that circuits can be designed with an emphasis on optimizing efficiency rather than linearity.

This describes a new architecture for implementing multi-mode multi-band transmitters, which provides all the necessary functionality for both linear and non-linear modulation schemes. There are a number of different architectures that can be employed to implement a multi-mode transmitter. However, the trade-offs between power consumption, linearity, baseband complexity, and implementation issues have resulted in a favored architecture for supporting both non-linear and linear modulation. In Figure 4.24, the architecture of a multi-mode transmitter supporting the linear and non-linear modulation scheme is shown.

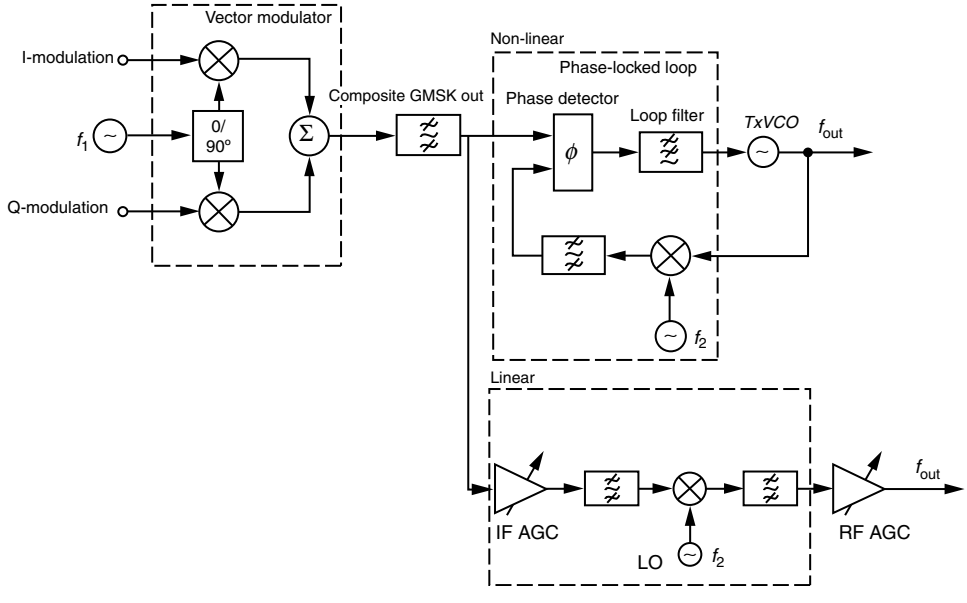


Figure 4.24 Multi-mode linear-nonlinear transmitter

As discussed there are several problems associated with the power amplifier design, and below some advantages and disadvantages of using a direct conversion linear transmitter are indicated.

Advantages: (1) The circuit BW should not exceed the signal BW. (2) Wide output power dynamic range. (3) No problem if the RF signal amplitude becomes zero. (4) There is wide experience in design and manufacturing. (5) Generally it supports any signal type.

Disadvantages: (1) RF circuits are generally not linear leading to compensating design complexity. (2) Low dc to RF energy efficiency due to back off requirement. (3) High PA operating temperature from internal power dissipation. (4) High broadband output noise floor. (5) Difficulty maintaining modulation accuracy. (6) The possibility of self-oscillation. (7) Gain that is dependent on frequency.

To address the PA efficiency problem with envelope varying signals, polar modulation techniques were proposed.

4.6.4 Polar Transmitter

The polar transmitter transforms the digital I–Q baseband signals from the Cartesian domain to the polar domain. The quadrature equivalent signal representation is: $S(t) = I(t) \cos \omega_c t + Q(t) \sin\omega_c t$, which can be represented in the polar domain as $A(t) \cos [\omega_c t + \Phi(t)]$. In this case the amplitude and phase components of the signal are processed independently. The phase information extracted from the original signal (either constant envelope or non-constant envelope) is transformed into a constant envelope signal. This is achieved by phase modulating with the help of a phase lock loop designed to output the desired transmit frequencies. The resulting signal may now be amplified by compressed amplifiers without any concern of distorting the amplitude information.

The extracted amplitude information is quantized into control bits. These bits are used to modulate a digital power amplifier (DPA). Each bit is a digital representation of the amplitude envelope. The control bits are used to switch amplifier elements of the DPA into on or off states. The examples use a 7-bit control word offering 128 unique amplitude modulation states. Fewer quantization states can be implemented if decreased amplitude modulation resolution is acceptable. More quantization states can also be implemented for greater resolution. The digitized amplitude envelope and the phase-modulated RF carrier are synchronized and recombined within the DPA to produce linear and efficient RF transmission. Existing developments of polar modulated transmitters generally fall within three major categories: polar loop, polar modulator, direct polar.

Polar Loop: Feedback control is used to correct the output signal into its desired form (Figure 4.25). One advantage of such a polar loop is improved PA efficiency over the best linear systems, gained from operating the power amplifier much closer to saturation. Additional benefits from this compressed PA operation include a low wideband output noise floor and also usually a reduction of circuit oscillation tendencies with varying output load impedance. Disadvantages include the need for a precision receiver within the transmitter, control loop bandwidths which must greatly exceed the signal bandwidth, restricted output power dynamic range, maintaining stability of the feedback control loops across the output dynamic range, and the lack of circuit design techniques when operating with strong circuit non-linearity.

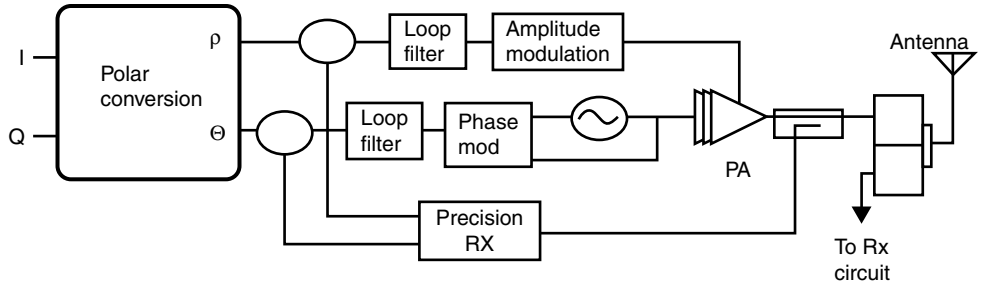


Figure 4.25 Polar loop transmitter

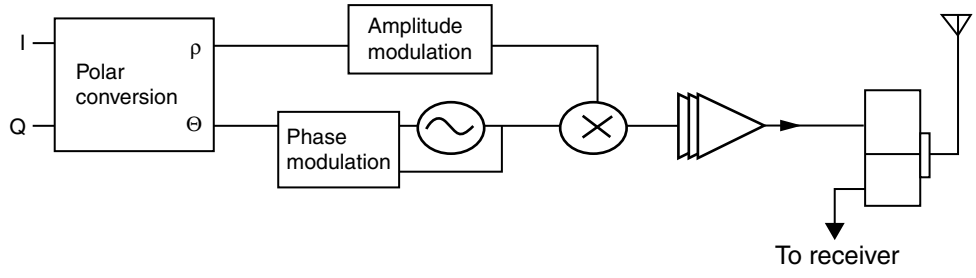


Figure 4.26 Polar modulator transmitter

Polar Modulator: As shown in Figure 4.26, in this type of transmitter, the output signal is amplified from the output of this polar modulator using conventional linear amplifier devices. The advantages and disadvantages of the polar modulator transmitter are given in Table 4.5.

Table 4.5 Advantages and disadvantages of polar modulator transmitter

Advantages	Disadvantages
The modulator noise is much lower than a quadrature modulator	Linear PA provides no efficiency benefit
Pre-distortion applied to the polar modulator	Time alignment of amplitude modulation and phase modulation is required
Generating output power = 0 is not an issue	Lack of manufacturing experience for polar modulator transmitter
Sigma-delta ADC can be used along with this	
Modulation accuracy is good	

Direct Polar Transmitter: Another method that removes the feedback from the PA is a direct polar transmitter. This is shown in Figure 4.27. The advantages and disadvantages of the direct polar transmitter are shown in Table 4.6.

All signals have constant-envelope phase components, so linear RF circuitry is not required in either architecture. Between the UMTS and GSM modes, most of the things are implemented digitally within the polar modulator, and therefore have little impact on how the radio design is actually implemented.

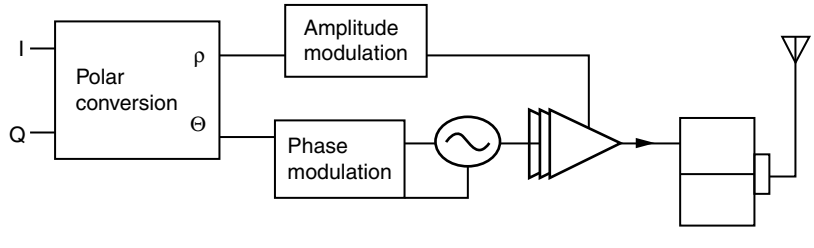


Figure 4.27 Direct polar transmitter

Table 4.6 Advantages and disadvantages of a direct polar transmitter

Advantages	Disadvantages
Efficiency is higher than the best linear transmitter	Power amplifier (PA) characterization required within the transmitter
Linear RF circuit is not required	AM/AM and AM/PM distortion
Unconditional power amplifier stability	Time alignment of AM and PM paths
Modulation accuracy	Lack of manufacturing experience

4.7 Transmitter Performance Measure

The performance of a transmitter is generally measured taking the following factors into account: amplifier power efficiency, amplifier non-linear distortion, modulated signal power efficiency, and signal bandwidth efficiency. This is based on the total energy needed per transmitted bit, and bandwidth efficiency.

The average total power consumed by the amplifier is:

$$\begin{aligned} P_t &= P_{dc} + P_{in} \\ &= P_{rf}[1 + (1 - P_{ae})P_{dc}/P_{rf}] \end{aligned} \tag{4.5}$$

where

P_{ae} = the instantaneous power-added efficiency of the amplifier
$P_{ae}(t) = [P_{rf}(t) - P_{in}(t)]/P_{dc}(t)$

Transmitter figures of merit are dependent on the following parameters: (1) spectral efficiency, (2) power efficiency, (3) spurious emission, and (4) power level.

4.7.1 Design Challenges

Transmitter architectures for multiple standards include direct up-conversion, translation loop, modulation through a phase-locked loop and a polar loop. The trend has been towards further digitizing to reduce the analog content in the total transmitter chain. Key challenges include current drain, dynamic-range requirements, and cost. Loop-phase modulation using a sigma-delta modulator shows promise in terms of low power consumption and a simpler architectural approach. For systems such as CDMA and W-CDMA, separation of AM and PM components is required. This leads to polar loop architectures, which are gaining wider use. However, challenges remain in their use for wideband systems, where alignment of the AM and PM components and the effect on spectral distortion are critical. While direct modulation has the advantage of compatibility with multiple standards, the challenges of meeting noise-floor requirements remain. Multimode phones require several bulky SAW filters to attenuate the received band noise. Signal digitization in the transmitter could include I and Q over-sampled D/A converters to ease the requirements of the reconstruction filter. As there are no blockers in the transmitter, this eases the design of the converter somewhat. Sufficient dynamic range to meet spectral mask requirements still has to be considered in the transmitter chain.

The final stage in the transmitter chain is the power amplifier, which transmits close to 3 W at maximum output power in certain systems. Maintaining efficiency at such power is critical; traditionally, PAs have been designed in GaAs or InGaP.

Recent trends point toward CMOS power amplifiers that can potentially enable on-chip integration with the rest of the transmitter and lower system costs. However, challenges remain in terms of efficiency, thermal behavior and isolation.

Further Reading

Das, S.K. (2000) *Microwave Signals and Systems Engineering*, Khanna, New Delhi.
Stern, H.P.E. and Mahmoud, S.A. (2004) *Communication Systems: Analysis and Design*, Pearson Education, Upper Saddle River, NJ.
Varrall, G. and Belcher, R. (2003) *3G Handset and Network Design*, John Wiley Publishing, Hoboken.

5

Wireless Channel Multiple Access Techniques for Mobile Phones

5.1 Introduction to Multiple Access Techniques

As discussed in Chapter 1, for wireless communication we use air or free-space as the medium and an EM wave as the information carrier. However, there is a problem with this – the air channel can be considered as a single line or one communication link, so how can so many users be served using a single line? The answer is – we need to multiplex the same air channel between many different simultaneous channel users, so we will investigate the various technologies that are available for multiplexing the same air medium between different users. These are mainly: time division multiple access (TDMA), frequency division multiple access (FDMA), code division multiple access (CDMA), and orthogonal frequency division multiple access (OFDMA). We will discuss these access technologies in detail, and in later chapters we will explain how these are used in various mobile communication standards.

5.1.1 Time Division Multiplexing

In this technique, the same channel is shared by various users using different time slots. Each user has a specific time slot number, and when that particular time slot arrives, it uses the slot to send or receive its data. This is depicted in Figure 5.1.

In Figure 5.1, the same channel is time multiplexed between 8 users. When user-1 on the left-hand side (called-party) is connected to the channel, then at the same time user-1 (the calling party) on the right-hand side is also connected to the channel via the switch. So for a specific time duration these two users (left side user-1 and right side user-1) will be connected via the channel. Then in the next time slot, it will switch to slot 2, where the right side user-2 and the left side user-2 will be connected. Similarly, it will switch step by step up to 8, then again it will revert back to the slot-1 position. This technique of channel multiplexing is known as time division multiplexing (TDM) and this access technique is called time division multiple access (TDMA).

In time-division multiplexing (TDM), each signal is transmitted for a certain period of time, and then is "turned off" and the chances are then given to the next user's signal.

In particular, for each time slot, only one user is allowed either to transmit or receive. After N number of slots, the pattern repeats again, this time period is referred to as one frame, for example, one time frame contains several time slots (Figure 5.2). The characteristics of TDMA are: (a) non-continuous

Mobile Handset Design Sajal Kumar Das

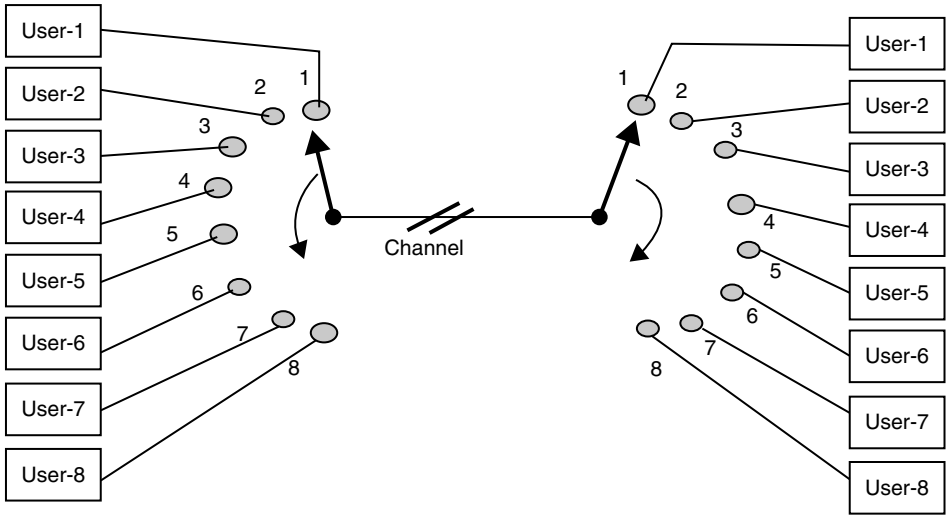


Figure 5.1 Time division multiplexing of the same channel

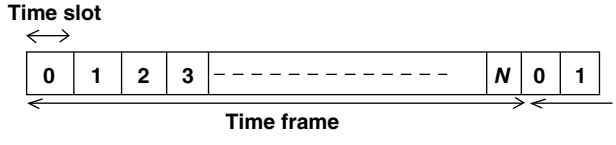


Figure 5.2 Different time slots

transmission – buffer and burst, (b) generally, digital data and digital modulation techniques are used, (c) a certain amount of guard time is required, which allows for different propagation delays between the mobile and BS, (d) 20–30% data overheads for controlling and maintaining the channel and multiplexing, and (e) trade-offs in data overhead, size of data payload, and latency.

Some advantages of TDMA techniques are: (a) they can share a single carrier frequency with multiple users, (b) non-continuous transmission makes handoff simpler, (c) slots can be assigned on demand (concatenation and reassignment) – bandwidth supplied on demand, for example, many slots can also be allocated to the same user according to the demand (for example, GPRS multi-slot class), (d) less stringent power control due to reduced inter-user interference, and (e) suitable for digital system implementation.

Some disadvantages of TDMA techniques are: (a) higher synchronization overhead, (b) equalization necessary for high data rates due to multi-path, (c) slot allocation complexity, and (d) pulsating power envelop interferes with other devices.

5.2 Frequency Division Multiplexing

As discussed in Chapter 1, when an EM wave is generated from the source during this time, based on the oscillation of the charge, the frequency of the generated wave will be different. So by using different oscillations, we can generate EM waves of different frequencies. When these waves pass via the channel,

if they have different frequencies then ideally their energies will not mix with each other. Thus, waves with different frequencies can carry different information through the same channel at the same time (Figure 5.3). This method of many users using the same channel at the same time using different frequencies is called frequency division multiplexing (FDM), and this multiple access technique is called frequency division multiple access (FDMA).

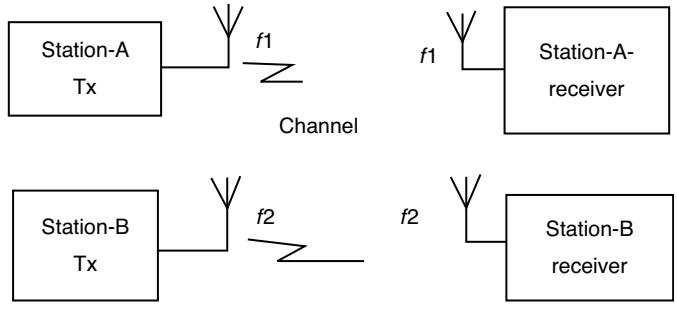


Figure 5.3 Various stations transmit and receive at the same time using different frequencies

Some specific characteristics of FDMA are: (a) if the channel is not in use, it remains idle, (b) based on the assigned channel bandwidth, the systems are classified into two categories – narrow-band systems (such as GSM) and wide-band systems (such as WCDMA), (c) generally, if symbol time $\gg$ average delay spread, in that case little or no equalization is required, (d) if only FDMA is used, then it is best suited for analog links, (e) requires sharp filtering to minimize interference, and (f) usually combined with FDD for duplexing.

All EM waves travel with the same velocity ($c = 3 \times 10^{10}$ cm/s) in free space, but their frequencies are different. The arrangement of the EM radiation according to wavelength or frequency ($c = f^{*}\lambda$) is called the EM spectrum. The EM spectrum has no definite upper or lower limit and various regions of the EM spectrum do not have sharp defined boundaries. The EM spectrum and their applications are given in Table 5.1, whereas in Table 5.2, the various frequency bands used for cellular and wireless systems are mentioned.

Sectorization, directional antennas, powerfull modulation, coding schemes, information compression, bit rate control, and improved algorithms in channel assignments allow an increase in the cell capacity of a cellular system that uses TDMA or FDMA based multiplexing.

Besides the fact that both TDMA and FDMA have some advantages as well as some disadvantages, TDMA is still preferred, because it allows us to take advantage of the digital processing technology. TDMA is actually the great competitor of a newly popular multiple access technique, called CDMA (code division multiple access).

5.3 Duplexing Techniques

For bi-directional communication, many users want to send data as well as receive data. A half-duplex system provides communication in both directions, but only one direction at a time (not simultaneously). An example of a half-duplex system is a two-party system, such as a "walkie-talkie." A full-duplex system allows communication in both directions, and unlike the half-duplex, it allows this to happen simultaneously.

For two-way communication, the user requires two channels simultaneously, one for sending (uplink/reverse link) and the other for receiving (downlink/forward link) the information. So any mobile phone

Table 5.1 Electromagnetic spectrum

Name of EM wave	Frequency range (Hz)	Source	Application
Electric wave	60–50	Weak radiation from ac circuits	Lighting
Radio waves	3^*10^9–3^*10^4	Oscillating circuits	Radio comm., television, mobile comm.
Microwave waves	3^*10^{11}–3^*10^8	Oscillation circuits, Gunn, IMPATT, tunnel diodes	Radar, TV, satellite, remote sensing, mobile comm.
Infra-red	4^*10^{14}–1^*10^{13}	Excitation of atom and molecules	Low range data comm., remote sensing
Visible	8^*10^{14}–4^*10^{14}	Excitation of atoms and vacuum spark	To study information on structure of molecules, optical comm..
UV rays	1^*10^{16}–8^*10^{14}	Excitation of atoms and vacuum spark	External atomic electron shell study, remote sensing, night vision
X-rays	3^*10^{19}–1^*10^{16}	Bombardment of high atomic number target by electrons	X-ray therapy, radiography, crystallography
Gamma rays	5^*10^{20}–3^*10^{19}	Emitted by radioactive substance	Study of information about atom, nuclear reactor
Cosmic rays	$>10^{20}$	From other stars and galaxies	Not used yet

user requires a duplex channel. The duplex channel can be provided to that user by multiplexing the available channels for sending and receiving. This technique is called duplexing and this is done by time or frequency multiplexing.

5.3.1 Frequency Division Duplexing (FDD)

Frequency division duplexing (FDD) provides two distinct bands of frequency for every user – one for sending and the other for receiving. The forward band provides the traffic from the base station (discussed later) to the mobile and the reverse band provides the traffic from the mobile to the base station.

5.3.2 Time Division Duplexing (TDD)

Time division duplexing (TDD) uses time instead of frequency to provide both a forward and a reverse link, for example, sending and receiving are separated by a particular time duration (this duration is very small which is why the users do not apparently experience it).

5.4 Spectral Efficiency

The spectral efficiencymeasures how the multiple access technique used in a wireless system allows a better use of the bandwidth with respect to the available time or frequency resource. Another parameter is spectral efficiency with respect to modulation, which measures how efficient the radio plan or modulation scheme is. The link spectral efficiency is measured in bit/s/Hz, and is the channel capacity

Table 5.2 Different wireless communication systems/standards and their respective frequency bands and associated parameters

Communication standard	Allocated frequency bands (MHz)	Data rate (over air interface)	Medium access technologies	Modulation used	Number of channels	Number of carriers	Channel BW
AMPS	824–849 869–894	10 kbps	FDMA/FDD	FM	1	832	30 kHz
PCS (IS-54/IS-136)	824–849 869–894 1850–1910 1930–1990	48 kbps	TDMA/FDD	DQPSK	3	832	30 kHz
IS-95	824–849 869–894	1.2288 Mbps	DS-CDMA	QPSK, O-QPSK	256	1	1.25 MHz
GSM	890–915 935–960 1710–1785 1805–1880	270 kbps	FDMA/TDMA FDD	GMSK	8	124	200 kHz
PDC	810–826 940–956 1429–1453 1477–1501	42 kbps	TDMA	DQPSK	3	640	25 kHz
DECT	1880–1930	1.152 Mbps	TDMA/TDD	GFSK	12	10	1.728
PHS	1895–1918.1	384 kbps	TDMA/TDD	DQPSK	4	77	300 kHz
Bluetooth	2400–2483.5	1 Mbps	FH-CDMA TDD	GFSK	79	79	1 MHz
UMTS/WCDMA	1900–2025 2110–2200	3.84 Mbps	DS-CDMA	QPSK	4...256	1	5 MHz
CDMA-2000	422.5–457.475 462.5–467.475	1.2288 Mbps 3.6864 Mbps	DS-CDMA	QPSK/BPSK	4...128 4...256	1 3	1.25 MHz
WLAN 802.11a	5150–5350 5425–5675 5725–5875	54 Mbps	OFDM	BPSK, QPSK, 16/64 QAM	12	52	20 MHz

(*continued*)

Table 5.2 (*Continued*)

Communication standard	Allocated frequency bands (MHz)	Data rate (over air interface)	Medium access technologies	Modulation used	Number of channels	Number of carriers	Channel BW
WLAN 802.11b	2400–2484	11 Mbps	DS-CDMA	CCK	13	13	5 MHz
WLAN 802.11g	2400–2497	54 Mbps	OFDM	BPSK/QPSK CCK	12	52	30 MHz
Hiperlan 2	5150–5350 5470–5725	54 Mbps	OFDM	BPSK/QPSK 16/64 QAM	5	52	20 MHz
802.15.4 (ZigBee)	868.0–868.6	20 kbps	DS-CDMA	BPSK	1	3	2 MHz
	902.0–928.0	40 kbps		BPSK	10		5 MHz
	2400–2483.5	250 kbps		OQPSK	16		
UWB	3100–10 600	480 Mbps	OFDM	QPSK	255	3	528 MHz
WiMAX (802.16d-WMAN/802.16e -Mobile WMAN)	2–11 GHz/2–6 GHz	Up to 75 Mbps/Up to 30 Mbps	OFDMA	QPSK, 16 QAM, 64 QAM			1.25–20 MHz/ 1.25–20 MHz

or maximum throughput of a logical point-to-point link with a given modulation method. If a forward error correction code is combined with the modulation, a "bit" refers to a user data bit; the FEC overhead is always excluded.

For example: GSM allows a reuse factor of 3–4 cells per cluster. The number of physical channels in a 25 MHz bandwidth ($2 \times 25 = 50$ MHz for uplink and downlink) GSM-900 is 124 carriers $\times$ 8 channel/carrier = 992.

$$\text{Spectral efficiency} = \text{total number of channels available in the system/}$$
$$(\text{cells*spectrum BW}) = 992/(4 \times 50) = 4.96 \text{ or } 5 \text{ calls/MHz/cell}$$

5.5 Code Division Multiple Access

Code division multiple access (CDMA), as the name indicates, is where the communication medium (for example, air) is simultaneously accessed by various users using different codes, for example, each user's data are separated by different codes. Put more simply, we can say that each user has a separate code and whenever he/she wants to send information, the user information is multiplied by that special code and then transmitted through the medium. The intended receiver, which knows about that code, can only extract the information from the received signal. These special codes are known as orthogonal codes.

So, code division multiplexing (CDM) allows signals from a series of independent sources to be transmitted at the same time over the same frequency band. As mentioned, this is accomplished by using orthogonal codes to spread each signal over a wide, common frequency band. Then at the receiver, the appropriate orthogonal code is used to recover the particular signal intended for a particular user.

5.5.1 Spectrum Spreading Concepts

Generally, the digital data pulse in the time-domain looks like a square wave pulse as shown in Figure 5.4a. If we see this square wave in the frequency domain, the spectrum will look like a sync pulse (using a Fourier transform). This is shown in Figure 5.4b.

$$F(\omega) = \int f(x)\mathrm{e}^{-j2\pi\omega t}\mathrm{d}t = \int_0^T A\mathrm{e}^{-j2\pi\omega t}\mathrm{d}t = \frac{-A}{2\pi j\omega}\langle \mathrm{e}^{-j2\pi\omega t}\rangle_0^T = \frac{A}{\pi\omega}\sin(\pi\omega T)\mathrm{e}^{-j\pi\omega T} \tag{5.1}$$

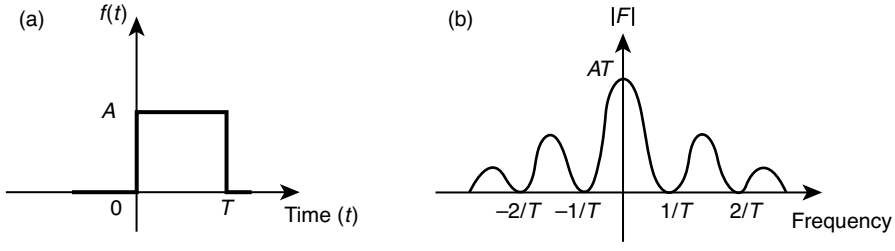


Figure 5.4 Square wave pulse and its Fourier transform (power spectrum)

Apart from the main lobe of the sync pulse, there should also be some side lobes and theoretically the wave continues to infinity. We know the total energy (here it can be taken as the total area under the pulse curve) in both these cases should be constant. In the transformed sync pulse, most of the energy lies in the main lobe curve only. One important thing to notice here is that if the data pulse's period (T) is decreased, for

example, the data rate is increased (smaller pulse width) then $1/T = f$ increases, which implies the sync pulse's cut-off frequency, for example, bandwidth or the position where the main lobe touches the zero on the frequency axis, also increases. Hence the sync pulse is stretched along the frequency axis. As mentioned earlier, to keep the total energy constant in the sync pulse, the peak amplitude value ($A.T$) will also reduce and the spectrum will be flatter as compared with the previous one. This indicates the spectrum is spread across the frequency scale, and the bandwidth is increasing. The higher the data rate the more the spreading and the more the bandwidth over which it will be spread. This property of spreading the signal energy across the frequency band is known as spreading (Figure 5.5). This technique is known as spread spectrum technology.

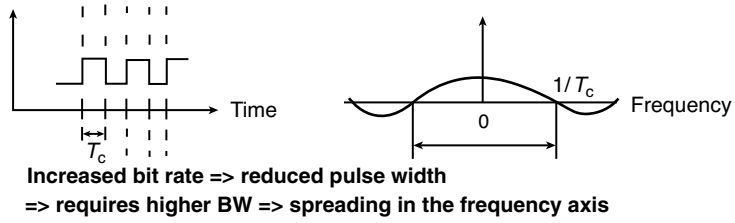


Figure 5.5 Concept of spreading

The next question is how to utilize this concept in a practical scenario. This is very simple: just multiply your main data stream (which has a lower rate, for example, wider pulse) with a high data rate code (narrower pulse). We know that if we have data D (with a rate of $1/T_d$) and high bit rate code C (with a rate of $1/T_c$) and if they are multiplied, then the rate of the resultant signal will be the maximum of these two, for example, $1/T_c$ (as $1/T_c > 1/T_d$). So the resultant signal will follow the high rate code signal (Figure 5.6), that means, the data signal will be spread by the code signal.

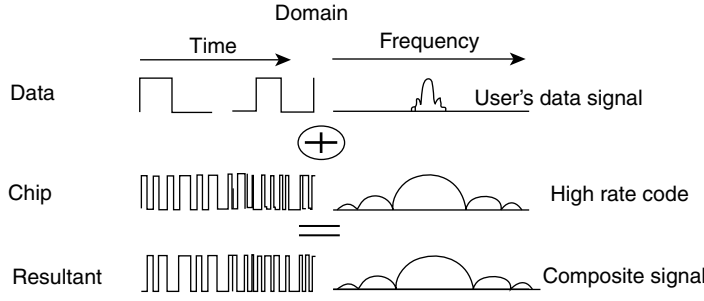


Figure 5.6 Data, chip, and spread signal

Each bit of the code used for spreading is known as a chip and the rate of the code is called the chip rate. Generally, the higher the chip rate, the higher the resultant data spreading in the frequency domain. The ratio of the chip rate to the data rate is known as the spreading factor: spreading factor = chip rate/data rate.

The spreading factor is an indication of how well the data are spread across the frequency. If they are spread more (for example, spreading factor is high as in a high chip rate) then the total energy for each data signal of any transmitter will be spread over a large frequency band. This requires a wide bandwidth, and is known asa wideband spread spectrum. If the chip rate is not very high then the spreading factor will also not be very high and the required bandwidth will not be particularly large; this is known as a narrow-band

spread spectrum. The advantages and disadvantages of spreading over a narrow bandwidth compared with a wide bandwidth will be discussed later. Before that, we will see how this spreading code not only helps to spread the energy over a bandwidth, but also assists in sending and receiving the data signal over the medium without being mixed up with the others user's send signal (at the same time and at the same frequency).

As stated earlier, spread spectrum systems afford protection against jamming (intentional interference) and interference from other users in the same band as well as noise, by "spreading" the signal to be transmitted and performing the reverse operation "de-spreading" on the signal that arrives at the receiver.

5.5.2 Mathematical Concepts

CDMA exploits the core mathematical properties of orthogonality. Let us represent data signals as vectors. For example, data = 1011, as the binary string "1011" would be represented by the vector (1, 0, 1, 1). We may give this vector a name, for example **a**. We can use an operation of dot product of vectors, to "multiply" the vectors, by summing the product of the components. For example, the dot product of vector **a** (1, 0, 1, 1) and **b** (1, −1, −1, 0) would be (1)(1) + (0)(−1) + (1)(−1) + (1)(0) = 1 + (−1) = 0. If the dot product of vectors **a** and **b** is 0, then we say that the two vectors are orthogonal. The dot product has a number of properties, and understanding this will help us to know how CDMA works.

For vectors **a**, **b**, **c**: a.(b + c) = a.b + a.c and a.kb = k(a.b), where k is a scalar. The square root of **a**.**a** is a real number: $\|a\| = \sqrt{(a.a)}$. Suppose vectors **a** and **b** are orthogonal, then, **a.b** = **0**, and **a.a** = **a**, these are very important properties used in the CDMA. This implies that if two different orthogonal codes are multiplied the resultant will be zero, but if two same orthogonal codes are multiplied then the maxima will be achieved.

$$a.(a+b) = \|a\|^2 \text{ as } a.a. + a.b = \|a\|^2 + 0$$

$$a.(-a+b) = -\|a\|^2 \text{ as } -a.a + a.b = -\|a\|^2 + 0$$

$$b.(a+b) = \|b\|^2 \text{ as } b.a + b.b = 0 + \|b\|^2$$

$$b.(a-b) = -\|b\|^2 \text{ as } b.a - b.b = 0 - \|b\|^2$$

Thus if the code vectors are orthogonal then their dot product will be zero. In digital communication language we can say that if two codes are orthogonal then their cross correlation will be zero.

5.5.3 Correlation

Correlation is a measure of how two entities are related. A high correlation means that there is a lot of resemblance between the two compared entities.

5.5.4 Auto-Correlation

This is the measure of correlation between the same signals but where one is a delayed version. The auto-correlation for a periodic signal of period T is defined as follow:

$$R_i(\tau) = \frac{1}{T} \int_{-T/2}^{T/2} S_i(t) S_i(t-\tau) \mathrm{d}t \tag{5.2}$$

It defines how much a function correlates with a time-shifted version of itself, with respect to that time shift.

5.5.4.1 Cross-Correlation

The cross-correlation for periodic signals of period T is defined as:

$$C_{ij}(\tau) = \frac{1}{T} \int_{-T/2}^{T/2} S_i(t) S_j(t-\tau) \mathrm{d}t \tag{5.3}$$

It measures how much two different signals, S_i and S_j (where one is time shifted from the other), correlate with each other.

5.5.5 Orthogonality

Orthogonality between signals is a very important concept in communication systems. We say that two periodic signals of period T are orthogonal, when their cross-correlation is null for a zero time shift:

$$\int_{-T/2}^{T/2} S_i(t) S_i(t-\tau) \mathrm{d}t = 0 \tag{5.4}$$

Therefore, two orthogonal signals can be transmitted at the same time and will not interfere with each other. This principle is largely applied in CDMA. The orthogonal codes are derived by using a Hadamard matrix.

5.5.5.1 Hadamard Matrix

A Hadamard matrix is a square matrix whose entries are either $+1$ or -1 and where the rows are mutually orthogonal. The definition that a Hadamard matrix $\boldsymbol{H}$ of order n satisfies:

$$\boldsymbol{H}^T \boldsymbol{H} = nI_n$$

If $\boldsymbol{H}$ is a complex matrix of order n, whose entries are bounded by $|\boldsymbol{H}_{ij}| \leq 1$, then Hadamard's determinant bound states that $|\det(\boldsymbol{H})| < \text{or} = n^{n/2}$. Equality in this bound is attained for a real matrix $\boldsymbol{H}$ if and only if $\boldsymbol{H}$ is a Hadamard matrix. The order of a Hadamard matrix must be 1, 2, or a multiple of 4.

Sylvester first constructed Hadamard matrices in 1867. Let $\boldsymbol{H}$ be a matrix of order n. Then the matrix of order 2^n will be:

$$\begin{bmatrix} \boldsymbol{H} & \boldsymbol{H} \\ \boldsymbol{H} & -\boldsymbol{H} \end{bmatrix} \tag{5.5}$$

This can be repeated, which then leads to Walsh matrices.

$$\boldsymbol{H}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$\boldsymbol{H}_{2^k} = \begin{bmatrix} \boldsymbol{H}_{2^{k-1}} & \boldsymbol{H}_{2^{k-1}} \\ \boldsymbol{H}_{2^{k-1}} & -\boldsymbol{H}_{2^{k-1}} \end{bmatrix} = \boldsymbol{H}_2 \otimes \boldsymbol{H}_{2^{k-1}}$$

For $2 \leq k \varepsilon N$, where $\otimes$ denotes the Kronecker product.

5.5.6 Implementation

In a simpler way, this behavior can be modeled digitally as below. Suppose there are two orthogonal codes C1 = (0,0) and C2 = (0,1), and two users A and B have data A = (1,0) and B = (0,1) respectively. If C1 is used by A and C2 by B then:

```
C1 XOR A = (0,0) XOR (1,0) = 1,0
C2 XOR B = (0,1) XOR (1,1) = 1,0
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
          SUM = 1,0
```

which will be sent to the receiving party and after receiving it, the intended receiver, who has code C1 will get the data = SUM XOR C1 = (1,0) XOR (0,0) = (1,0) = a = data of A. Similarly, the receiver who has code C2 will get the data = SUM XOR C2 = (1,0) XOR (0,1) = (0,1) = b = B. So, the transmitted data by both the parties A and B, which was sent simultaneously (and was mixed up) are separated correctly at the receiver.

In practice, the data level *zero* is represented by 1 and data level *one* is represented by -1, and the resultant signals are summed up (modulo two addition). On this basis, some examples are shown below.

1. Suppose we have two senders A and B, both sending simultaneously. The data from A is 1 (=> -1) and from B is 0 (=> $+1$). The orthogonal code for A is Ca = (0,0) and for B is Cb = (0,1). Thus, the spreading sequence for A = $+1$, $+1$ and for B = $+1$, $-1$is:

   ```
   Transmitted sequence from A will be => (-1).(+1, +1) = -1,-1
   Transmitted sequence from B will be => (+1).(+1,-1) = +1,-1
   - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
   ```

   The resultant signal in the medium during both transmissions will be = 0, -2
   Once it has arrived at the intended receiver of A, it will multiply the resultant signal by the code Ca ($+1, +1$), for example., (0,-2). ($+1, +1$) = $0 - 2 = -2$. As it is negative then the data signal will be 1. Similarly, for the intended receiver of B the signal will be (0,-2).($+1, -1$) = $0 + 2 = +2$, As it is positive then the data will be 0. Hence the data is recovered correctly from the resultant mixed signal that was transmitted over the air.
2. Orthogonal code (0,0) and (0,1) and data (1,0) and (0,0). The converted sequence will be: code =>(1,1) and (1,-1) and data => (-1,1) and (1,1). Each data bit is treated as a symbol and the chips of the orthogonal code are multiplied to that. After the multiplication, the resultant bit numbers for each data bit will be the same as orthogonal code's bit number, for example, here each data bit will converted into two bits as two bit orthogonal codes are used. Because two data bits are taken in the sequence, so a total of $2 \times 2 = 4$ bits will be generated for each sender.

   ```
   (-1,1).(1,1) =-1,-1, 1,1
   (1,1).(1,-1) = 1,-1, 1,-1
   - - - - - - - - - - - - - - - - - - - - - - - - -
   SUM = 0,-2 , +2,0
   ```

   The signal levels will simply be added into the medium by adding the signals from many senders. At the intended receiver A side, multiply this with the orthogonal code:

   ```
   0,-2 , 2,0
   1, 1, 1,1 (Ca)
   - - - - - - - - - - - - - - - - - - - - - (add)
   0,-2 , +2, 0 =(-2, 2).
   ```

   So the data will be => 1, 0.
   Similarly, for the intended receiver B side,

```
0, -2 , 2,0
1, -1, 1,-1 (Ca)
- - - - - - - - - - - - - - - - - - - - (add)
0,+2 , +2, 0 => (2, 2).
```

As both are positive this means that the data will be => 0, 0.

3. Data sequence of A = (. . .. . ., 1,0) => (. . .. . ., −1,1), data sequence of B = (. . .. . .. .1,1) = (−1,−1)

```
Orthogonal code séquence 1=(0,1,0,1)=>(1,-1,1,-1),
Code séquence 2 = (0,1,1,0)=>(1,-1,-1,1)
(..... . .,-1,1) . (1,-1,1,-1)=(.....,-1,1,-1,1,     1,-1,1,-1)
(..... . .,-1,-1) . (1,-1,-1,1)=>(.....,-1,1,1,-1,  -1,1,1,-1)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
(..... .,-2,2,0,0,   0,0,2,-2) (Summed in the air medium).
At intended receiver of A (code seq)-> .(1,-1,1,-1), (1,-1,1,-1)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Resultant sequence at receiver A=>(..... .(-2-2+0+0) , (0+0+2+2))
(.......... -4 , +4 ) => (..... . .,1,0)
At intended receiver of B -> (..... . .,-2,2,0,0,    0,0,2,-2)
(summed in the air medium).
Code sequence of B =>(1,-1,-1,1), (1,-1,-1,1)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Resultant sequence at receiver B=>(..... .(-2-2+0+0) , (0+0-2-2))
(..... . .,-4 ,-4 ) => (.....,1,1).
```

In the above examples only two senders have been taken, and only two orthogonal codes are used. This can be extended to an increased number of senders and receivers.

5.5.6.1 Digital Circuit Implementation

We can implement this above mathematical concept into a digital circuit and illustrate the spreading and de-spreading process in the frequency domain. As shown in Figure 5.7, the sender's data source is for example A, and this is passed through the XOR circuit for spreading of the source code by orthogonal chip

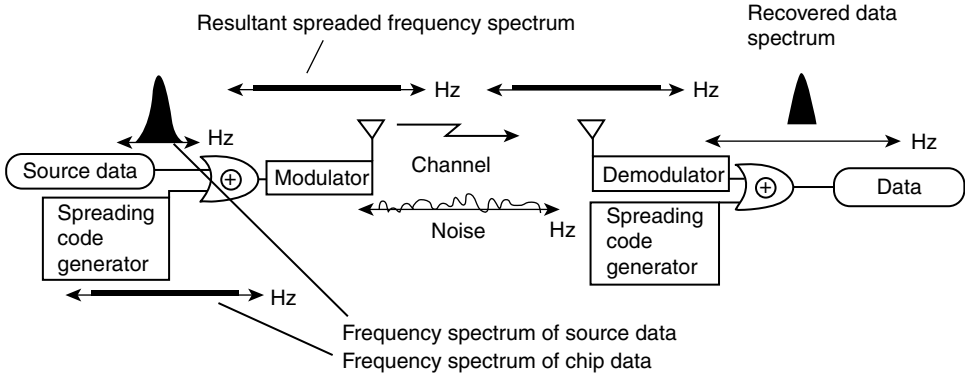


Figure 5.7 CDMA access techniques

code, such as C. The resultant digital signal R is digitally modulated then RF up-converted and sent via the air. So R = A XOR C. When it is transmitted through the air, apart from the noise signal (N), signals from some other simultaneous senders (for example, S1, S2, ...) will also add up in the air medium. On the receiver side this composite signal is received and first RF down-converted, then digitally demodulated after that the de-spreading operation is done through the XOR operation on the resultant received signal (R + N + S1 + S2) with the appropriate same orthogonal code C as used by the transmitter:

```
Received Data at receiver = (R + N + S1 + S2) XOR C = R XOR C + N XOR C + S1
XOR C + S2 XOR C = (A XOR C ) XOR C + N XOR C + (A1 XOR C1) XOR C + (A2 XOR C2)
XOR C = A . 1 + 0 + 0 + 0 = A.
```

where C1, C2 are the orthogonal code used by the other senders.

As discussed earlier, from the orthogonal property C XOR C = 1 but C XOR C1 = 0 or C XOR C2 = 0, for example, multiplying by another code will give a zero result whereas by the same code will give 1.

The spreading, transmitting, and de-spreading processes for a general signal corrupted in the channel by additive white Gaussian noise is shown in the Figure 5.7. A PN spreading code is used. Note that the spreading process flattens (as well as spreads) the spectrum of the information signal (Figure 5.8a), and that the spread transmitted signal is essentially buried in the noise by the time it arrives at the receiver. The de-spreading process narrows and heightens the spectrum of the received signal, yet leaves the spectrum of the received noise essentially unchanged (Figure 5.8b). The receiver can now extract the de-spread signal from the noise using a band-pass filter.

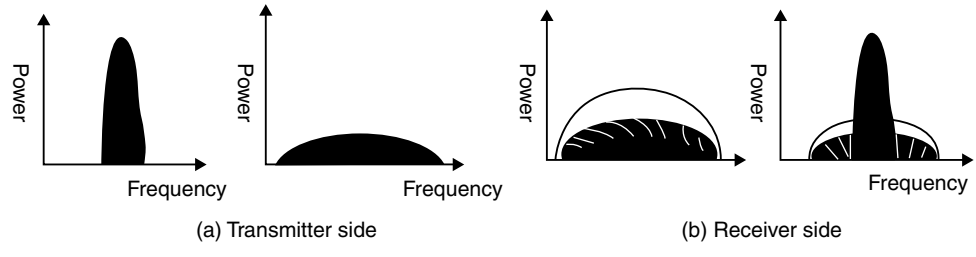


Figure 5.8 (a) The power spectrum of the transmitted data and spectrum after spreading. (b) Received power spectrum which is the sum of its own plus the other user's spread power and de-spread data's power spectrum

5.5.7 Multiple Access Using CDMA

Spreading can be used as a multiplexing technique by developing a series of orthogonal spreading codes. These are PN codes with the additional feature that if any two different orthogonal spreading codes are Exclusive-ORed bit by bit, the resulting series of bits will itself be a PN code, that is, it will look like a noise signal. Thus, if a signal is spread using one code and then de-spread using another orthogonal code, the result will be like a PN code and will have a power spectral density similar to wideband white Gaussian noise.

Consider the system shown in Figure 5.9, where we have taken three calling and called parties (for example, each sender has an intended receiver on the other side). Each of these three user pairs is employing mutually orthogonal spreading codes to transmit information over a common channel. Source A uses one spreading code (let's call it spreading code A) and transmits the spread spectrum signal as shown by medium black color in the Figure 5.9. This signal is wideband. From the viewpoint of the channel and any

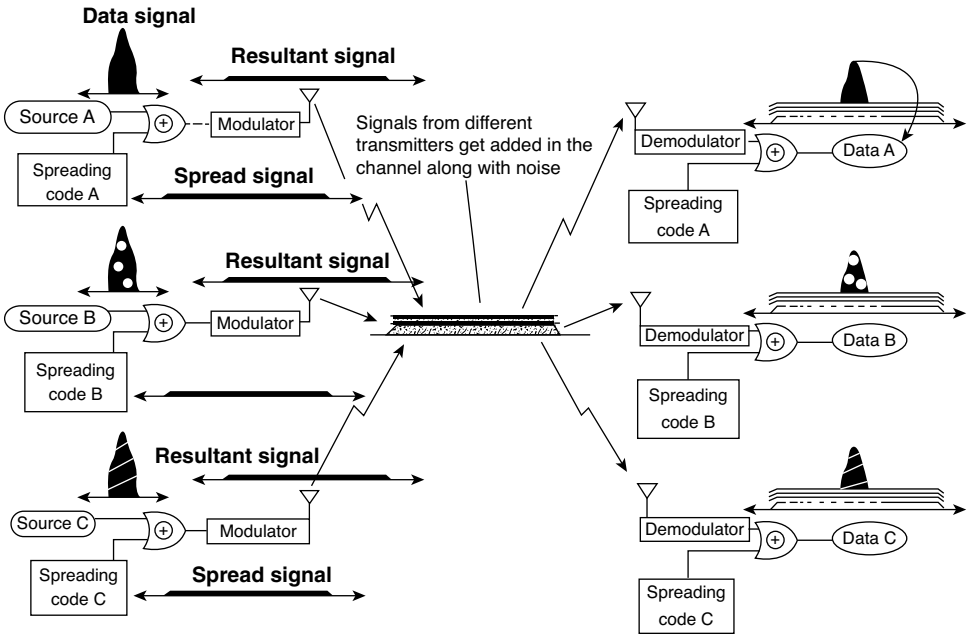


Figure 5.9 Transmission and reception using CDMA user multiplexing technique

receiver that does not know the code, the signal will look like additive white Gaussian noise. Source B uses a second, orthogonal spreading code (spreading code B) and transmits another spread spectrum signal, shown by the light black color in the figure. The same condition also applies here. Source C uses a third, orthogonal spreading code (spreading code C) to transmit a message, shown by the black color with white lines, across the channel. In the channel, all these signals will be summed up as they are transmitted at the same time using the same frequency. Some noise will also be added by the channel itself (this is the brick portion of the power spectral density in the channel in Figure 5.9).

Let us now look at the receivers. The receiver associated with user B (upper right corner in the figure) applies spreading code B to the total received signal. This de-spreads the portion of the signal transmitted from source B (the light gray part of the spectrum in the upper right corner of figure) but leaves all other portions of the received signal with a wideband noise-like power spectral density (the cross-hatched part of the spectrum). Using a narrow bandpass filter, user B may now extract the portion of the signal associated with source B, with the channel noise and the interference from users A and C significantly reduced. Similarly, user C may use spreading code C to extract its intended message from source C (see the spectrum in the center right part of the figure), and user A may use spreading code A to extract its intended message from source A.

Hence, as discussed, when using different orthogonal codes many user pairs can send and receive their data through the same channel, using the same frequency band, and at the same time. This technique is multiplexing of the air channel using different orthogonal codes.

5.5.8 Commercialization of CDMA

A CDMA signal uses many chips to convey just one bit of user information. The chip rate decides on the spreading of information and capacity of the system. In a commercial system there will be many users and

each user requires transmission and reception channels. So we need to separate the cells (for example, base stations), user's mobiles, and the various channels. Different codes need to be used, to create downlink (also called forward link) and uplink (reverse link) channels for different users. This implies that we require a huge set of orthogonal codes for this purpose, which is very expensive and difficult to design. So, for channelization and user separation, different sets of PN sequence codes are derived and used along with the orthogonal codes. Original CDMA channels are composed of combinations of three codes. A forward channel flows in the form of a specific Walsh code assigned to the user, and a specific PN offset for the sector. A reverse channel flows in the form of the mobile's specific offset of the long PN sequence code (Figure 5.10).

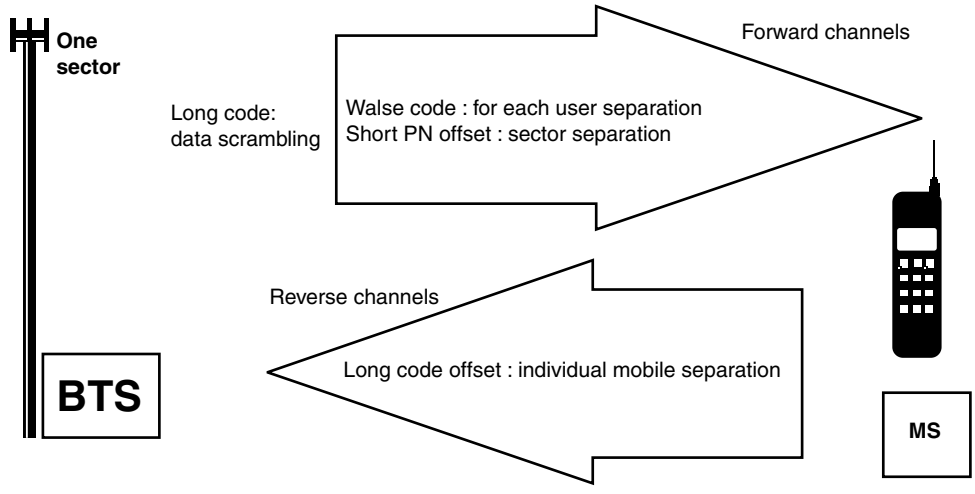


Figure 5.10 CDMA channels

Requirements for the spreading codes:

- **Good auto-correlation properties** – for separating different paths
- **Good cross-correlation properties** – for separating different channels.

Therefore the orthogonality can be achieved by first multiplying each user's binary input data by a short spread sequence, which is orthogonal to that of all users of the same cell. Then this spread signal is followed by multiplication of a long PN sequence, which is cell specific but common to all users of that cell in the forward link and user specific in the reverse link. The short orthogonal codes are called channelization codes; long PN sequence codes are called scrambling codes. Thus each transmission code is distinguished by the combination of a channelization code and a scrambling code.

In WCDMA standards two levels of spreading are used (see Figure 5.11):

1. Channelization code (short code) provides spreading, which means increasing bandwidth. Channelisation codes are used for channel separation from the same source. The same set of codes are generally used for all the cells.
   - Short codes: used for channel separation in uplink and downlink. This is basically *Walsh code* (using a Hadamard matrix) and derived from OVSF code structure of different lengths. However, these do not have good correlation properties, so need additional long code on top of them.

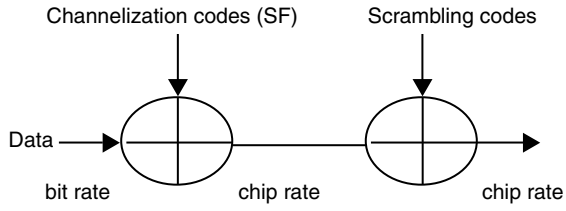


Figure 5.11 Scrambling and channelization

2. Scrambling code (long code) provides separation of users/cells, and does not affect the transmission bandwidth. These have good correlation properties. In uplink these are used for different user's separations, and in the downlink direction it is used for separating different base stations (cells). These are basically *gold codes.*

5.5.8.1 Spreading Code and Spreading Code Synchronization

There are certain requirements for spreading code: autocorrelation peak must have acute synchronization (time shift = 0), autocorrelation must be minimal in terms of absolute value when the time shift is not 0, and autocorrelation must be minimal in absolute value between different codes at all timings. A code that meets these requirements is gold code.

The Walsh code generated through a Walsh–Hadamard transform is also an orthogonal code with a period of the power of 2^n ($n >= 1$). The respective number of Walsh codes and orthogonal gold codes with a code length of SF is equal to SF. The application of these codes in a cellular system requires spreading code cell iteration, because of frequency reuse in the case of FDMA based system. As a result, the number of spreading codes that can be used in one cell will be limited, and therefore the system capacity can not be expanded. To make it possible and to use the same orthogonal code sequence repeatedly in each cell, two layers of spreading codes are assigned by multiplying the orthogonal code sequence, by scrambling codes with an iteration period that is substantially longer than the information symbol rate. The iteration period of the scrambling codes is one radio frame long (=10 ms in the case of UMTS), which is 38 400 chips long. It is assigned uniquely to each cell in downlink and to each user in uplink.

In order to extract the information data components, the destination mobile phone needs to execute the spreading code synchronization, which consists of two processes, namely, acquisition and tracking, in which tracking maintains the synchronization timing within ± 1 chip of acquisition. The de-spreader may be a sliding correlator or a matched filter (MF) with high speed synchronization capabilities. In WCDMA, a sliding correlator is generally used, while MF is often used at the first step of the three step cell search process (refer to Chapter 13).

5.5.9 Generation of a Scrambling Code

The spreading code is generated from a binary shift register. The pseudo random codes used are cyclic in nature. PN sequences are periodic sequences that have a noise-like behavior. Figure 5.12 illustrates the generation of PN sequences using shift registers, modulo-2 adders (XOR gates), and feedback loops.

The maximum length of a PN sequence is determined by the length of the register and the configuration of the feedback network. An N bits register can take up to 2^N different combinations of zeros and ones. As the feedback network performs linear operations, if all the inputs (that is, the content of the flip-flops) are zero, the output of the feedback network will also be zero. Therefore, the all-zero combination will always give zero output for all subsequent clock cycles, so we do not include it in the sequence. Thus, the

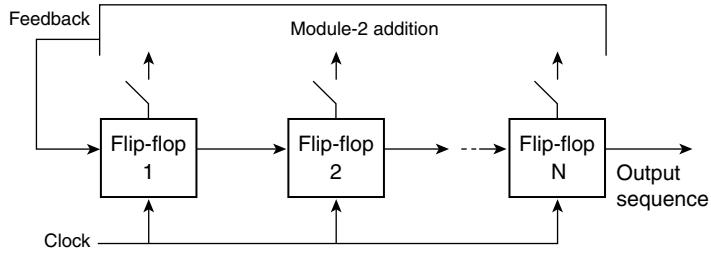


Figure 5.12 Binary shift register

maximum length of any PN sequence is $2^N - 1$ and sequences of that length are called maximum-length sequences or m-sequences. They are useful because longer sequences have better properties.

5.5.9.1 Gold Codes

In CDMA a multi-user environment needs a set of codes with the same length and with good cross-correlation properties. Gold codes are product codes generated by the XORing (modulo-2 addition) of two maximum-length sequences with the same length. The code sequences are added chip by chip by synchronous clocking. As m-sequences are of the same length, the code generators maintain the same phase relationship and the codes generated are of the same length as the two base codes, which are added together, but are non-maximal (Figure 5.13). Every change in the phase position between the two generated m-sequences causes a new sequence to be generated.

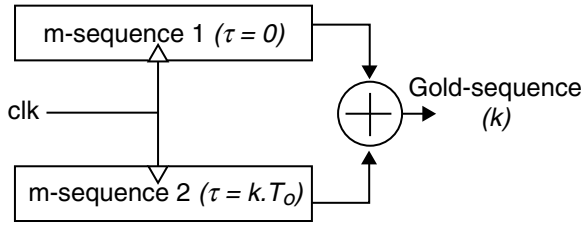


Figure 5.13 Gold code generator

Any two registered gold code generators of length L can generate $2^L - 1$ sequences plus the two base m-sequences, giving of a total of $2^L + 1$ sequences. In addition to their advantage in generating the large numbers of codes, the gold codes may be chosen so that over a set of codes available from a given generator the autocorrelation and the cross-correlation between the codes is uniform and bounded.

5.5.10 Process Gain

In a spread spectrum, the data are modulated by a spreading signal, which uses more bandwidth than the data signal. As multiplication in the time domain corresponds to convolution in the frequency domain, a narrowband signal multiplied by a wide band signal ends up being a wide band. In a general sense, we will see that the increase in bandwidth above the minimum bandwidth in a spread spectrum system can be thought of as applying gain to the desired signal with respect to the undesirable signals. We can now define the processing gainGP as, $GP = BWRF/BW_{\text{info}}$, where $BWRF$ is the bandwidth that the resultant signal

occupies and BW_{info} is the minimum bandwidth necessary to transmit the information or data signal. Processing gain can be thought of as the improvement over conventional communication schemes due to the spreading operation on the signal.

5.5.11 Different Types of Spread Spectrum Techniques

There are two types of spread spectrum techniques usually used in the communication system design. One is a direct sequence spread spectrum and other is a frequency hopping spread spectrum. These are discussed below.

5.5.11.1 Direct Sequence Spread Spectrum Techniques (DSSS)

The direct sequence spread spectrum (DSSS) technique is based on directly spreading and de-spreading the baseband data by means of a pseudorandom noise (PN) sequence. Here each bit to be transmitted is modulated by the PN sequence. A typical direct sequence transmitter is represented by the block diagram as shown in Figure 5.14.

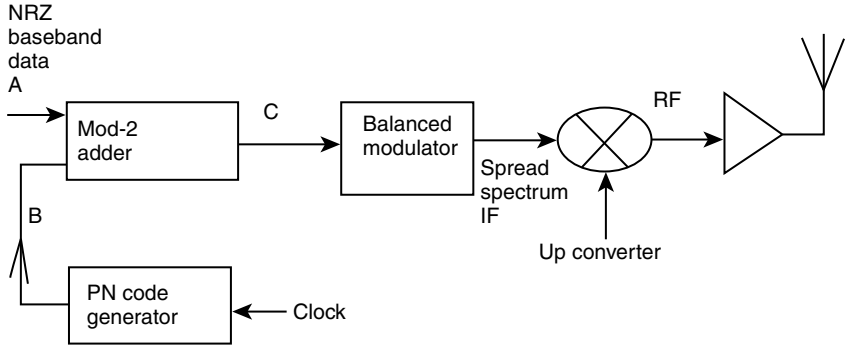


Figure 5.14 DS transmitter

Thus a DSSS transmitter is composed of a high-speed PN code generator, mod-2 adder, modulator, up-converter, power amplifier, and transmitting antenna. A DSSS transmitter converts the data stream into a symbol stream, where each symbol represents a group of 1 or more bits. These NRZ baseband data are represented, as "A" in Figure 5.15 and "B" is the high-speed PN sequence. The output of the mod-2 adder (exclusive OR gate) is C, which in a Boolean expression can be represented by:

$$\mathrm{C} = \mathrm{A}\bar{\mathrm{B}} + \bar{\mathrm{A}}\mathrm{B}$$

The respective waveforms and power spectrums are also shown in Figure 5.16.

Let the baseband data have a duration T_{d} in the time domain, which can be represented by:

$$V(t) = V \quad \text{for} \quad 0 < t < T_d$$
$$V(t) = 0 \quad \text{elsewhere}$$

This can be transformed into the frequency domain by Fourier transform:

$$S(\omega) = \int_0^T V(t).\mathrm{e}^{-j\omega t}.\mathrm{d}t = V.T_{\text{d}}\,[\sin(\omega T_d/2)/(\omega T_{\text{d}}/2)] \quad \text{for} \quad 0 < 1/f < 1/f_d \tag{5.6}$$

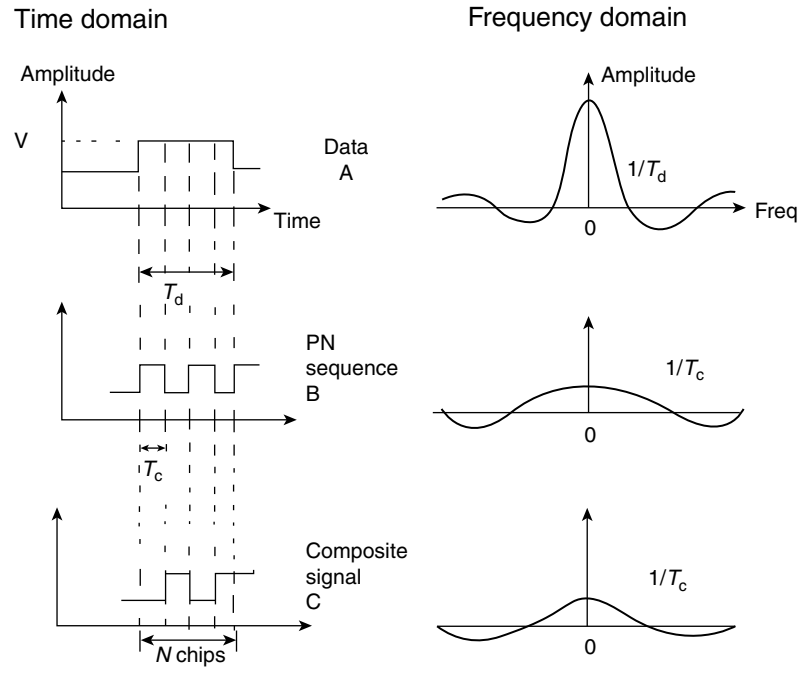


Figure 5.15 Time and frequency domain representation of data, PN sequence, and composite data

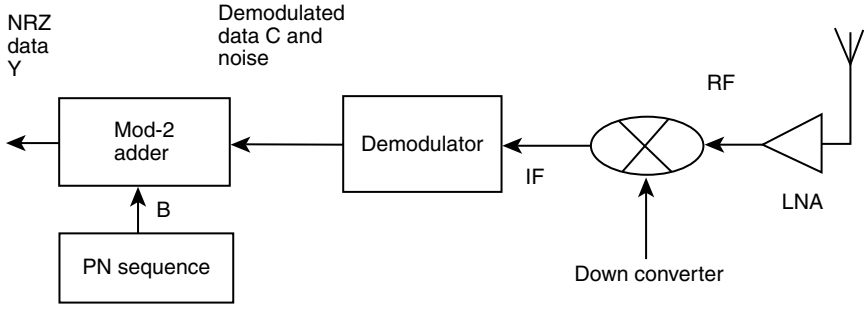


Figure 5.16 DS receiver system

and the power spectrum:

$$P(\omega) = (1/T_\mathrm{d})|S(\omega)|^2 = V^2\, T_d\, [\sin(\omega T_d/2)/(\omega T_d/2)]^2$$

Then the energy delivered within the time interval $t = 0$ to $t = T_\mathrm{d}$ will be:

$$E = \int_0^T P(\omega).\mathrm{d}t = (1/T) \int_0^T |S(\omega)|2.\mathrm{d}t \tag{5.7}$$

In the power spectrum, the main lobe is the power for the fundamental frequency and the other side lobes are the power for the other higher harmonics. In the frequency domain the signal crosses the zero value at $1/T$. Now if the time period decreases (that is, frequency $= 1/T$ increases), then it will cross zero at a latter instant. However, as the total energy remains constant, so the amplitude of the main lobe decreases, that is, the spectrum becomes more flat. The information signal is multiplied by the PN sequence with narrower pulses of time duration T_c, which is known as the chip time and code is called chip code. If the multiplicity factor is N then $T_d = N.T_c$. In Figure 5.16 the input signal A is low-speed NRZ data having a narrower power spectrum and the second input, B, is a high-speed PN sequence, which has a wider power spectrum and generates a composite signal C at the output, which will also has a wide-band power spectrum. However, as the total energy over a specified bandwidth is constant, so the amplitude of the main lobe in the composite signal power spectrum will be lower. The amplitude level can be decreased more by reducing the chip time (T_c) or increasing the value of N. As the resulting spectrum is spread over the available frequency band, if the BW increases then the amplitude of the signal reduces accordingly.

Ideally, the spreading code should be designed so that the chip amplitudes should be statistically independent of each other. The entire period of the PN sequence consists of N time chips. The total number of random sequences that can be generated by means of an m-bit shift register, will be $(2^m - 1) = N$. The mathematical properties of the PN sequence plays an important role in tha case of DSSS systems. A PN sequence consists of a series of plus and minus ones. It must posses certain auto-correlation properties. Periodic sequences that meet the criteria of pseudo-randomness are called Barker sequences and are known to exist only for short sequences. For this reason such sequences are typically too short for appropriate spreading of signals. In general, only the periodic sequences are of interest to the designers of DSSS systems.

The reverse process happens in the receiver side.

The received RF signal from the receiver antenna is amplified by LNA and then down converted by means of a mixer to obtain the IF signal. This IF signal is then demodulated to obtain composite high-speed data. This composite high-speed data are then added in the mod-2 adder with the same PN sequence code, B, to recover the desired signal A.

Using the Boolean expression – the output of the receiver mod-2 adder will be:

$$Y = C \overline{\oplus} \overline{B} = CB + \overline{B}\,\overline{C}.$$

$$= (\overline{A}B + A\overline{B})\,\overline{B} + (AB + \overline{A}\,\overline{B})\,B = A \tag{5.8}$$

Thus if both the transmitter and receiver PN sequences are identical then only the data can be recovered. One of the major advantages of this technique is that it reduces the effect of narrow-band sources of interference. One parameter that is used in specifying the performance of a spread spectrum signal in the presence of interference is known as processing gain (GP), as mentioned previously. This is defined as the ratio of signal bandwidth B_s (that is, transmission BW) to the message bandwidth B_m (information bandwidth) – $GP = B_s/B_m = 2t_m/t_c$, where t_c is the chip duration. The mean square value of the output interference signal (j_0^2) can be expressed as:

$$j_0^2 = J/GP \tag{5.9}$$

where J is the interference power. So if GP increases then j_0^2 decreases, which means intentional or unintentional interference can be reduced by increasing the length of the PN sequence, where B_j is the BW of the interfering signal. Now the output signal-to-noise ratio can be calculated as:

$$(S/N)_0 = P_r/(n_0^2 + j_0^2) \tag{5.10}$$

where n_0^2 is the mean square output noise and P_r is the receiver power:

$$(S/N)_0 = P_r/(N_0/2t_m + J/GP) = GP.P_r/((N_0/2).(B_s/B_m).B_m + J) = GP.P_r/(N_0B_s/2 + J) \tag{5.11}$$

The output SNR is proportional to the ratio of received signal power to the sum of interference and noise power in half of the SS bandwidth. Now if we define input signal-to-noise ratio $(S/N)_I = P_r/(N_0B_s/2 + J)$, then $(S/N)_0 = (GP)\,(S/N)_I$. Owing to the spreading of the signal there is a loss, which is called the process loss (PL). Here each user uses the same frequency band at the same time with different PN code, the resultant power spectrum in the frequency domain will appear as a low power wideband noise to an unintentional receiver. However, the noise-like power spectrum from each user will be added and the desired signal can be recovered after de-spreading from the noise-like power spectrum, as long as the power spectrum of the signal is greater than the sum of the interference and noise power level of the resultant spectrums.

One of the greatest advantages of the DSSS is its ability to reject jamming.

The jamming margin is expressed as:

$$AJ\ (\text{in dB}) = (GP)\ (\text{in dB}) - L\ (\text{in dB}) - (S/N)_0\ (\text{in dB}) \tag{5.12}$$

Wideband CDMA

From Figure 5.18, it is evident that in the time domain, the resultant bit rate, after spreading the data with chip becomes the same as the chip rate. So in the frequency domain the resultant signal spectrum follows the chip signal's spectrum. As the chip has a higher rate than data, so it occupies more bandwidth than data signal. As shown in Figure 5.15, the chip signal will occupy a bandwidth (BW_c) of almost $\sim 2^*1/T_c$, whereas the data signal will occupy a bandwidth (BW_d) of $\sim 2^*1/T_d$. From Figure 5.17, we can say that provided $P_d > (P1 + P2 + \ldots + P_n)$, for example, up until that point the data signal can be correctly recovered from the receivedsignal at the receiver. However, if the chip rate is less, then the signals from different users will not be stretched much in the frequency scale. Hence when only a few users are added into the system, the above equation will be violated, for example, the sum of the noise and interference levels from different users will be more than the energy from a single user's data. This issue can be resolved by stretching each user's signal energy further along the frequency axis, which leads to the reduction of noise and interference energy level from the various users, as shown in Figure 5.18. This is because now the same amount of energy should be residing over a greater area. As the length is now increased, so the thickness of each layer will reduce and this will allow to add on more layers. To stretch the signal energy for each user along the frequency axis, we need to extend the bandwidth and this technique is called wideband CDMA technology. To support many users (and a high data rate), the chip rate is thus increased, which in turn increases the required bandwidth.

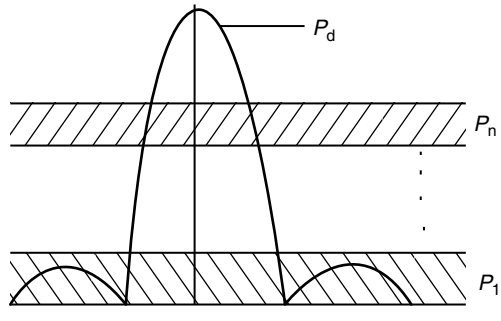


Figure 5.17 Interference power level from different users and the power spectrum desired data

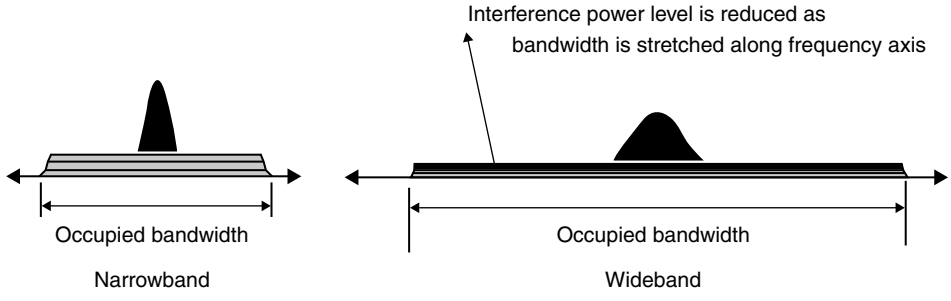


Figure 5.18 Narrowband CDMA (DSSS) and wideband system's power spectrum

5.5.11.2 Frequency Hopping Spread Spectrum (FHSS)

FHSS uses a narrowband carrier that changes the frequency in a pattern known to both the transmitter and the receiver. The transmitter changes the carrier frequency of the modulated information signal according to a certain "hopping" pattern. The hopping pattern is decided by the spreading code, where the spectrum of the transmitted signal is spread sequentially rather than instantaneously. As an example an FH signal is shown in Figure 5.19.

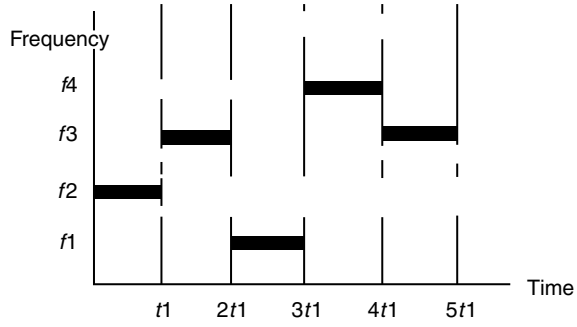


Figure 5.19 Various hopping frequencies generated at different time instances by the FHSS system

The sequence of frequencies appears to be random, but is predetermined and tracked by both the transmitting and receiving stations. A block diagram of a typical FHSS transmitter and receiver system is shown in Figure 5.20. Frequency hopping systems can be divided into fast hop or slow hop. In a slow hop system the hopping rate is smaller than the message bit rate and in a fast hopping system the hopping rate is greater than the message bit rate.

With m as the length of the PN sequence, the data signal is first baseband modulated. The modulation techniques that are commonly used for this are: FSK, MFSK, or GFSK. The resulting modulated wave and the output from a digital frequency synthesizer are then applied to a mixer followed by a filter. The filter is designed to transmit the sum frequencies component resulting from the multiplication process. The frequency synthesizer accepts m binary digits, which are mapped into one of $M = 2^m$ frequencies. The frequency synthesizers are unable to maintain phase coherence over successive hops, thus most of the frequency hop spread spectrum communication systems use non-coherent modulation. In the receiver, using a locally generated code sequence, the received signal is converted down to the baseband. The output of the frequency synthesizer is synchronously controlled in the same manner as in the transmitter.

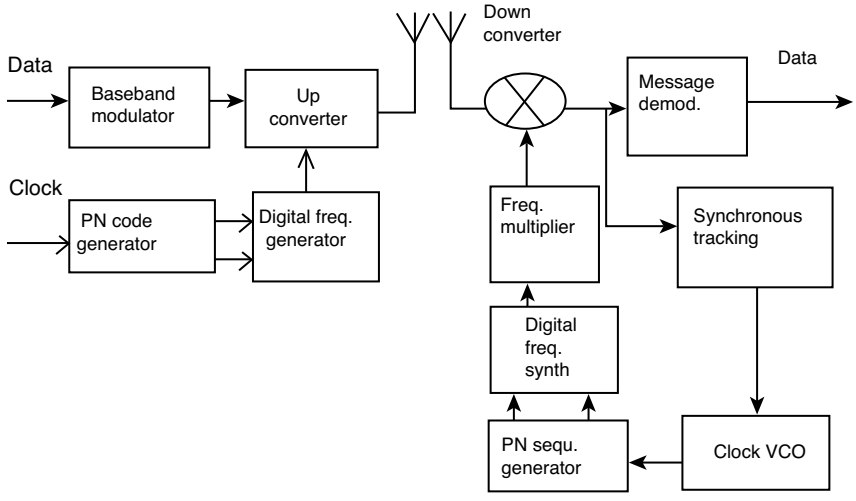


Figure 5.20 FHSS transmitter and FHSS receiver

Then the output is demodulated to obtain the desired data. An individual modulated tone of the shortest duration is referred to as a chip.

For slow hopping the information bandwidth is the main factor that decides the occupied bandwidth, and in the case of fast hopping the pulse shape of the hopping signal will decide the occupied bandwidth at one hopping frequency. If ωc is the bandwidth and R_s is the symbol rate here, then the process gain will be: $GP = \omega c/R_s = 2^m$, where m is the length of the PN segment employed to select a frequency hop.

5.5.11.3 Comparison Between DSSS and FHSS

1. The frequency occupation of an FHSS system differs considerably from a DSSS system. A DSSS system occupies the whole frequency band when it transmits, whereas an FH system uses only a small part of the available whole bandwidth during transmission, but location of this part differs in time. DSSS radios require no guard band associated with the frequency channel. For this reason it is possible to have larger numbers of DSSS radios operating in the same bandwidth at the same time, that is, a more efficient channel bandwidth.
2. FHSS systems are much easier to implement than DSSS systems and use a low cost microprocessor to control the frequency hopping functions. In comparison, DSSS systems employ extensive digital circuitry to implement the encoding and decoding algorithms. However, FHSS systems need a highly sophisticated frequency synthesizer for operation, whereas the DSSS frequency synthesizer is very simple, as only one carrier frequency has to be generated.
3. Synchronization is much easier in FHSS systems than in DSSS systems. FHSS synchronization has to be done within a fraction of hop time. As the spectral spreading is not obtained by using a very high hopping frequency but by using a large hop-set, the hop time will be much larger than the chip time of a DSSS system. Thus an FH-CDMA system allows a larger synchronization error.
4. As FHSS systems are essentially narrowband, so the rejection of interfering signals up to 70 dB higher than the desired signal can be attained, whereas DSSS systems reject interfering signals as a function of their "processing gain", which is about 20–30 dB.
5. The near–far problem is severe in DSSS systems. In FHSS the probability of multiple users transmitting using the same frequency band at the same time is less, so the base station will be able to receive the signal from both the far as well as the near receivers.

6. Coherent modulation is more suitable for DSSS. As maintaining the phase relationship during hopping is difficult in the case of FHSS, so FHSS systems are generally non-coherent. The DSSS modulation technique can be modified to support higher data transfer rates (at the cost of efficiency), while the FHSS can not. Because the usual FHSS modulation technique is less efficient than the normal DSSS technique, the FHSS transmitter will have a smaller range for a given power consumption level.
7. The FHSS systems offer better receiver sensitivity than DSSS systems. The sensitivity of an FHSS and a DSSS modem operating at 115 kbaud are around −103 dBm and −92 dBm, respectively. So in this respect FHSS can offer a better range.
8. FHSS system shows better performance in multipath environments than DSSS system. An FHSS system inherently provides a solution to this problem by hopping to a different frequency, which is not attenuated. The DSSS system also causes some delay in the received signal.
9. Frequency hopping also allows for the placement of numerous wireless LAN segments in a single area. However, for DSSS, allowing multiple users to access the channel at the same time is difficult. In dense user areas, greater loadss can be supported by connecting multiple access points, with over lapping coverage, to the same ether net network. Because of its better efficiency, an FHSS network is inherently able to provide three to four times more total network capacity than a DSSS network.
10. FHSS networks offer better network scalability. On trading floors, airports, and other environments where multiple organizers may want to operate wireless LANs, DSSS is not a good choice. FHSS allows a larger number of non-overlapping channels.
11. A DSSS system consumes more power than an FHSS system, thus requiring heavier batteries, which is not very convenient for mobile applications.
12. In the case of a DSSS system an intruder would have to first know which part of the frequency range was used for the transmission, and then establish the chipping code to spread the data in order to decrypt the original stream. Whereas in FHSS, an intruder must know both the current transmission frequency and the hopping pattern that decides the next frequency to which the system will jump.
13. A DSSS system has better throughput.
14. An FHSS system requires error correction.

5.5.11.4 Applications

From the above discussion, it is obvious that both systems have advantages as well as disadvantages. So, of these, which one is better? This depends on the goals and the needs of the wireless network. DSSS solutions tend towards having greater transmission speeds and the best noise and anti-jam performance. On the other hand, FH solutions are simple to implement and are well suited for moderately to highly dense user populations. Therefore the design engineer has to choose the correct one for a particular application. FHSS is used in Bluetooth systems, whereas DSSS is used for wireless LAN and UMTS systems.

5.6 Orthogonal Frequency Division Multiplex Access (OFDMA)

For multimedia communication in 4G application scenarios, the demands for bandwidth and quality of service (QoS) is very high compared with what is available today to the mobile user. The bit rates for multimedia span from a few kbps, for voice, to about 20 Mbps for HDTV, or even more in the peaks. Also, several problem exist in wireless channels, which arise because of the mobility of the users. The question now is how to satisfy those requirements?

Which modulation technique can compromise all contradicting requirements and provide the best solution?

In the air channel, as the data rate increases in a multipath environment, the channel fading goes from flat fading to frequency selective fading (the last reflected component arrives after the symbol period).

Also, the channel delay spread can cause ISI. These lead to heavy degradation of the bit error rate in the signal transmission as a result of frequency selective fading and inter symbol interference.

The most popular solutions to compensate for these above problems are:

a. **Use of Equalizers** – Adaptive compensation of time-variant channel distortion. However, as we move to higher data rates (that is, >1 Mbps), equalizer complexity grows to a level where the channel changes before you can compensate for it. Thus there are practical difficulties in operating this equalization in real-time at several Mbps with compact, low-cost hardware.
b. **Adaptive Array Antenna** – Considers delayed waves as interference waves and eliminates them to avoid overlapping waves. This is very complex and expensive solution.
c. **Multi-carrier Transmission** – This is an alternative promising solution.

Multi-carrier modulation (MCM) is where the channel is broken up into sub-bands such that the fading over each sub-channel becomes flat and thus helps to eliminate the ISI problem.

Single carrier systems transfer data streams using a serial transmission, while a multi-carrier system uses parallel transmission. The ability to sustain a higher throughput in a single carrier system becomes diminished as the symbol duration becomes smaller in order to support a higher data rate. A higher data rate can be achieved by placing an increased number of bits in a symbol, for example, use of higher order modulation, or by placing a higher number of symbols in the defined time frame, for example, use of smaller symbol duration. This is why a single carrier system becomes more susceptible to ISI and multipath effects.

However instead, the serial data stream can be divided into parallel data streams and each of them are individually modulated by a narrow-band carrier, and then summed up and transmitted in parallel from the same source. As a single stream of data is split up to individually modulate these multiple carriers, then these systems are sometimes referred to as multi-carrier systems. As the high speed data stream is divided into several parallel paths, the data rate at each of the parallel path will be reduced, for example, the symbol duration can be increased. Thus the BW requirement will be reduced, for example, signal BW < coherence BW, which means this will be robust against the multipath frequency selective fading and ISI.

Figure 5.21a shows single carrier system [bandwidth ($f4 - f1$)], where due to fading there is a heavy loss of information. However, Figure 5.21b shows a multi-carrier system [with the same overall bandwidth ($f4 - f1$)], only one carrier is affected due to fading. The information can be recovered using good coding and an interleaving scheme. The individual carriers are called sub-carriers.

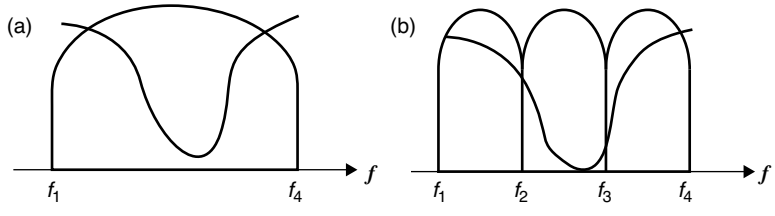


Figure 5.21 Effect of fading for (a) single carrier and (b) multi-carrier systems

Several typical implementation problems arise with the use of a large number of sub-carriers.

1. When we have large numbers of sub-carriers, then we will have to assign the sub-carrier's frequencies very close to each other. We know that the receiver needs to synchronize itself to every sub-carrier frequency in order to recover data related to that particular sub-carrier. When there is very little

spacing, then the receiver synchronization components need to be very accurate, which is not yet possible with low-cost RF hardware. So the bandwidth utilization will be very poor.
2. Also, at the transmitter side arrays of sinusoidal generators and at the receiver side arrays of coherent demodulators are required to support this multi-carrier system. This makes the system very complex and expensive.

So how are these problems overcome?

The solution to the first problem is – use orthogonal frequency carriers – known as OFDM, and the solution to the second problem is use of the FFT technique. We will now discuss these solutions.

5.6.1 Importance of Orthogonality

As the carriers are all sine/cosine waves, we know that area under one period of a sine or cosine wave is zero, as shown in the Figure 5.22.

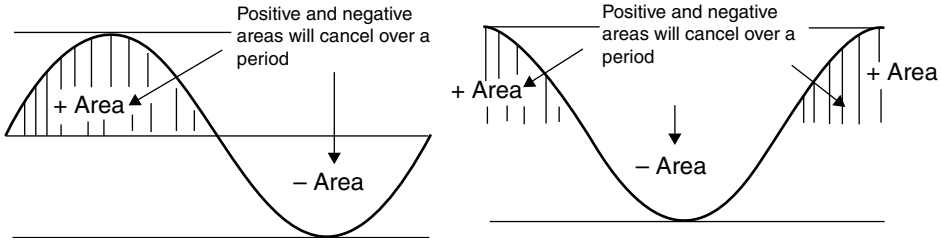


Figure 5.22 Sine and cosine wave area over a period

The sine and cosine waves are orthogonal, because at the time when the sine is at its maximum the cosine wave is at its minimum. Mathematically two signals are called orthogonal if the following condition is met:

$$\int_0^T S_1(t).S_2^*(t)\mathrm{d}t = 0 \tag{5.13}$$

If we take a sine wave of frequency m and multiply with a sinusoid of frequency n, where both m and n are integers, then the integral over one period, for example, the area will be:

$$E(t) = \sin m\omega t * \sin n\omega t = {}^1\!/_2[\cos{(m-n)}\,\omega t + \cos{(m+n)}\,\omega t]$$

As these two components are also sinusoids then the integral or area under one period will also be zero.

$$\int_0^{2\pi} \frac{1}{2}\cos(m-n)\omega t - \int_0^{2\pi} \frac{1}{2}\cos(m+n)\omega t = 0 - 0 \tag{5.14}$$

We can conclude that when we multiply a sinusoid of frequency n by a sinusoid of frequency m (where m and n are integers), the area under the product is zero. In general for all integer values of m and n, $\sin nx$, $\sin mx$, $\cos nx$, $\cos mx$ are all orthogonal with each other. These frequencies are called harmonics. However, remember, when $m = n$, the above result is not zero, for example, the area is not zero. This principle is used in OFDM, where the orthogonality between the carriers allows overlapping of the carriers and transmitting simultaneously.

The receiver acts as a bank of demodulators, translating each carrier down to dc, with the resulting signal integrated over a symbol period to recover the raw data. If the other carriers all beat down the frequencies (in the time domain, take a whole number of cycles in the symbol period T) then the integration process results in zero contribution from all these other carriers. Thus, the carriers are linearly independent (that is, orthogonal) if the carrier spacing is a multiple of $1/T$. To maintain orthogonality between carriers, it is necessary to ensure that the symbol time contains one or multiple cycles of each sinusoidal carrier waveform. Generally, the carrier frequencies are chosen as integer multiples of the symbol period.

The main concept in OFDM is orthogonality of the sub-carriers. This special property prevents adjacent carriers in OFDM systems from interfering with one another; in the same manner that the human ear can clearly distinguish between each of the tones created by the adjacent keys of a piano. OFDM can also be considered a multiple access technique as an individual carrier or groups of carriers can be assigned to different users. Each user can be assigned a predetermined number of carriers when they have information to send, or alternatively, a user can be assigned a variable number of carriers based on the amount of information they have to send. The media access control (MAC) layer controls the assignments and schedules the resources based on user demand.

OFDM is similar to FDM but much more spectrally efficient because the sub-channels are spaced much closer together. This is done by finding frequencies that are orthogonal, which means that they are perpendicular in a mathematical sense, allowing the spectrum of each sub-channel to overlap with the other, without interfering with it. Consider a simple example: let us consider that a bandwidth of "5f" is available for transmission. Firstly we will see that if we use the FDM technique and if each of the narrow-band carriers takes a bandwidth of f, then all of the five carriers can be accommodated using the FDM technique. This is shown in Figure 5.23.

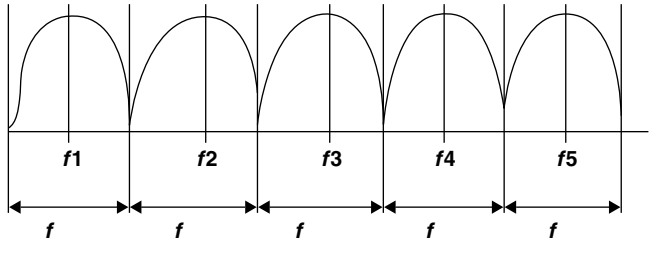


Figure 5.23 Carrier assignment using FDM technique

However, if the frequencies are integer multiples of each other, for example, $f_5 = 5f_1, f_4 = 4f_1, f_3 = 3f_1$, $f_2 = 2f_1$, then these can be closely spaced, as they will not interfere with each other and when one of the power spectrums is at the maximum position, at that time the other's power spectrum will be at the minimum position. The signal spacing can be shown in Figure 5.24.

Now for each carrier of bandwidth f, the total spacing required for five carriers will be approximately $\sim 3f$. This clearly indicates better usage of the spectrum. We have almost 50% bandwidth saving.

The waveform of carriers in OFDM transmission and the transmission power spectrum are shown in Figure 5.25. Notice that the peak of each tone corresponds to a zero level, or null, of every other carrier. The result of this is that there is no interference between the carriers. When the receiver samples at the center frequency of each carrier, then only the desired carrier's energy will be present at that point (along with the noise signal).

The power spectrum for a single carrier, multi-carrier, and an OFDM based multi-carrier system is shown in the Figure 5.26.

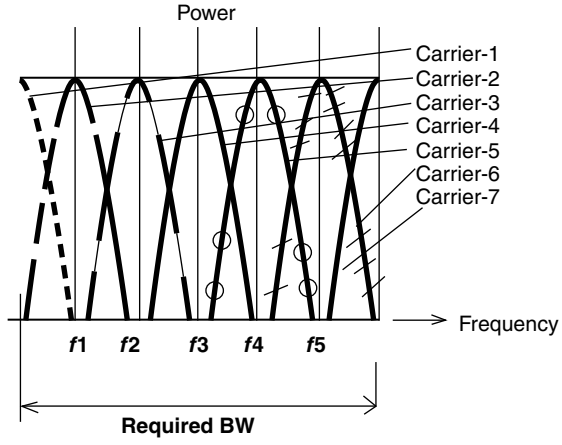


Figure 5.24 OFDM spectrum

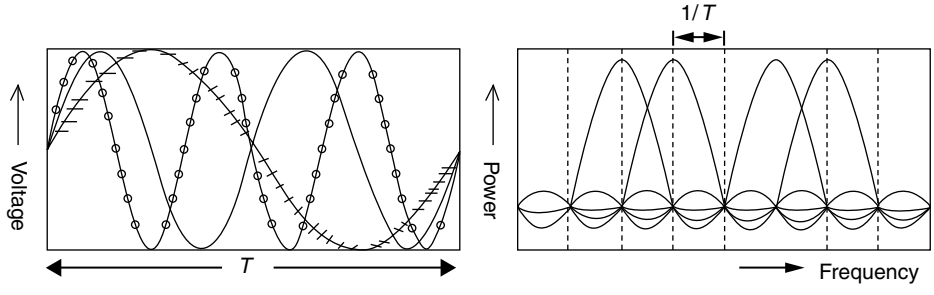


Figure 5.25 OFDM transmission wave and power spectrum

5.6.2 Mathematical Description of OFDM

OFDM transmits a large number of narrow-band carriers, closely spaced in the frequency domain. Mathematically, each carrier can be described as a complex wave [1]:

$$S(t) = A_c(t)\left[e^{j(2\pi f_c t + \varphi_c(t))}\right] \tag{5.15}$$

$A_c(t)$ and $\Phi_c(t)$ are the amplitude and phase of the carrier. The amplitude and phase can vary on a symbol by symbol basis. The values of the parameters are constant over the symbol duration period T.

OFDM consists of many carriers. Thus the complex signals $S_n(t)$ is represented by:

$$S_n(t) = \frac{1}{N}\sum_{n=0}^{N-1} A_n(t)\left[e^{j(2\pi f_n t + \varphi_n(t))}\right] \tag{5.16}$$

where $\omega_n = \omega_0 + n.\Delta\omega$. This is of course a continuous signal. If we consider the waveforms of each component of the signal over one symbol period, then the variables $A_n(t)$ and $\Phi_n(t)$ take fixed values, which depend on the frequency of that particular carrier, and can be rewritten as: $\Phi_n(t) = \Phi_n$ and

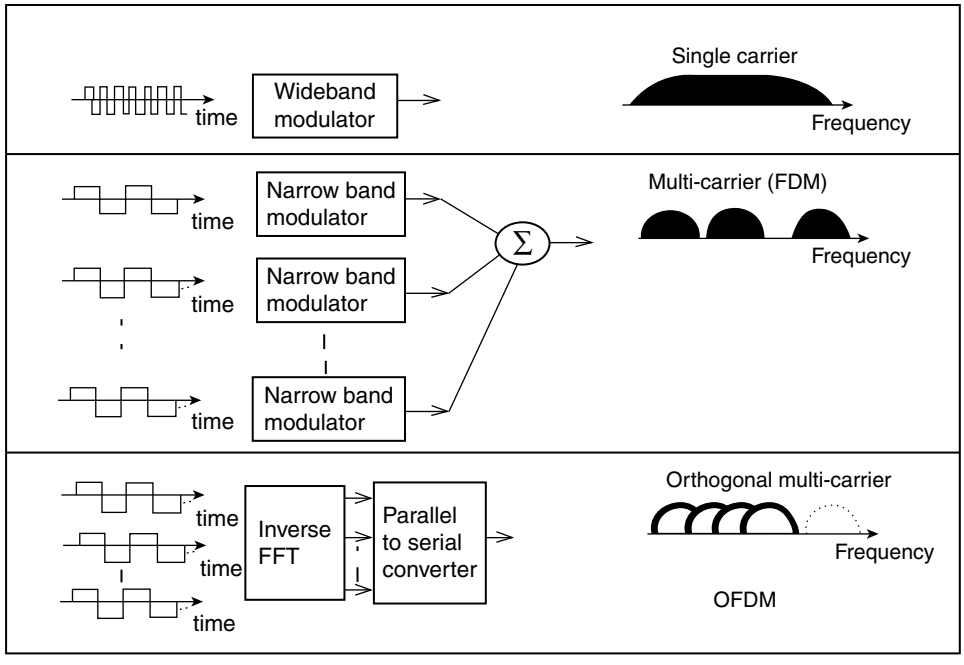


Figure 5.26 Single-carrier versus multiple-carrier versus OFDM spectrum

$A_n(t) = A_n$. If the signal is sampled using a sampling frequency of $1/T$, then the resulting signal is represented by:

$$S_n(kT) = \frac{1}{N}\sum_{n=0}^{N-1} A_n(t)\left[e^{j(2\pi(f_0 + n\Delta f)kT + \phi_n)}\right] \tag{5.17}$$

At this point, we have restricted the time over which we analyze the signal to N samples. It is convenient to sample over the period of one data symbol. Thus we have a relationship: $\tau = NT$. Now, if we simplify the above equation, without a loss of generality by letting $f_o = 0$, then the signal becomes:

$$S_n(kT) = \frac{1}{N}\sum_{n=0}^{N-1} A_n(t)\left[e^{j2\pi.n.\Delta f.kT'}\right].e^{j\phi_n} \tag{5.18}$$

In above equation, the function $A_n e^{j\phi_n}$ is no more than a definition of the signal in the sampled frequency domain and $S(kT)$ is the time domain representation. The above equation can be compared with the general form of the inverse Fourier transform:

$$s(kT) = \frac{1}{N}\sum_{n=0}^{N-1} S\left(\frac{n}{NT}\right).e^{j2\pi nk/N} \tag{5.19}$$

The above two equations are equivalent if $\Delta f = (\Delta\omega/2\pi) = (1/NT) = 1/\tau$. This is the same condition that was required for orthogonality. Thus, one consequence of maintaining orthogonality is that the OFDM signal can be defined by using Fourier transform procedures. Hence, according to its mathematical distribution, on the transmitter side, inverse digital Fourier transform (IDFT) summarizes all sine and cosine waves of amplitudes stored in an $S[k]$ array, forming a time domain signal (Figure 5.27):

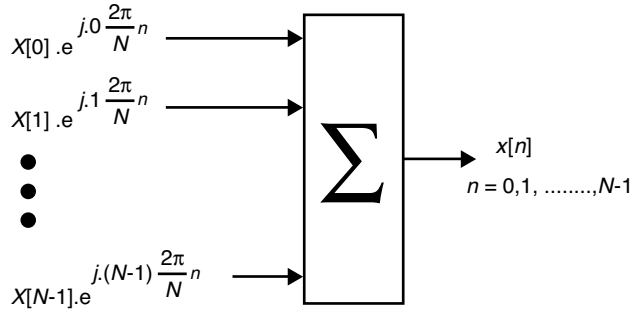


Figure 5.27 IDFT operation from different complex exponential carriers

$$x[n] = \frac{1}{N}\sum_{k=0}^{N-1} X(k).\mathrm{e}^{j2\pi nk/N} = \frac{1}{N}\sum_{k=0}^{N-1} X(k).\left[\cos\left(k.\frac{2\pi}{N}n\right) + j.\sin\left(k.\frac{2\pi}{N}n\right)\right] \tag{5.20}$$

where

$n = 0, 1, \ldots, N-1$

We can observe from the above equation that IDFT takes a series of complex exponential carriers, modulates each of them with a different symbol from the information array $S[k]$, and multiplexes all this to generate N samples of a time domain signal. These carriers are orthogonal and frequency spaced with $\Delta f = 2\pi\ N$.

At the receiver side, the inverse process is performed. The time domain signal constitutes the input to a DFT block, which is implemented using the FFT algorithm. The FFT demodulator takes the N time domain transmitted samples and determines the amplitudes and phases of sine and cosine waves forming the received signal, according to the equation below:

$$X[k] = \frac{1}{N}\sum_{n=0}^{N-1} x(n).\mathrm{e}^{-j2\pi nk/N} = \frac{1}{N}\sum_{n=0}^{N-1} x(n).\left[\cos\left(k.\frac{2\pi}{N}n\right) - j.\sin\left(k.\frac{2\pi}{N}n\right)\right] \tag{5.21}$$

where

$k = 0,1, \ldots, N-1$

5.6.3 Mathematics to Practice

Let us take an example, consider that there are three users and the total frequency band of B Hz is available for transmission as shown in Figure 5.28. Suppose this bandwidth is divided into N sub-carriers using the OFDM technique. Now, these N sub-carriers are distributed between these three users according to their needs, where, for example, the first user has two sub-carriers with frequency f_1 and f_2. The serial data stream from the user-1 source is combined together to form symbols (each symbol will contain the appropriate number of bits according to the modulation used – QAM, QPSK. . .), then these respective symbols will be divided into two parallel sub-streams (as there are two sub-carriers assigned to user-1). The symbol is then bifurcated to the I and Q path and multiplied by the sine and cosine of the respective assigned orthogonal frequency (f_1 for path 1 and f_2 for path-2). The same is applied for other users; and finally these signals are added together, up-converted and transmitted. The reverse process is performed at the receiver side.

Thus our first problem is solved, now let us see how to solve the second problem.

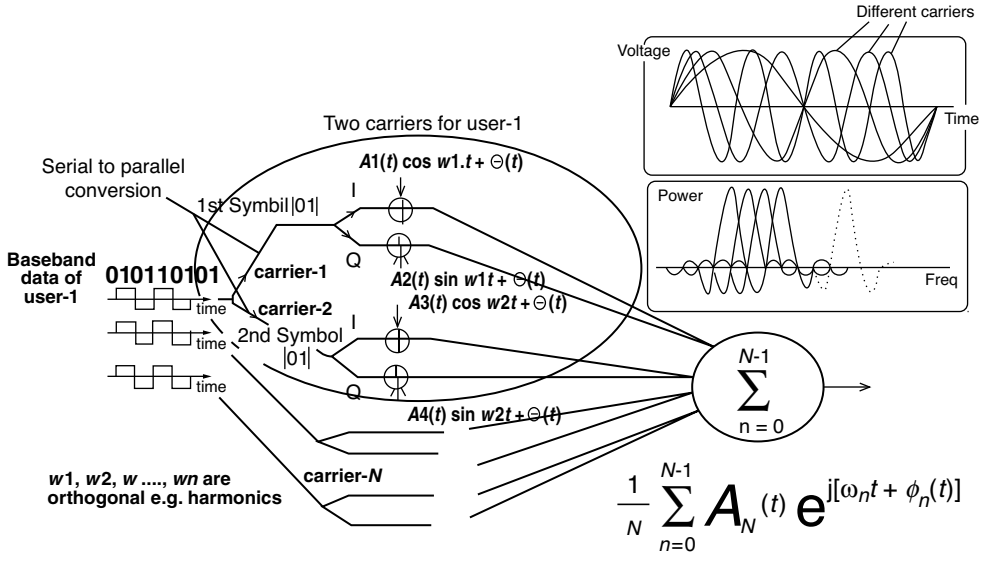


Figure 5.28 Data transmission using OFDM

5.6.4 Digital Implementation of Fourier Transform

Digital signals are discrete in nature and require storage space associated with the large volume of data. If the data volume is large, then it becomes difficult for the digital signal processor (DSP) to handle this. DSP can work only with the Fourier transform of a discrete wave that is finite in length, which means that discrete Fourier transform (DFT) is suitable to do this. However, as the processing of DFT takes lots of time, so the use of digital techniques was not popular for this implementation. The DFT of a continuous time signal sampled over the period of T, with a sampling rate of Δt can be given as:

$$s(m\Delta f) = \frac{T}{N}\sum_{n=0}^{N-1} S(n\Delta t).\mathrm{e}^{j2\pi.m.\Delta f.n.\Delta t} \tag{5.22}$$

where

$\Delta f = 1/T$, and is valid at frequencies up to $f_{max} = 1/(2\Delta t)$

Alhough OFDM is an old technology, it was not very popular early on because of this difficulty. In 1965 a paper was published by Cooley and Turkey describing a very different algorithm to implement the DFT [2], which is known as fast Fourier transform (FFT). Subsequently, at the transmitter and receiver side the OFDM signal generation and reception have been implemented using efficient FFT techniques that reduce the number of operations from N^2 (in DFT) down to N log N. The ability to define the signal in the frequency domain, in software on VLSI processors, and to generate the signal using the inverse Fourier transform is the key to its current popularity.

5.6.5 OFDM History

A brief history of OFDM technology is described in Table 5.3.

Today, OFDM enables the creation of a very flexible system architecture that can be used efficiently for a wide range of services, including both voice and data. In order to create a rich user experience for any

Table 5.3 Brief history of OFDM technology

Year	Events
1950	Concept of multi-carrier modulation with non-overlapping sub-channels
1960	Orthogonal sub-channels: the first OFDM scheme was proposed by Chang in 1966 for dispersive fading channels
1970	A US patent filed, used in military applications
1980	OFDM employing QAM with DFT technique is developed
1990	Various standards developed for wire line and wireless systems based on OFDM
2000	Applications in cellular environments

mobile system, it must provide ubiquitous, fast and user-friendly connectivity. OFDM has several unique properties that make it especially well suited to handle the challenging environmental conditions that mobile wireless data applications must operate in, which is why, it has been chosen as a potential physical layer solution for LTE and 4G system.

5.6.6 Key Advantages of the OFDM Transmission Scheme

1. Makes efficient use of the spectrum by allowing frequency spectrum overlap.
2. By dividing the channel into narrow-band flat fading sub-channels, OFDM is more resistant to frequency selective fading than single carrier systems are.
3. Eliminates ISI and IFI through use of a cyclic prefix.
4. Using adequate channel coding and interleaving one can recover symbols lost due to the frequency selectivity of the channel.
5. Channel equalization becomes simpler than by using adaptive equalization techniques with single carrier systems.
6. It is possible to use maximum likelihood decoding with reasonable complexity. Thus OFDM is computationally efficient, when the FFT technique is used to implement the modulation and demodulation functions.
7. In conjunction with differential modulation there is no need to implement a channel estimator.
8. Less sensitive to sample timing offsets than single carrier systems.
9. Provides good protection against co-channel interference and impulsive parasitic noise.

5.6.7 Drawbacks of OFDM

1. **Interference Between OFDM Symbols** – Usually a block is referred to as an "OFDM symbol." When OFDM transmits data in blocks through the wireless channel, due to multipath and other related channel impairments the blocks of signal will overlap, for example, interfere with each other as shown in the Figure 5.29. This type of interference is called inter-block interference (IBI), which will eventually lead to ISI, as two adjacent blocks will overlap, causing the distortion of the symbol affected by overlapping.
   In order to combat this interference, one of the possible approaches was to introduce "a silence period" between the transmitted frames. Known as "zero prefix," this silence period consists of a number of zero samples, added to the front of each symbol. However, the zero-padding does not seem to be the ideal solution, because the zero prefix will destroy the periodicity of the carriers. The demodulation process that uses FFT will be facilitated by keeping this periodicity, as we will see next. Instead of this "quiet period" zero prefix, we could use a cyclic prefix (CP) at the beginning of each symbol. The cyclic

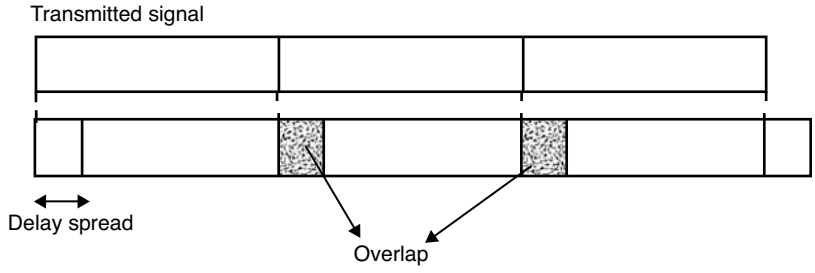


Figure 5.29 Inter-block interference

prefix consists of the last L samples of the OFDM symbol that are copied to the front of the data block. Of course, the price to pay is a decrease in the transmission efficiency with a factor of $T/(T + T_g)$, representing a ratio between the useful and the total transmission time for an OFDM symbol.

2. **Peak Power Problem of OFDM** – In a multi carrier modulation, the combined signal is the sum of modulated signals with sub-carriers. When the phase of each modulated signal is in-phase the multi-carrier modulated signal will have a very high peak power. This will influence the linearity of the power amplifier and the amplifier may go to a non-linear region and signal clipping can happen. Therefore, it requires RF power amplifiers with a high peak to average power ratio.
3. **DFT Leakage** – It is more sensitive to carrier frequency offset and drift than single carrier systems are due to DFT leakage.

References

[1] Edfors, O. and Sandell M. (1996) An Introduction to Orthogonal Frequency Division Multiplexing, technical report September 1996, Lulea University of Technology.

[2] Cooley, J.W. and Tukey, J.W. (1965) An algorithm for the machine calculation of complex Fourier series, *Math. Comput.* **19**, 297–301.

Further Reading

Langton, C. Intuitive Guide to Principle of Communications, February 2002, www.complextoreal.com.

Proakis, J.G. (1995) *Digital Communications*, 3rd edn. McGraw-Hill, New York.

Richardson, A. (2005) *WCDMA Design Handbook*, Cambridge University Press, Cambridge, ISBN 10: 0521828155.

Schulze, H. and Christian, L. (2005) *Theory and Applications of OFDM and CDMA Wideband Wireless Communications*, John Wiley & Sons, Inc., Hoboken, ISBN 978 0 470 85069 5.

6

GSM System (2G) Overview

6.1 Introduction

GSM (Global System for Mobile Communications) is the world's first cellular system to specify digital modulation, network level architectures and services. Today, it is the most popular second generation (2G) technology, having more than one billion subscribers worldwide.

6.2 History of GSM

During the early 1980s, analog cellular telephone systems were experiencing rapid growth in Europe, particularly in Scandinavia and the United Kingdom. Each country was developing its own system, which was incompatible with other network's equipment and operation. This was not a desirable situation, because the operation of such mobile equipment was limited to within the national boundaries, and due to this incompatibility issue, the equipment had very limited markets. Soon the limitation of this economic scale and opportunities for this market potential were realized. In 1982, the main governing body of the European telecommunication operators, known as CEPT (Conference Europe des Posts et Telecommunications) was formed. To overcome the above issue, the task of specifying a common mobile communication system for Europe in the 900 MHz frequency band (initially) was given to the Group Special Committee (GSM), which was a working group of CEPT. This group was formed to develop a pan-European public land mobile system. The proposed system had to meet several criteria, such as: (1) good subjective speech quality, (2) ability to support handheld terminals, (3) low terminal and service costs, (4) support for a range of new services and facilities, (5) support for international roaming, (6) ISDN compatibility, and (7) good spectral efficiency.

In 1989, the GSM responsibility was transferred to the European Telecommunication Standards Institute (ETSI), and in 1990 phase I of the GSM standard's specifications were published. Commercial service was started in mid-1991, and by 1993 about 36 GSM networks were operational in 22 countries. In 1992, GSM changed its name to "Global System for Mobile Communications" for marketing reasons. In Phase II of the GSM specifications, which were frozen in June 1993, the GSM 900 and the DCS 1800 (Digital Cellular System – at the request of the UK a version of GSM operating in the 1800 MHz band was included in the specification process) were combined into the same set of documents. Today, GSM has become very popular and over more than 400 GSM networks (including DCS1800 and PCS1900) are operational in 130 countries around the world. A brief history of GSM development is included in Table 6.1.

Mobile Handset Design Sajal Kumar Das

Table 6.1 GSM history

Year	Events
1982	CEPT establishes GSM group in order to develop the standards for a pan-European cellular mobile system.
1985	Adoption of a list of recommendations to be generated by the group.
1986	Field tests were performed in order to test the different radio techniques proposed for the air interface.
1987	TDMA (in combination with FDMA) is chosen as access method. Initial Memorandum of Understanding signed by the telecommunication operators (representing 12 countries). GSM spec drafted.
1988	Validation of the GSM system. The European Telecommunications Standards Institute (ETSI) was founded.
1989	The responsibility of the GSM specifications is passed to the ETSI.
1990	Appearance of phase I of the GSM specifications. DCS adaptation starts.
1991	Commercial launch of the GSM service in Europe.
1992	Actual launch of commercial service, and enlargement of countries that signed the GSM. GSM changed its name to Global System for Mobile Communication.
1993	Several non-European countries in South America, Asia, and Australia adopted GSM.
1995	Phase II of the GSM specifications. Coverage of rural areas. GSM 1900 was implemented in USA.

6.3 Overview of GSM Network Architecture

A GSM network is composed of several functional entities, whose functions and interfaces are properly defined in the GSM specification. Figure 6.1 shows the architecture of the GSM network. The GSM network can be broadly divided into three parts: (1) the mobile station (MS) – this is the mobile part and is carried by the user; (2) the base station subsystem (BSS) – this controls the radio link with the mobile station; (3) the network subsystem (NSS) – the main part of the NSS is the mobile services switching center (MSC), which performs the switching of calls between the mobile and other fixed or mobile network users, as well as management of mobile services, such as authentication, ciphering , and so on. Another part, which is also shown in the Figure 6.1, is the operations and maintenance center (OMC), which oversees the correct operation and setup of the network. The mobile station and the base station subsystem communicate via the Um interface, also known as the air interface or radio link. The base station subsystem communicates with the mobile service switching center via the A interface.

6.3.1 Mobile Station (MS)

The MS is the mobile unit, which consists of the physical equipment used by the subscriber to access a network in order to use the services offerd by this network. The MS is composed of two distinct functional entities: the subscriber identity module (SIM) and mobile equipment (ME) (see Figure 6.2).

6.3.1.1 SIM

The SIM is a credit card sized smart card, which can be used by the subscriber to personalize an ME. Inserting a valid SIM card into any GSM mobile equipment (ME), the user will be able to receive or make calls using that mobile phone. In the first generation analog cellular systems, a user's unique electronic serial number (ESN) is programmed directly into the mobile phone. This makes it difficult to switch to any

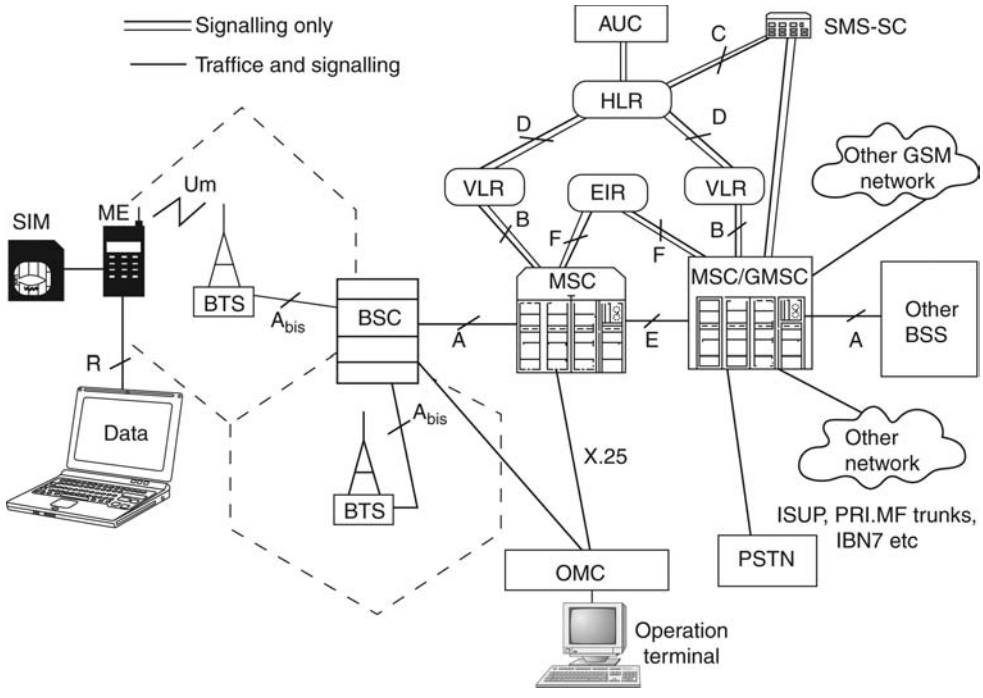


Figure 6.1 GSM network architecture

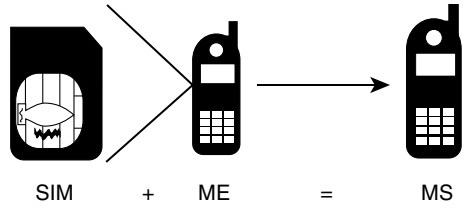


Figure 6.2 SIM, ME, and MS

other networks (operators). In such a situation, the subscriber needs to exchange or reprogram the mobile phone. The introduction of a SIM card provides subscribers with the complete freedom to switch between different network operators. The SIM provides personal mobility, so that the user can have access to all subscribed services irrespective of both the location of the terminal and the use of specific mobile equipment. The introduction of SIM also allows the subscribers to change the ME without changing the number or subscription details. Internal details about the SIM are discussed in Chapter 10.

6.3.1.2 ME

The mobile equipment (ME) can be subdivided into three functional blocks. (1) Terminal equipment (TE) – this performs functions specific to a particular service, such as a FAX machine, but does not handle any functions specific to the operation of the GSM system. (2) Mobile terminal (MT) – this contains all the

functionalities related to the transmission and reception of information over the GSM radio interface, for example, GSM radio modem part. (3) Terminal adapter (TA) – this is used to ensure compatibility between the MT and TA , for example, a TA would be required to interface between an ISDN-compatible MT and a TA with a modem interface.

6.3.2 Base Station Subsystem (BSS)

The base station subsystem acts like a local exchange of a wire-line system. This is composed of two parts, the base transceiver station (BTS) and the base station controller (BSC).

6.3.2.1 Base Transceiver Station (BTS)

A base transceiver station (BTS) performs all the transmission and reception functions with MS relating to GSM via a Um (air) radio interface and on the other side it communicates with BSC via an A-bis interface. The BTS houses the radio transceivers that define a cell and handles the radio-link protocols with the mobile station. A BTS is comprised of radio transmitters and receivers, antennas, interface to the PCM facility, and so on, and the tasks include RF transmission and reception, channel coding/decoding, and encryption/decryption and so on.

6.3.2.2 Base Station Controller (BSC)

A group of BTSs are connected to a particular base station controller (BSC), which manages the radio resources for them. The management functions include: the allocation of radio channels to the MSs on call set up, determining when the handover is required, identifying suitable BTS, and controlling the transmitted power of an MS to ensure that it is sufficient to reach the serving BTS. The mobile stations normally send a measurement report about their received signal strength and quality every 480 ms to the BSC. With this information the BSC takes decision about- when to initiate the handovers to other cells, when to change the BTS transmitter power, and so on. The BSS is the connection between the mobile and the mobile service switching center (MSC). The BSC also translates the 13 kbps voice channel used over the radio link to the standard 64 kbps channel used by the public switched telephone network or ISDN. Typically, a BSC may control up to 40 BTSs and the capability of the BSCs varies from manufacturer to manufacturer. The functions of BTS and BSC are specified in Table 6.2.

Table 6.2 BTS and BSC Functions

Functions	BTS	BSC
Management of radio channels		Yes
Frequency hopping	Yes	Yes
Management of terrestrial channels		Yes
Mapping of terrestrial onto radio channels		Yes
Channel coding and decoding	Yes	
Rate adaptation	Yes	
Encryption and decryption	Yes	Yes
Paging	Yes	Yes
Uplink signal measurements	Yes	
Traffic measurement		Yes
Authentication		Yes
Location registry and location update		Yes
Handover management		Yes

6.3.3 Network Subsystem (NSS)

6.3.3.1 Mobile Services Switching Center (MSC)

The central component of the network subsystem is the mobile services switching center (MSC). It acts like a normal switching node of the PSTN or ISDN, and in addition it provides all the functionality needed to handle a mobile subscriber, such as registration, authentication, ciphering, location updating, handovers, generation of call records, and call routing to a roaming subscriber. These services are provided in conjunction with several functional entities, which together form the network subsystem. The MSC provides the connection to the public fixed network (PSTN or ISDN). Signaling between functional entities uses the ITUT signaling system number 7 (SS7), which is used in ISDN and widely used in current public networks. The network operator may also select one or more MSCs to act as gateway MSCs (GMSC). This provides the interface between the PLMN and external networks. MSC does not contain information about particular mobile stations, so this information is stored in the location registers.

6.3.3.2 Home Location Register (HLR)

A Home Location Register (HLR) is a database that contains semi-permanent mobile subscriber information for a wireless operators' entire subscriber base. Responsibilities of the HLR include: management of service profiles, mapping of subscriber identities (MISDN, IMSI), supplementary service control and profile updates, execution of supplementary service logic, for example, incoming calls barred and passing subscription records to the VLR. Two types of information are stored in the HLR: the subscriber information and part of the mobile information to allow incoming calls to be routed to the MSC for the particular MS. HLR subscriber information includes the international mobile subscriber identity (IMSI), location information, service restrictions, and supplementary services information, service subscription information, and so on. The HLR contains all the administrative information of each subscriber registered in the corresponding GSM network, along with the current location of the mobile. The current location of the mobile is in the form of a mobile station roaming number (MSRN, please refer to Section 6.8.5). Request information from the HLR or update the information contained in the HLR is handled by SS7 transactions with the MSCs and VLRs. The HLR also initiates transactions with VLRs to complete incoming calls and to update subscriber data. Traditional wireless network design is based on the utilization of a single Home Location Register (HLR) for each GSM network, but growth considerations are prompting operators to consider multiple HLR topologies and this can also be implemented as a distributed database.

6.3.3.3 Visitor Location Register (VLR)

A Visitor Location Register (VLR) is a database that contains temporary information concerning the mobile subscribers that are currently located in a given MSC serving area, but whose HLR is elsewhere. The information in VLR includes MSRN, TMSI, MS ISDN number, IMSI, HLD address, local MS identity (if any), the location area in which the MS has been registered, data related to supplementary services, and so on. When a mobile subscriber roams away from his home location into a remote location, SS7 messages are used to obtain information about the subscriber from the HLR, and to create a temporary record for the subscriber in the VLR. There is usually one VLR per MSC. The HLR and VLR, together with the MSC, provide the call routing and (possibly international) roaming capabilities of GSM.

6.3.3.4 Equipment Identity Register (EIR)

Each mobile station is identified by its International Mobile Equipment Identity (IMEI) number. Equipment Identity Register (EIR) is a database that contains a list of all valid IMEI numbers. This is used for security purposes and to prevent any illegal usage (see Section 8.1).

6.3.3.5 Authentication Center (AuC)

The authentication center (AuC) is an intelligent database concerned with the regulation of access to the network, ensuring that services can only be used by those who are entitled to do so and that the access is achieved in a secure way. The AuC authenticates each user (SIM card) that attempts to connect to the GSM core network (typically when the phone is powered on). It is a protected database that stores a copy of the secret key stored in each subscriber's SIM card, which is used for authentication and ciphering of the radio channel. Generally, it contains the subscriber's secret key (*Ki*) and the A3 and A8 security algorithms. This is discussed in detail in Chapter 9.

6.3.4 Operation and Maintenance Subsystem (OMSS)

The Operations and Maintenance Center (OMC) provides the means by which operators control the networks. The Network Management Center (NMS) is concerned with the management of the entire network and generally has a wider operational role than an OMC. The OMC is a management system that oversees the GSM functional blocks. The OMC assists the network operator in maintaining satisfactory operation of the GSM network. It can be in charge of an entire public land mobile network (PLMN) or just some parts of the PLMN.

6.4 PLMN and Network Operators

The GSM system is divided into a number of separate operational networks, each being operated independently to a large extent from the others. Each of these networks is called a PLMN (Public Land Mobile Network). The licenses for operating a GSM network in a country have been granted by Government agencies or some other authority. The operator may be a private company (such as Orange, Airtel, Vodaphone, AT&T), a public company or an administration, who buy the frequency licenses to deploy the GSM network. So the PLMNs are operated by different operators and again each PLMN is interconnected with other PLMNs, PSTNs or data networks and provide global communication access to a mobile user. The access to PLMN services is achieved by means of air interface (discussed in the Chapter 7) involving radio communications between MS and land based base stations (BTS). Most countries have several PLMNs, whose coverage areas can overlap partially or completely through appropriate frequency planning. This may cause problems in the border areas of a country. So, one restriction that has been imposed by CEPT, is that the commercial coverage area of each PLMN should be confined within the borders of one country.

6.4.1 Hierarchy of GSM Network Entities

Typically based on a geographical area, different cellular system providers deploy their own GSM networks. Again in the same area, several GSM networks (belonging to different operators) can co-exist, as shown in the Figure 6.3, where over the same geographical area, operator A and operator B deploy their services. They have taken licenses for different radio frequencies (in a GSM band) to operate in the same zone. The MS belonging to operator A will have SIM-A whereas MS belonging to operator B will have SIM-B inside it.

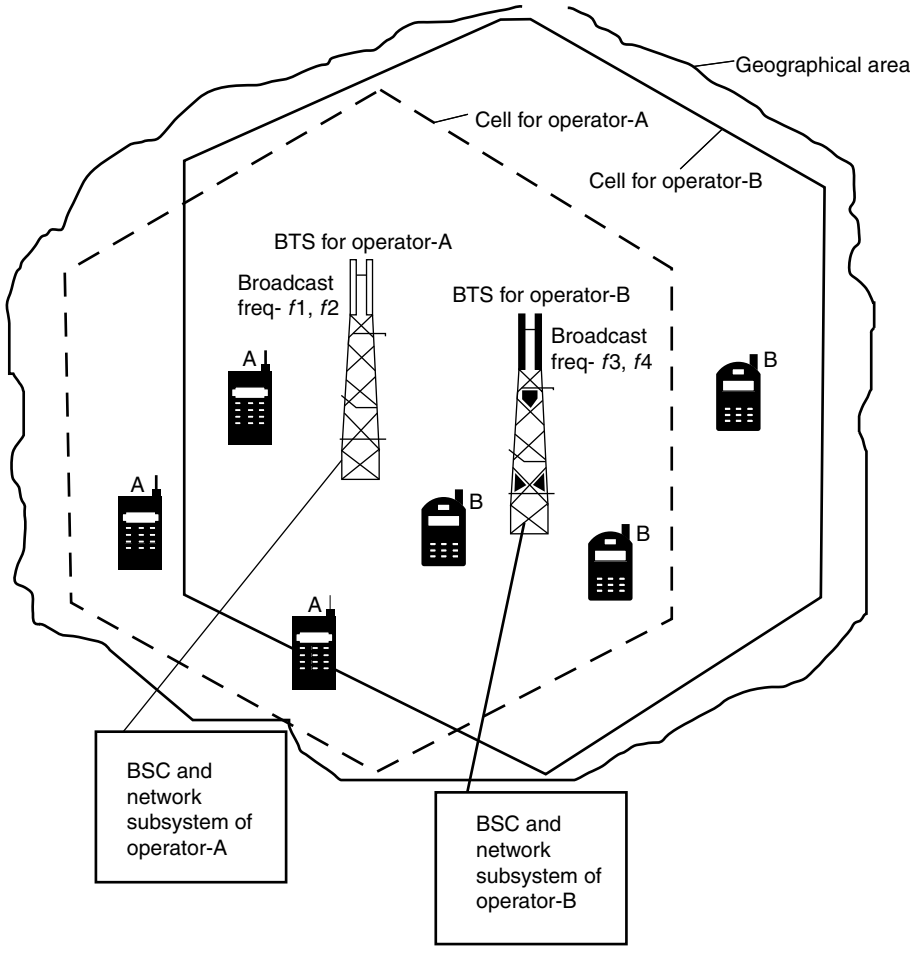


Figure 6.3 GSM network deployment by different operators

Cell sizes vary from 100 m up to 35 km depending on user density, geography, transceiver power, and so on. A cell site of any operator may typically contain a single BTS with one antenna subsystem (Omni-directional antenna, transmitting power equally in all directions) or a cell is split into several sectors (this is called sectorization) and involves dividing the cell into number of sectors (see Figure 6.4). One way to think about sectors is to consider them as separate smaller cells covering particular zones using directional antennas. Thus the advantage here is that the base stations corresponding to these divided sectors are co-located, which leads to saving of space, resources, and cabling. Sectorization is achieved by having a directional antenna at the base station that focuses the transmissions into the sector of interest and is designed to be null in other sectors. The ideal end result is an effective creation of new smaller cells without the added burden of new base stations and network infrastructure. This can help to increase network capacity and also to reduce the required transmission power.

The hierarchy of GSM network entities is shown in the Figure 6.5.

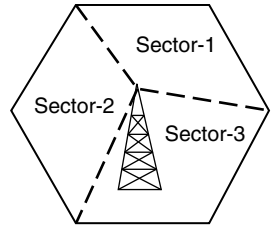


Figure 6.4 Sectorization of a cell

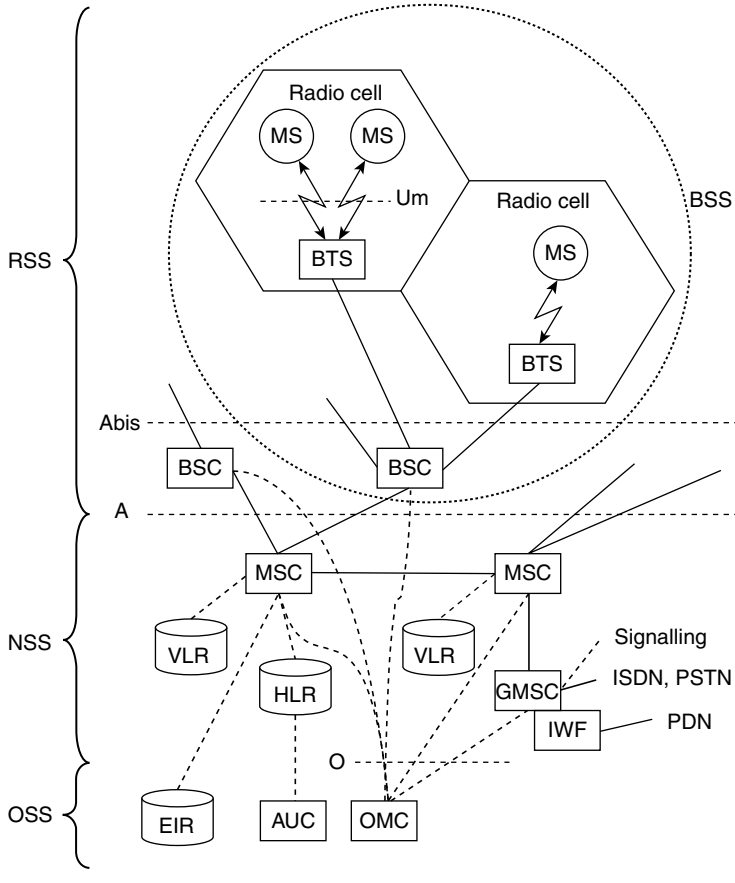


Figure 6.5 Hierarchy of GSM network entities

6.4.2 GSM Network Areas

The GSM network is made up of geographic areas. As shown in Figure 6.6, these areas include cells, location areas (LAs), MSC/VLR service areas, and public land mobile network (PLMN) areas. The cell is the area, where radio coverage is given by one base transceiver station (BTS). The GSM network identifies each cell via the cell global identity (CGI) number assigned to each cell. The location area is a group of

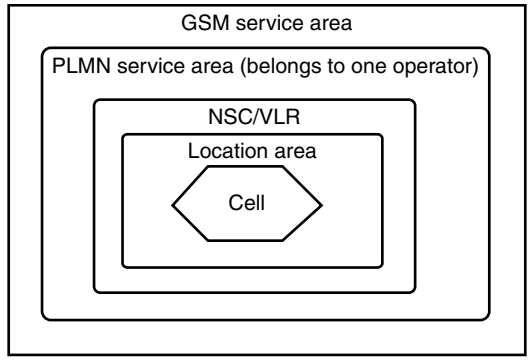


Figure 6.6 GSM network areas

cells. It is the area in which the subscriber will be paged. Each LA is served by one or more base station controllers, but only by a single MSC. Each LA is assigned a location area identity (LAI) number. An MSC/VLR the service area represents the part of the GSM network that is covered by one MSC and which is reachable, as it is registered in the VLR of the MSC. The PLMN service area is an area served by one network operator.

As described earlier, there can be several network operators for a GSM service area. Hence there can be several PLMNs belonging to different operators in a GSM service area.

6.4.3 Objectives of a GSM PLMN

A GSM PLMN cannot establish calls autonomously other than local calls between mobile subscribers. In most cases, the GSM PLMN depends upon the existing wire-line networks to route the calls via PSTN. Most of the time the service provided to the subscriber is a combination of the access service by a GSM PLMN and the service by some existing wire-line network. The general objectives of a GSM PLMN are: (1) to provide the subscriber a wide range of services and facilities, both voice and data, that are compatible with those offered by existing networks (PSTN, ISDN, etc.); (2) to introduce a mobile service system that is compatible with the ISDN; (3) to provide facilities for automatic roaming, locating, and updating of mobile subscribers; and (4) efficient use of the frequency spectrum.

6.4.4 PLMN

As discussed previously, a GSM network is a public land mobile network (PLMN). GSM uses the following sub-division of the PLMN.

1. **Home PLMN (HPLMN)** – The HPLMN is the GSM network to which a GSM user originally subscribed to. GSM user's subscription data reside in the HLR of HPLMN. During registration in another PLMN, the HLR may transfer the subscription data to a VLR or to a GMSC during mobile terminating call handling. The HPLMN may also contain various service nodes, such as a short message service center (SMSC), service control point (SCP), and so on.
2. **Visitor PLMN (VPLMN)** – The VPLMN is the GSM network where a subscriber is currently registered. Originally the subscriber may be registered in HPLMN (or in another PLMN).
3. **Interrogating PLMN (IPLMN)** – The IPLMN is the PLMN containing the GMSC that handles mobile terminating (MT) calls. GMSC always handles the MT calls in the PLMN, regardless of the

origin of the call. For most operators, MT call handling is done by a GMSC in the HPLMN; in this case, the HPLMN is the same as IPLMN. Once the call has arrived in the HPLMN, the HPLMN acts as IPLMN. When basic optimal routing (BOR) is applied, the IPLMN is not the same PLMN as the HPLMN.

6.5 GSM Mobility and Roaming

Roaming in the GSM network is possible through the separation of switching capability and subscription data. It should be noted that the grouping of several operationally independent PLMNs and forwarding data among them, enables the roaming service, in which the user can move across areas but keeping access to their subscribed services linked to their SIM. This means that the PLMNs should communicate between themselves to offer the user mobility. A GSM subscriber's subscription data is permanently registered in the HLR of HPLMN. The GSM operator is responsible for provisioning this data in the HLR. However, whatever the subscription conditions may be, an emergency call is the only service that is available anywhere in the system. The MSC and GMSC in a PLMN are not specific for one subscriber group. The switching capability of the MSC in a PLMN may be used by that PLMN's own subscribers, and also by inbound roaming subscribers. For example, one GSM user, who is a subscriber of PLMN-1, roams to PLMN-2. The HLR in PLMN-1 transfers the user's subscription data to the MSC/VLR in PLMN-2. The subscriber's subscription data remain in that MSC/VLR as long as the user is served by a BSS that is connected to that MSC. Even when the user switches the MS OFF and later ON again, the subscription data remain in the MSC. When MS is being switched off for an extended period of time, then the subscription data will be purged from the MSC. When the subscriber switches MS on again, the subscriber has to re-register with the MSC, which entails the MSC asking the HLR in the HPLMN to re-send the subscription data for that subscriber.

When a subscriber moves from one MSC service area (such as MSC-1) to another MSC service area (such as MSC-2), the HLR will instruct MSC-1 to purge the subscription data of this subscriber and will send the subscription data to MSC-2.

6.6 GSM PLMN Services

The GSM PLMN defines a group of communication capabilities that the service providers can offer to its subscribers. Features that can be supported in the GSM network, such as establishing a voice call, establishing a data call, sending a short message, and so on, are classified as basic services. The user needs to have a subscription in order to use the GSM basic service. The handling of basic services is fully standardized. Thus, when the subscriber roams into another GSM network, the user may use the basic services (which he/she subscribed to) in that network (provided that that those basic services are also supported in that new network). The HLR will send a list of subscribed basic services to the MSC/VLR during registration. When a GSM subscriber initiates a call, the MS supplies a set of parameters describing the circuit switched connection that is requested to the serving MSC. The MSC uses these to derive the basic service for this call. The rules for deriving the basic service are specified in GSM TS 09.07. The MSC then checks whether the subscriber has a subscription to the requested basic service, that is, whether the subscription data in the VLR contains that basic service. If the service is not subscribed to, then the MSC does not allow the call. The basic service is not transported over ISUP.

1. **Basic services** are divided into two groups: tele-services and bearer services.
   a. **Bearer Services** – These services give the subscriber the capacity required to transmit appropriate signals between certain access points (mobile user and network interfaces.), such as asynchronous data and synchronous data bearer services.

   b. **Tele-services** – The tele-services are telecommunication services as well as functions that enable communication between users, and are based on protocols agreed on by the network operators. Examples include speech transmission, SMS, e-mail, facsimile, teletext transmission. Please refer to GSM TS 02.03 for the available tele-services (TS).
2. **Supplementary Services** are the services offered to enrich the user experiences, and are modified or supplement the basic telecommunication services. They are offered together or in association with basic communication services. For example, the ability to put calls on hold, call waiting, and caller-ID, and so on. Supplementary services may be provisioned for an individual basic service or for a group of basic services, for example, a subscriber may have the barring of all outgoing calls for all tele-services and all bearer services, except SMS (tele-service group 20). Such a subscriber is barred from establishing outgoing calls (except for emergency calls), but may still send short messages. Some supplementary services may be activated or deactivated by the user. Examples include call forwarding and call barring. An operator may decide to bar certain subscribers or subscriber groups from modifying their supplementary services.

In addition, we have the value added services. Value added services are supplied by the respective service provider or network operator, and can be transmitted either via a normal telephone call or via SMS: examples include reserving a hotel room, a flight or a hire car.

6.7 GSM Interfaces

6.7.1 Radio Interface (MS to BTS)

The air interface between the BTS and MS is known as the Um interface. The manufacturers of network and MS might not be same, but these have to be compliant with each other, in order to work together in a GSM system. The air interface is defined, so that MS and network manufacturers can design their equipment independently following the standards so that the outcomes will be compatible. More about this and the radio transmitter design aspect will be discussed in the next chapter.

6.7.2 Abis Interface (BTS to BSC)

The interface between BTS and BSC is known as Abis standard interface. The primary functions carried over this interface are traffic channel transmissions, radio channel management, and terrestrial channel management. This interface mainly supports two types of communication links: (1) traffic channels at 64 kbps, which carry speech or user data for a full or half rate radio traffic channel, and (2) signaling channels at 16 kbps, which carry information for BSC-BTS and BSC-MSC signaling. The BSC handles LAPD channel signaling for every BTS carrier. The lower three layers are based on the OSI/ITU-T recommendation: physical layer (ITU-T recommendation G.703 and GSM recommendation-08.54), data link layer [GSM recommendation 08.56 (LAPD)], and network layer (GSM recommendation 08.58). Transparent and non-transparent are the two types of messages handled by the traffic management procedure part of the signaling interface. BTS does not analyze the transparent messages between the MS and BSC-MSC.

6.7.3 A Interface (BSC to MSC)

The “A” interface is used for interconnections between the BSS radio subsystem and MSC. The physical layer of the “A” interface supports a 2 Mbps standard CCITT digital connection. The signaling transport uses the message transfer part and the signaling connection control part of SS7. The data transfer and protocol on these interfaces are discussed in detail in the Chapter 8.

6.8 GSM Subscriber and Equipment Identity

The GSM system distinguishes explicitly the user and the devices and deals with these accordingly. The user and mobile equipment in the system separately get their own internally unique identifiers. The user identity is associated with a subscriber identity module (for example, IMSI associated with SIM) and the device identity is associated with the equipment number (for example, IMEI associated with mobile equipment). These are described in the next section. An MS has a number of identities including the International Mobile Equipment Identity (IMEI), International Mobile Subscriber Identity (IMSI), and the ISDN number. These are needed for management of subscriber mobility and for addressing all the network elements.

6.8.1 International Mobile Equipment Identity (IMEI)

The IMEI (International Mobile Equipment Identity) is a unique 15-digit code used to identify an individual GSM mobile station in a GSM network. It is stored inside the mobile device by programming the EPROM inside the MS and should not be changed subsequently. When new mobile equipment is registered for the first time for use in a network, its IMEI number is stored in the Equipment Identity Register (EIR) of the network.

$$\text{IMEI} = \text{TAC} + \text{FAC} + \text{SNR} + \text{spare}$$

where the TAC (type approval code) is determined by a central GSM/PCS body (6 digits), the FAC (final assembly code) identifies the manufacturer (2 digits), the SNR (serial number) uniquely identifies all equipment within each TAC and FAC (6 digits), and a spare (1 digit).

The format of an IMEI is AABBBB–CC-DDDDDD-E. The significance of each digit is explained in the Table 6.3.

Table 6.3 IMEI format

Digit	Significance
AA	Country code
BBBB	Final assembly code
CC	Manufacturer code. This varies according to the manufacturer, such as for NOKIA it is 10 or 20 and MOTOROLA it is 07
DDDDDD	Serial number
E	Unused

An IMEI is marked as invalid if it has been reported stolen or is not type approved. IMEI numbers are classified as follows. (1) White – valid GSM mobile stations. The WHITE list contains the series of IMSIs that have been allocated to MEs and can be used legally on the GSM network. (2) Grey – GSM mobile stations to be tracked. The network operator use a GREY list to hold the IMSIs of MEs that must be tracked by the network for evaluation purpose. (3) Black – barred mobile stations. The BLACK list contains the IMSIs of all MEs that must be barred from accessing the GSM network. This will contain the IMSIs of stolen and malfunctioning MEs.

The EIR is used to store three different lists of IMSIs. The network commands the MS to send its IMEI number during a call, or access attempt. Once it receives the IMEI number, the IMEI is passed to the EIR by the serving MSC and the IMEI check is performed in the EIR (black or white listed) and the result of the IMEI check is returned by the EIR to the serving MSC. EIR checks whether this is in a black or white list and if it is found that it is included in the black list, then the network simply send an “illegal ME” message and terminate the call or access attempt.

The IMEI number of most mobile phones can be displayed by dialing the code *# 06 #. Usually this is printed on the compliance plate under the battery.

6.8.2 International Mobile Subscriber Identity (IMSI)

The IMSI is a unique non-dialable number allocated to each mobile subscriber in the GSM system that identifies the subscriber and user subscription within the GSM network. IMSI is assigned to an MS at the time of subscription time by the network provider, when the subscriber receives a new SIM card. This is stored inside the subscriber identity module (SIM) and in the network side it is also stored in the HLR. The IMSI is a unique 15-digit code used to identify an individual user on a GSM network. It consists of three components: (1) mobile country code (MCC) – a 3 digits (home country), (2) mobile network code (MNC) – 2 digits (home GSM PLMN), and (3) mobile subscriber identity number (MSIN) – 10 digits.

6.8.3 Temporary International Mobile Subscriber Identity (TIMSI)

The TIMSI is a pseudo-random number generated from the IMSI number. The TIMSI is introduced in order to avoid the need to transmit the IMSI over-the-air, which helps to keep the IMSI more secure. The TMSI is assigned to an MS by VLR after the initial registration. This only has local significance in the area handled by the VLR. It is not passed to HLR. The maximum number of bits that can be used for the TMSI is 32. The TMSI is also stored temporarily in the SIM. Before the mobile is switched off, the current TMSI is stored into the SIM, so that during the next registration process this same number can be used to make the initial process faster.

6.8.4 Mobile Subscriber ISDN Number (MSISDN)

MSISDN is the mobile station's real telephone number, through which it is called by another party. Primarily the MSISDN and IMSI are separated, because of the confidentiality of the IMSI, as the IMSI should not be made public. One cannot derive the subscriber identity from the MSISDN, unless the association of IMSI and MSISDN as stored in the HLR has been made public. So using of a false identity is difficult. In addition to this, a subscriber can hold several MSISDN numbers for the selection of different services depending on the SIM. Each MSISDN of a subscriber is reserved for a specific service (voice, data, fax, etc.). The MSISDN categories follow the international ISDN numbering plan and therefore have the following structure: (1) country code (CC) – up to 3 digits in place; (2) national destination code (NDC) – typically 2–3 decimal places; and (3) subscriber number (SN) – maximum 10 decimal places. The MSISDN has a maximum length of 15 decimal digits. It is also stored in the HLR of the network. The country is internally standardized, complying with ITU-T E.164 series. For example, India has the country code 091, the USA 001, and so on. The national operator or regulatory administration assigns the NDC as well as the subscriber number SN.

6.8.5 Mobile Station Roaming Number (MSRN)

The mobile station roaming number (MSRN) is a temporary location dependent ISDN number. It is assigned by the locally responsible VLR to each mobile station in its area. Calls are routed to the MS by using the MSRN. On request the MSRN is passed to the HLR then to the GMSC. The MSRN has the same structure as the MSISDN: MSRN = CC + NDC + SN. The components of CC and NDC are determined by the visited network and depend on the current location. The SN is assigned by the current VLR and is unique within the mobile network. The assignment of MSRN is done in such a way that the

currently responsible switching node MSC in the visited network (CC + NDC) can be determined from the subscriber number. The MSRN can be assigned in two ways by the VLR: either at the registration when the MS enters into a new location area (LA) or each time when the HLR requests it for setting up a connection for incoming calls to the MS. In the first case, MSRN is also passed on from the VLR to HLR, where it is stored for routing. In the case of the incoming call, the MSRN is first requested from the HLR of the mobile station. This way currently responsible MSC can be determined, and the call can be routed to this switching node. Additional localization information can be obtained from responsible VLR. In the second case, the MSRN can not be stored in the HLR, as it is only assigned at the time of call set-up. Therefore the address of the current VLR must be stored in the table of the HLR. Once the routing information is requested from the HLR, the HLR itself goes to the current VLR and uses unique subscriber identification (IMSI and MSISDN) to request a valid roaming number MSRN. This allows further routing of a call.

6.8.6 Location Area Identity (LAI)

Each LA of a PLMN has its own identifier. The location area identifier (LAI) is also structured hierarchically and internationally unique, with LAI again consisting of an internationally standardized part and an operator dependent part: (1) country code (CC) – 3 decimal digits; (2) mobile network code (MNC) – 2 decimal places; and (3) location area code – maximum 5 decimal places. The LAI is broadcasted regularly by the BTs on the BCCH channel. Thus each cell is identified uniquely on the radio channel as belonging to an LA, and each MS can determine its current location through the LAI. If the LAI that is heard by the MS changes, the MS notices this LA change and requests the updating of it location information in the VLR and HLR – this is known as location update. The mobile station itself is responsible for monitoring the local conditions for signal reception, to select the base station that can be received best and to register with the VLR of that LA which the current base station belongs to. The LAI is requested from the VLR, if the connection for an incoming call has been routed to the current MSC using MSRN. This determines the precise location of the mobile station where the mobile can be subsequently paged. When the mobile station answers the exact cell and the base station becomes known, this information then can be used for call switching.

6.8.7 Local Mobile Subscriber Identity (LMSI)

The VLR can assign an additional searching key to each mobile station within its area to accelerate the database access. This is the local mobile station identity. Generally, an LMSI contains of 4 octets. The LMSI is assigned when mobile station registers with the VLR and is also sent to the HLR. The LMSI is not used any further by the HLR, but each time messages are sent to the VLR concerning a mobile station, the LMSI is added, so the VLR can use the short searching key for transactions concerning this MS. This type of additional identification is only used when the MSRN is newly assigned with each call. In this case, fast processing is very important to achieve short times for call set-up. As for the TMSI, an LMSI is also assigned in an operator specific way, and it is only unique within the administrative area of a VLR.

6.8.8 Cell Identifier (CI)

Within an LA, the individual cells are uniquely identified with a cell identifier (CI), which contains a maximum of 2×8 bits. Together with the global cell identity (LAI + CI), cells are thus also internationally defined in a unique way. In GSM, during execution of handover or after the handover is done, BSS informs the core network (CN) about the new cell that is being used by the MS.

6.8.9 Base Station Identity Code (BSIC)

In order to distinguish neighboring base stations in the GSM network, the BTSs are assigned a unique base transceiver station identity code (BSIC) which consists of two parts: (1) network color code (NCC) – color code within a PLMN (3 bits); and (2) base station color code (BCC) – BTS color code (3 bits). The BSIC is broadcasted periodically by the base station via the synchronization channel (SCH).

6.8.10 Identification of MSCs and Location Registers

MSCs and location registers (HLR and VLR) are addressed with ISDN numbers. In addition, they may have signaling point code (SPC) within a PLMN, which can be used to address them uniquely within the signaling number 7 network.

6.8.11 PIN and PUK

PIN stands for personal identification number. A PIN code is a numeric value used in certain systems as a password to gain access, and for authentication. A PIN is a 4–8 digit access code which can be used to secure your mobile telephone from use by others. PIN2 (personal identity number 2) is a 4–8 digit access code which can be used to access the priority number memory and the cost of calls. The PUK (personal unblocking key) and PUK2 are used to unlock the PIN and PIN 2 codes, respectively, if your SIM card is blocked. Generally, to change SIM card PIN the user has to dial ** 04 * old PIN * new PIN * new PIN #.

Further Reading

3GPP specification GSM TS 02.03. Teleservices Supported by a GSM Public Land Mobile Network (PLMN).

Mehrotra, A. (1997) *GSM System Engineering*, Artech House, Boston.

Mouley, M. and Pautet, M.-B. (1992) *The GSM System for Mobile Communications*, F-99120, Telecom Publisher, Palaiseau, France.

Steele, R. Lee, C.-C., and Gould, P. (2001) *GSM, cdmaOne and 3G Systems*, John Wiley & Sons Ltd., Chichester

Yacoub, M.D. (2002) *GSM Overview, Wireless Technology*, CRC Press, Boca Raton, ISBN 0-8493-0969-7.

7

GSM Radio Modem Design: From Speech to Radio Wave

7.1 Introduction

Earlier in Chapter 5, we discussed about the various medium access techniques. In a GSM system, a combination of time and frequency division multiple access (TDMA and FDMA) techniques are used to multiplex the air medium among the users. For uplink and downlink separation, GSM uses a frequency division duplex (FDD) technique. So the uplink and downlink frequencies are different. As mentioned earlier, GSM 900 was the first GSM system to be deployed, and uses a 25 MHz frequency band (935–960 MHz) in the downlink and another 25 MHz frequency band (890–915 MHz) in the uplink direction (Figure 7.1). The uplink and downlink bands are separated by 45 MHz. Again, FDMA divides the frequency bandwidth of the 25 MHz (maximum) in each direction (uplink and downlink) into 125 carrier frequencies, where each carrier frequency has a 200 kHz bandwidth. Although GSM 900 is most popular, currently there are several other types of networks available in the world using a GSM standard at different frequency bands, and these are given in Table 7.1.

As mentioned earlier, in the GSM 900 system, the total available number of uplink and downlink carriers = 25 MHz/200 kHz = 125. However, the last one cannot be used, as it is kept as a guard band with other adjacent wireless systems. So the total available carriers in each direction = 124, and of these again the top and button ones are used for additional guard band purposes with other wireless systems. Thus the suggested carrier frequency number = 122. Each base station (BS) is assigned carrier frequencies based on the licenses brought by the operator. Out of these, one carrier frequency will be used as the broadcast frequency and this is unique to that cell (BTS), which helps to identify it.

GSM uses frequency division duplexing (FDD) and supports the full duplex mode of operation, which means that two links (uplink and downlink) are connected simultaneously (more about this is discussed in Chapter 11). Uplink and downlink RF carriers are paired to allow simultaneous data flow in both directions. Each RF carrier frequency pair (uplink and corresponding downlink frequency) is assigned an absolute radio frequency channel number (ARFCN). The RF carrier pairs

Mobile Handset Design Sajal Kumar Das

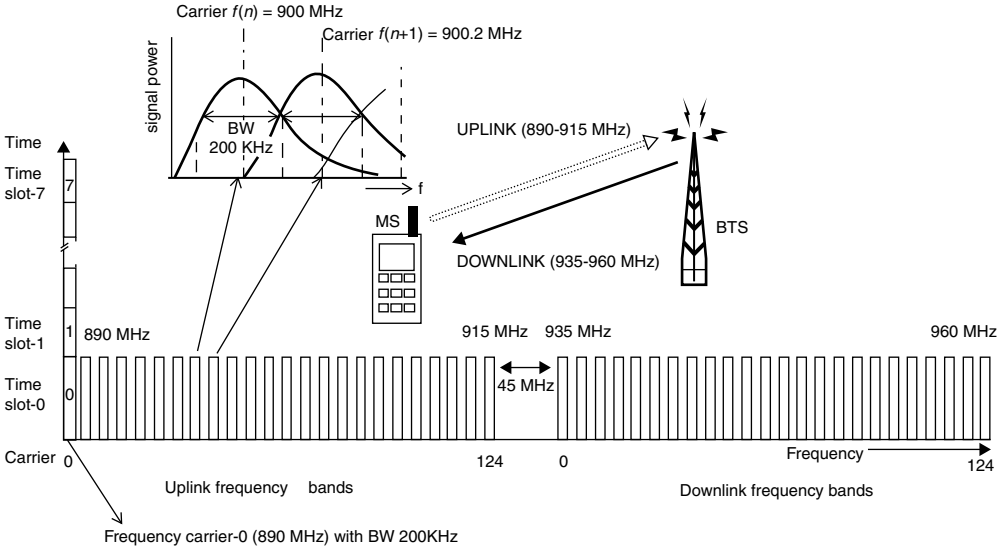


Figure 7.1 GSM uplink and downlink frequency bands

must have a separation of 45 MHz, so the uplink frequency (F_u) and the downlink frequency (F_d) are related by:

$$F_u = 890.2 + 0.2(N-1)\ \text{MHz}$$

$$F_d = 935.2 + 0.2(N-1)\ \text{MHz}$$

where

$N = 1, 2, \ldots, 124$

As mentioned, GSM also uses time division multiple access (TDMA) along with FDMA. Using a TDMA scheme each carrier frequency is again sub-divided in time, which forms logical channels (Figure 7.2). There are eight slots per frequency carrier. When the mobile is assigned to an information channel, an ARFCN (one uplink and one downlink frequency) and time slot (one in the uplink and one in the downlink) are also assigned. We can say, that each pair of radio frequency channel supports up to eight time slots, so the GSM 900 system can support up to $8 \times 124 = 992$ simultaneous channels, out of which some channels are used for broadcasting system specific information (discussed next).

In addition to frequency separation between uplink and downlink carriers, the downlink and uplink bursts of a duplex link are separated by three time slots. This removes the necessity for MS to transmit and receive simultaneously. Where the propagation delay between the MS and BTS is very small, the MS will receive a downlink burst from BTS, which then retunes to the uplink frequency and transmits an uplink burst three time slots later. This is shown in Figure 7.3.

Thus for technical reasons (the mobile handset design must be made simple, because at any given instant of time, it will perform either transmission or reception, but simultaneous Tx and Rx are not

Table 7.1 Different GSM network systems and associated frequency bands (there are 1 bands defined in 3GPP TS 45.005)

System	Band	Uplink frequency Band (MHz)	Downlink frequency Band (MHz)	Channel number
T-GSM-380	380	380.2–3 89.8	390.2–399.8	Dynamic
T-GSM-410	410	410.2–419.8	420.2–429.8	Dynamic
GSM-450	450	450.4–457.6	460.4–467.6	259–293
GSM-480	480	478.8–486.0	488.8–496.0	306–340
GSM-710	710	698.0–716.0	728.0–746.0	Dynamic
GSM-750	750	747.0–762.0	777.0–792.0	438–511
T-GSM-810	810	806.0–821.0	851.0–866.0	Dynamic
GSM-850	850	824.0–849.0	869.0–894.0	128–251
P-GSM-900 (primary GSM-900 band)	900	890.0–915.0	935.0–960.0	1–124
E-GSM-900 (extended GSM-900 band, which also includes standard GSM-900 band)	900	880.0–915.0	925.0–960.0	975–1023, 0–124
R-GSM-900 (railways GSM-900 band, which also includes standard and extended GSM-900 band)	900	876.0–915.0	921.0–960.0	955–1023, 0–124
T-GSM-900 (TETRA-GSM)	900	870.4–876.0	915.4–921.0	Dynamic
DCS-1800	1800	1710.0–1785.0	1805.0–1880.0	512–885
PCS-1900	1900	1850.0–1910.0	1930.0–1990.0	512–810

required in the GSM receiver, but in GPRS it depends on the class of mobile handset), it is necessary that MS and BS do not transmit simultaneously. Therefore, the MS is transmitting three time slots after receiving data from the BS. The time between sending and receiving data is used by the MS to perform various tasks such as processing data, measuring signal quality of the receivable neighbor cells, and so on. As shown in Figure 7.3, the MS does not actually send exactly three time slots later, after receiving the data from BS. Rather, based on the distance between MS and BS, a considerable propagation delay from MS to BS needs to be taken into account. That propagation delay, known as timing advance (TA), requires the MS to transmit its data a little earlier which is determined by the "three time slot delay rule." Without timing advance, bursts transmitted by two different MSs in slots adjacent in time, would overlap and interfere with each other. Some GSM system parameters are listed in the Table 7.2.

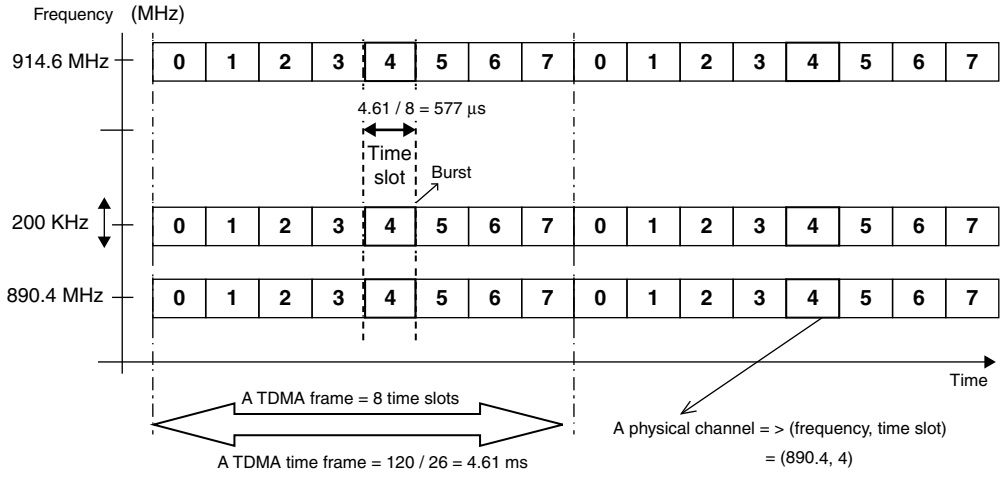


Figure 7.2 Time division multiplexing

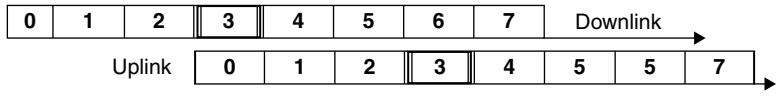


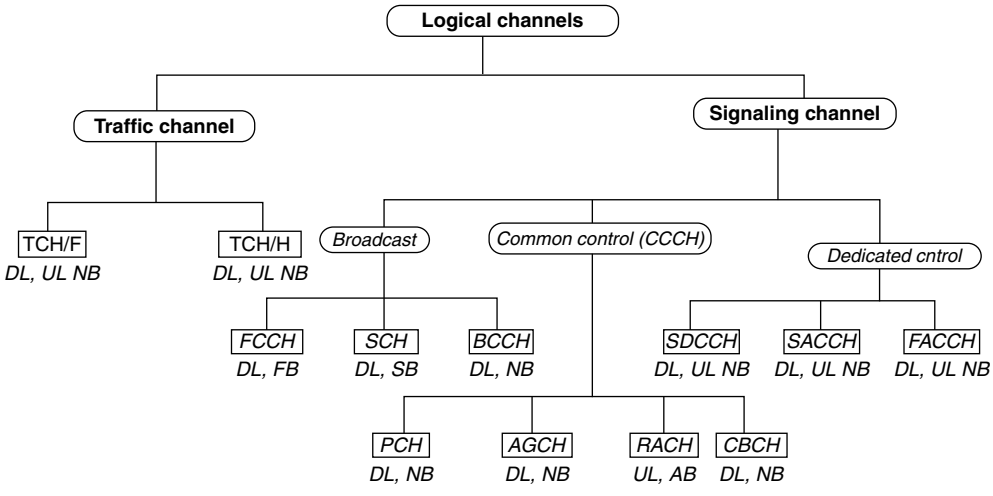
Figure 7.3 Uplink and downlink separation by approximately three time slots

Table 7.2 Some GSM system parameters

Multiple access method	TDMA and FDMA
Uplink frequencies (MHz)	890–915 (basic GSM)
Downlink frequencies (MHz)	935–960 (basic GSM)
Transmission method	8 time slots per carrier
Duplexing	FDD
Channel spacing (kHz)	200
Modulation	Gaussian minimum shift keying (GMSK) with normalized bandwidth 0.3
Bits per symbol	1
Error protection methods	Interleaving, channel coding – block and convolutional, slow frequency hopping 217 hops/s, adaptive equalization 16 s propagation time dispersion
Portable TX power, maximum/average (mW)	1000/125
Power control, handset and BSS	Yes
Speech coding and rate (kbps)	RPE-LTP/13
Speech/user channels per RF carrier	8
Channel rate (kbps)	270.833
Channel coding	Rate 1/2 convolutional
Frame duration (ms)	4.615
Overall channel bit rate	22.8 kbps

7.2 GSM Logical Channels

To establish and maintain a communication link between the two parties, apart from the user's information, some amount of signaling information also has to be exchanged over the link. Hence the data from MS would be either the user's real data (known as traffic data, such as coded voice data) or the signaling data (which is needed for establishing and maintaining the data exchange protocol over the air channel). Data, whether user traffic or signaling information, are mapped onto the physical channels by defining a set of logical channels. There are two types of GSM logical channels: (1) traffic channels (TCHs), which carry user speech or data and (2) signaling channels, which carry signaling and synchronization messages. The GSM channel structure is shown in the Figure 7.4.



UL- uplink, DL- downlink, TCH/F- traffic channel (full rate), TCH/H- traffic channel (half rate), FCCH- frequency correction channel, SCH- synchronization channel, AGCH- access grant channel, CBCH- cell broadcast channel, RACH- random aceess channel, BCCH broadcast control channel, PCH- paging channel, SDCCH- standalone dedicated control channel, SACCH- slow associated control Channel, FACCH- fast associated control channel,
NB- normal burst, SB- synchronization burst, AB- access burst, FB- frequency correction burst.

Figure 7.4 GSM logical channels structure

7.2.1 Traffic Channels

A traffic channel is required to carry the user's speech or data traffic. GSM uses two types of traffic channels.

7.2.1.1 Full-Rate Channel (TCH/FS)

The full-rate traffic channel (TCH/F) allows speech transmission at 13 kbps and in the GSM specification it is referred to as a TCH/FS channel to indicate that it is full-rate TCH carrying speech information. The full-rate TCH occupy a complete time slot in each TDMA frame, and also allows user data transmission at the primary user data rate of 9.6, 4.8, and $\leq$2.4 kbps, which is represented as TCH/9.6, TCH/F4.8, and TCH/F2.4, respectively. If a TCH/F is used for data communications, the usable data rate drops to 9.6 kbps

due to the enhanced security algorithms and data protection. The following full-rate speech and data channels are supported.

1. **Full-rate speech channel (TCH/FS)** – The full-rate speech channel carries user speech, which is digitized at a raw data rate of 13 kbps. After channel coding this rate becomes 22.8 kbps.
2. **Full-rate data channels for 9.6 (TCH/F9.6), 4.8 (TCH/F4.8), and 2.4 (TCH/F2.4) kbps** – The full-rate data traffic channel carries raw data, which is sent at a rate of 9.6, 4.8, or 2.4 kbps, respectively. After the addition of FEC bits, the rate for all these becomes 22.8 kbps.

7.2.1.2 Half-Rate Channel (TCH/HS)

In addition to these full-rate TCHs (TCH/F), half-rate traffic channels (TCH/H) are also defined. The half-rate channel is primarily intended to support the GSM half-rate speech coder, the design of which was finalized in January 1995. Half-rate TCHs double the capacity of a system effectively by making it possible to transmit two calls in a single channel. Thus two half-rate channels share one physical channel (discussed in the next section). The half-rate channel uses one time slot in every other (alternate) TDMA frame, and this means that each physical channel can support two half-rate TCHs. The half-rate TCH allows speech transmission at around 7 kbps (TCH/H) and data at primary user rates of 4.8 and $\leq$2.4 kbps, referred to as TCH/H4.8 and TCH/H2.4, respectively. For speech data, improved codecs have rates of 6.5 kbps, plus FEC and for packet data, it can be transmitted at 3 or 6 kbps. Figure 7.5 shows the concept of the FS and HS techniques.

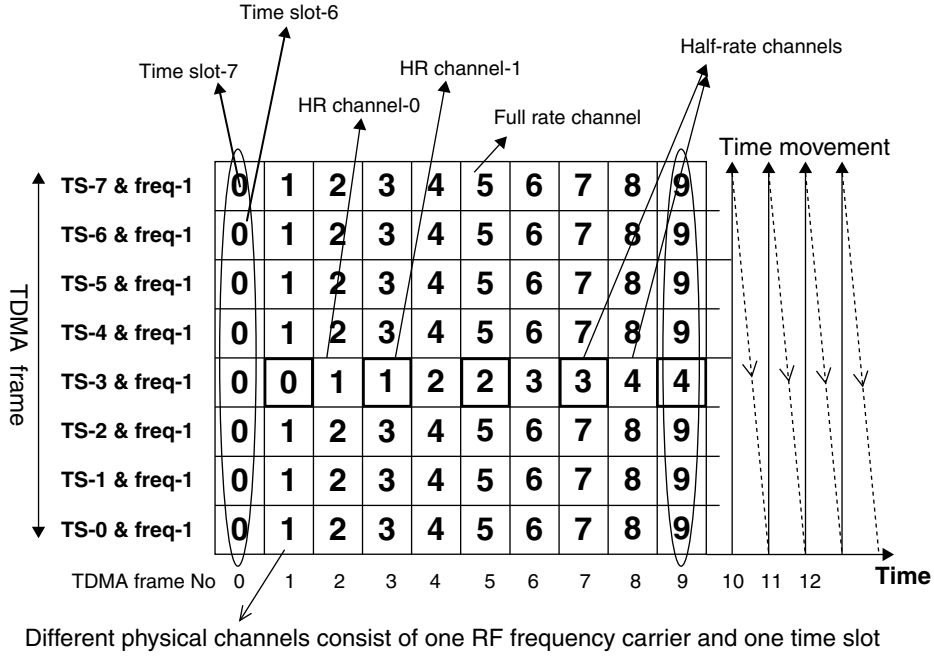


Figure 7.5 Time division technique in GSM (on the same carrier frequency) and concept of full-rate and half-rate channels

7.2.2 Signaling Channels

The signaling channels on the air interface are used for initial camp on, synchronization, call establishment, paging, call maintenance, and so on. There are three groups of signaling channels.

7.2.2.1 Broadcast Channels

The broadcast channels are used to broadcast synchronization and general network specific information to all MSs within a cell. They carry only downlink information and are responsible mainly for synchronization and frequency correction.

The broadcast channel operates on the downlink of a specific ARFCN within each cell, and transmits data only in the first time slot (TS0), for example, the physical channel for a broadcast channel is ($f_{\text{broadcast}}$, TS0). This channel is always ON, irrespective of whether any MS in the cell is listening to it or not. This is always transmitted with higher power. For better channel usage this physical channel ($f_{\text{broadcast}}$, TS0) is again time multiplexed (based on the TDMA frame number) between the following three different types of channels.

Frequency Correction Channel (FCCH)

This is a downlink only channel and mainly used for frequency correction. MS corrects its local clock frequency by using this channel. It is also used for synchronization of an acquisition by providing the boundaries between time slots and the position of the first time slot of a TDMA frame. The frequency correction channel (FCCH) is the simplest GSM logical channel, because all its information bits are set to zero. The FCCH consists of a frequency correction burst (FB), which consists of an all-zero bit pattern, which is why after GMSK modulation, these bursts produce a pure sine wave at a frequency of around 67.7 kHz (1625/24 kHz) above the carrier frequency. The FCCH is used by the MS in the initial stages of BTS acquisition to correct its internal frequency sources and recover the carrier phase of the BTS transmissions. This occupies TS0 for every first GSM frame (TDMA frame number 0) in broadcast frequency and is repeated on every 10th TDMA frame within a control channel multi-frame structure.

Synchronization Channel (SCH)

This is a downlink only channel and is mainly used for frame synchronization (TDMA frame number) and identification of the base station. The synchronization channel (SCH) information contains the network and BTS (cell) identification number, and all the necessary information needed to synchronize with a BTS. The synchronization channel contains full details of its own position within the GSM framing structure. The frame number (FN) which ranges from 0 to 2 715 647, is sent along with the base station identity code on the SCH burst. Using the information supplied on SCH, an MS can fully synchronize its frame counters with those of a BTS. The SCH information is transmitted using synchronization burst (SB).

As mentioned earlier, in addition to frame synchronization information, the SCH also contains a 6-bit base station identity code (BSIC). The BSIC consists of a 3-bit network color code and 3-bit BS color code (BCC), which is assigned during the network frequency planning. The term color is used because the assignment of these codes may be achieved by coloring the regions on a map according to the code that is in use within a particular area. Where two networks sharing the same frequency band have overlapping coverage, each of these will be assigned a different NCC, for example, their BCC may be the same but the NCC will be different. This ensures that the two networks can not use the same BSIC. In general, two PLMNs within the same country will use different frequency bands, and the use of the NCC only becomes important at international boundaries, where the coverage of two PLMNs using the same frequency bands may overlap. The BCC is used by the individual operators to ensure that co-channel BTSs have different BSICs. There is direct relationship between the BCC number and the training sequence (out of eight possible TSC) used in the normal burst (NB). This means that by decoding the BSIC for a particular BTS, the MS can determine which of the eight possible training sequences to expect in the normal bursts

transmitted by that particular BTS. SCH is broadcast in the TS0 of the broadcast frequency and appears on the TDMA frame immediately following the FCCH TDMA frame, for example, next to the frame containing FCCH.

Broadcast Channel (BCCH)

This is a downlink only channel and carries the information regarding general cell specific information such as local area code (LAC), network operator, access parameters, list of neighboring cells, details of control channel configuration used at BTS, a list of carrier frequencies used at the neighboring BTSs, and so on to the MS. The MS receives signals via the broadcast control channel (BCCH) from many BTSs within the same network and/or different networks. The TS0 of the broadcast frequency channel will contain BCCH channel data (apart from SCH, FCCH, and CCCHs) during its specific TDMA frames (discussed later in this chapter). It appears next to the SCH channel TDMA frame number (only for the first time) in a group of four consecutive frames in the control frame multi-frame structure.

System Information Broadcasting

The network continuously broadcasts information on the BCCH, irrespective of whether any MS is listening to it or not. When an MS camps into a cell, it reads the broadcast information in order to determine, if it can camp on that cell and uses these parameters to access the network. As a part of the broadcast information, a set of system information (SI) messages is broadcast using this. Some of these SI messages are mandatory (S12-4), while the others are optional to transmit. The following SI type messages provide all the necessary information.

1. **SI 1 message** – This optional message provides information about random access channel (RACH) control parameters and the frequencies that can be used in a particular cell.
2. **SI 2 message** – This is mandatory message and provides information about the RACH control parameters and BCCH allocations of the neighbor cells.
3. **SI 2 bis and 2 ter message** – These are optional message that provide the information on the extension of the BCCH allocation of the neighboring cells.
4. **SI 2 quat message** – This is an optional message and provides information about the additional measurement and reporting parameters.
5. **SI 3 message** – This is mandatory message and provides information about the RACH control parameters, the cell identity, the location area identifier, and cell selection parameters.
6. **SI 4 message** – This mandatory message provides information about the cell broadcast channel (CBCH) parameters in addition to the cell identity, the location area identification, and cell selection parameters.
7. **SI 7 and 9** – These are optional, and provide information about the cell re-selection parameters to be used in the cell.
8. **SI 9 message** – This is an optional message and provides the information about the BCCH scheduling.
9. **SI 13 message** – This message is transmitted on the BCCH, if GPRS (discussed later) is supported in the cell.
10. **SI 16 and 17 messages** – These provide the information about the cell selection and re-selection parameters.

Additionally, there are other SI messages, which are transmitted on a slow associated control channel.

7.2.2.2 Common Channels (CCCH)

These are a group of common uplink and downlink channels between the MS and the BTS, which are used to convey information from the network to the MSs and provide access to the network for MSs. The CCCHs include the following channels.

Paging Channel (PCH)

The paging channel (PCH) is a downlink only channel and BTS uses this channel to inform the MS about any incoming call. There are two different PCHs, a full-rate PCH and a reduced-rate PCH for use in cells with a limited capacity. The normal burst is always used for the PCH information transmission.

Access Grant Channel (AGCH)

The access grant channel (AGCH) is a downlink only channel and is used by the network to grant or deny an MS access request to the network by supplying it with the details of a dedicated channel (that is, TCH or SDCCH), to be used for subsequent communications. BTS allocates a TCH or SDCCH to the MS, thus allowing the MS access to the network. It also uses normal burst.

Random Access Channel (RACH)

This is an uplink only channel and is used by the MS initially to access the network, for example, at call set up or prior to a location update. The random access channel (RACH) is referred to as random because whenever any MS wants to send an RACH message, it just schedules to do so. There is no mechanism to ensure that at a given time instant not more than one MS should transmit in each RACH time slot. So there is a finite probability that two mobiles could attempt to transmit the RACH at the same time. This could result in neither of the access attempts being successful, as the two signals from different MSs collide at the channel and reach the BTS. If an MS receives no response from the BTS, it will attempt to access the BTS again after waiting a certain amount of time and this waiting period is random.

7.2.2.3 Dedicated Control Channels

There are three types of dedicated control channels, and these are bi-directional and user specific.

Slow Associated Control Channel (SACCH)

In a slow associated control channel (SACCH), the name "associated" indicates that it is meant to associate with some other channel, which is TCH (or SDCCH). SACCH is always associated with a traffic channel or SDCCH and maps onto the same physical channel (time slot, frequency combination). This is used for both uplink and downlink channels. On downlink, SACCH is used to send slow, but regularly changing control information to the MS, such as transmit power level instructions, or timing advance instructions for each users. However, on uplink, SACCH carries information such as the received signal strength, quality of TCH, as well as the BCH measurement results for different neighboring cells, and so on. When the MS is engaged in a call, during that time a certain amount of signaling information also needs to flow in order to maintain the call. For example, MS will continuously report the received signal level of BCCH carriers of its neighboring BTSs in order to make a handover decision. Non-urgent information such as measurement data, is transmitted using the slow associated control channel (SACCH). This channel is always present when a dedicated link (traffic) is active between the BTS and MS. This occupies one time slot in every 26 slots in the case of Full-rate TCH. SACCH messages may be sent once every 480 ms, for example, approximately every 2s.

Fast Associated Control Channel (FACCH)

As with SACCH, a fast associated control channel (FACCH) is also an associated channel, but it is used in urgent situations, as the name fast indicates. This channel is known as the fast associated control channel, because of its ability to transfer information between the BTS and MS more quickly than SACCH. The FACCH is used only when the information that needs to be transmitted cannot wait till the next SACCH frame (for example, SACHH appears after the 26 TDMA frame). An FACCH signaling block is used to replace a single (20 ms) speech block exactly and a complete FACCH message may be sent once every

20 ms, as the TCH. If the data rate of the FACCH is insufficient, a borrowing mode is used, for example, an additional bandwidth is borrowed from the TCH; this happens for messages associated with call establishment authentication of the subscriber, handover decisions, and so on. So, more urgent information such as a handover command is sent using time slots that are stolen from the traffic channel. The TCH slot is stolen and FCCH data burst is placed over there, and this stealing of the TCH slot is indicated by stealing a bit flag in the normal burst. Two stealing flags are there in an NB, these are set based on odd numbered bits or even numbered bits (or both) of the NB are stolen. This is a downlink and uplink channel.

Standalone Dedicated Control Channel (SDCCH)

When the traffic channel is already allocated, during this time associated channels such as SACCH and FACCH help to carry some of the dedicated signaling between BTS and MS, but this is also required when the traffic channel is not established (signaling during call set up, location update, etc.). This can be addressed by allocating a TCH and its associated SACCH to carry the information. However, this is not a perfect solution, as the data transfer requirements for such a process (for example, location updating) are much less compared with speech transmissions. So the radio channel (TCH) will be wasted and not of much use. For this reason a lower data rate channel has been defined, which has around one-eighth of the TCH full-rate capacity. This channel is known as a standalone dedicated control channel (SDCCH). The name standalone indicates it can exist independently of any TCH. The SDCCH also has an associated channel SACCH (as in the TCH case). Because the SDCCH always carries signaling traffic, there is no frame stealing and consequently it does not require an FACCH. Alternately, one could argue that the SDCCH is constantly in the FACCH mode. SDCCH is both an uplink and a downlink channel and this is the basic channel that is used for exchange of signaling messages between the MS and the network.

7.2.3 Cell Broadcast Channel

The CBCH is the common control channel, implementation of which is optional. It is used to provide cell broadcast services (CBS). The cell broadcast service provides the means to broadcast a number of unacknowledged general CBS messages. CBCH is used to transmit short alpha numeric messages to all the MSs within a particular cell. These messages appear on the MSs display and a subscriber may choose to receive different messages by selecting alternative paging, similar to the way the teletext system works on broadcast television.

7.3 GSM Physical Channel

When an MS and a BTS communicate, they are assigned a specific pair of radio frequency (RF) carriers one for uplink (MS transmits) and other for downlink (BTS transmits) with a separation of 45 MHz. Also, they choose a given time slot in each consecutive TDMA frame. This combination of time slot and carrier frequency forms the physical channel. As discussed earlier, one RF carrier will support eight physical channels from time slots 0 to 7.

7.3.1 Mapping of Logical Channel to Physical Channel

The logical channel data is mapped to the physical channel. The physical channel data is known as burst. The data from logical channels first undergo various physical layer processings (discussed later) and are next combined with tail bits, a training sequence, and guard bits to form a burst. Then it is mapped to a specific (time, frequency) slot, as shown in the Figure 7.6. Depending on the specific logical channel different burst types are used.

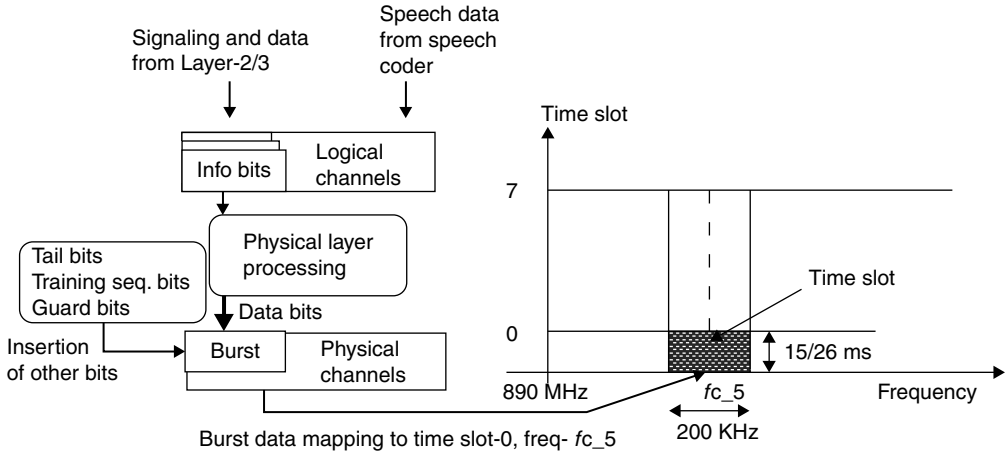


Figure 7.6 Mapping of logical to physical channels

7.4 GSM Bursts

Earlier, we saw that each TDMA slot (0–7) contains the data that needs to be transmitted or received over the air. The data transmitted or received in the form of bursts, are designed to fit within those slots. A burst has a finite duration, and occupies a finite part of the radio spectrum. They are sent in time and frequency windows.

General requirement is: 1 multi-frame (=26 TDMA traffic frame, discussed later) needs to be transmitted in 120 ms. Thus the duration of each TDMA frame = 120/26 = 4.615 ms.

Again, each TDMA frame consists of eight time slots, so each slot duration = (120/26) × (1/8) = 15/26 ms = 576.9 µs. The duration of a time slot is equal to the duration of a burst period (BP), which is 576.9 µs. Each time slot in a TDMA frame is numbered from 0 to 7 and these numbers repeat for each consecutive frame. In this 576.9 µs time slot interval, a series of 156.25 bits (called bursts) are sent.

We will now analyze the content of these bursts. The user data, speech, signaling data, handover commands, and so on, are mapped onto these TDMA bursts as information bits. The information bits are passed to the physical layer from the higher layer at every defined time interval (Transmit Time Interval = 20 ms). This is then processed in the physical layer (channel coding, interleaving, ciphering, and then burst data forming) and a predefined training sequence is added for channel estimation purposes, and also the tail bits and guard bits are inserted to form a complete burst.

7.4.1 Burst Structure

Generally, each burst contains:

1. **Data Bits:** The information bits are passed from a higher layer to the physical layer, which are then processed and segmented to form burst data bits.
2. **Training Sequence:** The training sequence bits in the middle of the burst are used by the receiver to synchronize, to compensate for time dispersion produced by multi-path propagation, and for channel estimation purposes. This is a fixed set of sequences known to both the transmitter and receiver. The training period in the middle is sometimes called midamble (minimum distance to useful bits), and is mainly used for channel estimation, equalization, and demodulation purposes.

The training sequence allows digital phones to overcome one of the problems that plague analog phones. Radio waves bounce off things such as buildings and hills. This can cause interference in analog phones, because it means the waves from the base station follow different paths of varying lengths on their way to the phone, so some arrive later than others. Digital phones combat this problem by comparing the training sequence they receive with a copy of the sequence stored in their local memory. The phone can then work out how the channel has corrupted the training sequence of the incoming signal and on knowing this, it corrects the data bits (as the training sequence and data bits are embedded inside a burst). The equalizer works by first looking at the channel-filtered training sequence and then adjusting its own filter response to yield the original undistorted training sequence. Normally the equalization is done on a burst by burst basis.

3. **Tail Bits:** Indicates the start and end of the burst real data part, after that the guard bit will follow it. The tail bits in the normal burst are always set to zero to ensure that the Viterbi decoder begins and ends in a known state. This leaves time available for transmission power ramp-up/down at the start and end of the frame.
4. **Guard Bits:** The guard bit is inserted at the end of each burst to avoid the overlap of two consecutive transmitted bursts in the channel. The guard bit number varies based on the burst type. For a normal burst, the 8.25-bit guard bit time allows for some propagation time delay in the arrival of bursts and prevents overlapping of bursts with the previous ones. An access burst (AB) is the first burst, which will be sent by the MS to gain access to the network, so a long guard time of 68.25-bit time is used in case of AB. This guard time corresponds to a propagation distance of 75 km, or a maximum cell radius of 37.5 km.

The GSM specification defines five different types of bursts.

7.4.1.1 Frequency Correction Burst (FB)

The frequency correction burst (FB) is used by the MS to detect a special carrier, which is transmitted by every BTS in a GSM network. The carrier is called the broadcast control channel carrier. This is also used by the MS as a frequency reference for their internal time bases. The burst structure of FB is shown in the Figure 7.7. It contains 3-bit tail bits, 142 info bits, 3-bit tail bits, and 8.25 bits of guard bits (0.25 bits signify one quarter bit). All bits in the frequency correction burst are set to zero (including the tail bits). After GMSK modulation, this results in a pure sine wave at a frequency around 67.7 kHz (1625/24 kHz) higher than the RF carrier center frequency.

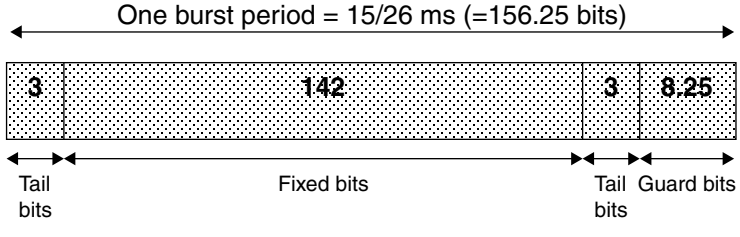


Figure 7.7 Frequency correction burst

7.4.1.2 Synchronization Burst (SB)

The basic structure of a synchronization burst (SB) is shown in the Figure 7.8. It carries 78 bits of coded data formed into two blocks of 39 bits on either side of a 64-bit training sequence. The burst carries details

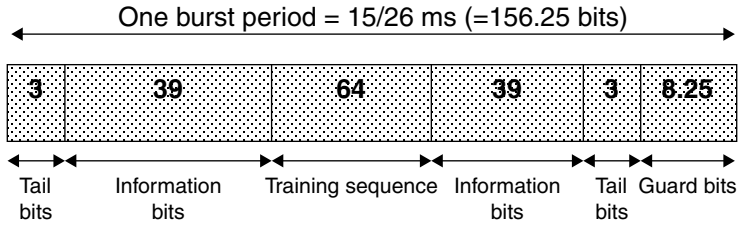


Figure 7.8 Synchronization burst

of GSM frame structures and allows an MS to fully synchronize with the BTS. The synchronization burst is the first burst that the MS has to demodulate and because of this the training sequence is extended to 64 bits to provide a larger autocorrelation peak than the 26 bit sequence of the normal burst. It also helps resolving larger multipath delay spreads.

The synchronization bursts (SB) transmitted by every BTS in the GSM system use the same training sequence, form bit number 42 to bit number 105 in the burst. The arrangement for the training sequence is shown below:

```
b42,b43,b44,.....,b105 =
(1,0,1,1,1,0,0,1,0,1,1,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,
  1,1,1,1,0,0,1,0,1,1,0,1,0,1,0,0,0,1,0,1,0,1,1,1,0,1,1,0,0,
  0,0,1,1,0,1,1)
```

An MS can use this training sequence to synchronize to the BTS transmission to stay within a quarter-bit of accuracy.

7.4.1.3 Normal Burst

This burst is normally used by most of the logical channels. It consists of a 26 bit training sequence surrounded by two 58-bit information blocks. The detailed content of the normal burst is shown in Figure 7.9.

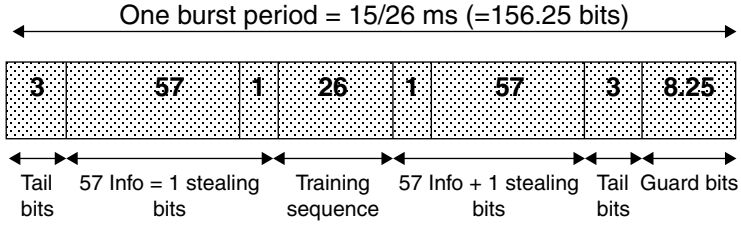


Figure 7.9 Normal burst

The known training sequence (TS) is placed at the middle of each burst. This will help to estimate the time shift (towards the left or right in the time scale) of the received burst at the receiver due to delay and multipath in the channel. This also helps to minimize the error in the information bits farthest from the TS. Consequently, the first section of the burst must be stored before demodulation can proceed. The training sequence in an NB consists of a 16-bit core sequence, and this is extended in both directions by copying the

first five bits to the back of the sequence and the last five bits to the front of the sequence. The central 16 bit is chosen to have a highly peaked autocorrelation function, when passed through the GMSK modulator, and the repeated bits at either end ensure that the resulting channel estimate may be up to 5 bits wide before being corrupted by the information bits. The specification defines eight different training sequences for use in normal burst (NB), each with low cross-correlation properties following GMSK modulation. Each training sequence is described by a training sequence code (TSC). A list of these sequences is given in Table 7.3. Generally, the training sequence used in NB and the BTS number (base station color code) are linked, which indicates that in a cellular structure each adjacent cell (BTS) should use different TSC in NB. Potential co-channel cells (for example, their own and all surrounding neighboring BTSs) will use different training sequences for transmitting NB to prevent the channel estimate being corrupted by an interference signal.

Table 7.3 Training sequences for different BTSs

Training sequence code	Training sequence bits (bit_61, . . .bit_86)
0	(00100101110000100010010111)
1	(00101101110111100010110111)
2	(01000011101101000100011110)
3	(01000111101101000100011110)
4	(00011010111001000001101011)
5	(01001110101100000100111010)
6	(10100111110110001010011111)
7	(11101111000100101110111100)

In the previous section, we noted that SB contains a long length training sequence (64 bit) compared with other burst types, because SB is the first burst that needs to be demodulated (although FB is searched first, it does not require demodulation) and all BTSs use the same training sequence bits for SB. Once SB is decoded the base station identity code (BSIC) is known, and from that the base station color code (BCC) is known. NB contains a 26-bit training sequence, and a training sequence used by any BTS depends on its BCC number. Hence once the BCC is known, the training sequence used inside the NB will be known to the MS, as BCC and TSC are mapped.

7.4.1.4 Access Burst

This is mainly used by the MS to access the network initially and is the first uplink burst that a BTS will have to demodulate from a particular MS. An AB burst structure is shown in the Figure 7.10. This consists of a 41-bit training sequence followed by 36 information bits. Here the number of tail bits at the beginning

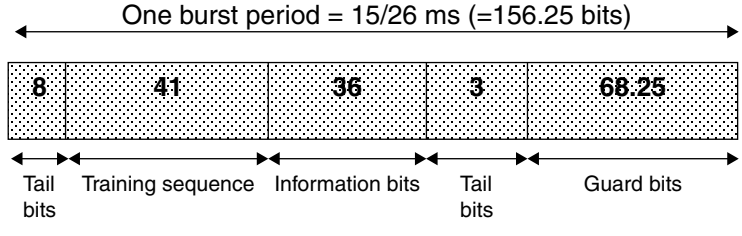


Figure 7.10 Access burst

of the burst is increased to eight. The extended tail bits at the front of the burst are: b0,b1,b2,b3, . . ., b7 = (0,0,1,1,1,0,1,0). This is then followed by a training sequence:

```
b8,b9,b10,.......,b48 =
(0,1,0,0,1,0,1,1,0,1,1,1,1,1,1,1,1,0,0,1,1,0,0,1,1,0,1,0,
  1,0,1,0,0,0,1,1,1,1,0,0,0).
```

Then 36 bits of information are added. All the tail bits at the end of the burst are set to zero. In the end there is large number of guard bits (=68.25 bits) padded to avoid overlapping of successive bursts.

The access burst is shorter compared with any other bursts with respect to the total length of information bits inside the burst, due to the space being filled with a long guard period. This is included to compensate for the propagation delay between the MS and BTS. Once a duplex link has been established, a closed loop timing advance mechanism is activated to ensure that the MS uplink burst arrives at the BTS within the correct time slots. However, this is not possible by only using this mechanism, as a guard period of 68.25-bit periods, equivalent to 252 μs, which allows the MS to be up to 38 km from the BTS before its uplink bursts, will spill into the next time slot. This puts a limit on the cell size.

7.4.1.5 Dummy Burst (DB)

A dummy burst (DB) is similar to an NB. It uses the same structure and the same training sequences as NB, but here the information bits are set to a predefined bit sequence. The DB is used to fill inactive time slots on the BCCH carrier, as the BCCH carrier should be always transmitted continuously and at a constant power.

Different channels use different types of burst. This is detailed in Table 7.4.

Table 7.4 Burst type usage for different channels

Logical channel	Abbreviation	Uplink/ downlink	Channel used for	Burst type used
Broadcast channel (BCH)	BCCH	DL	System information broadcast	NB
	FCCH	DL	Cell frequency synchronization	FB
	SCH	DL	Cell time synchronization and identification	SB
Common control channel (CCCH)	PCH	DL	MS paging	NB
	RACH	UL	MS random access	AB
	AGCH	DL	Resource allocation	NB
	CBCH	DL	Short messages broad cast	NB
Dedicated control channel	SDCCH	UL/DL	General signaling	NB
	SACCH	UL/DL	Signaling associated with the TCH	NB
	FACCH	UL/DL	Handover signaling	NB
Traffic channel (TCH)	TCH/FS	UL/DL	Full-rate voice channel	NB
	TCH/HS	UL/DL	Half-rate voice channel	NB
	TCH/F2.4 TCH/F4.8 TCH/F9.6 TCH/F14.4	UL/DL	Full-rate data channels	NB
	TCH/H2.4 TCH/H4.8	UL/DL	Half-rate data channels	NB

7.5 Burst RF Output Spectrum

The bursts are bounded in a time scale of slot duration (577 μs) and bounded in a frequency scale of 200 kHz bandwidth. Thus the design should be such that the burst should not spread more than the defined limit in time or frequency scale. In a TDMA system, the RF output spectrum of the transmitted signals is not only determined by the modulation process used, but also by the switching transients. Switching transient occurs when the bursts of RF energy are transmitted. This tends to widen the RF spectrum of the transmitted signal. So instead of an abrupt start and end to the burst-out power, a ramp shape at the start and at the end of the burst will help to reduce this effect. This why, during the transmission of a burst, instead of just keying the transmitter sharp ON and OFF, the output power is ramped up during the start and ramped down during the end of the burst. The transmitted information bits in the burst (which are kept in the middle of the burst) must not be affected by this process of power ramping, which is performed at the beginning and end of the slot. The variation of the transmitted power with respect to time for NB is shown in Figure 7.11, and it is also shown that the useful part of a burst (=147 bits) is one bit shorter than the active part of NB (=148 bits), and it begins halfway through the first bit period. When this part is transmitted, the amplitude of the modulated RF signal should remain at a constant level. The same mask also applies in the case of SB and FB. However, for AB, the useful part is reduced to 87 bits, and the constant level time duration will be $87 \times 3.69 = 321.2\,\mu s$.

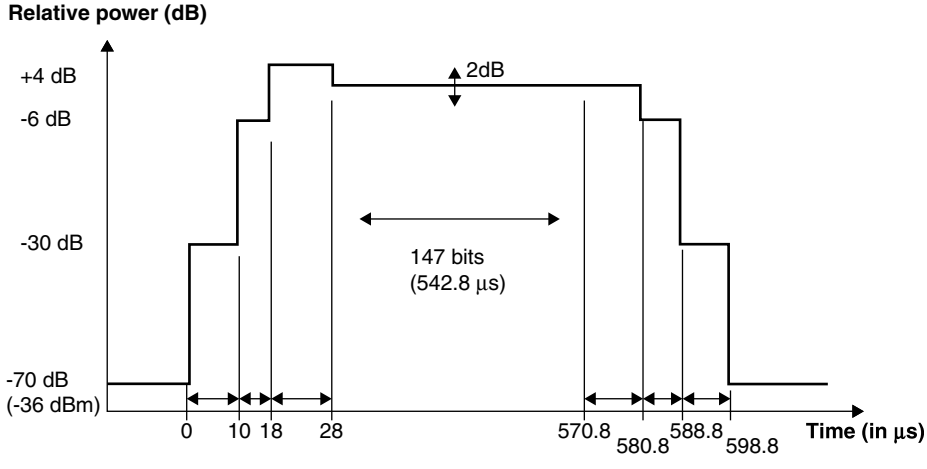


Figure 7.11 Power ramping for normal burst

7.5.1 RF Characteristics

7.5.1.1 Transmission Characteristics

As shown in Table 7.5, mobile devices are classified into several categories according to their maximum allowable transmitted output power capability. In GSM 900, most of the mobiles available on the market are class 4 handheld terminals, whereas class 2 terminals are used as vehicle-mounted equipment. The class 4 and 5 MSs are denoted as "small MS." In DCS 1800, the typical class is class 1.

These output power levels are maximum permitted values. However, the maximum power value is not used all the time, rather, the transmitted power values are reduced according to the commands that are sent by the network to the MSs. Each MS has the ability to reduce (or increase) the output power in steps of 2 dB, when it receives the command from the BTS. The output power can be reduced from its maximum

Table 7.5 MS power class

Power class	GSM-400, GSM-900, GSM-850 Nominal maximum output power	DCS-1800 Nominal maximum output power	PCS-1900 Nominal maximum output power
1	20 W (43 dBm)	1 W (30 dBm)	1 W(30 dBm)
2	8 W (39 dBm)	0.25 W (24 dBm)	0.25 W (24 dBm)
3	5 W (37 dBm)	4 W (36 dBm)	2 W (33 dBm)
4	2 W (33 dBm)		
5	0.8 W (29 dBm)		

value (as defined in the table) down to a minimum value of 5 dBm (3.2 mW) for a GSM 900 MS or 0 dBm (1 mW) for a DCS 1800. The output power level should be such that it is sufficient (optimum) to maintain an acceptable signal quality with the BTS (as more transmitted power means more interference in the system, more battery power consumption and less power means poor link quality with the BTS). The power up or down commands are sent on the basis of measurements, which are performed by the MS and by the BTS. For instance, with a class 4 MS, the range of transmission can be several kilometers, but if the MS is moving closer to the BTS, it may receive a request from the network to decrease its output power level. This procedure is called power control. The power control helps to improve the system performance by reducing the interference level caused to the other users. Above all, it helps to prolong the battery life of the MS. As with MS, for the BTS transceiver (TRX) the power classes are also defined and are given in Table 7.6.

Table 7.6 BTS TRX power level

TRX Power class	GSM-400, GSM-900, GSM- 850 Maximum output power	DCS-1800 and PCS-1900 Maximum output power
1	320 (<640) W	20(<40) W
2	160(<320) W	10(<20) W
3	80(<160) W	5(<10) W
4	40(<80) W	2.5(<5) W
5	20(<40) W	
6	10(<20) W	
7	5(<10) W	
8	2.5(<5) W	

Another option is that the BSS can utilize downlink RF power control, with up to 15 steps of power control levels with a step size of 2 dB. One thing to be noted is that the power control on the downlink is not used on the broadcast frequency, which is always transmitted with constant output power.

7.5.1.2 Reception Characteristics

The minimum link budget requirement needs to be satisfied in order to establish a communication link, which has already been discussed in Chapter 2. However, owing to fading, when the link quality degrades the BER in the receiver tends to increase. Here BER stands for bit error rate, FER for frame erasure ratio, that is, incorrect-speech-frames ratio, and RBER for residual BER, which is defined as the ratio of the number of errors detected over the frames defined as "good" to the number of transmitted bits in the good

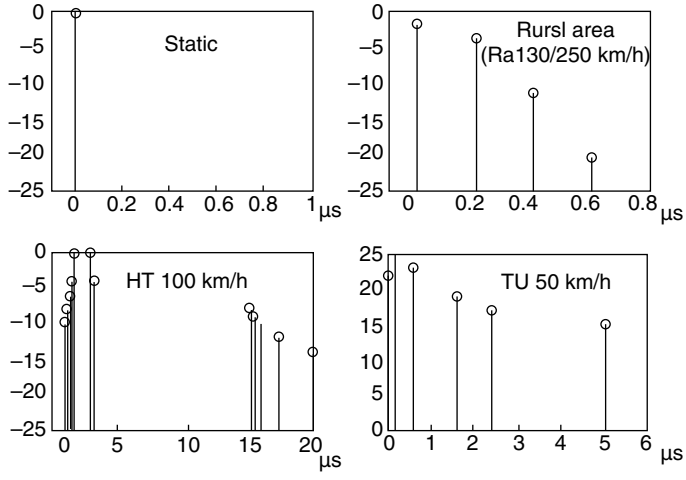


Figure 7.12 Channel profiles (channel tap gains with respect to time for different types of channels)

frames. For a given set of radio channel conditions, the performance of MS and BTS receivers are specified by defining a maximum permissible BER for each of the different GSM logical channels. The specification defines a minimum performance, in terms of BER, for a given set of receiver input conditions. The receiver performance is defined as a maximum channel BER or FER, for a given logical channel received at reference sensitivity level for various radio channel propagation conditions. As the GSM system is designed to operate in many different environments, from rural areas to dense urban settings, so it is important that the specifications reflect this by specifying the performance of the MS and BTS receivers over a wide range of different operational environments. GSM specifications (GSM 05.05) define four different channel models. Each channel model consists of a number of independently fading impulses or paths at different time delays. These models represent several environments (Figure 7.12) and are classified as:

1. Typical urban (TU*x*)
2. Rural area (RA*x*)
3. Hilly terrain (HT*x*)
4. Static channel

In the above definitions, the *x* stands for the velocity of the mobile, in km/h. Channel fading characteristics are governed by the speed of the MS, which is included in the channel type nomenclature, such as TU50 means typical urban with MS speed of 50 km/h. The various propagation models are represented by a number of taps, each determined by their time delay and average power. Each simulated channel model is defined for a six-tap channel simulator; however, some channels also include a 12-tap setting for use with larger simulators. In Figure 7.13, the block diagram of a wide band channel simulator is shown. The delay blocks are essentially simulating the path delay for different multipaths, and the tap gains are the path gains, which are adjusted dynamically and help to find the channel transfer functions based on maximum energy or tap method.

The Rayleigh distributed amplitude of each tap varies according to a Doppler spectrum. In addition to these three multipath fading channels, the static channel is also defined. This is a simple single-path constant channel. With this channel, the only perturbation comes from the receiver noise of the measured equipment. The specifications also define a further channel model, which is used to test the performance of

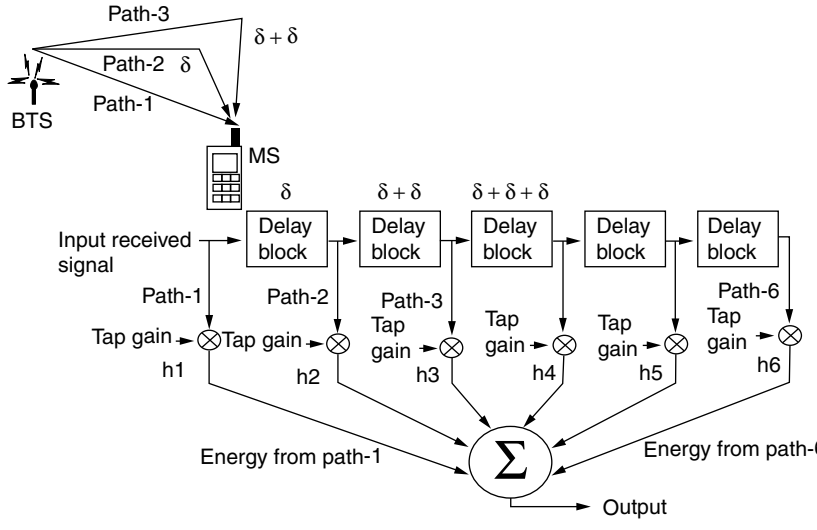


Figure 7.13 Block diagram of a wide band channel simulator (6 taps)

the Viterbi equalizer (EQ). The specifications state that the receiver must achieve a BER of $\leq 3\%$, without channel coding, over an EQ50 channel where the received signal power is 20 dB above the receiver's minimum sensitivity level, which means the GSM system should operate in the presence of delay spreads of up to 16 μs. The noise performance of a receiver is defined in terms of sensitivity and selectivity of the receiver (see Chapter 4). Table 10.6 (Chapter 10) shows the reference sensitivity performance for GSM 900. Note that frequency hopping may be used for the sensitivity performance measurements.

Apart from noise and channel fading, another important characteristic which affects the receiver performance is the presence of an interferer. The interference limited performance of the receiver is defined for a number of reference levels, which are specified relative to the desired signal. This is specified either for co-channel interference or adjacent channel interference. The level of the useful signal is set 20 dB higher than for the sensitivity evaluation, and a GMSK interfering signal is added, either at the same frequency or with an offset of 200 or 400 kHz from the carrier. The reference carrier to interference (C/I) ratios for both co-channel and adjacent channel interference are as mentioned below;

- Co-channel interference (C/I) = 9 dB
- First adjacent channel interference (200 kHz) – (C/I) = −9 dB
- Second adjacent channel interference (400 kHz) – (C/I) = −41 dB
- Third adjacent channel interference (600 kHz) – (C/I) = −49 dB

With the above given input conditions, in specification "GSM 05.05" the receiver performance is defined in terms of BER and FER for different logical channels and propagation conditions.

7.6 Channel Allocation

When a cellular provider applies for a GSM network deployment license, it buys several frequencies for traffic and a few for broadcast channel transmission. Among these, one broadcast channel frequency along with several other frequencies (for traffic) are allocated for a specific cell (for example, BTS). The available broadcast frequencies are distributed efficiently between the different adjacent cells by using

frequency planning, so that no two adjacent cells use the same broadcast frequency. Obviously these frequencies lie in the band of 935–960 MHz for downlink and similarly there will be a set of corresponding uplink frequencies (which lie in the band of 890–915 MHz) that the BTS will instruct the MS to use later during the call set up.

For example, say "f_{b1}," "f_{b2}," "f_{b3}," "f_{t1}" "f_{t2}" are the frequencies (in downlink) licensed by operator A and out of these one broadcast frequency-"f_{b1}," and other two traffic frequencies-"f_{t1},""f_{t2}" are used in cell-1. Each BTS will only have one broadcast channel (BCCH) carrier frequency, but can have several traffic channel (TCH) carrier frequencies. Using "f_{b1}" a total of eight (time slots TS0 to TS7) physical channels can be supported. However, out of the eight, TS0 of the broadcast frequency is always reserved for the broadcast channel (FCCH, SCH, BCCH, CCCH) information transmission. In addition to this channel, based on the cell size (and some other conditions), other time slots of this frequency are also used for broadcast or used for traffic (discussed in the next section) channel information transmission.

Now, using these three frequencies a total of $3 \times 8 = 24$ physical channels will be supported by this particular cell BTS, but of these, the TS0 of the broadcast frequency (for example, f_{b1}, TS0) will be always used for the broadcast channel's information transmission. Thus the remaining 23 channels can be used for user specific data transmission. The physical channel arrangement from the BTS (downlink) side will be – Physical Channel-0: (f_{b1}, TS0) is used for the broadcast channel information transmission. This is again time shared (time/frame multiplexed) between FCCH, SCH, BCCH, CCCH (and SDCCH in a combined mode). This means that in the (f_{b1}, TS0) physical channel, FCCH, SCH, BCCH, CCCH logical channels appear in a predefined manner based on the TDMA frame number in the signaling channel multi-frame structure. Physical Channel-1: (f_{b1}, TS1) can be used for full-rate traffic channel-1 (TCH + SACCH), or SDCCH (+ SACCH). Physical Channel-2: (f_{b1},TS2) can be used for full-rate traffic channel-2 (TCH + SACCH), or SDCCH (+ SACCH) or time multiplexed among BCCH, CCCH, and SDCCH in a combined mode. Physical Channel-3: (f_{b1},TS3) can be used for full-rate traffic channel-3. Physical Channel-4: (f_{b1},TS4) can be used for full-rate traffic channel-4 (TCH + SACCH), or SDCCH (+ SACCH) or time multiplexed among BCCH, CCCH, and SDCCH in a combined mode. Physical Channel-5: (f_{b1},TS5) can be used for full-rate traffic channel-5. Physical Channel-6: (f_{b1},TS6) can be used for full-rate traffic channel-6, (TCH + SACCH), or SDCCH (+ SACCH) or time multiplexed among BCCH, CCCH, and SDCCH in a combined mode. Physical Channel-7: (f_{b1},TS7) can be used for full-rate traffic channel-7. Physical Channel-8: (f_{t1}, TS0) can be used for full-rate traffic channel-8. Physical Channel-9: (f_{t1},TS2) can be used for full-rate traffic channel-9. Physical Channel-10: [f_{t2}, TS0 (odd)] can be used for half-rate traffic channel-12. It goes on like this, but some of the traffic channels may be unused at that moment. The unused slots of broadcast frequency are filled with dummy bursts.

Remember that the broadcast frequency is only required in the downlink direction and each BTS (of any operator) has one unique broadcast frequency (also known as the BCCH frequency, broadcast channel frequency). As discussed earlier, each PLMN consists of several cells and each cell has one BCCH frequency. One operator buys several BCCH frequencies and as a frequency license is expensive they try to reuse the same BCCH frequency again in a distant cell, based on the cell planning (Figure 7.14).

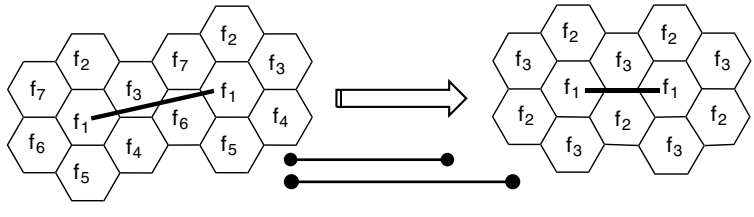


Figure 7.14 Cell layout, with frequency reuse factor of 7 and 3 (here "f_n" denotes the BCCH frequencies)

Typically, cells are hexagonal, but in practice, it depends on the available cell sites, radio propagation conditions, and subscriber density. The hexagon is an ideal choice for macro cellular coverage areas, because it closely approximates a circle and offers a wide range of tessellating reuse cluster sizes. If D is the reuse distance–distance to cell using the same frequency, r is cell radius, and N is frequency reuse factor, then the relationship between D and r can be computed as:

- $D/r = (3N)^{0.5}$
- $N = i^2 + ij + j^2$, where i,j are positive integers

The distance between (the centers of) any two cells in a hexagonal packing of the plane is:

$$r\,(l,\ m, \theta) = 2d\sqrt{(3m^2 + 1^2 + 2\sqrt{3}.1.m.\cos\theta)}$$

where $2d$ is the distance between the centers of two adjacent cells and the triplet (l, m, θ) uniquely specifies the relative positions of the two cells with respect to one another. Co-channel interference is main problem, which limits the frequency reuse by degrading the system performance, which is discussed in Chapters 2 and 3.

7.7 GSM Frame Structure

Each TDMA frame on a particular carrier frequency is divided into eight time slots. A physical channel occupies (maximum of one, in the case of full-rate TCH; minimum of $^1/_2$, as it is shared by two channels, in the case of half-rate TCH, where TCH/H0 and TCH/H1 will appear alternatively in two consecutive TDMA frames under the same slot number) one time slot in each TDMA frame.

The hierarchical topology is as below (Figure 7.15):

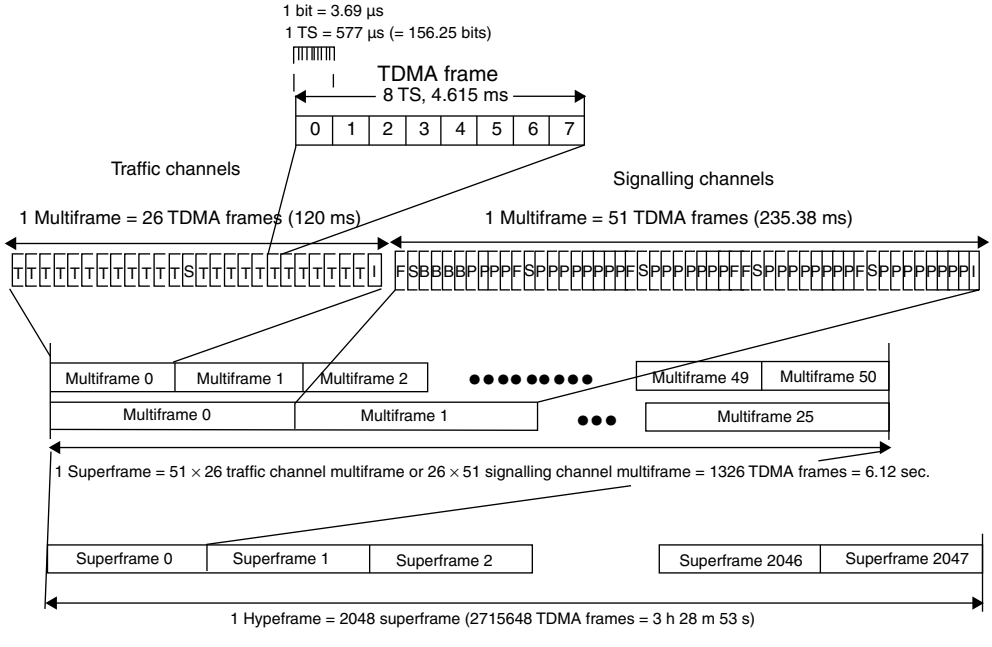


Figure 7.15 Hierarchy of GSM TDMA frame

1. The lowest level is a **bit** (symbol), which has a duration of 3.69 μs.
2. 156.25 bits form a **time slot**.
3. 8 time slots together form a **TDMA frame**.
4. 26 traffic or 51 signaling channel's TDMA frames form a **multi-frame**.
5. 51 traffic or 26 signaling channel's multi-frames forms a **super-frame**.
6. 2048 numbers of super-frames forms a **hyper-frame**.

So a hyper-frame = 2048 super-frames = 2048 × 51 × 26 = 2 715 648 number of TDMA frames (of either signaling or traffic channel's).

Timing Value Computation GSM full-rate (FR) speech codec operates at 13 kbps, for example, it delivers 260 bits at every 20 ms, which need to be transmitted. This is passed to the channel coder for insertion of protection bits and after that it increases to 456 bits, which is subdivided into eight blocks of 57 bits and places two such blocks inside a burst. So, at burst level 456 bits need to be transmitted per 20 ms. This means, over 120 ms the total bits that need to be transmitted will be 456 × 6 = 2736 bits. Each burst contains 114 information bits, so total burst number required to transmit 2736 bits over 120 ms 2736/114 = 24. Apart from the 24 TCH frames, 1 idle and a SACCH frame are added, so the total TDMA frame number becomes 26. One multi-frame (which contains 26 TDMA traffic frames) needs to be transmitted in 120 ms. Hence the duration of each TDMA frame = 120/26 = 4.615 ms. Again, each TDMA frame consists of eight time slots, so each slot duration = (120/26) × (1/8) = 15/26 ms = 576.9 μs. The duration of a time slot is equal to the duration of a burst period (BP), which is 576.9 μs. In one time slot duration there are 156.25 bits accommodated as discussed in Section 7.4.1 on Burst Structure. Therefore each bit duration will be 576.9/156.25 = 3.69 μs. Again one quarter bit duration will be 3.69/4 = 12/13 μs = 0.923 μs.

Thus in a burst there will be 625 numbers of quarter bit and 156.25 numbers of bit.

The multi-frame for a traffic channel contains 26 TDMA frames and is of 120 ms duration, whereas for the signaling channel the multi-frame contains 51 frames and has a duration of 235.38 ms.

Data rate at the burst level or when it is passed to modulator level will be: 156.25/576.9 μs = 1/0.00000369 = 270.833 × 10^3/s. So ideally in air, bit rate = 270.833 × 10^3/s, for example, 270.833 kHz. On the receiver side, during sampling, at least one sampling should be performed over a bit period, for example, the minimum sample rate will be 1/0.00000369 = 270.833 kHz.

7.8 Combination of Logical Channels

The physical channels (TS_n, $Freq_n$) are precious resources, so sometimes several logical channels are again time multiplexed, for example, different data for the logical channel will appear in the same physical channel after a defined number of TDMA frames (for example, different time intervals). Before mapping into a physical channel, various logical channels may be combined in one of the six different ways, as described below.

1. Every TCH has an associated SACCH channel, which carries the measurement information during the call when TCH is established. The simplest mapping is the full rate traffic channel (TCH/F) and its associated SACCH. When combined these channels fit exactly into a single physical channel. The TCH is not distinguished with respect to data or speech.
2. A single physical channel will also support two half-rate traffic channels (TCH/H) and their associated SACCHs.
3. A single physical channel will also support eight SDCCHs and their associated SACCHs.

   The remaining three logical channel combinations are complicated and are explained in more detail in Section 7.8.3.

4. The basic broadcast and common control channels combination consists of FCCH, SCH, and BCCH on the downlink along with CCCH (a full rate PCH and full rate AGCH). These are essential for any cell-broadcast and paging or granting a channel information. Similarly, in the uplink direction it is entirely dedicated for full rate RACH. This type of channel configuration is used in medium capacity or large capacity cells, where the access capacity of full rate PCH, AGCH, and RACH channel is justified, for example, less wastage of channel slot resources. This control/signaling channel combination may occur only in time slot zero of the broadcast frequency.
5. In smaller capacity cells (for example, a cell with a lesser number of licensed RF carriers) the physical channels (TS_n, $Freq_n$) are limited and very precious. So, the use of full rate PCH, AGCH, and RACH may not be justified. For this reason, a second combination of access channels is introduced. The downlink continues to support an FCCH, SCH, and BCCH, but the rate of downlink PCH and AGCH is reduced to around one-third of their full rate, as that many paging and granting channels might not be required in a smaller capacity cell. Hence, the rate of these two channels is reduced, so as a result some free slots will be created. These free slots are used to support four SDCCHs and their associated SACCHs. This control channel combination may only occur on time slot zero of the BCCH carrier.

   Similarly, in the uplink direction, the RACH rate has been reduced and the SDCCHs and its associated SACCH are placed in those created vacant slots. This will effectively halve the number of time slots allocated to the RACH.
6. The final combination is defined for use in high capacity cells, where the access capacity of single PCH, AGCH, and RACH is insufficient and requires an increase in the number of PCH, AGCH, and RACH. As many subscribers are supported here, so only one physical channel (f_{BCCH}, TS0) would not be sufficient, and an increased number of physical channels are needed to carry the logical channel information for the PCH, AGCH, and RACH. This combination consists of a BCCH and a full-rate PCH and AGCH on the downlink and a full-rate RACH on the uplink. This channel combination may only occur on (slot two), or (slots two and four) or (slots two, four, and six) of the BCCH carrier frequency. This is given in Table 7.7. The CBCH has been omitted here. If the channel is required it may replace a single SDCCH.

Table 7.7 Possible combinations of channels

Possible time slots of BCCH carrier frequency	Downlink channels	Uplink channels
1–7	1 TCH/F + SACH	1 TCH/F + SACCH
1–7	2 TCH/H + SACCH	2 TCH/H + SACCH
1–7	8 SDCCH + SACCH	8 SDCCH + SACCH
0 (broadcast freq) (non-combined configuration)	1 SCH + 1 FCCH + 1 BCCH + 1 AGCH + 1 PCH	1 RACH
0 (broadcast freq) (combined channel configuration)	1 SCH + 1 FCCH + 1 BCCH + 1 reduced rate AGCH + 1 reduced rate PCH + 4 SDCCH + SACCH	1 Reduced RACH + 4 SDCCH + SACCH
2,4,6 (broadcast freq)	1 BCCH + 1 AGCH + 1 PCH	1 RACH

7.8.1 Mapping of Traffic Channels and SACCH

As discussed earlier, the multi-frame for traffic channels consists of 26 TDMA frames. The frame structure for a full-rate traffic channel (TCH/F) occupying the time slot number one in each consecutive TDMA frame of that carrier frequency is shown in Figure 7.16. The first 12 time slots, located at time slot

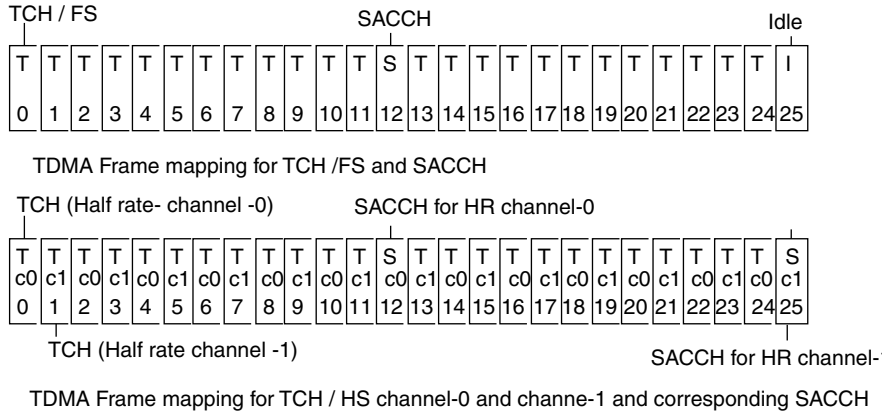


Figure 7.16 TCH channel mapping

"1" position of each TDMA frame from number 0 to 11 are used by the TCH/F itself. Similarly, 12 time slots, located at time slot "1" position of each TDMA frame from number 13 to 25 are used by TCH/F. The time slot "1" of 12th and 25th TDMA frame is not used by the TCH/F, rather these are used by "idle" and SACCH. For the odd numbered time slots (here time slot "1" in the TDMA frame), the 12th TDMA frame's respective (here slot one) slot position will contain "idle" and 25th TDMA frame's respective (here slot one) slot position will contain SACCH. In the case of even numbered time slots (and time slot zero), the position of the idle and SACCH time slots are just inter-changed, for example, 12th TDMA frame's slot will be SACCH and 25th one will be idle. To understand the reason for this, we need to examine the way the information is carried on SACCH. The SACCH messages are interleaved over four bursts and in a traffic multi-frame only one SACCH burst occurs. So four traffic multi-frames (or $120 \times 4 = 480$ ms) time period need to be waiting for SACCH decoding. That indicates BTS must receive four SACCH bursts before the information can be successfully de-interleaved and decoded, for example, SACCH information can only be decoded once on every four frames. Now, if the SACCH bursts occur at the same point (slot) in the multi-frame for each physical channel, then the SACCH message from every MS would arrive within the same TDMA frame and this would put a large processing load on the BSC. This is why in order to spread the SACCH processing load more evenly in time, the position of the SACCH bursts are changed from time slot to time slot and the interleaving period is also arranged such that the BTS will have to decode a maximum of one SACCH message per TDMA frame. This also gives the BSC a period of around 12 TDMA frames to process the information in the SACCH message before another message arrives from another MS using the same carrier but in a different time slot.

The organization of half-rate traffic channel (TCH/H) is a little more complicated than TCH/F. In Figure 7.16, two half-rate channels TCH/H0 and TCH/H1 are shown. TCH/H0, TCH/H1 and their associated SACCH0 and SACCH1 use, on average, every slot within the traffic multi-frame structure.

```
TCH/H0 | TCH/H1 | TCH/H0 | TCH/H1 | ...| SACCH0 | TCH/H0 |
TCH/H1 | ....|SACCH1 | ...
```

7.8.2 Mapping of SDCCH

In Figure 7.17, the mapping of eight SDCCHs into a single physical channel is shown. The burst mapping is based on two multi-frame cycles ($2 \times 51 = 102$ TDMA frames), for example, it repeats every two

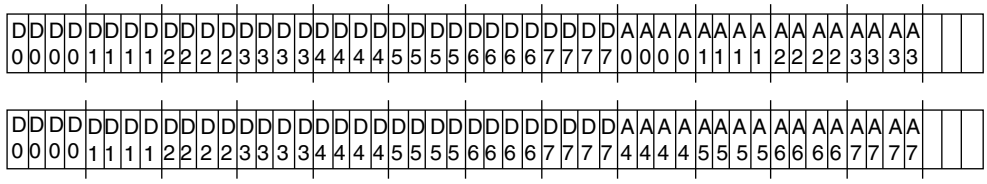


Figure 7.17 Mapping of SDCCH

multi-frames. The diagram shows two multi-frames with frames numbered 0–50 and 51–101. The time slot that are used by a particular SDCCH are labeled as SDCCH n ($n = 0$–7) and the time slot used by associated SACCH are labeled as SACCH n ($n = 0$–7). This control channel arrangement is known as SDCCH/8. Figure 7.17 shows that each SDCCH/8 occupies a total of eight time slots in two multi-frames. So a total of $8 \times 8 = 64$ SDCCH slots will be present in two multi-frame (102 TDMA frames). The associated SACCH, which is referred to as an SACCH/C8 in the specifications, occupies a further four time slots in the same period. Thus a total of $4 \times 8 = 32$ slots of SACCH will be present in two multi-frames. Therefore, a total of $64 + 32 = 96$ time slots are occupied and the remaining six are left idle. Figure 7.17 shows the situation for the down link slots, the uplink configuration may be derived by shifting each slot 15 places to the right, for example, SDCCH/0 will occupy slots 15–18 and slots 66–69 and its associated SACCH will occupy slot 47–50 on the uplink. Except for time slot zero of the BCCH carrier, this control channel arrangement may be used on any time slot and any carrier.

7.8.3 Mapping of Broadcast and Common Channels

In the downlink direction, broadcast channels (FCCH, SCH, and BCCH) and control channels (PCH AGCH) are always transmitted using the cell's broadcast frequency. PCH and AGCH are collectively referred to as CCCH. In the case of non-combined configuration on 51 signaling multi-frame structures (on the $\text{Freq}_{\text{BCCH}}$) – the dedicated signaling channels SDCCH and BCCH/CCCH are not combined together at TS0. This means that SDCCH does not appear in the TS0 of the broadcast frequency. Generally, FCCH, SCH, BCCH, and CCCH are present on TS0 and SDCCH/8 on TS1, whereas, in the case of the combined configuration, FCCH, SCH, BCCH, CCCH, and SDCCH appear on TS0.

Figure 7.18, shows for the case of a non-combined configuration, how the FCCH, SCH, BCCH, and CCCH are mapped onto the signaling channel multi-frame at broadcast frequency with TS0.

As discussed earlier, there are three possible combinations of broadcast and common channels.

1. Figure 7.18 shows the basic arrangement of channels for a non-combined configuration on time slot zero of the broadcast carrier. The downlink signaling multi-frame (51 slots) is sub-divided into five groups of ten time slots with a single idle slot at the end ($5 \times 10 + 1 = 51$). The first time slot of each group is then assigned to FCCH and contains a frequency correction burst. The second time slot is assigned to SCH and contains a synchronization burst. In the first group, the four time slots following the FCCH and SCH slots are assigned to the BCH. The remaining time slots, except for the idle at the end of the multi-frame, is assigned to CCCH (PCH, AGCH). The PCH/AGCH time slots may be assigned to either channel on a block by block basis, where a block consists of four time slots as defined by the interleaving block size. In a non-combined mode, there is a capacity to carry nine CCCH slots on a 51 mufti-frame indicated by CCCH/0 to CCCH/9. The internal division of the CCCH capacity is controlled by a parameter referred to as BS_AG_BLK_RES. Its value ranges from 0 to 7. This parameter indicates the number of blocks on each CCCH that are reserved for AGCH. If the system information is more than that which can be sent on the BCCH block, then the first CCCH

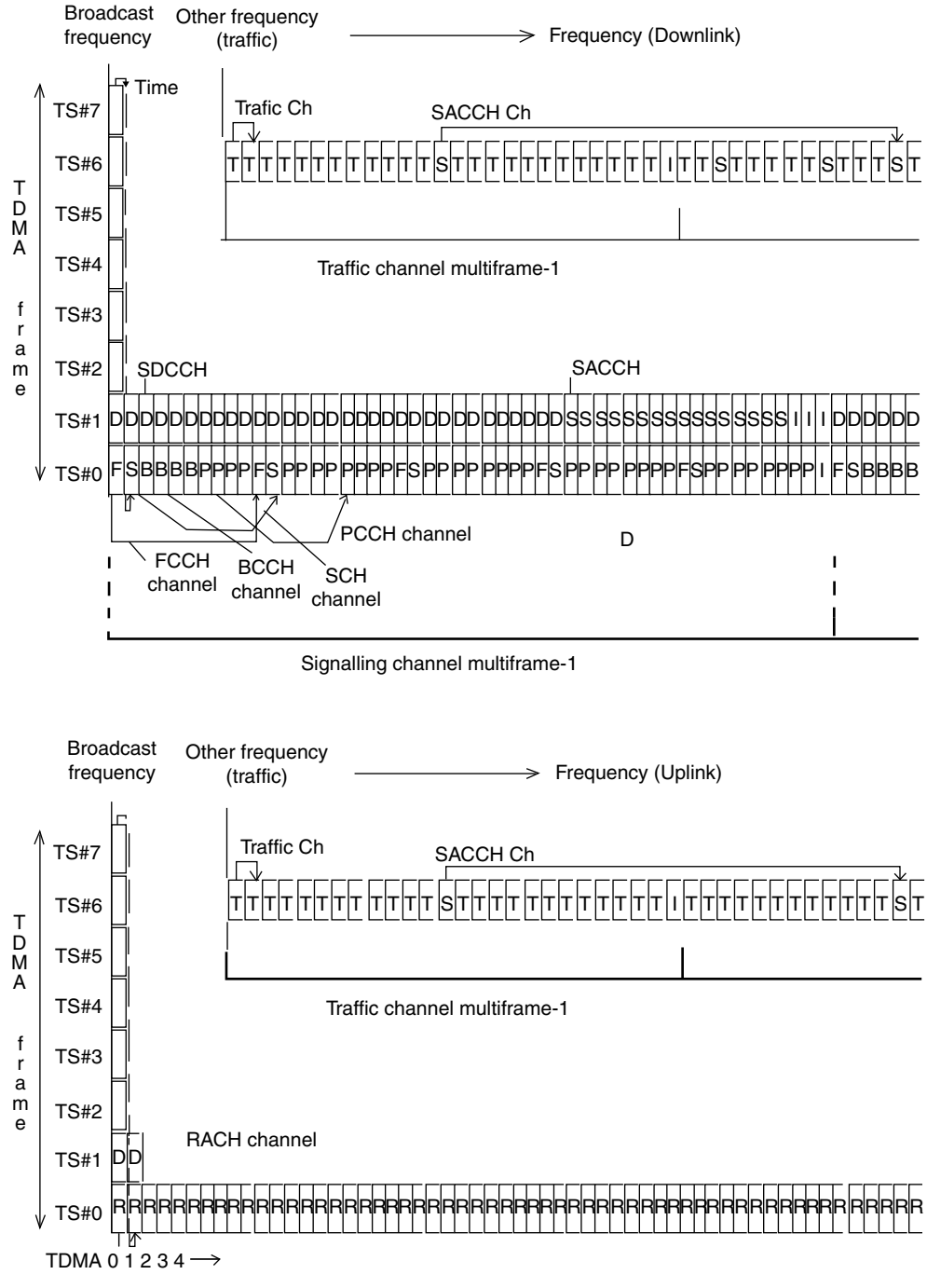


Figure 7.18 Broadcast channel and traffic channel time multiplexing (downlink and uplink)

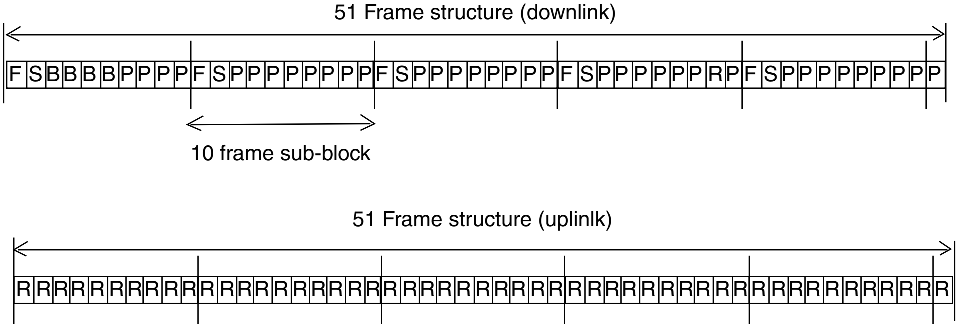


Figure 7.19 Mapping of broadcast, control, and RACH in scenario-1

block is taken for system information and this is referred to as extended BCCH. On the uplink all the time slots (on the broadcast frequency's corresponding uplink pair frequency) are assigned to RACH (Figure 7.19).

2. The control and broadcast channel arrangement for small capacity cells for combined configuration are shown in Figure 7.20, where the CCCH capacity is reduced to accommodate four SDCCHs on the TS0 of the broadcast frequency. The downlink multi-frame is again sub-divided into five groups of ten slots and each group begins with an FCCH and an SCH. In the first group after one FCH and SCH, the next four are assigned to BCCH and the next four are assigned to CCCH (PCH/AGCH). In the second group after one FCCH and one SCH, the remaining eight slots are assigned to PCH/AGCH. In the subsequent groups, except for the first (FCCH) and second slot position all other positions are used by the four SDCCHs and their associated SACCHs, the specification refers to them as SDCCH/4 and the associated control channel as SACCH/C4. There is one idle slot at the end of the multi-frame.

   On the uplink, 24 slots are assigned to the four SDCCHs and their SACCHs and the remaining 27 slots are used by the RACH. This control channel arrangement may only be used on slot zero of the BCCH carrier.

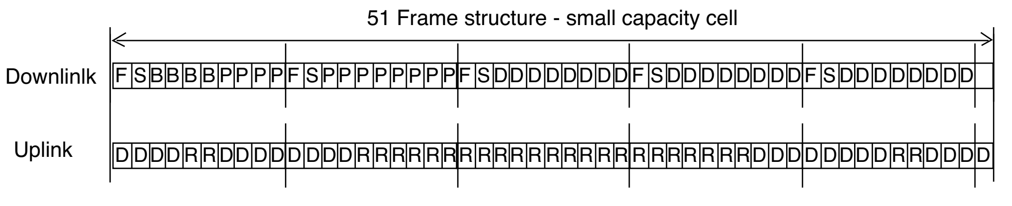


Figure 7.20 Mapping of broadcast, control, and RACH in scenario-2

3. For high capacity cells, many CCCH and RACH are required. So, apart from the TS0 of broadcast frequency TS2, TS4, and TS6 can also be used for those channel transmission. This channel configuration is shown in Figure 7.21, which is effectively the same as the basic control channel arrangement given in Figure 7.19, but in this case the FCCH and SCH slots are replace by idle slots, as these are already transmitted at slot zero of the BCCH frequency and a BTS should not transmit the FCCH and SCH multiple times. This arrangement can only be used on slots 2, 2 and 4 or 2 and 4 and 6 of the BCCH frequency. We also note that this is an extension set and can only be used when there are no SDCCHs on time slot zero of BCCH carrier.

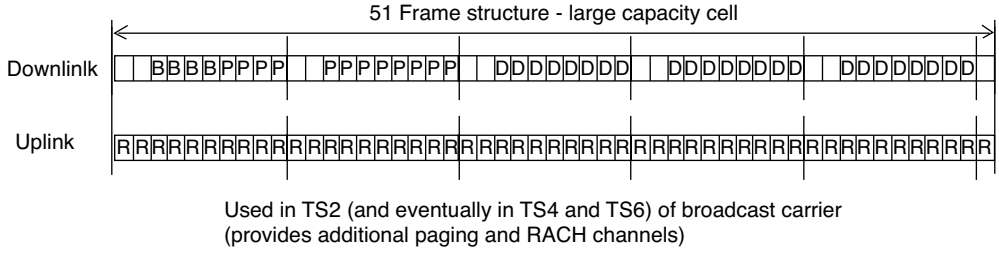


Figure 7.21 Mapping of broadcast, control, and RACH in scenario-3

Points to remember:

- The specific time slots of a multi-frame (total 51) are classified into five sub-blocks each having ten slots ($10 \times 5 = 50$). The last slot (slot-50) is left idle.
- Each sub-block starts with a frequency correction channel (FCCH) immediately followed by a synchronization channel (SCH).
- One BCCH carrier frequency per cell (beacon) and FCCH and SCH must be only on time slot 0 of that frequency.
- Other time slots of the broadcast frequency may be used by TCH provided all empty slots are filled with dummy bursts and the downlink power control must be disabled.
- Frame number distinguishes logical channels in the same physical channel.
- Multi-frame determines how BCCH is constructed, that is, which specific information is transmitted on BCCH during a given multi-frame.
- Super-frame used as input parameter by a ciphering algorithm.

7.9 Physical Layer Processing for Logical Channel Transmission and Reception Procedures

7.9.1 Traffic Channel Transmission Procedures (from Speech to Radio Waves)

The user speaks in front of the phone's microphone and the analog speech signal is converted into a digital signal and source coded. Then error correction (FEC) coding is applied to the speech data in order to protect it from the channel disturbances. Next the coded data are interleaved to protect them from burst error and then ciphered for security purposes. These are then assembled into bursts along with a training sequence and tail and guard bits. It is now ready for transmission over the radio interface. Digital-modulation is performed and then up-converted to the GSM RF channel frequency band and transmitted via the air. The reverse process is performed at the receiver side. The block diagram of this process is shown in Figure 7.22.

7.9.1.1 Full-Rate Speech Coding

In the mobile handset, the speech signal is first sampled at a rate of 8 kHz and each sample is represented by a 13-bit format. So the total generated bits per second will be $8000 \times 13 = 104\,000$. The GSM coder belongs to the family of regular pulse excited linear predictive codec. It also employs long-term prediction in addition to the conventional short-term prediction and, accordingly, it is known as RPE-LTP speech coder. The 20 ms block (which will contain $104\,000 \times 20/1000 = 2080$ bits) is stored first and then encoded to $2080/8 = 260$ bits, using 8 : 1 compression. This produces a bit rate of $(260/20) \times 1000 =$

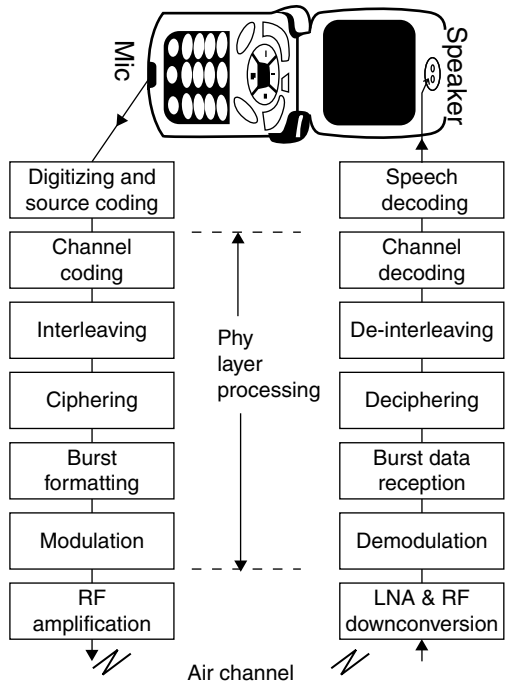


Figure 7.22 Traffic channel transmission and reception procedures

13 kbps digital stream (as 260 bits every 20 ms) then forward error correction is added by a convolutional encoder, as will be discussed next.

The situation is slightly different on the network side, as the speech, which is coming to MSC from other networks or PSTN/ISDN lines, uses an 8-bit A-law pulse code modulation format. This means that before the speech signal is passed to the speech coder on the network side, it must first undergo an 8-bit A-law PCM to 13-bit uniform PCM conversion. After this, the situation is the same as that discussed above.

7.9.1.2 Channel Coding

The above encoded speech data are then passed through the channel coder unit in an effect to reduce the system BER. Some extra bits are added to shield the speech data, when it passes through the noisy air channel. Channel coding helps to move from 10^{-1}–10^{-3} radio channel native BER down to an acceptable range (10^{-5}–10^{-6}) BER.

The FEC coding introduces a level of redundancy into the transmitted data, thereby increasing the transmitted data rate and channel bandwidth. The speech coder delivers 260 bits at every 20 ms, so the data rate is 13 kbps. These 260 bits input block (over a time of 20 ms) is divided into two classes:

- **Class I:** important bits (**182**)
  - **Class Ia** Most important 50
  - **Class Ib** Less important 132
- **Class II:** low importance bits (**78**)

First step: block coding is performed for error detection in class Ia bits.
Second step: convolutional coding for error correction.

The class Ia bits are so important that the speech frame must be discarded if any of these bits are received in error. Thus it is important for the receiver to be able to detect when an error occurs in class Ia bits, and this is achieved by using a weak error detecting block code. The code used is a shortened cyclic code (53, 50, 2) with a generator polynomial: $g(D) = D^3 + D + 1$ (see Table 7.8 for polynomials used for different logical channels). This coding process generates three parity bits, $p0$, $p1$, and $p2$, which are then appended to the end of the class Ia bits. Then the class Ib bits are appended at the end of these. Four all zero tail bits are then added to the end of class Ib bits and all the bits are re-ordered as shown in Figure 7.23. The resulting 189 bit block (50 + 3 + 132 + 4) is then convolutionally encoded using a half-rate code with the following generator polynomials: $g0 = 1 + D^3 + D^4$ and $g1 = 1 + D + D^3 + D^4$.

Table 7.8 Various generator polynomials used for different channels

Channel type	Generator polynomials
TCH/FS, TCH/EFS, TCH/F9.6, TCH/H4.8, SDCCH, BCCH, PCH, SACCH, FACCH, AGCH, RACH, SCH	$G0 = 1 + D^3 + D^4$
TCH/FS, TCH/EFS, TCH/F9.6, TCH/H4.8, TCH/F4.8, TCH/F4.2, TCH/H2.4, SACCH, FACCH, SDCCH, BCCH, PCH, AGCH, RACH, SCH	$G1 = 1 + D + D^3 + D^4$
TCH/F4.8, TCH/F2.4, TCH/H2.4	$G2 = 1 + D^2 + D^4$
TCH/F4.8, TCH/F2.4, TCH/H2.4	$G3 = 1 + D + D^2 + D^3 + D^4$
TCH/HS	$G4 = 1 + D^2 + D^3 + D^5 + D^6$
TCH/HS	$G5 = 1 + D + D^4 + D^6$
TCH/HS	$G6 = 1 + D + D^2 + D^3 + D^5 + D^6$

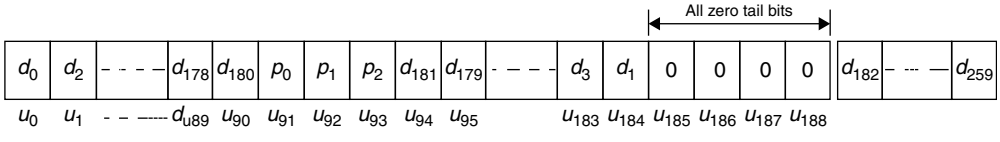


Figure 7.23 Coded block before convolution coding

It produces a code block of $189 \times 2 = 378$ bits. Then 78 bits the class II bit are just added to this to produce 456 bits block. The output from the channel coder unit is 456 bits block over a 20 ms frame. This gives a rate of 22.8 kbps.

7.9.1.3 Interleaving

Owing to the mobility of the MS, the errors in the transmitted bits tend to occur in bursts, when MS moves into and out of deep fades. The convoutional error correcting code described earlier is most effective when the errors occur randomly and are distributed throughout the bit stream. Hence the convolution coders will not work properly if the errors are present in a cluster. Therefore to randomize the probability of bit error, the coded data are interleaved again before being transmitted over the air.

The output from the channel coder unit is a 20 ms block, which contains 456 bits. For full-rate TCH, carrying speech information, the 456 coded speech block is portioned into 57 bit sub-blocks – B0, B1, . . ., B7 by assigning coded bit c_k to sub-block B_i, based on the relationship $i = k \bmod 8$, for example, every 8th bit is assigned to the same sub-block. Each block then forms one half of eight consecutive transmission bursts on the radio interface. The remaining half of each burst is occupied by sub-blocks from either the previous speech frame or the next frame, where B_i^n refers to sub-block i of the speech frame n.

In addition to block interleaving, the data bits are interleaved within the burst. One sub-block will occupy either the odd or even bit positions within the burst. Where a sub-block from a speech frame shares its burst with a sub-block from the previous speech frame, this will use even numbered bit positions. On the other hand, when a sub-block shares its burst with a sub-block from the next speech frame, it will use odd numbered bit positions, for example, B0, . . . B3 will use the even numbered bit positions and B4,. . . B7 will use the odd numbered bit positions. The bits within the sub-blocks are also reordered to increase the maximum distance between the consecutive bits. The interleaving procedure is described in GSM spec 05.03.

7.9.1.4 Ciphering

To make the user data secure from fraud over the air the encryption is applied on the transmitted data. The encryption process involves performing the module-2 addition of a 114-bit wide encryption word and the 114 data bits in a burst. This is described in more detail in the GSM security procedure of Chapter 9.

7.9.1.5 Burst Forming

With these 114 information bits, the guard bits, tail bits, and training sequence bits are added to form a normal burst. Two flag bits (h_l and h_u) are added on both sides of the training sequence, these are known as stealing flags. If h_l is set to 1, this indicates that even bits of the traffic channel are stolen to carry signaling information (FACCH), similarly, an h_u flag is set for stealing odd bits. Each time-slot burst contains: two 57-bit blocks ($2 \times 57 = 114$ bits) of traffic information bits + 2 stealing bits + 26 bit training sequence bits + 3×2 tail bits + 8.25 guard bits = 156.25 bits.

7.9.1.6 RF Conversion

The modulation scheme used in GSM is Gaussian minimum shift keying (GMSK) with a normalized bandwidth product, BT of 0.3 and modulation symbol rate around 270.833 kbps, where 0.3 describes the 3 dB bandwidth of the Gaussian pulse shaping filter with relation to the bit rate ($BT = 0.3$). As described earlier, GMSK is a special type of MSK modulation. The digital square wave pulse contains sharp edges, which contains lots of high frequency harmonics, so it requires a higher bandwidth to transmit. This is why in order to reduce this and to increase the spectral efficiency, the rectangular pulses are shaped to Gaussian shaped after passing through a Gaussian filter. This is then multiplied by the cos (for the I path) or sin (for the Q path) waveform of the frequency as required by the channel ARFCN and summed to form a composite waveform. In the case of a polar transmitter, the amplitude and phase terms are computed and accordingly the resultant waveform is generated (as discussed in Chapter 4).

Based on the input binary zeros and ones, the modulated frequencies are represented in GSM by shifting the RF carriers by 67.708 kHz upwards or downwards. The modulating bit rate is $1/T = 1\,625/6$ kbps (that is, approximately 270.833 kbps). The channel data rate of GSM is 270.83333 kbps, which is exactly four times the RF frequency shift. Please refer to GSM spec 05.02.

Start and Stop of the Burst

Before the first bit of the bursts enters the modulator, the modulator has an internal state, as if a modulating bit stream consisting of consecutive ones ($d_i = 1$) had entered the differential encoder. Also, after the last bit of the time slot, the modulator has an internal state as if a modulating bit stream consisting of consecutive ones ($d_i = 1$) had continued to enter the differential encoder. These bits are called dummy bits and define the start and the stop of the active and the useful part of the burst. Nothing is specified about the actual phase of the modulator output signal outside the useful part of the burst.

Differential Encoding

Burst data bits (d_i) are differentially encoded by performing modulo-2 addition of the current and previous bits. The output of the differential encoder is:

$$\hat{d}_i = d_i \oplus d_{i-1} \quad (d_i \in \{0, 1\})$$

where d_i may take 0 or 1, and $\oplus$ denotes modulo-2 addition. Then the even and odd bits are separated into two data streams for the I and Q paths (for a linear transmitter). Next, the resultant data bit on each path is mapped onto $+1$ (for 0) and -1 (for 1) values, to form the modulating data value α_i input to the GMSK modulator, as follows:

$$a_i = 1 - 2\hat{d}_i \quad (a_i \in \{+1, -1\})$$

So the signal is now converted from non-return to zero signal ($+1$, -1) format.

Filtering

The modulating data values αi as represented by Dirac pulses excite a linear filter with impulse response defined by:

$$g(t) = h(t) rect\left(\frac{t}{T}\right)$$

where the function $rect(x)$ is defined by:

$$rect\left(\frac{t}{T}\right) = \frac{1}{T}, \quad \text{for } |t| < \frac{T}{2}$$

and the * means convlotion. $h(t)$ is defined by:

$$h(t) = \frac{\exp\left(\frac{-t^2}{2\delta^2 T^2}\right)}{\sqrt{(2\pi)}.\delta T}$$

where $\delta = \frac{\sqrt{\ln(2)}}{2\pi BT}$ and $BT = 0.3$, B is the 3 dB bandwidth of the filter with impulse response $h(t)$, and T is the duration of one input data bit. This theoretical filter is associated with tolerances defined in GSM 05.05.

Output Phase

The phase of the modulated signal is:

$$\phi(t') = \sum_i \alpha_i \pi h \int_{-\infty}^{t'-iT} g(u)\mathrm{d}u$$

where the modulating index h is 1/2 (maximum phase change in radians is $\pi/2$ per data interval).

The time reference $t' = 0$ is the start of the active part of the burst as shown in Figure 7.24. This is also the start of the bit period of bit number 0 (the first tail bit) as defined in GSM 05.02.

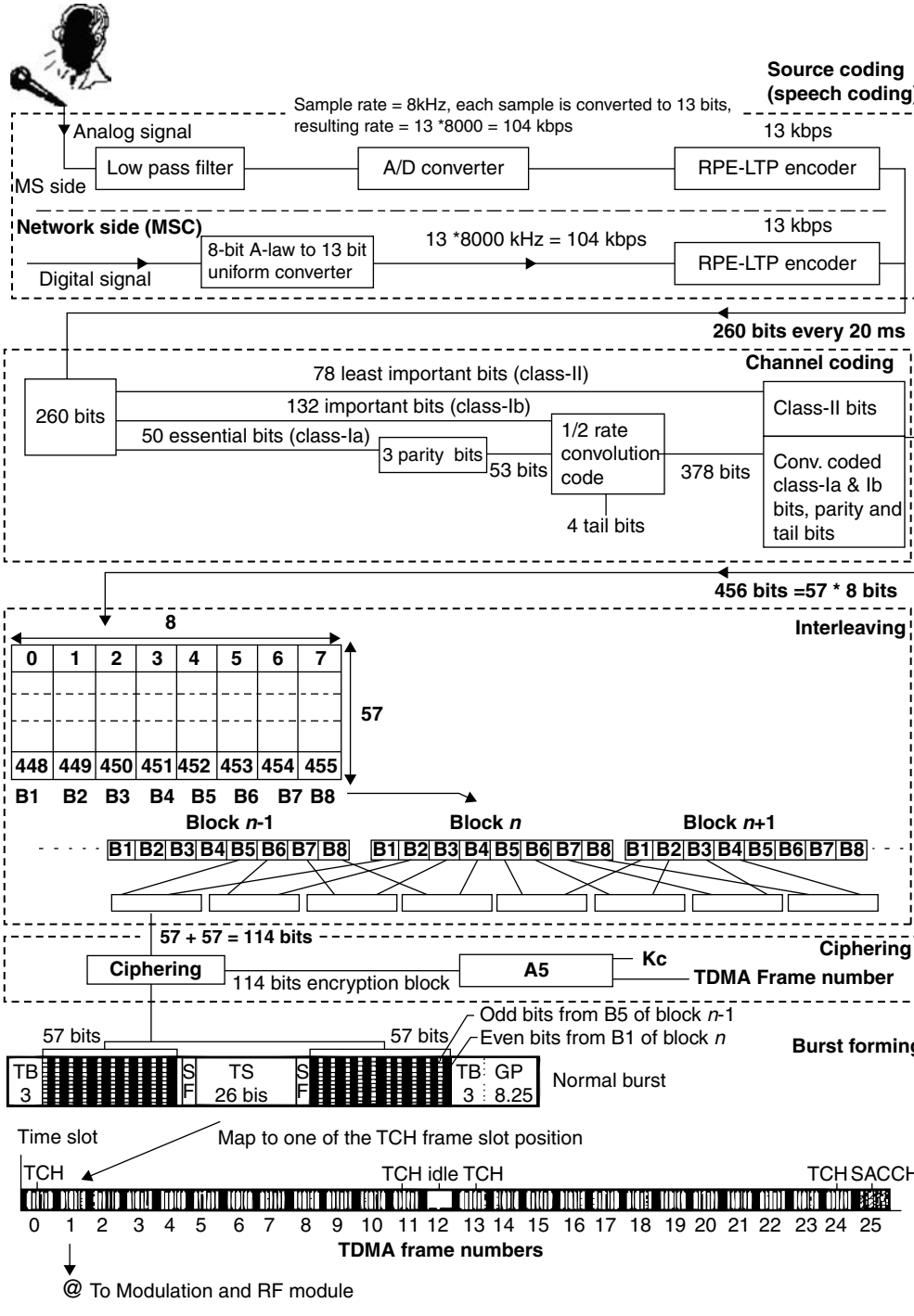


Figure 7.24 Different stages of TCH processing from speech to radio wave

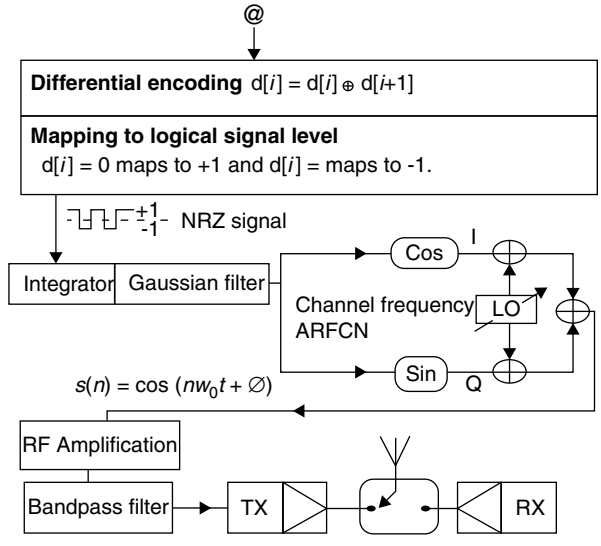


Figure 7.24 (*Continued*)

Modulation

The modulated RF carrier, except for the start and stop of the TDMA burst may therefore be expressed as:

$$x(t') = \sqrt{\frac{2E_c}{T}}.\cos(2\pi f_0 t' + \varphi(t') + \varphi_0)$$

where E_c is the energy per modulating bit, f_0 is the center frequency of the RF radio channel (generated according to the required channel frequency) and φ_0 is a random phase that is constant during one burst.

7.9.2 User Data Transmission Using TCH

Apart from speech, a GSM system also supports data transmission at rates of 9.6, 4.8, and $\leq$2.4 kbps on full-rate channels (TCH/F9.6, TCH/F4.8, and TCH/F2.4) and rates of 4.8 kbps and $\leq$2.4 kbps on the half-rate channels (TCH/H4.8 and TCH/H2.4). Various processes involved in the user data transmission are described below. In many instances the blocks are similar to those for speech data transmission, which was discussed in the previous section.

1. **Channel Coding** – The data to channel coder unit comes from the upper layers (layer-2). GSM supports both the transparent and non-transparent modes of operation. In the transparent mode the error correction is performed using only FEC techniques, whereas in the non-transparent mode, if error occurs the information will be repeated. The non-transparent connection is only available on the TCH/F9.6 and TCH/H4.8 channels. Based on the type of the channel used, the channel coding procedure and the correspondingly used polynomials will also vary. See the coding methods for different logical channels in Table 7.8.

a. **TCH/F9.6 channel** – The data at rate 9.6 kbps is delivered to the protocol layer; then in this layer a certain amount of auxiliary information is added to produce an intermediate data rate of 12 kbps. These data are delivered to the channel coding unit in the form of a 60-bit block on every 5 ms, and the coder operates on a group of four blocks, that is, $60 \times 4 = 240$ bits. Four "all zero" tails bits are then added at the end of the block and the data are convolutionally encoded using a one-half rate code through the following generator polynomials:

$$G0 = 1 + D^3 + D4 \quad \text{and} \quad G1 = 1 + D + D^3 + D^4$$

This produces a coded data block of length 488 bits, which is then punctured by removing 32 coded bits using the following rule: removed bits = C(11 + 15*j*), where $j = 0.31$. This means that bits C(11), C(26), C(41), and so on are removed. Puncturing is used to precisely tailor the rate of a convolutional code to the requirements of a transmission link. At the receiver, the convolutional decoder will effectively treat the deleted bits as errors and they will be corrected in the conventional way within the Viterbi decoder.

b. **TCH/F4.8 channel** – The data are delivered to the coding unit at a rate of one 60-bit block on every 10 ms. Each block is extended to 76 bits by the addition of 16 all zero bits, which are inserted in blocks of four, once every 15 bits. Two of these blocks are then concatenated to form a single block of 152 bits, which is 1/3 rate convolutionally encoded using the following generator polynomials:

$$G1 = 1 + D + D^3 + D^4,\ \text{G2} = 1 + D^2 + D^4, \quad \text{and} \quad G3 = 1 + D + D^2 + D^3 + D^4$$

This results in a coded block of 456 bits.

c. **TCH/F2.4** – The data are delivered to the coding unit at an intermediate bit rate of 3.6 kbps in the form of 36-bit blocks at every 10 ms. The coding unit operates on blocks of 72 bits formed by the concatenation of two 36-bit blocks. Initially four all-zero tail bits are added at the end of the block to produce a block of 76 bits. This block is then one-sixth rate convolutionally encoded using the following generator polynomials:

$$G1 = 1 + D + D^3 + D^4,\ G2 = 1 + D^2 + D^4, \quad \text{and} \quad G3 = 1 + D + D^2 + D^3 + D^4$$

This results in a coded data block of 456 bits.

2. **Interleaving** – Two types of interleaving techniques used here: block diagonal and inter-burst interleaving. The TCH/H2.4 channel uses the same interleaving scheme as used for TCH/FS. The remaining channels use a different interleaving scheme: the interleaving scheme for the TCH/F9.6 channel is discussed below.
   a. In this case, a 456-bit block is sub-divided into four 114-bit sub-blocks, each of which is evenly distributed over 19 bursts with six bits in each. The sub-blocks are block diagonally interleaved with a shift of one burst between each sub-block.
   b. The 456-bit block is sub-divided into: two blocks of six bits, which are placed in bursts 0 and 21; two blocks of 12 bits, which are placed in bursts 1 and 20; two blocks of 18 bits, which are placed in bursts 2 and 19; and 16 blocks of 24 bits, which are placed in bursts 3–18. Therefore, each 456-bit block is interleaved over 22 bursts. The data blocks are also diagonally interleaved with a new data block beginning every fourth burst. We also note that the data bits are reordered within each burst.
3. **Ciphering** – The interleaved data bits are then encrypted in the same manner as the full-rate speech traffic channel, which was described in the previous section.

After this, similar steps to those in speech transmission are followed.

7.9.3 Signaling Channel Transmission Procedures

7.9.3.1 BCH, PCH, AGCH, CBCH, SACCH, SDCCH Channel

1. **Information bits:** The signaling information comes from the upper layers. For these channels, the signaling information from layer-2 contains a maximum of 184 bits (Figure 7.25). It does not make a difference whether the type of signaling information to be transmitted is mapped onto a BCCH, PCH, SDCCH or SACCH. The format always stays the same.

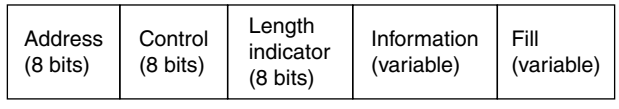


Figure 7.25 Message format

2. **Channel coding:** The data are delivered to the channel coder unit in fixed blocks of 184 bits of information. For SACCH, SDCCH, BCCH, AGCH, and PCH channel, the block code is used instead of convolution coding and the main block code in use is a fire code.

   The result of the fire coding process is the generation of 40 parity bits that are appended to the end of the data block to form a 224-bit block. This data block is then covolutionally encoded using a one-half rate convolution code with the below specified generator polynomials. The output is 456 coded bits (Figure 7.26).

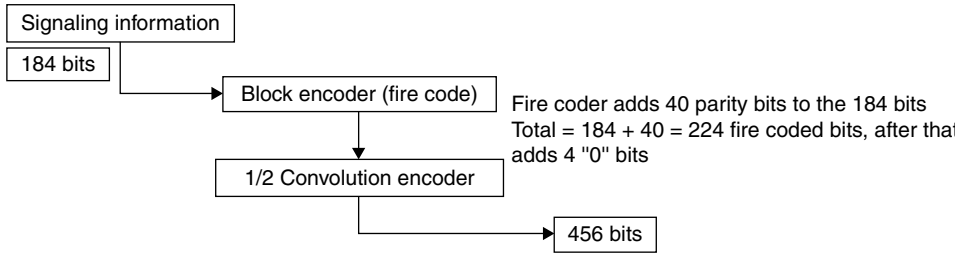


Figure 7.26 Information processing steps for BCH, PCH, SACCH, SDCCH channels

In the case of an SACCH channel, the initial encoding is performed on the delivered fixed blocks of 184 information bits using a shortened binary cyclic code defined by the following generator polynomial:

$$G(D) = (D^{23} + 1)\,(D^{17} + D^3 + 1)$$

This type of code is commonly referred to as a fire code and is used to detect "bursty" residual errors that are not corrected by the convolutional decoder. The result of this coding process is the generation of 40 parity bits that are appended to the end of the data block to form a 224-bit block. This block is extended to 228 bits with the addition of four all zero tail bits at the end of the block. This data block is then convolutionally encoded using a one-half rate convolutional code with the following generator polynomials:

$$G0 = 1 + D^3 + D^4,\ G1 = 1 + D + D^3 + D^4$$

This provides a block of 456 coded bits. The bits are reordered and divided into eight, 57-bit sub-blocks in the same way as the 456-bit speech block on the TCH/FS, that is, some blocks occupy the even-numbered bits and some blocks occupy the odd-numbered bits.

3. **Interleaving:** As discussed in the previous section, after channel coding, the block rectangular interleaving is applied. In case of the SACCH block, interleaving occurs over four full bursts with each burst containing bits from the same block in both the odd and even bit positions. This technique is termed block interleaving and a new data block begins once every fourth burst and is interleaved over four bursts.

   For SACCH, bursts occur once every 26 bursts or 120 ms with the full-rate traffic channel, which means that the delay caused by interleaving and the channel coding process will be $4 \times 120\,\text{ms} = 480\,\text{ms}$.

   The broadcast control channel (BCCH), the paging channel (PCH), the access grant channel (AGCH), the cell broadcast channel (CBCH) and the stand-alone dedicated control channel (SDCCH) all use the same coding and interleaving scheme as the SACCH, described above.
4. **Ciphering:** Generally, the SCH, BCCH, PCH, AGCH and CBCH are not encrypted, because they must be available to every MS within a cell. Only the SACCH and SDCCH data are encrypted in the same process as for the traffic channel.
5. **Burst formatting:** These data are then mapped into a normal burst, as described earlier.
6. **Modulation:** The GMSK modulation applied on the data.
7. **Frequency conversion:** Finally this modulated signal is up-converted to the proper GSM frequency and transmitted via the air.

The SACCH processing steps is described in the Figure 7.27.

7.9.3.2 RACH Channel

1. **Information bits and channel coding:** The RACH information contains eight information bits. Six parity bits are generated using a simple systematic cyclic code with the following feedback polynomial: $g(D) = D^6 + D^5 + D^3 + D^2 + D + 1$. The six parity bits are then added, bit-wise modulo-2 to the 6-bit base station identity code (BSIC) of the BTS for which the RACH message is intended. Only the BTS with the same BSIC as that used in the RACH burst generation will be able to successfully decode the information. This process results in a 14-bit block to which four all-zero tail bits are added to form an 18-bit block. This block is then one-half rate convolutionally encoded using the same generator polynomial as discussed earlier and produces a 36-bit coded block. This exactly fits into the data portion of a single RACH (access) burst (Figure 7.28).
2. **Interleaving:** There is no interleaving applied to RACH data.
3. **Ciphering:** There is no ciphering applied to the RACH information.
4. **Burst forming:** These data are then mapped into *access burst* as described earlier.
5. **Modulation:** The GMSK modulation is applied on the data.
6. **Transmission:** Finally this modulated signal is up-converted to the proper GSM frequency (broadcast beacon frequency) and transmitted via the air.

7.9.3.3 FACCH Channel

The FACCH channel is used for urgent signaling purposes, such as for call set up, handover and so on. The physical layer processing for the FACCH channel is as described below (see Figure 7.29):

1. **Information bits:** The FACCH information block from layer-2 contains 184 bits.
2. **Channel coding:** FACCH uses block code for channel coding similar to SACCH. The main block code in use is a fire code, which adds 40 bits of redundancy to a layer-2 data block on FACCH.

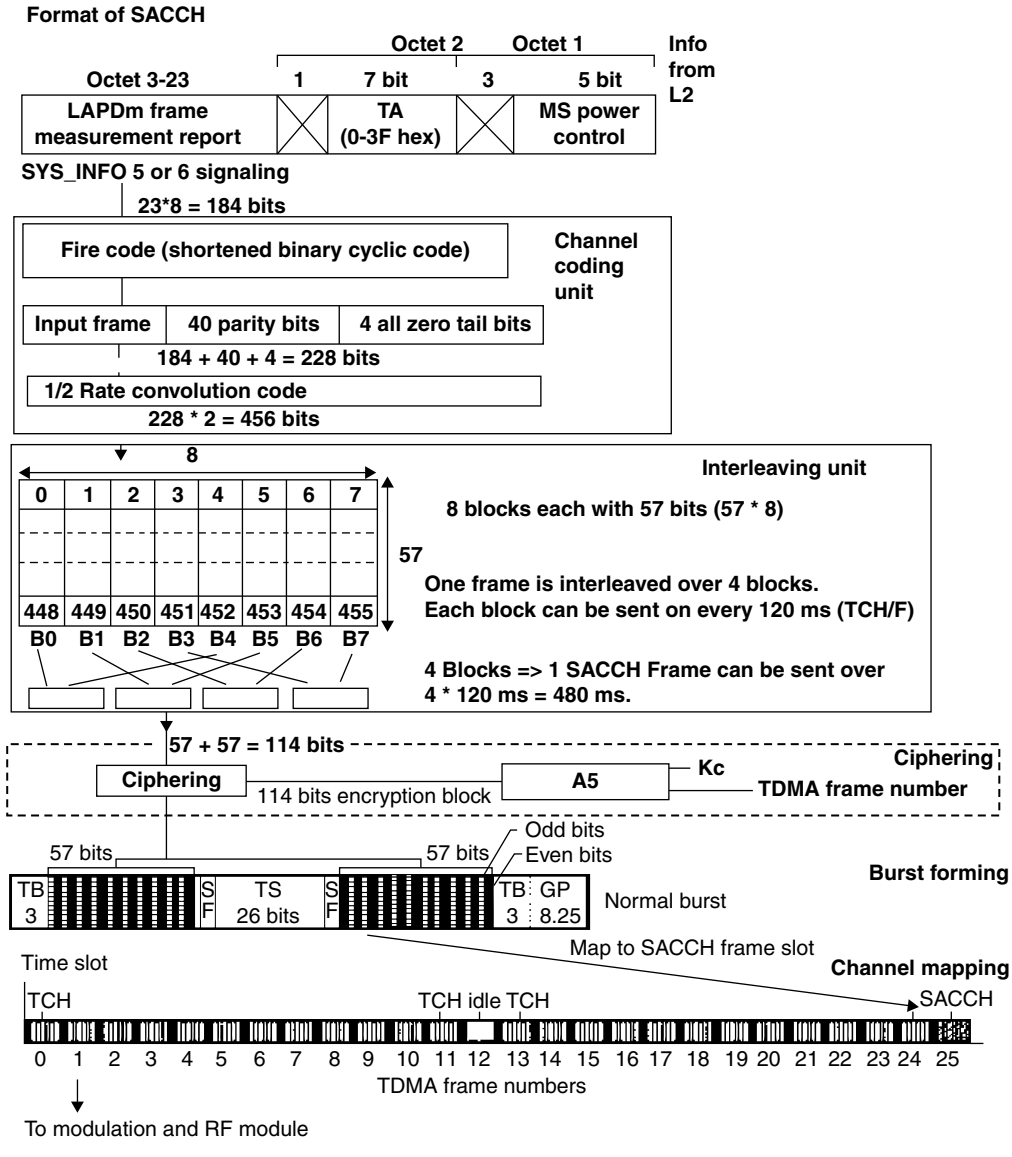


Figure 7.27 SACCH processing

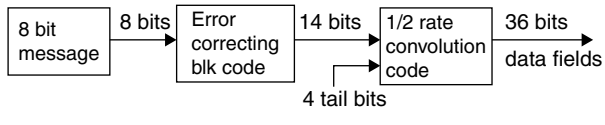


Figure 7.28 RACH processing

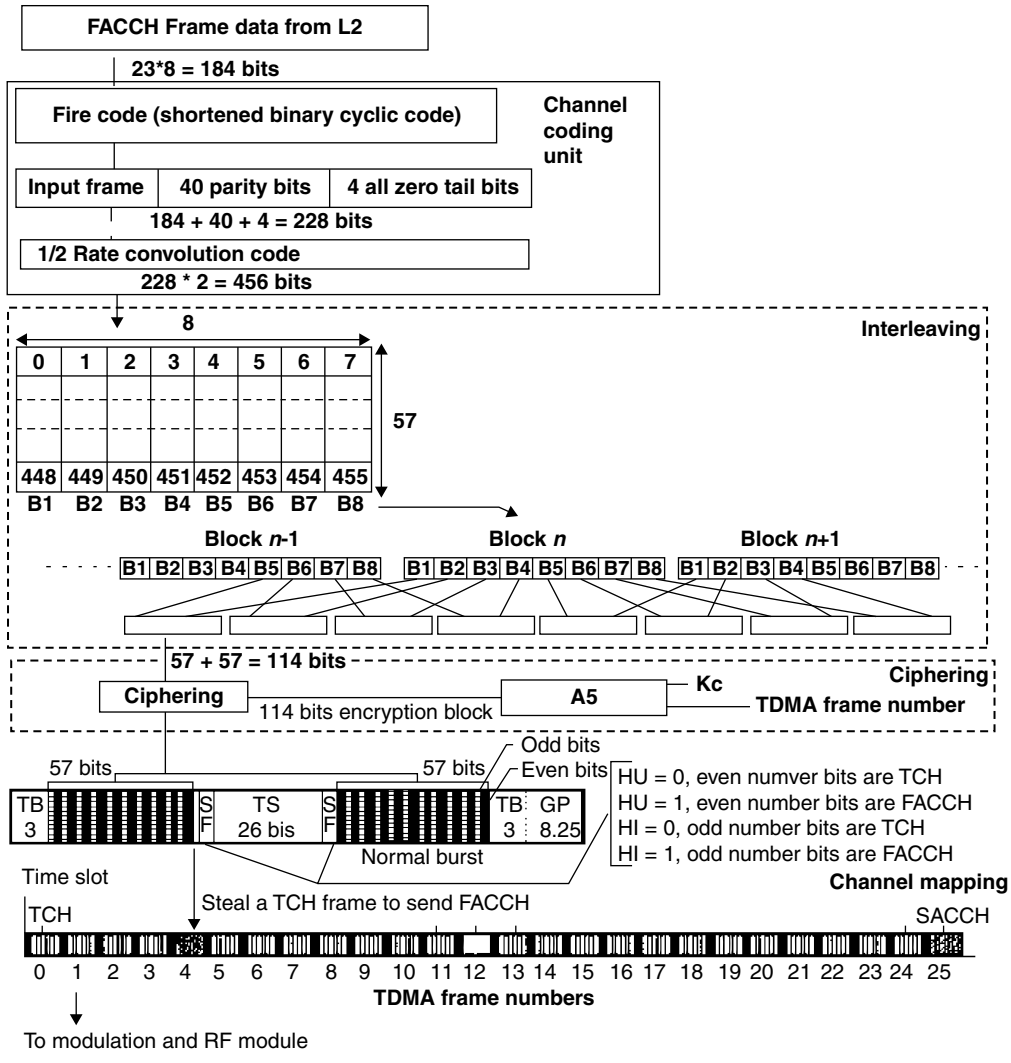


Figure 7.29 FACCH processing

3. **Interleaving:** The interleaving scheme for FACCH is different for half-rate and full-rate channels. For a full-rate channel, the interleaving scheme is identical with that used for full-rate 456-bit coded speech frames. In tthis instance, each block is divided into eight 57-bit sub-blocks and then the block is diagonally interleaved over eight consecutive bursts with the first four sub-blocks occupying only the even-numbered bit positions and the last four sub blocks occupying only the odd numbered bit positions. We know there are two stealing flags inside an NB, when the even numbered bits of a burst are stolen by the FACCH, then the h_u flag is set to a logical 1 and when odd numbered bits have been stolen, then h_l flag is set to 1. As the FACCH and full-rate speech interleaving is similar, thus a 456-bit FACCH block will completely replace a 456-bit coded speech block on a one-for-one basis, but in this instance, due to insertion of an FACCH block, there will a loss of a 20 ms speech block. When an

FACCH is inserted on a half-rate traffic channel, the 184-bit FACCH block is convolutionally encoded in the same manner as the SACCH block and produces a 456-bit coded data block. The block is then interleaved over six bursts using the following method. The 456-bit is block is divided into eight sub-blocks with first four sub-blocks occupying the even numbered bit positions and last four blocks occupying the odd numbered bit positions. The sub-block (2) and (3) are combined with sub-block (4) and (5) to fill two complete bursts and the remaining sub-blocks fill the half bursts. So, effectively the blocks are block diagonally interleaved over six bursts with a new data block beginning once every fourth burst. Therefore, we can say, an FACCH block steals the even numbered bits of the first two bursts of the TCH/H (for example, *hu* set to 1), all of the bits of the next two bursts (for example, both h_u and h_l set to 1) and the odd numbered bits of the next two bursts (for example, h_l set to 1), in effect there will be a loss of two consecutive speech frames.

7.9.3.4 Frequency Correction Channel (FCCH)

In a GSM/GPRS system, frequency synchronization is achieved through the detection of the presence of a frequency correction burst (FB) and then the frequency is estimated from the detected FB. In this situation, the frequency synchronization information is sent through a frequency correction channel, which is time multiplexed in the same BCCH carrier frequency along with the SCH, PCH, AGCH, and BCCH control channels based on the cell configuration. The FCCH is a downlink only channel, mainly used for frequency correction. It is also used for synchronization and acquisition by providing the boundaries between time slots and the position of the first time slot of a TDMA frame. The FCCH consists solely of the frequency correction burst (FB), which consists of an all-zero bit pattern. This occupies TS0 for every first GSM frame (frame 0) and is repeated every ten frames within a control channel multi-frame.

The frequency correction burst is used by the MS to detect a special carrier, which is transmitted by every BTS in a GSM network. The burst has a duration of 576.923 µs, which is equivalent to 156.25 symbol periods with 270.83 kHz symbol rate. As all bits in the frequency correction burst are set to zero (including the tail bits) so, after GMSK modulation, this results in a pure sine wave at a frequency around 67.7 kHz (1625/24 kHz) higher than the RF carrier center frequency. An FB is a single tone signal at 67.7 kHz relative to the carrier center. In GMSK, a logical "1" will cause the carrier phase to increase by 90° over a bit period and a logical "0" will cause the carrier phase to decrease by 90°. This phase change is produced by instantaneously switching of the carrier frequency between two different values $-f1$ and $f2$ according to the input data. $f1$ and $f2$ are given by:

$$f1 = f_c + R_b/4$$
$$f2 = f_c - R_b/4$$

where R_b is the modulation symbol rate and f_c is the nominal carrier frequency. It is interesting that in MSK the carrier frequency f_c is never sent.

Now, as the frequency correction burst consists of an all-zero bit sequence, after modulation by a GMSK (Gaussian minimum shift keying) modulator, it manifests as a complex sine wave, of frequency $R_b/4$, where, R_b is 270.833 kHz.

As each bit of the transmitted sequence eventually adds $\pi/2$ to the phase of the signal, so four bits are required before the signal returns to its initial phase state. The rate of the input sequence then determines the speed of this phase rotation. Hence, on delivery of such a sequence, the modulator should return a sinusoidal signal of frequency $R_b/4$ for both the I and Q channels. Owing to the sine/cosine relationship, the Q channel should trail the I channel by an amount of one. This is illustrated in Figure 7.30.

Owing to limitations in the accuracy of the local oscillator, the frequency of the signal changes to 67.7033 kHz $\pm\Delta f$. For example, for a 20 ppm oscillator at 900 MHz, Δf is 18 kHz. Also, the presence

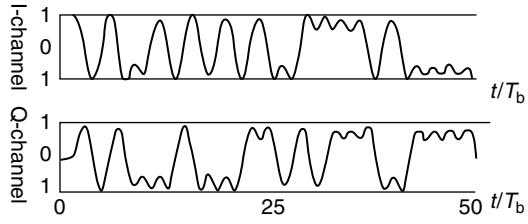


Figure 7.30 I- and Q-channel information after modulation

of multi-path and noise in the system enhances the ambiguity. Thus, the frequency burst detection involves the identification of the time of occurrence of the narrow band tone with the uncertainties mentioned above.

7.9.3.5 Synchronization Channel (SCH)

1. **Information bits:** The synchronization channel information contains 25 bits. The data include the BSIC (3-bit network color code + 3-bit base station color code) and the frame number of the current frame within the hyper-frame (19-bit reduced frame number). The BSIC (base station identity code) consists of a 3-bit network color code (NCC) and a 3-bit base station (BS) color code.
2. **Channel coding:** The information is fed to the channel coder unit and ten parity bits are generated and then four all-zero tail bits are added to produce a 39-bit data block. The block is then one-half rate convolutionally encoded using the same code as the TCH/FS to produce a coded data block of 78 bits (Figure 7.31).

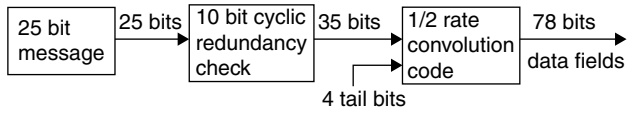


Figure 7.31 SCH processing

Ten parity bits are generated using the following polynomial:

$$g(D) = D^{10} + D^8 + D^6 + D^5 + D^4 + D^2 + 1$$

Four "all zero" tail bits are added to produce a 39-bit data block. The block is then one-half rate convolutionally encoded using the same code as the TCH/FS to produce a coded data block of 78 bits. This information block exactly fits into a single SB burst information section and interleaving is not used on the SCH.

3. **Interleaving:** Interleaving is not used for SCH. So, only one received burst is enough to start the decoding of SCH.
4. **Burst forming:** This is mapped into the synchronization burst and modulated and transmitted, as described for earlier channels.

Table 7.9 gives the coding methods for different logical channels.

Table 7.9 Coding methods for different logical channels

Channel type	Bit/block data + parity + tail	Convolutional coding rate	Bit/block	Interleaving depth
TCH/FS			456	8
Class I	182 + 3 + 4	1/2	(378)	
Class II	78 + 0 + 0	—	(78)	
TCH/F9.6	4 × 60 + 0 + 4	244/456	456	19
TCH/F4.8	60 + 0 + 16	1/3	228	19
TCH/H4.8	4 × 60 + 0 + 4	244/456	456	19
TCH/F2.4	72 + 0 + 4	1/6	456	8
TCH/H2.4	72 + 0 + 4	1/3	228	19
FACCHs	184 + 40 + 4	1/2	456	8
SDCCHs, SACCHs	184 + 40 + 4	1/2	456	4
BCCH, AGCH, PCH	184 + 40 + 4	1/2	456	4
SCH	25 + 10 + 4	1/2	78	1
RACH	8 + 6 + 4	1/2	36	1

7.10 Design of Transmitter and Receiver Blocks for GSM Radio Modem

In the previous section, we discussed the various physical layer processing steps for the different logical channel information before it is transmitted via the air. In Figure 7.32, transmission flows of the FCCH, SCH, and BCCH channels information is shown, which takes place in the BTS side (downlink). The information bits are processed and bursts are formed for different channels, which are passed though the modulator and transmitted from the BTS antenna using a broadcast carrier frequency.

MS receives the data from these channels from the air and tries to process it through its receiver circuits. The signal reception flow inside a GSM mobile receiver is shown in the Figure 7.33. After switch ON and the initial boot on procedure, the mobile will read the FCCH channel information from the sampled (I, Q) incoming data from the ADC. FCCH does not require any channel estimation or equalization process, so the signal directly goes to baseband and generally one algorithm is used to detect an all-zero sequence in the received samples to establish the presence of the FCCH channel data (which is nothing but sampled values of a pure sine wave of frequency 67.7 kHz). Once the FB is found, the frequency is estimated and frequency correction is applied to the local crystal. Next, the SCH channel is scheduled to read, and this requires the channel estimation and decoder modules, but SCH does not require ciphering and de-interleaving modules (as SCH data are not ciphered or interleaved). Next, the BCCH channel is scheduled to read and decode, which comes using normal burst. For this, channel estimation, equalization and de-interleaving modules are used. The channel estimation block computes the inverse of the channel transfer function using the known training sequence bits in the burst and passes to the equalizer. Generally, the output from the channel estimation block is the filter co-efficient, which will be used for equalization purposes. Then the equalizer produces the soft-bits with a confidence value and burst SNR. This is de-interleaved and then passed to the Viterbi decoder for producing the hard-bits, which are passed to the higher protocol layer.

For traffic channel reception (which also uses NB, and where burst data are interleaved, ciphered) all the blocks used are as shown in the Figure 7.33, and the decoded data blocks are passed to the speech decoder to produce a voice signal. All these processes are described in detail in Chapter 9 and different blocks are described in Chapter 10.

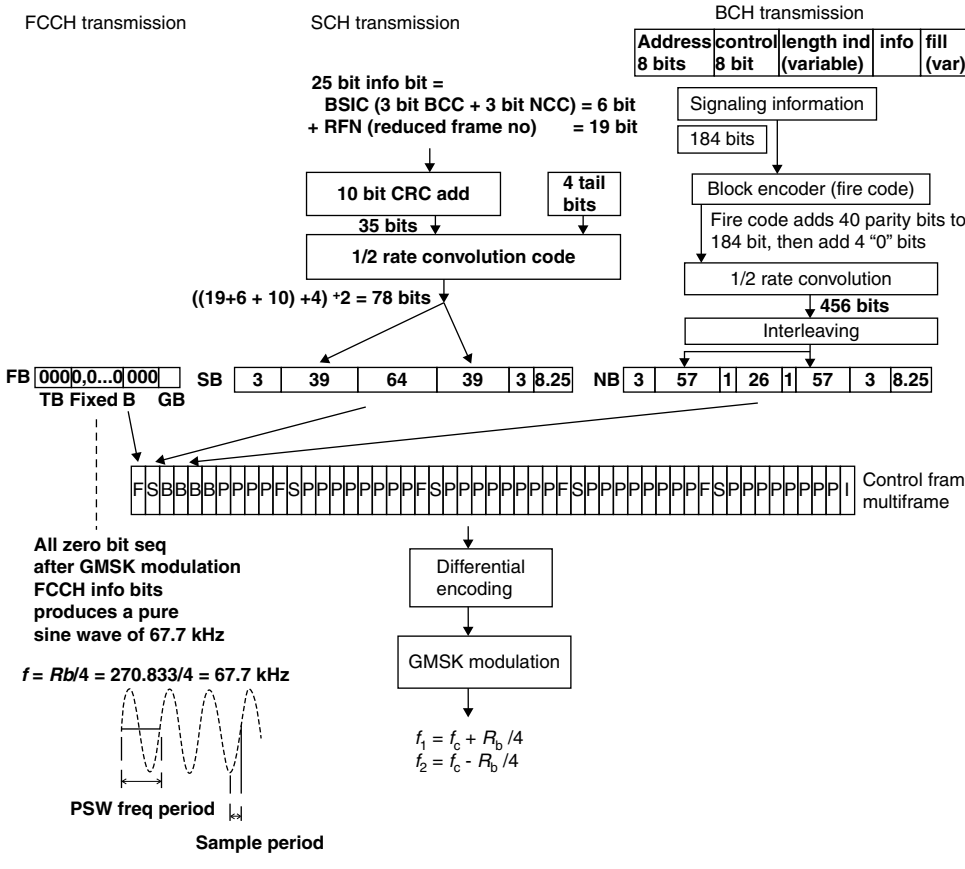


Figure 7.32 FCCH, SCH, and BCH information transmission flow

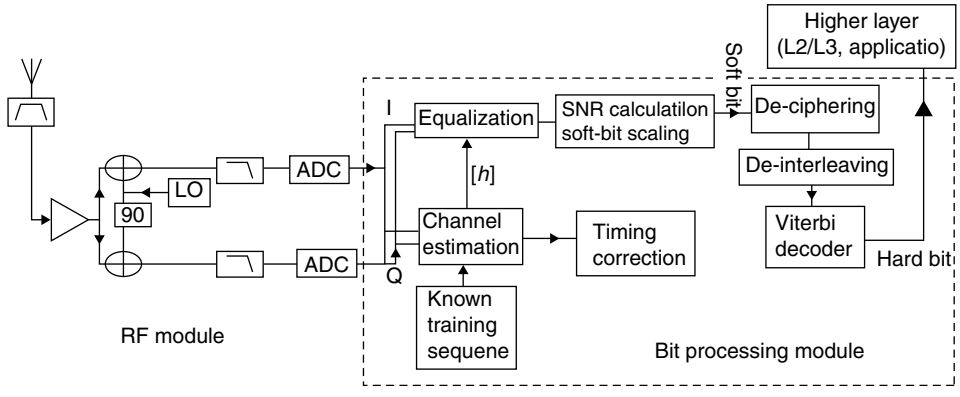


Figure 7.33 Signal reception flow inside a mobile receiver

Further Reading

GSM Technical Specification (1998) 05.02 (ETS 300 574). *European Digital Cellular Telecommunications System (Phase 2); Multiplexing and Multiple Access on the Radio Path*, ETSI TC-SMG, Sophia-Antipolis Cedex.

GSM Technical Specification (1998) 05.03 (ETS 300 575). *European Digital Cellular Telecommunications System (Phase 2); Channel Coding*, ETSI TC-SMG, Sophia-Antipolis Cedex.

GSM Technical Specification (1998) 05.04 (ETS 300 576). *European Digital Cellular Telecommunications System (Phase 2); Modulation*, ETSI TC-SMG, Sophia-Antipolis Cedex.

GSM Technical Specification (1998) 05.05 (ETS 300 577). *European Digital Cellular Telecommunications System (Phase 2); Radio Transmission and Reception*, ETSI TC-SMG, Sophia-Antipolis Cedex.

Redl, S.M., Weber, M.K., and Oliphant, M.W. (1995) *An Introduction to GSM*, Artech House, Norwood, MA.

Steele, R., Lee, C.-C., and Gould, P. (2001) *GSM, cdmaOne and 3G Systems*, John Wiley & Sons Ltd., Chichester.

8

GSM Mobile Phone Software Design

8.1 Introduction to GSM Mobile Handset Software

As discussed in the earlier chapters, a GSM mobile handset system consists of several essential hardware blocks, the necessary software for driving those hardware blocks, protocol stack for governing communication, and application software for running different applications. A high level picture of different software blocks, which are required for a typical GSM mobile phone system, is shown in Figure 8.1.

Typically the software part consists of several modules, such as the boot loader, initialization code, protocol stack, device drivers, and RTOS. In addition, audio–video related software, Bluetooth stack, and other application software (such as gaming, calculator, etc.) are also in-housed in a mobile phone device.

8.1.1 Boot Loader and Initial Power on Software Module

Once the reset button is pressed the processor jumps to the reset location (the data book of each microcontroller/microprocessor specifies the reset vector/memory location) and the jump to the next to address location is then specified. Boot loader is a small program that loads the operating system into memory of the mobile, when the mobile is booted. Generally, the boot loader is stored and executed from a special ROM area of the chip, and this uses the crypto capabilities and makes sure that only certified software can be downloaded. The boot loader allows programming the of the flash via a USB/UART connection. An embedded system also uses a flash boot loader, which resides in flash memory.

As an example we will discuss a simple boot-on procedure for a mobile phone. After the reset, startup/init code is executed, where static variables are initialized and memory areas are cleaned up. Also, the power-up sequence and initialization of the necessary hardware blocks are managed in this code, and then interrupts are enabled as are the wait-state for accessing of Flash and the external RAM is setup. Next, the SW drivers, services, and the RTOS will be initialized, so that the scheduling of threads can be started and the system starts up. Device drivers are in charge of mapping the hardware functionality of several blocks for inter-IC communication, user interaction, and storage/retrieval of data into a software API, which can be used by other drivers, services or SW blocks in the system. Examples of services/middleware are: JPEG encoding/decoding, unified file system on different storage media, MP3 player, MIDI player. A camera system controls a camera sensor via two different hardware interfaces and manages the programming of

Mobile Handset Design Sajal Kumar Das

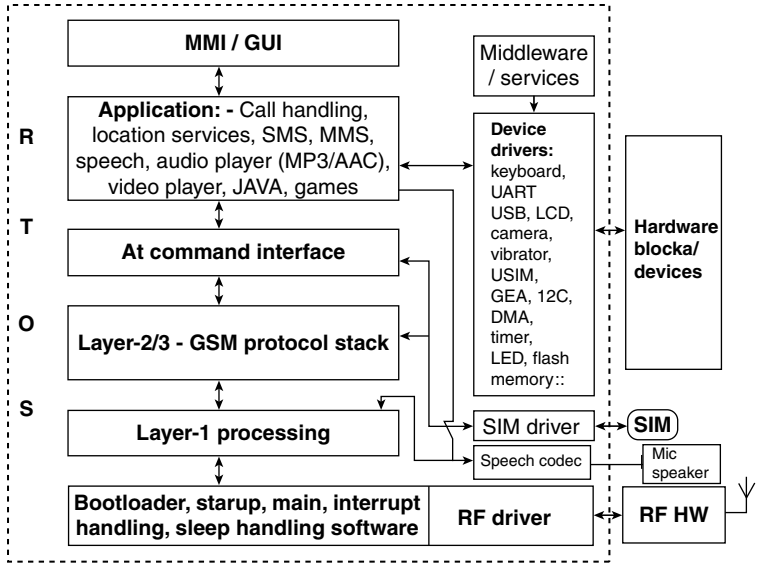


Figure 8.1 High level block diagram of a mobile phone software architecture

the DMA between the sensor, external memory, and the LCD in order to support camera preview and snapshot. Most of the time a real time operating system (RTOS) is used for hardware abstraction, real-time scheduling and to avail all OS functionalities. The software driver is used to program and configure the RF module for transmission and reception at a scheduled time. The L1, protocol stack L23, and ACI represent the modem software of the phone. On top of ACI an AT-command based interface is provided to enable application to control the modem. Applications are seen as separate functional blocks, which take care of several areas in the phone user interface (UI). For example browsers for WAP, editors for SMS and MMS, personal information manager, phonebook manager, games, camera UI and idle screen. Applications are controlled by a man–machine interface (MMI), which controls the activation, suspension and deactivation of the various applications and which arbitrates the human interface I/O blocks keypad, microphone, loudspeaker, display, and so on between applications.

8.2 Operating System Software

Today, every mobile must have an operating system to run the various programs. The Operating system performs basic tasks, such as recognizing input from the keyboard, sending output to the display screen, hardware abstraction, resource management, memory management, user interface, driving applications, and so on. Most operating system designs involve a software component called the kernel, which is responsible for hardware abstraction and resource management. Most of the architectural decisions in operating system design concern exactly what role the kernel should play in the operating system architecture.

A real-time operating system (RTOS) performs its functions and responds to external events (such as interrupts) within a specified period of time ($\sim$20 µs). It is usually more efficient, predictable, easier to maintain, and less buggy. The RTOS should implement task-priority levels, so that important tasks may be executed at a higher priority and they should allow changing a task priority during the run-time. There are several candidate OSs for digital cellular handset applications and modem operation. A brief discussion of some of these is given below.

8.2.1 Symbian

In 1998, Symbian was formed by Nokia, Ericsson, Motorola, and Psion Software with the aim of developing the software and hardware standards for the next generation of wireless smart phones and portable terminals. The Symbian operating system (EPOC) has been developed and used in palmtop computers such as the Psion Series 5. This is an open operating system, designed for mobile devices, with associated libraries, user interface frameworks, and reference implementations of common tools, and runs exclusively on ARM processors. It is structured like many desktop operating systems with pre-emptive multitasking and memory protection. EPOC was inspired by an openVMS-type of approach for multitasking with server-based asynchronous serialized access, based on events.

8.2.2 RT-Linux

Being a multi-user system, UNIX enjoys several other advantages over the MSDOS. UNIX, which is a trademark of Novel, which is very expensive and not suitable for a single PC or mobile user. To overcome this difficulty, many developers across the world contributed and came up with new free shareware that is known as Linux. The Linux is a freely distributable version of UNIX developed primarily by Linus Torvalds at the University of Helsinki in Finland. Linux is a clone of the UNIX operating system. Compared with UNIX OS, the Linux is a small, portable, fast and flexible operating system. Linux is not a trademark; it has been publicly available since November 1991. Linux is not designed to be a hard real-time operating system. The main reason is the non-preemptive behavior of the Linux kernel. Non-preemptive means that in same cases it may be that the Linux kernel will not execute a higher priority task even if it is required to do so. The sources of unpredictability in the Linux OS includes a Linux scheduling algorithm, device drivers, the use of interrupt disabling, uninterruptible system calls, and virtual memory operation.

So, to make the Linux kernel behave as a real-time kernel either we have to rewrite the Linux kernel, which is very difficult, or add some extra modules to the existing Linux to run as a real time OS. The best way to avoid these problems is to make a small predictable kernel separate from the Linux kernel. This simple technique gave rise to the idea of RT-Linux. However, the question still remains of how to implement this.

Linux supports the dynamic loading and unloading of kernel modules (pieces of code) and once the module is loaded, it becomes the part of the kernel. So, if we can load a module that could take over the system and run all Linux native processes as a lowest priority task, then we will be able to use Linux as a real-time operating system. This is the approach used by real-time Linux. Here, these modules take over the system and implement their own task schedulers. Under this new task scheduler, Linux itself runs at the lowest priority. In this case we can run real-time tasks under this newly installed scheduler, which can support preemption and all the user-programs are scheduled to run only when no real-time task is scheduled. Real-time Linux, or RT-Linux, was initially developed as an educational project. Now FMSLabs develops and maintains this. The RT-Linux kernel is shown in Figure 8.2.

8.2.3 Palm

In 1996 the Palm OS was developed by US Robotics, which is owned by Palm Computing Inc., for personal digital assistants (PDAs). Palm OS is designed for ease of use with a touch screen-based GUI (graphical user interface). The key features of the current Palm OS (Garnet 5.4) are: simple, single-tasking environment to allow launching of full-screen applications with a basic, common GUI set, handwriting recognition input system known as Graffiti 2, and so on.

Generally, in mobile handsets, a range of RTOSs are used for protocol, radio modem, and application processing activities.

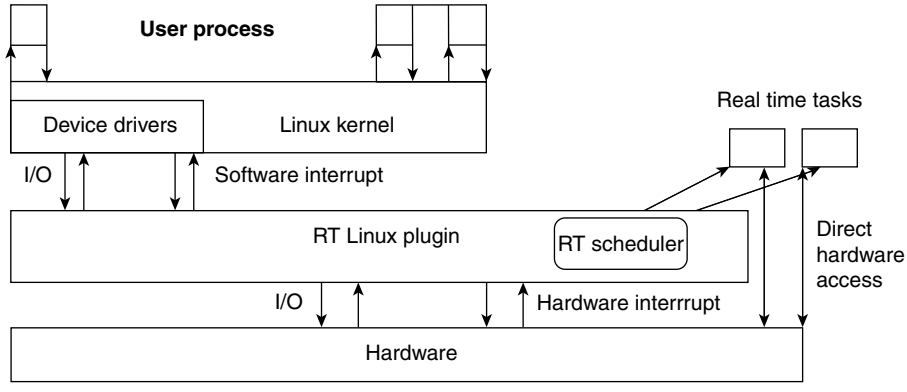


Figure 8.2 RT-Linux kernel

8.3 Device Driver Software

Protocol software, radio modem processing software (layer-1 part), and application software usually run on processors on top of an RTOS and there are several hardware blocks, which are meant to perform specific tasks (such as keypad, microphone, display, RF HW module, etc.). These hardware blocks are configured and controlled by the processors as required. It is risky to work with bare hardware and also difficult to make the hardware work according to the command. Device driver is a software program that controls devices such as the keyboard, LCD, camera, USB, and so on. A device driver acts like a translator between the device and programs that use the device. Each device has its own set of specialized commands that are known to its driver. In contrast, most programs access devices by using generic commands. The driver, therefore, accepts generic commands from a program and then translates them into specialized commands for that device. Now, instead of putting code in each application, we write controls for each device and then share the code between different applications in order to share the same hardware among the various applications.

8.4 GSM System Protocol Software

As mentioned earlier, for various reasons (for example, to connect to the network, to set up, maintain, release a call, and to route a call seamlessly), there is a need to define a set of rules or protocols between various network entities, which should be followed for communicating or passing information among the entities. For the GSM system these are called GSM protocols. The protocol stack for GSM follows the same basic concepts of ISO OSI-7 layer architecture, but this has been modified at the lower three layers to suit the specific requirements. The very first aim of communication is to transport user information, but in order to support this, in parallel there is also a need for signaling data transmission. We have discussed earlier, for GSM, that the TCH is used for user specific traffic information sending and the SACCH and FACCH channels are mainly used to transmit the signaling information during the call (for example, along with the TCH) and SDCCH is used outside a call, for example, when the call is not established.

The overall protocol architecture can be broadly divided into three planes: (1) user plane (speech, data); (2) control plane (signaling); and (3) management plane (management of network elements, such as configuration, faults, and so on). Within a GSM network, different protocols are needed to enable the flow of data and signaling between the varioust GSM subsystems. Figure 8.3 shows the interfaces that link different GSM subsystems and the protocols used to communicate on each interface. The GSM protocol

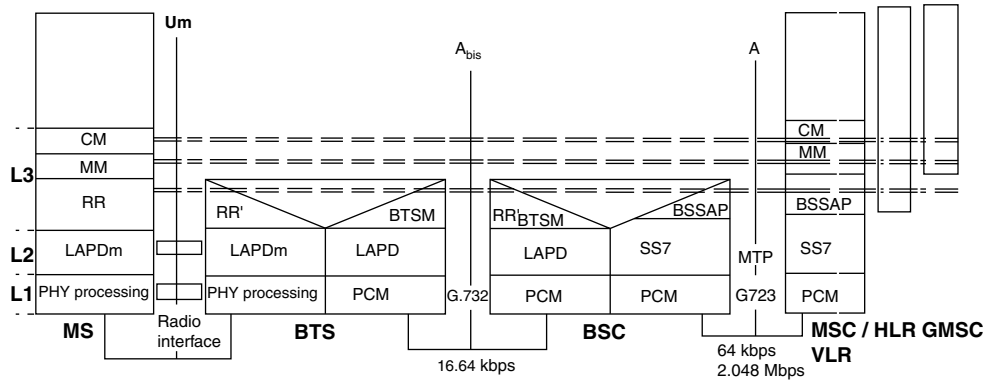


Figure 8.3 GSM system entities and protocol interfaces

architecture is designed such that the MS communicates with various protocol entities at different levels of abstraction.

GSM specific protocols are mainly divided into three layers:

a. **Layer-1** is the physical layer, which uses the channel structures over the air interface (Um) as discussed in Chapter 6, Section 6.7.1 and Chapter 7. This is responsible for channel encoding/decoding, interleaving, ciphering, burst forming, for example, bit (or symbol) transmission and reception over the air link.
b. **Layer-2** is the data link layer. The functionalities of this layer are: multiplexing of one or more layer-2 connections on control/signaling channels, error detection (based on HDLC), flow control, transmission quality assurance, routing, and so on [1].
c. **Layer-3** is the network layer. The functionalities of this layer are: connection management (air interface), subscriber identification, management of location data, and management of added services (SMS, call forwarding, conference calls, etc.).

The user plane protocol architecture is much simpler and involves only physical layer and data-link layers.

8.4.1 GSM Mobile Handset (MS) Protocol Stack

Inside a GSM mobile handset, entities for all the layers are residing and interacting with their corresponding counterparts, which are spread across the GSM subsystems as shown in Figure 8.3. Different sub-layers in layer-2 and layer-3 of the GSM phone protocol stack are shown in Figure 8.4. Layer-2 and layer-3 consists of several entities and generally, entities run as separate threads in an OS environment and use a queue based communication model, where the primitive contents are stored in dynamically allocated partition memory sections. Generally, the interface to L1 consists of the primitive based service access points.

DL (data link layer) – This provides layer-2 functionality to the RR on different logical GSM channels.
RR (radio resource) – This sub-layer makes sure that a suitable cell is selected, the surrounding neighbor cells are observed and the serving cell with the best radio quality is used, in idle mode and in a call or data connection.

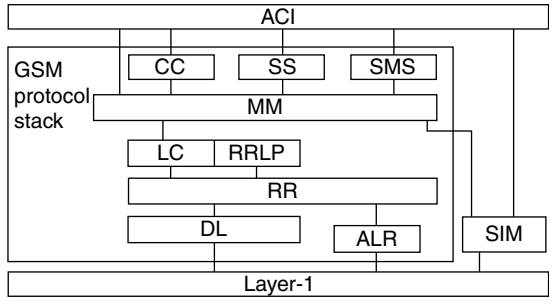


Figure 8.4 Mobile phone GSM protocol stack (L2/L3) architecture

ALR (adaptation layer) – This is an adaptation sub-layer that exists between the RR and L1 and helps to support the RR in cell-selection/re-selection, SI pre-processing, and paging detection.

MM (mobility management) – This is a user protocol between the mobile station and the network switching subsystem (NSS), for which the base station subsystem (BSS) is transparent. This sub-layer makes sure that the mobile stays in the registered state in the home or roaming network. It is responsible for the functions relating to location registration, paging, attachment/detachment, handover, dynamic channel allocation and management. MM maintains the full or limited service of the phone. The messages defined in MM allow for roaming and security functions in GSM.

CC (call control) – This is in charge of circuit switched call handling. It performs the required signaling between MS and network in order to establish, receive, maintain, and end a call. Handling of conferences and a second call is also in the scope of this entity. CC is part of the connection management (CM) layer.

SS (supplementary service) – This is in charge of getting/setting and querying services in the network, such as call-forwarding, call-deflection, and call-barring. SS is part of the connection management layer.

SMS (short message service) – This provides the capability to send and receive SMS. The reception of a cell broadcast messages (CBM) is also handled here. SMS is part of the connection management layer.

SM (session management) – This sub-layer controls the activation/deactivation and modification of the PDP (packet data profile) context in the GPRS system (discussed in Chapter 13). A context defines the QoS for a packet oriented connection, the used NSAPI/SAPI (service access point identifier), the IP addresses of the MS, gateway and the DNS. Up to seven contexts can be handled by SM. It is part of the connection management layer.

Each entity consists of several state machines, which process incoming events such as primitives, timeouts, events, and use static storage to store state relevant data between events. Entities can interact with other entities by sending events or primitives and they could maintain their own timers. Several entities could be grouped and run in a single OS thread as: CC – SMS – SS – SM.

As the real-time requirements and the protocol itself allow that only one entity of the groups is active at a particular time, the entities can thus share the same input queue and the same stack.

SIM (subscriber identity module) – This entity controls the SIM driver that manages the access to different SIM data fields and provides an SAP (service access point is the interface point between two layers) towards MM, GMM, and ACI.

ACI (application control interface) – The application control interface consists of several state machines, which are triggered by AT-commands or incoming primitives from the underlying L23 entities. ACI controls the L23 functions also by primitive exchange on the connected SAPs.

MMI (man–machine interface) – MMI is used to interact with the user. It takes the input from the user through a touch screen or key pad and displays the output on the LCD screen or invokes the proper operations.

In the next sections, the protocols and message structure used in different layers for communication among different entities in the system are described.

8.4.2 Air Interface (Um) Protocol

In the Figure 8.5, the interfaces between L1, L2, and L3 layers are shown. In the signaling plane, as RR is an L3 entity, it can interact with L1 and L2.

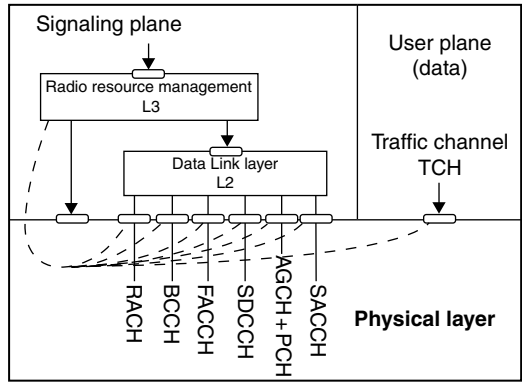


Figure 8.5 L1, L2, and L3 interfaces

8.4.2.1 Layer-1

The physical properties of the Um interface have already been described in detail in the previous chapter.

8.4.2.2 Layer-2

The main functionalities of the link layer are structuring in the frame, segmentation and reassembly, error detection and correction, multiplexing and flow control. Across the Um interface, the data link layer used is the LAPDm (modified LAPD). This is a modified version of the LAPD (link access protocol for ISDN "D" channel) protocol, which is used in ISDN. This LAPDm protocol is related to two other layer-2 protocols, such as HDLC and LAPB. For the development of the LAPDm protocol, the LAPD protocol is taken and all dispensable parts were removed to save resources. The frame formats defined for LAPDm are based on those defined for LAPD. However, there are some important differences between LAPDm and LAPD, in particular with regard to frame delimitation methods and transparency mechanisms. These differences are necessary for operation within the constraints set by the radio path. LAPDm supports two modes of operation: (1) unacknowledged mode operation using UI frames (no flow control and error control); and (2) acknowledged mode operation using the multiple frame procedure (positive

acknowledgment, error correction based on ARQ). For BCCHs and CCCHs only the unacknowledged mode of operation is implemented. LAPDm is used for information sent on the control channels BCCH, AGCH, NCH, PCH, FACCH, SACCH, and SDCCH as defined in the GSM standard.

LAPDm uses three frame formats:

1. **A-format** – This frame format can be used for any DCCH and it does not carry any higher layer data (no information field), but is is used for filling.
2. **B-format** – This frame carries the actual signaling data on the radio interface. It is transmitted in every DCCH and ACCH. The maximum length of layer-3 information is restricted based on the logical channel and is defined by parameter *N*201. A-format and B-format frames are sent both in uplink and downlink. SACCH, FACCH, SDCCH uses frame types A or B.
3. **Bbis-format** – This frame does not have any address field, as this is not required on a broadcast channel and this frame format is used for transmission on BCCH, PCH, and AGCH. These frames are only sent on the downlink.

The maximum frame length is 23 bytes = 184 bits, which is the length of the layer-2 data block passed to the layer-1 channel coding unit. Figure 8.6 shows the LAPDm frame format and coding of fields for the several message types used by the protocol. The LAPDm address field has as its main element the SAPI, through which the layer-3 message is received. On the radio interface two values of SAPI are

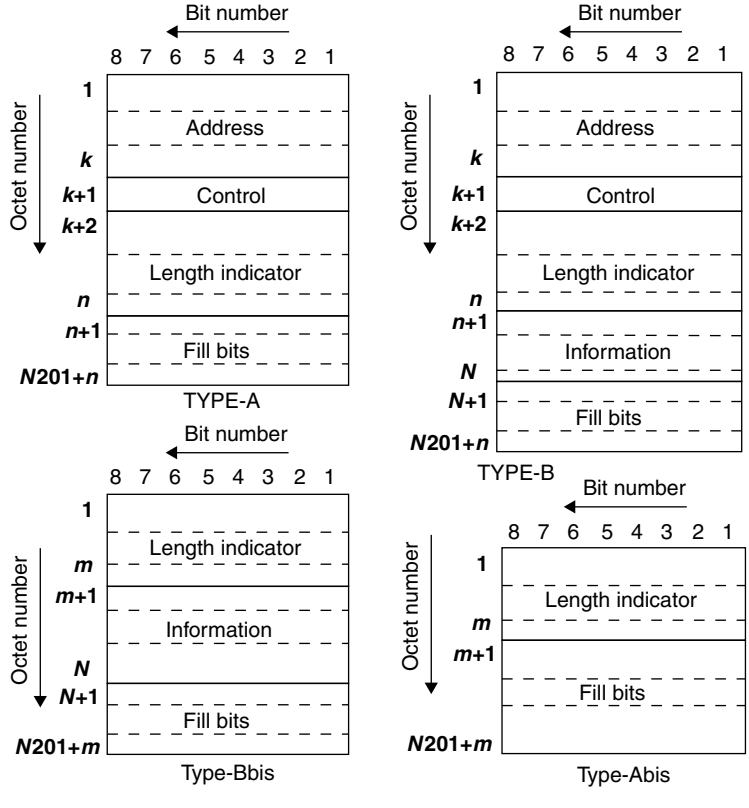


Figure 8.6 LAPDm frame format

used: (1) SAPI = 0 for messages from the radio resource management (RR), mobility management (MM), and call control (CC); and (2) SAPI = 3 for messages from the SMS and supplementary services (SS) messages. The control field is used in the same fashion as in HDLC or LAPB and contains the sequence and retransmission counters *N*(S) and *N*(R), respectively. The frame length field contains the length of the layer-3 message within the information field of the LAPDm frame. If the message is less than the length specified in parameter *N*201 of the radio interface, fill-in octets are used to make up for the space. If the layer-3 message to be transmitted is longer than *N*201, segmentation occurs. Whether segmentation has occurred or not, is indicated in the M-bit of the length field.

For link layer signaling, two physical layer channels are used (SACCH and FACCH). The frame structure of the SACCH message is shown in Figure 8.7. The important functionality of a link layer is to improve the quality of transmission, by detecting frames that have been subjected to transmission errors, and probably asking for repetition. It supports both acknowledgment (acknowledged back indicating correctly reached) and un-acknowledgment modes. The link layer offers the possibility of multiplexing independent flows on the same channel. On the radio interface, two independent flows can co-exist. The first one is devoted to transfer of signaling messages and the second one is for SMS. These two flows are distinguished by a link identifier SAPI, as discussed earlier.

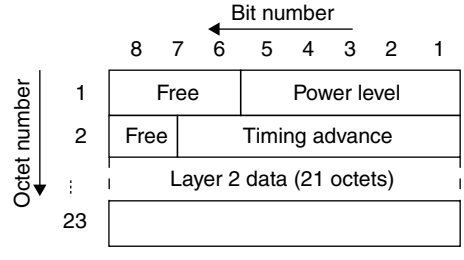


Figure 8.7 Frame structure for SACCH block

8.4.2.3 Layer-3

Layer-3 contains several sub-layers (Figure 8.8), which control signaling channel functions (BCH, CCCH, and dedicated channels). These sub-layers are radio resource management (RR), mobility management (MM), call control (CC) as well as short message service (SMS) management, and supplementary services (SS) management.

1. **Radio Resource (RR) Sub-Layer**
   RR is the heart of the protocol. This controls the setup, maintenance, and termination of radio and fixed channels, supports measurements and handovers. The main procedures in anRR layer are:channel assignment, channel release, channel change and handover, change of channel frequencies, hopping sequences (algorithms) and frequency tables, measurement reports from the MS, power control discontinuous transmission reception, time advance, modification of channel modes (speech and data), and cipher mode setting.
   An RR session is always initiated by an MS, through the access procedure, either for an outgoing call, or in response to a paging message. The access and paging procedures (such as when a dedicated channel is actually assigned to the mobile and the paging sub-channel structure) are handled in the RR sub-layer.
2. **Mobility Management (MM) Sub-Layer**
   This sub-layer manages the location update and registration procedures, as well as security and authentication. The procedures in the MM sub-layer are: location update, periodic updating, authentication procedure, IMSI attach procedure (on power up an MS will present its IMSI to

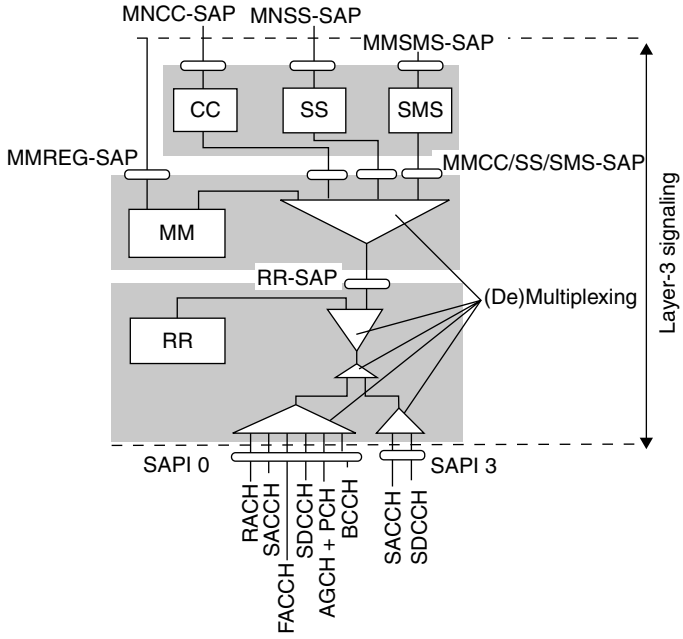


Figure 8.8 Layer-3 signaling

network and get a TMSI), IMSI detach (on power off of an MS, detach procedure to tell network it is no longer in service), TMSI reallocation, and identification.

3. **Communication Management (CM)**
   The communication management layer (CM) is responsible for call control (CC), supplementary service management (SS), and short message service management (SM). Again each of these may be considered as a separate sub-layer within the CM layer. Call control attempts to follow the ISDN procedures are specified in Q.931, although routing to a roaming mobile subscriber is obviously unique in GSM.
   a. **Call Control (CC) Sub-Layer**
      This sub-layer manages all functions necessary for circuit-switched call control in GSM PLMN, call establishment for mobile-originated calls, call establishment for mobile-terminated calls, changes of transmission mode during an ongoing call, call re-establishment after interruption of an MM connection, dual-tone multi-frequency (DTMF) control procedure for DTMF transmission.
   b. **Supplementary Service Management (SS)**
      This sub-layer contains functions such as call waiting, call forwarding, group call, called party identity, and so on.
   c. **SMS Management**
      The short message service – point to point (SMS-PP) is defined in GSM recommendation 03.40. Messages are sent to a short message service center (SMSC) which provides a store-and-forward mechanism. Both mobile terminated (MT) (for messages sent to a mobile handset) and mobile originating (MO) (for those that are sent from the mobile handset) operations are supported. Transmission of short messages between the SMSC and the handset is done using the mobile application part (MAP) of the SS7 protocol. Messages are sent with the MAP MO- and MT-forward SM operations, whose payload length is limited by the constraints of the signaling protocol to precisely 140 octets (140 octets = 140 × 8 bits = 1120 bits). The maximum single text message size is either 160, 7-bit characters or 140, 8-bit characters, or 70, 16-bit characters.

Neither the BTS nor the BSC interrupt CM and MM messages, these messages are simply exchanged with the MSC or MS using the direct transfer application part (DTAP) protocol on the A interface.

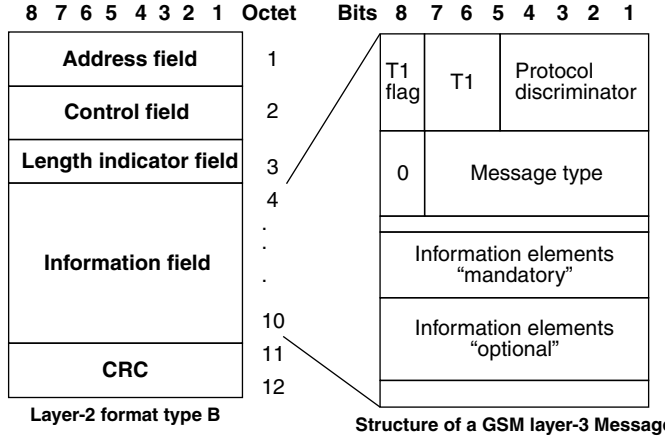


Figure 8.9 Structure of layer-3 message

Layer-3 Message Format

The structure of layer-3 message format is shown in Figure 8.9. A layer-3 message consists of three fields: the first field is type ID, and the second field is the message type, and the third field is the data field. The type ID consists of a 4-bit protocol discriminator (PD), and a transaction identifier (TI). A protocol discriminator links layer3-protocol to the entity to which the message is addressed. This identifies six protocols, as defined in Table 8.1.

Table 8.1 Six different protocols

Protocol	PD binary
RRM	110
MM	101
CC	11
SMS	1001
SS	1011
Test procedure	1111
All other values reserved	

A TI is used to distinguish between possible (multiple) parallel CC connections and between various transactions taking place over these simultaneous CC connections. The message type field consists of an 8-bit code that identifies the type of message sent. The data field is of variable length and contains information elements (IE), which convey the data to the receiver. Each message data consists of mandatory and optional IEs. Depending on MT, it may or may not have one or more IE. Types of IE are: mandatory fixed length (MF), mandatory variable length (MV), and optional fixed length (OF), and optional variable length (OV).

8.4.3 Abis Interface

An Abis interface is used between the BTS and BSC. The BSC and BTS can be connected using leased lines, radio links or metropolitan area networks (MANs). Generally, two channel types exist between the BSC and BTS:

Traffic Channels (TCH): These transport user data and can be configured in 8, 16, and 64 kbps formats.

Signaling Channels: These are used for signaling purposes between the BTS and BSC and can be configured in 16, 32, 56, and 64 kbps formats.

Generally, in a BSC, each transceiver (TRX) requires a signaling channel on the Abis interface. The positioning of the user data frames (T = traffic) and signaling data frames (S = signaling) varies from manufacturer to manufacturer and system to system. The only design requirement is that the frame alignment signal (FAS)/non-frame alignment signal (N-FAS) frame must be in time slot 0. A signaling channel can run either at rate 16 kbps (sub-channel signaling) or 64 kbps. The TRAU (transcoder rate adapter unit) has the GSM specific language coding and decoding (from 64 kbps to 16 kbps per voice channel in both directions). The TRAU is located at the BSC or directly at the MSC (mobile switching center). The TRAU frame is the transport unit for a 16 kbps traffic channel (TCH) on the Abis interface. It uses 13.6 kbps for user data and 2.4 kbps for inband signaling, timing and synchronization. In addition to the radio signaling procedures the Abis interface also provides a means of transport for operation and maintenance procedures for BTSs. A transport mechanism for layer-2 management procedures is also inherited directly from ISDN standards.

8.4.3.1 Protocol on Abis Interface

In the Abis interface the following protocols are used:

1. **Layer-1** 2.048 Mbps (ITU-T: E1) or 1.544 Mbps (ANSI: T1) PCM facility with 64/32/16 kbps signaling channels and 16 kbps traffic channels (four per time slot).
2. **Layer-2** The LAP-D protocol is used as the transport mechanism for data message sending between the BTS and BSC. Within GSM the SAPI refers to the link identifier transmitted in the LAPD protocol that is inherited from ISDN.
3. **Layer-3** The BTS management (BTSM) mainly works in layer-3. This distinguishes three logical signaling connections with the SAPI. SAPI 0 is used by all messages coming from or going to the radio interface. SAPI 62 provides OM message transport between the BTS and BSC. SAPI 63 is used for layer-2 management functions, as well as dynamic management of TEIs (terminal endpoint identifier). The addition of another field to the LAPD link layer address is for the TEIs.

8.4.4 A Interface

The interface between the BSC and MSC is the A interface. If the BSC contains the transcoder equipment (TCE), a traffic channel (TCH) occupies a complete 64 kbps time slot in the 2 Mbps or 1.544 Mbps PCM link. Out of 32 available time slots on the PCM link, at least two time slots are needed for control and signaling purposes (TS0 for FAS/NFAS and another TS for signaling, usually TS16) and a maximum of 30 traffic channels can be operated simultaneously on PCM facilities. Normally, two active 64 kbps time slots are used for signaling purposes and one signaling channel supports many 64 kbps PCM facilities between one BSC and the MSC. If the MSC is equipped with a TCE, the TCHs are converted from 64 to 16 kbps in the transcoder equipment. If the BCS does not contain a TCE, then the TCHs are 16 kbps on the A interface. Between the BSC and MSC, the TCHs are recorded from 64 to 16 kbps in the TCE.

8.4.4.1 Protocol on A Interface

The signaling protocol (layer-2 and layer-3) between the BSC and MSC is based on the SS7 standard, but is transmitted along with the user data within the PCM facility. Normally time slot 16 (TS16) of the 64 kbps frame is used. The following protocols are used for this purpose:

Layer 1: 2.048 Mbps (ITU-T: E1) or 1.544 Mbps (ANSI: T1) PCM link.

Layer 2: SS7-based protocols are used for layer-2. The message transfer part (MTP) protocol is responsible for transmission security between the BCS and MSC and the signaling connection control part (SCCP) protocol allows global addressing of network elements and offers a service corresponding to the exchange layer. MTP and SCCP also perform layer-3 functions. SCCP is used to transport DTAP and the base station management application part (BSSMAP) messages on the A interface, ensuring both connectionless and connection-oriented message flows. The connections can be related to a specific MS or radio channel. An SCCP connection can be initiated by a mobile station (MS) or an MSC.

Layer 3: Containing the base station system application part (BSSAP) protocol, this layer has multiple parts on the MSC end. The base station management application part (BSSMAP) protocol is the counterpart to the RR protocol on the air interface. The direct transfer application part (DTAP) protocol transmits CC and MM messages and is transmitted transparently through the BTS and BSC.

In Table 8.2, the various ETSI standards are tabulated.

Table 8.2 ETSI-GSM standards (the numbers in parentheses indicates the relevant GSM recommendations)

Level-3	CM (04.08) MM (4.08) RR (4.08)	RR′ (4.08)	BTSM (8.58)	RR′ (4.08) BTSM (8.58)	CM, MM(4.08), DTAP, BSS MAP BSSAP (8.06)	TUP, ISUP, INAAP, MAP, TACP, SCCP, MTP	MUP, INUP, ISUP, TUP, TACP, SCCP, MTP
Level-2	LAPDm (4.05/4.06)	LAP-Dm (4.05/4.06)	LAP-D (8.56)	LAP-D (8.56)	SCCP MTP (8.06)		
Level-1	Radio (4.04)	Radio (4.04)	64 kbps (8.54)	64 kbps (8.54)	64 kbps (8.54)	64 kbps (8.54)	64 kbps (8.54)

8.5 Speech and Multimedia Application Software

Today's GSM mobile phones support voice, audio as well as video playback features. Ideally, each of these applications generates a huge number of bits from the information source, which needs to be reduced and controlled at the source level using different source coding techniques. Uncompressed multimedia (graphics, audio and video) data require considerable storage space and transmission bandwidth. Despite the rapid progress in mass-storage density, processor speeds, and digital communication system performance, demand for data storage capacity and data-transmission bandwidth

continues to outstrip the capabilities of available technologies. The recent growth of data intensive multimedia-based applications needs more efficient ways to encode signals and images for storage and communication technology.

So far we have discussed modem software, which helps to transfer data from one mobile (or network) to the other distant mobile. Thus the data generated at the speech encoder side are transmitted and on the receiver side the received bits are passed to the speech decoder to reproduce the voice. When we talk in front of a microphone, the voice is turned into a digital bit stream using waveform encoding (source coding). The vocoder's main job is to reduce the data rate. In wire-line systems this is achieved in the time domain by using time domain compression techniques, but in digital cellular handsets, speech synthesis codec is used in the frequency domain. At the transmitter side the source encoding is done by describing each sample in terms of frequency coefficients and compression is achieved by exploiting the similarity between the samples. At the receiver side the decoder uses the frequency coefficients to rebuild or synthesize the harmonic structure of the original voice sample. In order to carry out these source encoding and decoding functions, the mobile contains a source codec unit. This can be implemented in software (generally runs on DSP) or in hardware logic.

8.5.1 Speech Codec

A speech codec is a particular type of audio codec designed specifically for human voice encoding and decoding. By analyzing vocal tract sounds, instead of sending the sound waves, a recipe is sent to the receiver end for rebuilding the sound. The speech codec is able to achieve a much higher compression ratio, which results in a smaller amount of digital data transmission. Speech quality as produced by a codec is a function of transmission bit rate, complexity, delay, and bandwidth. Speech coding differs from other forms of audio coding, as speech is a much simpler signal than most other audio signals, a speech signal is limited to a bandwidth of 300–3400 kHz (whereas audio signal is limited to a bandwidth of 0–20 000 Hz, for example, audible range); there is lot of statistical information available about the properties of speech. A speech signal varies fairly infrequently, resulting in a high degree of correlation between consecutive samples. This short-term correlation is due to the nature of the vocal tract. Long-term correlation also exists due to the periodic nature of the speech. This statistical redundancy can be exploited by introducing prediction schemes, which quantize the prediction error instead of the speech signal itself. On the other hand, shortcomings in the human capability to receive sounds lead to the fact that a lot of information in a speech signal is perceptually irrelevant. This means that a human ear can not differentiate between the changes in magnitude below a certain level and can not distinguish frequencies below 16 Hz or above 20 kHz. This can be exploited by designing optimum quantization schemes, where only a finite number of levels are necessary.

Speech coding methods can be classified as:

a. Waveform coding
b. Source coding
c. Hybrid coding

Source codecs try to produce a digital signal by modeling the source of the codec, whereas waveform codecs do not use any knowledge of the source of the signal but instead try to produce a digital signal, whose waveform is as identical as possible to the original analog signal. Pulse code modulation (PCM) is the most simple and purest waveform codec. Hybrid codecs attempt to fill the gap between the waveform and source codecs. In the early 1960s, when telephones were using first digital signal transmission techniques, the PCM was used to generate a 64 kbps digital bit stream from the analog voice signal. In pulse code modulation (PCM) codec, a voice signal is sampled at a rate of 8 kHz, with each sampled voltage level being converted to 8 bit, so that the total bits generated per second are: $8 \times 8000 = 64$ kbits.

Narrowband speech is typically sampled 8000 times per second, and then each speech sample must be quantized. If linear quantization is used then about 12 bits per sample are needed, giving a bit rate of about (12×8 kbps = 96 kbps). However, this can be easily reduced by using non-linear quantization. For coding speech it was found that with non-linear quantization, 8 bits per sample were sufficient for speech quality, which is almost indistinguishable from the original. This gives a bit rate of 64 kbps, and two such non-linear PCM codecs were standardized in the 1960s. In America μ-law coding standard, and in Europe the slightly different A-law compression is used. Because of their simplicity, excellent quality and low delay, both of these codecs are still widely used today. Sun Microsystems Inc. has released code to implement the G711 A-law and μ-law codes into the public domain and this has been modified by Lindberg [2].

The most common speech coding scheme is code excited linear prediction (CELP) coding, which is used for example in the GSM standard [3]. In CELP, the modeling is divided into two stages, a linear predictive stage that models the spectral envelope and a code-book based model of the residual of the linear predictive model.

8.5.1.1 GSM Codecs

Full-Rate Codec (FR)

The full-rate speech codec in GSM is described as regular pulse excitation (RPE) with long-term prediction (LTP) (LPC-RPE codec: GSM 06.10 RPE-LTP). It is a full-rate speech codec (FR) and operates at 13 kbps. The encoder has three major parts: (1) linear prediction analysis (short-term prediction), (2) long-term prediction, and (3) excitation analysis.

The encoder processes 20 ms blocks of speech and each speech block contains 260 bits (188 + 36 36 = 260), which is depicted in Figure 8.10. So the rate is 260 bits per 20 ms = 13 000 bps = 13 kbps.

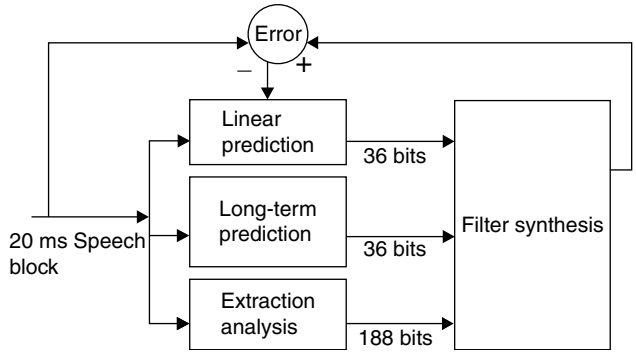


Figure 8.10 Diagram presentation of the GSM full-rate LPC-RPE codec

Generally, the input speech is split up into frames of length 20 ms, and for each frame a set of eight short-term predictor coefficients are computed. Each frame is then further split into four 5 ms sub-frames ($4 \times 5 = 20$ ms) and for each sub-frame the encoder finds a delay and a gain for the codec's long-term predictor. The linear predictor part of the codec uses 36 bits and linear prediction uses a transfer function of the order of 8. The long-term predictor estimates pitch and gain four times at 5 ms intervals. Each estimate provides a lag coefficient and a gain coefficient of 7 bits and 2 bits, respectively. Together these four estimates require $4 \times (7 + 2)$ bits = 36 bits. The gain factor in the predicted speech sample ensures that the synthesized speech has the same energy level as the original speech signal.

The remaining 188 bits are derived from the regular pulse excitation analysis. After both short- and long-term filtering, the residual signal, that is the difference between the predicted signal and the actual signal, is quantized for each 5 ms sub-frame.

At the decoder the reconstructed excitation signal is fed through the long-term and then the short-term synthesis filters to give the reconstructed speech as shown in Figure 8.11. A post filter is used to improve the perceptual quality of this reconstructed speech.

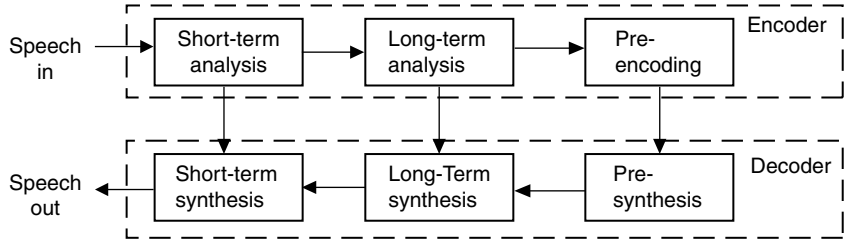


Figure 8.11 Block diagram of GSM speech encoder and decoder

However, on the network side the situation is slightly more complicated, as speech signals are usually coded using an 8-bit A-law pulse code modulation (PCM) format in order to be compatible with the PSTN or ISDN. Thus before the speech signal is passed to the speech coder on the network side, it must first undergo an 8-bit A-law PCM to 13-bit uniform PCM conversion.

Half-Rate (HR) Codec

GSM has also defined the half-rate version of the GSM codec. This is a vector self-excited linear predictor (VSELP) codec at a bit rate of 5.6 kbps. It is a close relative of the CELP codec family. The difference is that VSELP uses more than one separate excitation codebook, which are separately scaled by their respective excitation gain factors. The GSM half-rate vocoder operates in one of four different modes (0, 1, 2, 3) based on the grade of the voice detected in the speech. The speech spectral envelope is encoded by using 28 bits per 20 ms frame for vector quantization of the LPC coefficient and the four synthesis modes corresponds to different excitation modes.

Enhanced Full-Rate (EFR) Speech Codec

The enhanced full rate (EFR) speech codec is defined by the European Telecommunications Standards Institute (ETSI). It has a bit rate of 12.2 kbps and uses the algebraic code excited linear prediction (ACELP) algorithm, which is an analysis-by-synthesis algorithm.

8.5.1.2 AMR Codec (AFS/AHS)

The adaptive multi-rate (AMR) codec is the speech codec standard for GSM phase 2+, which adaptively changes the source rate based on the quality of the wireless channel. The AMR speech codec was proposed by ETSI in June 1996 to improve speech quality in mobile phones and to compensate for the GSM slow power control. AMR is based on EFR speech codec; it incorporates multiple sub-modes for use in full-rate or half-rate modes that are determined by the channel quality. The two options for AMR logical speech channels are: adaptive full-rate speech (AFS) and adaptive half-rate speech (AHS). In order to provide the best speech quality, variable portioning between speech and channel coding bit rates is selected based on the variation of the channel conditions. According to the channel quality, the receiver can request (or command) the transmitter AMR to adjust the speech coding rate to allow for a higher or lower channel coding rate in response. Thus if the channel quality deteriorates, progressively lower codec rates are requested, otherwise, if channel conditions improve, higher codec rates are requested. The codec rate requests and commands are transmitted as often as every 40 ms, using in-band signaling. AMR rate requests/commands give an indication of the channel quality and they are transmitted more often than RXQUAL and RXLEV.

The AMR codec (narrowband) uses a set of eight codec rates (4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.2, and 12.2 kbps) for speech encoding. For every 20 ms input speech frame, the codec rate can be switched to a different codec rate. In GSM, only a subset of the possible codec rates is used during a connection. This subset is referred to as the active codec set (ACS) and contains at least one and at most four of the possible AMR codec rates. The network decides on an active codec set of up to four code modes in AFS and AHS. This active code set is initially signaled to the MS during call set-up via a layer-3 signaling channel assignment/immediate assignment/channel mode modify/handover message (see GSM spec. 4.08). It is possible to change the ACS during a connection.

Codec Mode Information

The codec mode information sent on the downlink are:

a. **Codec Mode Indications (CMI)** – CMI is used for indicating to the peer AMR which codec rate is to be used for decoding the received speech frame.
b. **Codec Mode Commands (CMC)** – CMC instructs the AMR on the MS side which codec mode is to be applied.

The codec mode information sent on the uplink are:

a. **Codec Mode Indications (CMI)** – As mentioned above.
b. **Codec Mode Requests (CMR)** – The CMR informs the other end (that is, BTS end) of the preferred codec mode. This means, based on the channel quality, that the MS requests the preferred rate to the network.

The codec mode indications (CMIs) and codec mode commands (CMCs) or codec mode requests (CMRs) are sent alternately, on consecutive speech frames. The codec mode changes only every second speech frame, for example, the signaling of CMI and CMR messages (CMI and CMC) is alternated in the uplink (downlink) resulting in a 40 ms signaling interval for each type of message. Codec mode information is transmitted in-band in the speech traffic channel. The details of the in-band coding can be found in GSM Standard 5.03, Section 3.10.7 [4]. The codec rate to be applied for encoding of each input speech frame needs to be provided to the codec every 20 ms. Also, the codec rate to be used for decoding every frame has to be provided. In GSM, the codec rate information is transmitted in-band every 20 ms with the encoded speech frames. The robust AMR traffic synchronized control channel (RATSCCH) mechanism is used to modify the AMR configuration on the radio interface without interruption of the speech transmission. During regular speech transmission (in the middle of a speech burst) RATSCCH replaces (steals) one TCH/AFS (or two TCH/AHS) speech frames [5]. Also, in all non-speech cases the RATSCCH will be handled the same as for speech.

Channel Quality Measure and Link Adaptation

The receiver side performs link quality measurements, which are known as quality indicators. The details concerning the reference performance are in 3GPP Standard 5.05 [6]. The CMC/CMR messages are generated based on a particular channel quality metric such as the estimated received C/I and then the estimated C/I is compared against codec mode switching thresholds, as shown in Figure 8.12. The quality indicator is defined as a normalized C/I ratio based on actual C/I estimates from the equalizer or raw BER estimates. The equalizer either estimates C/I (SNR) or this has to be derived from the raw bit errors from the channel decoder. If the equalizer estimates C/I (SNR), then it should communicate this information to the vo-coder for post-processing and to determine the code mode requests by comparing with the threshold levels. The MS and BTS will continuously update their quality indicator on a frame by frame basis.

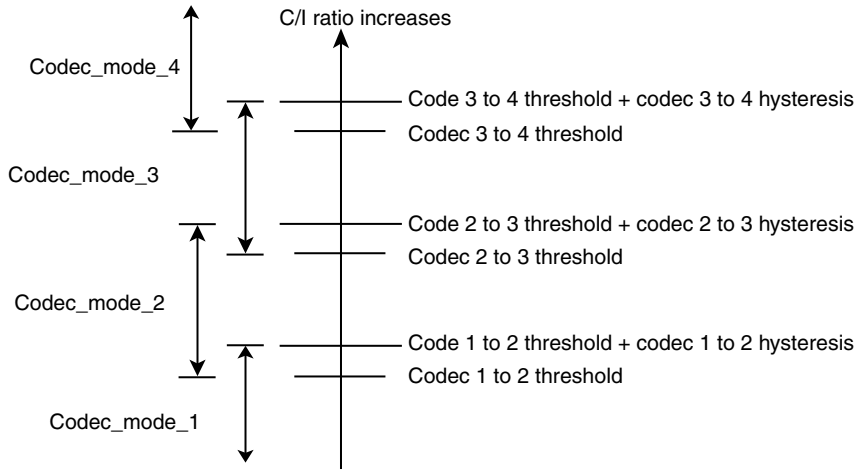


Figure 8.12 Codec mode adaptation based on estimated C/I levels

The quality indicator is directly fed into the UL mode control unit in the case of an uplink adaptation. This unit compares the quality indicator with certain thresholds and generates a codec mode command indicating the codec mode to be used on the uplink. The codec mode command is then transmitted in-band to the mobile side, where the incoming speech signals are encoded using the corresponding codec mode.

However, for downlink adaptation, the DL mode request generator within the mobile compares the DL quality indicator with certain thresholds and generates a codec mode request indicating the preferred codec mode for the downlink. The codec mode request is transmitted in-band to the network side, where it is fed into the DL mode control unit. This unit generally grants the requested mode or sometimes ignores it. The resulting codec mode is then applied for encoding of the incoming speech signal in the downlink direction.

Both for uplink and downlink, the presently applied codec mode is transmitted in-band as a codec mode indication together with the coded speech data. At the decoder, the codec mode indication is decoded and applied for decoding of the received speech data.

In-band signaling is the term used to refer to an embedded data of information in addition to the block of data transmitted (speech data block, RATSCH data block or FACCH data block). The speech coder delivers to the channel encoder a sequence of data blocks. One data block corresponds to one speech frame and the block length is different in each of the eight channel codec modes. As shown in Figure 8.13, the in-band data id (0,1) representing mode indication or mode command/mode request based on the current frame number, is sent along with the data block, with information of the channel codec mode to use when encoding the block.

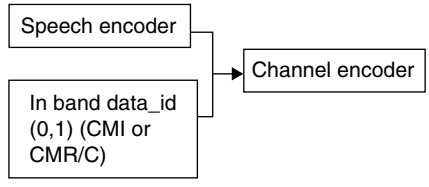


Figure 8.13 Speech and in-band data input to the channel encoder unit

After reception of the code mode command, the MS should apply the codec mode in the next possible speech frame. If the codec mode requires the change of mode in more than one step within the active code set, this should be performed in several steps, with one step every second speech frame, except for the call set-up and handover. In such a case, the MS should start with the initial code mode (ICM), irrespective of what the previously used code mode by layer-3 signaling or by a RATSCCH message was.

Frame Formats

Traffic frames are blocks of 95 . . . 244 information bits and are transmitted on the TCH/AFS or TCH/AHS speech traffic channels. SID (silence indicator) frames are the frames characterized by the SID (silence descriptor) gross bit patterns. It may convey information on the acoustic background noise. A VAD flag is a Boolean flag, generated by the VAD algorithm defined in GSM 06.94 indicating the presence ("1") or the absence ("0") of a speech frame.

The following are the frame formats available under full-rate channels (AFS) – SID_UPDATE, SID_FIRST, ONSET, SPEECH, and RATSCCH.

The following are the frame formats available under half-rate channels (AHS) – SID_UPDATE, SID_UPDATE_INH, SID_FIRST_P1, SID_FIRST_P2, SID_FIRST_INH, ONSET, SPEECH, RATSCCH_MARKER, and RATSCCH_DATA.

In a normal conversation, the speech is usually separated by pauses. The GSM provides the option of discontinuous transmission (DTX), whereby if there is no voice activity then only comfort noise parameters are transmitted. Discontinuous transmission (DTX) is a mechanism that allows the radio transmitter to be switched off most of the time during speech pauses, which helps to save power in the mobile station (MS) and to reduce the overall interference level over the air interface. Hence the transmission is stopped during this period by ramping the transmission power down to the minimum.

During the DTX mode, codec mode information, CMC and CMR, are transmitted along with comfort noise parameters at every 160 ms. Thus when DTX is enabled:

a. For SID_FIRST frames, the codec mode indication or codec mode command/request that is in phase with the alternating transmission will be transmitted (same phase as in speech frames).
b. Both the codec mode indication and codec mode command/request will be transmitted together in every SID_UPDATE frame (as in RATSCCH frames).
c. For ONSET frames the codec mode indication for the subsequent speech frame shall be transmitted, regardless of the phase of the in-band signaling. The general phase of the in-band signaling will not be changed by this.

Transmitter and Receiver Synchronization – The alternating transmission of CMI requires the synchronization of transmitting and receiving ends, so that the CMI and CMR/C are decoded in the correct order. To ensure appropriate synchronization, the CMI are transmitted aligned to the SACCH multiframe structure (26 TDMA frame) of the GSM.

For TCH/AHS, if the sub-channel is 0: The default transmission phase on the uplink should be such that the CMIs are aligned with the TDMA frame 0 (in the 26 multiframe), that is, CMIs are sent with speech frames having their first burst sent on TDMA frames 0, 8, 17 (modulo-26). The default transmission phase on the downlink should be such that the CMIs are aligned with TDMA frame 4 (in the 26 multiframe), that is, CMIs are sent with speech frames having their first burst sent on the TDMA frames 4, 13, 21 (modulo-26).

For TCH/AHS, if the sub-channel is 1: The default transmission phase on the uplink should be such that the CMIs are aligned with the TDMA frame 0 (in the 26 multiframe), that is, CMIs are sent with speech frames having their first burst sent on TDMA frames 1, 9, 18 (modulo-26).

The default transmission phase on the downlink should be such that the CMIs are aligned with TDMA frame 4 (in the 26 multiframe), that is, CMIs are sent with speech frames having their first burst sent on the TDMA frames 5, 14, 22 (modulo-26). These details are defined in 3GPP TS 5.02 [7].

Frame Detection in AFS

The following could be a generic sequence for frame detection in AFS.

1. Test the stealing flags (h_u and h_l) to determine if the frame is an FACCH, process the FACCH.
2. Perform the IM_1 (identification marker) detection to determine if a RATSCCH frame is present, if so, process the RATSCCH.
3. Perform the IM_0, location 1 detection to determine if a SID_FIRST frame is present, if so, process the SID_FIRST.
4. Perform the IM_0, location 0 detection to determine if a SID_UPDATE frame is present, if so, re-extract using $\times 4$ de-interleaving and process the SID_UPDATE.
5. Perform the in-band data detection, one for each of the four combinations of the 2 bits, to determine if a correlation peak exists and an ONSET frame is present, if so, process the ONSET frame.
6. If none of these frames have been detected, the frame is a speech frame. Detect the in-band data to recover the CMI, or CMC/CMR, and process the speech frame according to the current ACS and recovered CMI flag.

Frame Detection in AHS

The following could be a generic sequence for frame detection in AHS.

1. Test the stealing flags (h_u and h_l) to determine if the frame is an FACCH, process the FACCH.
2. Perform the IM_1 detection to determine if a RATSCCH frame is present, if so, process the RATSCCH.
3. Sum and convert to HD the $\sim$12 copies of the 9-bit marker ($!IM_0$ or IM_0) in the first two bursts of the four-burst block. Call this detected 9-bit sequence A.
4. Sum and convert to HD the $\sim$12 copies of the 9-bit marker ($!IM_0$ or IM_0) in the last two bursts of the four-burst block. Call this detected 9-bit sequence B.
5. Compute C = A and B.
6. If $C \sim = !IM_0$, the frame is an SID_UPDATE, re-extract using $\times 4$ block de-interleaving and process the SID_UPDATE frame.
7. If $C \sim = IM_0$, the frame is an SID_FIRST_P1/P2 combo. Process the SID_FIRST_P1 frame. Extract and process the SID_FIRST_P2 from the odd bits of the last two bursts.
8. If $C \sim = 0$, $A \sim = !IM_0$, and $B \sim = IM_0$, the frame is an SID_UPDATE/INH combo. Process the combo to recover in-band data channel 1 (CMI). The coded bits and in-band data channel 0 (CMC/CMR) are lost in this instance.
9. If $C \sim = 0$, $A \sim = IM_0$, and $B \sim = !IM_0$, the frame is an SID_FIRST_P1/INH combo. Process the combo to recover the in-band data (CMI, or CMC/CMR).
10. Perform the in-band data detection, one for each of the four combinations of the 2 bits, to determine if a correlation peak exists and an ONSET frame is present, if so, process the ONSET frame.
11. If none of these frames have been detected, the frame is a speech frame. Detect the in-band data to recover the CMI, or CMC/CMR, and process the speech frame according to the current ACS and recovered CMI flag.

WideBand (AMR-WB)

Wideband (AMR-WB) is a speech coding standard developed after the AMR using the same technology as for ACELP. This provides excellent speech quality due to a wider speech bandwidth of 50–7000 Hz compared with narrowband speech codecs, which in general are optimized for POTS wire-line quality of 300–3400 Hz. AMR-WB is codified as G.722.2, an ITU-T standard speech codec. The wide version of AMR supports codec rates of 6.6, 8.85, 12.65, 15.25, 15.85, 18.25, 19.85, 23.05, and 23.85 kbps.

8.5.2 Audio Codec

The audio codec uses a time domain to frequency domain transform to expose redundancy in the input signal. Contents are converted using various different compression algorithms, such as those from Microsoft, advanced streaming format (ASF), real audio (rm) or MPEG-1 audio layer-3 protocol (MP3), and so on. Some commonly used file formats are mentioned below, and of these only the MP3 file format is discussed in any detail [8]: WAV – waveform audio; MIDI – music instrument digital interface; AAC – advanced audio coding; ASF – advanced streaming format; MP3 – MPEG-1 audio layer-3 protocol.

8.5.2.1 MP3

The MP3 is a special format used to compress digital audio, keeping the audio quality as good as possible. Though this is a lossy compression technique, this loss can hardly be noticed because the compression method tries to control it. By using mathematical algorithms, it will only loose those parts of the sound that are hard to hear even in the original form. This way the audio can be compressed up to 12 times, which is really significant. MP3 encoding tools analyze the incoming source signal, break it down into mathematical patterns, and compare these patterns with psychoacoustic models stored in the encoder itself. The encoder can then discard most of the data that does not match the stored models, keeping that which matches. This configuration is based on a "tolerance" level. The lower the data storage allotment, the more data will be discarded, which leads to poorer audio sound quality.

MP3 encoded files are composed of a series of very short frames, one after another, much like a filmstrip. Each frame of data is preceded by a *header*. The header contains extra information about the compressed data frame. In some encodings, consecutive frames may hold information for each other. For example, if one frame has leftover storage space, whereas the next frame is running short of free space, then they may team up for optimal results. At the beginning or end of an MP3 file, extra information about the file itself, such as the name of the artist, the track title, the name of the album from which the track came, the recording year, genre, and personal comments may be stored. This is called "ID3" data.

The frame header is constituted by the very first four bytes (32 bits) in a frame. The first 11 bits of a frame header are always set and they are called the "frame sync." The exact meaning of each bit in the header id is defined in standard ISO/IEC 11172-3, see Table 8.3. Frames may have a 16-bits long CRC check just after the frame header. Next to this, the audio data are stored. We may calculate the length of the frame and use it if we also need to read other headers or just want to calculate the CRC of the frame, to compare it with the one we read from the file.

8.5.3 Image

Image compression can be lossy or lossless (lossless compression involves compressing data which, when decompressed, will be an exact replica of the original data). Various techniques, such as run-length encoding, entropy coding, and deflation are used for lossless compression and chroma sub-sampling, transform coding, and fractal compression are used for lossy compression. A common characteristic of

Table 8.3 MP3 header format

Sign	Length (bits)	Position (bits)	Description
A	11	(31–21)	Frame sync (all bits set)
B	2	(20,19)	MPEG audio version ID 00 – MPEG version 2.5 01 – reserved 10 – MPEG version 2 (ISO/IEC 13818-3) 11 – MPEG version 1 (ISO/IEC 11172-3)
C	2	(18,17)	Layer description 00 – reserved 01 – Layer-3 10 – Layer-2 11 – Layer-1
D	1	−16	Protection bit 0 – Protected by CRC (16 bit CRC follows header) 1 – Not protected
E	4	(15,12)	Bit-rate index- 8–448 kbps
F	2	(11,10)	Sampling rate frequency index- 8000–44100 (values are in Hz)
G	1	−9	Padding bit 0 – Frame is not padded 1 – Frame is padded with one extra slot Padding is used to fit the bit rates exactly. As an example: 128 k 44.1 kHz layer-2 uses a lot of the 418 bytes and some of the 417 bytes long frames to get the exact 128 k bit rate. For layer-1 the slot is 32-bits long, for layer-2 and layer-3 the slot is 8 bits long
H	1	−8	Private bit. It may be freely used for specific needs of an application, that is, if it has to trigger some application specific events
I	2	(7,6)	Channel mode 00 – Stereo 01 – Joint stereo (stereo) 10 – Dual channel (stereo) 11 – Single channel (mono)
J	2	(5,4)	Mode extension (only if joint stereo)

most images is that the neighboring pixels are correlated and therefore these contain redundant information. Therefore we need to find out the less correlated representation of the image using spatial redundancy (correlation between neighboring pixel values), spectral redundancy (correlation between different color planes or spectral bands), temporal redundancy (correlation between adjacent frames in a sequence of images) methods.

For still image compression, the "Joint Photographic Experts Group" or JPEG standard has been established by ISO (International Standards Organization) and the IEC (International Electro-Technical Commission) in 1992. Generally, in JPEG [9] the encoders and decoders are usually DCT-based. DCT can be computed with a fast Fourier transform (FFT) type of algorithm in $O(n \log n)$ operations. The JPEG standard specifies three modes, namely, sequential, progressive, and hierarchical for lossy encoding, and one mode of lossless encoding. "Baseline JPEG coder" uses sequential encoding. In Figure 8.14, the key processing steps in such an encoder and decoder are shown for gray-scale images, and for color image compression these can be approximately regarded as compression of the multiple gray-scale images.

The DCT-based encoder can be thought of as essentially a compression of a stream of 8×8 blocks of image samples. Each 8×8 block makes its way through each processing step, and yields output in a compressed form into the data stream. Because adjacent image pixels are highly correlated, the "forward" DCT (FDCT) processing step lays the foundation for achieving data compression by concentrating most of the signal in the lower spatial frequencies. After output from the FDCT, each of the 64 DCT coefficients is uniformly quantized in conjunction with a carefully designed 64-element quantization table (QT). A quantizer simply reduces the number of bits needed to store the transformed coefficients by reducing the precision of those values. An entropy encoder further compresses the quantized values losslessly to give a better overall compression. At the decoder, the quantized values are multiplied by the corresponding QT elements to recover the original unquantized values. After quantization, all of the quantized coefficients are ordered into the "zig-zag" sequence as shown in Figure 8.14. This ordering helps to facilitate entropy encoding by placing low-frequency non-zero coefficients before high-frequency coefficients. The dc coefficient, which contains a significant fraction of the total image energy, is differentially encoded. The JPEG proposal specifies both Huffman coding and arithmetic coding.

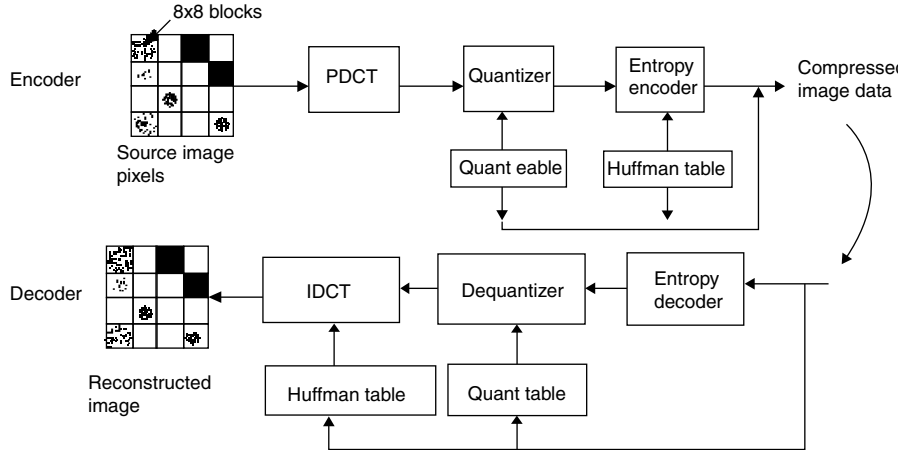


Figure 8.14 JPEG encoder and decoder blocks

The performance of these coders generally degrades at low bit-rates mainly because of the underlying block-based discrete cosine transform (DCT) scheme. More recently, the wavelet transform has emerged as a cutting edge technology, within the field of image compression.

8.5.4 Video

Video compression is a combination of image compression and motion compensation. Video is basically a three-dimensional array of color pixels, where two dimensions represent the spatial (horizontal and

vertical) directions of the moving pictures and the third dimension represents the time domain. A data frame is a set of all pixels that correspond to a single time moment. Basically, a frame is same as a still picture. As for JPEG images, the spatial encoding is performed by taking advantage of the fact that the human eye is unable to distinguish small differences in color as easily as it can changes in brightness, so very similar areas of color can be "averaged out." With temporal compression, only the changes from one frame to the next are encoded, as, often a large number of the pixels will be the same on a series of frames. The steps which are commonly followed for encoding are signal analysis are quantization and variable length encoding. There are four methods of compression: discrete cosine transform (DCT), vector quantization (VQ), fractal compression, and discrete wavelet transform (DWT).

In 1993 the Motion Picture Expert Group (MPEG) was founded; this originally focused on producing non-interactive video compression but, later was extended as MPEG-4 and MPEG-5.

- **MPEG-1** – CD-ROM storage compression standard. Designed for bit rate up to 1.5 Mbps.
- **MPEG-2** – DVB and DVD compression standard. Designed for bit rate between 1.5 and 15 Mbps.
- **MPEG-3** – MPEG-2 layer-3 audio streaming standard.
- **MPEG-4** – Audio and video streaming and complex media manipulation.
- **MPEG-5** – Multimedia hypermedia standard.

The MPEG-4 standard was created to be the next major standard in the world of multimedia. Unlike MPEG-1 and MPEG-2, where more focus was given to better compression efficiency, in the case of MPEG-4 the emphasis was on new functionality. The new MPEG-4 standard facilitates the growing interaction and the convergence of the previously separate worlds of telecommunications, computing, and mass media.

MPEG-4 runs on the MP4 file format. It is the next generation beyond MP3. AS with MP3, MPEG-4 will become the accepted standard because it extends the success of MP3 in several important ways.

a. MPEG-4 enables video, even at bit rates as low as 9.6 kbps.
b. MPEG-4 enables digital rights management to protect the precious intellectual property of the content provider.
c. The MPEG-4 solution provides mobile users access to full-motion news and financial stories, sports highlights, short entertainment clips and music videos, weather and traffic reports, home or work security cameras and corporate communications, from any location.

Some other commonly used video standards are-

- **H.261** – This is an ITU standard designed for two-way communication over ISDN lines (video conferencing) and supports data rates that are multiples of 64 kbps [10, 11]. The algorithm is based on DCT and can be implemented in hardware or software and uses intra-frame and inter-frame compression. H.261 supports CIF and QCIF resolutions.
- **H.263** – This is based on H.261 with enhancements that improve video quality over modems. It supports CIF, QCIF, SQCIF, 4CIF, and 16CIF resolutions.

References

[1] GSM Technical Specification (2005) 3GPP TS 05.08 (v8.23.0). *Radio Subsystem Link Control*. ETSI TC-SMG, Sophia-Antipolis Cedex.
[2] Lindberg, B. (1994) μ-law, A-law and Linear PCM Conversions, December 30, 1994, Center for PersonKommunikation, Aalborg University.

[3] Vary, P., Hellwig, K., Hofmann, R., and Sluyter, R., Speech codec for European Mobile Radio System. Paper presented at the Proceedings ICASSP (227-30), April 1988.
[4] GSM Technical Specification (1999) GSM 05.03 (ETS 300 575). *European Digital Cellular Telecommunications System (Phase 2); Channel Coding*. ETSI TC-SMG, Sophia-Antipolis Cedex.
[5] GSM Technical Specification (2005) 3GPP TS 05.09 (v8.6.0), (2005-06). *Link Adaptation (www.3gpp.org)*. ETSI TC-SMG, Sophia-Antipolis Cedex.
[6] GSM Technical Specification (2005) 3GPP TS 05.05 (v8.82.0), (2005-11). *Radio Transmission Reception (http://www.3gpp.org/ftp/Specs/2006-12/R1999/05_series/)*. ETSI TC-SMG, Sophia-Antipolis Cedex.
[7] GSM Technical Specification (2003) 3GPP TS 05.02 (v8.11.0), (2003-06). *Multiplexing and Multiple Access on the Radio Path*. ETSI TC-SMG, Sophia-Antipolis Cedex.
[8] ITU-T and ISO/IEC JTC 1 (1994) ITU-T Rec. H.262 and ISO/IEC 13818-2 (MPEG-2). *Generic Coding of Moving Pictures and Associated Audio Information – Part 2: Video (Nov. 1994, with several subsequent amendments and corrigenda)*. International Telecommunications Union, Geneva.
[9] ISO/IEC (2002) Intl. Std. 15444. *Information Technology – JPEG 2000 Image Coding System, Part 3: Motion JPEG 2000 (Sept. 2002, with subsequent amendments), MPEG-1, MPEG-2, MPEG-4 ASP, MPEG-4/AVC*. International Organization for Standards, Geneva.
[10] ITU-T (2000) ITU-T Rec. H.263; v1: Nov. 1995, v2: Jan. 1998, v3: Nov. 2000. *Video coding for low bit rate communication*. International Telecommunications Union, Geneva.
[11] ITU-T (1990/1993) ITU-T Rec. H.261 v1: Nov 1990, v2: Mar. 1993. *Video codec for audiovisual services at px64 kbits/s*. International Telecommunications Union, Geneva.

9

GSM Mobile Phone Operations and Procedures

9.1 Initial Procedures after Mobile Power ON

When the mobile is switched ON, its first task is to find a suitable BTS through which it can gain access to the network. As discussed earlier, all BTSs broadcast their allocated BCCH carrier [various neighboring BTSs are allocated different radio beacon frequencies to transmit according to frequency planning and each BTS has a single broadcast (BCCH) frequency] in the respective cell. Again, in the same area there may be one or more BTSs installed by other cellular operators, and they also have different radio beacon frequencies. The mobile starts searching the relevant frequency band for BCCH carriers. This process is called cell selection. Generally, the upper layer of the protocol stack commands layer-1 to program the RF module to measure the RSSI (received signal strength indication – which is usually measured as the square root of I^2 and Q^2) for different carrier frequencies, and then once it is performed, layer-1 indicates the result to the upper layer. On the basis of this result the upper layer decides on which carrier frequency it should search for the FB and SB burst (generally it selects the carrier frequency on the basis of the highest RSSI value).

9.1.1 Cell Selection

The mobile may implement the cell selection algorithm by one of two different search algorithms depending on its knowledge of the BCCH carriers in use.

The first algorithm is applied when the mobile has no knowledge of the BCCH carriers deployed in a particular PLMN. Initially, the mobile scans through the entire downlink frequency band (for example, 124 downlink carriers for GSM-900, etc.) and measures the received signal strength of each carrier. The received signal strength for each carrier is determined from the average of at least five measurements spread evenly over a time period of 3–5 s. Once the frequency scanning is done and the list is prepared by ordering the carriers according to the received signal's strength, then this list is reported to the higher layers. Next, the higher layer commands layer-1 to tune the RF to the strongest carrier in the list and waits for the frequency correction (FB) burst, which is a burst of pure sine wave, for example, it contains data samples of a pure sine wave of frequency 67.7 kHz. If an FCCH burst, which occurs every 10 (or 11 in the case when one idle slot comes in the 51 multiframe) frame intervals on a time slot of zero of a BCCH carrier (time slot $= 0, f =$ broadcast frequency), is not detected (because it may not be a BCCH frequency,

Mobile Handset Design Sajal Kumar Das

for example, it may be a TCH carrier frequency), then the mobile retunes to the next strongest carrier in the list and repeats the same process. Once the mobile finds the FCCH burst on a specific carrier frequency of a particular BTS, then it tries to decode the SCH information, which appears just next to the FCCH (on the same physical channel, for example, TS-0 and the BCCH frequency), for example, in the next TDMA frame's slot zero location it could be eight slots latter on the same frequency.

The second cell selection algorithm exists for situations where the MS has the prior knowledge of the BCCH carriers used within the network. This may be because the mobile was switched OFF and ON again after some time and during the switch OFF, the necessary data were stored in the SIM. In this case, the mobile will first search only the carriers stored in its BCCH carrier list and apply the cell selection algorithm. If it finds that the FB and SB are not detected on these stored carriers, for example, if none of the stored BCCH carriers are now valid, then the mobile will revert back to the full cell search algorithm as described earlier. This situation may occur when the mobile is switched OFF, and switched ON again in some new location area, or if the mobile battery is unplugged without switching the mobile off properly (because if the mobile is switched off correctly by pressing the power down button, then the necessary relevant information is stored in the SIM for later usage).

This procedure is known as the *initial synchronization* procedure and the following steps are usually performed sequentially:

1. **Mobile switch on** – Power on the mobile system.
2. **Boot on** – Mobile performs initial boot-on procedure, hardware initialization, OS load, self test, and so on.
3. **Received power measurement for various carrier frequencies in different bands** – Next, the higher layer commands layer-1 for RSSI measurement over the entire frequency bands or selected frequency bands for the various carrier frequencies. Thus, L2 provides an RSSI scan message indicating the frequency band as an input parameter to L1.
4. **RF programming for signal strength measurement on different carriers** – Layer-1 programs the RF modules for this measurement and once the measurement is done, then it reports to the upper layer on the result of the RSSI values for different carrier frequencies in a band with an ordered list of carriers from highest to lowest received signal strength.
5. **FB search** – After receiving the signal strength of different carriers in various bands, the upper layer decides which carrier should be looked at first, based on the data stored in the SIM, or based on other rules. From the prioritized list of allowed PLMNs in the SIM, the MS selects the one with the highest priority available. If the home network is available it is selected; otherwise the MS selects a foreign PLMN. Then it commands L1 to search for the frequency correction burst (FB in the FCCH channel) providing the carrier frequency (ARFCN) number. The FB contains all-zero and after GMSK modulation it creates a pure sine wave (PSW), as discussed in Chapter 7. So, using a specific algorithm, the layer-1 program tries to search the pure sine wave (FB) in the received samples from the RF module. The FCCH channel does not require any decoding process (at this stage the receiver is not frequency or time synchronized), as the sine wave data pattern can easily be found from the received samples. Once the sine wave is detected, from the peak position of the sine wave data in a time slot, the relative frequency offset between the MS and this BTS is found and applied to the AFC module for local frequency correction (Figure 9.1). Then the MS is frequency synchronized. Generally, layer-1 intimates to the upper layer whether the FB is detected on that carrier frequency or not, and if it is detected then what is the approximate timing offset value (slot beginning) and frequency error. For FB reception, the MS usually opens the RF reception window for at least nine continuous time slots. This is because, if that frequency is a broadcast frequency, then a complete FB slot data will be inside the collected nine frames sampled data (as FB repeats after nine frames in 51 signaling multi-frame structure). If FB is not detected on that carrier (as it may be the TCH frequency), then the upper layer commands layer-1 to do an FB search in the next highest RSSI carrier frequency in the list, and this process continues until FB detection is successful.

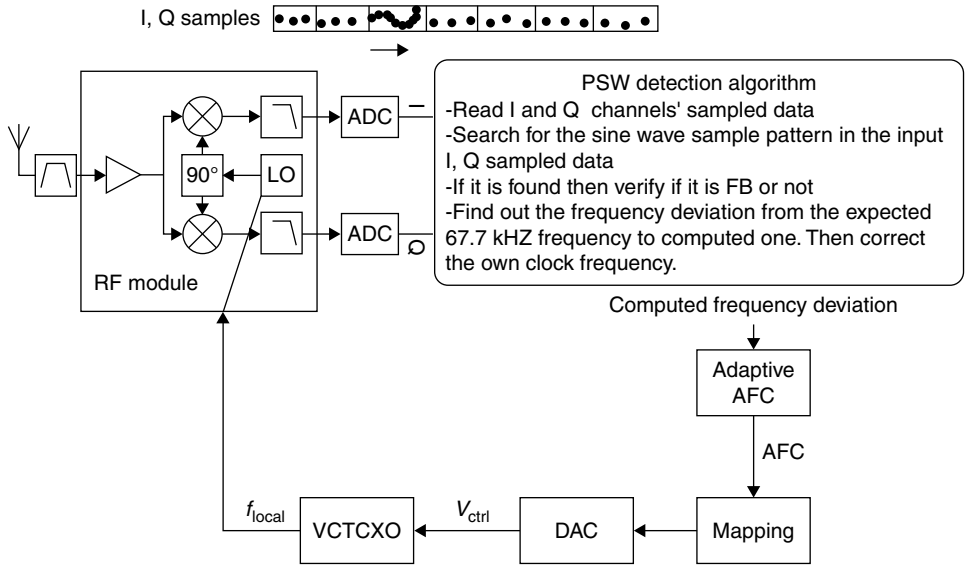


Figure 9.1 FB detection inside a GSM mobile phone

6. **SB decoding and own BTS identification** – Once the FB is detected, the upper layer commands layer-1 for an SB search (for time or slot synchronization). This is the first channel after power on that is required to be decoded. In the broadcast channel (combination of broadcast frequency and time slot-0), SB follows next to FB. Now the location of the FB is already known, so the SB location can be approximately judged (after 1 TDMA frame). Collect the received samples on that window from RF and search for SB. SB contains a 64-bit training sequence (this SB training sequence is the same for all BTSs in the network) placed in the middle. Autocorrelation of the received SB burst data with the known SB burst training sequence helps to find the peak. As SB uses a long training sequence, thus it provides a long auto-correlation peak. From this timing correction and channel transfer function estimation, equalization and decoding is performed. Once the SB is decoded and the information content of the SB [=19 bits reduced frame number and 6-bit BISC number (19 + 6 = 25 bits) and after channel coding it is 39 bits] is detected and then passed to higher layer. The BSIC (base station identity code) consists of a 3-bits base station color code and 3-bits network color code used for BS identification and network identification. A reduced frame number is used to identify the slot location of the MS in the entire hyper frame slot structure for time synchronization. At this point one thing needs to be remembered – once the SB is decoded successfully then the BSC code is known, so the intended own base station is known. Each base station uses a specific training sequence for normal burst sending (used for other channels including BCCH, PCH, TCH). At this stage, as the base station is known so the training sequence used for NB transmission by the intended BTS is also known. The MS is now ready for any NB burst decoding.
7. **Read system information (SI)** – The upper layer checks whether the intended BTS (that is just found) is allowed to camp on or not. If it is, then it commands layer-1 for reading the BCCH information (which is passed in the BCCH channel and uses normal burst) to know the system specific information.
8. **Paging group identification and enter idle mode** – Once the BCCH information is read and passed to the upper layer, the MS then sets the values accordingly and decides the paging group and enters into the idle mode. In this mode, it wakes up periodically and listens to the paging and also does the cell re-selection as required.

9.1.1.1 Measurements for Normal Cell Selection

Measurements for normal cell selection are performed by an MS, which has no prior knowledge of the BCCH carriers in GSM or DCS 1800 or PCS 1900 RF bands. In which case, the MS searches all RF channels in the system (124 for P-GSM, 174 for E-GSM and 374 for DCS 1800), take readings of received RF signal strength on each RF channel, and calculates the received level average for each. The averaging is based on at least five measurement samples per RF carrier spread over 3–5 s, the measurement samples from the different RF carriers being spread evenly during this period. A multi-band MS will search all channels within its bands of operation as specified above. The number of channels searched will be the sum of the channels on each band of operation. BCCH carriers can be identified by searching for frequency correction bursts (FB). Once FB and SB are found in a beacon carrier (BCCH carrier), the MS should next attempt to synchronize it with BTS and then read the BCCH data. Generally, the maximum time allowed for synchronization to a BCCH carrier is 0.5 s, and the maximum time allowed to read the BCCH data, when being synchronized to a BCCH carrier, is 1.9 s.

9.1.2 Synchronization

As we know, GSM is a TDMA and FDMA based system, so system frequency as well as time synchronization is vital for correct operation of the system. The FCCH burst is used by the MS (for frequency correction) to correct its internal time base to ensure that its carrier frequency is accurate and frequency deviation is within 0.1 ppm compared with the signal received from the BTS. The MS employs its internal time base to generate both the local versions of the RF carriers for demodulation, and the clock signals for its internal counters and baseband operations.

After the relevant frequency correction, the mobile attempts to decode the synchronization burst contained in the SCH time slot (remember the SB burst is the first burst to be decoded by the MS after power ON, as the FB is not required to be decoded – it contains all zeros and after modulation it is a pure sine wave). The SB is easily located, as it always follows immediately after the FCCH time slot on the same physical channel, for example, eight time slots later. Sufficient information is contained in the SB for the mobile to identify its frame counter position in the complete GSM frame structure. As we have seen earlier, the burst contains 25 bits of information prior to channel coding, and of these 6 bits are used to transmit the base station identity code (BSIC) and the remaining 19 bits are used to transmit the reduced TDMA frame number (RFN) of the time slot containing the SB. The RFN consists of three parameters – T1 (11 bits), T2 (5 bits), and T3′ (3 bits), which are determined using the full frame number (FN) unique to each TDMA frame within the hyper frame. FN ranges from 0 to $(2048 \times 51 \times 26) - 1 = 2\,715\,647$ and RFN parameters are defined as follows:

$T1 = FN \text{ div } (26 \times 51)$ – this is 11 bits and ranges from 0 to 2047.
$T2 = FN \bmod 26$ – this is 5 bits and ranges from 0 to 25.
$T3' = (T3 - 1) \text{ div } 10$ – this is 3 bits and ranges from 0 to 4.
T3 is a number in the range from 0 to 50.
$T3 = FN \bmod 51$.

In the above equations, the mod and div operators return the integer result and the remainder of an integer division, respectively. It is evident from the above relationships that T1 provides the position of the super frame containing the SB within the hyper frame and T2 provides the position of the multiframe in the super frame. In the control channel multiframe structure there are 51 TDMA frames and the position of the frame containing the SB within the multiframe is given by T3, which requires 6 bits. However, a synchronization burst can only occupy one of the five different positions within the multiframe structure (it repeats on every 10 TDMA frame, $51/10 \sim 5$) and consequently this information is transmitted using three bits as T3′.

Apart from the frame number (FN), the mobile must also maintain the counters for the time slot number (TN) and the quarter bit number (QN). The QN counts the quarter bit periods and its value ranges from 0 to 624 ($156.25 \times 4 = 625$). The quarter bit number counter is incremented every 12–13 μs, and this is set using an extended training sequence located in the middle of the SB. Every time SB is received, the TN counter is set to zero and is incremented each time the QN count changes from 624 to 0. The TN count is used to hold the position of the time slot within the TDMA frame and its value ranges from 0 to 7. FN is incremented when the value of TN changes from 7 to 0.

After successfully synchronized with respect to frequency and time with the own BTS, the mobile may proceed to decode the system information (SIB) contained on the BCCH. The BCCH is easily located as, it always occupies the same position within the 51-TDMA frame control channel multiframe. This channel's information contains several parameters that influence the cell selection, maximum allowable mobile transmitted power, minimum received power level at MS, and so on.

9.1.3 Flow Diagram of Initial Mobile Acquisition

The flow diagram of initial mobile acquisition is shown in Figure 9.2.

Once a mobile is camped into a network, then it operates in two modes based on the usage: (1) idle mode or (2) access mode (dedicated mode).

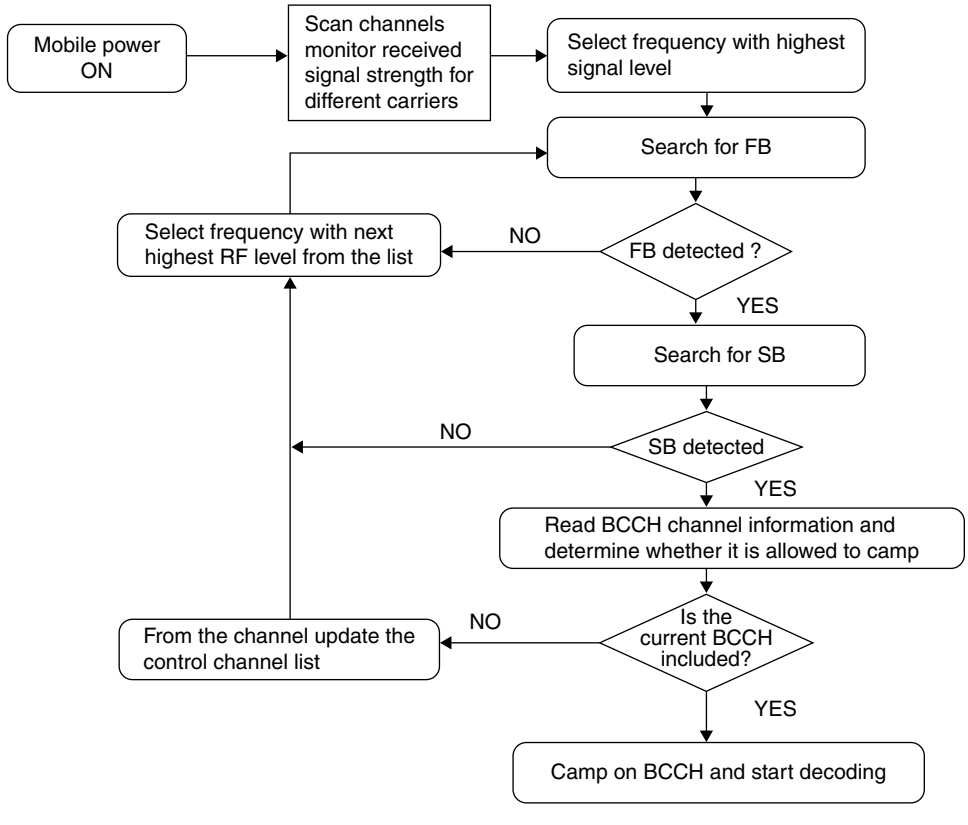


Figure 9.2 Initial cell selection flow diagram

9.2 Idle Mode

After the appropriate cell is selected and the mobile is camped to the BTS, then the MS enters the idle mode, where it must monitor the BTS paging channel (for any incoming call), and runs a procedure periodically to check whether it has camped to the most appropriate cell or not (for example, the cell with the highest signal strength and quality) and this procedure is called cell re-selection.

The idle mode includes the reception of BCCH and CCCH, transmission of RACH, cell re-selection, and measurements. Measurements can be made on any idle frame except on:

- PCH, PPCH frames (these frames may be used by the neighbor monitoring task).
- FB, SB, CBCH, neighboring BCCH, serving PBCCH (these bursts may be situated anywhere in the frame and in order to simplify the software no other radio window is allowed).

The idle mode is normally exited when the layer-1 is configured for the SDCCH or TCH by the upper layer to switch to a dedicated mode.

9.2.1 Paging and Discontinuous Reception (DRX)

MSC/VLR initiate the paging message (CS paging) to alert any MS about the incoming call (to establish the RR connection) using the paging channel (PCH), which is part of the downlink CCCH. An MS in idle mode must continuously monitor the relevant PCH for paging calls containing its unique address. The GSM system also supports a slotted paging mode, whereby the PCH is divided into a number of paging sub-blocks and the MS is required to listen to the channel during its assigned paging sub-block. The MS may be powered down during the periods it is not monitoring the PCH and this helps to prolong the battery lifecycle of the mobile, this technique is known as discontinuous reception (DRX). Thus, in order to reduce power consumption, paging groups are defined and an MS listens to paging sub-channels corresponding to its paging group. In order to operate in DRX mode on CCCH, the mobile needs to calculate the CCCH_GROUP and paging group for CCCH. In order to calculate the CCCH_GROUP and the paging group, the mobile requires details about the control channel information provided by the network in System Information-3 on the BCCH channel.

The following parameters are provided. (1) **CCCH_CONF** = 0, indicates SDCCH and BCCH/CCCH are not combined, 1 indicates SDCCH and BCCH/CCCH are combined, 4 indicates two CCCH on the BCCH frequency, 6 indicates three CCCH on the BCCH frequency, 7 indicates four CCCH on the BCCH frequency. The number of CCCH timeslots on the BCCH frequency is referred as BS_CC_CHANS. (2) **BS_AG_BLKS_RES** indicates the number of blocks reserved for the AGCH. This parameter indicates the number of blocks on each CCCH reserved for AGCH. Its value remains between 0 and 7. 0, which implies no channel is reserved for AGCH and 7 implies 7 are reserved out of a total of 9. (3) **BS_PA_MFRMS-** (=2–9) this is used for computation of the paging group. It defines the number of 51-frame multiframes between the transmission paging messages to MS belonging to the same CCCH_GOUP. MS calculates CCCH_GROUP and paging group using the following equations: *CCCH_GROUP* (0. . .(BS_CC_CHANS-1)) = (IMSI mod 1000) mod (BS_CC_CHANS * N)) div N.

$$Paging\ group = ((\text{IMSI} \bmod 1000) \bmod (\text{BS\_CC\_CHANS} * \text{N}) \bmod \text{N}$$

The message sent on PCH will include the IMSI of the MS paged. Once the MS recognizes its own IMSI, it will answer the paging call by entering the access mode. Unanswered paging messages will be repeated to overcome channel variation; however, the exact reception policy is implementation specific for the network operator.

9.2.2 Cell Re-Selection

While MS is in idle mode, the MS must continue to monitor the downlink signal strength of the neighboring BTSs to ensure that it is always camped to the most appropriate BTS. The MS should monitor the received signal strength of the BCCH carriers from the serving BTS, in addition to six other neighboring BTSs and make a list. Owing to environmental changes or the MS movement, the received signal strength may vary from time to time. Becasue of this, the MS is required to apply the cell selection algorithm to identify whether the idle mode handover is required or not at an interval of 5 s or so. If the new target cell (whose signal strength is found to be relatively stronger) is in the same location area, then the MS may simply switch to the new target cell and starts decoding the PCH data. However, if the new cell belongs to a different location area, then in that case the MS must perform a location update procedure first and then begin to decode the PCH data.

9.2.3 PLMN Selection

The goal of this PLMN selection is to search for available PLMN between the 30 (or more) strongest carriers. This process is executed in idle mode, when the MS looses any radio link or when the MS user requires a PLMN re-search. There are two mandatory modes for PLMN selection: automatic mode and manual mode. The MS performs registration on the PLMN, if the MS is capable of services, that require registration. MS selects and attempts to perform a location registration on the registered PLMN (if it exists) at switch-on or on recovery from a lack of coverage area (or where necessary). If successful registration is achieved, the MS indicates the selected PLMN in the display. Where there is no registered PLMN, or if registration is not possible due to PLMN being unavailable or registration failure, the MS follows one of the following two procedures depending on its operating mode.

1. **Automatic Network Selection Mode Procedure:**
   The MS selects and attempts registration on other PLMNs (if available and allowable), in all of its bands of operation in the following order:
   a. HPLMN (if not previously selected);
   b. each PLMN in the "PLMN selector" data field in the SIM (in priority order);
   c. other PLMNs with received signal level above −85 dBm in random order;
   d. all other PLMNs in order of decreasing signal strength.
2. **Manual Network Selection Procedure:**
   The MS indicates, whether there are any PLMNs (including "forbidden PLMNs"), in all of its bands of operation that are available. The user may select their desired PLMN and the MS then initiates registration on that PLMN.

Next, let us have a closer look at the location update procedure.

9.3 Location Updating

The MS is informed about any incoming call by a paging message sent over the PCCH channel of a cell. If the location of the MS is not known, then the question is in which cells will the message be paged? One solution would be to page in every cell in the network for each call, which is obviously wastage of the radio bandwidth. The other solution would be that the mobile will notify the system about its current location at the individual cell level via location update messages. Then paging messages will be sent exactly to one cell, but this would also be very wasteful due to the large number of location updating messages from the MS. This is why a compromise solution is used. In GSM cells are grouped into *location areas*. Updating messages are required when moving between location areas. MSs are

frequently paged in the cells of their current location area. The location updating procedures, and subsequent call routing, use the MSC and two location registers: HLR and VLR. When a mobile station is switched on in a new location area, or it moves to a new location area or different operator's PLMN, it must register with the network to indicate its current location. In the normal case, a location update message is sent to the new MSC/VLR, which records the location area information, and then sends the location information to the subscriber's HLR. The information sent to the HLR is normally the SS7 address of the new VLR, although it may be a routing number. A routing number is not normally assigned (even though it would reduce signaling). The reason is that there is only a limited number of routing numbers available in the new MSC/VLR and they are allocated on demand for incoming calls. If the subscriber is entitled to service, the HLR sends a subset of the subscriber information, needed for call control, to the new MSC/VLR, and sends a message to the old MSC/VLR to cancel the old registration.

GSM also has a periodic location updating procedure due to reliability reasons. If an HLR or MSC/VLR fails, then to have each mobile register simultaneously to bring the database up to date would cause overloading. Therefore, the database is updated as location updating events occur. The enabling of periodic updating, and the time period between periodic updates, is controlled by the operator, and is a trade-off between signaling traffic and speed of recovery. If a mobile does not register after the updating time period, it is de-registered. The message flow for location update procedure is shown in Figure 9.3.

A procedure related to location updating is the IMSI attach and detach. A detach lets the network know that the mobile station is unreachable and avoids unnecessary allocation of channels and send of paging messages. An attach is similar to a location update, and informs the system that the mobile is reachable again. The activation of IMSI attach/detach is up to the operator on an individual cell basis. As the radio medium is a public medium, it can be accessed by anyone, which is why, authentication of users is very important in a mobile network and this is performed after the registration (or any other time). This is described in detail in the next section.

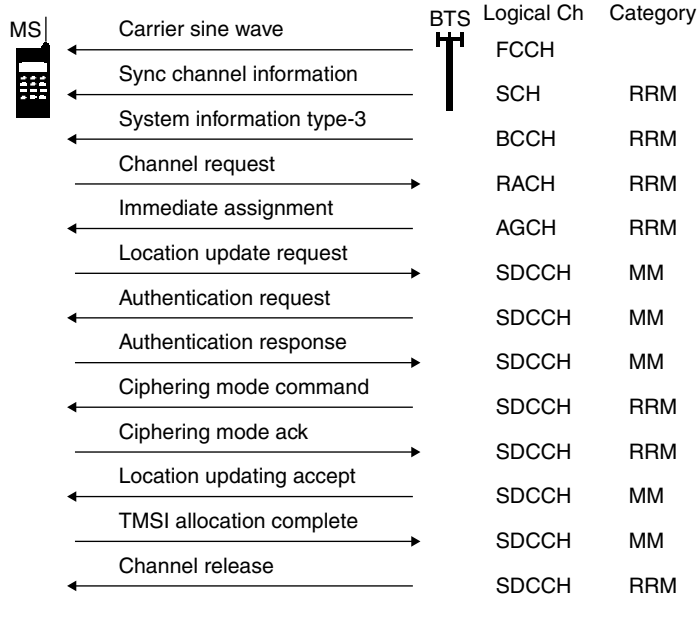


Figure 9.3 Location update procedure message flow

9.4 Security Procedure

All cellular communications use air as the channel to send or receive information. Air is not a private channel, so a wireless network is less secure than a wired network and it opens the door to eavesdroppers who have an appropriate receiver.

First generation analog cellular phones did not usually contain much in terms of security aspects and protection. Hence, it was possible to eavesdrop on the analog radio path and thereby listen to other user's calls, or to program the identities of the mobile phones such that the accessing cost appears on another user's bill.

Moving to second generation systems, the GSM security was designed keeping 1G threat scenarios in mind in order to achieve two primary goals: firstly, protecting the network against unauthorized access, and secondly, protecting the privacy of the users. Thus some of the security aspects were introduced to provide user related security features for authentication, confidentiality, and anonymity and protecting the network against un-authorized access. These features were designed to afford both the subscriber and the network operator a greater level of protection against fraudulent activities.

Several security functions were built into GSM to safeguard subscriber privacy. These include:

- authentication of the registered subscribers only;
- secure data transfer through the use of encryption;
- subscriber identity protection;
- mobile phones are inoperable without an SIM;
- duplicate SIMs are not allowed on the network;
- securely stored K_i in the SIM.

GSM security features provide PIN code protection, authentication, confidentiality and anonymity, and confidentiality of voice and data by appropriate encryption and decryption.

9.4.1 PIN Code Protection

Access to the SIM card is protected by using the personal identification number (PIN) code. In the SIM, the PIN takes a 4–8 decimal digit code. Once the right PIN is entered, then only ME will have the access to the data stored inside the SIM card. Generally, after three consecutive incorrect PIN entries, the SIM will be blocked. The SIM may be unblocked by entering a further eight digit code known as the PIN unblocking key (PUK), which is also stored in the SIM. After ten incorrect attempts to enter the PUK, the unblocking key itself becomes blocked and then there is no way to unblock the SIM. The user has the option of disabling the level of PIN protection or a second PIN2 code can be stored.

9.4.2 Anonymity

When a new GSM subscriber turns the phone on for the first time, its IMSI is transmitted to the AuC on the network to intimate its identity. As each IMSI is associated with a unique user (SIM), if the IMSI is known, an eavesdropper can determine the location of a subscriber by intercepting the message. This problem is reduced by introducing a temporary mobile subscriber identity (TMSI) number. After the initial registration, a TMSI is assigned to the subscriber. The IMSI is rarely transmitted unless it is absolutely necessary. This prevents a potential eavesdropper from identifying a GSM user by reading their IMSI. The user continues to use the same TMSI, depending on how often location updates occur. Every time a location update occurs, the network assigns a new TMSI to the mobile phone. The TMSI is stored along with the IMSI in the network HLR. The MS uses the TMSI when reports to the network or during call initiation. Similarly, the network uses the TMSI to communicate with the mobile station. The visitor location register

(VLR) performs the assignment, the administration, and the update of the TMSI. When, the mobile is switched off, the MS stores the TMSI in the SIM card, so that it is available when it is switched on again.

As shown in Figure 9.4, initially of course the phone will have no TMSI, and thus is addressed by its IMSI. Once ciphering has been successfully completed, then the initial TMSI is allocated. The VLR controlling the LA in which the TMSI is valid maintains a mapping between the TMSI and IMSI. If the MS moves into a new VLR area, the new VLR can ask the old VLR to whom the TMSI (which is not valid in the new VLR) belonged to.

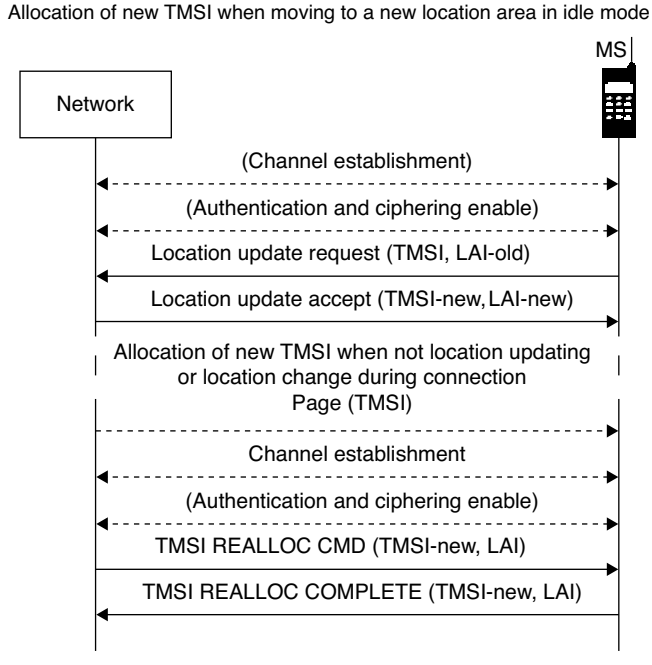


Figure 9.4 Allocation of TMSI

9.4.3 Authentication

The authentication procedure checks the validity of the subscriber's SIM card and then decides whether the mobile station is allowed on that particular network or not. The network authenticates the subscriber through the use of a challenge-response method as shown in Figure 9.5.

The authentication procedure is triggered when the MS attempts one of the following:

1. on the first access to the network;
2. accessing the network for the purpose of making or receiving a call;
3. location update process and the change of subscriber-related information stored in either HLR or VLR.

As shown in Figure 9.6, the authentication is initiated by the network by sending an authentication request message to the MS. This message contains RAND, which is a 128-bit random number. The A3 is the authentication algorithm used in the GSM system. Both the A3 algorithm and subscriber authentication key (K_i, is unique to the subscriber) are stored in the SIM. When the subscriber is

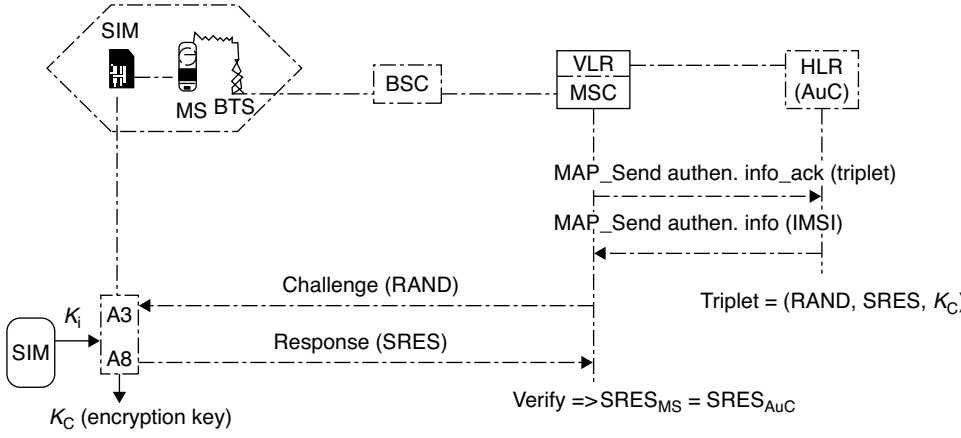


Figure 9.5 GSM authentication and key agreement

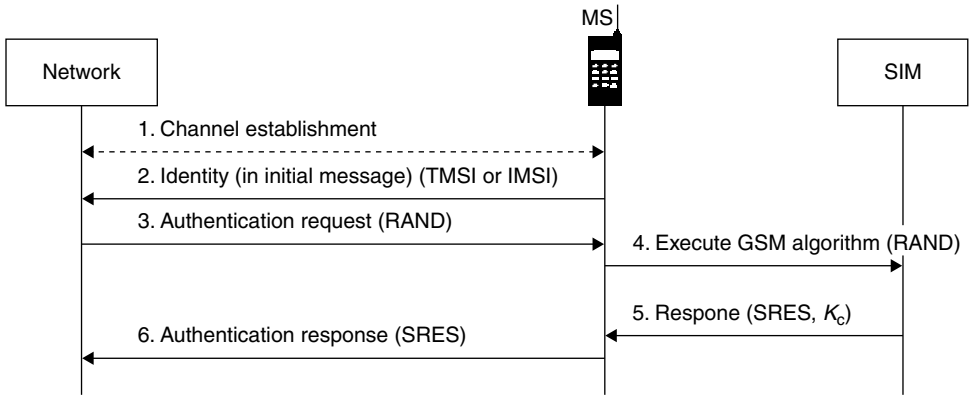


Figure 9.6 Message flow during authentication

added to the home network for the first time, a subscriber authentication key (K_i) is assigned in addition to the IMSI to enable the verification of the subscriber identity in the SIM. The RAND and K_i are the input to the A3 algorithm and it generates a 32-bits length output, known as SRES (signed RESult). This is returned back to the network in the form of an authentication response message. On the network side, the authentication center (AuC) also stores the user's key (K_i – its length is operator dependent and the maximum key length is 128 bits) and the A3 algorithm, and it also generates the SRES. If the computed SRES value matches with the SRES value sent by the MS, only then is authentication successful and the subscriber joins the network.

Authentication Failure If authentication fails the first time, and the TMSI was used, the network may choose to repeat the authentication with the IMSI. If that also fails, then the network releases the radio connection and the mobile should consider that particular SIM to be invalid (until switch-off or the SIM is re-inserted).

9.4.3.1 A3 Algorithm

The A3 algorithm does not refer to a particular algorithm, rather the algorithm the operator has chosen to be implemented for authentication. The most common implementations for A3 are COMP128v1 and COMP128v2. In fact, both of these algorithms perform the function of both A3 and A8 (the ciphering key generation algorithm – discussed later) in the same stage.

Whenever the SIM is asked to compute the SRES (with the RUN GSM ALGORITHM command) it also computes a new K_c (ciphering key – discussed later). Thus the authentication procedure is not only used to verify a user, but it is also used whenever the network wishes to change the keys.

9.4.4 Encryption and Decryption

Once the user is authenticated, the RAND (delivered from the network) together with the K_i (from the SIM) is sent through the A8 ciphering key generating algorithm, to produce a ciphering key (K_c – 64-bits long). The A8 algorithm is also stored in the SIM card. The K_c (generated by A8 algorithm) is then used by the A5 ciphering algorithm to encipher or decipher the data. The A5 algorithm is implemented in the hardware of the mobile phone, as it has to encrypt and decrypt data during transmission and reception of information, which must be fast enough. The A5 algorithm takes the 64-bit long K_c key and a 22-bit long representation of the TDMA frame number and produces two 114-bit long encryption words, BLOCK1, BLOCK2, for use on the uplink and downlink, respectively.

The encryption words are EXORed with the 114 data bits in each burst. Because the encrypted data are computed using the TDMA frame number, the words change from burst to burst and are not repeated over the hyperframe cycle (around 3.5 h). This is summarized in Figure 9.7.

Anonymity is used to identify the users, and authentication is used for billing purposes, and signal and data encryption are used for signal and data protection.

9.4.4.1 Ciphering Algorithms

As mentioned above, the network can choose one from up to seven different ciphering algorithms (or no ciphering). However, it must select an algorithm that the phone can support. Currently there are three algorithms defined – A5/1, A5/2, and A5/3. A5/1 and A5/2 were the original algorithms defined by the GSM standard and are based on simple clock controlled LFSRs. A5/2 was a deliberate weakening of the algorithm for certain export regions, where A5/1 is used in countries such as the USA, UK and Australia.

A5/3 was added in 2002 and is based on the open Kasumi algorithm defined by 3GPP.

9.4.5 Weaknesses of GSM Security

Although the basic objective of security has been achieved, there are some weaknesses in GSM security, such as:

1. In the first stage of the authentication process, when serving VLR/SGSN requests security data (triplets) from the HLR, HLR computes up to five triplets and forwards them to the requesting node. This is executed over SS7 based mobile application part (MAP) protocol. The MAP protocol had no security mechanism and consequently the sensitive data in the triplet (RAND, SRES, K_c) is transmitted clearly from HLR to VLR/SGSN.
2. Only the network can initiate the authentication and it is an optional procedure.

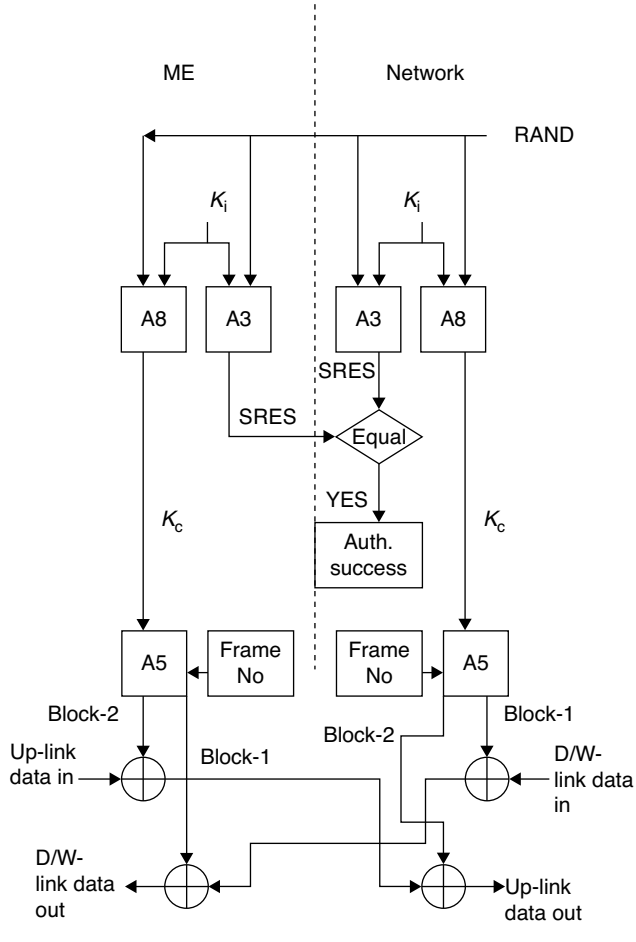


Figure 9.7 GSM authentication and encryption process

3. Only the subscriber (SIM) is authenticated, with the network not being authenticated (no mutual authentication). There is no mechanism to check the validity of the network. So the use of a false base station is possible, this is also called as active attacks.
4. Data integrity is not provided.
5. Lack of anonymity (protect the identity and location of a user) the user does not know whether the encryption is ON or OFF.
6. Week encryption because of short key lengths and hard to upgrade the algorithms.

9.5 Access Mode

As discussed, in the idle mode the MS wakes up periodically to listen to the incoming paging and if there is no paging message, then it goes into the sleep state again, keeping itself idle. When the MS wants to gain dedicated access to the network, to perform a location update, to answer an incoming

paging call or to make a mobile-originated (MO) call, MS initially accesses a BTS using an RACH channel on the uplink of the broadcast carrier. Although an MS transmits an access request in the uplink slots assigned to the RACH, there is no restriction as to which slot it should use, so it is transmitted randomly. If collision occurs, then it may not be unanswered from BTS, so this must be repeated. In this, the MS transmits a 5-bit random number along with the 3-bit number indicating the network about the reason for this access attempt using the access burst power control step mechanism. Some of the bits in the access burst are EXORed with the BSIC number of the intended BTS, so that only the intended BTS decode the message correctly. The initial assignment message is sent to MS by BTS using AGCH.

When a call is flowing, the speech data is mapped to the TCH slot of the traffic channel and sent to the BTS continuously and during that time measurement information is sent using the SACCH channel, as described earlier.

9.5.1 Mobile Originating (MO) Call Procedure

The message flow for mobile originated (MO) call setup is shown in Figure 9.8. The user keys in the phone number for the landline subscriber and presses the send button. RR connection establishment is triggered by sending the channel request message through RACH channel. This message requests the base station system (BSS) for allocation of radio resources for the RR connection setup. The mobile now waits for an assignment on the access grant channel (AGCH). At this point the mobile is listening to the AGCH for a reply. The BSS allocates a traffic channel (TCH) to the MS. The TCH allocation assigns a traffic frequency and a time slot on that frequency. After the mobile receives this message, the mobile must only use the specified resources for communication with the mobile network. The BSS transmits the radio resource assignment to the mobile via the AGCH channel.

The message also contains the ARFCN, slot number, time, and frequency corrections. The time corrections allow the mobile to adjust it's time for transmission. The frequency corrections correct for the Doppler shift caused by the mobile's motion. The mobile detunes from the AGCH and tunes to the specified radio channel. BSS sends to the MS – RR SABM + MM CM SERVICE REQUEST. It is the first message that is sent after tuning to the channel. The mobile initiates an LAPm connection with the BSC by sending a set asynchronous balanced mode (SABM) message. The service request message meant for the MSC is also sent in this message. The BSS replies with Unnumbered Acknowledge (UA) to complete the LAPm setup handshake.

The BSS receives the CM service request message from the mobile and forms a layer-3 message. The BSS then piggy backs it onto the SCCP connection request message, then checks subscriber authentication. MSC checks if the subscriber has been authenticated. If the subscriber has been successfully authenticated, the MSC initiates ciphering of the data being sent on the channel by BSSMAP CIPHER MODE COMMAND. BSS replies back to the MSC, indicating that ciphering has been successfully enabled by BSSMAP CIPHER MODE COMPLETE. The mobile sends the setup message to establish a voice call. The message contains the dialed digits and other information needed for call establishment by CC SETUP. The mobile is informed that the call setup is in progress and MS displays *connecting* on the screen. The MSC allocates a voice circuit on one of the digital trunks between the MSC and the BSS. MSC informs the BSS about the allocated voice circuit. The call is also switched from signaling to voice by BSSMAP ASSIGNMENT REQUEST. The BSS notifies the mobile about the changeover to voice mode by RR CHANNEL MODE MODIFY. Then the mobile acknowledges. The BSS responds back to the MSC by BSSMAP ASSIGNMENT COMPLETE. The MSC routes the call and sends the call towards the called subscriber by ISUP INITIAL ADDRESS MESSAGE SS7 dialed digit. The PSTN indicates to the MSC that it has received all the digits and the called subscriber is being rung by ISUP ADDRESS COMPLETE MESSAGE. The MSC informs the mobile that the called subscriber is being alerted via a ring by CC ALERTING message.

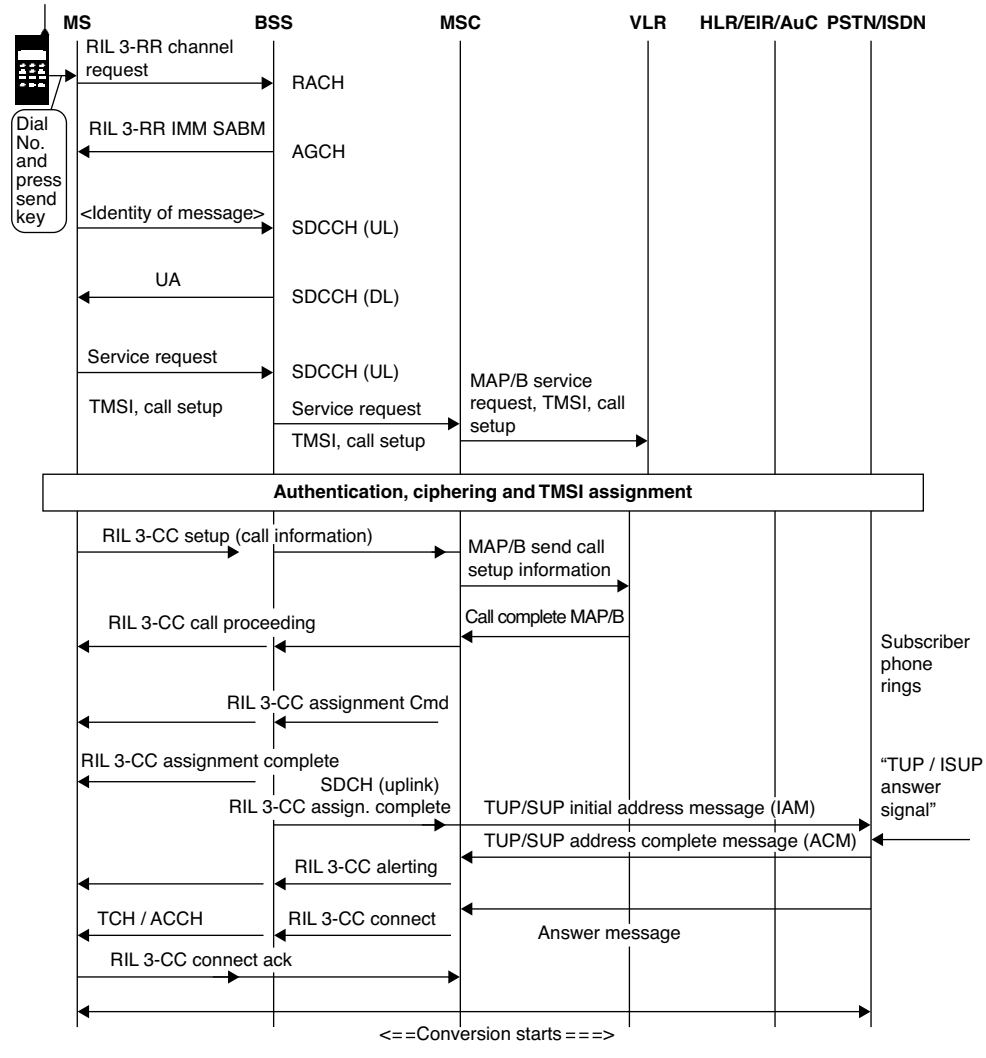


Figure 9.8 Mobile originated call setup messages

When the called subscriber answers the call, the MSC informs the mobile that the call has been answered by CC CONNECT. The MS acknowledges the receipt of CC CONNECT and displays that the call has been connected. The call has entered the conversation phase. The speech path has been setup between the mobile subscriber and the landline subscriber.

Once the call is over, the mobile subscriber hits END to end the call. The mobile sends the disconnect message to the MSC. The MSC initiates release on the PSTN side by ISUP RELEASE. The MSC disconnects the voice path and also releases the voice circuit between the BSS and the MSC. The MSC informs the mobile that it has initiated call release by CC RELEASE. The MSC informs the PSTN that the call release has been completed. The PSTN informs that call release has been completed at its end. The mobile indicates that the call has been released. Call release has been completed, now the RR connection

is released by the MSC. The BSS initiates RR release with the mobile. The BSS informs the MSC that the RR connection has been released. The mobile sends a disconnect message to release the LAPm connection. The BSS replies with an Unnumbered Acknowledge message. The BSS releases the TCH channel. The mobile goes back to the default display to indicate that call has been completely released.

9.5.2 Channel usage for Incoming Call Establishment

In Figure 9.9, the channel usage and the various procedures for incoming call (mobile terminated call – MT call) establishment is shown in sequence.

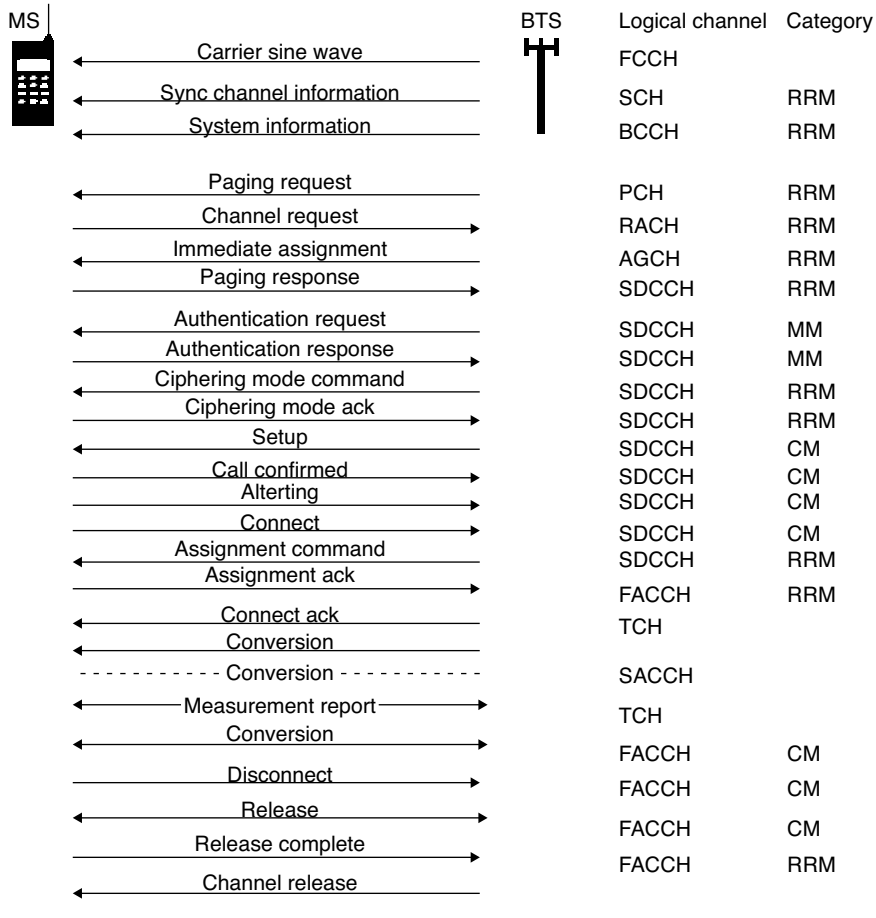


Figure 9.9 Channel usage for incoming call establishment

Figure 9.10 describes a call setup sequences from a fixed network subscriber to a mobile subscriber in a GSM network.

(1) The incoming call is passed from the fixed network to the gateway MSC (GMSC) of the GSM network. (2) Then, based on the IMSI numbers of the called party, its HLR is determined. The HLR checks

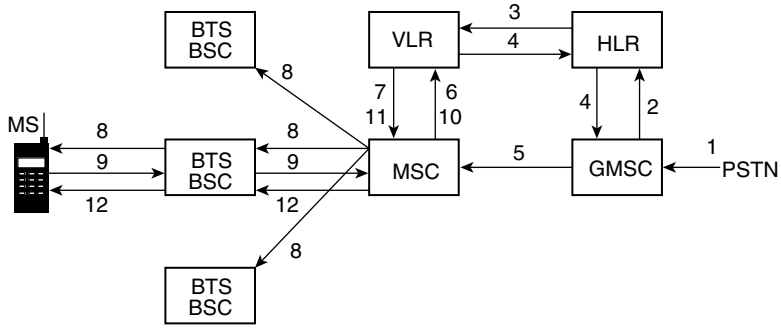


Figure 9.10 Entities for call setup in GSM network

for the existence of the called number. (3) The relevant VLR is requested to provide a mobile station roaming number (MSRN). (4) This is transmitted back to the GMSC. (5) The connection is then switched through to the responsible MSC. (6) Now the VLR is queried for the location range and reachability status of the mobile subscriber. (7) If the MS is marked reachable, a radio call is enabled and (8) executed in all radio zones assigned to the VLR. (9) When the mobile subscriber telephone responds to the page request from the current radio cell, then (10) all necessary security procedures are executed. (11) If this is successful, the VLR indicates to the MSC that the call can be completed and (12) BSC intimates this to the MS.

9.5.2.1 Measurements Performed by MS on Access Mode

When TCH or SDCCH is assigned, during that time when the idle slot is found in this slot, the MS performs measurements on all the adjacent BCCH frequencies. These measurement reports are then sent to the network on SACCH, and are interpreted by the network for the power control and handover procedures. Measurements are performed in each TDMA frame, and are referred to as monitoring, which consists of estimating the receive signal strength on a given frequency. The list of frequencies to be monitored is broadcast on the BCCH information, by means of the BCCH allocation (BA) list, which contains up to 32 frequencies. The frequencies are monitored one after the other, and the measured samples are averaged over the SACCH reporting period prior to reporting to the network, on an uplink SACCH block, in the form of a value referred to as RXLEV. The MS therefore measures the received signal level from surrounding cells by tuning and listening to their BCCH carriers. This can be achieved without inter base station synchronization. The measurements are reported at every reporting period.

For a TCH/FS, the reporting period duration is 104 TDMA frames (480 ms). It is essential that the MS identify which surrounding BSS is being measured in order to ensure reliable handover. Because of frequency reuse with small cluster sizes, the BCCH carrier frequency may not be sufficient to uniquely identify a surrounding cell. The cell in which the MS is situated may have more than one surrounding cell using the same BCCH frequency. It is therefore necessary for the MS to synchronize to and demodulate surrounding BCCH carriers to identify the BSIC in the SB. In order to do so, the MS uses the idle frames. Note that a window of nine consecutive slots is needed to find time slot 0 on the BCCH frequency (remember that time slot 0 carries the SCH and FCCH), as the beacon channels are not necessarily synchronized with one another. Here, one important characteristic to be noted is that the SCH and FCCH are mapped onto the 51 control multiframe structure, and the idle frame of the mobile where measurement is performed appears during the dedicated mode communication and occurs on the 26 traffic multiframe

structure. As 26 and 51 are mutually prime numbers, this means a search frame will be available every 26 modulo 51 frame on the beacon channel.

For instance, let us imagine that an idle frame occurs in the frame 0 of the 51 multiframe. The next idle frames will be programmed on frames 26, 1, 27, 2, and so on. Therefore, after a certain number of search frames, the MS will necessarily decode an FB and an SB. Another measured parameter during a TCH or SDCCH is the RXQUAL, which represents an indication of the quality of the received link, in terms of BER. For each channel, the measured received signal quality is averaged on that channel over the reporting period of length one SACCH multiframe.

9.6 Handover

The GSM network uses a cellular structure. In a cellular mobile network, the user moves from one location to other locations during the call (or in the idle mode), so the radio (and fixed) links cannot be permanently allocated for the entire duration of the call (or throughout the whole connection). In order to maintain a good link quality during the call, the switching of an on-going call to a different channel or cell (BTS or other network resources) is known as handover or handoff (as it is known in North America).

Handover aims to provide continuity of mobile services or seamless connection to a mobile user traveling over cell boundaries in a cellular infrastructure. When a call is going on, during that time, if a user crosses the cell boundaries, then it will be more favorable to use the radio resources in the new cell (target cell) compared with the old cell, because the signal strength provided by the old cell worsens as the user goes away from the old cell coverage area and enters into the new target cell. The whole process of tearing down the existing connection with the current cell BTS and establishing a new connection with the appropriate cell BTS is called "handover." The ability of a cellular network to perform efficient handovers is critical to offering attractive services and seamless connectivity experience to the users. Sometimes handover procedures cannot be completed for several reasons, such as signaling failures due to the lack of resources in the new target cell, synchronization issues, and so on, and then this is called "handover failure." In high performance networks, where there is a trend towards the use of smaller cells to increase the subscriber capacity, in such situations the handover process becomes even more important, as due to the smaller cell size, more frequent handovers are necessary.

The basic function of the RR (layer-3) is the execution of handover using measurement reports. Handover is required for several reasons and handover scenarios are categorized based on the location of the MS and the clock synchronization between the source cell and destination cell. The handover can happen in idle mode or in the dedicated mode when the call is going on. Normally, based on location and usage there are four different types of handover used in the GSM system:

1. Handover between channels (time slots) in the same cell – inter carrier handover. In this case, a subscriber is diverted to a different traffic channel within the same cell. Generally, this channel is generated with a different frequency or time slot. The decision about the handover is made by the BSC that controls the cell.
2. Handover between cells (base transceiver stations) under the control of the same base station controller (BSC) – intra BSC handover. This takes place when a mobile user moves from one cell into a neighboring cell, both controlled by the same BSC. The traffic connection to the old cell is discontinued as soon as the connection setup to the new cell is successfully completed. This process is controlled by the BSC.
3. Handover between cells under the control of different BSCs, but belonging to the same mobile services switching center (MSC) – inter BSC handover.
4. Handover between cells under the control of different MSCs – inter MSC handover.

The first two types of handover, called internal handovers, involve only one base station controller (BSC). To save the signaling bandwidth, they are managed by the BSC without involving the mobile services switching center (MSC), except for notifying it at the completion of the handover. The last two types of handover, also called external handovers, are handled by the MSCs. An important aspect of GSM is that the original MSC (the anchor MSC) remains responsible for most call-related functions, with the exception of subsequent inter-BSC handovers under the control of the new MSC, called the relay MSC.

The GSM system uses mobile assisted handover. Handovers can be initiated by either the mobile or the MSC (as a means of traffic load balancing and optimum network resource usage). During its idle time slots, the mobile scans the broadcast control channel of up to 16 neighboring cells (as mentioned in the BA list in BCCH channel information), and forms a list of the six best candidates for possible handover, based on the received signal strength. This information is passed to the BSC and MSC, at least once per second, and this is used by the handover algorithm.

Handover categorization based on source and destination BTSs system clocks are: (1) synchronized, (2) non-synchronized, (3) pseudo synchronized, and (4) blind handover.

The algorithm for when a handover decision should be taken and what the algorithm would be are not specified in the GSM standards recommendations. Generally, there are two basic algorithms used, both closely tied to the power control. This is because the BSC usually does not know whether the poor signal quality is due to multi-path fading or because of the mobile's movement into another cell.

What triggers a handover process?

Handover may occur either because of deterioration of radio parameters or network parameters, as listed below.

1. Radio criteria/radio parameters:
   i. Received quality (RXQUAL) too low or bit error rate too high, for example, BER has increased above the threshold level.
   ii. Received level too low (RXLEV on uplink and downlink), for example, RSSI has been dropped below the expected level.
   iii. MS–BS distance handover (timing advance), for example, the distance between MS and BTS has been increased above the threshold distance, which is measured from the timing advance.
   iv. Power budget handover (handover to a better cell with regard to relative received level.).
2. Network criteria/network parameters:
   i. Serving cell congestion.
   ii. MS–BS distance too high in extended cells.

9.6.1 Handover Process

When making a handover decision the BSS will process, store, and compare certain measurement parameters that are received on the SACCH channel with the predefined thresholds.

9.6.1.1 Measurement Information

In Chapter 2, we saw that the wireless channel is very unpredictable and its characteristics changes with respect to time. The channel gain ($|h|^2$) changes based on the fading type being experienced, velocity of the mobile, and presence of reflectors. If the channel gain or the changing behavior of the channel is known beforehand, then on both sides (BTS and MS) transmitters can adjust the parameters (power level, coding, etc.) to mitigate this effect or can take a decision to make handover to the other cell. Thus, in order to do this, the MS and the BTS have to monitor the channel

periodically (at a defined interval and for a fast fading channel more frequently than for a slow fading channel). The base station and mobile periodically measure different parameters of the radio link (wireless channel).

1. **The measurement performed at the BTS**
   a. The power level of uplink signal received from MS (RXLEV_UL).
   b. The quality (BER) of the uplink received signal from MS (RXQUAL_UL).
   c. The distance between the MS and the BTS based on the adaptive timing advance parameters.
   d. The interference level in unallocated time slots.
2. **Measurements performed by the MS**
   a. The power level of downlink signal received from the serving cell (RXLEV_DL).
   b. The quality (BER) of the downlink signal received from the serving cell (RXQUAL_DL).
   c. The power level of the downlink signal received from different (*n* cells) neighbor cells (RXLEV_NCELL(*n*)).

The exact measured values are not sent to the other party; rather these are mapped to different values based on the level. This is shown in Table 9.1.

Table 9.1 Measurement values averaged over 1 SACCH block (104 frame = 480 ms)

RXLEV (measured power level (on neighboring cell + serving cell BCCH))			RXQUAL (raw bit error rate)		
RX signal level	From (dBm)	To (dBm)	Bit error ratio	From (%)	To (%)
RXLEV_0	—	−110	RXQUAL_0	—	0.2
RXLEV_1	−110	−109	RXQUAL_1	0.2	0.4
RXLEV_2	−109	−108	RXQUAL_2	0.4	0.8
RXLEV_3	−108	−107	RXQUAL_3	0.8	1.6
RXLEV_62	−49	−48			
RXLEV_63	−48	—	RXQUAL_6	6.4	12.8

9.6.1.2 Measurement Schedule

As discussed above, the MS should not only measure the downlink signal from its serving BTS, but also downlink signals from neighboring BTSs. Each BTS transmits a list, which contains the BCCH carrier frequencies of its neighboring BTSs, this is known as BA (BCCH allocation) list via the BCCH channels. The MS will read the list, which is given by its serving BTS and find out the neighboring BTS's of the BCCH frequencies, on which it has to perform the measurement. The GSM specification requires that the BCCH information in a BCCH carrier should be read at least once every 10 s. Ideally, examining the transmission and reception schedule at the MS, there are three slots/windows available during these periods where measurements can be performed. As we have seen earlier, there is a gap of at least two time slots (minus the timing advance) between the reception of the downlink burst and the transmission of the uplink burst. Considering the timing advance at its maximum, the window is 920 μs in duration, and that looks too short for measurement purposes.

The second window occurs between the transmission of the uplink burst and the reception of the downlink burst and the minimum duration is four time slots or 2.3 ms (with no timing advance). This

window is used by the MS to measure the downlink signal strength of the BCCH carriers for neighboring cells. During this time period, the MS must retune to the respective neighbor cell's BCCH carrier, which needs to be measured. It performs the measurement, and then again retunes to the current downlink frequency, in time to receive the next burst from the serving BTS. This tight schedule does not allow the MS to wait for an active burst, and consequently every slot on the BCCH carrier must remain active. This is achieved by using dummy bursts to fill the slots of the BCCH carrier that would normally be inactive. There is also a requirement that the BCCH carrier is transmitted at full power (higher power level) and therefore downlink power control must not be applied and DTX may not be used on any slots on the BCCH carrier.

The third measurement window is obtained from the idle frame, which is included in each traffic multiframe. This window is a minimum of 12 time slots or around 6.92 ms + timing advance. This measurement window is used to ensure that the BCCH carrier measurements described above are always associated with the correct BTS. Simply the measuring of the RSSI value for different BCCH carrier frequencies (BTSs) is not enough, as BCCH carrier frequencies are reused by distant cells (BTSs) throughout the network. Thus, MS has to periodically verify the identity of the BTS by decoding the SB and checking the 6-bit BSIC number. In the case of co-channel interference, for example, two far BTSs using the same BCCH carrier frequency, they will be distinguished because they have different BSICs. During this period (12 time slots), the MS is required to retune to the BCCH carrier, identify and decode a valid SB, and then again retune to its current downlink frequency to receive the next burst from the serving BTS. Here it is assumed that the SB of the neighboring BCCH carrier falls within this 12-slot measurement window. The SB occurs at an interval of 10 or 11 (where ther is an idle frame in between) TDMA frames within the 51 frame control multiframe structure. Hence, there is no guarantee that this measurement window will align with an SB every time. However, the time frame structure means that an idle frame window in the 26-frame traffic multiframe gradually slips past the 51-frame control multiframe ensuring that the idle window will coincide with an SB.

From the decoded SB, the MS will find out and store the multiframe, super frame and hyper frame parameters, T1, T2, and T3′, and it may use this to schedule the decoding of BSIC. This synchronization information may also be employed to reduce the switching times at handover. The MS reports the BCCH carrier measurements to the network against the BSIC number and this allows the network to ensure that the measurements are associated with the correct BTSs.

9.6.1.3 Measurement Averaging

The MS sends the measurement results to the networks as a measurement report message via the SACCH, which contains a report of up to six neighboring cells, in addition to the serving cell. The information carried on the SACCH is interleaved over four bursts and this represents a delay of $4 \times 120\,\text{ms} = 480\,\text{ms}$; this time period is referred to as a reporting period. Thus, to send or receive the interleaved or de-interleaved SACCH data, the MS or BTS has to wait for 4 SACCH periods (26*4.615 ms). Generally, over this period, the MS averages out the measurement values, before they are reported to the network. Further averaging will take place once the measurement reports arrive at the BSS. The BSS must be able to store at least 32 measurement samples, where a sample is defined as the value evaluated by the MS during the measurement reporting period of 480 ms. The BSS may weight the samples to attach more importance to the more recent measurements or may perform an unweighted average of the 32 samples to take handover decisions using a handover decision algorithm. Although the standard does not specify the algorithm, GSM 05.08 provides an example of how to design a handoff algorithm.

Handover margin (Figure 9.11) is a parameter used in order to prevent repetitive handover between adjacent cells. It may also be used as a threshold in handover cause.

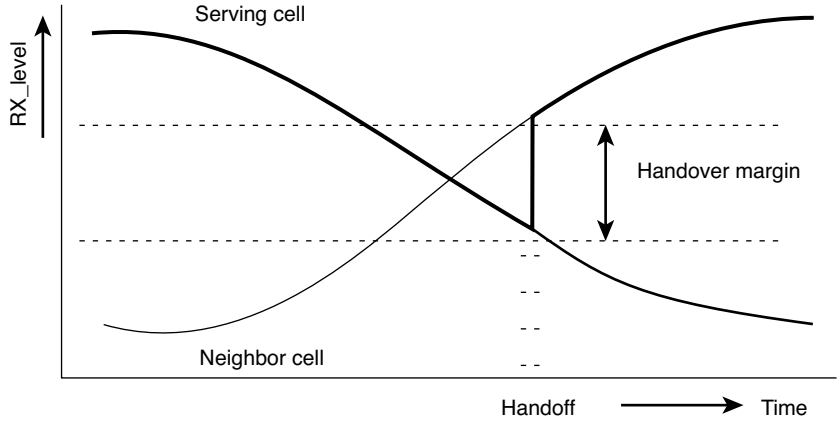


Figure 9.11 Handover threshold margin

9.6.2 *Example Handover Procedure*

A typical handover procedure (inter BSC, blind handover) is shown in Figure 9.12. The different steps involved in this sequence are:

1. Handover request from old BSC_O to MSC.
2. Request is forwarded to new BSC_N.

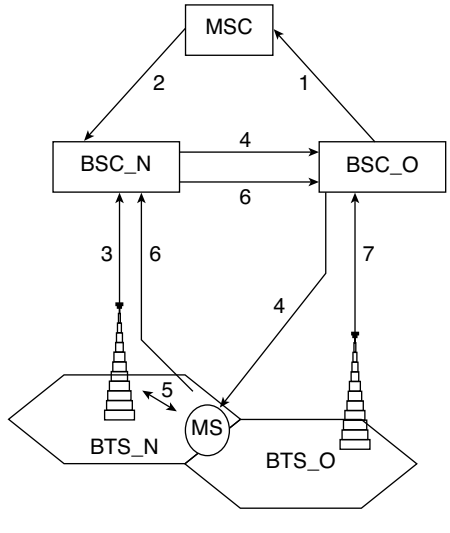


Figure 9.12 Inter BSC handover

3. BSC_N prepares to receive new MS.
4. BSC_O sends handover command to MS.
5. MS sends RACH messages continuously to BTS_N and MS starts T3124 timer, BTS new sends PHYSICAL INFO to MS and starts timer T3105.
6. On success, HO_COMPLETE is sent to BSC_O through BSC_N.
7. BSC_O releases resources.

This process is described in detail below.

During an active call, the mobile periodically reports the signal level and quality of the serving and neighboring cells to the network via the measurement report message in every SACCH frame with a periodicity of 480 ms. When the mobile is reporting good signal quality, no further action is taken. However, when the mobile moves to the edge of the serving cell and its report shows poorer signal strength from serving cell, then the serving BSC decides to initiate a handover as the mobile could be served better by another cell. The BSC analyzes the measurement reports to determine which cell will serve the mobile better. The BSC decides to request for a handover. The serving BSC sends a handover request message to the MSC indicating a rank-ordered list of the target cells that are qualified to receive the call. Then the T7 timer is started to wait for the reply handover command from the MSC. The MSC reviews the global cell identity associated with the best candidate to determine the target cell. The MSC passes on the handover request to the new target BSC. The MSC-VLR starts a timer (T101) to wait for the response from the new BSC. The handover request is treated as a new call. A traffic channel (TCH) (if available) is allocated for the call that will be handed-in soon. At this point the new BSC prepares the handover command that needs to be sent to the mobile. This message contains all the information the mobile will need to handover to this cell. The new BSC includes the RR HANDOVER COMMAND message as a payload in the HANDOVER REQUEST ACK that is sent back to the MSC. As the MSC has heard back from the destination BSC, so the T101 timer is stopped. The MSC delivers the handover command to the old BSC. This command encapsulates the RR HANDOVER COMMAND from the destination BSC. Thus the T7 timer is stopped. Now, T102 is started to track the completion of the handover. The RR HANDOVER COMMAND will be delivered to the mobile via the old BSC. The old BSC extracts the RR HANDOVER COMMAND message from the BSSMAP message and sends it to the mobile. T8 is started to wait for the success of this call from the MSC. If the handover to the target cell is successful, the MSC will initiate a resource release to the old BSC.

The mobile extracts the destination channel information from the message and tunes to the assigned channel. After tuning to the assigned channel, the mobile starts sending the handover accept message. Note that this message is sent as an access burst (RACH) as the mobile is not completely synchronized to send normal bursts. The T3124 timer is started to await the PHYSICAL INFORMATION message from the network. The BSC receives the HANDOVER ACCEPT from the terminal. The actual call is identified using the handover reference. The BSC informs the MSC that the handover has been detected. At this point the MSC can switch the voice path. The MSC switches the voice path. The new BSC sends the PHYSICAL INFORMATION message to the mobile via RR PHYSICAL INFORMATION message. This message contains a time and frequency correction (based on the handover type). T3105 is started to await the receipt of the SABM for the signaling connection. The mobile applies the received corrections and can now send TCH bursts on the channel. TCH bursts contain the speech from the user. T3124 is stopped as PHYSICAL INFORMATION message has been received. The mobile sends an SABM to establish the signaling connection. Receipt of the SABM stops the T3105 timer. The BSC replies with a UA message. The mobile uses the signaling connection to indicate that the handover has been completed. The BSC forwards the handover completion event to the MSC. Now, handover has been completed, so T102 is stopped. Call release has been completed, and the RR connection is released by the MSC. The T8 timer is stopped as the resources for the handed over call in the source BSC are released. The BSS informs the MSC that the RR connection has been released.

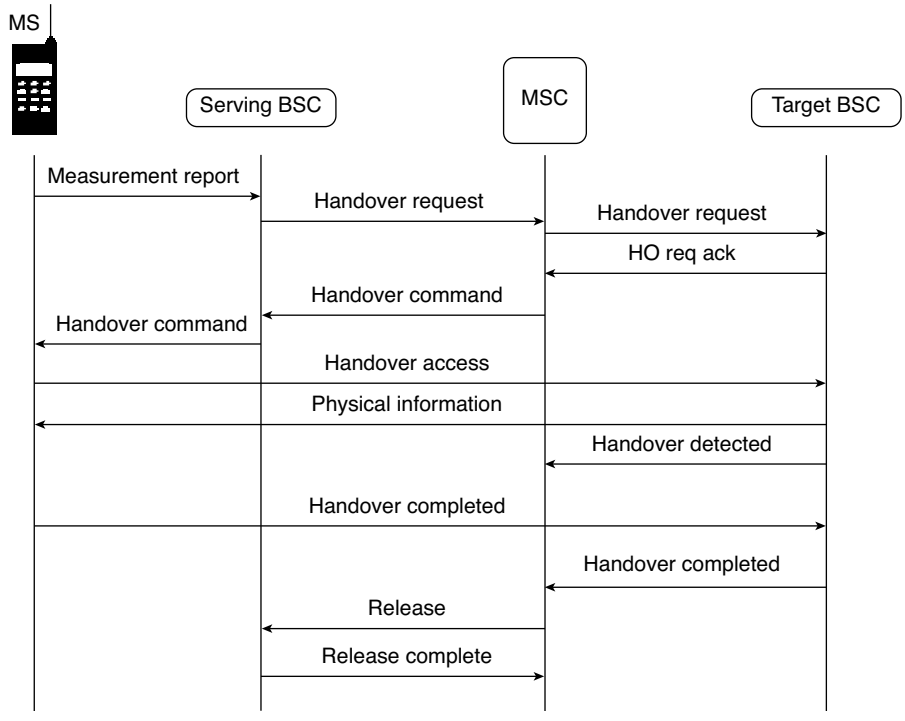


Figure 9.13 Inter MSC handover message flow

In the case of an inter-MSC handover (Figure 9.13), when a call is handed over from the serving MSC to the target MSC via PSTN, the serving MSC sets us an inter-MSC voice connection by placing a call to the directory number that belongs to the target MSC. When the serving MSC routes the call, the PSTN is unaware of the handover and follows the normal call routine procedures, delivering the call to target MSC. The serving MSC sends a prepared handover message to target MSC. Then the target MSC sends an allocated handover message to its VLR for TMSI assignment. Next, target VLR sends the TMSI in the handover report message. After this the process is more or less similar to intra-MSC handover.

Based on handover type (with respect to clock synchronization between the BTSs) the handover procedure varies a little, which is described below.

9.6.2.1 Synchronous Handover

In the case of synchronized handover, the network commands a synchronous handover on the downlink main DCCH. Layer-3 posts an SYNC_HO_REQ message to layer-1. Layer-1 tries to find the new cell's FB and SB information. The L1 search result is reported and this checks whether the BSIC matches with the BSIC passed inside the handover command. Here, sending RACH is not required, as BTSs are synchronized. In many implementations, the FB, SB search is done during the received signal strength monitoring of different neighbors (and send BSIC again via measurement report), so in that case again a separate search is not required. For an AMR feature, the protocol stack must fill all the multi-rate configuration parameters only if the channel mode is TCH/AFS or TCH/AHS. The previous multi-rate configuration must be sent again to the L1 with the handover message.

9.6.2.2 Non-Synchronized Handover

Networks issue non-synchronized handovers if old and new BTSs are not synchronized. The MS switch to assigned channels, starts repeating RACH messages to BTS. MS starts the T3124 timer. BTS sends PHYSICAL INFO to MS and starts timer T3105. MS stops the timer T3124 on reception of PHYSICAL INFO. BTS waits for layer-2 frame or TCH from MS. If there is decoder error on the frame received from MS, BTS repeats PHY INFO and restarts timer T3105 (maximum number of repetitions is Ny1).

T3124 – 675 ms if the channel type of the channel allocated in the HANDOVER COMMAND is an SDCCH (+SACCH), otherwise its value is set to 320 ms. T3105 and Ny1 are network dependent.

9.6.2.3 Pseudo-Synchronized

Sending RACH is optional and three different parameters are defined: propagation delay, RTD (real-time difference), and OTD (observed-time difference). BTS_O sends the RTD value in the handover message. MS calculates the timing advance of BTS_N with the relationship: $t_N = \text{OTD} - \text{RTD} + t_O$. MS sends OTD + t_O to BTS_N allowing it to make a non-biased estimate of RTD.

9.6.2.4 Blind Handover

This is an R99 feature. Handover is ordered to the cell to which the MS is not synchronized. The MS starts searching for FB–SB information. A timeout of 300 ms is allowed for the FB–SB search. The handover procedure is more or less the same as for a non-synchronized handover.

9.7 Radio Resource Control Procedure

The RR in BSS communicates the RR counter part in the MS. RR is used for signaling between the GSM network and MS. The RR state machine primarily consists of two states, namely idle and dedicated. In the idle mode, as the name suggests, the MS is not actively involved in any communication and no dedicated resource is assigned to it. In the dedicated mode, the resource is reserved to communicate with the BSS. The RR protocol is specified in detail in 3GPP TS 44.018.

As the MS does not have dedicated resources to communicate to the network, all communication first requires that the MS establishes an RR connection. The RR connection is allocated and supervised by the BSC. Once the RR connection is established the MS moves from the idle connection state to the dedicated connection state (Figure 9.14).

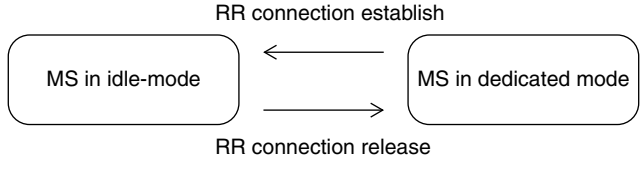


Figure 9.14 RR state machine

9.8 Mobility Management Procedure

The mobility management (MM) layer is built on top of the RR layer, and handles the functions that arise from the mobility of the subscriber (such as keeping track of the present location of the MS), in addition to the authentication and security aspects. MM procedures are used between the MS and GSM network subsystem. BSS transparently relays MM procedure messages. The simplified state diagram of the MM state model is shown in Figure 9.15.

- Detached state – the network does not know the location of the MS.
- Idle state – the location of the MS is known to the network, but there is no active session for the MS.
- Connected state – the location of the MS is known and there is an ongoing active session for the MS.

Location management is concerned with the procedures that enables the system to know the current location of a power-on mobile station so that incoming call routing can be completed.

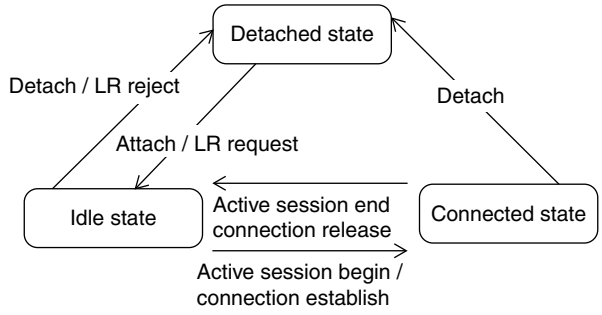


Figure 9.15 Mobility management state procedure

9.9 Call Routing

Generally, in a fixed network a terminal is semi-permanently wired to a central office, but the GSM user can roam nationally and even internationally. As discussed in Chapter 6, the directory number dialed to reach a mobile subscriber is called the mobile subscriber ISDN (MSISDN), which is defined by the ITU-T recommendation E.164 numbering plan. An incoming mobile terminating (MT) call is directed to the gateway MSC (GMSC). The GMSC is basically a switch, which interrogates the subscriber's HLR to obtain the routing information, and thus contains a table linking MSISDNs to their corresponding HLR. It should be noted that the GMSC function is distinct from the MSC function, but is usually implemented in an MSC.

The routing information that is returned to the GMSC is the mobile station roaming number (MSRN), which is also defined by the E.164 numbering plan. MSRNs are related to the geographical numbering plan, and not assigned to subscribers, nor are they visible to subscribers.

Then, as shown in Figure 9.16, the most general routing procedure starts with the GMSC querying to the called subscriber's HLR for an MSRN. The HLR typically stores only the SS7 address of the subscriber's current VLR, and does not have the MSRN. The HLR must therefore query the subscriber's current VLR, which will temporarily allocate an MSRN from its pool for the call. This MSRN is returned to the HLR and back to the GMSC, which can then route the call to the new MSC. At the new MSC, the IMSI corresponding to the MSRN is looked up, and then the mobile is paged in its current location area.

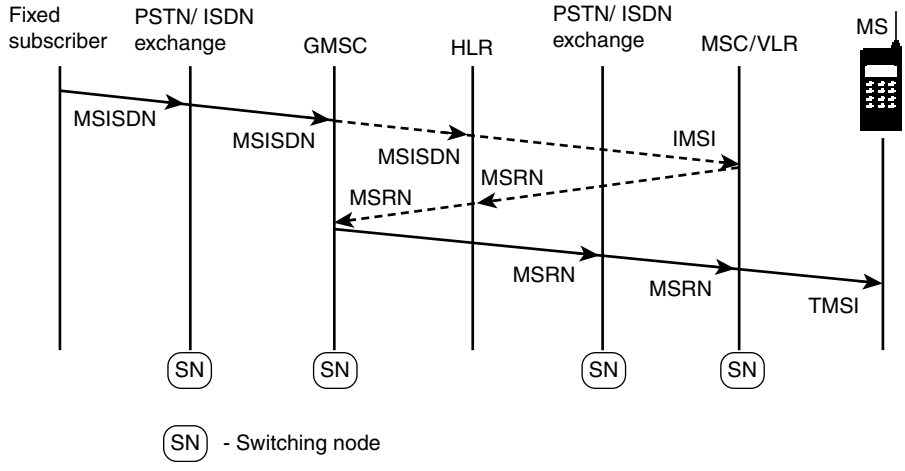


Figure 9.16 Call routing for a mobile terminating call

9.10 Power Control

As mentioned earlier, the transmitted power should always be the optimum, as it helps to minimize the interference level in the system and also helps to improve battery life. The GSM system employs power control to ensure that the MS and BTS only transmit sufficient power to maintain an acceptable link, thereby reducing interference to neighboring cells and improving spectral efficiency.

When the BTS commands the MS to increase or decrease power, the MS has the ability to reduce or increase the power in steps of 2 dB up to its maximum power class limit (Figure 9.17), as defined in Chapter 7, Section 7.5. The transmission power of the MS is controlled by the network conveying messages over the slow associated control channel (SACCH). Power level values range from 0 to 15, where, 0 means −43 dBm (20 W) and 15 means −13 dBm (20 mW). After receiving a power control command, an MS adjusts its transmitted power to the requested power level at a maximum rate of 2 dB every 60 ms. This power control algorithm is manufacturer-specific and runs on the BSC. The operator has the freedom to decide whether to use power control or not on both uplink and downlink, and it may also be applied independently on either link. However, downlink power control may not be applied to any slots on

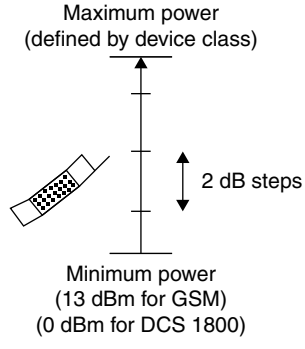


Figure 9.17 Power control steps

the BCCH carrier as it must be transmitted at a constant power, because it is measured by the MSs in the surrounding cells for the handover decision purposes.

9.11 Discontinuous Transmission and Reception

One goal in a GSM cellular system is to minimize co-channel interference, as it allows better service and increases the overall capacity of the system. Generally, during a typical telephone conversation, a person speaks for around 40% of the time and the remaining 60% of the time remains silent. Discontinuous transmission (DTX) is a method that takes the advantage of this fact by turning the transmitter off during periods of silence. It helps in two ways: it reduces the co-channel interference and also saves battery power by switching off the mobile transmitter.

The silent period is detected using voice activity detection (VAD), where the energy in the speech signal is computed for each speech block and a decision is taken about the block that contains the speech or background noise by using an adaptive threshold. Distinguishing between voice and noise (silent) inputs is not a trivial task, because if a voice signal is misinterpreted as noise, then the transmitter is turned off and a very annoying effect called clipping is heard at the receiving end, but on the other hand, if noise is misinterpreted as a voice signal (which happens too often), then the efficiency of DTX will be dramatically decreased.

When the transmitter is turned off, total silence will be heard at the receiving end, which can not be disabled by the user. This, to assure the receiver that the connection is not dead, a comfort noise is inserted during the silence period at the receiving end, by trying to match the characteristics of the transmitting end's background noise.

There is also another method that is used to conserve power at the mobile station, which is discontinuous reception. We have already discussed this earlier in Section 9.2.1 in detail.

Normally, downlink power control and discontinuous transmission (DTX) is not used on the BCCH carrier. A traffic channel or dedicated control channel may use the BCCH carrier frequency as part of its hopping sequence, but it must obey the above rules while it is using that carrier, that is, power control and DTX are not permitted.

9.12 Frequency Hopping

Frequency hopping is a technique for improving the signal to noise and interference ratio in a link by using frequency diversity. Whenever it is required the base station commands the MS to activate frequency hopping. When frequency hopping is activated in the MS, the base station assigns the mobile station a set of RF channels, rather than a single RF channel. A frequency hopping algorithm is also assigned to the mobile, which will indicate the pattern of the available frequencies that it has to use. In a GSM/GPRS/EGPRS network, frequency hopping is specified differently in individual cells, based on the number of frequencies offered by a specific cell. Each burst of a particular physical channel will be transmitted on a different carrier frequency in each TDAM frame. So, the hopping rate is equal to the frame rate (that is, 217 frames/s). The only physical channels that are not allowed to hop are the broadcast and common control channels (that is, the FCH, SCH, BCCH, PCH, and AGCH). Because an MS must be able to locate these channels easily on power-up, this process would become more complex if frequency hopping were allowed on those channels.

9.12.1 Frequency Hopping Sequences

The order of sequence in which different carrier frequencies will be used on the uplink and the downlink is known as the frequency hopping sequence. Only a single hopping sequence is required to describe the

complete duplex link (both for up- and down-link sequences), because the uplink and downlink frequencies are always separated by the duplex channel spacing (45 MHz). On the BCCH, each BTS transmits the details of all the carriers that it is using, in a cell channel description message. This message takes the form of a 124-bitmap, where each bit represents a carrier and the bit value of "1" or "0" indicates whether each particular carrier is in use by that BTS or not. In the idle mode, the MS usually decodes and stores this information. Once the list of carrier frequencies are known and assigned to the frequency hopping channel, the MS must also determine the sequence in which each frequency has to be used, which is the hopping sequence. On initial assignment, the mobile allocation is described as a subset of the cell allocation, thus reducing the signaling overhead on the AGCH. The initial assignment messages are sent on the common access grant channel (AGCH), where the message size should be kept short to preserve the access capacity of the system. A two-step approach is used to avoid transmitting the full mobile allocation parameter at initial assignment. During the channel assignment, the hopping sequence is derived by the mobile from the broadcast parameters, namely, the mobile allocation (set of N frequencies on which to hop), the hopping sequence number (HSN) of the cell (which allows different sequences on the co-channel cells), and the index offset (to distinguish the different mobiles of the cell using the same mobile allocation) or mobile allocation index offset (MAIO). Based on these parameters and on the FN, the MS knows which frequency to hop in each TDMA frame. The HSN selects one of the 64 predefined "random" hopping sequences, while the MAIO selects the start point within the sequence, and the MAIO may take as many values as there are frequencies in the mobile allocation. If the value of $\text{HSN} = 0$, this indicates that it will choose a cyclic sequence, where the frequencies in the mobile allocation are used one after another. Frequency hopping channels with the same HSN, but having different MAIOs, will never use the same frequency simultaneously as they are orthogonal, for example, all frequency hopping channels within a cell employ the same HSN but have different MAIOs.

Further Reading

Erbespächer, J., Vogel, H., and Bettstetter, C. (2001) *GSM: Switching Services and Protocols*, John Wiley & Sons, Inc., New York, ISBN 0-471-499903-X.

Garg, V.K. and Wilkes, J.E. (1999) *Principles and Applications of GSM*, Pearson Education, Upper Saddle River, NJ.

Gudmundson, M. (1991) Analysis of handover algorithms. Proc. IEEE VTC '91, St. Louis, May 1991 (537–542).

Redl, S.M., Weber, M.K., and Oliphant, M.W. (1995) *An Introduction to GSM*, Artech House, Norwood, MA.

10

Anatomy of a GSM Mobile Handset

10.1 Introduction to the GSM Handset

The GSM mobile handset has evolved over a period of time and its efficiency with respect to size, weight, complexity, application support, performance, and battery life has improved. In Chapter 1, we briefly discussed the internal components of any mobile phone. The basic phone architecture and the associated peripherals such as display (LCD), keypad, speaker, microphone, and so on, remain almost the same with respect to the air interface technology or mobile standard used. However, based on the particular mobile standard, the front-end RF unit, the baseband processing unit, and the protocol stack used (especially up to layer-3) will be different.

As discussed in Chapter 1, a GSM mobile contains several components, such as a microphone, speaker, LCD display, keypad, battery, LED, baseband processors, CODEC, SIM card, memory, RF unit, antennas, connectors, and so on. Generally only a few ICs are mounted in the PCB to carry out these functional activities. In Figure 10.1 an IC mounted PCB of a typical GSM mobile phone is shown as an example.

Nowadays, as a result of technological advancements, higher integration is possible which iswhy only a few ICs are sufficient to make a complex mobile phone. A single-chip solution (system-on-a-chip, SOC) is also available, which will not only reduce the size of the device but also reduce the cost and battery consumption. Deep submicron CMOS technologies enable sophisticated SOC technologies, moving beyond the traditional integration of digital functionality to include RF, mixed-signal, and power management capabilities. A typical block diagram of a GSM single-chip solution is shown in Figure 10.2. This integrates an application and protocol processor, digital signal processor for L1 (baseband) and speech processing, SRAM, RF transceiver, audio processing, audio driver, RF and audio amplifiers, A/D and D/A converters, loudspeaker driver, dc to dc converter for display backlight, regulators for power supplies, battery charger, SIM interface, clock generation unit, RTC, and so on. This high level of integration inside the SOC results in a minimal BOM (bill of material) for the GSM handset. Thus only a few components (such as the LCD display, keypad, loudspeaker, microphone, RF power amplifier, RF receiver filters, and flash memory) need to connect externally to make the complete phone.

Earlier, we described the details ofthe baseband processing and protocol processing parts of a GSM phone, in Chapters 7–9, and we also discussed the different RF transmitter and receiver architectures in Chapter 4. In this chapter we will explain the internal details of a GSM mobile phone and some of its components in more detail.

Mobile Handset Design Sajal Kumar Das

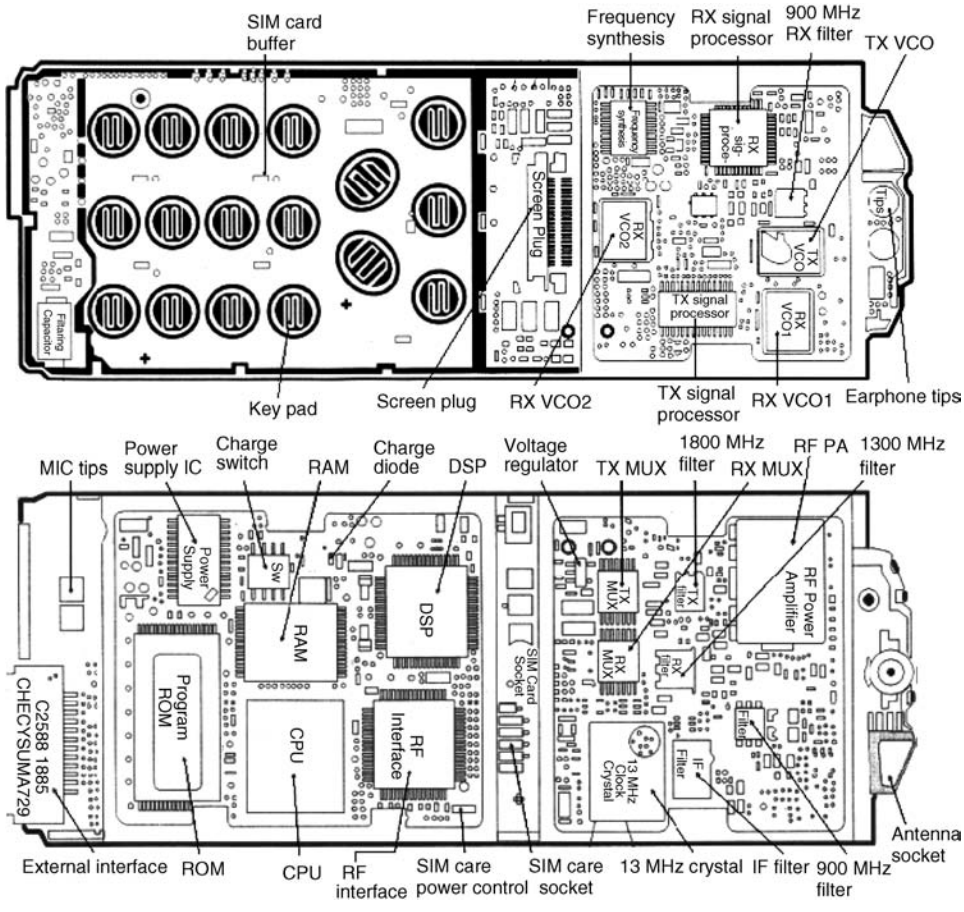


Figure 10.1 GSM mobile phone PCB populated with components (top and bottom part of the PCB)

10.2 Functional Blocks Inside a GSM Mobile Phone

Today, apart from the speech communication, the GSM mobile phone supports many other multimedia applications, such as audio-video players, games, and so on. These functional blocks can basically be divided into two categories: the communication processing unit (which helps to transfer data over the channel, for example, modem part) and the application processing unit. Firstly, we will consider the communication processing unit, for example, the modem functionalities that need to be implemented inside a GSM mobile, and then we will see how to map those functionalities into different hardware or software blocks inside the mobile. Later in this chapter we will discuss some of the application processing units. The functional blocks inside a GSM handset are shown in Figure 10.3. There are two main planes: (1) the control plane, on which the protocol specific control information flows, and (2) the user plane, on which user data flows.

In the GSM user plane, the user speaks in front of the phone's microphone, which converts the acoustic pressure activity into an analog electrical signal. This electrical signal is then converted into a digital signal and given to speech coder. The output from the speech coder is a 13 kbps digital stream, which is

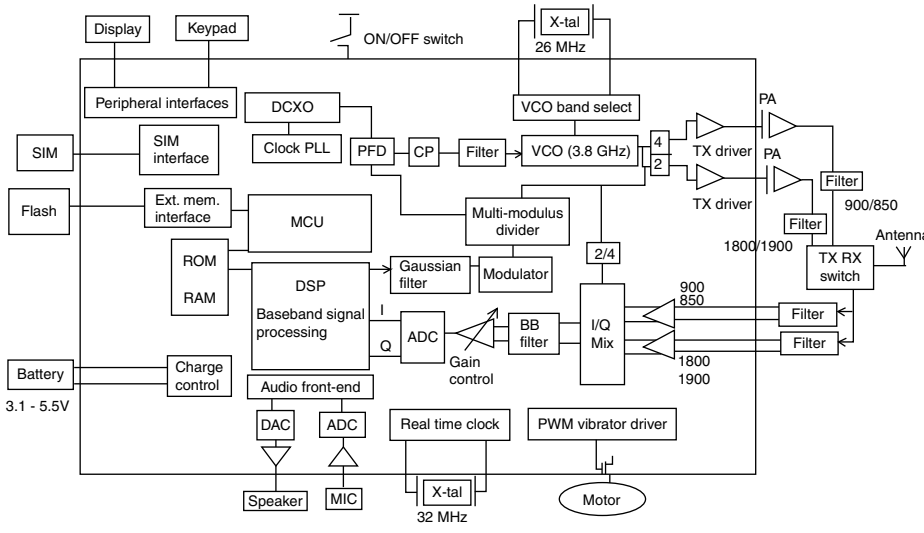


Figure 10.2 Block diagram of a GSM single-chip solution based phone

passed to the channel coder unit. Similarly for data transmission, data from data applications (web applications, ftp, and so on, or a pc with a USB interface to a phone) is passed to the protocol stack (at a rate of 9.6 or 4.8 or $<$2.4 kbps) and a certain amount of auxiliary information is added to produce an intermediate data rate of 12 kbps. These data are also delivered to the channel coding unit. After that, the channel coded speech data or application data are passed to the interleaving, ciphering, and burst forming module, respectively. Normal burst is used for this. This burst data are then mapped to TCH channel multiframe.

However, for the GSM control plane, data are passed to the channel coding unit from the protocol stack, then interleaved and ciphered as required and mapped to different bursts according to the channel configuration (BCCH, SACCH, FACCH, SDCCH, etc.).

The burst data (whether from the user plane or control plane) is differentially encoded and passed to the modulator (GSMK). Then it is passed to the RF module for carrier frequency multiplication, power amplification, and transmission via the antenna.

On the other hand, at the receiver side, the EM waves impinge on the metal antenna mounted inside the mobile and tries to penetrate via this, which creates a very feeble surface current that is band-pass filtered and passed via LNA in the RF block for the desired amplification. Then it is passed through the equalization and demodulation unit. Usually the channel estimation is based on the known sequence of bits, that is unique for a certain transmitter and which is repeated in every transmission burst, and is known as the training sequence. Thus, the channel estimator is able to estimate the channel impulse response (CIR) for each burst separately by exploiting the known transmitted bits and the corresponding received samples. This channel impulse response is used by the equalizer to equalize the received burst energy against the channel impairment. Then the equalized burst data (normally softbits) are given to a Viterbi decoder for decoding of the bits (hardbits). These are then de-ciphered, de-interleaved, and passed to the higher layer. Most commonly, the channel estimation and equalization block provide the softbit values on which the deciphering and de-interleaving are performed, then they are passed to a Viterbi decoder for hardbit decoding (for more details, please refer to Chapters 3 and 7).

The functionalities are implemented by using special hardware modules with software control from the processor or implemented on software running on the processors. During the initial system architecture

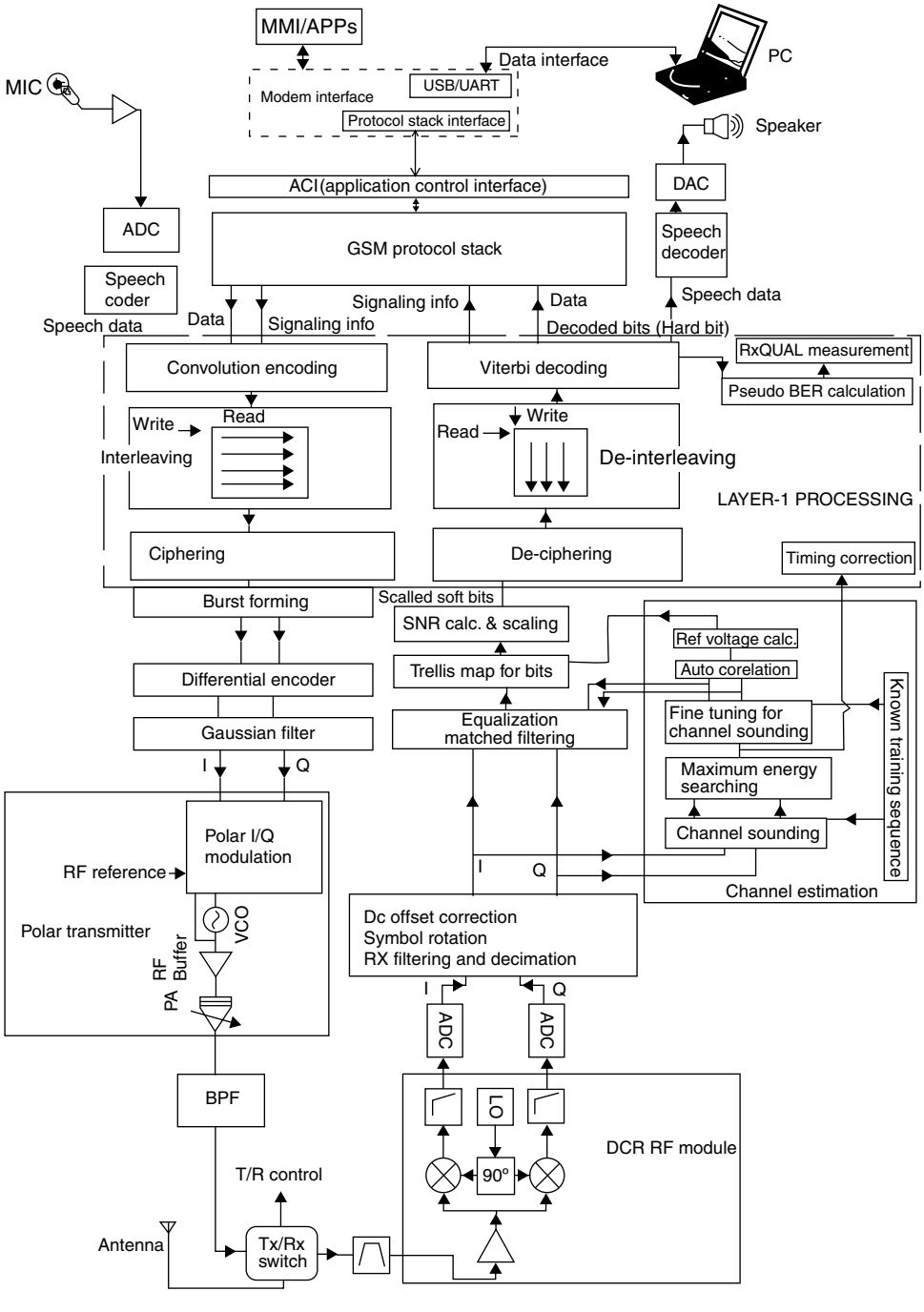


Figure 10.3 Functional block diagram of a GSM phone (modem part)

specification, the functionality should be mapped appropriately to the hardware or software blocks, which is known as hardware–software partitioning (refer to Chapter 17 for more details).

10.3 Hardware Block Diagram of a Mobile Phone

The typical hardware block diagram of a GSM mobile phone is shown in Figure 10.4. The heart of this is the baseband processor module. Generally, two processors are used for this module: one MCU unit

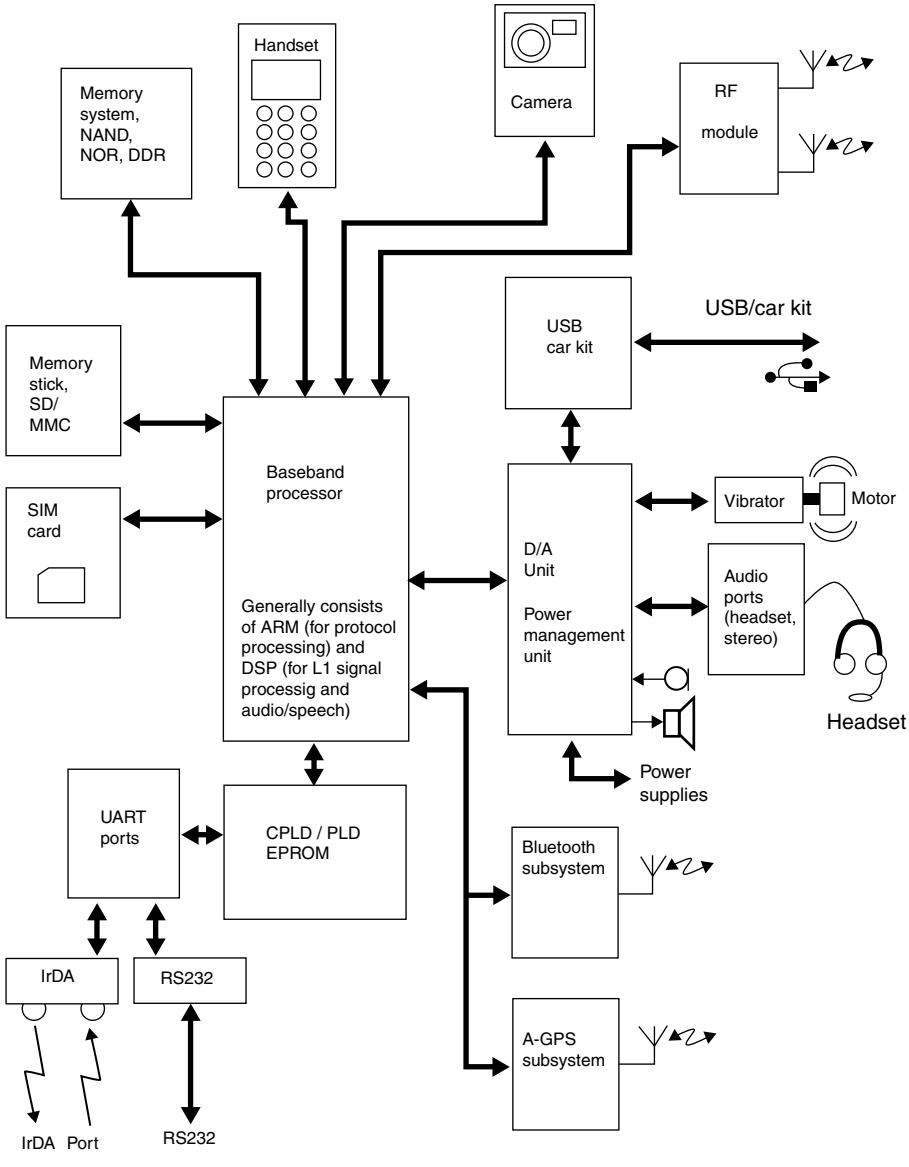


Figure 10.4 Internal hardware blocks of a GSM phone

normally an ARM (advanced RISC machine[1]) processor is used for the protocol related processing and one DSP processor [2] is used for layer-1 information and speech/audio signal processing. All the other modules are connected to this baseband module via different interfaces. For high-end phones one more application processor is used to drive the other applications.

10.4 GSM Transmitter and Receiver Module

Although the GSM specifications do not define the manner in which the transmitted information should be recovered at the BTS or MS receiver, the bursts are specifically designed for a Viterbi equalizer. The performance of the GSM receiver is tested, such that a minimum performance standard can be maintained across all GSM type approved equipment. The mobile wireless channel suffers from the various problems that we have already discussed in Chapter 2. In GSM phone, usually a polar transmitter is used as an RF transmitter (see Chapter 4) and either low-IF or zero-IF receiver architecture is used (see Figure 10.3).

Earlier in Chapter 3, we discussed GMSK modulation and demodulation in detail. Here we will discuss the bit detection module used in a typical GSM system. At the receiver side, once the digital I–Q samples are passed to the digital receiver, first it tries to estimate and compensate for the dc part. Then it goes to a burst scaling unit, which multiplies the samples by an estimated scale factor to ensure a fairly consistent dynamic range of I–Q samples from burst to burst. It tries to nullify the I–Q imbalance to make the I- and Q-branch data perfectly orthogonal. Next, it rotates the complex input signal (I–Q) by $-\pi/2$ radian per symbol (de-rotation operation to undo the rotation operation by the GMSK modulator; see Chapter 3, where for EDGE 8-psk modulation it should be $-3\pi/8$).

Now from the whole burst, pick out the training sequence (TSC) part (but at this moment as the burst is not completely time aligned, we need to take not only the 26-bit TSC part but also some more samples on both sides of the TSC position) for channel estimation to estimate the complex channel taps (see Chapter 3). Generally, the channel estimation is performed in two stages. In the first stage, channel estimation is done on non-burst time aligned received TSC data. Here, the input to the module is the received TSC and the locally stored TSC. From this, it finds out the possible set of h [] using the correlation or least square error method. Next, it does the burst timing alignment for finding out the exact burst starting position in the received burst samples from the opened RF window. Now, using this computed start position adjust the received burst data and TSC part (shifting it left or right).

Next it does the 2nd level channel estimation by taking the time-aligned received TSC and known TSC, and then estimating the fine set of channel taps from the candidate taps. The noise and interference power are estimated to further improve the channel estimate. Noise variance is needed by the LMMSE channel estimation. In LMMSE each tap is weighted by a function of tap power and the noise variance. Then the time and frequency corrected received I–Q burst data and estimated channel taps are forwarded to the equalizer unit. Maximum likelihood sequence estimation finds out the sequence of bits that maximizes the probability for best soft decisions. To handle interference limited scenarios, SAIC modules are incorporated into the receiver structure. In this instance, the sensitivity detector detects the sensitivity or interference limited scenario, based on empirical analysis of the auto regressive (AR) co-efficient or quadratic discriminator analysis (QDA), and if the interference limited scenario is detected then AR and I–Q whitening is performed before enhanced channel estimation, as shown in Figure 10.5. In the interference limited scenario the whitening filter coefficient is computed and multiplied with the channel matrix to compute the composite filter taps. The prefiltering is done on the whole burst data and passed to the equalizer. Generally, the equalizer is DFSE (non-linear with MLSE as core) or MAP type equalizer.

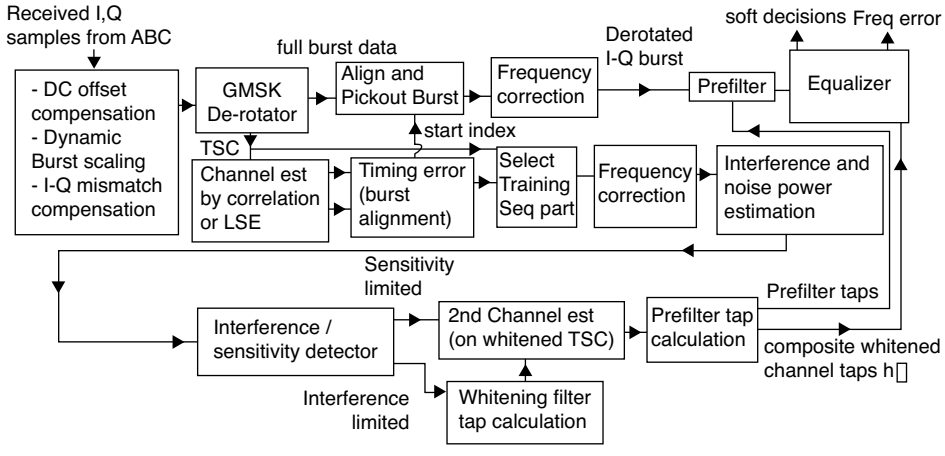


Figure 10.5 Bit detection unit for GMSK digital receiver inside a GSM phone

10.4.1 Channel Equalization

The received signal contains the original and several other reflected signals, which causes inter-symbol interference (ISI). The channel may therefore be modeled as an FIR filter, with impulse response $h_c(t)$. The output of this FIR filter will be a sum of its delayed and attenuated inputs. In GSM, except for the frequency correction burst (which is not decoded), all other bursts contain data and training sequence bits, which is represented as $s_r(t)$. This embedded training sequence inside every burst enables the receiver to perform channel estimation and equalization by computing the impact of the channel on the training sequence bits. Let us represent $r_r(t)$ as the received training sequence [3]. This can be expressed as a convolution of the transmitted training sequence, $s_r(t)$, and the channel's impulse response, $h_c(t)$: $r_r(t) = s_r(t) \otimes h_c(t)$. The samples of received training sequence in the digital domain can be represented as $r_r[n]$. This is fed into a digital matched filter with an impulse response, $h_{mf}[n]$, that is matched to $s_r[n]$. Then the matched filter output, $h_e[n]$, can be written as:

$$h_e[n] = r_r[n] \otimes h_{mf}[n] = s_r[n] \otimes h_c[n] \otimes h_{mf}[n] = R_s[n] \otimes h_c[n]$$

where $R_s[n]$ represents the auto-correlation function of $S_r[n]$. In GSM, the training sequences are designed such that $R_s[n]$ is a highly peaked real function. Therefore, $h_e[n]$ is a good estimation of the complex value $h_c[n]$. Now, let us consider that if L denotes the channel memory, then this means that the channel has $L + 1$ taps and $x[n]$ is the complex valued sample of the received signal at time n. The Viterbi equalizer finds the sequence $a[n]$ belonging to $\{-1, +1\}$ that minimizes the Euclidian metric in the equation:

$$M[n] = \sum_{l=0}^{n} |x[l] - \sum_{l=0}^{L} h_c[i].a[l-i]|^2 \approx \sum_{l=0}^{n} |x[l] - \sum_{l=0}^{L} h_e[i].a[l-i]|^2$$

Thus, $a[n]$ is the maximum likelihood sequence estimation (MLSE), which is the optimal estimation of the input symbols to the channel (for AWGN channels). This Euclidian metric can be mathematically approximated to another metric known as the matched filter metric, $M[n]$. Owing to a change in sign, we now maximize $M[n]$ in the equation below:

$$M[n] = \sum_{l=0}^{n} a[l].R\left(y[l] - \sum_{i=1}^{L} S_i.a[l-i]\right)$$

where

$$y[n] = \sum_{l=0}^{L} h_e^*[i].x[i+n]$$

and

$$S_n = \sum_{i=0}^{L} h_e^*(i).h_e(i+n),\ n = 1, 2, 3, \ldots, L$$

Here $y[n]$ is the output of $x[n]$ applied to a matched filter with an impulse response $h_e[n]$ and S_n is the n-th tap of the autocorrelation of the channel impulse response estimation. The real part of the S_n series is known as the S-parameters. As $a[n]$ can be only $+1$ or -1 and the S-parameters are known from the following equation:

$$R\left(\sum_{i=1}^{L} S_i.a[l-i]\right)$$

Hence this can take only one 2^L value and these values are denoted as the Viterbi parameters (VP).

To maximize the matched filter metric $M[n]$, all the possible $a[n]$ sequences should be tried. The Viterbi algorithm does exactly this in a very efficient way using 2^L states. In each stage, all the possible VP values are calculated (one per state). Also, at each stage two branch metrics are calculated: the first is $+(y - VP)$ and the second is $-(y - VP)$, these are known as Ungerboeck metrics. Each path in the trellis denotes the value of $M[n]$ using different $a[n]$. The MLSE path represents the $a[n]$ that maximizes $M[n]$. We know that the convolution encoder is used for information encoding during transmission and the Viterbi algorithm is used in the GSM receiver for channel decoding. The theory behind convolutional encoding and channel decoding was explained earlier in Chapter 3. The channel decoder performance significantly improves when soft input symbols are used instead of hard symbols. Therefore, the equalizer implements a soft output, which will be used as input to the Viterbi algorithm.

10.4.1.1 Impulse Response Calculation

The training sequence (TSC) is used to estimate the impulse response of the channel. Then the estimated impulse response is used for equalization, matched filtering, timing measurement, and calculating the reference values for bit-detection of the Viterbi decoder. As discussed in Chapter 7, the TS in NB consists of a 16-bit sounding sequence with 5 bits appended to either end. As TS is known at the receiver, so all possible data sequences are generated locally within the receiver then passed through the local baseband modulator, which produces a number of GMSK symbols. This is convolved with the estimated impulse response to produce a number of waveform templates. Then this is compared with the received waveform to estimate error and correct the estimated impulse response.

10.4.1.2 Matched Filtering

Complex FIR filtering is used for matched filtering. This tries to remove the inter-symbol interference caused by the channel. The same filter coefficients are estimated from the training sequence and are used for filtering the entire burst, for example, for other information bits in the burst.

10.4.1.3 Symbol Detection

In general, the symbol detection is done using a modified Viterbi algorithm. The algorithm tries to find the optimal path from start to end without trace-back. The inputs to this are matched filtered samples,

reference values from the channel estimation unit, and maximum power of the impulse response, then it outputs softbits.

10.4.1.4 De-Ciphering and De-Interleaving

These generated softbit values in the burst data are passed to de-ciphering and de-interleaving blocks.

10.4.1.5 Channel Decoding

The decoder has to find the bit sequence that generates the state sequence that is nearest to the received sequence $y[n]$. Each transition in the trellis depends only on the starting state and the end state. In GSM there are introductory logical 0 tail bits at the beginning and end of a burst data that ease the equalization process. The Viterbi algorithm is being used in convolutional decoding, which takes the soft decisions as input data, and calculates the maximum likelihood estimate for the bit sequence transmitted by the BTS and outputs the hard decisions (1 or 0 as decoded bits). This helps to correct bit-errors.

10.5 Antenna

As discussed in Chapter 1, the antenna is one of the essential components in a mobile phone. it is actually a transducer, which converts the EM wave into an electrical signal or vice versa. Generally it is a metallic object, often a rod or wire. In a mobile phone, the most commonly used main antenna is a fixed helix type and an external antenna connection is provided by the rear RF connector. Nowadays the antenna is mounted inside the mobile phone.

10.5.1 Antenna Parameters

There are several critical parameters that affect an antenna's performance and can be adjusted during the design process. These are resonant frequency, impedance, directivity, gain, aperture or radiation pattern, polarization, efficiency, bandwidth, and so on.

10.5.1.1 Resonant Frequency

Owing to the presence of parasitic elements the effective length of an antenna becomes slightly larger than its physical length. The electrical length of an antenna is usually the physical length of the wire/dipole divided by its velocity factor (the ratio of the speed of wave propagation in the wire to the speed of light in a vacuum). The "resonant frequency" and "electrical resonance" is related to the electrical length of the antenna. Typically an antenna is tuned for a specific frequency and is effective for a range of frequencies (BW) usually centered on that resonant frequency. However, the other properties of the antenna (especially the radiation pattern and impedance) change with frequency, so the antenna's resonant frequency may merely be close to the center frequency to satisfy those other important properties. Antennas can be made resonant on harmonic frequencies with lengths that are fractions of the target wavelength. Some antenna designs have multiple resonant frequencies, and some are relatively effective over a very broad range of frequencies. The most commonly known type of wideband aerial is the logarithmic or log periodic, but its gain is usually much lower than that of a specific or narrower band aerial.

10.5.1.2 Polarization

The polarization of an electromagnetic wave is defined as the orientation of the electric field vector with respect to ground. The electric field vector is perpendicular to both the direction of propagation and the magnetic field vector. The "polarization" of an antenna is the orientation of the electric field of the radio wave with respect to the Earth's surface and is determined by the physical structure of the antenna and its orientation. Generally four types of polarization are used: horizontal, vertical, circular, and elliptical. A vertical antenna will have vertical polarization and a horizontal antenna will have horizontal polarization. In circular polarization, the antenna continuously varies the electric field of the radio wave through all possible values of its orientation with regard to the Earth's surface. Circular polarizations, as elliptical ones, are classified as right-hand polarized or left-hand polarized using a "thumb in the direction of the propagation" rule. In practice, regardless of any other parameters, it is important that linearly polarized transmitter and receiver antennas should be matched, for example, horizontal should be used with horizontal and vertical with vertical. Transmitters mounted on vehicles with large motional freedom commonly use circularly polarized antennas so that there will never be a complete mismatch with signals from other sources.

10.5.1.3 Impedance

When an incoming electromagnetic wave or outgoing RF signal travels through the different parts of the antenna system, it may encounter differences in impedance (E/H, V/I, etc.) between the input and output port. Then depending on the impedance match, some fraction of the input energy will reflect back towards the source, forming a standing wave in the feed line. The ratio of maximum power to minimum power in the wave is called the standing wave ratio (SWR). Although an SWR of 1 : 1 is ideal an SWR of 1.5 : 1 is also considered to be marginally acceptable in low power applications. The complex impedance of an antenna is related to its electrical length at the wavelength in use. The impedance of an antenna can be matched to the feed line and radio by adjusting the impedance of the feed line, using the feed line as an impedance transformer. More commonly, the impedance is adjusted at the load with an antenna tuner, a balun, which is a matching transformer, for matching networks composed of inductors and capacitors or matching sections.

10.5.1.4 Radiation Pattern

The radiation pattern of an antenna is a graphical representation of the radiated fields or power along different directions in the space. When radiation is expressed as field strength, E V/m, it is called a field strength pattern. If the power per solid angle is plotted in 3-D space, it is called power pattern. A power pattern will be the product of electric and magnetic field patterns or proportional to the square of the field strength pattern. The radiated field pattern looks like a conical lobe. This lobe has angular width. These are measured as the null beam width and half power beam width. As shown in Figure 10.6, the beam width goes to zero at point p. Now if we draw a tangent to this beam surface at both sides from point p, the angular separation between these two tangents (θ) will be the measure of the null beam width. The half power points are the points in the radiated field pattern where the field reduces to $1/\sqrt{2}$ times its peak value (the power becomes half). The angular separation between the two half power field points is called half power beam width.

10.5.1.5 Efficiency of an Antenna

The radiation efficiency of an antenna is defined as the ratio of radiated power to the total input power.

$$\eta = P_{\text{radiated}}/P_{\text{input}}$$

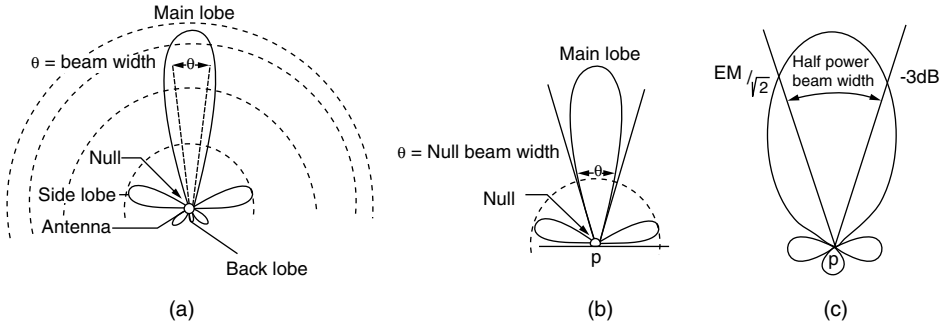


Figure 10.6 (a) Radiation pattern, (b) null beam width, and (c) half-power beam width

10.5.1.6 Directivity and Gain

The directive gain, $D(\theta, \Phi)$ of an antenna is a measure of the concentration of the radiated power along a particular direction θ, Φ. If $U(\theta, \Phi)$ is the radiation intensity, which is the measure of concentrated radiation, then the radiated power will be:

$$P(\text{rad}) = \oint_S U(\theta, \phi)\text{d}\Omega$$

where

$$\text{d}\Omega = \sin\theta \; \text{d}\theta \; \text{d}\Phi$$

is an element of solid angle Ω. The total radiated power is therefore the integral of the radiation intensity over a solid angle 4π. Thus, the average radiated intensity $U(\text{avg}) = P(\text{rad})/4\pi$. An isotropic antenna radiates EM power uniformly in all directions. However, a non-isotropic antenna concentrates power along a particular direction. If $U(\theta, \Phi)$ is the radiated power intensity along the direction (θ, Φ) then the directivity $D(\theta, \Phi)$ of the antenna is defined as the ratio of maximum radiation intensity along this particular direction to the average radiation intensity.

$$D(\theta, \Phi) = U(\theta, \Phi)_{\text{max}}/U(\text{avg}) = [P(\text{rad})/\Omega_{\text{A}})/(P(\text{rad})/4\pi] = 4\pi/\Omega_{\text{A}} \cdot \approx 4\pi/(HP_{\text{E}}{*}HP_{\text{H}})$$

where HP_{E} and HP_{H} are the half-power beam width of the electric and magnetic fields, respectively. The power gain is defined as the product of the directivity and efficiency of the antenna, $G(\theta, \Phi) = \eta\, D(\theta, \Phi)$. It accounts for the losses of the antenna. Generally, gain is expressed in dB, where $G(\text{dB}) = 10 \log_{10} G$. Equivalently, we can say that the gain of an antenna is the gain over an isotropic antenna, $G(\text{dB}) = 10 \log_{10}(G/G_{\text{o}})$.

10.5.2 Conventional Mobile Phone Antennas

In the early days, conventional mobile phones used either whip or helical antennas that extend from the top of the mobile handset, or else they were contained within the upper part of the handset. Today, there are several different types of antenna are used in mobile phones and some of these are discussed below.

10.5.2.1 Planar Inverted F Antennas

For many years, planar inverted F antennas (PIFAs) were used in mobile phone handsets. Figure 10.7 shows a simple single band PIFA. This has a low profile resonant element of about a quarter of a wavelength long. During operation, currents oscillate in the inverted L section. The impedance of this type of antenna is determined by the position where the feed is connected along the L section.

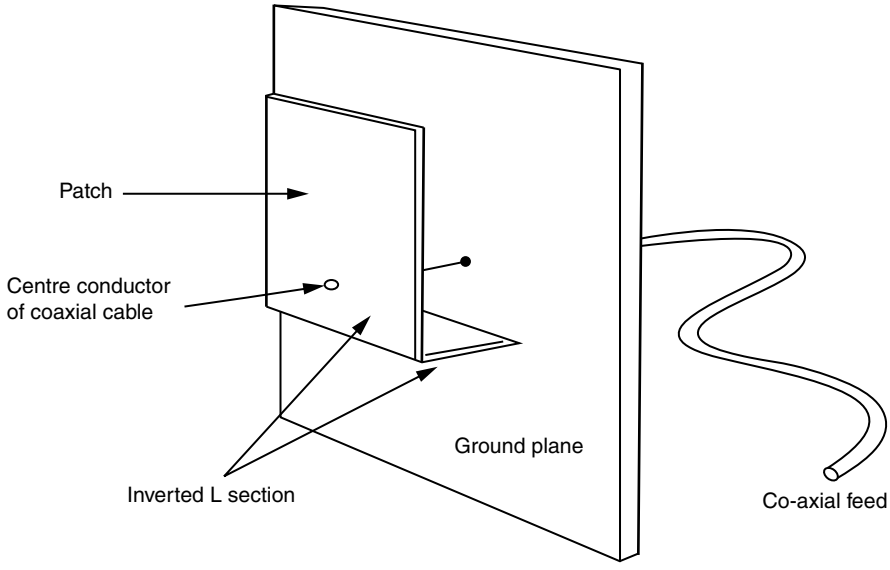


Figure 10.7 A planar inverted F antenna

10.5.2.2 Helical Antenna

A helical antenna consists of a conducting wire wound in the form of a helix as shown in Figure 10.8. Generally, helical antennas are mounted over a ground plane. It radiates when the circumference of the helix is of the order of at least one wavelength. The radiation along the axis of the helix is found to be the strongest. Generally, this type of antenna is directional. This is a simple antenna type and offers high-gain and broadband frequency characteristics. The radiation from a helical antenna is circularly polarized (clockwise or counter-clockwise). A helical antenna usually has two operating modes. (1) Normal mode (broadside) – in this mode the dimensions of the helix are small compared with the wavelength. The far field radiation pattern is similar to an electrically short dipole or monopole. These antennas tend to be inefficient radiators and are typically used for mobile communications where reduced size is a critical factor. (2) Axial mode (end fire) – in this mode the helix dimensions are at or above the wavelength of operation. This is the ame as waveguide antennas, and produces true circular polarized waves. These antennas are best suited for space communication, where the orientation of the sender and receiver cannot be easily controlled, or where the polarization of the signal may change. However, owing to the large antenna size these are not very popular in mobile handsets. Terminal impedance in the axial mode ranges between 100 and 200 Ω. The resistive part is approximated by: $R \approx 140\ (C/\lambda)$, where R is resistance in ohms, C is the circumference of the helix, and λ is the wavelength. Impedance is matched with the cable C by a short strip-line section between the helix and the cable termination. The maximum directive gain can be expressed as: $D_o \approx 15\,N\,(C^2S/\lambda^3)$. The approximate bandwidth for a helical antenna is: $0.75 * \lambda$–$1.3 * \lambda$.

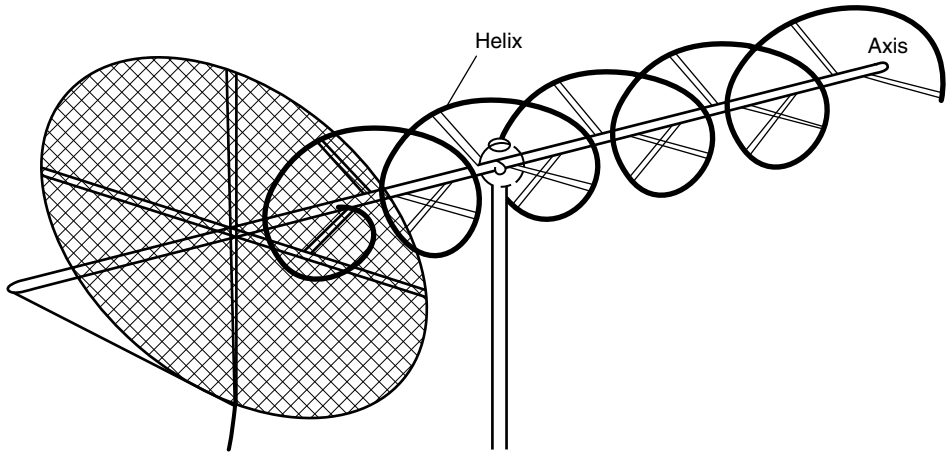


Figure 10.8 Helical antenna

10.5.2.3 Whip Antenna

A whip antenna is a single-element antenna that can be used with an unbalanced feed line, such as coaxial cable, or attached directly to a wireless transceiver. This antenna is mostly used on handheld two-way radios and cell phones. These are usually attached to a vehicle and designed to be flexible, so that they do not break when struck. The name is derived from the whip-like motion when perturbed. The whip resembles a ground-plane antenna without the radial system. Generally, the whip antennas are short, flexible "rubber duck" type, whereas in some cases long, flexible, stainless-steel material is also used. The whip antenna's electrical and mechanical design is very simple and it is very easy to install. However, they are not very efficient as most whip antennas are operated with a poor electrical ground system. The whip antenna is a stiff but flexible wire mounted, and generally vertical orientation is used, with one end adjacent to a ground plane. This can also be called a half-dipole antenna, and it creates a toroidal radiation pattern, where the axis of the toroid centers about the whip. The length of the whip determines its wavelength, although it may be shortened with a loading coil anywhere along the antenna. Whips are generally a fraction of their actual operating wavelength. There may be some uncertainty about the biological safety of this type of antenna.

10.5.2.4 Microstrip Patch Antennas

Microstrip patch antennas are commonly used in mobile communication terminals due to their many attractive features, such as simple structure, low production cost, light weight, smaller size, and robustness. These antennas are planar resonant cavities that leak from their edges and radiate as shown in Figure 10.9. Microstrips consists of a metal strip on a dielectric substrate (ε_r) covered by a ground plane on the other side. Unlike stripline, the single ground plane shields the circuit on one side only and on the other side there is air. In this inhomogeneous type of structure pure TEM cannot exist, as it is not possible to satisfy the boundary conditions for the TEM mode at the surface of the dielectric and air. The EM field lines in the microstrip are not contained entirely inside the substrate. Impedance match occurs when a patch resonates as a resonant cavity, and when it is matched the antenna achieves peak efficiency. A normal transmission line radiates little power because the fringing fields are matched by nearby

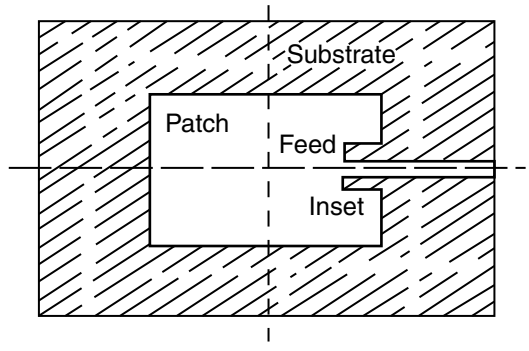


Figure 10.9 Micro-strip antenna

counteracting fields. Power radiates from open circuits and from discontinuities such as corners, but the amount depends on the radiation conductance load to the line relative to the patches. Without proper matching, only a small amount of power radiates into space.

A dielectric slab on a ground plane will support TM modes when it is thin and TE modes when it is thick. The TM mode is polarized normal to the slab surface, whereas the TE mode is polarized parallel to the slab surface. As today's standard mobile terminals operate in many frequency bands, for example, GSM850/900/1800/1900, so multi-band antenna elements are required. This is easier to achieve using microstrip antenna, as this type of antenna can also be easily fabricated by photolithographic process and is easily integrated with other passive and active microwave devices.

10.6 Analog to Digital Conversion (ADC) Module

The analog to digital conversion unit is one of the important components in a digital mobile phone. Generally the RF front-end unit processes the analog signal, which needs to be converted into a digital signal for baseband processing, and for this purpose one ADC circuit is used. On the transmit side, the source signal from the microphone is an analog signal, which is also converted into a digital signal using ADC and is given to the source coder (speech codec). As discussed in Chapter 1, there are many types of ADC available. However, of these the sigma delta converters have become very popular and most widely used in mobile phone receivers. The key feature of this converter is that it is the only low cost conversion method that provides both high dynamic range and flexibility in converting low bandwidth input signals. A simple block diagram of a first order sigma delta analog-to-digital converter (ADC) is shown in Figure 10.10. The input signal X comes into the modulator via a summing junction. After that, it passes through the integrator, which feeds a comparator that acts as a one-bit quantizer. Then, the comparator output is fed back to the input summing junction via a 1-bit digital-to-analog converter (DAC). The same signal also passes through the digital filter and emerges at the output of the converter (Y). The feedback loop forces the average of the signal W to be equal to the input signal X.

Noise Shaping – Moving the quantization noise from the band of interest to outside this band is referred to as noise shaping. Assuming that the noise type is AWGN, we can use feedback to remove the noise from low frequencies (for example, 0–4 kHz for the voice band) at the cost of increasing the noise at higher frequencies (out of the desired signal band). As shown in the Figure 10.10, this is done by imbedding a filter and the D/A converter in a feedback loop. The noise shaping filter or integrator of a sigma delta converter distributes the converter quantization error or noise such that it is very low in the band of interest.

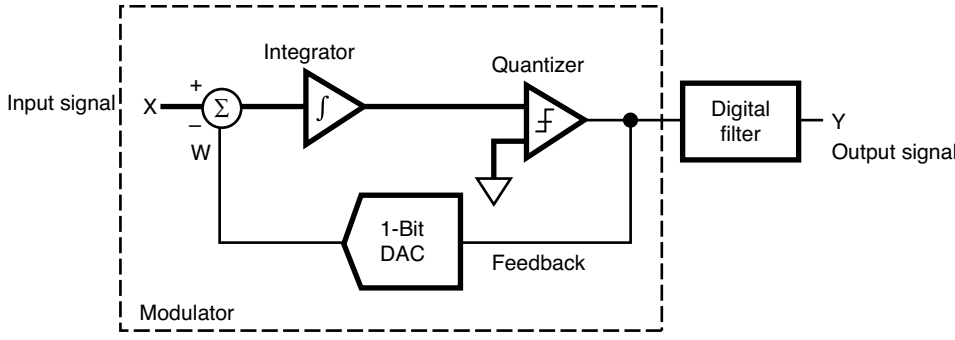


Figure 10.10 First order sigma–delta ADC

Over-Sampling – This is simply the act of sampling the input signal at a frequency much higher than the Nyquist frequency. As discussed in Chapter 1, over sampling decreases the quantization noise in the band of interest.

Digital Filter – As shown in Figure 10.10, an on-chip digital filter is used to attenuate signals and noise that are outside the band of interest.

Decimation – As the sigma-delta converter samples at a much higher rate, so the generated sampling data volume is much higher. The act of reducing the data rate down from the over-sampling rate without losing information is known as decimation. This process is used in a sigma-delta converter to eliminate redundant sampled data at the output. The sampling theorem tells us that the sample rate only needs to be twice the input signal bandwidth in order to reliably reconstruct the input signal without distortion. However, in this instance the input signal was heavily over-sampled by the sigma-delta modulator in order to reduce the quantization noise. Therefore, there are redundant data that can be eliminated without introducing distortion to the conversion result. The decimation process simply reduces the output sampling rate, while retaining the necessary information.

It should be noted that sigma-delta modulation [4] only alters the spectral properties of the quantization noise by shifting the noise power to the high frequency domain (Figure 10.11). However, it still needs to be removed from the output signal by means of low-pass filtering. In fact, the total amount of quantization noise increases for higher modulation orders. As discussed, the filtering is achieved by means of a decimation filter, which also reduces the sampling rate and thereby reduces the number of samples to be processed inside the baseband processor.

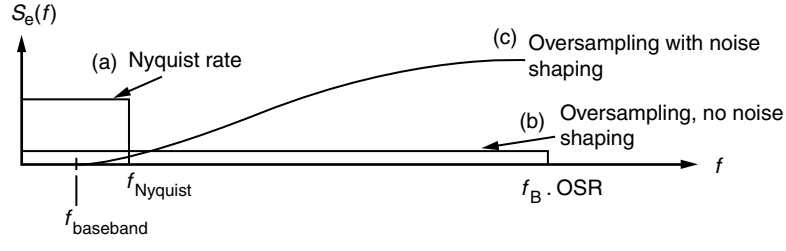


Figure 10.11 Spectral properties of quantized noise for: (a) Nyquist rate sampling, (b) oversampling with no noise shaping, and (c) Nyquist rate with oversampling

10.7 Automatic Gain Control (AGC) Module

Owing to noise and channel characteristics fluctuations, the received signal strength at the receiver changes from maximum to minimum or vice versa. AGC is widely used in communication systems to maintain constant signal strength by varying the amplifier gain. Apart from the amplitude variation in the speaker volume, the variation in the received signal strength leads to several other issues. (1) The performance of the amplifier (LNA and PA) circuit changes. This is because when the signal strength is high, this leads to saturation issues and when the signal strength is low this leads to poorer amplification. (2) The tolerances to the power levels in different components of the transmitter–receiver chain are not the same, so this may cause damage to the circuit. (3) ADC or DAC has a dynamic range, if the signal strength varies beyond that, then there will be an error. Under extreme conditions of voltage and temperature the mobile phone should work properly with AGC loop circuitry and with guaranteed component tolerances. In Figure 10.12, a typical AGC loop circuit is shown. This is a feedback system comprising a forward gain stage (A), feedback gain (β), and a signal comparison stage that generates a differential error signal. The AGC loop is analyzed in terms of its closed-loop gain (forward transfer function) and open-loop gain. R(s) and C(s) represent the input and output amplitude. There is a gain stage in the comparator and temperature compensation at the detector diode stage to compensate for diode detector forward voltage variation with temperature.

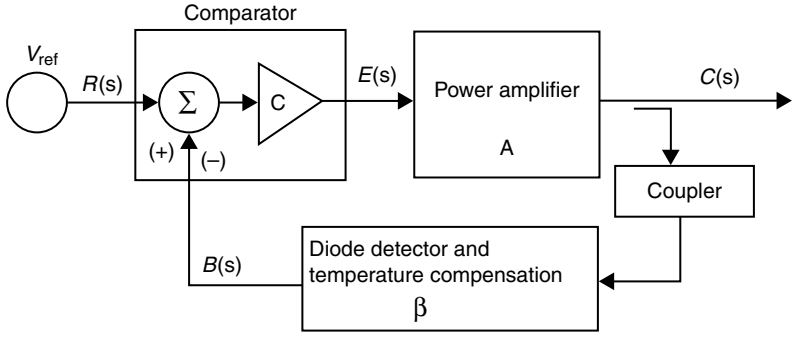


Figure 10.12 AGC loop diagram and control loop components used in a mobile

Most commonly, AGC is implemented using an IF amplifier, a voltage controlled amplifier controlled with analog voltage. In this instance, the amplifier attenuates more to the received signal with high amplitude or vice versa. The dynamic range is determined by the minimum carrier to noise ratio and blocking signal level at the ADC input. Generally, the dynamic range of a receiver is often limited by the dynamic range of the analog-to-digital converter (ADC). Automatic gain control (AGC) can be used to dramatically reduce the required dynamic range. The system designers know that adding variable gain is much less expensive than increasing the dynamic range of the ADC. The dynamic range of the receiver (reception window) is typically defined to be restricted above (15 dB) and below (20 dB) a specific reference level. The AGC function can be placed in the first stages of the receiver, after the RF conversion. However, today, the gain is controlled in several stages at the front-end and also at the baseband module through a software algorithm to produce a better result. In a typical mobile phone circuit, to maintain a constant output level, pre-monitoring is used. This pre-monitoring is done in three phases and determines the settling time for the RXAGC. The receiver is switched on approximately 150 s before the burst begins and processor measures the received signal level and then adjusts the AGC-DAC in accordance with the measured signal level and/or switches on/off the LNA with the front-end amplifier control line. The requirement for the received signal level under static conditions is that the MS should measure and report to the BS over the range -110 dBm to -48 dBm.

10.8 Automatic Frequency Correction Module

The primary requirement of the AFC is to keep the local transmitter (TX) frequency stable within certain limits and the frequency error should be low enough to ensure a good receiver performance. To establish and maintain a robust wireless connection, the reference oscillator frequency must attain high levels of precision and accuracy. As accuracy over time is very critical in a wireless phone design, the reference oscillator inside a mobile phone must be able to compensate for both static and dynamic errors. The frequency deviations can be caused primarily due to temperature drift, initial crystal offset, Doppler shifts, and aging.

Most of the GSM handsets use voltage-controlled temperature-compensated crystal oscillator (VC-TCXO) modules as the system reference oscillator. VC-TCXO modules use analog techniques to correct the frequency deviations. However, the problems are that VC-TCXOs bring a high price tag, a large footprint, and require some other external components along with them. These issues are creating real design challenges in the competitive mobile handset market. At present, one alternative solution to this problem is to use a digitally controlled crystal oscillator (DCXO). New GSM mobile phone transceiver architectures are being developed that house digitally controlled crystal oscillators (DCXOs) and helps to eliminate the headache of adding a VC-TCXO to the mobile phone architecture.

10.8.1 Analog VC-TCXO

The VC-TCXO module incorporates temperature compensation circuitry, which is typically comprised of a simple control loop using a thermistor circuit. The thermistor network biases an internal varactor diode to attain the correct crystal load capacitance in order to maintain the target oscillation frequency. The bias level of the varactor diode changes with temperature as the thermistor resistance changes. This helps to compensate for the temperature effect on the crystal frequency.

As shown in Figure 10.13a, VC-TCXO also includes an external tuning voltage input to finely control the varactor diode to compensate for frequency errors other than temperature drift. To ensure the specified

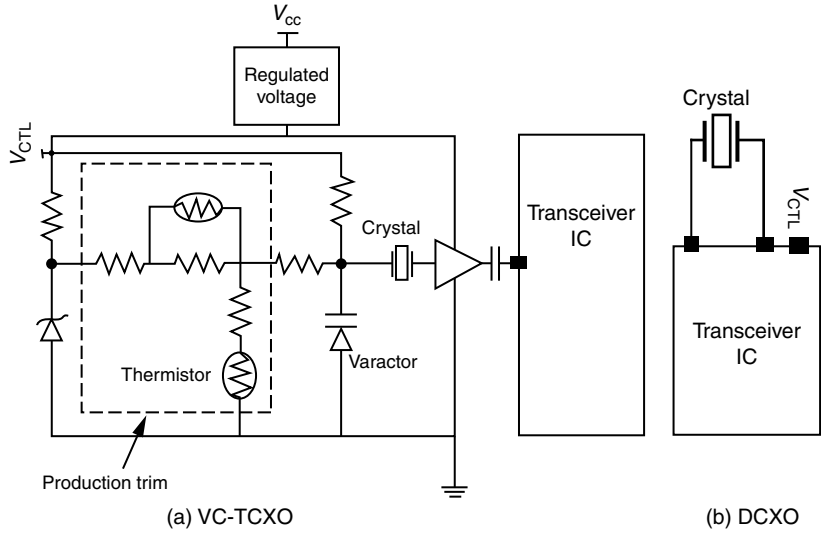


Figure 10.13 (a) Block diagram of a typical VC-TCXO circuit, and (b) block diagram of a typical DCXO implementation

precision and accuracy, each resistor must be production trimmed to offset each crystal's unique static error. In the VC-TXCO, a low dropout (LDO) voltage regulator is used for control and to stabilize the supply voltage. Depending on the implementation, the supply voltage may "push" the oscillation frequency away from the target and induce errors. The additional internal components and manufacturing steps significantly increase the cost of the module.

10.8.2 Digitally Controlled Crystal Oscillators – DCXO

As we are moving towards the digital world, the DCXO is replacing the analog VC-TXO. In the past, many handset designers have developed discrete DCXOs to avoid using costly VC-TCXO modules and to reduce the overall bill of material (BOM) cost. DCXOs compensate for frequency errors using a combination of digital and analog circuitry. In a GSM/GPRS mobile phone design, a DCXO can replace the VC-TCXO function with a standard AT-cut uncompensated crystal resonator. The digital circuit monitors the frequency deviation and constantly controls the deviation through circuit component adjustment. Frequency errors may be controlled by a software program using a control loop topology. Implementations are chip dependent and can be effected using a variety of configurations and different methods.

10.8.2.1 Working Principle of DCXO

Conceptually, based on the frequency measurement calculations by a transceiver software program (the deviation of the local frequency to the network frequency), a DCXO "pulls" the crystal frequency to the required target value. As shown in Figure 10.13b, a digitally configurable interface can be used to programmatically add or subtract load capacitance to the oscillator circuit to change the resonance frequency, which is particularly useful for correcting static errors. With the DCXO approach, the error compensation circuitry and voltage regulation are integrated into the IC.

DCXOs must compensate for both dynamic and static errors. In addition, the DCXO must be able to continuously adjust the frequency in incremental steps so that a final frequency error of 0.1 ppm or lower is achieved throughout. As the DCXO circuit does not include temperature sensors, they heavily rely on the frequency measurement in the digital controller. Generally, this involves three processes:

- **Frequency estimation** – This is the process of estimating the complex frequency components of a signal in the presence of noise or channel impairments. Various estimation methods can be used for frequency, such as: time-domain periodicity estimation, spectral pattern matching, frequency-domain periodicity, and so on. The different algorithms used commonly for frequency estimations are: maximum likelihood (ML) and approximate ML, Fourier coefficient, filtering techniques, signal subspace, noise subspace, phase weighted averagers.
- **Frequency compensation** – Once the frequency is estimated, the deviation of the local clock frequency from the estimated frequency is then computed and this computed frequency deviation is compensated by using various methods, such as: changing the varactor diode voltage, resistance value changes, and so on.
- **Frequency tracking** – After the frequency is estimated and deviation is computed, then it has to be constantly monitored and tracked to keep the deviation under a certain limit.

10.8.3 AFC Implementation for a Typical GSM Handset

The drift in frequency can cause several problems in a GSM phone, such as: (1) variation in the instant of sampling so that the signal is not sampled at the correct moment that it was supposed to be, and this will lead to bit decoding error; and (2) frequency and time synchronization issues, this will lead to improper

burst reception. Thus, to avoid the drift in output frequency, the GSM phone has to use VC-TCXO or VCXO (voice control crystal oscillator) for output frequency corrections. Generally, for high-end phones (expensive phones) VC-TXO and low-end phone DCXO are used. In typical GSM phone implementation, the frequency is estimated and tracked continuously to keep the frequency error within 0.1 ppm. During initial synchronizing (camp on), where the frequency error can be relatively large, the frequency control burst (FB) is mainly used to compute the frequency error. As discussed in Chapter 8, the BTS transmits the FB on the frequency control channel (FCCH). The handset receives the FB, calculates the frequency error, and adjusts the frequency accordingly. As the frequency adjustment is comprehensive, this helps to eliminate the need for special sensors in DXCO, provided that the DCXO can compensate for the full range of errors.

The frequency correction loop is illustrated in Figure 10.14. The carrier frequency is received by the RF module, but due to inaccuracy of the VC-TCXO/DCXO a frequency offset (or error) is generated at the baseband frequency. The baseband signal is sampled and then complex de-rotated. We know that after down-mixing the FB will be a pure sine wave (PSW) with a frequency of 67.71 kHz + frequency offset (as explained in Chapter 7). After a complex rotation the FB is seen as a PSW with a frequency equal to the frequency offset. The algorithm estimates this frequency offset and the adaptive AFC algorithm calculates the required correction value, which is then mapped to DAC to generate a control voltage for the VC-TCXO (or varactor tuning for DCXO). As the error value fluctuates randomly with time, so this error value is usually passed through a filter to smooth it before it is passed to the DAC.

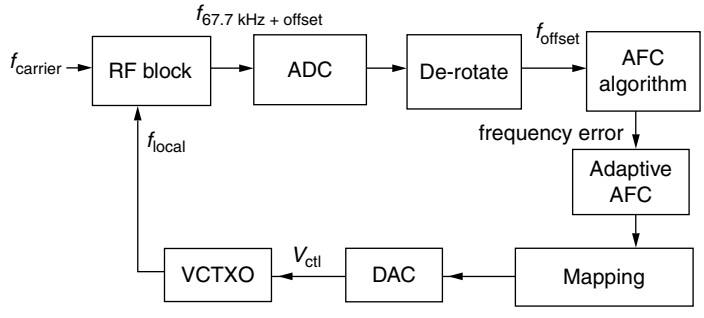


Figure 10.14 Automatic frequency corrections (AFC) loop

Once the initial synchronization is achieved, then this needs to be tracked periodically. For that purpose, the frequency error is generally estimated from every normal burst (NB) received via the BCCH, CCCH or TCH channel.

10.9 Loudspeaker

A speaker takes the electrical signal as input and translates it back into physical vibrations to create sound waves. A conventional diagram of a loudspeaker is shown in the Figure 10.15.

In 1876, Alexander Graham Bell patented the first loudspeaker as part of his telephone circuit. The modern design of moving-coil drivers based loud speakers was proposed by Oliver Lodge in 1898. Generally, a speaker uses a lightweight diaphragm connected to a rigid frame via flexible suspension. The flexible suspension part constrains a coil of fine wire to move axially through a cylindrical magnetic gap. The diaphragm is usually manufactured using paper, metal or plastic in a cone or dome shaped profile. The suspension system helps to keep the coil centered in the gap and provides a restoring force to make the speaker cone return to a neutral position after moving back and forth.

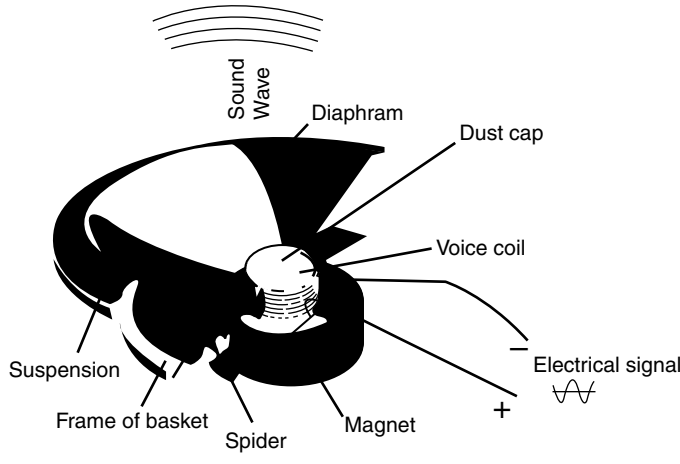


Figure 10.15 Internal diagram of a loudspeaker

A typical suspension system consists of two parts. (1) The spider – this connects the diaphragm or voice coil to the frame and provides the majority of the restoring force. It is usually made of a corrugated fabric disk. (2) The surround – this helps to keep the coil in the center and also allows free movement. The surround can be a roll of rubber or foam, or a corrugated fabric, attached to the outer circumference of the cone and to the frame. The narrow end of the cone is connected to the voice coil. The voice coil wire is round, rectangular or hexagonal in shape and is usually made of copper, aluminum, or silver. Passing an electrical current through the wire creates a magnetic field around the coil, magnetizing the metal it is wrapped around. The electromagnet and the permanent magnet interact with each other. The input alternating current causes the polar orientation of the electromagnet to reverse itself many times a second; this constantly reverses the magnetic forces between the voice coil and the permanent magnet, which pushes the coil back and forth rapidly, like a piston. This vibrates the air in front of the speaker, creating sound waves. A woofer is a driver that is capable of reproducing low (bass) frequencies. A tweeter is a driver that is capable of reproducing the high (treble) frequencies.

The efficiency of a loudspeaker is defined as the sound power output (usually specified in dB, and is known as the sensitivity of the speaker) divided by the electrical power input. The impedance of a speaker (typically 4 Ω, 8 Ω, etc.) is matched with the audio amplifier load to obtain the maximum power transfer. The rated power of a speaker is defined by two terms – the nominal power (continuous) and peak (or maximum short-term) power. These terms are important as they define the maximum input power that the loudspeaker can handle before it is thermally destroyed.

10.10 Microphone (MIC)

A microphone is an acoustic to electric transducer or sensor that converts sound (air pressure variations) into electrical signal (current variations). Sometimes it is also referred to as a mike or mic.

10.10.1 Principle of Operation

A sound wave is generated by a source; it creates contraction and rarefaction in the air medium and propagates. When this strikes a microphone surface, it produces vibration and from that a voltage/current

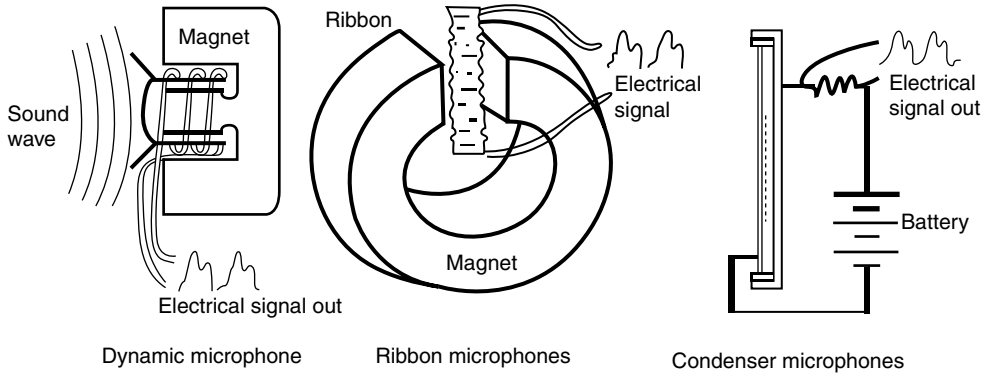


Figure 10.16 Dynamic microphone, ribbon microphone, and condenser microphone

is generated, which is proportional to the sound signal amplitude. A variety of mechanical techniques can be used for constructing microphones (Figure 10.16). The two most commonly used designs are the magneto-dynamic and the variable condenser designs. Typically, for speech/audio applications, we use dynamic, ribbon, or condenser microphones.

1. **Dynamic Microphone:** Sound wave vibrates the attached coil of wire in the field of a magnet. This produces a voltage that replicates the sound pressure variation – characterized as a pressure microphone.

   Advantages – (i) relatively cheap and rugged, and (ii) can be easily miniaturized.

   Disadvantages – the uniformity of response to different frequencies is worse than the ribbon or condenser microphones.
2. **Ribbon Microphone:** In this type of microphone, the air movement associated with the sound wave moves the metallic ribbon in the magnetic field, which generates an imaging voltage between the ends of the ribbon. This voltage is proportional to the velocity of the ribbon, hence this type of microphone is characterized as a "velocity" microphone.
3. **Condenser Microphone:** Here the sound pressure changes the spacing between a thin metallic membrane and the stationary back plate. The plates are charged to a total charge of:

$$Q = CV = [\alpha.\,(\text{area of plate}).\,\text{voltage}]/[\text{plate spacing}]$$

   where C is the capacitance and V the voltage of the biasing battery. A change in plate spacing will cause a change in charge Q and force a current through resistance R. This current replicates the sound pressure, making this a "pressure" microphone. Condenser microphones span the range from cheap throw-aways to high-fidelity quality instruments.

   Advantages – this offers the best overall frequency response, so it is selected for many recording applications.

   Disadvantages – (i) expensive, (ii) may pop and crack when it is close to the sound source, and (iii) requires a battery or external power supply to bias the plates.
4. **Carbon Microphone:** This type of microphone is a capsule containing carbon granules pressed between two metal plates. A voltage is applied across the metal plates, causing a small current to flow through the carbon. One of the plates is a diaphragm which vibrates when sound wave strike on it and produces a varying pressure to the carbon. The changing pressure deforms the granules and this causes

the change in contact area between each pair of adjacent granules, which leads to a change in electrical resistance of the mass of granules. The changes in resistance causes a corresponding change in the voltage across the two plates, and this is output from the mic as the electrical signal. Carbon microphones were formerly used in telephone handsets. This has extremely low-quality sound reproduction.

10.10.1.1 Characteristics of a Microphone

There is no inherent advantage in fidelity of one specific type of microphone over another. Condenser types require batteries or power from the mixing console to operate, and dynamic types require shielding from stray magnetic fields, which makes them a bit heavy sometimes. The most important factor in choosing a microphone is based on the application, size, and quality requirement. The following parameters must be considered for appropriate selection.

1. **Sensitivity:** Sensitivity of a mic is a measure of how much electrical output is produced by a given sound.
2. **Overload Characteristics:** When it is over driven by loud sound, the microphone produces distortion. This is caused by various factors such as loud sounds, coil pullout from the magnetic field, and amplifier clipping.
3. **Linearity or Distortion:** The distortion characteristics of a mic are determined mostly by the care with which the diaphragm is made and mounted.
4. **Frequency Response:** A flat frequency response is always desirable.
5. **Noise:** Microphones produce a very feeble current that requires amplification by a factor of more than 100. Now, any electrical noise produced by the microphone will also be amplified, so it should be noise free. Dynamic microphones are essentially noise free, but the electronic circuit built into condenser types is a potential source of trouble, and must be carefully designed and constructed of premium quality parts.
6. **Impedance Matching:** Microphones have an electrical characteristic termed impedance, which is measured in ohms (Ω) and this depends on the design. Typically, it can vary from 600 to 10 kΩ. To get the best sound, the impedance of the microphone must be matched with the load to which it is connected.

Typical GSM mobile phone implementation of a voice sampling circuit is shown in Figure 10.17.

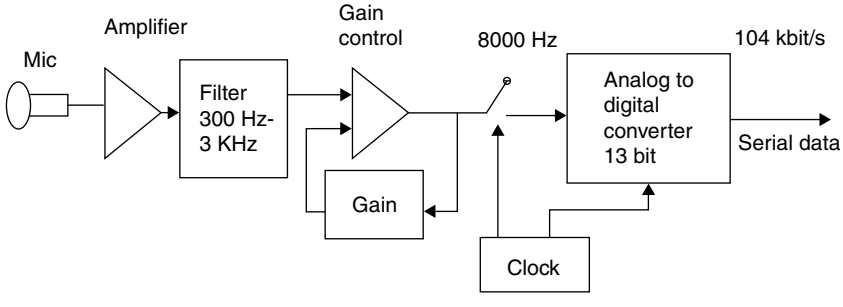


Figure 10.17 Voice sampling

10.11 Subscriber Identity Module (SIM)

The subscriber identity module (SIM) is a small smart card, which contains microprocessor, memory, programs, and information (Figure 10.18). The SIM card contains a microprocessor chip inside it, which stores unique information about the user's account, including the user's phone number, and identifies the user to the network. Therefore, it is not the cell phone that determines the telephone number, rather it is the SIM card. Subscribers activate their phones by inserting their SIM cards into the phone. Once the SIM is removed from the phone, then the phone can not be used for making calls except for some emergency calls. One of the advantages of the GSM system architecture is that the phone is not tightly coupled with the SIM, for example, the SIM can be moved from one mobile phone to another. This makes upgrades very simple for the GSM mobile phone user.

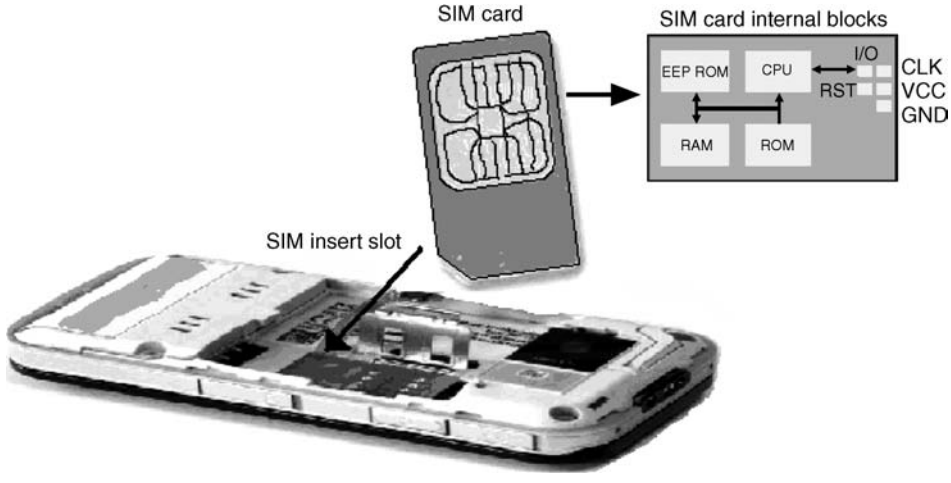


Figure 10.18 Blocks inside a SIM card

The subscriber identity module serial number (SSN) is a 19 or 20 digit unique number that identifies an individual SIM card. Typically, a SIM card has 16–64 kb of memory, which provides plenty of room for storing hundreds of personal phone numbers, text messages, and value-added services. SIM cards are available on a subscription basis, generally network operators in specific geographic locations provide those. A user can sign a contract with a provider and receive a monthly bill or cards are available on a prepaid basis, in which case users buy airtime as required. Generally, GSM specifies two types of SIM cards. The first type, an ID-1 card, is same size as a standard credit card and has embossing, a picture, lettering and a magnetic stripe similar to a credit card. Some larger GSM phones use this type. Another type is called a plug-in SIM card. An ID-1 card can be converted into a plug-in SIM card by removing the plastic holder. The ID-1 has the dimensions 54 mm $\times$ 85.6 mm, whereas a plug-in SIM has the dimension 15 mm $\times$ 25 mm.

The memory on a SIM card consists of a master file (MF), and this has two dedicated files (DF). The DFs contain elementary files (EF), which contain actual GSM data – each EF contains one record, which could be information such as a phone book or the IMSI (international mobile subscriber identity). Record sizes are measured in words, with one word containing 8 bits (1 byte). These records also contain information set by the operator that controls what feature services are enabled, such as SMS, ISDN, and fixed dialing.

The microprocessor-based SIM platform is designed to be secure. Attempts to reverse engineering may damage the card permanently. Certain data can be changed by the manufacturer only. Other data can be changed by the user by entering an appropriate PIN. The ciphering algorithms A3 and A8 are implemented in the subscriber identity module (SIM) and the ciphering key K_c is also stored in the SIM.

Generally, the SIM carries the following information: IMSI number, authentication key (K_i), subscriber information, access control class, cipher key (K_c), TMSI, additional GSM services, location area identity, forbidden PLMN, A3 and A8 algorithms, and BCCH information. The SIM card provides the storage capability for administrative information, ID card identification, recently dialed numbers, SMSs, and so on.

The interfaces between the mobile handset and the SIM card are fully standardized and there are already specifications in place. SIM card readers or editors are hardware–software combinations, which makes it possible to get access to the SIM card of a mobile phone right on a PC. With a SIM card reader it is possible to view, create, edit, and backup phonebook entries by using a PC and eliminates typing in the information using the mobile phone's keypad. PIN codes, transfer data from one SIM to another, backup, and export and import are all phonebook entries that can be managed. SIM card readers enable t the SIM card phonebook data to be backed up to the local memory and avoid data losses when the user loses or changes a SIM card or a GSM phone.

10.12 Application Processing Unit

As discussed earlier, today's GSM mobile phones not only contain the voice or data modem part, but also contain a bunch of applications and associated hardware and software. Typical examples are audio player, video player, GPS, and connectivity modules such as USB, Bluetooth, IrDA, and so on. Figure 10.19

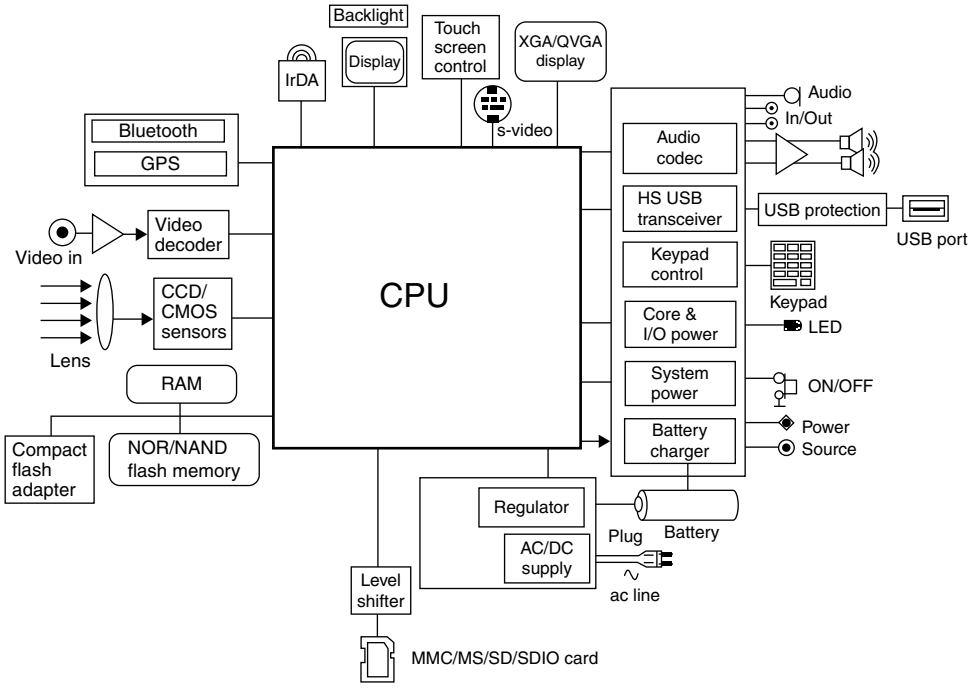


Figure 10.19 Application processing units

shows different hardware functional modules and their interface to a central application processor, which controls and executes the necessary software drivers for them. In Chapter 8, we have already discussed the software details for audio and video players.

10.13 Camera

Nowadays, in almost every mobile phone, a digital camera device is included. A digital camera is very similar to a conventional analog camera. It also contains most of the associated components that a conventional camera contains, such as a lens and a shutter. The lights fall on an array of image sensors or photosensitive cells. Most commonly, the image sensor is a charged-couple device (CCD) converting light into electric charges, and is essentially a silicon chip used to measure light. These charges are stored as analog data that are then converted into digital data via an analog-to-digital converter (ADC). These generated image data are too huge to store, so they are compressed using particular image coding techniques (as discussed in Chapter 7) and then stored on a memory card. The most common type of memory card is the compact flash card (CF card), and in addition to this, other popular formats are: memory stick (MS), multi-media-card (MMC), secure digital (SD), secure digital input–output (SDIO), and so on. The secure digital (SD) is a flash (non-volatile) memory card. Some digital cameras use CMOS (complementary metal oxide semiconductor) technology based microchips as image sensors; these are cheaper and easy to integrate. For the camera application, there are three major external components – a camera, an external RAM and an LCD.

10.14 LCD Display

A liquid crystal display (LCD) is usually used in mobile phones for display screen purposes. This is an electro-optical amplitude modulator realized as a thin, flat display device made up of any number of color or monochrome pixels arrayed in front of a light source or reflector. It uses a very small amount of electrical power for its operation. In Figure 10.20, the different parts of an LCD subsystem are shown.

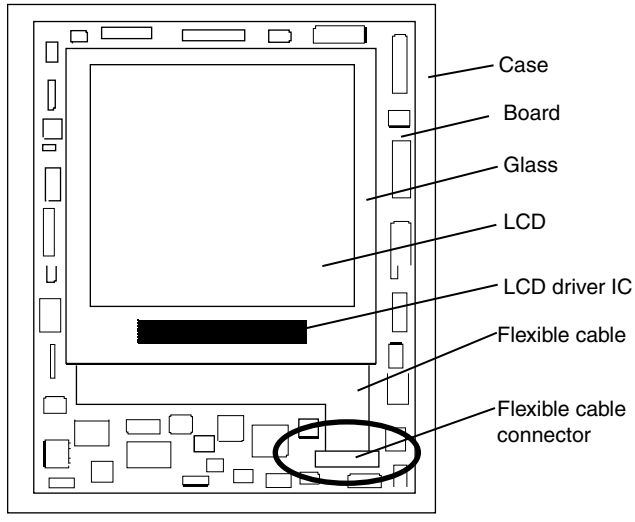


Figure 10.20 LCD module and associated components

In color LCDs, each individual pixel is divided into three cells, or sub-pixels, which are colored red, green, and blue, respectively, through the use of additional filters (pigment filters, dye filters, and metal oxide filters). Each sub-pixel can be controlled independently by the LCD driver software to yield thousands or millions of possible colors for each pixel. Active-matrix LCDs depend on thin film transistors (TFT), which are basically tiny switching transistors and capacitors and these are arranged in a matrix on a glass substrate. To address a particular pixel, the appropriate row is switched on, and then a charge is sent down the correct column. As all of the other rows that the column intersects are turned off, only the capacitor at the designated pixel receives a charge. The capacitor is able to hold the charge until the next refresh cycle. Passive-matrix LCDs use a simple grid to supply the charge to a particular pixel on the display. It starts with two glass layers termed substrates. One substrate is given columns and the other is given rows made from a transparent conductive material. This is usually indium–tin oxide.

10.15 Keypad

A keypad is a set of buttons arranged in a matrix form, which usually bear digits (0–9), the alphabet (a–z), alphanumeric characters (*, #), and also some special symbols for accept call, reject call, cursor movement, and so on. In Figure 10.21, the internal circuit diagram of a keypad is shown.

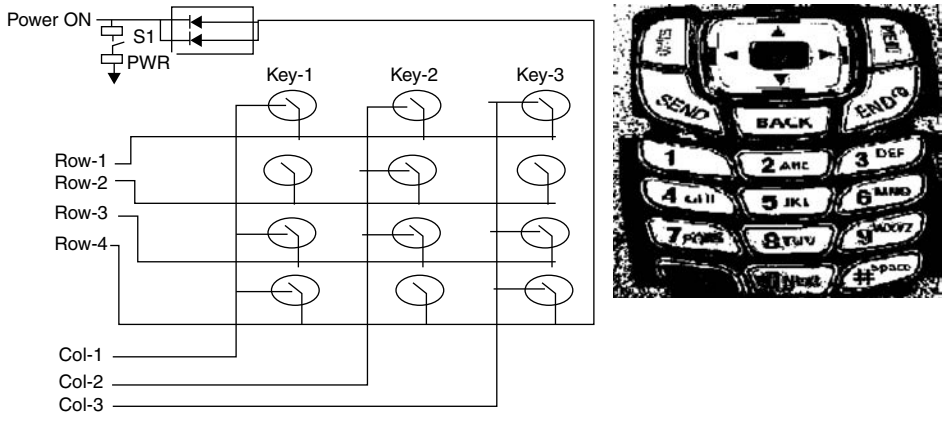


Figure 10.21 Keypad of a mobile phone

The keypad hardware may be polled by the keypad software driver to check if a key has been pressed by the user. In this case, the hardware must have a memory of the last key pressed so that the driver will detect that a keystroke has occurred on the next polling interval. The driver must reset the keystroke memory so that additional keystrokes can be detected on future polls of the hardware. To do this, no interrupt service routine (ISR) is required. Alternatively, the hardware may be implemented so that it generates an interrupt when any key is pressed. In this situation, an interrupt service routine (ISR) is required. Generally a keypad device will not use direct memory access (DMA) or shared buffers to transfer data, but will use programmed I/O instead.

Today many mobile phones are designed with touch screen based virtual keypads. A "touchpad" is a pointing device consisting of a specialized surface that can translate the motion and position of a user's

fingers to a relative position on the screen and that position (co-ordinate) can be used to take decisions on what key is pressed. It operates in many ways, such as capacitance and conductance sensing. George Gerpheide in April 1994 [5] developed the matrix approach, where in two layers, a series of conductors are arranged in an array of parallel lines, which are separated by an insulator and cross each other at right angles to form a grid. A high-frequency signal is applied sequentially between pairs in this two-dimensional grid array. The current that passes between the nodes is proportional to the capacitance. When a virtual ground, such as a finger, is placed over one of the intersections between the conductive layers, some of the electrical field is shunted to this ground point, resulting in a change in the apparent capacitance at that location. The "capacitive shunt method" senses the change in capacitance between a transmitter and receiver that are placed on opposite sides of the sensor. The transmitter creates an oscillating electric field. The capacitance value decreases when a finger (which is like a ground point) is placed between the transmitter and receiver, becuase some of the field lines will be shunted away through the fingers.

10.16 Connectivity Modules

10.16.1 Bluetooth

Bluetooth is a telecommunications industry specification for wireless personal area networks (PANs), which is a short-range (32 ft $\sim$ 10 m) radio-frequency technology that operates at 2.4 GHz (ISM band) and is capable of transmitting voice and also data. The name Bluetooth comes from the Danish King Harald "Bluetooth" Blaatand, who unified Denmark and Norway. At the beginning of the Bluetooth wireless technology era, Bluetooth was aimed at unifying the telecom and computing industries. Today, Bluetooth provides a way of connecting and exchanging information between a bunch of devices such as mobile phones, headsets, laptops, PCs, printers, digital cameras, and video game consoles over a secure, globally unlicensed short-range radio frequency (Figure 10.22). Bluetooth radios use a fast frequency-hopping spread spectrum (FHSS) technique, as discussed in Chapter 5. Up to eight data devices can be connected in an ad hoc piconet. Each piconet supports up to three simultaneous full duplex voice devices (CVSD). Technical parameters of Bluetooth are provided in Table 10.1 [6].

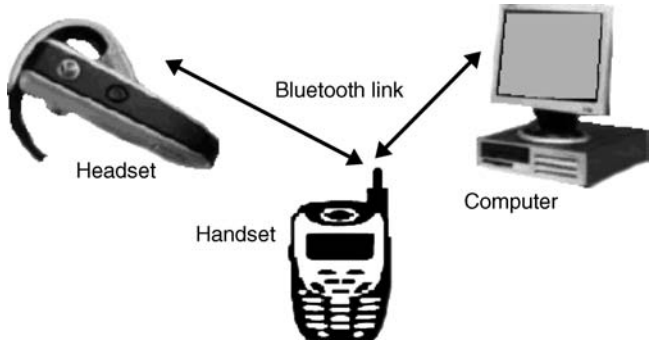


Figure 10.22 Bluetooth connected devices

In the first version of the standard, the gross data rate is 1 Mb/s. Bluetooth profiles are general behaviors through which Bluetooth enabled devices communicate with other BT devices. Bluetooth technology defines a wide range of profiles that describe many different types of use cases, such as: advanced audio

Table 10.1 Technical parameters for Bluetooth

Connection type	Spread spectrum (frequency hopping) and time division duplex (1600 hops/s)
Spectrum	2.4 GHz ISM open band (79 MHz of spectrum = 79 channels)
Modulation	Gaussian frequency shift keying
Transmission power	1–100 mW
Data rate	1 Mbps
Range	30 ft
Supported stations	8 devices
Data security – authentication key	128 bit key
Data security – encryption key	8–128 bits (configurable)
Module size	9×9 mm

distribution profile (A2DP), audio/video remote control profile (AVRCP), basic printing profile (BPP), common ISDN access profile (CIP), cordless telephony profile (CTP), Fax profile (FAX), file transfer profile (FTP), and so on.

The protocol stack for Bluetooth is shown in Figure 10.23. For more details please refer to Bluetooth specifications [5].

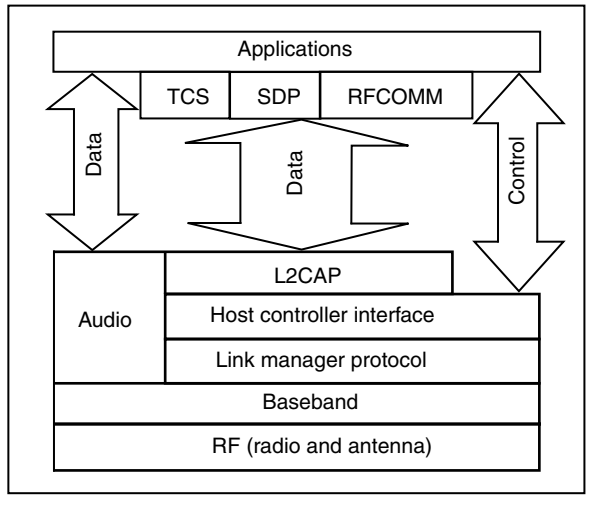


Figure 10.23 Bluetooth protocol stack

10.16.2 USB

A universal serial interface (USI) is a fast, bidirectional, isochronous/asynchronous, low-cost, dynamically attachable serial interface. A universal serial bus (USB) is specified to be an industry standard extension to the PC architecture with a focus on computer telephony integration (CTI), the

consumer, and productivity applications. An original intention of USB was to connect many devices to a PC host, as there was a shortage of serial and parallel ports in the PC. In 1994, an alliance of four Industrial Partners – Compaq, Intel, Microsoft, and NEC began to specify USB. The main goals for defining the USB are: plug-n-play, port expansion, low cost high performance, enabling seamless integration of new classes of devices, and open architecture.

The USB is a cable bus that supports data exchange between a host computer and a wide range of simultaneously accessible peripherals. The attached peripherals share the USB bandwidth through a host scheduled, token-based protocol. The bus allows peripherals to be attached, configured, used, and detached while the host and other peripherals are in operation. Several criteria were applied in defining the architecture for the USB, such as ease-of-use for peripheral expansion, full support for real time data for voice, audio and compressed video, protocol flexibility for mixed-mode isochronous data transfers and asynchronous messaging, support concurrent operation of many devices (multiple connections), up to 127 physical devices, and lower protocol overhead resulting in high bus utilization. The USB system consists of a single USB host and a number of USB devices and interconnects. The USB physical interconnect is a tiered star topology, and a hub is at the center of each star. This has some benefits, such as the power to each connected device can be monitored and even switched off independently. High, full and low speed devices can be supported. The main entities of an USB system are: (1) USB host, (2) USB function, and (3) interconnections.

10.16.2.1 USB Host

There is only one USB host in the USB bus chain and this host is the master of the USB system. The USB interface to the host computer system is referred to as the host controller. The host controller may be implemented in a combination of hardware, software and firmware. The root hub is integrated into the host system to provide one or more attachment points. The USB host controllers have their own specifications. With USB 1.1, there were two host controller interface specifications. (1) UHCI (universal host controller interface) developed by Intel, which puts more burden on the software (Microsoft) and allows for cheaper hardware. (2) OHCI (open host controller interface) developed by Compaq, Microsoft, and National Semiconductor, which places more burden on the hardware and makes the software simpler. In standard USB 2.0, another host controller interface is defined: (3) EHCI (enhanced host controller interface) – developed by Intel, Compaq, NEC, Lucent, and Microsoft, which is an enhanced version.

10.16.2.2 USB Device

USB devices are one of the following:

1. **Hub** – which provides additional attachment points to the USB. Hubs are wiring connectors and enable multiple break attachment characteristics of the USB. Attachment points are referred to as ports. The USB 2.0 hub consists of three portions: (a) hub controller, (b) hub repeater, and (c) the transaction translator. The USB specification does not limit the number of downstream connectors from a hub, but seven seems to be a practical limit, although four is most popular.
2. **Function** – which is a USB device that is able to transmit or receive data or control information over the bus. These follow USB protocol. A function is typically implemented as a separate peripheral device with a cable that plugs into a port on a hub. However, a physical package may implement multiple functions and an embedded hub with a single USB cable. This is known as a compound device. A compound device appears to the host as a hub with one or more non-removable USB devices.

10.16.2.3 Interconnects

To connect a host with hub or USB function devices there is a need for a cable, which is known as the USB cable. It carries power in addition to the data signal. The USB transfers signal and power through a four-wire cable, which is shown in Figure 10.24. The signaling occurs over two wires on each point-to-point segment. The power is transmitted over two wires: V_{bus} and Ground.

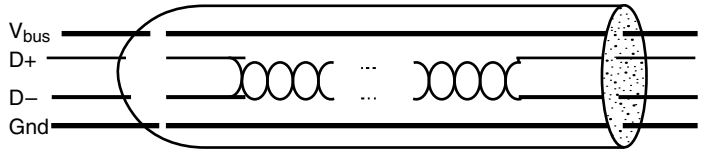


Figure 10.24 USB cable

The colors of each wire are given in the Table 10.2.

Table 10.2 USB interconnects

Pin number	Cable color	Function
1	Red	V_{bus} (+5 V)
2	White	D−
3	Green	D+
4	Black	Ground

USB allows for a variable length of cable up to 7 m. The cable length is decided after considering the delay and voltage drop.

10.16.2.4 Speed of USB

USB version 1.1 supported two speeds, a full speed mode of 12 Mb/s and a low speed mode of 1.5 Mb/s. The USB 2.0 standard has introduced speeds of up to 480 Mb/s, and this is known as the high speed mode.

1. High speed device – 480 Mb/s
2. Full speed device – 12 Mb/s
3. Low speed device – 1.5 Mb/s

10.16.2.5 USB On The Go (OTG)

As USB communication can only take place between a host PC and a peripheral, this imposes a limitation of having a PC host in the connection. In order to enable this limitation, a supplement to the USB 2.0 specification was developed that allows a portable device to take on the role of a limited USB host, without the burden of having a PC. This new standard basically defines:

1. A new type of device called a "dual role device" can (using the same connector) operate either as a normal USB peripheral or as a USB host. This single connector capability makes OTG (on the go) especially useful for handheld or other small devices that do not have the space for multiple connectors.

2. This also introduces two new receptacles and a new plug. Thus an OTG compliant device has only one USB connector (mini-AB receptacle). The ID pin makes it easy for a dual role device to determine if it should be default host or the default peripheral. On the mini-A plug, the ID pin is shorted to ground, whereas on a regular mini-B plug, the ID pin is left open (or in the case of a car-kit it is shorted to ground through 102 kΩ).
3. Full speed operation as peripheral (high speed optional) and full speed support as host (low-speed and high-speed optional).
4. Supports two new OTG protocols - (a) *Session Request Protocol:* This protocol allows the B-device to request that the A-device turn on V_{bus} and start a session. Once the session is started the host negotiation protocol can be invoked to give control to the B-device. The OTG supplement defines a session request protocol (SRP), which allows a B-device to request the A-device to turn on V_{bus} and start a session. This protocol allows the A-device, which may be battery powered, to conserve power by turning V_{bus} off when there is no bus activity while still providing a means for B-device to initiate bus activity. Any A-device, including a PC or laptop, is allowed to respond to SRP. Any B-device, including a standard USB peripheral, is allowed to initiate SRP. Any dual role device is allowed to respond as well as initiate to SRP. (b) *Host Negotiation Protocol:* This protocol allows the B-device to take control of the bus and become the host device with the A-device acting as peripheral. The host negotiation protocol (HNP) allows the host function to be transferred between two directly connected dual-role devices and eliminates the need for a user to switch the cable connections in order to allow a change in control communications between the devices. HNP will typically be initiated in response to input from the user or an application on the dual-role B-device. HNP may only be implemented through the mini-AB receptacle on a device.

10.17 Battery

The battery is the source of energy for the mobile phone circuitry and subsystems. Nowadays, there are so many power voracious applications run on the mobile handset, which consume too much of the power. Researches are moving ahead in reducing the battery size and increasing the battery life. The success of battery technologies greatly depends on the battery material, float life, temperature stability, and safety. Most of the mobile phone battery packs have a rating of 3.6 V, 650 mAh. Performance of a mobile battery is measured in terms of talk time (in dedicated mode) and standby time (in idle mode). It not only depends on the battery type used, but also on the sleep handling, clock, and some other system design parameters. Cell phone batteries have different weights, lifetimes, talk times, and thicknesses. All of these can have a significant impact on cell phone user experience. There are many types of battery available for a given phone and many factors should to be taken into account for the apposite battery selection.

10.17.1 Primary Cells

Some handsets can take primary (non-rechargeable) batteries, while others can also work on “ordinary” batteries, for example, the Motorola c520 works with 4 AA batteries.

10.17.2 Rechargeable Battery Types

Generally, there are three basic types of rechargeable battery used in mobile phones.

1. **NiCd (Nickel Cadmium):** This is the oldest types of battery. Generally, they are used in cordless phones and the old-generation mobile phones. These are very prone to the “memory effect.”

To maximize the performance, these batteries should be discharged and recharged completely every time.

Advantages – These are the cheapest variety of batteries, thus highly affordable and bring down the overall cost of the mobile handset.

Disadvantages – As mentioned, this type battery is very prone to the "memory effect." This is sometimes referred to as voltage depression. If it is not fully discharged before recharging it, after a few cycles, the battery "learns" this low water mark, and acts as if it is discharged to this point. This must be discharged and recharged fully on every recharge cycle. The chemicals in nickel cadmium are not environmentally friendly, and the disposal of cadmium-rich waste is an increasing problem.

2. **NiMH (Nickel Metal Hydride):** This is a better battery type than the NiCd and are much less prone to "memory effect."

   Advantages – They are cheaper than Li-Ion batteries, so are affordable and bring down the overall cost of the mobile phone. These batteries are less prone to the "memory effect" issue, and also have a higher capacity in relation to their size and weight.

   Disadvantages – The drawback of NiMH is that their longevity is less compared with NiCd cells. After a few hundred charge cycles, the crystals inside NiMH cells become coarser, and although they are able to provide the power for long standby times, when the extra current to sustain a call is needed, the voltage available drops rapidly, and suddenly shows low battery warnings. Once the call has ended, and after a few minutes rest, the battery is fine for many hours standby.

3. **Li-Ion (Lithium-Ion):** These are considered the most advanced and widespread cell phone batteries. A Li-ion gives exceptional capacity for its size and weight, and does not suffer from the "memory effect".

   Advantages – These are lighter and slimmer than the NiMH and NiCd batteries and are not subject to the "memory effect." Usually, they offer a longer standby time and talk-time.

   Disadvantages – These are expensive.

4. **Li-Polymer (Lithium-Polymer):** These are very similar to lithium-ion, except that they can be molded into more varied shapes, and so be squeezed into smaller phone casings. These are even thinner and lighter batteries.

Generally the baseband contains the components that control the power distribution to all the phone modules, except for those parts that require continuous battery supply. The battery feeds power directly to three parts of the system: charge controller, power amplifier, and user interfaces (buzzer, display, keyboard lights).

To find out the type of battery the mobile phone has, switch off the phone and remove the battery (normally on the back side of the phone). The battery type might be written on the label of the battery.

10.17.3 Battery Charger Circuit

Generally, the mobile-phone batteries are charged through a proprietary charging algorithm in the baseband controller. An example circuit is shown in Figure 10.25. Here, the phone's charger input is connected to the internal battery through a p-channel switch of low on-resistance, controlled by a pulse-width modulation (PWM) signal from the baseband controller. To minimize power dissipation and consequent thermal problems in the phone, the charging supply is current limited and specified according to the battery's chemistry and charge-recovery requirements. When the charger is connected it charges until the battery voltage level reaches 3.0 V. A control program changes the charging mode from start up to

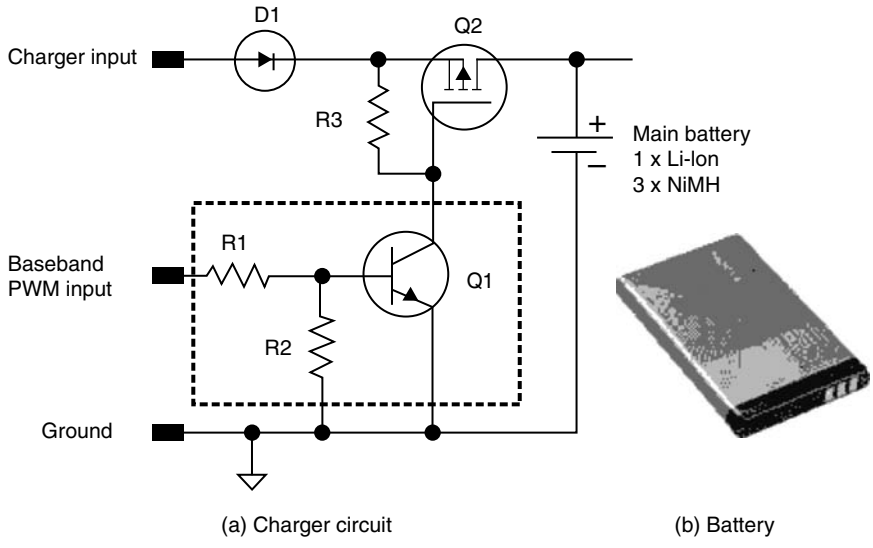


Figure 10.25 Mobile phone battery and charger circuit

pulse width modulation (PWM) changing mode. This is controlled by the baseband processor software program. If the battery voltages reaches >3.6 V before the program has taken control over the charging, the start up current is switched off. An output over voltage protection circuit is used to protect the phone from any over voltage damage.

An NTC (negative temperature coefficient) thermistor inside the battery pack is used for battery temperature measurement. Then this temperature is converted by an ADC and passed to a baseband processor for any action.

10.17.4 Sleep Mode

Sleep mode is introduced to save the battery power in mobile devices. In Chapter 9, we have seen that in the radio protocol level the sleep mode is also introduced so that the device can be put into power-down mode whenever there is nothing to do (no reception or transmission is scheduled). There are various types of sleep modes that can be introduced into the mobile device. Normally, the baseband architecture supports a power saving function called "sleep mode." This sleep mode shuts off the VC-TCXO, which is used as the system clock source for both RF and baseband. During the sleep mode the system runs from a 32 kHz crystal. Then periodically the phone wakes up through a timer running from this 32 kHz clock supply. As discussed, the sleeping time is determined by some network parameters. During the sleep mode the baseband processors (MCU, DSP) are in standby mode, the normal VC-TCXO clock is switched off, and the voltage regulators for the RF section are also switched off.

The sleep mode is exited either by the expiration of a sleep clock counter or by some external interrupt, generated by a charger connection, key press, headset connection.

10.18 Clocking Scheme

The clock signal is like the heart of the digital device, as this signal needs to be circulated among various modules inside the device. Different modules require various operating clock frequencies, so there are converters to do this job. Typically the system clock in a phone is 13 MHz. It is generated in the RF VC-TCXO circuit. The clock frequency is controlled by AFC. From the 13 MHz sine wave signal a square wave signal is generated. The derived clock goes to the sleep module (32 kHz), synthesizer module (3.25 MHz), RF unit, SIM module (3.25 MHz), and LCD driver. The real time clock logic consists of RTC logic, and the 32 kHz crystal. In a normal situation the real time clock takes the power from the switcher output. In case the main battery is removed, the RTC is powered by the output capacitors on the switcher until they are drained and the RTC loses its timing. The time must be set again upon power on.

10.19 Alert Signal Generation

A buzzer is used to provide various alerting audio signals as an indication of an incoming call. The user key press and other response beeps are also generated by the buzzer. The buzzer is controlled by a buzzer PWM output signal from the baseband processor. The mobile phone uses a dynamic type of buzzer. The low impedance buzzer is connected to an output transistor that obtains drive current from the PWM output. The volume can be adjusted either by charging the pulse width, causing the level to change or by changing the frequency to utilize the resonance frequency range of the buzzer. A vibra alert device, used for giving silent alert signals to the user about any incoming calls, is controlled by the vibra PWM output signal from the baseband processor. Generally, a specially designed motor is used for the vibra alert device. The vibra alert can be adjusted either by changing the pulse width or by changing the pulse frequency of the vibra PWM signal. The vibra device is placed inside a special vibra battery.

Backlight is used to provide the background light for the LCD display to see the screen or keypad buttons. Generally, LEDs are used for LCD back lighting. They are controlled by the signal voltage coming from the controller.

10.20 Memory

The memory is an essential component in a mobile phone. We can classify the memories into two categories – one is read-only memory and other is read–write memory. Read-only memories allow only reading from any location, but writing to any location is prohibited, an example of this type of memory is ROM. However, in the case of read–write memory, we can read or write to any specific location, an example is RAM. Some memories contain dual properties and these are known as hybrid memories, examples include Flash, EEPROM, and so on. Again, depending on the storing property, we can broadly divide the memories into two categories – volatile memory and non-volatile memory. With volatile memory, the memory content vanishes when the power supply to this memory device is switched off. So, if the power supply is off, the content of the memory will be lost forever. RAM is an example of this type of memory. With non-volatile memory, as the name suggests, the memory content is non-volatile. This means that the memory contents are not lost when the supply power to the memory is put off. ROM is an example of this type of memory.

10.20.1 Read Only Memory (ROM)

"ROM" is an acronym for "read only memory," which means that the data can only be read from it and can not be written into it. ROM is a non-volatile memory, so once the data are stored or written into it,

it remains there forever and does not lose the memory content once the power supply is put off. It is low cost, high speed, non-volatile memory, and is made up of arrays such as RAM. A ROM can be used to realize arbitrary truth tables, generate characteristics, convert codes or store system boot programs. The ROM is constructed by unipolar or bipolar devices. Some ROM devices are only one time programmable, that is, there is no way to alter the content or write the data into it – these are called one time (OT) programmable ROMs. Other types of ROM, which can be written using special techniques, such as UV light or electrical signals, are called field alterable ROMs. These are not on-system programmable, that is, in order to program these we have to use a special programming platform and device. Typical parameters of a ROM (uPD23C1000A) are: address access time, 200 ns; chip enable access time, 200 ns; operating voltage: 5 V.

10.20.1.1 Electrically Erasable Programmable ROM (EEPROM)

EEPROM (electrically erasable programmable ROM) is a user-modifiable read-only memory that can be erased and reprogrammed repeatedly by applying a higher electrical voltage. These are electrically programmable up to many thousand times. They can be packaged in simple plastic packages, which reduces the cost of the device compared with the EPROM. A common use is for holding BIOS programs. An EEPROM chip has to be erased and reprogrammed entirely. It also has a limited life – that is, the number of times it can be reprogrammed is limited. Differences between various types of ROM devices are given in Table 10.3.

Table 10.3 Comparison between different types of ROM

Memory device	Number of times programmable	Field erasable or on system erasable	Cost	Density
ROM	One time	NA	Low	High
EPROM	Many times	Field erasable	High	High
EEPROM	Many times	Field erasable and system erasable	Low	No

10.20.2 Flash Memory

Flash is a special form of an EEPROM type of device, which uses normal supply voltages (3.3–12 V) for erasing and reprogramming. It is always desirable to have non-volatile memory, as the power consumption becomes less and there are no worries about the loss of information. The basic problem with EEPROM is that it needs a special arrangement and high voltage to program, that is, to write into it, whereas Flash is system programmable and does not require any special platform to program it. It is a high density, truly non-volatile, high performance read–write solution. The Flash device also requires an external high voltage supply for programming as does a conventional EPROM. The electrical erase is by either hot electron or cold electron injection from a floating gate, with the oxide layer allowing the cell to be electrically erased through the source.

The Flash memory is available in a chip with several blocks. The blocks or sectors may be of the same or different size. When the boot code is stored in the Flash, there are some small sectors on the top or bottom side of the Flash memory to store the boot code, which usually has some level of protection from accidental overwrite. The boot block Flash memory family has asymmetrically blocked memory array

layouts to enable small parameter or boot code storage, along with efficient larger blocks for code and data file storage. The symmetrically blocked memory arrays, which enable the best code and data file management, have all the sectors of the same size.

10.20.2.1 Flash Erasing, and Programming

The Flash chip contains several blocks. A fixed address range (for example, 0000-0FFFF) defines the blocks. When a Flash memory is imported from the factory, all of its locations contain the same type of bit (either 1 or 0 depending on the Flash type, here for example we consider that all are in the 1 state). This is called the erased state of the entire Flash memory. During the write or program process, the Flash cells change from one binary voltage level to another (that is, 1 to 0). This means, if the Flash cells originally contain all 1, then during writing, we can change the cell contents from 1 to 0, wherever required, according to the input data pattern. However, the reverse is not possible. Thus, if the Flash cells are not erased before writing, then the cells will contain 0s and 1s. So, writing to the cells, where it contains 0s will not be possible. Hence, it is always required to erase the Flash before writing. In the erase process, the Flash cells are set back to their original binary voltage level or erase state. The erase process occurs on a block basis or entire chip erase. When a block is erased, all address locations within a block are erased in parallel, independent of other blocks in the Flash memory device. Flash components take a significant amount of time to erase a block or program, compared with RAM. To modify any data content in a block, it is required to copy the original content of that block in some other memory location, before erasing that block, then modifying the original data and again writing back to the Flash. There will be no problem in writing as it is already erased before freshly writing the modified data.

A Flash memory device is very useful to reduce system cost as well as improving data reliability, providing easy update capabilities, increasing battery life, and providing stability after power loss.

Data for a Typical Flash Memory – AMDAm29DL400B: write cycle time, 70–120 ns; sector erase operation time, 0.7 s (64 kbytes); chip erase time, 10 s; read cycle time, 70–120 ns. *Power consumption* – single power supply, 2.7–3.6 V; active read current, 12 mA; active write current, 25 mA; standby current, 5 μA. Recently Intel has developed a 1.8 V wireless Flash memory (28F640W18), which is the highest performance solution for Internet phones.

10.20.2.2 Different Types of Flash Memories

Two main technologies dominate the non-volatile Flash memory market today: NOR and NAND. NOR flash was first introduced by Intel in 1988. NAND flash architecture was introduced by Toshiba in 1989. Most Flash devices are used to store and run code (usually small), for which NOR Flash is the default choice. Some differences between NOR and NAND Flash memory are detailed in Table 10.4.

10.20.3 Random Access Memory (RAM)

The name "random access" indicates that each cell in the memory chip can be read or written in any order. All RAM (random access memory) memories are read–write memory. Some commonly used RAM memories are discussed below.

10.20.3.1 Static RAM (SRAM)

"SRAM" is an acronym for "static random access memory;" which means that once the data are written into the memory cell, they remain as they are, as long the power is not switched off. SRAM is available in

Table 10.4 Differences between NOR and NAND memory

Attributes	NOR Flash	NAND Flash
Capacity	132 MB	16512 MB
Performance	Very slow erase (~5 s)	Fast erase (~3 ms)
	Slow write	Fast write
	Fast read	Fast read
Reliability	Standard	Low
Erase Cycles	10 000–100 000	100 000–1 000 000
Life span	Less than 10% the life span of NAND.	Over 10 times more than NOR
Interface	Full memory interface	I/O only, CLE, ALE and OLE signals must be toggled
Access method	Random	Sequential
Ease-of-use (hardware)	Easy	Complicated
Full system integration	Easy	Hard; a simplistic SSFDC driver may be ported
Ideal usage	Code storage – limited capacity due to price in high capacity. May save limited data also	Data storage only – due to complicated Flash management. Code will usually not be stored in raw NAND Flash
	Some examples: simple home appliances, low-end set top boxes, low-end mobile handsets, PC BIOS chips	Some examples: PC cards, compact Flash, MP3 players, digital cameras.
Price	High	Low

many varieties starting from the super-fast bipolar and GaAs SRAM to the slow commodity CMOS variety. The early SRAM cells use NMOS technology and consist of six transistors; of these four are enhance mode transistors and two depletion mode resistors. The CMOS SRAM cells are very low power, wide noise margin but low speed.

Typical SRAM Data – 32 k × 8 bit low power CMOS static RAM; K6T0808C1D family; manufactured by SAMSUNG electronics. Access time: read cycle time, 70 ns; write cycle time, 70 ns. Power-supply voltage, 4.5 V; operating supply current, 5 mA standby current, 30 μA; density, generally 4–6 transistors per memory cell.

10.20.3.2 DRAM

"DRAM" is an acronym for "dynamic random access memory," which means that to remember the stored data, the memory chips require every bit to be refreshed within a certain period of time. When the power is removed from the DRAM the data are lost. The DRAM uses tiny built-in capacitors to store the data bits. The charge is stored in capacitor C. When the transistor is turned on, the information is refreshed (or charged).

Typical Data of DRAM – Mitsubishi Electric – M5M467400Dxx series; write cycle time, 90–110 ns; access time from CAS, 13 ns; access time from RAS, 50 ns; read cycle time, 90–110 ns; refresh cycle time, 64 ms; power dissipation. 300 mW; Vcc. + 3.3 V.

Generally, inside a mobile phone, Flash memory is used to store the processor program and application program or data. A series EEPROM is used for storing the system and tuning parameters, user settings and selections. The program is normally executed from SRAM after downloading it from Flash memory. This is also used for scratch pad memory.

10.21 GSM Receiver Performance

It is required that in a GSM system, every piece of equipment should comply with a minimum set of performance standards, regardless of its manufacturer or country of origin. In Chapter 4, we have already discussed about the different performance measurement parametersthat are normally used to characterize a transmitter and receiver. Digital receiver performance characteristics are often described by noting the receiver's ability to recover the modulation intelligence from an RF carrier, injected at low levels, into the antenna port in the presence of different types of interference. Co-channel rejection and adjacent-channel rejection specifications are indications of a receiver's robustness against interference from signals in the same or the adjacent channels, respectively, as the receiver tries to recover the intelligence from an on-channel GSM signal. The ability of the receiver to receive a desired GSM signal in the presence of a strong interfering signal on any frequency is described by its blocking specification. The performances of the MS and BTS receivers are specified by defining a maximum allowable bit error rate (BER) or frame eraser rate (FER) for each of the different GSM logical channels for a given set of radio channel conditions. Table 10.5 shows that the performance of the full-rate speech traffic channel (TCH/FS) is defined in terms of the FER and the residual BER (RBER) of the Class Ib and Class II bits.

Here the parameter α is defined as $1 \leq \alpha \leq 1.6$ and allows a tradeoff between the number of erased speech frames (these are decoded as wrong, so not passed to the voice decoder) and the quality of the non-erased frames.

10.21.1 Sensitivity and Noise Figure Requirements

The sensitivity requirements for different standards (P-GSM, E-GSM, DCS, PCS) are shown in Table 10.6 with the corresponding input SNR, required carrier-to-noise ratio (CNR) to maintain the minimum BER outlined in the standards, along with the required noise figure.

10.21.2 Reference Interference Level

The interference level is specified with respect to the desired signal level with the number of reference levels. In each case, the received signal level of the wanted signal is set to 20 dB above the reference sensitivity level. The reference carrier-to-interference (C/I) ratios for both co-channel and adjacent channel interference are given in Table 10.7.

Given the input conditions as shown in Figure 10.26, the receiver performance is defined in terms of maximum FERs and BERs for each logical channel and propagation condition similar to that shown for the sensitivity performance mentioned in Table 10.7. The interference performance of both GSM900 and DCS1800 is fully defined in the specifications. For more detail please refer to [7].

10.21.3 3GPP TS Requirements to TX Frequency

The main requirement in the specifications is that the frequency error must be less than 0.1 ppm, for example, for the 900 MHz band the frequency error should be below $0.1 \cdot 10^{-6} \times 900 \cdot 10^{6}$ Hz $\cong$ 90 Hz. GSM standard requirements for frequency error are given in Table 10.8.

Table 10.5 Reference sensitivity performance for various channels of GSM-850 and GSM-900 systems in different propagation conditions for wireless channel

GSM 850 and GSM 900						
Type of channel		Propagation conditions				
		Static	TU50 (no FH)	TU50 (ideal FH)	RA250 (no FH)	HT100 (no FH)
FACCH/H	(FER)	0.1%	6.9%	6.9%	5.7%	10.0%
FACCH/F	(FER)	0.1%	8.0%	3,.8%	3.4%	6.3%
SDCCH	(FER)	0.1%	13%	8%	8%	12%
RACH	(FER)	0.5%	13%	13%	12%	13%
SCH	(FER)	1%	16%	16%	15%	16%
TCH/F14,4	(BER)	10^{-5}	2.5%	2%	2%	5%
TCH/F9,6 and H4,8	(BER)	10^{-5}	0.5%	0.4%	0.1%	0.7%
TCH/F4,8	(BER)	—	10^{-4}	10^{-4}	10^{-4}	10^{-4}
TCH/F2,4	(BER)	—	$2\ 10^{-4}$	10^{-5}	10^{-5}	10^{-5}
TCH/H2,4	(BER)	—	10^{-4}	10^{-4}	10^{-4}	10^{-4}
TCH/FS	(FER)	0.1α%	6α%	3α%	2α%	7α%
	class Ib (RBER)	0.4/α%	0,4/α%	0,3/α%	0,2/α%	0,5/α%
	class II (RBER)	2%	8%	8%	7%	9%
TCH/EFS	(FER)	<0.1%	8%	3%	3%	7%
	(RBER Ib)	<0.1%	0.21%	0.11%	0.10%	0.20%
	(RBER II)	2.0%	7%	8%	7%	9%
TCH/HS	(FER)	0.025%	4.1%	4.1%	4.1%	4.5%
	Class Ib (RBER, BFI = 0)	0001%	0.36%	0.36%	0.28%	0.56%
	Class II (RBER, BFI = 0)	0.72%	6.9%	6.9%	6.8%	7.6%
	(UFR)	0.048%	5.6%	5.6%	5.0%	7.5%
	class Ib (RBER, (BFI or UFI) = 0)	0.001%	0.24%	0.24%	0.21%	0.32%
	(EVSIDR)	0.06%	6.8%	6.8%	6.0%	9.2%
	(RBER, SID = 2 and (BFI or UFI) = 0)	0.001%	0.01%	0.01%	0.01%	0.02%
	(ESIDR)	0.01%	3.0%	3.0%	3.2%	3.4%
	(RBER, SID = 1 or SID = 2)	0.003%	0.3%	0.3%	0.21%	0.42%

Table 10.6 Sensitivity, input SNR, and noise figure requirements for different standards

Wireless standard	Sensitivity (dBm)	Input noise (dBm)	Input SNR (dB)	Required C/N (dB)	Required NF (dB)
GSM	−102	−102.8	18.8	9	9.8
E-GSM	−102	−102.8	18.8	9	9.8
DCS1800	−100	−120.8	20.8	9	11.8
PCS1900	−102	−120.8	18.8	9	9.8

Table 10.7 *C/I* level requirements

For co-channel interference	C/I_c	9 B
For first adjacent channel (200 kHz) interference	C/I_{adj1}	−9 dB
For second adjacent channel (400 kHz) interference	C/I_{adj2}	−41 dB
For third adjacent channel (600 kHz) interference	C/I_{adj3}	−49 dB

Table 10.8 GSM standard requirement for frequency error

GSM 850 and GSM 900		DCS 1 800		PCS 1 900	
Propagation condition	Permitted frequency error (Hz)	Propagation condition	Permitted frequency error (Hz)	Propagation condition	Permitted frequency error (Hz)
RA250	±300	RA130	±400	RA130	±420
HT100	±180	HT100	±350	HT100	±370
TU50	±160	TU50	±260	TU50	±280

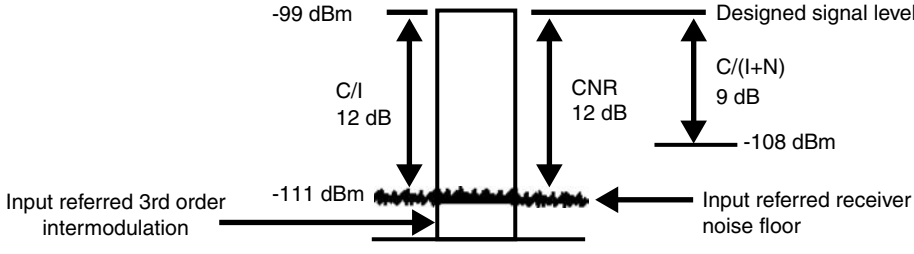


Figure 10.26 Maximum allowable input referred noise and distortion floors for GSM, PCS 1900

References

[1] Sloss, A., *ARM System Developer's Guide*, ARM, Sunnyvale, CA, ISBN 13: 978-1558608740.
[2] Lapsley, P., Bier, J., Shoham, A., and Lee, E.A., *DSP Processor Fundamentals: Architectures and Features,* IEEE Press Series on Signal Processing. IEEE Press, Los Alamitos.
[3] Das, S. Cavallaro, J.R., and Aazhang, B. (1997) Computationally efficient multiuser detectors. 8th IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications(PIMRC), Helsinki, Finland, pp. 62–67.
[4] Boser, B.E. and Wooley, B.A. (1988) The design of sigma-delta modulation analog-to-digital converters. *IEEE Journal of Solid-State Circuits*, **23**, 1298–1308.
[5] Gerpheide, G. (1994) US Patent No. 6680731.
[6] Bluetooth specification (version 2.1) (2008) Bluetooth SIG. Retrieved 2008-02-04. http://www.bluetooth.com/Bluetooth/Technology/Building/Specifications/.
[7] GSM Technical Specification (1998) 05.05 (ETS 300 577). *European Digital Cellular Telecommunications System (Phase 2); Radio Transmission and Reception.* ETSI TC-SMG, Sophia-Antipolis Cedex

Further Reading

Albrecht, S. (2005) Sigma-Delta Based Techniques for Future Multi-Standard Wireless Radios, Doctoral Thesis. KTH Library. ISBN 91-7178-170-6.

Eskelinen, P. (2001) Mobile antenna systems handbook. *Aerospace and Electronic Systems Magazine, IEEE*, **16** (10), 20–21.

Proakis, J.G. (1995) *Digital Communications*, 3rd edn, McGraw-Hill, New York.

Steele, R., Whitehead, J., and Wong, W.C. (1995) System aspects of cellular radio. *IEEE Communications Magazine*, **33** (1), 80–86.

11

Introduction to GPRS and EDGE (2.5G) Supported Mobile Phones

11.1 Introduction

The wireless data services offered by GSM are based on the circuit switched radio transmission. In this instance, at the air interface, a complete traffic channel is allocated for a single user for the entire call duration. Hence with bursty traffic (for example, Internet traffic) this results in highly inefficient resource utilization. It is obvious that for bursty traffic, packet switched bearer services will result in much better utilization of the traffic channels, because a channel will only be allocated whenever it is needed and will be released immediately after the transmission of the packets. Using this principle, multiple users can share one physical channel, for example, one physical channel can be multiplexed among several simultaneous users as required.

In order to address the inefficiencies of circuit switched radio transmission, two cellular packet data technologies have been developed: cellular digital packet data (CDPD) (for AMPS, IS-95, and IS-136) and the general packet radio service (GPRS). Here we will focus only on GPRS. This was originally developed in 1990 for GSM to support data, but later was also integrated within IS-136. We treat GPRS as an evolution from GSM to support packet based services.

Basically GPRS is based on the packet radio principle. Packets can be directly routed from the GPRS mobile (MS) to packet switch networks. Networks based on the Internet protocol (IP) and X.25 networks are supported in the current version of GPRS. Billing can be based on the amount of transmitted data volume, instead of the conventional billing method which is based on the connection time duration.

In short, GPRS improves the utilization of the scarce radio resources, offers volume-based billing, higher data transfer rates, shorter access times, QoS based service, point-to-point in addition to point-to-multi-point services, and simplifies the access to packet data networks.

11.2 System Architecture

In order to integrate GPRS into the existing GSM system architecture, new classes of network nodes, called GPRS support nodes (GSN), have been introduced. GSNs are responsible for the delivery and routing of data packets between the MSs and the external packet data networks (PDN). Figure 11.1, illustrates the system architecture of GPRS.

A serving GPRS support node (SGSN) is responsible for the delivery of data packets from/to the MSs within its service area. The tasks of the SGSN are packet routing and transfer, mobility management

Mobile Handset Design Sajal Kumar Das

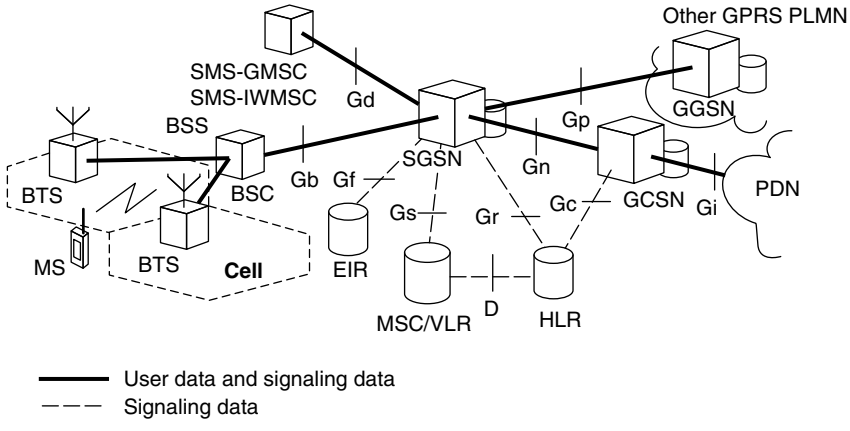


Figure 11.1 GPRS system architecture

(attach/detach and location management), logical link management, and authentication and billing functions, and so on. It manages packet communication sessions between individual mobile stations and the GGSN. The SGSN also interacts with the GSM databases to achieve mobility management functions and obtain subscription information to offer the required services to the mobile station. It is also involved in counting of data packets for billing purpose.

A gateway GPRS support node (GGSN) acts as an interface between the GPRS backbone network and the external packet data networks. It converts the GPRS packets coming from the SGSN into the appropriate packet data protocol (PDP) format (for example, IP or X.25) and sends these out on the corresponding packet data network. The GGSN also performs authentication and billing functions. The GGSN is similar to a GMSC (in GSM) as it provides a gateway between the GPRS network and the public PDN (packet data network) or other GPRS networks. It provides authentication and location management functions in addition to firewall functions on the Gi interface to the PDN. All GSNs are connected via an IP-based GPRS backbone network. Within this backbone, the GSNs encapsulate the PDN packets and transmit (tunnel) them using the GPRS tunneling protocol GTP. There are two types of GPRS backbones:

1. Intra-PLMN backbone networks connect GSNs of the same PLMN and are therefore private IP-based networks of the GPRS network provider.
2. Inter-PLMN backbone networks connect GSNs of different PLMNs. A roaming agreement between two GPRS network providers is necessary to install such a backbone.

In addition, the MSC/VLR may be extended with functions and register entries that allow efficient coordination between packet switched (GPRS) and circuit switched (conventional GSM) services. Examples of this are combined GPRS and non-GPRS location updates and combined attachment procedures. Moreover, paging requests of circuit switched GSM calls can be performed via the SGSN. For this purpose, the Gs interface connects the databases of SGSN and MSC/VLR. To exchange messages in the short message service (SMS) via GPRS, the Gd interface is defined. It interconnects the SMS gateway MSC (SMS-GMSC) with the SGSN.

A packet control unit (PCU) is also added into the BSC to control packet channels and to separate data flow for circuit switched and packet switched calls. The BSC of GSM is also given a new functionality for mobility management and handling GPRS paging. The PCU takes care of the radio resource functionality, such as allocation of the air interface channel blocks. The functional view of a GPRS system is shown in the Figure 11.2.

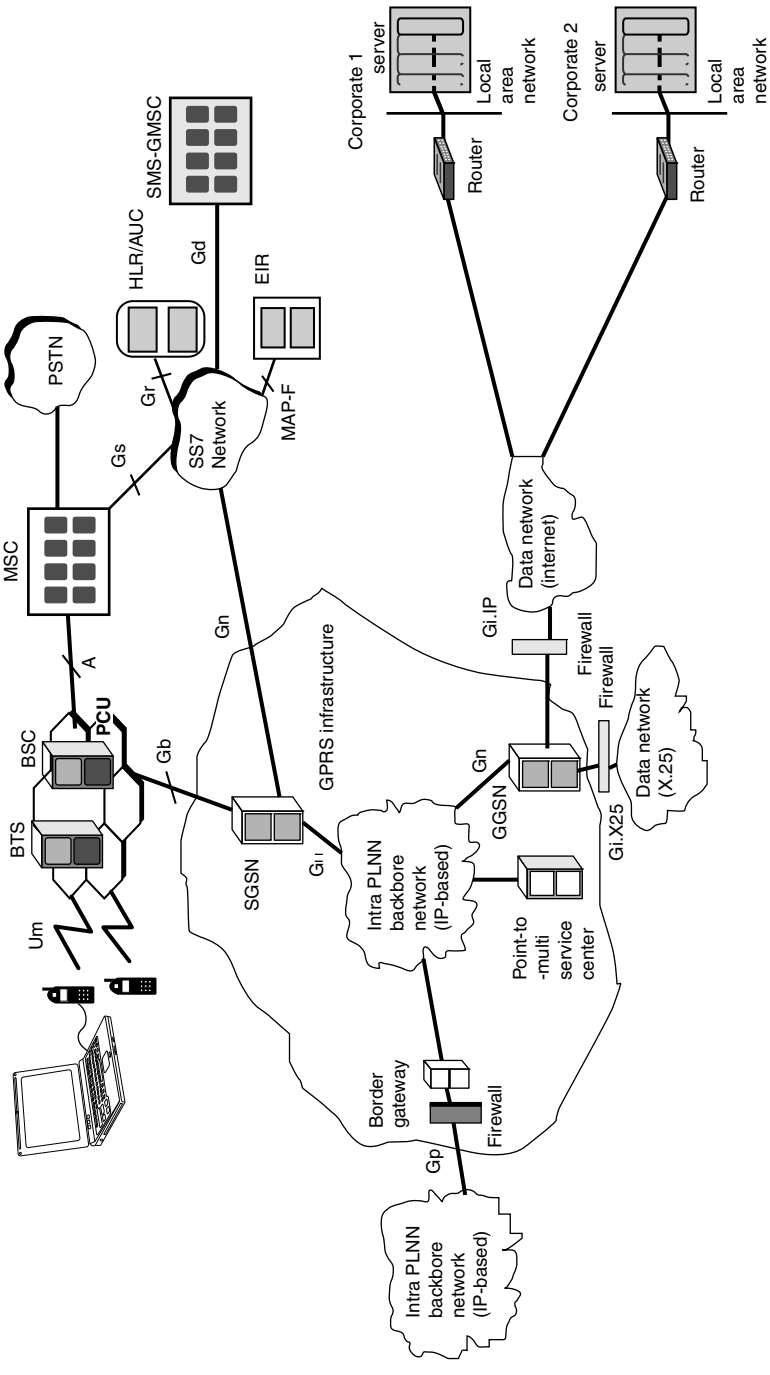


Figure 11.2 Functional view of GPRS

11.3 Services

Generally, there are two types of services defined for GPRS: the point-to-point (PTP) and the point-to-multi-point (PTM) service. A PTM service is available in later releases. The PTP service offers transfer of user packet data in two modes: connection-less mode (for example, IP) and connection-oriented mode (for example, X.25). On the other hand, the PTM service offers transfer of data packets from one user to multiple users.

GPRS allows defining QoS profiles using the parameters service precedence, reliability, delay, and throughput. The billing of the service is then based on the transmitted (and/or received) data volume, the type of service, and the chosen QoS profile. In a GSM/GPRS network, simultaneous use of conventional circuit switched and packet switched services can be supported based on the classes of MSs as defined below:

Class A: Supports simultaneous operation of GPRS and conventional GSM services.
Class B: Able to register simultaneously with the network for both GPRS and conventional GSM services but can only use one of the two services at a given time.
Class C: MSs can only attach for either GPRS or conventional GSM services. Simultaneous registration and usage is not possible. An exception is SMS messages, which can be sent and received at any time.

11.4 Session Management, Mobility Management, and Routing

In this section, we will see how the data call is set up. The GPRS attach and PDP context activation are required for data call setup. So, before using GPRS services an MS has to register with an SGSN of the GPRS network. When the mobile is first powered ON (in the GPRS default mode), the GPRS attach procedure takes place and this can also occur afterwards depending on the network settings. This is always initiated by MS. Only SGSN is involved in the GPRS attach process and this is transparent to the BSS. In a GPRS attach process the mobile informs the SGSN about its identity IMSI or P-TMSI (packet TMSI, which is allocated by the SGSN on GPRS attach), old routing area identification, mobile class-mark, CKSN, and so on, along with the attach type, which indicates to the SGSN whether this mobile wants to do a GPRS, GSM or both attaches. For MSs using both circuit switched and packet switched services, it is possible to perform a combined GPRS/IMSI attach procedure. The SGSN will attach the mobile and inform the HLR if there is a change in the routing area. If the attach type is both, then the SGSN will also do a location update with the VLR. The disconnection from the GPRS network is known as GPRS detach, which can be initiated either by the MS or the network (SGSN or HLR).

Before the MS can transmit or receive any information, it needs to activate a packet data protocol (PDP) context. A mobile can establish multiple PDP context sessions for different applications. A PDP context activates a packet communication session with the SGSN (Figure 11.3). It provides information for mapping and routing information between the MS and the GGSN. The SGSN asks the GGSN for a PDP context. Then the GGSN will create a new entry in its PDP context table and send confirmation to the SGSN, including the address, if it is dynamic. The mobile in a PDP context activation procedure either specifies a static IP address or requests an IP address. It also specifies the access point with which it wants to communicate, such as some intranet network or an Internet service provider (ISP). It also aks for the desired QoS and a network service access point identifier (NSAPI), which is used to discriminate data packets of different applications. When the SGSN obtains the mobile information, it decides on the GGSN connected to the APN and forwards the request to the GGSN. Then, if a static address exists, the GGSN connects the MS to the desired access point; unless it gets an IP address from the APN. The GGSN also provides some transaction identifiers for communication of data to reference this mobile between the GGSN and SGSN. The SGSN in its request also provides a negotiated QoS based on the subscription information of the user and availability of services. Once the communication and activation at GGSN is successful, the appropriate information is forwarded to the mobile.

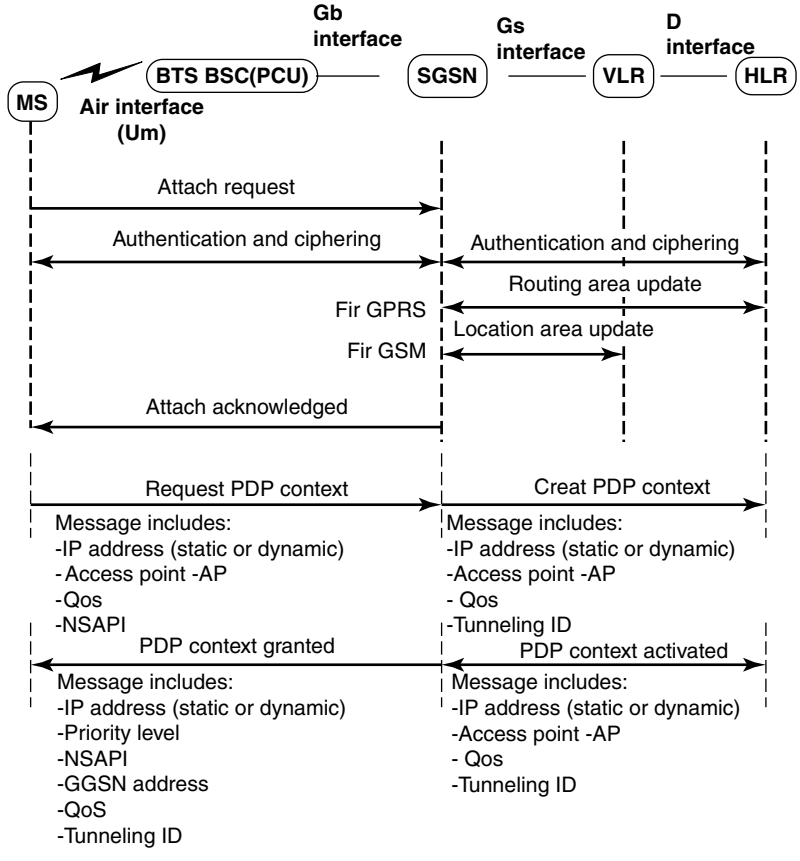


Figure 11.3 GPRS attach and PDP context activation process

As GPRS uses packet-based architecture, where a connection is granted during traffic and released after completion, location update is thus essential here. However, in order to avoid too much consumption of the resources and power for frequent updates during idle time and extra delay for regular paging for each downlink packet, a state model has been defined for location management of GPRS, as shown in Figure 11.4. A mobile can be in any of three states (idle, standby, and ready) based on its need and the location update frequency depends on the MS state.

1. **Idle State**: The mobile is powered on in the idle state, but not attached to the GPRS. In this state the mobile is not reachable, and no location updating is performed, that is, the current location of the MS is unknown to the network.
2. **Ready State**: Upon performing a GPRS attach, the mobile enters into the ready state. Here the mobile is currently engaged in packet transfer or has recently terminated a packet transfer. With a GPRS detach the mobile may disconnect from the network and return to the idle state again and all PDP contexts will be deleted. An MS in the ready state informs its SGSN with respect to its every movement to a new cell.
3. **Standby State**: The mobile is powered on and attached to the GPRS but no packet transfer is in progress. The standby state is reached when a mobile does not send any packets for a long period of

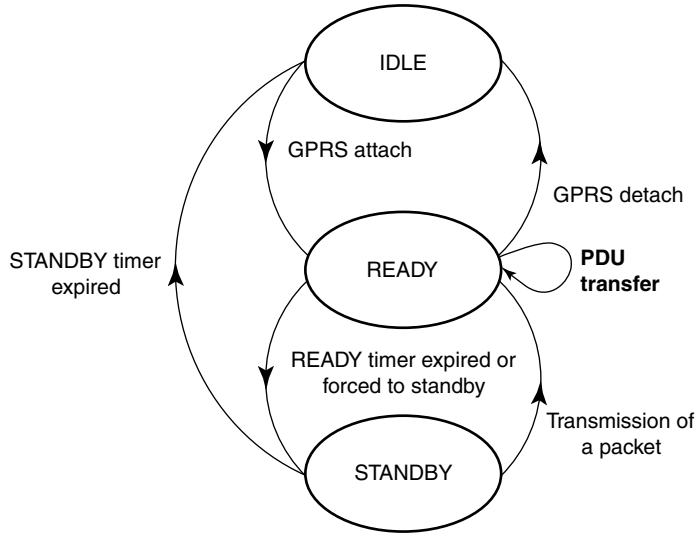


Figure 11.4 State model of a GPRS mobile station

time. This causes the ready timer to expire. In this state routing area updates are sent as needed. In the standby state, a GSM location area (LA) is divided into several routing areas (RA). In general, an RA consists of several cells. The SGSN will only be informed when an MS moves to a new RA. To find out the current cell of an MS in the standby state, paging of the MS within a certain RA must be performed.

In general, GPRS mobility management consists of two levels:

1. Micro-mobility management performed by SGSN tracks the current routing area or cell of the mobile station.
2. Macro-mobility management keeps track of the mobile station's current SGSN and stores it in the HLR, VLR, and GGSN.

The routing area consists of several cells, and one location area includes several routing areas, as shown in the Figure 11.5.

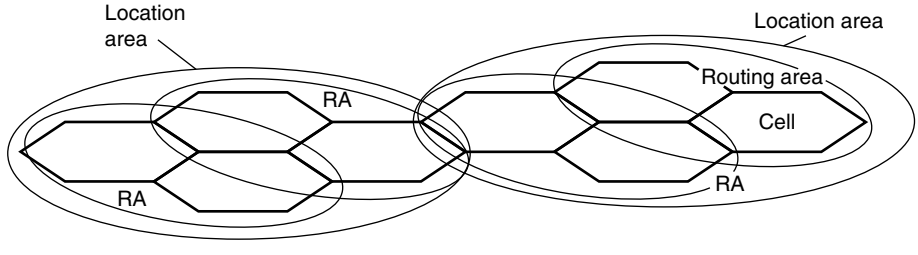


Figure 11.5 Location area and routing area

11.5 GPRS Protocol Architecture

The protocol layer is split into two planes. On one side there is the transmission plane, which is mainly used for the transfer of user data. The signaling plane, on the other side, is used for the control and support of the transmission plane functions.

11.5.1 Transmission Plane

Figure 11.6 depicts the GPRS network protocol architecture for the transmission plane. This provides transmission of user data and its associated signaling, such as for flow control, error detection, and error correction.

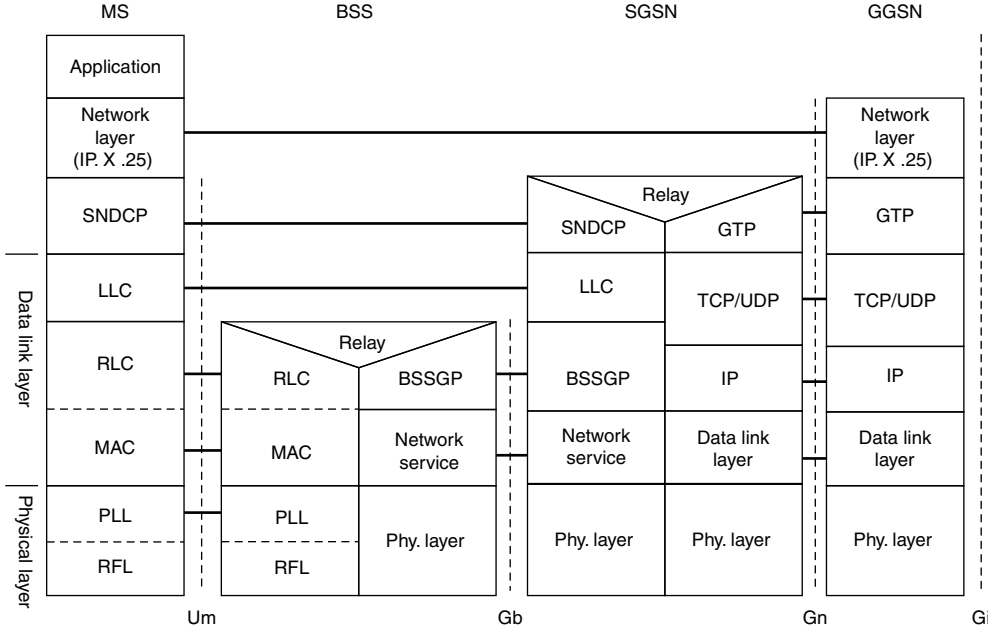


Figure 11.6 Protocol architecture of the GPRS transmission plane

11.5.1.1 Air Interface

The air interface is located between the BSS and MS, and it uses the following protocols.

1. **Radio Link Control (RLC)**: The main purpose of the RLC layer is to establish a reliable link between the MS and the BSS. This includes the segmentation and reassembly of LLC frames into RLC data blocks and ARQ of uncorrectable code words.
2. **Medium Access Control (MAC)**: The main functions of the MAC layer include controlling the access signaling procedures to the GPRS radio channel and the multiplexing of signaling and RLC blocks from different users onto the GSM physical channel. The MAC layer controls the access attempts of an MS on the radio channel shared by several MSs. It employs algorithms for contention resolution, multi-user multiplexing on a PDTCH, and scheduling and prioritizing based on the

negotiated QoS. The GPRS MAC protocol is based on the principle of slotted Aloha. In the RLC/MAC layer, both the acknowledged and unacknowledged modes of operation are supported.
3. **Physical and RF Layer (PLL, RFL)**: This layer can be split into two sub-layers: the radio frequency layer (RFL), which handles the radio and baseband part (physical channel management, modulation, demodulation, and transmission and reception of radio blocks), and the physical link layer (PLL), which manages control of the RFL (power control, synchronization, measurements, and channel coding/decoding).

A relay function is implemented in the BSS to relay the LLC PDUs between the air interface and the Gb interface.

11.5.1.2 Gb Interface

The interface between BSS and SGSN is the Gb interface. This supports data transfer in the transmission plane along with the following protocols.

1. **BSS GPRS Protocol (BSSGP)**: This is data link control on the radio link level. The BSS GPRS application protocol (BSSGP) delivers routing and QoS-related information between the BSS and SGSN.
2. **Network Service (NS)**: This transports BSSGP PDUs and is based on a frame relay connection between the BSS and SGSN.

A relay function is implemented in the SGSN to relay the packet data protocol (PDP) PDUs between the Gb and Gn interfaces.

11.5.1.3 Gn/Gp Interface

The Gn interface is defined between two GSNs (SGSN or GGSN) within the same PLMN, while the Gp interface is between two GSNs belong to different PLMNs. The Gn/Gp interface is used to transfer packets between the SGSN and the GGSN in the transmission plane. The following protocols are supported in Gn/Gp interface.

1. **GPRS Tunneling Protocol (GTP)**: This is a data link control protocol on a logical link level. The GPRS tunneling protocol (GTP) tunnels the user data packets and related signaling information between the GSNs. GTP packets carry the user's IP or X.25 packets. In the GPRS backbone, we have an IP/X.25-over-GTP-over-UDP/TCP-over-IP transport architecture.
2. **User Datagram Protocol (UDP)**: This carries GTP packet data units (PDUs) in the GPRS core network for protocols that do not require a reliable data link.
3. **Internet Protocol (IP)**: IP is employed in the network layer to route packets through the backbone.

11.5.1.4 Interface between MS and SGSN

This interface between MS and SGSN supports the following protocols.

1. **Subnet Work-dependent Convergence Protocol (SNDCP)**: This is used to transfer data packets between SGSN and MS. This protocol maps the IP protocol to the underlying network. SNDCP also

provides other functions such as compression, segmentation, and multiplexing of network layer messages.

2. **Logical Link Control (LLC)**: The logical link control (LLC) layer provides a highly reliable logical link between an MS and its assigned SGSN. Its functionality is based on the well known HDLC protocol that includes sequence control, in-order delivery, flow control, detection of transmission errors, and retransmission (automatic repeat request (ARQ). LLC frames are variable in size (with temporary ID) maximum 1600 bytes.

11.5.2 Signaling Plane

The signaling plane consists of protocols for control and support of the transmission plane functions. The signaling plane architecture between the MS and the SGSN is shown in Figure 11.7.

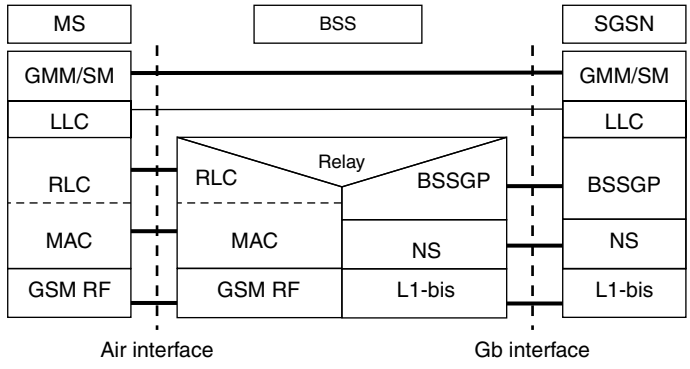


Figure 11.7 Signaling plane between MS-SGSN

11.5.2.1 Interfaces between MS and SGSN

The signaling plane is made up of the following protocols.

1. **GPRS Mobility Management (GMM)**: This supports mobility management functionalities such as GPRS attach/detach, security functions, RA update, PDP context activation, and location update.
2. **Session Management (SM)**: This supports functionalities such as PDP context activation, PDP context modification, and PDP context deactivation and so on.

The signaling architecture between SGSN and the registers HLR, VLR, and EIR uses the same protocols as conventional GSM and extends them with GPRS-specific functionality. An enhanced mobile application part (MAP) is employed between SGSN and HLR as well as between SGSN and EIR. The MAP is a mobile network specific extension of the signaling system SS7. It transports the signaling information related to location updates, routing information, user profiles, and handovers. The exchange of MAP messages is accomplished over the transaction capabilities application part (TCAP) and the signaling connection control part (SCCP). The base station system application part (BSSAP+) includes functions of the GSM's BSSAP. It is applied to transfer signaling information between the SGSN and the VLR (Gs interface). This includes signaling of the mobility management,

when coordination of GPRS and conventional GSM functions is necessary (for example, combined GPRS and non-GPRS location update, combined GPRS/IMSI attach, or paging of an MS via GPRS for an incoming GSM call).

The overall protocol software architecture inside the mobile handset of a GSM protocol stack is modified to support GPRS, which is shown in Figure 11.8.

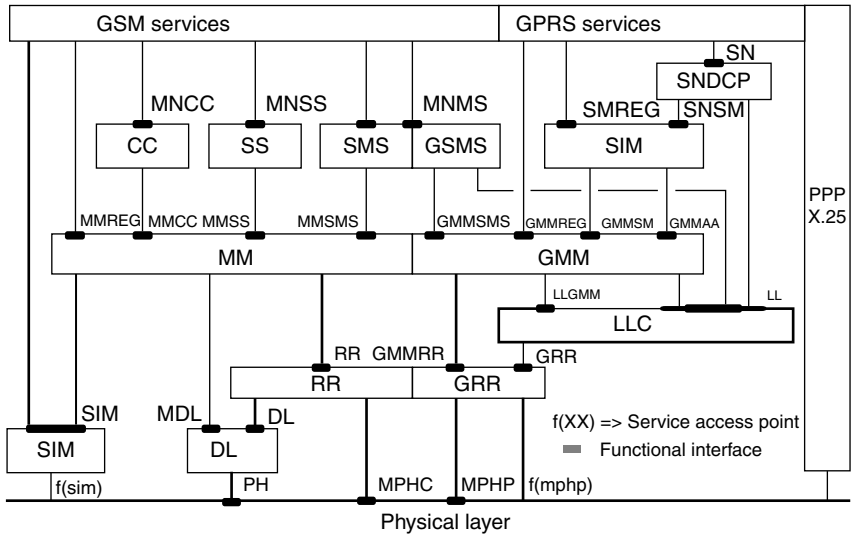


Figure 11.8 Protocol stack architecture of GPRS/GSM mobile

As discussed above, GPRS defines additional layers to adapt packets from application and for optimum resource utilization and these are SNDCP, SM, GMM, LLC, and RLC/MAC. GMM is responsible for registration, authentication, location management, and attach. SM is responsible for setting up data call, and negotiates QoS (XID parameters). RLC is responsible for packet segmentation/reassembly, flow control error correction, and so on, whereas MAC is responsible for ARFCN, TLLI, frame number, USF, TFI, RRBP, block number, radio priority, and so on.

11.6 Air Interface–Physical Layer

As discussed earlier, on the physical layer, GSM uses a combination of FDMA and TDMA for multiple access. Each of the 200 kHz frequency channels is divided into eight TDMA channels by dividing each of them into eight time slots. The eight time slots in these TDMA channels form one TDMA frame. The recurrence of one particular time slot defines a physical channel. The channel allocation in GPRS is different from the original GSM. GPRS allows a single MS to transmit or receive on multiple time slots on the same TDMA frame (multi-slot operation). It results in a very flexible channel allocation: ideally one to eight time slots per TDMA frame can be allocated for one MS (Figure 11.9). So, theoretically a maximum data rate of 11.2 $\times$ 8 kbps rates can be achieved. Moreover, uplink and downlink channels are allocated separately, which efficiently supports asymmetric data traffic. A multi-frame structure for GPRS packet data channel consists of 52 TDMA frames, whereas a GSM traffic channel uses a 26-frame structure.

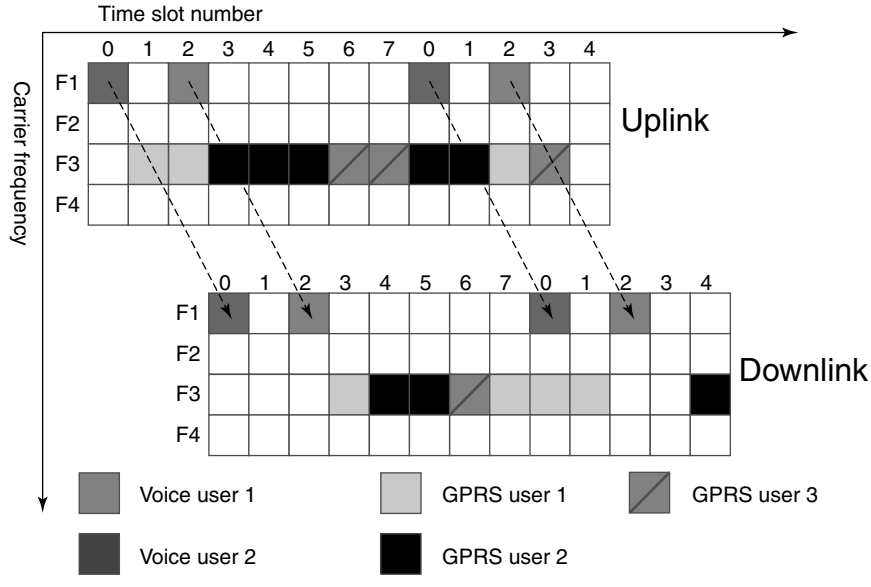


Figure 11.9 Slot allocation for users

11.6.1 Physical Channels

In GPRS the channels are allocated only when the data packets need to be sent or received, and channels are released after the transmission. With respect to the usage of scarce radio resources, this is more efficient for bursty traffic. According to this principle, multiple users can share one physical channel. A cell supporting GPRS may allocate physical channels for GPRS traffic and such a physical channel is denoted as the packet data channel (PDCH). The radio resources of a cell are shared by all GPRS and non-GPRS MSs located in this cell. The PDCHs are taken from the common pool of all channels available in the cell. The mapping of physical channels to either packet switched (GPRS) or circuit switched (conventional GSM) services can be performed dynamically (capacity on demand principle), depending on the current traffic load, the priority of the service, and the multi-slot class. PDCHs are dynamically allocated in the cell by the network. The PDCH is mapped on a 52-multi-frame structure as shown in Figure 11.10. The 52-multi-frame consists of 12 radio blocks (B0 to B11) of four consecutive TDMA frames and four idle frames (frames 12, 25, 38, and 51), leading to a total of 52 ($=3 \times 4 + 1 + 3 \times 4 + 1 + 3 \times 4 + 1 + 3 \times 4 + 1$) frames. One block essentially consists of four slots (bursts).

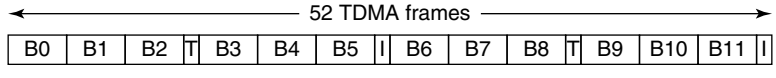


Figure 11.10 One multi-frame for PDCH

11.6.2 Logical Channels

As with GSM, GPRS uses the concept of logical channels mapped on top of the physical channels (Figure 11.11). Two types of logical channels have been defined, namely traffic channels and control

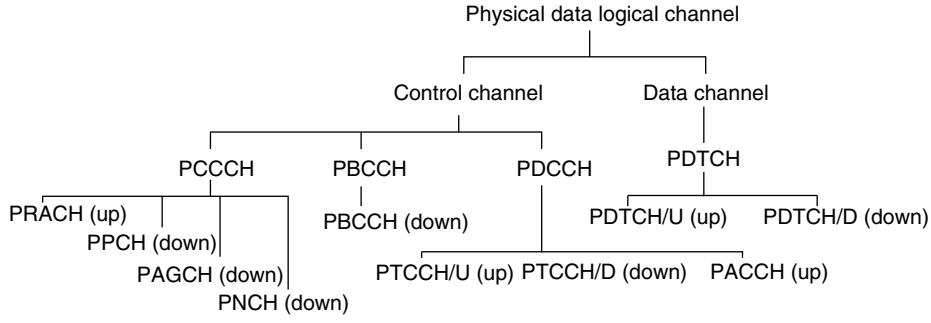


Figure 11.11 Channel structure in GPRS

channels. Three subtypes of control channels have been defined for GPRS: broadcast, common control, and associated.

Table 11.1 lists the packet data logical channels defined for GPRS.

Table 11.1 Logical channels in GPRS

Group	Channel	Function	Direction
Packet data traffic channel	PDTCH	Data traffic	MS ↔ BSS
Packet broadcast control channel	PBCCH	Broadcast control	MS ← BSS
Packet common control channel	PRACH	Packet random access	MS → BSS
(PCCCH)	PAGCH	Packet access grant	MS ← BSS
	PPCH	Paging	MS ← BSS
	PNCH	Notification	MS ← BSS
Packet dedicated control channel	PACCH	Associated control	MS ↔ BSS
	PTCCH	Timing advance control	MS ↔ BSS

In 52-multi-frame structure, there are four idle frames, out of which two TDMA frames are reserved for transmission of the PTCCH, and the remaining two frames are idle frames. The mapping of the logical channels onto the blocks B0–B11 of the multi-frame can vary from block to block and is controlled by parameters that are broadcast on the PBCCH.

Besides the 52-multi-frame, which can be used by all logical GPRS channels, a 51-multi-frame structure is also defined. It is used for PDCHs carrying only the logical channels PCCCH and PBCCH.

The packet data traffic channel (PDTCH) is employed for the transfer of user data. It is assigned to one mobile station (or in the case of PTM, to multiple mobile stations). One mobile station can use several PDTCHs simultaneously.

PBCCH is used by the BSS to broadcast specific information about the organization of the GPRS radio network to all GPRS MSs of a cell. It also broadcasts important system information about circuit switched services, so that a GSM/GPRS MS does not have to listen to the broadcast control channel (BCCH). The presence of PBCCH in the cell is optional. When there is no PBCCH in the cell, the information needed by the mobile to access the network for a packet transfer is broadcast on BCCH.

PCCCH is a bi-directional point-to-multipoint signaling channel that transports signaling information for network access management, for example, for allocation of radio resources and paging. It consists of four subchannels:

- PRACH is used by an MS to request one or more PDTCH.
- PAGCH is used to allocate one or more PDTCH to an MS.
- PPCH is used by the BSS to find out the location of an MS (paging) prior to downlink packet transmission.
- PNCH is used to inform an MS about incoming PTM messages (multicast or group call).

Packet dedicated control channel is a bi-directional PTP signaling channel. It consists of two subchannels:

- PACCH is always allocated in combination with one or more PDTCH that are assigned to one MS. It transports signaling information related to one specific MS (for example, power control information).
- PTCCH is used for adaptive frame synchronization.

11.6.3 Channel Allocation

Figure 11.12 shows the principle of the uplink channel allocation (mobile originated packet transfer). An MS requests radio resources for uplink transfer by sending a "packet channel request" on the PRACH or RACH and the network answers on PAGCH or AGCH, respectively. It tells the MS which PDCHs it can use. A so-called uplink state flag (USF) is transmitted in the downlink to tell the MS which are the time slots that the MS should use to transmit on uplink, based on whether the uplink channel is free or not.

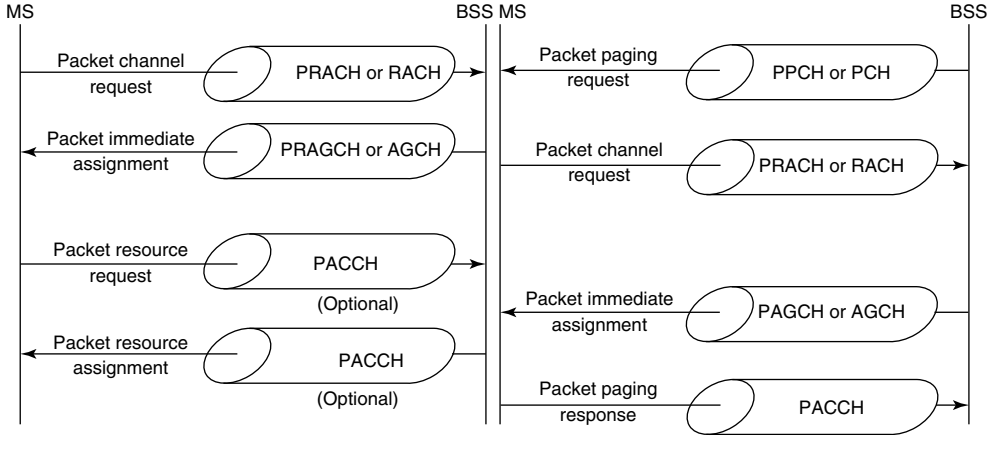


Figure 11.12 Uplink and downlink channel allocation (mobile originated packet transfer and mobile terminated packet transfer)

11.6.3.1 Reservation and Release of Radio Resources

A temporary block flow (TBF) is a physical connection between two radio resource (RR) entities (L3) to support the unidirectional transfer of LLC frames using PDCHs. The TBF is an allocated radio resource on one or more PDCHs and contains a number of RLC/MAC blocks carrying one or more LLC frames. A TBF is temporary and is maintained only for the duration of the data transfer, which means until there are no more blocks to be transmitted and all the transmitted blocks have been acknowledged by the receiving RR entity. Concurrent TBFs may be established in opposite directions (Figure 11.13). Once the data transfer is finished, the TBF is released.

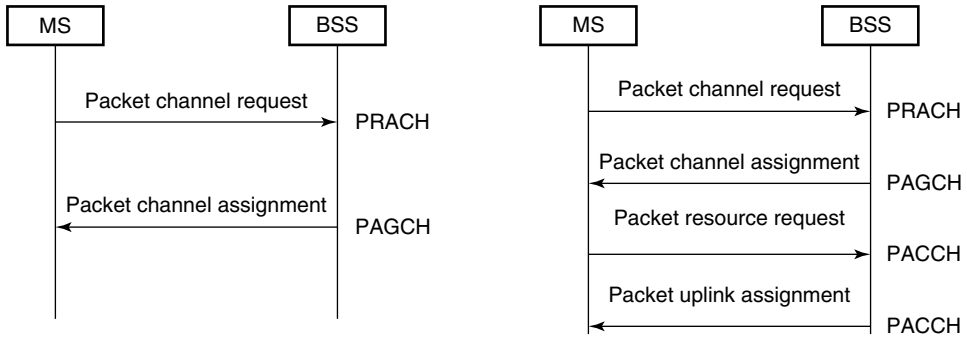


Figure 11.13 TBF establishment for 1 phase and 2 phase access

When the mobile sends continuous data to the network, it requests the establishment of an uplink TBF by sending signaling information over CCCH or PCCCH. When the network wants to send data to the mobile, it assigns a downlink TBF between the two RR entities. The number of TBFs per mobile and per direction is limited to one. However, TBFs belonging to different mobiles can share the same PDCH.

In a resource assignment message, which comes from the network to the MS, each TBF is assigned a unique temporary flow identity (TFI) by the network. The TFI is unique in both directions. It allows for the multiplexing of several users over the same time slot, and it can provide the assignment of priority classes. In the RLC/MAC layer, TFI is used instead of the MS identity, and it is included in every RLC header, which belongs to the corresponding TBF. Allocation of radio resources can happen in one or two phases. In one-phase access, the network may not know exactly, which MS owns the allocation until the first block from the MS is received by the network. In two-phase access, the MS which requested resource allocation is automatically uniquely defined. Both the network and the MS can require the two-phase access.

In principle, three different medium access modes are supported. These are called fixed allocation, dynamic allocation, and extended dynamic allocation. According to 3GPP Release 1997, the support for extended dynamic allocation is optional.

Dynamic Allocation

Dynamic allocation allows unused channels to be allocated as packet data channels (PDCHs) and if a higher priority application requires resources the PDCHs can be released. The mobile station monitors the downlink to determine when to send data on the uplink. The uplink state flag (USF) is assigned to the mobile station during the establishment of a TBF. The USF is included in the header of each RLC/MAC data block sent on the downlink. It designates which mobile is allowed to transmit data in that particular PDCH of the next uplink radio block. When the mobile station detects its assigned USF it can transmit either a single RLC/MAC block or a set of four RLC/MAC blocks. Because all the mobile stations constantly monitor the USF, the allocation scheme can be altered dynamically.

In principle, dynamic allocation allows uplink transmission to mobiles sharing the same PDCH, on a block-by-block basis. During the uplink TBF establishment, an uplink state flag (USF) is given to the MS for each allocated uplink PDCH. The USF is used as a token given by the network to allow transmission of one uplink block.

Whenever the network wants to allocate one radio block occurrence on one uplink PDCH, it includes, on the associated downlink PDCH, the USF in the radio block immediately preceding the allocated block occurrence. When the mobile decodes its assigned USF value in a radio block sent on a downlink PDCH

associated with an allocated uplink one, it transmits an uplink radio block in the next uplink radio block occurrence, which means at the B(n) radio block, if the USF was detected in the B($n-1$) radio block. The USF coding (3 bits) enables eight mobiles to be multiplexed on the same uplink PDCH. Dynamic allocation implies the constant monitoring (radio block decoding) of the downlink PDCHs associated with the allocated uplink PDCHs. The USF allows the sending of one block in the next uplink occurrence. However, dynamic allocation can also be used in such a way that the decoding of one USF value allows the mobile to send four consecutive uplink blocks on the same PDCH. The choice between one block or four blocks is indicated during the TBF establishment by the network to the mobile.

Extended Dynamic Allocation

The extended dynamic allocation scheme offers an improvement over the dynamic allocation scheme. Some RR configurations are not compliant with all MS multi-slot classes in the dynamic allocation scheme. In the dynamic allocation scheme, the MS must decode all USF values on all downlink PDCHs associated with the allocated uplink PDCHs.

Extended dynamic allocation allows the mobile station to be allocated multiple time slots in a radio block without having to monitor the USF value for each time slot. It differs from dynamic allocation in that when a mobile station sees its USF value in a particular downlink time slot, it assumes that it can use that time slot and all higher numbered time slots in the allocated set during the next uplink radio block.

The mobile monitors its assigned PDCHs starting from the lowest numbered one (the one that is mapped on the first allocated time slot in the TDMA frame), then it monitors the next lowest numbered time slot, and so on. Whenever the MS detects its assigned USF value on a PDCH, it transmits one radio block or a sequence of four radio blocks on the same PDCH and all higher-numbered assigned PDCHs. The mobile does not need to monitor the USF on these higher PDCHs. This is of particular interest in some RR configurations that are not compliant with all MS multi-slot classes in the dynamic allocation scheme.

Let us take the example of a class 12 MS, which is defined by: a maximum number of four receive time slots per TDMA frame and a maximum number of four transmit time slots per TDMA frame, but the total number of transmit and receive time slots per TDMA frame less than or equal to 5.

The network cannot allocate four uplink PDCHs to a MS multi-slot class-12 with the dynamic allocation. Indeed, the MS must decode the USF fields on the four associated downlink PDCHs. This means that the MS would have to receive on four time slots to be able to transmit on four time slots. This gives a total number of eight receive and transmit time slots, which is not compliant with a multi-slot class-12 MS. In the case of extended dynamic allocation, the network can allocate four uplink PDCHs without exceeding a total number of five receive and transmit time slots.

Fixed Allocation

Fixed allocation assigns the mobile station exclusive use of certain channels. The network commands the mobile station to use fixed allocation via the packet uplink assignment message. This message also contains a bitmap indicating the specific PDCHs, which may be used to transfer data. The network allocates uplink radio blocks using bitmaps (a series of zeros and ones). A 0 indicates that the mobile is not allowed to transmit, and a 1 indicates a transmission occurrence. The bitmaps are sent during the establishment of the uplink TBF. If more uplink resources are required during the uplink TBF, the network sends a bitmap in the downlink on the PACCH.

Fixed allocation enables a given MS to be signaled with the predetermined uplink block occurrences on which it is allowed to transmit. The network assigns to each mobile a fixed uplink resource allocation of radio blocks onto one or several PDCHs.

A fixed allocation TBF operates as an open-ended TBF when an arbitrary number of octets are transferred during the uplink TBF. When the allocated bitmap ends, the MS requests a new bitmap if it wishes to continue the TBF.

A fixed allocation TBF operates as a close-ended TBF when the MS specifies the number of octets to be transferred during the uplink TBF establishment.

11.7 Packet Data Transport Across Layers

In Figure 11.14, the PDU flow over the GPRS transmission plane is shown.

A radio block consists of one byte MAC header, followed by RLC data or an RLC/MAC control block and is terminated by a 16-bit block check sequence (BCS). It is carried by four normal bursts (that is, 114 bits long). GPRS allows a maximum of eight slots per frame to be allocated to the PDTCH on the

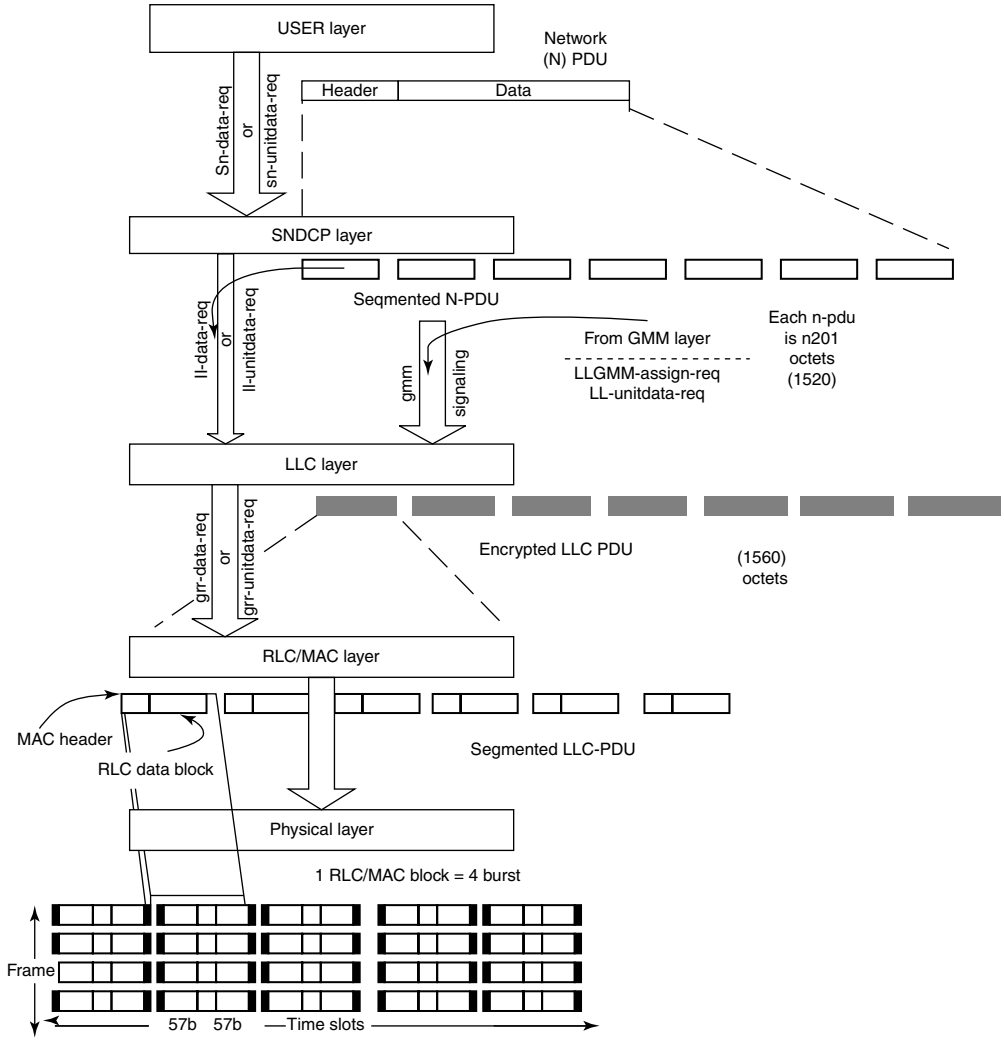


Figure 11.14 Illustrates the PDU flow through transmission plane of the GPRS protocol stack

downlink and uplink on all radio blocks B0-B11. On the downlink, an IP datagram of 1500 bytes to be transmitted as an LLC PDU must first be fragmented into 29 RLC blocks. These blocks can be transmitted using a total of 116 consecutive bursts.

Each radio block is 20 ms and contains 4 bursts, spread over 4 TDMA frames; 12 radio blocks on a 52-multi-frame is spread over 240 ms. The number of radio blocks is 50 per secondand the frame length is 4.6 ms.

As shown in Figures 11.15 and 11.16, the bursts are formed from the LLC data packet information after several processings, and finally they are mapped to normal burst of the PDTCH. As shown in the Figure 11.17, the block data of four frames each with 114 bits (57 + 57) of information is mapped into four successive PDCH. So, in the physical layer, on every 20 ms time interval (TTI) 456 bits are available after encoding and puncturing. These data are then put in four bursts each contains 57 + 57 bits of data.

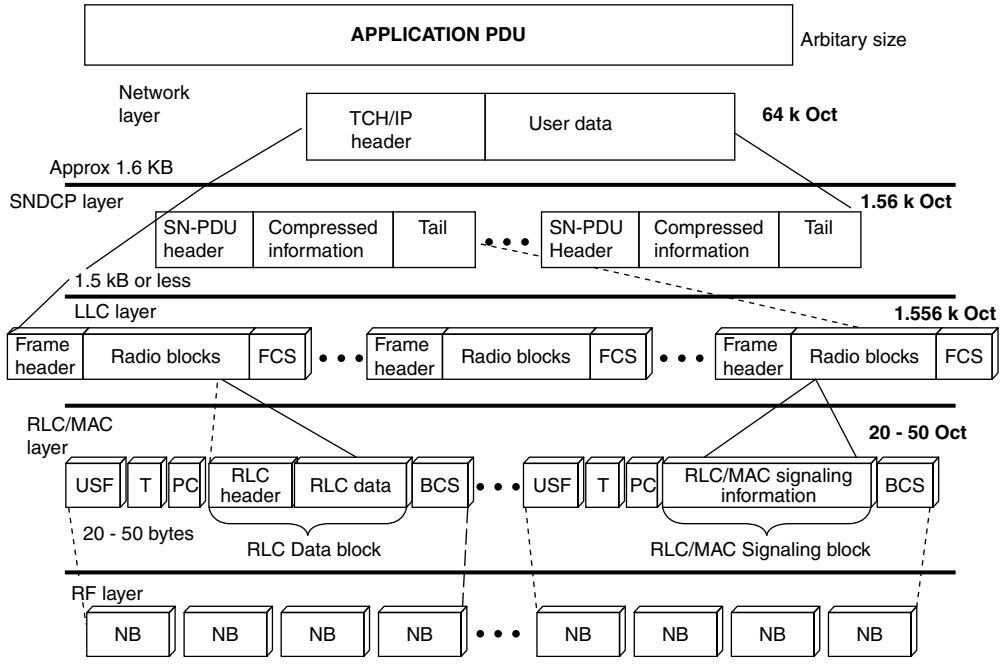


Figure 11.15 Segmentation and encapsulation of GPRS data packets. PH, packet header; FH, frame header; FCS, frame check sequence; and BCS, block check sequence

11.8 Channel Coding and Puncturing

The channel coding technique in GPRS is fairly similar to the one employed in conventional GSM. An outer block coding, an inner convolution coding, and an interleaving scheme are used here. Four different coding schemes are defined as shown in Table 11.2.

Based on the quality of the channel, one of these four coding schemes is chosen for the coding of the traffic channel (PDTCH). CS-1 offers the lowest rate but has the highest robustness. Thus, under very bad channel conditions, we may use CS-1 and obtain a data rate of 9.05 kbps per GSM time slot. However, under good channel conditions, we transmit without convolutional coding using the CS-4

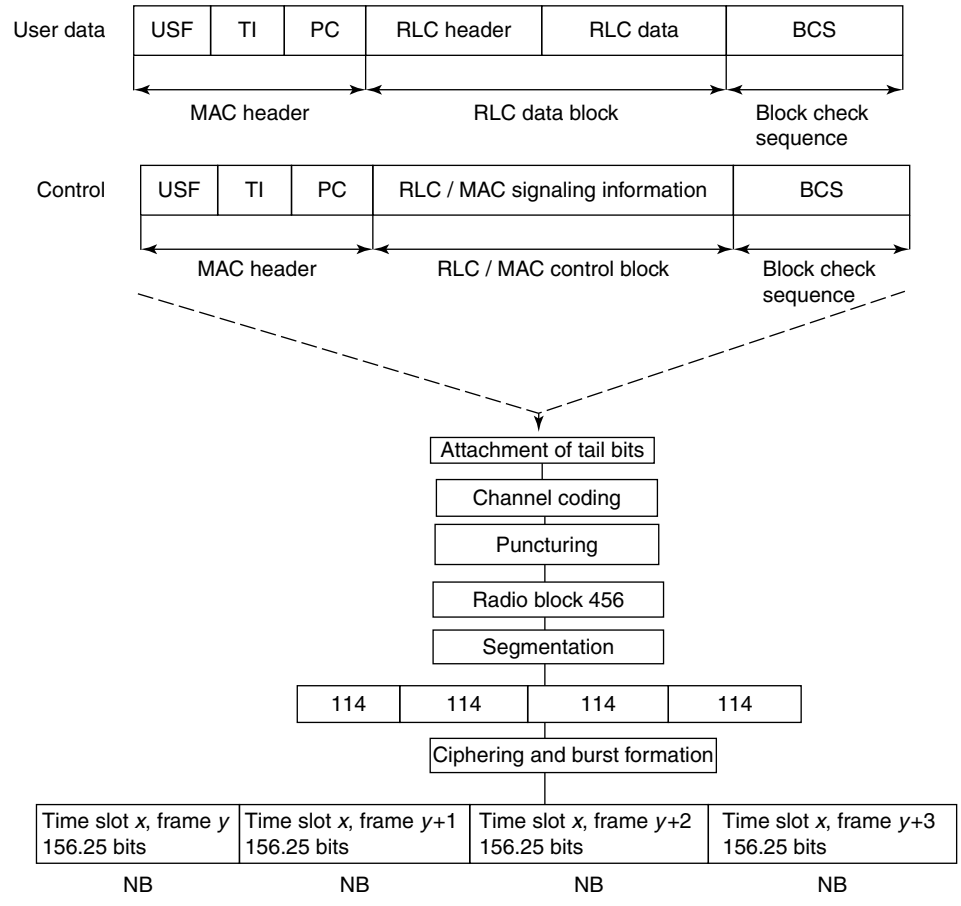


Figure 11.16 Data processing in GPRS

scheme (no protection) and achieve a data rate of 21.4 kbps per time slot. The CS-1 scheme is used for coding of the signaling channels. Using all eight time slots, we may obtain a maximum data rate of 171.2 kbps. In practice, multiple users share the time slots, hence a much lower bit rate is available to the individual user.

Table 11.2 Channel coding schemes for logical traffic channels in GPRS

Coding scheme	Pre-coded USF	Info bits without USF	Parity bits BC	Tail bits	Output convolution encoder	Punctured bits	Code rate	Data rate (kbps)
CS-1	3	181	40	4	456	0	1/2	9.05
CS-2	6	268	16	4	588	132	~2/3	13.4
CS-3	6	312	16	4	676	220	~3/4	15.6
CS-4	12	428	16	—	456	—	1	21.4

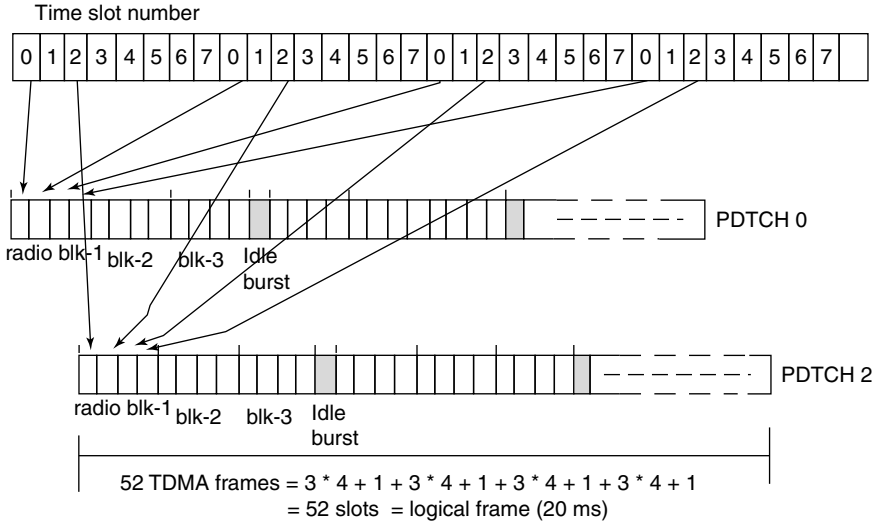


Figure 11.17 Packet data channel logical frame

The bit processing for any coding scheme is shown in Figure 11.18. With the payload, a block check sequence (BCS) is added, and thereafter the USF, either in its original form or with extra redundancy, is prefixed (pre-coded USF). Tail bits are then added and the resultant bit stream is convolutionally coded. Next, the output bits are punctured to give 456 bits (Figure 11.19). CS-1: 181-bits payload, along with the 3-bits USF, 40-bits USF and 4-bit tail bits are convolutionally coded to provide 456 bits. CS-2: 268-bits payload, 6-bits pre-coded USF, 16-bits BCS, 4-bit tail bits add up to 294, after a $^1/_2$ convolution provides a 588-bits stream, which is punctured to give 456 bits. CS-3: 312-bits payload, 6-bits pre-coded USF, 16-bits BCS, 4-bit tail bits adds up to 338 after a $^1/_2$ convolution provides a 676-bits stream, which is punctured to give 456 bits. CS-4: 428-bits payload, 12-bits precoded USF, 16-bits BCS adds up to 456 bits. Using CS-4, the three USF bits are mapped to 12 bits. No convolutional coding is applied.

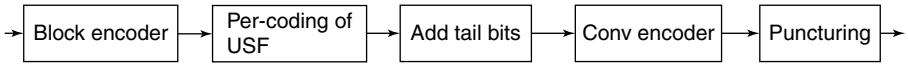


Figure 11.18 GPRS data packet processing flow

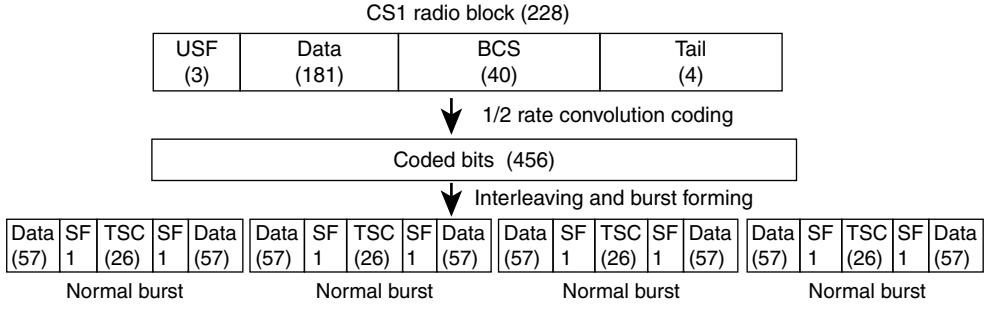


Figure 11.19 Channel coding in CS1

For the convolution coding, a non-systematic rate $^1/_2$ encoder with constraint length 4 is used. It is defined by the generator polynomials:

$$g1(D) = 1D^3 + D^4$$
$$g2(D) = 1 + D + D^3 + D^4$$

In order to intimate to the receiver about which coding scheme is used by the transmitter, the stealing flags are used. Four different coding schemes can be identified by 2 stealing flags in a burst, but a total of 8 stealing flags are used instead: CS-1 – 11111111, CS-2 – 11001000, CS-3 – 00100001, and CS-4 uses stealing flags 00010110.

11.8.1 Puncturing

The puncturing scheme used in GPRS is shown in Figure 11.20.

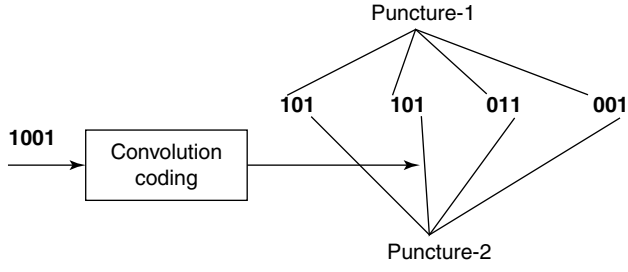


Figure 11.20 Puncturing schemes in GPRS

In Figure 11.21, the data bit processing for the CS-4 coding scheme is shown. The FBI (final bit indicator) is inside the RLC header for final block poll setting.

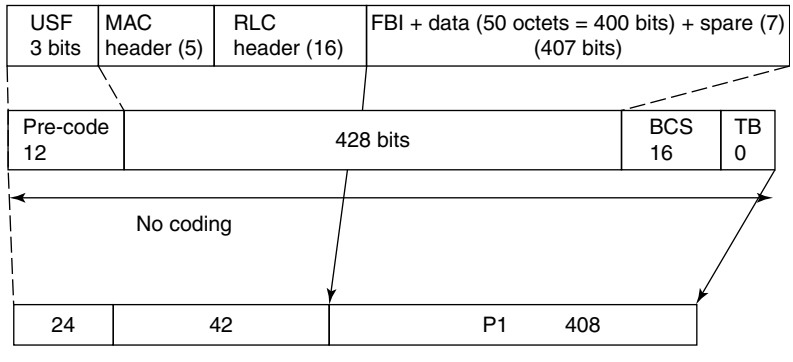


Figure 11.21 Information bit processing for CS4 coding scheme (1 symbol = 1 bit – GSMK)

After encoding, the code words are input into a block interleaver of depth 4. On the receiver side, the code words are de-interleaved. The decoding is performed by using the Viterbi algorithm.

11.9 Cell Re-selection

As in the GSM, in GPRS the mobile performs cell re-selection, but there are some differences compared with GSM. In GPRS the mobile performs cell re-selection when it is in idle mode as well as in packet transfer mode. The cell re-selection is either performed by the mobile autonomously or is optionally controlled by the network. As such, there is no handover in GPRS, just the cell re-selection. So, when there is a re-selection during a packet transfer, data transfer is interrupted and it has to be started again in the new cell. There is an interruption of the packet transfer during the re-selection phase. Although the GPRS cell re-selection algorithms used by the mobile are based on the same principles as those used in GSM, they have been slightly modified in order to provide more flexibility.

11.9.1 Routing Area Update Procedure

The MS sends an RA update request containing the cell identity and the identity of the previous routing area to the new SGSN. Then the new SGSN asks for the context (GGSN address and tunneling information) of the MS from the old SGSN. The new SGSN informs the GGSNs of the new SGSN address and tunneling information. It also informs the HLR, which cancels the MS information context in the old SGSN. The HLR loads the subscriber data to the new SGSN, which acknowledges this to the MS. The previous SGSN is requested to transmit the undelivered data to the new SGSN.

11.10 Radio Environment Monitoring

In GPRS the radio environment monitoring is essential, as it is in GSM. The MS performs different types of radio measurements and the results are reported to the network and this is used by the RLC. These estimations are also used by the mobile itself to compute its transmission power (open-loop power control), for cell selection and cell re-selection purpose.

The mobile performs the following types of measurements.

- **Received Signal Level (RXLEV) Measurements**: For the purpose of cell re-selection, the RXLEV measurements are performed on both the serving cell and neighboring cells. The serving cell RXLEV measurements can also be used for downlink coding scheme adaptation, network-controlled cell re-selection, and power control during the packet-transfer mode.
- **Quality (RXQUAL) Measurements**: During the packet-transfer mode the mobile estimates the quality of the received downlink blocks. The RXQUAL is computed from the average BER before channel decoding. The RXQUAL can be used by the network for network-controlled cell re-selection, dynamic coding scheme adaptation, and downlink power control. In the packet idle mode, no quality measurements are performed.
- **Interference Measurements**: Interference measurements have been introduced for GPRS. These correspond to a received signal level measurement performed on a frequency that is different to a beacon frequency. The interest is in having an estimation of the interference level on the PDTCH. It can be used by the network to optimize the mobile RR allocation, to select a more appropriate coding scheme, to trigger a network-controlled cell re-selection, and for power control or for network statistics.

The RXLEV and RXQUAL measurements are also performed on the BSS side. These are used for network-controlled cell re-selection, uplink power control (closed loop), and dynamic coding scheme adaptation.

11.10.1 Principles of Power Control

In a wireless environment, the power control is very important in order to reduce the co-channel interference and to improve the spectrum efficiency, while maintaining an optimum radio link quality and

to reduce the power consumption in the MS. Power control can be performed in both uplink and downlink directions. In GPRS, as the channels are not necessarily a continuous two-way connection, studying the channel is very difficult. Because of this, power control in GPRS is more complicated than for a circuit-switched connection.

11.11 Multi-Slot Class

In order to provide higher throughputs, a GPRS MS may transmit or receive using several time slots in a TDMA frame, for example, it can use one or more slots (multi-slot) of the same TDMA frame. The multi-slot capability of the GPRS MS is indicated by the multi-slot class (MSC). The implementation complexity of an MS varies, based on the number of support transmission and reception slots in a TDMA frame. To have a reasonable impact on the MS design complexity, it has been decided to allow the RF transmission on several slots of a TDMA frame, with a number of restrictions, as listed below.

1. Although several bursts can be transmitted within a TDMA frame, all should be on the same carrier frequency (ARFCN).
2. The MS needs to perform the adjacent cell measurements, or monitoring, in between the transmission and reception. Based on the mobile capability, delay constraints are needed between the transmission and reception of bursts.
3. If m time slots are allocated to an MS for reception and n time slots are allocated for transmission, then the system requires a minimum of (m,n) reception–transmission time slots that have the same time slot number (TN) within the TDMA frame.

Two types of MSs are defined.

1. **Type-1 Mobiles**: This type of mobiles are not capable of transmitting and of receiving at the same time, for example, at a given instant of time it either transmits or receives.
2. **Type-2 Mobiles**: This type of mobiles are capable of transmitting and receiving at the same time. Thus they are more complex in design, and the transmitter and receiver circuits are completely separate and expensive.

For these two types of mobile, there are various multi-slot classes, based on the capability of the MS in terms of complexity. Depending on the multi-slot class number, the mobile is able to transmit on a maximum of Tx time slots, and to receive on a maximum of Rx time slots within a TDMA frame, but the sum of (Tx + Rx) slots is limited. The maximum number of Tx slots and the maximum Rx slots are not active at the same time (see Tables 11.3 and 11.4).

Based on the following time constraints, the different types of type-1 multi-slot classes can be defined.

- $\mathbf{T_{ta}}$ – is used to set the minimum allowed delay between the end of a transmit or receive time slot and the next transmit time slot, with an adjacent cell received signal measurement to be performed in between. This is the maximum number of time slots allowed to the MS to measure an adjacent cell received signal and to get ready for transmission.
- $\mathbf{T_{rb}}$ – is the minimum delay between the end of a transmit or receive time slot and the first next receive time slot.
- $\mathbf{T_{ra}}$ – is the minimum number of allowed time slots between the end of a transmit or receive time slot and the next receive time slot, in between an adjacent cell measurement.
- $\mathbf{T_{tb}}$ – is the minimum number of time slots between the end of a receive or transmit time slot and the first next transmit time slot, without adjacent cell measurement in between.

Table 11.3 Type-1 MS

	Maximum number of slots			Minimum number of slots			
Multi-slot class	R*x*	T*x*	Sum R*x* + T*x*	T_{ta}	T_{tb}	T_{ra}	T_{rb}
1	1	1	2	3	2	4	2
2	2	1	3	3	2	3	1
3	2	2	3	3	2	3	1
4	3	1	4	3	1	3	1
5	2	2	4	3	1	3	1
6	3	2	4	3	1	3	1
7	3	3	4	3	1	3	1
8	4	1	5	3	1	2	1
9	3	2	5	3	1	2	1
10	4	2	5	3	1	2	1
11	4	3	5	3	1	2	1
12	4	4	5	2	1	2	1
19	6	2	N/A[a]	3	X[b]	2	Y[c]
20	6	3	N/A	3	X	2	Y
21	6	4	N/A	3	X	2	Y
22	6	4	N/A	2	X	2	Y
23	6	6	N/A	2	X	2	Y
24	8	2	N/A	3	X	2	Y
25	8	3	N/A	3	X	2	Y
26	8	4	N/A	3	X	2	Y
27	8	4	N/A	2	X	2	Y
28	8	6	N/A	2	X	2	Y
29	8	8	N/A	2	X	2	Y

[a] N/A indicates not applicable.
[b] X = 1 with frequency hopping or change from R*x* to T*x*; and X = 0 without frequency hopping and no change from R*x* to T*x*.
[c] Y = 1 with frequency hopping or change from T*x* to R*x*; and Y = 0 without frequency hopping and no change from T*x* to R*x*.

Table 11.4 Type-2 MS

	Maximum number of slots			Minimum number of slots[a]			
Multi-slot Class	R*x*	T*x*	Sum R*x* + T*x*	T_{ta}	T_{tb}	T_{ra}	T_{rb}
13	3	3	N/A[b]	N/A	Z[c]	3	Z
14	4	4	N/A	N/A	Z	3	Z
15	5	5	N/A	N/A	Z	3	Z
16	6	6	N/A	N/A	Z	2	Z
17	7	7	N/A	N/A	Z	1	0
18	8	8	N/A	N/A	0	0	0

[a] Note that only one monitoring window (that is, an adjacent cell power measurement window) is needed in a TDMA frame, so that only a couple (T_{ra},T_{tb}) or (T_{ta},T_{rb}) are needed to define a valid configuration of a given multi-slot class.
[b] N/A indicates not applicable.
[c] Z = 1 with frequency hopping, 0 without frequency hopping.

11.12 Dual Transfer Mode (DTM)

Dual transfer mode (DTM) is a 3GPP baseline R99 feature. It is a protocol based on the GSM standard that allows simultaneous transfer of circuit switched (CS) voice and packet switched (PS) data over the same radio channel (ARFCN). A DTM capable mobile phone can be simultaneously engaged in both CS and PS calls for voice and packet data connection in GSM/EDGE networks. A simultaneous voice and data call implies that a data call might start during an ongoing voice call or a voice call might start during an ongoing data call. If a PS call begins first, and next a CS call needs to be started, then the TBF (data call) is released. A dedicated connection for the voice call is initiated and finally, the mobile phone uses DTM for re-establishing the data connection. One common class implemented by the mobile phone vendors is the DTM Multi-slot Class 11. For example, the technical specification of a Nokia N95 states a speed of DL/UL of 177.6/118.4 kbps.

11.13 EDGE (Enhanced Data Rates for GSM Evolution) Overview

In the previous section, we noted that to support higher data rates, GPRS employs variable-rate coding schemes and multi-slot operation. The use of packet access further enhances system throughput and spectrum efficiency. However, the peak date rate for GPRS is limited to about 115 kbps, which is not sufficient for supporting popular Internet applications such as web browsing, e-mail, video services, surveillance, voice over Internet, and so on. So, even higher rates are desirable. Therefore, ETSI has developed EDGE (enhanced data rates for GSM evolution) technologies to improve the existing 2G network capacity. EDGE is generally classified as 2.75G, although it is part of ITU's 3G definition. EDGE was introduced into GSM networks in 2003, initially by the operator network Cingular in the USA.

This technology is compatible with TDMA and GSM networks. The EDGE system employs a dynamic adaptation between a number of modulation and coding schemes, as a means of providing several hundred kbps peak data rates in a macro-cellular environment, while supporting adequate robustness for impaired channels. The Hybrid ARQ (Type II) is considered to improve the performance. When it is combined with the GPRS, it allows a data flow of 384 kbps (limited to 200 kbps for the EDGE Compact) and a theoretical maximum flow of 474 kbps. The evolution of GPRS towards EDGE is known as E-GPRS (enhanced GPRS). It is also known as EDGE classic.

A major impact for supporting EDGE on an existing GSM/GPRS system has been on BSS and on MS. In this section, we will briefly discuss the EDGE/EGPRS air interface in terms of the physical layer and the RLC/MAC layer modifications to support EDGE.

The term EDGE is used to refer to both EDGE Classic and EDGE Compact.

- **EDGE Classic** – allows a total compatibility with the current GSM system.
- **EDGE Compact** – allows implementations with limited frequency spectrum (less than 1 MHz).

11.13.1 Physical Layer

The EDGE radio interface is similar to the GPRS interface. The concept of multi-frames (52 frames), physical channels (PDTCH), logical channels (PBCCH, PCCCH, PDTCH, PACCH, and PTCCH) and their mapping into physical channels is the same for GPRS. Similar to GPRS, EDGE uses a rate adaptation algorithm that adapts the modulation and coding scheme (MCS) according to the quality of the radio channel. This means that when the radio channel condition is good, it utilizes the coding schemes, which provides the higher throughput (for example, less redundant bits are added with the data bits). However, during poor channel conditions, the number of error protection bits is increased in order to reduce the bit error rate (BER) and thereby reduce the need for retransmissions. Apart from the coding scheme change, EDGE has the capability of also changing the modulation technique. EDGE uses both GMSK and 8-PSK modulation techniques, whereas GPRS uses only GMSK. One of the main improvements in EDGE is the

introduction of nine modulation–coding schemes (MCS 1–9), whereas in GPRS only four coding schemes (CS 1–4) are used. Out of these nine MCS schemes, MCS1–MCS4 use GMSK modulation, while MCS5–MCS9 use 8-PSK modulation. The former would allow the transmission of 3 bits over one symbol, whereas in GMSK this is limited to one bit per symbol. The use of 8-PSK would result in an eight point constellation diagram as show in Figure 11.22. The coding scheme to be used for any data transfer is determined by the network, depending on the radio channel conditions.

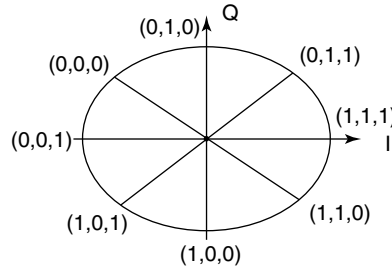


Figure 11.22 8-PSK constellation

The modulating symbol rate is the same as GSM $= 1/T = 1\,625/6$ ksymb/s (that is, approximately 270.833 ksymb/s), which corresponds to $3 \times 1\,625/6$ kbps (that is, 812.5 kbps), where T is the symbol period.

The input bits are grouped into three to make 8-PSK symbols (Figure 11.23) and then Gray mapped and 8-PSK modulated using the following rule:

$$s_i = \mathrm{e}^{j2\pi l/8}$$

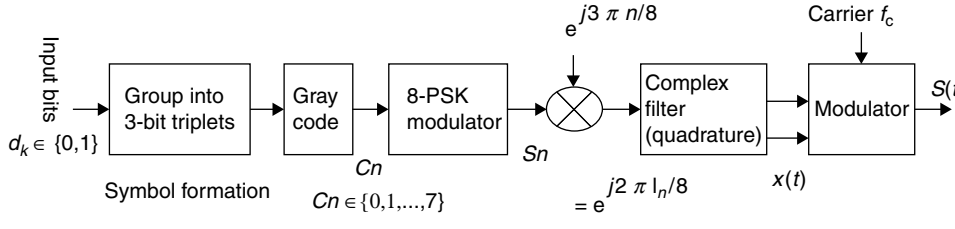


Figure 11.23 Generation of the EDGE signal (8-PSK)

The value of l is given in Table 11.5.

Table 11.5 Mapping between modulating bits and the 8PSK symbol parameter l

d_{3n}, d_{3n+1}, d_{3n+2}	0,0,0	0,0,1	0,1,0	0,1,1	1,0,0	1,0,1	1,1,0	1,1,1
C_n	3	4	2	1	6	5	7	0

The 8-PSK rotated symbols are defined as:

$$\hat{s}_i = s_i \cdot \mathrm{e}^{ji3\pi/8}$$

In 8-PSK, symbols are continuously rotated with $3.\pi/8$ radians per symbol before pulse shaping. The modulating 8-PSK symbols $\hat{s}_i$, as represented by Dirac pulses, excite a linear pulse shaping filter. This filter is a linearized GMSK pulse. The impulse response is defined by:

$$c_0(t) = \begin{cases} \prod_{i=0}^{3} S(t+iT), & \text{for} \quad 0 \le t \le 5T \\ 0, & \text{for any other } t \text{ value} \end{cases}$$

The modulated RF carrier during the useful part of the burst is therefore:

$$x(t') = \sqrt{\frac{2E_s}{T}} \mathrm{Re}\left[y(t') \cdot \mathrm{e}^{j(2\pi f_0 t' + \varphi_0)} \right]$$

where Es is the energy per modulating symbol, f_0 is the center frequency and φ_0 is a random phase, which is constant during one burst. Higher protection is required for 8-PSK, because of the threefold increase in bit rate and the higher number of transition states in the constellation diagram would also result in an increase in the symbol error rate, and thus the increase in the block error rate (BLER). Table 11.6 shows the various MCS used in EDGE. GPRS uses a $^1/_2$ rate convolutional coder and then employs different amounts of puncturing (removal of bits) to yield a code rate that is appropriate for the channel characteristics. However, EDGE uses a $^1/_3$ rate convolutional coder and selects a puncturing rate that will maximize the net throughput. Here, interleaving is performed over four frames.

Table 11.6 EDGE modulation coding schemes (MCS)

Modulation coding scheme (MCS)	Modulation	Max. throughput (kbps)	Code rate	RLC blocks/radio blocks	Family
MCS1	GMSK	8.8	0.37	1	C
MCS2	GMSK	11.2	0.49	1	B
MCS3	GMSK	14.8	0.53	1	A
MCS4	GMSK	17.6	0.66	1	C
MCS5	8-PSK	22.4	0.76	1	B
MCS6	8-PSK	29.6	0.85	1	A
MCS7	8-PSK	44.8	0.92	2	B
MCS8	8-PSK	54.4	1	2	A
MCS9	8-PSK	59.2	1	2	A

For MCS-9–MCS-5, 1 symbol contains 3 bits and for MCS-4–MCS-1, 1 symbol contains 1 bit. The higher layer PDUs are transmitted in the form of RLC/MAC data blocks, which are different in uplink and downlink directions. The block structures for GPRS and EDGE are different. The blocks are transmitted over four consecutive radio bursts on four TDMA frames of a given packet data channel (PDCH). A 20 ms EDGE radio block consists of one RLC/MAC header and either one or two RLC data blocks (the second

part is conditional). In order to support the incremental redundancy feature the header is coded and punctured independently from the data. In GPRS a radio block is interleaved and transmitted over four bursts; each one must be received correctly in order to decode the entire radio block otherwise it needs to be retransmitted.

11.13.1.1 Concept of Coding Family

The MCS coding scheme is divided into four families – A, A′, B, and C. Each of this family consists of a set of MCS and associated data unit of fixed size as shown in Table 11.7.

Table 11.7 Different MCS families and associated data unit

Family	MCS belongs to this family	Size of each data unit (bytes)
A	MCS9, MCS6, MCS3	37
A′	MCS8, MCS6, MCS3	34
B	MCS7, MCS5, MCS2	28
C	MCS4, MCS1	22

Three block sizes are defined for the nine modulation and coding schemes. This is done to facilitate the re-transmission process. The same MCS (or another MCS from the same family of MCSs) can be selected for the re-transmission of data. Three RLC block sizes and their corresponding MCSs are shown in Figure 11.24. MCS3 and MCS6 are shared between family A and A′ with data unit sizes of 37 and 34 bytes, respectively. MCS7, MCS8, and MCS9 actually transmit two radio blocks over the four bursts and the interleaving occurs over two bursts instead of four. This reduces the number of bursts that must be re-transmitted in errors. When a radio block is sent using MCS9, it consists of four data units of size 37 bytes each. However, when this block needs to be re-transmitted, the MCS3 or MCS6 (as they

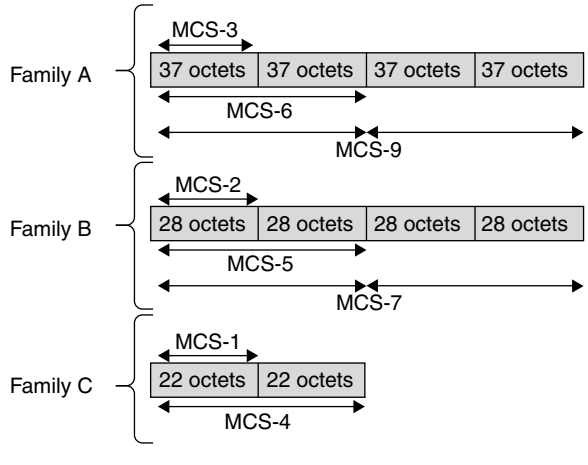


Figure 11.24 Relationship of the three RLC block sizes to the EGPRS modulation coding schemes

are from the same family) can be used. Now if MCS6 is used for the re-transmissions, then two radio blocks are required in order to transmit four data units. Hence the data rate is reduced here. Whereas, if MCS3 is used, then four blocks are needed to transfer the data units, for example, a fourfold reduction in data rate.

As an example, consider the scenario where MCS9 carries two RLC blocks each of 74 bytes in size. If the signal to interference ratio gets too low or the noise gets too high a transmission error may occur and a re-transmission will be requested. The 74-bytes blocks may then be re-transmitted using MCS6 with one block per four GSM physical layer bursts. If additional coding is required, this can be further segmented into two 37-bytes sub-blocks, and each can be transmitted using MCS3. The header would indicate that this is a segmented portion of a 74-byte RLC block and not a re-transmission using 37-byte blocks. Thus, EDGE provides plenty of flexibility for block-by-block rate adaptation.

11.13.1.2 Incremental Redundancy (IR)

Incremental redundancy (IR), also known as hybrid automatic repeat request (ARQ) type II, is achieved by puncturing a different set of bits each time a block is re-transmitted, thus gradually decreasing the effective code rate for every new transmission of the block. The principle involves the re-transmission of the data block until it is correctly decoded by the receiver. If the receiver fails to decode, it stores the soft bits at the output of the demodulator. Then these bits are used along with the newly re-transmitted bits to decode the next. The re-transmission bits are determined by the puncturing scheme used. The problem with an IR scheme is the memory requirement for storing soft decisions.

11.13.2 Link Adaptation

The dynamic selection of modulation and coding scheme according to the radio link quality is referred to as link adaptation. The EDGE standard supports a dynamic-selection algorithm that includes:

1. Downlink quality measurement and report.
2. Order for new modulation and coding for the uplink.

Link adaptation, incremental redundancy, and combinations of the two are commonly referred to as link quality control.

The radio link quality is measured in the downlink by the mobile station and in the uplink by the base station. Based on this measured radio link quality, the most appropriate coding scheme for the current prevailing radio channel conditions is determined. Ideally, the MCS can be changed for each radio block but the practical adaptation rate is usually dependent upon the measurement interval.

11.13.2.1 Measurements for Link Adaptation

The decision for which MCS should be used under different radio channel conditions is determined from the measurement report. In the case of GPRS, the decision is based on the parameter RXQUAL (0–7). In EGPRS, a new parameter called the bit error probability (BEP) has been introduced to determine the radio conditions and hence to decide the coding scheme to be used. The BEP is determined on a burst by burst basis and this is used to determine two important parameters: CV_BEP and MEAN_BEP. The CV_BEP indicates the quality from one burst to the other and thus reflects the effect of frequency hopping and the interleaving loss or gain. It is basically a coefficient of the channel quality. Alternatively, the MEAN_BEP

indicates the C/I ratio and velocity. It is evaluated by averaging the BEP values calculated on all of the four bursts of a radio block. The MS measures the GMSK_CV_BEP and GMSK_MEAN_BEP or 8_PSK_MEAN_BEP and 8_PSK_CV_BEP and sends the values for the blocks that have been received since the last measurement was sent.

The CV_BEP is computed from MEAN_BEP using the following equation:

$$\text{CV (BEP)} = \text{standard deviation of the BEP calculated within a radio block/mean BEP}$$

$$= \sqrt{1/3}\left\{\sum_{i=1} i = 4\left[\text{BEP}_i - \text{mean(BEP)}^2\right]\right\}/\text{mean BEP}$$

The measured BEP values are mapped to different logical values in the measurement report. The reported values of these parameters in the measurement report are mapped from the actual computed mean BEP and CV BEP values using a predefined table.

11.13.2.2 Adaptation Mechanism

The dynamic change of MCS schemes (in the same family) helps to adapt the data rates according to the radio channel conditions. For re-transmission a lower data rate is chosen. One example scenario is shown in Figure 11.25, where an MCS9 data block is re-transmitted using an MCS6 scheme. An MCS9 data block consists of two RLC data blocks of different BSN values, N1 and N2. When this block is split into two MCS6 data blocks, one of these data blocks consists of the first BSN value N1 and the other data block consists of second BSN value N2.

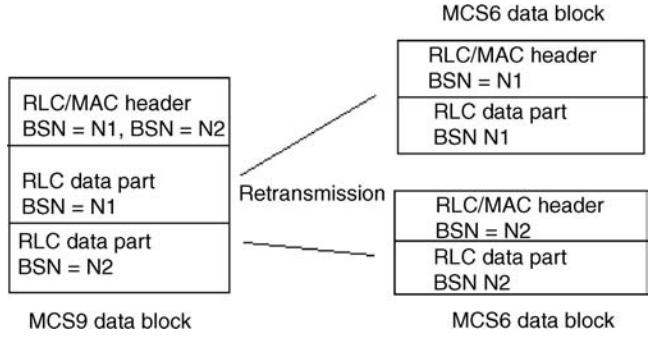


Figure 11.25 Re-transmission of RLC data block for MCS9

11.13.3 RLC Layer

In EGPRS the RLC/MAC layer and the structure of the RLC data block have been modified to support the introduction of new coding schemes and increased throughputs. Some changes have also been introduced in the procedures for TBF establishment in the EGPRS mode. RLC tasks include the segmentation and reassembly of logic link control (LLC) PDUs, LQC (link quality control) and ARQ (automatic repeat request). It is important to correct radio link errors before these are passed up to higher layers. For error

free reception, the RLC layer uses selective re-transmission to correct errors. This scheme only requires that erroneous frames to be re-transmitted. The correctly received frames are buffered until the erroneous frame is received correctly and then all the frames are placed in the correct order and sent to the upper layer, which is the logical link control (LLC) layer. Block sequence numbers (BSNs) are assigned in order to complete this reassembly task as well as to detect missing radio blocks.

11.13.3.1 RLC Data Blocks

As mentioned earlier, the RLC data block consists of one header and is followed by one or two data blocks based on the MCS scheme (MCS 7, 8, 9 schemes use two data blocks). Both the data and header part are encoded using different coding schemes but at the same rate. For nine different coding schemes, three header types are defined: (1) header type-1, defined for MCS7–MCS9; (2) header type-2, defined for MCS5 and MCS6; and (3) header type-3, defined for MCS1–MCS4. Headers are different for downlink and uplink. The parameters for downlink header are:

a. **Temporary Flow Identifier (TFI)** – This is a 5-bit field and spread over octets 1 and 2. This identifies the TBF to which the data blocks belong.
b. **Relative Reserved Block Period (RRBP)** – This indicates the position of the uplink radio block on which the mobile should transmit the message.
c. **Block Sequence Number (BSN)** – The transferred data blocks are numbered in-sequence which is known as BSN. For EGPRS this is an 11-bit field. In the case two RLC data blocks carried in the data block, the BSN of the second block BSN2 is encoded relative to the BSN of the first block BSN1.
d. **EGPRS Supplementary/Polling Bit (E S/P)** – This field indicates, whether the RRBP field is valid or not.
e. **Uplink State Flag (USF)** – This was discussed earlier, and is used in the dynamic allocation of the resources to determine which mobile would transmit in the next uplink radio block.
f. **Power Reduction (PR)** – This indicates the power reduction of the current RLC block.
g. **Coding and Puncturing Scheme** – This is used to indicate the coding and puncturing scheme used for the current data block.
h. **Split Block Indicator (SPB)** – This is used in the case of header type-3. It indicates whether the particular data block is segmented and re-transmitted using two data blocks.

In the uplink direction, some of the parameters in the uplink header are the same as in the downlink header. Some other parameters are:

i. **Countdown Value (CV)** – This parameter allows the network to compute the remaining RLC data blocks to be transferred in the ongoing data transfer.
j. **PFI Indicator (PI)** – This indicates the presence or absence of a PFI field.
k. **Stall Indicator (SI)** – The value "0" indicates the window is not stalled, "1" indicates stalled.
l. **Retry (R)** – This indicates whether the channel request message is sent once or more than once in the most recent access request.
m. **Resent Block Bit (RSB)** – This indicates whether an RLC data block has been transmitted previously or not.

The channel coding and burst formation for the MCS1–4 schemes in EGPRS are shown in Figure 11.26. For MCS1–3, first the USF is pre-coded with a block code, and then coded with a convolution code. However, for MCS4, the USF is pre-coded with block code only. The MS can detect the stealing flag and decode the USF from the EGPRS data block, because in all the coding schemes the same stealing flag as the CS4 coding scheme is used. Then 8 bits of the header check sequence (HCS) and 12-bits BCS are added to the header and data part. This is then encoded with the rate $^1/_3$ convolution code. Both the header

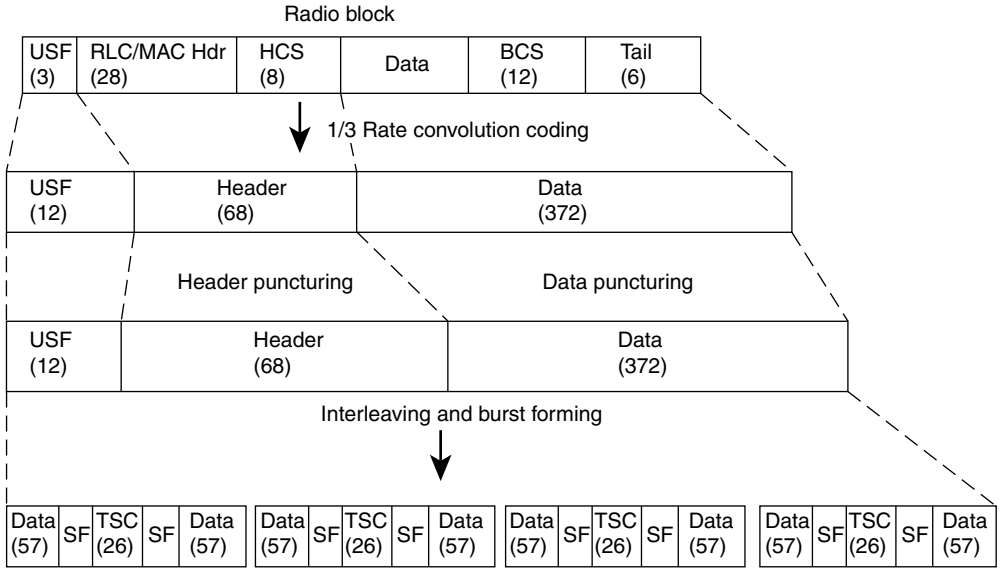


Figure 11.26 Channel coding for MCS1-4 data blocks

and the data part are punctured and this results in a data block of size 452 bits. Apart from the 452 bits (comprising of USF, header, and data), there are $8 + 4 = 12$ bits stealing flags inserted. Eight stealing flags indicate the coding scheme used, four other extra stealing flags are spread across the four normal burst and are kept for future use and presently set to 0. In the case of MCS7–9, the data blocks contain two RLC data blocks and the USF is pre-coded with 36 bits. The block check sequence of 8 bits for the header and 12 bits for each data part are added. The encoding of the data block is done using $^1/_3$ convolution code and then punctured using different puncturing schemes. The encoded bits are interleaved and transmitted over the air interface.

As in the GPRS, the CS1 scheme is used to control channels, similarly in EGPRS also the same coding scheme is used for all signaling procedures.

The RLC layer supports two modes of operation.

Unacknowledged Operation

Unacknowledged operation does not guarantee the arrival of the transmitted RLC blocks and there is constant delay. The receiver attempts to preserve the length of the data blocks it receives. This is useful for real time applications such as video.

Acknowledged Operation

The acknowledged operation guarantees the arrival of the transmitted RLC blocks. Selective re-transmission is used to re-transmit erroneous data blocks. For each RLC peer-to-peer entity there is a transmit and receive window size established that allows a limited number of blocks to be transmitted prior to receiving an acknowledgement. The window size for EDGE is set according to the number of time slots allocated in the direction of the TBF and ranges from 64 to 192 for single time slot operation or from 64 to 1024 for 8-time slot operation. In GPRS the window size is set at 64.

11.13.4 Data Transfer

Once the PDP context activation has been completed, the data session may begin. Communication between the SGSN and the GGSN is achieved through the use of tunneling. This is the process of adding a header to the existing packet, so that it can be routed through the backbone network. When the packet reaches the far side of the GPRS network, the additional header is discarded and the packet continues on its route based on the original header. The use of tunneling helps solve the problem of mobility for the packet networks and eliminates the complex task of protocol interworking. The GPRS system employs tunneling when sending packets from the mobile station to fixed nodes and also when sending from fixed nodes to mobile stations. This is a distinction from mobile IP, which only uses tunneling in the second case.

11.13.5 Medium Access Control (MAC)

The MAC controls channel access, resource allocation, resource management, and thus enables fixed or multiplexed use of multiple time slots (TS). The MAC layer provides the capability for multiple mobile stations to share the same transmission medium through the use of contention resolution and scheduling procedures. A reservation protocol based on the Slotted Aloha protocol is used for contention resolution among several mobile stations. The MAC layer uses three modes to control the transfer of data in the uplink. The initial mode is specified when the temporary block flow (TBF) is established. The MAC layer assigns temporary block flows (TBFs) for data and signaling transfer between the MS and the network. The TBF is used by each entity to communicate LLC PDUs on the packet data physical channels (PDCH) between the RLC entities on each side of the communication link, as discussed earlier. In general the MAC functionality is similar to that of GPRS. The EGPRS packet channel request message has been defined for initiating the EGPRS data transfer. This is sent on RACH or PRACH. MS specifies its EGPRS capabilities during this message indicating two training sequences. These training sequences are defined in addition to the existing ones. One training sequence indicates that the MS is EGPRS capable and supports 8-PSK in both uplink and downlink directions, while the second training sequence indicates MS is EGPRS capable but supports 8-PSK only in downlink. As for GPRS, if a TBF has to be established in an RLC acknowledged mode, the two phase access procedure is mandatory. For a two phase access procedure, the RLC mode is the acknowledged mode by default.

11.13.6 Impact of EDGE Support on Air Interface and Equipments

The support of EDGE has a direct influence on the design of base stations as well as mobile terminals. New terminals and base station transceivers must be developed that can transmit and receive EDGE-modulated information.

11.13.6.1 Mobile Stations

GSM mobile stations must be designed with the appropriate protocol layers for them to support GPRS and EDGE. They must also be modified to operate on shared traffic channels and the coding schemes must be added. If the MS is EDGE-capable this means it must also implement a new modulation scheme (8-PSK). 8-PSK is a linear modulation type and requires a linear power amplifier. This is especially true for high-output power equipment. Indeed, the designer's challenge is to build a cost-effective transmitter while fulfilling the GSM spectrum mask. EDGE transceiver performance must be acceptable in terms of both transmit spectrum and heat dissipation. Compared with GMSK, the

average power decrease (APD) in for 8-PSK could be between 2 and 5 dB. The design of a good sub-optimum equalizer for 8-PSK will be slightly more complex than that of a standard GSM equalizer. The increased bit rate (compared with standard GPRS) also reduces robustness in terms of time dispersion and mobile-terminal velocity.

Today, the GSM standard includes several classes of mobile terminals, ranging from single-slot devices with low complexity to eight-slot devices with high bit rates. EDGE technology has introduced several new classes, with different combinations of modulation and multi-slot capabilities, such as MCS 45 with a maximum number of R$x = 6$ and T$x = 6$ with sum $= 7$. There are three classes of mobile stations:

- **Class A**: Allows for simultaneous use of GPRS/EDGE and other GSM services (such as voice).
- **Class B**: Alternate use of GPRS/EDGE or GSM services is possible. Only one can be used at a time but it is possible to toggle back and forth.
- **Class C**: Designed for GPRS/EDGE only. This class provides no voice service.

The EDGE capable MS is divided into two categories: type-1, which supports both GMSK and 8-PSK modulation schemes in downlink and only GMSK in the uplink direction, whereas the type-2 MS supports both GMSK and 8-PSK in both uplink and downlink directions.

EDGE evolution continues in Release 7 and 8 of the 3GPP standard for providing doubled performance, for example, to complement high-speed packet access (HSPA).

11.13.6.2 Impact on GSM Network Architecture

The introduction of EDGE has very limited impact on the core network, and because the GPRS nodes, SGSNs, and gateway GPRS support nodes (GGSN) are more or less independent of user data rates, no new hardware is required. However, the BTS should support the transmission and reception of a new modulation (8-PSK) and coding scheme. An apparent bottleneck is the Abis interface, which currently supports up to a 16 kbps per traffic channel and time slot. With EDGE, the bit rate per traffic channel will approach or exceed 64 kbps, which makes it necessary to allocate multiple Abis slots to each traffic channel.

11.14 Latest Advancements in GERAN (GSM/GPRS/EDGE Radio Access Network) Standard

11.14.1 EDGE Evolution

The standardization for EDGE (enhanced data rates for global evolution) was finalized by the 3GPP in year 2000. Since then EDGE has achieved market maturity in terms of networks, terminals and business models. Today it offers user bit-rates of around 250 kbps, with end-to-end latency of less than 150 ms. EGPRS2 is a term introduced in GERAN 3GPP Release 7, which specifies a two-level support – level A always refers to the use of the legacy symbol rate (270 833 ksymb/s), whereas level B always refers to the use of the increased symbol rate (325 ksymb/s). Level B will have a more significant impact on the existing network deployments.

EGPRS2 Coding and Modulation: For downlink 16 additional modulation and coding schemes (DAS-5–DAS-12 for 2A and DBS-5–DBS-12 for 2B) and for uplink 13 additional modulation and coding schemes (UAS-7–UAS-11 for 2A and UBS-5–UBS-12 for 2B) are defined for EGPRS2 packet data traffic channels. The Turbo coding is used in downlink only and the scheme of the Turbo

Coder is a Parallel Concatenated Convolutional Code (PCCC) with two 8-state constituent encoders and one Turbo code internal interleaver. The coding rate of the Turbo Coder is $^1/_3$. Based on MS feedback, the coding and modulation scheme is adapted to current propagation and signal-to-noise/interference conditions.

Latency Reduction: The latency is reduced by introduction of a reduced transmission time interval (RTTI) and additional protocol enhancements. The blocks are currently transmitted over four consecutive bursts on one time slot using a TTI of 20 ms, whereas in RTTI, it is reduced to 10 ms. In a reduced TTI configuration, a radio block consisting of four bursts is sent using two PDCHs, that is, a PDCH-pair, in each of two consecutive TDMA frames. In a reduced TTI configuration, the time to transmit a radio block is half of a basic radio block period.

Dual Carriers: The introduction of dual carriers doubles the available bandwidth (to 400 kHz) as well as the practical peak bit-rate. Now, this is part of the 2B standard only, and it allows the two downlink carriers to be used simultaneously for one channel. Using dual carriers and five time slots on each carrier provides bit-rates of almost 600 kbps, with no other changes to EDGE.

Dual-Antenna Terminals: By combining signals from the two antennas (mobile station receive diversity, MSRD), a large proportion of the interference can be cancelled out (dual-antenna interference cancellation, DAIC) significantly, improving the average bit-rates and spectrum efficiency. SAIC is a part of the DARP Phase-I requirement, whereas MSRD is a part of DARP Phase-II requirement.

11.14.2 Voice Services Over Adaptive Multi-User Orthogonal Subchannels (VAMOS)

Today, the GSM network is the most successful commercial cellular mobile communication system having more than 3.0 billion subscribers all over the world. It is still growing continuously, because of the growing demand for mobile voice services in emerging markets, especially in China and India, where the subscriber density is very high. As day by day the voice service cost is getting cheaper and new subscribers are registering for the service, this is why most of operators are now facing a real challenge to support so many simultaneous voice calls using limited radio resource.

In the GSM system, the total available physical radio channels in both the directions are: $124 * 8 = 992$. Out of these total available frequency carriers, some are used as broadcast frequencies. Thus these usages of physical radio channels limits the number of available traffic channels. During the initial rollout of GSM networks, the main concern was to ensure sufficient coverage at a reasonable cost. Even though coverage is still important, now many networks are limited by the number of users they can serve simultaneously with a sufficient quality. As the capacity of the wireless network increases, the network operators and vendors are challenged to find various creative ways to increase the capacity using the limited spectrum and resources of the network.

One option to accommodate an increased number of users is to introduce smaller cells and a tighter frequency re-use to increase the number of physical channels over a geographical area. However, this approach also leads to higher interference levels and today the capacity of many networks is in fact limited by interference. Smaller cells lead to an increase in the physical channels per user ratio (also lesser transmit power), but as in this case, cells will be closely spaced (more frequency re-use factor), so this will create more co-channel interference. Also, the solution to this situation is the use of the single antenna interference cancellation (SAIC) technique, which is discussed in Chapter 3. Another innovative option is to use the same radio physical channel (frequency, time slot) for multiple users without sacrificing the service quality. The multi-user re-using-one-slot (MUROS) technique originates a new idea to enhance the capacity of both voice and data service.

In MUROS, we create an orthogonal subchannel (OSC), where two users use one time slot with the same ARFCN. The initial idea was proposed by NSN under the name of OSC. It was termed

MUROS when it became a study item (SI) by 3GPP. Later, MUROS became a 3GPP work item (WI) at GERAN#40 and denoted voice services over adaptive multi-user orthogonal subchannels (VAMOS), which will be incorporated into 3GPP GERAN Release 9, scheduled for December 2009. An important motivation for MUROS is that it can take advantage of the widely available DARP Phase I capable MS, that is handsets supporting single antenna interference cancellation (SAIC) technology.

Basic Principle: As discussed earlier, DARP was specified to provide improved reception on the mobile station side, when there is co-channel or adjacent channel interference. However, when the downlink signal quality is good, then there is little benefit from DARP. This fact was exploited when the idea of MUROS was conceived, where we intentionally created the co-channel interference scenario in the system by assigning the same physical channel (f_{TS}) to two different mobile stations, as we know that the DARP receiver is capable of interference cancellation. Therefore, in the downlink the network can assign the same physical resources (for example, the same frequency and time-slot combination) to two different mobile stations, but allocates them different training sequence codes (TSC) for channel estimation. As the frequencies of both mobiles are the same, so both these signals will be passed transparently via the RF filter circuits of both receivers, and as these are placed at the same time slot then their energy is just mixed with each other in the channel (own signal with the co-TCH user). This can be treated as a synchronous co-channel interference scenario. Now, each mobile's digital receiver has to use an interference cancellation technique to reject the other user's signal energy from the received signal before decoding. On the uplink each mobile station would use a different training sequence code. The network can use techniques such as joint detection to separate the two users on the uplink.

Introduction of New Training Sequence: The FCCH channel uses FB, SCH uses SB, and RACCH uses AB, and other logical channels, including TCH, use NB. Presently eight different types of TSCs are defined to be used inside NB in the GSM system. Out of these eight, each training sequence code (TSC) is assigned to a cell, and all the adjacent cells must be assigned with different TSCs. The cell BTS number and used TSC are inter-linked and mapped accordingly. The MUROS concept introduces eight (or more) new TSCs (Figure 11.27) in order that each cell can be assigned two or even more TSCs without changing the current frequency planning. Each BSS uses the TSCs to pack two or even more users onto one slot.

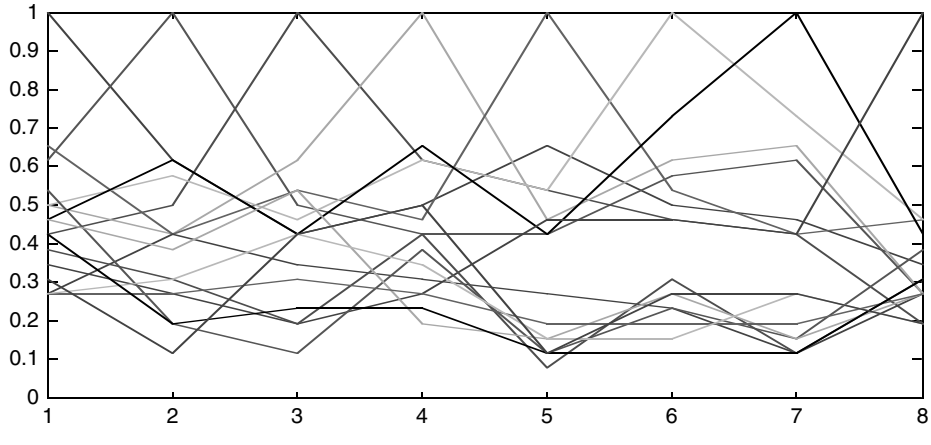


Figure 11.27 Illustration of cross correlation properties between existing training sequences (solid line) and between new and existing training sequences (gray line)

A BTS is assigned two TSCs, one legacy TSC and one new TSC. These must be low cross-correlated so that receivers can differentiate their own signal from MUROS partners in the same slot, with interference cancelling (IC) technology, such as space–time interference rejection combining (STIRC), successive interference cancellation (SIC) or joint detection (JD) in uplink, or single antenna interference cancellation (SAIC) in downlink.

A number of candidate techniques using a new set of training sequences were proposed for speech capacity enhancement under the MUROS study item. In total seven different sets of new training sequences were proposed by Ericsson, Nokia, Motorola, China Mobile, RIM, and Huawai. The set of new training sequences dedicated to the second subchannel are paired with current training sequences for the lowest cross-correlation with optimal autocorrelation and are listed in Table 11.8.

Table 11.8 Set of new training sequences (TSCs) paired with current ones

Training sequence code	Training sequence bits
0	0 0 1 0 1 1 0 1 1 1 0 1 1 1 0 1 0 0 0 1 1 1 1 0 1 1
1	0 0 0 1 0 0 0 1 1 1 1 0 1 0 0 1 0 0 1 0 0 0 1 0 0 0
2	0 1 1 1 0 1 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1 1 1 1 0
3	0 1 0 0 0 1 0 0 0 1 1 1 0 0 0 0 1 0 1 1 0 1 1 1 0 1
4	0 1 0 0 0 1 0 1 1 0 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 0
5	0 1 0 1 1 1 1 1 0 0 1 0 0 1 1 1 0 0 1 0 1 0 0 0 0 0
6	0 1 1 1 0 1 1 1 1 0 0 1 0 1 1 1 1 0 0 1 0 0 0 1 0 1
7	0 0 1 0 1 0 1 1 0 0 1 1 1 1 1 1 0 0 1 1 1 1 0 1 0 1

Different Options for VAMOS Support: Four candidate techniques are proposed for VAMOS, but it is assumed that VAMOS will be based on one of the three candidate proposals: OSC, alpha-QPSK or Co-TCH. However, from a terminal point of view the VAMOS signals from these three candidate techniques look similar. Three different types of MUROS modulation schemes for DL signals and their spectrums have been analyzed and these are: (1) linear sum of two GMSK signals (90° phase separation), (2) QPSK with linear Gaussian filter, and (3) QPSK with RRC (roll-off 0.3) filter. The QPSK with RRC filter (roll-off 0.3) spectrum is 10 dB higher than the GMSK spectrum between 140 and 200 kHz. The linear combination of two GMSK signals has the same spectrum (curve overlapping) as normal GMSK signals, as expected. It is within the specification defined GMSK mask for the useful part of the burst. QPSK with a linear Gaussian filter (8PSK pulse shaping) is similar to GMSK. Based on these observations, it is proposed that two GMSK linear combinations and QPSK with a linear Gaussian filter should be the candidates for a MUROS DL modulation scheme.

VAMOS Support for Different Logical Channels: VAMOS allows multiplexing of two users simultaneously on the same radio resource in the CS domain both in downlink and in uplink, using the same time-slot number, ARFCN, and TDMA frame number. The logical channels used for VAMOS are TCH/FS, TCH/HS, TCH/EFS, TCH/AFS, TCH/AHS, TCH/WFS and the corresponding types of associated control channels (FACCH and SACCH).

11.14.2.1 Impact on Various Entities in the GSM System

In theory, MUROS can double the voice capacity (or even more) with negligible impact to the handsets or to the networks. The co-TCH concept requires that one of the two mobile stations must support DARP

Phase-I. It is not necessary for the two mobiles to support the new training sequence codes provided the network assigns a different TSC to each of the two mobile stations and these two training sequence codes are not used by neighbouring cells, which use the same frequency (ARFCN).

Impact on Mobile Station

Support of Legacy Mobile Stations: No implementation impacts are required for legacy MS types. The first priority is supporting the legacy DARP Phase-I capable terminals, whilst the second priority is given to supporting the legacy GMSK terminals not supporting DARP Phase-I capability.

Implementation Impacts for New Mobile Stations: For new mobiles it is desirable to support the new TSC set in addition to the existing TSC set so that network has more flexibility in selecting the TSCs. In order to support the use of this new TSC set, radio resource signalling changes would be required. There would be a very minimum impact on MUROS supported MS hardware. Also, the additional complexity in terms of processing power and memory should be kept as minimum as possible.

MS Support Levels for VAMOS: A terminal supporting VAMOS supports a new set of GSM TSC and may be based on either DARP Phase-I or an advanced receiver architecture. Hence two different terminal support levels will be specified for VAMOS aware terminals.

- **Level 1**: These terminals are basically DARP Phase-I terminals updated to support the new VAMOS TSC set. One major difference between the old SAIC receiver and the Vamos Level-1 receiver is that 8 new training sequences are introduced.
- **Level 2**: These terminals should have a more advanced RX, fulfilling some tightened 3GPP performance requirements. VAMOS aware terminals will have to indicate their support level to the network in order for the network to exploit the capacity to the highest extent.

Impacts on BSS (BTS and BSC)

The introduction of the candidate techniques proposed under MUROS should change BSS hardware as little as possible and HW upgrades to the BSS should be avoided. For MUROS operation each BTS should have two TSCs. During the channel assignment it should use these TSCs and support the QPSK modulation. When both OSCs are activated, the baseband modulator can take both streams of binary data with two different TSCs applied to the corresponding payload streams, and modulate them in such a way that it satisfies GMSK as defined in the specification, and they are effectively the linear sum of the two independent burst signals that can be well received by legacy mobile stations. In addition, the receiver needs to be able to decode the two GMSK modulated signals that are separated by a TSC. There are a number of BTS receiver techniques that can be employed to provide adequate performance on the uplink, such as duel antenna, joint detection. Abis bandwidth should now be doubled, so that it can carry twice as much as the voice data could before MUROS deployment.

Impact on Radio Resource Management: Radio resource management (RRM) is considered the most vital component in voice capacity enhancement. RRM has to do the following. (1) Determine the most appropriate users to pair together. This may involve the power requirements of each user; the rate of power change, or signal quality. (2) Power control is crucial to providing maximum benefit from the MUROS mode of operation. Power control can allow users with varying channel conditions to be kept in the MUORS mode for longer. Fast power control (that is, enhanced power control) can be valuable for the MUROS mode operation. (3) Determine the most appropriate point to un-pair users.

This has to be a balance between maintaining call quality and spectral efficiency. In order to support pairing and un-paring of users BSS can use existing procedures to move users from one TCH to another (that is, intra-cell handover command or assignment command). It is down to the BSS implementation that is used.

Impact on Network Planning

The impacts on network planning and frequency re-use are minimized. Impacts to legacy MS interfered on downlink by the MUROS candidate technique should be avoided in case of usage of a wider transmit pulse shape on downlink.

$$\text{Network capacity gain} = \frac{capacity\,(with\_MUROS)}{capacity\,(without\_MUROS)}$$

Impacts on the Specification: The impact on specifications is shown in Table 11.9.

Table 11.9 Impact on specification with co-TCH

3GPP specification	Impact
TS 44.018 – RR signaling	Signaling changes to support new TSC for use with CS connections
TS 24.008	Signaling changes for MS to indicate support for new TSC set
TS 45.002	Defines new TSC set
TS 45.004	Defines new modulation scheme for downlink
TS 45.005	Defines performance requirements for MUROS type modulation
TS 51.010	Defines new performance and signaling tests cases for MUROS capable mobiles

11.14.2.2 MUROS Basic Operations at the Network and MS Side

Now, in a MUROS supported scenario, when the network assigns a channel to the MS, that time tries to pair up two MSs. Of these two MSs, one should be a VAMOS supported phone and the other one may be a VAMOS supported phone or any legacy DARP supported phone. During the RRC signaling, the MS provides the information about its VAMOS support capability. During the channel assignment, the network will assign the same ARFCN and TS to both the MSs, but to the VAMOS supported MS, it will assign TSC_n1 from the new set and to the old legacy MS it will assign TSC_o1. TSC_n1 and TSC_o1 have minimum cross-correlation properties. This means the first sub-channel can use an existing TSC and the second sub-channel should use the corresponding new one for both downlink and uplink.

The MUROS supported transmission and reception scenario is shown in Figure 11.28.

BTS Transmission: The data from user-1 and -2 are channel encoded (and ciphered individually using A5/1 or A5/3) and then bursts are formed using TSC_n1 and TSC_o1, respectively. Then it is QPSK modulated.

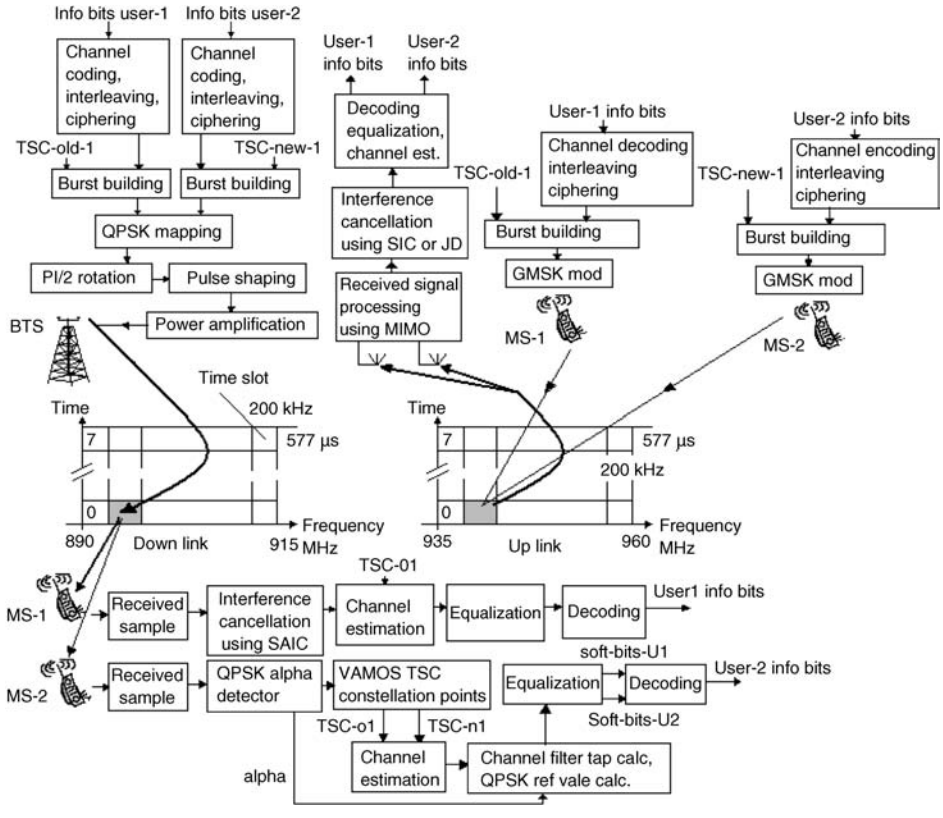


Figure 11.28 Transmission and reception by MS and BTS in MUROS supported scenario

Modulation: Two modulation schemes (GMSK and 8PSK) are supported by the GSM/EDGE system. To introduce an OSC solution in downlink, four points on the 8PSK constellation (Table 11.10) are selected by a BTS transmitter to form the QPSK constellation. The pair symbols from MUROS are mapped, respectively, onto the I- and Q-channels of the QPSK constellation. The first subchannel (OSC-0) is mapped to MSB and the second subchannel (OSC-1) is mapped to the LSB as shown in Figure 11.29.

Support for legacy GMSK MSs can be provided with a modified QPSK modulation. A parameter $0 \leq \alpha \leq \sqrt{2}$ is chosen to create a quaternary constellation. The constellation is termed an α-QPSK constellation. The extreme values $\alpha = 0$ and $\alpha = \sqrt{2}$ yield BPSK constellations, while for $\alpha = 1$ an ordinary QPSK constellation is obtained. As α changes, the power in the I-channel is changed by $10 \log_{10}(\alpha^2)$ dB, relative to the power of the I-channel when using ordinary QPSK. Similarly, the power in the Q-branch is changed by $10 \log_{10}(2 - \alpha^2)$ dB relative to the power of the Q-branch for ordinary QPSK. The cross power ratio χ, between the I- and Q-branches, is determined through α as:

$$\chi = 10 \log_{10}\left(\frac{\alpha^2}{2-\alpha^2}\right)$$

Table 11.10 QPSK symbol mapping on 8PSK constellation

Original Gray mapped 8PSK modulating bits d_{3i}, d_{3i+1}, d_{3i+2}	Mapping of bits for orthogonal subchannels to 8PSK symbols OSC_0, OSC_1	Symbol parameter l for rule $s_i = e^{j2\pi l/8}$
(1,1,1)	-	0
(0,1,1)	(1,1)	1
(0,1,0)	-	2
(0,0,0)	(0,1)	3
(0,0,1)	-	4
(1,0,1)	(0,0)	5
(1,0,0)	-	6
(1,1,0)	(1,0)	7

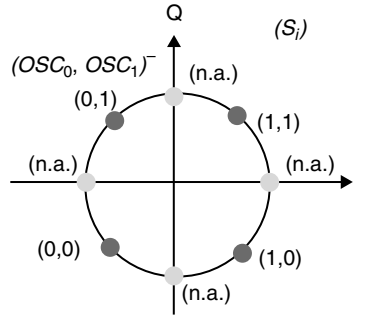


Figure 11.29 QPSK mapping on 8PSK constellation

It is expected that legacy GMSK mobiles will be able to demodulate one of the subchannels, provided α is chosen such that $|\chi|$ is large enough. Note that the energy in an α-QPSK constellation is always 1, independent of α. To determine the symbol constellation, that is α, the modulator may receive feedback from the MSs. For example, α may depend upon the reported RXQUAL, or upon the capabilities of the MSs, for example, legacy/legacy SAIC/α-QPSK aware. When an additional rotation is introduced for the α-QPSK modulation, the MS needs to detect it blindly.

A symbol rotation of $\pi/2$ can provide subchannels to imitate GMSK. Alternatively, the QPSK signal constellation can be designed so that it appears the same as a legacy GMSK modulated symbol sequence on at least one subchannel, for example, it is legacy compliant. The symbol rotation of $\pi/2$ used in downlink allows multiplexing with legacy handsets and alos enables GMSK to be used in the case of DTX and FACCH/SACCH signaling. Different TX pulse shapes may be used in uplink, as proposed for downlink. Whilst the re-use of the GMSK pulse shape is proposed for the initial OSC concept, investigations on an optimized TX pulse in uplink are FFS.

Power Control in Co-TCH MUROS Operation: In the MUROS mode the power given to each user is based on their need, provided that the power difference is within a suitable range (that is, 10 dB) to

provide sufficient signal quality for reliable reception by each mobile. From the path loss of each of the two co-TCH mobiles, the required power level $P1$ for user 1 and $P2$ for user 2 are derived. Both $P1$ and $P2$ are linear quantities. Using $P1$ and $P2$, obtain the I-Q amplitude ratio of the two users as follows:

$$R = \sqrt{P2/P1}(P1 > 0,\ P2 > 0)$$

Determine the digital gains for each of the two co-TCH mobiles: for user 1, $G1 = \cos(\alpha)$, and for user 2, $G2 = \sin(\alpha)$, where $\alpha = \text{arc tan}(R)$ and $\alpha \in [0, \pi/2]$.

MS Receiver: Generally, single antenna interference cancellation (SAIC) is used in the receiver. As discussed earlier, the legacy MS should support DARP and will cancel the interference using SAIC algorithm without any change to the existing system. However, the VAMOS supported receiver requires some changes in the receiver architecture to support QPSK de-modulation and new TSCs. Assume that t_k and c_k are two TSCs of the two MUROS subchannels. The two sequences are in-phase with the corresponding binary subchannel symbols. The discrete signal of the TSC part at the receiver side is:

$$r_k = \sum_l h_l t_{k-l} + j \sum_l h_l c_{k-1} + n_k$$

Correlation is the general operation at the receiver side for channel estimation. The correlation of received r_k with TSC t_k yields:

$$\sum_m t_m^* r_{k+m} = \sum_m \left| t_m \right|^2 h_k + \sum_{l \neq 0} h_{k-1} \sum_m t_m^* t_{m+l} + j \sum_l h_{k-1} \sum_m t_m^* c_{m+l} + \sum_m t_m^* n_{k+m}$$

The first item on the right-hand side is the desired channel response. The second item relates the autocorrelation of TSC t_k with delay. The third item of is related to the cross-correlation between the two subchannel TSCs. Usually good TSC will ensure the second and third items to be zero.

The VAMOS receiver chain co-exists with the existing DARP receiver chain. In the VAMOS receiver chain, first the received samples are de-rotated by $-\pi/2$ radian per symbol, which is identical with the GMSK signal complex input signal rotation. Then the TSC part is passed through the α-QPSK detector, and VAMOS constellation points for calculation of the complex valued α-QPSK constellation points that correspond to the VAMOS training sequence symbols for a particular α-value. Each training sequence symbol is identified by a bit pair $b_1 b_0$ where b_1 is the bit from the desired user training sequence and b_0 is the bit from the orthogonal user training sequence (both bits $\in \{0, 1\}$). This module should be made general, so that both the desired user training sequence and the orthogonal user training sequence can be any of the legacy GMSK training sequences and any of the new training sequences for VAMOS. Next, burst timing (using maximum energy) is performed and enhanced channel estimation is done to find out the channel taps. The rotated and frequency corrected whole burst and the channel taps are passed to the equalizer unit for equalization. The softbits of user 1 and user 2 are generated and then the noise variance scaled user 1 softbits are passed for decoding.

MS Transmitter: Modulation (that is, GMSK) and burst structure (normal burst) are the same as for legacy traffic channels. Each MS uses GMSK modulation for burst transmission. Inside the normal burst,

the legacy MS uses the old TSC, whereas the VAMOS supported MS uses the new TSC, which is assigned by the network during the RRC signaling.

BTS Receiver: For compatibility reason, UL OSC allows MS to use a normal GMSK transmitter with good cross-correlation properties. IC technology is necessary at a BTS receiver to detect signals of the MUROS pair simultaneously (Figure 11.30). It is assumed that BTS uses, for example, an STIRC or SIC receiver to receive orthogonal subchannels used by different MSs. A BTS receiver may use, for example, successive interference cancellation (SIC) or joint detection (JD) to receive signals from two mobiles on simultaneous subchannels with individual propagation paths. Thus, the uplink scheme can be seen as a 2×2 multi-user MIMO, where different propagation paths from two users provide the basis to fully utilize the degree of freedom of two receive antennas in a typical BTS.

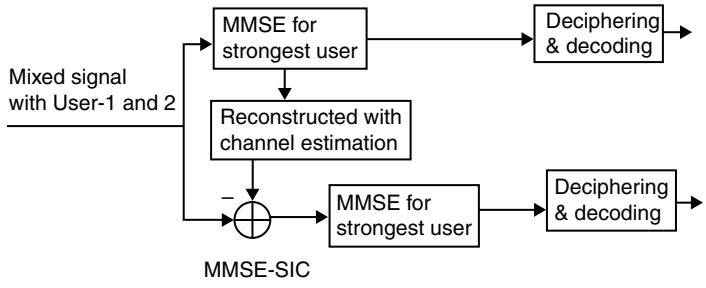


Figure 11.30 BTS receiver processing

SIC is a process of multi-user detection (MUD), when the SIC receiver detects signals of several users, it will demodulate them in decreasing order of signal power. The strongest signal is detected and demodulated first, and then removed from the mixed signals, then it is the second strongest signal, and so on. The detection algorithms for SIC receiver are zero forcing (ZF), minimum mean square error (MMSE), and so on. The MMSE method could be adopted by the BTS receiver with two or more receive antennas. Let $\boldsymbol{x}$ be the transmitting signal, and $\hat{\boldsymbol{x}}$ be the estimated signal, and $\boldsymbol{r}$ be the received signal, the criterion of the MMSE algorithm is illustrated as follows:

$$\min(\varepsilon^2) = \min\{E[\|x - \hat{x}\|^2]\} = \min\{E[\|x - W \cdot r\|^2]\}$$

where ε^2 is the mean square value and $\boldsymbol{W}$ is the weighted matrix.

MUROS should be considered as the potential feature for future GERAN evolution capacity improvement, due to its ability to double the capacity of GSM/EDGE networks without degrading the speech quality very much. In particular, a voice quality better than for GSM HR should be ensured.

Further Reading

3GPP TS 45.002. *GPRS Logical Channels and Mapping to Physical Channels.* ETSI TC-SMG, Sophia-Antipolis Cedex.

3GPP TS 45.003. *Specifies the Channel-Coding Rules for Different Logical Channels.* ETSI TC-SMG, Sophia-Antipolis Cedex.

3GPP TS 45.004. *8-PSK Modulation for EDGE*. ETSI TC-SMG, Sophia-Antipolis Cedex.
3GPP TS 45.008. *Radio Sub-System Link Control*. ETSI TC-SMG, Sophia-Antipolis Cedex.
3GPP TS 44.060. *RLC/MAC layer and Signaling Pprocedures*. ETSI TC-SMG, Sophia-Antipolis Cedex.
3GPP TS 43.064. *GPRS/EDGE air Interface and RLC/MAC Procedures*. ETSI TC-SMG, Sophia-Antipolis Cedex.
Grant, S.J. and Cavers, J.K. (2008) Performance enhancement through joint detection of cochannel signals using diversity arrays. *IEEE Trans. Commun.* **46**, 1038–1049.
Klang, G. and Ottersten, B. (2002) Space-time interference rejection cancellation in transmit diversity systems. Paper presented at The 5th International Symposium on Wireless Personal Multimedia Communications, **2**, 706–710.

12

UMTS System (3G) Overview

12.1 Introduction

Second generation (2G) mobile communication systems have several limitations including network capacity and data rate. To satisfy the increasing demand for higher data rate, tighter data security, larger network capacity and support of various multimedia applications, the International Telecommunication Union (ITU) has defined a set of requirements, which specify what is needed from the next generation (3G) mobile systems. Emphasis was given to the following points: (a) high data rate, greater than 2 Mbps, (b) simultaneous support of voice and data capability, (c) high speech quality, (d) channel switching and packet switching transfer, (e) symmetrical and asymmetrical data transfer (IP-services), (f) low round-trip packet delay (below 200 ms), (g) seamless mobility for voice as well as for packet data applications, (h) high spectrum efficiency, and (i) inter-working with the existing networks (GSM/GPRS). To satisfy these requirements, 15 different proposals came up worldwide, out of these ten were related to the terrestrial segment; the other five were for satellite systems. These system proposals were tested and evaluated by the ITU and finally six different systems were incorporated into the International Mobile Telecommunications at 2000 MHz (IMT-2000) family.

For the terrestrial segment, these proposals can be divided roughly into four categories.

1. **W-CDMA Systems:** These include the frequency division duplex (FDD) components of the UMTS standard in Europe and Japan, along with the CDMA2000 in the USA.
2. **TD-CDMA Systems:** This group contains the time division duplex (TDD) components of UMTS and the Chinese TD-SCDMA, which has now also been integrated into the UMTS-TDD mode.
3. **TDMA Systems:** As a further development of IS-136 and GSM, the UWC-136 system has been incorporated into the IMT-2000 family. EDGE is also a member of IMT-2000 family through the UWC-136 path and uses the GSM frequency spectrum.
4. **FDTDMA Systems:** The further development of the European cordless telephone standard digital enhanced cordless telecommunications (DECT) has also been adopted for applications with low mobility.

There have been several competing proposals for a global 3G standard from a large number of organizations, each pursuing their own goals and interests, and who are involved in the ongoing standardization of the 3G system. Consequently, the standardization job is rather complex and can sometimes also be influenced by political processes, in which technical decisions are made in an environment full of varied and contradictory interests.

Mobile Handset Design Sajal Kumar Das

Based on the parameters defined by the ITU, the Third Generation Partnership Project (3GPP) is currently standardizing the Universal Mobile Telecommunication Systems (UMTS) system. Rel'99 was the last release specified by ETSI SMG in summer 2000, after that it was moved to 3GPP. The Third Generation Partnership Project 2 (3GPP2) has taken over similar tasks for the Code Division Multiple Access 2000 (CDMA2000) evolution. Direct members of the 3GPP include the standardization bodies of the different regions, such as the European Telecommunications Standards Institute (ETSI) (Europe), the Association of Radio Industries and Businesses (ARIB) (Japan), Tl (USA), the Telecommunication Technology Association (TTA) (Korea), the Chinese Wireless Telecommunications Standards (CWTS) (China), and the Telecommunication Technology Committee (TTC) (Japan).

UMTS is European concept and is a part of the International Telecommunication Union's "IMT-2000" vision of a global family of third generation mobile communication. UMTS has the support of many major telecommunications operators and manufacturers, as it represents a unique opportunity to create a mass market for highly personalized and user-friendly mobile access to the information society. The aim is to bring mobile networks significantly closer to the capabilities of fixed networks, providing mobile users with full interactive multimedia capabilities at data rates up to 2 Mbps (in-door environment), in conventional voice, fax, and data services. Hence, the improvements in coding and data compression technology will provide better speech quality and more reliable data transmission with improved quality of service.

UMTS was conceived as a global system, comprising of both terrestrial and global satellite components. Today, GSM is very popular and is used by over 1 billion customers worldwide. Thus, it was realized that although 3G/UMTS offers much more bandwidth and a wide range of complex applications with a more secure environment,it still can not wipe out the 2G systems from the market overnight. This is why both GSM and 3G/UMTS mobile technologies will co-exit and that there are demands for a multimode environment and multimode handset, which can be switched either to GSM or UMTS mode dynamically. Similarly, from the network point of view, operators have spent plenty of money and deployed the GSM network, so they want the GSM network to co-exist with the new network, so that they can do business using this old network also. This is why the UMTS network architecture is built on top of the existing GSM network. The GSM multimode terminals that are able to operate via 2G systems will also further extend the reach of many UMTS services. UMTS will also allow a subscriber to roam among different networks, from a 3G network to a 2G network and then to a satellite one, with seamless transfer.

12.2 Evolution of the 3G Network

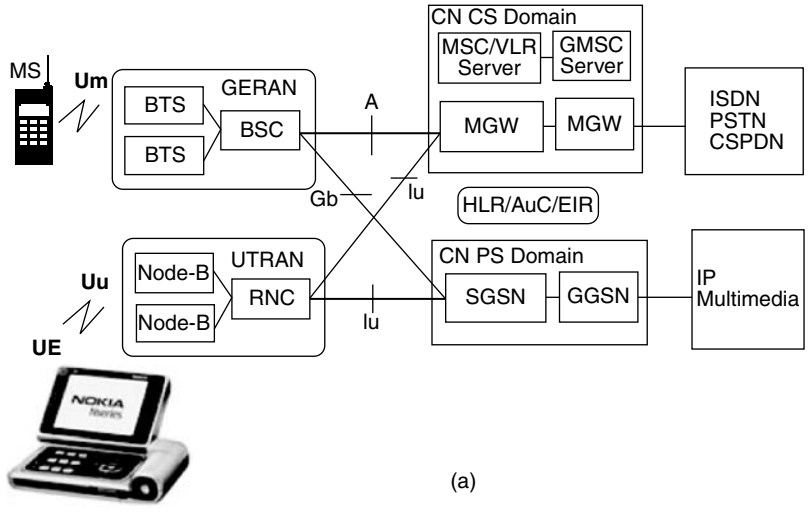
The 3G frequency licenses have already been assigned in most of the European countries and many operators are currently involved in building the networks. In some countries, the licenses were issued by means of an auction. As with GSM, UMTS standardization and deployment is also taking place in several stages. For a first stage, Release 99 (R99) is based on the first systems deployed in Japan and Europe, which does not include the availability of all options eventually planned for UMTS.

3G introduces a new radio access network – UMTS Radio Access Network (UTRAN). UTRAN mainly uses WCDMA as the radio access technology. In early UMTS networks, the UTRAN is attached to the existing GSM core network via a new interface, called Iu. Some new components at the transmission–-reception front-end side of the network are added into the existing GSM networks, which are discussed in the next section.

To reach global acceptance, 3GPP is introducing UMTS in several phases; 3GPP standards are structured as Releases.

3GPP Release 99 – For this phase most of the specifications were frozen in March of 2000. It laid the foundations for high-speed traffic transfer in both circuit switched and packet switched modes by defining enhancements and transitions for existing GSM networks and specifying the development of new radio access network.

3GPP Release 4 – Most of the core technical specifications were frozen in March 2001. It is a minor release with the evolutions including UTRAN access with QoS enhancement, CS domain evolution along with introducing the MSC server and MGWs (Media Gateway) based on IP protocols, enhancements in LCS, MMS, MExE, and so on. With these modifications, the basic architecture of a 3G network (3GPP R4) is as shown in Figure 12.1a.



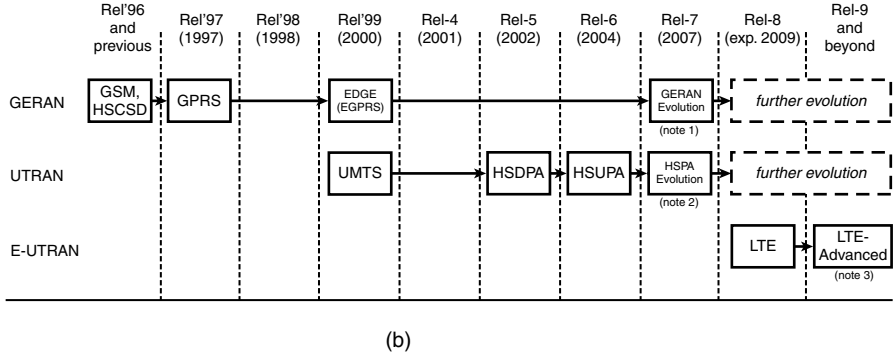


Figure 12.1 (a) 3G network (Release 4). (b) Evolution of 3GPP radio access networks

3GPP Release 5 – Most of the specifications and technical reports for this phase were frozen in June 2002. This was a major release aiming to utilize IP networking as much as possible. IP and the overlying protocols is used in both networks control and user data flows, that is, they implement an "All IP" network, but the IP-based network should still support circuit switched networks. The main features of this release include the introduction of IMS, enhancement in WCDMA, MMS, and LCS. In 3GPP R4/R5, GSM/EDGE radio access network (GERAN) is specified as an alternative for radio access to build a UMTS mobile network.

3GPP Release 6 – For this phase most of the specifications were defined by June 2003. In this release, numerous enhancements and improvements in IMS, MBMS, MMS, QoS, and GERAN are specified.

Also, many new services such as digital rights management, speech recognition and speech enabled services, and priority service are specified. Japan launched the world's first commercial WCDMA network in 2001. Nokia and AT&T Wireless completed the first live 3G EDGE call on 1 November 2001. Telenor launched the first commercial UMTS network in Norway in 1 December 2001. On 20 February 2002, Nokia and Omnital Vodafone made the first rich call in an end-to-end all IP mobile network. In 2002, many of the main UMTS vendors announced their progresses in the battle of pushing their 3G networks and technologies forwards.

Several releases have already appeared and some are also in pipeline, these are tabulated in the Table 12.1.

Table 12.1 Summary of 3GPP release information

Version	Release time	Information
Release 98	1998	This release and the earlier releases specify pre-3G GSM networks.
Release 99	2000 Q1	Specified the first UMTS 3G networks, with WCDMA air interface.
Release 4	2001 Q2	Originally called Release 2000 – added features including an all-IP core network.
Release 5	2002 Q1	Introduced IP multimedia system (IMS) and high-speed downlink packet access (HSDPA).
Release 6	2004 Q4	Inter-operability with wireless LAN networks and adds HSUPA, MBMS, enhancements to IMS such as push-to-talk over cellular (PoC), generic access network (GAN or UMA).
Release 7	2007 Q4	Focuses on decreasing latency, improvements to QoS and real-time applications such as VoIP. This specification will also focus on HSPA + (high-speed packet access evolution), SIM high-speed protocol and contactless front-end interface (near-field communication enabling operators to deliver contactless services such as mobile payments), EDGE Evolution.
Release 8 and onwards	December 2008	Long term evolution (LTE), all-IP network (SAE). Release 8 constitutes a refactoring of UMTS as an entirely IP-based fourth-generation network.
Release 9	In progress (December 2009)	VAMOS, local call local switch, self-organizing network, LTE Improvements, support of WiMAX-LTE and UMTS-WiMAX mobility, support of IMS emergency calls over GPRS and EPS, SAES enhancements.
Release 10 and onwards	In progress	LTE advanced, cooperative base stations, coordinated multi-site beam forming, multi-cell MIMO, cooperative relaying.

Recently, work on the Evolved Universal Terrestrial Radio Access Network (EUTRAN), also known as long term evolution (LTE), has been initiated in 3GPP (Figure 12.1b). The objective of EUTRAN is to develop a framework for the evolution of the 3GPP radio-access technology towards a wider bandwidth, lower latency and packet-optimized radio-access technology with a peak data rate capability of up to 100 Mbps.

12.2.1 Synchronous and Asynchronous Network

The 3G WCDMA radio interface proposals are divided into two groups: synchronous network and asynchronous network. In a synchronous network all base stations are time synchronized to each other. This results in a more efficient radio interface, but requires more expensive hardware in base stations for synchronization. Generally, global positioning systems (GPS) are used in all base stations for this purpose. CDMA2000 is an example of such a network and offers an attractive technology choice, as it can coexist with IS-95 systems. Also, CDMA2000 uses the ANSI-41 as its core network instead of GSM. In the case of an asynchronous network, there is no need to synchronize the networks. An example is the UMTS-FDD network. As CDMA2000 employs a synchronous network, the increased efficiency is attractive to new operators, or to existing GSM operators who are more concerned with deploying an efficient network than attending to the needs of their legacy subscribers. These operators may jump off the GSM track and deploy CDMA2000 instead of upgrading to the UTRAN-FDD mode. The differences between WCDMA and CDMA2000 systems are discussed in Table 12.2.

Table 12.2 Difference between WCDMA and CDMA2000

Parameters	WCDMA	CDMA2000
Frequency band	2 GHz band	It is defined to operate at 450, 700, 800, 900, 1700, 1800, 1900, and 2100 MHz
Bandwidth	1.25/5/10/20 MHz (DSCDMA)	1.25/5/10/20 MHz (DSCDMA), 3.75/5 MHz (MCCDMA)
Chip rate	3.84 Mcps	3.84 Mcps (DSCDMA-FDD)
Data rate	144 kbps (under high mobility environment), 384 kbps (low mobility with a velocity of 30 km/h) and 2 Mbps (indoor or stationary environment)	Depends on the type
Synchronization between base stations	Asynchronous/synchronous	Synchronous
Exchange	GSM-MAP based	ANSI-41 based

As discussed earlier, International Mobile Telecommunication (IMT)-2000 has become the standard. IMT-2000 aims to realize 144 kbps (under high mobility), 384 kbps (low mobility with a velocity of 30 KMPH), and 2 Mbps (under stationary/indoor environments). Thus, on the basis of CDMA technology, three radio-access schemes have been standardized: (1) direct sequence CDMA (DSCDMA)-frequency division duplex (FDD) – this is known as DSCDMA-FDD, (2) DSCDMA-TDD, and (3) multi-carrier CDMA (MCCDMA)-FDD – this is known as MCCDMA-FDD (Figure 12.2).

12.2.2 The UMTS Network Structure

The third-generation mobile communication system, known as the universal mobile telecommunication system (UMTS), is aimed at providing a full range of services to users in the different types of operating environments. As with any other public land mobile system, UMTS consists of three parts: core network (CN), UMTS terrestrial access network (UTRAN), and user equipment (UE).

The external networks can be divided into two groups: (1) CS networks – these provide the circuit switched connections, and (2) PS networks – these provide connections for packet data services.

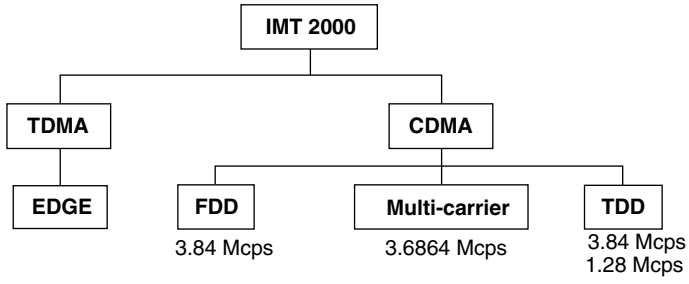


Figure 12.2 Different radio access schemes for 3G

12.3 UTRAN Architecture

In the UMTS network, a radio network system (RNS) is incorporated, which is equivalent to a base station subsystem (BSS = BTS + BSC) in the GSM network. It is mainly responsible for the connection between the UE (user equipment) and the CN (core network). An RNS consists of a radio network controller (RNC) (equivalent to BSC) and one or more logical entities called Node B (equivalent to BTS). Node Bs are connected to the RNC through the Iub interface. Inside the UTRAN, different RNCs of the radio network subsystem can be interconnected through the logical interface Iur. Iur can be conveyed over a physical direct connection between the RNCs or via any suitable transport network. Figure 12.3 shows the UTRAN architecture. In comparison with the parts of the GSM system, here the 3G radio access network parts can be listed as below:

Iu <=> A interface; Iub <=> Abis interface; RNC <=> base station controller; Node B <=> BTS

In UTRAN, a Node B can support FDD mode, TDD mode or dual-mode operation.

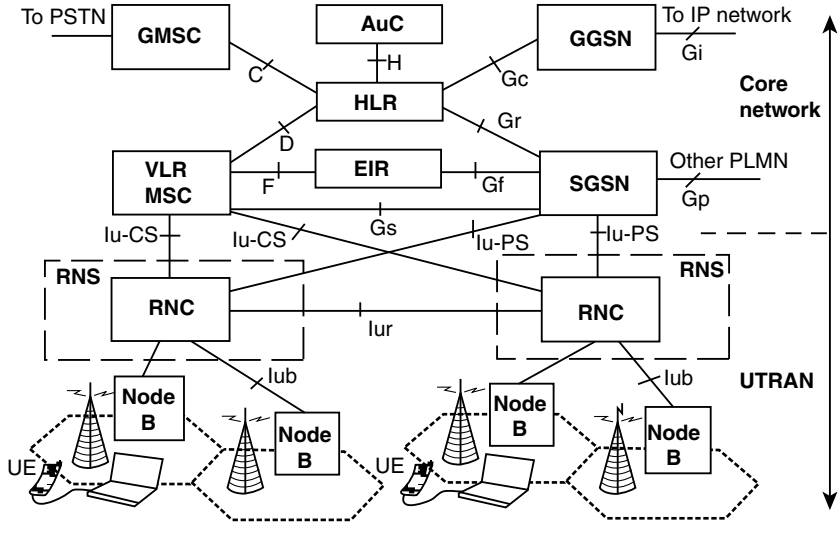


Figure 12.3 UTRAN architecture

The GSM base station subsystem (BSS) and the WCDMA radio access network (RAN) are both connected to the GSM core network to provide a radio connection to the handset. Hence, both technologies can share the same core network (Figure 12.4). The main elements of the GSM/GPRS core networks have already been discussed.

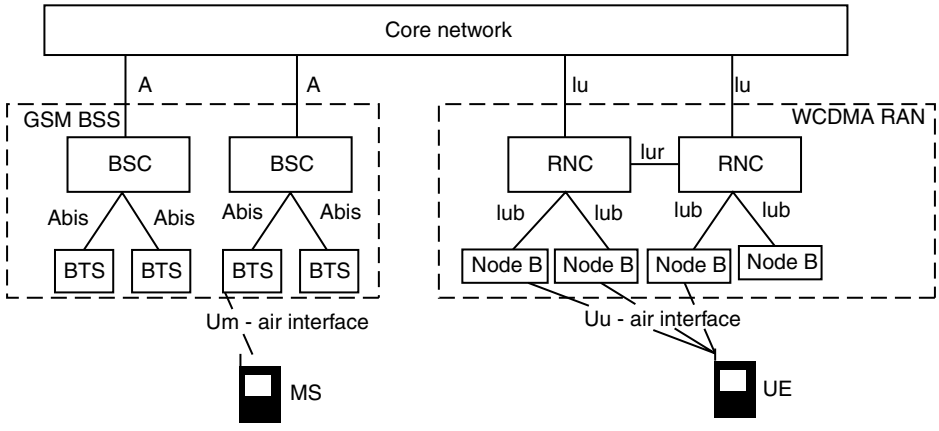


Figure 12.4 RNC and BSC interface with core network in a 3G network

12.3.1 Radio Network Controller (RNC)

The GSM base station controller (BSC) corresponds to the WCDMA radio network controller (RNC). The GSM radio base station (RBS) corresponds to the WCDMA RBS, and the A interface of GSM was the basis of the development of the Iu interface of WCDMA, which mainly differs in the inclusion of the new services offered by WCDMA.

Two RNCs have been defined: serving RNC and drift RNC. The serving RNC has overall control of the handset that is connected to the WCDMA radio access network (Figure 12.5). From the UE

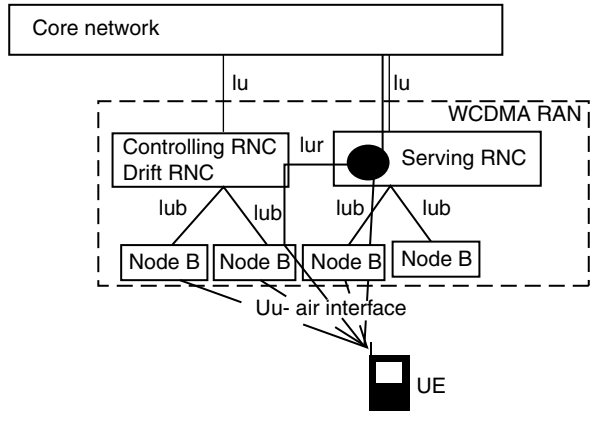


Figure 12.5 WCDMA RAN

(the mobile user equipment) point of view, the serving RNC terminates the mobiles link layer communications. From the CN (core network) point of view, the serving RNC terminates the Iu for this UE. The serving RNC also exerts admission control for new mobiles or services attempting to use the core network over its Iu interface. Drift RNCs, where the physical layer communications of the mobile terminate. One or more drift RNCs communicate with the serving RNC via the Iur interface. Where no soft handover activity is in progress, a drift RNC may also be the serving RNC. The controlling RNC has the overall control of a particular set of cells, and their associated base stations. When a handset must use resources in a cell not controlled by its serving RNC, the serving RNC must ask the controlling RNC for those resources. This request is made via the Iur interface, which connects the RNCs with each other. In this case, the controlling RNC is also said to be a drift RNC for this particular handset. This type of operation is primarily required to be able to provide soft handover throughout the network.

12.3.2 Node B

This unit is meant for radio transmission and reception within a cell. The main tasks of Node B are the transmission and reception of data over air, which includes forward error correction (channel coding), rate adaptation, spreading/de-spreading, modulation, and RF conversion. The Node B is also responsible for softer handover. Depending on sectorization, one or more cells may be served by a Node B. The GSM BTS and Node B can be co-located to reduce the infrastructure and implementation costs. Node B is connected to UE via the Uu radio interface and Iub (ATM) based interface with the RNC.

12.3.3 User Equipment (UE)

The UE consists of two parts. (1) The mobile equipment (ME) is the radio terminal used for radio communication over the Uu interface. (2) The UMTS subscriber identity module (USIM) is a smartcard that holds the subscriber identity, stores the keys to perform authentication and encryption, and some subscription information. The UMTS mobile station can operate in one of the following three modes of operation.

1. **PS/CS Mode of Operation:** The MS is attached to both the PS domain and CS domain, and the MS is capable of simultaneously operating PS services and CS services.
2. **PS Mode of Operation:** The MS is attached to the PS domain only and may only operate services of the PS domain. However, this does not prevent CS-like services to be offered over the PS domain (such as VoIP).
3. **CS Mode of Operation:** The MS is attached to the CS domain only and may only operate services of the CS domain.

12.4 Different Interfaces in the UMTS System

A high level UMTS architecture with different interfaces is shown in Figure 12.6. The openness of the interfaces allows network components from different suppliers to fit in the same network seamlessly. The interfaces between the UE and the UTRAN (Uu interface) and also between the UTRAN and the CN (Iu) are open multi-vendor interfaces.

WCDMA is the air interface used by UMTS terrestrial radio access (UTRA), developed by the third-generation partnership project (3GPP). WCDMA has two modes characterized by the duplex method: FDD (frequency division duplex) and TDD (time division duplex), for operating with paired and unpaired bands, respectively.

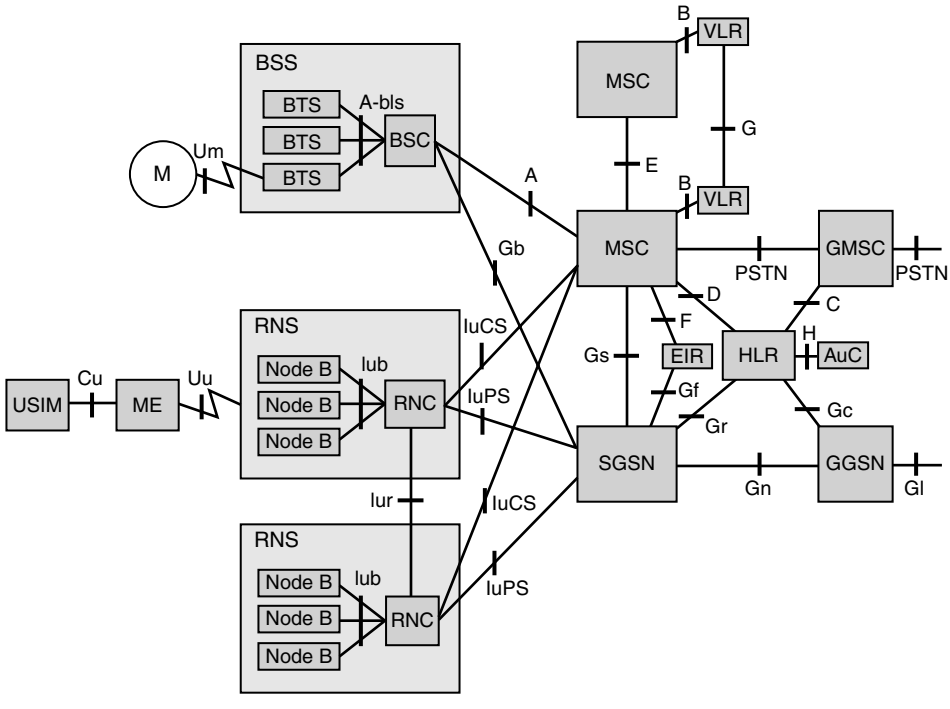


Figure 12.6 Different interfaces in a UMTS network structure (GSM and WCDMA existing together)

12.5 Data Rate Support

The UMTS has the following capabilities to support circuit and packet data at high bit rates.

- **High Mobility:** 144 kbps for rural outdoor mobile use. This data rate is available for environments in which the 3G user is traveling more than 120 km/h in outdoor environments.
- **Full Mobility:** 384 kbps for pedestrian users traveling less than 120 km/h in urban outdoor environments.
- **Limited Mobility:** At least 2 Mbps with low mobility (less than 10 km/h) in stationary indoor and short range outdoor environments. These types of maximum data rates that are often referred to when illustrating the potential for 3G technology will only therefore be available in stationary indoor environments.

12.6 Service Requirement and Frequency Spectrum

UMTS offers teleservices and bearer services, which provide the capability for information transfer between access points. It is possible to negotiate and renegotiate the characteristics of a bearer service at session or connection establishment and during ongoing session or connection. Both connection-oriented and connectionless services are offered for point-to-point and point-to-multipoint communication. Bearer

services have different QoS parameters for maximum transfer delay, delay variation, and bit error rate. This provides following services:

- Speech service in any environment;
- At least 384 kbps with high mobility (mobile speed greater than 120 km/h) in suburban outdoor environment;
- At least 2 Mbps with low mobility (mobile speed less than 10 km/h) in indoor and short range outdoor environments.

Radio Access Bearers – The main service offered by WCDMA RAN is the radio access bearer (RAB). To establish a call connection between the handset and the base station an RAB is needed. Its characteristics are different depending on what type of service/information is to be transported. The RAB carries the subscriber data between the handset and the core network. It consists of one or more radio access bearers between the handset and the serving RNC, and one Iu bearer between the serving RNC and the core network.

3GPP has defined four different quality classes of radio access bearers:

1. conversational (used for example in voice telephony) – low delay, strict ordering;
2. streaming (used for example when watching a video clip) – moderate delay, strict ordering;
3. interactive (used for example in web surfing) – moderate delay;
4. background (used for example in file transfer) – no delay requirement.

Both the conversational and streaming RABs require a certain reservation of resources in the network, and are primarily meant for real-time services. They differ mainly in that the streaming RAB tolerates a higher delay, appropriate for one-way real-time services. The interactive and background RABs are so called "best effort," that is, no resources are reserved and the throughput depends on the load in the cell. The only difference is that the interactive RAB provides a priority mechanism. The RAB is characterized by certain quality of service (QoS) parameters, such as bit rate and delay. The core network will select an RAB with appropriate QoS based on the service request from the subscriber, and ask the RNC to provide such an RAB.

In order to satisfy the service requirement while maintaining a maximum spectral efficiency, a high bandwidth is needed. In addition, two operational modes, frequency division duplex mode (FDD) and time division duplex mode (TDD), should be employed in the wideband CDMA air interface to achieve efficient radio resource utilization.

The FDD mode is intended for applications in macro and micro cell environments. It is more adaptable to symmetrical service. The TDD mode, which is suitable for asymmetrical data service, is used for applications in micro and pico cell environments. The main driving force for asymmetrical service will be the use of the mobile network.

In order to support the FDD and TDD modes in a third-generation mobile communication system, ITU has allocated both paired and unpaired frequency bands (Figure 12.7).

12.7 Cell Structure

To provide a full range of services, the UMTS terrestrial radio access network (UTRAN) normally employ a hierarchical cell structure. Under this structure, three different types of cells exist: macro, micro and pico cells. Various types of cells are needed for different requirements, which can be explained through Figure 12.8.

Normally, macro cells guarantee continuous coverage, while micro cells and pico cells are needed for high traffic regions. In addition, macro cells are used for high mobility terminals, and micro/pico cells are used by low mobility and high capacity terminals. The different types of cells overlay and operate upon

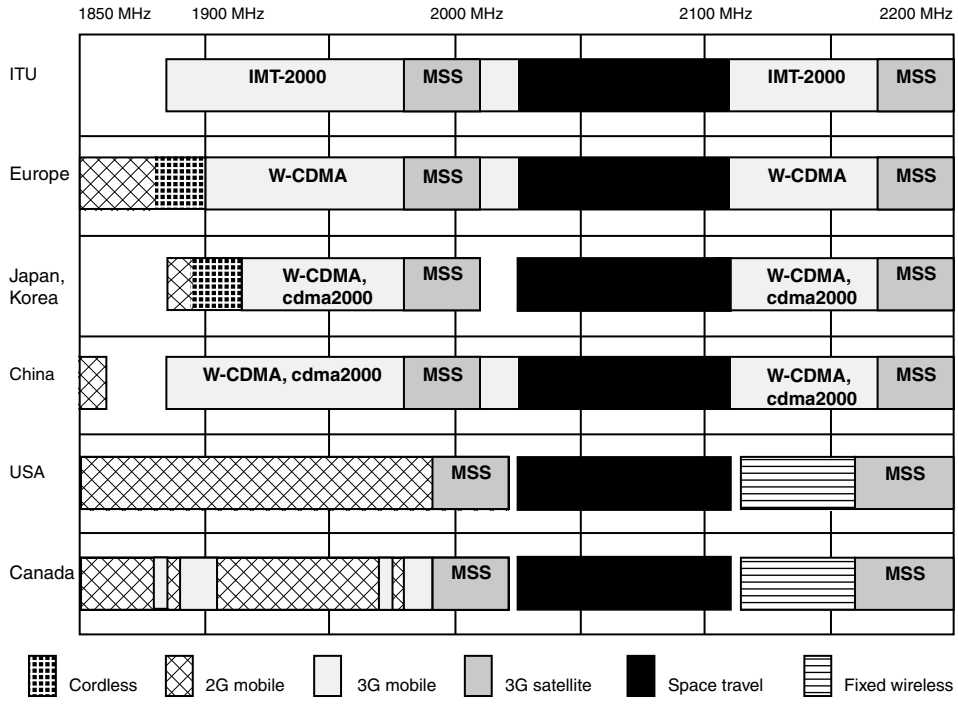


Figure 12.7 Frequency resource for third-generation mobile communication systems

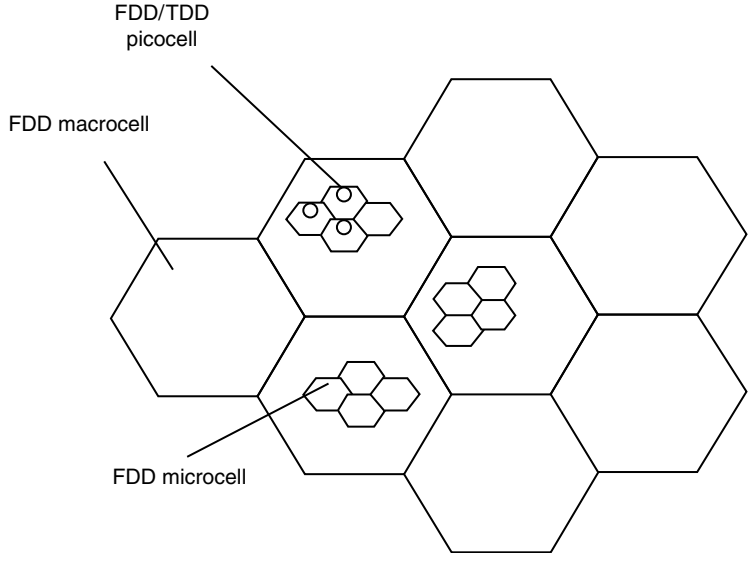


Figure 12.8 Hierarchical cell structure

Table 12.3 Characteristics of different types of cells

Cell type	Radius	Mobility	Max. available data rate
Macro	$\geq$10 km	High	384 kbps
Micro	0.1–1 km	High/low	384 kbps
Pico	$<$100 m	Low	2 Mbps

each other. The characteristics for each cell, cell radius, mobility of user, and maximum available data rate for user, are listed in Table 12.3.

12.8 UTRAN Function Description

As a radio access network, UTRAN should have four major functions: (1) overall system access control, (2) security and privacy; (3) handover; and (4) radio resource management and control.

12.8.1 Overall System Access Control

System access is the means by which a UMTS user is connected to the UMTS in order to use the UMTS services and facilities. User system access may be initiated either from the mobile side or the network side.

12.8.1.1 Admission Control

The purpose of the admission control is to admit or deny new users, new radio access bearers (RAB) or new radio links. The admission control should try to avoid overload situations and base its decisions on interference and resource measurements. The admission control is employed at initial UE access, RAB assignment/reconfiguration, and at handover. The function of admission control is implemented by controlling the RNC based on uplink (UL) interference and downlink (DL) power. The Node B should be able to report UL interference measurements and DL power information over the Iub. The controlling RNC controls this reporting function, that is, if this information needs to be reported and the period of these reports.

12.8.1.2 Congestion Control

The task of congestion control is to monitor, detect, and handle situations when the system is reaching a near overload or an overload situation with the already connected users. This means that some part of the network has run out, or will soon run out, of resources. The congestion control should then bring the system back to a stable state as seamlessly as possible.

12.8.1.3 System Information Broadcasting

This function provides the mobile station with the information that is needed to camp on a cell and to set up a connection in idle mode and to perform a handover or route packets in the communication mode. Because of its close relationship to the basic radio transmission and the radio channel structure, the basic control and synchronization of this function should be located in UTRAN.

12.8.2 Security and Privacy

To secure the user data and signals, UTRAN uses several security produces.

1. UTRAN – should use a temporary identifier such as TMSI in GSM, to replace the permanent CN assigned identity for UE. This function is located in the UE and in the UTRAN.
2. Ciphering and deciphering – this function is a pure computation function whereby the radio transmitted data can be protected against a non-authorized third-party. Both ciphering and deciphering functions are located in the UE and UTRAN, triggered by RNC, implemented at Node B.

12.8.3 Handover

The radio environment survey function performs measurements on radio channels (current and surrounding cells) and translates these measurements into radio channel quality estimates. Measurements may include: received signal strengths (current and surrounding cells), estimated bit error ratios (current and surrounding cells), estimation of propagation environments (for example, high-speed, low-speed, satellite, etc.), transmission range (for example, through timing information), Doppler shift, synchronization status, received interference level, and total DL transmission power per cell.

The handover decision function consists of gathering estimates of the quality of the radio channels (including estimates from surrounding cells) from the measuring entities and to assess the overall quality of service of the call. The overall quality of service is compared with requested limits and with estimates from surrounding cells. Depending on the outcome of this comparison, the macro-diversity control function or the handover control function may be activated.

This handover completion function will free up any resources that are no longer needed. A re-routing of the call may also be triggered in order to optimize the new connection. The function is located both in the UTRAN and in the CN.

12.8.3.1 SRNS Relocation

The SRNS relocation function coordinates the activities when the SRNS role is to be taken by another RNS. SRNS relocation implies that the Iu interface connection point is moved to the new RNS. This function is located in RNC and the CN.

12.8.4 Radio Resource Management and Control

Radio resource management is concerned with the allocation and maintenance of radio communication resources. UMTS radio resources must be shared between circuit transfer mode services and packet transfer modes services (that is, connection-oriented and/or connectionless-oriented services).

12.8.4.1 Radio Bearer Control

This function is responsible for the control of connection element setup and release in the radio access sub-network. The purpose is to participate in the processing of the end-to-end connection setup and release, and to manage and maintain the element of the end-to-end connection, which is located in the radio access sub-network. In the former case, this function will be activated by request from other functional entities at call setup/release. In the latter case, that is, when the end-to-end connection has already been established, this function may also be invoked to cater for in-call service modification or at handover execution. The radio bearer control function interacts with the reservation and release of physical (radio) channels function. It is located both in the UE and in the RNC.

12.8.4.2 Reservation and Release of Physical Radio Channels

This function consists of translating the connection element setup or release requests into physical radio channel requests, reserving or releasing the corresponding physical radio channels, and acknowledging this reservation/release to the requesting entity. It may also perform physical channel reservation and release in the case of a handover. Moreover, the amount of radio resource required may change during a call, due to service requests from the user or macro-diversity requests. Therefore, this function must also be capable of dynamically assigning physical channels during a call. It is located in the RNC and Node B.

12.8.4.3 Allocation and De-allocation of Physical Radio Channels

This function is responsible, once the physical radio channels have been reserved, for actual physical radio channel usage, and allocating or de-allocating the corresponding physical radio channels for data transfer. Acknowledging this allocation/de-allocation to the requesting entity is the subject of further study. This function is located in the RNC and Node B.

12.8.4.4 Allocation of Downlink Channelization Codes

Allocation of downlink channelization codes of cells belonging to Node B is performed by the controlling RNC.

12.8.4.5 Packet Data Transfer Over Radio Function

This function provides packet data transfer capability across the UMTS radio interface, and includes procedures which: provide packet access control over radio channels, provide packet multiplexing over common physical radio channels, provide packet discrimination within the mobile terminal, provide error detection and correction, provide flow control procedures. It is located in both the UE and in the UTRAN.

12.8.4.6 RF Power Control

This group of functions controls the level of the transmitted power in order to minimize interference and keep the quality of the connections. It consists of the following functions: UL outer loop power control, DL outer loop power control, UL inner loop power control, DL inner loop power control, UL open loop power control, and DL open loop power control.

UL Outer Loop Power Control – The UL outer loop power control located in the RNC sets the target quality value for the UL inner loop power control located in Node B. It receives input from quality estimates of the transport channel. The UL outer loop power control is mainly used for a long-term quality control of the radio channel.

DL Outer Loop Power Control – The DL outer loop power control sets the target quality value for the DL inner loop power control. It receives input from quality estimates of the transport channel, which is measured in the UE. The DL outer loop power control is mainly used for a long-term quality control of the radio channel. This function is located mainly in the UE, but some control parameters are set by the UTRAN. The RNC, regularly (or under some algorithms), sends the target downlink power range based on the measurement report from the UE.

UL Inner Loop Power Control – The UL inner loop power control sets the power of the uplink dedicated physical channels. It receives the quality target from the UL outer loop power control and quality estimates of the uplink dedicated physical control channel. The power control commands are sent on the downlink dedicated physical control channel to the UE. This function is located in both the UTRAN and the UE.

DL Inner Loop Power Control – The DL inner loop power control sets the power of the downlink dedicated physical channels. It receives the quality target from the DL outer loop power control and quality estimates of the downlink dedicated physical control channel. The power control commands are sent on the uplink dedicated physical control channel to the UTRAN. This function is located in both the UTRAN and the UE.

UL Open Loop Power Control – The UL open loop power control sets the initial power of the UE, that is, at random access. The function uses UE measurements and broadcasted cell/system parameters as input. This function is located in both the UTRAN and the UE.

DL Open Loop Power Control – The DL open loop power control sets the initial power of the downlink channels. It receives downlink measurement reports from the UE. This function is located in both the UTRAN and the UE.

12.8.4.7 Radio Channel Coding

This function introduces redundancy into the source data flow, increasing its rate by adding information calculated from the source data, in order to allow the detection or correction of signal errors introduced by the transmission medium. The channel coding algorithm(s) used and the amount of redundancy introduced may be different for the different types of logical channels and different types of data. This function is located in both the UE and in Node B.

12.8.4.8 Radio Channel Decoding

This function tries to reconstruct the source information using the redundancy added by the channel coding function to detect or correct possible errors in the received data flow. The channel decoding function may also employ a priori error likelihood information generated by the demodulation function to increase the efficiency of the decoding operation. The channel decoding function is the complement function to the channel coding function. This function is located in both the UE and in Node B.

12.8.4.9 Channel Coding Control

This function generates control information required by the channel coding/decoding execution functions. This may include a channel coding scheme, code rate, and so on. It is located in both the UE and in Node B.

12.8.4.10 Initial (Random) Access Detection and Handling

This function will have the ability to detect an initial access attempt from a mobile station and will respond appropriately. The handling of the initial access may include procedures for a possible resolution of colliding attempts, and so on. The successful result will be the request for allocation of appropriate resources for the requesting mobile station. It is located in the UTRAN.

12.9 Function Partition Over Iub

The Iub connects n RNC and a Node B. As with an Abis interface, Iub should be an open interface, which means two things:

1. inter-connection of RNCs and Node Bs from different manufacturers;
2. separation of the Iub interface radio network functionality and transport network functionality to facilitate introduction of future technology.

12.9.1 Iub Interface Function

The Iub interface should implement these following functions: management of Iub transport resources, logical O&M of Node B, Iub link management, cell configuration management, radio network performance measurement, resource event management, common channels management, radio resource management, implementation specific O&M transport, traffic management of common channels, admission control, power management, data transfer, traffic management of dedicated channels, channel allocation/de-allocation, measurement reporting, dedicated transport channel management, traffic management of downlink shared channels, channel allocation/de-allocation, transport channel management, timing and synchronization management.

Further Reading

3GPP TS 23.002. *Network Architecture*. ETSI TC-SMG, Sophia-Antipolis Cedex.
3GPP TS 23.101, *General UMTS Architecture*. ETSI TC-SMG, Sophia-Antipolis Cedex.
3GPP TS 25.301, *Radio Interface Protocol Architecture*. ETSI TC-SMG, Sophia-Antipolis Cedex.
3GPP TS 25.401, *UTRAN Overall Description*. ETSI TC-SMG, Sophia-Antipolis Cedex.
Release information www.3gpp.org

13

UMTS Radio Modem Design: From Speech to Radio Wave

13.1 Introduction

We have already discussed CDMA techniques in detail in Chapter 5. Before reading this chapter, readers should consult Chapter 5 for a better understanding. The present chapter discusses the WCDMA air interface (also referred as UMTS terrestrial radio access (UTRA), which has been developed by the third-generation partnership project – 3GPP, and the associated radio modem design aspects. 3GPP has the goal of harmonizing and standardizing, in detail, the similar types of proposals that have been prepared by ETSI, ARIB, TTC, TTA, and T1.

Direct-sequence code division multiple access (DS-CDMA) is used as the air medium multiple access technique. We have seen earlier that spread spectrum techniques can be used to achieve a greater data rate, but the chip rate needs to be increased, which in turn demands for more bandwidth. This is why in WCDMA the bandwidth is widened further to accommodate more users or to support an increased data rate. WCDMA has two modes characterized by the duplex method: FDD (frequency division duplex) and TDD (time division duplex), for operating with paired and unpaired bands, respectively. In the FDD version of UMTS, a physical channel is defined by its code and carrier frequency, while in TDD it is in terms of its code, carrier frequency, and time slot.

1. **FDD:** In this duplex method the uplink and downlink transmission use two separate radio frequency bands. A pair of frequency bands, which are separated by a specified frequency range, will be assigned for this system.
2. **TDD:** Here uplink and downlink transmissions are carried over the same radio frequency by using synchronized time intervals. In TDD, time slots in a physical channel are divided into transmission and reception parts. Information on uplink and downlink are transmitted reciprocally. UTRA TDD has two options, with chip rates of 3.84 and 1.28 Mcps. In UTRA TDD, there is a TDMA component in the multiple access in addition to DS-CDMA. Thus the multiple access has also often been denoted as TDMA/CDMA due to the added TDMA nature. A physical channel is therefore defined as a code (or number of codes) and additionally in the TDD mode the sequence of time slots completes the definition of a physical channel.

Mobile Handset Design Sajal Kumar Das

13.1.1 FDD System Technical Parameters

Before going to the detailed discussion, the main technical parameters of WCDMA-FDD mode system are summarized here.

1. **Frequency band:** (1920–1980 MHz) and (2110–2170 MHz) – Frequency division duplex for uplink and downlink, respectively
2. **Minimum frequency band required (bandwidth):** ~5 MHz (for uplink and downlink)
3. **Frequency re-use factor:** 1
4. **Carrier spacing:** 4.4–5.2 MHz
5. **Maximum number of (voice) channels on 2 × 5 MHz:** ~196 (spreading factor 256 UL, AMR 7.95 kbps)/~98 (spreading factor 128 UL, AMR 12.2 kbps)
6. **Voice coding:** AMR codecs (4.75–12.2 kHz, GSM EFR = 12.2 kHz) and SID (1.8 kHz)
7. **Channel coding:** Convolutional coding, Turbo code for high data rate
8. **Antenna multiplexing:** Duplexer needed (190 MHz separation), asymmetric connection supported
9. **Tx/Rx isolation:** MS – 55 dB, BS – 80 dB
10. **Receiver type:** Rake
11. **Receiver sensitivity:** Node B − 121 dBm, mobile − 117 dBm at BER of 10^{-3}
12. **Data type:** Packet and circuit switch
13. **Modulation:** QPSK (downlink) and BPSK (uplink)
14. **Pulse shaping:** Root raised cosine, roll-off = 0.22
15. **Chip rate:** 3.84 Mcps
16. **Channel raster:** 200 kHz
17. **Maximum user data rate (physical channel):** ~2.3 Mbps [spreading factor 4, parallel codes (3 DL/6 UL), $^1/_2$ rate coding], but interference limited
18. **Maximum user data rate (offered):** 384 kbps (in mobile environment – outdoor), higher rates (~2 Mbps) in the near future. HSPDA will offer data speeds up to 8–10 Mbps (and 20 Mbps for MIMO systems)
19. **Channel bit rate:** 5.76 Mbps
20. **Frame length:** 10 ms
21. **Number of slots/frame:** 15
22. **Number of chips/slot:** 2560 chips
23. **Handovers:** Soft, softer, (inter-frequency: hard)
24. **Power control period:** Time slot = 1500 Hz rate
25. **Power control step size:** 0.5, 1, 1.5 and 2 dB (variable)
26. **Power control range:** UL 80 dB, DL 30 dB
27. **Mobile peak power:** Power class 1, + 33 dBm (+ 1 dB/−3 dB) = 2W; class 2, + 27 dBm; class 3, + 24 dBm; class 4 + 21 dBm
28. **Number of unique base station identification codes:** 512/frequency
29. **Physical layer spreading factors:** 4 . . . 256 UL, 4 . . . 512 DL.

13.2 Frequency Bands

UTRA/FDD is designed to operate in the following paired frequency bands (Figure 13.1) and with the following T*x*–R*x* frequency separation, as described in Table 13.1.

The chip rate of the system is 3.84 Mcps, for example, 3 840 000 chips will always be transmitted over the air per second. As we know, the data will be multiplied with the chip signal. Now, based on the transmission data rate needed (which varies from channel to channel), the number of data bits per second will be different. The ratio of the number of chips to the number of data bits is defined as spreading factor

Table 13.1 UTRA FDD frequency bands

Operating band	UL frequencies UE transmit, Node B receive (MHz)	DL frequencies UE receive, Node B transmit (MHz)	T*x*-R*x* frequency separation (MHz)
I	1920–1980	2110–2170	190
II	1850–1910	1930–1990	80
III	1710–1785	1805–1880	95
IV	1710–1755	2110–2155	400
V	824–849	869–894	45
VI	830–840	875–885	45

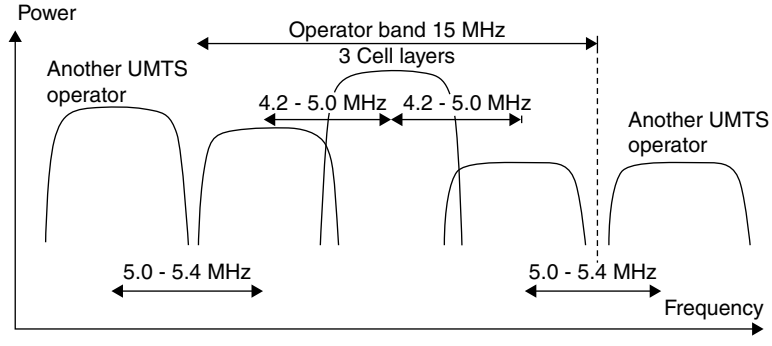


Figure 13.1 Frequency bands allocation

(SF). The spreading factor ranges from 256 to 4 in the uplink and from 512 to 4 in the downlink. Thus, the respective modulation symbol rates vary from 960 to 15 ksymb/s (7.5 ksymb/s) for FDD uplink.

13.3 Radio Link Frame Structure

The data carried by the UMTS/WCDMA transmissions is organized into frames, slots and channels. In this way all the payload data, and the control data, can be carried in an efficient manner. UMTS uses CDMA techniques (WCDMA) as its multiple access technology, but it additionally uses time division techniques with a slot and frame structure to provide the full channel structure.

Physical channels are defined by a specific carrier frequency, scrambling code, channelization code (optional), time start and stop (giving a duration), and on the uplink, relative phase (0 or $\pi/2$). Time durations are defined by start and stop instants, measured in integer multiples of chips. Based on this, it is divided into three sub-categories.

1. **Radio Frame:** A radio frame is a processing duration of 10 ms. The chip rate used here is 3 840 000 chips per second, so the total number of chips in a 10 ms duration (for example, in one radio frame) will be 38 400 chips. This radio frame is again divided into smaller time intervals, which are known as slots. One radio frame consists of 15 slots (Figure 13.2).

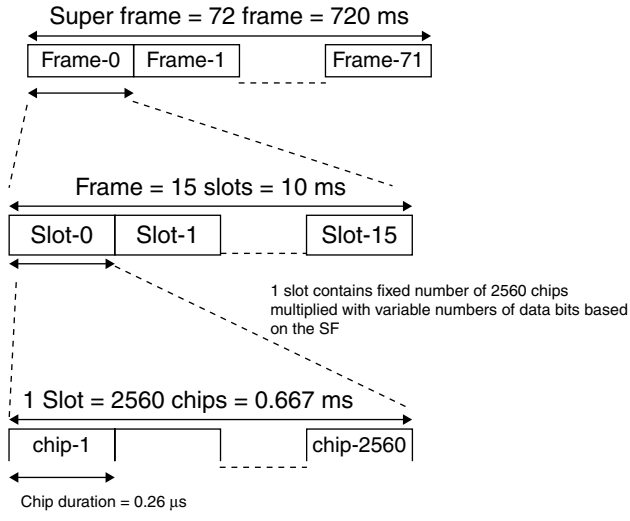


Figure 13.2 Radio link frame structure

2. **Slot:** A slot is a duration of 10 ms/15 = 0.667 ms. In one slot, the total number of chips will be = 38 400/15 = 2560 chips.
3. **Super Frame:** This consists of 72 radio frames and has a duration = 72 × 10 ms = 720 ms.

One symbol consists of a number of chips. The number of chips per symbol is equivalent to the spreading factor of the physical channel. The default time duration for a physical channel is continuous from the instant when it is started to the instant when it is stopped. Physical channels which are not continuous will be discussed in detail later.

Several channels are required in the forward and reverse link for communication between BS and UE. In Figure 13.3, the requirement for different types of channel is indicated.

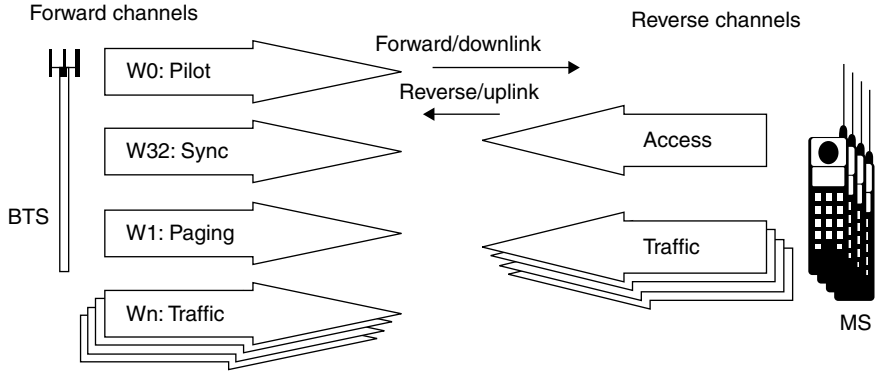


Figure 13.3 Different channel requirement for uplink and downlink

What do we need to separate or distinguish in the system?

- sectors/cells (BTSs)
- mobile terminals (UE)
- directions (for uplink and downlink directions)
- channels

Uplink and downlink are separated by two separate 5 MHz bands. Sectors/cells are separated by primary scrambling code. Mobile terminals are separated by scrambling code. Channels are separated by spreading/channelization code (and scrambling code).

13.4 Channel Structure

The channels are categorized into three different levels, according to the information carried by the channel: (1) logical level channels; (2) transport level channels; and (3) physical level channels. The logical channels are related to what is transported; transport channels are the way in which data are transported, and the physical channels deal with how data are transported and govern the physical characteristics of the signal. Figure 13.4 shows the scope of these channels over the entities in UMTS architecture.

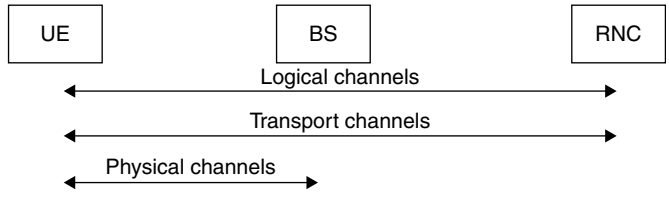


Figure 13.4 Visibility of different channels in the system

13.4.1 Logical Channels

A set of logical channel types are defined for different types of data transfer services as offered by MAC (see Table 13.2). Logical channels are broadly classified into two groups:

- **Control Channels** – for the transfer of control plane information.
- **Traffic Channels** – for the transfer of user plane information.

The air interface protocol architecture is shown in Figure 13.5, and this is discussed in more detail in the next chapter. Introduction of different channels and usage of channels are shown in this architecture.

13.4.2 Transport Channels

Transport channels are channels supplied from the physical layer to the MAC sublayer (L2) or vice versa. A transport channel is defined by how and with what characteristics data are transferred over the air interface. Generally, it is classified into two groups:

1. **Dedicated channels** – using inherent addressing of UE.
2. **Common channels** – using explicit addressing of UE if addressing is needed.

Table 13.2 Logical channels

Broadcast bontrol channel (BCCH)	This is a downlink channel. This channel broadcasts information relevant to the cell to UEs.
Paging control channel (PCCH)	This is a downlink channel, and is used for paging messages and notification of information.
Dedicated control channel (DCCH)	This is an up- and downlink channel. It is used to carry dedicated control information in both directions.
Common control channel (CCCH)	This is an up- and downlink channel, and is for transfer of control information.
Shared control channel (SHCCH)	This channel is bidirectional and only found in the TDD form of WCDMA, used to transport shared control information.
Dedicated traffic channel (DTCH)	Uplink and downlink channel, used to carry user data or traffic.
Common traffic channel (CTCH)	A unidirectional (downlink) channel used to transfer dedicated user information to a group of UEs.

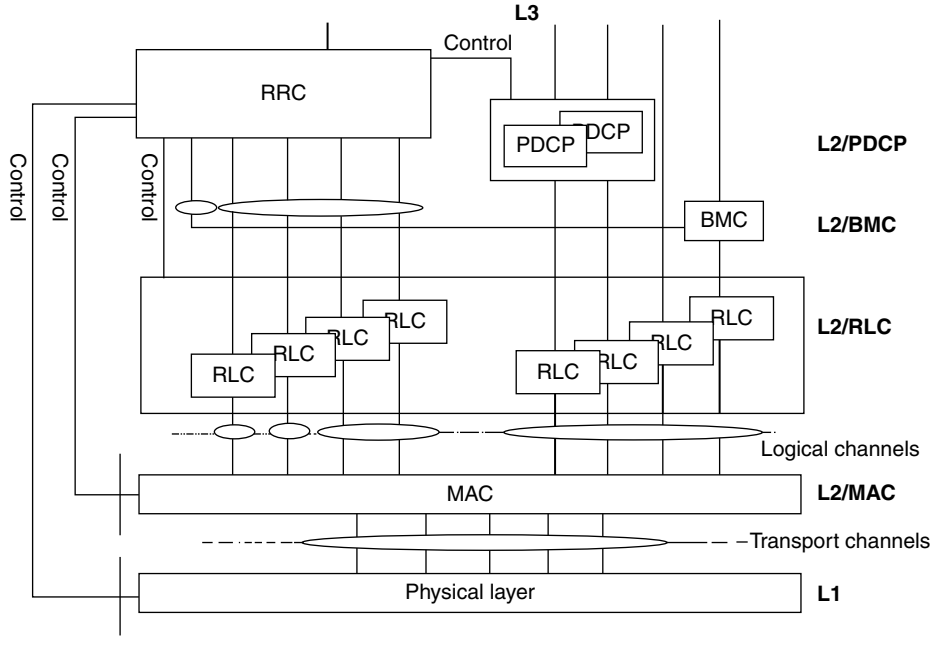


Figure 13.5 Air interface protocol architecture and introduction of different channel hierarchy

The main difference between them is that a common channel is a resource divided between all or a group of users in a cell, whereas a dedicated channel resource, identified by a certain code on a certain frequency, is reserved for a single user only.

13.4.2.1 Dedicated Transport Channels

Only one type of dedicated transport channel exists, the dedicated channel (DCH).

DCH – Dedicated Channel

The dedicated channel (DCH) is a downlink or uplink transport channel. The DCH is transmitted over the entire cell or over only a part of the cell using beam-forming antennas. It is assigned individually to each UE. The dedicated transport channel carries all the information intended for the given user coming from layers above the physical layer, including data for the actual service, as well as higher layer control information. The content of the information carried on the DCH is not visible to the physical layer, thus higher layer control information and user data are treated in the same way. Naturally the physical layer parameters set by UTRAN may vary between control and data. The dedicated transport channel is characterized by features such as fast power control, fast data rate change on a frame-by-frame basis. The dedicated channel supports soft handover.

13.4.2.2 Common Transport Channels

There are six types of common transport channels: BCH, FACH, PCH, RACH, CPCH, and DSCH.

BCH – Broadcast Channel

The broadcast channel (BCH) is a downlink transport channel that is used to broadcast system- and cell-specific information. It is transmitted at a fixed rate and always transmitted over the entire cell and has a single transport format. As the terminal cannot register to the cell without the possibility of decoding the broadcast channel, so this channel is transmitted with relatively high power in order to reach all the users within the intended coverage area.

FACH – Forward Access Channel

The forward access channel (FACH) is a downlink transport channel that carries control information to terminals known to be located in the given cell. The FACH is transmitted over the entire cell or over only a part of the cell using beam-forming antennas. This is used, for example, after a random access message has been received by the base station. It is also possible to transmit packet data on the FACH. There can be more than one FACH in a cell. With more than one FACH, the additional channels can have a higher data rate. This may be shared by multiple UEs. The FACH can be transmitted using slow power control.

PCH – Paging Channel

The paging channel (PCH) is a downlink transport channel used for transmitting paging information, that is, when the network wants to initiate communication with the mobile terminal. The identical paging message can be transmitted in a single cell or in several cells, depending on the system configuration. The terminals must be able to receive the paging information in the whole cell area. The PCH is always transmitted over the entire cell. The transmission of the PCH is associated with the transmission of physical-layer generated paging indicators, to support efficient sleep-mode procedures. The design of the paging channel also affects the terminal's power consumption in the standby mode. The less often the terminal has to tune to listen for a possible paging message, the longer the terminal's battery life.

RACH – Random Access Channel

The random access channel (RACH) is an uplink transport channel used for transmitting control information and user data, such as requests to set up a connection. It can also be used to send small amounts of packet data from the terminal to the network. The RACH is always received from the entire cell, applied in random access, and used for low-rate data transmissions from a higher layer. The RACH is characterized by a collision risk and by being transmitted using open loop power control.

CPCH – Common Packet Channel

The common packet channel (CPCH) is an uplink transport channel. CPCH is associated with a dedicated channel on the downlink which provides power control and CPCH control commands (for example, emergency stop) for the uplink CPCH. The uplink CPCH is an extension to the RACH channel that is intended to carry packet-based user data in the uplink direction. The reciprocal channel providing the data in the downlink direction is the FACH. In the physical layer, the main differences to the RACH are the use of fast power control, a physical layer-based collision detection mechanism, and a CPCH status monitoring procedure. The uplink CPCH transmission may last several frames in contrast with one or two frames for the RACH message. The CPCH is characterized by initial collision risk and by being transmitted using inner loop power control. It is applied in random access and used primarily for high rate bursty data transmissions.

DSCH – Downlink Shared Channel

The downlink shared channel (DSCH) is a downlink transport channel shared by several UEs. The DSCH is associated with one or several downlink DCH. The DSCH is transmitted over the entire cell or over only a part of the cell using beam-forming antennas.

The downlink shared channel (DSCH) is a transport channel intended to carry dedicated user data and/or control information; it can be shared by several users. In many respects it is similar to the forward access channel, although the shared channel supports the use of fast power control as well as variable bit rate on a frame-by-frame basis. The DSCH does not need to be heard in the whole cell area and can employ the different modes of transmit antenna diversity methods that are used with the associated downlink DCH. The downlink shared channel is always associated with a downlink DCH.

The mapping of logical to transport channels is shown in Figure 13.6.

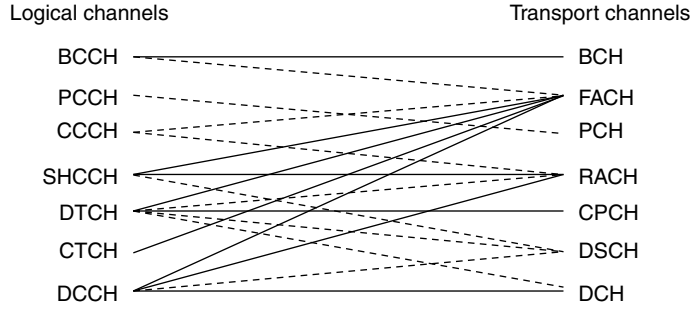


Figure 13.6 Mapping between logical and transport channels

13.4.2.3 Frame Structure of Transport Channels

As discussed earlier, UTRA channels use a 10 ms radio frame structure. The frame structure also employs a longer period, called the system frame period. The system frame number (SFN) is a 12-bit number and is used by procedures that span more than a single frame. Physical layer procedures, such as the paging procedure or random access procedure, are examples of procedures that need a longer period than 10 ms for correct definition.

13.4.2.4 Transport Channels Mapping on to the Physical Channels

Each transport channel is accompanied by the transport format indicator (TFI) and appears at each time event at which data are expected to arrive for the specific transport channel from the higher layers. The physical layer combines the TFI information from different transport channels into the transport format

combination indicator (TFCI). The TFCI is transmitted in the physical control channel to inform the receiver about the transport channels that are active for the current frame; the exception to this is the use of blind transport format detection (BTFD), which will be covered in connection with the downlink dedicated channels. The TFCI is decoded appropriately in the receiver and the resulting TFI is given to higher layers for each of the transport channels that can be active for the connection.

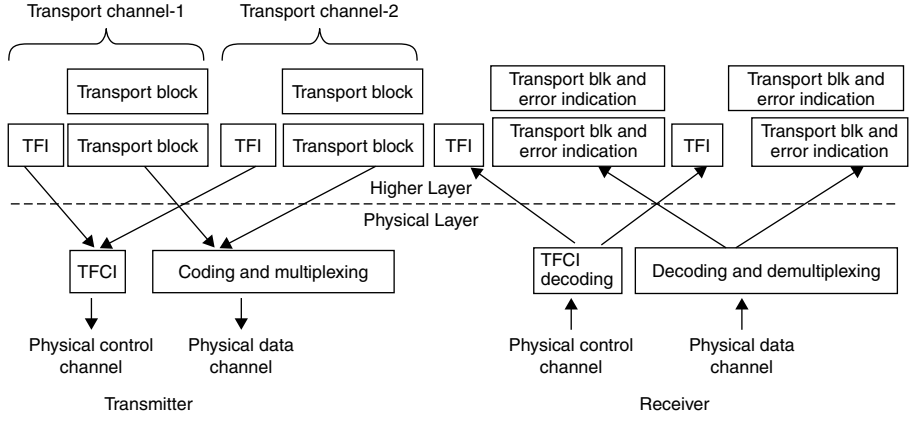


Figure 13.7 The interface between higher layers and the physical layer

As an example, Figure 13.7 shows two transport channels, which are mapped to a single physical channel, and error indication is also provided for each transport block. The transport channels may have a different number of blocks, and at any instant of time, all the transport channels will not necessarily have to be active. One physical control channel and one or more physical data channels form a single coded composite transport channel (CCTrCh). There can be more than one CCTrCh on a given connection but only one physical layer control channel is transmitted in such a situation.

Transport Formats/Configurations

The following parameters are used for transport formats or configurations.

- **Transport block (TB):** Basic unit of data exchanged between L1 and MAC for L1 processing.
- **Transport block size:** Number of bits in a TB.
- **Transport block set (TBS):** A set of TBs exchanged between L1 and MAC at the same time instant using the same transport channel.
- **Transport block set size:** Number of bits in a TBS.
- **Transmission time interval (TTI):** Periodicity at which a TBS is transferred by the physical layer on to the radio interface – {10, 20, 40, 80 ms}. MAC delivers one TBS to the physical layer every TTI.
- **Transport format (TF):** Format offered by L1 to MAC (and vice versa) for the delivery of a TBS during a TTI on a given transport channel (TrCH) – dynamic part (TB size, TBS size), semi-static part (TTI, type/rate of coding size of CRC). TB size, TBS size, TTI define the TrCH bit rate before L1 processing.
- **Transport format set (TFS):** A set of TFs associated with a TrCH. Semi-static part of all TFs in a TFS is the same.
- **Transport format combination (TFC):** Multiple TrCHs each having a TF. Authorized combination of the currently valid TFs that can be submitted to L1 on a CCTrCH, containing one TF from each TrCH.

- **Transport format combination set (TFCS):** This is a set of TFCs on a CCTrCH. Produced by RNC. TFCS is given to MAC by L3 for control. MAC chooses between the different TFCs specified in the TFCS. MAC has control over only the dynamic part of the TFs. Semi-static part relates to QoS (for example, quality) and is controlled by RNC admission control. Bit rate can be changed quickly by MAC with no need for L3 signaling.
- **Transport format indicator (TFI):** A label for a specific TF within a TFS. Used between MAC and L1.
- **Transport format combination indicator (TFCI):** Used to inform the receiving side of the currently valid TFC.

13.4.3 Physical Channels

Physical channels typically consist of a layered structure of radio frames and time slots, although this is not true for all physical channels. Depending on the channel bit rate of the physical channel, the configuration of the slots varies.

The Physical Resource – The basic physical resource is the code/frequency plane. In addition, on the uplink, different information streams may be transmitted on the "I" and "Q" branch. Consequently, a physical channel corresponds to a specific carrier frequency, code, and, on the uplink, the relative phase (0 or $\pi/2$) also.

The uplink and downlink physical channels are shown in Figure 13.8. For uplink and downlink different frequency bands are used (WCDMA-FDD), and inside uplink or downlink, for the various channels different spreading codes are used.

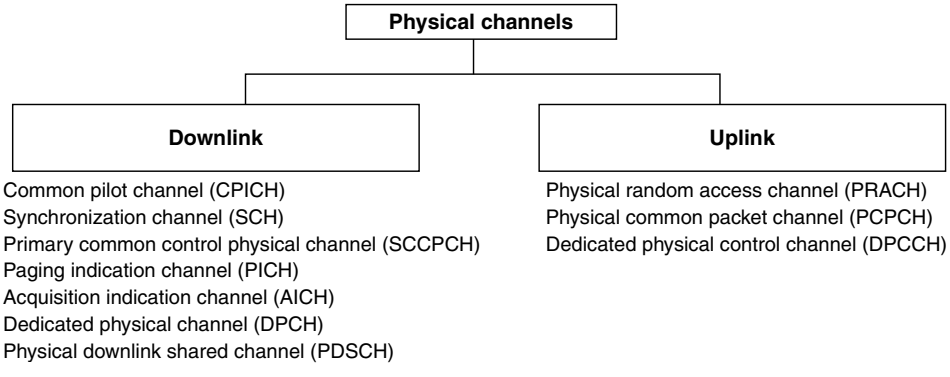


Figure 13.8 Uplink and downlink physical channels

13.4.3.1 Mapping of Transport Channels onto Physical Channels

Figure 13.9 summarizes the mapping of various transport channels onto different physical channels.

The DCHs are coded and multiplexed as will be described later, and the resulting data stream is mapped sequentially (first-in-first mapped) directly to the physical channel(s). The mapping of BCH and FACH/PCH is equally straightforward, where the data stream, after coding and interleaving, is mapped sequentially to the primary and secondary CCPCH, respectively. Also, for the RACH, the coded and interleaved bits are sequentially mapped to the physical channel, in this case the message part of the PRACH. Some channels (such as SCH, CPICH, etc.) are generated at the physical layer itself.

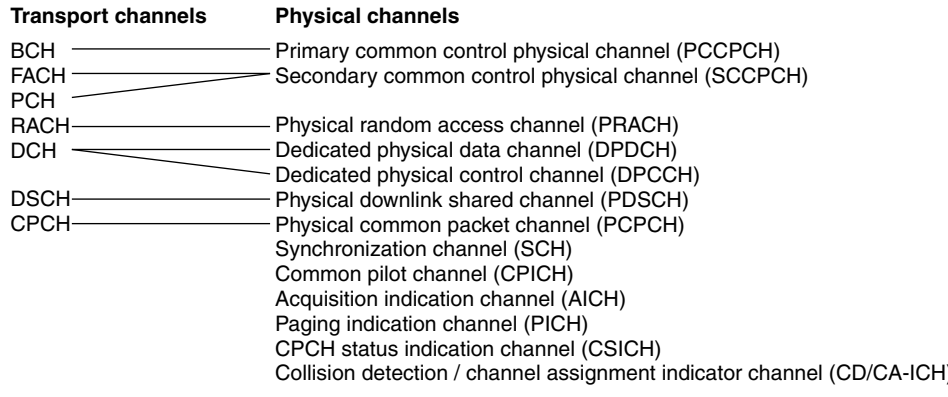


Figure 13.9 Transport-channel to physical-channel mapping

13.5 Spreading, Scrambling, and Modulation

As discussed in Chapter 5, in WCDMA, in addition to spreading, the scrambling operation is performed in the transmitter (Figure 13.10). This is required in order to separate the user terminals and base stations (cells/sectors) from each other. Scrambling is used on top of spreading, and it does not change the rate or signal bandwidth but only makes the signals from different sources separable from each other. With scrambling, it would not matter if the actual spreading were performed with identical codes for several transmitters.

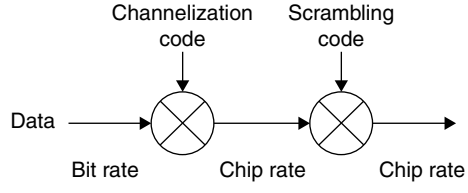


Figure 13.10 Relationship between spreading and scrambling operations

Transmissions from a single source are separated by channelization codes, that is, downlink connections within one sector and the dedicated physical channel in the uplink from one terminal are separated by this. The spreading/channelization codes of UTRA are based on the orthogonal variable spreading factor (OVSF) technique. The use of OVSF codes allows the spreading factor to be changed and orthogonality between different spreading codes of different lengths to be maintained. The codes are picked from the code tree, based on the data rate and channel number requirements (as we know, a lesser spreading factor, for example, more towards the left side of the tree, provides more actual data rate, and more towards the right-hand side of the tree provides more spreading factor value, such as more branches, which can be used for many simultaneous orthogonal channels), which is illustrated in Figure 13.11. Where the connection uses a variable spreading factor, the correct use of the code tree also allows despreading according to the smallest spreading factor. This requires that channelization code should only be used from the branch indicated by the code used for the smallest spreading factor.

There are certain restrictions as to which of the channelization codes can be used for a transmission from a single source. Another physical channel may use a certain code in the tree if no other physical

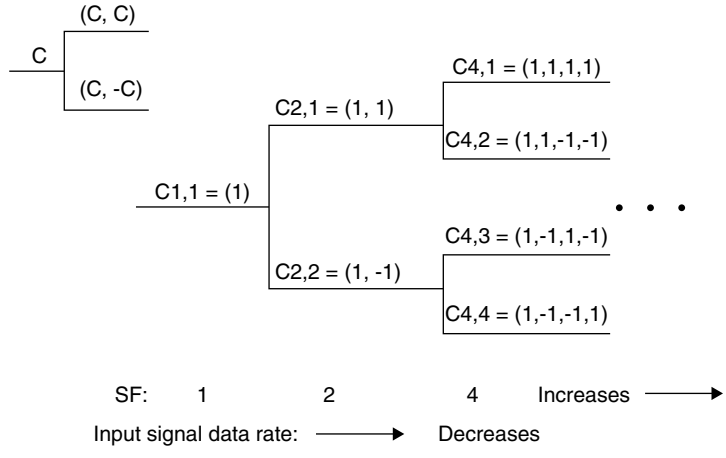


Figure 13.11 Channelization code tree

channel to be transmitted using the same code tree is using a code that is on an underlying branch. The downlink orthogonal codes within each base station are managed by the radio network controller (RNC) in the network.

The functionality and characteristics of the scrambling and channelization codes are summarized in Table 13.3. The definition for the same code tree means that for transmission from a single source, from either a terminal or a base station, one code tree is used with one scrambling code on top of the tree. This means that different terminals and different base stations may operate their code trees totally independent of each other; there is no need to coordinate the code tree resource usage between different base stations or terminals.

13.5.1 Down Link (DL) Spreading and Modulation

The support of DL (Node B side) code generation and knowledge of the DL modulation scheme are mandatory at UE. The code generation and allocation scheme for the downlink direction is described below.

13.5.1.1 Channelization Codes

The DL channelization codes are derived from the channelization code tree. The code tree under a single scrambling code is shared by several users (all channels from the one sector are separated by different spreading codes and sectors are separated by sector specific unique primary scrambling code). As typically only one scrambling code is used per sector so only one code tree is used per sector in the base station. The common channels and dedicated channels share the same code tree resource. There is one exception for the physical channels: the synchronization channel (SCH) is not under a downlink scrambling code. In the downlink, the dedicated channel spreading factor does not vary on a frame-by-frame basis; the data rate variation is taken care of either by a rate matching operation or with discontinuous transmission, where the transmission is off during part of the slot. The channelization code for the P-CPICH is fixed to $C_{\text{ch},256,0}$ and the channelization code for the P-CCPCH is fixed to $C_{\text{ch},256,1}$.

Table 13.3 Functionality of channelization and scrambling code

Code	Channelisation code	Scrambling code	Synchronization code
Property	Orthogonal property helps to reduce interference. Does not have good correlation properties. Needs additional long codes. Walsh code is an example.	Has good correlation properties. These are basically Gold codes. Long scrambling code and short scrambling code.	Good auto-correlation properties
Usage	Uplink – separation of DPDCH and DPCCH from same UE Downlink – separation of DL connections to different users within one cell.	Uplink – separations of UEs Downlink – separation of sectors/cells	Downlink – initial cell search and synchronization
Length	Uplink – 4–256 Downlink- 4–512	Uplink – 10 ms 38 400 chips. Downlink – 10 ms 38 400 chips	256-chip
Number of codes	Number of codes under one scrambling code = spreading factor	Uplink – several millions Downlink – 512	Every cell across the system (regardless of network operator) will transmit the same primary synchronization code. There are 64 sets from which the secondary synchronization channels can be selected.
Code family	OVSF	Long codes – 10 ms Gold codes Short codes: extended S(2) code family	The primary synchronization code is constructed from generalized hierarchical Golay sequence. The secondary synchronization code words requires Hadamard sequence.
Spreading	Increases bandwidth	Does not increase transmission bandwidth	Increases bandwidth but predefined

Spreading and modulation for CPICH, S-CCPCH, P-SCCCH, PDSCH, PICH, and AICH channels are done in an identical way as for the DL DPCH. Spreading/modulation for P-CCPCH is done in exactly the same way as for the DL DPCH, except that the P-CCPCH is time multiplexed after spreading. P-SCH and S-SCH are code multiplexed and transmitted simultaneously during the first 256 chips of each slot. The SCH is *non-orthogonal* to the other DL physical channels.

13.5.1.2 Scrambling Code

The downlink scrambling uses long codes, the same Gold codes as in the uplink. The complex-valued scrambling code is formed from a single code by simply having a delay between the I- and Q-branches (Figure 13.12). The code period is truncated to 10 ms; no short codes are used in the downlink direction. A total of (2^{18}–1) number of scrambling codes, numbered from 0 to 262 142 can be generated, however not all are used. The downlink set of the (primary) scrambling codes is limited to 512 codes; otherwise the cell search procedure (described in Chapter 15, Section 15.5.1) would become excessive. The scrambling codes must be allocated to the sectors during the network planning, as is done for GSM for broadcast frequency planning in cells.

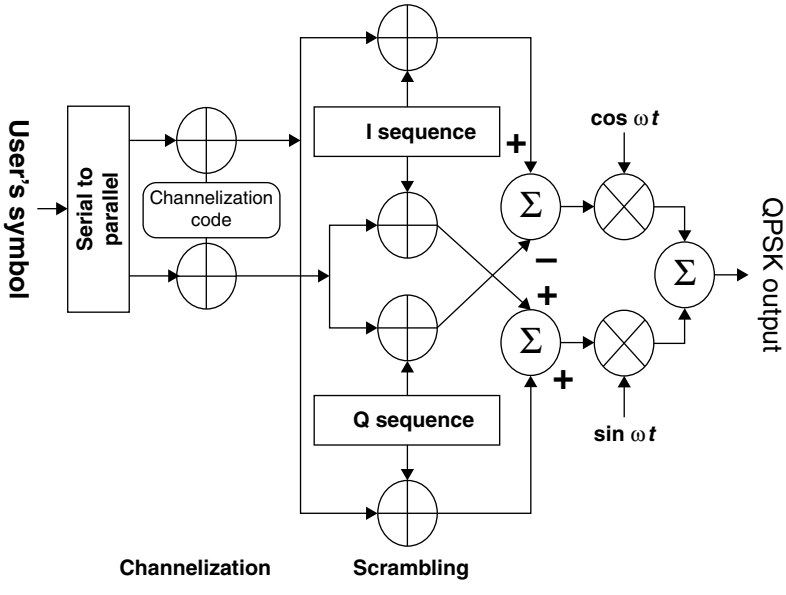


Figure 13.12 Complex scrambling

The downlink scrambling codes are divided into 512 sets. Each set consists of a primary scrambling code and 15 secondary scrambling codes. Overall, 8192 codes are allocated. In a compressed mode, each of these codes are associated with an even alternative scrambling code and an odd alternative scrambling code. The even alternative scrambling code corresponds to $k + 8192$; the odd alternative scrambling code corresponds to $k + 16\,384$, where k corresponds to the code used from the 8192 scrambling code set. The set of primary scrambling codes is further divided into 64 scrambling code groups, each group consisting of 8 primary scrambling codes. The 64 groups have a one-to-one mapping to the sequence of secondary synchronization codes.

Each cell is allocated one, and only one, primary scrambling code and the PCCPCH + PCPICH is always transmitted using this primary scrambling code. The broadcast information is conveyed in the

PCCPCH. The other downlink physical channels are transmitted with either the primary scrambling code or an associated secondary scrambling code. The x sequence is constructed using the polynomial $1 + X^7 + X^{18}$, while the y sequence is constructed using the polynomial $1 + X^5 + X^7 + X^{10} + X^{18}$. The resulting sequences thus constitute segments of a set of Gold sequences. The scrambling codes are repeated for every 10 ms radio frame. A mixture of primary and secondary scrambling codes can be used in a transmission of CCTrCH. The code generator used to generate the downlink scrambling code is different to that of the code generator used to generate the uplink scrambling code. Refer to TS 25.213 Section 13.5.2.2 for details of the scrambling code generator.

The UE will contain a function for generating the downlink scrambling code. There will be multiple generators, one associated with each finger in the rake, to allow the UE to connect to multiple downlink transmissions derived from different cells. By virtue of the synchronization channels, the UE will derive the downlink scrambling code in use by the cell that it intends to camp onto. This code shall be reported to the higher layers. From this point, the system information decoded from the PCCPCH will then inform the UE of the primary scrambling codes for the neighboring/hierarchical cells. If more capacity is needed, then a secondary scrambling code needs to be introduced in the cell, and only those users not fitting under the primary scrambling code should use the secondary code. The biggest loss in orthogonality occurs when the users are shared evenly between two different scrambling codes.

13.5.1.3 Synchronization Codes

For synchronization purposes the network uses the synchronization channel (SCH), which uses synchronization codes (SC). SCH is introduced at the physical layer only and is not visible above the physical layer. SCH channels are treated in a different manner to the normal channels and as a result they are not spread using the OVSFs and PN scrambling codes. Instead they are spread using synchronization codes (SC). SCH are not under the cell specific primary scrambling code. The terminal (UE) must be able to synchronize to the cell before knowing the downlink scrambling code.

The synchronization channel for a cell is organized into a primary synchronization channel and secondary synchronization channel. The primary synchronization channel (or code) is a 256-chip sequence transmitted every slot, and every cell across the system (regardless of network operator) will transmit the same code. From the primary synchronization code, the UE will obtain slot and chip synchronization. The secondary synchronization channel is a sequence of 256-chip codes. A different code is transmitted on every slot, and this repeats over every 15 slots (or 10 ms), for example, periodic over a frame. There are 64 sets from which the secondary synchronization channels can be selected. These 64-sets correspond to the 64 primary scrambling code generators. (Sometimes the same code is used across two or more consecutive slots. Please refer to TS 25.213 for details of the code assignments.)

From the secondary synchronization channel the UE will obtain frame synchronization and be able to identify the primary scrambling code (here sectors are identified by primary scrambling, the same as for GSM, where BTSs are identified by broadcast frequency used by different BTS). Refer to TS 25.213 Section 5.2.3 for details of the synchronization codes used. The primary synchronization code is constructed as a so-called generalized hierarchical Golay sequence. The construction of secondary synchronization code words requires a Hadamard sequence, which can be obtained as the rows in a matrix $\boldsymbol{H}_8$ constructed recursively by:

$$\boldsymbol{H}_0 = (0)$$
$$\boldsymbol{H}_k = \begin{pmatrix} \boldsymbol{H}_{k-1} & \boldsymbol{H}_{k-1} \\ \boldsymbol{H}_{k-1} & \overline{\boldsymbol{H}_{k-1}} \end{pmatrix}, k \geq 1 .$$

There are altogether 16 secondary synchronization codes and 1 primary synchronization code (this primary synchronization code is common to all cells). The 16 secondary synchronization codes are

arranged into 64 sequences (to identify the scrambling code group) such that the cyclic-shifts of these sequences are unique.

The modulating chip rate is 3.84 Mcps and in the downlink direction QPSK modulation is used.

13.5.2 Uplink Spreading and Modulation

13.5.2.1 Uplink Channelization Codes

OVSF channelization codes are used to preserve orthogonality between a user's different physical channels. The OVSF can be generated using the code tree in Figure 13.11. Alternatively, one may also use the following representation for generating OVSF codes.

$$C_{\mathrm{ch},1,0} = 1$$

$$\begin{bmatrix} C_{ch,2,0} \\ C_{ch,2,1} \end{bmatrix} = \begin{bmatrix} C_{ch,1,0} & C_{ch,1,0} \\ C_{ch,1,0} & -C_{ch,1,0} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$\begin{bmatrix} C_{ch,2^{(n+1)},0} \\ C_{ch,2^{(n+1)},1} \\ C_{ch,2^{(n+1)},2} \\ C_{ch,2^{(n+1)},3} \\ \vdots \\ C_{ch,2^{(n+1)},2^{(n+1)}-2} \\ C_{ch,2^{(n+1)},2^{(n+1)}-1} \end{bmatrix} = \begin{bmatrix} C_{ch,2^n,0} & C_{ch,2^n,0} \\ C_{ch,2^n,0} & -C_{ch,2^n,0} \\ C_{ch,2^n,1} & C_{ch,2^n,1} \\ C_{ch,2^n,1} & -C_{ch,2^n,1} \\ \vdots & \vdots \\ C_{ch,2^n,2^n-1} & C_{ch,2^n,2^n-1} \\ C_{ch,2^n,2^n-1} & -C_{ch,2^n,2^n-1} \end{bmatrix}$$

In the downlink direction, the OVSF is used to convey different users and different channels to a single user, whereas in the uplink direction, the OVSF is used to convey different channels from the user.

The UE will generate the channelization codes when given appropriate information by the higher protocol layers, and then it will use the same until that code is no longer required.

In the uplink direction the spreading factor on the DPDCH may vary on a frame-by-frame basis. The spreading codes are always taken from the earlier described code tree. When the channelization code is used for spreading, it is always taken from the same branch of the code tree, the despreading operation can take advantage of the code tree structure and avoid chip level buffering. The terminal provides data rate information, or more precisely the transport format combination indicator (TFCI), on the DPCCH, to allow data detection with a variable spreading factor on the DPDCH.

13.5.2.2 Uplink Scrambling Codes

There are 2^{24} UL scrambling codes. All channels will use either long or short scrambling codes, except for PRACH, where only the long scrambling code is used. The UE will be able to generate long and short scrambling codes as dictated by the higher layers. Two separate scrambling code generators will be required to allow uplink communications with two cells (Figure 13.13). If communications is only with one cell, then the other scrambling code generator shall be switched off. The UE will initiate the uplink-scrambling generators with a 24-bit value that is provided by the higher layers. Note that two sets of scrambling code generator need to be initiated.

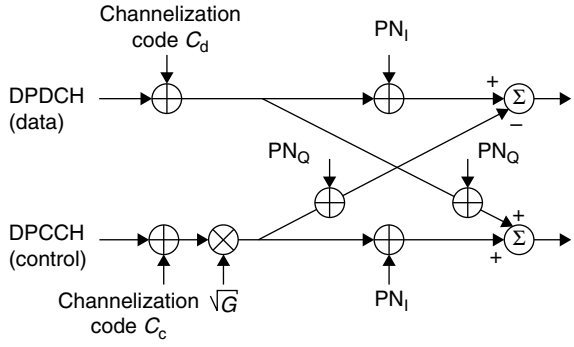


Figure 13.13 I/Q multiplexing with complex spreading

Long Scrambling Code

The long scrambling code has a period of 38 400 chips and repeats once in a 10 ms frame. The x sequence is constructed using the polynomial $X^{25} + X^3 + 1$, while the y sequence is constructed using the polynomial $X^{25} + X^3 + X^2 + X + 1$. The resulting sequences thus constitute segments of a set of Gold sequences.

Short Scrambling Code

The short scrambling code has a period of 256 chips and repeats 150 times in a 10 ms frame. The long and short scrambling codes are generated as described by TS 25.213, Section 4.3.2. The UL short codes $S_v(n)$, $n = 0,1, \ldots, 255$, of length 256 chips are obtained by one chip periodic extension of $S(2)$ sequences of length 255. This means that the first chip $[S_v(0)]$ and the last chip $[S_v(255)]$ of any UL short scrambling code are the same. The modulating chip rate is 3.84 Mcps. The uplink modulation of both DPCCH and DPDCH is BPSK. The complex scrambling codes are formed in such a way that the rotations between consecutive chips within one symbol period are limited to $\pm 90^\circ$. The full 180° rotation can happen only between consecutive symbols.

13.6 Uplink Physical Channels

13.6.1 Dedicated Uplink Physical Channels

There are two types of uplink dedicated physical channels defined; (1) the uplink dedicated physical data channel (uplink DPDCH) and (2) the uplink dedicated physical control channel (uplink DPCCH). In the uplink, DPDCH and DPCCH are transmitted in parallel. The DPDCH and the DPCCH are I/Q code multiplexed within each radio frame. The uplink DPDCH is used to carry the DCH transport channel. At any point of time, there may be zero, one, or several uplink DPDCHs and only one uplink DPCCH on each radio link.

The uplink DPCCH is used to carry control information generated at layer-1. Layer-1 control information consists of known pilot bits to support channel estimation for coherent detection, transmit power-control (TPC) commands, feedback information (FBI), and an optional transport-format combination indicator (TFCI). The transport-format combination indicator informs the receiver about the instantaneous transport format combination of the transport channels mapped to the simultaneously transmitted uplink DPDCH radio frame. Figure 13.14 shows the frame structure of the uplink dedicated physical channels. Each radio frame of length 10 ms is split into 15 slots, each of length $T_{\text{slot}} = 2560$ chips, corresponding to one power-control period.

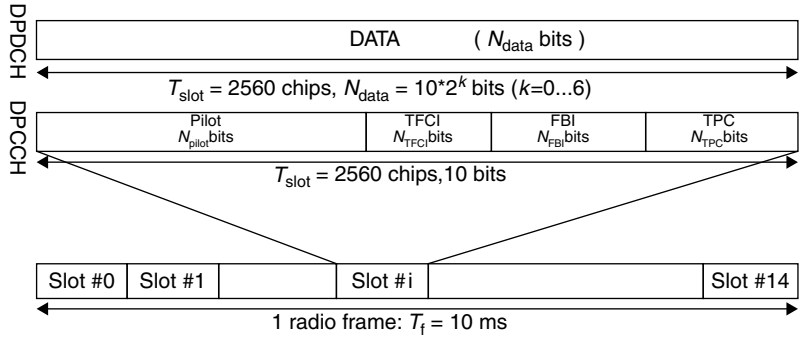


Figure 13.14 Frame structure for uplink DPDCH/DPCCH

In the uplink direction BPSK modulation is used, where each symbol is one bit. The parameter k in Figure 13.14 determines the number of bits per uplink DPDCH slot. It is related to the spreading factor SF of the DPDCH as $SF = 2560/(2^{k*}10)$. As k ranges from 0 to 6, so the DPDCH spreading factor may range from 256 down to 4. The spreading factor of the uplink DPCCH is always equal to $2560/10 = 256$, that is, there are 10 bits per uplink DPCCH slot. The exact number of bits of the uplink DPDCH and the different uplink DPCCH fields (N_{pilot}, N_{TFCI}, N_{FBI}, and N_{TPC}) is given in the standard. Which slot format to use is configured by the higher layers. The FBI bits are used to support techniques requiring feedback from the UE to the UTRAN access point, including closed loop mode transmit diversity and site selection diversity transmission (SSDT).

13.6.1.1 Transmission of Uplink Dedicated Physical Control Channel (DPCCH) and Dedicated Physical Data Channel (DPDCH)

In uplink, DPDCP and DPCCH are transmitted in parallel. In the uplink direction the DPCCH and DPDCH output from the channel codec are processed in L1 as below, and which is shown in Figure 13.15.

1. **Mapping** – A binary "0" is mapped to a real valued $+1$ and a binary "1" is mapped to a real valued -1.
2. **Channelization/Spreading** – As different mobiles use different uplink scrambling codes, the uplink channelization codes may be allocated without coordination between different connections. The uplink channelization codes are always allocated in a predetermined manner. The DPCCH is always spread to chip rate (3.84 Mcps) by the channelization code $C_{ch,256,0}$ ($SF = 256$), whereas the nth DPDCH$_n$ ($0 <= n <= 6$) is spread to the chip rate by the channelization code $C_{ch,SF,n}$.
3. **Combination** – One DPCCH and up to six parallel DPDCHs can be transmitted simultaneously. The DPCCH spreading factor (SF) is always 256, while the DPDCH SF can vary from 4 to 256. After the channelization the real valued spread signals are weighted by gain factor β_c (for DPCCH) and β_d (for DPDCH), then the stream of real valued chips on the I and Q branches are summed and treated as a complex valued stream of chips. This is then scrambled by complex valued scrambling code C_{scramb}.
4. **Scrambling** – The uplink scrambling code is decided by the network only. The mobile is informed in the downlink access grant message about what scrambling code to use. The scrambling code may, in rare cases, be changed during the duration of the connection. The change of uplink scrambling code is

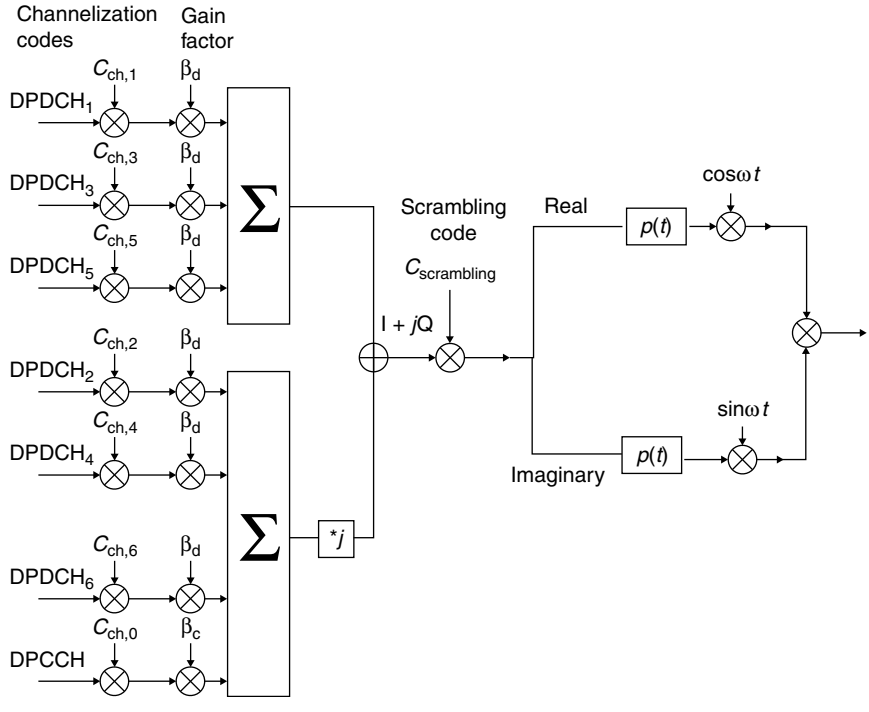


Figure 13.15 Spreading/modulation for uplink DPCCH and DPDCHs

negotiated over the DCCH. There are 2^{24} uplink scrambling codes. Either short (used with MUD only) or long scrambling codes can be used for the uplink transmission (as described in the previous section).

5. **Modulation** – This is then BPSK modulated.

For hand-over purposes, the UE should be able to establish an uplink connection with two cells.

13.6.2 Common Uplink Physical Channels

13.6.2.1 Physical Random Access Channel (PRACH)

Physical random access channel (PRACH) is an uplink channel used by UE for connection request purposes. PRACH is used to carry the RACH transport channel data. The random-access transmission is based on a slotted ALOHA approach with fast acquisition indication. The UE can start the random-access transmission at the beginning of a number of well defined time intervals, denoted access slots. There are 15 access slots per two frames and they are spaced 5120 chips apart, see Figure 13.16 (each slot is 2560 chips). The timing of the access slots and the acquisition indication is described later. Information about the available access slots for random-access transmission is given by the network higher layers via BCH.

The structure of the random-access transmission is shown in Figure 13.17. The random-access transmission consists of one or several preambles of length 4096 chips and a message of length 10 or 20 ms. The mobile station indicates the length of the message part to the network by using specific signatures.

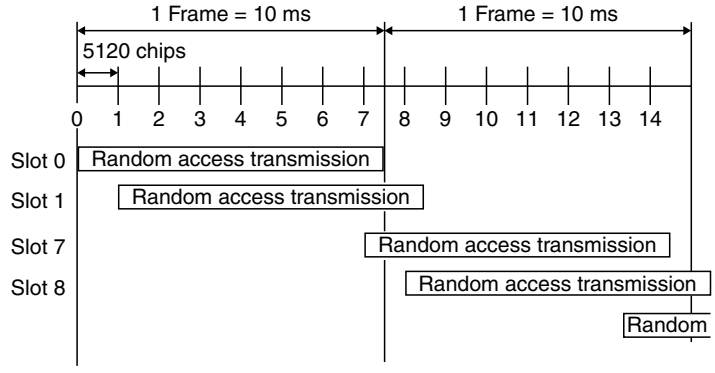


Figure 13.16 RACH access slot numbers and their spacing

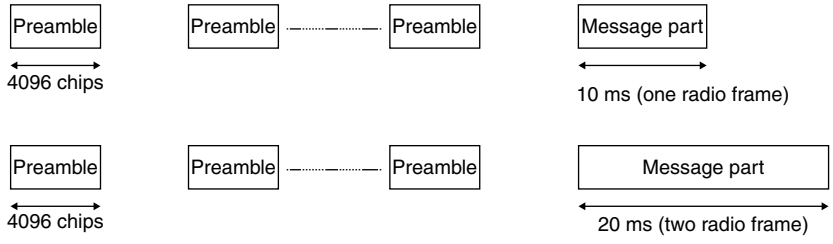


Figure 13.17 Structure of the random-access transmission

UE transmits the preamble by random access before sending the message part. When it receives the acquisition indication corresponding to the preamble from the network, UE sends the message part.

RACH Preamble Part: Each preamble is of length 4096 chips and consists of 256 repetitions of a signature of length16 chips Walsh code. The term for the preamble having a repetitive Walsh code is a preamble signature. There are a maximum of 16 available signatures. UE randomly selects one of them prior to each access attempt. The characteristics of the preamble part are:

Channel coding: 256 repetitions of complex signature of length 16
Scrambling: real part of long Gold sequence.

For scrambling code of the preamble part, the code generating method is the same as for the real part of long codes on dedicated channels. Only the first 4096 chips of the code are used for preamble spreading with the chip rate of 3.84 Mcps. The long code C_1 for the in-phase components is used directly on both in-phase and quadrature branches without offset between branches. The preamble scrambling code is defined as position-wise mod-2 sum of 4096 chips segments of two binary m-sequences generated by means of two generator polynomials of degree 25.

RACH Message Part: Figure 13.18 shows the structure of the random-access message part in a radio frame. The 10 ms message part radio frame is split into 15 slots, each of length $T_{\text{slot}} = 2560$ chips. Each slot consists of two parts, a data part to which the RACH transport channel is mapped and a control part that carries layer-1 control information. The data and control parts are transmitted in parallel. A 10 ms

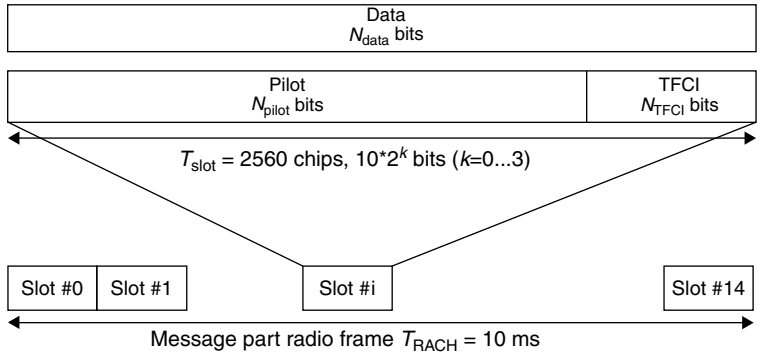


Figure 13.18 Structure of the random-access message part radio frame

message part consists of one message part radio frame, while a 20 ms message part consists of two consecutive 10 ms message part radio frames. The message part length can be determined from the used signature and/or access slot, as configured by higher layers. The data part consists of 10×2^k bits, where $k = 0, 1, 2, 3$. This corresponds to a spreading factor of 256, 128, 64, and 32 respectively, for the message data part. The control part consists of 8 known pilot bits to support channel estimation for coherent detection and 2 TFCI bits. This corresponds to a spreading factor of 256 for the message control part. The pilot bit pattern is described in the standard. The total number of TFCI bits in the random-access message is $15 \times 2 = 30$. The TFCI of a radio frame indicates the transport format of the RACH transport channel mapped to the simultaneously transmitted message part radio frame. For a 20 ms PRACH message part, the TFCI is repeated in the second radio frame.

The message part is scrambled with a 10 ms complex code and the scrambling code is cell specific (Figure 13.19). The parameter for the data and control part is shown in Table 13.4.

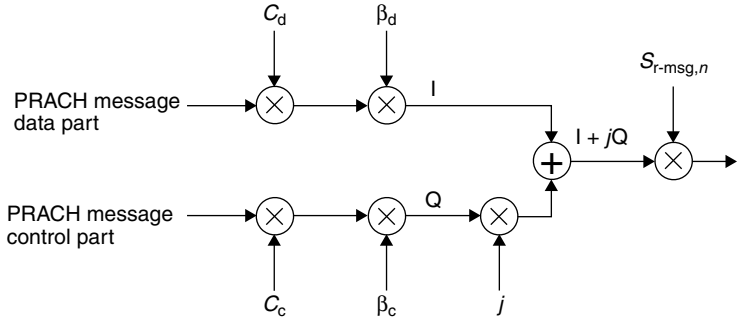


Figure 13.19 Spreading of PRACH message part

PRACH Transmission Procedure

The following steps are carried out during a random access burst.

1. The terminal decodes the BCH (PCCPCH) of the target cell to find out: the cell specific spreading codes available for preamble and message parts, the signatures and access slots available in the cell, the spreading factor allowed for message part, and the PCCPCH transmit power level.

Table 13.4 Parameters for data and control part

Data part		Control part	
Symbol rate	15/30/60/120 ksymb/s	Symbol rate	15 ksymb/s
Spreading factor	32/64/128/256	Spreading factor	256
Channel coding	CRC, convolutional code	Pilot symbol	8 bits per slot
Coding rate	$^1/_2$	TFCI	2 bits per slot
Modulation	BPSK		
Spreading	Orthogonal Gold codes		
Scrambling	Gold sequence, 3.84 Mcps, 10 ms periodic		

2. The mobile randomly selects the signature and access slot to be used for the RACH burst.
3. The mobile estimates the downlink path loss and calculates the required uplink transmit power to be used for the random access burst.
4. A 1 ms preamble is sent with the selected signature.
5. The terminal decodes the AICH to see whether the base station has detected the preamble.
6. In case no AICH is detected, the terminal increases the preamble transmission power by a step given by the station, as multiples of 1 dB and transmits in the next available access slot.
7. If AICH is received with the signature *S* of the PRACH, then the message part is sent.

13.6.2.2 Physical Common Packet Channel (PCPCH)

The physical common packet channel (PCPCH) is used to carry the transport channel-CPCH.

CPCH Transmission – The CPCH transmission is based on the DSMA-CD approach with fast acquisition indication. The UE can start the transmission at the beginning of a number of well defined time-intervals, relative to the frame boundary of the received BCH of the current cell. The access slot timing and structure is identical to RACH. The structure of the CPCH access transmission is shown in Figure 13.20. The PCPCH access transmission consists of one or several access preambles (A-P) of length 4096 chips, one collision detection preamble (CD-P) of length 4096 chips, a DPCCH power control preamble (PC-P), which is either 0 slots or 8 slots in length, and a message of variable length $N \times 10$ ms.

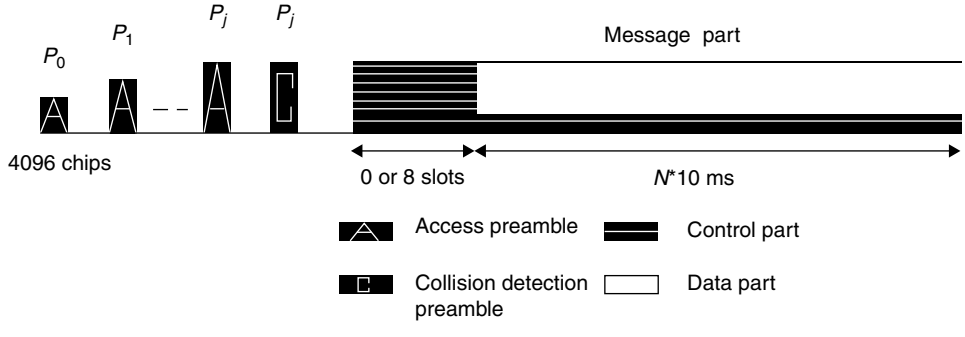


Figure 13.20 Structure of the CPCH access transmission

CPCH Access Preamble Part – Similar to the RACH preamble part, the CPCH preamble signature sequences are used. The number of sequences used could be less than the ones used in the RACH preamble. The scrambling code could either be chosen to be a different code segment of the Gold code

used to form the scrambling code of the RACH preambles, or could be the same scrambling code in situations where the signature set is shared.

CPCH Power Control Preamble Part – The power control preamble segment is called the CPCH power control preamble (PC-P) part. The power control preamble length is a parameter which will take the values 0 or 8 slots, as set by the higher layers.

CPCH Message Part – Each message consists of up to N_Max_frames 10 ms frames. N_Max_frames is a higher layer parameter. Each 10 ms frame is split into 15 slots, each of length $T_{\text{slot}} = 2560$ chips. Each slot consists of two parts, a data part that carries higher layer information and a control part that carries layer-1 control information. The data and control parts are transmitted in parallel. The spreading factor for the control part of the CPCH message part will be 256. The slot format of the control part of the CPCH message part will be the same as the control part of the CPCH PC-P. Each frame of length 10 ms is split into 15 slots, each of length $T_{\text{slot}} = 2560$ chips, corresponding to one power-control period.

The data part consists of 10×2^k bits, where $k = 0, 1, 2, 3, 4, 5, 6$, corresponding to spreading factors of 256, 128, 64, 32, 16, 8, and 4, respectively.

The spreading factor for the control part of the CPCH message part will be 256. The slot format of the control part of the CPCH message part will be the same as the control part of the CPCH PC-P. There are two types of uplink dedicated physical channels; those that include TFCI (for example, for several simultaneous services) and those that do not include TFCI (for example, for fixed-rate services). It is the UTRAN that determines if a TFCI should be transmitted and it is mandatory for all UEs to support the use of TFCI in the uplink.

In the compressed mode, DPCCH slot formats with TFCI fields are changed. There are two possible compressed slot formats for each normal slot format. They are labeled A and B, and the selection between them is dependent on the number of slots that are transmitted in each frame in compressed mode.

Multi-code operation is possible for the uplink dedicated physical channels. When multi-code transmission is used, several parallel DPDCH are transmitted using different channelization codes, however, there is only one DPCCH per radio link.

The spreading and modulation of the message part of the PRACH and PCPCH are basically the same as for the uplink dedicated channel consisting of a data channel DPDCH and control channel DPCCH (Figure 13.21). Table 13.5 provides the details for access preamble, CD preamble, CPCH power control and CPCH message part.

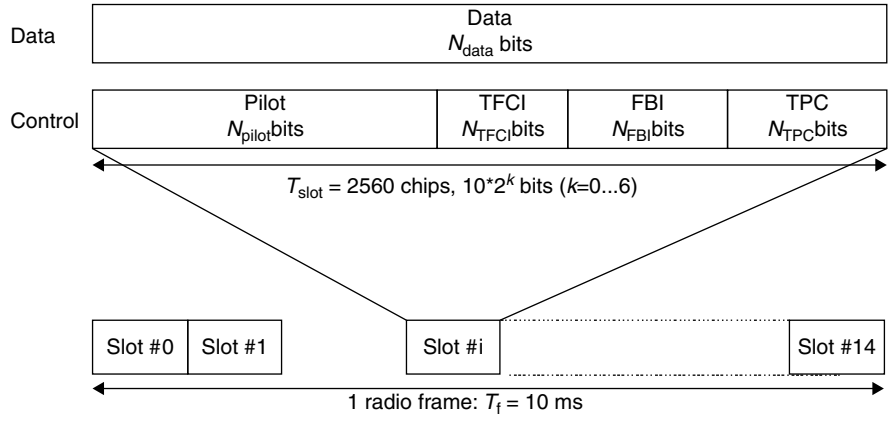


Figure 13.21 Frame structure for uplink data and control parts associated with PCPCH

Table 13.5 Access preamble, CD preamble, CPCH power control and CPCH message part

Access preamble part	CD preamble	CPCH power control part	CPCH message part
The complex-valued access preamble consists of access preamble codes and its corresponding scrambling code. The construction of the access preamble codes requires the knowledge of the preamble signature.	The complex-valued CD preamble consists of CD preamble codes and its corresponding scrambling code. The construction of the CD preamble codes requires the knowledge of the preamble signature.	This consists of access preamble signatures and CD preamble signatures.	The signature in the preamble points to channelization codes of length 16.
Channel coding – 256 repetitions for one of 16 signatures.	Channel coding – 256 repetitions for one of 16 signatures.	Symbol rate 15 ksymb/s	The message part is scrambled with a 10 ms complex code. The scrambling code is cell specific.
Scrambling real part of Gold sequence (first 4096 chips).	Scrambling real part of Gold sequence (chip index 4096.8191).	Spreading factor 256	Each message consists of a number of 10 ms (15 slot) frames. Each slot consists of a data part and control part. Spreading factor 4–64 for data part, and 256 for control part.

13.7 Downlink Physical Channels

13.7.1 Dedicated Downlink Physical Channels

There is only one type of downlink dedicated physical channel, which is known as the downlink dedicated physical channel (downlink DPCH). Within one downlink DPCH, dedicated data generated at layer-2 and above, that is, the dedicated transport channel (DCH), is transmitted in time-multiplex with control information generated at layer-1 (known pilot bits, TPC commands, and an optional TFCI). In the downlink direction the DPDCH and DPDCH are time multiplexed as shown in Figure 13.22. Each frame of length 10 ms is split into 15 slots, each of length $T_{slot} = 2560$ chips, corresponding to one power-control period.

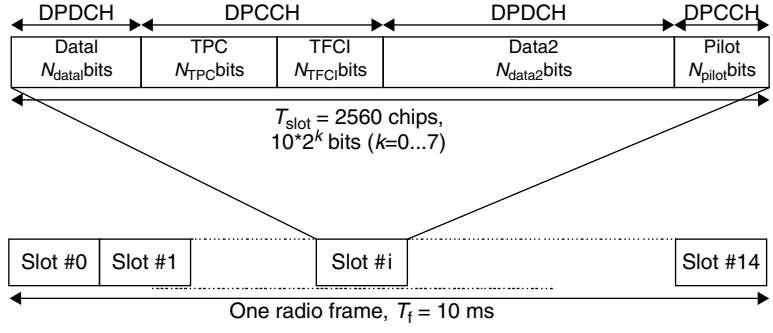


Figure 13.22 Frame structure for downlink DPCH

The QPSK modulation, where each symbol contains two bits is used in the downlink direction. The parameter k in Figure 13.22 determines the total number of bits per downlink DPCH slot. It is related to the spreading factor SF of the physical channel by $SF = 512/2^k$ (where $k = 0, \ldots, 7$). The spreading factor may thus range from 512 down to 4. Hence, in each slot of length 2560 chips, we are able to put twice the number of bits compared with uplink (BPSK) as QPSK modulation is used. The exact number of bits of the different downlink DPCH fields (N_{pilot}, N_{TPC}, N_{TFCI}, N_{data1}, and N_{data2}) is given in the standard. Which slot format to use is configured by the higher layers, and can also be reconfigured by the higher layers.

There are basically two types of downlink dedicated physical channels: those that include TFCI (for example, for several simultaneous services), and those that do not include TFCI (for example, for fixed-rate services). It is the UTRAN that determines if a TFCI should be transmitted and it is mandatory for all UEs to support the use of TFCI in the downlink.

DPCH Transmission Procedure

Data modulation is QPSK where each pair of 2 bits is serial-to-parallel converted and mapped to the I- and Q-branch, respectively. The I- and Q-branches are then spread with the same channelization code (real spreading) and subsequently scrambled by the scrambling code (complex scrambling). The channelization code is derived from the OVSF code tree based on the SF used, and the scrambling code is the sector (base station) specific scrambling code (Figure 13.23). Parameters for downlink DPCH are given in Table 13.6.

Table 13.6 Parameters for downlink DPCH

Channel coding	CRC, convolutional code, Turbo code (according to Qos) in DPDCH
Symbol rate	7.5/15/30/60/120/240/480/960 ksymb/s
Spreading factor	4/8/16/32/64/128/256/512
Modulation	QPSK
Spreading	OVSF codes
Scrambling	Gold sequence, 3.84 Mcps, 10 ms periodic
Power control period	0.625 ms
Pilot symbol	Include
TFCI bits	Include
TPC bit	Include

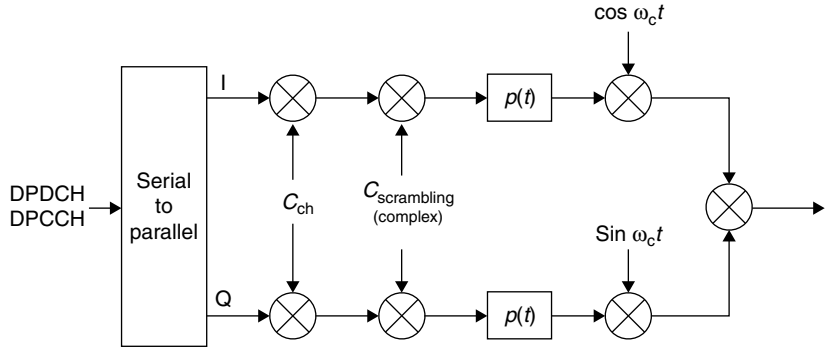


Figure 13.23 Downlink spreading arrangement and modulation

In compressed mode, a different slot format is used compared with normal mode. There are two possible compressed slot formats, which are labeled A and B. Format B is used for compressed mode by spreading factor reduction and format A is used for all other transmission time reduction methods.

13.7.2 Common Downlink Physical Channels

13.7.2.1 Common Pilot Channel (CPICH)

The pilot channel is used by the base station to provide a reference to all mobile stations and to aid the channel estimation at the terminals. It provides phase reference for coherent demodulation at the mobile receiver to enable coherent detection. This is an unmodulated code channel, which is scrambled with the cell-specific primary scrambling code. It has a predefined bit sequence, which for a single transmit antenna is an all logical 1 sequence. The CPICH is a fixed rate (30 kbps, $SF = 256$) downlink physical channel that carries a pre-defined bit/symbol sequence. Figure 13.24 shows the frame structure of the CPICH.

There are two types of CPICH: *primary* and *secondary*. The P-CPICH provides a coherent reference to obtain the SCH, P-CCPCH, AICH and PICH at the UEs, as these channels do not carry their own pilot information. The channelization code used by the P-CPICH is $C_{ch,256,0}$, an all logical 1 code, while its scrambling code is the cell's primary scrambling code. In the case of a single transmit antenna, the CPICH

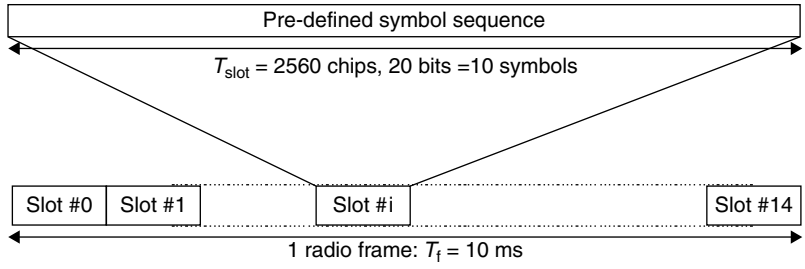


Figure 13.24 Frame structure for common pilot channel

is the unmodulated primary scrambling code. There is only one P-CPICH in each cell, and this is broadcast over the entire cell.

UE estimates the channel impulse response from the received pilot signal, and armed with this response the data may be recovered. Thus the pilot and data must be transmitted over the same radio channel (which includes the transmitter and receiver antenna). Consequently, as the CPICH is transmitted over the entire cell or sector, it cannot be used to recover data from a narrow beam of a smart antenna because the radio channels for the data and CPICH may be very different. A smart antenna with its narrow beams will create radio channels with few or no significant multipath components, unlike a wide angle beam.

The secondary common pilot channel (S-CPICH) provides a common coherent reference within part of a cell or sector. The antenna have narrow beams, for example, from a smart antenna, and may be used to target individual UEs or groups of UEs in close proximity to one another.

The Node B (a BS) may use any channelization code having a length of 256 chips. The SCPICH may be used as reference for the S-CCPCH (which transmits paging messages) and the downlink dedicated channels.

There are two types of common pilot channels, the primary and secondary CPICH. They differ in their use and the limitations placed on their physical features.

Primary Common Pilot Channel (P-CPICH)

The primary common pilot channel (P-CPICH) has the following characteristics.

- The same channelization code is always used for the P-CPICH.
- The P-CPICH is scrambled by the primary scrambling code.
- There is one and only one P-CPICH per cell.
- The P-CPICH is broadcast over the entire cell.

The primary CPICH is the phase reference for the following downlink channels: SCH, primary CCPCH, AICH, and PICH. The primary CPICH is also the *default* phase reference for all other downlink physical channels.

Secondary Common Pilot Channel (S-CPICH)

A secondary common pilot channel (S-CPICH) has the following characteristics.

- An arbitrary channelization code of $SF = 256$ is used for the S-CPICH.
- An S-CPICH is scrambled by either the primary or a secondary scrambling code.
- There may be zero, one, or several S-CPICH per cell.
- An S-CPICH may be transmitted over the entire cell or only over a part of the cell.

- An S-CPICH may be the reference for the S-CCPCH and the downlink DPCH. If this is the case, the UE is informed about this by higher-layer signaling.

13.7.2.2 Primary Common Control Physical Channel (P-CCPCH)

The primary CCPCH is a fixed rate (30 kbps, $SF = 256$) downlink physical channel used to carry the BCH transport channel. Figure 13.25 shows the frame structure of the primary CCPCH. The frame structure differs from the downlink DPCH in that no TPC commands, no TFCI, and no pilot bits are transmitted. The primary CCPCH is not transmitted during the first 256 chips of each slot. Instead, primary SCH and secondary SCH are transmitted during this period. Table 13.7 tabulates the parameters for PCCPCH.

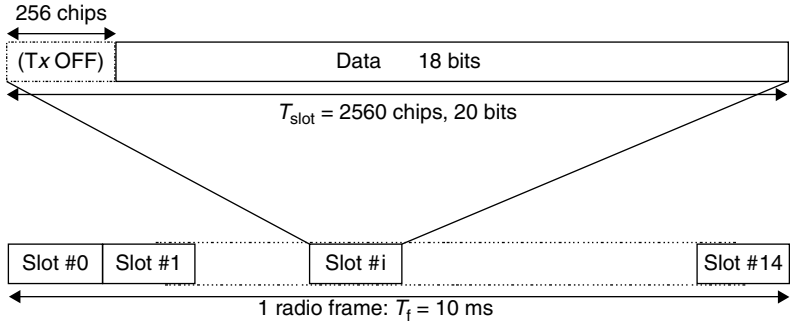


Figure 13.25 Frame structure for primary common control physical channel

Table 13.7 Parameters for PCCPCH

Channel coding	CRC, convolutional code
Symbol rate	30 ksymb/s
Spreading factor	256
Modulation	QPSK
Spreading	Predefined code ($C_{ch,256.1}$)
Scrambling	Gold sequence, 3.84 Mcps, 10 ms periodic, primary scrambling code of the sector
Power control	Not supported
Pilot symbol	Include
TFCI bits	Not included

13.7.2.3 Secondary Common Control Physical Channel (S-CCPCH)

The secondary CCPCH is used to carry the FACH and PCH. There are two types of secondary CCPCH: those that include TFCI and those that do not include TFCI. It is the UTRAN that determines if a TFCI should be transmitted, hence making it mandatory for all UEs to support the use of TFCI. The set of possible rates for the secondary CCPCH is the same as for the downlink DPCH. The frame structure of the secondary CCPCH is shown in Figure 13.26.

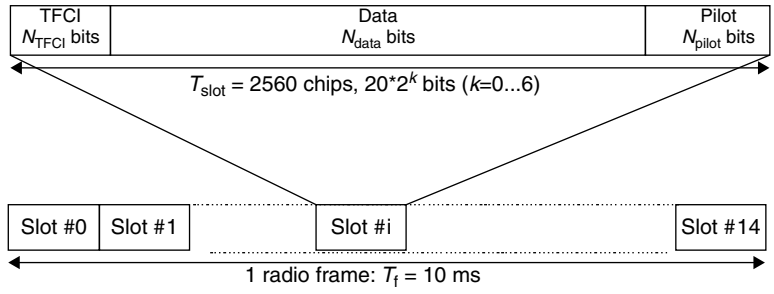


Figure 13.26 Frame structure for secondary common control physical channel

The parameter k in Figure 13.26 determines the total number of bits per downlink secondary CCPCH slot. It is related to the spreading factor SF of the physical channel as $SF = 256/2^k$. The spreading factor range is from 256 down to 4.

The FACH and PCH can be mapped to the same or to separate secondary CCPCHs. If FACH and PCH are mapped to the same secondary CCPCH, they can be mapped to the same frame. The main difference between a CCPCH and a downlink dedicated physical channel is that a CCPCH is not inner-loop power controlled. The main difference between the primary and secondary CCPCH is that the transport channel mapped to the primary CCPCH (BCH) can only have a fixed predefined transport format combination, while the secondary CCPCH support multiple transport format combinations using TFCI. Furthermore, a primary CCPCH is transmitted over the entire cell while a secondary CCPCH may be transmitted in a narrow lobe in the same way as a dedicated physical channel (only valid for a secondary CCPCH carrying the FACH). Parameters for SCCPCH are mentioned in Table 13.8.

Table 13.8 Parameters for SCCPCH

Channel coding	CRC, convolutional code
Symbol rate	15/30/60/120/240/480/960 ksymb/s
Modulation	QPSK
Spreading	Predefined code broadcast on the BCH
Scrambling	Gold sequence, 3.84 Mcps, 10 ms periodic
Power control	Not supported
Pilot symbol	Include
TFCI bits	Included/not include

13.7.2.4 Synchronization Channel (SCH)

The synchronization channel (SCH) is a downlink channel used for initial synchronization purposes and cell search. The SCH consists of two subchannels, the primary and secondary SCH. The 10 ms radio frames of the primary and secondary SCH are divided into 15 slots, each of length 2560 chips. Figure 13.27 illustrates the structure of the SCH radio frame.

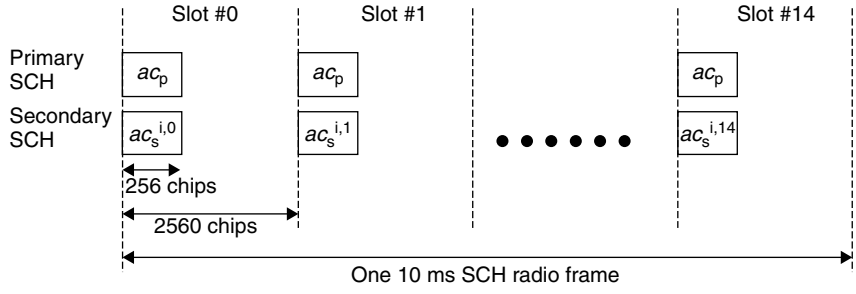


Figure 13.27 Structure of synchronization channel (SCH)

The primary SCH consists of a modulated code of length 256 chips, the primary synchronization code (PSC) denoted c_p in Figure 13.27, is transmitted once every slot. The PSC is the same for every cell in the system.

The secondary SCH consists of repeatedly transmitting a length of 15 sequences of modulated codes of length 256 chips, the secondary synchronization codes (SSC), transmitted in parallel with the primary SCH. The SSC is denoted $ac_s^{i,k}$ in Figure 13.27, where $i = 0, 1, \ldots, 63$ is the number of the scrambling code group, and $k = 0, 1, \ldots, 14$ is the slot number. Each SSC is chosen from a set of 16 different codes of length 256. This sequence on the secondary SCH indicates which of the code groups the cell's downlink scrambling code belongs to (Figure 13.28).

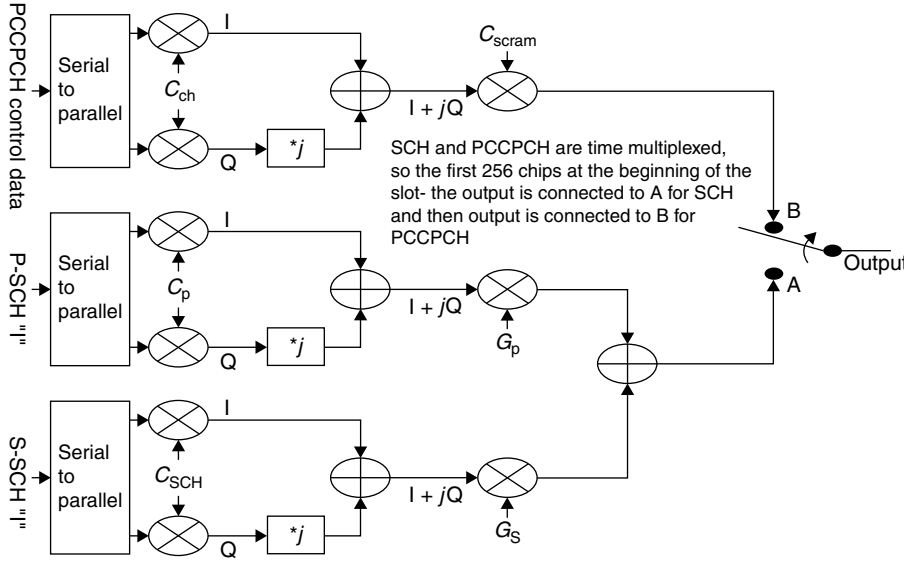


Figure 13.28 Spreading and scrambling for PCCPCH, SCCPCH, and SCH channels

13.7.2.5 Physical Downlink Shared Channel (PDSCH)

The physical downlink shared channel (PDSCH), used to carry the downlink shared channel (DSCH) transport channel, is shared by users based on code multiplexing. As the DSCH is always associated with one or several DCHs, the PDSCH is always associated with one or several downlink DPCHs. More exactly, each PDSCH radio frame is associated with one downlink DPCH.

The frame and slot structure of the PDSCH are shown on Figure 13.29.

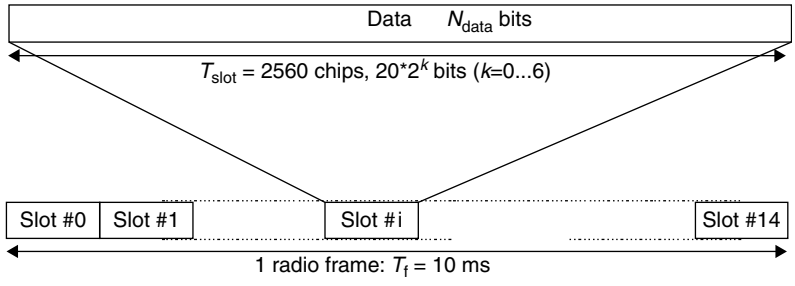


Figure 13.29 Frame structure for the PDSCH

To indicate for UE that there are data to decode on the DSCH, two signaling methods are used, either the TFCI field, or higher layer signaling.

The PDSCH transmission with associated DPCH is a special case of multi-code transmission. The PDSCH and DPCH do not =necessarily have the same spreading factors. Furthermore, the PDSCH spreading factors may vary from frame to frame. All relevant layer-1 control information is transmitted on the DPCCH part of the associated DPCH, that is, the PDSCH does not carry physical layer information. For PDSCH, the allowed spreading factors may vary from 256 to 4.

If the spreading factor and other physical layer parameters can vary on a frame-by-frame basis, the TFCI will be used to inform the UE what are the instantaneous parameters of PDSCH, including the channelization code from the PDSCH OVSF code tree. A DSCH may be mapped to multiple parallel PDSCHs. In such a case the parallel PDSCHs will be operated with frame synchronization between each other and the spreading factors of all PDSCH codes will be the same. PDSCH parameters are mentioned in Table 13.9.

Table 13.9 Parameters for PDSCH

Channel coding	CRC, convolutional code, and Turbo code (according to Qos)
Symbol rate	15/30/60/120/240/480/960 ksymb/s
Modulation	QPSK
Spreading	OVSF code
Scrambling	Gold sequence, 3.84 Mcps, 10 ms periodic, primary scrambling code
Power control	Supported by associated DCH
Pilot symbol	Not include
TFCI bits	Supported by associated DCH

13.7.2.6 Acquisition Indication Channel (AICH)

The acquisition indication channel (AICH) is a physical channel used to carry the acquisition indicators (AI). Acquisition indicator AI_s corresponds to the signature S of the PRACH.

Figure 13.30 shows the structure of the AICH, which consists of a repeated sequence of 15 consecutive access slots (AS), each of length 40-bit intervals. Each access slot consists of two parts, an acquisition indicator (AI) part consisting of 32 real-valued symbols $a_0, \ldots, a_{31}$ and a part of 1024 chips duration with no transmission.

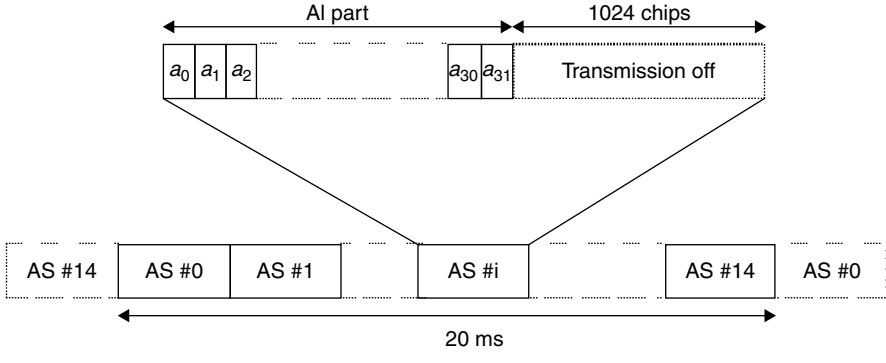


Figure 13.30 Structure of acquisition indication channel (AICH)

The phase reference for the AICH is the primary CPICH.

The real-valued symbols $a_0, a_1, \ldots, a_{31}$ in Figure 13.30 are given by:

$$a_j = \sum_{s=0}^{15} \mathrm{AI}_s b_{s,j}$$

where AI_s, taking the values $+1, -1$, and 0, is the acquisition indicator corresponding to signature S and the sequence $b_{s,0}, \ldots, b_{s,31}$ is given in the standard. Parameters for AICH are shown in Table 13.10.

Table 13.10 Parameters for AICH

Channel coding	Orthogonal code
Symbol rate	15 ksymb/s
Spreading factor	256
Spreading code	OVSF
Scrambling	Gold sequence, 3.84 Mcps, 10 ms periodic

13.7.2.7 Paging Indicator Channel (PICH)

The paging indicator channel (PICH) is always associated with a paging channel (PCH) on S-CCPCH. The PICH carries the page indicators (PIs), where PI indicates the subset of UEs within a cell whether they should check the next S-CCPCH frame for paging messages. UE in idle mode receives nothing but the PI. UE receives PCH in the radio frame of the S-CCPCH corresponding to the PI, only when it is informed of an incoming call by the PI. PIs are divided into several groups. This helps to save battery life. The paging

indicator channel (PICH) is a fixed rate ($SF = 256$) physical channel used to carry the paging indicators (PI). The PICH is always associated with an S-CCPCH to which a PCH transport channel is mapped.

Figure 13.31 illustrates the frame structure of the PICH. One PICH radio frame of length 10 ms consists of 300 bits ($b_0, b_1, \ldots, b_{299}$). Of these, 288 bits ($b_0, b_1, \ldots, b_{287}$) are used to carry paging indicators. The remaining 12 bits ($b_{288}, b_{289}, \ldots, b_{299}$) are undefined.

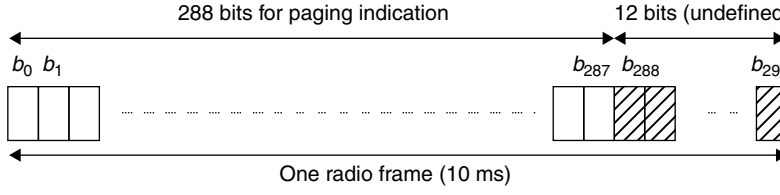


Figure 13.31 Structure of paging indicator channel (PICH)

N paging indicators $\{\mathrm{PI}_0, \ldots, \mathrm{PI}_{N-1}\}$ are transmitted in each PICH frame, where $N = 18, 36, 72$, or 144.

The PI calculated by higher layers for use for a certain UE, is mapped to the paging indicator PI_p, where p is computed as a function of the PI computed by higher layers, the SFN of the P-CCPCH radio frame during which the start of the PICH radio frame occurs, and the number of paging indicators per frame (N):

$$p = \left\{ PI + \left[(\{18 \times [SFN + (SFN/8) + (SFN/64) + (SFN/512)]\} \bmod 144) \times \frac{N}{144} \right] \right\} \bmod N$$

If a paging indicator in a certain frame is set to "1," it is an indication that UEs associated with this paging indicator should read the corresponding frame of the associated S-CCPCH. PICH parameters are mentioned in Table 13.11.

Table 13.11 Parameters for PICH

Symbol rate	15 ksymb/s
Spreading factor	256
Spreading code	OVSF
Scrambling	Gold sequence, 3.84 Mcps, 10 ms periodic
Number of PI	18/36/72/144

There are some other channels are also defined, such as: the CPCH status indicator channel (CSICH), which is a fixed rate ($SF = 256$) physical channel used to carry CPCH status information; the collision detection channel assignment indicator channel (CD/CA-ICH), which is a physical channel used to carry the CD indicator (CDI) only if the CA is not active, or CD indicator/CA indicator (CDI/CAI) at the same time if the CA is active, and so on.

13.8 Timing Relationship between Physical Channels

The P-CCPCH, on which the cell SFN is transmitted, is used as a timing reference for all the physical channels, directly for downlink and indirectly for uplink. SCH (primary and secondary), CPICH (primary

and secondary), P-CCPCH, and PDSCH have identical frame timings. The S-CCPCH timing may be different for different S-CCPCHs, but the offset from the P-CCPCH frame timing is a multiple of 256 chips. The PICH timing is $t_{\text{PICH}} = 7680$ chips prior to its corresponding S-CCPCH frame timing, that is, the timing of the S-CCPCH carrying the PCH transport channel with the corresponding paging information. AICH access slots 0 start at the same time as P-CCPCH frames with (SFN modulo 2) = 0. The DPCH timing may be different for different DPCHs, but the offset from the P-CCPCH frame timing is a multiple of 256 chips.

13.8.1 Channel Number and Bands

The carrier frequency is designated by the UTRA absolute radio frequency channel number (UARFCN). For each operating band, the UARFCN values are defined as follows:

$$\text{Uplink: } N_U = 5^*(F_{\text{UL}} - F_{\text{UL\_offset}}), \text{ for the carrier frequency range } F_{\text{UL\_low}} \leq F_{\text{UL}} \leq F_{\text{UL\_high}}$$

$$\text{Downlink: } N_D = 5^*(F_{\text{DL}} - F_{\text{DL\_offset}}), \text{ for the carrier frequency range } F_{\text{DL\_low}} \leq F_{\text{DL}} \leq F_{\text{DL\_high}}$$

13.9 Transmitter Characteristics

UE Maximum Output Power – Different power classes are defined for nominal maximum output power: power class-1, -2, -3, and -4. For band-I, the power is defined as $+33$ dBm for class-1, $+27$ dBm for class-2, $+24$ dBm for class-3, and $+21$ dBm for class-4. The nominal power defined is the broadband transmit power of the UE, that is, the power in a bandwidth of at least $(1 + \alpha)$ times the chip rate of the radio access mode.

Frequency Error – The UE modulated carrier frequency should be accurate within ± 0.1 ppm, observed over a period of one time slot compared with the carrier frequency received from Node B. For the PRACH preambles the measurement interval is lengthened to 3904 chips (this being the 4096 chip nominal preamble period less a 25 μs transient period allowance at each end of the burst). The UE willl use the same frequency source for both RF frequency generation and the chip clock.

Power Control – Open loop power control is the ability of the UE transmitter to set its output power to a specific value. The UE open loop power is defined as the mean power in a time slot or ON power duration, whichever is available. In normal conditions the tolerance is ± 9 dB. Inner loop power control in the uplink is the ability of the UE transmitter to adjust its output power in accordance with one or more TPC commands received in the downlink.

Diversity Characteristics – Three forms of diversity are considered to be available in UTRA/FDD. (1) Time diversity – channel coding and interleaving in both uplink and downlink. (2) Multi-path diversity – Rake receiver or other suitable receiver structure with maximum combining; additional processing elements can increase the delay-spread performance due to increased capture of signal energy. (3) Antenna diversity – antenna diversity with maximum ratio combing in Node B and optionally in the UE.

Reference Sensitivity Level – The reference sensitivity level is the minimum mean power received at the UE antenna port at which the bit error ratio (BER) should not exceed a specific value. The minimum requirement is that the BER should not exceed 0.001 for different bands as specified in the standard. For example, DPCH_Ec < reference sensitivity > -117 for operating bands 1 and 4 unit dBm/3.84 MHz.

13.10 Different Channel Usage in Various Scenarios

As discussed earlier, in the downlink direction some channels are transmitted continuously in each cell, these are treated as overhead channels (Figure 13.32). They are used by the UE to synchronize to the cell, to identify the cell and the network, and to obtain information about how to access the cell.

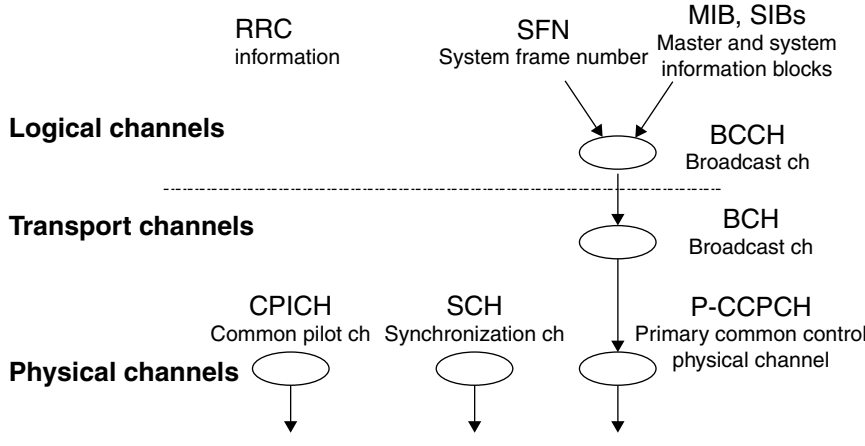


Figure 13.32 Downlink channel usage

The common pilot channel (CPICH) is used to transmit timing and frequency reference information to UEs (mobile stations) that are used by the UE (mobile station) to find the primary scrambling code and to help determine its transmit power during open loop power control. The synchronization channel (SCH) includes the primary and secondary synchronization channels (PSC and SSC) that contain timing information to allow the UE to synchronize to the base station. It is time multiplexed with P-CCPCH. The primary common control physical channel (P-CCPCH) is used to transmit the broadcast transport channel (BCH), which provides system information to the UE. It is time multiplexed with the synchronization channel (SCH), which is used to aid the UE synchronization to the network. In the FDD test operating mode, this channel consists of pseudo-random bit sequence (PRBS) data and a valid system frame number (SFN).

13.10.1 Channel Used for Call Setup

After the UE has identified a cell that it wants to access and has read the access information from the broadcast channel, it must register. Registration informs the network of the presence of the UE and is performed using the location update procedure.

In order to make the communication between the mobile and the network a control connection must be established between the RRC entities of the network side and the UE side. This step is the same whether the purpose is registration, mobile initiated call setup, or network originated call setup. These channels are shown in Figure 13.33 as W-CDMA connection setup channels.

Physical Random Access Channel – Used by the UE to make its initial transmissions to the network.
Acquisition Indication Channel (AICH) – Used to acknowledge UE access request.
Paging Indication Channel (PICH) – Used to alert the UE of a forthcoming page message. In FDD test mode, the test set only provides a user specified bit pattern to allow the operator to verify that the UE is correctly decoding this channel.

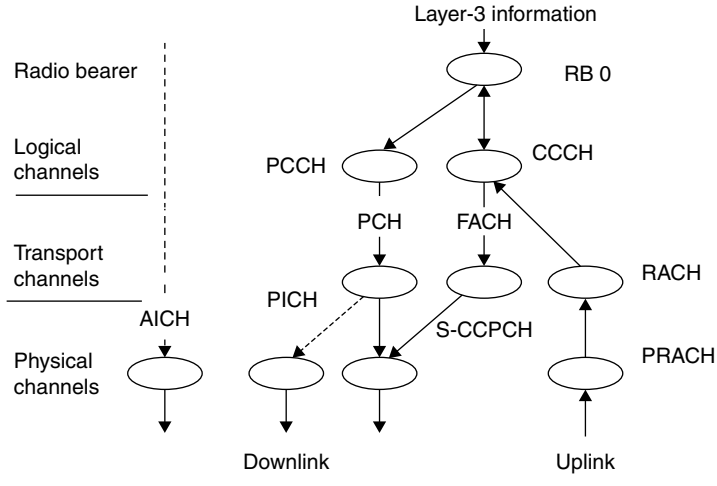


Figure 13.33 Channel usage for call setup

Secondary Common Control Physical Channel (S-CCPCH) – Used to transmit pages and signaling to idling UEs.

13.11 Compressed Mode

Compressed mode is a function that enables the measurement of cells with different frequencies for the purpose of carrying out handover between the different frequencies. The support of the downlink compressed mode is essential for a single carrier UE. The decision to migrate to compressed mode is made by UTRAN, which informs UE of the parameters required for compressed mode. In compressed mode no data transmission takes place in the slot referred to as the transmission gap (Figure 13.34). In a frame of compressed mode, the transmission power is raised temporarily to prevent the degradation in quality due

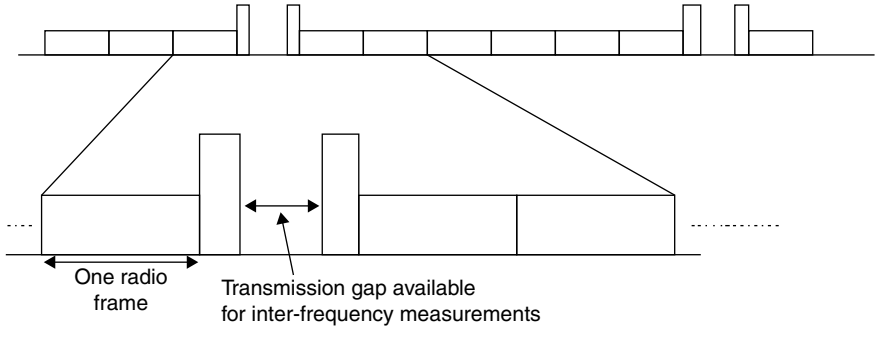


Figure 13.34 Compressed mode

Table 13.12 Different methods of compressed mode

Method	Overview
Compressed mode by puncturing	A way to reduce the number of transmitted bits by using rate matching function. The same *SF* is used in compressed mode as in the normal mode.
Compressed mode by reducing the *SF* by 2	A way to temporarily increase the transmission speed by halving *SF* so that the same number of bits can be transmitted as in the normal mode in slots other than the transmission gap.
Compressed mode by higher layer scheduling	A way to limit the transport format set by the higher layer according to the number of bits that can be transmitted in slots other than the transmission gap. The same *SF* is used in compressed mode as in normal mode. It is basically applicable to non-real time services such as packet transmission.

to the lower gain caused by the suspension of transmission. Various methods for creating the compressed mode are given in the Table 13.12.

The rate and type of compressed frames is variable and depends on the environment and the measurement requirements. There are two different types of frame structures defined for downlink compressed frames. Type A maximizes the transmission gap length and type B is optimized for power control. The frame structure type A or B is set by higher layers independent of the downlink slot format type A or B.

Further Reading

3GPP Technical Specification Group Radio Access Network. *Physical Channels and Mapping of Transport Channels (FDD), 3GPP*, TS 25.211, Version 3.0.0. ETSI TC-SMG, Sophia-Antipolis Cedex.

3GPP Technical Specification Group Radio Access Network. *Spreading and Modulation (FDD), 3GPP*, TS 25.213, Version 3.0.0. ETSI TC-SMG, Sophia-Antipolis Cedex.

3GPP Technical Specification Group (TSG) RAN WG4 UTRA (BS) FDD. *Radio Transmission and Reception, 3GPP*, TS 25.104, Version 3.0.0, www.3gpp.org/Specifications.

Goodman, D.J. (1997) *Wireless Personal Communication Systems*, Addison-Wesley, London.

Zvonar, Z., Jung, P., and Kammerlander, K. (1999) *GSM, Evolution Towards 3rd Generation Systems*, Kluwer Academic Press, Dordrecht.

14

UMTS Mobile Phone Software and Operations

14.1 Introduction to UMTS Protocol Architecture

UTRAN protocol architecture consists of a set of horizontal and vertical layers. The requirements are addressed in horizontal radio network layer across different types of control and user planes. Control planes are used to control a link or for connection, whereas user planes are used for transmitting user data from higher layers. Signaling bearers are used to transmit higher layer's signaling and control information. Data bearers are the frame protocol used to transport user data. The application protocols are used to provide UMTS specific signaling and control within UTRAN, for example bearer setup in a radio network. A complete UTRAN and UE control plane architecture is shown in Figure 14.1.

The protocol architecture in the main interfaces from UE to ISP across the UMTS network for packet switched traffic is illustrated in Figure 14.2a and b. From the UE to UTRAN (RNC), the IP data packets are carried as PDCP packets. PDCP provides either an acknowledged/unacknowledged or transparent transfer service. It also performs a compression/decompression function. From the UTRAN (RNC) to SGSN IP, packets are tunneled using GTP-U. Another GTP-U tunnel then runs from the SGSN to GGSN. Using GTP-U, UMTS can carry a number of different packets, such as IPv4, IPv6, PPP, and X.25 over a common infrastructure. GTP-U packets are formed by adding a header to the underlying PDP packet and are sent using UDP over the IP using the IP address of the tunnel end point, for example, the GGSN for traffic sent from the SGSN to an external network. In the UMTS core network, IP layer-3 routing is typically supported by ATM switching networks. It is the operator's choice to implement QoS at the IP or ATM level.

The communication resources and services in UMTS are controlled via the protocols located in the control plane. The GTP-C protocol takes care of the setting up, modifying, and tearing down of GTP tunnels. It runs between SGSN and GGSN and also carries the messages to set up and delete PDP contexts. GTP-C does not run over the Iu interface between UTRAN (RNC) and SGSN.

The GTP tunnel from UTRAN (RNC) to SGSN is setup by part of the RANAP, which provides the signaling across the Iu interface and is also responsible for: radio access bear setup, modification and release, control of the UTRAN security modes, management of RNC relocation procedures, exchanging user information between RNC and CN, and transport MM and CC information between the UE and CN.

Mobile Handset Design Sajal Kumar Das

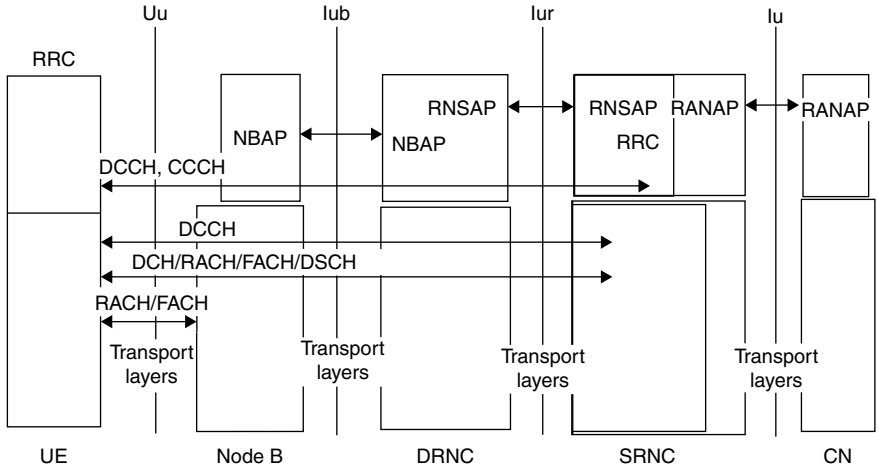


Figure 14.1 UE RANAP protocol architecture

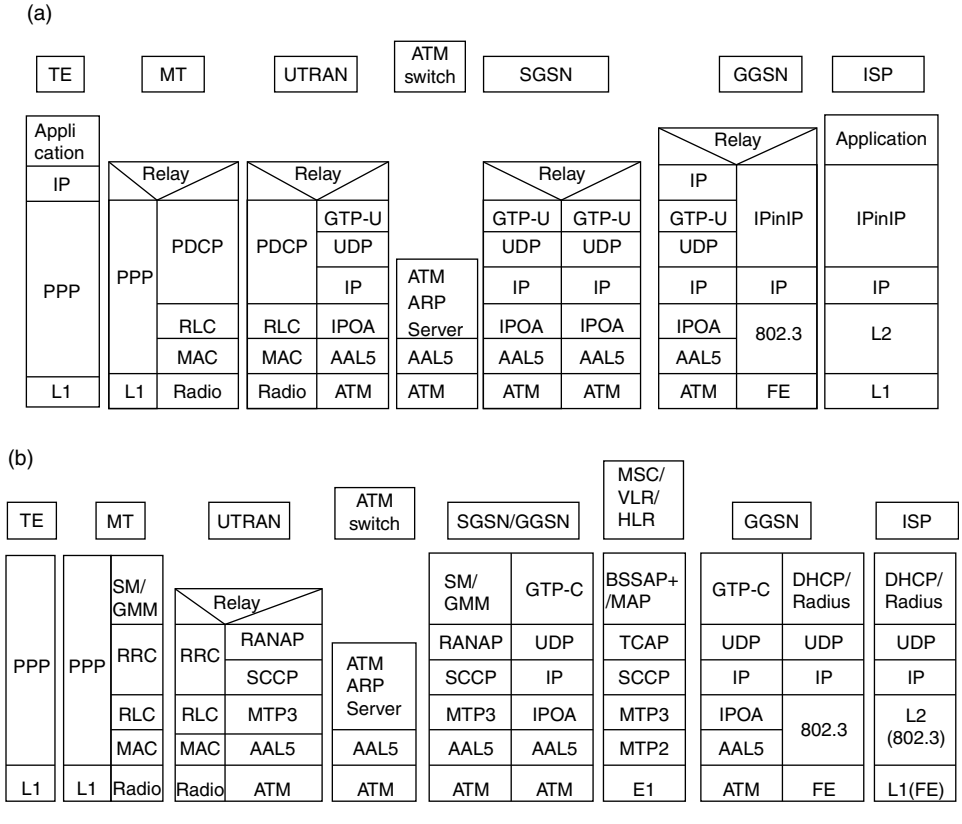


Figure 14.2 (a) UMTS user plane protocol stack, and (b) UMTS control plane protocol stack

Through the Uu interface, the RRC sets up a signaling connection from UE to RNC. It covers the assignment, re-configuration and release of radio resources. RRC also handles handover, cell re-selection, paging updates, and notifications. It decouples the terminal packet data protocol from the network transport through the use of tunneling and can transport IPv4 and IPv6 packets without modification. The underlying UMTS CN can also be IPv4 or IPv6 networks, and it has no interaction with the user data being tunneled over it.

14.2 Protocol Structure

As for GPRS, the W-CDMA protocol has a layered structure designed to give the system a great deal of flexibility. The WCDMA structure is divided vertically into an "access stratum" and a "non-access stratum," and horizontally into a "control plane" and a "user plane." Protocol layers-1 and -2 are in the access stratum, whereas protocol layer-3 is divided between the access and non-access strata (Figure 14.3). In layers-2 and -3 the control plane and user plane information is carried on separate channels. As discussed in the previous chapter, within layer-1 some channels carry only control plane information, while others carry both user and control plane data.

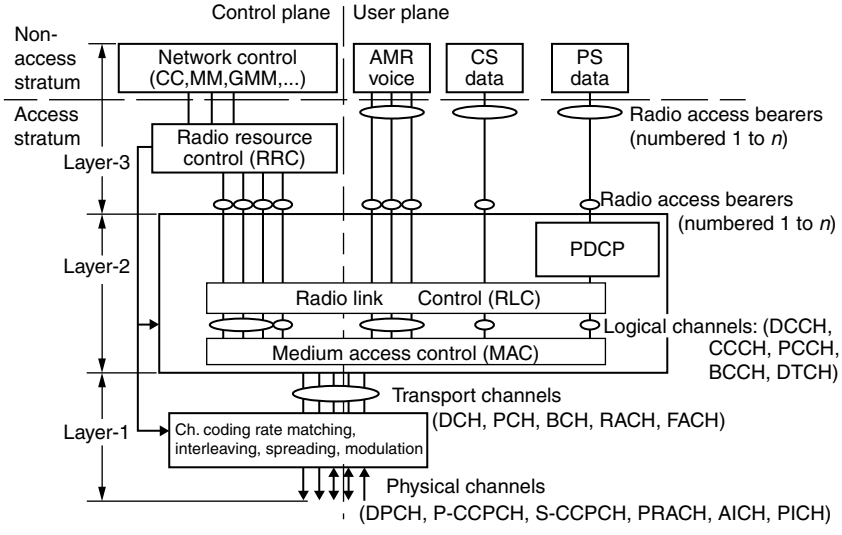


Figure 14.3 WCDMA protocol structure

The non-access stratum (NAS) refers to the protocol layers that are independent of the radio access technology, that is, the CC, MM, GMM, SM, SS, and SMS, whereas the access stratum (AS) refers to the layers of the protocol stack that have some dependency upon the radio technology, that is, RR, RLC, MAC, and PHY.

14.3 UE Protocol Architecture

The protocol entities in the UE software stack are shown in Figure 14.4. As in the GSM phone software, the WCDMA based phone also contains all the necessary pieces of software modules, but the protocol stack software for the WCDMA based phone is different from the GSM mobile phone. In Figure 14.4, the main modules and layers of the WCMDA based mobile phone are shown.

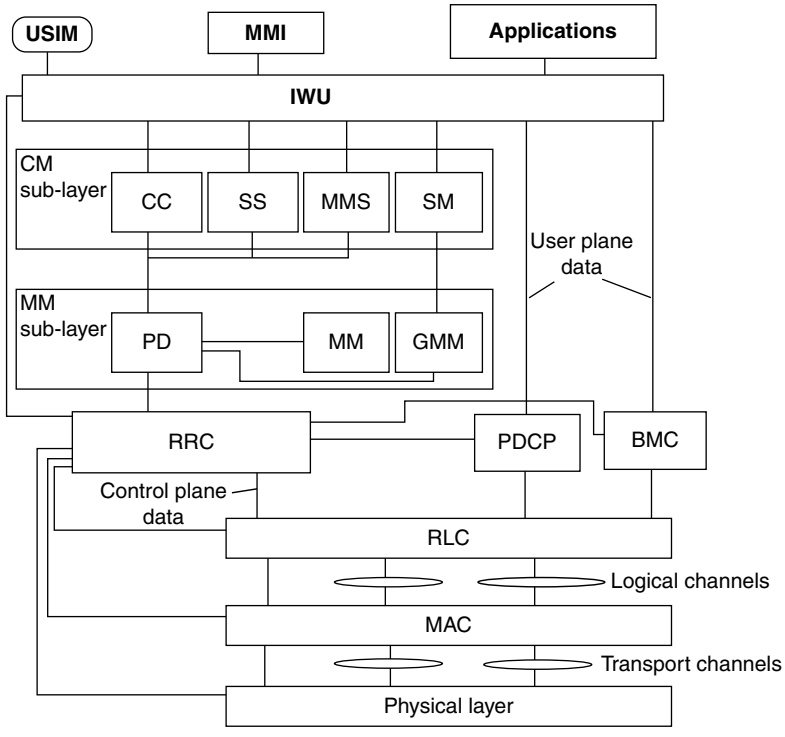


Figure 14.4 UE protocol architecture

The design of the protocol stack is guided by the 3GPP specifications. A detailed description about the different protocol layers is described below.

14.3.1 Physical Layer

This section will provide an overview on services and functions provided by the physical layer for UMTS-FDD.

Physical Layer Services to Higher Layers – The physical layer should offer data transport services to the higher layers. The access to these services will be through the use of transport channels via the MAC sublayer. The characteristics of a transport channel will be defined by its transport format (or transport format set), specifying the physical layer processing to be applied to the transport channel, such as error protection scheme, size of CRC, and resulting code ratio after rate matching.

Physical Layer Functions – The physical layer will perform the following main functions: (1) FEC encoding/decoding of transport channels, (2) physical layer measurements and indication to higher layers, for example, received signal quality, channel decoding quality, (3) macro-diversity distribution/combining and soft handover execution, (4) error detection on transport channels, (5) multiplexing of transport channels and de-multiplexing of coded composite transport channels, (6) rate matching, (7) mapping of coded composite transport channels on physical channels, (8) modulation and

spreading/demodulation and de-spreading of physical channels, (9) frequency and time synchronization, (10) closed-loop power control, (11) power weighting and combining of physical channels, and (12) RF processing.

14.3.2 Medium Access Control (MAC)

This section will provide an overview on services and functions provided by the MAC sublayer for UMTS-FDD.

MAC Services to Higher Layers – The transport channels are the interface between the MAC and layer-1, whilst the logical channels are the interface between the MAC and RLC (Figure 14.5). The MAC layer provides data transfer services on logical channels, where a set of logical channel types is defined for different types of data transfer services as offered by the MAC. The type of information to be transferred defines each logical channel type.

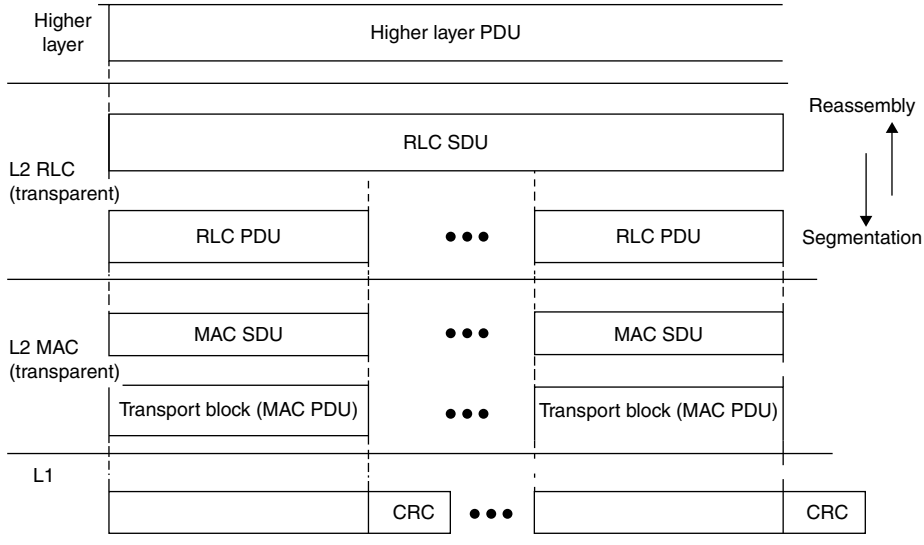


Figure 14.5 Data flow for transparent RLC and MAC

The basic services provided by the MAC layer are as follows.

1. **Data transfer** – This service provides unacknowledged transfer of MAC SDUs between peer-to-peer MAC entities. This service does not provide any data segmentation. Therefore, a higher layer will achieve the segmentation and re-assembly functions.
2. **Reallocation of radio resources and MAC parameters** – This service performs, on request from the RRC, execution of radio resource reallocation, a change of the MAC parameters that includes reconfiguration of MAC functions, a change of UE identity, a change of transport format (combination) sets, and a change of the transport channel type.
3. **Reporting of measurements** – Local measurements are reported to the RRC, that is, traffic volume, quality indication, and MAC status indication.

MAC Functions – The functions of MAC sub layer are as follows.

1. **Mapping between logical channels and transport channels** – The MAC is responsible for mapping of logical channel(s) onto the appropriate transport channel(s).
2. **Selection of the appropriate transport formats for each transport channel depending on instantaneous source rate** – Given the transport format combination set, which is assigned by the RRC, the MAC selects the appropriate transport format within an assigned transport format set for each active transport channel depending on the source rate. The control of transport formats ensures efficient use of the transport channels.
3. **Priority handling between data flows of one UE** – When selecting between the transport format combinations in the given transport format combination set, priorities of the data flows to be mapped onto the corresponding transport channels are taken into account. Priorities are given by attributes of the radio access bearer services and RLC buffer status. The priority handling is achieved by selecting a transport format combination for which high priority data are mapped onto L1 with a "high bit rate" transport format, at the same time letting lower priority data be mapped with a "low bit rate" (which could be a zero bit rate) transport format.
4. **Identification of UEs on common transport channels** – When a particular UE is addressed on a common downlink channel, or when a UE is using the RACH, there will be an inband identification of the UE. As the MAC layer handles the access and multiplexing onto the transport channels, the MAC performs the identification functionality.
5. **Multiplexing/de-multiplexing on common transport channels** – Multiplexing/de-multiplexing of higher layer PDUs into/from transport blocks delivered to/from the physical layer on common transport channels. The MAC supports service multiplexing for common transport channels, as the physical layer does not support multiplexing of these channels.
6. **Multiplexing/de-multiplexing on dedicated transport channels** – Multiplexing/de-multiplexing of the higher layer PDUs into/from transport block sets delivered to/from the physical layer on dedicated transport channels. The MAC allows service multiplexing for dedicated transport channels. This function is utilized when several higher layer services, for example, the RLC, are required to be mapped efficiently on the same transport channel. In this case the identification of multiplexing will be contained in the MAC protocol control information.
7. **Traffic volume monitoring** – The MAC performs measurement of traffic volume on logical channels and reports to RRC. Based on the reported traffic volume information, the RRC performs transport channel switching decisions.
8. **Maintenance of a MAC signaling connection between peer MAC entities** – The MAC supports unacknowledged transfer of MAC-internal messages between peer-to-peer MAC entities.
9. **Dynamic transport channel type switching** – The MAC supports execution of switching between common and dedicated transport channels based on a switching decision derived by RRC.
10. **Ciphering** – This function prevents unauthorized acquisition of data. Ciphering is performed in the MAC layer for the transparent RLC mode.

14.3.3 Radio Link Control (RLC)

This section provides an overview on the services and functions provided by the RLC sublayer for UMTS-FDD. The RLC supports three types of connection: transparent, unacknowledged, and acknowledged (Figure 14.6).

RLC Services – The RLC sublayer provides the following services to the higher layers by a combination of RLC–MAC functions and layer-1 services.

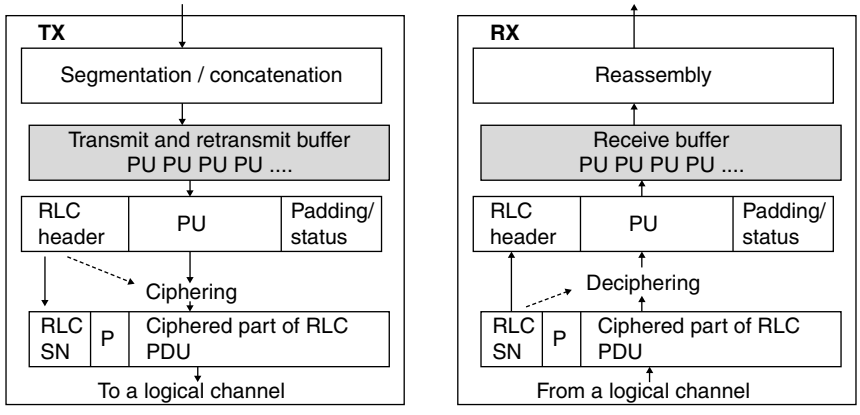


Figure 14.6 RLC in acknowledged mode

1. **RLC connection establishment/release** – This service performs establishment and release RLC connections.
2. **Transparent mode service** – This service transmits higher layer PDUs without adding any protocol information, possibly including segmentation/reassemble functionality.
3. **Unacknowledged mode service** – This service transmits higher layer PDUs without guaranteeing delivery to the peer entity. The unacknowledged data transfer mode has the following characteristics. (a) Detection of erroneous data: the RLC sublayer delivers only those SDUs to the receiving higher layers which are free of transmission errors, by using the sequence number check function. (b) Unique delivery: the RLC sublayer delivers each SDU only once to the receiving higher layer using the duplication detection function. (c) Immediate delivery: the receiving RLC sublayer entity delivers an SDU to the higher layer receiving entity as soon as it arrives at the receiver.
4. **Acknowledged mode service** – This service transmits higher layer PDUs and guarantees delivery to the peer entity. Should the RLC be unable to deliver the data correctly, the user of the RLC at the transmitting side is notified. For this service, both in-sequence and out-of-sequence delivery is supported. The acknowledged data transfer mode has the following characteristics. (a) Error-free delivery: this is ensured by means of retransmission. The receiving RLC entity delivers only error-free SDUs to the higher layer. (b) Unique delivery: the RLC sublayer delivers each SDU only once to the receiving higher layer using duplication detection function. (c) In-sequence delivery: the RLC sublayer provides support for in-order delivery of SDUs, that is, the RLC sublayer delivers SDUs to the receiving higher layer entity in the same order that the transmitting higher layer entity submits to the RLC sublayer. (d) Out-of-sequence delivery: as compared with in-sequence delivery, it is also possible to allow the receiving RLC entity to deliver SDUs to a higher layer in a different order than that submitted to the RLC sublayer at the transmitting side.
5. **QoS setting** – The retransmission protocol should be configurable by layer-3 to provide a different level of QoS. Adjusting the maximum number of retransmissions according to different delay requirements controls this.
6. **Notifications of unrecoverable errors** – The RLC notifies the higher layer of errors that cannot be resolved by the RLC layer itself, by normal exception error handling procedures.

RLC Functions – The RLC sub-layer performs the following functions to deliver the layer-2 services listed in the above section. The functions listed below can be used individually or in combination to deliver the services identified.

1. **Connection control** – This function performs the establishment, release, and maintenance of an RLC connection.
2. **Segmentation and re-assembly** – This function performs segmentation/re-assembly of variable-length higher layer PDUs, into/from smaller RLC payload units (PUs). One RLC PDU carries one PU, except iwhen the header compression is applied, where there are several RLC PUs. The size of the smallest re-transmission unit is determined by the smallest possible bit rate. The RLC PDU size is adjustable to the actual set of transport formats.
3. **Header compression** – The feature to include several payload units into one RLC PDU is referred to as the RLC header compression. RLC header compression should be applied for acknowledged data transfer service. Its applicability is negotiable between UTRAN and UE.
4. **Concatenation** – If the content of an RLC SDU does not fill an integer number of RLC PUs, the first segment of the next RLC SDU is put into the RLC PU in concatenation with the last segment of the previous RLC SDU.
5. **Padding** – When concatenation is not applicable and the remaining data to be transmitted do not fill an entire RLC PDU of given size, the remainder of the data field is filled with padding bits.
6. **Transfer of user data** – This function is used for conveyance of data between users of RLC services. The RLC supports acknowledged, unacknowledged, and transparent data transfer. QoS setting control transfer of user data.
7. **Error correction** – This function provides error correction by re-transmission, for example, Selective Repeat, Go Back N, or a Stop-and-Wait ARQ), in the acknowledged data transfer mode.
8. **In-sequence delivery of higher layer PDUs** – This function preserves the order of higher layer PDUs that were submitted for transfer by RLC using the acknowledged data transfer service. If this function is not used, out-of-sequence delivery will be provided.
9. **Duplicate detection** – This function detects received RLC PDUs that are duplicated and ensures the resultant higher layer PDU is delivered only once to the higher layer.
10. **Flow control** – This function allows an RLC receiver to control the rate at which the peer RLC transmitting entity sends information.
11. **Sequence numbering check (unacknowledged data transfer mode)** – This function guarantees the integrity of reassembled PDUs and provides a mechanism for the detection of corrupted RLC SDUs through checking the sequence number in RLC PDUs when they are reassembled into an RLC SDU. A corrupted RLC SDU will be discarded.
12. **Protocol error detection and recovery** – This function detects and attempts to recover from errors in the operation of the RLC protocol.
13. **Ciphering** – This is discussed later.

14.3.4 Radio Resource Control (RRC)

The following sections provide an overview of the services and functions provided by the RRC sublayer for UMTS-FDD.

RRC Services – The services provided by RRC are as follows.

1. Reception of broadcast information related to non-access stratum and access stratum.
2. Reception of paging broadcast information addressed to UE.
3. Reception of notification broadcast information addressed to UE will include to the point where the UE received the broadcast information, when the access stratum delivers broadcast information to the non-access stratum.
4. Establishment of connection.
5. Transmission of UE capability information.
6. Setup of radio access bearers as required by the quality of service class, using this connection.

7. Transfer of information via the appropriate radio-access bearers.
8. Reconfiguration of radio-access bearers, reconfiguration of transport channel, reconfiguration of physical channel.
9. Soft handover, hard handover, inter-system handover.
10. Measurement and monitoring of control parameters.
11. Release of radio-access bearers.
12. Release of connection.

A radio-access bearer refers to a dedicated link, channel or service between the UE and UTRAN. One example of a radio-access bearer is an SMS service, another example is a speech call.

RRC Functions – The radio resource control (RRC) layer handles the control plane signaling of layer-3 between the UEs and the UTRAN. The RRC performs the following functions.

1. **Reception of broadcast information provided by the non-access stratum (core network)** – The RRC layer handles system information broadcasting from the core network. The system information is normally repeated on a regular basis.
2. **Reception of broadcast information related to the access stratum** – The RRC layer handles system information broadcasting from the UTRAN.
3. **Establishment, maintenance and release of an RRC connection between the UE and UTRAN** – The establishment of an RRC connection is initiated by a request from higher layers at the UE side to establish the first signaling connection for the UE. The establishment of an RRC connection includes an optional cell re-selection, an admission control, and a layer-2 signaling link establishment. The release of an RRC connection can be initiated by a request from higher layers to release the last signaling connection for the UE or by the RRC layer itself in the case of RRC connection failure. The RRC layer detects the loss of an RRC connection, then releases the resources assigned for connection failure.
4. **Establishment, reconfiguration and release of radio-access bearers** – The RRC layer, on request from higher layers, performs the establishment, reconfiguration and release of radio-access bearers in the user plane. It is possible to establish a number of radio-access bearers to the UE at the same time. At establishment and reconfiguration, the RRC layer performs admission control and selects the parameters describing the radio-access bearer processing in layer-2 and layer-1, based on information from higher layers.
5. **Assignment, reconfiguration and release of radio resources for the RRC connection** – The RRC layer handles the assignment of radio resources (code, frequency, time slot, etc.) needed for the RRC connection including the needs from both the control and user plane. The RRC layer has the capability to reconfigure radio resources during an established RRC connection. This function includes coordination of the radio resource allocation between multiple radio bearers related to the same RRC connection.
6. **RRC connection mobility functions** – The RRC layer performs evaluation, decision and execution related to RRC connection mobility during an established RRC connection, such as handover, cell re-selection and cell/paging area update procedures. These functions are based on measurements from the lower layers.
7. **Paging/notification** – The RRC layer handles broadcast paging information from the UTRAN addressed to the UE. The RRC layer also handles paging during an established RRC connection.
8. **Routing of higher layer PDUs** – This function performs routing of higher layer PDUs to the correct higher layer entity.
9. **Control of requested QoS** – This function ensures that the QoS requested for the radio-access bearers can be met. This includes the allocation of a sufficient number of radio resources.
10. **UE measurement reporting and control of the reporting** – The measurements performed by the lower layers are controlled by the RRC layer, in terms of what to measure, when to measure and

how to report. The RRC layer performs the reporting of the measurements from the UE to the UTRAN.

11. **Outer loop power control** – The RRC layer controls setting of the target of the closed loop power control.
12. **Control of ciphering** – The RRC layer provides procedures for setting of ciphering (on/off) between the UE and UTRAN.
13. **Initial cell selection and re-selection in idle mode** – The RRC selects the most suitable cell based on idle mode measurements and cell selection criteria.
14. **Congestion control** – The RRC manages the internal data buffer during information transfer.

The RRC states, when the UE is in connected mode, are shown in Figure 14.7. Once a UE is switched on, it selects a PLMN and looks for a suitable cell in that PLMN. A UE stays in idle mode until it is able to transmit a request to establish an RRC connection. If the RRC connection is released or it fails, the UE moves from connected mode to idle mode. There can only be zero or one RRC connection between UE and UTRAN. If there are multiple signaling connections between UE and CN, they all share the same RRC connection.

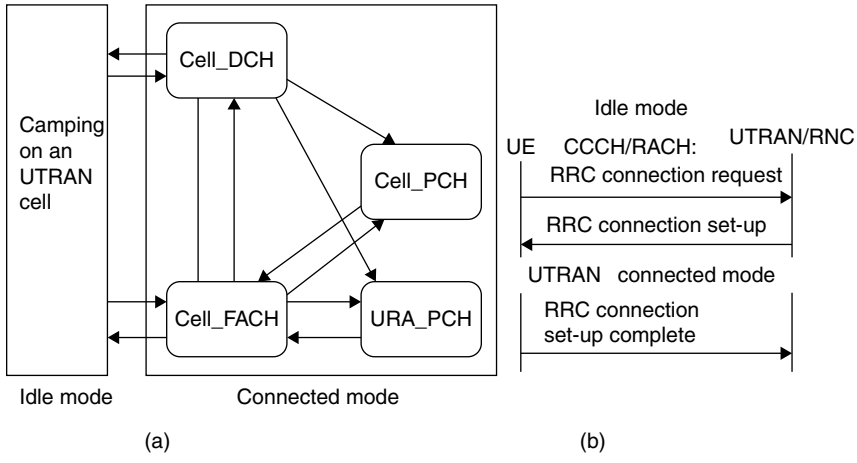


Figure 14.7 (a) RRC states when UE is in connected mode, and (b) RRC connection establishment procedure

In the connected mode, there are different RRC service states: Cell_DCH, Cell_FACH, Cell_PCH and URA_PCH, where each state defines the physical channel that the UE is using. In the Cell_DCH state, a dedicated physical channel is allocated to the UE and the UE is known by its serving RNC on a cell. DSCH can also be used in this state. In the Cell_FACH state, RACH and FACH channels are used by the UE. In this state, the UE performs cell re-selection and after a re-selection, it sends a Cell Update message to the RNC. If the new cell belongs to other system, such as GPRS, the UE enters the idle mode and accesses the other system according to that system's access procedure. In the Cell_PCH, UE is known at the cell level in SRNC but can only be reached via the paging channel. In this state, UEs battery consumption is less than that in the Cell_FACH state. If UE performs cell re-selection, it moves to Cell_FACH, sends a Cell Update message and moves back to the Cell_PCH state if no other activity is triggered during the cell update

procedure. In URA_PCH, after doing a cell re-selection, UE reads the UTRAN registration area (URA) from the broadcast channel and sends a URA update if URA changes.

14.3.5 Packet Data Convergence Protocol (PDCP)

One of the PDCP functions is compression and decompression of protocol control information. Typical VoIP packets are small (for example, 20 bytes – depending upon the codec size). An RTP/UDP/IP header is at least 40 bytes for IPv4 and at least 60 bytes for IPv6. It is not efficient to send packets with full RTP/UDP/IP or TCP/IP headers over the air interface. IP header compression and robust header reduction protocol (ROHC) are some methods for this purpose.

14.3.6 Call Control (CC)

Call control (CC) is one of the protocols in the communication management (CM) sublayer. Every UE supports the call control protocol. If a UE does not support any bearer capability at all, then it responds to a SETUP message with a RELEASE COMPLETE message. In the call control protocol, it is possible to define more than one CC entity. Each CC entity is independent of the other and communicates with the correspondent peer entity using its own MM connection. Different CC entities use different transaction identifiers.

The call control entities are described as communicating finite state machines which exchange messages across the radio interface and communicate internally with other protocol sublayers. Certain sequences of actions of the two peer entities compose "elementary procedures" which are used as a basis for the description in this section. These elementary procedures are grouped into the following classes: (1) call establishment procedures, (2) call clearing procedures, (3) call information phase procedures, and (4) miscellaneous procedures.

The terms "mobile originating" or "mobile originated" (MO) is used to describe a call initiated by the UE. The terms "mobile terminating" or "mobile terminated" (MT) are used to describe a call initiated by the UTRAN.

Call Establishment Procedures – Establishment of a call is initiated by the request of a higher layer in either the UE or UTRAN. It consists of: (1) the establishment of a CC connection between the UE and UTRAN, and (2) the activation of the codec or inter-working function. The UE supports the following types of call establishment. – (a) Mobile originating call establishment – There are two types of mobile originating call, a basic call and an emergency call. The request to establish an MM connection contains a parameter to specify whether the call is a basic or an emergency call. (b) Mobile terminating call establishment – It is possible to terminate a call from a UE, provided that an MM connection is already established by the UTRAN.

14.3.7 Mobility Management (MM)

The main function of the mobility management sublayer is to support the mobility of UEs, such as informing the UTRAN of its present location and providing user identity confidentiality. Other functions of the MM sublayer are to provide connection management services to the different entities of the higher connection management (CM) sublayer. There are two sets of procedures defined for the MM: (1) MM procedures for non-GPRS services, performed by the MM entity of the MM sublayer, and (2) GMM procedures for GPRS services, performed by the GMM entity and GMM-AA entity of the MM sublayer.

Depending on how they are initiated, there are three types of MM procedures.

1. **Common procedures** – It is always possible to initiate an MM common procedure. Procedures that belong to this category are: TMSI re-allocation, authentication, identification, IMSI detach, MM information procedure.
2. **Specific procedures** – It is possible to initiate an MM specific procedure only if no other MM specific procedure is running, or no MM connection exists. The procedures belonging to this type are: normal location updating, periodic updating, IMSI attach procedure.
3. **MM Connection management procedures** – These procedures are used to establish, maintain and release an MM connection between the UE and the UTRAN, over which an entity of the higher CM layer can exchange information with its peer. It is possible to perform an MM connection establishment only if no MM specific procedure is running. It is possible for a multiple MM connection to be active at the same time.

There are two types of GMM procedures for GPRS services.

1. **GMM common procedures** – There are four types of GMM common procedures as follows: P-TMSI re-allocation, GPRS authentication and ciphering, GPRS identification, GPRS information.
2. **GMM specific procedures** – Two types of GMM specific procedures are supported in the UE in the GMM context. One is initiated by UE and the other is initiated by UTRAN.

14.3.8 Session Management (SM)

The session management (SM) provides management services to the GPRS point-to-point data services at the UE radio interface. The SM supports PDP context handling of the UE. The SM procedures for identified access are performed only if a GMM context has been established between UE and UTRAN. For anonymous access the SM procedures are performed without a GMM context being established. The SM procedures are as follows.

1. **PDP context activation** – This procedure is used to establish a PDP context between UE and UTRAN for specific QoS on a specific NSAPI. The PDP context is initiated by UE, or upon request, by the network.
2. **PDP context modification** – This procedure is used to change the QoS negotiated during the PDP context activation procedure or at a previously performed PDP context modification procedure. The network initiates the procedure at any time when a PDP context is active.
3. **PDP context deactivation** – This procedure is used to deactivate any existing PDP context between the UE and the network. The context deactivation is initiated by the UE or network.
4. **Anonymous PDP context activation** – This procedure is used to anonymously establish a PDP context between the UE and network for a specific QoS on a specific NSAPI. The procedure is initiated by UE only.
5. **Anonymous PDP context deactivation** – This procedure is used to deactivate any anonymous PDP context that exists between the UE and the network. The context deactivation is initiated by the UE or network.
6. **Broadcast/multicast control (BMC)** – In the UMTS system, this protocol adapts broadcast and multicast services on the radio interface. The infrastructure offers an option to use an uplink channel for interaction between the service and the user, which is not a straightforward issue in usual broadcast networks.
7. **Packet data convergence protocol (PDCP)** – This is responsible for IP header compression and decompression, transfer of user data and maintenance of sequence numbers for radio bearers which are configured for lossless SRNS relocation.

14.3.9 Universal Subscriber Identity Module (USIM) Interface

The USIM interface provides the transmission protocol for retrieving information elements that are stored in the USIM for 3GPP network operations. The transmission protocol is in accordance with ISO/IEC 7816-3 standards. The USIM interface retrieves the following USIM related information upon request from the UE: (a) administration information – mode of operation of USIM, for example, normal, type approval; (b) USIM service table – optional services provided by the USIM; (c) IMUI; (d) language indication; (e) location information; (f) cipher key, K_c, and cipher key sequence number; (g) access control class(es); (h) forbidden PLMN; (i) phase identification; (j) ciphering key for GPRS; (k) GPRS location information; (l) cell broadcast related information; (m) emergency call codes; (n) capability and related parameters; (o) HPLMN search period; (p) BCCH information – list of carrier frequencies to be used for cell selection; and (q) phone numbers – abbreviated dialing number, fixed dialing number.
In addition the USIM interface, via directions from the UE, provides the functions to manage and provide storage for the following information: PIN, PIN enabled/disabled indicator, PIN error counter, unblocked PIN, unblocked PIN error counter, data integrity keys, and subscriber authentication keys.

14.3.10 Man Machine Interface (MMI)

The MMI interfaces with the user and provides user procedures for call control, physical input and output, such as indications and displayed information. The MMI is positioned above the protocol stack and has interfaces with the keypad, display, and USIM. For all the features mentioned, the MMI uses the services of the protocol stack, keypad drivers, and LCD drivers. The following features are supported by the MMI: (1) display of called number; (2) indication of call progress signals; (3) country/PLMN indication; (4) country/PLMN selection; (5) basic key pad entry – physical means of entry of 0–9, +, * and #; (6) service indicator; (7) call control – SEND and END function keys for call initiation and termination, respectively; (8) call acceptance – the call is accepted when the user presses the SEND-function key; (9) off-hook call initiation; and (10) call termination.

14.3.11 Inter-Working Unit (IWU)

The IWU interfaces with the MMI and application on the one side, and the CM and MM sublayers on the other side. The IWU is defined as a type of formal adapter between the application and the top of the protocol stack or in short a driving engine. The IWU consists of several inter-working functions (IWFs), which are invoked by the IWU as a result of service requirements for inter-working.

14.4 Procedures in the UE

This section will describe the UE procedures in different operating modes during the operation of the UE. When the UE is in an inactive state or in a mode when it does not expect to make or receive any communication with the network, it makes no attempt to gain access with a UTRAN. When the UE is active, it attempts to camp-on to any existing UTRAN (mandatory). The UE camps-on to a UTRAN that is selected either manually or automatically (UTRAN selection). Once the UE recognizes the UTRAN, it looks for a suitable cell in that UTRAN and tunes to it (cell selection). On camped on, the UE enters the idle mode.

14.4.1 Procedures in Idle Mode

The UE in idle mode performs location registration (mandatory). The CM asks the MM of the UE to initiate a location registration procedure. During the location registration procedure, information related

to the location of the UE is transferred to the UTRAN. The first time location registration of UE, and subsequent location registration without changing the registered area, is referred to as "attach" (attaching and detaching is to inform the UTRAN that the UE is ready to receive incoming calls). So the attach procedure is the same as the location registration, which is used to indicate to the UTRAN that the UE is active.

After successful location registration, the UE in idle mode is able to initiate and receive calls. It should be noted that the UE always tries to camp to the best available cell in that UTRAN or other UTRAN (cell re-selection). The location registration procedure is initiated compulsorily whenever the UE moves to a new registration (location) area, that is, new UTRAN (UTRAN re-selection). Prior to the power off, the UE runs a detaching procedure (the reverse of attach), if necessary.

If the UE fails to find a suitable cell to camp on, the USIM is not inserted, or if the location registration is unsuccessful, it attempts to camp on to any cell, irrespective of the UTRAN. In this situation, the UE has limited services (only emergency calls).

When camped on, attached to a UTRAN and in the idle mode, the UE monitors the system information and paging message from the UTRAN. When the registered UE receives a paging message from UTRAN, it indicates that there is call (usually considered as a UTRAN initiated call or incoming call) for that UE. The UE responds to this paging message by setting up an RRC connection, and enters connected mode when the RRC connection is established. The idle mode tasks are sub-divided into three procedures: UTRAN selection and re-selection, cell selection and re-selection, and location registration.

14.4.2 UTRAN Selection and Re-selection

The non-access stratum selects a suitable UTRAN. Normally the UE operates on its home UTRAN (HUTRAN). However, a visited UTRAN (VUTRAN) is selected if the UE loses coverage with its HUTRAN. There are two modes for UTRAN selection:

1. **Automatic mode** – This mode utilizes a list of UTRANs in priority order. The highest priority UTRAN that is available and allowable will be selected.
2. **Manual mode** – Here the UE indicates which UTRANs are available to the user. Only when the user makes a manual selection will the UE try to obtain normal service on the UTRAN.

In the automatic mode, the UE always look for more suitable UTRAN and camps onto them if a more suitable candidate is found, particularly on a home UTRAN. This is referred to as UTRAN-re-selection.

14.4.3 Cell Selection and Re-selection

The UE selects the most suitable cell and the radio access mode based on idle mode measurements and cell selection criteria. The non-access stratum controls the cell selection, for instance in terms of a list of forbidden registration area(s) and a list of non-access stratum defined service area(s) in priority order.

When camped on a cell, the UE always searches for a better cell according to the cell re-selection criteria. When a more suitable cell is found, the UE connects to the UTRAN via that cell. The non-access stratum is informed if the cell selection and re-selection results in changes in the received system information. For normal service, the UE camps onto a suitable cell, and receives the following information from the UTRAN via the control channel: (1) receive registration area information from the UTRAN, for example, location area and routing area, (2) identify the non-access stratum defined service areas(s) to which the serving cell belongs, and (3) other access-stratum and non-access stratum information.

If registered, it receives paging and notification messages from the UTRAN, and initiates call setup for outgoing calls or other actions from the UE.

14.4.4 Location Registration

When first camped onto a suitable cell after power on, the non-access stratum registers the UE as active and presents in the registration area of the chosen cell, if necessary. The non-access stratum registers the UEs presence in a registration area, for instance regularly and when entering a new registration area. Prior to power off, the non-access stratum de-registers the UE, if necessary.

14.4.5 Procedures in Connected Mode

For UE initiated calls, the higher layers of the UE send a request to the RRC to establish a signaling connection establishment. Upon receiving this, the RRC of the UE initiates a connection establishment procedure with the peer RRC of the UTRAN by sending an RRC connection request message through the MAC. The UE enters connected mode once the RRC connection is established (when the UE-RRC receives the RRC connection setup message). During this connection establishment procedure, the RRC configures the layers L1 and L2 using the parameters received from the UTRAN. Once the configuration is complete the RRC of the UE initiate the RLC signaling link. It should be noted that once the RRC connection establishment is successful, the signaling connection establishment will be resumed. Once the signaling connection is established there will be higher layer peer-to-peer signaling data transfer.

The UE enters connected mode once the RRC connection is established. Within the connected mode the level of UE connection to UTRAN will be determined by the QoS requirements of the active radio access bearers and the characteristics of the traffic on those bearers.

The main procedures in connected mode are as follows.

1. **Radio access bearer establishment** – This procedure is used for establishing a new radio access bearer depending on the QoS parameters, assignment of RLC parameters, multiplexing priority for the DTCH, scheduling priority for the DCH, transport format set (TFS) for the DCH and updates of transport format combination set (TFCS). (Note that the TFS is defined as the set of transport formats associated with a transport channel and is the format of communication between the MAC and L1 entities.)
2. **Radio access bearer release** – This procedure is used to release a radio access bearer by releasing the RLC entity, release of the DCH which affects the TFCS, release of physical channel(s) and change of the used transport channels types and RRC state.
3. **Radio access bearer and signaling link configuration** – This procedure is used to reconfigure parameters for a radio access bearer or the signaling link to reflect the changes in QoS. This procedure has the option of reconfiguring either the synchronized or unsynchronized radio access bearer.
4. **Transport channel reconfiguration** – This procedure is used to configure the parameters related to a transport channel such as the TFS. This procedure also assigns TFCS and changes physical channel parameters to reflect a reconfiguration of a transport channel in use.
5. **Transport format combination control** – This procedure is used to control which transport format combinations (within the transport format combination set) will be used in the UE in the uplink as requested by the network.
6. **Physical channel reconfiguration** – This procedure is used to assign, or replace a set of physical channels used by UE.

7. **Data transmission** – This procedure is used for controlling data transmission. The procedure caters for two types of data transmission: acknowledged-mode data transmission in DCH/DCH + DSCH, and acknowledged-mode data transmission in CPCH/FACH.

14.5 Mobility Procedures in Connected Mode

The mobility related procedures play a significant role in ensuring the smooth data transmission in lieu of the mobile nature of the UE in the connected mode.

URA Update – This procedure is used by the UE to inform UTRAN that the UE has switched to a new UTRAN registration area (URA). This procedure is triggered after change of cell and after the UE has read information broadcast by the UTRAN indicating the change of URA.

Cell Update – This procedure is used by the UE to inform the UTRAN that the UE has switched to the new cell. This procedure is considered as a forward handover procedure and is triggered after the change of cell and after the UE has read information broadcast by UTRAN. This procedure is also triggered by expiry of a cell update periodicity timer in the UE.

Handover Measurement Reporting – This procedure is used by the UE for reporting the measurement results to UTRAN for its use in handover measurements. The procedure caters for measurement of the evaluation of radio link quality (measurement), intra-frequency measurements, inter-frequency measurements, inter-system measurements, and traffic volume measurements as requested by UTRAN.

Soft Handover Procedures – This procedure is used by UE during the soft handover in FDD mode. The UE supports three sets of procedures for soft handover which are:

1. **Radio link addition** – The UE uses this procedure to configure the layer-1 to begin reception on a new radio link upon request by the RCC of the UTRAN.
2. **Radio link removal** – The UE uses this procedure to configure the layer-1 to terminate and remove the radio link upon request by the RCC of the UTRAN.
3. **Combined radio link addition and removal** – The UE uses this procedure to configure the layer-1 for the replacement of the radio link. The UE terminates and removes the old radio link and starts a reception on a new radio link upon request by the RRC of the UTRAN.

Hard Handover Procedures for FDD Mode – This handover procedure is used for inter-frequency hard handover in the FDD mode. The UE, upon reception of the handover message from network, configures the L1 of the UE with the radio resources as provided by the network, terminates the old radio link and starts reception on the new radio link. Upon the L1 of the UE achieving downlink synchronization on the new frequency, an L2 link is established and the RRC of the UE sends a handover completion message to the RRC of the UTRAN.

14.6 Other Procedures during Connected Mode

In addition to the above procedures the UE supports the following procedures in the connected mode:

Transmission of UE Capability Information Procedure – The UE transfers its capability information to the network. This procedure is performed either after RRC connection setup procedure or during the lifetime of the RRC connection if the UE capability changes (for example, due to a change in UE power class).

System Information Procedure – The UE is capable of receiving the system information in order to update the neighboring cell and MM information. The RRC of the UEUE forwards the received MM information to the MM sublayer of the UE.

Direct Transfer Procedure – Upon reception of this message, the higher layer PDU is routed by using the CN domain identifier parameter set in the UE to indicate the destination CN node of the non-access stratum message.

RRC Status Procedure – If the UE is signaling connection to two core networks CN1 and CN2, a request of RRC release from one of the nodes will be received by the RRC of the UE from the RRC of the UTRAN. After receiving this message the UE informs it's corresponding MM entity of RRC connection release and sends an acknowledgement back to UTRAN.

RRC Connection Re-Establishment Procedure – This procedure is used when a UE loses radio connection due to radio link failure or other. After selecting a new cell, the RRC of the UE sends to the RRC of the UTRAN an RRC connection re-establishment message. The required acknowledgment message is sent by the RRC of the UE upon completion of the connection re-establishment procedure.

14.7 Security Procedures

First generation analog cellular phones, generally did not contain much in terms of security aspects and protection. Thus, it was possible to eavesdrop on the analog radio path and thereby listen to other user's calls, or to program the identities of the mobile phones such that the accessing cost appears in other user's bill. Moving to the second generation GSM system, some security aspects were introduced, such as subscriber authentication, user confidentiality, identity, and confidentiality of voice and data are provided. Its goal was to provide user related security features for authentication, confidentiality and anonymity and to protect the network against un-authorized access. However, there are some weaknesses in the GSM mobile security system, which made it vulnerable to security attacks. Whereas in the third generation system (3G/UMTS), to overcome these drawbacks some security functions have been added and some existing ones have been improved over GSM security functions by introducing stronger encryption algorithms, encryption in the base station (BTS) to radio network controller (RNC) transmission path, stricter authentication algorithms and tighter subscriber confidentiality. It follows three principles: (1) keep the proven GSM security features to ensure the compatibility for inter-working and handover; (2) improve the weaknesses of GSM security; and (3) add security features for new 3G radio access networks and services.

14.7.1 UMTS Security Overview

The following is a brief description of the UMTS security mechanism.

14.7.1.1 Authentication and Key Agreement (AKA) Mechanism

The protocol machinery of the UMTS authentication and key agreement (AKA) scheme is similar to GSM AKA, but one added feature is the mutual authentication. AKA takes place between USIM and HLR/AuC (Figure 14.8). In practice SN is the counterpart to the USIM on behalf of HLR/AuC. In some cases, SN has enough information to perform the authentication without involving the HLR/AuC. These cases include: authentication based on knowledge of a previously derived cipher/integrity key pair, and an authentication by using an authentication vector, which was previously transferred from HLR/AuC to the VLR/SGSN in the SN. As with GSM, here also when any new user is added to a home network for the first time, a subscriber authentication key (K) is assigned in addition to IMSI to enable the verification of subscriber identity. The key is stored in AuC at the network side and at the subscriber side in the USIM. In addition, USIM and HE keep track of counters SQN (sequence number) (MS) and SQN (HE), respectively, to support network authentication. The sequence number SQN (HE) is an

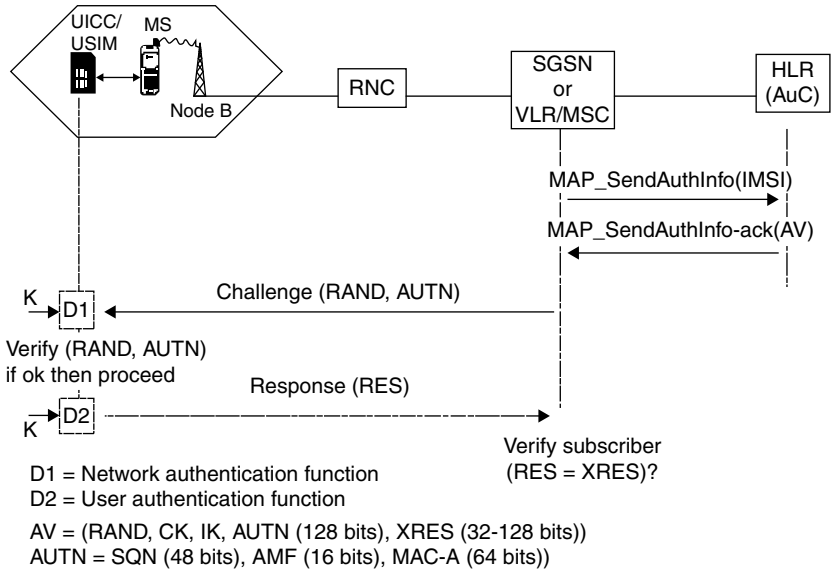


Figure 14.8 UMTS authentication and key agreement

individual counter for each user and the sequence number SQN (MS) denotes the highest sequence number the USIM has accepted. The purpose of this procedure is to authenticate the user and establish a new pair of cipher and integrity keys between VLR/SGSN and the USIM. The steps involved in this process are described below.

VLR (in the case of CS) or SGSN (in case of PS) sends an authentication request to HLR/AuC, which includes IMSI to indicate the user's identity as each key (K) is related to the IMSI. Upon the receipt of the authentication request from VLR/SGSN, HLR/Auc generates an authentication vector (AV). HLR/AuC may have pre-computed the required number of authentication vectors and retrieve them from the HLR database or may compute these on demand. The authentication vectors are ordered according to the sequence number. For production of AV, five functions are required; these are $f1$, $f2$, $f3$, $f4$, and $f5$. The inputs to $f1$ are K, SQN and the authentication management field (AMF) parameter and output from it is MAC (message authentication code computed). HLR/AuC generates a fresh sequence number SQN (HE) and an unpredictable challenge RAND. For each user (this means for each IMSI, that is, each K) HLR/AuC keeps track of the counter SQN (HE). The SQN is very important because it ensures that the authentication vector used for ciphering the authentication process is unique and has not been used before. The input to $f2$, $f3$, $f4$, and $f5$ are RAND and K and output from them are XRES, ciphering key (CK), integrity key (IK), and acknowledgement key (AK), respectively. Each authentication vector (the equivalent of GSM "triplet") consists of the following components: a random number (RAND), an expected response (XRES), a cipher key (CK), an integrity key (IK) and an authentication token (AUTN). Each authentication vector (AV) is good for one authentication and key agreement between VLR/SGSN and the USIM. Then the HRL/AuC sends an authentication response back to the VLR/SGSN that contains an ordered array of n authentication vectors AV (1. . .n). VLR/SGSN stores the AV. When the VLR/SGSN initiates an authentication and key agreement, it selects the next unused authentication vector from the ordered array in the database (authentication vectors in a particular node are used on a FIFO basis) and sends the RAND and AUTN parameters to

the USIM. Upon receipt of this, USIM first computes the anonymity key AK and retrieves the sequence number SQN then it computes the XMAC and compares this with MAC, which is included in the AUTN. If they are different the user sends authentication reject back to VLR/SGSN with the failure cause indication and the user abandons the procedure. However, if these are same, network authentication is successful and USIM next verifies the received SQN and checks whether it is in the correct range or not. If the USIM considers the SQN to be not in the correct range, it sends synchronization failure back to the VLR/SGSN with cause and abandons the procedure. If this is successful, then the USIM produces CK, IK and a response RES, which is sent back to the VLR/SGSN. VLR/SGSN (also known as the serving network SN) compares the expected response (XRES), which was produced by AuC with the received response (RES). If these match then VLR/SGSN considers the AKA is successfully completed.

14.7.1.2 Authentication and Key Agreement Procedure

The detailed operational flow of authentication and key agreement procedure in UMTS is depicted in Figure 14.9.

14.7.1.3 Security Mode Command

The purpose of this procedure is to trigger the start of ciphering or to command the restart of ciphering with new ciphering configuration for the radio bearer of one CN domain and for all signaling radio bearers. It is also used to start integrity protection or to modify integrity protection configuration for all signaling radio bearers. Initially, during the RRC connection establishment procedure, UE transfers to RNC the UE security capabilities [which includes ciphering capabilities (UEAs) and the integrity capabilities (UIAs)] and START values for the CS service domain and the respective PS service domain. The UE sends the initial L3 message to the VLR/SGSN. This contains user identity and the key set identifier (KSI), which is allocated by the CS service domain or PS service domain during the last AKA procedure for this CN domain. The VLR/SGSN initiates integrity and ciphering by sending the RANAP message "security mode command" to SRNC, which contains a list of allowed UIAs and the IK to be used. SRNC decides which algorithms to use by selecting from the list of allowed algorithms (UIA1, Kasumi, . . .) and the list of algorithms supported by the UE. The SRNC generates the random value FRESH and initiates the downlink integrity protection by sending a "security mode command" to the UE. However, if the requirement received from VLR/SGSN cannot be fulfilled, then it sends a "security mode reject" to the requesting VLR/SGSN. If the requirements are met then SRNC sends an RRC message "security mode command" to UE. The message includes the UE security capability, the UIA and FRESH to be used and if ciphering should be started then it also contains UEA. Some addition information, such as when the ciphering will be started, is also included in the message. Because of this the UE can have two integrity and ciphering keys. The network must indicate which key set should be used. This is obtained by including a CN type indicator in the security mode command message. If the GSM MS classmark exists, then the message also contains that. Before sending this message to the UE, SRNC first generates the MAC-I (message authentication code for integrity) and attaches this information to the message for integrity protection.

At reception of the security mode command message, UE checks that the UE security capabilities received is equal to the UE security capability transferred in the connection setup message. The same applies to the GSM MS classmark. The UE computes XMAC-I on this message by using the indicated UIA, stored COUNT-I and received FRESH parameter. The MS verifies the integrity of the message by comparing the received MAC-I with the generated XMAC-I. If all controls are successful, the UE

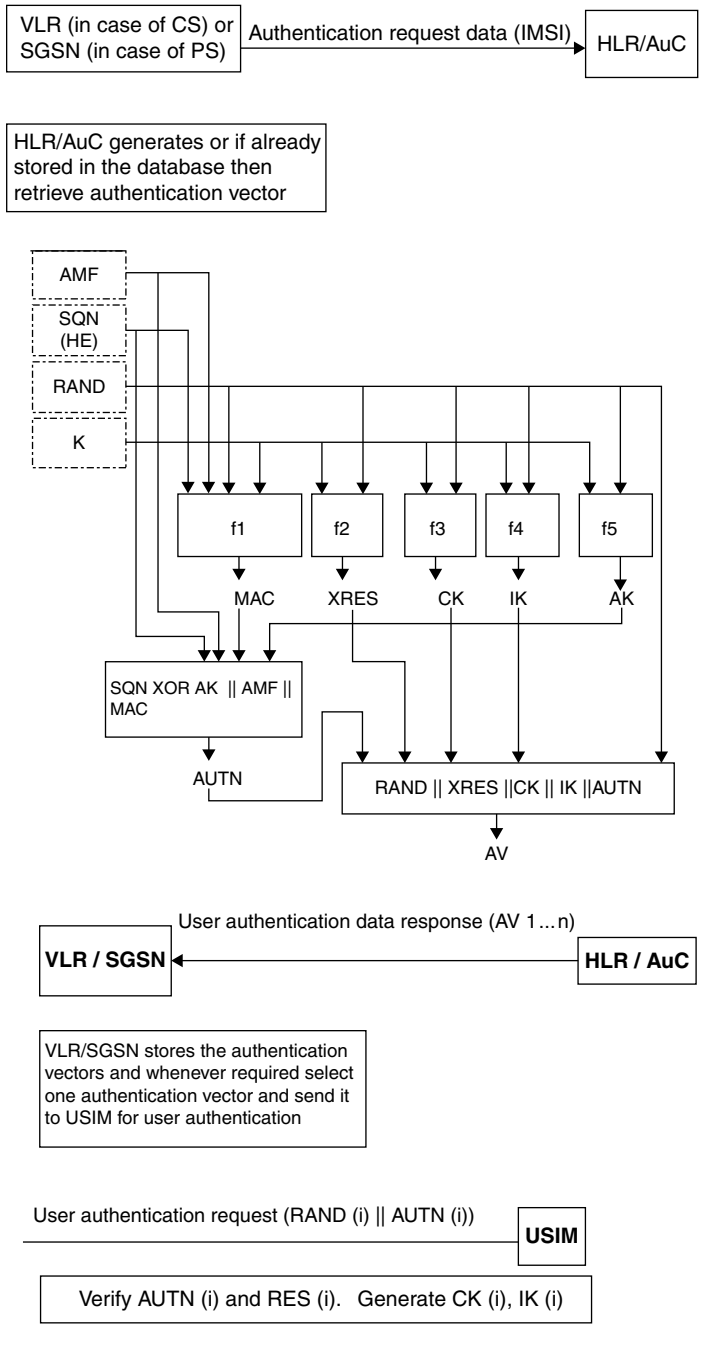


Figure 14.9 Authentication and key agreement procedure

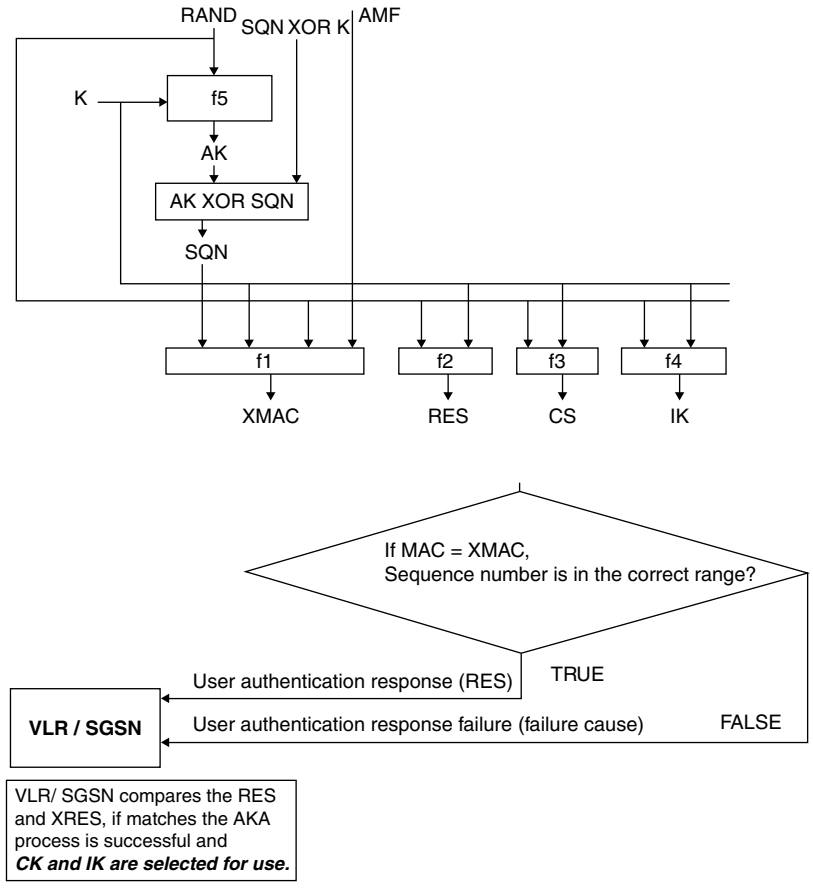


Figure 14.9 *(Continued)*

compiles the RRC message security mode complete and generates the MAC-I for this message. If any control is not successful, the procedure ends in the UE. When SRNC receives the "security mode complete" message, the SRNC computes the XMAC-I on the message. The SRNC verifies the data integrity of the message by comparing the received MAC-I with the generated XMAC-I. Then RANAP message "security mode complete" is sent from SRNC to VLR/SGSN with the complete response and selected algorithms. The procedure ends here.

14.7.2 Integrity Protection

The security mode command starts the downlink integrity protection, that is, this message and after this all other (except-) messages sent to UE are integrity protected using the new integrity configuration. The signaling radio bearers are used for transfer of signaling data for services delivered by both CS and PS service domains. These signaling RBs are data integrity protected by the IK of the service domain for which the most recent security mode negotiation took place. This may require that the integrity key of an ongoing signaling (already integrity protected) connection has to be changed, when a new connection is established with another service domain, or when a security mode negotiation follows a re-authentication

during an ongoing connection. Figure 14.10 illustrates the use of integrity algorithm f9 to authenticate the data integrity of a signaling message.

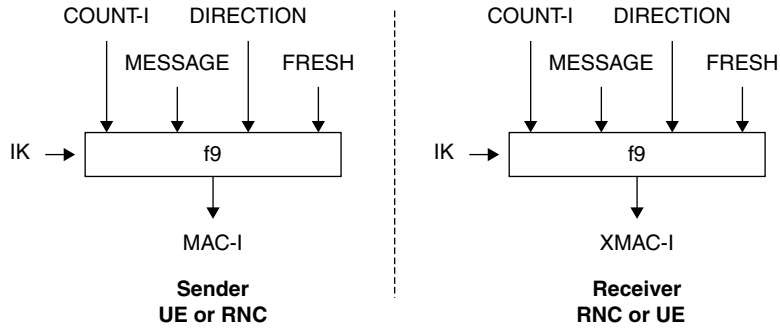


Figure 14.10 Derivation of MAC-I (or XMAC-I) on a signaling message

Input parameters to this algorithm are integrity key (IK), the integrity sequence number (COUNT-I), a random value generated by the network side (FRESH), the direction bit DIRECTION, and signaling data message MESSAGE. Based on these input parameters the user computes message authentication code for data integrity MAC-I using integrity algorithm f9. This MAC-I is appended to the security mode complete message from the UE to UTRAN. The receiver computes the XMAC-I on the message received in the same way and verifies the data integrity of the message by comparing MAC-I and XMAC-I.

Details of the input parameter to the integrity algorithm are as follows.

COUNT-I: The integrity sequence number COUNT-I is 32-bits long. For signaling RB (RB 0–4) there is one COUNT-I value per uplink signaling radio bearer and one COUNT-I value per downlink signaling radio bearer. It is composed of two parts, a short sequence number "short," a long sequence number "long." The "short" sequence number forms the least significant bits of COUNT-I, while "long" sequence number forms the most significant bits of COUNT-I. The short sequence number is the 4-bit RRC sequence number (RRC SN) that is available in each RRC PDU. The "long" sequence number is the 28-bit RRC hyperframe number (RRC HFN), which is incremented at each RRC SN cycle. The RRC HFN is initialized by means of the parameter START. The UE and RNC initialize the 20 most significant bits of the RRC HFN to START, and the remaining bits of RRC HFN are initialized to 0.

Integrity Key: The integrity key is 128-bits long. There may be one IK (CS) for CS service domain and the user and one IK (PS) for PS service domain and the user. For UMTS subscribers the IK is established during the AKA process and is stored in the USIM in the user side and in AuC/HLR in the network side. For GSM subscribers that access the UTRAN, IK is established following GSM AKA and is derived from the GSM cipher key K_c. Generally IK is stored in the USIM and also its copy is stored in the UE. IK is sent from USIM to UE upon request from UE. The USIM sends IK under the condition that a valid IK is available. The UE triggers a new authentication procedure if the current value of START (cs), or START (ps), in the USIM are not up-to-date or START values have reached THRESHOLD. The UE deletes IK from its memory after power off and after removal of USIM. IK is sent from the HLR/AuC to VLR/SGSN as a part of a quintet. It is sent from VLR/SGSN to RNC in the RANAP security mode command. At handover, the IK is transmitted within the network infrastructure from the old RNC to the new RNC, to enable the communication to proceed, and the synchronization procedure is resumed. The IK remains unchanged at handover.

FRESH: This is 32-bits long and is sent by the network. There is one FRESH parameter value per user. The network parameter FRESH protects the network against replay of the signaling message by the user. At the connection setup the RNC generates a random value FRESH and sends it to the user in RRC security mode command. The value FRESH is subsequently used both by the network and the user throughout the duration of a single connection. At handover with relocation of SRNC, the new SRNC generates its own values for FRESH parameter and sends it to the UE in the RRC message that indicates a new UTRAN radio network temporary identity due to SRNC relocation.

DIRECTION: The direction identifier is one bit long. The direction identifier is input to avoid the integrity algorithm used to compute the message authentication codesfrom using an identical set of input parameter values for the up-link messages.

MESSAGE: This is the signaling message itself with the radio bearer identity. The latter is appended in front of the message. Note the RB identity is not transmitted with the message but it is needed to avoid the same set of input parameter being used for different instances of message authentication codes.

The security mode complete message from UE starts the uplink integrity protection.

14.7.3 Ciphering

The ciphering function is performed either in the RLC sublayer or in the MAC sublayer, according to the following rules.

If an RB is using a non-transparent RLC mode (AM or UM), ciphering is performed in the RLC sublayer. If an RB is using the transparent RLC mode, ciphering is performed in the MAC sublayer (MAC-d entity). Ciphering is performed in the UE and SRNC and the context needed for ciphering is only known to S-RNC and ME.

Figure 14.11 illustrates the use of ciphering algorithm f8 to encrypt plaintext by applying a key stream using a bit per bit binary addition of the plaintext and key stream. The plaintext may be recovered by generating the same key stream using the same input parameters and applying a bit per bit binary addition with the cipher text.

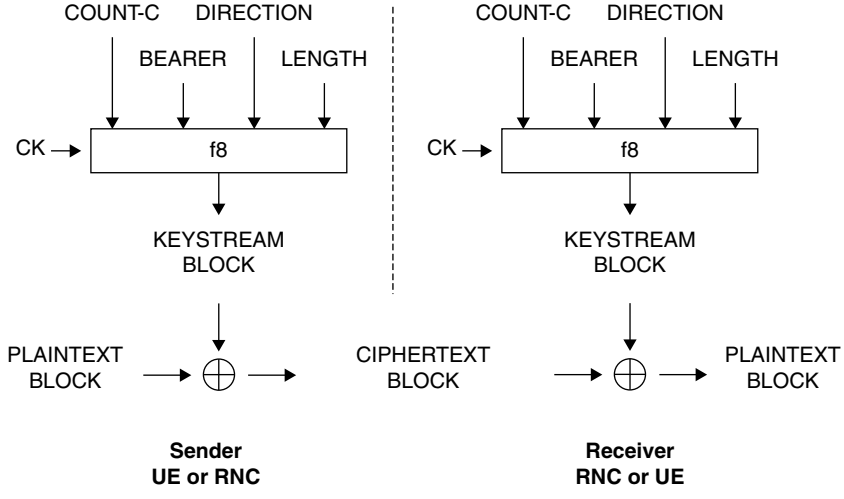


Figure 14.11 Ciphering of user and signaling data transmitted over the radio access link

The input parameters to this algorithm are the cipher key CK, a time dependent input COUNT-C, the bearer identity BEARER, the direction of transmission DIRECTION and length of keystream required LENGTH. Based on these input parameters the algorithm generates the output keystream block KEYSTREAM, which is used to encrypt the input plaintext block PLAINTEXT to produce the output ciphertext block CIPHERTEXT. The input parameter LENGTH will only affect the length of the KEYSTREAM BLOCK, not the actual bits in it.

The details of the input parameters to the cipher algorithm are as follows.

COUNT-C – There is one COUNT-C for uplink RB and one for downlink RB using RLC AM or RLC UM (Figure 14.12). There is one uplink and one downlink COUNT-C value for all radio bearers using transparent RLC mode that are connected to the same CN domain (and mapped onto DCH). The COUNT-C is composed of two parts: a "short" sequence number and a "long" sequence number. The "short" sequence number forms the least significant bits of COUNT-C while "long" sequence number forms the most significant bits of the COUNT-C. The update of COUNT-C depends on the transmission mode as described below. (a) For RLC TM on DCH, the "short" sequence number is the 8-bit connection frame number CFN of COUNT-C. It is independently maintained in the UE MAC-d entity and the SRNC MAC-d entity. The "long" sequence number is the 24-bit MAC-d HFN, which is incremented at each CFN cycle. (b) For RLC UM mode, the "short" sequence number is the 7-bit RLC sequence number (RLC SN) and this is part of the RLC UM PDU header. The "long" sequence number is the 25-bit RLC UM HFN, which is incremented at each RLC SN cycle. (c) For RLC AM mode, the "short" sequence number is the 12-bit RLC sequence number (RLC SN) and this is part of the RLC AM PDU header. The "long" sequence number is the 20-bit RLC AM HFN, which is incremented at each RLC SN cycle.

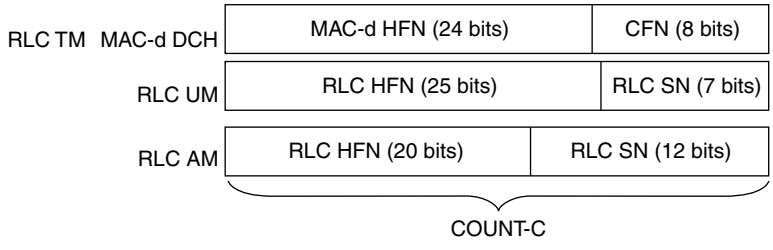


Figure 14.12 The structure of COUNT-C for all transmission modes

The hyperframe number (HFN) is initialized by means of the parameter START. The ME and the RNC then initialize the 20 most significant bits of the RLC AM HFN, RLC UM HFN, and MAC-d HFN to START. The remaining bits of the RLC AM HFN, RLC UM HFN and MAC-d HFN are initialized to zero.

When a new radio bearer is created during an RRC connection in ciphered mode, the HFN is initialized by the current START value.

Ciphered Key – The cipher key CK is 128-bits long. There may be one CK for CS connections (CK_{CS}), established between the CS service domain and the user and one CK for PS connections (CK_{PS}) established between the PS service domain and the user. For UMTS subscribers, CK is established during UMTS AKA, as the output of the cipher key derivation function *f*3, available in the USIM and in HLR/AuC. For GSM subscribers that access the UTRAN, CK is established following GSM AKA and is derived from the GSM cipher key K_c. CK is stored in the USIM and a copy is stored in the ME. CK is sent from the USIM to the ME upon request of the ME. The USIM sends CK under the condition that a valid CK is available. The ME triggers a new authentication procedure if the current values of $START_{CS}$ or $START_{PS}$ in the USIM have reached THRESHOLD. The ME deletes CK from the memory after

power-off and after removal of the USIM. CK is sent from the HLR/AuC to the VLR/SGSN and stored in the VLR/SGSN as part of the quintet. It is sent from the VLR/SGSN to the RNC in the (RANAP) security mode command.

At handover, the CK is transmitted within the network infrastructure from the old RNC to the new RNC, to enable the communication to proceed. The cipher CK remains unchanged at handover. There is one CK for CS connections (CK_{CS}), established between the CS service domain and the user and one CK for PS connections (CK_{PS}) established between the PS service domain and the user. The radio bearers for CS user data are ciphered with CK_{CS}. The radio bearers for PS user data are ciphered with CK_{PS}.

The signaling radio bearers are used for transfer of signaling data for services delivered by both CS and PS service domains. These signaling radio bearers are ciphered by the CK of the service domain for which the most recent security mode negotiation took place. This may require that the cipher key of an (already ciphered) ongoing signaling connection has to be changed, when a new connection is established with another service domain, or when a security mode negotiation follows a re-authentication during an ongoing connection. This change should be completed within five seconds after the security mode negotiation.

Each UEA will be assigned a 4-bit identifier. Currently the following values have been defined: "00002:" UEA0, no encryption, "00012":UEA1, Kasumi, the remaining values are not defined.

BEARER – The radio bearer identifier BEARER is 5-bits long. There is one BEARER parameter per radio bearer associated with the same user and multiplexed on a single 10 ms physical layer frame. The radio bearer identifier is input to avoid a different keystream being used for an identical set of input parameter values.

DIRECTION – The direction identifier DIRECTION is 1-bit long. The direction identifier is input to avoid the keystreams for the uplink and for the downlink using the identical set of input parameter values. The value of the DIRECTION is 0 for the messages from UE to RNC and 1 for the messages from RNC to UE.

LENGTH – The length indicator LENGTH is 16 bits long. The length indicator determines the length of the required key stream block. LENGTH shall affect only the length of the KEYSTREAM BLOCK, not the actual bits in it.

14.7.4 Weakness in UMTS Security

These are weakness of the UMTS security. (1) IMSI is sent in clear text when allocating TMSI to the user. (2) The transmission of IMEI is not protected. (3) A user can be enticed to camp on a false BS. Once the user camps on the radio channels of a false BS, the user is out of reach of the paging signals of the SN. (4) Hijacking outgoing/incoming calls in networks with disabled encryption is possible. The intruder poses as a man-in-the-middle and drops the user once the call is setup.

In practice, protection of SS7-based protocols is best done at the application layer as there are no security mechanisms available at the network layer of SS7. The drawback of implementing protection at the application layer is that the target protocol itself will have to be modified. This process is both expensive and time consuming, and it must be repeated for every target protocol. After careful analysis, it was found that one could only afford to protect the mobile application part (MAP) protocol in some way.

14.8 Measurement Procedures

The purpose of the measurement control procedure is to setup, modify or release a measurement in the UE. The UTRAN may request a measurement (taking UE capabilities into account) by the UE to be setup,

modified or released with a MEASUREMENT CONTROL message, which is transmitted on the downlink DCCH using AM RLC. The UE supports a number of measurements running in parallel and each measurement is controlled and reported independently of every other measurement by using "measurement identity." Upon reception of a MEASUREMENT CONTROL message the UE performs and reports to UTRAN accordingly.

The following information is used to control the UE measurements and the measurement results reporting.

1. **Measurement identity:** A reference number that should be used by the UTRAN when setting up, modifying or releasing the measurement and by the UE in the measurement report.
2. **Measurement command:** One out of three different measurement commands. (a) Setup: Setup a new measurement. (b) Modify: Modify a previously defined measurement, for example, to change the reporting criteria. (c) Release: Stop a measurement and clear all information in the UE that are related to that measurement.
3. **Measurement type:** One of the types listed below describing what the UE measures. The different types of measurements are as follows. (a) *Intra-frequency measurements*: Measurements on downlink physical channels at the same frequency as the active set. A measurement object corresponds to one cell. (b) *Inter-frequency measurements*: Measurements on downlink physical channels at frequencies that differ from the frequency of the active set and on downlink physical channels in the active set. A measurement object corresponds to one cell. (c) *Inter-RAT measurements*: Measurements on downlink physical channels belonging to another radio access technology such as GSM. (d) *Traffic volume measurements*: Measurements on uplink traffic volume. A measurement object corresponds to one cell. (e) *Quality measurements*: Measurements of downlink quality parameters, for example, downlink transport block error rate. A measurement object corresponds to one transport channel in the case of BLER. A measurement object corresponds to one time slot for SIR (TDD only). (f) *UE-internal measurements*: Measurements of UE transmission power and UE received signal level. (g) *UE positioning measurements*: Measurements of UE position.
4. **Measurement objects:** The objects on which the UE measures measurement quantities, and the corresponding object information.
5. **Measurement quantity:** The quantity the UE measures on the measurement object. This also includes the filtering of the measurements.
6. **Reporting quantities:** The quantities the UE includes in the report in addition to the quantities that are mandatory to report for the specific event.
7. **Measurement reporting criteria:** The triggering of the measurement report, for example, periodical or event-triggered reporting.
8. **Measurement Validity:** Defines in which UE states the measurement is valid.
9. **Measurement reporting mode:** This specifies whether the UE transmits the measurement report using AM or UM RLC.

Cells that the UE is monitoring are grouped in the UE into three mutually exclusive categories.

a. **Active set cells:** User information is sent from all these cells. In FDD, the cells in the active set are involved in soft handover. In TDD the active set always comprises one cell only.
b. **Monitored set cells:** Cells that are not included in the active set but are included in the CELL_INFO_LIST belong to the monitored set.
c. **Detected set cells:** Cells detected by the UE that are neither in the CELL_INFO_LIST nor in the active set belong to the detected set. Reporting of measurements of the detected set is only applicable to intra-frequency measurements made by UEs in CELL_DCH state.

14.9 Handover Procedure

The handover procedure is the key factor to support roaming. In UMTS there are five different types of handover defined. The following categories of handover (also referred to as handoff) are defined in UMTS.

1. **Hard Handover**
   This is where the UE needs to break the current radio link and then establish a fresh radio link. Such circumstances are: changing the radio frequency of the channel connecting the UE to the UTRAN, changing the FDD mode to TDD mode, and changing to a cell on the same frequency but no support of diversity. Hard handover can be seamless or non-seamless. Seamless hard handover means that the handover is not perceptible to the user. In practice a handover that requires a change of the carrier frequency (inter-frequency handover) is always performed as hard handover.

   In UMTS hard handovers are used to, for example, change the radio frequency band of the connection between the UE and the UTRAN. During the frequency allocation process for UMTS, it is planned that each UMTS operator will have the possibility to claim additional spectrum (each of 5 MHz bandwidth) to enhance the capacity, when a certain usage level is reached (for example, a single operator can buy several 5 MHz bands of license to support an increased capacity). In this case several bands of approximately 5 MHz will be in use by one operator, resulting in the need for handovers between them. Also, it could switch the frequency between two operators (who might be using two different 5 MHz bands of frequency).

   Hard handovers are the so-called inter-mode handovers. This allows for changes between the FDD and the TDD UTRA modes. This handover type is sometimes also classified as inter-system handover, as the measuring methods used are very similar to WCDMA-GSM handovers.

   a. **Inter-system handover**
      Inter-system handover are necessary to support compatibility with other system architectures. It is mainly the handovers between UTRAN and the GSM radio access network that will be vital during the rollout of UMTS networks. In the initial deployment phase of 3G networks it is very likely that rural areas will not yet be covered by the WCDMA network. Thus GSM networks will still be used to provide coverage in those areas. On the other hand, it looks probable that the additional capacity provided by WCDMA networks will be used to unload the urban GSM network. In the later releases of the 3GPP specifications, handovers to systems other than GSM are included.

      The signaling procedure for handing over a UMTS user to the GSM system is shown in Figures 14.13 and 14.14. This example is illustrative of the general procedure followed during

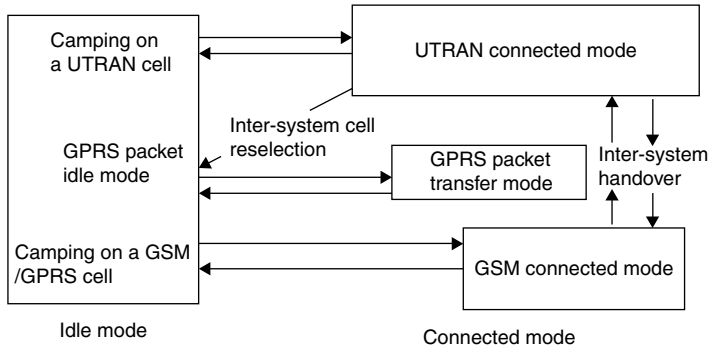


Figure 14.13 Dual mode UTRA FDD – GSM/GPRS

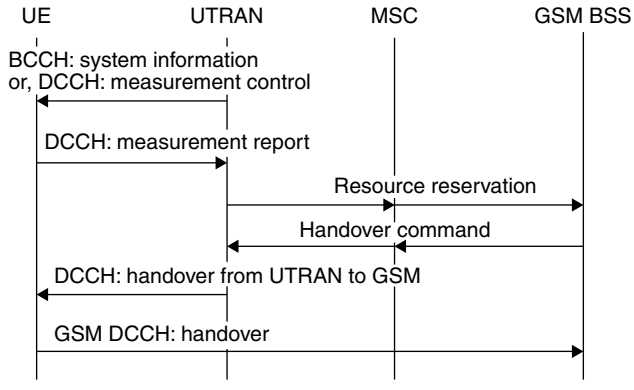


Figure 14.14 UMTS to GSM handover

handovers. The procedure generally consists of carrying out the measurements, reserving resources, and performing the actual handover. When switching the connection to another system architecture there is a need for a measurement on the frequency used by the other system. Using compressed mode the transmission gap is generated for this measurement, which was discussed in the previous chapter.

A dual mode (GSM-UMTS) mobile terminal would support handover from UMTS to GSM and vice versa.

b. **Inter-frequency handover**
Inter-frequency handovers are needed for utilization of hierarchical cell structures; macro, micro, and indoor cells. Several carriers and inter-frequency handovers may also be used for taking care of the high capacity needs in hot spots. Inter-frequency handovers will be needed also for handovers to second-generation systems such as GSM or IS-95. In order to complete interfrequency handovers, an efficient method is needed for making measurements on other frequencies while still having the connection running on the current frequency.

2. **Soft Handover**
This is when radio links are added and abandoned and the UE has more than one physical link to the UTRAN via one or more cells (sectors). Soft handover means that the radio links are added and removed in a way that the UE always keeps at least one radio link to the UTRAN. Soft handover is performed by means of macro diversity, which refers to the condition that several radio links are active at the same time. Normally soft handover can be used when cells operated on the same frequency are changed.
3. **Softer Handover**
Softer handover is a special case of soft handover, where the radio links that are added and removed belong to the same Node B (that is, the site of co-located base stations from which several sector cells are served). In softer handover, macro diversity with maximum ratio combining can be performed in Node B, whereas generally in soft handover on the downlink, macro diversity with selection combining is applied.

A soft or softer handover occurs when the mobile station is in the overlapping coverage area of two adjacent cells. The user has two simultaneous connections to the UTRAN part of the network using different air interface channels concurrently. In the case of soft handover, the mobile station is in the overlapping cell coverage area of two sectors belonging to different base stations; softer handover is the situation where one base station receives two user signals from two adjacent sectors it serves. Although there is a high degree of similarity between the two handover types there are some significant differences.

Thus, in summary we can say that softer handover involves the UE communicating with two sectors on the same cell site using a common carrier frequency. Eventually the UE changes from its original sector to the new sector. When two or more BSs forming different cells communicate simultaneously via a common carrier frequency with a UE as it roams between cells, then a soft handover is said to be in process. The process ends when the UE is communicating with a single BS. Sometimes a UE entering a different cell has to have its carrier frequency changed. This handover procedure is called a hard handover.

In UMTS systems the major part of the control signaling between UE and UTRAN is done by the radio resource control (RRC) protocol. Most of the RRC functionality is implemented in the RNC in the network side. Accurate measurements of the E_c/I_0 of the pilot channel (CPICH) form the main input for obtaining the RRC measurement report, necessary for making handover decisions. Usually three parameters can be measured. Besides the E_c/I_0 of the CPICH, also the received signal code power (RSCP) and the received signal strength indicator (RSSI) are measured (Figure 14.15). RSCP is the power carried by the decoded pilot channel and RSSI is the total wideband received power within the channel bandwidth. E_c/I_0 is defined as: $(E_c/I_0) = (\text{RSCP/RSSI})$.

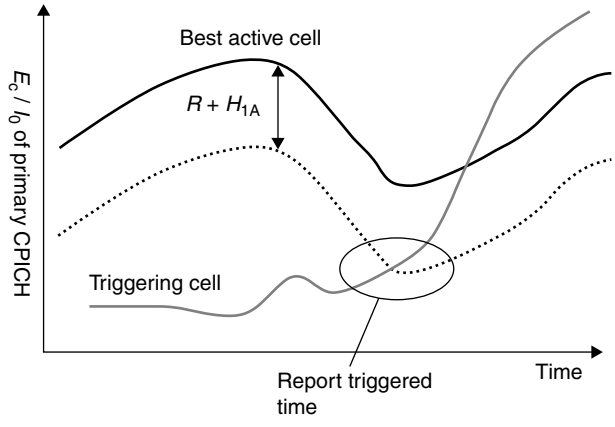


Figure 14.15 $(E_c = I_0)$ dB versus time showing curves of: the best active cell; a candidate entering cell; a dotted curve of $R + H_{1A}$ below the best active cell, with $S = 0$. The report is triggered when the two curves intersect

It is important to apply filtering on the handover measurements to average out the effect of fast fading. Measurement errors can lead to unnecessary handovers. Appropriate filtering can increase the performance significantly. As long filtering periods can cause delays in the handovers, the length of the filtering period has to be chosen as a trade-off between measurement accuracy and handover delay. Also, the speed of the user matters, the slower the user equipment is moving the harder it is to average out the effects of fast fading. Often a filtering time of 200 ms is chosen. Other essential information needed during the so-called intra-mode handovers – soft and softer handover – is timing information. As the WCDMA network is of asynchronous nature relative timing differences exist between the cells. To allow easy combining in the Rake receiver and avoid delays in the power control loops, the transmissions have to be adjusted in time. After the UE has measured the timing difference between the CPICH channels of the serving cell and the target cell, the RNC sends DCH timing adjustment information to the target cell.

The WCDMA soft handover algorithm as described in the 3GPP TR 25.922 specifications differs slightly from the IS 95A algorithm as used in cdmaOne, the standard for North American cellular

systems, also based on CDMA. Even though the significance of the latter cannot be ignored, this discussion is restricted to the analysis of the WCDMA algorithm only. Based on the E_c/I_0 measurements of the set of cells monitored, the mobile station decides which of three basic actions to perform; it is possible to add, remove or replace a Node B in the active cell. These tasks are, respectively, called radio link addition and radio link removal, while the last is combined radio link addition and removal. The example below is directly taken from the original 3GPP specifications. Discussing this scenario gives a good insight into the algorithm itself and forms an introduction to the illustrating simulations included in the next paragraph. This scenario can be based on a user following a trajectory as shown below.

The handover algorithm to make the handover decision needs different types of measurement information. Table 14.1 lists the measurements that can be carried out for handover purposes. The actual handover algorithm implementation is left to the equipment manufacturers.

Table 14.1 Measurements for handover

Received signal code power (RSCP)	This is received power on one code measured on the pilot bits of the primary CPICH. The reference point for it is the antenna connector at the mobile station.
SIR	Signal to interference ratio.
RSSI	Received signal strength indicator.
GSM carrier RSSI	Measurements are performed on a GSM BCCH carrier.
CPICH E_c/I_0	The received energy per chip divided by the power density in the band.
Transport channel BLER	Estimation of transport channel block error rate (BLER).
Physical channel BLER	The physical channel bit error rate (BER).
Mobile station transmit power	The total transmit power of the mobile on one carrier.

Base stations in WCDMA are asynchronous, and therefore no external source of synchronization, as in GPS, is needed for the base stations. Asynchronous base stations must be considered when designing soft handover algorithms and when implementing location services.

Before entering soft handover, the mobile station measures observed timing differences of the downlink SCHs from the two base stations. The mobile station reports the timing differences back to the serving base station. The timing of a new downlink soft handover connection is adjusted with a resolution of one symbol (that is, the dedicated downlink signals from the two base stations are synchronized with an accuracy of one symbol). This enables the mobile Rake receiver to collect the macro diversity energy from the two base stations. Timing adjustments of dedicated downlink channels can be carried out with a resolution of one symbol without losing orthogonality of the downlink codes.

14.10 Cell Update

The cell update procedure can be triggered by several reasons, including cell re-selection, expiry of periodic cell update timer, initiation of uplink data transmission, UTRAN originated paging, and radio link failure in Cell_DCH state. The cell update confirm may include UTRAN mobility information elements (new URNTI and C-RNTI) for the UE. In this case, it responds with a UTRAN mobility information confirm message so that the RNC knows that the new identities are taken into use. The cell update confirm may also include a radio bearer release, radio bearer reconfiguration, transport channel reconfiguration or physical channel reconfiguration. In these instances, the UE responds with a suitable "complete" message, see Figure 14.16.

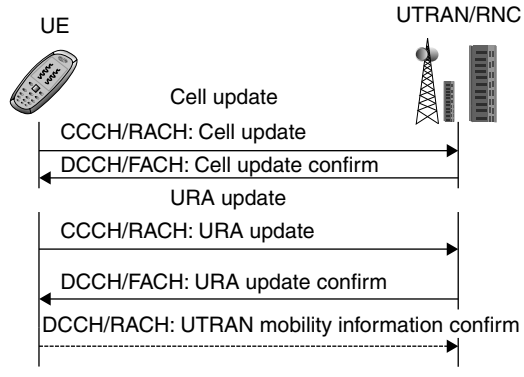


Figure 14.16 Cell update procedure and URA update procedure

The URA update confirm may assign a new URA identity that the UE has to follow. It may also assign new RNTIs for the UE. In these cases UE responds with a UTRAN mobility information confirm message so that the RNC knows that the new identities are taken into use.

14.11 High-Speed Downlink Packet Access (HSDPA)

High-speed downlink packet access is a new protocol for mobile telephone data transmission. The high-speed downlink packet access (HSDPA) channel was introduced by 3GPP in Release 5. It supports up to 14 Mbps peak data rate in downlink (2 Mbps on the uplink), which is a fivefold improvement in data spectral efficiency. HSDPA is an improvement over W-CDMA as it uses different modulation and coding techniques. It creates a new channel within W-CDMA called HS-DSCH, or high-speed downlink shared channel (only used for downlink). This channel performs differently than other channels and allows for faster downlink speeds. Release 5 supports very dynamic adaptive modulation and coding, adaptive scheduling and hybrid automatic retransmission request (H-ARQ). In addition to HSDPA, Release 5 introduces the IP multimedia system (IMS) architecture that promises to greatly enhance the end-user experience for integrated multimedia applications and offers the wireless operators an efficient means of providing such IP based multimedia services. Release 5 also introduces the IP UTRAN concept to reduce network costs and to increase network efficiencies. Release 99 UMTS carriers can be upgraded to support Release 99 as well as Release 5 terminals in the same 5 MHz band. The three key features of Release 5 – HSDPA, IMS and IP UTRAN will be discussed briefly.

HSDPA defines a new transport channel type, known as high-speed downlink shared channel (HS-DSCH) that allows several users to share the air interface channel dynamically with peak channel rates up to 14 Mbps. HS-DSCH resources can be shared between all users in a particular sector. It supports QPSK and 16-QAM modulations, link adaptation and combining of re-transmission at the physical layer with HARQ. Primary channel multiplexing occurs in the time domain where each TTI consists of three slots (or 2 ms). TTI has been reduced from 80/40/20/10 ms TTI sizes as supported in Release 99. It helps to support delay sensitive applications better and improves the efficiency of adaptive modulation and coding. Within each 2 ms TTI, a constant spreading factor of 16 is used for code multiplexing and a maximum of 15 parallel codes are allocated for HS-DSCH. All these 15 parallel codes may be assigned to one user in a TTI or may be split across several users. The number of codes assigned to each UE depends upon several factors, such as UE code capabilities, QoS requirements, and cell loading.

Release 5 also defines two new control channels for HSDPA operation. A high-speed shared control channel (HS-SCCH) informs all the terminals how to decode the HS-DSCH (for example, modulation, codes, re-transmission information, etc.). A high-speed dedicated physical channel (HS-DPCCH) carries the channel

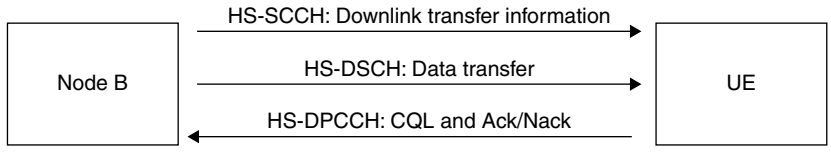


Figure 14.17 HSDPA channels

quality indicator (CQI) from UE to Node B (Figure 14.17). HS-DPCCH also carries ACK/NACK signaling for HARQ. The base station uses the CQI of each active user to take an AMC and scheduling decision.

HSDPA supports up to 15 parallel codes that can be shared dynamically by different users. Various combinations of modulation schemes and channel coding rates can be used to provide different peak rates.

H-ARQ: Re-transmission requests are managed by the base station instead of the RNC as in Release 99. If decoding of initial transmission fails, a re-transmission is sent that can be combined with the initial transmission or is self-decodable. UE does not discard a failed re-transmission but stores it and later combines it with re-transmissions to increase probability of successful decoding. Combining of different transmissions provides improved decoding efficiency while minimizing the need for additional repeat requests over the air-interface. HSDPA supports chase combining (CC) and incremental redundancy (IR). In CC, if a UE detects that it has received an erroneous packet, it sends an NAQ to Node B. Node B transmits the packet again with the same coding scheme. If this is also received in error, this packet is combined with previous packet in an attempt to recover from errors. Eventually this packet with be received without error or a re-transmission limit will be reached and error recovery will be left to higher layers (such as RLC and TCP for TCP based applications).

IR is similar to CC but re-transmitted data are coded with additional redundant information to improve the chances that the packet will be received either without errors or with enough errors removed to allow combining with previous packets to allow error re-correction. In order to better utilize the waiting time between acknowledgements, multiple processes are allowed to run for the same UE using different TTIs. This is called *N*-channel stop and wait protocol. If one channel is waiting for an acknowledgement, the remaining *N*-1 channels continue to transmit. *N* is up to six for advanced Node B implementations.

MAC functionality in Release 99 was located at the RNC. Now it is split between the RNC and Node B (that is, base station) and thus brought closer to the air-interface. MAC-HS deals with the functions critical to delay and performance and is located at Node B (Figure 14.18). For non-HSDPA

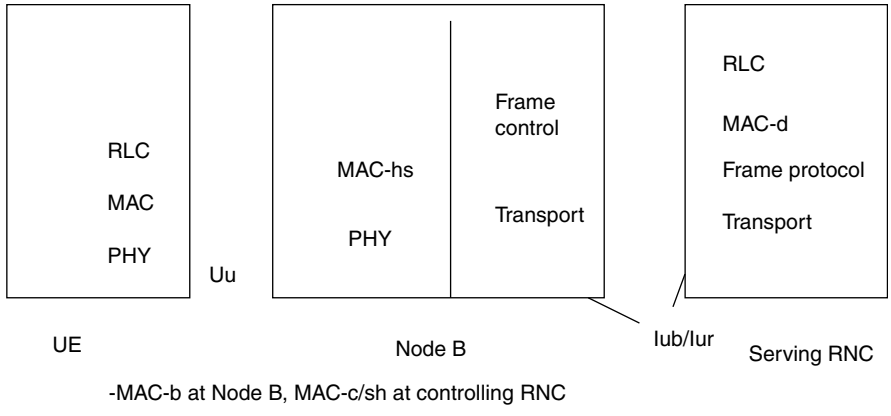


Figure 14.18 MAC-hs for HSDPA

support, Node B stations are connected to RNC that provide scheduling, coding parameters, and retransmission services to UE devices. To support HSDPA, these parameters are determined based on the instantaneous channel conditions as reported by the UEs. The delays associated with the forwarding of channel data to the RNC for processing coupled with the burden of an RNC having to service multiple Node B stations requires that the NodeB and not the RNC perform these services for HSDPA.

Adaptive Scheduling: Release 5 moves the scheduling decision to the base station (that is, Node B) while Release 99 does scheduling of packet data at the RNC. The base station uses terminal feedback information about channel quality, terminal capabilities, QoS needs, and air-interface resource availability to take the best decision in real-time. The proportional fair scheduler is one example of a scheduler that prioritizes applications in the best channel conditions while also prioritizing users that have received lower throughout than other users. Reduction of TTI to 2 ms for HSDPA leads to significantly lower latencies than Release 99.

HSDPA in Release 5 does not support soft handover. UE continuously monitors all the Node Bs in its active set and reports to UTRAN when a change in the best cell occurs. UTRAN then reconfigures the serving HS-DSCH cell using either synchronous or asynchronous re-configurations. Both inter-Node B and intra-Node B handovers are supported.

UE Capabilities: There are 12 different categories of mobiles defined for HSDPA that specify the following parameters:

- maximum number of HS-DSCH codes that a UE can simultaneously receive;
- minimum inter-TTI time that is defined as the minimum time between the beginning of two consecutive transmissions to this UE;
- supported modulations (QPSK, 16QAM);
- maximum number of transport block bits received within an HS-DSCH TTI;
- maximum number of soft channel bits over all the HARQ processes.

HSDPA UE categories are given in Table 14.2.

Table 14.2 HSDPA UE categories (3GPP TS25.306)

Category	Codes	Inter-TTI	TB size	Total # of soft bits	Modulation	Data rate (Mbps)
1	5	3	7300	19200	QPSK/16QAM	1.2
2	5	3	7300	28800	QPSK/16QAM	1.2
3	5	2	7300	28800	QPSK/16QAM	1.8
4	5	2	7300	38400	QPSK/16QAM	1.8
5	5	1	7300	57600	QPSK/16QAM	3.6
6	5	1	7300	67200	QPSK/16QAM	3.6
7	10	1	14600	115200	QPSK/16QAM	7.2
8	10	1	14600	134400	QPSK/16QAM	7.2
9	15	1	20432	172800	QPSK/16QAM	10.2
10	15	1	28776	172800	QPSK/16QAM	14.4
11	5	2	3650	14400	QPSK	0.9
12	5	1	3650		QPSK	1.8

QoS Scheduling and Admission Control Algorithms: These are typically vendor dependent algorithms. They may use several parameters such as downlink and uplink air-interface load, traffic, and QoS requirements for an application, channel conditions for a user, observed QoS, and so on.

An application specific scheduler part controls the utilization of RRC states, allocation of transport channels and their bit rates. A cell specific scheduler part controls sharing of radio resources between various applications belonging to different users. Dedicated channels can be used for real-time applications with constraints on delay bound and jitter. A cell specific packet scheduler is used for other applications such as Web browsing and background applications.

Transition from one transport channel to another can be controlled by various factors such as buffer length, inactivity timer, and so on. Based on traffic volume threshold, it could move from Cell_FACH to Cell_DCH and depending upon an inactivity timer, it could move from Cell_DCH to Cell_FACH. For example, RACH/FACH may be used for TCP connection establishment. A DCH may be allocated during TCP slow start and the bit rate of a DCH channel may be increased during the download process.

Both the mobile receiver and the transmitter are active in the Cell_DCH state and it causes relatively high power consumption. In the Cell_FACH state, only the mobile receiver is active and thus it consumes less power. In the Cell_PCH (paging) state, discontinuous reception is allowed and it results in much lower power consumption than Cell_FACH where continuous reception is required.

14.12 High-Speed Uplink Packet Access (HSUPA)

Release 6 features include the following: multimedia broadcast multicast service (MBMS), enhanced dedicated channels (E-DCH), advanced receiver performance specifications (receive diversity at the terminal), IMS enhancements, enhancements to support interworking with WLAN, wideband AMR speech codec, IP flow based bearer level charging, push-to-talk over cellular (PoC), and support for emergency services. To support a high-speed uplink packet access (HSUPA), features similar to HSDPA are used. These include adaptive modulation and coding, HARQ, shorter TTI, and improved QoS mechanisms. Release 6 IMS allows provision of a CS domain-like service via the PS domain.

14.13 IP Multimedia Subsystem (IMS)

IMS increases the functionality of 3G mobile networks by supporting IP-based applications and services through the SIP protocol. Rapid spread of fixed-network broadband and the offering of services such as transactions, content distribution, and VoIP over all-IP networks have also made IMS increasingly relevant to fixed operators. Effectively, IMS provides a unified architecture that supports a wide range of IP-based services over both packet- and circuit-switched networks, employing a range of different wireless and fixed-access technologies. A user could, for example, pay for and download a video clip to a chosen mobile or fixed device and subsequently use some of this material to create a multimedia message for delivery to friends on many different networks. A single IMS presence-and-availability engine could track a user's presence and availability across mobile, fixed, and broadband networks, or a user could maintain a single integrated contact list for all types of communications. A key point of IMS is that it is intended as an open-systems architecture. Services are created and delivered by a wide range of highly distributed systems (real-time and non-real-time) cooperating with each other.

In order to achieve access independence and to maintain a smooth inter-operation with wireline terminals across the Internet, the IP multimedia subsystem attempts to be conformant with IETF "Internet standards." Therefore, the interfaces specified conform as far as possible to IETF "Internet standards" for the cases where an IETF protocol has been selected, for example, SIP. IMS is potentially the base of a new telecom business model for both fixed and mobile networks and is a key enabler of fixed/mobile convergence. In principle, it replaces the traditional walled-garden approach of a single operator, offering a limited range of services from within a highly controlled network, with an almost limitless range of highly functional services that span multiple operator and service-provider domains – fixed and mobile.

MSC was split into MSC server and media gateway (MGW) in Release 4. GMSC was also split into GMSC server and MGW in Release 4 to allow for better scalability as features with higher data rates are introduced. User data goes through MGW that performs several functions such as packet switching, echo cancellation, and speech encoding/decoding. One MSC/GMSC server can control several MGWs. The IP multimedia subsystem (IMS) was introduced in Release 5, which enables IP-based service provision via PS domain. SIP is the protocol between UE and IMS. Call session control function (CSCF) is introduced, which acts as the first contact point to the terminal in the IMS. Media gateway control function (MGCF) handles protocol conversions and media resource function (MRF) controls media stream resources.

Call Session Control Function: The CSCF can act as proxy CSCF (P-CSCF), serving CSCF (S-CSCF), or interrogating CSCF (I-CSCF). The CSCF serves as a centralized routing engine, policy manager, and policy enforcement point to facilitate the delivery of multiple real-time applications using IP transport. It is application-aware and uses dynamic session information to manage network resources (servers, media gateways, and edge devices) and to provide advance allocation of these resources depending on the application and user context. The P-CSCF is the first contact point within the IMS for the subscriber. It accepts requests and serves them internally or forwards them. The I-CSCF is the contact point within an operator's network for all connections destined for a user of that network, or for a roaming user currently located within that network's service area. There may be multiple I-CSCFs within an operator's network. The S-CSCF is responsible for identifying the user's service privileges, selecting access to the home network application server, and providing access to that server. Further definitions of the P-, S- and I-CSCF are provided in 3GPP WRAN TS 23.228.

Media Gateway Control Function (MGCF): The MGCF controls the parts of the call state that pertain to connection control for media channels in an IMS-MGW and communicates with CSCF. It selects the CSCF depending on the routing number for incoming calls from legacy networks. MGCF also performs protocol conversion between ISUP and the IM subsystem call control protocols.

IMS – Media Gateway Function (IMS-MGF): A IMS-MGW may terminate bearer channels from a switched circuit network and media streams from a packet network (for example, RTP streams in an IP network). It may support media conversion, bearer control, and payload processing (for example, codec, echo canceller, conference bridge). It interacts with MGCF for resource control and handles resources such as echo cancellations, and so on. It may also have codecs.

Multimedia Resource Function Processor (MRFP): It mixes media streams (for example, for multiple parties) and processes media streams (audio transcoding, media analysis) and manages access rights to resources in a conferencing environment.

Messaging Services in IMS: There are many different types of messaging services available both in the wired and wireless worlds. Some services are supported in both, others are only found in one. Various services are designed to be used in what is perceived as "real time" and others are designed as a "mailbox" service, where the message is stored ready for collection or delivery at a later date. Introducing real-time services into the wireless world brings many challenges not currently experienced in the wired environment. These include limited air-interface capacity, limited memory in the terminal, charging and

billing. Examples of these messaging services are SMS, MMS, Instant Messaging, Chat, e-mail, messaging in the IMS.

For the SMS service, a user can send a message to the originators home SMSC, where it is stored until it is possible for the SMSC to deliver the message to the recipient. A multimedia messaging service (MMS) allows users to send and receive messages exploiting the whole array of media types available today, for example, text, images, audio, video. The recipient may be notified that a new MMS message has arrived in their inbox, from which the recipient can then connect to their mailbox to retrieve the message or have the message pushed to them. The expectation of the Instant Messaging service from a user's viewpoint is to be able to communicate with other users in real time. This service relies on a communications association between the originator and the recipient in order to meet this expectation. The service is primarily text based, although most services allow for various attachments to be added. Most applications include access to a "Presence" service to allow users to see who is available for Instant Messaging, however this is not mandatory. Chat service enables people to send text to a central point (chat server) allowing all of those users who are connected to the central point to view the text.

As 3GPP has developed the concept of IMS, it's useful to consider how an SIP based IP network can be utilized to provide messaging capabilities. SIP allows creation of real-time sessions between groups of users. Therefore it is possible that SIP-based messaging could be a potential candidate to provide the equivalent of "Chat Room" and "Instant Messaging" (IM) type services found on the Internet today. Typical characteristics of instant messaging are instant delivery of the messages to the targeted recipient (s) and interaction with presence information, where users are able to see who is on-line as well as their status.

Processing granularity will be a major consideration for the efficient implementation of an HSDPA-compliant base station. Systems based on a small number of high-performance DSPs tend to demand large buffers and, to reduce the overhead of switching between tasks, will tend to work on large groups of data at any one time. This makes things "clumpy" with high latency. Fine-grained control will be necessary to implement features such as fast scheduling and per-user coding and modulation adaptation.

Further Reading

3GPP TS 21.111 *USIM and IC Card Requirements*, and TS 25.101, *UE Radio Transmission and Reception (FDD)*. http://www.3gpp.org/Specifications.

3GPP TS 22.129. *Hand-over Requirements between UMTS and GSM or Other Radio Systems*.http://www.3gpp.org/Specifications.

3GPP TS 23.060. *General Packet Radio Service (GPRS) Service Description-Stage 2*. http://www.3gpp.org/Specifications.

3GPP TS 24.007. *Mobile Radio Interface Signalling layer 3-General Aspects*. http://www.3gpp.org/Specifications.

3GPP TS 24.008. *Mobile Radio Interface Layer 3 Specification; Core Network Protocols-Stage 3*. http://www.3gpp.org/Specifications.

3GPP TS 25.103. *RF Parameters in Support of Radio Resource Management*. http://www.3gpp.org/Specifications.

3GPP TS 25.201. *Physical Layer – General Description*, and TS 25.301, *Radio Interface Protocol Architecture*. http://www.3gpp.org/Specifications.

3GPP TS 25.211. *Physical Channels Mapping of Transport Channels onto Physical Channels (FDD)*. http://www.3gpp.org/Specifications.

3GPP TS 25.215. *Physical Layer – Measurements (FDD)*, and TS 25.303, *UE Functions and Interlayer Procedures in Connected Mode*. http://www.3gpp.org/Specifications.

3GPP TS 25.302. *Services Provided by the Physical Layer*, and TS 25.303, *UE Functions and Inter-Layer Procedures in Connected Mode*. http://www.3gpp.org/Specifications.

3GPP TS 25.304. *UE Procedures in Idle Mode*. http://www.3gpp.org/Specifications.

3GPP TS 25.321. *Medium Access Control (MAC) Protocol Specification.* http://www.3gpp.org/Specifications.
3GPP TS 25.322. *Radio Link Control (RLC) Protocol Specification*, and TS 25.331, *Radio Resource Control (RRC) Protocol Specification.* http://www.3gpp.org/Specifications.
3GPP TS 25.331. *RRC Protocol Description*, and 3GPP 25.331*RRC Procedures*. http://www.3gpp.org/Specifications.
3GPP TS 25.401. *UTRAN Overall Description.* http://www.3gpp.org/Specifications.
3GPP TR 25.832. *Manifestations of Hand-over and SRNS Relocation.* http://www.3gpp.org/Specifications.

15

Anatomy of a UMTS Mobile Handset

15.1 Introduction

3G mobile terminals should be capable of supporting many complex applications, including multimedia. There is a wide variety of types, ranging from portable phones, PC-cards, video phones, to personal digital assistance (PDA) types. The UE functionalities are implemented in the software and in the hardware. The hardware components are divided into modem hardware and application hardware. The modem hardware is needed to support the radio modem functionality and the application hardware is needed for vocoders, imaging, video encoders, display, and so on. A typical 3G handset includes a microphone, CMOS imager and MPEG-4 encoder, keyboard, a smart card, speaker, display, RF module, baseband processing transceiver modules, and so on. The software components can be divided into several modules as described for the case of a GSM mobile phone. However, the radio modem physical layer processing software and radio protocol software modules will be different from a GSM phone.

In Chapter 10 we have discussed the various components of a GSM mobile phone. In the present chapter, we will discuss the WCDMA based 3G mobile phone architecture and some of its internal modules.

15.2 Mobile System Architecture

We discussed the CDMA techniques in detail in Chapter 5. As an example, in Figure 15.1 the simple architecture of a CDMA based mobile phone is shown (for IS-95 based system).

Although many modules are similar to the GSM phone, spreading and scrambling blocks are newly introduced here. Figure 15.2 shows the basic blocks for the spreading and scrambling units. OVSF codes are applied on the transmit side and then the same ones are used to decorrelate the signal energy of interest on the receiver side. The generated source code of the I and Q data streams of the PSK (QPSK or BPSK) modulated signal is passed through the filters to constrain the required bandwidth. Generally a particular Nyquist filter is chosen because it is easier to implement than other configurations, such as a root raised cosine filter.

Mobile Handset Design Sajal Kumar Das

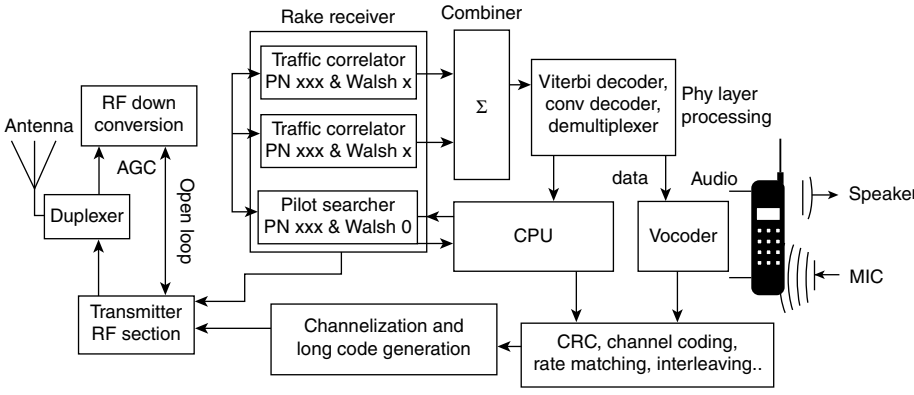


Figure 15.1 Internal blocks of a CDMA based mobile phone

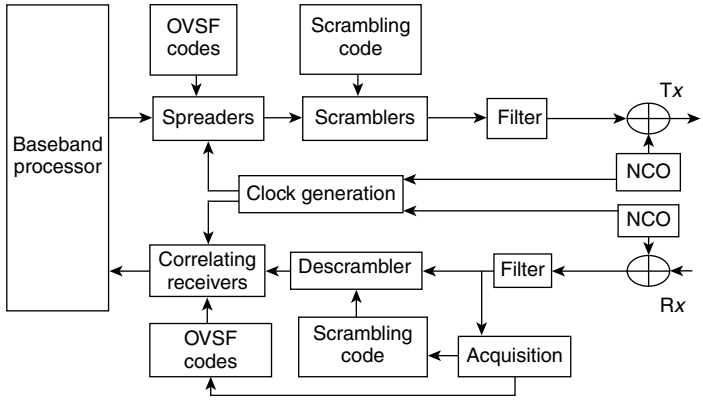


Figure 15.2 Spreader and scrambling units

15.2.1 Configuration of a WCDMA Radio Transmitter and Receiver

A generic block diagram of a WCDMA transmitter and receiver is shown in Figure 15.3. From the protocol stack layer-2, layer-1 receives the data block (transport block) for the various transport channels at different time intervals (based on the TTI of those channels). After this, it adds CRC bits to each transport block (TB), which is the basic unit of data that is subject to processing. Next, channel coding and interleaving procedures takes place. The interleaved bit sequence is subject to overhead additions, for example, pilot bits for channel estimation and TPC bits for power control. Then after bit multiplexing, data modulation takes place, where data are mapped to the I-phase and Q-phase components and spread across the spectrum by two layers of spreading code sequences (channelization and then scrambling). The resulting chip data sequence is restricted to a 5 MHz band by a pulse shaping filter (roll-off factor 0.22) and then converted into an analog signal by a D/A converter. The orthogonally modulated IF signals are further up converted to RF signals in the 2 GHz band and are subjected to power amplification and then send via an antenna.

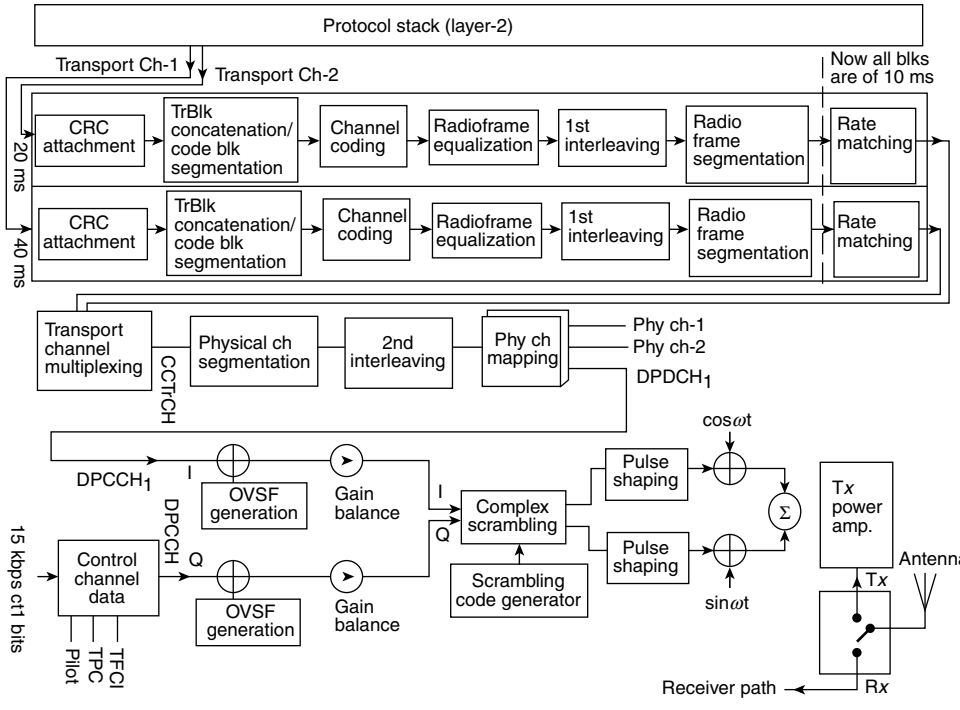


Figure 15.3 WCDMA UE transmission blocks

In uplink the transport channel DCH is mapped to the DPDCH and the DPCCH physical channels. So, from the upper layer of UE, the traffic and signaling data arrive at the physical layer at an interval of 20 ms (for traffic) or 40 ms (for signaling) or any other time interval based on the TTI of that channel. This is then processed and mapped on the I-path of the BPSK modulation. DCCH (the control data Pilot, TFCI, power control bits are generated at the physical layer itself) is mapped on the Q-path. This has already been shown in Chapter 13. The DPCCH is always spread to the chip rate (3.84 Mcps) by the channelization code $C_{\text{ch},256,0}$ ($SF = 256$), whereas the nth DPDCH$_n$ ($0 <= n <= 6$) is spread to the chip rate by the channelization code $C_{\text{ch},SF,n}$. This complete transmission process is shown in Figure 15.3.

On the other hand at the receiver side, the received signal is amplified by an LNA and down converted into IF signals (based on the used RF down conversion technique). The amplified signals are subjected to quadrature detection to generate I- and Q-components. This is then converted into a digital signal by using an ADC. The digitized I- and Q-components are bound within the specified band by a square root raised cosine Nyquist filter and are time divided into a number of multipath components with different propagation delay times through a despreading process that uses the same spreading code as the one used for spreading transmitted signals (Figure 15.4). The data signal energy is recovered in the spread spectrum process by multiplying synchronously or despreading the received signal with the copy of the code sequence that was used to spread it in the transmitter. Owing to the multipath, there will be several time delayed versions at the receiver, so the signal is applied simultaneously to a number of synchronous receiver blocks and each of these can be allocated to a separate multipath signal. There will be separate, despread, time delayed recovered data streams. The data streams have to be equalized in phase (time) and

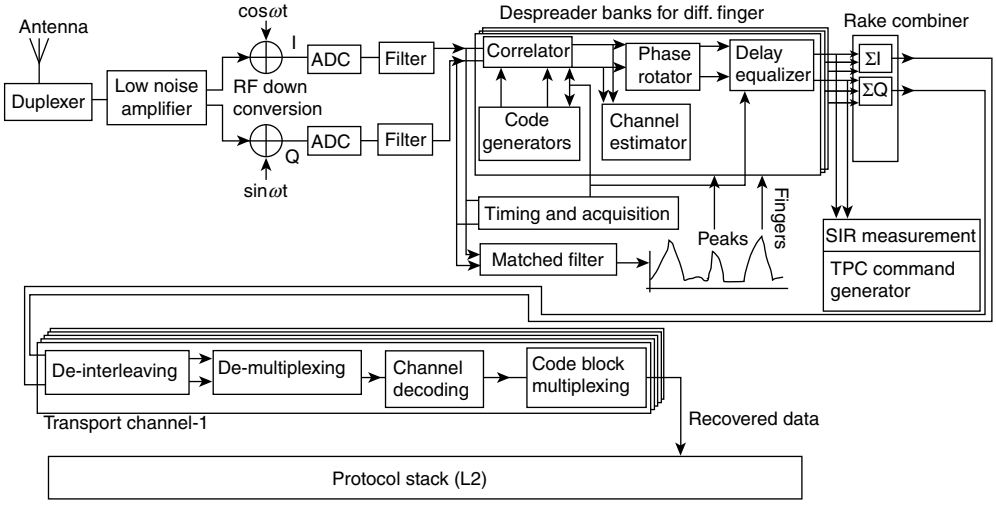


Figure 15.4 WCDMA UE reception blocks

combined using a Rake combiner. To identify the accurate signal phase, the SCH channel is used. The received signal has been processed with both scrambling and spreading codes. The code generators and correlators will generate scrambling codes to descramble the signal and then OVSF codes to despread the signal. This is performed by a parallel Rake finger processing unit.

After this the resulting data sequences are deinterleaved and channel decoded. The transported data sequence is recovered by binary data decision, which is then divided into transport channels and is subjected to block error detection and forwarded to the higher layer.

15.3 UE Hardware Architecture and Components

A typical block diagram of a UE hardware module is shown in Figure 15.5.

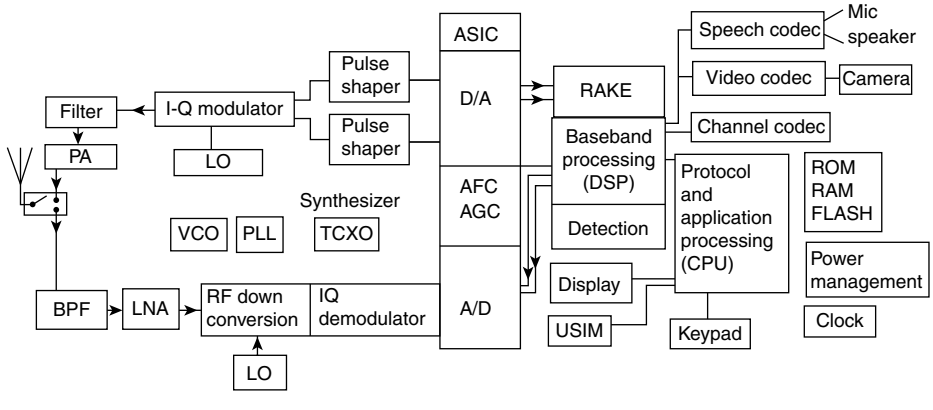


Figure 15.5 Typical hardware block diagram of a 3G UE

15.3.1 RF Front-End Architecture

Most commonly the super-heterodyne architecture or DCR architecture is used in the analog RF front-end part (Figure 15.6). The receiver with super-heterodyne architecture will be discussed as an example. The received frequencies at 2110–2170 MHz are down-converted to the first intermediate frequency (IF) of 270 MHz. The desired 5 MHz channel is then selected by an IF band-pass filter. An AGC amplifier provides variable gain control to the IF signal with 90 dB of dynamic range, from 0.45 to 45 dB. The IF signal is fed to the demodulator circuits, where it is mixed with a fixed local oscillator frequency to produce zero IF baseband quadrature I- and Q-signals. To avoid the unwanted dc and low frequency signals, simple dc blocks are used. The differential I- and Q-signals emerging after dc blocking are filtered by low-pass filters and sent to the 10-bit A/D converters.

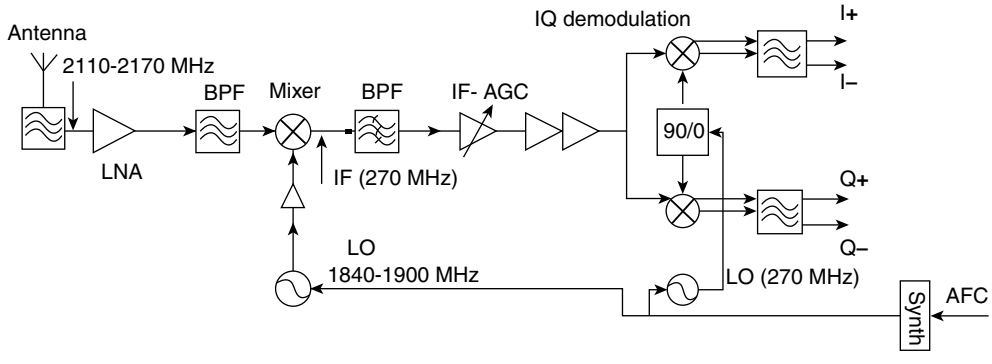


Figure 15.6 Block diagram of the RF section of a WCDMA mobile phone

15.3.2 Baseband Architecture

The baseband modules of the WCDMA receiver are shown in Figure 15.7. Different units will be described in detail later.

15.3.2.1 Rake Management and Synchronization

Demodulation of the received downlink WCDMA signal is performed by means of a Rake receiver. The main tasks of this module include Rake management, synchronization, and tracking. The Rake receiver is time aligned to the PN sequence in the spread signal and will de-spread the signal by means of correlating the received signal with the scrambling code (10 ms based) and the channelization code.

There are a number of correlation processes; each one termed a finger, which is time-aligned to the various components of the received signal after it has been passed through a multipath channel environment. Each finger tracks its signal independently of the others. There is an algorithm that monitors the timing of all fingers so that the average tracking is known, and thus all receiver timings derived from this timing. This average tracking is termed the RX system timing. It is possible to program each finger with different scrambling and channelization codes, so as to receive signals from different cells. Each finger that is programmed with the same scrambling/channelization code is combined after de-spreading. This combination of signals is coherent. This means that the multipath coefficients for the channel are thus required and will be estimated using the DPCCH pilot bits. Whereas, each finger that is

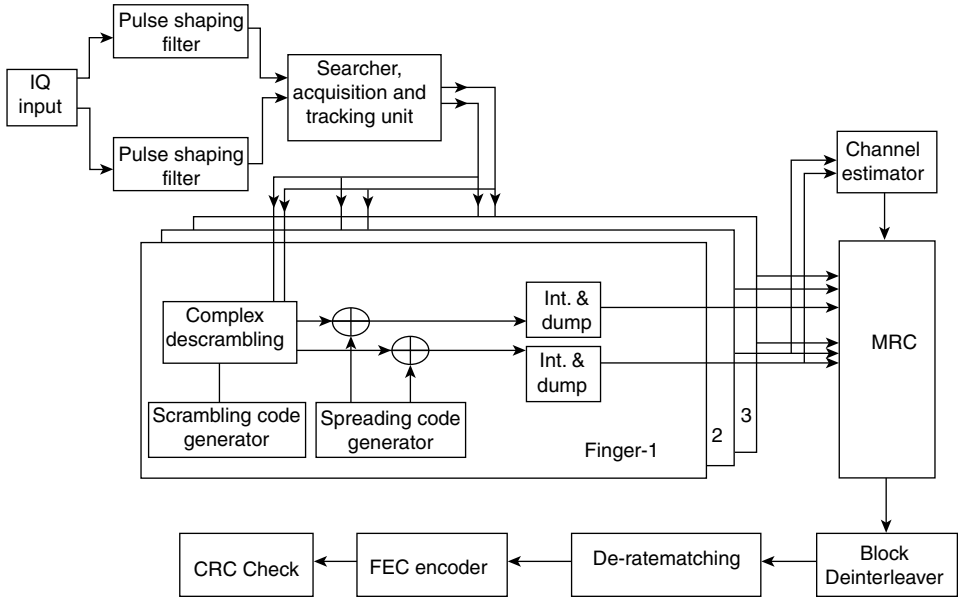


Figure 15.7 Reverse link receiver

programmed with different scrambling/channelization codes is not combined. Un-combined signals are treated as separate signal sources and have their own channel decoding.

When engaged in communications with a cell, one set of fingers will be dedicated to receive the DPCCH and DPDCH, whilst another set of fingers will be dedicated to search for new cells, updating measurements of the serving cells and the new cells and, in general, maintaining a set of statistics in the event that a handover is necessary.

Rake Management

Functions carried out by Rake management are as follows.

- **Tracking/RX System Timing** – Determines RX system timing of the serving cell and active cells.
- **Finger Control** – Decides when the signal from a finger of the Rake is no longer producing a useful signal and reassigns another timing based on the output of channel estimation, or disables that finger.
- **Channel Estimation** – One or more of the fingers of the Rake is used to periodically establish the channel response to $^1/_4$ chip resolution for both the serving cell and active cells.
- **AFC and Frequency Measurement** – Based on the decoded pilot signal from the Rake, the UE determines the frequency offset, and feeds a correction voltage back to the frequency synthesizer in the radio subsystem. This control system is termed the automatic frequency control (AFC). Owing to the inherent fact that the UE can be moving in a multipath environment, and is subject to a temperature variation, the averaging of this signal is done over a very long period.
- **AGC and Power Measurement** – Based on the received signal before the RRC filter and de-spreading, the UE measures the power so as to determine the gain setting on a frame (10 ms) basis. This control system is termed the automatic gain control (AGC).
- **Inner-Loop Power Control** – After de-spreading, the TPC bit is extracted on a slot basis and decoded, so as to increment/decrement the level of transmitted power. This control system is on a slot basis.

- **Other Measurements** – The system measurements that are reported to the L2 and L3 entities.
- **Compressed Mode** – The Rake management does not pass information to the channel decoder when it is making power or other measurements during the idle slots.
- **Handover** – The Rake management performs assignments of fingers to establish new links and to delete old ones.

15.3.2.2 Hardware Implementation of Rake Receiver

As discussed in Chapter 3, the Rake receiver is essentially a diversity receiver designed and suitable for CDMA, where the diversity is provided by the fact that the multipath components are practically uncorrelated from one another when their relative propagation delays exceed a chip period. The Rake receiver is an efficient method of utilizing most of the received signal energy from the different path signals that carry the same information in one symbol in the spread spectrum.

Figure 15.8 shows a simple block diagram of a Rake receiver with multiple fingers for multipath combining. The path delay will be controlled by the Rake management. The digital correlation is mainly for extracting data signals by despreading with the scrambling codes and channelization codes. It is multiplied by the estimated weights for each path after summing for a symbol period and routed to the Rake combiner. The combining is done for each symbol period and the symbol timing is controlled by the protocol layers.

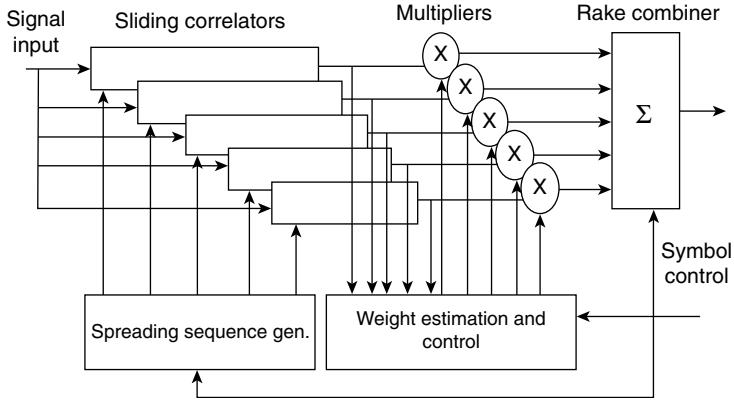


Figure 15.8 Rake module

The scrambling code sequences are constructed by combining two real sequences into a complex sequence and repeating for every 10 ms radio frame (Figure 15.9). Each of the two real sequences are constructed as position-wise modulo-2 sums of 38 400 chip segments of two binary m-sequences generated by means of two generated polynomials of degree 18. The resulting sequences will be segments of a set of Gold sequences. If x and y are the sequences, x is constructed using the polynomial $1 + X^7 + X^{18}$ and the y sequence is constructed using the polynomial $1 + X^5 + X^7 + X^{10} + X^{18}$.

The configuration of a scrambling code generator is shown in Figure 15.9. The sequence is generated using sequential logic circuits. The binary sequences are shifted through the shift registers in response to the chip clock inputs. The output of required stages are logically combined and then feedback to the first stage. The binary codes are converted into real valued sequences by transforming "0" to "+1," "1" to −1. The initial values will have to be loaded, which should be from protocol.

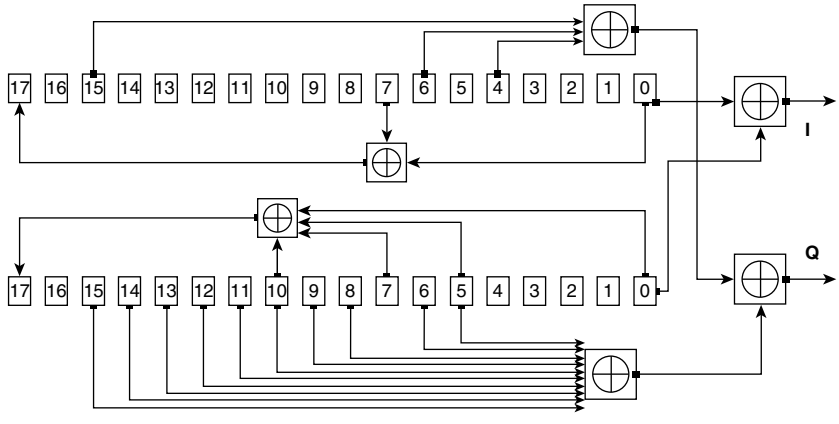


Figure 15.9 Scrambling code generators

This combiner accumulates all the values from the fingers for a symbol period. The SF and symbol information are known to the protocol layers on negotiation with the base station. So the control for the accumulator and the information for the channel estimator are passed from the respective protocol layer.

Channel tap coefficients are constant for a symbol period. As the channel impulse response is randomly changing these weights cannot be determined exactly, so it is estimated. This estimation is based on the received signal; either the data signal itself or the pilot signal is used for the estimation.

Generally the estimation of the signal strength is done by low pass filtering because the signal is contaminated with noise.

15.3.2.3 Searcher

The searcher is used to search the CPICH of other cells in order to make preparations for a soft handover. The searcher normally implements the following functions: makes slot, frame, and super-frame synchronization with the CPICH and PCCPCH of other cells, de-correlates the BCH information of other cells. For the slot/frame synchronization function, it could be implemented by a slot and frame synchronization block when the current cell synchronization process is finished. The super-frame synchronization will be implemented after the super-frame synchronization of the current cell is finished. When the whole synchronization process is finished, the BCH information could be read.

15.3.2.4 Code Synchronization and Tracking

Code timing is established by an acquisition subsystem and maintained by a tracking subsystem. The pilot channel is a "structural beacon" that does not contain a character stream. It is a timing source used in system acquisition and as a measurement device during handoffs. This carries a data stream of system identification and parameter information used by mobiles during system acquisition (see Figures 15.10 and 15.11).

Timing information of the transmitted signal is acquired in two steps: code acquisition and tracking. Code acquisition involves testing all likely hypotheses for the correct value of the parameter (code phases/sample timing). Tracking is to maintain the receiver timing synchronization within a fraction of a chip time (Figure 15.12).

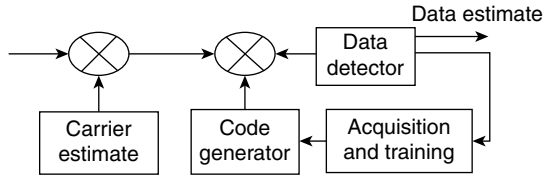


Figure 15.10 Acquisition block

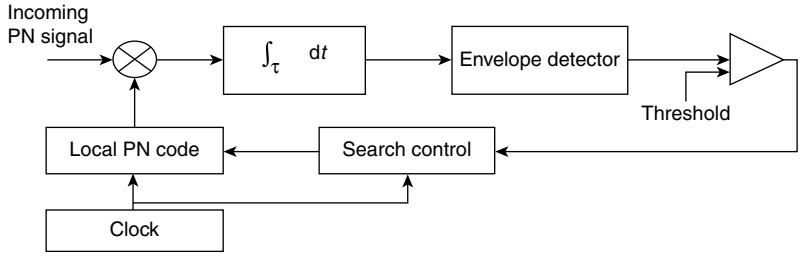


Figure 15.11 Generic acquisition circuit

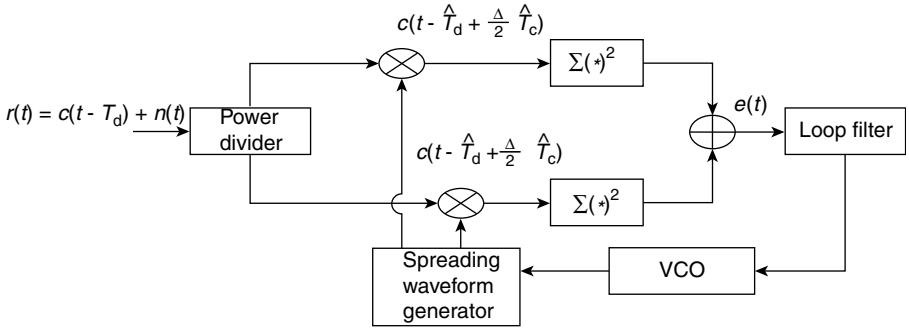


Figure 15.12 Tracking loop

The integration time of the integrator is referred to as the dwell time. The spreading sequence is usually designed to have a small out-of-phase autocorrelation magnitude, but the out-of-phase partial autocorrelation magnitude is not guaranteed to be small in standard sequence design methods. Both the dwell time and the threshold of the acquisition circuit should be designed to minimize the false alarm rate and the miss detection rate.

15.4 Multirate User Data Transmission

WCDMA has a flexible multirate transmission scheme that enables transmission of different types of services using various data rates and quality of service parameters. For example, channel coding type, interleaving depth, and data rate can be varied to achieve the desired quality of service. Please refer to

3GPP standards TS 25.212 and 25.213 for the multirate transmission and multiplexing schemes for the uplink and downlink. Data from transport channels are encoded and thereafter mapped to the physical channels and transmitted over the radio transmission link. The channel coding scheme is a combination of error detection, and error correction scheme then rate matching, interleaving, and transport channels mapping onto the physical channels steps are performed.

Data arrives to the coding/multiplexing unit in the form of transport block sets once every transmission time interval (TTI), which is transport-channel specific and can be 10, 20, 40, or 80 ms. The TTI indicates how often data arrive from the higher layers to physical layer. Multirate transmission consists of following steps: (1) addition of cyclic redundancy check (CRC) to each transport block; (2) concatenation of transport block and segmentation of code block; (3) channel coding; (4) rate matching; (5) insertion of discontinuous transmission (DTX) indication bits; (6) interleaving; (7) segmentation of radio frames; (8) multiplexing of transport channels; (9) segmentation of physical channel; and (10) mapping to physical channels.

Error detection is provided on transport blocks through the CRC. The CRC is 24, 16, 12, 8, or 0 bits, and the higher layers signal what CRC length should be used for each transport channel. After CRC addition, transport block concatenation and code block segmentation are performed. All transport blocks are serially concatenated. If the number of bits in the transmission time interval is larger than the maximum size of the used code block, then code block segmentation is performed after the concatenation of the transport blocks. The maximum size of the code blocks depends on whether convolutional coding, Turbo coding, or no coding is used. The maximum code block sizes are: (a) convolutional coding – 504; (b) Turbo coding – 5114; and (c) no channel coding – unlimited.

Radio frame size equalization is padding the input bit sequence in order to ensure that the output can be segmented in consecutive radio frames of the same size. It is only performed in the uplink. In the downlink, rate matching the output block length is already suitable for radio frame segmentation. When the transmission time interval is longer than 10 ms, the input bit sequence is segmented and mapped onto consecutive radio frames. This enables interleaving over several radio frames improving spectrum efficiency. Because WCDMA provides flexible data rates, the number of bits on a transport channel can vary between the different transmission time intervals. The rate matching adapts this resulting symbol rate to the limited set of possible symbol rates of a physical channel. Rate matching means that bits on a transport channel are repeated or punctured according to the defined rate matching attribute, which is semi-static and can only be changed through higher layer signaling.

In the downlink the transmission is interrupted if the number of bits is lower than the maximum (that is, DTX is used to fill up the radio frame with bits). The insertion point of the DTX indication bit depends on whether fixed or flexible positions of the transport channels in the radio frame are used. It is up to the network to decide for each transport channel whether fixed or flexible positions are used during the connection. DTX indication bits only indicate when the transmission should be turned off, they are not transmitted.

One or more physical channels can be used to transmit the result. When more than one physical channel is used, physical channel segmentation divides the bits among the different channels. After the second interleaving, physical channel mapping is performed.

Transport Format Detection Transport format detection can be performed both with and without transport format combination indicator (TFCI). If a TFCI is transmitted, the receiver detects the transport format combination from the TFCI. When no TFCI is transmitted, so-called blind transport format detection may be used (that is, the receiver side detects the transport format combination using some information, for example, received power ratio of DPDCH to DPCCH or CRC check results).

Channel Coding The channel coding parameters for different transport channel types are mentioned in Table 15.1. The following channel coding schemes can be applied: (1) convolutional coding with constraints of length 9 and coding rate $^1/_3$ or $^1/_2$; (2) Turbo coding; and (3) no channel coding.

The first and second interleavings are both block interleavers with intercolumn permutations.

Table 15.1 Error correction coding parameters

Transport channel type	Coding scheme	Coding rate
BCH	Convolution code	$^1/_2$
PCH		
RACH		
CPCH, DCH, DSCH, FACH		$^1/_3$, $^1/_2$
	Turbo code	$^1/_3$
	No coding	

The turbo coding scheme is a parallel concatenated convolutional code (PCCC) with eight-state constituent encoders. The initial value of the shift registers of the PCCC encoder are all zeros (Figure 15.13). The output of the PCCC encoder is punctured to produce coded bits corresponding to the desired code rate. For rate $^1/_3$, none of the systematic or parity bits are punctured. The Turbo code internal interleaver consists of mother interleaver generations and pruning (Figure 15.14). For an arbitrary given block length of K, one mother interleaver is selected from the 134 mother interleavers set. After the mother interleaver generation, l-bits are pruned in order to adjust the mother interleaver to the block length K. Tail bits T1 and T2 are added for constituent encoders RSC1 and RSC2, respectively.

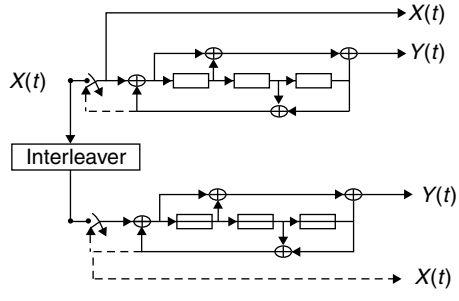


Figure 15.13 Structure of 8-state PCCC encoder

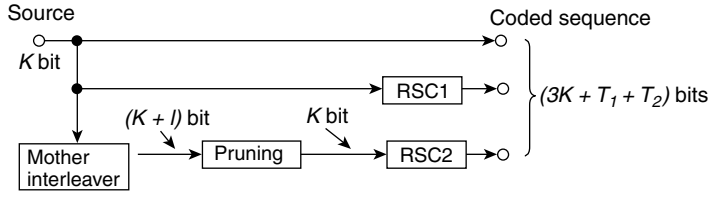


Figure 15.14 Overall 8-state PCCC Turbo coding

15.5 Implementation of UE System Procedures

In this section, implementations for different UE procedures are described in detail. After the switch on and initial boot-up, before any communications can take place, UE needs to synchronize to the transmissions of the UTRAN. This process is called acquisition and search. The higher layers will initiate synchronization. Synchronization to the PCCPCH is achieved as follows.

1. The UE makes measurements on all the frequency bands (for example, 2010–2070 MHz each of 5 MHz BW), unless instructed by higher layers to look at a specific set of frequency channels, to establish where the active cells are located.
2. The UE then proceeds starting with the strongest signal, next strongest and so on, until synchronization is achieved. This procedure is known as cell search and is described in the next section.
3. The UE first achieves slot synchronization and chip synchronization by correlating the received signal against the primary synchronization code (PSC). The primary synchronization code is transmitted every slot and is the same for each cell across the system.
4. On slot synchronization, the UE achieves frame synchronization. Frame synchronization is achieved by correlating the time-aligned received signal with the secondary synchronization codes (SSC). The SSC consists of 16 codes, which are arranged into 64 code groups. Each code group contains a different sequence of 15 out of the 16 codes of the SSC. Hence over one frame, the UE has to determine which subgroup is in use by the cell. Once frame synchronization is achieved, the UE knows which code group the scrambling code is derived from.

   There are 8 primary scrambling codes per code group. The UE performs symbol-by-symbol correlation over the CPICH to determine which primary scrambling code the cell uses.
5. Once it has the primary scrambling code, then it can decode the system information that is carried on the PCCPCH.

The calculations for achieving slot/frame synchronization are intensive. Therefore the UE considers algorithms such as sign-bit correlation, correlation against partial sequences, 1x, 2x sampling, and so on, to minimize the design complexity.

- UE will be in idle mode when power is turned on, until the establishment of RRC connection. The connection of UE is limited to the access stratum. UE in idle mode is identified by identifiers (IMSI, TMSI) and packet TMSI in NAS. UTRAN has no information on UE. The establishment of RRC connection is activated in response to the request from the higher layer of UE or paging request from the network. When UE receives a confirmation message about the establishment of RRC connection from UTRAN, UE goes into the UTRAN connected mode (CELL_FACH OR CELL_DCH). If it fails to establish RRC connection, UE stays in idle mode.
- UE in UTRAN connected mode is assigned a radio network temporary identity (RNTI), which is used to identify UE on the common transport channel. The status of RRC in the UTRAN connected mode depends on the level of the transport channel that can be used by UE, which may be CELL_PCH, CELL_FACH, CELL_DCH OR URA_PCH. If UE in the UTRAN connected mode releases RRC connection, it goes into the idle mode. Figure 15.15 shows the RRC protocol status.
- CELL_DCH – In this state a DPCH is assigned to UE. The UE is identified at the cell level by the current active set. The dedicated transport channel, downlink shared transport channel and the combination of these have also been identified in this state.
- CELL_FACH – No DPCH is assigned to UE. In this state UE receives FACH in downlink, and in uplink. UTRAN is aware of the location of UE at the cell level.
- CELL_PCH – No dedicated channel is assigned to UE. In downlink UE receives PCH via PICH.
- URA_PCH – No dedicated channel is assigned to UE. UE receives in downlink PCH via PICH. UTRAN aware of the UE at the UTRAN registration level.

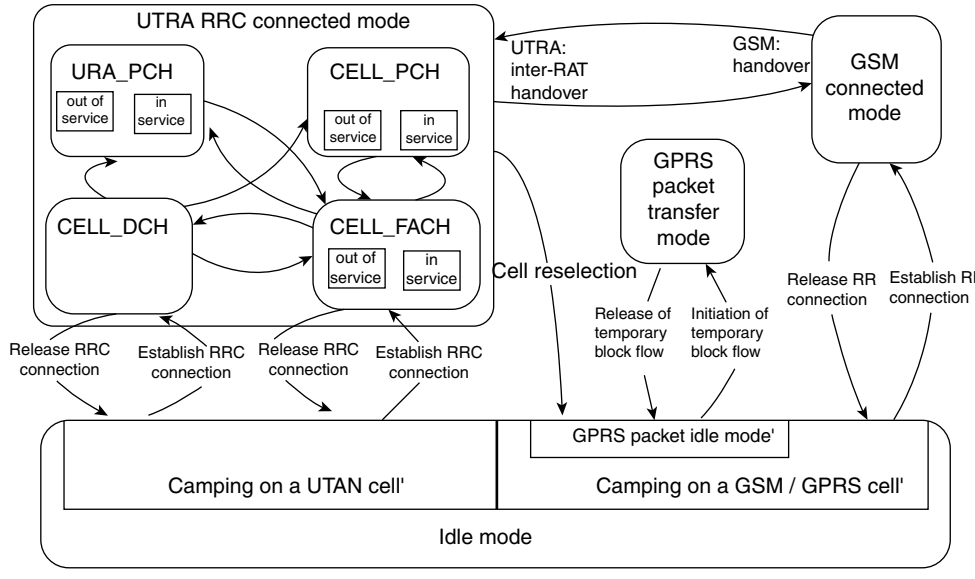


Figure 15.15 Protocol states

In the URA_PCH or CELL_PCH state the UE performs the following actions–

If the UE is "in service area" – maintains up-to-date system information as broadcast by the serving cell, performs cell reselection process, performs a periodic search for higher priority PLMNs, monitors the paging occasions and PICH monitoring occasions, acts on RRC messages received on PCCH and BCCH, performs measurement process according to measurement control information, maintains up-to-date BMC data if it supports cell broadcast service (CBS), runs timer T305 for periodical URA update if the UE is in URA_PCH or for periodical cell update if the UE is in CELL_PCH.

In the CELL_FACH state the UE performs the following actions–

If the UE is "in service area" – maintains up-to-date system information as broadcast by the serving cell, performs cell reselection process, performs measurements process according to measurement control information, runs timer T305 (periodical cell update), selects and configures the RB multiplexing options applicable for the transport channels to be used in this RRC state, listens to all FACH transport channels mapped on the S-CCPCH selected by the UE accordingly, acts on RRC messages received on BCCH, CCCH and DCCH, acts on RRC messages received on, if available, SHCCH (TDD only).

If the UE is "out of service area" – performs cell selection process, runs timers T305 (periodical cell update), and T317 (cell update when re-entering "in service") or T307 (transition to idle mode).

In the CELL_DCH state the UE performs the following actions–

Reads system information broadcast on FACH (applicable only to UEs with certain capabilities and in FDD mode), performs measurement process according to measurement control information, selects and configures the RB multiplexing options applicable for the transport channels to be used in this RRC state, acts on RRC messages received on DCCH, acts on RRC messages received on BCCH (applicable only to UEs with certain capabilities and in FDD mode), acts on RRC messages received on BCCH (TDD only) and, if available, SHCCH (TDD only).

If the UE is "out of service area" – performs cell selection process, runs timer T316, runs timer T305.

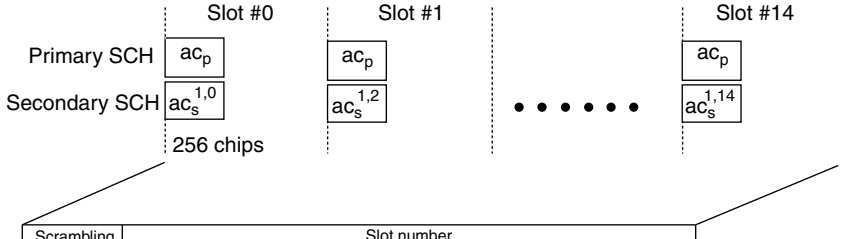
The system information elements are broadcast in system information blocks (SIB). A system information block groups together system information elements of the same nature. Various system

information blocks may have different characteristics, for example, regarding their repetition rate and the requirements on UEs to re-read the system information blocks.

On receiving a message the UE checks that the message is addressed to the UE (for example, by checking the IE "initial UE identity" or the IE "U-RNTI" for messages on CCCH) and discards the messages addressed to other UEs. If it is for itself, then applies integrity check as appropriate, proceeds with error handling, acts upon the IE "RRC transaction identifier," continues with the procedure as specified in the relevant subclause of 3GPP 25.331 standards.

15.5.1 Cell Search Procedure

This procedure is initiated each time the UE is switched on or if UE loses contact with the network. The primary objectives of the cell search are slot synchronization, frame synchronization, and scrambling code group identification (Figure 15.16). The basic function of the primary and secondary SCH is to find out the slot boundary and to enable the identity of scrambling code group used by the BS to be determined by a UE, which will be used to decode other channels from BS.



Scrambling Code Group	Slot number														
	#0	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14
Group 0	1	1	2	8	9	10	15	8	10	16	2	7	15	7	16
Group 1	1	1	5	16	7	3	14	16	3	10	5	12	14	12	10
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Group 62	9	11	12	15	12	9	13	13	11	14	10	16	15	14	16
Group 63	9	12	10	15	13	14	9	14	15	11	11	13	12	16	10

Figure 15.16 Cell search procedure

As discussed in the previous chapter, the synchronization channels has the following structure.

- **PSCH**: This uses PSC (a sequence C_p) and periodically transmits the same C_p sequence in every 15 slots of a frame, and this code is the same for all cells in the network. This helps to detect the presence of nearby BS and to find out the slot boundary. The PSC is generated by modulating a 16-chip code running at 3.84 Mcps by another 16-chip code generated at 240 kcps. The result is a 256 chip sequence at 3.84 Mcps whose auto correlation function can be rapidly found.
- **SSCH**: This uses SSC and SSC changes from slot to slot and repeats over every frame with the same sequence, which is dependent on the cell or sector. This is periodically transmitted in a cell (i) with a different repeating sequence {Csi0 Csi1, . . ., Csi14} aligned with PSCH to indicate both slot number and code group.

In the 256-chip long zone at the beginning of each slot, the primary SCH and secondary SCH are sent and after this, in the same slot, the remaining parts, for example, excluding the 256-chip long part, in each slot

P-CCPCH is transmitted. So, 10% of the slot (256 chips) is occupied by SCH (both PSCH and SSCH transmitted in parallel) and 90% of a slot is occupied by PCCPCH.

A secondary SCH sequence is constructed from an allocation table of SSCs and each element in the table belongs to a set of sequences.

The scrambling code employed in UTRA FDD is a 38 400 chip segment of a 2^{18-1} length Gold code. The scrambling code has I- and Q-components, and there are a total of 8192 codes. For the 3.84 Mcps transmitted chip rate the 38 400 chip code lasts for 10 ms, for example, it lasts for 15 slots. The 8192 codes are divided into 512 sets, each set having 16 codes. These 16 codes are accompanied by a primary and 15 secondary scrambling codes.

The PSC sequence C_p and SSC sequence $C_s^{i,j}$ ($i = 1, \ldots, 512$ and $j = 0,1,2, \ldots, 14$) are used for PSCH and SSCH, respectively, in the downlink direction. Then after this they are not subjected to multiplication by either channelization or scrambling code.

There are 512 codes in the code domain and 38 400 chips (1 frame = 15 slots) in the time domain. We perform 512*38 400 searches to find a cell, then partition the 512 cell-specific codes into 64 groups, each having 8 codes. The two-layer search can reduce a 512 times search into a 64 + 8 times search. Partition the 38 400 timing candidates (chip) into 15 groups (slot), each having 2560 timings candidates. The two-layer search can reduce a 38 400 times search into a 2560 + 15 times search.

The arrangements of scrambling codes are shown in Figure 15.17.

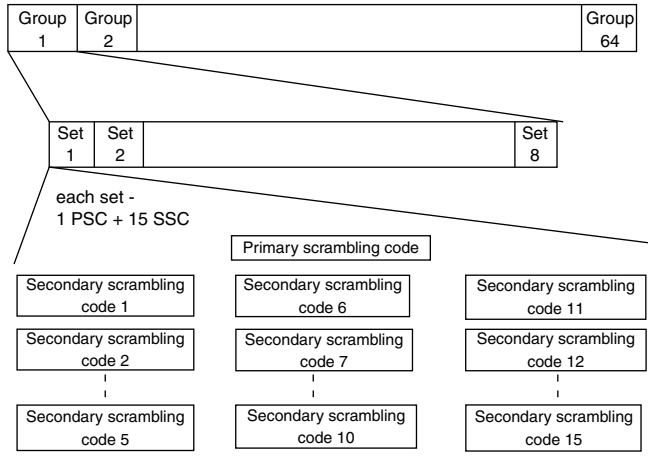


Figure 15.17 Scrambling code hierarchy

In summary we can say the total of 512 groups is divided into 64 groups each having 8 sets. Again each set has 16 scrambling codes, and out of these 1 is a primary scrambling code and 15 are secondary scrambling codes.

Thus, each cell transmits primary SCH and secondary SCH. The primary SCH contains same the data sequence (C_p) and repeats the same code sequence over every slot and this is same for all cells. The secondary SCH repeats over 15 different code sequences, for example, repeats after every frame (15 slots = 1 frame). There are a total of 512 scrambling codes and these are divided into 64 groups, then each group into 8 slots and each slot into 1 primary and 15 secondary codes to reduce the cell search time. Each sector (BS or cell) has a different scrambling code and is identified by this.

Step 1: *Slot synchronization*

During the first step of the cell search procedure the UE uses the SCHs primary synchronization code to acquire slot synchronization for a cell. This is typically done with a single matched filter (or any similar

device) matched to the primary synchronization code (which is the same for all cells and repeats the sequence over every slot). So, when the peak is detected this indicates the slot boundary (Figure 15.18).

The correlation values of PSCH and SSCH are time-averaged to calculate the maximum correlated peak.

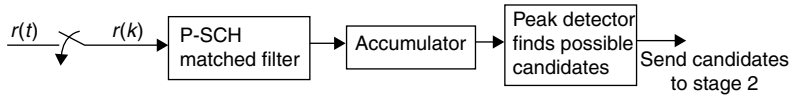


Figure 15.18 Peak detector

Step 2: *Frame synchronization and code-group identification*

There are 15 different 256-chip long secondary synchronization codes which change from slot to slot in a frame as a defined sequence that is associated with the scrambling code group used by the cell. There are 64 scrambling code groups, and with the aid of the second synchronization code, UE knows the actual group number. The received signal is correlated with all 64 possible SSC sequences (over a frame, for example, 15 slots). Knowing the sequence that gives the maximum correlation helps to find the code sequence which identifies the code group used by BS. As a consequence of identifying the SSC (which repeats every frame in the same manner), the receiver knows the frame boundaries and hence frame synchronization is obtained (Figure 15.19).

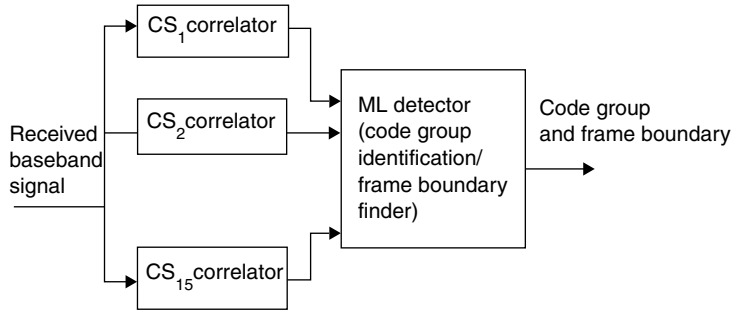


Figure 15.19 ML detector

Step 3: *Scrambling-code identification*

Once the receiver knows the code group, it knows the eight sets codes associated with this code group. Now it has to identify the set that is used by the BS. Each set is associated with a primary scrambling code and 15 secondary codes (each set = 1 PSC + 15 SSC and in a group there are a total of 8 sets, so there are a total of 8 PSC per group). The UE receiver cross-correlates the pilot code with all the 8 PSC of the set in a group. Accordingly, it cross-correlates the common pilot channel with each of the 8 scrambling codes and thereby deems which code is most probable. In this way the receiver determines the correct primary scrambling code. As this code also scrambles the BCH data, so now this can be recovered.

The channelization code used by the P-CPICH is $C_{ch;\,256;0}$, an all logical 1 code, while its scrambling code is the cell's primary scrambling code (Figure 15.20). After the primary scrambling code has been

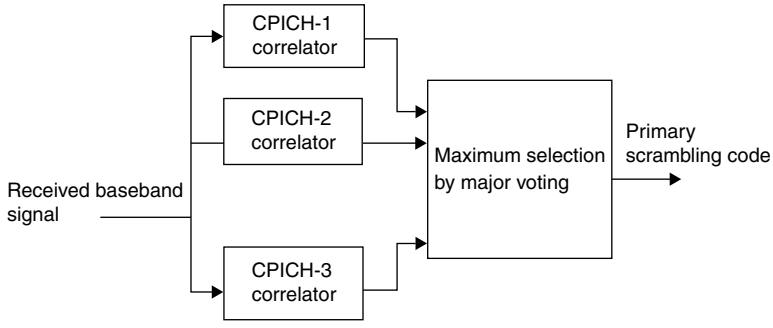


Figure 15.20 Primary scrambling code detection

identified, the primary CCPCH can be detected, and the system and cell specific BCH information can be read.

15.5.2 Power Control

CDMA is an interference limited multiple access system, where all the users in a cell are transmitting using the same frequency band, for example, these can look like co-channel interference. Here the interference level increases as the number of subscribers increases and each user's signal is treated as interference to the other users. Thus is why minimization or controlling the level interference is one of the most required features here, which helps to increase the number of subscribers supported and the bit error rate of the system. Thus in the WCDMA system, transmission power for all the connections is intended to be kept at as minimum a level as possible and later it is increased based on demand.

In order to keep the received power at a suitable level, WCDMA has a fast power control that updates power levels 1500 times every second. By doing this the rapid change in the radio channel is handled. To ensure good performance, power control is implemented in both the uplink and the downlink, which means that both the output powers of the handset and the base station are frequently updated.

Power control also gives rise to a phenomenon called "cell breathing." This is the trade-off between coverage and capacity, which means that the size of the cell varies depending on the traffic load. When the number of subscribers in the cell is low (low load), good quality can be achieved even at a long distance from the base station. On the other hand, when the number of users in the cell is high, the large number of subscribers generates a high interference level and subscribers have to get closer to the base station to achieve good quality.

The nominal power control step size is 1 dB, but multiples of the nominal step sizes can also be used. Power control commands can only be sent every second slot. Open loop power control is used before initiating transmission on the RACH or CPCH. The TPC scheme in the WCDMA system is designed in view of increasing the radio link capacity, quality, and battery life cycle. TPC (transmit power control) used in WCDMA can be broadly divided into two groups – open loop TPC and closed loop TPC (Figure 15.21).

Open loop TPC – Here UE estimates the downlink propagation loss and determines the uplink transmission power using common control channel (CCCH) estimation. In dedicated channels to which closed loop TPC is applied, the initial transmission power is normally decided by open loop TPC. In particular, closed loop TPC can not be applied to uplink CCCH as it is not a channel in which uplink and downlink are used in pairs, so open loop TPC is used.

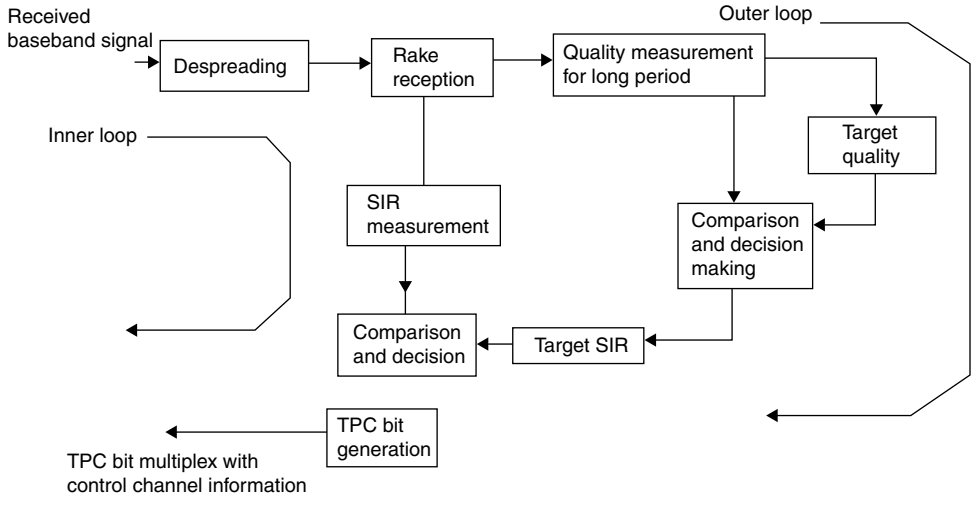


Figure 15.21 TPC control diagram

Closed Loop TPC – In this case the quality of the communication channel is measured at the point of reception and on the basis of measurement results. TPC bits are transmitted using the loop back channel (DPCCH). This consists of two steps – inner loop control and outer loop control.

1. **Inner loop TPC** – Under inner loop TPC in the uplink (downlink) communication channel, BS (UE) measures the received SIR (signal to interference ratio) and compares it with the target SIR, and then it sends the TPC command bits indicating UP (if SIR estimated is above the target SIR value so, increment is required) or DOWN (to decrease). UE (BS) receives the TPC bits and changes the transmission power by 1 dB according to the TPC bit decoding result. Such closed loop control is performed at a slot cycle of 0.667 ms.
2. **Outer loop TC** – The outer loop controls the target SIR so that the bit error rate and block error rate would meet the target values. This involves the measurement of communication link quality over a relatively long period and setting the target SIR at an adequate level to achieve the target quality. Frequency of outer loop control is typically 10–100 Hz.

During soft handover, multiple base stations send commands to a single mobile terminal. The terminal combines the commands and also takes into account the reliability of each command decision in deciding whether or not to increase or decrease the power.

15.5.2.1 Paging Group and Sleep

The paging channel reception is divided into groups or subchannels. The actual number of the paging subchannel to be used by a particular UE is assigned by the network. In this way the UE has to listen for the paging in the assigned time interval. To achieve this, the paging indicator channel (PICH) is split into 10 ms frames, each of which has 300 bits – 288 for paging data and 12 idle bits. At the beginning of each paging channel frame there is a paging indicator (PI), which identifies the paging group data being transmitted. By synchronizing with the paging channels being transmitted the UE is able to turn the receiver circuitry on only when it needs to monitor the paging channel.

15.6 Design of the UMTS Layer-1 Operation States

In this section the design of different layer-1 functionalities in different operational states are described. The UE can be in one of many possible states, indicated by Figure 15.22, at any time, and a different set of L1 functions need to be activated depending upon the state as follows (see Table 15.2):

1. **Idle or Null** – The L1 core control is waiting for a command from the higher layers to enter the RX_P_SCH state, where it shall search for the primary synchronization code.
2. **Rx_P_SCH** – The L1 core control shall search for the strongest transmission of primary synchronization code to establish slot and chip timing, it shall then enter the RX_S_SCH state. The main function performed here shall be the correlation of the received signal with the primary synchronization code.

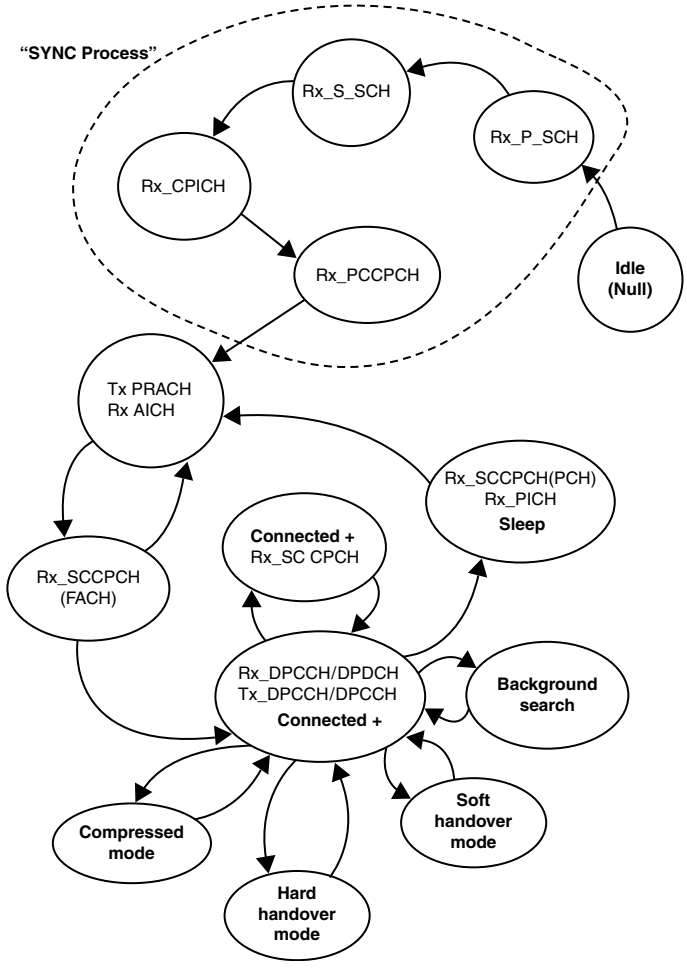


Figure 15.22 States of the layer-1 kernel

Table 15.2 Active functions in various L1K states

L1K state	L1 functions
RX P_SCH	Acquisition and search, AFC, AGC
RX S_SCH	Acquisition and search, AFC, AGC
RX CPICH	Rake management, channel decoding (optional), AFC, AGC
RX PCCPCH	Rake management, channel estimation, channel decoding, AFC, AGC
TX PRACH	Channel encoding, power control, Rake management, AFC, AGC, modulation/ spreading
RX AICH	
RX SCCPCH (FACH)	Rake management, channel decoding, AFC, AGC
Connected states	Rake management, channel estimation, channel decoding, (speech decoding), AFC, AGC, channel coding, (speech coding), power control, modulation and spreading
RX DPCCH/DPDCH	
TX DPCCH/DPDCH	
+ Soft handover	
Sleep states	Rake management, channel estimation, channel decoding, AFC, AGC
RX SCCPCH (PCH)	
RX PICH	

3. **Rx_S_SCH** – The L1 core control shall search for the secondary synchronization code to establish frame timing and identify the code group of the cell. It shall enter the RX_CPICH once the secondary synchronization code has been found.
4. **Rx_CPICH** – The L1 core control shall determine the exact primary scrambling code used by the cell through symbol-by-symbol correlation over the primary CPICH. The primary scrambling code is identified by the code that yields the strongest correlation power. Further confirmation can by performed by channel decoding one of the sub-channels conveyed in the PCCPCH and checking the CRC result for pass/fail. On establishing the primary scrambling code, the L1 core control shall enter the RX_PCCPCH state.
5. **Rx_PCCPCH** – The L1 core control shall remain in this state until the UE has received all of the system information conveyed in the BCH. At this stage the UE is said "to be camped" onto the cell. The higher layer protocols shall notify the L1K to enter the Tx_PRACH & Rx_AICH state to perform location update. The main L1 functions performed here are the Rake management and channel decoding.
6. **Tx_PRACH & Rx_AICH** – The PRACH and AICH are used to establish a link with the UTRAN when no link is active, the UE must be camped onto a cell first. The L1 core control shall transmit the preamble part of the PRACH until it receives an AICH from the UTRAN. Once the AICH has been received, it shall transmit the message part of the PRACH (one frame length) and go to the Rx_SCCPCH state to receive the FACH (logical Channel) from the UTRAN. At this stage most of the L1 processing functions are being used.
7. **Rx_SCCPCH** – The L1 core control shall remain in this state until it receives an acknowledgment from the UTRAN or for a time-out, where it shall return to the previous state. Once the UE receives the FACH message, it shall have all the necessary information to make a dedicated connection with the UTRAN and enter the Connected state. In this mode only receive functions of the L1 are active.
8. **Rx_DPCCH/DPDCH & Tx_DPCCH/DPDCH or Connected** – This is the longest steady state of the L1 core control. All L1 processing functions are active. From time-to-time the L1 core control shall be instructed by higher layers to temporarily enter other states such as Compressed_Mode,

Background_Search, Soft_Hand-over and Hard_Hand-over. In these other states, the L1 functions are more or less the same as the Connected state, with the exception for additional functions required by the particular states. The L1K shall go into "Sleep State," when instructed by the higher layers to disconnect from UTRAN.

9. **Background Search** – In this state, the L1 core control shall search for the neighboring cells P_SCH, S_SCH and CPICH in the same carrier frequency for soft hand-over purposes. In order to achieve a soft hand-over the L1K has made a number of measurements, and reported these to the higher layers. In particular the L1K needs to know the primary scrambling codes and timing offset of the active cells.
10. **Compressed Mode State** – The L1 core control shall enter this state to make power measurements on other carrier frequencies for the purpose of intra-frequency hand-over. Compressed mode can be either in the TX or RX directions, or in both (UTRAN decision). When in compressed mode RX direction, when making a power measurement on another carrier, the information received from the rake shall not be passed to the channel decoder. Similarly the L1 core control shall take into account that different spreading factors are used in this mode to achieve the same traffic throughput in both TX and RX directions.
11. **Soft Hand-over State** – In this state the L1 core control shall start receiving DPCCH/DPDCH from neighboring cells, thus shall assign and manage the rake fingers correspondingly.
12. **Hard Hand-over State** – In this state the L1 core control shall drop the current DPCCH/DPDCH connection with the current cell, and establish a new DPCCH/DPDCH connection with a new cell in a different carrier frequency.
13. **Connected + Rx_SCCPCH** – UE may have to support this state depending on its service capability.
14. **Sleep State (Rx_SCCPCH & Rx_PICH)** – The L1 core control shall enter a power saving mode with this state, and shall periodically wake up to receive the paging message from the UTRAN. All unnecessary L1 functions shall be switched off.

Further Reading

Das, S., Sengupta, C., and Cavallaro, J.R. (1998) Hardware design issues for a mobile unit for next generation CDMA communication systems. SPIE, Advanced Signal Processing Algorithms, Architectures, and Implementations VIII, July 1998, San Diego, CA, pp. 476–487.

3GPP TS 25.211 V2.1.0 (June 1999) *Physical Channels and Mapping of Transport Channels Onto Physical Channels (FDD)*. http://www.3gpp.org/Specifications.

3GPP TS 25.212 V2.0.0 (June 1999) *Multiplexing and Channel Coding (FDD)*. http://www.3gpp.org/Specifications.

3GPP TS 25.213 V2.1.0 (June 1999) *Spreading and Modulation (FDD)*. http://www.3gpp.org/Specifications.

3GPP TS 25.401 V1.0.0 *UTRAN Overall Description*. http://www.3gpp.org/Specifications.

3GPP TS 25.430 V0.1.0 *UTRAN Iub Interface: General Aspects and Principles*. http://www.3gpp.org/Specifications

3GPP TS 26.071. *AMR Speech Codec; General Description*. http://www.3gpp.org/Specifications.

3GPP TS 26.090. *AMR Speech Codec; Transcoding Functions*. http://www.3gpp.org/Specifications.

3GPP TS 26.092. *AMR Speech Codec; Comfort Noise Aspects*. http://www.3gpp.org/Specifications.

3GPP TS 26.093. *AMR Speech Codec; Source Controlled Rate Operation*. http://www.3gpp.org/Specifications.

3GPP TS 26.094. *AMR Speech Codec; Voice Activity Detector (VAD)*. http://www.3gpp.org/Specifications.

3GPP TS 26.101. *AMR Speech Codec; Frame Structure*. http://www.3gpp.org/Specifications.

16

Next Generation Mobile Phones

16.1 Introduction

The mobile phone is becoming the most ubiquitous digital device across the world, and promises to reach two-thirds of the world's population by the end of this decade. Owing to the ever increasing demand for higher data rate, support of more complex applications (such as interactive TV, mobile video blogging), and seamless handover between various networks, the mobile system is continuously evolving from one generation to the next. Recent trends in the wireless market show some specific requirements, such as: (1) reconfigurability, adaptability, programmability, flexibility of the user terminal devices, (2) introduction of more and more power voracious applications, (3) shift towards bursty, high-speed multimedia (for example, streaming video) data traffic, (4) IP-based (wireless Internet), (5) demand for high spectral efficiency, (6) increased demand for mobility (higher speeds + wider range), (7) seamless, ubiquitous wireless (and wired) access across heterogeneous networks, (8) multi-layered *ad hoc* network structures, and (9) cooperation across terminals and sub-networks (for example, multi-hop relaying).

16.1.1 Limitation of Legacy and Current Generation Wireless Technologies (1G, 2G, and 3G)

2G systems are designed for voice centric services and are not suitable for data services. Hence 3G technology was introduced, which offers higher data rate and satisfies reasonable quality of service requirements. However, it has several limitations:

1. Recent research shows that the current UMTS standard has fundamental capacity limitations for high user loads. Also, when the number of active users increases beyond a certain point, the aggregate system capacity starts to decrease. In principle, this deficiency can be fixed by modifying the current standard, but that may increase the complexity by ten times just to double the capacity.
2. 3G performance may not be sufficient to meet the needs of future high-performance applications, such as multimedia, full-motion video, wireless teleconferencing. We need a network technology that extends the 3G capacity by an order of magnitude.
3. There are multiple standards, which are making it difficult to roam and inter-operate across networks. We need global mobility and service portability.
4. 3G is based primarily on a wide-area concept. We need hybrid networks that utilize both the wireless LAN (hot spot) concept and cell or base-station wide area network design.

Mobile Handset Design Sajal Kumar Das

5. Wider bandwidth is needed.
6. Researchers have come up with spectrally more efficient modulation schemes that can not be retrofitted into the 3G infrastructure.
7. A need for all digital packet networks that utilize IP in its full form with converged voice and data capability.
8. Present system can not support seamless handover and mobility amongst heterogeneous IP networks with both cellular networks and wireless local area networks (WLANs), which is one of the driving factors for 4G in telecommunications networks and systems.

16.1.2 Need for 4G Wireless Technology

4G is an initiative to move beyond the limitations and problems of 3G, which is having trouble with deployment and meeting its promised performance and throughput. However, 4G is intended to provide high-speed, high-capacity, low-cost per bit, IP-based services. The 4G mobile communication systems are projected to solve the still-remaining problems of 3G systems and to provide a wide variety of new services, from high-quality voice to high-definition video to high-data-rate wireless channels. One term used to describe 4G is:

$$4G = C.A^5$$

where, C is communication, A is anytime, anywhere, with anyone, on any device, through any network.

16.1.3 Evolution of 4G

The 4G (fourth generation) mobile communication systems are projected to solve the still-remaining problems of 3G (third generation) systems and to provide a wide variety of new services, from high-quality voice to high-definition video to high-data-rate wireless channels. Based on the developing trends of mobile communication, 4G will have broader bandwidth, higher data rate, and smoother and quicker handoff and will focus on ensuring seamless service across a multitude of wireless systems and networks. The key concept is integrating the 4G capabilities with all of the existing mobile technologies through the use of advanced technologies. The different key parameters associated with the 3G and 4G systems are mentioned in Table 16.1.

Generally speaking, 4G is an evolution not only to move beyond the limitations and problems of 3G, but also to enhance the quality of services, to increase the bandwidth and to reduce the cost of the resource. 4G will be supported by IPv6, OFDM, MC-CDMA, LAS-CDMA, UWB, and Network-LMDS. 4G technologies are typified by high rates of data transmission and packet-switched transmission protocols. 3G technologies, by contrast, are a mix of packet and circuit-switched networks. Today, the 4G system is evolving mainly through 3G LTE and WiMAX systems.

The Third Generation Partnership Project (3GPP) is presently developing 3GPP LTE technology, which is an upgrade to existing GSM networks. This is one of the most advanced mobile communication technologies to date, and is currently undergoing 4G technology standardization by the 3GPP. This is the most likely technology to become the 4G standard, as many of the world's major operators and telecommunications companies are members of the LTE/SAE (Long Term Evolution/System Architecture Evolution) Trial Initiative (LSTI). It has the following advantages:

1. Supports high data rate, with peak data rates of 100 Mbps downlink and 50 Mbps uplink.
2. It makes CDMA and GSM debates moot.
3. It offers both FDD and TDD duplexing. This means the upload and download speeds do not have to be synchronous, so operators can better optimize their networks to use more upload channels.
4. It will offer lower latency, which makes real-time interaction feasible.

Table 16.1 Key differences between 3G and 4G systems

Parameter	3G system	4G system
Major requirement driving architecture	This is predominantly voice driven and data is always add on	Converged data and voice over IP
Network architecture	Wide area cell-based	Hybrid – integration of wireless LAN (WiFi, Bluetooth) and wide area cellular network
Frequency band	1.8–2.5 GHz	2–8 GHz
Bandwidth	5–20 MHz	100 MHz (or more)
Data rate	384 Kbps to 2 Mbps	20–100 Mbps in mobile mode
Mobile top speeds	200 km/h	200 km/h
Forward error correction	Convolutional rate $^1/_2$, $^1/_3$	
Access technologies	W-CDMA	MC-CDMA or OFDM (TDMA)
Switching design basis	Circuit and packet	All digital with packetized voice
IP	A number of air link protocols, including IP 5.0	All IP (IP6.0)
Component design	Optimized antenna design, multi-band adapters	Smarter antennas, software multi-band and wideband radios

On the other hand, WiMAX has certain advantages mainly over the fiber to the home (FTTH) technology and offers broadband Internet access and IPTV to residential subscribers. It offers a cost effective infrastructure with efficient use of the spectrum. Given the QoS, security and reliability mechanisms built into WiMAX, the users will find WiMAX VoIP as good as or even better than voice services from the telephone company. However, for WiMAX, there is a limitation of the wireless bandwidth, which may be an issue in high density areas.

We will next briefly discuss these two upcoming technologies: 3GPP LTE and WiMAX.

16.2 3GPP LTE

Currently, 3GPP is defining the long-term evolution (LTE), which allows UMTS operators to use new and wider spectra (up to 20 MHz), with higher data rates, lower latency and higher capacity, to provide an enhanced broadband experience in high-data demand and densely populated urban areas. Many of the targets for LTE are similar to those for the continuing development of high speed packet access (HSPA) – generally known as HSPA + – although LTE has some specific additional capabilities, such as flexible channel bandwidths and the advantages of orthogonal frequency division multiple access (OFDMA). Both LTE and HSPA + are being developed in 3GPP Release 8 specifications, which also include work on the evolution of EDGE. LTE utilizes a new core network, the evolved packet core (EPC), which allows for a flatter IP-based architecture. Throughout the design of LTE and EPC, emphasis has been placed on interoperability with existing 3GPP technologies, such as UMTS and GSM. This will ensure that HSPA + and LTE co-exist seamlessly. Again, within the 3GPP specifications for LTE, the evolved radio access network is split into two parts: the evolved UMTS terrestrial radio access (E-UTRA), which describes the mobile part of LTE, and the evolved UMTS terrestrial radio access network (E-UTRAN), which describes the base station part containing the evolved Node B (eNB).

LTE incorporates many key features that enable operators to provide an enhanced broadband experience and these are: (1) OFDMA on the DL and SC-FDMA on the UL using advanced antenna techniques, (2) MIMO, (3) SDMA, (4) beam-forming, (5) single frequency network multicast services, (6) all-IP packet-optimized network architecture, and (7) enhanced interference control.

16.2.1 LTE Design Goals

As discussed earlier, the LTE system is designed to meet some specific objectives:

1. **Support of scalable bandwidths** – LTE systems should support bandwidths of 1.25, 2.5, 5.0, 10.0, and 20.0 MHz.
2. **Data rate support** – The system should support the following peak data rate that scales with the system bandwidth:
   a. Downlink (2 Ch MIMO) peak rate of 100 Mbps in 20 MHz channel
   b. Uplink (single Ch TX) peak rate of 50 Mbps in 20 MHz channel.
   This is given in Table 16.2.

Table 16.2 Peak data rate in LTE

Downlink peak data rates (64-QAM)			
Antenna configuration	SISO	2 × 2 MIMO	4 × 4 MIMO
Peak data rate (Mbps)	100	172.8	326.4
Uplink peak data rates (single antenna)			
Modulation depth	QPSK	16-QAM	64-QAM
Peak data rate (Mbps)	50	57.6	86.4

Support is intended for even higher data rates, up to 326.4 Mbps in the downlink, using multiple antenna configurations.

3. **Supported antenna configurations** – This will support different antenna configuration, such as (a) downlink: 4 × 2, 2 × 2, 1 × 2, 1 × 1, and (b) uplink: 1 × 2, 1 × 1.
4. **Spectrum efficiency** – (a) downlink: 3 to 4 × HSDPA Release 6, and (b) uplink: 2 to 3 × HSUPA Release 6.
5. **Latency** – (a) Control-plane: <50–100 ms to establish U-plane, and (b) user-plane: <10 ms from UE to server.
6. **Mobility** – LTE system is optimized for low speeds (0–15 km/h), but will still provide high performance up to 120 km/h and with support for mobility it can be maintained up to 350 km/h. 3GPP are considering the support for even higher speeds up to 500 km/h.
7. **Coverage** – (a) Full performance up to 5 km, (b) slight degradation 5–30 km, and (c) operation up to 100 km should not be precluded by the standard.

16.3 LTE System Design

To address these objectives, several changes are required in the different parts of the LTE system infrastructure.

16.3.1 RF

The variable channel bandwidths are specified in LTE to increase the system's flexibility and capability, but this adds more complexity to the RF system design. The use of multiple antenna configurations and

OFDMA to support high data rates adds a further complication. There should be careful design trade-offs to cover each critical part of the transmitter and receiver chain.

16.3.2 Layer-1/Baseband

To support the high data rate, exceptionally large amounts of processing power are needed, particularly in the baseband, where all the error handling and signal processing occurs.

16.3.2.1 LTE Physical Layer Design

The LTE PHY is a highly efficient means of conveying both data and control information between an enhanced base station (eNodeB) and the mobile user equipment (UE). The LTE PHY employs orthogonal frequency division multiplexing (OFDM) and multiple input multiple output (MIMO) data transmission. The LTE PHY uses orthogonal frequency division multiple access (OFDMA) on the downlink (DL) and single carrier – frequency division multiple access (SC-FDMA) on the uplink (UL). Single carrier frequency division multiple access (SC-FDMA) combines the low peak-to-average power ratio (PAPR) of single-carrier systems with the multi-path resistance and flexible subcarrier frequency allocation offered by OFDMA. In OFDMA, users are allocated a specific number of subcarriers for a predetermined amount of time. In the LTE specification, these are referred to as physical resource blocks (PRBs). The PRBs have both a time and frequency dimension. The allocation of PRBs is handled by a scheduling function at the 3GPP base station (eNodeB).

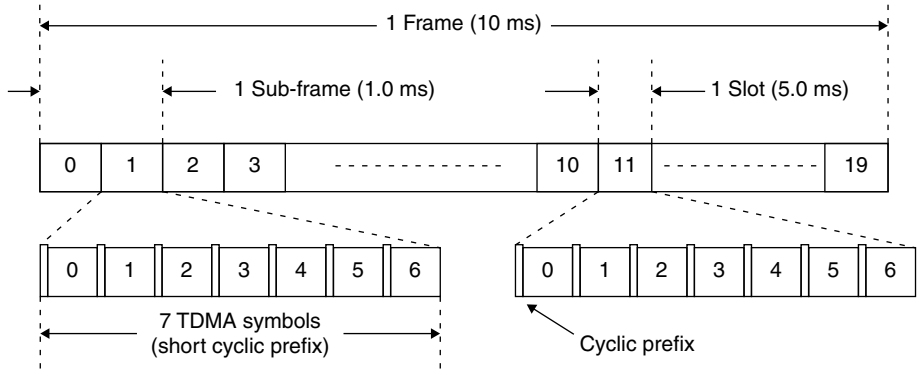


Figure 16.1 LTE generic frame structure

The LTE frames are 10 ms in duration. Then each frame is again divided into 10 subframes, and each subframe is 1.0 ms long. Each subframe is further divided into two slots, each of 0.5 ms duration (Figure 16.1). Each slot consists of either 6 or 7 ODFM symbols, depending on whether the normal or extended cyclic prefix is employed. The total number of available subcarriers depends on the overall transmission bandwidth of the system. As mentioned earlier, the LTE specifications define parameters for system bandwidths from 1.25 to 20 MHz and the subcarrier bandwidth = 15 kHz; the physical resource block (PRB) bandwidth = 180 kHz. A PRB is defined as consisting of 12 consecutive

subcarriers for one slot of duration (0.5 ms), and is the smallest element of resource allocation assigned by the base station scheduler.

The transmitted downlink signal consists of N_{BW} subcarriers for a duration of N_{symb} OFDM symbols. It can be represented by a resource grid. Each box within the grid represents a single subcarrier for one symbol period and is referred to as a resource element. Note that in MIMO applications, there is a resource grid for each transmitting antenna.

Downlink Channel Structure

The service access points for L2/3 layers are transport channels. These are then mapped to physical channels. The DL supports physical channels, which convey information from higher layers in the LTE stack or vice versa, and physical signals which are for the exclusive use of the PHY layer. Depending on the assigned task, physical channels and signals use different modulation and coding parameters.

Downlink Physical Channels

Three different types of physical channels are defined for the LTE downlink. One common characteristic of physical channels is that they all convey information from higher layers in the LTE stack. However, physical signals convey information that is used exclusively within the PHY layer. The LTE downlink physical channels are described below.

1. **Physical Downlink Shared Channel (PDSCH)** – This channel is designed for high data rate support and is utilized basically for data and multimedia transport. The modulation techniques used for this channel are QPSK, 16QAM or 64QAM. Spatial multiplexing is also used in the PDSCH. In fact, spatial multiplexing is exclusive to the PDSCH. It is not used on either the PDCCH or the CCPCH.
2. **Physical Downlink Control Channel (PDCCH)** – This channel conveys UE-specific control information so robustness rather than maximum data rate is considered here. The modulation used for this channel is QPSK. The PDCCH is mapped onto resource elements in up to the first three OFDM symbols in the first slot of a subframe.
3. **Common Control Physical Channel (CCPCH)** – For this channel robustness is also more important than maximum data rate. The CCPCH carries cell-wide control information. The modulation used for this channel is QPSK. In addition, the CCPCH is transmitted as close to the center frequency as possible. CCPCH is transmitted exclusively on the 72 active subcarriers centered on the DC subcarrier. Control information is mapped to resource elements (k, l) where k refers to the OFDM symbol within the slot and l refers to the subcarrier. CCPCH symbols are mapped to resource elements in increasing order of index k first, then l.

Physical Signals

Although physical signals use assigned resource elements, unlike physical channels, physical signals do not convey information to/from higher layers. There are two types of physical signals: (1) reference signals used to determine the channel impulse response (CIR), (2) synchronization signals, which convey network timing information.

Physical channels are mapped to specific transport channels.

Downlink Transport Channels

Transport channels are included in the LTE PHY and act as service access points (SAPs) for higher layers. Transport channels provide the following functions: (1) structure for passing data to/from higher layers, (2) mechanism by which higher layers can configure the PHY, (3) status indicators (packet error, CQI, etc.) to higher layers, and (4) support for higher-layer peer-to-peer signaling. The different downlink transport channels are described below and given in Table 16.3.

Table 16.3 Logical and transport channel structure

Logical channels (characterized by the information that is transferred)	Control (carry control plane information)	Broadcast control channel (DL channel for broadcasting system control info)
		Paging channel (DL channel for transferring paging)
		Common control channel (UL channel for transferring control info. and used by UE without RRC connection)
		Multi-cast control channel (DL point-to-multipoint channel for transmitting MBMS control info.)
		Dedicated control channel (DL point-to-point bi-directional channel for exchanging control info. and used by UEs with RRC connection)
	Traffic (carry user plane information)	Dedicated traffic channel (bi-directional channel dedicated to single UE)
		Multi-cast traffic channel (DL point-to-multipoint channel for transmission of MBMS data)
Transport channels (characterized by how the data are transferred over the radio interface)	Downlink channels	Broadcast channel (fixed transport format)
		Downlink shared channel (HARQ, dynamic link adaption, support for UE DRX, dynamic and semi-static resource allocation)
		Paging channel (required to be broadcast)
		Multicast channel (support for SFN combining and semi-static resource allocation)
	Uplink channels	Uplink shared channel (HARQ, dynamic link adaption, support for UE DRX, dynamic and semi-static resource allocation)
		Random access channel (limited control information, collision risk)
Physical Channels	Downlink	– Physical broadcast channel (PBCH)
		– Physical control format indicator channel (PCFICH)
		– Physical downlink control channel (PDCCH)
		– Physical hybrid ARQ indicator channel (PHICH)
		– Physical downlink shared channel (PDSCH)
		– Physical multicast channel (PMCH)
	Uplink	– Physical uplink control channel (PUCCH)
		– Physical uplink shared channel (PUSCH)
		– Physical random access channel (PRACH)

1. **Broadcast Channel (BCH)** – This is of fixed format and must be broadcast over the entire coverage area of the cell.
2. **Downlink Shared Channel (DL-SCH)** – This supports hybrid ARQ (HARQ), dynamic link adaptation by varying modulation, coding and transmit power, dynamic and semi-static resource allocation and discontinuous receive (DRX) for power save. This is suitable for transmission over the entire cell coverage area and for use with beam forming.
3. **Paging Channel (PCH)** – This is for paging intimation and helps to support UE DRX. This is required for broadcast over the entire cell coverage area and mapped to dynamically allocated physical resources.
4. **Multicast Channel (MCH)** – This is required to broadcast over the entire cell coverage area. It helps to support for MB-SFN and semi-static resource allocation.

In Figure 16.2, the mapping of transport channels to downlink physical channels is shown.

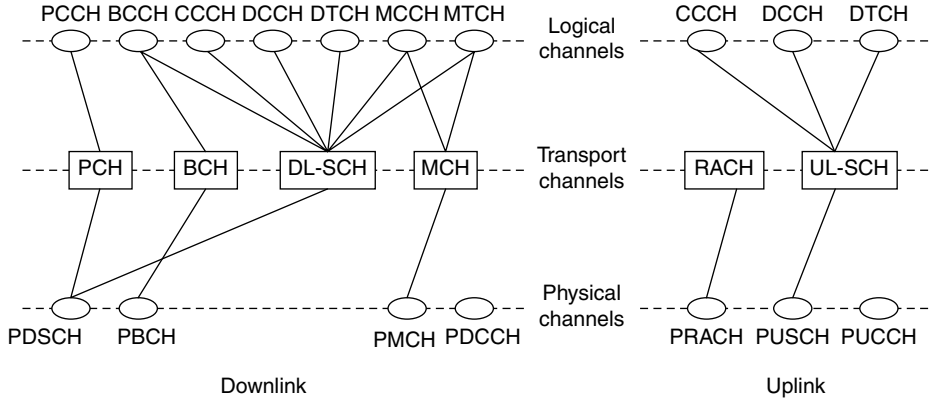


Figure 16.2 Mapping of DL and UL logical to transport channels to physical channels

Uplink Channel Structure

In FDD applications, the uplink uses the same generic frame structure as the downlink. It also uses the same subcarrier spacing of 15 kHz and PRB width (12 subcarriers). Uplink PRBs are assigned to UE by the base station scheduler via the downlink CCPCH. Uplink PRBs consist of a group of 12 contiguous subcarriers for one slot time duration.

Uplink Physical Channels

An uplink physical channel is used to transmit information originating in layers above the PHY. The different UL physical channel is described below.

Physical Uplink Shared Channel (PUSCH) – Resources for the PUSCH are allocated on a subframe basis by the UL scheduler. Subcarriers are allocated in multiples of 12 (PRBs) and may be hopped from subframe to subframe. The PUSCH can employ QPSK, 16QAM or 64QAM modulation.

Uplink Physical Signals

Uplink physical signals are used within the PHY and do not convey information from higher layers. Two types of UL physical signals are defined: the reference signal and the random access preamble.

1. **Uplink Reference Signal** – Actually, there are two variants of the UL reference signal. The demodulation signal facilitates coherent demodulation. It is transmitted in the fourth SC-FDMA symbol of the slot and is the same size as the assigned resource. There is also a sounding reference signal used to facilitate frequency dependent scheduling. Both variants of the UL reference signal are based on Zadhoff–Chu sequences.
2. **Random Access Preamble** – The random access procedure involves the PHY and higher layers. At the PHY layer, the cell search procedure is initiated by transmission of the random access preamble by the UE. If successful, a random access response is received from the base station. The random access preamble format is shown in Figure 16.3. It consists of a cyclic prefix, a preamble, and a guard time during which no signal is transmitted.

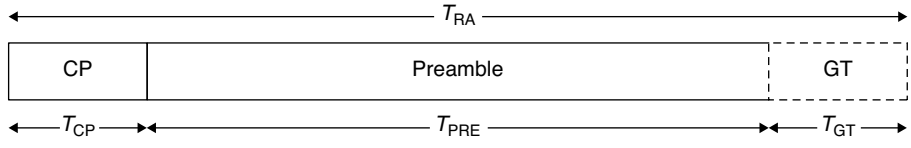


Figure 16.3 Random access preamble part

For the generic frame structure, the timing parameters are: TRA – 30720 *TS*, TGT – 3152 *TS*, TPRE – 24576 *TS*, where *TS* = period of a 30.72 MHz clock. Random access preambles are derived from Zadoff–Chu sequences. They are transmitted on blocks of 72 contiguous subcarriers allocated for random access by the base station. In FDD applications, there are 64 possible preamble sequences per cell. The exact frequency used for transmission of the random access preamble is selected from available random access channels by higher layers in the UE. Other information provided to the PHY by higher layers includes: (1) available random access channels, (2) preamble format (the preamble sequences), (3) initial transmission power, (4) power ramp step size, and (5) maximum number of retries.

Uplink Transport Channels

As in downlink, uplink transport channels act as service access points for higher layers. Different types of UL transport channels and their characteristics are described below.

1. **Uplink Shared Channel (UL-SCH)** – This channel supports the possible use of beam forming, dynamic link adaptation (varying modulation, coding and/or TX power), HARQ, and dynamic as well as semi-static resource allocation.
2. **Random Access Channel (RACH)** – This channel supports transmission of limited control information and the possible risk of collision.

Mapping of uplink physical channels to transport channels is shown in Figure 16.2.

Physical Channel Processing

The physical channel information is processed by using different functional blocks (as shown in Figure 16.4) and algorithms, such as bit scrambling, modulation, layer mapping, CDD precoding, and resource element assignment. Layer mapping and precoding are related to MIMO applications. Basically, a layer corresponds to a spatial multiplexing channel. MIMO systems are defined in terms of $N_{\text{transmitters}} \times N_{\text{receivers}}$. For LTE, defined configurations are 1×1, 2×2, 3×2, and 4×2. This means there could be as many as four transmitting antennas and a maximum of two receiver antennas for spatial multiplexing the data streams. Precoding is also used in conjunction with spatial multiplexing. Recall that MIMO exploits the multipath to resolve independent spatial data streams. In other words, MIMO systems require a certain degree of multipath for reliable operation. In a noise-limited environment with low multipath distortion, MIMO systems can actually become impaired.

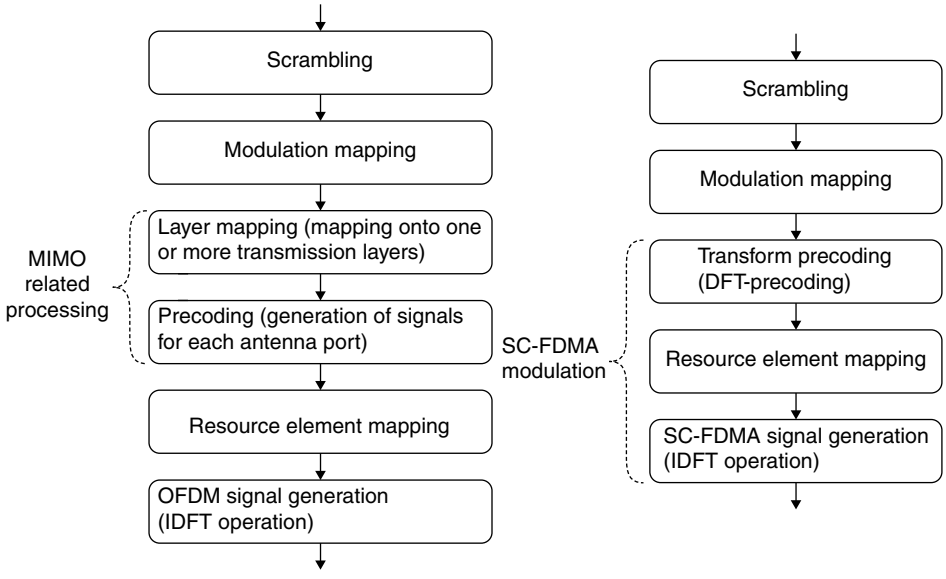


Figure 16.4 Physical layer processing blocks

Uplink Coding

The UL-SCH uses the same rate $^1/_3$ Turbo encoding scheme (two 8-state constituent encoders and one internal interleaver) as the DL-SCH.

Downlink Channel Coding

Different coding algorithms are employed for the DL physical channels. For the common control channel (CCPCH), modulation is restricted to QPSK. For CCPCH the convolutional coding has been selected. The PDSCH uses up to 64 QAM modulation. On the PDSCH, higher-complexity modulation is employed to achieve the highest possible downlink data rates. It uses QPSK, 16QAM, or 64QAM depending on channel conditions. As a result, coding gain is emphasized over latency. Rate $^1/_3$ Turbo coding has been selected for the PDSCH.

Modulation Parameters

The downlink and uplink modulation parameters (including normal and extended CP length) are identical. In the uplink, depending on the channel quality the data is mapped onto QPSK, 16QAM, or 64QAM signal constellations. However, rather than using the QPSK/QAM symbols to directly modulate subcarriers (as is the case in OFDM), uplink symbols are sequentially fed into a serial/parallel converter and then into an FFT block. The result at the output of the FFT block is a discrete frequency domain representation of the QPSK/QAM symbol sequence. The discrete Fourier terms at the output of the FFT block are then mapped to subcarriers before being converted back into the time domain (IFFT). The final step prior to transmission is appending a CP.

16.3.3 Protocol Architecture

The control plane and the user plane protocol stacks layers are shown in Figure 16.5. In the control plane, the NAS protocol, which runs between the MME and the UE, is used for control purposes, such as network attach, authentication, setting up of bearers, and mobility management. All NAS messages are ciphered and integrity protected by the MME and UE. Layer-3 handles the main service connection protocols. The RRC layer in the eNB makes handover decisions based on neighbor cell measurements sent by the UE, pages for the UEs over the air, broadcasts system information, controls UE measurement reporting such as the periodicity of channel quality information (CQI) reports, and allocates cell-level temporary identifiers to active UEs. It is also responsible for setting up and maintenance of radio bearers. LTE layer-2 is split into three sublayers, including the medium access control (MAC) and packet data convergence protocol (PDCP). In the user plane, the PDCP layer is responsible for compressing/decompressing the headers of user plane IP packets using robust header compression (ROHC) to enable efficient use of the air-interface bandwidth.

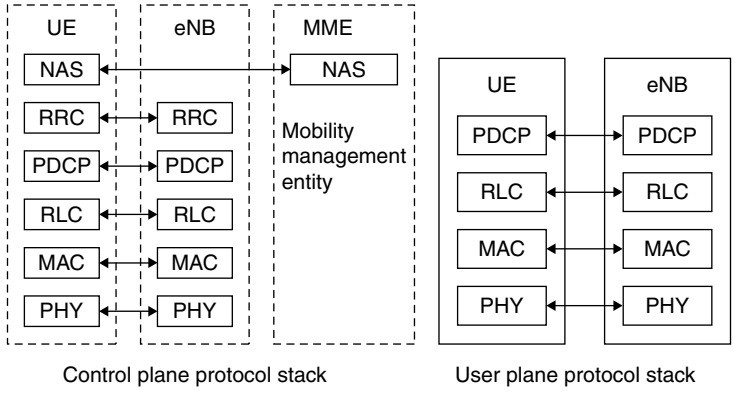


Figure 16.5 Protocol architecture

Design challenges at this layer will be the handling of significant amounts of data in the PDCP and implementation of the 2 ms MAC turnaround time. Detailed specifications for both of these layers are still under discussion.

16.3.4 Procedures

1. **Cell search procedure:** Based on BCH (broadcast channel) signal and hierarchical SCH (synchronization channel) signals, the mobile terminal or user equipment (UE) acquires time and frequency

synchronization with a cell and detects the cell ID of that cell. P-SCH (primary-SCH) and S-SCH (secondary-SCH) are transmitted twice per radio frame (10 ms) for FDD. The following steps are performed during the cell search procedure:
   a. 5 ms timing identified using P-SCH
   b. radio timing and group ID found from S-SCH
   c. full cell ID found from DL RS
   d. Decode BCH.
2. **Synchronization procedures:** This helps to perform the following:
   a. radio link monitoring
   b. inter-cell synchronization for MBMS
   c. transmission timing adjustments.

There are some other procedures executed in different optional modes, such as power control for DL and UL, UE procedure for CQI (channel quality indication) reporting, UE procedure for MIMO feedback reporting, UE sounding procedure, and so on.

16.4 IEEE 802.16 System

In recent years there has been increasing interest shown in wireless technologies for subscriber access, as an alternative to a traditional twisted-pair local loop. By enabling quick and relatively inexpensive deployment of the broadband services infrastructure, the IEEE 802.16 Standards for wireless broadband access have the potential to solve the long-standing "last mile" problem that has plagued the data and telecom carrier industries. The term WiMAX (Worldwide Interoperability for Microwave Access) has become synonymous with the IEEE 802.16 Wireless Metropolitan Area Network (MAN) air-interface standard. Work on 802.16 started in July 1999. Four years into its mission, the IEEE 802.16 Working Group on Broadband Wireless Access delivered a base and some follow-up standards.

IEEE 802.16: (*Air Interface for Fixed Broadband Wireless Access Systems*) Standard was approved in December 2001 for wireless MAN. This defines an MAC layer and several physical layer specifications. The MAC supports frequency-division-duplex (FDD) and time-division-duplex (TDD), as well as real-time adaptive modulation, coding and single-carrier modulation. The physical layer of the standard covers the spectrum from 10 to 66 GHz, which includes the LMDS bands.

IEEE 802.16.2: This was published in 2001, specifies a "recommended practice" to address the operation of multiple, different broadband systems in the 10–66 GHz frequency range.

IEEE 802.16a: In January of 2001, the IEEE approved an amendment to 802.16, termed 802.16a, which adds to the original standard operation in licensed and unlicensed frequency bands. It is a completed amendment that extends the physical layer to the 2–11 GHz spectrum range and specifies three possible modulations: single carrier, 256 OFDM, and orthogonal frequency division multiple access (OFDMA).

IEEE 802.16b: This is proposed to have a license exempt frequency band of 5–6 GHz.

IEEE 802.16c: This was approved in December 2002, and is aimed at improving interoperability by specifying system profiles in the 10–66 GHz range.

IEEE 802.16d: This contains system profiles for 802.16a (2–11 GHz) implementations.

IEEE 802.16e: Mobile wireless broadband up to vehicular speeds in licensed bands from 2 to 6 GHz. Enables roaming for portable clients within and between service areas. This is still very early in the process.

A detailed comparison between the various IEEE802.16 Standards is shown in the Table 16.4.

Table 16.4 Comparison of different IEEE 802.16 standards

Attribute	802.16	802.16a	802.16d	802.16e
Evolution	Freq. range 10–66 GHz, line of sight, data rate up to 134 Mbps	Freq. range 2–11 GHz, line of sight, data rate up to 75 Mbps	Interoperability and system profile improved MAC to support portability and idle mode, enhanced sub-channlization of OFDMA, HARQ, enhanced MIMO, and AAS	Complete mobility management handoff, scalable OFDMA, sleep mode, BCMCS, paging, enhanced idle mode, enhanced MIMO
Spectrum	10–66 GHz	2–11 GHz	2–11 GHz	2–6 GHz
Channel conditions	Line of sight	Non-line of sight	Non-line of sight	Non-line of sight
Raw bit rate	32–134.4 Mbps	1–75 Mpbs	1–75 Mbps	12.5 Mbps
Modulation	QPSK, 16-QAM, 64-QAM; single carrier	QPSK, 16-QAM, 64-QAM, (256-QAM); single carrier, 256 OFDM, 2048 OFDMA	QPSK, 16-QAM, 64-QAM, (256-QAM); single carrier, 256 OFDM, 2048 OFDMA	QPSK, 16-QAM, 64-QAM, (256-QAM); single carrier, 256 OFDM, scalable OFDMA
UL/DL duplexing	TDD/FDD	TDD/FDD	TDD/FDD	TDD/FDD
Channel bandwidth	20, 25 and 28 MHz	1.25–20 MHz	1.25–20 MHz	1.25–20 MHz
Cell radius	Typical: 2.5 km	Typical: 6–9 km	Typical: 6–9 km	Typical: 6–9 km

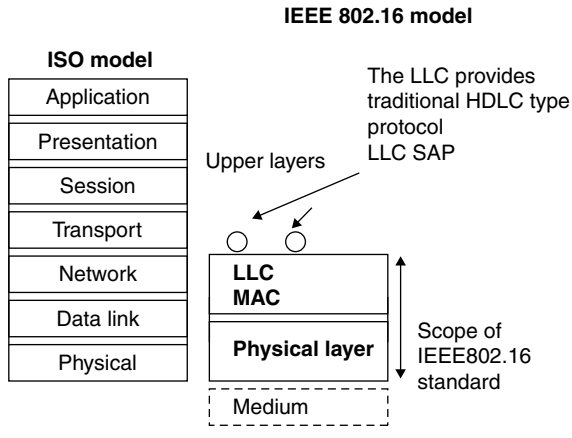


Figure 16.6 Layered architecture

In WiMAX the customer premises equipment (CPE) is a telephone or other service provider equipment that is located on the customer's premises (physical location) rather than on the provider's premises or in between. Telephone handsets, cable TV set-top boxes, and digital subscriber line routers are examples of such devices. Historically, this term referred to equipment placed at the customer's end of the telephone line and was usually owned by the telephone company. Today, almost any end-user equipment can be called customer premise equipment and it can be owned by the customer or by the provider.

16.4.1 IEEE 802.16 Architecture Overview

The IEEE 802.16 Standard is organized into a three-layer architecture as shown in Figure 16.6. Only the physical and data link layer are within the scope of this standard.

16.4.1.1 Physical Layer

The physical layer is the lowest layer, which specifies the frequency band, the modulation scheme, error-correction techniques, synchronization between transmitter and receiver, data rate, and the time-division multiplexing (TDM) structure. A simple block diagram of a wireless MAN transmitter–receiver is shown in Figure 16.7a.

Error Correction Scheme

The forward error correction (FEC) used in 802.16 is Reed–Solomon GF (256), with variable block size and error correction capabilities. This is paired with an inner block convolution code to robustly transmit critical data, such as frame control and initial access. The FEC options are paired with quadrature phase shift keying (QPSK), 16-state quadrature amplitude modulation (16-QAM), and 64-state QAM (64-QAM) to form burst profiles of varying robustness and efficiency. If the last FEC block is not filled, that block may be shortened. Shortening in both the uplink and downlink is controlled by the BS and is implicitly communicated in the uplink map (ULMAP) and the downlink map (DL-MAP).

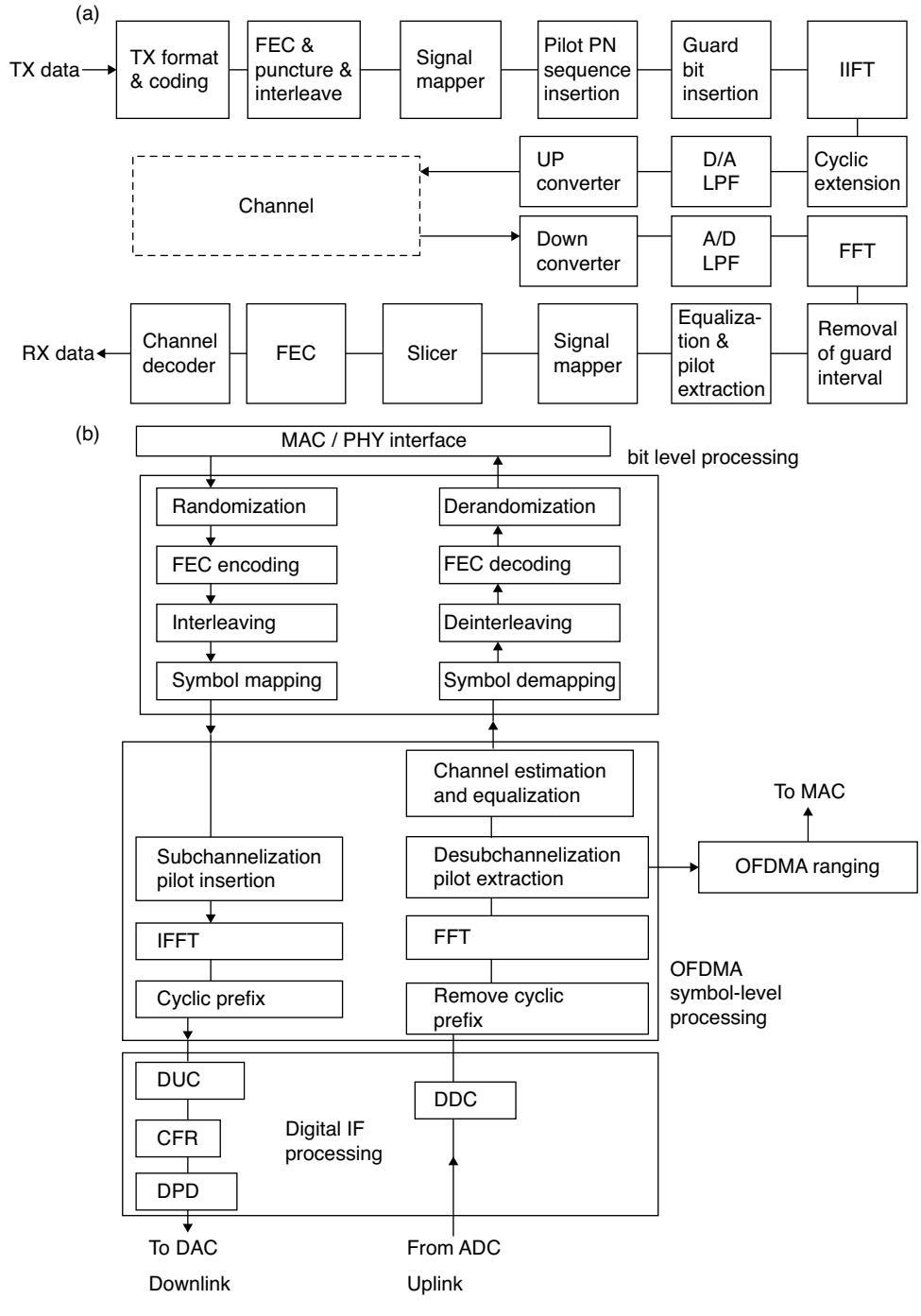


Figure 16.7 (a) Block diagram of a wireless MAN transmitter–receiver; and (b) WiMAX physical layer processing blocks

Forward error correction (FEC) in the IEEE 802.16a and other subsequent standards b, c, d, e is implemented in three phases: randomization, forward error correction, and interleaving.

Randomization

This stage of error correction ensures that there is a high level of entropy in the data. As the receiver uses a maximum likelihood detector, the decision regions are optimum as a result of randomization. A pseudo random binary sequence generator is used to implement randomization. It uses a 15-stage shift register with a generator polynomial of $1 + x^{14} + x^{15}$. The bits issued from the randomizer are fed to the FEC block, which is discussed next.

Forward Error Correction

This block consists of two layers – an outer Reed–Solomon code and an inner convolutional code. Reed–Solomon codes are block codes that are good for correcting burst errors. Convolutional codes are good for correcting random errors. Together, the combination effectively corrects most errors caused by the hostile wireless channel. The Reed–Solomon code uses a block length of 255 bytes and a payload length of 239 bytes. With this, 8 bytes of error can be corrected. The convolutional encoder is of rate $^1/_2$ and constraint length 7. If rates other than $^1/_2$ are desired, code shortening and puncturing are employed. A Viterbi decoder is used at the receiving end to decode data and correct errors.

Interleaving

The encoded data are now passed through a block interleaver. The interleaver is defined by a two-step permutation. The first ensures that neighboring data bits are mapped onto non-adjacent OFDM carriers. This ensures that if a deep fade affects a bit, its neighboring bits are likely to remain unaffected by the fade, and therefore FEC is sufficient to correct the effects of the fade.

The second step maps adjacent bits onto less or more significant constellation bits. This makes detection accurate and long runs of low reliability bits are avoided.

Data Modulation

After the data bits are interleaved, they are entered serially into a data modulator with the option of Gray coded QPSK, 16-QAM, and 64-QAM. In Standard 802.16 QPSK, 16-QAM, and 64-QAM data modulation schemes are used, whereas in the case of Standards 802.16a/b/c/d/e QPSK, 16-QAM, 64-QAM, and 256-QAM modulation schemes are used.

Adaptive modulation allows a WiMAX system to adjust the channel modulation scheme, according to SNR in the radio link. If a good SNR is achieved, the system can switch to the highest throughput modulation (64-QAM). If fading occurs the system can shift to other low-throughput modulation scheme without dropping the connection.

Channel Estimation and Equalization

In order to overcome the effects of the channel, channel estimation and frequency domain equalization should be done at the receiver. Estimation can be performed in two possible ways. The first method inserts pilot tones in all of the subcarriers of the OFDM symbols with a specific period. The second method inserts pilot tones in each of the OFDM symbols. These are known as block type and comb type channel estimation, respectively. The channel estimation for this model can be based on least squares and minimum mean square error estimates.

After estimating the channel, the received signal needs to be equalized. If the cyclic prefix is longer than the maximum delay spread of the channel, we can model the effect of the channel as a complex multiplication in the frequency domain. The equalization thus simplifies to a complex division of the received signal by the estimated channel.

Frequency Band

In the original release of Standard 802.11 the line-of-sight (LOS) environments at high frequency bands operates in the 10–66 GHz range, whereas the recently adopted amendment, the Standards 802.16a,b, c, d or e, are designed for systems operating in bands between 2 and 11 GHz. The significant difference between these two frequency bands lies in the ability to support non-line-of-sight (NLOS) operation in the lower frequencies, something that is not possible in higher bands.

Carrier Modulation

In Standard 802.16 single carrier modulation was chosen, because of the low complexity of the system. The downlink channel is shared among users with TDM signals. Subscriber units are being allocated by individual time slots. Access in uplink is being realized with TDMA. Channel bandwidths are 20 or 25 MHz in the USA and 28 MHz in Europe. Duplex can be realized with either TDD or FDD schemes, in which the uplink and downlink operate on separate channels.

The Standard 802.16 PHY specification uses burst single-carrier modulation with adaptive burst profiling in which transmission parameters, including the modulation and coding schemes, may be adjusted individually to each subscriber station (SS) on a frame-by-frame basis. Here both TDD and burst FDD variants are defined. The channel bandwidth is defined with Nyquist square-root raised-cosine pulse shaping with a roll-off factor of 0.25. Randomization is performed for spectral shaping and to ensure bit transitions for clock recovery. The OFDM frame is divided into DL and UL subframes (Figure 16.8). The DL subframe consists of preamble, frame control header, and a number of data bursts. The FCH specifies the burst profile and the length of one or more DL bursts that immediately follow the FCH. The DLMAP, UL-MAP, DL channel descriptor (DCD), UL channel descriptor, and other broadcast messages that describe the content of the frame are sent at the beginning of these first bursts.

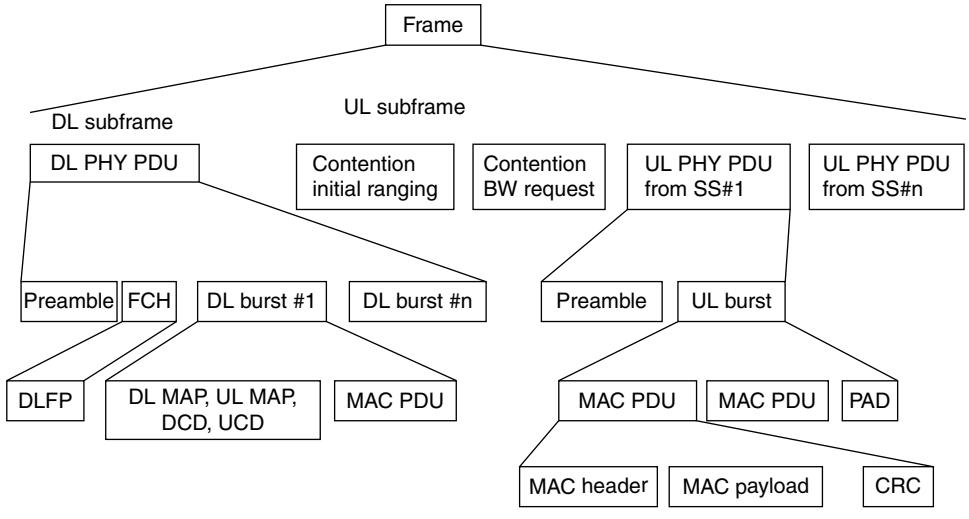


Figure 16.8 OFDM frame structure

The remainder of the DL subframe is made up of data bursts to individual SS. Each burst consists of an integer number of OFDM symbols and is assigned a burst profile that specifies the code algorithm, code

rate, and modulation level that are used for those data transmitted within the burst. However, the UL frame contains a contention interval for initial ranging and bandwidth allocation purposes and UL PHY PDUs from different SS.

16.4.1.2 MAC Layer

The main focus of the MAC layer is to manage the resources of the air-link in an efficient manner. The IEEE 802.16 MAC layer performs the function of providing a medium-independent interface to the 802.16 physical (PHY) layer. This MAC layer supports different types of physical layers, as discussed earlier. The 802.16 MAC protocol is designed to support very high bit rates both in uplink and in downlink, and also to support point-to-multipoint (PMP) and mesh network models. The 802.16 MAC protocol is connection oriented. Once the SS enters the network then each SS creates one or more connections over which their data are transmitted to and from the base station (BS). The MAC layer schedules the usage of the air-link resources and provides quality-of-service (QoS) differentiation. It performs link adaptation and automatic repeat request (ARQ) functions to maintain target bit error rates (BER), while maximizing the data throughput. The MAC layer also handles network entry for SSs that enter and leave the network, and it performs standard protocol data unit (PDU) creation tasks. Finally, the MAC layer provides a convergence sublayer that supports asynchronous transfer mode (ATM) cell and packet-based network layers.

Depending on these functionalities the 802.16 MAC layer is divided into several sublayers:

1. **Service specific convergence sublayer (SSCS) layer:** This provides an interface to the upper layer entities through CS SAP. The primary task of this sublayer is to classify service data units (SDUs) to the proper MAC connection, preserve or enable QoS, and enable bandwidth allocation. The mapping takes various forms depending on the type of service. IEEE Standard 802.16 defines two general service-specific convergence sublayers for mapping services to and from 802.16 MAC connections.
   a. ATM convergence sublayer: This is defined for ATM services.
   b. Packet convergence sublayer: This is defined for mapping packet services such as IPv4, IPv6, Ethernet, and virtual local area network (VLAN).

   In addition to these basic functions, the convergence sublayers can also perform more sophisticated functions such as payload header suppression and reconstruction to enhance air link efficiency.
2. **MAC common part sublayer (CPS) layer:** This provides the core MAC functions, including uplink scheduling, bandwidth request and grant, connection control, ARQ and ranging. All services, including inherently connectionless services, are mapped to a connection. This provides a mechanism for requesting bandwidth, associating QoS and traffic parameters, transporting and routing data to the appropriate convergence sublayer, and all other actions associated with the contractual terms of the service. Connections are referenced with 16-bit connection identifiers (CIDs) and may require a continuously granted bandwidth or bandwidth on demand. Each SS has a standard 48-bit MAC address, but this serves mainly as an equipment identifier, as the primary addresses used during operation are the CIDs. Upon entering the network, the SS is assigned three management connections in each direction. These three connections reflect the three different QoS requirements used by different management levels. The first of these is the basic connection, which is used for the transfer of short, time-critical MAC and radio link control (RLC) messages.

MAC PDU Formats –The MAC PDU is the data unit exchanged between the MAC layers of the BS and its SSs. An MAC PDU consists of a fixed-length MAC header, a variable-length payload, and an optional cyclic redundancy check (CRC). There are two types of MAC header: (1) a generic header, used to transmit data or MAC messages, and (2) a bandwidth (BR) request header, used by the SS to request more

bandwidth on the UL. The maximum length of the MAC PDU header is 2048 bytes, including header, payload, and cyclic redundancy check (CRC). MAC PDU does not contain any payload, except for bandwidth request MAC PDUs, which contain either MAC management messages or convergence sublayer data.

Transmission of MAC PDUs – The IEEE 802.16 MAC supports various higher layer protocols such as ATM or IP. Incoming MAC SDUs from corresponding convergence sublayers are formatted according to the MAC PDU format, possibly with fragmentation and/or packing, before being conveyed over one or more connections in accordance with the MAC protocol. After traversing the air link, MAC PDUs are reconstructed back into the original MAC SDUs, so that the format modifications performed by the MAC layer protocol are transparent to the receiving entity. IEEE 802.16 takes advantage of incorporating the packing and fragmentation processes with the bandwidth allocation process to maximize the flexibility, efficiency, and effectiveness of both. Fragmentation is the process in which an MAC SDU is divided into one or more MAC SDU fragments. Packing is the process in which multiple MAC SDUs are packed into a single MAC PDU payload (Figure 16.9). Both processes may be initiated by either a BS for a downlink connection or an SS for an uplink connection. IEEE 802.16 allows simultaneous fragmentation and packing for efficient use of the bandwidth.

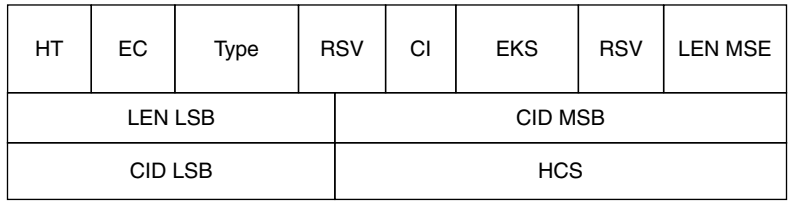


Figure 16.9 Generic header format for MAC PDU

16.4.1.3 Radio Link Control (RLC)

The RLC layer of Standard 802.16 is more advanced and has the capability of PHY transition from one burst profile to another, power control, and ranging functions. RLC begins with a periodic BS broadcast of the burst profiles that have been chosen for the uplink and downlink. Burst profiles for the downlink are each tagged with a downlink interval usage code (DIUC) and for the uplink are each tagged with an uplink interval usage code (UIUC).

During initial access, the SS performs initial power leveling and ranging using ranging request (RNG-REQ) messages transmitted in the initial maintenance windows. The adjustments to the transmit time advance of the SS, in addition to power adjustments, are returned to the SS in ranging response (RNG-RSP) messages. For ongoing ranging and power adjustments, the BS may transmit unsolicited RNG-RSP messages commanding the SS to adjust its power or timing. During initial ranging, the SS also requests to be served in the downlink via a particular burst profile by transmitting its choice of DIUC to the BS. The choice is based on received downlink signal quality measurements performed by the SS before and during initial ranging. The BS may confirm or reject the choice in the ranging response. Similarly, the BS monitors the quality of the uplink received signal from the SS. Using ULMAP messages with the appropriate burst profile UIUC with the grants of the SS, the BS commands the SS to use a particular uplink burst profile. After initial determination of uplink and downlink burst profiles between the BS and a particular SS, RLC continues to monitor and control the burst profiles. Harsher environmental conditions can force the SS to request a more robust burst profile. However, exceptionally good weather may allow an SS to temporarily operate with a more efficient burst profile.

In the downlink, the SS monitors the quality of the received signal and knows when its downlink burst profile should change. However, BS controls the change. There are two methods available to the SS to request a change in downlink burst profile, depending on whether the SS operates in the grant per connection (GPC) (which applies to only GPC SSs) or the grant per SS (GPSS) mode. In the first instance, the BS may periodically allocate a station maintenance interval to the SS. The SS can use the RNG-REQ message to request a change in downlink burst profile. The preferred method is for the SS to transmit a downlink burst profile change request (DBPC-REQ). In this case, which is always an option for GPSS SSs and can be an option for GPC SSs, the BS responds with a downlink burst profile change response (DBPC-RSP) message confirming or denying the change. Because of irrecoverable bit errors, messages may be lost, so the protocols for changing a downlink burst profile of an SS must be carefully structured and is well taken care in the standard.

16.4.1.4 Privacy Sublayer (PS)

The main functions of this sublayer are authentication and data encryption. The privacy protocol of IEEE 802.16 is based on the privacy key management (PKM) protocol of the DOCSIS BPI + specification, but has been enhanced to fit seamlessly into the IEEE 802.16 MAC protocol and to better accommodate stronger cryptographic methods.

16.4.1.5 Convergence Layer

A convergence layer is placed above the MAC layer. This provides functions specific to the service being provided. For IEEE 802.16.1, bearer services include digital audio/video multicast, digital telephony, ATM, Internet access, wireless trunks in telephone networks, and frame relay.

16.4.2 Service Classes

The 802.16 MAC provides QoS differentiation for different types of applications that might operate over the 802.16 networks. Standard 802.16 defines the following types of services:

1. **Unsolicited grant services (UGS):** UGS is designed to support constant bit rate (CBR) services, such as T1/E1 emulation, and voice over IP (VoIP) without silence suppression.
2. **Real-time polling services (rtPS):** rtPS is designed to support real-time services that generate variable size data packets on a periodic basis, such as MPEG video or VoIP with silence suppression.
3. **Non-real-time polling services (nrtPS):** nrtPS is designed to support non-real-time services that require variable size data grant burst types on a regular basis.
4. **Best effort (BE) services:** Today, BE services are typically provided by the Internet for Web surfing.

16.4.3 Mobility Support

The mobility feature is supported in Standard 802.16e considering the Doppler effect, multipath, fading and multicell interference environment. Several factors are taken care of, such as the dedicated control channels for critical MAC and PHY functions, for example power control, timing control, ARQ acknowledgements, uplink requests, and so on. In supporting handover, MS collects information related to potential handover (HO) and transfers it to the network; the network collects the relevant information (for example, PHY measurements from BSs), makes a decision and executes the handover. Different types of handover are supported. (1) Inter-channel HO: this is between channels (sectors) at the same BS. In this

instance BS makes the HO decision and executes the handover. (2) Inter-cell soft HO: this is between two BSs. Here the serving BS makes the HO decision and executes the HO. (3) Inter-cell hard HO: this is where the MS fails to communicate to the serving BS, it performs a complete NW entry procedure with the best possible BS; the new BS informs the old BS about the HO.

16.4.4 Power Control

In this standard, the power control mechanisms are used to improve the overall system work and this is implemented at the base station, which sends the steering signal to the subscriber station (SS) to achieve a pre-determined signal level at the BS. It is also required to reduce interferences with neighboring cells.

16.4.4.1 Operational Procedures

The operational sequences when the SS is switched on are described below and are shown in Figure 16.10.

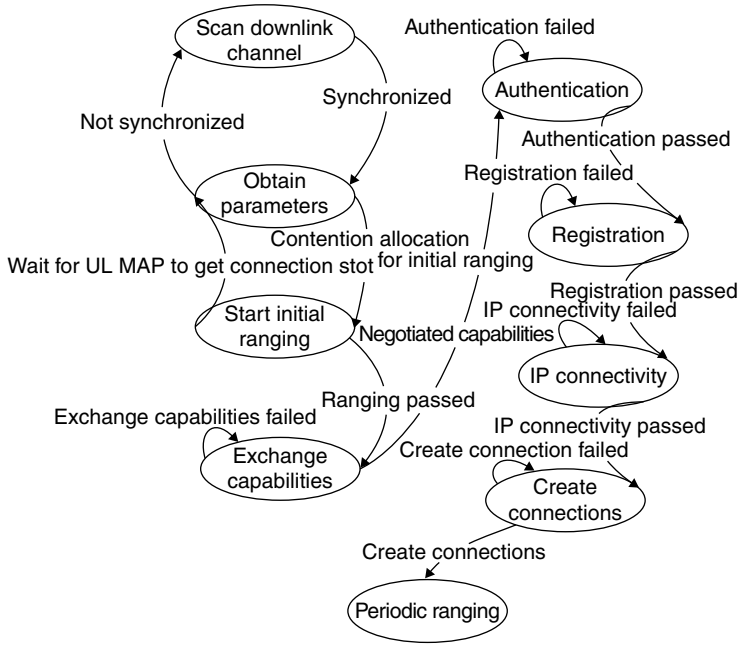


Figure 16.10 Network entry process

1. **Network Entry** – To get the access to the network, SS needs to successfully complete the network entry process with the desired BS. The network entry process is divided into several steps – DL channel synchronization, initial ranging, capabilities negotiation, authentication message exchange, registration, and IP connectivity stages. The network entry state machine moves to the

reset state if it fails to succeed from a state. Once the network entry process is completed, the SS creates one or more service flows to send data to the BS. Each of these stages are discussed below in more detail.

a. **Downlink channel synchronization –** SS scans for a channel in the defined frequency list. Generally, an SS is configured to use a specific BS with a given set of operational parameters, when operating in a licensed band. If the SS finds a DL channel and is able to synchronize at the PHY level (it detects the periodic frame preamble), then the MAC layer looks for DCD and UCD to get information on modulation and other DL and UL parameters.
b. **Initial ranging** – Once the SS is synchronized with the DL channel and has received the DL and UL MAP for a frame, it begins the initial ranging process by sending a ranging request MAC message on the initial ranging interval using the minimum transmission power and waits for the response. If it does not receive a response, it sends the request again in a subsequent frame, using higher transmission power. Eventually the SS receives a ranging response. The response either indicates power and timing corrections that the SS must make or indicates success. If the response indicates corrections, the SS makes these corrections and sends another ranging request. If the response indicates success, the SS is ready to send data on the UL.
c. **Capabilities negotiation** – Once the initial ranging is successfully completed, next SS sends a capability request message to the BS describing its capabilities in terms of the supported modulation levels, coding schemes and rates, and duplexing methods. The BS accepts or denies the SS, based on its capabilities.
d. **Authentication** – After capability negotiation, the BS authenticates the SS and provides the required data for enabling the data ciphering. The SS sends the X.509 certificate of the SS manufacturer and a description of the supported cryptographic algorithms to its BS. The BS validates the identity of the SS, determines the cipher algorithm and protocol that should be used, and sends an authentication response to the SS. The response contains the key material to be used by the SS. The SS is required to periodically perform the authentication and key exchange procedures to refresh its key material.
e. **Registration** – Once the authentication is successfully complete then the SS registers with the network. After that the SS sends a registration request message to the BS, and the BS sends a registration response to the SS. The registration exchange includes IP version support, SS managed or non-managed support, ARQ parameters support, classification option support, CRC support, and flow control.
f. **IP connectivity** – Next the SS starts DHCP (IETF RFC 2131) to get the IP address and other parameters to establish IP connectivity.

2. **Transport Connection Creation** – After completion of registration and the transfer of operational parameters, transport connections are created. For pre-provisioned service flows, BS initiates the connection creation process. The BS sends a dynamic service flow addition request message to the SS and the SS sends a response to confirm the creation of the connection. However, SS initiates a connection creation request for non-pre-provisioned service flows by sending a dynamic service flow addition request message to the BS and BS responds to it with a confirmation.

16.5 4G Mobile System

The basic block diagram of a typical OFDM based multimedia mobile device is shown in Figure 16.11. The band–band unit is the heart of the modern digital mobile device, which implements modulation–demodulation, channel coding, error control, and other physical layer procedures, different transmission protocol, user application processing, and system control.

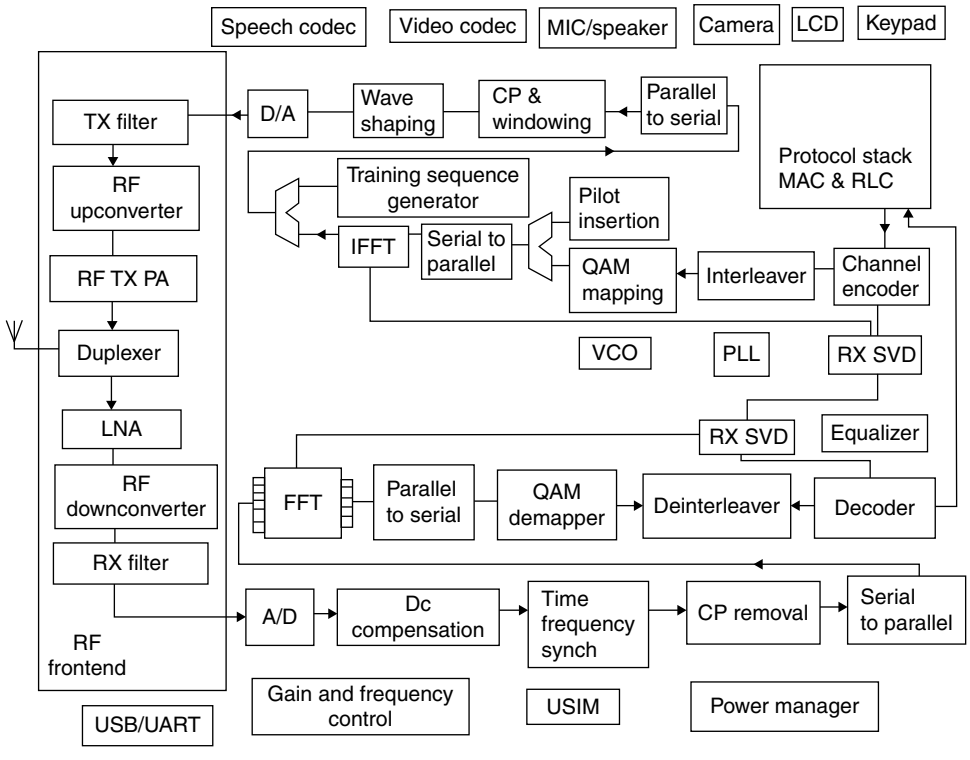


Figure 16.11 Block diagram of an OFDM based multimedia mobile device

16.6 Key Challenges in Designing 4G Mobile Systems and Research Areas

Over the past decade the complexity of wireless systems has expanded tremendously to support many complex applications and algorithms. Figure 16.12 shows the processing complexity explosions over the wireless generations.

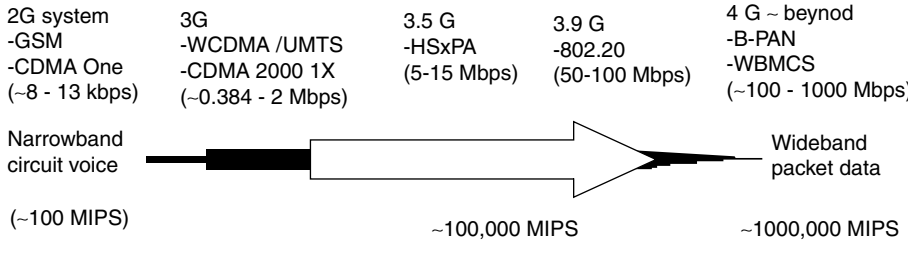


Figure 16.12 Processing complexity explosions over wireless generations

The baseband SoC in a modern cell phone typically contains one or more general purpose processor cores and one or more DSP cores. The processor core(s) may control 20–40 peripherals (many in the form of intellectual property blocks) for such tasks as multimedia functions, digital camera interfaces, cryptographic functions, 2D and 3D graphics processing and acceleration, and interfaces such as WiFi, USB, and UART. The DSP core(s) typically have a number of associated accelerator peripherals for tasks such as modulation, baseband filtering, channel decoding, and so forth.

Moore's Law states that a chip's transistor density will double every 18 months. This sets an upper limit on the rate that system performance can improve. Claude Shannon defined the theoretical limit of information transmission in the presence of noise. Much of the complexity in next generation wireless technology results from sophisticated signal-processing requirements. The algorithmic complexity increases according to Shannon's law while the silicon technology increases according to Moore's law. It is clear that there is an increasing gap between the algorithmic complexity and the processor performance (Figure 16.13). The increased algorithmic complexities demands fhigher processing power, which requires very high MIPS. The gap between the algorithmic complexity and the battery capacity is even bigger. This critically demands efficient design of both more compact and more power efficient architectures.

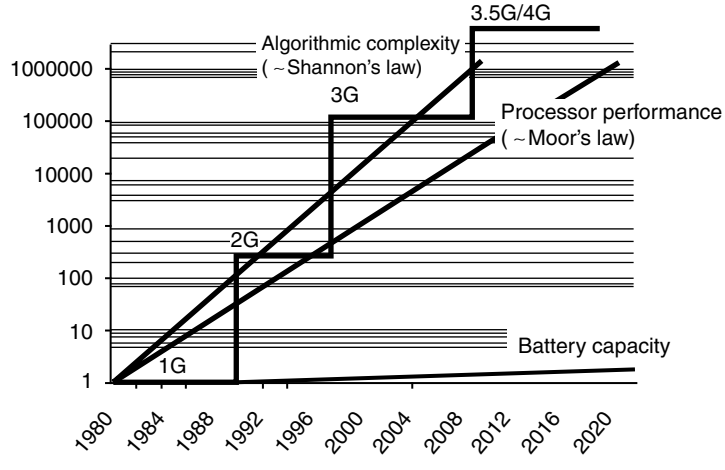


Figure 16.13 SoC architecture design challenges: future trend comparison of algorithmic complexity (shown in steps), silicon capacity, and battery capacity

There are several challenges associated with the next generation (>3G systems) development and among these the foremost one is the design of a 4G mobile device. The next generation applications are looking for unique SoC solutions, which can offer low power, low cost, smaller size, high performance, and tremendous flexibility.

- **Why high performance?**
  With an ever increasing demand for more and more complex applications, support of high speed, high capacity, real time, guaranteed QoS, and seamless communications requires a system with high performance.
- **Why low power?**
  In the case of a mobile power-voracious system, such as a multimedia phone, laptop, and so on, the power consumption is high and there is limited battery capacity. So, the design of the system should be

such that the power consumption should be reduced or at least optimized at every step of the system design, wherever feasible. Reduction in power consumption provides several benefits, such as less heat will be generated (which reduces the problems associated with the rise in temperature), prolongs the battery life, and device reliability increases.

- **Why flexibility?**
  There are applications (for example, 4G system) where devices are required to adapt their functionality to the changing parameters of the communication link available at a given time (that is, bandwidth, error rates, protocols, etc.). Therefore, these devices have to be flexible enough to accommodate various multimedia services (for example, different video decompression schemes) and communication capabilities (for example, cellular GSM, PCS, pico-cellular). At the same time, low-power consumption will continue to be the predominant design challenge of mobile embedded systems.

Thus, the bottom line is – 4G needs highly flexible, high performance, low power mobile user equipment, which is a major challenge.

16.7 Cognitive Radio

Over the last few years, there has been a considerable resurgence of interest in wireless communication. This step forward has not only expanded the wireless communications market, but has also created opportunities for newer products operating at different frequency bands. The main drawback of this explosion in services is that there is an increase in electromagnetic waves and a crisis with respect to the spectrum availability at frequencies that can be economically used. Access to the radio spectrum is presently regulated either as licensed, where the rights to use specific spectral bands are granted in exclusivity to an individual operator, or as unlicensed, where certain spectral bands are declared open for free use by any operator or individual following specific rules. The drawbacks of this static frequency band allocation to any single service are: (a) inefficient use of the spectrum, (b) licensing for a new spectrum is very expensive, slow and involved political process, and (c) more electromagnetic radiation leads to health risks and potential biological effects.

Apart from the spectrum congestion issue, there are many other issues associated with today's radios. Among these, the most important issues are the hardware flexibility and lack of compatibility between the various types of devices and different communications and network infrastructures. In particular, in different countries, incompatible wireless network technologies make it difficult to deploy global roaming facilities effectively.

Another problem is trying to keep up with evolving link-layer communication (transmission) protocol standards, such as 2.5G, 3G, and 4G. The legacy handsets currently on the market cannot be easily upgraded to new features and/or services. Most are not even compatible with the new standards.

What is needed is a generic, programmable hardware base that would allow software to enable various features, depending on the radio environments, where users would get tremendous mobility. The communications industry is now looking for a way to create radios that can handle multiple frequency bands, understand multiple transmission protocols, be reconfigured on the fly, and be easily upgraded – all in a single device design.

This would be in the form of a radio that can sense its environment, location, and intended use and then alter its power, frequency, modulation and other parameters, so as to dynamically reuse the available spectrum.

The move towards the re-configurability concept was initiated as an evolution of software defined radio. Its aim is to provide essential mechanisms to terminals and networks, so as to enable them to adapt dynamically, transparently, and securely to the most appropriate RAT depending on the current situation. Cognitive radio was initially thought of as an extension to software-defined radio (full cognitive radio), but it is much smarter than SDR, as it dynamically senses and utilizes the available spectrum. We can distinguish many types of cognitive radios.

1. **Full cognitive radio ("Mitola radio"):** In which every possible parameter observed by the wireless node and/or network is taken into account while making decision on transmission and/or reception parameter change.
2. **Spectrum sensing cognitive radio:** This is a special example of full cognitive radio in which only the radio frequency spectrum is observed.

   Also depending on the parts of the spectrum available for cognitive radio we can distinguish it as:
3. **Licensed band cognitive radio:** When cognitive radio is capable of using bands assigned to licensed users, apart from the utilization of unlicensed bands, such as the UNII band or ISM band. One such example is the IEEE 802.15 Task Group specification.
4. **Unlicensed band cognitive radio:** When cognitive radio can only utilize unlicensed parts of the radio frequency spectrum. An example of unlicensed band cognitive radio is IEEE 802.19.

Full cognitive radio has great potential as the next generation wireless radio to integrate these different evolving and emerging wireless access technologies in a common flexible/adaptable platform and to solve the spectrum congestion problem, providing a multiplicity of possibilities for current and future services and applications to users in a single terminal.

At present cognitive radio is in the research phase, and many more challenges still need to be solved before the final product is launched.

16.7.1 System Overview

The functional model of cognitive radio is shown in Figure 16.14.

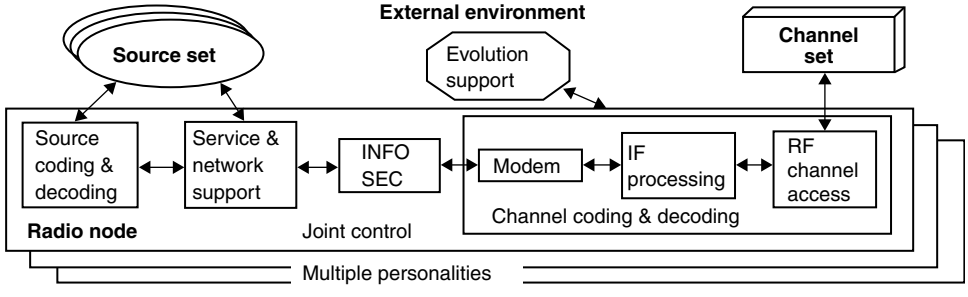


Figure 16.14 Functional model

The radio hardware consists of a set of modules: antenna, RF transceiver, modem, INFOSEC module, baseband/protocol processor, and user interface. This could be a software radio, SDR, or PDR. In Figure 16.14, the baseband processor hosts the protocol and control software. The modem software includes the modem with equalizer, among other physical layer processes. In addition, however, a cognitive radio contains an internal model of its own hardware and software structure. The model of the equalizer shown would contain the codified knowledge about equalizers, including how the taps represent the channel impulse response. Variable bindings between the equalizer model and the software equalizer establish the interface between the reasoning capability and the operational software. Cognitive radio enhances the flexibility of personal services through a radio knowledge representation language (RKRL). The model-based reasoning capability that applies these RKRL frames to solve radio control problems gives the radio its "cognitive" ability.

16.7.2 System Architecture

The cognitive radio system architecture is shown in Figure 16.15.

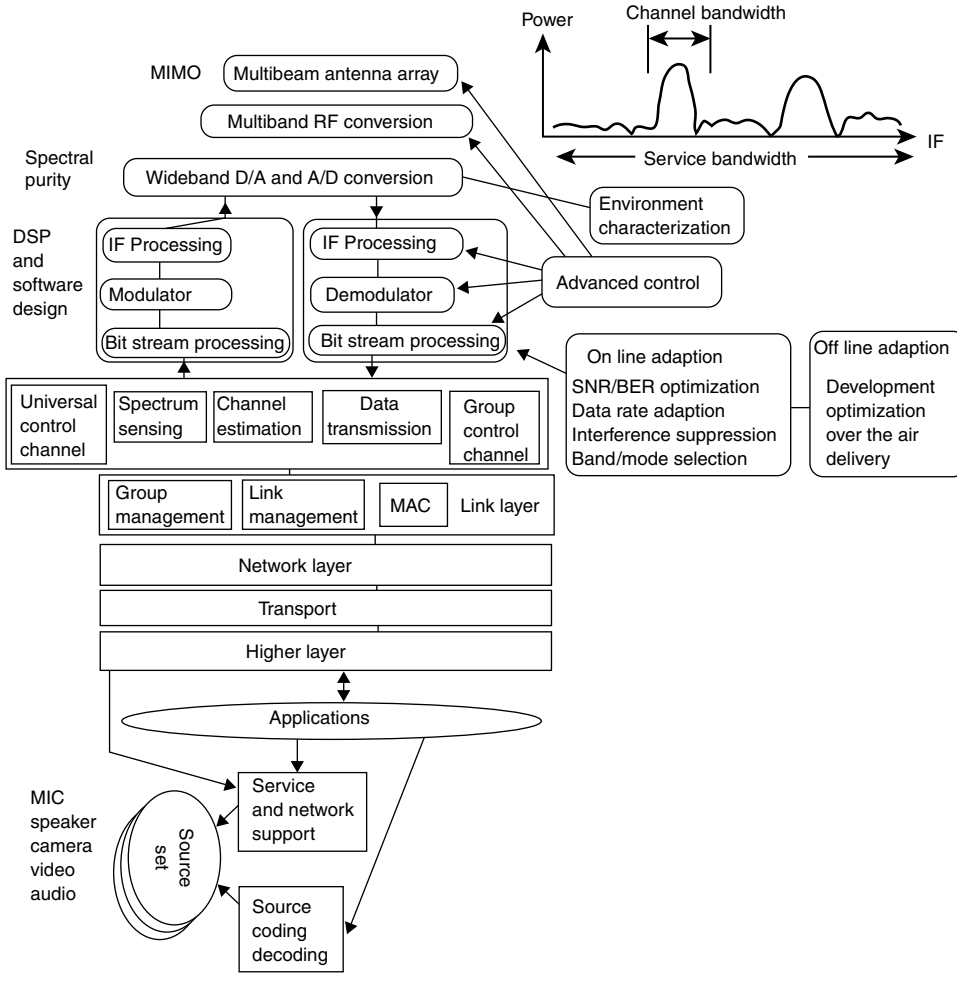


Figure 16.15 System architecture

The system design only covers the ISO/OSI layers – one (physical layer) and layer-2 (link layer).

Higher layers will be implemented as standard protocols such as UMTS, GSM, WLAN 802.11a/b/g/n, 802.15, and so on, and not specific to cognitive radios.

Generally, six main functions and two control channels will implement the core functionality of a cognitive radio. The six system functions can be split between the physical and the link layer. Whereas the physical layer is responsible for sensing the spectrum, detecting active primary users and estimating the quality of the subchannels, the link layer has to deal with group management, link setup and maintenance, and handle the medium access of the subchannels.

16.7.3 Key Challenges and Research Areas

The essential problem with spectrum sensing cognitive radio is the design of high quality spectrum sensing devices and algorithms for exchanging spectrum sensing data between nodes. It has been shown that a simple energy detector cannot guarantee accurate detection of signal presence. This calls for more sophisticated spectrum sensing techniques. The information about spectrum sensing must be exchanged between nodes regularly. The cognition cycle implies a large area of hard research problems for cognitive radio. Parsing incoming messages requires natural language text processing. Scanning the user's voice channels for content that further defines the communications context requires speech processing. Planning technology offers a wide range of alternatives in temporal calculus, constraint based scheduling, task planning, causality modeling, and so on. Resource allocation includes algebraic methods for wait-free scheduling protocols, open distributed processing (ODP), and parallel virtual machines (PVM). Finally, machine learning remains one of the core challenges in artificial intelligence research. The focus of this cognitive radio research, then, is not on the development of any one of these technologies per se. Rather, it is on the organization of cognition tasks and on the development of cognition data structures needed to integrate contributions from these diverse disciplines for the context sensitive delivery of wireless services by software radio.

Another obvious issue is the security of downloads. For example, given a script that describes a link-layer protocol, there must be a phase in which the protocol is downloaded to the hardware and run as a configurable protocol. The question is, how is that download secured? Obviously, it must be signed and have digital authorization. Otherwise, downloads might be made to devices that could then broadcast on unauthorized bands. Security issues facing SDR technology include encryption, user identification, device authentication, and others.

Above all, the MIMO based OFDM multi-band cognitive radio requires a very complex platform with high digital signal processing capability to execute this intelligence, which needs a major breakthrough in the signal processing and semiconductor technology to offer a high processing power, high integration, low power consumption, and low cost based system.

Some of the basic challenges in this radio design include: (1) an immediate, cost-effective solution that doesn't require agencies to buy new radios, for example, the same radio will be used everywhere with any networks; (2) energy efficient and will conserve battery power; (3) should be portable, lightweight, and small; (4) powerful, cost-effective programmable digital signal processors (DSPs); (5) adaptive computing machines (ACMs), which handle multiple protocols by adjusting themselves to the algorithm or mathematical equation being executed; (6) high-performance analog-to-digital converters (ADCs); and (7) ultra-fast data-transfer interfaces.

At present research is continuing in the following areas: (1) cognitive radio system architectures; (2) platforms and hardware implementations for the support of cognitive radio systems; (3) cognition cycle optimization; (4) multi-band, spectrum-agile, and adaptive radio transceivers; (5) cognitive radio resource management and dynamic spectrum sharing; (6) signal detection and interference avoidance (management/algorithms); (7) spectral estimation and radio environment characterization; (8) efficient spectrum utilization; (9) soft-spectrum adaptation technology; (10) radio access protocols and algorithms for the PHY, MAC, and network layer; (11) cross-layer optimization for cognitive algorithms; (12) software defined radio systems; (13) ultra-wideband (UWB) cognitive radio systems; (14) cognitive impulse radio systems; (15) bio-inspired cognitive radio; and (16) linear network coding, cooperative coding and MIMO techniques for cognitive radio.

Further Reading

3GPP LTE Standards. http://www.3gpp.org/ftp/Specs/html-info/36-series.htm.

Bryan, A. and Ivan, S. High Performance Cognitive Radio Platform with Integrated Physical & Network Layer Capabilities, WINLAB, Rutgers University, www.winlab.rutgers.edu.

Guo, Y. (2005) Advanced MIMO-CDMA Receiver for Interference Suppression: Algorithms, System-on-Chip Architectures and Design Methodology. Thesis submitted in partial fulfillment of the requirements for the degree doctor of philosophy. Houston, Texas.

IEEE Broadband Wireless Access Working Group (June 2002) 802.16. *A Technical Overview of the WirelessMAN™ Air Interface for Broadband Wireless Access.* Carl Eklund, Nokia Research Center, Roger B. Marks, National Institute of Standards and Technology, Kenneth L. Stanwood and Stanley Wang, Ensemble Communications Inc., IEEE Communications Magazine.

Koutsopoulos, I. (2002) Resource Allocation Issues in Broadband Wireless Networks with OFDM Signaling, Doctor of philosophy thesis. University of Maryland.

Krenik, W. and Batra, A. (2005) Cognitive Radio Techniques for Wide Area Networks, Texas Instruments Incorporated, Dallas, Texas. 214-480-6448.

Lawrey, E. (1997) The suitability of OFDM as a modulation technique for wireless telecommunications, with a CDMA comparison. Thesis. James Cook University.

Mitola, J. III (May 8, 2000) Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio. Dissertation. Royal Institute of Technology (KTH).

Sahai, A. (May 15, 2005) Some fundamental limits on cognitive radios and their implications, Wireless Foundations, UC Berkeley EECS.

Working Group 6 White Paper (2004) Cognitive Radio, Spectrum and Radio Resource Management, Wireless World Research Forum.

Younus, A. (June 2006) WiMAX-Broadband Wireless Access, Technical University of Munich, Germany.

17

Competitive Edge in Mobile Phone System Design

17.1 Introduction

In the previous chapters, we discussed the various aspects and internal details of mobile phones. In this chapter we will discuss the system design issues of a mobile phone. Today's mobile handset system is not just a piece of hardware or a bunch of software; rather, it is a combination of both hardware and software. In the present competitive market the key factors to success are designing a system that can work with minimum resources (such as memory size, MIPS, etc.) and can offer high performance in terms of execution speed, low power consumption, high robustness, and reliability. It is not always a big deal to write a piece of software to work on a system to be logically correct, but it is a really big deal to write a piece of software that will work in an environment with limited resources (such as memory, MIPS) with greater speed of operation and that is logically correct, and so this is a real challenge. This chapter examines various factors contributing towards the development of a competitive mobile phone hardware and software protocol stack. Both technical and non-technical aspects are considered. The key issues addressed include protocol architecture, system performance in terms of memory, CPU, operating system (OS), electrical power consumption, processing power (MIPS), cost, optimum hardware/software partitioning, testing, and productization.

17.2 Key Challenges in Mobile Phone System Design

The key challenges of an efficient mobile phone system design are as follows.

- How much hardware do we need? Which CPU to use? Which and how much memory to use?
- How much to put in the hardware?
- How do we meet our deadlines?
- Faster hardware or cleverer software?
- How do we minimize power consumption? Turn off unnecessary logic? Reduce memory accesses?
- How to reduce the system cost?
- How to increase system efficiency?

Mobile Handset Design Sajal Kumar Das

17.3 System Design Goal

The primary goal of a system designer should be to design a system that must confirm size and weight limits, consume much less power, satisfy safety and reliability requirements, meet tight cost targets and above all guarantee real time operation that is reactive to external events. Some requirements are functional and some are non-functional requirements.

- **Functional requirements:** output as a function of input.
- **Non-functional requirements:** such as time required to compute output, size, weight, power consumption; reliability, and so on.

The mobile phone system is a complex embedded system and the design goals of an embedded system vary based on the system's application area, such as:

a. **Real time operation** – Real time operation means the computation of the data should be completed within a time limit. So, during the design of the system, the worst-case situation should always be taken into account. With reactive computation the software executes in response to external events. These events may be periodic, thus scheduling of events to guarantee performance may be possible.
b. **Portability** – The size and weight of the system plays an important role in the system design. In many cases, especially in the mobile environment, it is always desirable that the system should be as small as possible in size and weight.
c. **Power consumption** – For mobile devices, power consumption is a vital issue. Sufficient care should be taken in the system design to reduce the power consumption in all possible ways.
d. **Safety and reliability** – The systems have associated risks with failure. However, the probability of failure will be reduced, if the system is designed properly and tested rigorously before delivery. The safe and reliable product enhances the customer's faith and satisfaction. After taking all precautions, failure may still happen, so there should be some way to quickly recover or to debug and make it work again.
e. **Cost** – The system may be designed to meet all the challenges and satisfy all the requirements. However, it will not attract the customer's attention if the cost is not low enough. Putting attention into a particular stage cannnot drastically reduce the cost, so it has to be done at each and every stage of the system design. A good designer has to manage all the issues to deliver a cost-effective system at the right time with the right price.

17.4 Protocol Architecture Design Optimization

The most commonly used protocol architectural model is layered architecture. The layered architecture specifies partitioning of the protocol into clearly defined layers based on functionality. It also defines the interaction between the layers. Implementation of the layered architecture can vary and choosing a suitable method depends on many factors, which need to be identified and analyzed. Alternative architectures could also be considered.

17.4.1 Various Alternative Solutions

Simple protocols could be implemented without an RTOS. Such protocols will be implemented as a single task with only a few interrupts, which may or may not be time critical. The same cannot be said for complex protocols, including wireless protocols. This is where the concept of ISO-OSI layered architecture comes into the picture. In this architecture, the protocol is partitioned into seven layers, each with a specific function. Wireless protocols on user equipment will probably not have all seven layers

of the OSI layering. Nevertheless, it is possible to group certain functions of protocol into a specific layer of the OSI architecture. This is the work of standards bodies, such as ETSI and 3GPP.

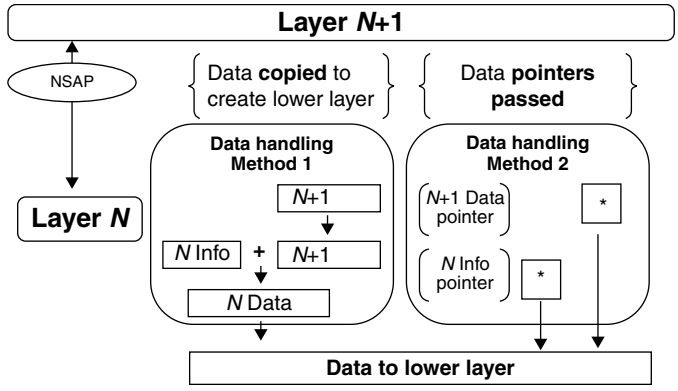


Figure 17.1 Layered architecture – comparing data copying and passing pointers

It is clear that each layer has certain functions to perform, and also that each layer communicates to a higher layer with a view to providing defined services to the latter. These services are provided via service access points (SAP), which are the points of interface between the layers. The actual provision of services is executed by means of exchange of messages, which are known as primitives. This concept is illustrated in Figure 17.1. Table 17.1 explores the different ways in which layers, SAPs, and primitives can be implemented. The table presents five options:

A. Each layer is implemented as a process with a certain priority. Processes are scheduled for execution according to their priorities by the RTOS. Each process can share the same stack space or each one can use a different stack space. The latter is the better method as it minimizes the context switching time. SAPs are implemented as signals between processes carrying parameters. Signals are put in the receiving process queue. Too many processes can result in excessive signaling. On the other hand, too few processes can result in increased complexity, compromising better efficiency.
B. All layers are combined into a single process. The layered concept is maintained by logically separating the code and data within the same process. States are maintained as variables, one for each layer or instance. Increased complexity is the result and programmers must be extremely adept at their job at designing, testing, and debugging. SAPs are implemented internally as variables, whose scope spans more than one layer and it results in less signaling. In fact, the only signals are interrupts from sources that are external to the protocol and the remainder of the SAPs are simply procedure calls.
C. Each layer is implemented as a process with a certain priority but they share a single queue. There is no clear reason why one would want to do this.
D. All layers are combined into a single process. The same variables could be duplicated across a few layers. This redundancy could be removed by making them global to save data space. The compromise made is increased complexity in managing these variables. The difference between this method and option B is that in the latter option a variable that belongs to a particular layer is not accessible to other layers. In essence, this is encapsulation, which is one of the cornerstones of object-oriented methodology.
E. This option is a compromise between options A and B.

Table 17.1 Different protocol architectures compared

Protocol	Implementation	Advantages	Disadvantages
A	Each layer = one task/process	Finer granularity – a lower priority process can be interrupted by a higher priority process	Process overhead is higher
	SAP/primitives = signals	Project is easier to manage	More context switching
		Object oriented methodology and design are possible	Memory requirements are higher
		Development is easier	More signaling is involved when communicating between layers
		Easy to debug	More processing of signals is also needed at every layer
		SDL implementation and simulation and validation becomes easier	
B	All layers = one task/process	Less overhead and more efficient in terms of memory as well as execution speed	Granularity of task pre-emption is now limited: this can be overcome if a customized RTOS is written to handle memory partitioning within the single task
	SAP/primitives = signals or procedure calls	Faster execution	Highly complex
		Lower data size and code size	Maintenance and upgrading is difficult
		Layered concept is implemented as logical partitioning within the single task	Top down knowledge needed from the outset
		Scaled down RTOS is sufficient – RTOS may not even be needed	Coding methods must be strictly followed by all designers – design is highly vulnerable to deviations from agreed coding practices
		Good for prototyping and benchmarking	Adaptability and portability is lower
		Less signaling across layers and less processing at layers	Difficult to debug – debug routines must be separately written and probably only textual

C	Each layer = one task/process SAP/primitives = signals sharing a single queue	Similar to A	Single queue across processes is not a common RTOS feature • No real advantage over option [A]
D	All layers = one task/process SAP/primitives = global variables or variables shared within the single task/process	Similar to B The added advantage here is that the same variables are used by different layers and this reduces data size	More co-ordination needed among designers Difficult to partition task among designers Maintenance is extremely complex Difficult to debug and prone to run-time errors This will only work if design is firm from the start
E	Group some layers = one task/process (based on priority of layers)	Compromise between A and B Gets the best of both A and B	Some disadvantages of A and B remain

The second architectural detail was considered with regard to managing the data through the protocol stack, where a few points have to be noted.

When an SDU is delivered to the RLC, the RLC breaks it into several PDUs and passes it to MAC. MAC adds its required header to the PDU and passes it to layer-1. The RLC has two options:

1. Make a copy of the PDU and pass it to MAC.
2. Take a pointer indexed into the original SDU and pass the pointer to MAC.

Similarly, MAC has two options – either make a separate copy or pass the pointers to layer-1. MAC has to pass the pointer to the header and also the pointer to the PDU. Where RLC headers are involved, RLC will pass the pointers to the header and the PDU to the MAC. What layer-1 interface sees is a series of pointers, length and offset for each, and the correct order for processing the pointers. Passing pointers violates the layered concept as the MAC functions of bit-shifting and adding headers to the PDU will be done at layer-1 and managed by layer-1 interface. However, this design is efficient as it requires less copying of actual data. The overheads on protocol are reduced. Only one copy of the data is maintained within the protocol stack. The time critical operations are now done in layer-1, which may be implemented either in hardware or software. The pointers are finally released at layer-1 interface based on the context of the pointers. If it is a pointer to an MAC header it is released immediately. If it is a pointer to the RLC PDU, only the pointer to the RLC SDU can be released and this is done only by the RLC.

Passing by pointers brings up some design issues that could argue against this particular option:

1. For each RLC PDU there may be an acknowledgement. RLC should wait for a particular time and then discard that SDU if no proper acknowledgment is received for the PDUs of the SDU. Discarding the SDU could pose a problem as layer-1 interface could still be using some pointers to PDUs of the SDU. Handling this scenario requires extra control within the stack.
2. Passing by the pointers mechanism is applicable only in the uplink.

Pointer passing works well for speech applications, where there is no segmentation and RLC operates in the transparent mode – data can directly flow to the MAC and layer-1. It is also possible that data need not go through the MAC and can go to layer-1 directly. If this happens, choosing the correct TFIs for all transport channels on the CCTrCH now becomes a layer-1 function. No MAC headers are required.

Some recommendations are as follows.

1. Each layer need not be a process. Whether certain layers can be combined into a single task without resulting in increased complexity must be explored. Sharing of variables across layers must also be explored.
2. Protocol stack should preferably maintain only one copy of the data which is referenced by relevant layers through pointers. Pointer management and signaling becomes complex. The viability of this method must be analyzed in detail before it is implemented.

17.5 Hardware/Software Partitioning

The required functionality of a system can be designed by hardware and by software. The decision whether to implement a function in hardware or software is usually a tradeoff between expense, performance, and complexity. Everything that is implemented by software is flexible to being changed at any stage of the design phase, because in today's changing world the customer requirements also change up until the end stage of the system design. So, if the design specification changes the software can be easily changed to accommodate the new requirements or modify the required changes.

However, the hardware is not so flexible, once the device is chosen and the circuit schematic, lay out, PCB fabrication, and assembling of the devices on the board are done then that is it, no more changes can be made, although the FPGA, EPLD, CPLD code can be changed.

Thus from the flexibility point of view obviously the software may be a good choice. However, this is not enough, the software will run on the processor, so processor will require more processing power, for example, MIPS to run the software. Plus we also have to look at the speed of operation. Obviously, if it runs in dedicated hardware, it will be faster. In addition, use of several hardware accelerators to perform dedicated fixed tasks can help to reduce the load on the processor. Some procedures or operations are better done in hardware than in software or vice versa. The partitioning depends closely on the system architecture, time criticality, available processor MIPS, fixed or changeable specification, and several other factors. Therefore there is a need to identify an efficient partitioning before the design of the individual hardware and software.

A hardware implementation is good where less decision making is done and the tasks execute frequently and are repetitive in nature. They should also be simple to implement with as few gates as possible. Some operations are such that inherently there is some parallelism involved. This is also beneficial towards a hardware implementation as gates can be clocked in parallel.

Shifting functions to hardware will reduce the load on the processor (ARM7, ARM9 or any DSP). Another design consideration is that some functions are better run on a DSP than on ARM. Ciphering/deciphering is one such function. Figure 17.2 gives a simple interface architecture between the processor and the hardware blocks via DPRAM or DMA.

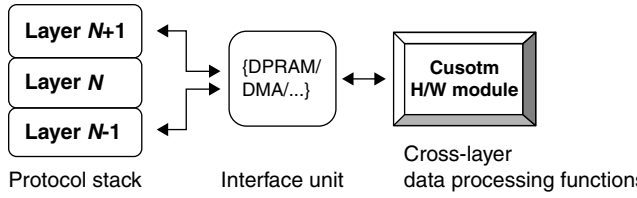


Figure 17.2 Hardware and software interfacing

Two functions of the protocol stack are considered for hardware implementation:

1. **Ciphering/deciphering:**
   a. This relates to 3G security where bits at MAC and higher layers are encrypted. The operation is repetitive. A few instances of the implementation are required as many logical channel paths within the protocol can be ciphered in parallel.
   b. The advantage is the speed, and a lesser load on the processor.
   c. The disadvantage is that hardware costs will increase as many instances are required. Another problem is that this solution is not easily scalable, if capabilities of the protocol stack changes. More DPRAM will be required, the overall system power consumption may increase, and management of data between the stack and the hardware blocks will not be easy to design. Correct ciphering parameters must be given to the hardware blocks every time ciphering is required. These interface overheads must be weighed against the advantages.
2. **Bit-shifting:**
   a. It is the function of MAC to add headers to the bit packets delivered to it by RLC. These headers need not be octet-aligned and in this case shifting of bits is required. This can be a hardware function that could be done in parallel. Parallelism can be based on bytes, words, double words or MAC SDUs.

b. The advantage is the MAC will have time to perform its other time-critical functions. Greater co-ordination between MAC and the layer-1 block doing the bit-shifting is needed.
c. The challenge is to pass the pointers and indices to the hardware block accurately. It must also be possible to power down the hardware block at the control of the protocol stack.

It is possible to combine both operations discussed above into a single hardware block if message passing is done via pointers across layers all the way to layer-1. This works well in the uplink but may not be feasible in the downlink as MAC and RLC have to decode the received information before deciphering and bit-shifting can be performed. In this case, it may be inefficient to do the operations in hardware as the decoded data must be passed back to the hardware. The basic design strategy should be such that the hardware blocks do not care if the data are meant for uplink or downlink. The hardware blocks take input data, the required parameters, then do the necessary generic operations and generate the output. The TFC selection algorithm, which is heavily used by MAC in the uplink could be implemented in the hardware. This requires closer co-ordination between MAC designers and layer-1 implementers. Layer-1 designers also need to understand the details of the algorithm for accurate implementation. The algorithm depends on lots of input parameters and therefore the method by which these are passed to the hardware blocks need to be studied.

Many modules are such that they have special processing needs and input parameters that are frequently changing. This means that although a module may be implemented in hardware, the hardware has to be interfaced carefully to the software controlling it. The inherent delays in hardware and software handshaking must be analyzed to decide if such an implementation is efficient.

Some recommendations are as follows.

1. Generally, it is difficult to decide on the right partitioning unless appropriate performance metrics are available for hardware and software implementations.
2. At the system level correct partitioning must lead to lower consumption, lower memory requirements, be easy to maintain and interfaces and also be a cost effective design.

17.6 System Performance

System performance is measured in terms of the MIPS requirement, memory requirement for code and data, power consumption, power management, and cost. These performance metrics are closely related to each other and therefore it is important to understand the interdependencies and the associated tradeoffs.

17.6.1 CPU Selection

Correct selection of the CPU is very important with respect to processing the power requirement, power consumption, and systems cost. We need to find out, for example: What MIPS (millions of instruction per second) is actually required? Is the application requiring more data multiplication and addition logic? What is the on-chip peripheral that is needed? Is there any need of a barrel shifter? Depending on the answers, the right processor should be chosen. Generally in a mobile phone we execute different applications in addition to modem processing. So, for application processing and protocol processing one or two RISC (reduced instruction set) processor/s (such as ARM) and for physical layer modem processing one DSP (digital signal processor) can be used.

17.6.2 Memory Selection

It is recognized that a system will no longer use the same type of memory for the entire system. This is because each memory type has certain advantages and each one is best suited to some specific system

requirement. The choice will be made on the basis of the memory requirement, access speed, power consumption, memory size, and cost. Table 17.2 compares the different memory technologies in terms of the relevant performance metrics.

Table 17.2 Comparison of different memory technologies

Memory	Access time (ns)	Power consumption	Cost	Internal/external
Flash	70–200	Moderate	High	Ext. (may be internal also).
SRAM	10–40	Low	Moderate	Int. (may be external also)
Cache	2–6	Low	Very high	Int. (may be external also).
DRAM	50–100	High	Low	Int. (may be external also).
ROM	60–80	Very low	Moderate	Int. (may be external also).
DPRAM	10–40	Very high	High	Ext. (may be internal also).

Several options can be considered for memory selection:

1. ISR, high priority tasks, time critical tasks will be put in memories with fast access times, such as on-chip SRAM.
2. Repetitive module codes and look-up tables can be placed in on-chip ROM to reduce the overall system power consumption.
3. Less time critical functions, rarely called functions, heap and stack spaces can be placed in DRAM for low cost and size.
4. The data section can be put in an external Flash and code in the internal memory.

The following possibilities for memory management are identified:

1. Store the data and code in the external Flash. Load the code and data into the on-chip memory (SRAM/DRAM) and run it from there. This results in faster operation but more internal memory space is required. Power consumption for the chip will increase but power consumption for the system as a whole may reduce.
2. It is also possible to store a compressed version of the program in the external memory instead of storing an uncompressed version. At run time the program will be uncompressed and copied into the internal RAM (SRAM/DRAM). Execution will be performed from the internal RAM. In this way the external Flash size requirement will be less. However, initial loading of the program code for execution will be delayed. Internal RAM space requirements to contain the program code are likely to increase.
3. The usual method is to run the code from the external memory. This results in lesser internal memory requirement and lower chip power consumption. However, the operation will be slower and the system power consumption may increase. Speed can be improved with the use of cache. Cache management could be controlled to store the most recently used instructions or to store instructions that are time critical and used often. The feasibility of implementing the latter option needs to be researched. Instructions that are part of time-critical modules but are not often used could be in internal SRAM instead of being stored in the cache.
4. We could have a separate memory management program residing in the internal memory space. The function of this program is to copy the required software code every few clock cycles from the external Flash to internal memory just before execution. This results in less internal memory

requirement. The complexity involved is in the implementation of the memory manager, which has to be intelligent enough to copy the right code at the right time without adversely affecting the speed of operation.

17.6.3 Operating System Selection

The use of an operating system is not mandatory for small embedded system programming. However, when the software becomes complex, task numbers increase, time criticality requirements come into the picture, and in such situations use of an OS become essential. The embedded OS are small in size because they lack many of the things that you find in desktop computing systems, such as a disk drive, graphical displays.

Selection of a proper OS makes the system more efficient, similarly improper selection may create an extra burden to the system. Several issues have to be considered here like:

- **Real-time requirement** – The OS should help to process the functionality on time to meet the real-time requirement for any application. The time elapsed between an interrupt is flagged and the interrupt is serviced and is known as interrupt latency. This should be as minimum as possible (for an RTOS generally it is of the order of 50 μs).
- **Stability and robustness** – Probably this is the most important feature in an OS. There are OSs which will not work unless you reboot them twice a day. However, with an embedded system, you are not supposed to reboot the machine again and again.
- **Memory management** – In an embedded system, where we can not afford to have lots of memory, as it increases cost and size, so, we have to utilize the limited memory in a more efficient way. This is why we use an OS which has a good memory management unit and a small code volume.
- **Memory leaks** – A memory leak occurs when a process requests a chunk of memory for temporary usage, but then does not subsequently release it. Thus the system cannot use that memory until it is released. This leads to a memory crisis.
- **Sharing the memory** – An OS with good memory sharing capabilities reduces memory consumption considerably.
- **Cost and support** – Of course cost is also an issue. It might happen that the cost of the OS itself becomes the major part of the entire system's cost. Sometime we look for cheaper and free OS. A free OS does not mean it is bad. Check for its support issues and see what exactly fits your needs.
- **Choosing hardware** – Choosing an efficient and economic OS does not solve your problem. Before this we should choose the correct hardware to run the selected OS, because, all OSs cannot run on all processors.

17.6.4 Power Down Mode

Power down modes can be used to turn off the device when it is not involved in any task. Several types of power down modes are possible: (1) disable the processor clock to the CPU, but allow the on chip peripherals to remain active; (2) disable the clock to the CPU, and also disable some of the on chip peripherals such as the timer, standard serial ports, TDM serial ports, but the buffered serial ports remain active; and (3) disable the clock to the CPU, disable on chip peripherals, and PLL.

Sleep mode is an extension of the selective power down strategy. The activity of the entire system is monitored rather than that of individual modules. If the system has been idle for some predetermined time out of a duration, then the entire system is shut down and this mode is called sleep mode. During sleep mode the systems' inputs are monitored for activity, which will then trigger the systems to wake up and resume processing. As there are some overheads in time and power associated with entering and leaving the sleep mode, there are some tradeoffs to be made in setting the length of the desired time out of the period.

17.6.5 Adaptive Clocking/Voltage Schemes

Adapting clocking frequency and/or supply voltage can be used to meet system performance requirements. As performance requirements of a system typically vary over time, so the number of tasks and the nature of the tasks changes over time. Hence all of the time it is not required to run the system at high speed, rather it truly depends on the operation load at that instant. It is wasteful to run the system at maximum performance all the time. Adapting the clocking frequency and adaptive supply voltage schemes help to adjust dynamically the clock frequency and supply voltage of the system based on the need, which helps to reduce power consumption.

17.6.6 Algorithm Selection

The choice of algorithm is the most highly leveraged decision in meeting the power constraints. Selecting the right algorithm reduces the power consumption and increases the operation speed dramatically. The total power consumed by the device varies based on the program activity. The ability for an algorithm to be parallelized is critical with respect to the speed of execution. The number of operations, operation type, and memory access numbers are the primary factors that contribute to power consumption. Basic computation complexity must be optimized in order to reduce MIPS requirements and the power consumption.

17.6.7 MIPS Requirement

The MIPS requirement of the system is closely related to the size, cost, and power consumption. This is the critical factor in any real-time system. MIPS requirements can be reduced through the following possible techniques:

1. Shift process intensive operations such as ciphering, bit-shifting, and TFC selection from software to hardware.
2. Optimize the code with respect to time and space.
   a. Time – frequently called functions can be made inline, provided they are reasonably small.
   b. Space – repetitive codes could be replaced with functions, especially if they are huge blocks of code.
3. Avoid division, if possible.
4. Avoid floating point operations by using fixed point libraries.
5. Use global variables to avoid message passing or parameter passing across protocol layers.

Some recommendations are as follows.

1. The peak processing load should be reduced. Simpler algorithms should be used. Memory accesses should be minimized.
2. MIPS requirement, power consumption, memory requirement, size, and cost are inter-related. All of these must be considered in any analysis.

17.7 Adaptability

Different customers may want different solutions. The design should be adaptable to varying requirements with minimum rework and modifications. Adaptability is a general word. Different types of adaptability are possible and these are addressed here. Figure 17.3 gives an illustration.

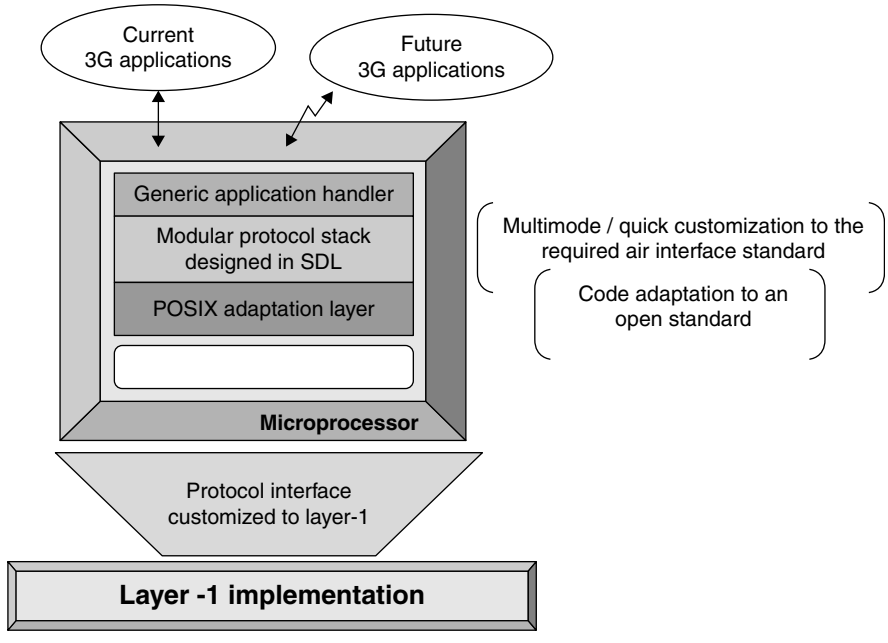


Figure 17.3 Adaptability of protocol stack

17.7.1 Adaptable to Different Physical Layer Solutions

Layer-1 could be from the customer or a third party. Different vendors may develop the physical layer, but with no change or with only minimal change the protocol software must be able to run on different physical layers.

The layer-1 interface software can be divided into two parts:

1. **Fixed portion** – which is independent of the hardware related parameters such as memory locations, size formatting, and so on.
2. **Flexible portion** – in which some parameters and settings can be changed according to the physical layer properties. In this case, layer-1 interface has to be designed specifically to that particular layer-1 under consideration. Different layer-1 designs will need different drivers. A complete adaptability to all layer-1 solutions is not feasible.

Given the above partitioning of the layer-1 interface, the objective should be to keep the design within the fixed portion as much as possible and keep within the flexible portion as little as possible.

17.7.2 Adaptable to Different Applications

This protocol should support different real time applications (video, speech, etc.) and non-real time applications (data, packet service, etc.). Instead of directly interfacing the IWU to the applications, the IWU should be interfaced to the API of the application. The API can be developed using Java. We have to define a generic interface between the API and IWU. Application developers will have to interface to the stack through these generic interface calls. This is represented in Figure 17.3 as the "Generic application

handler." Of course this method will compromise on speed to some extent. In fact, the usual solution is to have the API within the IWU. IWU isolates the specific requirements of each application from the protocol stack. For example, for speech, the application may require the corrupted data to be delivered to the higher layers. A generic signaling between IWU and lower layers should be able to control this function without lower layers knowing the type of application they are servicing.

17.7.3 Adaptable to Different OS

The protocol stack should be independent of the chosen operating system. The commonly used operating systems are Nucleus, eCOS, VxWorks, pSos, μNtron, Embedded Linux, Windows CE, and EPOC. There are two possible solutions:

1. Each RTOS has a specific adaptation layer to the protocol stack that will translate the calls of the stack to those recognized by the RTOS and vice versa. Depending on the need, the chosen RTOS can be conditionally compiled.
2. Many commercial RTOS support POSIX system calls. The adaptation layer can be designed in terms of POSIX calls, which can be used with any RTOS that supports POSIX. If an RTOS does not support POSIX calls, a separate adaptation layer can be built for it and conditionally compiled as explained earlier.

17.7.4 Adaptable to Different Air-Interface Standards

The protocol should be adaptable to different types of similar air-interface standards: WCDMA–TDD and FDD mode, CDMA-2000, IS-95, TD-SCDMA, GSM, GPRS, DECT, and AMPS. If these standards belong to the same technology group and the same RF band then the same physical devices can be used. If the devices are of completely different technologies it may be difficult to support all of them using the same protocol software. If they belong to the same category then the protocol software can be designed to support different air interface standards. This adaptability will need changes in the MAC and layer-1. Layer-1 interface software may require lesser changes.

Some recommendations are as follows.

1. Adaptability generally involves a compromise in terms of speed, extra code, and more work in the initial phases of the project.
2. Complete adaptability to all platforms and solutions is not possible but generic interfaces can be designed to reduce time to do porting.
3. A top down approach with clear partitioning and modularity are required before the protocol stack can be considered for adaptability.
4. In general, it is also easier to implement adaptability efficiently at C-code level rather than at SDL-level.

17.8 Verification, Validation, and Testing

Once the product has been developed, how can you be sure or be assured that the right product has been built or delivered? Probably the appropriate answer to that is the system is verified and validated throughout the product life cycle. This is why today advanced validation techniques are playing a pivotal role in system development. Broadly speaking, validation is concerned with building the right product and verification is concerned with building the product right. Verification involves reviews and meetings to evaluate documents, plans, code, requirements and specifications, and so on. Validation ensures that

functionality, as defined in the requirements, is the intended behavior of the product. Validation typically involves actual testing and takes place after verifications are completed. The right amount of testing helps to lower the development risks, improve product performance, ensure a high quality product, and reduce development costs. For proper testing, different test models need to studied and evaluated, such as functional requirements testing, compliance testing, interoperability testing, non-functional requirements testing, reliability and robustness testing, and so on.

Broadly, two types of tests are identified:

- **Conformance testing:** The whole protocol software should be extensively tested to ensure conformance to all the 3GPP requirements. Test scenarios must follow those specified in 3GPP documents. To start with, the tests will be performed in-house but eventually it is better to perform these tests against a UTRAN/GERAN model that is developed by a third party to ensure full 3GPP conformance. The tests performed ensure the correct functionality of each layer. It is best to do unit testing of each module and then testing after integration of all or some modules, where all the interdependent modules can be tested.
- **Performance testing:** Tests should be performed to specifically measure the performance of the system. This can be done in a modular fashion, the performance being measured at every layer or process of the stack and also being measured at the system level for different configurations and bit-rates. Scenarios to be considered are the worst case that the stack can support and a typical case.

Some recommendations are as follows.

1. Tests should be performed at every stage of development. Modular testing should be done followed by complete protocol testing after integration.
2. If the intention is to go into productization, device self-tests should be performed to enhance device reliability.

The main objective of this is to build the right product that provides consistent behavior and a quality user experience. So, use of extensive testing in product development will offer the right product to the customer along with a feeling of satisfaction. This is the key to winning the market share and brand value.

17.9 Productization

Productization is a complex stage of any project, as it involves various activities, technical and otherwise. It is essential to identify and understand the steps involved towards productization. Briefly, the following criteria are necessary when moving towards productization:

1. Adaptability and easy customization to different customer needs configuration management and version control
2. Exhaustive testing, debugging and validation
3. Product reliability
4. Advertisement and marketing
5. Customer support
6. Good documentation

Requirements from customers must be accurately recorded and be easily accessible by all designers. Any changes in the requirements must be explicitly made known to all affected designers. This also means that the person responsible for each of the deliverables must be clearly named.

A configuration management tool (such as ClearCase) is essential at all stages of the project. Any bugs detected internally or reported by customers can be fixed at short notice based on the previously

released version. Once the bugs are fixed, new releases or patches can be delivered. This greatly contributes towards shorter response times in servicing customer needs. It is also recommended that if productization is the end goal, it must be pursued from the outset. The industry must acknowledge the work of the development team and this is only possible if the team actively contributes to the 3GPP standardization process.

Some recommendations are as follows.

1. Productization needs more than just technical people.
2. Close contact with customers is crucial at all stages of development.
3. Configuration management is indispensable.
4. Defect tracking system is necessary.

At present, the mobile handset industry is the largest consumer electronics market in the world, with almost one billion mobile handsets sold every year. The major players in the market are Nokia, Motorola, Samsung, Sony Ericsson, Apple, and LG, who together account for over 80% of the shipments of the devices. Their success is largely a result of their performance on a number of criteria which include: proper market segmentation, introduction of killer applications, and R&D and cost control. Whether the market grows or shrinks and competition increases still further, these factors will remain the key to success. However, reduction in ASPs (average selling price) and convergence are obliging device vendors to review their businesses in order to hold onto the lion's share part of today's market.

Further Reading

Goodenough, J. Bruce, A., and Nightingle, A. (2001) A unified validation methodology for system level co-design and co-implementation. Paper presented at ASIC/SOC Conference, 2001. Proceedings. 14th Annual IEEE International, ARM IP Solution Division. ISBN 0-7803-6741-3.

Wallance, D.R. and Fugii, R.U. (1989) Software verification and validation: an overview. *IEEE Software*.

Zurawski, R. (2009) *Embedded Systems Handbook*, 2nd edn, CRC Press, Boca Raton, ISBN 978-1-4398-0755-2.

Index