# Overview of Wireless Communications

Wireless communications is the fastest growing segment of the communications industry. Cellular systems have experienced exponential growth over the last decades and there are currently around seven billion users worldwide. Indeed, cellular phones have become a critical business tool and part of everyday life. In addition, wireless local area networks currently supplement or replace wired networks in many homes, businesses, and campuses. Many new applications, including wireless sensor networks, automated highways and factories, smart homes and appliances, and remote telemedicine, are emerging from research ideas to concrete systems. The explosive growth of wireless systems coupled with the proliferation of laptop and palmtop computers indicate a bright future for wireless networks, both as stand-alone systems and as part of the larger networking infrastructure. However, many technical challenges remain in designing robust wireless networks that deliver the performance necessary to support emerging applications.

The first wireless networks were developed in the pre-industrial age. These systems transmitted information over line-of-sight distances (later extended by telescopes) using smoke signals, torch signaling, flashing mirrors, signal flares, or semaphore flags. An elaborate set of signal combinations was developed to convey complex messages with these rudimentary signals. Observation stations were built on hilltops and along roads to relay these messages over large distances. These early communication networks were replaced first by the telegraph network (invented by Samuel Morse in 1838) and later by the telephone. In 1895, a few decades after the telephone was invented, Marconi demonstrated the first radio transmission from the Isle of Wight to a boat 18 miles away, and radio communications was born. Radio technology advanced rapidly to enable transmissions over larger distances with better quality, less power, and smaller, cheaper devices, thereby enabling public and private radio communications, television, and wireless networking.

Early radio systems transmitted analog signals. Today most radio systems transmit digital signals composed of binary bits, where the bits are obtained directly from a data signal or by digitizing an analog signal. A digital radio can transmit a continuous bit stream or it can group the bits into packets. The latter type of radio is called a **packet radio** and is characterized by bursty transmissions: the radio is idle except when it transmits a packet. The first network based on packet radio, ALOHANET, was developed at the University of Hawaii in 1971. This network enabled computer sites at seven campuses spread out over four islands to communicate with a central computer on Oahu via radio transmission. The network architecture used a star topology with the central computer at its hub. Any two computers could establish a bi-directional communications link between them by going through the central hub. ALOHANET incorporated the first set of protocols for channel access and routing in packet radio systems, and many of the underlying principles in these protocols are still in use today. The U.S. military was extremely interested in the combination of packet data and broadcast radio inherent to ALOHANET. Throughout the 1970's and early 1980's the Defense Advanced Research Projects Agency (DARPA) invested significant resources to develop networks using packet radios for tactical communications in the battlefield. The nodes in these ad hoc wireless networks had the ability to self-configure (or reconfigure) into a network without the aid of any established infrastructure. Packet radio networks also found commercial application in supporting wide-area wireless data services. These services, first introduced in the early 1990's, enable wireless data access (including email, file transfer, and web browsing) at fairly low speeds, on the order of 20 Kbps. A strong market for these wide-area wireless data services never really materialized, due mainly to their low data rates, high cost, and lack of "killer applications". These services mostly disappeared in the 1990s,

supplanted by the wireless data capabilities of cellular telephones and wireless local area networks (LANs).

The introduction of wired Ethernet technology in the 1970's steered many commercial companies away from radio-based networking. Ethernet's 10 Mbps data rate far exceeded anything available using radio. In 1985 the Federal Communications Commission (FCC) enabled the commercial development of wireless LANs by authorizing the public use of the Industrial, Scientific, and Medical (ISM) frequency bands for wireless LAN products. The ISM band was very attractive to wireless LAN vendors since they did not need to obtain an FCC license to operate in this band. However, the wireless LAN systems could not interfere with the primary ISM band users, which forced them to use a low power profile and an inefficient signaling scheme. Moreover, the interference from primary users within this frequency band was quite high. As a result these initial wireless LANs had very poor performance in terms of data rates and coverage. This poor performance, coupled with concerns about security, lack of standardization, and high cost resulted in weak sales. Few of these systems were actually used for data networking: they were relegated to low-tech applications like inventory control. Despite the big data rate differences, wireless LANs are becoming the preferred Internet access method in many homes, offices, and campus environments due to their convenience and freedom from wires. However, most wireless LANs support applications such as email and web browsing that are not bandwidth-intensive. The challenge for future wireless LANs will be to support many users simultaneously with bandwidth-intensive and delay-constrained applications such as video. Range extension is also a critical goal for future wireless LAN systems.

By far the most successful application of wireless networking has been the cellular telephone system. The roots of this system began in 1915, when wireless voice transmission between New York and San Francisco was first established. In 1946 public mobile telephone service was introduced in 25 cities across the United States. These initial systems used a central transmitter to cover an entire metropolitan area. This inefficient use of the radio spectrum coupled with the state of radio technology at that time severely limited the system capacity: thirty years after the introduction of mobile telephone service the New York system could only support 543 users.

A solution to this capacity problem emerged during the 50's and 60's when researchers at AT&T Bell Laboratories developed the cellular concept [4]. Cellular systems exploit the fact that the power of a transmitted signal falls off with distance. Thus, two users can operate on the same frequency at spatially-separate locations with minimal interference between them. This allows very efficient use of cellular spectrum so that a large number of users can be accommodated. The evolution of cellular systems from initial concept to implementation was glacial. In 1947 AT&T requested spectrum for cellular service from the FCC. The design was mostly completed by the end of the 1960's, the first field test was in 1978, and the FCC granted service authorization in 1982, by which time much of the original technology was out-of-date. The first analog cellular system deployed in Chicago in 1983 was already saturated by 1984, at which point the FCC increased the cellular spectral allocation from 40 MHz to 50 MHz. The explosive growth of the cellular industry took almost everyone by surprise. Throughout the late 1980's, as more and more cities became saturated with demand for cellular service, the development of digital cellular technology for increased capacity and better performance became essential.

The second generation of cellular systems, first deployed in the early 1990's, were based on digital communications. The shift from analog to digital was driven by its higher capacity and the improved cost, speed, and power efficiency of digital hardware. While second generation cellular systems initially provided mainly voice services, these systems gradually evolved to support data services such as email, Internet access, and short mes-

saging. Unfortunately, the great market potential for cellular phones led to a proliferation of second generation cellular standards: three different standards in the U.S. alone, and other standards in Europe and Japan, all incompatible. The fact that different cities have different incompatible standards makes roaming throughout the U.S. and the world using one cellular phone standard impossible. Moreover, some countries have initiated service for third generation systems, for which there are also multiple incompatible standards. As a result of the standards proliferation, many cellular phones today are multi-mode: they incorporate multiple digital standards to facilitate nationwide and worldwide roaming.

Satellite systems are typically characterized by the height of the satellite orbit, low-earth orbit (LEOs at roughly 2000 km altitude), medium-earth orbit (MEOs at roughly 9000 km altitude), or geosynchronous orbit (GEOs at roughly 40,000 km altitude). The geosynchronous orbits are seen as stationary from the Earth, whereas the satellites with other orbits have their coverage area change over time. The concept of using geosynchronous satellites for communications was first suggested by the science fiction writer Arthur C. Clarke in 1945. However, the first deployed satellites, the Soviet Union's Sputnik in 1957 and the NASA/Bell Laboratories' Echo-1 in 1960, were not geosynchronous due to the difficulty of lifting a satellite into such a high orbit. The first GEO satellite was launched by Hughes and NASA in 1963. GEOs then dominated both commercial and government satellite systems for several decades.

Geosynchronous satellites have large coverage areas, so fewer satellites (and dollars) are necessary to provide wide-area or global coverage. However, it takes a great deal of power to reach the satellite, and the propagation delay is typically too large for delay-constrained applications like voice. These disadvantages caused a shift in the 1990's towards lower orbit satellites. The goal was to provide voice and data service competitive with cellular systems. However, the satellite mobile terminals were much bigger, consumed much more power, and cost much more than contemporary cellular phones, which limited their appeal. The most compelling feature of these systems is their ubiquitous worldwide coverage, especially in remote areas or third-world countries with no landline or cellular system infrastructure. Unfortunately, such places do not typically have large demand or the resources to pay for satellite service either. As cellular systems became more widespread, they took away most revenue that LEO systems might have generated in populated areas. With no real market left, most LEO satellite systems went out of business.

A natural area for satellite systems is broadcast entertainment. Direct broadcast satellites operate in the 12 GHz frequency band. These systems offer hundreds of TV channels and are major competitors to cable. Satellite delivered digital radio has also become popular. These systems, operating in both Europe and the US, offer digital audio broadcasts at near-CD quality.

## Course Syllabus

- Cellular Concept
- Radiomobile Channel
  - Large-scale Propagation Models
  - Small-scale Propagation Models
- Multiple Access Techniques
- Radio Resource Management
- 2G standards
  - GSM, DECT, cdmaOne
- 2.5G standards
  - HSCSD, GPRS, EDGE

- 3G standards
  - UMTS, cdma2000, TD-SCDMA
- Bluetooth Standard
- Cellular Network Design Principles
- Location Techniques

CHAPTER 2

# The Cellular Concept — System Design Fundamentals

The design objective of early mobile radio systems was to achieve a large coverage area by using a single, high powered transmitter with an antenna mounted on a tall tower. While this approach achieved very good coverage, it also meant that it was impossible to reuse those same frequencies throughout the system, since any attempts to achieve frequency reuse would result in interference. For example, the Bell mobile system in New York City in the 1970s could only support a maximum of twelve simultaneous calls over a thousand square miles [Cal88]. Faced with the fact that government regulatory agencies could not make spectrum allocations in proportion to the increasing demand for mobile services, it became imperative to restructure the radio telephone system to achieve high capacity with limited radio spectrum, while at the same time covering very large areas.

## 2.1 Introduction

The cellular concept was a major breakthrough in solving the problem of spectral congestion and user capacity. It offered very high capacity in a limited spectrum allocation without any major technological changes. The cellular concept is a system level idea which calls for replacing a single, high power transmitter (large cell) with many low power transmitters (small cells), each providing coverage to only a small portion of the service area. Each base station is allocated a portion of the total number of channels available to the entire system, and nearby base stations are assigned different groups of channels so that all the available channels are assigned to a relatively small number of neighboring base stations. Neighboring base stations are assigned different groups of channels so that the interference between base stations (and the mobile users

under their control) is minimized. By systematically spacing base stations and their channel groups throughout a market, the available channels are distributed throughout the geographic region and may be reused as many times as necessary, so long as the interference between co-channel stations is kept below acceptable levels.

As the demand for service increases (i.e., as more channels are needed within a particular market), the number of base stations may be increased (along with a corresponding decrease in transmitter power to avoid added interference), thereby providing additional radio capacity with no additional increase in radio spectrum. This fundamental principle is the foundation for all modern wireless communication systems, since it enables a fixed number of channels to serve an arbitrarily large number of subscribers by reusing the channels throughout the coverage region. Furthermore, the cellular concept allows every piece of subscriber equipment within a country or continent to be manufactured with the same set of channels, so that any mobile may be used anywhere within the region.

## 2.2 Frequency Reuse

Cellular radio systems rely on an intelligent allocation and reuse of channels throughout a coverage region [Oet83]. Each cellular base station is allocated a group of radio channels to be used within a small geographic area called a *cell*. Base stations in adjacent cells are assigned channel groups which contain completely different channels than neighboring cells. The base station antennas are designed to achieve the desired coverage within the particular cell. By limiting the coverage area to within the boundaries of a cell, the same group of channels may be used to cover different cells that are separated from one another by distances large enough to keep interference levels within tolerable limits. The design process of selecting and allocating channel groups for all of the cellular base stations within a system is called *frequency reuse* or *frequency planning* [Mac79].

Figure 2.1 illustrates the concept of cellular frequency reuse, where cells labeled with the same letter use the same group of channels. The frequency reuse plan is overlaid upon a map to indicate where different frequency channels are used. The hexagonal cell shape shown in Figure 2.1 is conceptual and is a simplistic model of the radio coverage for each base station, but it has been universally adopted since the hexagon permits easy and manageable analysis of a cellular system. The actual radio coverage of a cell is known as the *footprint* and is determined from field measurements or propagation prediction models. Although the real footprint is amorphous in nature, a regular cell shape is needed for systematic system design and adaptation for future growth. While it might seem natural to choose a circle to represent the coverage area of a base station, adjacent circles can not be overlaid upon a map without leaving gaps or

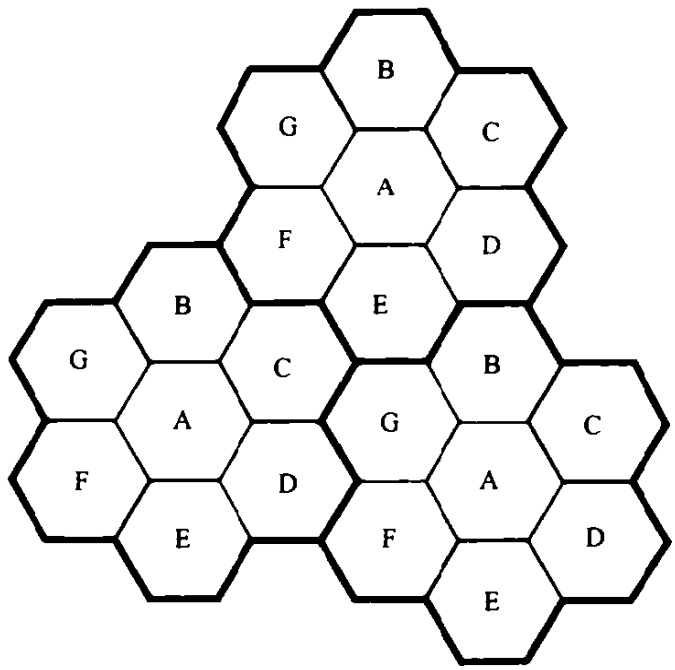



Figure 2.1
Illustration of the cellular frequency reuse concept. Cells with the same letter use the same set of frequencies. A cell cluster is outlined in bold and replicated over the coverage area. In this example, the cluster size, $N$, is equal to seven, and the frequency reuse factor is 1/7 since each cell contains one-seventh of the total number of available channels.

creating overlapping regions. Thus, when considering geometric shapes which cover an entire region without overlap and with equal area, there are three sensible choices: a square; an equilateral triangle; and a hexagon. A cell must be designed to serve the weakest mobiles within the footprint, and these are typically located at the edge of the cell. For a given distance between the center of a polygon and its farthest perimeter points, the hexagon has the largest area of the three. Thus, by using the hexagon geometry, the fewest number of cells can cover a geographic region, and the hexagon closely approximates a circular radiation pattern which would occur for an omni-directional base station antenna and free space propagation. Of course, the actual cellular footprint is determined by the contour in which a given transmitter serves the mobiles successfully.

When using hexagons to model coverage areas, base station transmitters are depicted as either being in the center of the cell (center-excited cells) or on three of the six cell vertices (edge-excited cells). Normally, omni-directional antennas are used in center-excited cells and sectored directional antennas are used in corner-excited cells. Practical considerations usually do not allow base stations to be placed exactly as they appear in the hexagonal layout. Most system designs permit a base station to be positioned up to one-fourth the cell radius away from the ideal location.

To understand the frequency reuse concept, consider a cellular system which has a total of $S$ duplex channels available for use. If each cell is allocated a group of $k$ channels ($k < S$), and if the $S$ channels are divided among $N$ cells into unique and disjoint channel groups which each have the same number of channels, the total number of available radio channels can be expressed as

$$S = kN \tag{2.1}$$

The $N$ cells which collectively use the complete set of available frequencies is called a *cluster*. If a cluster is replicated $M$ times within the system, the total number of duplex channels, $C$, can be used as a measure of capacity and is given

$$C = MkN = MS \tag{2.2}$$

As seen from equation (2.2), the capacity of a cellular system is directly proportional to the number of times a cluster is replicated in a fixed service area. The factor $N$ is called the *cluster size* and is typically equal to 4, 7, or 12. If the cluster size $N$ is reduced while the cell size is kept constant, more clusters are required to cover a given area and hence more capacity (a larger value of $C$) is achieved. A large cluster size indicates that the ratio between the cell radius and the distance between co-channel cells is large. Conversely, a small cluster size indicates that co-channel cells are located much closer together. The value for $N$ is a function of how much interference a mobile or base station can tolerate while maintaining a sufficient quality of communications. From a design viewpoint, the smallest possible value of $N$ is desirable in order to maximize capacity over a given coverage area (i.e., to maximize $C$ in equation (2.2)). The *frequency reuse factor* of a cellular system is given by $1/N$, since each cell within a cluster is only assigned $1/N$ of the total available channels in the system.

Due to the fact that the hexagonal geometry of Figure 2.1 has exactly six equidistant neighbors and that the lines joining the centers of any cell and each of its neighbors are separated by multiples of 60 degrees, there are only certain cluster sizes and cell layouts which are possible [Mac79]. In order to tessellate — to connect without gaps between adjacent cells — the geometry of hexagons is such that the number of cells per cluster, $N$, can only have values which satisfy equation (2.3).

$$N = i^2 + ij + j^2 \tag{2.3}$$

where $i$ and $j$ are non-negative integers. To find the nearest co-channel neighbors of a particular cell, one must do the following: (1) move $i$ cells along any chain of hexagons and then (2) turn 60 degrees counter-clockwise and move $j$ cells. This is illustrated in Figure 2.2 for $i = 3$ and $j = 2$ (example, $N = 19$).

---

**Example 2.1**

If a total of 33 MHz of bandwidth is allocated to a particular FDD cellular telephone system which uses two 25 kHz simplex channels to provide full duplex

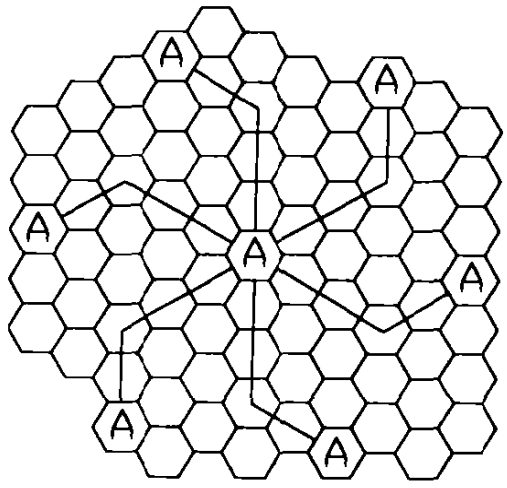



Figure 2.2
Method of locating co-channel cells in a cellular system. In this example, $N = 19$ (i.e., $i = 3, j = 2$). [Adapted from [Oet83] © IEEE].

voice and control channels, compute the number of channels available per cell if a system uses (a) 4-cell reuse, (b) 7-cell reuse (c) 12-cell reuse. If 1 MHz of the allocated spectrum is dedicated to control channels, determine an equitable distribution of control channels and voice channels in each cell for each of the three systems.

## Solution to Example 2.1

Given:
Total bandwidth = 33 MHz
Channel bandwidth = 25 kHz × 2 simplex channels = 50 kHz/duplex channel
Total available channels = 33,000/50 = 660 channels
(a) For $N = 4$,
total number of channels available per cell = $660/4 \approx 165$ channels.
(b) For $N = 7$,
total number of channels available per cell = $660/7 \approx 95$ channels.
(c) For $N = 12$,
total number of channels available per cell = $660/12 \approx 55$ channels.

A 1 MHz spectrum for control channels implies that there are 1000/50 = 20 control channels out of the 660 channels available. To evenly distribute the control and voice channels, simply allocate the same number of channels in each cell wherever possible. Here, the 660 channels must be evenly distributed to each cell within the cluster. In practice, only the 640 voice channels would be allocated, since the control channels are allocated separately as 1 per cell.
(a) For $N = 4$, we can have 5 control channels and 160 voice channels per cell. In practice, however, each cell only needs a single control channel (the control

channels have a greater reuse distance than the voice channels). Thus, one control channel and 160 voice channels would be assigned to each cell.
(b) For $N = 7$, 4 cells with 3 control channels and 92 voice channels, 2 cells with 3 control channels and 90 voice channels, and 1 cell with 2 control channels and 92 voice channels could be allocated. In practice, however, each cell would have one control channel, four cells would have 91 voice channels, and three cells would have 92 voice channels.
(c) For $N = 12$, we can have 8 cells with 2 control channels and 53 voice channels, and 4 cells with 1 control channel and 54 voice channels each. In an actual system, each cell would have 1 control channel, 8 cells would have 53 voice channels, and 4 cells would have 54 voice channels.

## 2.3 Channel Assignment Strategies

For efficient utilization of the radio spectrum, a frequency reuse scheme that is consistent with the objectives of increasing capacity and minimizing interference is required. A variety of channel assignment strategies have been developed to achieve these objectives. Channel assignment strategies can be classified as either *fixed* or *dynamic*. The choice of channel assignment strategy impacts the performance of the system, particularly as to how calls are managed when a mobile user is handed off from one cell to another [Tek91], [LiC93], [Sun94], [Rap93b].

In a fixed channel assignment strategy, each cell is allocated a predetermined set of voice channels. Any call attempt within the cell can only be served by the unused channels in that particular cell. If all the channels in that cell are occupied, the call is *blocked* and the subscriber does not receive service. Several variations of the fixed assignment strategy exist. In one approach, called the *borrowing strategy*, a cell is allowed to borrow channels from a neighboring cell if all of its own channels are already occupied. The mobile switching center (MSC) supervises such borrowing procedures and ensures that the borrowing of a channel does not disrupt or interfere with any of the calls in progress in the donor cell.

In a dynamic channel assignment strategy, voice channels are not allocated to different cells permanently. Instead, each time a call request is made, the serving base station requests a channel from the MSC. The switch then allocates a channel to the requested cell following an algorithm that takes into account the likelihood of future blocking within the cell, the frequency of use of the candidate channel, the reuse distance of the channel, and other cost functions.

Accordingly, the MSC only allocates a given frequency if that frequency is not presently in use in the cell or any other cell which falls within the minimum restricted distance of frequency reuse to avoid co-channel interference. Dynamic channel assignment reduce the likelihood of blocking, which increases the trunking capacity of the system, since all the available channels in a market are accessible to all of the cells. Dynamic channel assignment strategies require the MSC

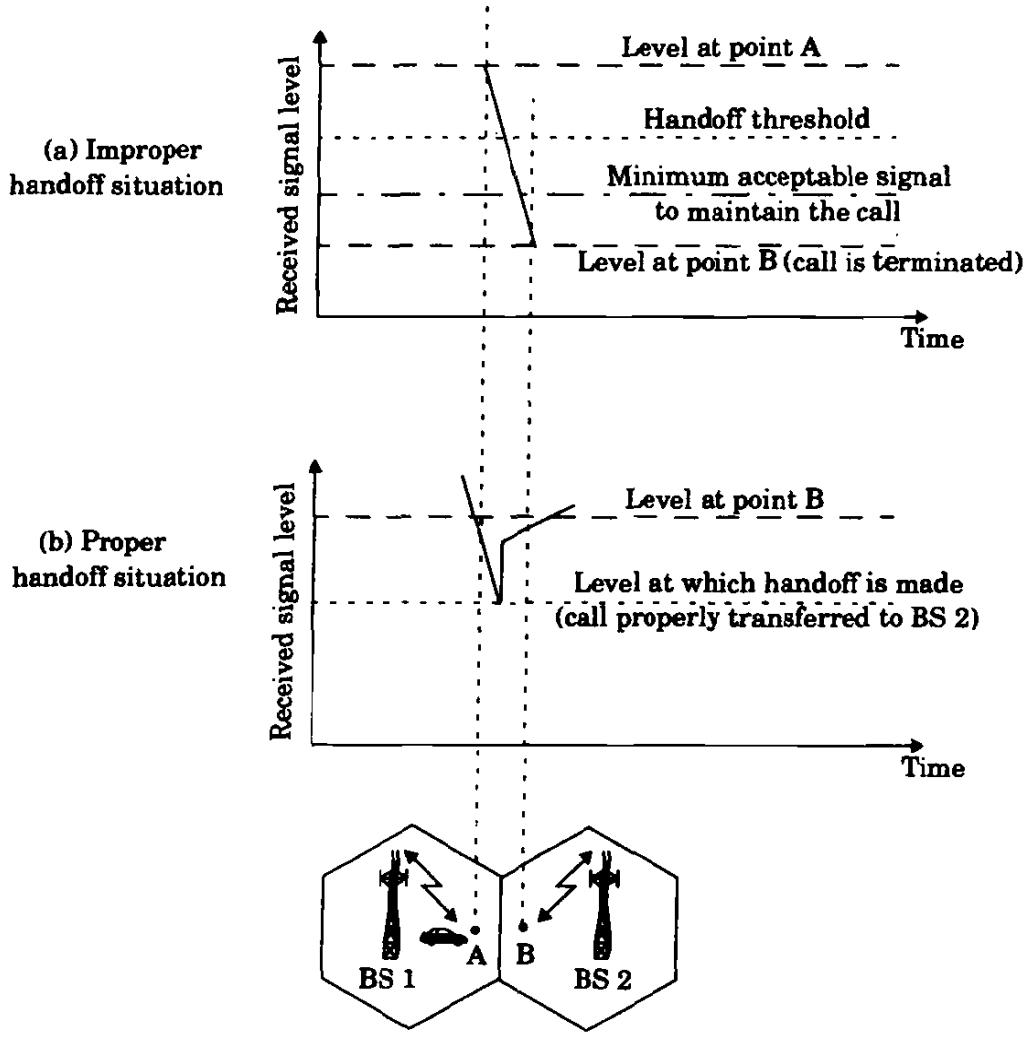



Figure 2.3
Illustration of a handoff scenario at cell boundary.

to collect real-time data on channel occupancy, traffic distribution, and *radio signal strength indications* (RSSI) of all channels on a continuous basis. This increases the storage and computational load on the system but provides the advantage of increased channel utilization and decreased probability of a blocked call.

## 2.4 Handoff Strategies

When a mobile moves into a different cell while a conversation is in progress, the MSC automatically transfers the call to a new channel belonging to the new base station. This handoff operation not only involves identifying a new base station, but also requires that the voice and control signals be allocated to channels associated with the new base station.

Processing handoffs is an important task in any cellular radio system. Many handoff strategies prioritize handoff requests over call initiation requests when allocating unused channels in a cell site. Handoffs must be performed successfully and as infrequently as possible, and be imperceptible to the users. In order to meet these requirements, system designers must specify an optimum signal level at which to initiate a handoff. Once a particular signal level is specified as the minimum usable signal for acceptable voice quality at the base station receiver (normally taken as between –90 dBm and –100 dBm), a slightly stronger signal level is used as a threshold at which a handoff is made. This margin, given by $\Delta = P_{r\ handoff} - P_{r\ minimum\ usable}$, cannot be too large or too small. If $\Delta$ is too large, unnecessary handoffs which burden the MSC may occur, and if $\Delta$ is too small, there may be insufficient time to complete a handoff before a call is lost due to weak signal conditions. Therefore, $\Delta$ is chosen carefully to meet these conflicting requirements. Figure 2.3 illustrates a handoff situation. Figure 2.3(a) demonstrates the case where a handoff is not made and the signal drops below the minimum acceptable level to keep the channel active. This dropped call event can happen when there is an excessive delay by the MSC in assigning a handoff, or when the threshold $\Delta$ is set too small for the handoff time in the system. Excessive delays may occur during high traffic conditions due to computational loading at the MSC or due to the fact that no channels are available on any of the nearby base stations (thus forcing the MSC to wait until a channel in a nearby cell becomes free).

In deciding when to handoff, it is important to ensure that the drop in the measured signal level is not due to momentary fading and that the mobile is actually moving away from the serving base station. In order to ensure this, the base station monitors the signal level for a certain period of time before a handoff is initiated. This running average measurement of signal strength should be optimized so that unnecessary handoffs are avoided, while ensuring that necessary handoffs are completed before a call is terminated due to poor signal level. The length of time needed to decide if a handoff is necessary depends on the speed at which the vehicle is moving. If the slope of the short-term average received signal level in a given time interval is steep, the handoff should be made quickly. Information about the vehicle speed, which can be useful in handoff decisions, can also be computed from the statistics of the received short-term fading signal at the base station.

The time over which a call may be maintained within a cell, without handoff, is called the *dwell time* [Rap93b]. The dwell time of a particular user is governed by a number of factors, which include propagation, interference, distance between the subscriber and the base station, and other time varying effects. Chapter 4 shows that even when a mobile user is stationary, ambient motion in the vicinity of the base station and the mobile can produce fading, thus even a stationary subscriber may have a random and finite dwell time. Analysis in

[Rap93b] indicates that the statistics of dwell time vary greatly, depending on the speed of the user and the type of radio coverage. For example, in mature cells which provide coverage for vehicular highway users, most users tend to have a relatively constant speed and travel along fixed and well-defined paths with good radio coverage. In such instances, the dwell time for an arbitrary user is a random variable with a distribution that is highly concentrated about the mean dwell time. On the other hand, for users in dense, cluttered microcell environments, there is typically a large variation of dwell time about the mean, and the dwell times are typically shorter than the cell geometry would otherwise suggest. It is apparent that the statistics of dwell time are important in the practical design of handoff algorithms [LiC93], [Sun94], [Rap93b].

In first generation analog cellular systems, signal strength measurements are made by the base stations and supervised by the MSC. Each base station constantly monitors the signal strengths of all of its reverse voice channels to determine the relative location of each mobile user with respect to the base station tower. In addition to measuring the RSSI of calls in progress within the cell, a spare receiver in each base station, called the locator receiver, is used to determine signal strengths of mobile users which are in neighboring cells. The *locator receiver* is controlled by the MSC and is used to monitor the signal strength of users in neighboring cells which appear to be in need of handoff and reports all RSSI values to the MSC. Based on the locator receiver signal strength information from each base station, the MSC decides if a handoff is necessary or not.

In second generation systems that use digital TDMA technology, handoff decisions are *mobile assisted*. In *mobile assisted handoff* (MAHO), every mobile station measures the received power from surrounding base stations and continually reports the results of these measurements to the serving base station. A handoff is initiated when the power received from the base station of a neighboring cell begins to exceed the power received from the current base station by a certain level or for a certain period of time. The MAHO method enables the call to be handed over between base stations at a much faster rate than in first generation analog systems since the handoff measurements are made by each mobile, and the MSC no longer constantly monitors signal strengths. MAHO is particularly suited for microcellular environments where handoffs are more frequent.

During the course of a call, if a mobile moves from one cellular system to a different cellular system controlled by a different MSC, an *intersystem handoff* becomes necessary. An MSC engages in an intersystem handoff when a mobile signal becomes weak in a given cell and the MSC cannot find another cell within its system to which it can transfer the call in progress. There are many issues that must be addressed when implementing an intersystem handoff. For instance, a local call may become a long-distance call as the mobile moves out of its home system and becomes a roamer in a neighboring system. Also, compati-

bility between the two MSCs must be determined before implementing an intersystem handoff. Chapter 9 demonstrates how intersystem handoffs are implemented in practice.

Different systems have different policies and methods for managing handoff requests. Some systems handle handoff requests in the same way they handle originating calls. In such systems, the probability that a handoff request will not be served by a new base station is equal to the blocking probability of incoming calls. However, from the user's point of view, having a call abruptly terminated while in the middle of a conversation is more annoying than being blocked occasionally on a new call attempt. To improve the quality of service as perceived by the users, various methods have been devised to prioritize handoff requests over call initiation requests when allocating voice channels.

### 2.4.1 Prioritizing Handoffs

One method for giving priority to handoffs is called the *guard channel concept*, whereby a fraction of the total available channels in a cell is reserved exclusively for handoff requests from ongoing calls which may be handed off into the cell. This method has the disadvantage of reducing the total carried traffic, as fewer channels are allocated to originating calls. Guard channels, however, offer efficient spectrum utilization when dynamic channel assignment strategies, which minimize the number of required guard channels by efficient demand-based allocation, are used.

Queuing of handoff requests is another method to decrease the probability of forced termination of a call due to lack of available channels. There is a tradeoff between the decrease in probability of forced termination and total carried traffic. Queuing of handoffs is possible due to the fact that there is a finite time interval between the time the received signal level drops below the handoff threshold and the time the call is terminated due to insufficient signal level. The delay time and size of the queue is determined from the traffic pattern of the particular service area. It should be noted that queuing does not guarantee a zero probability of forced termination, since large delays will cause the received signal level to drop below the minimum required level to maintain communication and hence lead to forced termination.

### 2.4.2 Practical Handoff Considerations

In practical cellular systems, several problems arise when attempting to design for a wide range of mobile velocities. High speed vehicles pass through the coverage region of a cell within a matter of seconds, whereas pedestrian users may never need a handoff during a call. Particularly with the addition of microcells to provide capacity, the MSC can quickly become burdened if high speed users are constantly being passed between very small cells. Several schemes have been devised to handle the simultaneous traffic of high speed and

low speed users while minimizing the handoff intervention from the MSC. Another practical limitation is the ability to obtain new cell sites.

Although the cellular concept clearly provides additional capacity through the addition of cell sites, in practice it is difficult for cellular service providers to obtain new physical cell site locations in urban areas. Zoning laws, ordinances, and other nontechnical barriers often make it more attractive for a cellular provider to install additional channels and base stations at the same physical location of an existing cell, rather than find new site locations. By using different antenna heights (often on the same building or tower) and different power levels, it is possible to provide "large" and "small" cells which are co-located at a single location. This technique is called the *umbrella cell* approach and is used to provide large area coverage to high speed users while providing small area coverage to users traveling at low speeds. Figure 2.4 illustrates an umbrella cell which is co-located with some smaller microcells. The umbrella cell approach ensures that the number of handoffs is minimized for high speed users and provides additional microcell channels for pedestrian users. The speed of each user may be estimated by the base station or MSC by evaluating how rapidly the short-term average signal strength on the RVC changes over time, or more sophisticated algorithms may be used to evaluate and partition users [LiC93]. If a high speed user in the large umbrella cell is approaching the base station, and its velocity is rapidly decreasing, the base station may decide to hand the user into the co-located microcell, without MSC intervention.

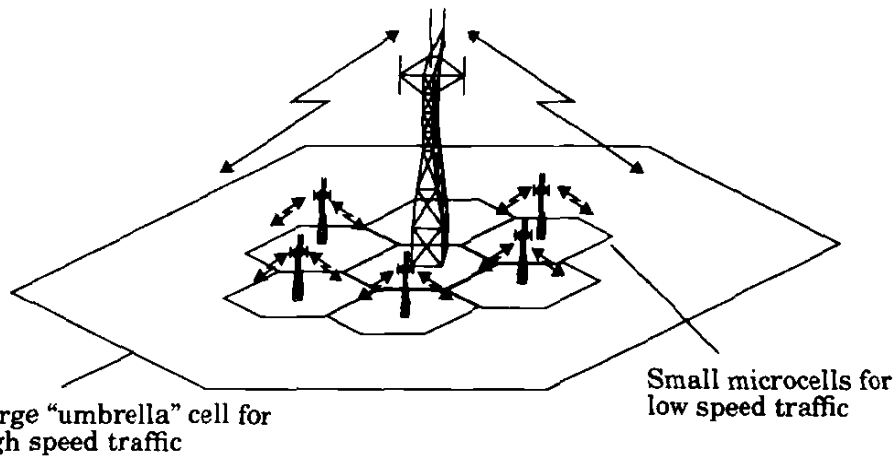



Figure 2.4
The umbrella cell approach.

Another practical handoff problem in microcell systems is known as *cell dragging*. Cell dragging results from pedestrian users that provide a very strong signal to the base station. Such a situation occurs in an urban environment when there is a line-of-sight (LOS) radio path between the subscriber and the base sta-

tion. As the user travels away from the base station at a very slow speed, the average signal strength does not decay rapidly. Even when the user has traveled well beyond the designed range of the cell, the received signal at the base station may be above the handoff threshold, thus a handoff may not be made. This creates a potential interference and traffic management problem, since the user has meanwhile traveled deep within a neighboring cell. To solve the cell dragging problem, handoff thresholds and radio coverage parameters must be adjusted carefully.

In first generation analog cellular systems, the typical time to make a handoff, once the signal level is deemed to be below the handoff threshold, is about 10 seconds. This requires that the value for $\Delta$ be on the order of 6 dB to 12 dB. In new digital cellular systems such as GSM, the mobile assists with the handoff procedure by determining the best handoff candidates, and the handoff, once the decision is made, typically requires only 1 or 2 seconds. Consequently, $\Delta$ is usually between 0 dB and 6 dB in modern cellular systems. The faster handoff process supports a much greater range of options for handling high speed and low speed users and provides the MSC with substantial time to "rescue" a call that is in need of handoff.

Another feature of newer cellular systems is the ability to make handoff decisions based on a wide range of metrics other than signal strength. The co-channel and adjacent channel interference levels may be measured at the base station or the mobile, and this information may be used with conventional signal strength data to provide a multi-dimensional algorithm for determining when a handoff is needed.

The IS-95 code division multiple access (CDMA) spread spectrum cellular system described in Chapter 10, provides a unique handoff capability that cannot be provided with other wireless systems. Unlike channelized wireless systems that assign different radio channels during a handoff (called a *hard handoff*), spread spectrum mobiles share the same channel in every cell. Thus, the term *handoff* does not mean a physical change in the assigned channel, but rather that a different base station handles the radio communication task. By simultaneously evaluating the received signals from a single subscriber at several neighboring base stations, the MSC may actually decide which version of the user's signal is best at any moment in time. This technique exploits macroscopic space diversity provided by the different physical locations of the base stations and allows the MSC to make a "soft" decision as to which version of the user's signal to pass along to the PSTN at any instance [Pad94]. The ability to select between the instantaneous received signals from a variety of base stations is called *soft handoff*.

## 2.5 Interference and System Capacity

Interference is the major limiting factor in the performance of cellular radio systems. Sources of interference include another mobile in the same cell, a call in progress in a neighboring cell, other base stations operating in the same frequency band, or any noncellular system which inadvertently leaks energy into the cellular frequency band. Interference on voice channels causes cross talk, where the subscriber hears interference in the background due to an undesired transmission. On control channels, interference leads to missed and blocked calls due to errors in the digital signaling. Interference is more severe in urban areas, due to the greater RF noise floor and the large number of base stations and mobiles. Interference has been recognized as a major bottleneck in increasing capacity and is often responsible for dropped calls. The two major types of system-generated cellular interference are co-*channel interference* and *adjacent channel interference*. Even though interfering signals are often generated within the cellular system, they are difficult to control in practice (due to random propagation effects). Even more difficult to control is interference due to out-of-band users, which arises without warning due to front end overload of subscriber equipment or intermittent intermodulation products. In practice, the transmitters from competing cellular carriers are often a significant source of out-of-band interference, since competitors often locate their base stations in close proximity to one another in order to provide comparable coverage to customers.

### 2.5.1 Co-channel Interference and System Capacity

Frequency reuse implies that in a given coverage area there are several cells that use the same set of frequencies. These cells are called *co-channel cells*, and the interference between signals from these cells is called *co-channel interference*. Unlike thermal noise which can be overcome by increasing the signal-to-noise ration (SNR), co-channel interference cannot be combated by simply increasing the carrier power of a transmitter. This is because an increase in carrier transmit power increases the interference to neighboring co-channel cells. To reduce co-channel interference, co-channel cells must be physically separated by a minimum distance to provide sufficient isolation due to propagation.

When the size of each cell is approximately the same, and the base stations transmit the same power, the co-channel interference ratio is independent of the transmitted power and becomes a function of the radius of the cell ($R$) and the distance between centers of the nearest co-channel cells ($D$). By increasing the ratio of $D/R$, the spatial separation between co-channel cells relative to the coverage distance of a cell is increased. Thus interference is reduced from improved isolation of RF energy from the co-channel cell. The parameter $Q$, called the *co-channel reuse ratio*, is related to the cluster size. For a hexagonal geometry

$$Q = \frac{D}{R} = \sqrt{3N} \tag{2.4}$$

A small value of $Q$ provides larger capacity since the cluster size $N$ is small, whereas a large value of $Q$ improves the transmission quality, due to a smaller level of co-channel interference. A trade-off must be made between these two objectives in actual cellular design.

**Table 2.1 Co-channel Reuse Ratio for Some Values of N**

| | Cluster Size ($N$) | Co-channel Reuse Ratio($Q$) |
|---|---|---|
| $i = 1, j = 1$ | 3 | 3 |
| $i = 1, j = 2$ | 7 | 4.58 |
| $i = 2, j = 2$ | 12 | 6 |
| $i = 1, j = 3$ | 13 | 6.24 |

Let $i_o$ be the number of co-channel interfering cells. Then, the signal-to-interference ratio ($S/I$ or $SIR$) for a mobile receiver which monitors a forward channel can be expressed as

$$\frac{S}{I} = \frac{S}{\sum_{i=1}^{i_0} I_i} \tag{2.5}$$

where $S$ is the desired signal power from the desired base station and $I_i$ is the interference power caused by the $i$ th interfering co-channel cell base station. If the signal levels of co-channel cells are known, then the $S/I$ ratio for the forward link can be found using equation (2.5).

Propagation measurements in a mobile radio channel show that the average received signal strength at any point decays as a power law of the distance of separation between a transmitter and receiver. The average received power $P_r$ at a distance $d$ from the transmitting antenna is approximated by

$$P_r = P_0\left(\frac{d}{d_0}\right)^{-n} \tag{2.6}$$

or

$$P_r(\text{dBm}) = P_0(\text{dBm}) - 10n\log\left(\frac{d}{d_0}\right) \tag{2.7}$$

where $P_0$ is the power received at a close-in reference point in the far field region of the antenna at a small distance $d_0$ from the transmitting antenna, and $n$ is the path loss exponent. Now consider the forward link where the desired signal is the serving base station and where the interference is due to co-channel base stations. If $D_i$ is the distance of the $i$ th interferer from the mobile, the received power at a given mobile due to the $i$ th interfering cell will be proportional to

$(D_i)^{-n}$ . The path loss exponent typically ranges between 2 and 4 in urban cellular systems [Rap92b].

When the transmit power of each base station is equal and the path loss exponent is the same throughout the coverage area, $S/I$ for a mobile can be approximated as

$$\frac{S}{I} = \frac{R^{-n}}{\sum_{i=1}^{i_0} (D_i)^{-n}} \tag{2.8}$$

Considering only the first layer of interfering cells, if all the interfering base stations are equidistant from the desired base station and if this distance is equal to the distance $D$ between cell centers, then equation (2.8) simplifies to

$$\frac{S}{I} = \frac{(D/R)^n}{i_0} = \frac{(\sqrt{3N})^n}{i_0} \tag{2.9}$$

Equation (2.9) relates $S/I$ to the cluster size $N$, which in turn determines the overall capacity of the system from equation (2.2). For example, assume that the six closest cells are close enough to create significant interference and that they are all approximately equal distance from the desired base station. For the U.S. AMPS cellular system which uses FM and 30 kHz channels, subjective tests indicate that sufficient voice quality is provided when $S/I$ is greater than or equal to 18 dB. Using equation (2.9) it can be shown in order to meet this requirement, the cluster size $N$ should be at least 6.49, assuming a path loss exponent $n = 4$. Thus a minimum cluster size of 7 is required to meet an $S/I$ requirement of 18 dB. It should be noted that equation (2.9) is based on the hexagonal cell geometry where all the interfering cells are equidistant from the base station receiver, and hence provides an optimistic result in many cases. For some frequency reuse plans (e.g. $N = 4$), the closest interfering cells vary widely in their distances from the desired cell.

From Figure 2.5, it can be seen for a 7-cell cluster, with the mobile unit is at the cell boundary, the mobile is a distance $D - R$ from the two nearest co-channel interfering cells and approximately $D + R/2$, $D$, $D - R/2$, and $D + R$ from the other interfering cells in the first tier [Lee86]. Using equation (2.9) and assuming $n$ equals 4, the signal-to-interference ratio for the worst case can be closely approximated as (an exact expression is worked out by Jacobsmeyer [Jac94]).

$$\frac{S}{I} = \frac{R^{-4}}{2(D-R)^{-4} + 2(D+R)^{-4} + 2D^{-4}} \tag{2.10}$$

Equation (2.10) can be rewritten in terms of the co-channel reuse ratio $Q$, as

$$\frac{S}{I} = \frac{1}{2(Q-1)^{-4} + 2(Q+1)^{-4} + 2Q^{-4}} \tag{2.11}$$

For $N = 7$, the co-channel reuse ratio $Q$ is 4.6, and the worst case $S/I$ is approximated as 49.56 (17 dB) using equation (2.11), whereas an exact solution using equation (2.8) yields 17.8 dB [Jac94]. Hence for a 7-cell cluster, the $S/I$ ratio is slightly less than 18 dB for the worst case. To design the cellular system for proper performance in the worst case , it would be necessary to increase $N$ to the next largest size, which from equation (2.3) is found to be 12 (corresponding to $i = j = 2$). This obviously entails a significant decrease in capacity, since 12-cell reuse offers a spectrum utilization of 1/12 within each cell, whereas 7-cell reuse offers a spectrum utilization of 1/7. In practice, a capacity reduction of 7/12 would not be tolerable to accommodate for the worst case situation which rarely occurs. From the above discussion it is clear that co-channel interference determines link performance, which in turn dictates the frequency reuse plan and the overall capacity of cellular systems.

**Example 2.2**

If a signal to interference ratio of 15 dB is required for satisfactory forward channel performance of a cellular system, what is the frequency reuse factor and cluster size that should be used for maximum capacity if the path loss exponent is (a) $n = 4$, (b) $n = 3$? Assume that there are 6 co-channels cells in the first tier, and all of them are at the same distance from the mobile. Use suitable approximations.

**Solution to Example 2.2**

(a) $n = 4$

First, let us consider a 7-cell reuse pattern.

Using equation (2.4), the co-channel reuse ratio $D/R = 4.583$.

Using equation (2.9), the signal-to-noise interference ratio is given by

$$S/I = (1/6) \times (4.583)^4 = 75.3 = 18.66 \text{ dB}.$$

Since this is greater than the minimum required $S/I$, $N = 7$ can be used.

b) $n = 3$

First, let us consider a 7-cell reuse pattern.

Using equation (2.9), the signal-to-interference ratio is given by

$$S/I = (1/6) \times (4.583)^3 = 16.04 = 12.05 \text{ dB}.$$

Since this is less than the minimum required $S/I$, we need to use a larger $N$.

Using equation (2.3), the next possible value of $N$ is 12, ($i = j = 2$).

The corresponding co-channel ratio is given by equation (2.4) as

$$D/R = 6.0.$$

Using equation (2.3) the signal-to-interference ratio is given by

$$S/I = (1/6)\times(6)^3 = 36 = 15.56 \text{ dB}.$$

Since this is greater than the minimum required $S/I$, $N = 12$ can be used.

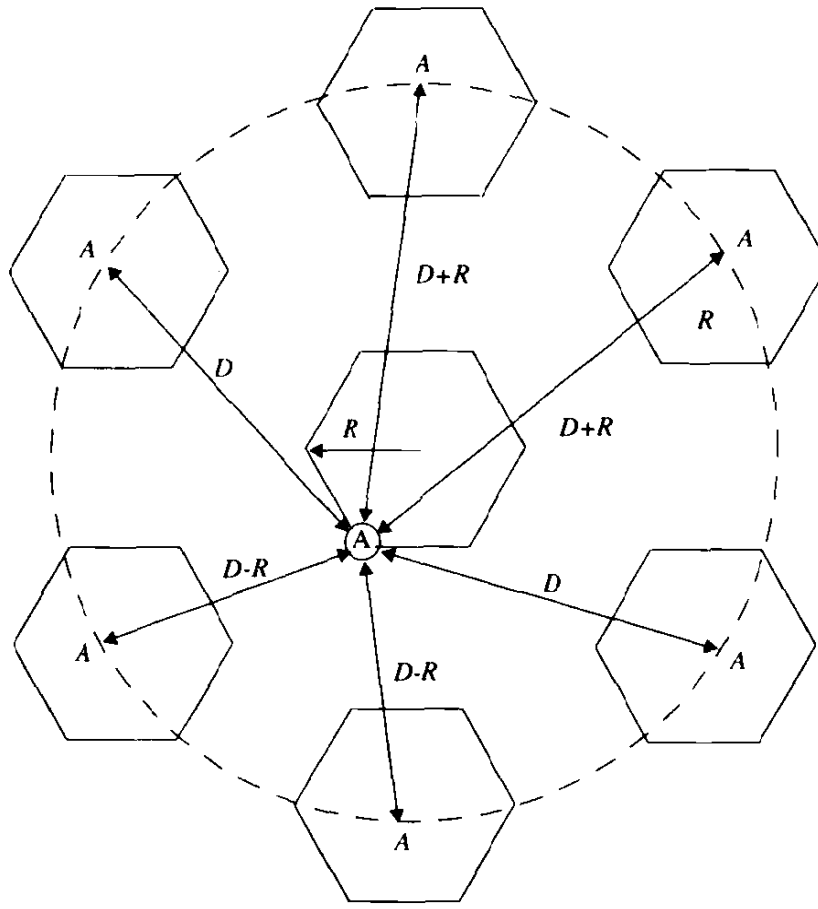



Figure 2.5
Illustration of the first tier of co-channel cells for a cluster size of $N$=7. When the mobile is at the cell boundary (point $A$), it experiences worst case co-channel interference on the forward channel. The marked distances between the mobile and different co-channel cells are based on approximations made for easy analysis.

### 2.5.2 Adjacent Channel Interference

Interference resulting from signals which are adjacent in frequency to the desired signal is called *adjacent channel interference*. Adjacent channel interference results from imperfect receiver filters which allow nearby frequencies to leak into the passband. The problem can be particularly serious if an adjacent channel user is transmitting in very close range to a subscriber's receiver, while the receiver attempts to receive a base station on the desired channel. This is referred to as the *near-far* effect, where a nearby transmitter (which may or may

CHAPTER 3

# Mobile Radio Propagation: Large-Scale Path Loss

The mobile radio channel places fundamental limitations on the performance of wireless communication systems. The transmission path between the transmitter and the receiver can vary from simple line-of-sight to one that is severely obstructed by buildings, mountains, and foliage. Unlike wired channels that are stationary and predictable, radio channels are extremely random and do not offer easy analysis. Even the speed of motion impacts how rapidly the signal level fades as a mobile terminal moves in space. Modeling the radio channel has historically been one of the most difficult parts of mobile radio system design, and is typically done in a statistical fashion, based on measurements made specifically for an intended communication system or spectrum allocation.

## 3.1 Introduction to Radio Wave Propagation

The mechanisms behind electromagnetic wave propagation are diverse, but can generally be attributed to reflection, diffraction, and scattering. Most cellular radio systems operate in urban areas where there is no direct line-of-sight path between the transmitter and the receiver, and where the presence of high-rise buildings causes severe diffraction loss. Due to multiple reflections from various objects, the electromagnetic waves travel along different paths of varying lengths. The interaction between these waves causes multipath fading at a specific location, and the strengths of the waves decrease as the distance between the transmitter and receiver increases.

Propagation models have traditionally focused on predicting the average received signal strength at a given distance from the transmitter, as well as the variability of the signal strength in close spatial proximity to a particular loca-

tion. Propagation models that predict the mean signal strength for an arbitrary transmitter-receiver (T-R) separation distance are useful in estimating the radio coverage area of a transmitter and are called *large-scale* propagation models, since they characterize signal strength over large T-R separation distances (several hundreds or thousands of meters). On the other hand, propagation models that characterize the rapid fluctuations of the received signal strength over very short travel distances (a few wavelengths) or short time durations (on the order of seconds) are called *small-scale* or *fading* models.

As a mobile moves over very small distances, the instantaneous received signal strength may fluctuate rapidly giving rise to small-scale fading. The reason for this is that the received signal is a sum of many contributions coming from different directions, as described in Chapter 4. Since the phases are random, the sum of the contributions varies widely; for example, obeys a Rayleigh fading distribution. In small-scale fading, the received signal power may vary by as much as three or four orders of magnitude (30 or 40 dB) when the receiver is moved by only a fraction of a wavelength. As the mobile moves away from the transmitter over much larger distances, the local average received signal will gradually decrease, and it is this local average signal level that is predicted by large-scale propagation models. Typically, the local average received power is computed by averaging signal measurements over a measurement track of $5\lambda$ to $40\lambda$. For cellular and PCS frequencies in the 1 GHz to 2 GHz band, this corresponds to measuring the local average received power over movements of 1 m to 10 m.

Figure 3.1 illustrates small-scale fading and the slower large-scale variations for an indoor radio communication system. Notice in the figure that the signal fades rapidly as the receiver moves, but the local average signal changes much more slowly with distance. This chapter covers large-scale propagation and presents a number of common methods used to predict received power in mobile communication systems. Chapter 4 treats small-scale fading models and describes methods to measure and model multipath in the mobile radio environment.

## 3.2 Free Space Propagation Model

The free space propagation model is used to predict received signal strength when the transmitter and receiver have a clear, unobstructed line-of-sight path between them. Satellite communication systems and microwave line-of-sight radio links typically undergo free space propagation. As with most large-scale radio wave propagation models, the free space model predicts that received power decays as a function of the T-R separation distance raised to some power (i.e. a power law function). The free space power received by a receiver antenna which is separated from a radiating transmitter antenna by a distance $d$, is given by the Friis free space equation,

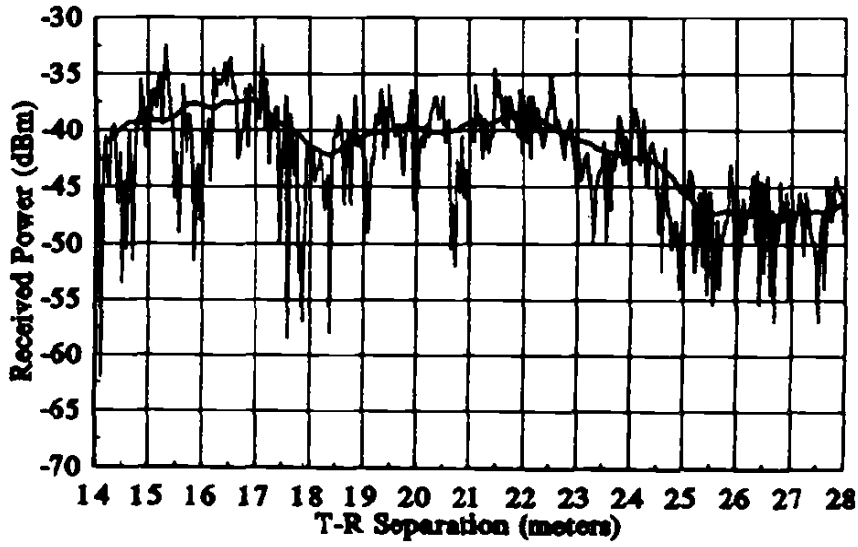



Figure 3.1
Small-scale and large-scale fading.

$$P_r(d) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^2 L} \tag{3.1}$$

where $P_t$ is the transmitted power, $P_r(d)$ is the received power which is a function of the T-R separation, $G_t$ is the transmitter antenna gain, $G_r$ is the receiver antenna gain, $d$ is the T-R separation distance in meters, $L$ is the system loss factor not related to propagation ($L \geq 1$), and $\lambda$ is the wavelength in meters. The gain of an antenna is related to its effective aperture, $A_e$, by

$$G = \frac{4\pi A_e}{\lambda^2} \tag{3.2}$$

The effective aperture $A_e$ is related to the physical size of the antenna, and $\lambda$ is related to the carrier frequency by

$$\lambda = \frac{c}{f} = \frac{2\pi c}{\omega_c} \tag{3.3}$$

where $f$ is the carrier frequency in Hertz, $\omega_c$ is the carrier frequency in radians per second, and $c$ is the speed of light given in meters/s. The values for $P_t$ and $P_r$ must be expressed in the same units, and $G_t$ and $G_r$ are dimensionless quantities. The miscellaneous losses $L$ ($L \geq 1$) are usually due to transmission line attenuation, filter losses, and antenna losses in the communication system. A value of $L = 1$ indicates no loss in the system hardware.

The Friis free space equation of (3.1) shows that the received power falls off as the square of the T-R separation distance. This implies that the received power decays with distance at a rate of 20 dB/decade.

An *isotropic* radiator is an ideal antenna which radiates power with unit gain uniformly in all directions, and is often used to reference antenna gains in wireless systems. The *effective isotropic radiated power (EIRP)* is defined as

$$EIRP = P_t G_t \tag{3.4}$$

and represents the maximum radiated power available from a transmitter in the direction of maximum antenna gain, as compared to an isotropic radiator.

In practice, *effective radiated power (ERP)* is used instead of EIRP to denote the maximum radiated power as compared to a half-wave dipole antenna (instead of an isotropic antenna). Since a dipole antenna has a gain of 1.64 (2.15 dB above an isotrope), the ERP will be 2.15 dB smaller than the EIRP for the same transmission system. In practice, antenna gains are given in units of dBi (dB gain with respect to an isotropic source) or dBd (dB gain with respect to a half-wave dipole) [Stu81].

The *path loss*, which represents signal attenuation as a positive quantity measured in dB, is defined as the difference (in dB) between the effective transmitted power and the received power, and may or may not include the effect of the antenna gains. The path loss for the free space model when antenna gains are included is given by

$$PL\,(\text{dB}) = 10\log\frac{P_t}{P_r} = -10\log\left[\frac{G_t G_r \lambda^2}{(4\pi)^2 d^2}\right] \tag{3.5}$$

When antenna gains are excluded, the antennas are assumed to have unity gain, and path loss is given by

$$PL\,(\text{dB}) = 10\log\frac{P_t}{P_r} = -10\log\left[\frac{\lambda^2}{(4\pi)^2 d^2}\right] \tag{3.6}$$

The Friis free space model is only a valid predictor for $P_r$ for values of $d$ which are in the far-field of the transmitting antenna. The far-field, or *Fraunhofer region*, of a transmitting antenna is defined as the region beyond the far-field distance $d_f$, which is related to the largest linear dimension of the transmitter antenna aperture and the carrier wavelength. The Fraunhofer distance is given by

$$d_f = \frac{2D^2}{\lambda} \tag{3.7.a}$$

where $D$ is the largest physical linear dimension of the antenna. Additionally, to be in the far-field region, $d_f$ must satisfy

$$d_f \gg D \tag{3.7.b}$$

and

$$d_f \gg \lambda \tag{3.7.c}$$

Furthermore, it is clear that equation (3.1) does not hold for $d = 0$. For this reason, large-scale propagation models use a close-in distance, $d_0$, as a known received power reference point. The received power, $P_r(d)$, at any distance $d > d_0$, may be related to $P_r$ at $d_0$. The value $P_r(d_0)$ may be predicted from equation (3.1), or may be measured in the radio environment by taking the average received power at many points located at a close-in radial distance $d_0$ from the transmitter. The reference distance must be chosen such that it lies in the far-field region, that is, $d_0 \geq d_f$, and $d_0$ is chosen to be smaller than any practical distance used in the mobile communication system. Thus, using equation (3.1), the received power in free space at a distance greater than $d_0$ is given by

$$P_r(d) = P_r(d_0)\left(\frac{d_0}{d}\right)^2 \qquad d \geq d_0 \geq d_f \tag{3.8}$$

In mobile radio systems, it is not uncommon to find that $P_r$ may change by many orders of magnitude over a typical coverage area of several square kilometers. Because of the large dynamic range of received power levels, often dBm or dBW units are used to express received power levels. Equation (3.8) may be expressed in units of dBm or dBW by simply taking the logarithm of both sides and multiplying by 10. For example, if $P_r$ is in units of dBm, the received power is given by

$$P_r(d) \text{ dBm} = 10\log\left[\frac{P_r(d_0)}{0.001 \text{ W}}\right] + 20\log\left(\frac{d_0}{d}\right) \qquad d \geq d_0 \geq d_f \tag{3.9}$$

where $P_r(d_0)$ is in units of watts.

The reference distance $d_0$ for practical systems using low-gain antennas in the 1-2 GHz region is typically chosen to be 1 m in indoor environments and 100 m or 1 km in outdoor environments, so that the numerator in equations (3.8) and (3.9) is a multiple of 10. This makes path loss computations easy in dB units.

---

**Example 3.1**

Find the far-field distance for an antenna with maximum dimension of 1 m and operating frequency of 900 MHz.

**Solution to Example 3.1**

Given:

Largest dimension of antenna, $D = 1$ m

Operating frequency $f = 900$ MHz, $\lambda = c/f = \dfrac{3 \times 10^8 \text{ m/s}}{900 \times 10^6 \text{ Hz}}$ m

Using equation (3.7.a), far-field distance is obtained as

$$d_f = \frac{2(1)^2}{0.33} = 6 \text{ m}$$

**Example 3.2**

If a transmitter produces 50 watts of power, express the transmit power in units of (a) dBm, and (b) dBW. If 50 watts is applied to a unity gain antenna with a 900 MHz carrier frequency, find the received power in dBm at a free space distance of 100 m from the antenna. What is $P_r(10 \text{ km})$ ? Assume unity gain for the receiver antenna.

**Solution to Example 3.2**

Given:
Transmitter power, $P_t = 50$ W.
Carrier frequency, $f_c = 900$ MHz

Using equation (3.9),
(a) Transmitter power,

$$P_t(\text{dBm}) = 10\log[P_t(\text{mW})/(1 \text{ mW})]$$
$$= 10\log[50 \times 10^3] = 47.0 \text{ dBm}.$$

(b) Transmitter power,

$$P_t(\text{dBW}) = 10\log[P_t(\text{W})/(1 \text{ W})]$$
$$= 10\log[50] = 17.0 \text{ dBW}.$$

The received power can be determined using equation (3.1).

$$P_r = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^2 L} = \frac{50(1)(1)(1/3)^2}{(4\pi)^2(100)^2(1)} = 3.5 \times 10^{-6} \text{ W} = 3.5 \times 10^{-3} \text{ mW}$$

$$P_r(\text{dBm}) = 10\log P_r(\text{mW}) = 10\log\left(3.5 \times 10^{-3} \text{ mW}\right) = -24.5 \text{ dBm}.$$

The received power at 10 km can be expressed in terms of dBm using equation (3.9), where $d_0 = 100$ m and $d = 10$ km

$$P_r(10 \text{ km}) = P_r(100) + 20\log\left[\frac{100}{10000}\right] = -24.5 \text{ dBm} - 40 \text{ dB}$$
$$= -64.5 \text{ dBm}.$$

## 3.3 Relating Power to Electric Field

The free space path loss model of Section 3.2 is readily derived from first principles. It can be proven that any radiating structure produces electric and magnetic fields [Gri87], [Kra50]. Consider a small linear radiator of length $L$, that is placed coincident with the z-axis and has its center at the origin, as shown in Figure 3.2.

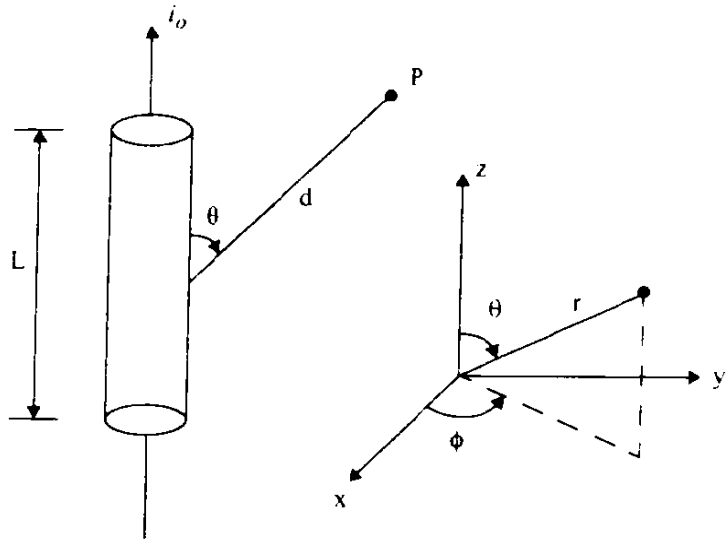



Figure 3.2
Illustration of a linear radiator of length $L$ ($L \ll \lambda$), carrying a current of amplitude $i_0$ and making an angle θ with a point, at distance $d$.

If a current flows through such an antenna, it launches electric and magnetic fields that can be expressed as

$$E_r = \frac{i_0 L \cos\theta}{2\pi\varepsilon_0 c}\left\{\frac{1}{d^2} + \frac{c}{j\omega_c d^3}\right\} e^{j\omega_c(t - d/c)} \tag{3.10}$$

$$E_\theta = \frac{i_0 L \sin\theta}{4\pi\varepsilon_0 c^2}\left\{\frac{j\omega_c}{d} + \frac{c}{d^2} + \frac{c^2}{j\omega_c d^3}\right\} e^{-j\omega_c(t - d/c)} \tag{3.11}$$

$$H_\phi = \frac{i_0 L \sin\theta}{4\pi c}\left\{\frac{j\omega_c}{d} + \frac{c}{d^2}\right\} e^{j\omega_c(t - d/c)} \tag{3.12}$$

with $E_\phi = H_r = H_\theta = 0$. In the above equations, all $1/d$ terms represent the radiation field component, all $1/d^2$ terms represent the induction field component, and all $1/d^3$ terms represent the electrostatic field component. As seen from equations (3.10) to (3.12), the electrostatic and inductive fields decay much faster with distance than the radiation field. At regions far away from the transmitter (far-field region), the electrostatic and inductive fields become negligible and only the radiated field components of $E_\theta$ and $H_\phi$ need be considered.

In free space, the *power flux density* $P_d$ (expressed in W/m$^2$) is given by

$$P_d = \frac{EIRP}{4\pi d^2} = \frac{P_t G_t}{4\pi d^2} = \frac{E^2}{R_{fs}} = \frac{E^2}{\eta} \text{ W/m}^2 \tag{3.13}$$

where $R_{fs}$ is the intrinsic impedance of free space given by $\eta = 120\pi\ \Omega$ ($377\Omega$). Thus, the power flux density is

$$P_d = \frac{|E|^2}{377\Omega} \text{ W/m}^2 \tag{3.14}$$

where $|E|$ represents the magnitude of the radiating portion of the electric field in the far field. Figure 3.3a illustrates how the power flux density disperses in free space from an isotropic point source. $P_d$ may be thought of as the *EIRP* divided by the surface area of a sphere with radius $d$. The power received at distance $d$, $P_r(d)$, is given by the power flux density times the effective aperture of the receiver antenna, and can be related to the electric field using equations (3.1), (3.2), (3.13), and (3.14).

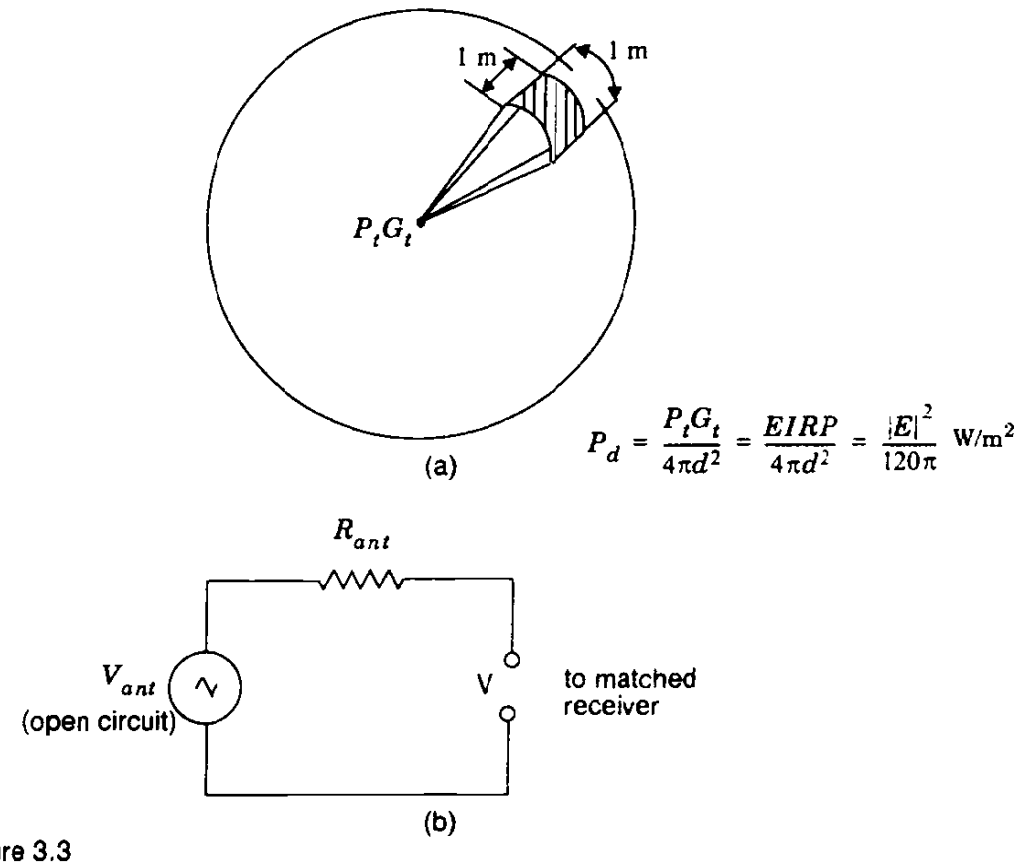



Figure 3.3
(a) Power flux density at a distance $d$ from a point source.
(b) Model for voltage applied to the input of a receiver.

$$P_r(d) = P_d A_e = \frac{|E|^2}{120\pi} A_e = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^2} \text{ Watts} \tag{3.15}$$

Equation (3.15) relates electric field (with units of V/m) to received power (with units of watts), and is identical to equation (3.1) with $L = 1$.

Often it is useful to relate the received power level to a receiver input voltage, as well as to an induced E-field at the receiver antenna. If the receiver

antenna is modeled as a matched resistive load to the receiver, then the receiver antenna will induce an rms voltage into the receiver which is half of the open circuit voltage at the antenna. Thus, if $V$ is the rms voltage at the input of a receiver (measured by a high impedance voltmeter), and $R_{ant}$ is the resistance of the matched receiver, the received power is given by

$$P_r(d) = \frac{V^2}{R_{ant}} = \frac{[V_{ant}/2]^2}{R_{ant}} = \frac{V_{ant}^2}{4R_{ant}} \tag{3.16}$$

Through equations (3.14) to (3.16), it is possible to relate the received power to the received E-field or the open circuit rms voltage at the receiver antenna terminals. Figure 3.3b illustrates an equivalent circuit model. Note $V_{ant} = V$ when there is no load.

**Example 3.3**

Assume a receiver is located 10 km from a 50 W transmitter. The carrier frequency is 900 MHz, free space propagation is assumed, $G_t = 1$, and $G_r = 2$, find (a) the power at the receiver, (b) the magnitude of the E-field at the receiver antenna (c) the rms voltage applied to the receiver input assuming that the receiver antenna has a purely real impedance of 50 Ω and is matched to the receiver.

**Solution to Example 3.3**

Given:

Transmitter power, $P_t = 50$ W
Carrier frequency, $f_c = 900$ MHz
Transmitter antenna gain, $G_t = 1$
Receiver antenna gain, $G_r = 2$
Receiver antenna resistance = 50 Ω

(a) Using equation (3.5), the power received at a distance $d = 10$ km is

$$P_r(d) = 10\log\left(\frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^2}\right) = 10\log\left(\frac{50 \times 1 \times 2 \times (1/3)^2}{(4\pi)^2 10000^2}\right)$$

$$= -91.5 \text{ dBW} = -61.5 \text{ dBm}$$

(b) Using equation (3.15), the magnitude of the received E-field is

$$|E| = \sqrt{\frac{P_r(d)120\pi}{A_e}} = \sqrt{\frac{P_r(d)120\pi}{G_r\lambda^2/4\pi}} = \sqrt{\frac{7\times10^{-10}\times120\pi}{2\times0.33^2/4\pi}} = 0.0039 \text{ V/m}$$

(c) Using equation (3.16), the open circuit rms voltage at the receiver input is

$$V_{ant} = \sqrt{P_r(d)\times4R_{ant}} = \sqrt{7\times10^{-10}\times4\times50} = 0.374 \text{ mV}$$

## 3.4 The Three Basic Propagation Mechanisms

Reflection, diffraction, and scattering are the three basic propagation mechanisms which impact propagation in a mobile communication system. These mechanisms are briefly explained in this section, and propagation models which describe these mechanisms are discussed subsequently in this chapter. Received power (or its reciprocal, path loss) is generally the most important parameter predicted by large-scale propagation models based on the physics of reflection, scattering, and diffraction. Small-scale fading and multipath propagation (discussed in Chapter 4) may also be described by the physics of these three basic propagation mechanisms.

*Reflection* occurs when a propagating electromagnetic wave impinges upon an object which has very large dimensions when compared to the wavelength of the propagating wave. Reflections occur from the surface of the earth and from buildings and walls.

*Diffraction* occurs when the radio path between the transmitter and receiver is obstructed by a surface that has sharp irregularities (edges). The secondary waves resulting from the obstructing surface are present throughout the space and even behind the obstacle, giving rise to a bending of waves around the obstacle, even when a line-of-sight path does not exist between transmitter and receiver. At high frequencies, diffraction, like reflection, depends on the geometry of the object, as well as the amplitude, phase, and polarization of the incident wave at the point of diffraction.

*Scattering* occurs when the medium through which the wave travels consists of objects with dimensions that are small compared to the wavelength, and where the number of obstacles per unit volume is large. Scattered waves are produced by rough surfaces, small objects, or by other irregularities in the channel. In practice, foliage, street signs, and lamp posts induce scattering in a mobile communications system.

## 3.5 Reflection

When a radio wave propagating in one medium impinges upon another medium having different electrical properties, the wave is partially reflected and partially transmitted. If the plane wave is incident on a perfect dielectric, part of the energy is transmitted into the second medium and part of the energy is reflected back into the first medium, and there is no loss of energy in absorption. If the second medium is a perfect conductor, then *all* incident energy is reflected back into the first medium without loss of energy. The electric field intensity of the reflected and transmitted waves may be related to the incident wave in the medium of origin through the *Fresnel reflection coefficient* ($\Gamma$). The reflection coefficient is a function of the material properties, and generally depends on the wave polarization, angle of incidence, and the frequency of the propagating wave.

In general, electromagnetic waves are *polarized*, meaning they have instantaneous electric field components in orthogonal directions in space. A polarized wave may be mathematically represented as the sum of two spatially orthogonal components, such as vertical and horizontal, or left-hand or right-hand circularly polarized components. For an arbitrary polarization, superposition may be used to compute the reflected fields from a reflecting surface.

### 3.5.1 Reflection from Dielectrics

Figure 3.4 shows an electromagnetic wave incident at an angle $\theta_i$ with the plane of the boundary between two dielectric media. As shown in the figure, part of the energy is reflected back to the first media at an angle $\theta_r$, and part of the energy is transmitted (refracted) into the second media at an angle $\theta_t$. The nature of reflection varies with the direction of polarization of the E-field. The behavior for arbitrary directions of polarization can be studied by considering the two distinct cases shown in Figure 3.4. The *plane of incidence* is defined as the plane containing the incident, reflected, and transmitted rays [Ram65]. In Figure 3.4a, the E-field polarization is parallel with the plane of incidence (that is, the E-field has a vertical polarization, or normal component, with respect to the reflecting surface) and in Figure 3.4b, the E-field polarization is perpendicular to the plane of incidence (that is, the incident E-field is pointing out of the page towards the reader, and is perpendicular to the page and parallel to the reflecting surface).

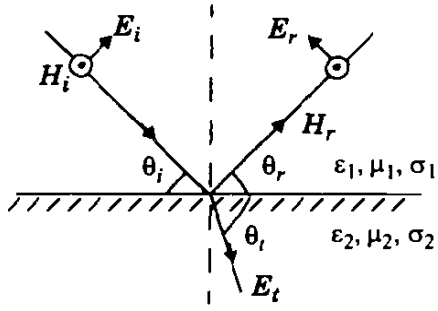
In Figure 3.4, the subscripts *i, r, t* refer to the incident, reflected, and transmitted fields, respectively. Parameters $\varepsilon_1, \mu_1, \sigma_1$, and $\varepsilon_2, \mu_2, \sigma_2$ represent the permittivity, permeability, and conductance of the two media, respectively. Often, the dielectric constant of a perfect (lossless) dielectric is related to a relative value of permittivity, $\varepsilon_r$, such that $\varepsilon = \varepsilon_0\varepsilon_r$, where $\varepsilon_0$ is a constant given by $8.85 \times 10^{-12}$ F/m. If a dielectric material is lossy, it will absorb power and may be described by a complex dielectric constant given by

$$\varepsilon = \varepsilon_0\varepsilon_r - j\varepsilon' \tag{3.17}$$

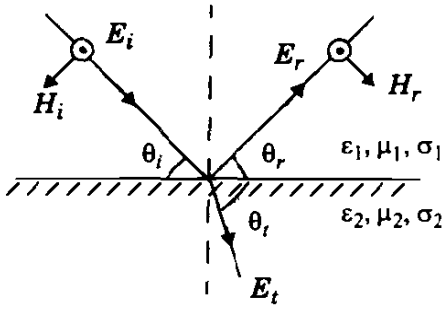
where,

$$\varepsilon' = \frac{\sigma}{2\pi f} \tag{3.18}$$

and $\sigma$ is the conductivity of the material measured in Siemens/meter. The terms $\varepsilon_r$ and $\sigma$ are generally insensitive to operating frequency when the material is a good conductor ($f < \sigma / (\varepsilon_0\varepsilon_r)$ ). For lossy dielectrics, $\varepsilon_0$ and $\varepsilon_r$ are generally constant with frequency, but $\sigma$ may be sensitive to the operating frequency, as shown in Table 3.1. Electrical properties of a wide range of materials were characterized over a large frequency range by Von Hipple [Von54].




(a) E-field in the plane of incidence




(b) E-field normal to the plane of incidence

Figure 3.4
Geometry for calculating the reflection coefficients between two dielectrics.

**Table 3.1 Material Parameters at Various Frequencies**

| Material | Relative Permittivity $\varepsilon_r$ | Conductivity $\sigma$ (s/m) | Frequency (MHz) |
|---|---|---|---|
| Poor Ground | 4 | 0.001 | 100 |
| Typical Ground | 15 | 0.005 | 100 |
| Good Ground | 25 | 0.02 | 100 |
| Sea Water | 81 | 5.0 | 100 |
| Fresh Water | 81 | 0.001 | 100 |
| Brick | 4.44 | 0.001 | 4000 |
| Limestone | 7.51 | 0.028 | 4000 |
| Glass, Corning 707 | 4 | 0.00000018 | 1 |
| Glass, Corning 707 | 4 | 0.000027 | 100 |
| Glass, Corning 707 | 4 | 0.005 | 10000 |

Because of superposition, only two orthogonal polarizations need be considered to solve general reflection problems. The reflection coefficients for the two cases of parallel and perpendicular E-field polarization at the boundary of two dielectrics are given by

$$\Gamma_{\parallel} = \frac{E_r}{E_i} = \frac{\eta_2 \sin\theta_t - \eta_1 \sin\theta_i}{\eta_2 \sin\theta_t + \eta_1 \sin\theta_i} \quad \text{(E-field in plane of incidence)} \tag{3.19}$$

$$\Gamma_{\perp} = \frac{E_r}{E_i} = \frac{\eta_2 \sin\theta_i - \eta_1 \sin\theta_t}{\eta_2 \sin\theta_i + \eta_1 \sin\theta_t} \quad \text{(E-field not in plane of incidence)} \tag{3.20}$$

where $\eta_i$ is the intrinsic impedance of the $i$ th medium ($i = 1, 2$), and is given by $\sqrt{\mu_i/\varepsilon_i}$, the ratio of electric to magnetic field for a uniform plane wave in the particular medium. The velocity of an electromagnetic wave is given by $1/(\sqrt{\mu\varepsilon})$, and the boundary conditions at the surface of incidence obey Snell's Law which, referring to Figure 3.4, is given by

$$\sqrt{\mu_1\varepsilon_1}\sin(90-\theta_i) = \sqrt{\mu_2\varepsilon_2}\sin(90-\theta_t) \tag{3.21}$$

The boundary conditions from Maxwell's equations are used to derive equations (3.19) and (3.20) as well as equations (3.22), (3.23.a), and (3.23.b).

$$\theta_i = \theta_r \tag{3.22}$$

and

$$E_r = \Gamma E_i \tag{3.23.a}$$

$$E_t = (1+\Gamma)E_i \tag{3.23.b}$$

where $\Gamma$ is either $\Gamma_{\parallel}$ or $\Gamma_{\perp}$, depending on polarization.

For the case when the first medium is free space and $\mu_1 = \mu_2$, the reflection coefficients for the two cases of vertical and horizontal polarization can be simplified to

$$\Gamma_{\parallel} = \frac{-\varepsilon_r \sin\theta_i + \sqrt{\varepsilon_r - \cos^2\theta_i}}{\varepsilon_r \sin\theta_i + \sqrt{\varepsilon_r - \cos^2\theta_i}} \tag{3.24}$$

and

$$\Gamma_{\perp} = \frac{\sin\theta_i - \sqrt{\varepsilon_r - \cos^2\theta_i}}{\sin\theta_i + \sqrt{\varepsilon_r - \cos^2\theta_i}} \tag{3.25}$$

For the case of elliptical polarized waves, the wave may be broken down (depolarized) into its vertical and horizontal E-field components, and superposition may be applied to determine transmitted and reflected waves. In the general case of reflection or transmission, the horizontal and vertical axes of the spatial coordinates may not coincide with the perpendicular and parallel axes of the propagating waves. An angle $\theta$ measured counter-clockwise from the horizontal axis is defined as shown in Figure 3.5 for a propagating wave out of the page (towards the reader) [Stu93]. The vertical and horizontal field components at a dielectric boundary may be related by

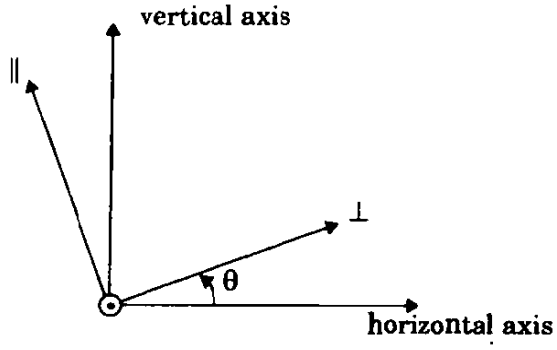



Figure 3.5
Axes for orthogonally polarized components. Parallel and perpendicular components are related to the horizontal and vertical spatial coordinates. Wave is shown propagating out of the page towards the reader.

$$\begin{bmatrix} E_H^d \\ E_V^d \end{bmatrix} = R^T D_C R \begin{bmatrix} E_H^i \\ E_V^i \end{bmatrix} \tag{3.26}$$

where $E_H^d$ and $E_V^d$ are the depolarized field components in the horizontal and vertical directions, respectively, $E_H^i$ and $E_V^i$ are the horizontally and vertically polarized components of the incident wave, respectively, and $E_H^d$, $E_V^d$, $E_H^i$, and $E_V^i$ are time varying components of the E-field which may be represented as phasors. $R$ is a transformation matrix which maps vertical and horizontal polarized components to components which are perpendicular and parallel to the plane of incidence. The matrix $R$ is given by

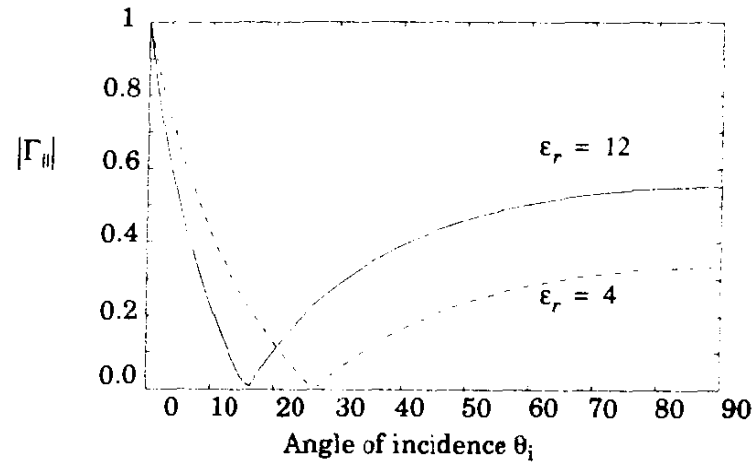
$$R = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

where $\theta$ is the angle between the two sets of axes, as shown in Figure 3.5. The depolarization matrix $D_C$ is given by

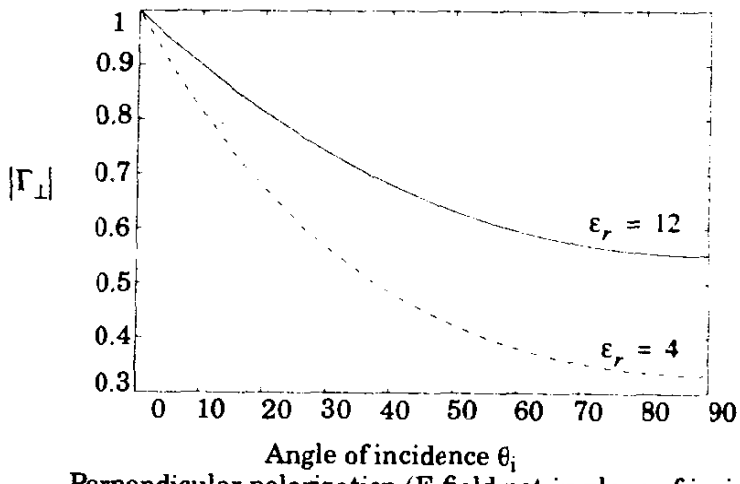
$$D_C = \begin{bmatrix} D_{\perp\perp} & 0 \\ 0 & D_{\parallel\parallel} \end{bmatrix}$$

where $D_{xx} = \Gamma_x$ for the case of reflection and $D_{xx} = T_x = 1 + \Gamma_x$ for the case of transmission [Stu93].

Figure 3.6 shows a plot of the reflection coefficient for both horizontal and vertical polarization as a function of the incident angle for the case when a wave propagates in free space ($\varepsilon_r = 1$) and the reflection surface has (a) $\varepsilon_r = 4$, and (b) $\varepsilon_r = 12$.




Parallel polarization (E-field in plane of incidence)




Perpendicular polarization (E-field not in plane of incidence)

Figure 3.6
Magnitude of reflection coefficients as a function of angle of incidence for $\varepsilon_r = 4, \varepsilon_r = 12$, using geometry in Figure 3.4.

## Example 3.4

Demonstrate that if medium 1 is free space and medium 2 is a dielectric, both $|\Gamma_{\parallel}|$ and $|\Gamma_{\perp}|$ approach 1 as $\theta_i$ approaches 0° regardless of $\varepsilon_r$.

## Solution to Example 3.4

Substituting $\theta_i = 0°$ in equation (3.24)

$$\Gamma_{\parallel} = \frac{-\varepsilon_r \sin 0 + \sqrt{\varepsilon_r - \cos^2 0}}{\varepsilon_r \sin 0 + \sqrt{\varepsilon_r - \cos^2 0}}$$

$$\Gamma_{\parallel} = \frac{\sqrt{\varepsilon_r - 1}}{\sqrt{\varepsilon_r - 1}}$$

$$= 1$$

Substituting $\theta_i = 0^o$ in equation (3.25)

$$\Gamma_{\perp} = \frac{\sin 0 - \sqrt{\varepsilon_r - \cos^2 0}}{\sin 0 + \sqrt{\varepsilon_r - \cos^2 0}}$$

$$\Gamma_{\perp} = \frac{-\sqrt{\varepsilon_r - 1}}{\sqrt{\varepsilon_r - 1}}$$

$$= -1.$$

This example illustrates that ground may be modeled as a perfect reflector with a reflection coefficient of unit magnitude when an incident wave grazes the earth, regardless of polarization or ground dielectric properties (some texts define the direction of $E_r$ to be opposite to that shown in Figure 3.4a, resulting in $\Gamma = -1$ for both parallel and perpendicular polarization).

### 3.5.2 Brewster Angle

The *Brewster angle* is the angle at which no reflection occurs in the medium of origin. It occurs when the incident angle $\theta_B$ is such that the reflection coefficient $\Gamma_{\parallel}$ is equal to zero (see Figure 3.6). The Brewster angle is given by the value of $\theta_B$ which satisfies

$$\sin(\theta_B) = \sqrt{\frac{\varepsilon_1}{\varepsilon_1 + \varepsilon_2}} \tag{3.27}$$

For the case when the first medium is free space and the second medium has a relative permittivity $\varepsilon_r$, equation (3.27) can be expressed as

$$\sin(\theta_B) = \frac{\sqrt{\varepsilon_r - 1}}{\sqrt{\varepsilon_r^2 - 1}} \tag{3.28}$$

Note that the Brewster angle occurs only for vertical (i.e. parallel) polarization.

**Example 3.5**

Calculate the Brewster angle for a wave impinging on ground having a permittivity of $\varepsilon_r = 4$.

**Solution to Example 3.5**

The Brewster angle can be found by substituting the values for $\varepsilon_r$ in equation (3.28).

$$\sin(\theta_i) = \frac{\sqrt{(4)-1}}{\sqrt{(4)^2-1}} = \sqrt{\frac{3}{15}} = \sqrt{\frac{1}{5}}$$

$$\theta_i = \sin^{-1}\sqrt{\frac{1}{5}} = 26.56^o$$

Thus Brewster angle for $\varepsilon_r = 4$ is equal to $26.56°$.

### 3.5.3 Reflection from Perfect Conductors

Since electromagnetic energy cannot pass through a perfect conductor a plane wave incident on a conductor has all of its energy reflected. As the electric field at the surface of the conductor must be equal to zero at all times in order to obey Maxwell's equations, the reflected wave must be equal in magnitude to the incident wave. For the case when E-field polarization is in the plane of incidence, the boundary conditions require that [Ram65]

$$\theta_i = \theta_r \tag{3.29}$$

and

$$E_i = E_r \qquad \text{(E-field in plane of incidence)} \tag{3.30}$$

Similarly, for the case when the E-field is horizontally polarized, the boundary conditions require that

$$\theta_i = \theta_r \tag{3.31}$$

and

$$E_i = -E_r \qquad \text{(E-field not in plane of incidence)} \tag{3.32}$$

Referring to equations (3.29) to (3.32), we see that for a perfect conductor, $\Gamma_{\parallel} = 1$, and $\Gamma_{\perp} = -1$, regardless of incident angle. Elliptical polarized waves may be analyzed by using superposition, as shown in Figure 3.5 and equation (3.26).

## 3.6 Ground Reflection (2-ray) Model

In a mobile radio channel, a single direct path between the base station and a mobile is seldom the only physical means for propagation, and hence the free space propagation model of equation (3.5) is in most cases inaccurate when used alone. The 2-ray ground reflection model shown in Figure 3.7 is a useful propagation model that is based on geometric optics, and considers both the direct path and a ground reflected propagation path between transmitter and receiver. This model has been found to be reasonably accurate for predicting the large-scale signal strength over distances of several kilometers for mobile radio systems

that use tall towers (heights which exceed 50 m), as well as for line-of-sight microcell channels in urban environments [Feu94].

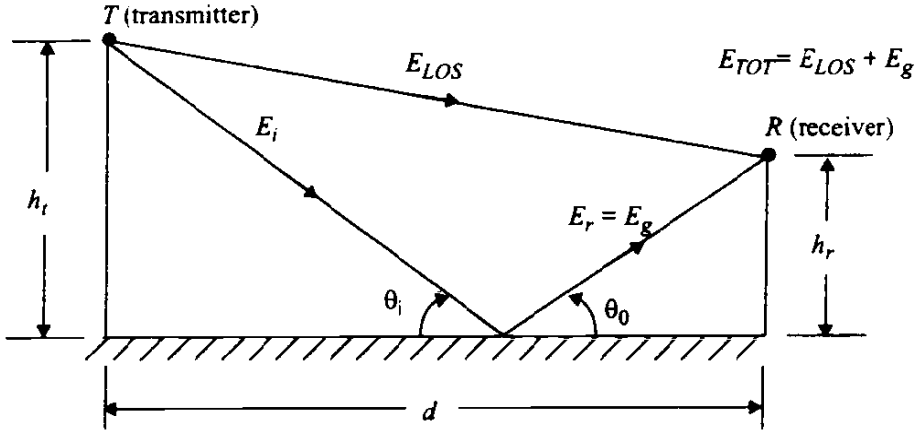



Figure 3.7
Two-ray ground reflection model.

In most mobile communication systems, the maximum T-R separation distance is at most only a few tens of kilometers, and the earth may be assumed to be flat. The total received E-field, $E_{TOT}$, is then a result of the direct line-of-sight component, $E_{LOS}$, and the ground reflected component, $E_g$.

Referring to Figure 3.7, $h_t$ is the height of the transmitter and $h_r$ is the height of the receiver. If $E_0$ is the free space E-field (in units of V/m) at a reference distance $d_0$ from the transmitter, then for $d > d_0$, the free space propagating E-field is given by

$$E(d,t) = \frac{E_0 d_0}{d}\cos\left(\omega_c\left(t - \frac{d}{c}\right)\right) \qquad (d > d_0) \tag{3.33}$$

where $|E(d,t)| = E_0 d_0 / d$ represents the envelope of the E-field at $d$ meters from the transmitter.

Two propagating waves arrive at the receiver: the direct wave that travels a distance $d'$; and the reflected wave that travels a distance $d''$. The E-field due to the line-of-sight component at the receiver can be expressed as

$$E_{LOS}(d',t) = \frac{E_0 d_0}{d'}\cos\left(\omega_c\left(t - \frac{d'}{c}\right)\right) \tag{3.34}$$

and the E-field for the ground reflected wave, which has a propagation distance of $d''$, can be expressed as

$$E_g(d'',t) = \Gamma\frac{E_0 d_0}{d''}\cos\left(\omega_c\left(t - \frac{d''}{c}\right)\right) \tag{3.35}$$

According to laws of reflection in dielectrics given in Section 3.5.1

$$\theta_i = \theta_0 \tag{3.36}$$

and

$$E_g = \Gamma E_i \tag{3.37.a}$$

$$E_t = (1+\Gamma)E_i \tag{3.37.b}$$

where $\Gamma$ is the reflection coefficient for ground. For small values of $\theta_i$ (i.e., grazing incidence), the reflected wave is equal in magnitude and $180^0$ out of phase with the incident wave, as shown in Example 3.4. The resultant E-field, assuming perfect ground reflection (i.e., $\Gamma = -1$ and $E_t = 0$) is the vector sum of $E_{LOS}$ and $E_g$, and the resultant total E-field envelope is given by

$$|E_{TOT}| = |E_{LOS} + E_g| \tag{3.38}$$

The electric field $E_{TOT}(d,t)$ can be expressed as the sum of equations (3.34) and (3.35)

$$E_{TOT}(d,t) = \frac{E_0 d_0}{d'}\cos\left(\omega_c\left(t-\frac{d'}{c}\right)\right) + (-1)\frac{E_0 d_0}{d''}\cos\left(\omega_c\left(t-\frac{d''}{c}\right)\right) \tag{3.39}$$

Using the *method of images*, which is demonstrated by the geometry of Figure 3.8, the path difference, $\Delta$, between the line-of-sight and the ground reflected paths can be expressed as

$$\Delta = d'' - d' = \sqrt{(h_t+h_r)^2 + d^2} - \sqrt{(h_t-h_r)^2 + d^2} \tag{3.40}$$

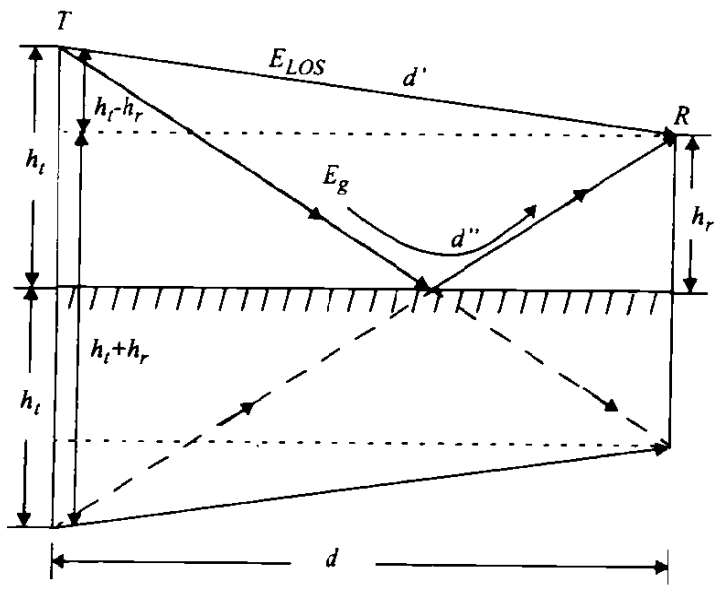



Figure 3.8
The method of images is used to find the path difference between the line-of-sight and the ground reflected paths.

When the T-R separation distance $d$ is very large compared to $h_t + h_r$, equation (3.40) can be simplified using a Taylor series approximation

$$\Delta = d'' - d' \approx \frac{2h_t h_r}{d} \tag{3.41}$$

Once the path difference is known, the phase difference $\theta_\Delta$ between the two E-field components and the time delay $\tau_d$ between the arrival of the two components can be easily computed using the following relations

$$\theta_\Delta = \frac{2\pi\Delta}{\lambda} = \frac{\Delta\omega_c}{c} \tag{3.42}$$

and

$$\tau_d = \frac{\Delta}{c} = \frac{\theta_\Delta}{2\pi f_c} \tag{3.43}$$

It should be noted that as $d$ becomes large, the difference between the distances $d'$ and $d''$ becomes very small, and the amplitudes of $E_{LOS}$ and $E_g$ are virtually identical and differ only in phase. That is

$$\left|\frac{E_0 d_0}{d}\right| \approx \left|\frac{E_0 d_0}{d'}\right| \approx \left|\frac{E_0 d_0}{d''}\right| \tag{3.44}$$

If the received E-field is evaluated at some time, say at $t = d''/c$, equation (3.39) can be expressed as



$$\begin{aligned} E_{TOT}\left(d, t = \frac{d''}{c}\right) &= \frac{E_0 d_0}{d'}\cos\left(\omega_c\left(\frac{d''-d'}{c}\right)\right) - \frac{E_0 d_0}{d''}\cos 0^\circ \\ &= \frac{E_0 d_0}{d'}\cos\theta_\Delta - \frac{E_0 d_0}{d''} \\ &\approx \frac{E_o d_0}{d}\left[\cos\theta_\Delta - 1\right] \end{aligned} \tag{3.45}$$

where $d$ is the distance over a flat earth between the bases of the transmitter and receiver antennas. Referring to the phasor diagram of Figure 3.9 which shows how the direct and ground reflected rays combine, the electric field (at the receiver) at a distance $d$ from the transmitter can be written as

$$|E_{TOT}(d)| = \sqrt{\left(\frac{E_0 d_0}{d}\right)^2 (\cos\theta_\Delta - 1)^2 + \left(\frac{E_0 d_0}{d}\right)^2 \sin^2\theta_\Delta} \tag{3.46}$$

or

$$|E_{TOT}(d)| = \frac{E_0 d_0}{d}\sqrt{2 - 2\cos\theta_\Delta} \tag{3.47}$$

Using trigonometric identities, equation (3.47) can be expressed as

$$|E_{TOT}(d)| = 2\frac{E_0 d_0}{d}\sin\left(\frac{\theta_\Delta}{2}\right) \tag{3.48}$$

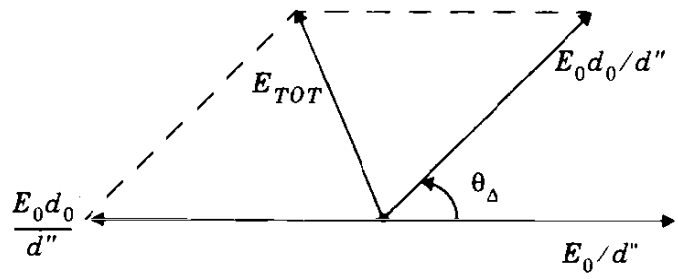



Figure 3.9
Phasor diagram showing the electric field components of the line-of-sight, ground reflected, and total received E-fields, derived from equation (3.45).

Note that equation (3.48) may be simplified whenever $\sin(\theta_\Delta/2) \approx \theta_\Delta/2$. This occurs when $\theta_\Delta/2$ is less than 0.3 radian. Using equations (3.41) and (3.42)

$$\frac{\theta_\Delta}{2} \approx \frac{2\pi h_t h_r}{\lambda d} < 0.3 \text{ rad} \tag{3.49}$$

which implies that

$$d > \frac{20\pi h_t h_r}{3\lambda} \approx \frac{20 h_t h_r}{\lambda} \tag{3.50}$$

Thus as long as $d$ satisfies (3.50), the received E-field can be approximated as

$$E_{TOT}(d) \approx \frac{2E_0 d_0}{d}\frac{2\pi h_t h_r}{\lambda d} \approx \frac{k}{d^2} \text{ V/m} \tag{3.51}$$

where $k$ is a constant related to $E_0$, the antenna heights, and the wavelength. The power received at $d$ is related to the square of the electric field through equation (3.15). Combining equations (3.2), (3.15), and (3.51), the received power at a distance $d$ from the transmitter can be expressed as

$$P_r = P_t G_t G_r \frac{h_t^2 h_r^2}{d^4} \tag{3.52}$$

As seen from equation (3.52) at large distances ($d \gg \sqrt{h_t h_r}$), the received power falls off with distance raised to the fourth power, or at a rate of 40 dB/decade. This is a much more rapid path loss than is experienced in free space. Note also that at large values of $d$, the received power and path loss become independent of frequency. The path loss for the 2-ray model (with antenna gains) can be expressed in dB as

$$PL(\text{dB}) = 40\log d - (10\log G_t + 10\log G_r + 20\log h_t + 20\log h_r) \tag{3.53}$$

At small distances, equation (3.39) must be used to compute the total E-field. When equation (3.42) is evaluated for $\theta_\Delta = \pi$, then $d = (4h_t h_r)/\lambda$ is where the ground appears in the first *Fresnel zone* between the transmitter and

receiver (Fresnel zones are treated in Section 3.7.1). The first Fresnel zone distance is a useful parameter in microcell path loss models [Feu94].

**Example 3.6**

A mobile is located 5 km away from a base station and uses a vertical $\lambda/4$ monopole antenna with a gain of 2.55 dB to receive cellular radio signals. The E-field at 1 km from the transmitter is measured to be $10^{-3}$ V/m. The carrier frequency used for this system is 900 MHz.

(a) Find the length and the gain of the receiving antenna.

(b) Find the received power at the mobile using the 2-ray ground reflection model assuming the height of the transmitting antenna is 50 m and the receiving antenna is 1.5 m above ground.

**Solution to Example 3.6**

Given:

T-R separation distance = 5 km

E-field at a distance of 1 km = $10^{-3}$ V/m

Frequency of operation, $f$ = 900 MHz

$$\lambda = \frac{c}{f} = \frac{3\times10^8}{900\times10^6} = 0.333 \text{ m.}$$

Length of the antenna, $L = \lambda/4 = 0.333/4 = 0.0833$ m = 8.33 cm.

Gain of $\lambda/4$ monopole antenna can be obtained using equation (3.2).

Gain of antenna =1.8 = 2.55 dB.

(b) Since $d \gg \sqrt{h_t h_r}$, the electric field is given by

$$E_R(d) \approx \frac{2E_0 d_0}{d}\frac{2\pi h_t h_r}{\lambda d} \approx \frac{k}{d^2} \text{ V/m}$$

$$= \frac{2\times10^{-3}\times1\times10^3}{5\times10^3}\left[\frac{2\pi(50)(1.5)}{0.333(5\times10^3)}\right]$$

$$= 113.1\times10^{-6} \text{ V/m.}$$

The received power at a distance $d$ can be obtained using equation (3.15)

$$P_r(d) = \frac{(113.1\times10^{-6})^2}{377}\left[\frac{1.8\,(0.333)^2}{4\pi}\right]$$

$$P_r(d = 5 \text{ km}) = 5.4\times10^{-13} \text{ W} = -122.68 \text{ dBW or } -92.68 \text{ dBm.}$$

## 3.7 Diffraction

Diffraction allows radio signals to propagate around the curved surface of the earth, beyond the horizon, and to propagate behind obstructions. Although the received field strength decreases rapidly as a receiver moves deeper into the

obstructed (shadowed) region, the diffraction field still exists and often has sufficient strength to produce a useful signal.

The phenomenon of diffraction can be explained by Huygen's principle, which states that all points on a wavefront can be considered as point sources for the production of secondary wavelets, and that these wavelets combine to produce a new wavefront in the direction of propagation. Diffraction is caused by the propagation of secondary wavelets into a shadowed region. The field strength of a diffracted wave in the shadowed region is the vector sum of the electric field components of all the secondary wavelets in the space around the obstacle.

### 3.7.1 Fresnel Zone Geometry

Consider a transmitter and receiver separated in free space as shown in Figure 3.10a. Let an obstructing screen of effective height $h$ with infinite width (going into and out of the paper) be placed between them at a distance $d_1$ from the transmitter and $d_2$ from the receiver. It is apparent that the wave propagating from the transmitter to the receiver via the top of the screen travels a longer distance than if a direct line-of-sight path (through the screen) existed. Assuming $h \ll d_1, d_2$ and $h \gg \lambda$, then the difference between the direct path and the diffracted path, called the *excess path length* $(\Delta)$, can be obtained from the geometry of Figure 3.10b as

$$\Delta \approx \frac{h^2}{2}\frac{(d_1 + d_2)}{d_1 d_2} \tag{3.54}$$

The corresponding phase difference is given by

$$\phi = \frac{2\pi\Delta}{\lambda} \approx \frac{2\pi}{\lambda}\frac{h^2}{2}\frac{(d_1 + d_2)}{d_1 d_2} \tag{3.55}$$

and when $\tan x \approx x$, then $\alpha = \beta + \gamma$ from Figure 3.10c and

$$\alpha \approx h\left(\frac{d_1 + d_2}{d_1 d_2}\right)$$

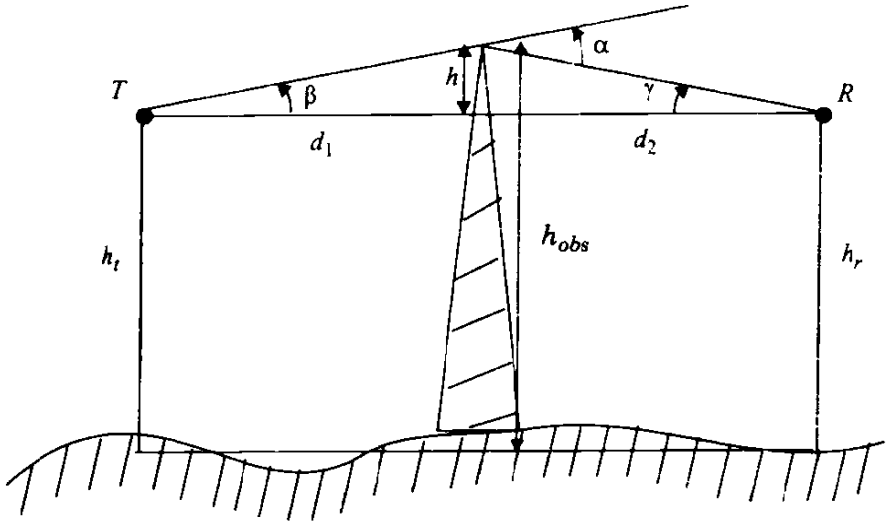
(a proof of equations (3.54) and (3.55) is left as an exercise for the reader).

Equation (3.55) is often normalized using the dimensionless *Fresnel-Kirchoff* diffraction parameter $v$ which is given by

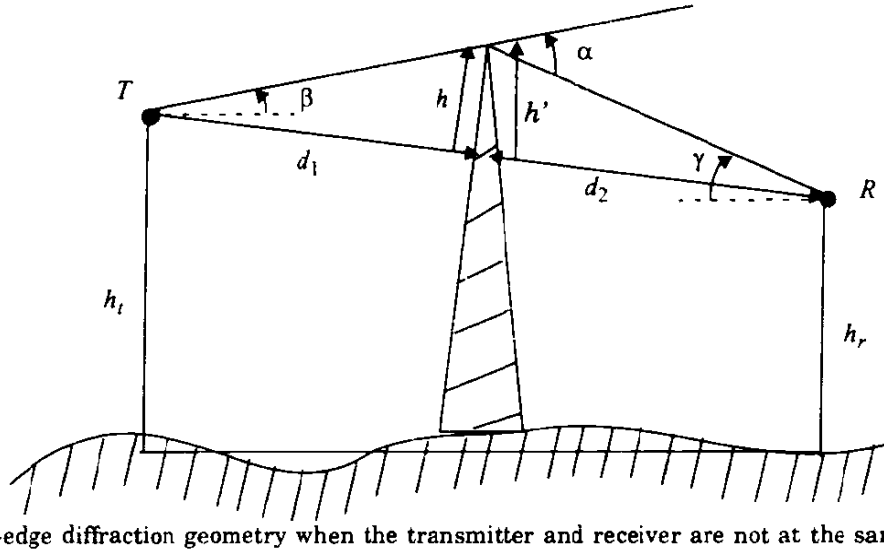
$$v = h\sqrt{\frac{2(d_1 + d_2)}{\lambda d_1 d_2}} = \alpha\sqrt{\frac{2d_1 d_2}{\lambda(d_1 + d_2)}} \tag{3.56}$$

where $\alpha$ has units of radians and is shown in Figure 3.10b and Figure 3.10c. From equation (3.56), $\phi$ can be expressed as

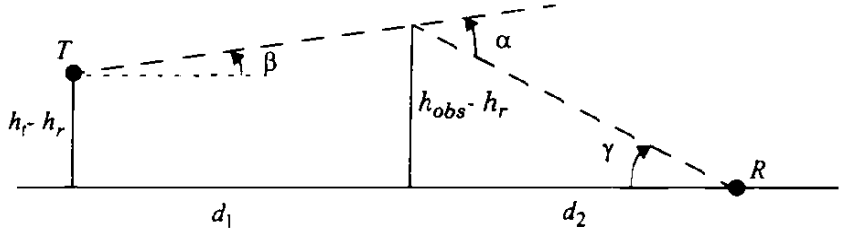
$$\phi = \frac{\pi}{2}v^2 \tag{3.57}$$




(a) Knife-edge diffraction geometry. The point $T$ denotes the transmitter and $R$ denotes the receiver, with an infinite knife-edge obstruction blocking the line-of-sight path.




(b) Knife-edge diffraction geometry when the transmitter and receiver are not at the same height. Note that if α and β are small and $h << d_1$ and $d_2$, then $h$ and $h'$ are virtually identical and the geometry may be redrawn as shown in Figure 3.10c.




(c) Equivalent knife-edge geometry where the smallest height (in this case $h_r$) is subtracted from all other heights.

Figure 3.10
Diagrams of knife-edge geometry.

From the above equations it is clear that the phase difference between a direct line-of-sight path and diffracted path is a function of height and position of the obstruction, as well as the transmitter and receiver location.

In practical diffraction problems, it is advantageous to reduce all heights by a constant, so that the geometry is simplified without changing the values of the angles. This procedure is shown in Figure 3.10c.

The concept of diffraction loss as a function of the path difference around an obstruction is explained by Fresnel zones. Fresnel zones represent successive regions where secondary waves have a path length from the transmitter to receiver which are $n\lambda/2$ greater than the total path length of a line-of-sight path. Figure 3.11 demonstrates a transparent plane located between a transmitter and receiver. The concentric circles on the plane represent the loci of the origins of secondary wavelets which propagate to the receiver such that the total path length increases by $\lambda/2$ for successive circles. These circles are called Fresnel zones. The successive Fresnel zones have the effect of alternately providing constructive and destructive interference to the total received signal. The radius of the $n$ th Fresnel zone circle is denoted by $r_n$ and can be expressed in terms of $n$, $\lambda$, $d_1$, and $d_2$ by

$$r_n = \sqrt{\frac{n\lambda d_1 d_2}{d_1 + d_2}} \tag{3.58}$$

This approximation is valid for $d_1, d_2 \gg r_n$.

The excess total path length traversed by a ray passing through each circle is $n\lambda/2$, where $n$ is an integer. Thus, the path traveling through the smallest circle corresponding to $n = 1$ in Figure 3.11 will have an excess path lengths of $\lambda/2$ as compared to a line-of-sight path, and circles corresponding to $n = 2, 3$, etc. will have an excess path length of $\lambda, 3\lambda/2$, etc. The radii of the concentric circles depend on the location of the plane. The Fresnel zones of Figure 3.11 will have maximum radii if the plane is midway between the transmitter and receiver, and the radii become smaller when the plane is moved towards either the transmitter or the receiver. This effect illustrates how shadowing is sensitive to the frequency as well as the location of obstructions with relation to the transmitter or receiver.

In mobile communication systems, diffraction loss occurs from the blockage of secondary waves such that only a portion of the energy is diffracted around an obstacle. That is, an obstruction causes a blockage of energy from some of the Fresnel zones, thus allowing only some of the transmitted energy to reach the receiver. Depending on the geometry of the obstruction, the received energy will be a vector sum of the energy contributions from all unobstructed Fresnel zones.

As shown in Figure 3.12, an obstacle may block the transmission path, and a family of ellipsoids can be constructed between a transmitter and receiver by joining all the points for which the excess path delay is an integer multiple of

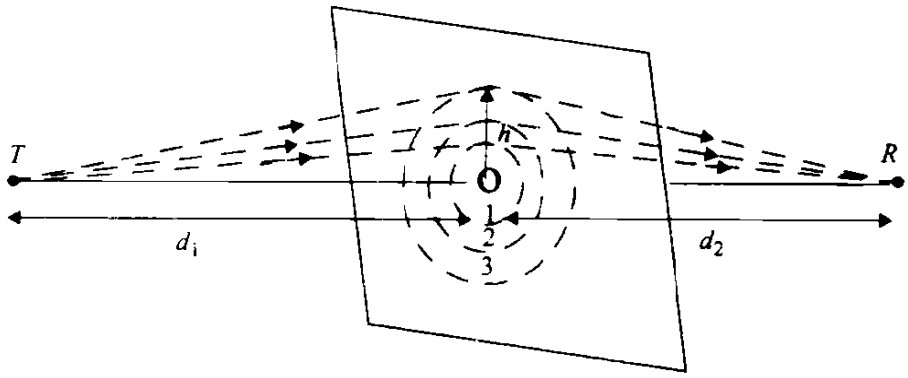



Figure 3.11
Concentric circles which define the boundaries of successive Fresnel zones.

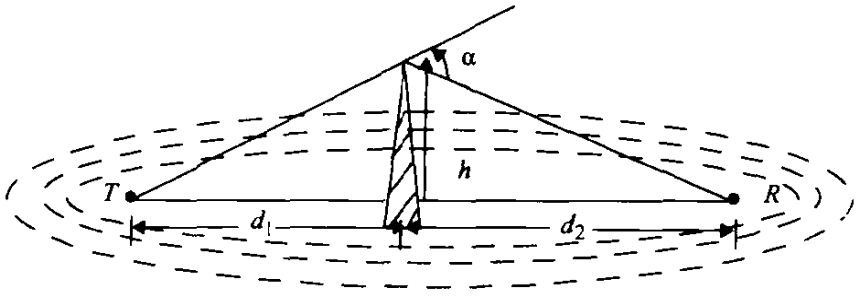
half wavelengths. The ellipsoids represent Fresnel zones. Note that the Fresnel zones are elliptical in shape with the transmitter and receiver antenna at their foci. In Figure 3.12, different knife edge diffraction scenarios are shown. In general, if an obstruction does not block the volume contained within the first Fresnel zone, then the diffraction loss will be minimal, and diffraction effects may be neglected. In fact, a rule of thumb used for design of line-of-sight microwave links is that as long as 55% of the first Fresnel zone is kept clear, then further Fresnel zone clearance does not significantly alter the diffraction loss.

### 3.7.2 Knife-edge Diffraction Model

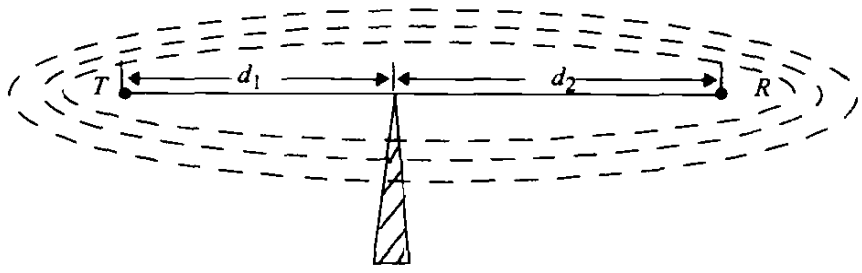
Estimating the signal attenuation caused by diffraction of radio waves over hills and buildings is essential in predicting the field strength in a given service area. Generally, it is impossible to make very precise estimates of the diffraction losses, and in practice prediction is a process of theoretical approximation modified by necessary empirical corrections. Though the calculation of diffraction losses over complex and irregular terrain is a mathematically difficult problem, expressions for diffraction losses for many simple cases have been derived. As a starting point, the limiting case of propagation over a knife-edge gives good insight into the order of magnitude of diffraction loss.

When shadowing is caused by a single object such as a hill or mountain, the attenuation caused by diffraction can be estimated by treating the obstruction as a diffracting knife edge. This is the simplest of diffraction models, and the diffraction loss in this case can be readily estimated using the classical Fresnel solution for the field behind a knife edge (also called a half-plane). Figure 3.13 illustrates this approach.

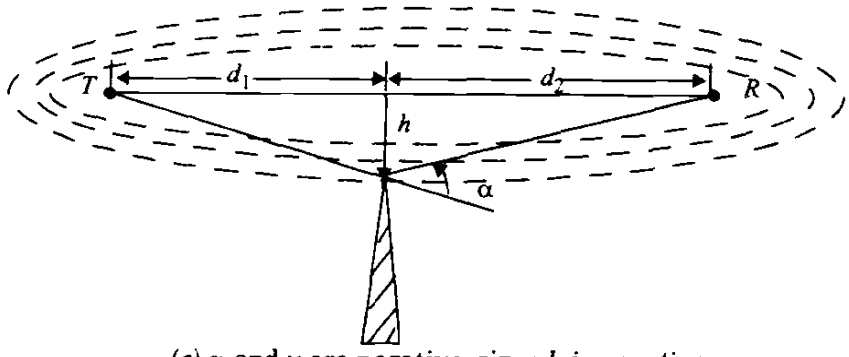
Consider a receiver at point $R$, located in the shadowed region (also called the *diffraction zone*). The field strength at point $R$ in Figure 3.13 is a vector sum of the fields due to all of the secondary Huygen's sources in the plane above the




(a) α and v are positive, since $h$ is positive




(b) α and v are equal to zero, since $h$ is equal to zero




(c) α and v are negative, since $h$ is negative

Figure 3.12
Illustration of Fresnel zones for different knife-edge diffraction scenarios.

knife edge. The electric field strength, $E_d$, of a knife-edge diffracted wave is given by

$$\frac{E_d}{E_o} = F(v) = \frac{(1+j)}{2}\int_v^{\infty} \exp\left(\left(-j\pi t^2\right)/2\right) dt \tag{3.59}$$

where $E_o$ is the free space field strength in the absence of both the ground and the knife edge, and $F(v)$ is the complex Fresnel integral. The Fresnel integral, $F(v)$, is a function of the Fresnel-Kirchoff diffraction parameter $v$, defined in equation (3.56), and is commonly evaluated using tables or graphs for given val-

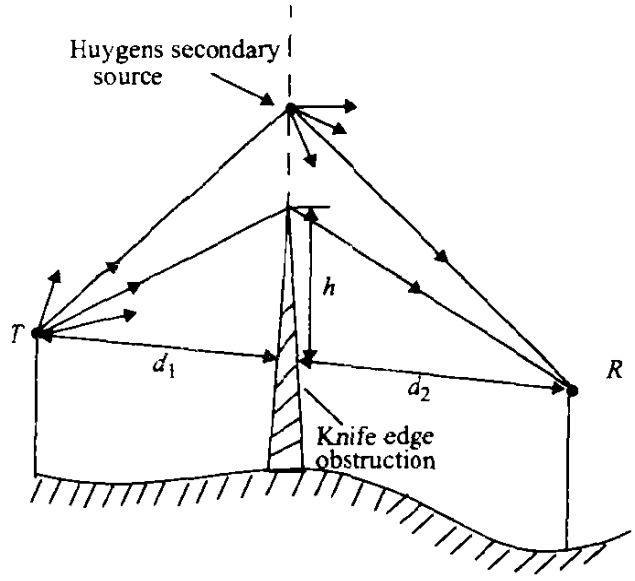



Figure 3.13
Illustration of knife-edge diffraction geometry. The receiver $R$ is located in the shadow region.

ues of $v$. The diffraction gain due to the presence of a knife edge, as compared to the free space E-field, is given by

$$G_d(\text{dB}) = 20\log|F(v)| \tag{3.60}$$

In practice, graphical or numerical solutions are relied upon to compute diffraction gain. A graphical representation of $G_d(\text{dB})$ as a function of $v$ is given in Figure 3.14. An approximate solution for equation (3.60) provided by Lee [Lee85] as

$$G_d(dB) = 0 \qquad v \le -1 \tag{3.61.a}$$

$$G_d(\text{dB}) = 20\log(0.5 - 0.62v) \qquad -1 \le v \le 0 \tag{3.61.b}$$

$$G_d(\text{dB}) = 20\log(0.5\exp(-0.95v)) \qquad 0 \le v \le 1 \tag{3.61.c}$$

$$G_d(\text{dB}) = 20\log\left(0.4 - \sqrt{0.1184 - (0.38 - 0.1v)^2}\right) \qquad 1 \le v \le 2.4 \tag{3.61.d}$$

$$G_d(\text{dB}) = 20\log\left(\frac{0.225}{v}\right) \qquad v > 2.4 \tag{3.61.e}$$

**Example 3.7**

Compute the diffraction loss for the three cases shown in Figure 3.12. Assume $\lambda = 1/3$ m, $d_1 = 1$ km, $d_2 = 1$ km, and (a) $h = 25$ m, (b) $h = 0$ (c) $h = -25$ m. Compare your answers using values from Figure 3.14, as well as the approximate solution given by equation (3.61.a) — (3.61.e). For each of

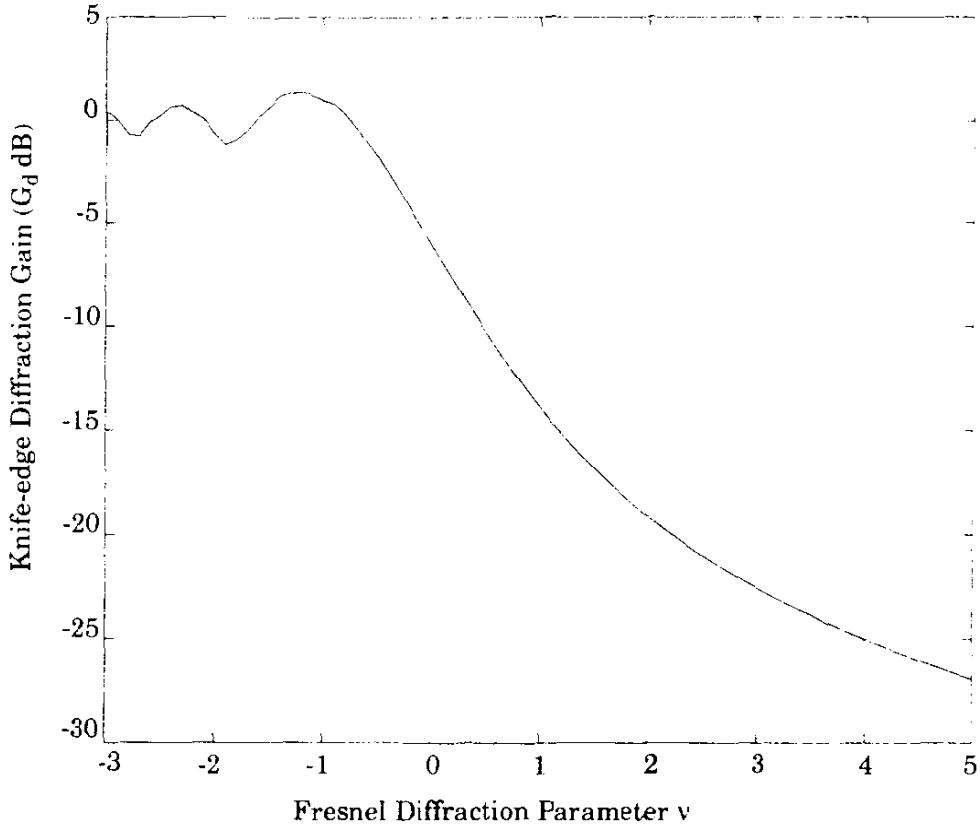



Figure 3.14
Knife-edge diffraction gain as a function of Fresnel diffraction parameter v

these cases, identify the Fresnel zone within which the tip of the obstruction lies.

## Solution to Example 3.7

Given:

$\lambda = 1/3$ m

$d_1 = 1$ km

$d_2 = 1$ km

(a) $h = 25$ m.

Using equation (3.56), the Fresnel diffraction parameter is obtained as

$$v = h\sqrt{\frac{2(d_1 + d_2)}{\lambda d_1 d_2}} = 25\sqrt{\frac{2(1000 + 1000)}{(1/3) \times 1000 \times 1000}} = 2.74.$$

From Figure 3.14, the diffraction loss is obtained as 22 dB.

Using the numerical approximation in equation (3.61.e), the diffraction loss is equal to 21.7 dB.

The path length difference between the direct and diffracted rays is given by equation (3.54) as

$$\Delta \approx \frac{h^2}{2}\frac{(d_1+d_2)}{d_1 d_2} = \frac{25^2}{2}\frac{(1000+1000)}{1000\times 1000} = 0.625 \text{ m}.$$

To find the Fresnel zone in which the tip of the obstruction lies we need to compute $n$ which satisfies the relation $\Delta = n\lambda/2$. For $\lambda = 1/3$ m, and $\Delta = 0.625$ m, we obtain

$$n = \frac{2\Delta}{\lambda} = \frac{2\times 0.625}{0.3333} = 3.75.$$

Therefore, the tip of the obstruction completely blocks the first three Fresnel zones.

(b) $h = 0$

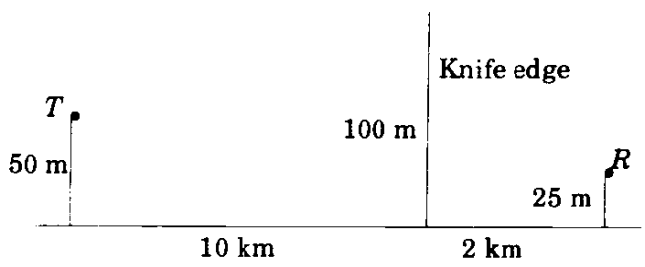
Therefore, the Fresnel diffraction parameter $v = 0$.
From Figure 3.14, the diffraction loss is obtained as 6 dB.
Using the numerical approximation in equation (3.61.b), the diffraction loss is equal to 6 dB.
For this case, since $h = 0$, we have $\Delta = 0$, and the tip of the obstruction lies in the middle of the first Fresnel zone.

(c) $h = -25$

Using equation (3.56),
the Fresnel diffraction parameter is obtained as -2.74.
From Figure 3.14, the diffraction loss is as approximately equal to 1 dB.
Using the numerical approximation in equation (3.61.a), the diffraction loss is equal to 0 dB.
Since the absolute value of the height $h$, is the same as part (a), the excess path length $\Delta$ and hence $n$ will also be the same. It should be noted that although the tip of the obstruction completely blocks the first three Fresnel zones, the diffraction losses are negligible, since the obstruction is below the line of sight ($h$ is negative).

## Example 3.8

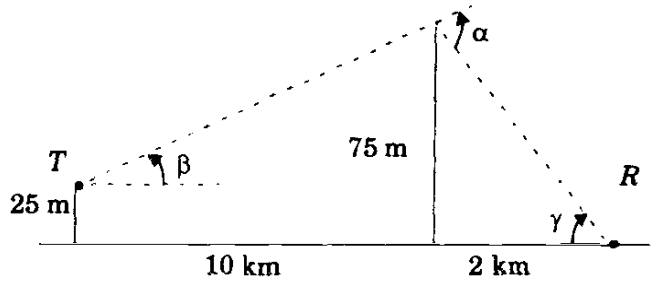
Given the following geometry, determine (a) the loss due to knife-edge diffraction, and (b) the height of the obstacle required to induce 6 dB diffraction loss. Assume $f = 900$ MHz.




## Solution to Example 3.8

(a) The wavelength $\lambda = \frac{c}{f} = \frac{3\times 10^8}{900\times 10^6} = \frac{1}{3}$ m.

Redraw the geometry by subtracting the height of the smallest structure.




$$\beta = \tan^{-1}\left(\frac{75-25}{10000}\right) = 0.2865^o$$

$$\gamma = \tan^{-1}\left(\frac{75}{2000}\right) = 2.15^o$$

and

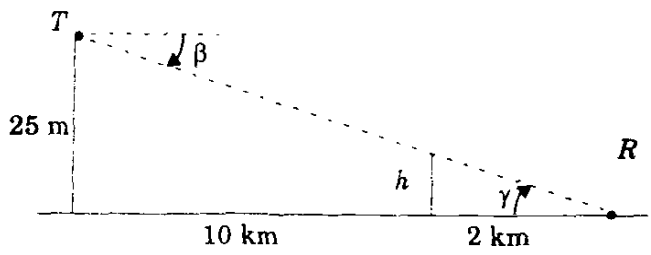
$$\alpha = \beta + \gamma = 2.434^o = 0.0424 \text{ rad}$$

Then using equation (3.56)

$$v = 0.0424\sqrt{\frac{2 \times 10000 \times 2000}{(1/3) \times (10000 + 2000)}} = 4.24.$$

From Figure 3.14 or (3.61.e), the diffraction loss is 25.5 dB.

(b) For 6 dB diffraction loss, $v = 0$. The obstruction height $h$ may be found using similar triangles ($\beta = -\gamma$) as shown below.




It follows that $\frac{h}{2000} = \frac{25}{12000}$, thus $h = 4.16$ m.

### 3.7.3 Multiple Knife-edge Diffraction

In many practical situations, especially in hilly terrain, the propagation path may consist of more than one obstruction, in which case the total diffraction loss due to all of the obstacles must be computed. Bullington [Bul47] suggested that the series of obstacles be replaced by a single equivalent obstacle so that the path loss can be obtained using single knife-edge diffraction models. This method, illustrated in Figure 3.15, oversimplifies the calculations and often provides very optimistic estimates of the received signal strength. In a more rigorous treatment, Millington et. al., [Mil62] gave a wave-theory solution for the field behind two knife edges in series. This solution is very useful and can be applied

easily for predicting diffraction losses due to two knife edges. However, extending this to more than two knife edges becomes a formidable mathematical problem. Many models that are mathematically less complicated have been developed to estimate the diffraction losses due to multiple obstructions [Eps53], [Dey66].

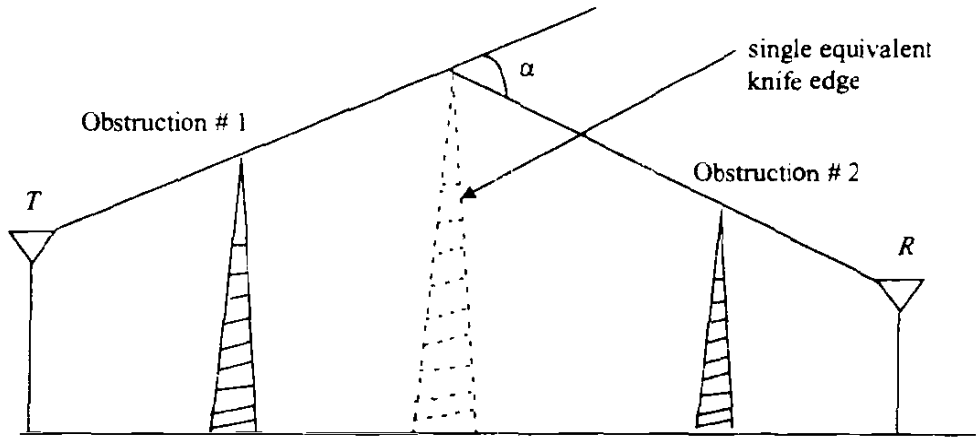



Figure 3.15
Bullington's construction of an equivalent knife edge [From [Bul47] © IEEE].

## 3.8 Scattering

The actual received signal in a mobile radio environment is often stronger than what is predicted by reflection and diffraction models alone. This is because when a radio wave impinges on a rough surface, the reflected energy is spread out (diffused) in all directions due to scattering. Objects such as lamp posts and trees tend to scatter energy in all directions, thereby providing additional radio energy at a receiver.

Flat surfaces that have much larger dimension than a wavelength may be modeled as reflective surfaces. However, the roughness of such surfaces often induces propagation effects different from the specular reflection described earlier in this chapter. Surface roughness is often tested using the Rayleigh criterion which defines a critical height $(h_c)$ of surface protuberances for a given angle of incidence $\theta_i$, given by

$$h_c = \frac{\lambda}{8 \sin \theta_i} \tag{3.62}$$

A surface is considered smooth if its minimum to maximum protuberance $h$ is less than $h_c$, and is considered rough if the protuberance is greater than $h_c$. For rough surfaces, the flat surface reflection coefficient needs to be multiplied by a scattering loss factor, $\rho_S$, to account for the diminished reflected field. Ament [Ame53] assumed that the surface height $h$ is a Gaussian distributed random variable with a local mean and found $\rho_S$ to be given by

$$\rho_S = \exp\left[-8\left(\frac{\pi\sigma_h \sin\theta_i}{\lambda}\right)^2\right] \tag{3.63}$$

where $\sigma_h$ is the standard deviation of the surface height about the mean surface height. The scattering loss factor derived by Ament was modified by Boithias [Boi87] to give better agreement with measured results, and is given in (3.63)

$$\rho_S = \exp\left[-8\left(\frac{\pi\sigma_h \sin\theta_i}{\lambda}\right)^2\right] I_0\left[8\left(\frac{\pi\sigma_h \sin\theta_i}{\lambda}\right)^2\right] \tag{3.64}$$

where $I_0$ is the Bessel function of the first kind and zero order.

The reflected E-fields for $h > h_c$ can be solved for rough surfaces using a modified reflection coefficient given as

$$\Gamma_{rough} = \rho_S \Gamma \tag{3.65}$$

Figure 3.16a and Figure 3.16b illustrate experimental results found by Landron et al [Lan96]. Measured reflection coefficient data is shown to agree well with the modified reflection coefficients of equations (3.64) and (3.65) for large exterior walls made of rough limestone.

### 3.8.1 Radar Cross Section Model

In radio channels where large, distant objects induce scattering, knowledge of the physical location of such objects can be used to accurately predict scattered signal strengths. The *radar cross section* (RCS) of a scattering object is defined as the ratio of the power density of the signal scattered in the direction of the receiver to the power density of the radio wave incident upon the scattering object, and has units of square meters. Analysis based on the geometric theory of diffraction and physical optics may be used to determine the scattered field strength.

For urban mobile radio systems, models based on the *bistatic radar equation* may be used to compute the received power due to scattering in the far field. The bistatic radar equation describes the propagation of a wave traveling in free space which impinges on a distant scattering object, and is then reradiated in the direction of the receiver, given by

$$\begin{aligned} P_R(\text{dBm}) = {} & P_T(\text{dBm}) + G_T(\text{dBi}) + 20\log(\lambda) + RCS[\text{dB m}^2] \\ & - 30\log(4\pi) - 20\log d_T - 20\log d_R \end{aligned} \tag{3.66}$$

where $d_T$ and $d_R$ are the distance from the scattering object to the transmitter and receiver, respectively. In equation (3.66), the scattering object is assumed to be in the far field (Fraunhofer region) of both the transmitter and receiver. The variable $RCS$ is given in units of $dB \cdot m^2$, and can be approximated by the surface area (in square meters) of the scattering object, measured in dB with respect to a one square meter reference [Sei91]. Equation (3.66) may be applied to scatterers in the far-field of both the transmitter and receiver (as illustrated in

[Van87], [Zog87], [Sei91]) and is useful for predicting receiver power which scatters off large objects, such as buildings, which are for both the transmitter and receiver.

Several European cities were measured from the perimeter [Sei91], and *RCS* values for several buildings were determined from measured power delay profiles. For medium and large size buildings located 5 - 10 km away, *RCS* values were found to be in the range of $14.1 dB \cdot m^2$ to $55.7\ dB \cdot m^2$.

## 3.9 Practical Link Budget Design using Path Loss Models

Most radio propagation models are derived using a combination of analytical and empirical methods. The empirical approach is based on fitting curves or analytical expressions that recreate a set of measured data. This has the advantage of implicitly taking into account all propagation factors, both known and unknown, through actual field measurements. However, the validity of an empirical model at transmission frequencies or environments other than those used to derive the model can only be established by additional measured data in the new environment at the required transmission frequency. Over time, some classical propagation models have emerged, which are now used to predict large-scale coverage for mobile communication systems design. By using path loss models to estimate the received signal level as a function of distance, it becomes possible to predict the SNR for a mobile communication system. Using noise analysis techniques given in Appendix B, the noise floor can be determined. For example, the 2-ray model described in section 3.6 was used to estimate capacity in a spread spectrum cellular system, before such systems were deployed [Rap92b]. Practical path loss estimation techniques are now presented.

### 3.9.1 Log-distance Path Loss Model

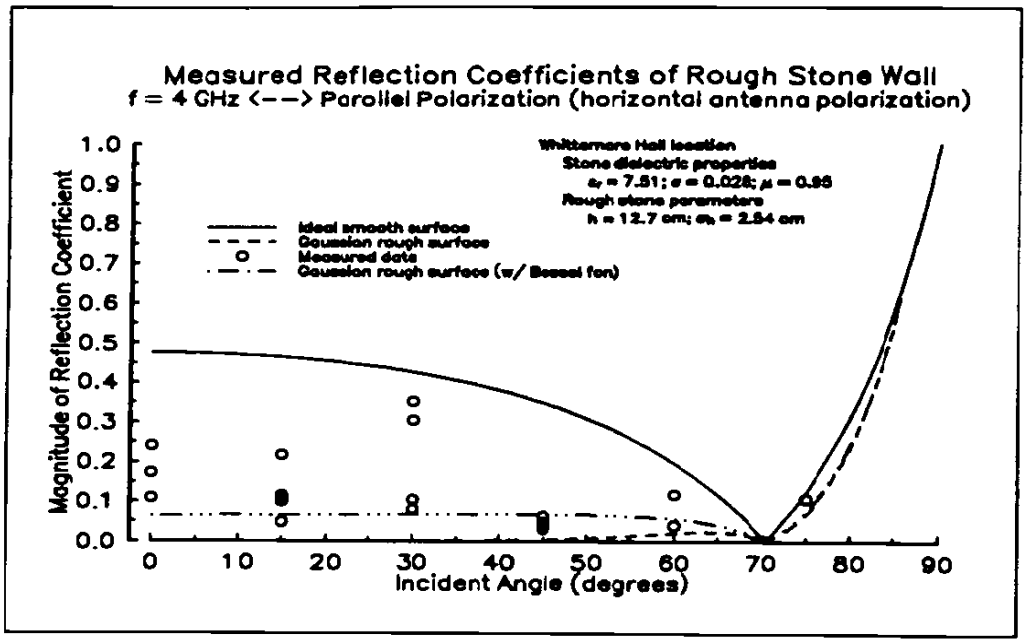
Both theoretical and measurement-based propagation models indicate that average received signal power decreases logarithmically with distance, whether in outdoor or indoor radio channels. Such models have been used extensively in the literature. The average large-scale path loss for an arbitrary T-R separation is expressed as a function of distance by using a path loss exponent, $n$.

$$PL(d) \propto \left(\frac{d}{d_0}\right)^n \tag{3.67}$$

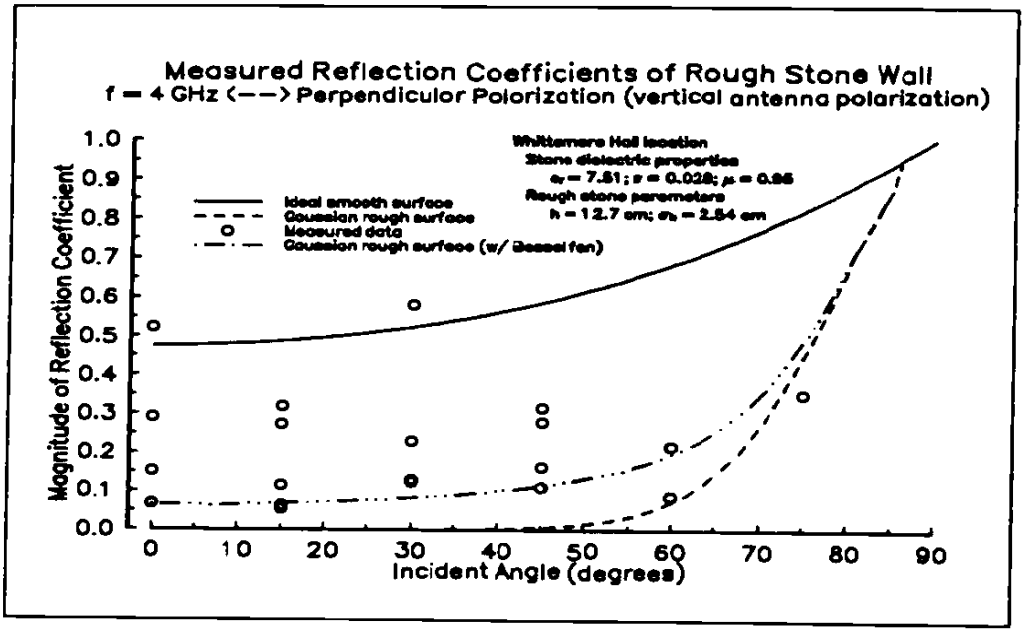
or

$$PL(\text{dB}) = PL(d_0) + 10n\log\left(\frac{d}{d_0}\right) \tag{3.68}$$

where $n$ is the path loss exponent which indicates the rate at which the path loss increases with distance, $d_0$ is the close-in reference distance which is determined from measurements close to the transmitter, and $d$ is the T-R separation




(a) E-field in the plane of incidence. (parallel polarization)




(b) E-field normal to plane of incidence. (perpendicular polarization)

Figure 3.16
Measured reflection coefficients versus incident angle at a rough stone wall site. In these graphs, incident angle is measured with respect to the normal, instead of with respect to the surface boundary as defined in Figure 3.4. These graphs agree with Figure 3.6 [Lan96].

distance. The bars in equations (3.67) and (3.68) denote the ensemble average of all possible path loss values for a given value of $d$. When plotted on a log-log scale, the modeled path loss is a straight line with a slope equal to $10n$ dB per decade. The value of $n$ depends on the specific propagation environment. For example, in free space, $n$ is equal to 2, and when obstructions are present, $n$ will have a larger value.

It is important to select a free space reference distance that is appropriate for the propagation environment. In large coverage cellular systems, 1 km reference distances are commonly used [Lee85], whereas in microcellular systems, much smaller distances (such as 100 m or 1 m) are used. The reference distance should always be in the far field of the antenna so that near-field effects do not alter the reference path loss. The reference path loss is calculated using the free space path loss formula given by equation (3.5) or through field measurements at distance $d_0$. Table 3.2 lists typical path loss exponents obtained in various mobile radio environments.

**Table 3.2 Path Loss Exponents for Different Environments**

| Environment | Path Loss Exponent, n |
|---|---|
| Free space | 2 |
| Urban area cellular radio | 2.7 to 3.5 |
| Shadowed urban cellular radio | 3 to 5 |
| In building line-of-sight | 1.6 to 1.8 |
| Obstructed in building | 4 to 6 |
| Obstructed in factories | 2 to 3 |

### 3.9.2 Log-normal Shadowing

The model in equation (3.68) does not consider the fact that the surrounding environmental clutter may be vastly different at two different locations having the same T-R separation. This leads to measured signals which are vastly different than the *average* value predicted by equation (3.68). Measurements have shown that at any value of $d$, the path loss $PL(d)$ at a particular location is random and distributed log-normally (normal in dB) about the mean distance-dependent value [Cox84], [Ber87]. That is

$$PL(d)[dB] = \overline{PL}(d) + X_\sigma = \overline{PL}(d_0) + 10n\log\left(\frac{d}{d_0}\right) + X_\sigma \qquad (3.69.a)$$

and

$$P_r(d)[dBm] = P_t[dBm] - PL(d)[dB] \text{ (antenna gains included in } PL(d)) \qquad (3.69.b)$$

where $X_\sigma$ is a zero-mean Gaussian distributed random variable (in dB) with standard deviation $\sigma$ (also in dB).

The log-normal distribution describes the random *shadowing* effects which occur over a large number of measurement locations which have the same T-R separation, but have different levels of clutter on the propagation path. This phenomenon is referred to as *log-normal shadowing*. Simply put, log-normal shadowing implies that measured signal levels at a specific T-R separation have a Gaussian (normal) distribution about the distance-dependent mean of (3.68), where the measured signal levels have values in dB units. The standard deviation of the Gaussian distribution that describes the shadowing also has units in dB. Thus, the random effects of shadowing are accounted for using the Gaussian distribution which lends itself readily to evaluation (see Appendix D).

The close-in reference distance $d_0$, the path loss exponent $n$, and the standard deviation $\sigma$, statistically describe the path loss model for an arbitrary location having a specific T-R separation, and this model may be used in computer simulation to provide received power levels for random locations in communication system design and analysis.

In practice, the values of $n$ and $\sigma$ are computed from measured data, using linear regression such that the difference between the measured and estimated path losses is minimized in a mean square error sense over a wide range of measurement locations and T-R separations. The value of $\overline{PL}(d_0)$ in (3.69.a) is based on either close-in measurements or on a free space assumption from the transmitter to $d_0$. An example of how the path loss exponent is determined from measured data follows. Figure 3.17 illustrates actual measured data in several cellular radio systems and demonstrates the random variations about the mean path loss (in dB) due to shadowing at specific T-R separations.

Since $PL(d)$ is a random variable with a normal distribution in dB about the distance-dependent mean, so is $P_r(d)$, and the $Q$-function or error function (*erf*) may be used to determine the probability that the received signal level will exceed (or fall below) a particular level. The $Q$-function is defined as

$$Q(z) = \frac{1}{\sqrt{2\pi}}\int_z^{\infty} \exp\left(-\frac{x^2}{2}\right)dx = \frac{1}{2}\left[1 - erf\left(\frac{z}{\sqrt{2}}\right)\right] \tag{3.70.a}$$

where

$$Q(z) = 1 - Q(-z) \tag{3.70.b}$$

The probability that the received signal level will exceed a certain value $\gamma$ can be calculated from the cumulative density function as

$$Pr[P_r(d) > \gamma] = Q\left(\frac{\gamma - \overline{P_r(d)}}{\sigma}\right) \tag{3.71}$$

Similarly, the probability that the received signal level will be below $\gamma$ is given by

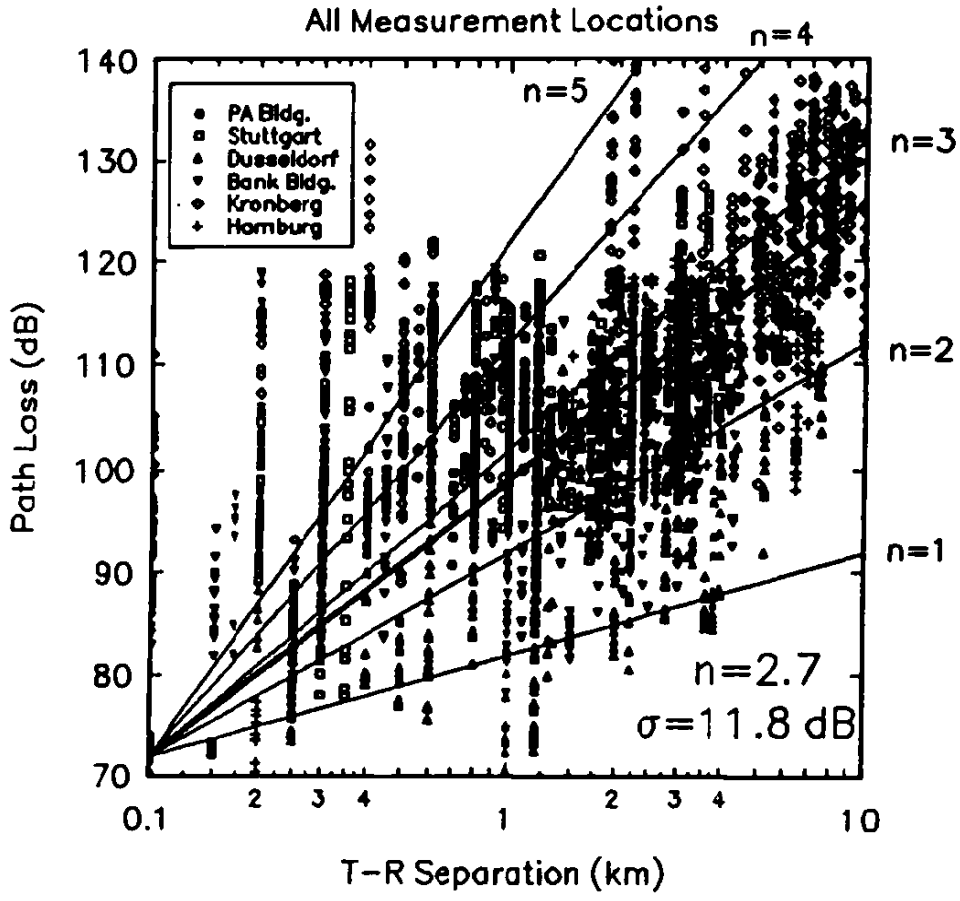



Figure 3.17
Scatter plot of measured data and corresponding MMSE path loss model for many cities in Germany. For this data, $n = 2.7$ and $\sigma = 11.8$ dB [From [Sei91] © IEEE].

$$Pr[P_r(d) < \gamma] = Q\left(\frac{\overline{P_r(d)} - \gamma}{\sigma}\right) \tag{3.72}$$

Appendix D provides tables for evaluating the $Q$ and *erf* functions.

### 3.9.3 Determination of Percentage of Coverage Area

It is clear that due to random effects of shadowing, some locations within a coverage area will be below a particular desired received signal threshold. It is often useful to compute how the boundary coverage relates to the percent of area covered within the boundary. For a circular coverage area having radius $R$ from a base station, let there be some desired received signal threshold $\gamma$. We are interested in computing $U(\gamma)$, the percentage of useful service area (i.e. the percentage of area with a received signal that is equal or greater than $\gamma$), given a

known likelihood of coverage at the cell boundary. Letting $d = r$ represent the radial distance from the transmitter, it can be shown that if $Pr[P_r(r) > \gamma]$ is the probability that the random received signal at $d = r$ exceeds the threshold $\gamma$ within an incremental area $dA$, then $U(\gamma)$ can be found by [Jak74]

$$U(\gamma) = \frac{1}{\pi R^2}\int Pr[P_r(r) > \gamma]\,dA = \frac{1}{\pi R^2}\int_0^{2\pi}\int_0^{R} Pr[P_r(r) > \gamma]\,r\,dr d\theta \tag{3.73}$$

Using (3.71), $Pr[P_r(r) > \gamma]$ is given by

$$\begin{aligned} Pr[P_r(r) > \gamma] &= Q\left(\frac{\gamma - \overline{P_r(r)}}{\sigma}\right) = \frac{1}{2} - \frac{1}{2}erf\left(\frac{\gamma - \overline{P_r(r)}}{\sigma\sqrt{2}}\right) \\ &= \frac{1}{2} - \frac{1}{2}erf\left(\frac{\gamma - [P_t - (\overline{PL}(d_0) + 10n\log(r/d_0))]}{\sigma\sqrt{2}}\right) \end{aligned} \tag{3.74}$$

In order to determine the path loss as referenced to the cell boundary ($r = R$), it is clear that

$$\overline{PL}(r) = 10n\log\left(\frac{R}{d_0}\right) + 10n\log\left(\frac{r}{R}\right) + \overline{PL}(d_0) \tag{3.75}$$

and equation (3.74) may be expressed as

$$\begin{aligned} &Pr[P_r(r) > \gamma] \\ &= \frac{1}{2} - \frac{1}{2}erf\left(\frac{\gamma - [P_t - (\overline{PL}(d_0) + 10n\log(R/d_0) + 10n\log(r/R))]}{\sigma\sqrt{2}}\right) \end{aligned} \tag{3.76}$$

If we let $a = (\gamma - P_t + \overline{PL}(d_0) + 10n\log(R/d_0))/\sigma\sqrt{2}$ and $b = (10n\log e)/\sigma\sqrt{2}$, then

$$U(\gamma) = \frac{1}{2} - \frac{1}{R^2}\int_0^{R} r\,erf\left(a + b\ln\frac{r}{R}\right)dr \tag{3.77}$$

By substituting $t = a + b\log(r/R)$ in equation (3.77), it can be shown that

$$U(\gamma) = \frac{1}{2}\left(1 - erf(a) + \exp\left(\frac{1-2ab}{b^2}\right)\left[1 - erf\left(\frac{1-ab}{b}\right)\right]\right) \tag{3.78}$$

By choosing the signal level such that $\overline{P_r}(R) = \gamma$ (i.e. $a = 0$), $U(\gamma)$ can be shown to be

$$U(\gamma) = \frac{1}{2}\left[1 + \exp\left(\frac{1}{b^2}\right)\left(1 - erf\left(\frac{1}{b}\right)\right)\right] \tag{3.79}$$

Equation (3.78) may be evaluated for a large number of values of $\sigma$ and $n$, as shown in Figure 3.18 [Reu74]. For example, if $n = 4$ and $\sigma = 8$ dB, and if the

boundary is to have 75% boundary coverage (75% of the time the signal is to exceed the threshold at the boundary), then the area coverage is equal to 94%. If $n = 2$ and $\sigma = 8$ dB, a 75% boundary coverage provides 91% area coverage. If $n = 3$ and $\sigma = 9$ dB, then 50% boundary coverage provides 71% area coverage.

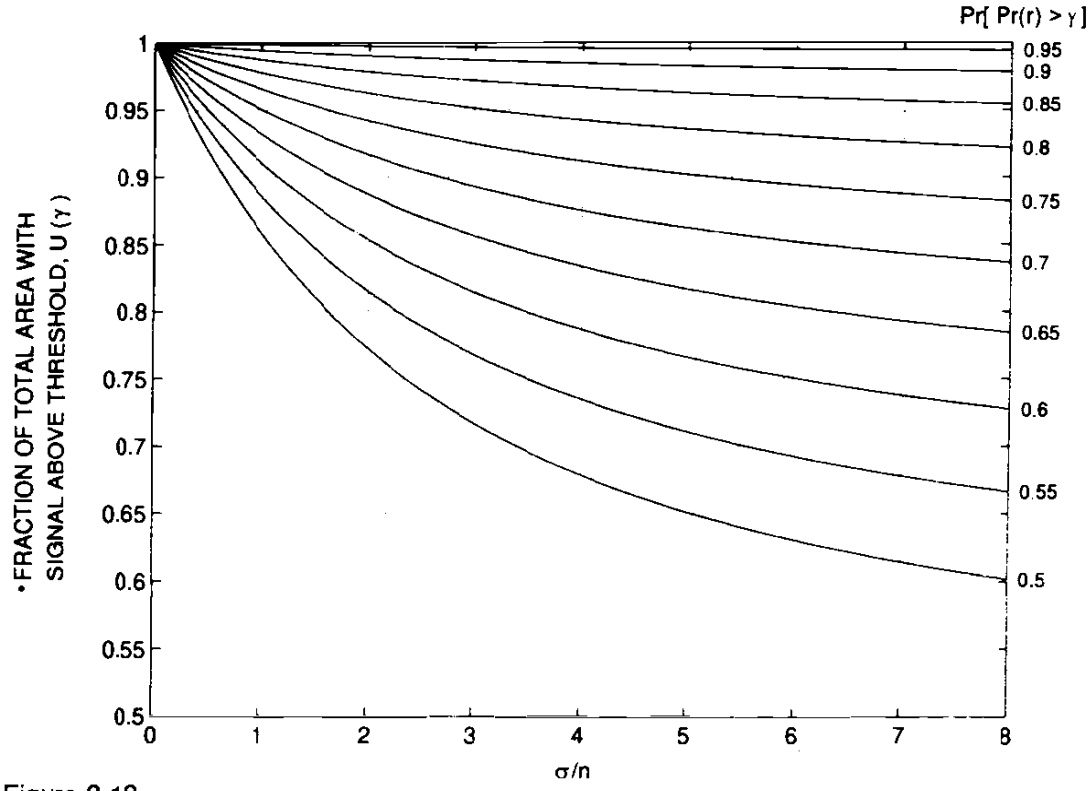



Figure 3.18
Family of curves relating fraction of total area with signal above threshold, $U(\gamma)$ as a function of probability of signal above threshold on the cell boundary.

**Example 3.9**

Four received power measurements were taken at distances of 100 m, 200 m, 1 km, and 3 km from a transmitter. These measured values are given in the following table. It is assumed that the path loss for these measurements follows the model in equation (3.69.a), where $d_0 = 100$ m: (a) find the minimum mean square error (MMSE) estimate for the path loss exponent, $n$; (b) calculate the standard deviation about the mean value; (c) estimate the received power at $d = 2$ km using the resulting model; (d) predict the likelihood that the received signal level at 2 km will be greater than -60 dBm; and (e) predict the percentage of area within a 2 km radius cell that receives signals greater than –60 dBm, given the result in (d).

| Distance from Transmitter | Received Power |
|---|---|
| 100 m | 0 dBm |
| 200 m | -20 dBm |
| 1000 m | -35 dBm |
| 3000 m | -70 dBm |

## Solution to Example 3.9

The MMSE estimate may be found using the following method. Let $p_i$ be the received power at a distance $d_i$ and let $\hat{p}_i$ be the estimate for $p_i$ using the $(d/d_0)^n$ path loss model of equation (3.67). The sum of squared errors between the measured and estimated values is given by

$$J(n) = \sum_{i=1}^{k} (p_i - \hat{p}_i)^2$$

The value of $n$ which minimizes the mean square error can be obtained by equating the derivative of $J(n)$ to zero, and then solving for $n$.

(a) Using equation (3.68), we find $\hat{p}_i = p_i(d_0) - 10n\log(d_i/100 \text{ m})$. Recognizing that $P(d_0) = 0$ dBm, we find the following estimates for $\hat{p}_i$ in dBm:

$$\hat{p}_1 = 0,\ \hat{p}_2 = -3n,\ \hat{p}_3 = -10n,\ \hat{p}_4 = -14.77n.$$

The sum of squared errors is then given by

$$\begin{aligned} J(n) &= (0-0)^2 + (-20-(-3n))^2 + (-35-(-10n))^2 \\ &\quad + (-70-(-14.77n))^2 \\ &= 6525 - 2887.8n + 327.153n^2 \end{aligned}$$

$$\frac{dJ(n)}{dn} = 654.306n - 2887.8.$$

Setting this equal to zero, the value of $n$ is obtained as $n = 4.4$.

(b) The sample variance $\sigma^2 = J(n)/4$ at $n = 4.4$ can be obtained as follows.

$$\begin{aligned} J(n) &= (0+0) + (-20+13.2)^2 + (-35+44)^2 + (-70+64.988)^2 \\ &= 152.36. \end{aligned}$$

$$\sigma^2 = 152.36/4 = 38.09$$

therefore

$\sigma = 6.17$ dB, which is a biased estimate. In general, a greater number of measurements are needed to reduce $\sigma^2$.

(c) The estimate of the received power at $d = 2$ km is given by

$$\hat{p}(d = 2 \text{ km}) = 0 - 10(4.4)\log(2000/100) = -57.24 \text{ dBm}.$$

A Gaussian random variable having zero mean and $\sigma = 6.17$ could be added to this value to simulate random shadowing effects at $d = 2$ km.

(d) The probability that the received signal level will be greater than –60 dBm is given by

$$Pr\,[P_r(d) > -60 \text{ dBm}] = Q\left(\frac{\gamma - \overline{P_r(d)}}{\sigma}\right) = Q\left(\frac{-60 + 57.24}{6.17}\right) = 67.4\ \%.$$

(e) If 67.4% of the users on the boundary receive signals greater than -60 dBm, then equation (3.78) or Figure 3.18 may be used to determine that 92% of the cell area receives coverage above -60 dBm.

## 3.10 Outdoor Propagation Models

Radio transmission in a mobile communications system often takes place over irregular terrain. The terrain profile of a particular area needs to be taken into account for estimating the path loss. The terrain profile may vary from a simple curved earth profile to a highly mountainous profile. The presence of trees, buildings, and other obstacles also must be taken into account. A number of propagation models are available to predict path loss over irregular terrain. While all these models aim to predict signal strength at a particular receiving point or in a specific local area (called a *sector*), the methods vary widely in their approach, complexity, and accuracy. Most of these models are based on a systematic interpretation of measurement data obtained in the service area. Some of the commonly used outdoor propagation models are now discussed.

### 3.10.1 Longley-Rice Model

The Longley-Rice model [Ric67], [Lon68] is applicable to point-to-point communication systems in the frequency range from 40 MHz to 100 GHz, over different kinds of terrain. The median transmission loss is predicted using the path geometry of the terrain profile and the refractivity of the troposphere. Geometric optics techniques (primarily the 2-ray ground reflection model) are used to predict signal strengths within the radio horizon. Diffraction losses over isolated obstacles are estimated using the Fresnel-Kirchoff knife-edge models. Forward scatter theory is used to make troposcatter predictions over long distances, and far field diffraction losses in double horizon paths are predicted using a modified Van der Pol-Bremmer method. The Longley-Rice propagation prediction model is also referred to as the *ITS irregular terrain model*.

The Longley-Rice model is also available as a computer program [Lon78] to calculate large-scale median transmission loss relative to free space loss over irregular terrain for frequencies between 20 MHz and 10 GHz. For a given transmission path, the program takes as its input the transmission frequency, path length, polarization, antenna heights, surface refractivity, effective radius of earth, ground conductivity, ground dielectric constant, and climate. The program also operates on path-specific parameters such as horizon distance of the anten-

nas, horizon elevation angle, angular trans-horizon distance, terrain irregularity and other specific inputs.

The Longley-Rice method operates in two modes. When a detailed terrain path profile is available, the path-specific parameters can be easily determined and the prediction is called a *point-to-point mode* prediction. On the other hand, if the terrain path profile is not available, the Longley-Rice method provides techniques to estimate the path-specific parameters, and such a prediction is called an *area mode* prediction.

There have been many modifications and corrections to the Longley-Rice model since its original publication. One important modification [Lon78] deals with radio propagation in urban areas, and this is particularly relevant to mobile radio. This modification introduces an excess term as an allowance for the additional attenuation due to urban clutter near the receiving antenna. This extra term, called the *urban factor* (UF), has been derived by comparing the predictions by the original Longley-Rice model with those obtained by Okumura [Oku68].

One shortcoming of the Longley-Rice model is that it does not provide a way of determining corrections due to environmental factors in the immediate vicinity of the mobile receiver, or consider correction factors to account for the effects of buildings and foliage. Further, multipath is not considered.

### 3.10.2 Durkin's Model — A Case Study

A classical propagation prediction approach similar to that used by Longley-Rice is discussed by Edwards and Durkin [Edw69], as well as Dadson [Dad75]. These papers describe a computer simulator, for predicting field strength contours over irregular terrain, that was adopted by the Joint Radio Committee (JRC) in the U.K. for the estimation of effective mobile radio coverage areas. Although this simulator only predicts large-scale phenomena (i.e. path loss), it provides an interesting perspective into the nature of propagation over irregular terrain and the losses caused by obstacles in a radio path. An explanation of the Edwards and Durkin method is presented here in order to demonstrate how all of the concepts described in this chapter are used in a single model.

The execution of the Durkin path loss simulator consists of two parts. The first part accesses a topographic data base of a proposed service area and reconstructs the ground profile information along the radial joining the transmitter to the receiver. The assumption is that the receiving antenna receives all of its energy along that radial and, therefore, experiences no multipath propagation. In other words, the propagation phenomena that is modeled is simply LOS and diffraction from obstacles along the radial, and excludes reflections from other surrounding objects and local scatterers. The effect of this assumption is that the model is somewhat pessimistic in narrow valleys, although it identifies isolated

weak reception areas rather well. The second part of the simulation algorithm calculates the expected path loss along that radial. After this is done, the simulated receiver location can be iteratively moved to different locations in the service area to deduce the signal strength contour.

The topographical data base can be thought of as a two-dimensional array. Each array element corresponds to a point on a service area map while the actual contents of each array element contain the elevation above sea level data as shown in Figure 3.19. These types of digital elevation models (DEM) are readily available from the United States Geological Survey (USGS). Using this quantized map of service area heights, the program reconstructs the ground profile along the radial that joins the transmitter and the receiver. Since the radial may not always pass through discrete data points, interpolation methods are used to determine the approximate heights that are observed when looking along that radial. Figure 3.20a shows the topographic grid with arbitrary transmitter and receiver locations, the radial between the transmitter and receiver, and the points with which to use diagonal linear interpolation. Figure 3.20b also shows what a typical reconstructed radial terrain profile might look like. In actuality, the values are not simply determined by one interpolation routine, but by a combination of three for increased accuracy. Therefore, each point of the reconstructed profile consists of an average of the heights obtained by diagonal, vertical (row), and horizontal (column) interpolation methods. From these interpolation routines, a matrix of distances from the receiver and corresponding heights along the radial is generated. Now the problem is reduced to a one-dimensional point-to-point link calculation. These types of problems are well-established and procedures for calculating path loss using knife-edge diffraction techniques described previously are used.

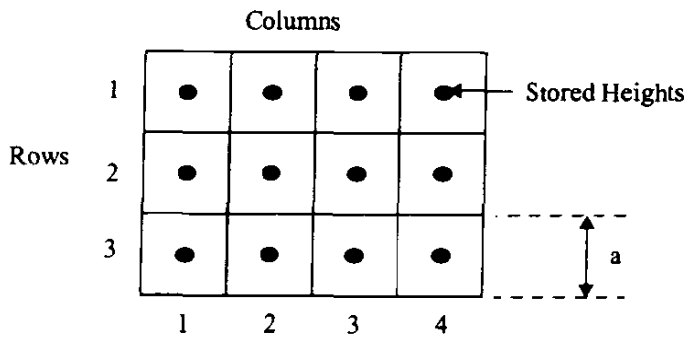
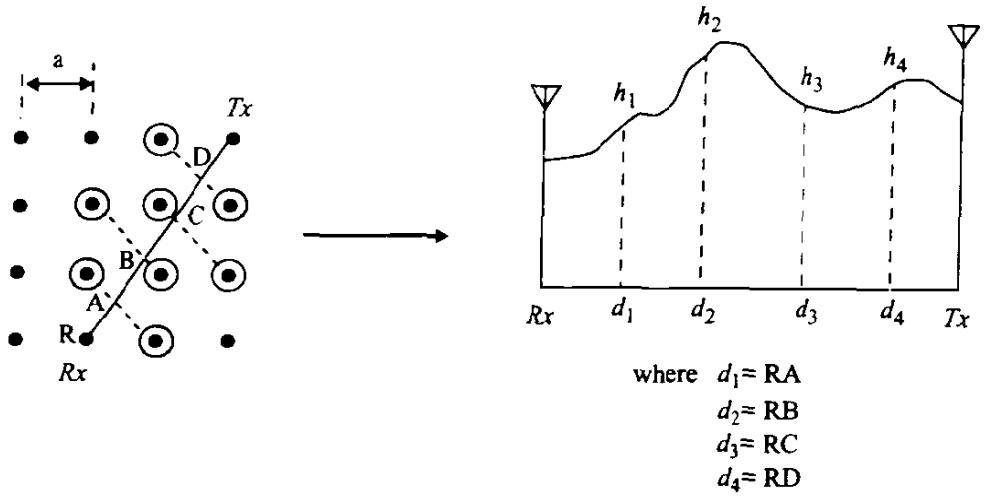



Figure 3.19
Illustration of a two-dimensional array of elevation information.

At this point, the algorithm must make decisions as to what the expected transmission loss should be. The first step is to decide whether a line-of-sight (LOS) path exists between the transmitter and the receiver. To do this, the program computes the difference, $\delta_j$, between the height of the line joining the

transmitter and receiver antennas from the height of the ground profile for each point along the radial (see Figure 3.21).

If any $\delta_j$ ($j = 1, \ldots, n$) is found to be positive along the profile, it is concluded that a LOS path does not exist, otherwise it can be concluded that a LOS path does exist. Assuming the path has a clear LOS, the algorithm then checks to see whether first Fresnel zone clearance is achieved. As shown earlier, if the first Fresnel zone of the radio path is unobstructed, then the resulting loss mechanism is approximately that of free space. If there is an obstruction that just barely touches the line joining the transmitter and the receiver then the signal strength at the receiver is 6 dB less than the free space value due to energy diffracting off the obstruction and away from the receiver. The method for determining first Fresnel zone clearance is done by first calculating the Fresnel diffraction parameter $v$, defined in equation (3.59), for each of the $j$ ground elements.




(a) Top view of interpolated map and line between Tx and Rx

(b) Side view showing reconstruced terrain profile between Tx and Rx

Figure 3.20
*Illustration of terrain profile reconstruction using diagonal interpolation.*

If $v_j \leq -0.8$ for all $j = 1, \ldots, n$, then free space propagation conditions are dominant. For this case, the received power is calculated using the free space transmission formula given in equation (3.1). If the terrain profile failed the first Fresnel zone test (i.e. any $v_j > -0.8$), then there are two possibilities:

a) Non-LOS

b) LOS, but with inadequate first Fresnel-zone clearance.

For both of these cases, the program calculates the free space power using equation (3.1) and the received power using the plane earth propagation equation given by equation (3.52). The algorithm then selects the smaller of the powers calculated with equations (3.1) and (3.52) as the appropriate received power

for the terrain profile. If the profile is LOS with inadequate first Fresnel zone clearance, the next step is to calculate the additional loss due to inadequate Fresnel zone clearance and add it (in dB) to the appropriate received power. This additional diffraction loss is calculated by equation (3.60).

For the case of non-LOS, the system grades the problem into one of four categories:

a) Single diffraction edge
b) Two diffraction edges
c) Three diffraction edges
d) More than three diffraction edges

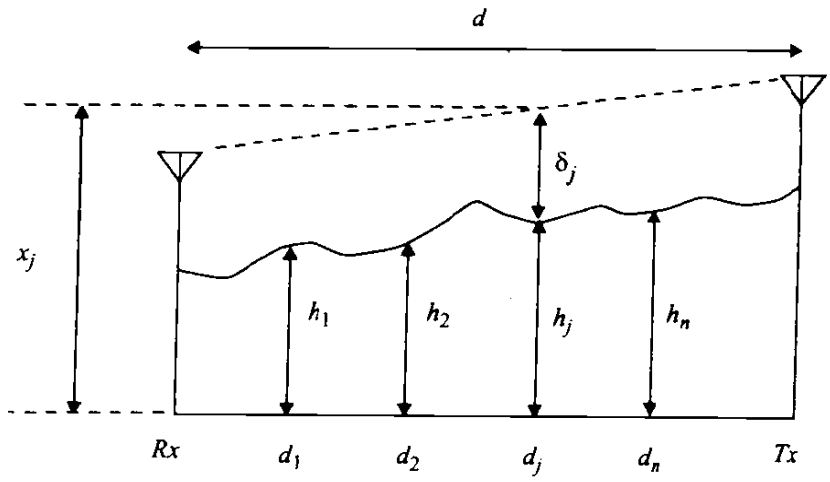



Figure 3.21
Illustration of line-of-sight (LOS) decision making process.

The method tests for each case sequentially until it finds the one that fits the given profile. A diffraction edge is detected by computing the angles between the line joining the transmitter and receiver antennas and the lines joining the receiver antenna to each point on the reconstructed terrain profile. The maximum of these angles is located and labeled by the profile point $(d_i, h_i)$. Next, the algorithm steps through the reverse process of calculating the angles between the line joining the transmitter and receiver antennas and the lines joining the transmitter antenna to each point on the reconstructed terrain profile. The maximum of these angles is found, and it occurs at $(d_j, h_j)$ on the terrain profile. If $d_i = d_j$, then the profile can be modeled as a single diffraction edge. The Fresnel parameter, $v_j$, associated with this edge can be determined from the length of the obstacle above the line joining the transmitter and receiver antennas. The loss can then be evaluated by calculating PL using the equation (3.60). This extra loss caused by the obstacle is then added to either the free space or plane earth loss, whichever is greater.

If the condition for a single diffraction edge is not satisfied, then the check for two diffraction edges is executed. The test is similar to that for a single dif-

fraction edge, with the exception that the computer looks for two edges in sight of each other (see Figure 3.22).

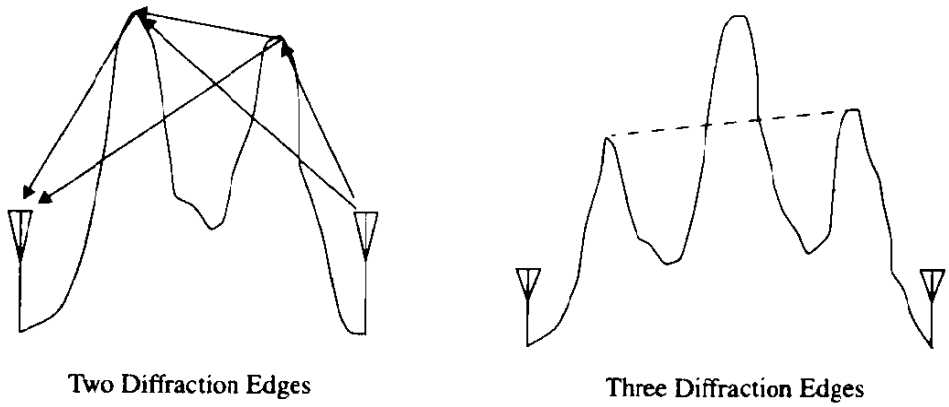



Figure 3.22
Illustration of multiple diffraction edges.

The Edwards and Durkin [Edw69] algorithm uses the Epstein and Peterson method [Eps53] to calculate the loss associated with two diffraction edges. In short, it is the sum of two attenuations. The first attenuation is the loss at the second diffraction edge caused by the first diffraction edge with the transmitter as the source. The second attenuation is the loss at the receiver caused by the second diffraction edge with the first diffraction edge as the source. The two attenuations sum to give the additional loss caused by the obstacles that is added to the free space loss or the plane earth loss, whichever is larger.

For three diffraction edges, the outer diffraction edges must contain a single diffraction edge in between. This is detected by calculating the line between the two outer diffraction edges. If an obstacle between the two outer edges passes through the line, then it is concluded that a third diffraction edge exists (see Figure 3.22). Again, the Epstein and Peterson method is used to calculate the shadow loss caused by the obstacles. For all other cases of more than three diffraction edges, the profile between the outer two obstacles is approximated by a single, virtual knife edge. After the approximation, the problem is that of a three edge calculation.

This method is very attractive because it can read in a digital elevation map and perform a site-specific propagation computation on the elevation data. It can produce a signal strength contour that has been reported to be good within a few dB. The disadvantages are that it cannot adequately predict propagation effects due to foliage, buildings, other man-made structures, and it does not account for multipath propagation other than ground reflection, so additional loss factors are often included. Propagation prediction algorithms which use terrain information are typically used for the design of modern wireless systems.

### 3.10.3 Okumura Model

Okumura's model is one of the most widely used models for sighal prediction in urban areas. This model is applicable for frequencies in the range 150 MHz to 1920 MHz (although it is typically extrapolated up to 3000 MHz) and distances of 1 km to 100 km. It can be used for base station antenna heights ranging from 30 m to 1000 m.

Okumura developed a set of curves giving the median attenuation relative to free space ($A_{mu}$), in an urban area over a quasi-smooth terrain with a base station effective antenna height ($h_{te}$) of 200 m and a mobile antenna height ($h_{re}$) of 3 m. These curves were developed from extensive measurements using vertical omni-directional antennas at both the base and mobile, and are plotted as a function of frequency in the range 100 MHz to 1920 MHz and as a function of distance from the base station in the range 1 km to 100 km. To determine path loss using Okumura's model, the free space path loss between the points of interest is first determined, and then the value of $A_{mu}(f, d)$ (as read from the curves) is added to it along with correction factors to account for the type of terrain. The model can be expressed as

$$L_{50}(\text{dB}) = L_F + A_{mu}(f, d) - G(h_{te}) - G(h_{re}) - G_{AREA} \tag{3.80}$$

where $L_{50}$ is the 50th percentile (i.e., median) value of propagation path loss, $L_F$ is the free space propagation loss, $A_{mu}$ is the median attenuation relative to free space, $G(h_{te})$ is the base station antenna height gain factor, $G(h_{re})$ is the mobile antenna height gain factor, and $G_{AREA}$ is the gain due to the type of environment. Note that the antenna height gains are strictly a function of height and have nothing to do with antenna patterns.

Plots of $A_{mu}(f, d)$ and $G_{AREA}$ for a wide range of frequencies are shown in Figure 3.23 and Figure 3.24. Furthermore, Okumura found that $G(h_{te})$ varies at a rate of 20 dB/decade and $G(h_{re})$ varies at a rate of 10 dB/decade for heights less than 3 m.

$$G(h_{te}) = 20\log\left(\frac{h_{te}}{200}\right) \qquad 1000 \text{ m} > h_{te} > 30 \text{ m} \tag{3.81.a}$$

$$G(h_{re}) = 10\log\left(\frac{h_{re}}{3}\right) \qquad h_{re} \leq 3 \text{ m} \tag{3.81.b}$$

$$G(h_{re}) = 20\log\left(\frac{h_{re}}{3}\right) \qquad 10 \text{ m} > h_{re} > 3 \text{ m} \tag{3.81.c}$$

Other corrections may also be applied to Okumura's model. Some of the important terrain related parameters are the terrain undulation height ($\Delta h$), isolated ridge height, average slope of the terrain and the mixed land-sea parameter. Once the terrain related parameters are calculated, the necessary correction

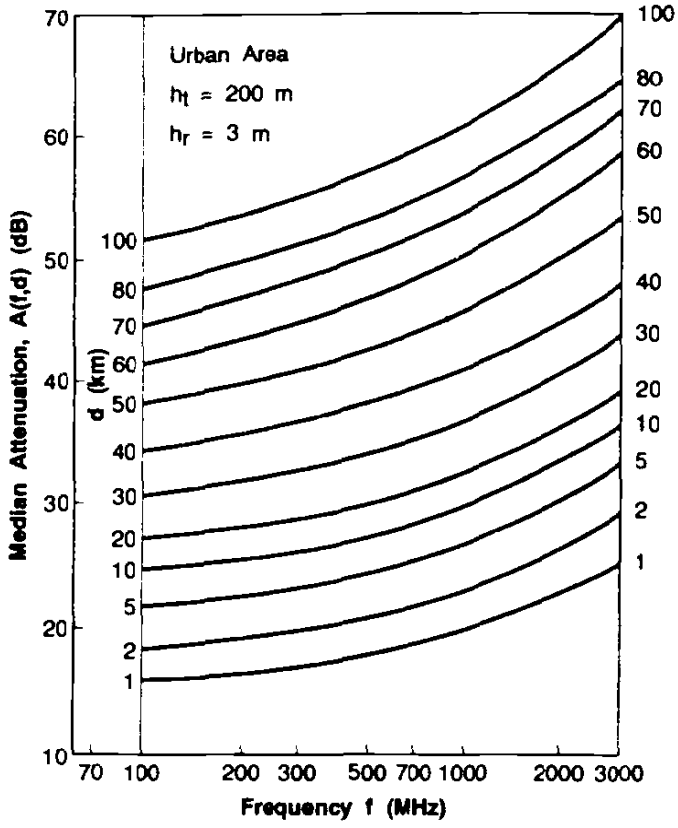



Figure 3.23
Median attenuation relative to free space ($A_{mu}(f,d)$), over a quasi-smooth terrain [From [Oku68] © IEEE].

factors can be added or subtracted as required. All these correction factors are also available as Okumura curves [Oku68].

Okumura's model is wholly based on measured data and does not provide any analytical explanation. For many situations, extrapolations of the derived curves can be made to obtain values outside the measurement range, although the validity of such extrapolations depends on the circumstances and the smoothness of the curve in question.

Okumura's model is considered to be among the simplest and best in terms of accuracy in path loss prediction for mature cellular and land mobile radio systems in cluttered environments. It is very practical and has become a standard for system planning in modern land mobile radio systems in Japan. The major disadvantage with the model is its slow response to rapid changes in terrain, therefore the model is fairly good in urban and suburban areas, but not as good in rural areas. Common standard deviations between predicted and measured path loss values are around 10 dB to 14 dB.

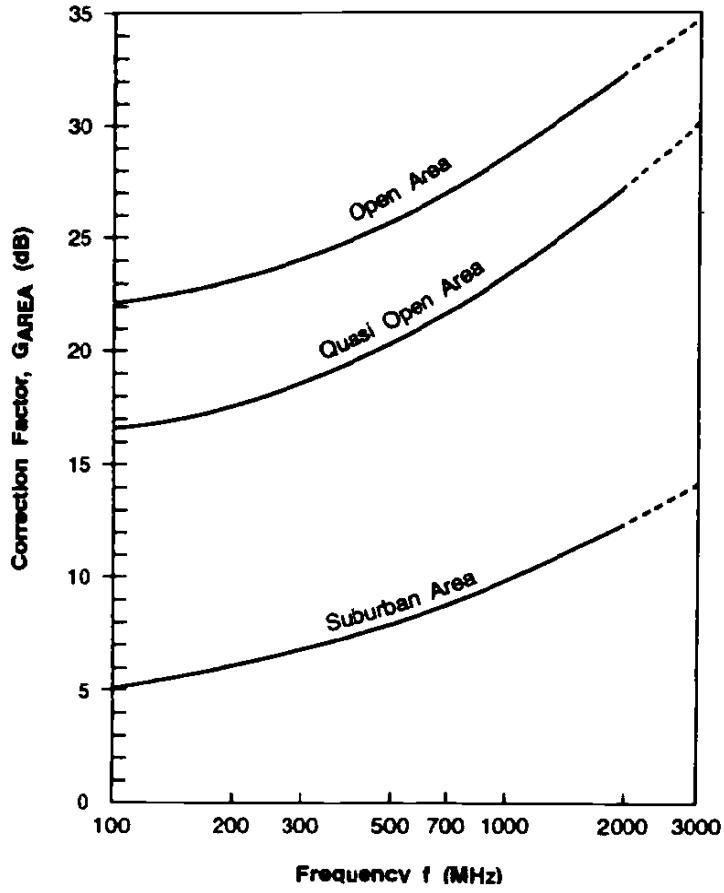



Figure 3.24
Correction factor, $G_{AREA}$, for different types of terrain [From [Oku68] © IEEE].

## Example 3.10

Find the median path loss using Okumura's model for $d$ = 50 km, $h_{te}$ = 100 m, $h_{re}$ = 10 m in a suburban environment. If the base station transmitter radiates an EIRP of 1 kW at a carrier frequency of 900 MHz, find the power at the receiver (assume a unity gain receiving antenna).

## Solution to Example 3.10

The free space path loss $L_F$ can be calculated using equation (3.6) as

$$L_F = 10\log\left[\frac{\lambda^2}{(4\pi)^2 d^2}\right] = 10\log\left[\frac{(3\times 10^8/900\times 10^6)^2}{(4\pi)^2\times(50\times 10^3)^2}\right] = 125.5\text{ dB}.$$

From the Okumura curves

$$A_{mu}(900\text{ MHz}(50\text{ km})) = 43\text{ dB}$$

and

$$G_{AREA} = 9\text{ dB}.$$

Using equation (3.81.a) and (3.81.c) we have

$$G(h_{te}) = 20\log\left(\frac{h_{te}}{200}\right) = 20\log\left(\frac{100}{200}\right) = -6\text{ dB}.$$

$$G(h_{re}) = 20\log\left(\frac{h_{re}}{3}\right) = 20\log\left(\frac{10}{3}\right) = 10.46\text{ dB}.$$

Using equation (3.80) the total mean path loss is

$$\begin{aligned} L_{50}(\text{dB}) &= L_F + A_{mu}(f,d) - G(h_{te}) - G(h_{re}) - G_{AREA} \\ &= 125.5\text{ dB} + 43\text{ dB} - (-6)\text{ dB} - 10.46\text{ dB} - 9\text{ dB} \\ &= 155.04\text{ dB}. \end{aligned}$$

Therefore, the median received power is

$$\begin{aligned} P_r(d) &= EIRP(\text{dBm}) - L_{50}(\text{dB}) + G_r(\text{dB}) \\ &= 60\text{ dBm} - 155.04\text{ dB} + 0\text{ dB} = -95.04\text{ dBm}. \end{aligned}$$

### 3.10.4 Hata Model

The Hata model [Hat90] is an empirical formulation of the graphical path loss data provided by Okumura, and is valid from 150 MHz to 1500 MHz. Hata presented the urban area propagation loss as a standard formula and supplied correction equations for application to other situations. The standard formula for median path loss in urban areas is given by

$$\begin{aligned} L_{50}(urban)(\text{dB}) = 69.55 + 26.16\log f_c - 13.82\log h_{te} - a(h_{re}) \\ + (44.9 - 6.55\log h_{te})\log d \end{aligned} \tag{3.82}$$

where $f_c$ is the frequency (in MHz) from 150 MHz to 1500 MHz, $h_{te}$ is the effective transmitter (base station) antenna height (in meters) ranging from 30 m to 200 m, $h_{re}$ is the effective receiver (mobile) antenna height (in meters) ranging from 1 m to 10 m, $d$ is the T-R separation distance (in km), and $a(h_{re})$ is the correction factor for effective mobile antenna height which is a function of the size of the coverage area. For a small to medium sized city, the mobile antenna correction factor is given by

$$a(h_{re}) = (1.1\log f_c - 0.7)h_{re} - (1.56\log f_c - 0.8)\text{ dB} \tag{3.83}$$

and for a large city, it is given by

$$a(h_{re}) = 8.29(\log 1.54h_{re})^2 - 1.1\text{ dB} \quad \text{for } f_c \le 300\text{ MHz} \tag{3.84.a}$$

$$a(h_{re}) = 3.2(\log 11.75h_{re})^2 - 4.97\text{ dB} \quad \text{for } f_c \ge 300\text{ MHz} \tag{3.84.b}$$

To obtain the path loss in a suburban area the standard Hata formula in equation (3.82) is modified as

$$L_{50}(\text{dB}) = L_{50}(urban) - 2[\log(f_c/28)]^2 - 5.4 \tag{3.85}$$

and for path loss in open rural areas, the formula is modified as

$$L_{50}(\text{dB}) = L_{50}(urban) - 4.78(\log f_c)^2 - 18.33\log f_c - 40.98 \tag{3.86}$$

Although Hata's model does not have any of the path-specific corrections which are available in Okumura's model, the above expressions have significant practical value. The predictions of the Hata model compare very closely with the original Okumura model, as long as $d$ exceeds 1 km. This model is well suited for large cell mobile systems, but not personal communications systems (PCS) which have cells on the order of 1km radius.

### 3.10.5 PCS Extension to Hata Model

The European Co-operative for Scientific and Technical research (EURO-COST) formed the COST-231 working committee to develop an extended version of the Hata model. COST-231 proposed the following formula to extend Hata's model to 2 GHz. The proposed model for path loss is [EUR91]

$$\begin{aligned} L_{50}(urban) = {} & 46.3 + 33.9\log f_c - 13.82\log h_{te} - a(h_{re}) \\ & + (44.9 - 6.55\log h_{te})\log d + C_M \end{aligned} \tag{3.87}$$

where $a(h_{re})$ is defined in equations (3.83), (3.84.a), and (3.84.b) and

$$C_M = \begin{cases} 0 \text{ dB} & \text{for medium sized city and suburban areas} \\ 3 \text{ dB} & \text{for metropolitan centers} \end{cases} \tag{3.88}$$

The COST-231 extension of the Hata model is restricted to the following range of parameters:

$f$ : 1500 MHz to 2000 MHz
$h_{te}$ : 30 m to 200 m
$h_{re}$ : 1 m to 10 m
$d$ : 1 km to 20 km

### 3.10.6 Walfisch and Bertoni Model

A model developed by Walfisch and Bertoni [Wal88] considers the impact of rooftops and building height by using diffraction to predict average signal strength at street level. The model considers the path loss, $S$, to be a product of three factors.

$$S = P_0 Q^2 P_1 \tag{3.89}$$

where $P_0$ represents free space path loss between isotropic antennas given by

$$P_0 = \left(\frac{\lambda}{4\pi R}\right)^2 \tag{3.90}$$

The factor $Q^2$ gives the reduction in the rooftop signal due to the row of buildings which immediately shadow the receiver at street level. The $P_1$ term is

based upon diffraction and determines the signal loss from the rooftop to the street.

In dB, the path loss is given by

$$S\,(\text{dB}) = L_0 + L_{rts} + L_{ms} \tag{3.91}$$

where $L_0$ represents free space loss, $L_{rts}$ represents the "rooftop-to-street diffraction and scatter loss", and $L_{ms}$ denotes multiscreen diffraction loss due to the rows of buildings [Xia92]. Figure 3.25 illustrates the geometry used in the Walfisch Bertoni model [Wal88], [Mac93]. This model is being considered for use by ITU-R in the IMT-2000 standards activities.

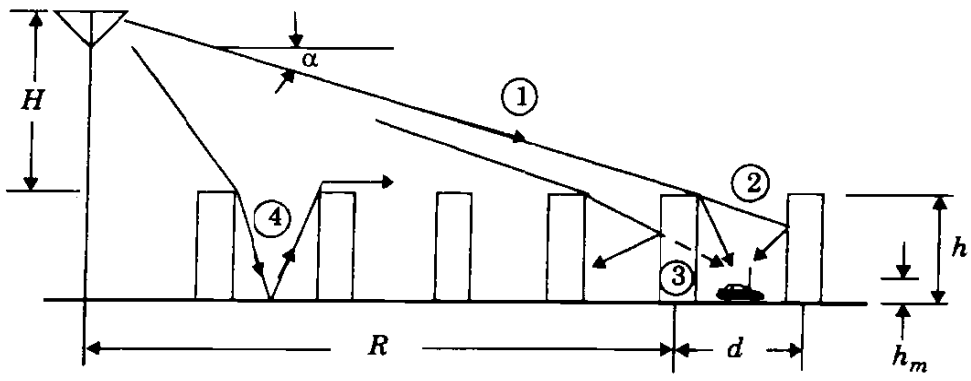



Figure 3.25
Propagation geometry for model proposed by Walfisch and Bertoni [From [Wal88] © IEEE].

### 3.10.7 Wideband PCS Microcell Model

Work by Feuerstein, et.al. in 1991 used a 20 MHz pulsed transmitter at 1900 MHz to measure path loss, outage, and delay spread in typical microcellular systems in San Francisco and Oakland. Using base station antenna heights of 3.7 m, 8.5 m, and 13.3 m, and a mobile receiver with an antenna height of 1.7 m above ground, statistics for path loss, multipath, and coverage area were developed from extensive measurements in line-of-sight (LOS) and obstructed (OBS) environments [Feu94]. This work revealed that a 2-ray ground reflection model (shown in Figure 3.7) is a good estimate for path loss in LOS microcells, and a simple log-distance path loss model holds well for OBS microcell environments.

For a flat earth ground reflection model, the distance $d_f$ at which the first Fresnel zone just becomes obstructed by the ground (first Fresnel zone clearance) is given by

$$d_f = \frac{1}{\lambda}\sqrt{(\Sigma^2-\Delta^2)^2 - 2(\Sigma^2+\Delta^2)\left(\frac{\lambda}{2}\right)^2 + \left(\frac{\lambda}{2}\right)^4} \tag{3.92.a}$$
$$= \frac{1}{\lambda}\sqrt{16h_t^2h_r^2 - \lambda^2\left(h_t^2+h_r^2\right) + \frac{\lambda^4}{16}}$$

For LOS cases, a double regression path loss model that uses a regression breakpoint at the first Fresnel zone clearance was shown to fit well to measurements. The model assumes omnidirectional vertical antennas and predicts average path loss as

$$\overline{PL}(d) = \begin{cases} 10n_1\log(d) + p_1 & \text{for } 1 < d < d_f \\ 10n_2\log(d/d_f) + 10n_1\log d_f + p_1 & \text{for } d > d_f \end{cases} \tag{3.92.b}$$

where $p_1$ is equal to $\overline{PL}(d_0)$ (the path loss in decibels at the reference distance of $d_0 = 1$ m), $d$ is in meters and $n_1$, $n_2$ are path loss exponents which are a function of transmitter height, as given in Figure 3.26. It can easily be shown that at 1900 MHz, $p_1 = 38.0$ dB.

For the OBS case, the path loss was found to fit the standard log-distance path loss law of equation (3.69.a)

$$\overline{PL}(d)\,[dB] = 10n\log(d) + p_1 \tag{3.92.c}$$

where $n$ is the OBS path loss exponent given in Figure 3.26 as a function of transmitter height. The standard deviation (in dB) of the log-normal shadowing component about the distance-dependent mean was found from measurements using the techniques described in Chapter 3, section 3.10.2. The log-normal shadowing component is also listed as a function of height for both the LOS and OBS microcell environments. Figure 3.26 indicates that the log-normal shadowing component is between 7 and 9 dB regardless of antenna height. It can be seen that LOS environments provide slightly less path loss than the theoretical 2-ray ground reflected model, which would predict $n_1 = 2$ and $n_2 = 4$.

| Transmitter Antenna Height | 1900 MHz LOS | | | 1900 MHz OBS | |
|---|---|---|---|---|---|
| | $n_1$ | $n_2$ | σ(dB) | $n$ | σ(dB) |
| Low (3.7m) | 2.08 | 3.29 | 8.76 | 2.58 | 9.31 |
| Medium (8.5m) | 2.17 | 3.36 | 7.88 | 2.56 | 7.67 |
| High (13.3m) | 2.07 | 4.16 | 8.77 | 2.69 | 7.94 |

Figure 3.26
Parameters for the wideband microcell model at 1900 MHz [From [Feu94] © IEEE].

CHAPTER 4

# Mobile Radio Propagation: Small-Scale Fading and Multipath

Small-scale fading, or simply *fading*, is used to describe the rapid fluctuation of the amplitude of a radio signal over a short period of time or travel distance, so that large-scale path loss effects may be ignored. Fading is caused by interference between two or more versions of the transmitted signal which arrive at the receiver at slightly different times. These waves, called *multipath waves*, combine at the receiver antenna to give a resultant signal which can vary widely in amplitude and phase, depending on the distribution of the intensity and relative propagation time of the waves and the bandwidth of the transmitted signal.

## 4.1 Small-Scale Multipath Propagation

Multipath in the radio channel creates small-scale fading effects. The three most important effects are:

- Rapid changes in signal strength over a small travel distance or time interval
- Random frequency modulation due to varying Doppler shifts on different multipath signals
- Time dispersion (echoes) caused by multipath propagation delays.

In built-up urban areas, fading occurs because the height of the mobile antennas are well below the height of surrounding structures, so there is no single line-of-sight path to the base station. Even when a line-of-sight exists, multipath still occurs due to reflections from the ground and surrounding structures. The incoming radio waves arrive from different directions with different propagation delays. The signal received by the mobile at any point in space may consist of a large number of plane waves having randomly distributed amplitudes,

phases, and angles of arrival. These multipath components combine vectorially at the receiver antenna, and can cause the signal received by the mobile to distort or fade. Even when a mobile receiver is stationary, the received signal may fade due to movement of surrounding objects in the radio channel.

If objects in the radio channel are static, and motion is considered to be only due to that of the mobile, then fading is purely a spatial phenomenon. The spatial variations of the resulting signal are seen as temporal variations by the receiver as it moves through the multipath field. Due to the constructive and destructive effects of multipath waves summing at various points in space, a receiver moving at high speed can pass through several fades in a small period of time. In a more serious case, a receiver may stop at a particular location at which the received signal is in a deep fade. Maintaining good communications can then become very difficult, although passing vehicles or people walking in the vicinity of the mobile can often disturb the field pattern, thereby diminishing the likelihood of the received signal remaining in a deep null for a long period of time. Antenna space diversity can prevent deep fading nulls, as shown in Chapter 6. Figure 3.1 shows typical rapid variations in the received signal level due to small-scale fading as a receiver is moved over a distance of a few meters.

Due to the relative motion between the mobile and the base station, each multipath wave experiences an apparent shift in frequency. The shift in received signal frequency due to motion is called the Doppler shift, and is directly proportional to the velocity and direction of motion of the mobile with respect to the direction of arrival of the received multipath wave.

### 4.1.1 Factors Influencing Small-Scale Fading

Many physical factors in the radio propagation channel influence small-scale fading. These include the following:

- **Multipath propagation** — The presence of reflecting objects and scatterers in the channel creates a constantly changing environment that dissipates the signal energy in amplitude, phase, and time. These effects result in multiple versions of the transmitted signal that arrive at the receiving antenna, displaced with respect to one another in time and spatial orientation. The random phase and amplitudes of the different multipath components cause fluctuations in signal strength, thereby inducing small-scale fading, signal distortion, or both. Multipath propagation often lengthens the time required for the baseband portion of the signal to reach the receiver which can cause signal smearing due to intersymbol interference.
- **Speed of the mobile** — The relative motion between the base station and the mobile results in random frequency modulation due to different Doppler shifts on each of the multipath components. Doppler shift will be positive or negative depending on whether the mobile receiver is moving toward or away from the base station.

- **Speed of surrounding objects** — If objects in the radio channel are in motion, they induce a time varying Doppler shift on multipath components. If the surrounding objects move at a greater rate than the mobile, then this effect dominates the small-scale fading. Otherwise, motion of surrounding objects may be ignored, and only the speed of the mobile need be considered.
- **The transmission bandwidth of the signal** — If the transmitted radio signal bandwidth is greater than the "bandwidth" of the multipath channel, the received signal will be distorted, but the received signal strength will not fade much over a local area (i.e., the small-scale signal fading will not be significant). As will be shown, the bandwidth of the channel can be quantified by the *coherence bandwidth* which is related to the specific multipath structure of the channel. The coherence bandwidth is a measure of the maximum frequency difference for which signals are still strongly correlated in amplitude. If the transmitted signal has a narrow bandwidth as compared to the channel, the amplitude of the signal will change rapidly, but the signal will not be distorted in time. Thus, the statistics of small-scale signal strength and the likelihood of signal smearing appearing over small-scale distances are very much related to the specific amplitudes and delays of the multipath channel, as well as the bandwidth of the transmitted signal.

### 4.1.2 Doppler Shift

Consider a mobile moving at a constant velocity $v$, along a path segment having length $d$ between points X and Y, while it receives signals from a remote source S, as illustrated in Figure 4.1. The difference in path lengths traveled by the wave from source S to the mobile at points X and Y is $\Delta l = d\cos\theta = v\Delta t\cos\theta$, where $\Delta t$ is the time required for the mobile to travel from X to Y, and $\theta$ is assumed to be the same at points X and Y since the source is assumed to be very far away. The phase change in the received signal due to the difference in path lengths is therefore

$$\Delta\phi = \frac{2\pi\Delta l}{\lambda} = \frac{2\pi v\Delta t}{\lambda}\cos\theta \tag{4.1}$$

and hence the apparent change in frequency, or Doppler shift, is given by $f_d$, where

$$f_d = \frac{1}{2\pi}\cdot\frac{\Delta\phi}{\Delta t} = \frac{v}{\lambda}\cdot\cos\theta \tag{4.2}$$

Equation (4.2) relates the Doppler shift to the mobile velocity and the spatial angle between the direction of motion of the mobile and the direction of arrival of the wave. It can be seen from equation (4.2) that if the mobile is moving toward the direction of arrival of the wave, the Doppler shift is positive (i.e., the apparent received frequency is increased), and if the mobile is moving away from the direction of arrival of the wave, the Doppler shift is negative (i.e. the

apparent received frequency is decreased). As shown in section 4.7.1, multipath components from a CW signal which arrive from different directions contribute to Doppler spreading of the received signal, thus increasing the signal bandwidth.

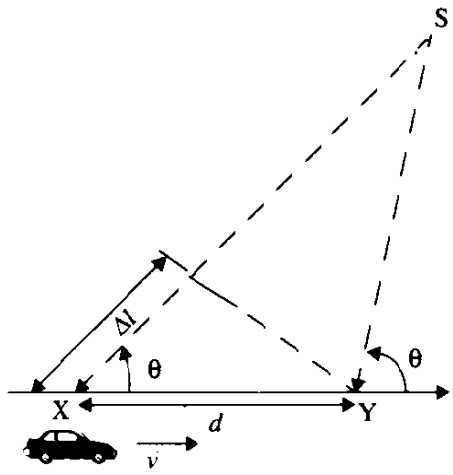



Figure 4.1
Illustration of Doppler effect.

**Example 4.1**

Consider a transmitter which radiates a sinusoidal carrier frequency of 1850 MHz. For a vehicle moving 60 mph, compute the received carrier frequency if the mobile is moving (a) directly towards the transmitter, (b) directly away from the transmitter, (c) in a direction which is perpendicular to the direction of arrival of the transmitted signal.

**Solution to Example 4.1**

Given:

Carrier frequency $f_c = 1850\,\text{MHz}$

Therefore, wavelength $\lambda = c/f_c = \dfrac{3\times10^8}{1850\times10^6} = 0.162\,\text{m}$

Vehicle speed $v = 60$ mph $= 26.82$ m/s

(a) The vehicle is moving directly towards the transmitter.
The Doppler shift in this case is positive and the received frequency is given by equation (4.2)

$$f = f_c + f_d = 1850\times10^6 + \frac{26.82}{0.162} = 1850.00016\ \text{MHz}$$

(b) The vehicle is moving directly away from the transmitter.
The Doppler shift in this case is negative and hence the received frequency is given by

$$f = f_c - f_d = 1850 \times 10^6 - \frac{26.82}{0.162} = 1849.999834 \text{ MHz}$$

(c) The vehicle is moving perpendicular to the angle of arrival of the transmitted signal.
In this case, $\theta = 90°$, $\cos\theta = 0$, and there is no Doppler shift.
The received signal frequency is the same as the transmitted frequency of 1850 MHz.

## 4.2 Impulse Response Model of a Multipath Channel

The small-scale variations of a mobile radio signal can be directly related to the impulse response of the mobile radio channel. The impulse response is a wideband channel characterization and contains all information necessary to simulate or analyze any type of radio transmission through the channel. This stems from the fact that a mobile radio channel may be modeled as a linear filter with a time varying impulse response, where the time variation is due to receiver motion in space. The filtering nature of the channel is caused by the summation of amplitudes and delays of the multiple arriving waves at any instant of time. The impulse response is a useful characterization of the channel, since it may be used to predict and compare the performance of many different mobile communication systems and transmission bandwidths for a particular mobile channel condition.

To show that a mobile radio channel may be modeled as a linear filter with a time varying impulse response, consider the case where time variation is due strictly to receiver motion in space. This is shown in Figure 4.2.


Figure 4.2
The mobile radio channel as a function of time and space.



In Figure 4.2, the receiver moves along the ground at some constant velocity $v$. For a fixed position $d$, the channel between the transmitter and the receiver can be modeled as a linear time invariant system. However, due to the different multipath waves which have propagation delays which vary over different spatial locations of the receiver, the impulse response of the linear time invariant channel should be a function of the position of the receiver. That is, the channel impulse response can be expressed as $h(d,t)$. Let $x(t)$ represent the transmitted signal, then the received signal $y(d,t)$ at position $d$ can be expressed as a convolution of $x(t)$ with $h(d,t)$.

$$y(d,t) = x(t) \otimes h(d,t) = \int_{-\infty}^{\infty} x(\tau)h(d,t-\tau)d\tau \tag{4.3}$$

For a causal system, $h(d,t) = 0$ for $t<0$, thus equation (4.3) reduces to

$$y(d,t) = \int_{-\infty}^{t} x(\tau)h(d,t-\tau)d\tau \tag{4.4}$$

Since the receiver moves along the ground at a constant velocity $v$, the position of the receiver can by expressed as

$$d = vt \tag{4.5}$$

Substituting (4.5) in (4.4), we obtain

$$y(vt,t) = \int_{-\infty}^{t} x(\tau)h(vt,t-\tau)d\tau \tag{4.6}$$

Since $v$ is a constant, $y(vt,t)$ is just a function of $t$. Therefore, equation (4.6) can be expressed as

$$y(t) = \int_{-\infty}^{t} x(\tau)h(vt,t-\tau)d\tau = x(t) \otimes h(vt,t) = x(t) \otimes h(d,t) \tag{4.7}$$

From equation (4.7) it is clear that the mobile radio channel can be modeled as a linear time varying channel, where the channel changes with time and distance.

Since $v$ may be assumed constant over a short time (or distance) interval, we may let $x(t)$ represent the transmitted bandpass waveform, $y(t)$ the received waveform, and $h(t,\tau)$ the impulse response of the time varying multipath radio channel. The impulse response $h(t,\tau)$ completely characterizes the channel and is a function of both $t$ and $\tau$. The variable $t$ represents the time variations due to motion, whereas $\tau$ represents the channel multipath delay for a fixed value of $t$. One may think of $\tau$ as being a vernier adjustment of time. The received signal $y(t)$ can be expressed as a convolution of the transmitted signal $x(t)$ with the channel impulse response (see Figure 4.3a).

$$y(t) = \int_{-\infty}^{\infty} x(\tau)h(t,\tau)d\tau = x(t) \otimes h(t,\tau) \tag{4.8}$$

If the multipath channel is assumed to be a bandlimited bandpass channel, which is reasonable, then $h(t,\tau)$ may be equivalently described by a complex baseband impulse response $h_b(t,\tau)$ with the input and output being the complex envelope representations of the transmitted and received signals, respectively (see Figure 4.3b). That is,

$$r(t) = c(t) \otimes \frac{1}{2}h_b(t,\tau) \tag{4.9}$$

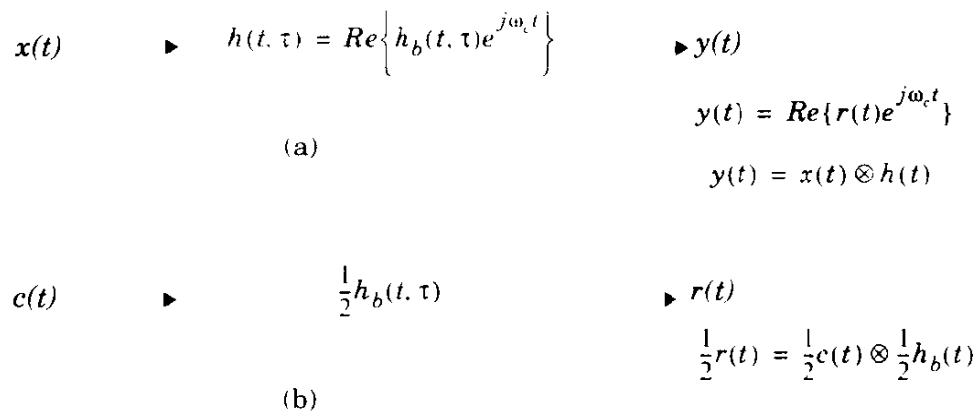



Figure 4.3
(a) Bandpass channel impulse response model.
(b) Baseband equivalent channel impulse response model.

where $c(t)$ and $r(t)$ are the complex envelopes of $x(t)$ and $y(t)$, defined as

$$x(t) = Re\{c(t)\exp(j2\pi f_c t)\} \tag{4.10}$$

$$y(t) = Re\{r(t)\exp(j2\pi f_c t)\} \tag{4.11}$$

The factor of $1/2$ in equation (4.9) is due to the properties of the complex envelope, in order to represent the passband radio system at baseband. The low-pass characterization removes the high frequency variations caused by the carrier, making the signal analytically easier to handle. It is shown by Couch [Cou93] that the average power of a bandpass signal $\overline{x^2(t)}$ is equal to $\frac{1}{2}\overline{|c(t)|^2}$, where the overbar denotes ensemble average for a stochastic signal, or time average for a deterministic or ergodic stochastic signal.

It is useful to discretize the multipath delay axis $\tau$ of the impulse response into equal time delay segments called *excess delay bins*, where each bin has a time delay width equal to $\tau_{i+1} - \tau_i$, where $\tau_0$ is equal to 0, and represents the first arriving signal at the receiver. Letting i = 0, it is seen that $\tau_1 - \tau_0$ is equal to the time delay bin width given by $\Delta\tau$. For convention, $\tau_0 = 0$, $\tau_1 = \Delta\tau$, and $\tau_i = i\Delta\tau$, for $i = 0$ to $N-1$, where $N$ represents the total number of possible equally-spaced multipath components, including the first arriving component. Any number of multipath signals received within the $i$ th bin are represented by a single resolvable multipath component having delay $\tau_i$. This technique of quantizing the delay bins determines the time delay resolution of the channel model, and the useful frequency span of the model can be shown to be $1/(2\Delta\tau)$. That is, the model may be used to analyze transmitted signals having bandwidths which are less than $1/(2\Delta\tau)$. Note that $\tau_0 = 0$ is the excess time delay

of the first arriving multipath component, and neglects the propagation delay between the transmitter and receiver. *Excess delay* is the relative delay of the $i$ th multipath component as compared to the first arriving component and is given by $\tau_i$. The *maximum excess delay* of the channel is given by $N\Delta\tau$.

Since the received signal in a multipath channel consists of a series of attenuated, time-delayed, phase shifted replicas of the transmitted signal, the baseband impulse response of a multipath channel can be expressed as

$$h_b(t,\tau) = \sum_{i=0}^{N-1} a_i(t,\tau)\exp[j(2\pi f_c\tau_i(t) + \phi_i(t,\tau))]\delta(\tau - \tau_i(t)) \tag{4.12}$$

where $a_i(t,\tau)$ and $\tau_i(t)$ are the real amplitudes and excess delays, respectively, of $i$ th multipath component at time $t$ [Tur72]. The phase term $2\pi f_c\tau_i(t) + \phi_i(t,\tau)$ in (4.12) represents the phase shift due to free space propagation of the $i$ th multipath component, plus any additional phase shifts which are encountered in the channel. In general, the phase term is simply represented by a single variable $\theta_i(t,\tau)$ which lumps together all the mechanisms for phase shifts of a single multipath component within the $i$th excess delay bin. Note that some excess delay bins may have no multipath at some time $t$ and delay $\tau_i$, since $a_i(t,\tau)$ may be zero. In equation (4.12), $N$ is the total possible number of multipath components (bins), and $\delta(\bullet)$ is the unit impulse function which determines the specific multipath bins that have components at time $t$ and excess delays $\tau_i$. Figure 4.4 illustrates an example of different snapshots of $h_b(t,\tau)$, where $t$ varies into the page, and the time delay bins are quantized to widths of $\Delta\tau$.

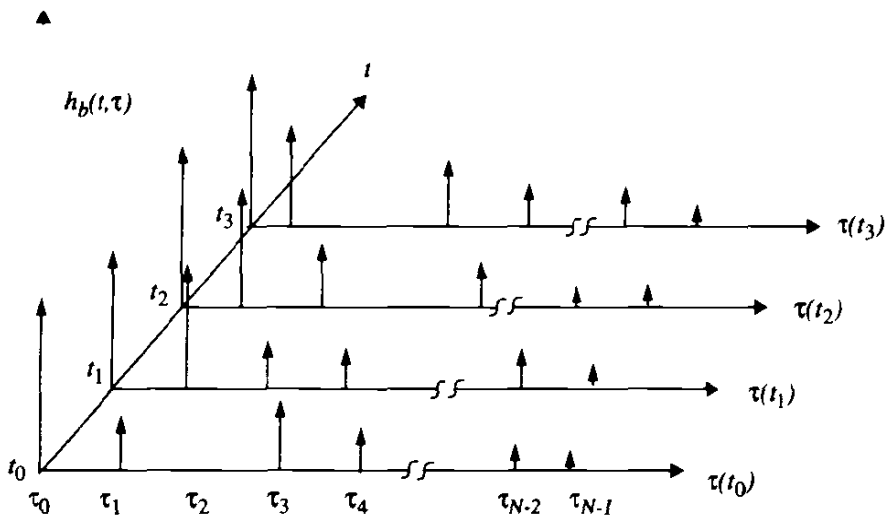



Figure 4.4
An example of the time varying discrete-time impulse response model for a multipath radio channel.

If the channel impulse response is assumed to be time invariant, or is at least wide sense stationary over a small-scale time or distance interval, then the channel impulse response may be simplified as

$$h_b(\tau) = \sum_{i=0}^{N-1} a_i \exp(-j\theta_i)\,\delta(\tau - \tau_i) \tag{4.13}$$

When measuring or predicting $h_b(\tau)$, a probing pulse $p(t)$ which approximates a delta function is used at the transmitter. That is

$$p(t) \approx \delta(t - \tau) \tag{4.14}$$

is used to sound the channel to determine $h_b(\tau)$.

For small-scale channel modeling, the *power delay profile* of the channel is found by taking the spatial average of $|h_b(t;\tau)|^2$ over a local area. By making several local area measurements of $|h_b(t;\tau)|^2$ in different locations, it is possible to build an ensemble of power delay profiles, each one representing a possible small-scale multipath channel state [Rap91a].

Based on work by Cox [Cox72], [Cox75], if $p(t)$ has a time duration much smaller than the impulse response of the multipath channel, $p(t)$ does not need to be deconvolved from the received signal $r(t)$ in order to determine relative multipath signal strengths. The received power delay profile in a local area is given by

$$P(t;\tau) \approx k|h_b(t;\tau)|^2 \tag{4.15}$$

and many snapshots of $|h_b(t;\tau)|^2$ are typically averaged over a local (small-scale) area to provide a single time-invariant multipath power delay profile $P(\tau)$. The gain $k$ in equation (4.15) relates the transmitted power in the probing pulse $p(t)$ to the total power received in a multipath delay profile.

### 4.2.1 Relationship Between Bandwidth and Received Power

In actual wireless communication systems, the impulse response of a multipath channel is measured in the field using channel sounding techniques. We now consider two extreme channel sounding cases as a means of demonstrating how the small-scale fading behaves quite differently for two signals with different bandwidths in the identical multipath channel.

Consider a pulsed, transmitted RF signal of the form

$$x(t) = \mathrm{Re}\{p(t)\exp(j2\pi f_c t)\}$$

where $p(t)$ is a repetitive baseband pulse train with very narrow pulse width $T_{bb}$ and repetition period $T_{REP}$ which is much greater than the maximum measured excess delay $\tau_{max}$ in the channel. Now let

$$p(t) = 2\sqrt{\tau_{max}/T_{bb}} \text{ for } 0 \le t \le T_{bb}$$

and let $p(t)$ be zero elsewhere for all excess delays of interest. The low pass channel output $r(t)$ closely approximates the impulse response $h_b(t)$ and is given by

$$r(t) = \frac{1}{2}\sum_{i=0}^{N-1} a_i(\exp(-j\theta_i)) \cdot p(t-\tau_i) \tag{4.16}$$
$$= \sum_{i=0}^{N-1} a_i \exp(-j\theta_i) \cdot \sqrt{\frac{\tau_{max}}{T_{bb}}}\ \mathrm{rect}\left[t - \frac{T_{bb}}{2} - \tau_i\right]$$

To determine the received power at some time $t_0$, the power $|r(t_0)|^2$ is measured. The quantity $|r(t_0)|^2$ is called the *instantaneous multipath power delay profile* of the channel, and is equal to the energy received over the time duration of the multipath delay divided by $\tau_{max}$. That is, using equation (4.16)

$$|r(t_0)|^2 = \frac{1}{\tau_{max}}\int_0^{\tau_{max}} r(t) \times r^*(t)\,dt \tag{4.17}$$
$$= \frac{1}{\tau_{max}}\int_0^{\tau_{max}} \frac{1}{4}\mathrm{Re}\left\{\sum_{j=0}^{N-1}\sum_{i=0}^{N-1} a_j(t_0)a_i(t_0)p(t-\tau_j)p(t-\tau_i)\exp(-j(\theta_j-\theta_i))\right\}dt$$

Note that if all the multipath components are resolved by the probe *p(t)*, then $|\tau_j - \tau_i| > T_{bb}$ for all $j \neq i$, and

$$|r(t_0)|^2 = \frac{1}{\tau_{max}}\int_0^{\tau_{max}} \frac{1}{4}\left(\sum_{k=0}^{N-1} a_k^2(t_0)p^2(t-\tau_k)\right)dt \tag{4.18}$$
$$= \frac{1}{\tau_{max}}\sum_{k=0}^{N-1} a_k^2(t_0)\int_0^{\tau_{max}}\left\{\sqrt{\frac{\tau_{max}}{T_{bb}}}\,\mathrm{rect}\left[t - \frac{T_{bb}}{2} - \tau_k\right]\right\}^2 dt$$
$$= \sum_{k=0}^{N-1} a_k^2(t_0)$$

For a wideband probing signal *p(t)*, $T_{bb}$ is smaller than the delays between multipath components in the channel, and equation (4.18) shows that the total received power is simply related to the sum of the powers in the individual multipath components, and is scaled by the ratio of the probing pulse's width and amplitude, and the maximum observed excess delay of the channel. Assuming that the received power from the multipath components forms a random process where each component has a random amplitude and phase at any time $t$, the average small-scale received power for the wideband probe is found from equation (4.17) as

$$E_{a,\theta}[P_{WB}] = E_{a,\theta}\left[\sum_{i=0}^{N-1} |a_i \exp(j\theta_i)|^2\right] \approx \sum_{i=0}^{N-1} \overline{a_i^2} \tag{4.19}$$

In equation (4.19), $E_{a,\theta}[\bullet]$ denotes the ensemble average over all possible values of $a_i$ and $\theta_i$ in a local area, and the overbar denotes sample average over a local measurement area which is generally measured using multipath measurement equipment. The striking result of equations (4.18) and (4.19) is that if a transmitted signal is able to resolve the multipaths, then the *small-scale received power is simply the sum of the powers received in each multipath component*. In practice, the amplitudes of individual multipath components do not fluctuate widely in a local area. Thus, the received power of a wideband signal such as $p(t)$ does not fluctuate significantly when a receiver is moved about a local area [Rap89].

Now, instead of a pulse, consider a CW signal which is transmitted into the exact same channel, and let the complex envelope be given by $c(t) = 2$. Then, the instantaneous complex envelope of the received signal is given by the phasor sum

$$r(t) = \sum_{i=0}^{N-1} a_i \exp(j\theta_i(t,\tau)) \tag{4.20}$$

and the instantaneous power is given by

$$|r(t)|^2 = \left|\sum_{i=0}^{N-1} a_i \exp(j\theta_i(t,\tau))\right|^2 \tag{4.21}$$

As the receiver is moved over a local area, the channel changes, and the received signal strength will vary at a rate governed by the fluctuations of $a_i$ and $\theta_i$. As mentioned earlier, $a_i$ varies little over local areas, but $\theta_i$ will vary greatly due to changes in propagation distance over space, resulting in large fluctuations of $r(t)$ as the receiver is moved over small distances (on the order of a wavelength). That is, since $r(t)$ is the phasor sum of the individual multipath components, the instantaneous phases of the multipath components cause the large fluctuations which typifies small-scale fading for CW signals. The average received power over a local area is then given by

$$E_{a,\theta}[P_{CW}] = E_{a,\theta}\left[\left|\sum_{i=0}^{N-1} a_i \exp(j\theta_i)\right|^2\right] \tag{4.22}$$

$$E_{a,\theta}[P_{CW}] \approx \overline{\left[\left(a_0 e^{j\theta_0} + a_1 e^{j\theta_1} + \ldots + a_{N-1} e^{j\theta_{N-1}}\right) \times \left(a_0 e^{-j\theta_0} + a_1 e^{-j\theta_1} + \ldots + a_{N-1} e^{-j\theta_{N-1}}\right)\right]} \tag{4.23}$$

$$E_{a,\theta}[P_{CW}] \approx \sum_{i=0}^{N-1} \overline{a_i^2} + 2\sum_{i=0}^{N-1} \sum_{i,j \neq i}^{N} r_{ij}\, \overline{\cos(\theta_i - \theta_j)} \tag{4.24}$$

where $r_{ij}$ is the path amplitude correlation coefficient defined to be

$$r_{ij} = E_a[a_i a_j] \tag{4.25}$$

and the overbar denotes time average for CW measurements made by a mobile receiver over the local measurement area [Rap89]. Note that when $\overline{\cos(\theta_i - \theta_j)} = 0$ and/or $r_{ij} = 0$, then the average power for a CW signal is equivalent to the average received power for a wideband signal in a small-scale region. This is seen by comparing equation (4.19) and equation (4.24). This can occur when either the multipath phases are identically and independently distributed (i.i.d uniform) over $[0, 2\pi]$ or when the path amplitudes are uncorrelated. The i.i.d uniform distribution of $\theta$ is a valid assumption since multipath components traverse differential path lengths that measure hundreds of wavelengths and are likely to arrive with random phases. If for some reason it is believed that the phases are not independent, the average wideband power and average CW power will still be equal if the paths have uncorrelated amplitudes. However, if the phases of the paths are dependent upon each other, then the amplitudes are likely to be correlated, since the same mechanism which affects the path phases is likely to also affect the amplitudes. This situation is highly unlikely at transmission frequencies used in wireless mobile systems.

Thus it is seen that the *received local ensemble average power of wideband and narrowband signals are equivalent.* When the transmitted signal has a bandwidth much greater than the bandwidth of the channel, then the multipath structure is completely resolved by the received signal at any time, and the received power varies very little since the individual multipath amplitudes do not change rapidly over a local area. However, if the transmitted signal has a very narrow bandwidth (e.g., the baseband signal has a duration greater than the excess delay of the channel), then multipath is not resolved by the received signal, and large signal fluctuations (fading) occur at the receiver due to the phase shifts of the many unresolved multipath components.

Figure 4.5 illustrates actual indoor radio channel measurements made simultaneously with a wideband probing pulse having $T_{bb} = 10$ ns, and a CW transmitter. The carrier frequency was 4 GHz. It can be seen that the CW signal undergoes rapid fades, whereas the wideband measurements change little over the $5\lambda$ measurement track. However, the local average received powers of both signals were measured to be virtually identical [Haw91].

**Example 4.2**

Assume a discrete channel impulse response is used to model urban radio channels with excess delays as large as 100 $\mu s$ and microcellular channels with excess delays no larger than 4 $\mu s$. If the number of multipath bins is fixed

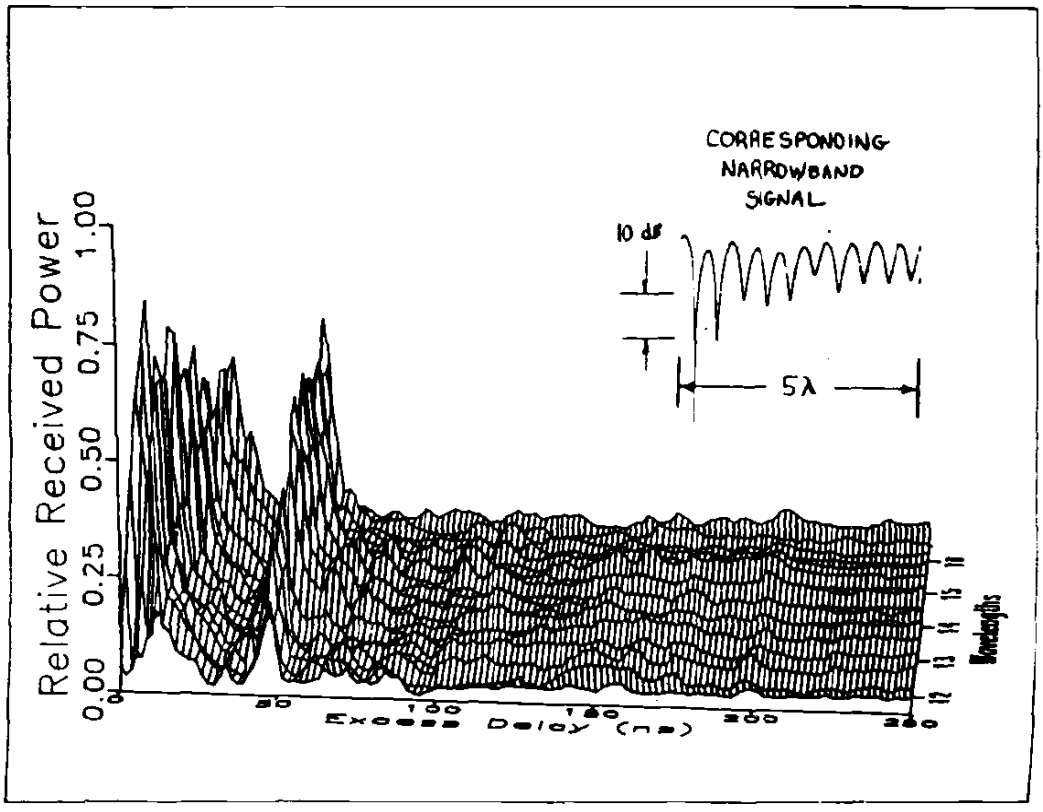



Figure 4.5
Measured wideband and narrowband received signals over a 5λ (0.375 m) measurement track inside a building. Carrier frequency is 4 GHz. Wideband power is computed using equation (4.19), which can be thought of as the area under the power delay profile.

at 64, find (a) $\Delta\tau$, and (b) the maximum bandwidth which the two models can accurately represent. Repeat the exercise for an indoor channel model with excess delays as large as 500 ns. As described in section 4.7.6, SIRCIM and SMRCIM are statistical channel models based on equation (4.12) that use parameters in this example.

## Solution to Example 4.2

The maximum excess delay of the channel model is given by $\tau_N = N\Delta\tau$. Therefore, for $\tau_N = 100\mu s$, and $N = 64$ we obtain $\Delta\tau = \tau_N/N = 1.5625\,\mu s$. The maximum bandwidth that the SMRCIM model can accurately represent is equal to

$$1/(2\Delta\tau) = 1/(2(1.5625\mu s)) = 0.32\text{ MHz}.$$

For the SMRCIM urban microcell model, $\tau_N = 4\mu s$, $\Delta\tau = \tau_N/N = 62.5$ ns. The maximum bandwidth that can be represented is

$$1/(2\Delta\tau) = 1/(2(62.5\text{ ns})) = 8\text{ MHz}.$$

Similarly, for indoor channels, $\Delta\tau = \tau_N/N = \frac{500 \times 10^{-9}}{64} = 7.8125$ ns.

The maximum bandwidth for the indoor channel model is

$$1/(2\Delta\tau) = 1/(2(7.8125 \text{ ns})) = 64 \text{ MHz}.$$

## Example 4.3

Assume a mobile traveling at a velocity of 10 m/s receives two multipath components at a carrier frequency of 1000 MHz. The first component is assumed to arrive at $\tau = 0$ with an initial phase of $0°$ and a power of -70 dBm, and the second component which is 3 dB weaker than the first component is assumed to arrive at $\tau = 1\mu s$, also with an initial phase of $0°$. If the mobile moves directly towards the direction of arrival of the first component and directly away from the direction of arrival of the second component, compute the narrowband instantaneous power at time intervals of 0.1 s from 0 s to 0.5 s. Compute the average narrowband power received over this observation interval. Compare average narrowband and wideband received powers over the interval.

## Solution to Example 4.3

Given $v = 10$ m/s, time intervals of 0.1 s correspond to spatial intervals of 1 m. The carrier frequency is given to be 1000 MHz, hence the wavelength of the signal is

$$\lambda = \frac{c}{f} = \frac{3 \times 10^8}{1000 \times 10^6} = 0.3 \text{ m}$$

The narrowband instantaneous power can be computed using equation (4.21). Note -70 dBm = 100 pW. At time $t = 0$, the phases of both multipath components are $0°$, hence the narrowband instantaneous power is equal to

$$|r(t)|^2 = \left|\sum_{i=0}^{N-1} a_i \exp(j\theta_i(t,\tau))\right|^2$$

$$= \left|\sqrt{100 \text{ pW}} \times \exp(0) + \sqrt{50 \text{ pW}} \times \exp(0)\right|^2 = 291 \text{ pW}$$

Now, as the mobile moves, the phase of the two multipath components changes in opposite directions.

At $t = 0.1$ s, the phase of the first component is

$$\theta_i = \frac{2\pi d}{\lambda} = \frac{2\pi v t}{\lambda} = \frac{2\pi \times 10 \text{ (m/s)} \times 0.1 \text{ s}}{0.3 \text{ m}}$$

$$= 20.94 \text{ rad} = 2.09 \text{ rad} = 120°$$

Since the mobile moves towards the direction of arrival of the first component, and away from the direction of arrival of the second component, $\theta_1$ is positive, and $\theta_2$ is negative.

Therefore, at $t = 0.1$ s, $\theta_1 = 120°$, and $\theta_2 = -120°$, and the instantaneous power is equal to

$$|r(t)|^2 = \left|\sum_{i=0}^{N-1} a_i \exp(j\theta_i(t,\tau))\right|^2$$

$$= \left|\sqrt{100\text{ pW}} \times \exp(j120°) + \sqrt{50\text{ pW}} \times \exp(-j120°)\right|^2 = 78.2\text{ pW}$$

Similarly, at $t = 0.2$ s, $\theta_2 = 240°$, and $\theta_2 = -240°$, and the instantaneous power is equal to

$$|r(t)|^2 = \left|\sum_{i=0}^{N-1} a_i \exp(j\theta_i(t,\tau))\right|^2$$

$$= \left|\sqrt{100\text{ pW}} \quad \exp(j240°) + \sqrt{50\text{ pW}} \times \exp(-j240°)\right|^2 = 81.5\text{ pW}$$

Similarly, at $t = 0.3$ s, $\theta_1 = 360° = 0°$, and $\theta_2 = -360° = 0°$, and the instantaneous power is equal to

$$|r(t)|^2 = \left|\sum_{i=0}^{N-1} a_i \exp(j\theta_i(t,\tau))\right|^2$$

$$= \left|\sqrt{100\text{ pW}} \times \exp(j0°) + \sqrt{50\text{ pW}} \times \exp(-j0°)\right|^2 = 291\text{ pW}$$

It follows that at $t = 0.4$ s, $|r(t)|^2 = 78.2$ pW, and at $t = 0.5$ s, $|r(t)|^2 = 81.5$ pW.

The average narrowband received power is equal to

$$\frac{(2)(291) + (2)(78.2) + (2)(81.5)}{6}\text{ pW} = 150.233\text{ pW}$$

Using equation (4.19), the wideband power is given by

$$E_{a,\theta}[P_{W,B}] = E_{a,\theta}\left[\sum_{i=0}^{N-1} |a_i \exp(j\theta_i)|^2\right] \approx \sum_{i=0}^{N-1} \overline{a_i^2}$$

$$E_{a,\theta}[P_{W,B}] = 100\text{ pW} + 50\text{ pW} = 150\text{ pW}$$

As can be seen, the narrowband and wideband received power are virtually identical when averaged over 0.5 s (or 5 m). While the CW signal fades over the observation interval, the wideband signal power remains constant.

## 4.3 Small-Scale Multipath Measurements

Because of the importance of the multipath structure in determining the small-scale fading effects, a number of wideband channel sounding techniques have been developed. These techniques may be classified as *direct pulse measurements*, *spread spectrum sliding correlator measurements*, and *swept frequency measurements*.

surements (e.g., indoor channel sounding). Another limitation with this system is the non-real-time nature of the measurement. For time varying channels, the channel frequency response can change rapidly, giving an erroneous impulse response measurement. To mitigate this effect, fast sweep times are necessary to keep the total swept frequency response measurement interval as short as possible. A faster sweep time can be accomplished by reducing the number of frequency steps, but this sacrifices time resolution and excess delay range in the time domain. The swept frequency system has been used successfully for indoor propagation studies by Pahlavan [Pah95] and Zaghloul, et.al. [Zag91a], [Zag91b].

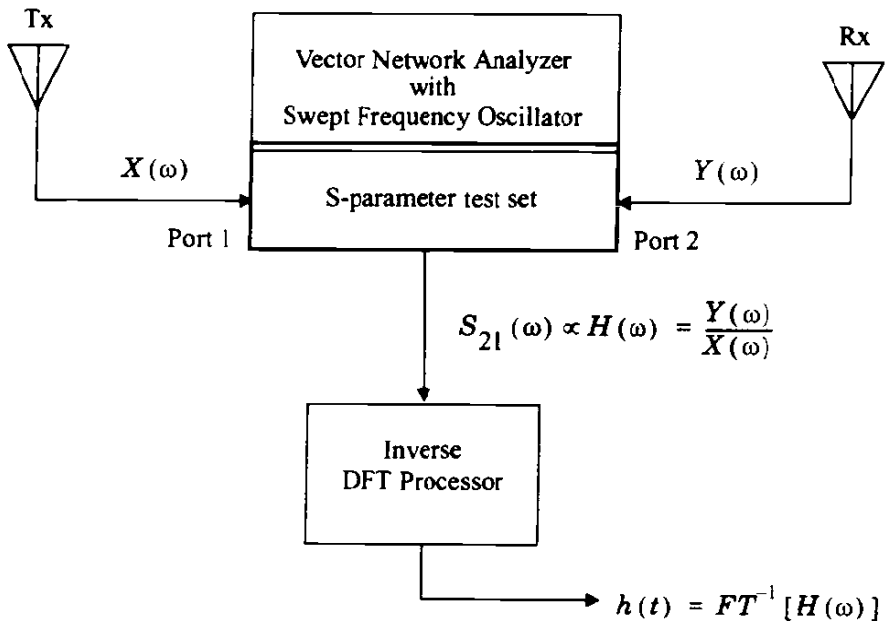



Figure 4.8
Frequency domain channel impulse response measurement system.

## 4.4 Parameters of Mobile Multipath Channels

Many multipath channel parameters are derived from the power delay profile, given by equation (4.18). Power delay profiles are measured using the techniques discussed in Section 4.4 and are generally represented as plots of relative received power as a function of excess delay with respect to a fixed time delay reference. Power delay profiles are found by averaging instantaneous power delay profile measurements over a local area in order to determine an average small-scale power delay profile. Depending on the time resolution of the probing pulse and the type of multipath channels studied, researchers often choose to sample at spatial separations of a quarter of a wavelength and over receiver movements no greater than 6 m in outdoor channels and no greater than 2 m in indoor channels in the 450 MHz - 6 GHz range. This small-scale sampling avoids

large-scale averaging bias in the resulting small-scale statistics. Figure 4.9 shows typical power delay profile plots from outdoor and indoor channels, determined from a large number of closely sampled instantaneous profiles.

### 4.4.1 Time Dispersion Parameters

In order to compare different multipath channels and to develop some general design guidelines for wireless systems, parameters which grossly quantify the multipath channel are used. The *mean excess delay, rms delay spread,* and *excess delay spread (X* dB*)* are multipath channel parameters that can be determined from a power delay profile. The time dispersive properties of wide band multipath channels are most commonly quantified by their mean excess delay ($\bar{\tau}$) and rms delay spread ($\sigma_\tau$). The mean excess delay is the first moment of the power delay profile and is defined to be

$$\bar{\tau} = \frac{\sum_k a_k^2 \tau_k}{\sum_k a_k^2} = \frac{\sum_k P(\tau_k)\,\tau_k}{\sum_k P(\tau_k)} \tag{4.35}$$

The rms delay spread is the square root of the second central moment of the power delay profile and is defined to be

$$\sigma_\tau = \sqrt{\overline{\tau^2} - (\bar{\tau})^2} \tag{4.36}$$

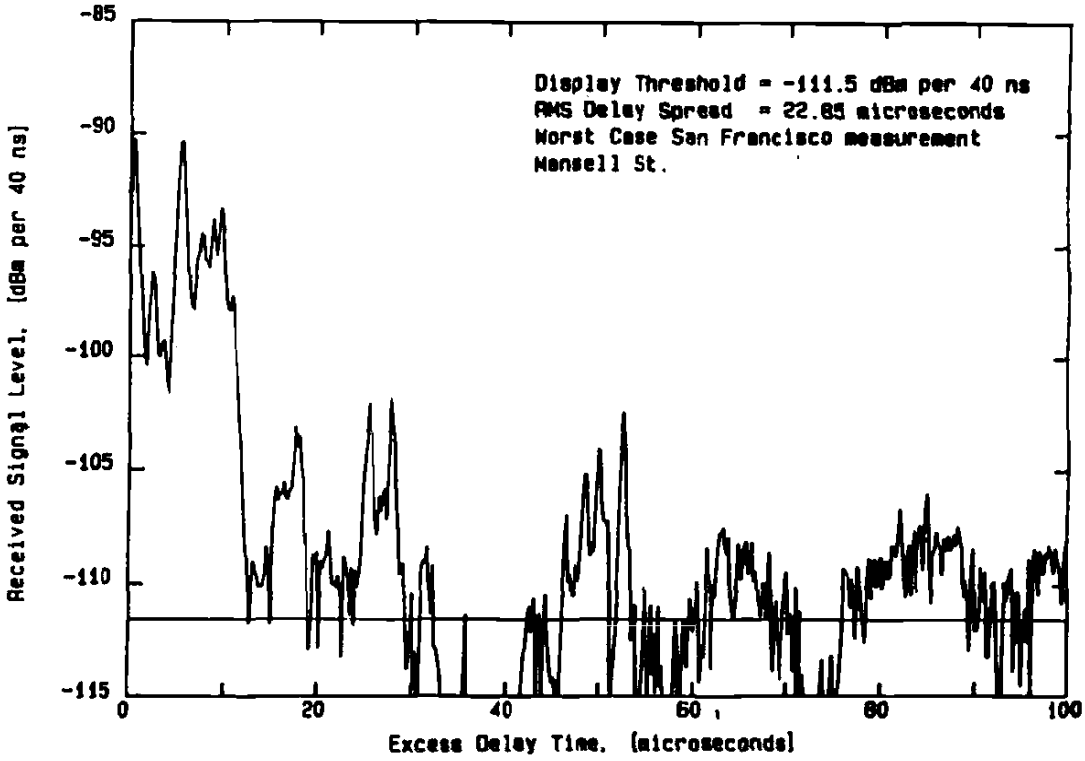
where

$$\overline{\tau^2} = \frac{\sum_k a_k^2 \tau_k^2}{\sum_k a_k^2} = \frac{\sum_k P(\tau_k)\,\tau_k^2}{\sum_k P(\tau_k)} \tag{4.37}$$

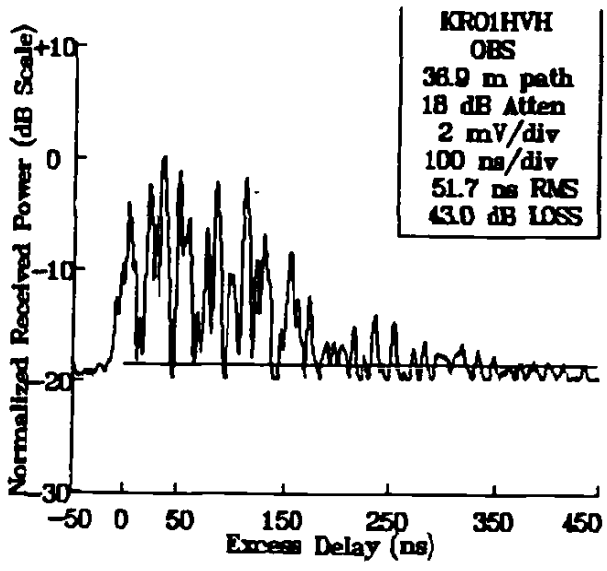
These delays are measured relative to the first detectable signal arriving at the receiver at $\tau_0 = 0$. Equations (4.35) - (4.37) do not rely on the absolute power level of $P(\tau)$, but only the relative amplitudes of the multipath components within $P(\tau)$. Typical values of rms delay spread are on the order of microseconds in outdoor mobile radio channels and on the order of nanoseconds in indoor radio channels. Table 4.1 shows the typical measured values of rms delay spread.

It is important to note that the rms delay spread and mean excess delay are defined from a single power delay profile which is the temporal or spatial average of consecutive impulse response measurements collected and averaged over a local area. Typically, many measurements are made at many local areas in order to determine a statistical range of multipath channel parameters for a mobile communication system over a large-scale area [Rap90].

The *maximum excess delay (X* dB*)* of the power delay profile is defined to be the time delay during which multipath energy falls to $X$ dB below the maxi-




(a)




(b)

Figure 4.9
Measured multipath power delay profiles
a) From a 900 MHz cellular system in San Francisco [From [Rap90] © IEEE].
b) Inside a grocerv store at 4 GHz [From [Haw91] © IEEE].

**Table 4.1 Typical Measured Values of RMS Delay Spread**

| Environment | Frequency (MHz) | RMS Delay Spread ($\sigma_\tau$) | Notes | Reference |
|---|---|---|---|---|
| Urban | 910 | 1300 ns avg.<br>600 ns st. dev.<br>3500 ns max. | New York City | [Cox75] |
| Urban | 892 | 10-25 μs | Worst case San Francisco | [Rap90] |
| Suburban | 910 | 200-310 ns | Averaged typical case | [Cox72] |
| Suburban | 910 | 1960-2110 ns | Averaged extreme case | [Cox72] |
| Indoor | 1500 | 10-50 ns<br>25 ns median | Office building | [Sal87] |
| Indoor | 850 | 270 ns max. | Office building | [Dev90a] |
| Indoor | 1900 | 70-94 ns avg.<br>1470 ns max. | Three San Francisco buildings | [Sei92a] |

mum. In other words, the maximum excess delay is defined as $\tau_X - \tau_0$, where $\tau_0$ is the first arriving signal and $\tau_X$ is the maximum delay at which a multipath component is within $X$ dB of the strongest arriving multipath signal (which does not necessarily arrive at $\tau_0$). Figure 4.10 illustrates the computation of the maximum excess delay for multipath components within 10 dB of the maximum. The maximum excess delay ($X$ dB) defines the temporal extent of the multipath that is above a particular threshold. The value of $\tau_X$ is sometimes called the *excess delay spread* of a power delay profile, but in all cases must be specified with a threshold that relates the multipath noise floor to the maximum received multipath component.

In practice, values for $\bar{\tau}$, $\overline{\tau^2}$, and $\sigma_\tau$ depend on the choice of noise threshold used to process $P(\tau)$. The noise threshold is used to differentiate between received multipath components and thermal noise. If the noise threshold is set too low, then noise will be processed as multipath, thus giving rise to values of $\bar{\tau}$, $\overline{\tau^2}$, and $\sigma_\tau$ that are artificially high.

It should be noted that the power delay profile and the magnitude frequency response (the spectral response) of a mobile radio channel are related through the Fourier transform. It is therefore possible to obtain an equivalent description of the channel in the frequency domain using its frequency response characteristics. Analogous to the delay spread parameters in the time domain, *coherence bandwidth* is used to characterize the channel in the frequency domain. The rms delay spread and coherence bandwidth are inversely proportional to one another, although their exact relationship is a function of the exact multipath structure.

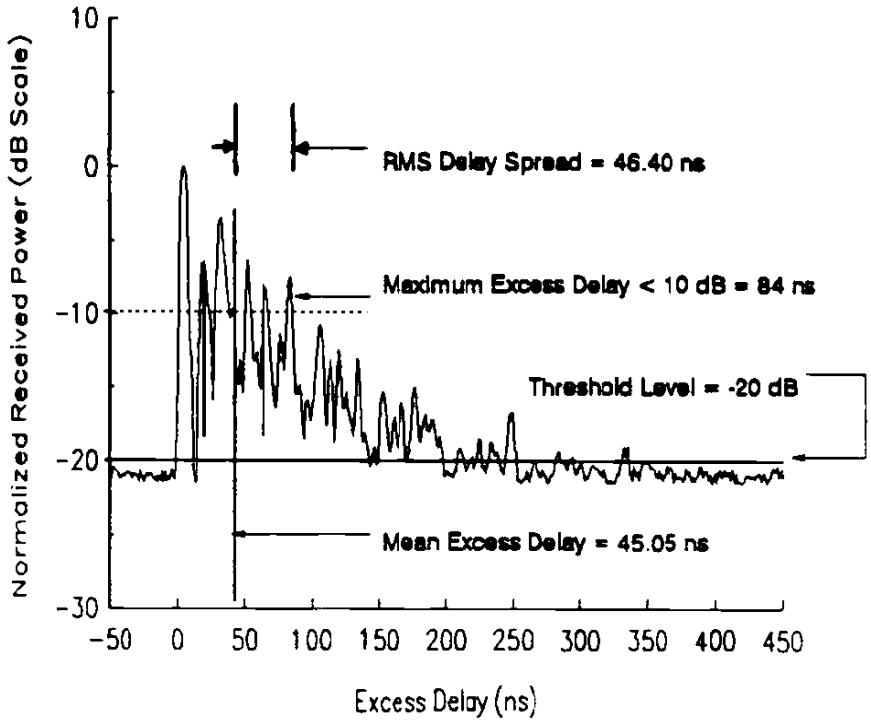



Figure 4.10
Example of an indoor power delay profile; rms delay spread, mean excess delay, maximum excess delay (10 dB), and threshold level are shown.

### 4.4.2 Coherence Bandwidth

While the delay spread is a natural phenomenon caused by reflected and scattered propagation paths in the radio channel, the coherence bandwidth, $B_c$, is a defined relation derived from the rms delay spread. Coherence bandwidth is a statistical measure of the range of frequencies over which the channel can be considered "flat" (i.e., a channel which passes all spectral components with approximately equal gain and linear phase). In other words, coherence bandwidth is the range of frequencies over which two frequency components have a strong potential for amplitude correlation. Two sinusoids with frequency separation greater than $B_c$ are affected quite differently by the channel. If the coherence bandwidth is defined as the bandwidth over which the frequency correlation function is above 0.9, then the coherence bandwidth is approximately [Lee89b]

$$B_c \approx \frac{1}{50\sigma_\tau} \tag{4.38}$$

If the definition is relaxed so that the frequency correlation function is above 0.5, then the coherence bandwidth is approximately

$$B_c \approx \frac{1}{5\sigma_\tau} \tag{4.39}$$

It is important to note that an exact relationship between coherence bandwidth and rms delay spread does not exist, and equations (4.38) and (4.39) are "ball park estimates". In general, spectral analysis techniques and simulation are required to determine the exact impact that time varying multipath has on a particular transmitted signal [Chu87], [Fun93], [Ste94]. For this reason, accurate multipath channel models must be used in the design of specific modems for wireless applications [Rap91a], [Woe94].

**Example 4.4**

Calculate the mean excess delay, rms delay spread, and the maximum excess delay (10 dB) for the multipath profile given in the figure below. Estimate the 50% coherence bandwidth of the channel. Would this channel be suitable for AMPS or GSM service without the use of an equalizer?

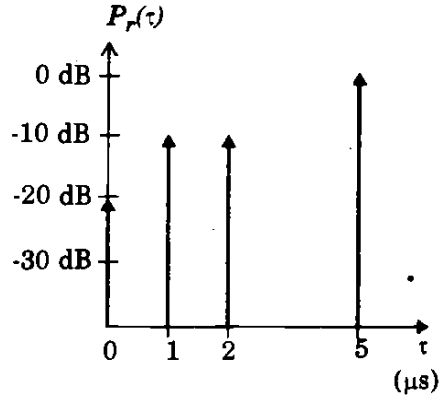



Figure E4.4

**Solution to Example 4.4**

The rms delay spread for the given multipath profile can be obtained using equations (4.35) – (4.37). The delays of each profile are measured relative to the first detectable signal. The mean excess delay for the given profile

$$\bar{\tau} = \frac{(1)(5) + (0.1)(1) + (0.1)(2) + (0.01)(0)}{[0.01 + 0.1 + 0.1 + 1]} = 4.38\mu s$$

The second moment for the given power delay profile can be calculated as

$$\overline{\tau^2} = \frac{(1)(5)^2 + (0.1)(1)^2 + (0.1)(2)^2 + (0.01)(0)}{1.21} = 21.07\mu s^2$$

Therefore the rms delay spread, $\sigma_\tau = \sqrt{21.07 - (4.38)^2} = 1.37\mu s$

The coherence bandwidth is found from equation (4.39) to be

$$B_c \approx \frac{1}{5\sigma_\tau} = \frac{1}{5(1.37\mu s)} = 146 \text{ kHz}$$

Since $B_c$ is greater than 30 kHz, AMPS will work without an equalizer. However, GSM requires 200 kHz bandwidth which exceeds $B_c$, thus an equalizer would be needed for this channel.

### 4.4.3 Doppler Spread and Coherence Time

Delay spread and coherence bandwidth are parameters which describe the time dispersive nature of the channel in a local area. However, they do not offer information about the time varying nature of the channel caused by either relative motion between the mobile and base station, or by movement of objects in the channel. *Doppler spread* and *coherence time* are parameters which describe the time varying nature of the channel in a small-scale region.

Doppler spread $B_D$ is a measure of the spectral broadening caused by the time rate of change of the mobile radio channel and is defined as the range of frequencies over which the received Doppler spectrum is essentially non-zero. When a pure sinusoidal tone of frequency $f_c$ is transmitted, the received signal spectrum, called the Doppler spectrum, will have components in the range $f_c - f_d$ to $f_c + f_d$, where $f_d$ is the Doppler shift. The amount of spectral broadening depends on $f_d$ which is a function of the relative velocity of the mobile, and the angle $\theta$ between the direction of motion of the mobile and direction of arrival of the scattered waves. *If the baseband signal bandwidth is much greater than $B_D$, the effects of Doppler spread are negligible at the receiver.* This is a *slow* fading channel.

Coherence time $T_c$ is the time domain dual of Doppler spread and is used to characterize the time varying nature of the frequency dispersiveness of the channel in the time domain. The Doppler spread and coherence time are inversely proportional to one another. That is,

$$T_C \approx \frac{1}{f_m} \tag{4.40.a}$$

Coherence time is actually a statistical measure of the time duration over which the channel impulse response is essentially invariant, and quantifies the similarity of the channel response at different times. In other words, coherence time is the time duration over which two received signals have a strong potential for amplitude correlation. If the reciprocal bandwidth of the baseband signal is greater than the coherence time of the channel, then the channel will change during the transmission of the baseband message, thus causing distortion at the receiver. If the coherence time is defined as the time over which the time correlation function is above 0.5, then the coherence time is approximately [Ste94]

$$T_C \approx \frac{9}{16\pi f_m} \tag{4.40.b}$$

where $f_m$ is the maximum Doppler shift given by $f_m = v/\lambda$. In practice, (4.40.a) suggests a time duration during which a Rayleigh fading signal may fluctuate

wildly, and (4.40.b) is often too restrictive. A popular rule of thumb for modern digital communications is to define the coherence time as the geometric mean of equations (4.40.a) and (4.40.b). That is,

$$T_C = \sqrt{\frac{9}{16\pi f_m^2}} = \frac{0.423}{f_m} \tag{4.40.c}$$

The definition of coherence time implies that two signals arriving with a time separation greater than $T_C$ are affected differently by the channel. For example, for a vehicle traveling 60 mph using a 900 MHz carrier, a conservative value of $T_C$ can be shown to be 2.22 ms from (4.40.b). If a digital transmission system is used, then as long as the symbol rate is greater than $1/T_C = 454$ bps, the channel will not cause distortion due to motion (however, distortion could result from multipath time delay spread, depending on the channel impulse response). Using the practical formula of (4.40.c), $T_C = 6.77$ ms and the symbol rate must exceed 150 bits/s in order to avoid distortion due to frequency dispersion.

**Example 4.5**

Determine the proper spatial sampling interval required to make small-scale propagation measurements which assume that consecutive samples are highly correlated in time. How many samples will be required over 10 m travel distance if $f_c = 1900$ MHz and $v = 50$ m/s. How long would it take to make these measurements, assuming they could be made in real time from a moving vehicle? What is the Doppler spread $B_D$ for the channel?

**Solution to Example 4.5**

For correlation, ensure that the time between samples is equal to $T_C/2$, and use the smallest value of $T_C$ for conservative design.
Using equation (4.40.b)

$$T_C \approx \frac{9}{16\pi f_m} = \frac{9\lambda}{16\pi v} = \frac{9c}{16\pi v f_c} = \frac{9 \times 3 \times 10^8}{16 \times 3.14 \times 50 \times 1900 \times 10^6}$$

$$T_C = 565\mu s$$

Taking time samples at less than half $T_C$, at $282.5\mu s$ corresponds to a spatial sampling interval of

$$\Delta x = \frac{vT_C}{2} = \frac{50 \times 565\mu s}{2} = 0.014125 \text{ m} = 1.41 \text{ cm}$$

Therefore, the number of samples required over a 10 m travel distance is

$$N_x = \frac{10}{\Delta x} = \frac{10}{0.014125} = 708 \text{ samples}$$

The time taken to make this measurement is equal to $\frac{10\text{m}}{50 \text{ m/s}} = 0.2$ s
The Doppler spread is

$$B_D = f_m = \frac{vf_c}{c} = \frac{50 \times 1900 \times 10^6}{3 \times 10^8} = 316.66 \text{ Hz}$$

## 4.5 Types of Small-Scale Fading

Section 4.3 demonstrated that the type of fading experienced by a signal propagating through a mobile radio channel depends on the nature of the transmitted signal with respect to the characteristics of the channel. Depending on the relation between the signal parameters (such as bandwidth, symbol period, etc.) and the channel parameters (such as rms delay spread and Doppler spread), different transmitted signals will undergo different types of fading. The time dispersion and frequency dispersion mechanisms in a mobile radio channel lead to four possible distinct effects, which are manifested depending on the nature of the transmitted signal, the channel, and the velocity. While multipath delay spread leads to *time dispersion* and *frequency selective fading*, Doppler spread leads to *frequency dispersion* and *time selective fading*. The two propagation mechanisms are independent of one another. Figure 4.11 shows a tree of the four different types of fading.

**Small-Scale Fading**
(Based on multipath time delay spread)

**Flat Fading**
1. BW of signal < BW of channel
2. Delay spread < Symbol period

**Frequency Selective Fading**
1. BW of signal > BW of channel
2. Delay spread > Symbol period

**Small-Scale Fading**
(Based on Doppler spread)

**Fast Fading**
1. High Doppler spread
2. Coherence time < Symbol period
3. Channel variations faster than baseband signal variations

**Slow Fading**
1. Low Doppler spread
2. Coherence time > Symbol period
3. Channel variations slower than baseband signal variations

Figure 4.11
Types of small-scale fading.

### 4.5.1 Fading Effects Due to Multipath Time Delay Spread

Time dispersion due to multipath causes the transmitted signal to undergo either flat or frequency selective fading.

#### 4.5.1.1 Flat fading

If the mobile radio channel has a constant gain and linear phase response over a bandwidth which is greater than the bandwidth of the transmitted signal, then the received signal will undergo *flat fading*. This type of fading is historically the most common type of fading described in the technical literature. In flat fading, the multipath structure of the channel is such that the spectral characteristics of the transmitted signal are preserved at the receiver. However the strength of the received signal changes with time, due to fluctuations in the gain of the channel caused by multipath. The characteristics of a flat fading channel are illustrated in Figure 4.12.

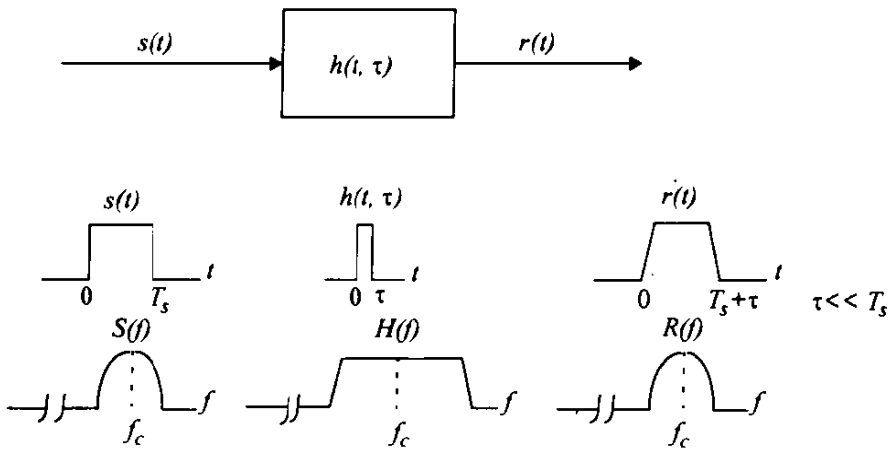



Figure 4.12
Flat fading channel characteristics.

It can be seen from Figure 4.12 that if the channel gain changes over time, a change of amplitude occurs in the received signal. Over time, the received signal $r(t)$ varies in gain, but the spectrum of the transmission is preserved. In a flat fading channel, the reciprocal bandwidth of the transmitted signal is much larger than the multipath time delay spread of the channel, and $h_b(t, \tau)$ can be approximated as having no excess delay (i.e., a single delta function with $\tau = 0$). Flat fading channels are also known as *amplitude varying channels* and are sometimes referred to as *narrowband channels*, since the bandwidth of the applied signal is *narrow* as compared to the channel flat fading bandwidth. Typical flat fading channels cause deep fades, and thus may require 20 or 30 dB more transmitter power to achieve low bit error rates during times of deep fades as

compared to systems operating over non-fading channels. The distribution of the instantaneous gain of flat fading channels is important for designing radio links, and the most common amplitude distribution is the Rayleigh distribution. The Rayleigh flat fading channel model assumes that the channel induces an amplitude which varies in time according to the Rayleigh distribution.

To summarize, a signal undergoes flat fading if

$$B_S \ll B_C \tag{4.41}$$

and

$$T_S \gg \sigma_\tau \tag{4.42}$$

where $T_S$ is the reciprocal bandwidth (e.g., symbol period) and $B_S$ is the bandwidth, respectively, of the transmitted modulation, and $\sigma_\tau$ and $B_C$ are the rms delay spread and coherence bandwidth, respectively, of the channel.

#### 4.5.1.2 Frequency Selective Fading

If the channel possesses a constant-gain and linear phase response over a bandwidth that is smaller than the bandwidth of transmitted signal, then the channel creates *frequency selective fading* on the received signal. Under such conditions the channel impulse response has a multipath delay spread which is greater than the reciprocal bandwidth of the transmitted message waveform. When this occurs, the received signal includes multiple versions of the transmitted waveform which are attenuated (faded) and delayed in time, and hence the received signal is distorted. Frequency selective fading is due to time dispersion of the transmitted symbols within the channel. Thus the channel induces *intersymbol interference* (ISI). Viewed in the frequency domain, certain frequency components in the received signal spectrum have greater gains than others.

Frequency selective fading channels are much more difficult to model than flat fading channels since each multipath signal must be modeled and the channel must be considered to be a linear filter. It is for this reason that wideband multipath measurements are made, and models are developed from these measurements. When analyzing mobile communication systems, statistical impulse response models such as the 2-ray Rayleigh fading model (which considers the impulse response to be made up of two delta functions which independently fade and have sufficient time delay between them to induce frequency selective fading upon the applied signal), or computer generated or measured impulse responses, are generally used for analyzing frequency selective small-scale fading. Figure 4.13 illustrates the characteristics of a frequency selective fading channel.

For frequency selective fading, the spectrum $S(f)$ of the transmitted signal has a bandwidth which is greater than the coherence bandwidth $B_C$ of the channel. Viewed in the frequency domain, the channel becomes frequency selective, where the gain is different for different frequency components. Frequency selec-

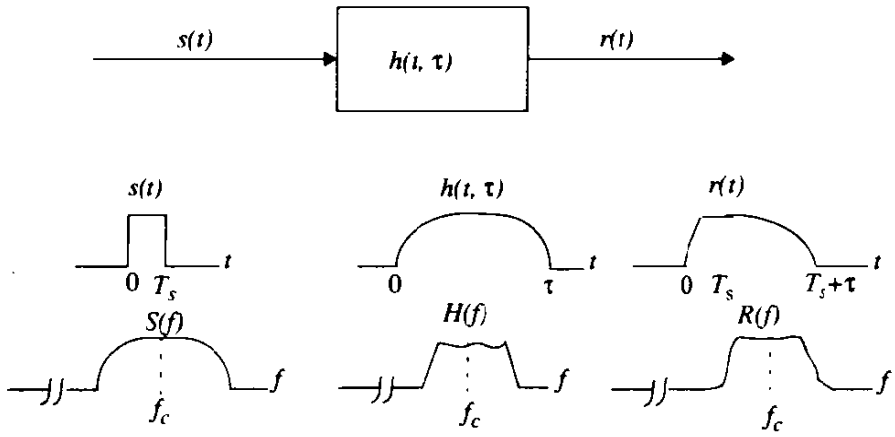



Figure 4.13
Frequency selective fading channel characteristics.

tive fading is caused by multipath delays which approach or exceed the symbol period of the transmitted symbol. Frequency selective fading channels are also known as *wideband channels* since the bandwidth of the signal $s(t)$ is wider than the bandwidth of the channel impulse response. As time varies, the channel varies in gain and phase across the spectrum of *s(t)*, resulting in time varying distortion in the received signal *r(t)*. To summarize, a signal undergoes frequency selective fading if

$$B_S > B_C \tag{4.43}$$

and

$$T_S < \sigma_\tau \tag{4.44}$$

A common rule of thumb is that a channel is frequency selective if $T_s \leq 10\sigma_\tau$, although this is dependent on the specific type of modulation used. Chapter 5 presents simulation results which illustrate the impact of time delay spread on bit error rate (BER).

### 4.5.2 Fading Effects Due to Doppler Spread

#### 4.5.2.1 Fast Fading

Depending on how rapidly the transmitted baseband signal changes as compared to the rate of change of the channel, a channel may be classified either as a *fast fading* or *slow fading* channel. In a *fast fading channel*, the channel impulse response changes rapidly within the symbol duration. That is, the coherence time of the channel is smaller than the symbol period of the transmitted signal. This causes frequency dispersion (also called time selective fading) due to Doppler spreading, which leads to signal distortion. Viewed in the frequency domain, signal distortion due to fast fading increases with increasing Doppler

spread relative to the bandwidth of the transmitted signal. Therefore, a signal undergoes fast fading if

$$T_S > T_C \tag{4.45}$$

and

$$B_S < B_D \tag{4.46}$$

It should be noted that when a channel is specified as a fast or slow fading channel, it does not specify whether the channel is flat fading or frequency selective in nature. Fast fading only deals with the rate of change of the channel due to motion. In the case of the flat fading channel, we can approximate the impulse response to be simply a delta function (no time delay). Hence, a *flat fading, fast fading* channel is a channel in which the amplitude of the delta function varies faster than the rate of change of the transmitted baseband signal. In the case of a *frequency selective, fast fading* channel, the amplitudes, phases, and time delays of any one of the multipath components vary faster than the rate of change of the transmitted signal. In practice, fast fading only occurs for very low data rates.

#### 4.5.2.2 Slow Fading

In a *slow fading channel*, the channel impulse response changes at a rate much slower than the transmitted baseband signal $s(t)$. In this case, the channel may be assumed to be static over one or several reciprocal bandwidth intervals. In the frequency domain, this implies that the Doppler spread of the channel is much less than the bandwidth of the baseband signal. Therefore, a signal undergoes slow fading if

$$T_S \ll T_C \tag{4.47}$$

and

$$B_S \gg B_D \tag{4.48}$$

It should be clear that the velocity of the mobile (or velocity of objects in the channel) and the baseband signaling determines whether a signal undergoes fast fading or slow fading.

The relation between the various multipath parameters and the type of fading experienced by the signal are summarized in Figure 4.14. Over the years, some authors have confused the terms fast and slow fading with the terms large-scale and small-scale fading. It should be emphasized that fast and slow fading deal with the relationship between the time rate of change in the channel and the transmitted signal, and not with propagation path loss models.

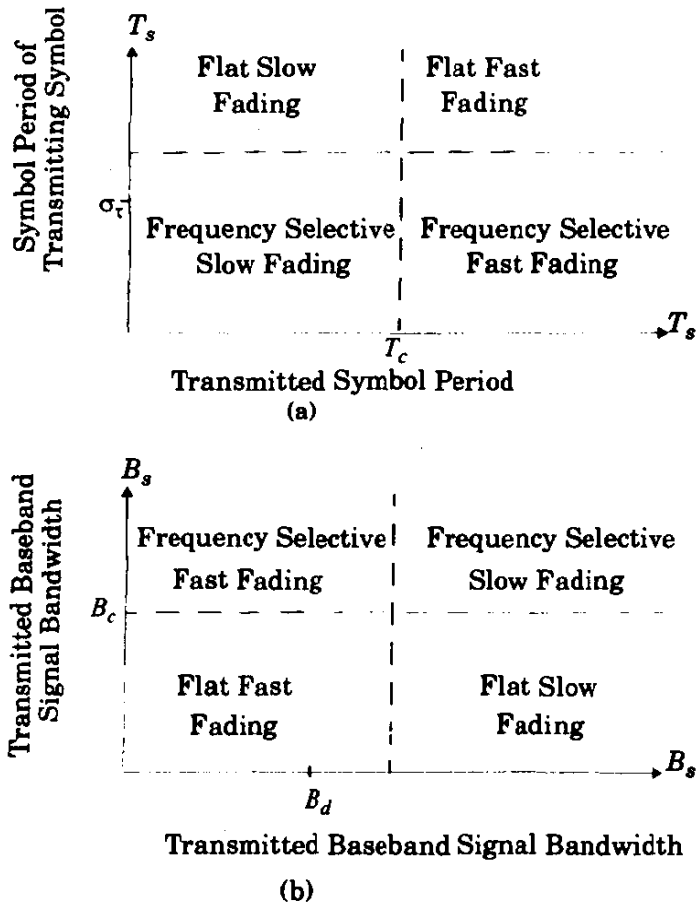



Figure 4.14
Matrix illustrating type of fading experienced by a signal as a function of
(a) symbol period
(b) baseband signal bandwidth.

## 4.6 Rayleigh and Ricean Distributions

### 4.6.1 Rayleigh Fading Distribution

In mobile radio channels, the Rayleigh distribution is commonly used to describe the statistical time varying nature of the received envelope of a flat fading signal, or the envelope of an individual multipath component. It is well known that the envelope of the sum of two quadrature Gaussian noise signals obeys a Rayleigh distribution. Figure 4.15 shows a Rayleigh distributed signal envelope as a function of time. The Rayleigh distribution has a probability density function (pdf) given by

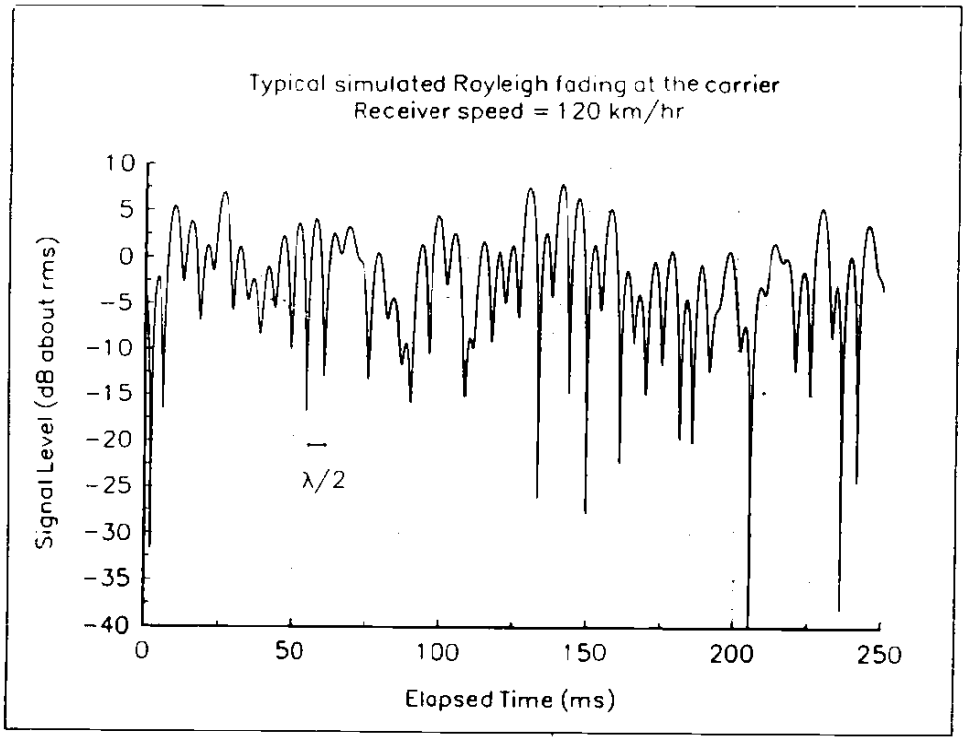



Figure 4.15
A typical Rayleigh fading envelope at 900 MHz [From [Fun93] © IEEE].

$$p(r) = \begin{cases} \dfrac{r}{\sigma^2}\exp\left(-\dfrac{r^2}{2\sigma^2}\right) & (0 \le r \le \infty) \\ 0 & (r < 0) \end{cases} \tag{4.49}$$

where $\sigma$ is the rms value of the received voltage signal before *envelope detection*, and $\sigma^2$ is the time-average power of the received signal *before* envelope detection. The probability that the envelope of the received signal does not exceed a specified value $R$ is given by the corresponding cumulative distribution function (CDF)

$$P(R) = Pr(r \le R) = \int_0^R p(r)\,dr = 1 - \exp\left(-\frac{R^2}{2\sigma^2}\right) \tag{4.50}$$

The mean value $r_{mean}$ of the Rayleigh distribution is given by

$$r_{mean} = E[r] = \int_0^\infty r p(r)\,dr = \sigma\sqrt{\frac{\pi}{2}} = 1.2533\sigma \tag{4.51}$$

and the variance of the Rayleigh distribution is given by $\sigma_r^2$, which represents the ac power in the signal envelope

$$\sigma_r^2 = E[r^2] - E^2[r] = \int_0^\infty r^2 p(r)\,dr - \frac{\sigma^2\pi}{2} \tag{4.52}$$

$$= \sigma^2\left(2 - \frac{\pi}{2}\right) = 0.4292\sigma^2$$

The rms value of the envelope is the square root of the mean square, or $\sqrt{2}\sigma$.

The median value of $r$ is found by solving

$$\frac{1}{2} = \int_0^{r_{median}} p(r)\,dr \tag{4.53}$$

and is

$$r_{median} = 1.177\sigma \tag{4.54}$$

Thus the mean and the median differ by only 0.55 dB in a Rayleigh fading signal. Note that the median is often used in practice, since fading data are usually measured in the field and a particular distribution cannot be assumed. By using median values instead of mean values it is easy to compare different fading distributions which may have widely varying means. Figure 4.16 illustrates the Rayleigh pdf. The corresponding Rayleigh cumulative distribution function (CDF) is shown in Figure 4.17.

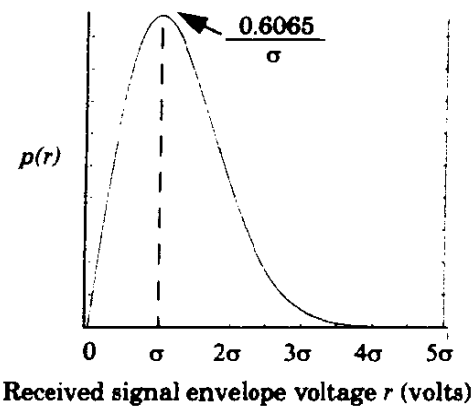



Figure 4.16
Rayleigh probability density function (pdf).

### 4.6.2 Ricean Fading Distribution

When there is a dominant stationary (nonfading) signal component present, such as a line-of-sight propagation path, the small-scale fading envelope

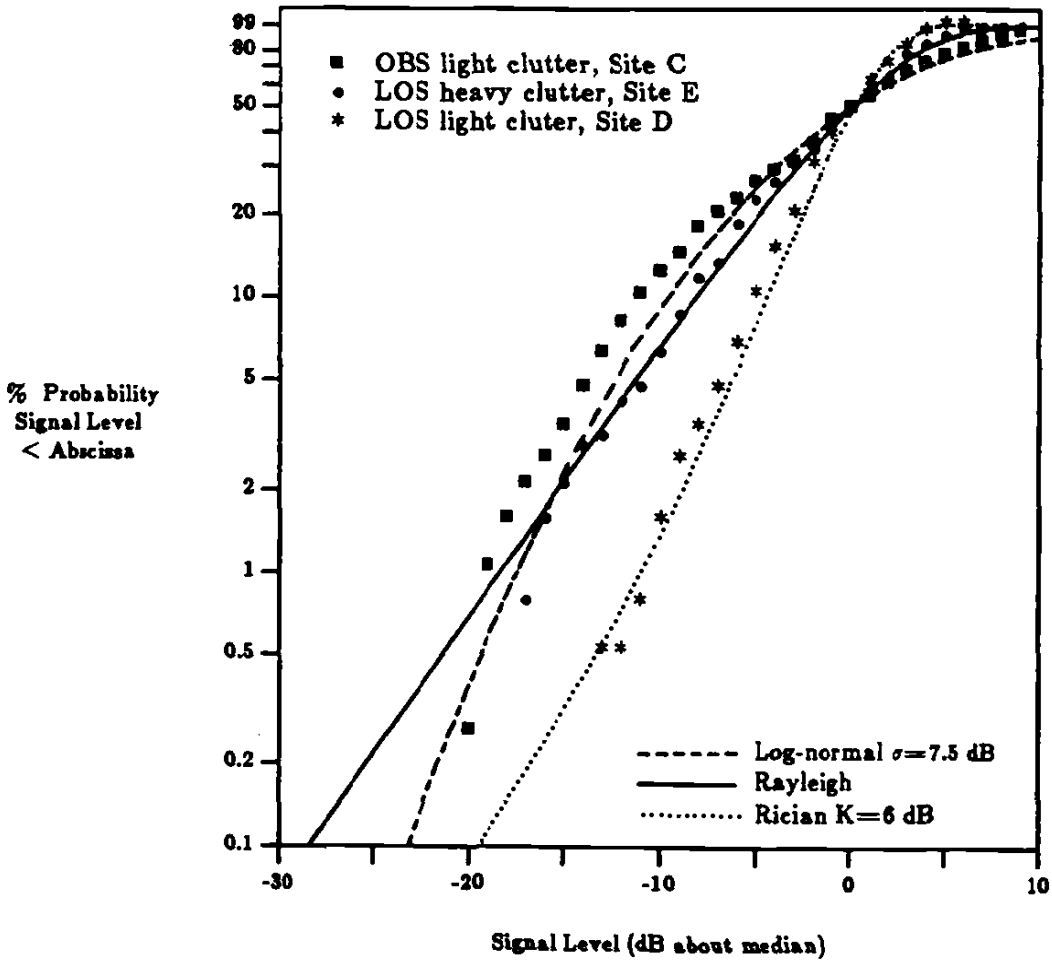



Figure 4.17
Cumulative distribution for three small-scale fading measurements and their fit to Rayleigh, Ricean, and log-normal distributions [From [Rap89] © IEEE].

distribution is Ricean. In such a situation, random multipath components arriving at different angles are superimposed on a stationary dominant signal. At the output of an envelope detector, this has the effect of adding a dc component to the random multipath.

Just as for the case of detection of a sine wave in thermal noise [Ric48], the effect of a dominant signal arriving with many weaker multipath signals gives rise to the Ricean distribution. As the dominant signal becomes weaker, the composite signal resembles a noise signal which has an envelope that is Rayleigh. Thus, the Ricean distribution degenerates to a Rayleigh distribution when the dominant component fades away.

The Ricean distribution is given by

$$p(r) = \begin{cases} \dfrac{r}{\sigma^2} e^{-\frac{(r^2+A^2)}{2\sigma^2}} I_0\left(\dfrac{Ar}{\sigma^2}\right) & \text{for } (A \geq 0, r \geq 0) \\ 0 & \text{for } (r < 0) \end{cases} \tag{4.55}$$

The parameter $A$ denotes the peak amplitude of the dominant signal and $I_0(\cdot)$ is the modified Bessel function of the first kind and zero-order. The Ricean distribution is often described in terms of a parameter $K$ which is defined as the ratio between the deterministic signal power and the variance of the multipath. It is given by $K = A^2/(2\sigma^2)$ or, in terms of dB

$$K(\text{dB}) = 10\log\frac{A^2}{2\sigma^2} \text{ dB} \tag{4.56}$$

The parameter $K$ is known as the Ricean factor and completely specifies the Ricean distribution. As $A \to 0, K \to -\infty$ dB, and as the dominant path decreases in amplitude, the Ricean distribution degenerates to a Rayleigh distribution. Figure 4.18 shows the Ricean pdf. The Ricean CDF is compared with the Rayleigh CDF in Figure 4.17.

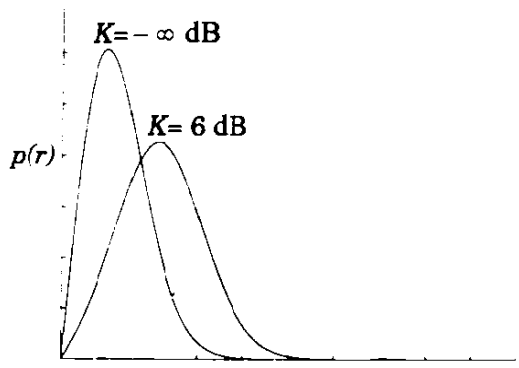



Figure 4.18
Probability density function of Ricean distributions: $K = -\infty$ dB (Rayleigh) and $K = 6$ dB. For K>>1, the Ricean pdf is approximately Gaussian about the mean.

## 4.7 Statistical Models for Multipath Fading Channels

Several multipath models have been suggested to explain the observed statistical nature of a mobile channel. The first model presented by Ossana [Oss64] was based on interference of waves incident and reflected from the flat sides of randomly located buildings. Although Ossana's model [Oss64] predicts flat fading power spectra that were in agreement with measurements in suburban

CHAPTER 8

# Multiple Access Techniques for Wireless Communications

Multiple access schemes are used to allow many mobile users to share simultaneously a finite amount of radio spectrum. The sharing of spectrum is required to achieve high capacity by simultaneously allocating the available bandwidth (or the available amount of channels) to multiple users. For high quality communications, this must be done without severe degradation in the performance of the system.

## 8.1 Introduction

In wireless communications systems, it is often desirable to allow the subscriber to send simultaneously information to the base station while receiving information from the base station. For example, in conventional telephone systems, it is possible to talk and listen simultaneously, and this effect, called *duplexing*, is generally required in wireless telephone systems. Duplexing may be done using frequency or time domain techniques. *Frequency division duplexing* (FDD) provides two distinct bands of frequencies for every user. The forward band provides traffic from the base station to the mobile, and the reverse band provides traffic from the mobile to the base. In FDD, any *duplex channel* actually consists of two simplex channels, and a device called a *duplexer* is used inside each subscriber unit and base station to allow simultaneous radio transmission and reception on the duplex channel pair. The frequency split between the forward and reverse channel is constant throughout the system, regardless of the particular channel being used. *Time division duplexing* (TDD) uses time instead of frequency to provide both a forward and reverse link. If the time split between the forward and reverse time slot is small, then the transmission and reception of data appears simultaneous to the user. Figure 8.1 illustrates FDD and TDD

techniques. TDD allows communication on a single channel (as opposed to requiring two simplex or dedicated channels) and simplifies the subscriber equipment since a duplexer is not required.

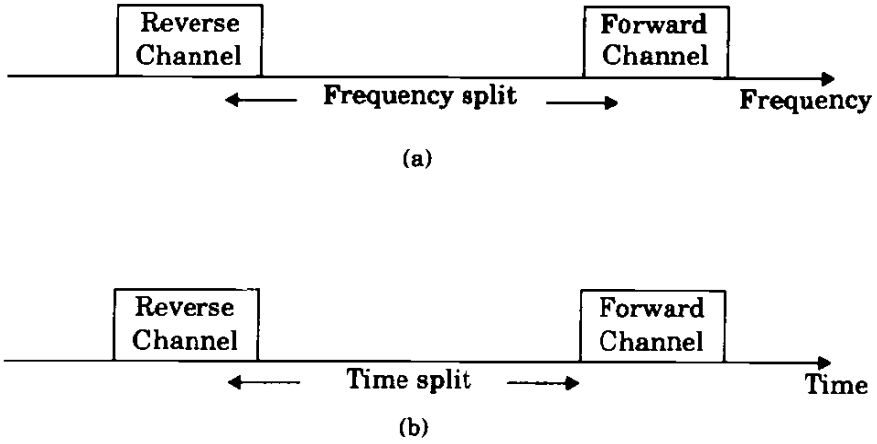



Figure 8.1
(a) FDD provides two simplex channels at the same time.
(b) TDD provides two simplex time slots on the same frequency.

There are several trade-offs between FDD and TDD approaches. FDD is geared toward radio communications systems that provide individual radio frequencies for each user. Because each transceiver simultaneously transmits and receives radio signals which vary by more than 100 dB, the frequency allocation used for the forward and reverse channels must be carefully coordinated with out-of-band users that occupy spectrum between these two bands. Furthermore, the frequency separation must be coordinated to permit the use of inexpensive RF technology. TDD enables each transceiver to operate as either a transmitter or receiver on the same frequency, and eliminates the need for separate forward and reverse frequency bands. However, there is a time latency due to the fact that communications is not full duplex in the truest sense.



### 8.1.1 Introduction to Multiple Access

Frequency division multiple access (FDMA), *time division multiple access* (TDMA), and *code division multiple access* (CDMA) are the three major access techniques used to share the available bandwidth in a wireless communication system. These techniques can be grouped as *narrowband* and *wideband* systems, depending upon how the available bandwidth is allocated to the users. The duplexing technique of a multiple access system is usually described along with the particular multiple access scheme, as shown in the examples below.

**Narrowband Systems** — The term *narrowband* is used to relate the bandwidth of a single channel to the expected coherence bandwidth of the channel. In a narrowband multiple access system, the available radio spectrum is

divided into a large number of narrowband channels. The channels are usually operated using FDD. To minimize interference between forward and reverse links on each channel, the frequency split is made as great as possible within the frequency spectrum, while still allowing inexpensive duplexers and a common transceiver antenna to be used in each subscriber unit. In narrowband FDMA, a user is assigned a particular channel which is not shared by other users in the vicinity, and if FDD is used (that is, each channel has a forward and reverse link), then the system is called FDMA/FDD. Narrowband TDMA, on the other hand, allows users to share the same channel but allocates a unique time slot to each user in a cyclical fashion on the channel, thus separating a small number of users in time on a single channel. For narrowband TDMA, there generally are a large number of channels allocated using either FDD or TDD, and each channel is shared using TDMA. Such systems are called TDMA/FDD or TDMA/TDD access systems.

**Wideband systems** — In wideband systems, the transmission bandwidth of a single channel is much larger than the coherence bandwidth of the channel. Thus, multipath fading does not greatly affect the received signal within a wideband channel, and frequency selective fades occur in only a small fraction of the signal bandwidth. In wideband multiple access systems, the users are allowed to transmit in a large part of the spectrum. A large number of transmitters are also allowed to transmit on the same channel. TDMA allocates time slots to the many transmitters on the same channel and allows only one transmitter to access the channel at any instant of time, whereas spread spectrum CDMA allows all of the transmitters to access the channel at the same time. TDMA and CDMA systems may use either FDD or TDD multiplexing techniques.

In addition to FDMA, TDMA, and CDMA, two other multiple access schemes are used for wireless communications. These are *packet radio* (PR) and *space division multiple access* (SDMA). In this chapter, the above mentioned multiple access techniques, their performance, and their capacity in digital cellular systems are discussed. Table 8.1 shows the different multiple access techniques being used in various wireless communications systems.

## 8.2 Frequency Division Multiple Access (FDMA)

Frequency division multiple access (FDMA) assigns individual channels to individual users. It can be seen from Figure 8.2 that each user is allocated a unique frequency band or channel. These channels are assigned on demand to users who request service. During the period of the call, no other user can share the same frequency band. In FDD systems, the users are assigned a channel as a pair of frequencies; one frequency is used for the forward channel, while the other frequency is used for the reverse channel. The features of FDMA are as follows:

- The FDMA channel carries only one phone circuit at a time.

**Table 8.1 Multiple Access Techniques Used in Different Wireless Communication Systems**

| Cellular System | Multiple Access Technique |
|---|---|
| Advanced Mobile Phone System (AMPS) | FDMA/FDD |
| Global System for Mobile (GSM) | TDMA/FDD |
| U.S. Digital Cellular (USDC) | TDMA/FDD |
| Japanese Digital Cellular (JDC) | TDMA/FDD |
| CT2 (Cordless Telephone) | FDMA/TDD |
| Digital European Cordless Telephone (DECT) | FDMA/TDD |
| U.S. Narrowband Spread Spectrum (IS-95) | CDMA/FDD |

- If an FDMA channel is not in use, then it sits idle and cannot be used by other users to increase or share capacity. It is essentially a wasted resource.
- After the assignment of a voice channel, the base station and the mobile transmit simultaneously and continuously.
- The bandwidths of FDMA channels are relatively narrow (30 kHz) as each channel supports only one circuit per carrier. That is, FDMA is usually implemented in narrowband systems.
- The symbol time is large as compared to the average delay spread. This implies that the amount of intersymbol interference is low and, thus, little or no equalization is required in FDMA narrowband systems.
- The complexity of FDMA mobile systems is lower when compared to TDMA systems, though this is changing as digital signal processing methods improve for TDMA.
- Since FDMA is a continuous transmission scheme, fewer bits are needed for overhead purposes (such as synchronization and framing bits) as compared to TDMA.
- FDMA systems have higher cell site system costs as compared to TDMA systems, because of the single channel per carrier design, and the need to use costly bandpass filters to eliminate spurious radiation at the base station.
- The FDMA mobile unit uses duplexers since both the transmitter and receiver operate at the same time. This results in an increase in the cost of FDMA subscriber units and base stations.
- FDMA requires tight RF filtering to minimize adjacent channel interference.

**Nonlinear Effects in FDMA** — In a FDMA system, many channels share the same antenna at the base station. The power amplifiers or the power combiners, when operated at or near saturation for maximum power efficiency, are nonlinear. The nonlinearities cause signal spreading in the frequency domain and generate *intermodulation* (IM) frequencies. IM is undesired RF radiation which can interfere with other channels in the FDMA systems. Spreading of the spec-

trum results in adjacent-channel interference. Intermodulation is the generation of undesirable harmonics. Harmonics generated outside the mobile radio band cause interference to adjacent services, while those present inside the band cause interference to other users in the mobile system [Yac93].

**Example 8.1**

Find the intermodulation frequencies generated if a base station transmits two carrier frequencies at 1930 MHz and 1932 MHz that are amplified by a saturated clipping amplifier. If the mobile radio band is allocated from 1920 MHz to 1940 MHz, designate the IM frequencies that lie inside and outside the band.

**Solution to Example 8.1**

Intermodulation distortion products occur at frequencies $mf_1 + nf_2$ for all integer values of m and n, i.e., $-\infty < m, n < \infty$. Some of the possible intermodulation frequencies that are produced by a nonlinear device are

$(2n + 1)f_1 - 2nf_2$, $(2n + 2)f_1 - (2n + 1)f_2$, $(2n + 1)f_1 - 2nf_2$,
$(2n + 2)f_2 - (2n + 1)f_1$, etc. for $n = 0, 1, 2, \ldots\ldots$

Table E8.1 lists several intermodulation product terms.

Table E 8.1: Intermodulation Products

| $n = 0$ | $n = 1$ | $n = 2$ | $n = 3$ |
|---|---|---|---|
| 1930 | 1926 | 1922 | 1918 |
| 1928 | 1924 | 1920 | 1916 |
| 1932 | 1936 | 1940 | 1944* |
| 1934 | 1938 | 1942* | 1946* |

The frequencies in the table marked with an asterisk (*) are the frequencies that lie outside the mobile radio band.

The first U.S. analog cellular system, the *Advanced Mobile Phone System* (AMPS), is based on FDMA/FDD. A single user occupies a single channel while the call is in progress, and the single channel is actually two simplex channels which are frequency duplexed with a 45 MHz split. When a call is completed, or when a handoff occurs, the channel is vacated so that another mobile subscriber may use it. Multiple or simultaneous users are accommodated in AMPS by giving each user a unique channel. Voice signals are sent on the forward channel from the base station to mobile unit, and on the reverse channel from the mobile unit to the base station. In AMPS, analog narrowband frequency modulation (NBFM) is used to modulate the carrier. The number of channels that can be simultaneously supported in a FDMA system is given by

$$N = \frac{B_t - 2B_{guard}}{B_c} \tag{8.1}$$

where $B_t$ is the total spectrum allocation, $B_{guard}$ is the guard band allocated at the edge of the allocated spectrum, and $B_c$ is the channel bandwidth.

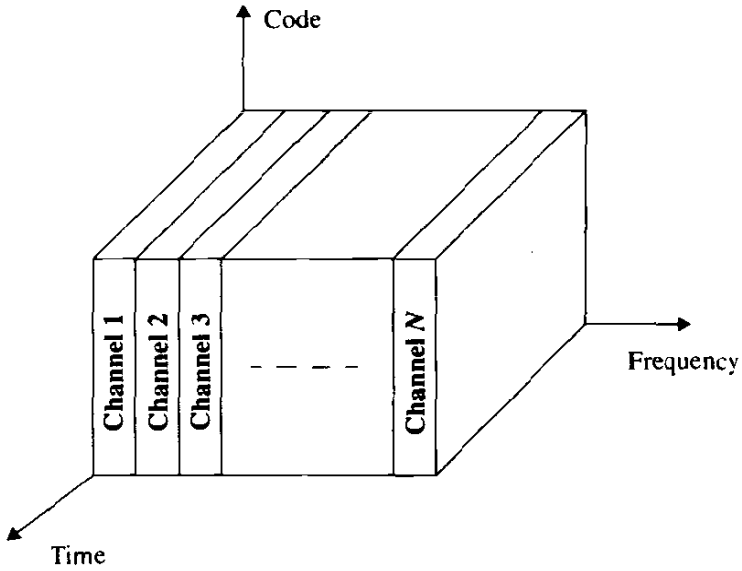



Figure 8.2
FDMA where different channels are assigned different frequency bands.

**Example 8.2**

If $B_t$ is 12.5 MHz, $B_{guard}$ is 10 kHz, and $B_c$ is 30 kHz, find the number of channels available in an FDMA system.

**Solution to Example 8.2**

The number of channels available in the FDMA system is given as

$$N = \frac{12.5 \times 10^6 - 2(10 \times 10^3)}{30 \times 10^3} = 416$$

In the U.S., each cellular carrier is allocated 416 channels.

## 8.3 Time Division Multiple Access (TDMA)

*Time division multiple access* (TDMA) systems divide the radio spectrum into time slots, and in each slot only one user is allowed to either transmit or receive. It can be seen from Figure 8.3 that each user occupies a cyclically repeating time slot, so a channel may be thought of as particular time slot that reoccurs every frame, where $N$ time slots comprise a frame. TDMA systems transmit data in a buffer-and-burst method, thus the transmission for any user is noncontinuous. This implies that, unlike in FDMA systems which accommodate analog FM, digital data and digital modulation must be used with TDMA. The transmission from various users is interlaced into a repeating frame structure as shown in Figure 8.4. It can be seen that a frame consists of a number of slots. Each frame is made up of a preamble, an information message, and tail bits. In TDMA/TDD, half of the time slots in the frame information message would be used for

the forward link channels and half would be used for reverse link channels. In TDMA/FDD systems, an identical or similar frame structure would be used solely for either forward or reverse transmission, but the carrier frequencies would be different for the forward and reverse links. In general, TDMA/FDD systems intentionally induce several time slots of delay between the forward and reverse time slots of a particular user, so that duplexers are not required in the subscriber unit.

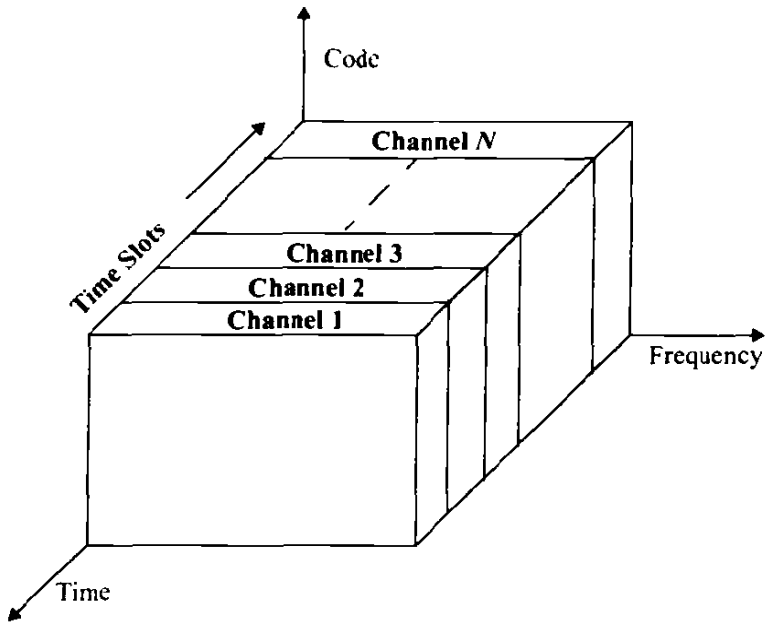



Figure 8.3
TDMA scheme where each channel occupies a cyclically repeating time slot.

In a TDMA frame, the preamble contains the address and synchronization information that both the base station and the subscribers use to identify each other. Guard times are utilized to allow synchronization of the receivers between different slots and frames. Different TDMA wireless standards have different TDMA frame structures, and some are described in Chapter 10. The features of TDMA include the following:

- TDMA shares a single carrier frequency with several users, where each user makes use of nonoverlapping time slots. The number of time slots per frame depends on several factors, such as modulation technique, available bandwidth, etc.
- Data transmission for users of a TDMA system is not continuous, but occurs in bursts. This results in low battery consumption, since the subscriber transmitter can be turned off when not in use (which is most of the time).
- Because of discontinuous transmissions in TDMA, the handoff process is much simpler for a subscriber unit, since it is able to listen for other base stations during idle time slots. An enhanced link control, such as that provided by *mobile assisted handoff* (MAHO) can be carried out by a subscriber

by listening on an idle slot in the TDMA frame.

- TDMA uses different time slots for transmission and reception, thus duplexers are not required. Even if FDD is used, a switch rather than a duplexer inside the subscriber unit is all that is required to switch between transmitter and receiver using TDMA.
- Adaptive equalization is usually necessary in TDMA systems, since the transmission rates are generally very high as compared to FDMA channels.
- In TDMA, the guard time should be minimized. If the transmitted signal at the edges of a time slot are suppressed sharply in order to shorten the guard time, the transmitted spectrum will expand and cause interference to adjacent channels.
- High synchronization overhead is required in TDMA systems because of burst transmissions. TDMA transmissions are slotted, and this requires the receivers to be synchronized for each data burst. In addition, guard slots are necessary to separate users, and this results in the TDMA systems having larger overheads as compared to FDMA.
- TDMA has an advantage in that it is possible to allocate different numbers of time slots per frame to different users. Thus bandwidth can be supplied on demand to different users by concatenating or reassigning time slots based on priority.

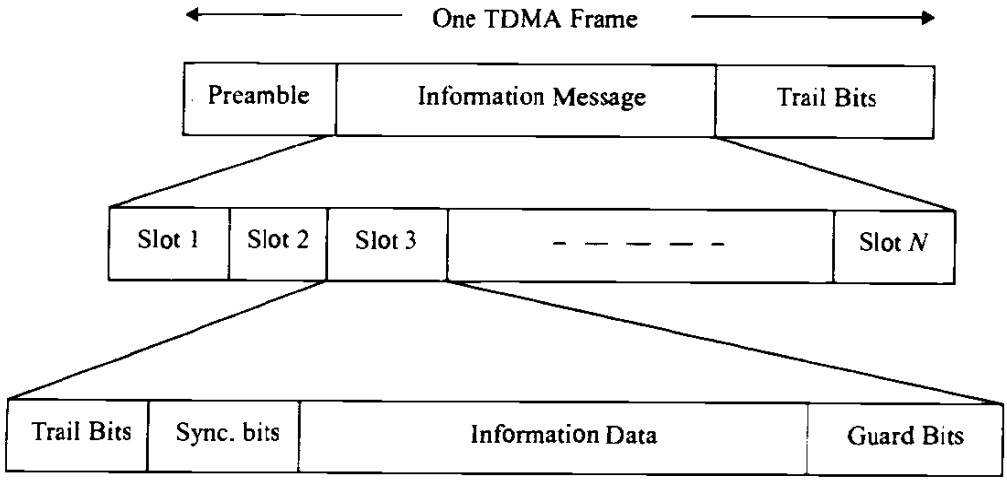



Figure 8.4
TDMA frame structure.

***Efficiency of TDMA*** — The efficiency of a TDMA system is a measure of the percentage of transmitted data that contains information as opposed to providing overhead for the access scheme. The frame efficiency, $\eta_f$, is the percentage of bits per frame which contain transmitted data. Note that the transmitted data may include source and channel coding bits, so the raw end-user efficiency of a system is generally less than $\eta_f$. The frame efficiency can be found as follows.

The number of overhead bits per frame is [Zie92],

$$b_{OH} = N_r b_r + N_t b_p + N_t b_g + N_r b_g \tag{8.2}$$

where, $N_r$ is the number of reference bursts per frame, $N_t$ is the number of traffic bursts per frame, $b_r$ is the number of overhead bits per reference burst, $b_p$ is the number of overhead bits per preamble in each slot, and $b_g$ is the number of equivalent bits in each guard time interval. The total number of bits per frame, $b_T$, is

$$b_T = T_f R \tag{8.3}$$

where $T_f$ is the frame duration, and $R$ is the channel bit rate. The frame efficiency $\eta_f$ is thus given as

$$\eta_f = \left(1 - \frac{b_{OH}}{b_T}\right) \times 100\% \tag{8.4}$$

***Number of channels in TDMA system*** — The number of TDMA channel slots that can be provided in a TDMA system is found by multiplying the number of TDMA slots per channel by the number of channels available and is given by

$$N = \frac{m\,(B_{tot} - 2B_{guard})}{B_c} \tag{8.5}$$

where $m$ is the maximum number of TDMA users supported on each radio channel. Note that two guard bands, one at the low end of the allocated frequency band and one at the high end, are required to ensure that users at the edge of the band do not "bleed over" into an adjacent radio service.

**Example 8.3**

Consider Global System for Mobile, which is a TDMA/FDD system that uses 25 MHz for the forward link, which is broken into radio channels of 200 kHz. If 8 speech channels are supported on a single radio channel, and if no guard band is assumed, find the number of simultaneous users that can be accommodated in GSM.

**Solution to Example 8.3**

The number of simultaneous users that can be accommodated in GSM is given as

$$N = \frac{25\text{ MHz}}{(200\text{ kHz})/8} = 1000$$

Thus, GSM can accommodate 1000 simultaneous users.

**Example 8.4**

If GSM uses a frame structure where each frame consists of 8 time slots, and each time slot contains 156.25 bits, and data is transmitted at 270.833 kbps in the channel, find (a) the time duration of a bit, (b) the time duration of a slot, (c) the time duration of a frame, and (d) how long must a user occupying a single time slot must wait between two simultaneous transmissions.

**Solution to Example 8.4**

(a) The time duration of a bit, $T_b = \frac{1}{270.833 \text{ kbps}} = 3.692 \ \mu\text{s}$.

(b) The time duration of a slot, $T_{slot} = 156.25 \times T_b = 0.577$ ms.

(c) The time duration of a frame, $T_f = 8 \times T_{slot} = 4.615$ ms.

(d) A user has to wait 4.615 ms, the arrival time of a new frame, for its next transmission.

**Example 8.5**

If a normal GSM time slot consists of 6 trailing bits, 8.25 guard bits, 26 training bits, and 2 traffic bursts of 58 bits of data, find the frame efficiency.

**Solution to Example 8.5**

A time slot has $6 + 8.25 + 26 + 2(58) = 156.25$ bits.
A frame has $8 \times 156.25 = 1250$ bits/frame.
The number of overhead bits per frame is given by

$$b_{OH} = 8(6) + 8(8.25) + 8(26) = 322 \text{ bits}$$

Thus, the frame efficiency

$$\eta_f = \left[1 - \frac{322}{1250}\right] \times 100 = 74.24\ \%$$

## 8.4 Spread Spectrum Multiple Access

*Spread spectrum multiple access* (SSMA) uses signals which have a transmission bandwidth that is several orders of magnitude greater than the minimum required RF bandwidth. A pseudo-noise (PN) sequence (discussed in Chapter 5) converts a narrowband signal to a wideband noise-like signal before transmission. SSMA also provides immunity to multipath interference and robust multiple access capability. SSMA is not very bandwidth efficient when used by a single user. However, since many users can share the same spread spectrum bandwidth without interfering with one another, spread spectrum systems become bandwidth efficient in a multiple user environment. It is exactly this situation that is of interest to wireless system designers. There are two main types of spread spectrum multiple access techniques; *frequency hopped multiple access* (FH) and *direct sequence multiple access* (DS). Direct sequence multiple access is also called *code division multiple access* (CDMA).

### 8.4.1 Frequency Hopped Multiple Access (FHMA)

*Frequency hopped multiple access* (FHMA) is a digital multiple access system in which the carrier frequencies of the individual users are varied in a pseudorandom fashion within a wideband channel. The digital data is broken into uniform sized bursts which are transmitted on different carrier frequencies. The

instantaneous bandwidth of any one transmission burst is much smaller than the total spread bandwidth. The pseudorandom change of the carrier frequencies of the user randomizes the occupancy of a specific channel at any given time, thereby allowing for multiple access over a wide range of frequencies. In the FH receiver, a locally generated PN code is used to synchronize the receivers instantaneous frequency with that of the transmitter. At any given point in time, a frequency hopped signal only occupies a single, relatively narrow channel since narrowband FM or FSK is used. The difference between FHMA and a traditional FDMA system is that the frequency hopped signal changes channels at rapid intervals. If the rate of change of the carrier frequency is greater than the symbol rate then the system is referred to as a *fast frequency hopping system*. If the channel changes at a rate less than or equal to the symbol rate, it is called *slow frequency hopping*. A fast frequency hopper may thus be thought of as an FDMA system which employs frequency diversity. FHMA systems often employ energy efficient constant envelope modulation. Inexpensive receivers may be built to provide noncoherent detection of FHMA. This implies that linearity is not an issue, and the power of multiple users at the receiver does not degrade FHMA performance.

A frequency hopped system provides a level of security, especially when a large number of channels are used, since an unintended (or an intercepting) receiver that does not know the pseudorandom sequence of frequency slots must retune rapidly to search for the signal it wishes to intercept. In addition, the FH signal is somewhat immune to fading, since error control coding and interleaving can be used to protect the frequency hopped signal against deep fades which may occasionally occur during the hopping sequence. Error control coding and interleaving can also be combined to guard against *erasures* which can occur when two or more users transmit on the same channel at the same time.

### 8.4.2 Code Division Multiple Access (CDMA)

In *code division multiple access* (CDMA) systems, the narrowband message signal is multiplied by a very large bandwidth signal called the *spreading signal*. The spreading signal is a pseudo-noise code sequence that has a chip rate which is orders of magnitudes greater than the data rate of the message. All users in a CDMA system, as seen from Figure 8.5, use the same carrier frequency and may transmit simultaneously. Each user has its own pseudorandom codeword which is approximately orthogonal to all other codewords. The receiver performs a time correlation operation to detect only the specific desired codeword. All other codewords appear as noise due to decorrelation. For detection of the message signal, the receiver needs to know the codeword used by the transmitter. Each user operates independently with no knowledge of the other users.

In CDMA, the power of multiple users at a receiver determines the noise floor after decorrelation. If the power of each user within a cell is not controlled

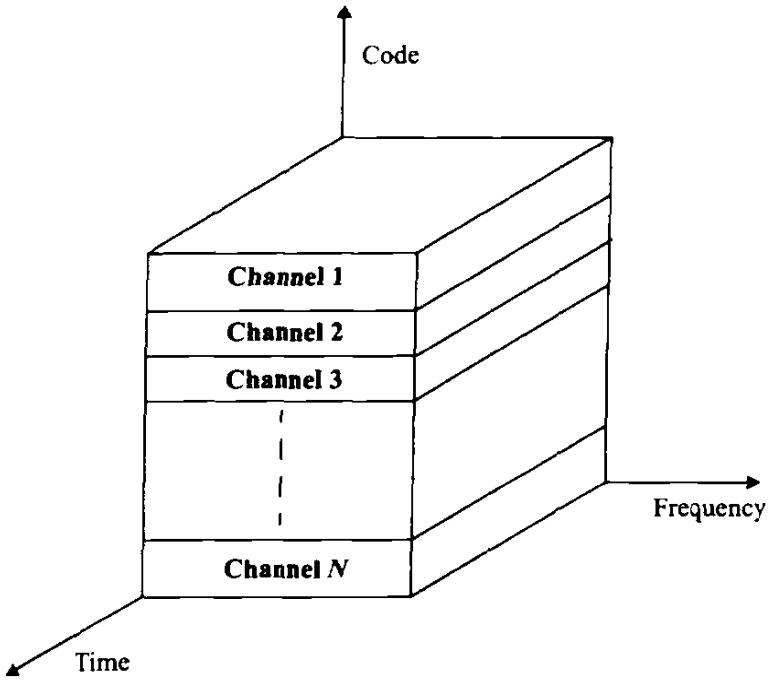



Figure 8.5
CDMA in which each channel is assigned a unique PN code which is orthogonal to PN codes used by other users.

such that they do not appear equal at the base station receiver, then the *near-far problem* occurs.

The near-far problem occurs when many mobile users share the same channel. In general, the strongest received mobile signal will *capture* the demodulator at a base station. In CDMA, stronger received signal levels raise the noise floor at the base station demodulators for the weaker signals, thereby decreasing the probability that weaker signals will be received. To combat the near-far problem, *power control* is used in most CDMA implementations. Power control is provided by each base station in a cellular system and assures that each mobile within the base station coverage area provides the same signal level to the base station receiver. This solves the problem of a nearby subscriber overpowering the base station receiver and drowning out the signals of far away subscribers. Power control is implemented at the base station by rapidly sampling the radio signal strength indicator (RSSI) levels of each mobile and then sending a power change command over the forward radio link. Despite the use of power control within each cell, out-of-cell mobiles provide interference which is not under the control of the receiving base station. The features of CDMA including the following:

- Many users of a CDMA system share the same frequency. Either TDD or FDD may be used.
- Unlike TDMA or FDMA, CDMA has a soft capacity limit. Increasing the

number of users in a CDMA system raises the noise floor in a linear manner. Thus, there is no absolute limit on the number of users in CDMA. Rather, the system performance gradually degrades for all users as the number of users is increased, and improves as the number of users is decreased.

- Multipath fading may be substantially reduced because the signal is spread over a large spectrum. If the spread spectrum bandwidth is greater than the coherence bandwidth of the channel, the inherent frequency diversity will mitigate the effects of small-scale fading.
- Channel data rates are very high in CDMA systems. Consequently, the symbol (chip) duration is very short and usually much less than the channel delay spread. Since PN sequences have low autocorrelation, multipath which is delayed by more than a chip will appear as noise. A RAKE receiver can be used to improve reception by collecting time delayed versions of the required signal.
- Since CDMA uses co-channel cells, it can use macroscopic spatial diversity to provide soft handoff. Soft handoff is performed by the MSC, which can simultaneously monitor a particular user from two or more base stations. The MSC may chose the best version of the signal at any time without switching frequencies.
- Self-jamming is a problem in CDMA system. Self-jamming arises from the fact that the spreading sequences of different users are not exactly orthogonal, hence in the despreading of a particular PN code, non-zero contributions to the receiver decision statistic for a desired user arise from the transmissions of other users in the system.
- The near-far problem occurs at a CDMA receiver if an undesired user has a high detected power as compared to the desired user.

### 8.4.3 Hybrid Spread Spectrum Techniques

In addition to the frequency hopped and direct sequence, spread spectrum multiple access techniques, there are certain other hybrid combinations that provide certain advantages. These hybrid techniques are described below.

**Hybrid FDMA/CDMA (FCDMA)** — This technique can be used as an alternative to the DS-CDMA techniques discussed above. Figure 8.6 shows the spectrum of this hybrid scheme. The available wideband spectrum is divided into a number of subspectras with smaller bandwidths. Each of these smaller subchannels becomes a narrowband CDMA system having processing gain lower than the original CDMA system. This hybrid system has an advantage in that the required bandwidth need not be contiguous and different users can be allotted different subspectrum bandwidths depending on their requirements. The capacity of this FDMA/CDMA technique is calculated as the sum of the capacities of a system operating in the subspectra [Eng93].

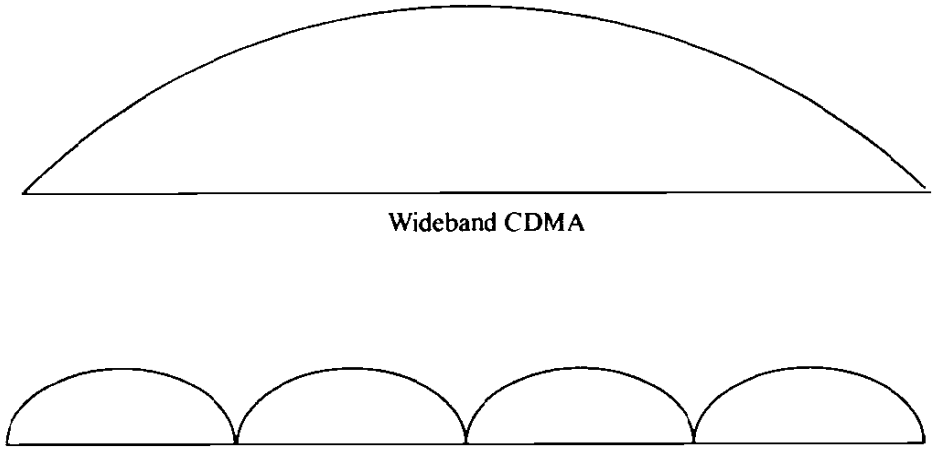



Figure 8.6
Spectrum of wideband CDMA compared to the spectrum of a hybrid, frequency division, direct sequence multiple access.

**Hybrid Direct Sequence/Frequency Hopped Multiple Access (DS/FHMA)** — This technique consists of a direct sequence modulated signal whose center frequency is made to hop periodically in a pseudorandom fashion. Figure 8.7 shows the frequency spectrum of such a signal [Dix94]. Direct sequence, frequency hopped systems have an advantage in that they avoid the near-far effect. However, frequency hopped CDMA systems are not adaptable to the soft handoff process since it is difficult to synchronize the frequency hopped base station receiver to the multiple hopped signals.

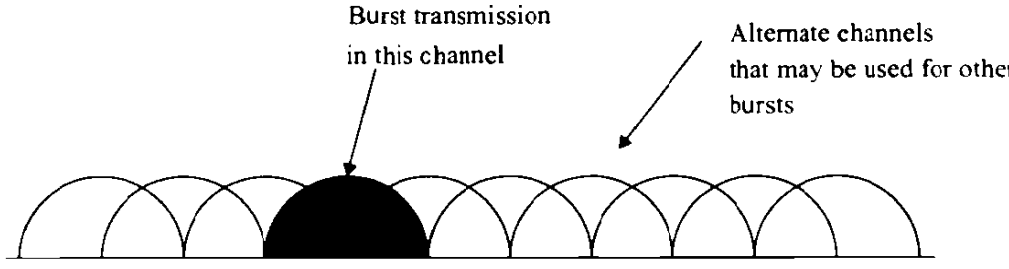



Figure 8.7
Frequency spectrum of a hybrid FH/DS system.

**Time Division CDMA (TCDMA)** — In a TCDMA (also called TDMA/CDMA) system, different spreading codes are assigned to different cells. Within each cell, only one user per cell is allotted a particular time slot. Thus at any time, only one CDMA user is transmitting in each cell. When a handoff takes place, the spreading code of the user is changed to that of the new cell. Using

TCDMA has an advantage in that it avoids the near-far effect since only one user transmits at a time within a cell.

**Time Division Frequency Hopping (TDFH)** — This multiple access technique has an advantage in severe multipath or when severe co-channel interference occurs. The subscriber can hop to a new frequency at the start of a new TDMA frame, thus avoiding a severe fade or erasure event on a particular channel. This technique has been adopted for the GSM standard, where the hopping sequence is predefined and the subscriber is allowed to hop only on certain frequencies which are assigned to a cell. This scheme also avoids co-channel interference problems between neighboring cells if two interfering base station transmitters are made to transmit on different frequencies at different times. The use of TDFH can increase the capacity of GSM by several fold [Gud92]. Chapter 10 describes the GSM standard in more detail.

## 8.5 Space Division Multiple Access (SDMA)

*Space division multiple access* (SDMA) controls the radiated energy for each user in space. It can be seen from Figure 8.8 that SDMA serves different users by using spot beam antennas. These different areas covered by the antenna beam may be served by the same frequency (in a TDMA or CDMA system) or different frequencies (in an FDMA system). Sectorized antennas may be thought of as a primitive application of SDMA. In the future, adaptive antennas will likely be used to simultaneously steer energy in the direction of many users at once and appear to be best suited for TDMA and CDMA base station architectures.

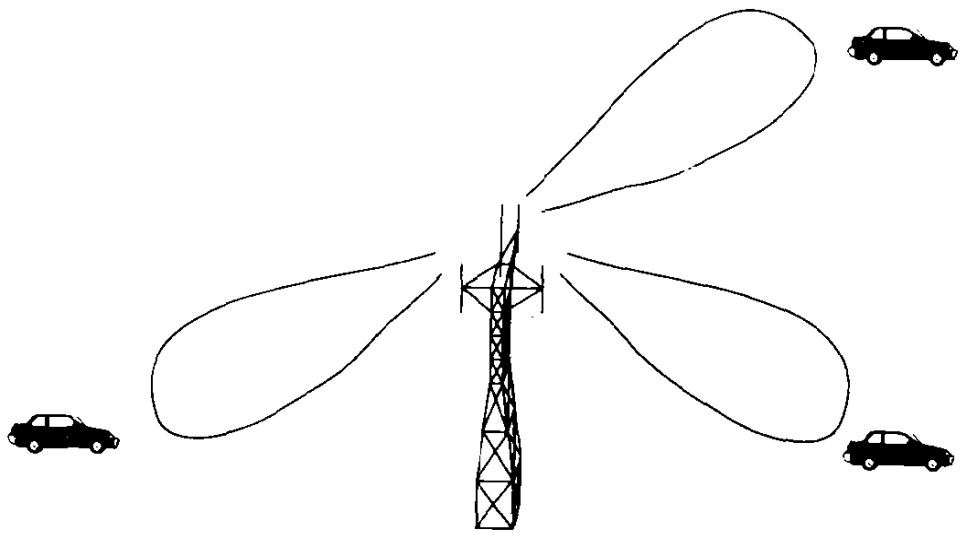


Figure 8.8
A spatially filtered base station antenna serving different users by using spot beams.

The reverse link presents the most difficulty in cellular systems for several reasons [Lib94b]. First, the base station has complete control over the power of all the transmitted signals on the forward link. However, because of different radio propagation paths between each user and the base station, the transmitted power from each subscriber unit must be dynamically controlled to prevent any single user from driving up the interference level for all other users. Second, transmit power is limited by battery consumption at the subscriber unit, therefore there are limits on the degree to which power may be controlled on the reverse link. If the base station antenna is made to spatially filter each desired user so that more energy is detected from each subscriber, then the reverse link for each user is improved and less power is required.

Adaptive antennas used at the base station (and eventually at the subscriber units) promise to mitigate some of the problems on the reverse link. In the limiting case of infinitesimal beamwidth and infinitely fast tracking ability, adaptive antennas implement optimal SDMA, thereby providing a unique channel that is free from the interference of all other users in the cell. With SDMA, all users within the system would be able to communicate at the same time using the same channel. In addition, a perfect adaptive antenna system would be able to track individual multipath components for each user and combine them in an optimal manner to collect all of the available signal energy from each user. The perfect adaptive antenna system is not feasible since it requires infinitely large antennas. However, section 8.7.2 illustrates what gains might be achieved using reasonably sized arrays with moderate directivities.

## 8.6 Packet Radio

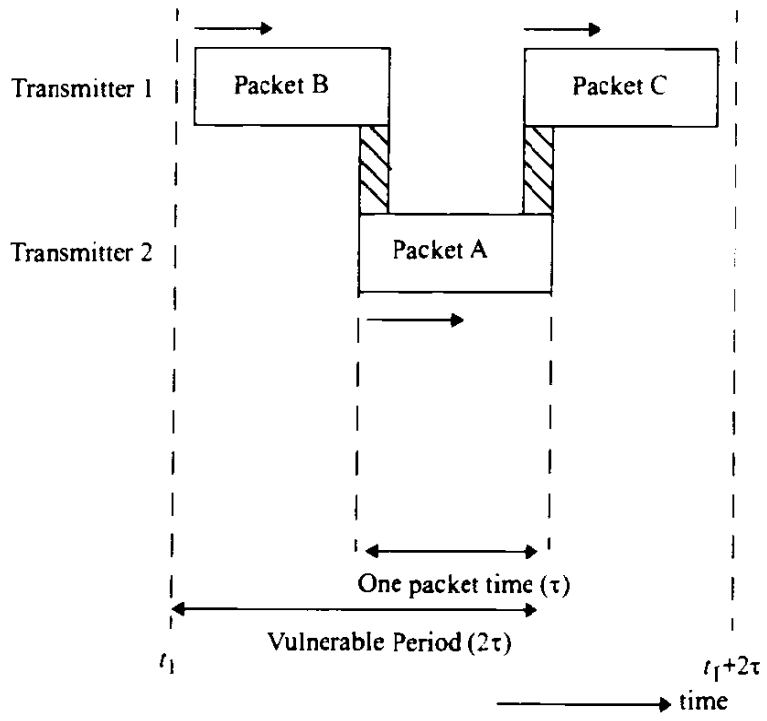
In *packet radio* (PR) access techniques, many subscribers attempt to access a single channel in an uncoordinated (or minimally coordinated) manner. Transmission is done by using bursts of data. Collisions from the simultaneous transmissions of multiple transmitters are detected at the base station receiver, in which case an *ACK* or *NACK* signal is broadcast by the base station to alert the desired user (and all other users) of received transmission. The ACK signal indicates an acknowledgment of a received burst from a particular user by the base station, and a NACK (negative acknowledgment) indicates that the previous burst was not received correctly by the base station. By using ACK and NACK signals, a PR system employs perfect feedback, even though traffic delay due to collisions may be high.

Packet radio multiple access is very easy to implement but has low spectral efficiency and may induce delays. The subscribers use a contention technique to transmit on a common channel. ALOHA protocols, developed for early satellite systems, are the best examples of contention techniques. ALOHA allows each subscriber to transmit whenever they have data to send. The transmitting subscribers listen to the acknowledgment feedback to determine if transmission has

been successful or not. If a collision occurs, the subscriber waits a random amount of time, and then retransmits the packet. The advantage of packet contention techniques is the ability to serve a large number of subscribers with virtually no overhead. The performance of contention techniques can be evaluated by the *throughput* ($T$), which is defined as the average number of messages successfully transmitted per unit time, and the average *delay* ($D$) experienced by a typical message burst.

### 8.6.1 Packet Radio Protocols

In order to determine the throughput, it is important to determine the *vulnerable period,* $V_p$, which is defined as the time interval during which the packets are susceptible to collisions with transmissions from other users. Figure 8.9 shows the vulnerable period for a packet using ALOHA [Tan81]. The Packet A will suffer a collision if other terminals transmit packets during the period $t_1$ to $t_1 + 2\tau$. Even if only a small portion of packet A sustains a collision, the interference may render the message useless.




Packet A will collide with packets B and C because of overlap in transmission time.

Figure 8.9
Vulnerable period for a packet using the ALOHA protocol.

To study packet radio protocols, it is assumed that all packets sent by all users have a constant packet length and fixed, channel data rate, and all other users may generate new packets at random time intervals. Furthermore, it is

assumed that packet transmissions occur with a Poisson distribution having a mean arrival rate of $\lambda$ packets per second. If $\tau$ is the packet duration in seconds, then the *traffic occupancy* or *throughput* $R$ of a packet radio network is given by

$$R = \lambda\tau \tag{8.6}$$

In equation (8.6), $R$ is the normalized channel traffic (measured in Erlangs) due to arriving and buffered packets, and is a relative measure of the channel utilization. If $R > 1$, then the packets generated by the users exceed the maximum transmission rate of the channel [Tan81]. Thus, to obtain a reasonable throughput, the rate at which new packets are generated must lie within $0 < R < 1$. Under conditions of normal loading, the throughput $T$ is the same as the total offered load, $L$. The load $L$ is the sum of the newly generated packets and the retransmitted packets that suffered collisions in previous transmissions. The normalized throughput is always less than or equal to unity and may be thought of as the fraction of time (fraction of an Erlang) a channel is utilized. The normalized throughput is given as the total offered load times the probability of successful transmission, i.e.

$$T = R\Pr[no\ collision] = \lambda\tau\Pr[no\ collision] \tag{8.7}$$

where $\Pr[no\ collision]$ is the probability of a user making a successful packet transmission. The probability that $n$ packets are generated by the user population during a given packet duration interval is assumed to be Poisson distributed and is given as

$$Pr(n) = \frac{R^n e^{-R}}{n!} \tag{8.8}$$

A packet is assumed successfully transmitted if there are no other packets transmitted during the given packet time interval. The probability that zero packets are generated (i.e., no collision) during this interval is given by

$$Pr(0) = e^{-R} \tag{8.9}$$

Based on the type of access, contention protocols are categorized as *random access*, *scheduled access*, and *hybrid access*. In random access, there is no coordination among the users and the messages are transmitted from the users as they arrive at the transmitter. Scheduled access is based on a coordinated access of users on the channel, and the users transmit messages within allotted slots or time intervals. Hybrid access is a combination of random access and scheduled access.

#### 8.6.1.1 Pure ALOHA

The pure ALOHA protocol is a random access protocol used for data transfer. A user accesses a channel as soon as a message is ready to be transmitted. After a transmission, the user waits for an acknowledgment on either the same channel or a separate feedback channel. In case of collisions, (i.e., when a NACK

is received), the terminal waits for a random period of time and retransmits the message. As the number of users increase, a greater delay occurs because the probability of collision increases.

For the ALOHA protocol, the vulnerable period is double the packet duration (see Figure 8.9). Thus, the probability of no collision during the interval of $2\tau$ is found by evaluating $Pr(n)$ given as

$$Pr(n) = \frac{(2R)^n e^{-2R}}{n!} \text{ at } n = 0 \tag{8.10}$$

One may evaluate the mean of equation (8.10) to determine the average number of packets sent during $2\tau$ (This is useful in determining the average offered traffic). The probability of no collision is $Pr(0) = e^{-2R}$. The throughput of the ALOHA protocol is found by using Equation (8.7) as

$$T = Re^{-2R} \tag{8.11}$$

#### 8.6.1.2 Slotted ALOHA

In slotted ALOHA, time is divided into equal time slots of length greater than the packet duration $\tau$. The subscribers each have synchronized clocks and transmit a message only at the beginning of a new time slot, thus resulting in a discrete distribution of packets. This prevents partial collisions, where one packet collides with a portion of another. As the number of users increase, a greater delay will occur due to complete collisions and the resulting repeated transmissions of those packets originally lost. The number of slots which a transmitter waits prior to retransmitting also determines the delay characteristics of the traffic. The vulnerable period for slotted ALOHA is only one packet duration, since partial collisions are prevented through synchronization. The probability that no other packets will be generated during the vulnerable period is $e^{-R}$. The throughput for the case of slotted ALOHA is thus given by

$$T = Re^{-R} \tag{8.12}$$

Figure 8.10 illustrates how ALOHA and slotted ALOHA systems trade-off throughput for delay.

---

**Example 8.6**

Determine the maximum throughput that can be achieved using ALOHA and slotted ALOHA protocols.

**Solution to Example 8.6**

The rate of arrival which maximizes the throughput for ALOHA is found by taking the derivative of Equation (8.11) and equating it to zero.

$$\frac{dT}{dR} = e^{-2R} - 2Re^{-2R} = 0$$

$$R_{max} = 1/2$$

Maximum throughput achieved by using the ALOHA protocol is found by substituting $R_{max}$ in Equation (8.11), and this value can be seen as the maximum throughput in Figure 8.10

$$T = \frac{1}{2}e^{-1} = 0.1839$$

Thus the best traffic utilization one can hope for using ALOHA is 0.184 Erlangs.

The maximum throughput for slotted ALOHA is found by taking the derivative of Equation (8.12) and equating it to zero.

$$\frac{dT}{dR} = e^{-R} - Re^{-R} = 0$$

$$R_{max} = 1$$

Maximum throughput is found by substituting $R_{max}$ in equation (8.12), and this value can be seen as the maximum throughput in Figure 8.10.

$$T = e^{-1} = 0.3679$$

Notice that slotted ALOHA provides a maximum channel utilization of 0.368 Erlangs, double that of ALOHA.

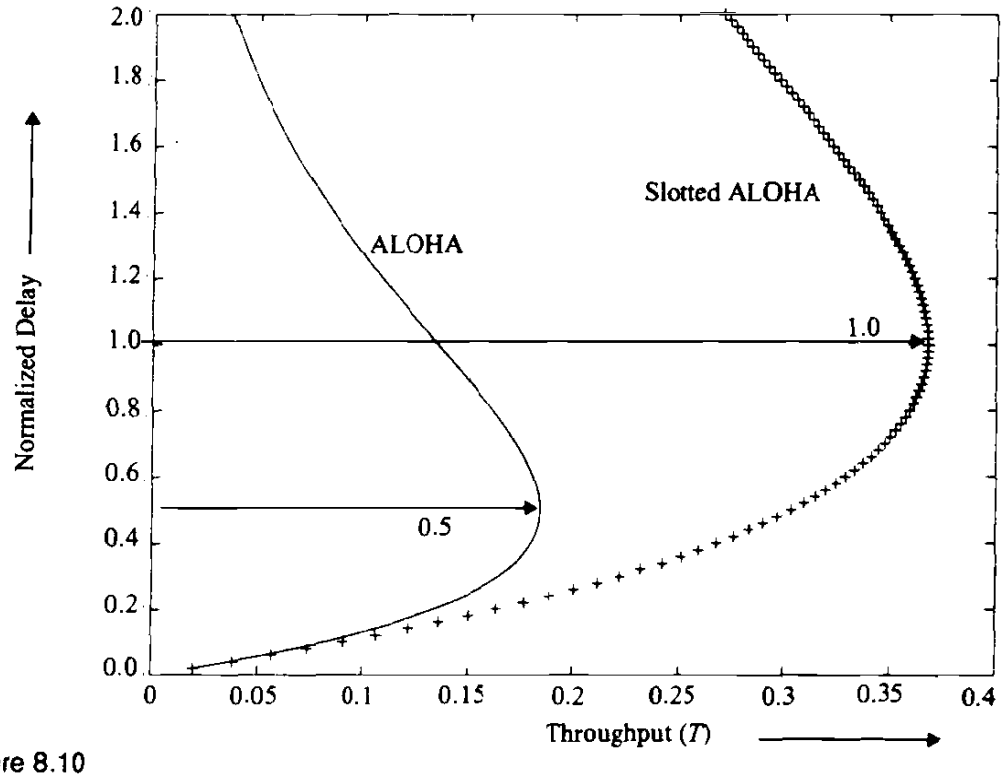



Figure 8.10
Trade-off between throughput and delay for ALOHA and slotted ALOHA packet radio protocols.

### 8.6.2 Carrier Sense Multiple Access (CSMA) Protocols

ALOHA protocols do not listen to the channel before transmission, and therefore do not exploit information about the other users. By listening to the channel before engaging in transmission, greater efficiencies may be achieved. CSMA protocols are based on the fact that each terminal on the network is able to monitor the status of the channel before transmitting information. If the channel is idle (i.e., no carrier is detected), then the user is allowed to transmit a packet based on a particular algorithm which is common to all transmitters on the network.

In CSMA protocols, *detection delay* and *propagation delay* are two important parameters. Detection delay is a function of the receiver hardware and is the time required for a terminal to sense whether or not the channel is idle. Propagation delay is a relative measure of how fast it takes for a packet to travel from a base station to a mobile terminal. With a small detection time, a terminal detects a free channel quite rapidly, and small propagation delay means that a packet is transmitted through the channel in a small interval of time relative to the packet duration.

Propagation delay is important, since just after a user begins sending a packet, another user may be ready to send and may be sensing the channel at the same time. If the transmitting packet has not reached the user who is poised to send, the latter user will sense an idle channel and will also send its packet, resulting in a collision between the two packets. Propagation delay impacts the performance of CSMA protocols. If $t_p$ is the propagation time in seconds, $R_b$ is the channel bit rate, and $m$ is the expected number of bits in a data packet [Tan81], [Ber92], then the propagation delay $t_d$ (in packet transmission units) can be expressed as

$$t_d = \frac{t_p R_b}{m} \tag{8.13}$$

There exist several variations of the CSMA strategy [Kle75], [Tob75]:

- ***1-persistent CSMA*** — The terminal listens to the channel and waits for transmission until it finds the channel idle. As soon as the channel is idle, the terminal transmits its message with probability one.
- ***non-persistent CSMA*** — In this type of CSMA strategy, after receiving a negative acknowledgment the terminal waits a random time before retransmission of the packet. This is popular for wireless LAN applications, where the packet transmission interval is much greater than the propagation delay to the farthermost user.
- ***p-persistent CSMA*** — $p$-persistent CSMA is applied to slotted channels. When a channel is found to be idle, the packet is transmitted in the first available slot with probability $p$ or in the next slot with probability $1\text{-}p$.
- ***CSMA/CD*** — In CSMA with collision detection (CD), a user monitors its transmission for collisions. If two or more terminals start a transmission at

the same time, collision is detected, and the transmission is immediately aborted in midstream. This is handled by a user having both a transmitter and receiver which is able to support *listen-while-talk* operation. For a single radio channel, this is done by interrupting the transmission in order to sense the channel. For duplex systems, a full duplex transceiver is used [Lam80].

***Data sense multiple access*** (DSMA) — DSMA is a special type of CSMA that relies on successfully demodulating a forward control channel before broadcasting data back on a reverse channel. Each user attempts to detect a *busy-idle* message which is interspersed on the forward control channel. When the busy-idle message indicates that no users are transmitting on the reverse channel, a user is free to send a packet. This technique is used in the cellular digital packet data (CDPD) cellular network described in Chapter 9.

### 8.6.3 Reservation Protocols

#### 8.6.3.1 Reservation ALOHA

Reservation ALOHA is a packet access scheme based on time division multiplexing. In this protocol, certain packet slots are assigned with priority, and it is possible for users to reserve slots for the transmission of packets. Slots can be permanently reserved or can be reserved on request. For high traffic conditions, reservations on request offers better throughput. In one type of reservation ALOHA, the terminal making a successful transmission reserves a slot permanently until its transmission is complete, although very large duration transmissions may be interrupted. Another scheme allows a user to transmit a request on a subslot which is reserved in each frame. If the transmission is successful (i.e, no collisions are detected), the terminal is allocated the next regular slot in the frame for data transmission [Tan81].

#### 8.6.3.2 Packet Reservation Multiple Access (PRMA)

PRMA uses a discrete packet time technique similar to reservation ALOHA and combines the cyclical frame structure of TDMA in a manner that allows each TDMA time slot to carry either voice or data, where voice is given priority. PRMA was proposed in [Goo89] as a means of integrating bursty data and human speech. PRMA defines a frame structure, much like is used in TDMA systems. Within each frame, there are a fixed number of time slots which may be designated as either "reserved" or "available", depending on the traffic as determined by the controlling base station. PRMA is discussed in Chapter 9.

### 8.6.4 Capture Effect in Packet Radio

Packet radio multiple access techniques are based on contention within a channel. When used with FM or spread spectrum modulation, it is possible for the strongest user to successfully *capture* the intended receiver, even when many other users are also transmitting. Often, the closest transmitter is able to capture a receiver because of the small propagation path loss. This is called the

*near-far effect*. The capture effect offers both advantages and disadvantages in practical systems. Because a particular transmitter may capture an intended receiver, many packets may survive despite collision on the channel. However, a strong transmitter may make it impossible for the receiver to detect a much weaker transmitter which is attempting to communicate to the same receiver. This problem is known as the *hidden transmitter* problem.

A useful parameter in analyzing the capture effects in packet radio protocols is the minimum power ratio of an arriving packet, relative to the other colliding packets, such that it is received. This ratio is called the *capture ratio*, and is dependent upon the receiver and the modulation used.

In summary, packet radio techniques support mobile transmitters sending bursty traffic in the form of data packets using random access. Ideal channel throughput can be increased if terminals synchronize their packet transmissions into common time slots, such that the risk of partial packet overlap is avoided. With high traffic loads, both unslotted and slotted ALOHA protocols become inefficient, since the contention between all transmitted packets exposes most of the offered traffic to collisions, and thus results in multiple retransmissions and increased delays. To reduce this situation CSMA can be used where the transmitter first listens either to the common radio channel or to a separate dedicated acknowledgment control channel from the base station. In a real world mobile system, the CSMA protocols may fail to detect ongoing radio transmissions of packets subject to deep fading on the reverse channel path. Utilization of an ALOHA channel can be improved by deliberately introducing differences between the transmit powers of multiple users competing for the base station receiver. The Table 8.2 below shows the multiple access techniques which should be used for different types of traffic conditions.

**Table 8.2 Multiple Access Techniques for Different Traffic Types**

| Type of Traffic | Multiple Access Technique |
|---|---|
| Bursty, short messages | Contention protocols |
| Bursty, long messages, large number of users | Reservation Protocols |
| Bursty, long messages, small number of users | Reservation protocols with fixed TDMA reservation channel |
| Stream or deterministic (voice) | FDMA, TDMA, CDMA |



## 8.7 Capacity of Cellular Systems

*Channel capacity* for a radio system can be defined as the maximum number of channels or users that can be provided in a fixed frequency band. Radio capacity is a parameter which measures spectrum efficiency of a wireless system. This parameter is determined by the required carrier-to-interference ratio $(C/I)$ and the channel bandwidth $B_c$.

In a cellular system the interference at a base station receiver will come from the subscriber units in the surrounding cells. This is called *reverse channel interference*. For a particular subscriber unit, the desired base station will provide the desired forward channel while the surrounding co-channel base stations will provide the *forward channel interference*. Considering the forward channel interference problem, let $D$ be the distance between two co-channel cells and $R$ be the cell radius. Then the minimum ratio of $D/R$ that is required to provide a tolerable level of co-channel interference is called the co-channel reuse ratio and is given by [Lee89a]

$$Q = \frac{D}{R} \tag{8.14}$$

The radio propagation characteristics determine the *carrier-to-interference ratio* $(C/I)$ at a given location, and models presented in Chapter 3 and Appendix B are used to find sensible $C/I$ values. As shown in Figure 8.11, the $M$ closest co-channel cells may be considered as first order interference in which case $C/I$ is given by

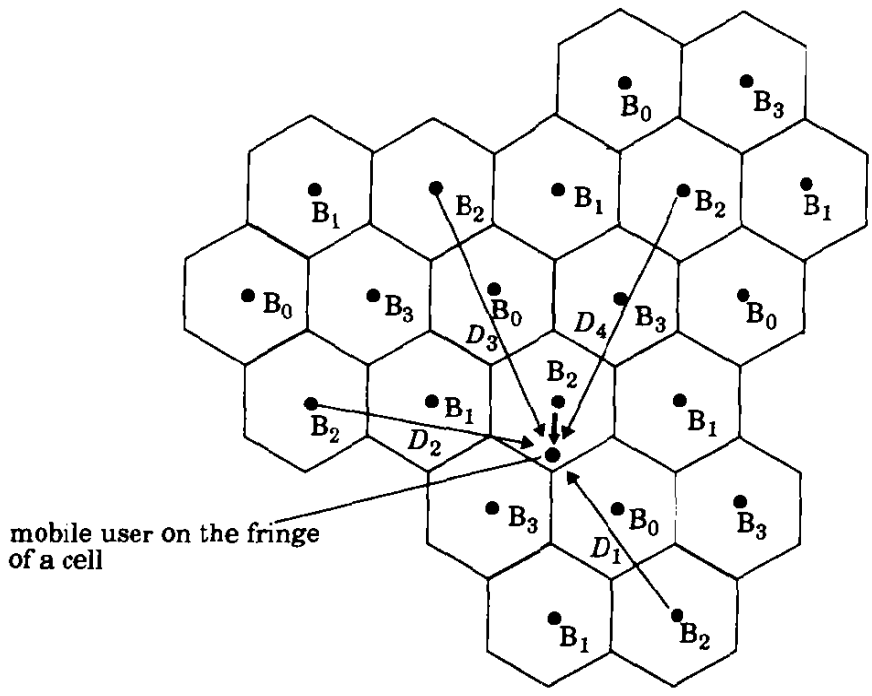



Figure 8.11
Illustration of forward channel interference for a cluster size of $N = 4$. Shown here are four co-channel base stations which interfere with the serving base station. The distance from the serving base station to the user is $D_0$, and interferers are a distance $D_k$ from the user.

$$\frac{C}{I} = \frac{D_0^{-n_o}}{\sum_{k=1}^{M} D_k^{-n_k}} \tag{8.15}$$

where $n_0$ is the path loss exponent in the desired cell, $D_0$ is the distance from the desired base station to the mobile, $D_k$ is the distance of the $k$ th cell from the mobile, and $n_k$ is the path loss exponent to the $k$ th interfering base station. If only the six closest interfering cells are considered, and all are approximately at the same distance $D$ and have similar path loss exponents equal to that in the desired cell, then $C/I$ is given by

$$\frac{C}{I} = \frac{D_0^{-n}}{6D^{-n}} \tag{8.16}$$

Now, if it is assumed that maximum interference occurs when the mobile is at the cell edge $D_0 = R$, and if the $C/I$ for each user is required to be greater than some minimum $(C/I)_{min}$, which is the minimum carrier-to-interference ratio that still provides acceptable signal quality at the receiver, then the following equation must hold for acceptable performance:

$$\frac{1}{6}\left(\frac{R}{D}\right)^{-n} \geq \left(\frac{C}{I}\right)_{min} \tag{8.17}$$

Thus, from Equation (8.14), the co-channel reuse factor is

$$Q = \left(6\left(\frac{C}{I}\right)_{min}\right)^{1/n} \tag{8.18}$$

The radio capacity of a cellular system is defined as

$$m = \frac{B_t}{B_c N} \text{ radio channels/cell} \tag{8.19}$$

where $m$ is the radio capacity metric, $B_t$ is the total allocated spectrum for the system, $B_c$ is the channel bandwidth, and $N$ is the number of cells in a frequency reuse pattern. As shown in Chapter 2, $N$ is related to the co-channel reuse factor $Q$ by

$$Q = \sqrt{3N} \tag{8.20}$$

From Equations (8.18), (8.19), and (8.20), the radio capacity is given as

$$m = \frac{B_t}{B_c \frac{Q^2}{3}} = \frac{B_t}{B_c\left(\frac{6}{3^{n/2}}\left(\frac{C}{I}\right)_{min}\right)^{2/n}} \tag{8.21}$$

As shown by Lee [Lee89a], when $n = 4$, the radio capacity is given by

$$m = \frac{B_t}{B_c\sqrt{\frac{2}{3}\left(\frac{C}{I}\right)_{min}}} \text{ radio channels/cell} \tag{8.22}$$

In order to provide the same voice quality, $(C/I)_{min}$ may be lower in a digital systems when compared to an analog system. Typically, the minimum required $C/I$ is about 12 dB for narrowband digital systems and 18 dB for narrowband analog FM systems, although exact values are determined by subjective listening tests in real-world propagation conditions. Each digital wireless standard has a different $(C/I)_{min}$, and in order to compare different systems, an equivalent $C/I$ must be used. If $B_t$ and $m$ are kept constant in Equation (8.22), then it is clear that $B_c$ and $(C/I)_{min}$ are related by

$$\left(\frac{C}{I}\right)_{eq} = \left(\frac{C}{I}\right)_{min}\left(\frac{B_c}{B_c'}\right)^2 \tag{8.23}$$

where $B_c$ is the bandwidth of a particular system, $(C/I)_{min}$ is the tolerable value for the same system, $B_c'$ is the channel bandwidth for a different system, and $(C/I)_{eq}$ is the minimum $C/I$ value for the different system when compared to the $(C/I)_{min}$ for a particular system. Notice that for a constant number of users per radio channel, the same voice quality will be maintained in a different system if $(C/I)_{min}$ increases by a factor of four when the bandwidth is halved. Equation (8.22) indicates that maximum radio capacity occurs when $(C/I)_{min}$ and $B_c$ are minimized, yet equation (8.23) shows that $(C/I)_{min}$ and $B_c$ are inversely related.

**Example 8.7**

Evaluate four different cellular radio standards, and choose the one with the maximum radio capacity.

System A: $B_c$ = 30 kHz, $(C/I)_{min}$ = 18 dB
System B: $B_c$ = 25 kHz, $(C/I)_{min}$ = 14 dB
System C: $B_c$ = 12.5 kHz, $(C/I)_{min}$ = 12 dB
System D: $B_c$ = 6.25 kHz, $(C/I)_{min}$ = 9 dB

**Solution to Example 8.7**

Consider each system for 6.25 kHz bandwidth, and use equation (8.23)

System A; $B_c$ = 6.25 kHz, $(C/I)_{eq}$ = 18 + 20 log (6.25/30) = 4.375 dB
System B; $B_c$ = 6.25 kHz, $(C/I)_{eq}$ = 14 + 20log (6.25/25) = 1.96 dB
System C; $B_c$ = 6.25 kHz, $(C/I)_{eq}$ = 12 + 20log (6.25/12.5) = 6 dB
System D; $B_c$ = 6.25 kHz, $(C/I)_{eq}$ = 9 + 20 log (6.25/6.25) = 9 dB

Based on comparison, the smallest value of $(C/I)_{eq}$ should be selected for maximum capacity in Equation (8.22). System B offers the best capacity.

In a digital cellular system, $C/I$ can be expressed as

$$\frac{C}{I} = \frac{E_b R_b}{I} = \frac{E_c R_c}{I} \tag{8.24}$$

where $R_b$ is channel bit rate, $E_b$ is the energy per bit, $R_c$ is the rate of the channel code and $E_c$ is the energy per code symbol. From equations (8.23) and (8.24), the ratio of $C/I$ to $(C/I)_{eq}$ is given as

$$\frac{\left(\frac{C}{I}\right)}{\left(\frac{C}{I}\right)_{eq}} = \frac{\frac{E_c R_c}{I}}{\frac{E_c' R_c'}{I}} = \left(\frac{B_c'}{B_c}\right)^2 \tag{8.25}$$

The relationship between $R_c$ and $B_c$ is always linear, and if the interference level $I$ is the same in the mobile environment for two different digital systems, then equation (8.25) can be rewritten as

$$\frac{E_c}{E_c'} = \left(\frac{B_c'}{B_c}\right)^3 \tag{8.26}$$

Equation (8.26) shows that if $B_c$ is reduced by half, then the energy code symbol increases eight times. This gives the relationship between $E_b/N_0$ and $B_c$ in a digital cellular system.

A comparison can now be made between the spectrum efficiency for FDMA and TDMA. In FDMA, $B_t$ is divided into $M$ channels, each with bandwidth $B_c$. Therefore, the radio capacity for FDMA is given by

$$m = \frac{B_t}{\frac{B_t}{M}\sqrt{\frac{2}{3}\left(\frac{C}{I}\right)}} \tag{8.27}$$

Consider the case where a multichannel FDMA system occupies the same spectrum as a single channel TDMA system with multiple time slots. The carrier and interference terms for the first access technique (in this case FDMA) can be written as, $C = E_b R_b$, $I = I_0 B_c$, whereas the second access technique (in this case TDMA) has carrier and interference terms represented by $C' = E_b R_b'$, $I' = I_0 B_c'$, where $R_b$ and $R_b'$ are the radio transmission rates of two digital systems, $E_b$ is the energy per bit, and $I_0$ represents the interference power per Hertz. The terms $C'$ and $I'$ are the parameters for the TDMA channels, and the terms $C$ and $I$ apply to the FDMA channels.

**Example 8.8**

Consider an FDMA system with three channels, each having a bandwidth of 10 kHz and a transmission rate of 10 kbps. A TDMA system has three time slots, channel bandwidth of 30 kHz, and a transmission rate of 30 kbps.

For the TDMA scheme, the received carrier-to-interference ratio for a single user is measured for 1/3 of the time the channel is in use. For example, $C'/I'$ can be measured in 333.3 ms in one second. Thus $C'/I'$ is given by

$$C' = E_b R_b' = \frac{E_b 10^4 \text{ bits}}{0.333 \text{ s}} = 3R_b E_b = 3C \qquad \text{(E.8.8.1)}$$

$$I' = I_0 B_c' = I_0 30 \text{ kHz} = 3I$$

It can be seen that the received carrier-to-interference ratio for a user in this TDMA system $C'/I'$ is the same as $C/I$ for a user in the FDMA system. Therefore, for this example, FDMA and TDMA have the same radio capacity and consequently the same spectrum efficiency. However, the required peak power for TDMA is $10\log k$ higher than FDMA, where $k$ is the number of time slots in a TDMA system of equal bandwidth.

***Capacity of Digital Cellular TDMA*** — In practice, TDMA systems improve capacity by a factor of 3 to 6 times as compared to analog cellular radio systems. Powerful error control and speech coding enable better link performance in high interference environments. By exploiting speech activity, some TDMA systems are able to better utilize each radio channel. Mobile assisted handoff (MAHO) allows subscribers to monitor the neighboring base stations, and the best base station choice may be made by each subscriber. MAHO allows the deployment of densely packed microcells, thus giving substantial capacity gains in a system. TDMA also makes it possible to introduce *adaptive channel allocation* (ACA). ACA eliminates system planning since it is not required to plan frequencies for cells. Various proposed standards such as the GSM, U.S digital cellular (USDC), and Pacific Digital Cellular (PDC) have adopted digital TDMA for high capacity. Table 8.3 compares analog FM based AMPS to other digital TDMA based cellular systems.

**Table 8.3 Comparison of AMPS With Digital TDMA Based Cellular Systems [Rai91]**

| Parameter | AMPS | GSM | USDC | PDC |
|---|---|---|---|---|
| Bandwidth (MHz) | 25 | 25 | 25 | 25 |
| Voice Channels | 833 | 1000 | 2500 | 3000 |
| Frequency Reuse (Cluster sizes) | 7 | 4 or 3 | 7 or 4 | 7 or 4 |
| Channels/Site | 119 | 250 or 333 | 357 or 625 | 429 or 750 |
| Traffic (Erlangs/sq. km) | 11.9 | 27.7 or 40 | 41 or 74.8 | 50 or 90.8 |
| Capacity Gain | 1.0 | 2.3 or 3.4 | 3.5 or 6.3 | 4.2 or 7.6 |

### 8.7.1 Capacity of Cellular CDMA

The capacity of CDMA systems is interference limited, while it is bandwidth limited in FDMA and TDMA. Therefore, any reduction in the interference will cause a linear increase in the capacity of CDMA. Put another way, in a CDMA system, the link performance for each user increases as the number of users decreases. A straightforward way to reduce interference is to use multisectorized antennas, which results in spatial isolation of users. The directional antennas receive signals from only a fraction of the current users, thus leading

to the reduction of interference. Another way of increasing CDMA capacity is to operate in a *discontinuous transmission mode* (DTX), where advantage is taken of the intermittent nature of speech. In DTX, the transmitter is turned off during the periods of silence in speech. It has been observed that voice signals have a duty factor of about 3/8 in landline networks [Bra68], and 1/2 for mobile systems, where background noise and vibration can trigger voice activity detectors. Thus, the average capacity of a CDMA system can be increased by a factor inversely proportional to the duty factor. While TDMA and FDMA reuse frequencies depending on the isolation between cells provided by the path loss in terrestrial radio propagation, CDMA can reuse the entire spectrum for all cells, and this results in an increase of capacity by a large percentage over the normal frequency reuse factor.

For evaluating the capacity of CDMA system, first consider a single cell system [Gil91]. The cellular network consists of a large number of mobile users communicating with a base station (In a multiple cell system, all the base stations are interconnected by the mobile switching center). The cell-site transmitter consists of a linear combiner which adds the spread signals of the individual users and also uses a weighting factor for each signal for forward link power control purposes. For a single cell system under consideration, these weighting factors can be assumed to be equal. A pilot signal is also included in the cell-site transmitter and is used by each mobile to set its own power control for the reverse link. For a single-cell system with power control, all the signals on the reverse channel are received at the same power level at the base station.

Let the number of users be $N$. Then, each demodulator at the cell site receives a composite waveform containing the desired signal of power $S$ and $(N-1)$ interfering users, each of which has power, $S$. Thus, the signal-to-noise ratio is [Gil91],

$$SNR = \frac{S}{(N-1)S} = \frac{1}{(N-1)} \tag{8.28}$$

In addition to SNR, bit energy-to-noise ratio is an important parameter in communication systems. It is obtained by dividing the signal power by the baseband information bit rate, $R$, and the interference power by the total RF bandwidth, $W$. The SNR at the base station receiver can be represented in terms of $E_b/N_0$ given by

$$\frac{E_b}{N_0} = \frac{S/R}{(N-1)(S/W)} = \frac{W/R}{N-1} \tag{8.29}$$

Equation (8.29) does not take into account the background thermal noise, $\eta$, in the spread bandwidth. To take this noise into consideration, $E_b/N_0$ can be represented as

$$\frac{E_b}{N_0} = \frac{W/R}{(N-1) + (\eta/S)} \tag{8.30}$$

The number of users that can access the system is thus given as

$$N = 1 + \frac{W/R}{E_b/N_0} - (\eta/S) \tag{8.31}$$

where $W/R$ is called the processing gain. The background noise determines the cell radius for a given transmitter power.

In order to achieve an increase in capacity, the interference due to other users should be reduced. This can be done by decreasing the denominator of equations (8.28) or (8.29). The first technique for reducing interference is antenna sectorization. As an example, a cell site with three antennas, each having a beam width of $120^o$, has interference $N_0'$ which is one-third of the interference received by an omni-directional antenna. This increases the capacity by a factor of 3 since three times as many users may now be served within a sector while matching the performance of the omni-directional antenna system. Looking at it another way, the same number of users in an omni-directional cell may now be served in 1/3rd the area. The second technique involves the monitoring of voice activity such that each transmitter is switched off during periods of no *voice activity.* Voice activity is denoted by a factor $\alpha$, and the interference term in equation (8.29) becomes $(N_s - 1)\alpha$, where $N_s$ is the number of users per sector. With the use of these two techniques, the new average value of $E_b/N_0'$ *within a sector* is given as

$$\frac{E_b}{N_0'} = \frac{W/R}{(N_s - 1)\alpha + (\eta/S)} \tag{8.32}$$

When the number of users is large and the system is interference limited rather than noise limited, the number of users can be shown to be

$$N_s = 1 + \frac{1}{\alpha}\left[\frac{W/R}{\frac{E_b}{N_0'}}\right] \tag{8.33}$$

If the voice activity factor is assumed to have a value of 3/8, and three sectors per cell site are used, Equation (8.33) demonstrates that the SNR increases by a factor of 8, which leads to an 8 fold increase in the number of users compared to an omni-directional antenna system with no voice activity detection.

**CDMA Power Control** — In CDMA, the system capacity is maximized if each mobile transmitter power level is controlled so that its signal arrives at the cell site with the minimum required signal-to-interference ratio [Sal91]. If the signal powers of all mobile transmitters within an area covered by a cell site are controlled, then the total signal power received at the cell site from all mobiles will be equal to the average received power times the number of mobiles operat-

ing in the region of coverage. A trade-off must be made if a mobile signal arrives at the cell site with a signal that is too weak, and often the weak user will be dropped. If the received power from a mobile user is too great the performance of this mobile unit will be acceptable, but it will add undesired interference to all other users in the cell.

**Example 8.9**

If $W = 1.25$ MHz, $R = 9600$ bps, and a minimum acceptable $E_b/N_0$ is found to be 10 dB, determine the maximum number of users that can be supported in a single-cell CDMA system using (a) omni-directional base station antennas and no voice activity detection, and (b) 3-sectors at the base station and activity detection with $\alpha = 3/8$. Assume the system is interference limited.

**Solution to Example 8.9**

(a) Using equation (8.31)

$$N = 1 + \frac{1.25 \times 10^6/9600}{10} = 1 + 13.02 = 14$$

(b) Using equation (8.33) for each sector we can find $N_s$.

$$N_s = 1 + \frac{1}{0.375}\left[\frac{1.25 \times 10^6/9600}{10}\right] = 35.7$$

The total number of users is given by $3N_s$, since three sectors exist within a cell; therefore $N = 3 \times 35.7 = 107$ users/cell.

### 8.7.2 Capacity of CDMA with Multiple Cells

In actual CDMA cellular systems that employ separate forward and reverse links, neighboring cells share the same frequency, and each base station controls the transmit power of each of its own in-cell users. However, a particular base station is unable to control the power of users in neighboring cells, and these users add to the noise floor and decrease capacity on the reverse link of the particular cell of interest. Figure 8.12 illustrates an example of how users in adjacent cells may be distributed over the coverage area. The transmit powers of each out-of-cell user will add to the in-cell interference (where users are under power control) at the base station receiver. The amount of out-of-cell interference determines the *frequency reuse factor*, $f$, of a CDMA cellular system. Ideally, each cell shares the same frequency and the maximum possible value of $f$ ($f = 1$) is achieved. In practice, however, the out-of-cell interference reduces $f$ significantly. In contrast to CDMA systems which use the same frequency for each cell, narrowband FDMA/FDD systems typically reuse channels every seven cells, in which case $f$ is simply 1/7 (see Chapter 2).

The frequency reuse factor for a CDMA system on the reverse link can be defined as [Rap92b]

CHAPTER 10

# Wireless Systems and Standards

This chapter describes many of the cellular radio, cordless telephone, and personal communications standards in use throughout the world. First, existing analog cellular standards in the United States and Europe are described. Then, a description of emerging digital cellular and PCS standards is presented. A convenient summary of world-wide standards and a discussion about the U.S. PCS and wireless cable frequencies are provided at the end of the chapter.

## 10.1 AMPS and ETACS

In the late 1970s, AT&T Bell Laboratories developed the first U.S cellular telephone system called the Advance Mobile Phone Service (AMPS) [You79]. AMPS was first deployed in late 1983 in the urban and suburban areas of Chicago by Ameritech. In 1983, a total of 40 MHz of spectrum in the 800 MHz band was allocated by the Federal Communications Commission for the Advanced Mobile Phone Service. In 1989, as the demand for cellular telephone services increased, the Federal Communications Commission allocated an additional 10 MHz (called the extended spectrum) for cellular telecommunications. The first AMPS cellular system used large cells and omni-directional base station antennas to minimize initial equipment needs, and the system was deployed in Chicago to cover approximately 2100 square miles.

The AMPS system uses a 7-cell reuse pattern with provisions for sectoring and cell splitting to increase capacity when needed. After extensive subjective tests, it was found that the AMPS 30 kHz channel requires a signal-to-interference ratio (SIR) of 18 dB for satisfactory system performance. The smallest reuse

factor which satisfies this requirement using 120 degree directional antennas is $N = 7$ (see Chapter 2), and hence a 7-cell reuse pattern has been adopted.

AMPS is used throughout the world and is particularly popular in the U.S., South America, Australia, and China. While the U.S. system has been designed for a duopoly market (e.g., two competing carriers per market), many countries have just a single provider. Thus, while U.S. AMPS restricts the A and B side carriers to a subset of 416 channels each, other implementations of AMPS allow all possible channels to be used. Furthermore, the exact frequency allocations for AMPS differ from country to country. Nevertheless, the air interface standard remains identical throughout the world.

The European Total Access Communication System (ETACS) was developed in the mid 1980s, and is virtually identical to AMPS, except it is scaled to fit in 25 kHz (as opposed to 30 kHz) channels used throughout Europe. Another difference between ETACS and AMPS is how the telephone number of each subscriber (called the mobile identification number or MIN) is formatted, due to the need to accommodate different country codes throughout Europe and area codes in the U.S.

### 10.1.1 AMPS and ETACS System Overview

Like all other first generation, analog, cellular systems, AMPS and ETACS use frequency modulation (FM) for radio transmission. In the United States, transmissions from mobiles to base stations (reverse link) use frequencies between 824 MHz and 849 MHz, while base stations transmit to mobiles (forward link) using frequencies between 869 MHz and 894 MHz. ETACS uses 890 MHz to 915 MHz for the reverse link and 935 MHz to 960 MHz for the forward link. Every radio channel actually consists of a pair of simplex channels separated by 45 MHz. A separation of 45 MHz between the forward and reverse channels was chosen to make use of inexpensive but highly selective duplexers in the subscriber units. For AMPS, the maximum deviation of the FM modulator is $\pm 12$ kHz ($\pm 10$ kHz for ETACS). The control channel transmissions and blank-and-burst data streams are transmitted at 10 kbps for AMPS, and at 8 kbps for ETACS. These wideband data streams have a maximum frequency deviation of $\pm 8$ kHz and $\pm 6.4$ kHz for AMPS and ETACS, respectively.

AMPS and ETACS cellular radio systems generally have tall towers which support several receiving antennas and have transmitting antennas which typically radiate a few hundred watts of effective radiated power. Each base station typically has one control channel transmitter (that broadcasts on the forward control channel), one control channel receiver (that listens on the reverse control channel for any cellular phone switching to set-up a call), and eight or more FM duplex voice channels. Commercial base stations support as many as fifty-seven voice channels. Forward voice channels (FVC) carry the portion of the telephone conversation originating from the landline telephone network caller and going to

the cellular subscriber. Reverse voice channels (RVC) carry the portion of the telephone conversation originating from the cellular subscriber and going to the landline telephone network caller. The actual number of control and voice channels used at a particular base station varies widely in different system installations depending on traffic, maturity of the system, and locations of other base stations. The number of base stations in a service area varies widely, as well, from as few as one cellular tower in a rural area to several hundred or more base stations in a large city.

Each base station in the AMPS or ETACS system continuously transmits digital FSK data on the forward control channel (FCC) at all times so that idle cellular subscriber units can lock onto the strongest FCC wherever they are. All subscribers must be locked, or "camped" onto a FCC in order to originate or receive calls. The base station reverse control channel (RCC) receiver constantly monitors transmissions from cellular subscribers that are locked onto the matching FCC. In the U.S. AMPS system, there are twenty-one control channels for each of the two service providers in each market, and these control channels are standardized throughout the country. ETACS supports forty-two control channels for a single provider. Thus any cellular telephone in the system only needs to scan a limited number of control channels to find the best serving base station. It is up to the service provider to ensure that neighboring base stations within a system are assigned forward control channels that do not cause adjacent channel interference to subscribers which monitor different control channels in nearby base stations.

In each U.S. cellular market, the nonwireline service provider (the "A" provider") is assigned an odd *system identification number* (SID) and the wireline service provider (the "B" provider) is assigned an even SID. The SID is transmitted once every 0.8 seconds on each FCC, along with other overhead data which reports the status of the cellular system. Transmitted data might include information such as whether roamers are automatically registered, how power control is handled, and whether other standards, such as USDC or narrowband AMPS, can be handled by the cellular system. In the U.S., subscriber units generally access channels exclusively on the A or B side, although cellular phones are capable of allowing the user to access channels on both sides. For ETACS, *area identification numbers* (AID) are used instead of SID, and ETACS subscriber units are able to access any control or voice channel in the standard.

### 10.1.2 Call Handling in AMPS and ETACS

When a call to a cellular subscriber originates from a conventional telephone in the public-switched telephone network (PSTN) and arrives at the mobile switching center (MSC), a paging message is sent out with the subscriber's mobile identification number (MIN) simultaneously on every base station forward control channel in the system. If the intended subscriber unit

successfully receives its page on a forward control channel, it will respond with an acknowledgment transmission on the reverse control channel. Upon receiving the subscriber's acknowledgment, the MSC directs the base station to assign a forward voice channel (FVC) and reverse voice channel (RVC) pair to the subscriber unit so that the new call can take place on a dedicated voice channel. The base station also assigns the subscriber unit a supervisory audio tone (SAT tone) and a voice mobile attenuation code (VMAC) as it moves the call to the voice channel. The subscriber unit automatically changes its frequency to the assigned voice channel pair.

The SAT, as described subsequently, has one of three different frequencies which allows the base and mobile to distinguish each other from co-channel users located in different cells. The SAT is transmitted continuously on both the forward and reverse voice channels during a call at frequencies above the audio band. The VMAC instructs the subscriber unit to transmit at a specific power level. Once on the voice channel, wideband FSK data is used by the base station and subscriber unit in a *blank-and-burst* mode to initiate handoffs, change the subscriber transmit power as needed, and provide other system data. Blank-and-burst signaling allows the MSC to send bursty data on the voice channel by temporarily omitting the speech and SAT, and replacing them with data. This is barely noticed by the voice users.

When a mobile user places a call, the subscriber unit transmits an origination message on the reverse control channel (RCC). The subscriber unit transmits its MIN, electronic serial number (ESN), station class mark (SCM), and the destination telephone number. If received correctly by the base station, this information is sent to the MSC which checks to see if the subscriber is properly registered, connects the subscriber to the PSTN, assigns the call to a forward and reverse voice channel pair with a specific SAT and VMAC, and commences the conversation.

During a typical call, the MSC issues numerous blank-and-burst commands which switch subscribers between different voice channels on different base stations, depending on where the subscriber is travelling in the service area. In AMPS and ETACS, handoff decisions are made by the MSC when the signal strength on the reverse voice channel (RVC) of the serving base station drops below a preset threshold, or when the SAT tone experiences a certain level of interference. Thresholds are adjusted at the MSC by the service provider, are subject to continuous measurement, and must be changed periodically to accommodate customer growth, system expansion, and changing traffic patterns. The MSC uses scanning receivers called "locate receivers" in nearby base stations to determine the signal level of a particular subscriber which appears to be in need of a handoff. In doing so, the MSC is able to find the best neighboring base station which can accept the handoff.

When a new call request arrives from the PSTN or a subscriber, and all of the voice channels in a particular base station are occupied, the MSC will hold the PSTN line open while instructing the current base station to issue a *directed retry* to the subscriber on the FCC. A directed retry forces the subscriber unit to switch to a different control channel (i.e., different base station) for voice channel assignment. Depending on radio propagation effects, the specific location of the subscriber, and the current traffic on the base station to which the subscriber is directed, a directed retry may or may not result in a successful call.

Several factors may contribute to degraded cellular service or dropped or blocked calls. Factors such as the performance of the MSC, the current traffic demand in a geographic area, the specific channel reuse plan, the number of base stations relative to the subscriber population density, the specific propagation conditions between users of the system, and the signal threshold settings for handoffs play major roles in system performance. Maintaining perfect service and call quality in a heavily populated cellular system is practically impossible due to the tremendous system complexity and lack of control in determining radio coverage and customer usage patterns. System operators strive to forecast system growth and do their best to provide suitable coverage and sufficient capacity to avoid co-channel interference within a market, but inevitably some calls will be dropped or blocked. In a large metropolitan market, it is not unusual to have 3-5 percent dropped calls and in excess of 10 percent blocking during extremely heavy traffic conditions.

### 10.1.3 AMPS and ETACS Air Interface

**AMPS and ETACS Channels**: AMPS and ETACS use different physical rate channels for transmission of voice and control information. A *control channel* (also called *setup* or *paging channel*) is used by each base station in the system to simultaneously page subscriber units to alert them of incoming calls and to move connected calls to a voice channel. The FCC constantly transmits data at 10 kbps (8 kbps for ETACS) using binary FSK. FCC transmissions contain either *overhead messages*, mobile station *control messages*, or *control file messages*.The FVC and RVC are used for voice transmissions on the forward and reverse link, respectively. Some of the air interface specifications for AMPS and ETACS are listed in Table 10.1.

While voice channels are in use, three additional signaling techniques are used to maintain supervision between the base station and subscriber unit. The supervisory signals are the *supervisory audio tone (SAT)* and the *signaling tone (ST)*, which will be described. In addition, *wideband data* signaling may be used on a voice channel to provide brief data messages that allow the subscriber and the base station to adjust subscriber power or initiate a handoff. The wideband data is provided by using a blank-and-burst technique, where the voice channel audio is muted and replaced with a brief *burst* of wide band signaling data sent

**Table 10.1 AMPS and ETACS Radio Interface Specifications**

| Parameter | AMPS Specification | ETACS Specification |
|---|---|---|
| Multiple Access | FDMA | FDMA |
| Duplexing | FDD | FDD |
| Channel Bandwidth | 30 kHz | 25 kHz |
| Traffic Channel per RF Channel | 1 | 1 |
| Reverse Channel Frequency | 824 - 849 MHz | 890 - 915 MHz |
| Forward Channel Frequency | 869 - 894 MHz | 935 - 960 MHz |
| Voice Modulation | FM | FM |
| Peak Deviation: Voice Channels<br>Control/Wideband Data | ±12 kHz<br>±8 kHz | ±10 kHz<br>±6.4 kHz |
| Channel Coding for Data Transmission | BCH(40,28) on FC<br>BCH(48,36) on RC | BCH(40,28) on FC<br>BCH(48,36) on RC |
| Data Rate on Control/Wideband Channel | 10 kbps | 8 kbps |
| Spectral Efficiency | 0.33 bps/Hz | 0.33 bps/Hz |
| Number of Channels | 832 | 1000 |

at 10 kbps using FSK (8 kbps for ETACS). Typical blank-and-burst events last for less than 100 ms, so they are virtually imperceptible to the voice channel users.

### Voice Modulation and Demodulation

Prior to frequency modulation, voice signals are processed using a compander, pre-emphasis filter, deviation limiter, and a postdeviation limiter filter. Figure 10.1 shows a block diagram of the AMPS modulation subsystem. At the receiver, these operations are reversed after demodulation.

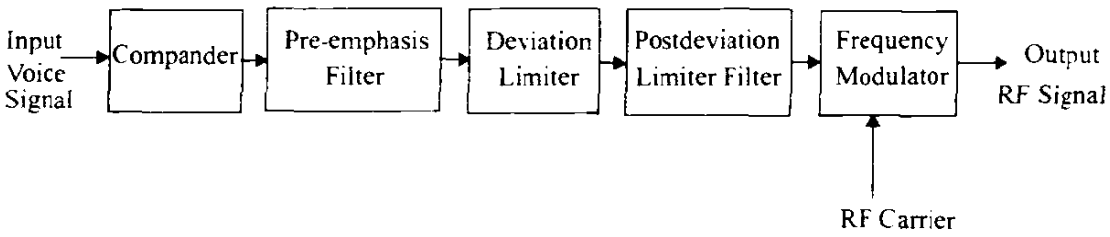



Figure 10.1
AMPS voice modulation process.

**Compander** — In order to accommodate a large speech dynamic range, the input signals need to be compressed in amplitude range before modulation. The companding is done by a 2:1 compander which produces a 1 dB increase in output level for every 2 dB increase in input level. The characteristics are specified

such that a nominal 1 kHz reference input tone at a nominal volume should produce a ±2.9 kHz peak frequency deviation of the transmitted carrier. Companding confines the energy to the 30 kHz channel bandwidth and generates a quieting effect during a speech burst. At the receiver the inverse of compression is performed, thus assuring the restoral of the input voice level with a minimum of distortion.

**Pre-emphasis** — The output of the compressor is passed through a pre-emphasis filter which has a nominal 6 dB/octave highpass response between 300 Hz and 3 kHz.

**Deviation Limiter** — The deviation limiter ensures that the maximum frequency deviation at the mobile station is limited to ±12 kHz (±10 kHz for ETACS). The supervisory signals and wideband data signals are excluded from this restriction.

**Postdeviation Limiter Filter** — The output of the deviation limiter is filtered using a postdeviation limiter filter. This is a low pass filter specified to have an attenuation (relative to the response at 1 kHz) which is greater than or equal to $40\log_{10}(f(Hz)/3000)$dB in the frequency ranges between 3 kHz to 5.9 kHz and 6.1 kHz to 15 kHz. For frequencies between 5.9 and 6.1 kHz, the attenuation (relative to the value at 1 kHz) must be greater than 35 dB, and for 15 kHz and above, the attenuation must be greater than 28 dB (above that at 1 kHz). The postdeviation limiter filter ensures that the specifications on limitations of emission outside the specified band are met, and it ensures that the 6 kHz SAT tones, which are always present during a call, do not interfere with the transmitted speech signal.

**Supervisory Signals (SAT and ST tones)**

The AMPS and ETACS systems provide supervisory signals during voice channel transmissions which allow each base station and its subscribers to confirm they are properly connected during a call. The SAT always exists during the use of any voice channel.

The AMPS and ETACS systems use three SAT signals which are tones at frequencies of either 5970 Hz, 6000 Hz, or 6030 Hz. A given base station will constantly transmit one of the three SAT tones on each voice channel while it is in use. The SAT is superimposed on the voice signal on both the forward and reverse links and is barely audible to a user. The particular frequency of the SAT denotes the particular base station location for a given channel and is assigned by the MSC for each call. Since a highly built up cellular system might have as many as three co-channel base stations in a small geographic region, the SAT enables the subscriber unit and the base station to know which of the three co-channel base stations is handling the call.

When a call is set up and a voice channel assignment is issued, the FVC at the base station immediately begins transmission of the SAT. As the subscriber unit begins monitoring the FVC, it must detect, filter, and demodulate the SAT

coming from the base station and then reproduce the same tone for continuous transmission back to the base station on the RVC. This "handshake" is required by AMPS and ETACS to dedicate a voice channel. If the SAT is not present or improperly detected within a one second interval, both the base station and subscriber unit cease transmission, and the MSC uses the vacated channel for new calls. Transmission of SAT by the mobile station is briefly suspended during blank-and-burst data transmissions on the reverse channel. The detection and rebroadcast of SAT must be performed at least every 250 ms at the subscriber unit. Dropped or prematurely terminated cellular calls can often be traced to interference or incorrect detection of the SAT at the subscriber unit or base station.

The signaling tone (ST) is a 10 kbps data burst which signals call termination by the subscriber. It is a special "end-of-call" message consisting of alternating 1s and 0s, which is sent on the RVC by the subscriber unit for 200 ms. Unlike blank-and-burst messages which briefly suspend the SAT transmission, the ST tone must be sent simultaneously with the SAT. The ST signal alerts the base station that the subscriber has ended the call. When a user terminates a call or turns the cellular phone off during a call, an ST tone is automatically sent by the subscriber unit. This allows the base station and the MSC to know that the call was terminated deliberately by the user, as opposed to being dropped by the system.

**Wideband Blank-and-Burst Encoding**

The AMPS voice channels carry wideband (10 kbps) data streams for blank-and-burst signaling. ETACS uses 8 kbps blank-and-burst transmissions. The wideband data stream is encoded such that each NRZ binary one is represented by a zero-to-one transition, and each NRZ binary zero is represented by a one-to-zero transition. This type of coding is called Manchester (or biphase) coding. The advantage of using a Manchester code in a voice channel is that the energy of the Manchester coded signal is concentrated at the transmission rate frequency of 10 kHz (see chapter 5), and little energy leaks into the audio band below 4 kHz. Therefore, a burst of data transmitted over a voice channel can be easily detected within a 30 kHz RF channel, is barely audible to a user, and can be passed over phone lines that have dc blocking circuits. The Manchester code is applied to both control channel and voice channel blank-and-burst transmissions.

The Manchester coded wideband data stream is filtered and channel coded using BCH block codes. Wideband data bursts on the voice channels occur in short blasts of repetitive blocks having the same length as the error correction code. The (40, 28) BCH codes are used on forward voice channel blank-and-burst transmissions and are able to correct 5 errors whereas (48, 36) BCH block codes are used on the reverse voice channel blank-and-burst transmissions. The encoded data are used to modulate the transmitter carrier using direct fre-

quency-shift keying. Binary ones correspond to a frequency deviation of +8 kHz and binary zeros correspond to a deviation of -8 kHz (±6.4 kHz for ETACS).

A wide array of commands may be sent to and from subscriber units using blank-and-burst signaling. These are defined in the AMPS and ETACS air interface specifications.

### 10.1.4 N-AMPS

To increase capacity in large AMPS markets, Motorola developed an AMPS-like system called N-AMPS (narrowband AMPS) [EIA91]. N-AMPS provides three users in a 30 kHz AMPS channel by using FDMA and 10 kHz channels, and provides three times the capacity of AMPS. By replacing AMPS channels with three N-AMPS channels at one time, service providers are able to provide more trunked radio channels (and thus a much better grade of service) at base stations in heavily populated areas. N-AMPS uses the SAT and ST signaling and blank-and-burst functions in exactly the same manner as AMPS, except the signaling is done by using subaudible data streams.

Since 10 kHz channels are used, the FM deviation is decreased. This, in turn, reduces the $S/(N+I)$ which degrades the audio quality with respect to AMPS. To counteract this, N-AMPS uses voice companding to provide a "synthetic" voice channel quieting.

N-AMPS specifies a 300 Hz high pass audio filter for each voice channel so that supervisory and signaling data may be sent without blanking the voice. The SAT and ST signaling is sent using a continuous 200 bps NRZ data stream that is FSK modulated. SAT and ST are called DSAT and DST in N-AMPS because they are sent digitally and repetitiously in small, predefined code blocks. There are seven different 24 bit DSAT codewords which may be selected by the MSC, and the DSAT codeword is constantly repeated by both the base station and mobile during a call. The DST signal is simply the binary inverse of the DSAT. The seven possible DSATs and DSTs are specially designed to provide a sufficient number of alternating 0's and 1's so that dc blocking may be conveniently implemented by receivers.

The voice channel signaling is done with 100 bps Manchester encoded FSK data and is sent in place of DSAT when traffic must be passed on the voice channel. As with AMPS wideband signaling, there are many messages that may be passed between the base station and subscriber unit, and these are transmitted in N-AMPS using the same BCH codes as in AMPS with a predefined format of 40 bit blocks on the FVC and 48 bit blocks on the RVC.

## 10.2 United States Digital Cellular (IS-54)

The first generation analog AMPS system was not designed to support the current demand for capacity in large cities. Cellular systems which use digital modulation techniques (called digital cellular) offer large improvements in

capacity and system performance [Rai91]. After extensive research and comparison by major cellular manufacturers in the late 1980s, the United States Digital Cellular System (USDC) was developed to support more users in a fixed spectrum allocation. USDC is a time division multiple access (TDMA) system which supports three full-rate users or six half-rate users on each AMPS channel. Thus, USDC offers as much as six times the capacity of AMPS. The USDC standard uses the same 45 MHz FDD scheme as AMPS. The dual mode USDC/AMPS system was standardized as Interim Standard 54 (IS-54) by the Electronic Industries Association and Telecommunication Industry Association (EIA/TIA) in 1990 [EIA90].

The USDC system was designed to share the same frequencies, frequency reuse plan, and base stations as AMPS, so that base stations and subscriber units could be equipped with both AMPS and USDC channels within the same piece of equipment. By supporting both AMPS and USDC, cellular carriers are able to provide new customers with USDC phones and may gradually replace AMPS base stations with USDC base stations, channel by channel, over time. Because USDC maintains compatibility with AMPS in a number of ways, USDC is also known as Digital AMPS (D-AMPS).

Currently, in rural areas where immature analog cellular systems are in use, only 666 of the 832 AMPS channels are activated (that is, some rural cellular operators are not yet using the extended spectrum allocated to them in 1989). In these markets, USDC channels may be installed in the extended spectrum to support USDC phones which roam into the system from metropolitan markets. In urban markets where every cellular channel is already in use, selected frequency banks in high traffic base stations are converted to the USDC digital standard. In larger cities, this gradual changeover results in a temporary increase in interference and dropped calls on the analog AMPS system, since each time a base station is changed over to digital, the number of analog channels in a geographic area is decreased. Thus, the changeover rate from analog to digital must carefully match the subscriber equipment transition in the market.

The smooth transition from analog to digital in the same radio band was a key force in the development of the USDC standard. In practice, only cities with capacity shortages (such as New York and Los Angeles) have aggressively changed out AMPS to USDC, while smaller cities are waiting until more subscribers are equipped with USDC phones. The introduction of N-AMPS and a competing, digital spread spectrum standard (IS-95, described later in this chapter) has delayed the widespread deployment of USDC throughout the U.S.

To maintain compatibility with AMPS phones, USDC forward and reverse control channels use exactly the same signaling techniques as AMPS. Thus, while USDC voice channels use 4-ary $\pi/4$ DQPSK modulation with a channel rate of 48.6 kbps, the forward and reverse control channels are no different than AMPS and use the same 10 kbps FSK signaling scheme and the same standard-

ized control channels. A recent standard, IS-136 (formerly IS-54 Rev.C) also includes $\pi/4$ DQPSK modulation for the USDC control channels [Pad95]. IS-54 Rev. C was introduced to provide 4-ary keying instead of FSK on dedicated USDC control channels in order to increase control channel data rate, and to provide specialized services such as paging and short messaging between private subscriber user groups.

### 10.2.1 USDC Radio Interface

In order to ensure a smooth transition from AMPS to USDC, the IS-54 system is specified to operate using both AMPS and USDC standards (dual mode) which makes roaming between the two systems possible with a single phone. The IS-54 system uses the same frequency band and channel spacing as AMPS and supports multiple USDC users on each AMPS channel. The USDC scheme uses TDMA which, as described in Chapter 8, has the flexibility of incorporating even more users within a single radio channel as lower bit rate speech coders become available. Table 10.2 summarizes the air interface for USDC.

**Table 10.2 USDC Radio Interface Specifications Summary**

| Parameter | USDC IS-54 Specification |
|---|---|
| Multiple Access | TDMA/FDD |
| Modulation | $\pi/4$ DQPSK |
| Channel Bandwidth | 30 kHz |
| Reverse Channel Frequency Band | 824 - 849 MHz |
| Forward Channel Frequency Band | 869 - 894 MHz |
| Forward and Reverse Channel Data Rate | 48.6 kbps |
| Spectrum Efficiency | 1.62 bps/Hz |
| Equalizer | Unspecified |
| Channel Coding | 7 bit CRC and rate 1/2 convolutional coding of constraint length 6 |
| Interleaving | 2 slot interleaver |
| Users per Channel | 3 (full-rate speech coder of 7.95 kbps/user)<br>6 (with half-rate speech coder of 3.975 kbps/user) |

**USDC Channels** — The USDC control channels are identical to the analog AMPS control channels. In addition to the forty-two primary AMPS control channels, USDC specifies forty-two additional control channels called the *secondary control channels*. Thus, USDC has twice as many control channels as AMPS, so that double the amount of control channel traffic can be paged throughout a market. The secondary control channels conveniently allow carriers to dedicate them for USDC-only use, since AMPS phones do not monitor or

decode the secondary control channels. When converting an AMPS system to USDC/AMPS, a carrier may decide to program the MSC to send pages for USDC mobiles over the secondary control channels only, while having existing AMPS traffic sent only on the AMPS control channels. For such a system, USDC subscriber units would be programmed to automatically monitor only the secondary forward control channels when operating in the USDC mode. Over time, as USDC users begin to populate the system to the point that additional control channels are required, USDC pages would eventually be sent simultaneously over both the primary and secondary control channels.

A USDC voice channel occupies 30 kHz of bandwidth in each of the forward and reverse links, and supports a maximum of three users (as compared to a single AMPS user). Each voice channel supports a TDMA scheme that provides six time slots. For full rate speech, three users utilize the six time slots in an equally spaced fashion. For example, user 1 occupies time slots 1 and 4, user 2 occupies time slots 2 and 5, and user 3 occupies time slots 3 and 6. For half-rate speech, each user occupies one time slot per frame.

On each USDC voice channel, there are actually four data channels which are provided simultaneously. The most important data channel, as far as the end user is concerned, is the digital traffic channel (DTC) which carries user information (i.e., speech or user data), and the other three channels carry supervisory information within the cellular system. The reverse DTC (RDTC) carries speech data from the subscriber to the base station, and the forward DTC (FDTC) carries user data from the base station to the subscriber. The three supervisory channels include the Coded Digital Verification Color Code (CDVCC), the Slow Associated Control Channel (SACCH), and the Fast Associated Control Channel (FACCH).

The CDVCC is a 12 bit message sent in every time slot, and is similar in functionality to the SAT used in AMPS. The CDVCC is an 8 bit number ranging between 1 and 255, which is protected with 4 additional channel coding bits from a shortened (12,8) Hamming code. The base station transmits a CDVCC value on the forward voice channel and each subscriber using the TDMA channel must receive, decode, and retransmit the same CDVCC value to the base station on the reverse voice channel. If the CDVCC "handshake" is not properly completed, then the time slot will be relinquished for other users and the subscriber transmitter will be turned off automatically.

The SACCH is sent in every time slot, and provides a signaling channel in parallel with the digital speech. The SACCH carries various control and supervisory messages between the subscriber unit and the base station. SACCH provides single messages over many consecutive time slots and is used to communicate power level changes or handoff requests. The SACCH is also used by the mobile unit to report the results of signal strength measurements of

neighboring base stations so that the base station may implement mobile assisted handoff (MAHO).

The FACCH is another signaling channel which is used to send important control or specialized traffic data between the base station and mobile units. The FACCH data, when transmitted, takes the place of user information data (such as speech) within a frame. FACCH may be thought of as a blank-and-burst transmission in USDC. FACCH supports the transmission of dual tone multiple frequency (DTMF) information from touch tone keypads, call release instructions, flash hook instructions, and MAHO or subscriber status requests. FACCH also provides tremendous flexibility in that it allows carriers to handle traffic internal to the cellular network if the DTC is idle during some of the TDMA time slots. As discussed subsequently, FACCH data is treated similarly to speech data in the way it is packaged and interleaved to fit in a time slot. However, unlike the speech data which protects only certain bits with channel coding in the USDC time slot, FACCH data uses a 1/4 rate convolutional channel code to protect all bits that are transmitted in a time slot.

**Frame Structure for USDC Traffic Channels** — As shown in Figure 10.2, a TDMA frame in the USDC system consists of six timeslots that support three full-rate traffic channels or six half-rate traffic channels. The TDMA frame length is 40 milliseconds. Since USDC uses FDD, there are forward and reverse channel time slots operating simultaneously. Each time slot is designed to carry interleaved speech data from two adjacent frames of the speech coder. (The frame length for the speech coder is 20 ms, half the duration of a TDMA frame). The USDC standard requires that data from two adjacent speech coder frames be sent in a particular time slot. The USDC speech coder, discussed in more detail below, produces 159 bits of raw, speech coded data in a frame lasting 20 ms, but channel coding brings each coded speech frame up to 260 bits for the same 20 ms period. If FACCH is sent instead of speech data, then one frame of speech coding data is replaced with a block of FACCH data, and the FACCH data within a time slot is actually made up of FACCH data from two adjacent FACCH data blocks.

In the reverse voice channel, each time slot consists of two bursts of 122 bits and one burst of 16 bits (for a total of 260 bits per time slot) from two interleaved speech frames (or FACCH data blocks). In addition, 28 sync bits, 12 bits of SACCH data, 12 bits of Coded Digital Verification Color Code (CDVCC) and 12 bits of guard and ramp-up time are sent in a reverse channel time slot.

On the forward voice channel, each time slot consists of two 130 bit bursts of data from two consecutive, interleaved speech frames (or FACCH data if speech is not sent), 28 sync bits, 12 bits of SACCH data, 12 bits of CDVCC, and 12 reserved bits. There are a total of 324 bits per time slot on both the forward and reverse channels, and each time slot lasts for 6.667 ms.

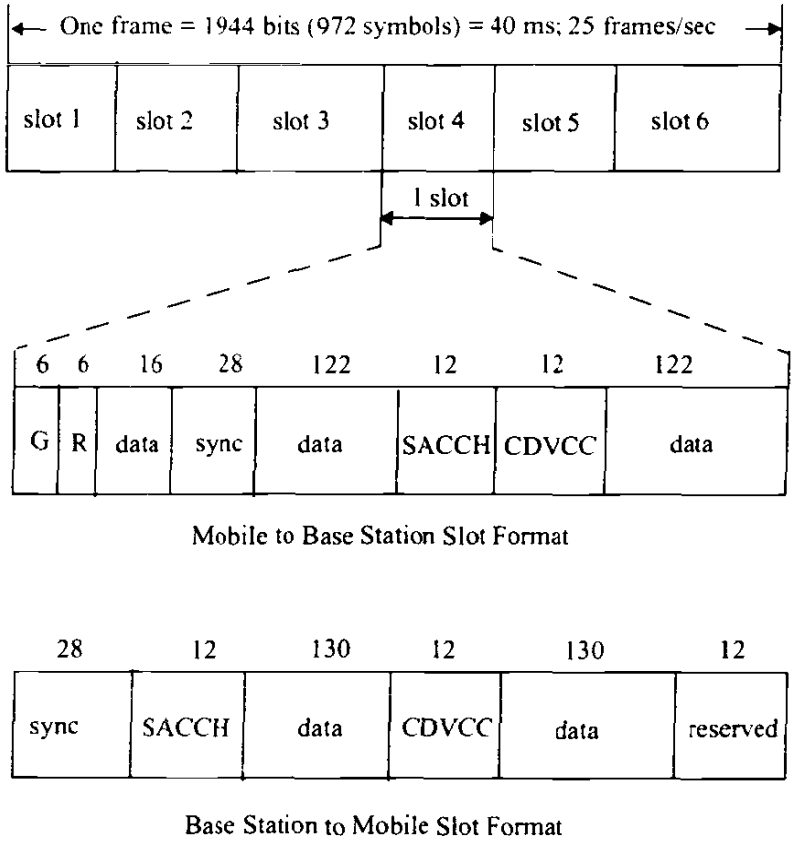



Figure 10.2
The USDC slot and frame structure on the forward and reverse link.

The time slots in the forward and reverse channels are staggered in time so that time slot 1 of the $N$th frame in the forward channel starts exactly one time slot plus 44 symbols (i.e., 206 symbols = 412 bits) after the beginning of time slot 1 of the $N$th frame on the reverse channel. As discussed in Chapter 8, this allows each mobile to simply use a transmit/receive switch, rather than a duplexer, for full duplex operation with the forward and reverse links. USDC provides the ability to adjust the time stagger between forward and reverse channel time slots in integer increments of half of a time slot so that the system may synchronize new subscribers that are assigned a time slot.

**Speech Coding** — The USDC speech coder is called the Vector Sum Excited Linear Predictive coder (VSELP). This belongs to the class of Code Excited Linear Predictive coders (CELP) or Stochastically Excited Linear Predictive coders (SELP). As discussed in Chapter 7, these coders are based upon codebooks which determine how to quantize the residual excitation signal. The VSELP algorithm uses a code book that has a predefined structure such that the number of computations required for the codebook search process is significantly

reduced. The VSELP algorithm was developed by a consortium of companies and the Motorola implementation was chosen for the IS-54 standard. The VSELP coder has an output bit rate of 7950 bps and produces a speech frame every 20 ms. In one second, fifty speech frames, each containing 159 bits of speech, are produced by the coder for a particular user.

**Channel Coding** — The 159 bits within a speech coder frame are divided into two classes according to their perceptual significance. There are 77 class-1 bits and 82 class-2 bits. The class-1 bits, being the most significant bits, are error protected using a rate 1/2 convolutional code of constraint length $K = 6$. In addition to convolutional coding, the twelve most significant bits among the class-1 bits are block coded using a 7 bit CRC error detection code. This ensures that the most important speech coder bits are detected with a high degree of probability at the receiver. The class-2 bits, being perceptually less significant, have no error protection added to them. After channel coding, the 159 bits in each speech coder frame are represented by 260 channel coded bits, and the gross bit rate of the speech coder with added channel coding is 13.0 kbps. Figure 10.3 illustrates the channel coding operations for the speech coded data.

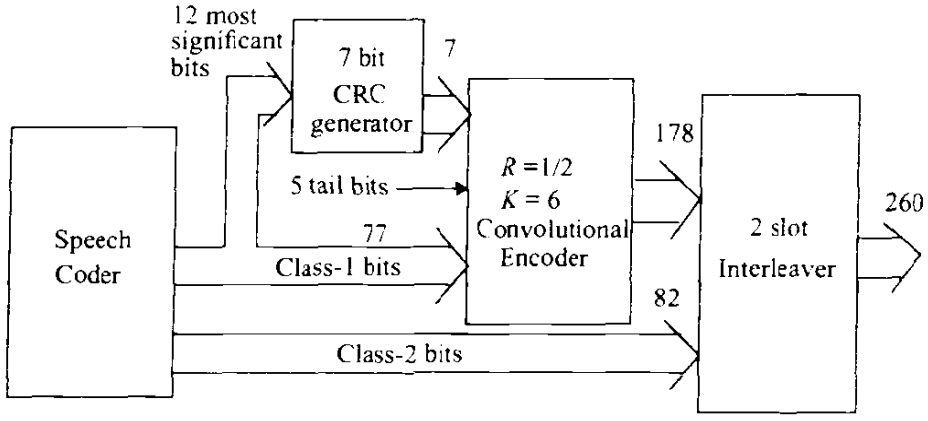



Figure 10.3
Error protection for USDC speech coder output.

The channel coding used for the FACCH data is different from that used for the speech coded data. A FACCH data block contains forty-nine bits of data per each 20 ms frame. A 16 bit CRC codeword is appended to each FACCH data block, providing a coded FACCH word of sixty-five bits. The 65 bit word is then passed through a rate 1/4 convolutional coder of constraint length six in order to yield 260 bits of FACCH data per each 20 ms frame. A FACCH data block occupies the same amount of bandwidth as a single frame of coded speech, and in this manner speech data on the DTC can be replaced with coded FACCH data. Interleaving of DTC and FACCH data is handled identically in USDC.

The SACCH data word consists of 6 bits during each 20 ms speech frame. Each raw SACCH data word is passed through a rate 1/2 convolutional coder of

constraint length five to produce twelve coded bits during every 20 ms interval, or twenty-four bits during each USDC frame.

**Interleaving** — Before transmission, the encoded speech data is interleaved over two time slots with the speech data from adjacent speech frames. In other words, each time slot contains exactly half of the data from each of two sequential speech coder frames. The speech data are placed into a rectangular $26 \times 10$ interleaver as shown in Figure 10.4. The data is entered into the columns of the interleaving array, and the two consecutive speech frames are referred to as $x$ and $y$, where $x$ is the previous speech frame and $y$ is the present or most recent speech frame. It can be seen from Figure 10.4 that only 130 of the necessary 260 bits are provided for frames $x$ and $y$. The encoded speech data for the two adjacent frames are placed into the interleaver in such a manner that intermixes the class-2 bits and the class-1 bits. The speech data is then transmitted row-wise out of the interleaver. The interleaving approach for coded FACCH blocks is identical to that used for speech data. A 6 bit SACCH message word, on the other hand, is coded using a rate 1/2 convolutional code and uses an incremental interleaver that spans over twelve consecutive time slots [EIA90].

$$\begin{bmatrix}
0x & 26x & 52x & 78x & 104x & 130x & 156x & 182x & 208x & 234x \\
1y & 27y & 53y & 79y & 105y & 131y & 157y & 183y & 209y & 235y \\
2x & 28x & 54x & 80x & 106x & 132x & 158x & 184x & 210x & 236x \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
12x & 38x & 64x & 90x & 116x & 142x & 168x & 194x & 220x & 246x \\
13y & 39y & 65y & 91y & 117y & 143y & 169y & 195y & 221y & 247y \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
24x & 50x & 76x & 102x & 128x & 154x & 180x & 206x & 232x & 258x \\
25y & 51y & 77y & 103y & 129y & 155y & 181y & 207y & 233y & 259y
\end{bmatrix}$$

Figure 10.4
Interleaving for two adjacent speech coder frames in USDC.

**Modulation** — To be compatible with AMPS, USDC uses 30 kHz channels. On control (paging) channels, USDC and AMPS use identical 10 kbps binary FSK with Manchester coding. On voice channels, the FM modulation is replaced with digital modulation having a gross bit rate of 48.6 kbps. In order to achieve this bit rate in a 30 kHz channel, the modulation requires a spectral efficiency of 1.62 bps/Hz. Also, to limit adjacent channel interference (ACI), spectral shaping on the digital channel must be used.

The spectral efficiency requirements are satisfied by conventional pulse-shaped, four-phase, modulation schemes such as QPSK and OQPSK. However,

as discussed in Chapter 5, symmetric differential phase-shift keying, commonly known as $\pi/4$-DQPSK, has several advantages when used in a mobile radio environment and is the modulation used for USDC. The channel symbol rate is 24.3 ksps, and the symbol duration is $41.1523\,\mu s$.

Pulse shaping is used to reduce the transmission bandwidth while limiting the intersymbol interference (ISI). At the transmitter, the signal is filtered using a square root raised cosine filter with a rolloff factor equal to 0.35. The receiver may also employ a corresponding square root raised cosine filter. Once pulse-shaping is performed on phase-shift keying, it becomes a linear modulation technique, which requires linear amplification in order to preserve the pulse shape. Nonlinear amplification results in destruction of the pulse shape and expansion of the signal bandwidth. The use of pulse shaping with $\pi/4$-DQPSK supports the transmission of three (and eventually six) speech signals in a 30 kHz channel bandwidth with adjacent-channel protection of 50 dB.

**Demodulation**— The type of demodulation and decoding used at the receiver is left up to the manufacturer. As shown in Chapter 5, differential detection may be performed at IF or baseband. The latter implementation may be done conveniently using a simple discriminator or digital signal processor (DSP). This not only reduces the cost of the demodulator, but also simplifies the RF circuitry. DSPs also support the implementation of the USDC equalizer as well as dual mode functionality.

**Equalization** — Measurements conducted in 900 MHz mobile channels revealed that the rms delay spreads are less than 15 $\mu s$ at 99% of all locations in four U.S. cities and are less than 5 $\mu s$ for nearly 80% of all locations [Rap90]. For a system employing DQPSK modulation at a symbol rate of 24.3 ksps, if the bit error rate due to intersymbol interference becomes intolerable for a $\sigma/T$ value of 0.1 (where $\sigma$ is the rms delay spread and $T$ is the symbol duration), then the maximum value of rms delay spread that can be tolerated is $4.12\mu s$. If the rms delay spreads exceeds this, it is necessary to use equalization in order to reduce the BER. Work by Rappaport, Seidel, and Singh [Rap90] showed that the rms delays spread exceeds $4\mu s$ at about 25% of the locations in four cities, so an equalizer was specified for USDC, although the specific implementation is not specified in the IS-54 standard.

One equalizer proposed for USDC is a Decision Feedback Equalizer (DFE) [Nar90] consisting of four feedforward taps and one feedback tap, where the feedforward taps have a 1/2 symbol spacing. This type of fractional spacing makes the equalizer robust against sample timing jitter. The coefficients of the adaptive filter are updated using the recursive least squares (RLS) algorithm described in Chapter 6. Many proprietary implementations for the USDC equalizer have been developed by equipment manufacturers.

### 10.2.2 United States Digital Cellular Derivatives (IS-94 and IS-136)

The added networking features provided in IS-54 have led to new types of wireless services and transport topologies. Because TDMA provides MAHO capability, mobiles are able to sense channel conditions and report these to the base station. This, in turn, allows greater flexibility in cellular deployment. For example, MAHO is used to support dynamic channel allocation which can be carried out by the base station. This allows an MSC to use a larger number of base stations placed in strategic locations throughout a service area and provides each base station with greater control of its coverage characteristics.

The IS-94 standard exploits capabilities provided by IS-54, and enables cellular phones to interface directly with private branch exchanges (PBX). By moving the intelligence of an MSC closer to the base station, it becomes possible to provide wireless PBX services in a building or on a campus, while using small base stations (microcells) that can be placed in closets throughout a building. IS-94 specifies a technique to provide private, or closed, cellular systems that use nonstandard control channels. IS-94 systems were introduced in 1994, and are proliferating throughout office buildings and hotels.

The IS-54 Rev.C standard provides 48.6 kbps control channel signaling on the USDC-only control channels and 10 kbps FSK control channels on the original AMPS channels. However, closed network capabilities are not fully developed under IS-54 Rev.C. A new interim standard, IS-136, has been developed to provide a host of new features and services that positions the cellular carriers for competition from PCS. IS-136 specifies short messaging capabilities and private, user group features, making it well-suited for wireless PBX applications and paging applications. Furthermore, IS-136 specifies a "sleep mode" that instructs compatible cellular phones to conserve battery power. IS-136 subscriber terminals are not compatible with those produced for IS-54, since IS-136 uses 48.6 kbps control channels exclusively on all control channels (the 10 kbps FSK is not supported). This allows IS-136 modems to be more cost effective, since only the 48.6 kbps modem is needed in each portable unit.

## 10.3 Global System for Mobile (GSM)

Global System for Mobile (GSM) is a second generation cellular system standard that was developed to solve the fragmentation problems of the first cellular systems in Europe. GSM is the world's first cellular system to specify digital modulation and network level architectures and services. Before GSM, European countries used different cellular standards throughout the continent, and it was not possible for a customer to use a single subscriber unit throughout Europe. GSM was originally developed to serve as the pan-European cellular service and promised a wide range of network services through the use of ISDN. GSM's success has exceeded the expectations of virtually everyone, and it is now the world's most popular standard for new cellular radio and personal communi-

cations equipment throughout the world. It is predicted that by the year 2000, there will be between 20 and 50 million GSM subscribers worldwide [Mou92], [Dec93].

The task of specifying a common mobile communication system for Europe in the 900 MHz band was taken up by the GSM (Groupe spe'cial mobile) committee which was a working group of the Confe'rence Europe'ene Postes des et Te'le'communication (CEPT). Recently, GSM has changed its name to the Global System for Mobile Communications for marketing reasons [Mou92]. The setting of standards for GSM is currently under the aegis of the European Technical Standards Institute (ETSI).

GSM was first introduced into the European market in 1991. By the end of 1993, several nonEuropean countries in South America, Asia, and Australia had adopted GSM and the technically equivalent offshoot, DCS 1800, which supports Personal Communication Services (PCS) in the 1.8 GHz to 2.0 GHz radio bands recently created by governments throughout the world.

### 10.3.1 GSM Services and Features

GSM services follow ISDN guidelines and are classified as either *teleservices* or *data services*. Teleservices include standard mobile telephony and mobile-originated or base-originated traffic. Data services include computer-to-computer communication and packet-switched traffic. User services may be divided into three major categories:

- **Telephone services**, including emergency calling and facsimile. GSM also supports Videotex and Teletex, though they are not integral parts of the GSM standard.
- **Bearer services** or **data services** which are limited to layers 1, 2, and 3 of the open system interconnection (OSI) reference model (see Chapter 9). Supported services include packet switched protocols and data rates from 300 bps to 9.6 kbps. Data may be transmitted using either a transparent mode (where GSM provides standard channel coding for the user data) or nontransparent mode (where GSM offers special coding efficiencies based on the particular data interface).
- **Supplementary ISDN services**, are digital in nature, and include call diversion, closed user groups, and caller identification, and are not available in analog mobile networks. Supplementary services also include the *short messaging service* (SMS) which allows GSM subscribers and base stations to transmit alphanumeric pages of limited length (160 7 bit ASCII characters) while simultaneously carrying normal voice traffic. SMS also provides *cell broadcast*, which allows GSM base stations to repetitively transmit ASCII messages with as many as fifteen 93-character strings in concatenated fashion. SMS may be used for safety and advisory applications, such as the broadcast of highway or weather information to all GSM subscribers within

reception range.

From the user's point of view, one of the most remarkable features of GSM is the *Subscriber Identity Module* (SIM), which is a memory device that stores information such as the subscriber's identification number, the networks and countries where the subscriber is entitled to service, privacy keys, and other user-specific information. A subscriber uses the SIM with a 4-digit personal ID number to activate service from any GSM phone. SIM's are available as smart cards (credit card sized cards that may be inserted into any GSM phone) or plug-in modules, which are less convenient than the SIM cards but are nonetheless removable and portable. Without a SIM installed, all GSM mobiles are identical and nonoperational. It is the SIM that gives GSM subscriber units their identity. Subscribers may plug their SIM into any suitable terminal — such as a hotel phone, public phone, or any portable or mobile phone — and are then able to have all incoming GSM calls routed to that terminal and have all outgoing calls billed to their home phone, no matter where they are in the world.

A second remarkable feature of GSM is the on-the-air privacy which is provided by the system. Unlike analog FM cellular phone systems which can be readily monitored, it is virtually impossible to eavesdrop on a GSM radio transmission. The privacy is made possible by encrypting the digital bit stream sent by a GSM transmitter, according to a specific secret cryptographic key that is known only to the cellular carrier. This key changes with time for each user. Every carrier and GSM equipment manufacturer must sign the Memorandum of Understanding (MoU) before developing GSM equipment or deploying a GSM system. The MoU is an international agreement which allows the sharing of cryptographic algorithms and other proprietary information between countries and carriers.

### 10.3.2 GSM System Architecture

The GSM system architecture consists of three major interconnected subsystems that interact between themselves and with the users through certain network interfaces. The subsystems are the *Base Station Subsystem* (BSS), *Network and Switching Subsystem* (NSS), and the *Operation Support Subsystem* (OSS). The *Mobile Station* (MS) is also a subsystem, but is usually considered to be part of the BSS for architecture purposes. Equipment and services are designed within GSM to support one or more of these specific subsystems.

The BSS, also known as the *radio subsystem*, provides and manages radio transmission paths between the mobile stations and the Mobile Switching Center (MSC). The BSS also manages the radio interface between the mobile stations and all other subsystems of GSM. Each BSS consists of many Base Station Controllers (BSCs) which connect the MS to the NSS via the MSCs. The NSS manages the switching functions of the system and allows the MSCs to commu-

nicate with other networks such as the PSTN and ISDN. The OSS supports the operation and maintenance of GSM and allows system engineers to monitor, diagnose, and troubleshoot all aspects of the GSM system. This subsystem interacts with the other GSM subsystems, and is provided solely for the staff of the GSM operating company which provides service facilities for the network.

Figure 10.5 shows the block diagram of the GSM system architecture. The Mobile Stations (MS) communicate with the Base Station Subsystem (BSS) over the radio air interface. The BSS consists of many BSCs which connect to a single MSC, and each BSC typically controls up to several hundred *Base Transceiver Stations* (BTSs). Some of the BTSs may be co-located at the BSC, and others may be remotely distributed and physically connected to the BSC by microwave link or dedicated leased lines. Mobile handoffs (called *handovers*, or HO, in the GSM specification) between two BTSs under the control of the same BSC are handled by the BSC, and not the MSC. This greatly reduces the switching burden of the MSC.

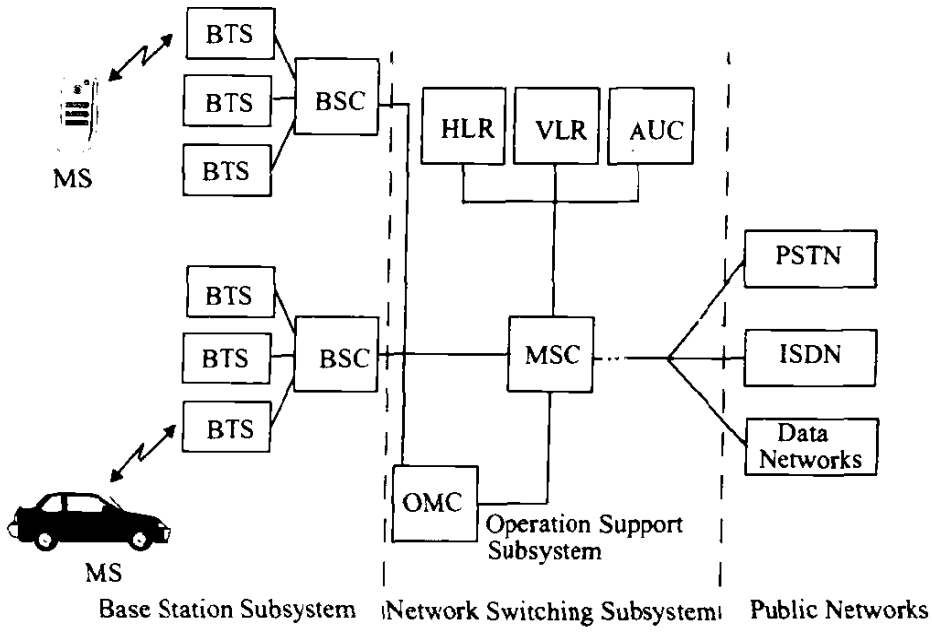



Figure 10.5
GSM system architecture.

As shown in Figure 10.6, the interface which connects a BTS to a BSC is called the *Abis interface*. The Abis interface carries traffic and maintenance data, and is specified by GSM to be standardized for all manufacturers. In practice, however, the Abis for each GSM base station manufacturer has subtle differences, thereby forcing service providers to use the same manufacturer for the BTS and BSC equipment.

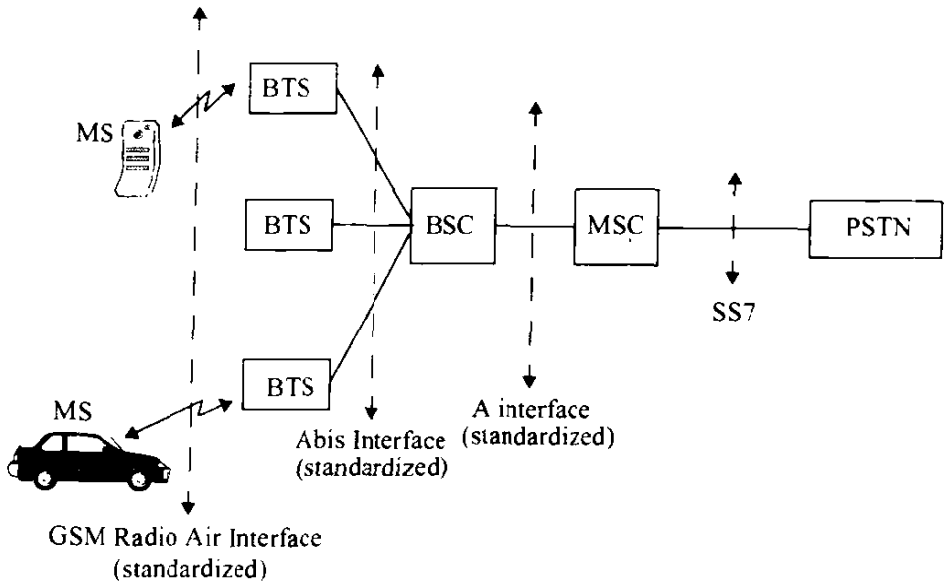



*Figure 10.6*
The various interfaces used in GSM.

The BSCs are physically connected via dedicated/leased lines or microwave link to the MSC. The interface between a BSC and a MSC is called the *A interface*, which is standardized within GSM. The A interface uses an SS7 protocol called the *Signaling Correction Control Part* (SCCP) which supports communication between the MSC and the BSS, as well as network messages between the individual subscribers and the MSC. The A interface allows a service provider to use base stations and switching equipment made by different manufacturers.

The NSS handles the switching of GSM calls between external networks and the BSCs in the radio subsystem and is also responsible for managing and providing external access to several customer databases. The MSC is the central unit in the NSS and controls the traffic among all of the BSCs. In the NSS, there are three different databases called the *Home Location Register* (HLR), *Visitor Location Register* (VLR), and the *Authentication Center* (AUC). The HLR is a database which contains subscriber information and location information for each user who resides in the same city as the MSC. Each subscriber in a particular GSM market is assigned a unique *International Mobile Subscriber Identity* (IMSI), and this number is used to identify each home user. The VLR is a database which temporarily stores the IMSI and customer information for each roaming subscriber who is visiting the coverage area of a particular MSC. The VLR is linked between several adjoining MSCs in a particular market or geographic region and contains subscription information of every visiting user in the area. Once a roaming mobile is logged in the VLR, the MSC sends the necessary information to the visiting subscriber's HLR so that calls to the roaming mobile can be appropriately routed over the PSTN by the roaming user's HLR. The

Authentication Center is a strongly protected database which handles the authentication and encryption keys for every single subscriber in the HLR and VLR. The Authentication Center contains a register called the *Equipment Identity Register* (EIR) which indentifies stolen or fraudulently altered phones that transmit identity data that does not match with information contained in either the HLR or VLR.

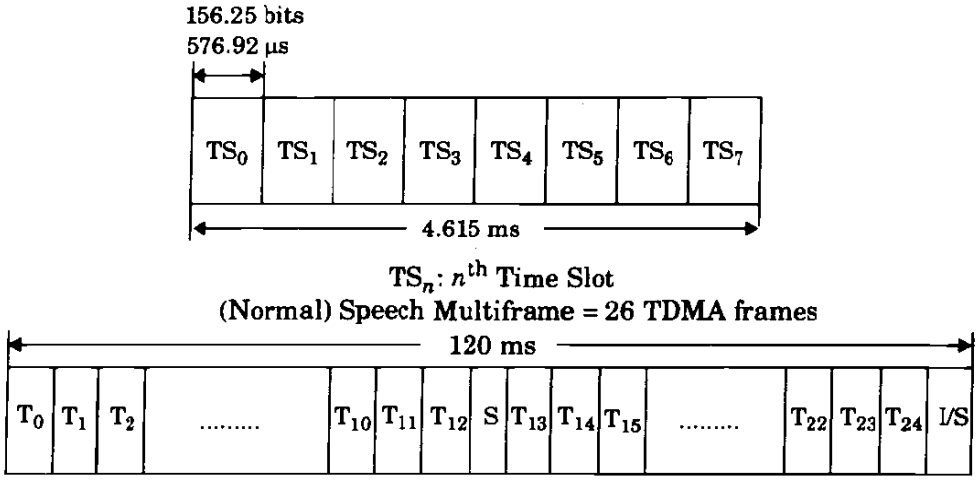
The OSS supports one or several *Operation Maintenance Centers* (OMC) which are used to monitor and maintain the performance of each MS, BS, BSC, and MSC within a GSM system. The OSS has three main functions, which are 1) to maintain all telecommunications hardware and network operations with a particular market, 2) manage all charging and billing procedures, and 3) manage all mobile equipment in the system. Within each GSM system, an OMC is dedicated to each of these tasks and has provisions for adjusting all base station parameters and billing procedures, as well as for providing system operators with the ability to determine the performance and integrity of each piece of subscriber equipment in the system.

### 10.3.3 GSM Radio Subsystem

GSM utilizes two bands of 25 MHz which have been set aside for system use in all member countries. The 890-915 MHz band is used for subscriber-to-base transmissions (reverse link), and the 935-960 MHz band is used for base-to-subscriber transmissions (forward link). GSM uses FDD and a combination of TDMA and FHMA schemes to provide base stations with simultaneous access to multiple users. The available forward and reverse frequency bands are divided into 200 kHz wide channels called ARFCNs (Absolute Radio Frequency Channel Numbers). The ARFCN denotes a forward and reverse channel pair which is separated in frequency by 45 MHz and each channel is time shared between as many as eight subscribers using TDMA.

Each of the eight subscribers uses the same ARFCN and occupies a unique timeslot (TS) per frame. Radio transmissions on both the forward and reverse link are made at a channel data rate of 270.833 kbps (1625.0/6.0 kbps) using binary $BT$=0.3 GMSK modulation. Thus, the signaling bit duration is 3.692 μs, and the effective channel transmission rate per user is 33.854 kbps (270.833 kpbs/8 users). With GSM overhead (described subsequently), user data is actually sent at a maximum rate of 24.7 kbps. Each TS has an equivalent time allocation of 156.25 channel bits, but of this, 8.25 bits of guard time and 6 total start and stop bits are provided to prevent overlap with adjacent time slots. Each TS has a time duration of 576.92 μs as shown in Figure 10.7, and a single GSM TDMA frame spans 4.615 ms. The total number of available channels within a 25 MHz bandwidth is 125 (assuming no guard band). Since each radio channel consists of 8 time slots, there are thus a total of 1000 traffic channels within GSM. In practical implementations, a guard band of 100 kHz is provided at the

upper and lower end of the GSM spectrum, and only 124 channels are implemented. Table 10.3 summarizes the GSM air interface.




$T_n$: $n^{th}$ TCH frame

S: Slow Associated Control Channel frame

I: Idle frame

Figure 10.7
The Speech Dedicated Control Channel Frame and multiframe structure.

**Table 10.3 GSM Air Interface Specifications Summary**

| Parameter | Specifications |
|---|---|
| Reverse Channel Frequency | 890 - 915 MHz |
| Forward Channel Frequency | 935 - 960 MHz |
| ARFCN Number | 0 to 124 and 975 to 1023 |
| Tx/Rx Frequency Spacing<br>Tx/Rx Time Slot Spacing | 45 MHz<br>3 Time slots |
| Modulation Data Rate | 270.833333 kbps |
| Frame Period | 4.615 ms |
| Users per Frame (Full Rate) | 8 |
| Time slot Period | 576.9 μs |
| Bit Period | 3.692 μs |
| Modulation | 0.3 GMSK |
| ARFCN Channel Spacing | 200 kHz |
| Interleaving (max. delay) | 40 ms |
| Voice Coder Bit Rate | 13.4 kbps |

The combination of a TS number and an ARFCN constitutes a *physical channel* for both the forward and reverse link. Each physical channel in a GSM system can be mapped into different *logical channels* at different times. That is, each specific time slot or frame may be dedicated to either handling traffic data (user data such as speech, facsimile, or teletext data), signaling data (required by the internal workings of the GSM system), or control channel data (from the MSC, base station, or mobile user). The GSM specification defines a wide variety of logical channels which can be used to link the physical layer with the data link layer of the GSM network. These logical channels efficiently transmit user data while simultaneously providing control of the network on each ARFCN. GSM provides explicit assignments of time slots and frames for specific logical channels, as described below.

### 10.3.4 GSM Channel Types

There are two types of GSM logical channels, called *traffic channels* (TCH) and *control channels* (CCH) [Hod90]. Traffic channels carry digitally encoded user speech or user data and have identical functions and formats on both the forward and reverse link. Control channels carry signaling and synchronizing commands between the base station and the mobile station. Certain types of control channels are defined for just the forward or reverse link. There are six different types of TCHs provided for in GSM, and an even larger number of CCHs, both of which are now described.

#### 10.3.4.1 GSM Traffic CHannels (TCH)

GSM traffic channels may be either full-rate or half-rate and may carry either digitized speech or user data. When transmitted as full-rate, user data is contained within one TS per frame. When transmitted as half-rate, user data is mapped onto the same time slot, but is sent in alternate frames. That is, two half-rate channel users would share the same time slot, but would alternately transmit during every other frame.

In the GSM standard, TCH data may not be sent in TS 0 within a TDMA frame on certain ARFCNs which serve as the broadcast station for each cell (since this time slot is reserved for control channel bursts in most every frame, as described subsequently). Furthermore, frames of TCH data are broken up every thirteenth frame by either slow associated control channel data (SACCH) or idle frames. Figure 10.7 illustrates how the TCH data is transmitted in consecutive frames. Each group of twenty-six consecutive TDMA frames is called a *multiframe* (or *speech multiframe*, to distinguish it from the control channel multiframe described below). For every twenty-six frames, the thirteenth and twenty-sixth frames consist of Slow Associated Control Channel (SACCH) data, or the idle frame, respectively. The twenty-sixth frame contains idle bits for the

case when full-rate TCHs are used, and contains SACCH data when half-rate TCHs are used.

**Full-Rate TCH**

The following full rate speech and data channels are supported:

- **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. With GSM channel coding added to the digitized speech, the full-rate speech channel carries 22.8 kbps.
- **Full-Rate Data Channel for 9600 bps (TCH/F9.6)** — The full-rate traffic data channel carries raw user data which is sent at 9600 bps. With additional forward error correction coding applied by the GSM standard, the 9600 bps data is sent at 22.8 kbps.
- **Full-Rate Data Channel for 4800 bps (TCH/F4.8)** — The full-rate traffic data channel carries raw user data which is sent at 4800 bps. With additional forward error correction coding applied by the GSM standard, the 4800 bps is sent at 22.8 kbps.
- **Full-Rate Data Channel for 2400 bps (TCH/F2.4)** — The full-rate traffic data channel carries raw user data which is sent at 2400 bps. With additional forward error correction coding applied by the GSM standard, the 2400 bps is sent at 22.8 kbps.

**Half-Rate TCH**

The following half-rate speech and data channels are supported:

- **Half-Rate Speech Channel (TCH/HS)** — The half-rate speech channel has been designed to carry digitized speech which is sampled at a rate half that of the full-rate channel. GSM anticipates the availability of speech coders which can digitize speech at about 6.5 kbps. With GSM channel coding added to the digitized speech, the half-rate speech channel will carry 11.4 kbps.
- **Half-Rate Data Channel for 4800 bps (TCH/H4.8)** — The half-rate traffic data channel carries raw user data which is sent at 4800 bps. With additional forward error correction coding applied by the GSM standard, the 4800 bps data is sent at 11.4 kbps.
- **Half-Rate Data Channel for 2400 bps (TCH/H2.4)** — The half-rate traffic data channel carries raw user data which is sent at 2400 bps. With additional forward error correction coding applied by the GSM standard, the 2400 bps data is sent at 11.4 kbps.

### 10.3.4.2 GSM Control Channels (CCH)

There are three main control channels in the GSM system. These are the *broadcast channel* (BCH), the *common control channel* (CCCH), and the *dedicated control channel* (DCCH). Each control channel consists of several logical

channels which are distributed in time to provide the necessary GSM control functions.

The BCH and CCCH forward control channels in GSM are implemented only on certain ARFCN channels and are allocated timeslots in a very specific manner. Specifically, the BCH and CCCH forward control channels are allocated only TS 0 and are broadcast only during certain frames within a repetitive fifty-one frame sequence (called the *control channel multiframe*) on those ARFCNs which are designated as broadcast channels. TS1 through TS7 carry regular TCH traffic, so that ARFCNs which are designated as control channels are still able to carry full-rate users on seven of the eight time slots.

The GSM specification defines thirty-four ARFCNs as standard broadcast channels. For each broadcast channel, frame 51 does not contain any BCH/CCCH forward channel data and is considered to be an idle frame. However, the reverse channel CCCH is able to receive subscriber transmissions during TS 0 of any frame (even the idle frame). On the other hand, DCCH data may be sent during any time slot and any frame, and entire frames are specifically dedicated to certain DCCH transmissions. GSM control channels are now described in detail.

- **Broadcast Channels (BCH)**— The broadcast channel operates on the forward link of a specific ARFCN within each cell, and transmits data only in the first time slot (TS 0) of certain GSM frames. Unlike TCHs which are duplex, BCHs only use the forward link. Just as the forward control channel (FCC) in AMPS is used as a beacon for all nearby mobiles to camp on to, the BCH serves as a TDMA beacon channel for any nearby mobile to identify and lock on to. The BCH provides synchronization for all mobiles within the cell and is occasionally monitored by mobiles in neighboring cells so that received power and MAHO decisions may be made by out-of-cell users. Although BCH data is transmitted in TS0, the other seven timeslots in a GSM frame for that same ARFCN are available for TCH data, DCCH data, or are filled with dummy bursts. Furthermore, all eight timeslots on all other ARFCNs within the cell are available for TCH or DCCH data.

The BCH is defined by three separate channels which are given access to TS 0 during various frames of the 51 frame sequence. Figure 10.8 illustrates how the BCH is allocated frames. The three types of BCH are now described.

(a) *Broadcast Control CHannel (BCCH)* — The BCCH is a forward control channel that is used to broadcast information such as cell and network identity, and operating characteristics of the cell (current control channel structure, channel availability, and congestion). The BCCH also broadcasts a list of channels that are currently in use within the cell. Frame 2 through frame 5 in a control multiframe (4 out of every 51 frames) contain BCCH data. It

should be noted from Figure 10.8 that TS 0 contains BCCH data during specific frames, and contains other BCH channels (FCCH and SCH), common control channels (CCCHs), or an idle frame (sent every 51st frame) during other specific frames.

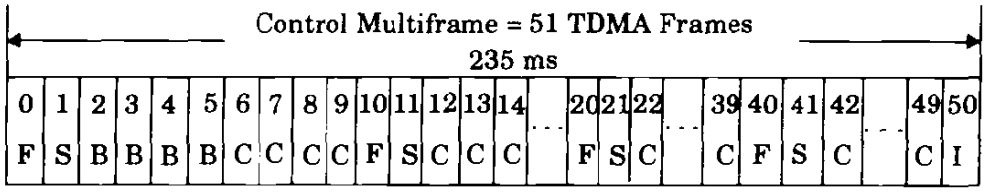
(b) *Frequency Correction CHannel (FCCH)* — The FCCH is a special data burst which occupies TS 0 for the very first GSM frame (frame 0) and is repeated every ten frames within a control channel multiframe. The FCCH allows each subscriber unit to synchronize its internal frequency standard (local oscillator) to the exact frequency of the base station.

(c) *Synchronization CHannel (SCH)* — SCH is broadcast in TS 0 of the frame immediately following the FCCH frame and is used to identify the serving base station while allowing each mobile to frame synchronize with the base station. The *frame number* (FN), which ranges from 0 to 2,715,647, is sent with the *base station identity code* (BSIC) during the SCH burst. The BSIC is uniquely assigned to each BST in a GSM system. Since a mobile may be as far as 30 km away from a serving base station, it is often necessary to adjust the timing of a particular mobile user such that the received signal at the base station is synchronized with the base station clock. The BS issues coarse *timing advancement* commands to the mobile stations over the SCH, as well. The SCH is transmitted once every ten frames within the control channel multiframe, as shown in Figure 10.8.

- **Common Control CHannels (CCCH)** — On the broadcast (BCH) ARFCN, the common control channels occupy TS 0 of every GSM frame that is not otherwise used by the BCH or the Idle frame. CCCH consists of three different channels: the paging channel (PCH), which is a forward link channel, the random access channel (RACH) which is a reverse link channel, and the access grant channel (AGCH), which is a forward link channel. As seen in Figure 10.8, CCCHs are the most commonly used control channels and are used to page specific subscribers, assign signaling channels to specific users, and receive mobile requests for service. These channels are described below.

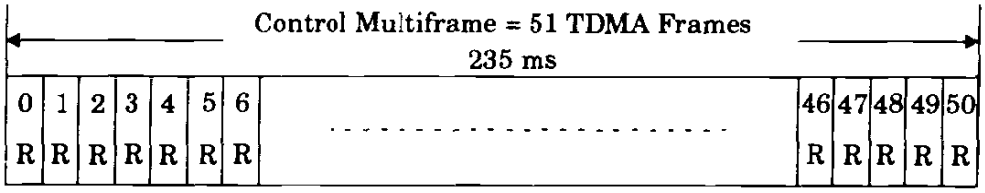
(a) *Paging CHannel (PCH)* — The PCH provides paging signals from the base station to all mobiles in the cell, and notifies a specific mobile of an incoming call which originates from the PSTN. The PCH transmits the IMSI of the target subscriber, along with a request for acknowledgment from the mobile unit on the RACH. Alternatively, the PCH may be used to provide *cell broadcast* ASCII text messages to all subscribers, as part of the SMS feature of GSM.

(b) *Random Access Channel (RACH)* — The RACH is a reverse link channel used by a subscriber unit to acknowledge a page from the PCH, and is also used by mobiles to originate a call. The RACH uses a slotted ALOHA access scheme. All mobiles must request access or respond to a PCH alert within TS




**F** : FCCH burst (BCH)
**S** : SCH burst (BCH)
**B** : BCCH burst (BCH)
**C** : PCH/AGCH burst (CCCH)
**I** : Idle

(a)




**R** : Reverse RACH burst (CCCH)

(b)

Figure 10.8
(a) The Control Channel Multiframe (Forward link for TS0)
(b) The Control Channel Multiframe (Reverse link for TS0)

0 of a GSM frame. At the BTS, every frame (even the idle frame) will accept RACH transmissions from mobiles during TS 0. In establishing service, the GSM base station must respond to the RACH transmission by allocating a channel and assigning a stand-alone dedicated control channel (SDCCH) for signaling during a call. This connection is confirmed by the base station over the AGCH.

(c) *Access Grant Channel (AGCH)* — The AGCH is used by the base station to provide forward link communication to the mobile, and carries data which instructs the mobile to operate in a particular physical channel (time slot and ARFCN) with a particular dedicated control channel. The AGCH is the final CCCH message sent by the base station before a subscriber is moved off the control channel. The AGCH is used by the base station to respond to a RACH sent by a mobile station in a previous CCCH frame.

- **Dedicated Control Channels (DCCH)** — There are three types of dedicated control channels in GSM, and, like traffic channels (see Figure 10.7), they are bidirectional and have the same format and function on both the forward and reverse links. Like TCHs, DCCHs may exist in any time slot and on any ARFCN *except TS0* of the BCH ARFCN. The stand-alone dedicated control channels (SDCCH) are used for providing signaling services required by the users. The Slow- and Fast- Associated Control Channels

(SACCH and FACCH) are used for supervisory data transmissions between the mobile station and the base station during a call.

(a) *Stand-alone Dedicated Control Channels (SDCCH)* — The SDCCH carries signaling data following the connection of the mobile with the base station, and just before a TCH assignment is issued by the base station. The SDCCH ensures that the mobile station and the base station remain connected while the base station and MSC verify the subscriber unit and allocate resources for the mobile. The SDCCH can be thought of as an intermediate and temporary channel which accepts a newly completed call from the BCH and holds the traffic while waiting for the base station to allocate a TCH channel. The SDCCH is used to send authentication and alert messages (but not speech) as the mobile synchronizes itself with the frame structure and waits for a TCH. SDCCHs may be assigned their own physical channel or may occupy TS0 of the BCH if there is low demand for BCH or CCCH traffic.

(b) *Slow Associated Control Channel (SACCH)* — The SACCH is always associated with a traffic channel or a SDCCH and maps onto the same physical channel. Thus, each ARFCN systematically carries SACCH data for all of its current users. As in the USDC standard, the SACCH carries general information between the MS and BTS. On the forward link, the SACCH is used to send slow but regularly changing control information to the mobile, such as transmit power level instructions and specific timing advance instructions for each user on the ARFCN. The reverse SACCH carries information about the received signal strength and quality of the TCH, as well as BCH measurement results from neighboring cells. The SACCH is transmitted during the thirteenth frame (and the twenty-sixth frame when half-rate traffic is used) of every speech/dedicated control channel multiframe (Figure 10.7), and within this frame, the eight timeslots are dedicated to providing SACCH data to each of the eight full-rate (or sixteen half-rate) users on the ARFCN.

(c) *Fast Associated Control Channels (FACCH)* — FACCH carries urgent messages, and contains essentially the same type of information as the SDCCH. A FACCH is assigned whenever a SDCCH has not been dedicated for a particular user and there is an urgent message (such as a handoff request). The FACCH gains access to a time slot by "stealing" frames from the traffic channel to which it is assigned. This is done by setting two special bits, called stealing bits, in a TCH forward channel burst. If the stealing bits are set, the time slot is known to contain FACCH data, not a TCH, for that frame.

### 10.3.5 Example of a GSM Call

To understand how the various traffic and control channels are used, consider the case of a mobile call origination in GSM. First, the subscriber unit must be synchronized to a nearby base station as it monitors the BCH. By receiving the FCCH, SCH, and BCCH messages, the subscriber would be locked on to the

system and the appropriate BCH. To originate a call, the user first dials the intended digit combination and presses the "send" button on the GSM phone. The mobile transmits a burst of RACH data, using the same ARFCN as the base station to which it is locked. The base station then responds with an AGCH message on the CCCH which assigns the mobile unit to a new channel for SDCCH connection. The subscriber unit, which is monitoring TS 0 of the BCH, would receive its ARFCN and TS assignment from the AGCH and would immediately tune to the new ARFCN and TS. This new ARFCN and TS assignment is physically the SDCCH (not the TCH). Once tuned to the SDCCH, the subscriber unit first waits for the SACCH frame to be transmitted (the wait would last, at most, 26 frames or 120 ms, as shown in Figure 10.7), which informs the mobile of any required timing advance and transmitter power command. The base station is able to determine the proper timing advance and signal level from the mobile's earlier RACH transmission and sends the proper value over the SACCH for the mobile to process. Upon receiving and processing the timing advance information in the SACCH, the subscriber is now able to transmit normal burst messages as required for speech traffic. The SDCCH sends messages between the mobile unit and the base station, taking care of authentication and user validation, while the PSTN connects the dialed party to the MSC, and the MSC switches the speech path to the serving base station. After a few seconds, the mobile unit is commanded by the base station via the SDCCH to retune to a new ARFCN and new TS for the TCH assignment. Once retuned to the TCH, speech data is transferred on both the forward and reverse links, the call is successfully underway, and the SDCCH is vacated.

When calls are originated from the PSTN, the process is quite similar. The base station broadcasts a PCH message during TS 0 within an appropriate frame on the BCH. The mobile station, locked on to that same ARFCN, detects its page and replies with an RACH message acknowledging receipt of the page. The base station then uses the AGCH on the CCCH to assign the mobile unit to a new physical channel for connection to the SDCCH and SACCH while the network and the serving base station are connected. Once the subscriber establishes timing advance and authentication on the SDCCH, the base station issues a new physical channel assignment over the SDCCH, and the TCH assignment is made.

### 10.3.6 Frame Structure for GSM

Each user transmits a burst of data during the time slot assigned to it. These data bursts may have one of five specific formats, as defined in GSM [Hod90]. Figure 10.9 illustrates the five types of data bursts used for various control and traffic bursts. Normal bursts are used for TCH and DCCH transmissions on both the forward and reverse link. FCCH and SCH bursts are used in TS 0 of specific frames (shown in Figure 10.8a) to broadcast the frequency and

time synchronization control messages on the forward link. The RACH burst is used by all mobiles to access service from any base station, and the dummy burst is used as filler information for unused time slots on the forward link.

Normal

| | | | | | |
|---|---|---|---|---|---|
| 3 start bits | 58 bits of encrypted data | 26 training bits | 58 bits of encrypted data | 3 stop bits | 8.25 bits guard period |

FCCH burst

| | | | |
|---|---|---|---|
| 3 start bits | 142 fixed bits of all zeroes | 3 stop bits | 8.25 bits guard period |

SCH burst

| | | | | | |
|---|---|---|---|---|---|
| 3 start bits | 39 bits of encrypted data | 64 bits of training | 39 bits of encrypted data | 3 stop bits | 8.25 bits guard period |

RACH burst

| | | | | |
|---|---|---|---|---|
| 8 start bits | 41 bits of synchronization | 36 bits of encrypted data | 3 stop bits | 68.25 bit extended guard period |

Dummy burst

| | | | | | |
|---|---|---|---|---|---|
| 3 start bits | 58 mixed bits | 26 training bits | 58 mixed bits | 3 stop bits | 8.25 bits guard period |

Figure 10.9
Time slot data bursts in GSM.

Figure 10.10 illustrates the data structure within a normal burst. It consists of 148 bits which are transmitted at a rate of 270.833333 kbps (an unused guard time of 8.25 bits is provided at the end of each burst). Out of the total 148 bits per TS, 114 are information-bearing bits which are transmitted as two 57 bit sequences close to the beginning and end of the burst. The midamble consists of a 26 bit training sequence which allows the adaptive equalizer in the mobile or base station receiver to analyze the radio channel characteristics before decoding the user data. On either side of the midamble there are control bits called stealing flags. These two flags are used to distinguish whether the TS contains voice (TCH) or control (FACCH) data, both which share the same physical channel. During a frame, a GSM subscriber unit uses one TS to transmit, one TS to receive, and may use the six spare time slots to measure signal strength on five adjacent base stations as well as its own base station.

As shown in Figure 10.10, there are eight timeslots per TDMA frame, and the frame period is 4.615 ms. A frame contains $8 \times 156.25 = 1250$ bits, although some bit periods are not used. The frame rate is 270.833 kbps/1250 bits/frame, or 216.66 frames per second. The 13th or 26th frame are not used for traffic, but for

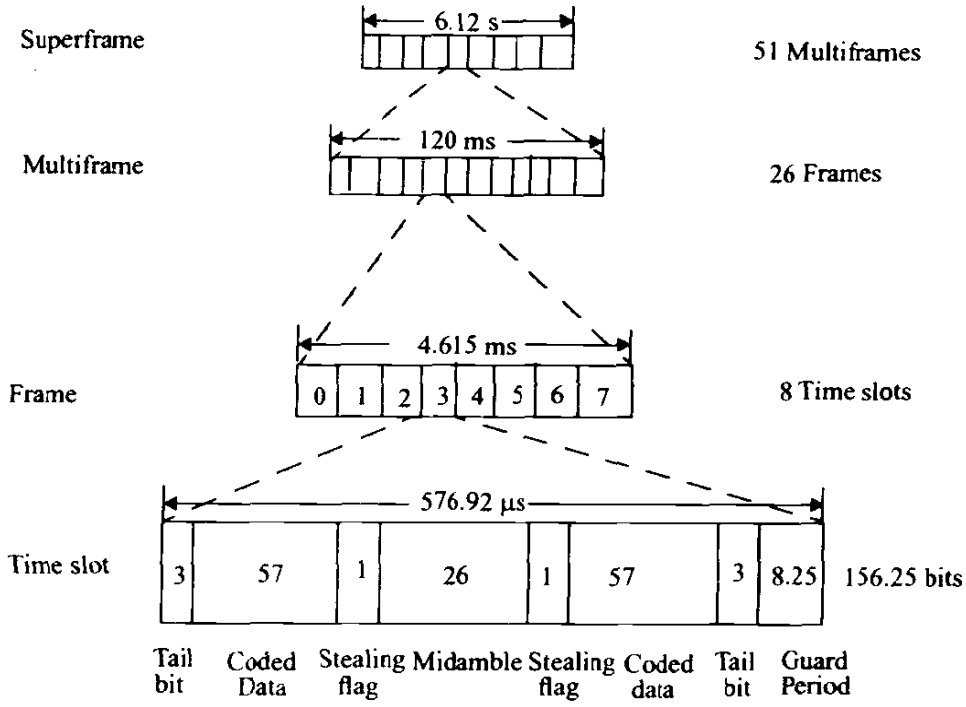



Figure 10.10
GSM frame structure.

control purposes. Each of the normal speech frames are grouped into larger structures called *multiframes* which in turn are grouped into *superframes* and *hyperframes* (hyperframes are not shown in Figure 10.10). One multiframe contains 26 TDMA frames, and one superframe contains 51 multiframes, or 1326 TDMA frames. A hyperframe contains 2048 superframes, or 2,715,648 TDMA frames. A complete hyperframe is sent about every 3 hours, 28 minutes, and 54 seconds, and is important to GSM since the encryption algorithms rely on the particular frame number, and sufficient security can only be obtained by using a large number of frames as provided by the hyperframe.

Figure 10.8 shows that the control multiframes span 51 frames (235.365 ms), as opposed to 26 frames (120 ms) used by the traffic/dedicated control channel multiframes. This is done intentionally to ensure that any GSM subscriber (whether in the serving or adjacent cell) will be certain to receive the SCH and FCCH transmissions from the BCH, no matter what particular frame or time slot they are using.

### 10.3.7 Signal Processing in GSM

Figure 10.11 illustrates all of the GSM operations from transmitter to receiver.

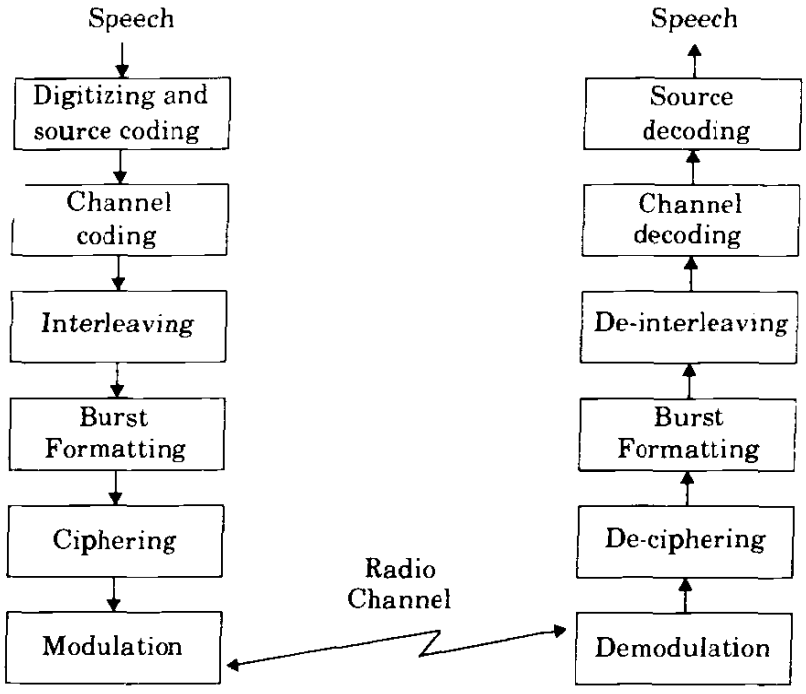



*Figure 10.11*
GSM operations from speech input to speech output.

**Speech Coding** — The GSM speech coder is based on the Residually Excited Linear Predictive Coder (RELP), which is enhanced by including a Long-Term Predictor (LTP) [Hel89]. The coder provides 260 bits for each 20 ms blocks of speech, which yields a bit rate of 13 kbps. This speech coder was selected after extensive subjective evaluation of various candidate coders available in the late 1980s. Provisions for incorporating half-rate coders are included in the specifications.

The GSM speech coder takes advantage of the fact that in a normal conversation, each person speaks on average for less than 40% of the time. By incorporating a voice activity detector (VAD) in the speech coder, GSM systems operate in a *discontinuous transmission mode* (DTX) which provides a longer subscriber battery life and reduces instantaneous radio interference since the GSM transmitter is not active during silent periods. A comfort noise subsystem (CNS) at the receiving end introduces a background acoustic noise to compensate for the annoying switched muting which occurs due to DTX.

**TCH/FS, SACCH, and FACCH Channel Coding** — The output bits of the speech coder are ordered into groups for error protection, based upon their significance in contributing to speech quality. Out of the total 260 bits in a frame, the most important 50 bits, called type Ia bits, have 3 parity check (CRC) bits added to them. This facilitates the detection of non-correctable errors at the

receiver. The next 132 bits along with the first 53 (50 type Ia bits + 3 parity bits) are reordered and appended by 4 trailing zero bits, thus providing a data block of 189 bits. This block is then encoded for error protection using a rate 1/2 convolutional encoder with constraint length $K = 5$, thus providing a sequence of 378 bits. The least important 78 bits do not have any error protection and are concatenated to the existing sequence to form a block of 456 bits in a 20 ms frame. The error protection coding scheme increases the gross data rate of the GSM speech signal, with channel coding, to 22.8 kbps. This error protection scheme as described is illustrated in Figure 10.12.

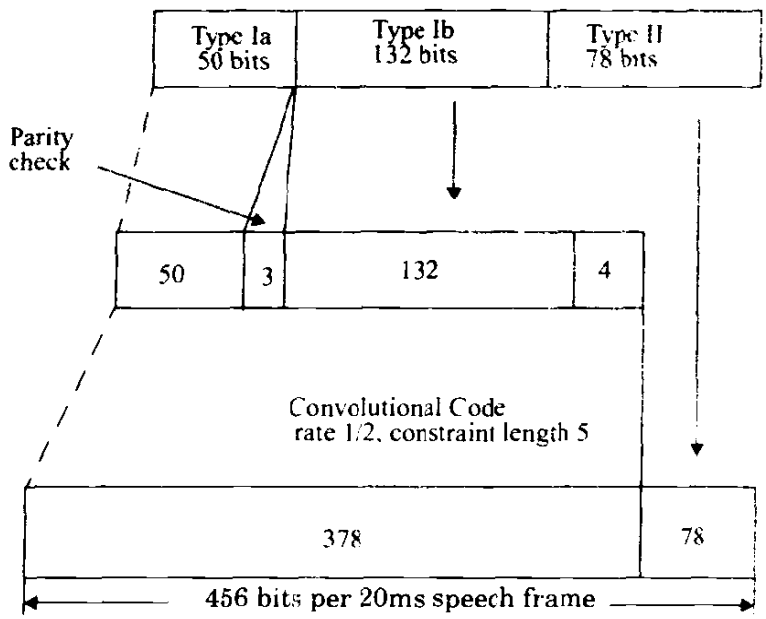



Figure 10.12
Error protection for speech signals in GSM.

**Channel Coding for Data Channels** — The coding provided for GSM full rate data channels (TCH/F9.6) is based on handling 60 bits of user data at 5ms intervals, in accordance with the modified CCITT V.110 modem standard. As described by Stallings [Ste92] (Chapter 8), 240 bits of user data are applied with 4 tailing bits to a half-rate punctured convolutional coder with constraint length $K = 5$. The resulting 488 coded bits are reduced to 456 encoded data bits through puncturing (32 bits are not transmitted), and the data is separated into four 114 bit data bursts that are applied in an interleaved fashion to consecutive time slots.

**Channel Coding for Control Channels** — GSM control channel messages are defined to be 184 bits long, and are encoded using a shortened binary cyclic fire code, followed by a half-rate convolutional coder.

The fire code uses the generator polynomial

$$G_5(x) = (x^{23}+1)(x^{17}+x^3+1) = x^{40}+x^{26}+x^{23}+x^{17}+x^3+1$$

which produces 184 message bits, followed by 40 parity bits. Four tail bits are added to clear the convolutional coder which follows, yielding a 228 bit data block. This block is applied to a half-rate $K = 5$ convolutional code (CC(2,1,5)) using the generator polynomials $G_0(x) = 1 + x^3 + x^4$ and $G_1(x) = 1 + x + x^3 + x^4$ (which are the same polynomials used to code TCH type Ia data bits). The resulting 456 encoded bits are interleaved onto eight consecutive frames in the same manner as TCH speech data.

**Interleaving** — In order to minimize the effect of sudden fades on the received data, the total of 456 encoded bits within each 20 ms speech frame or control message frame are broken into eight 57 bit sub-blocks. These eight sub-blocks which make up a single speech frame are spread over eight consecutive TCH time slots. (i.e., eight consecutive frames for a specific TS). If a burst is lost due to interference or fading, channel coding ensures that enough bits will still be received correctly to allow the error correction to work. Each TCH time slot carries two 57 bit blocks of data from two different 20 ms (456 bit) speech (or control) segments. Figure 10.13 illustrates exactly how the speech frames are diagonally interleaved within the time slots. Note that TS 0 contains 57 bits of data from the 0th sub-block of the $n$th speech coder frame (denoted as "a" in the figure) and 57 bits of data from the 4th sub-block of the ($n$ - 1)st speech coder frame (denoted as "b" in the figure).

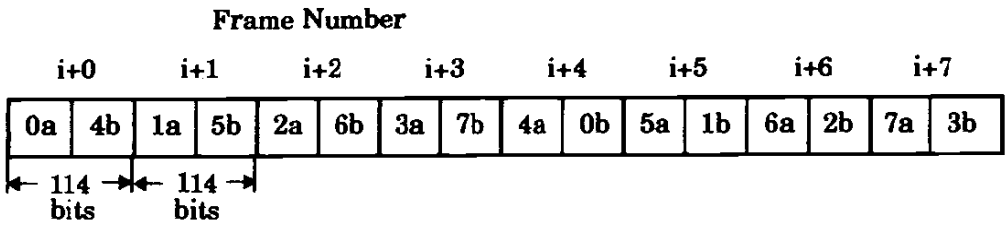



Figure 10.13
Diagonal interleaving used for TCH/SACCH/FACCH data. Eight speech sub-blocks are spread over eight successive TCH time slots for a specific time slot number.

**Ciphering** — Ciphering modifies the contents of the eight interleaved blocks through the use of encryption techniques known only to the particular mobile station and base transceiver station. Security is further enhanced by the fact that the encryption algorithm is changed from call to call. Two types of ciphering algorithms, called A3 and A5, are used in GSM to prevent unauthorized network access and privacy for the radio transmission respectively. The A3 algorithm is used to authenticate each mobile by verifying the users passcode within the SIM with the cryptographic key at the MSC. The A5 algorithm provides the scrambling for the 114 coded data bits sent in each TS.

**Burst Formatting** — Burst formatting adds binary data to the ciphered blocks, in order to help synchronization and equalization of the received signal.

**Modulation** — The modulation scheme used by GSM is 0.3 GMSK where 0.3 describes the 3 dB bandwidth of the Gaussian pulse shaping filter with rela-

tion to the bit rate (e.g., $BT = 0.3$). As described in Chapter 5, GMSK is a special type of digital FM modulation. Binary ones and zeros are represented in GSM by shifting the RF carrier by ±67.708 kHz. The channel data rate of GSM is 270.833333 kbps, which is exactly four times the RF frequency shift. This minimizes the bandwidth occupied by the modulation spectrum and hence improves channel capacity. The MSK modulated signal is passed through a Gaussian filter to smooth the rapid frequency transitions which would otherwise spread energy into adjacent channels.

**Frequency Hopping** — Under normal conditions, each data burst belonging to a particular physical channel is transmitted using the same carrier frequency. However, if users in a particular cell have severe multipath problems, the cell may be defined as a *hopping cell* by the network operator, in which case *slow frequency hopping* may be implemented to combat the multipath or interference effects in that cell. Frequency hopping is carried out on a frame-by-frame basis, thus hopping occurs at a maximum rate of 217.6 hops per second. As many as 64 different channels may be used before a hopping sequence is repeated. Frequency hopping is completely specified by the service provider.

**Equalization** — Equalization is performed at the receiver with the help of the training sequences transmitted in the midamble of every time slot. The type of equalizer for GSM is not specified and is left up to the manufacturer.

**Demodulation** — The portion of the transmitted forward channel signal which is of interest to a particular user is determined by the assigned TS and ARFCN. The appropriate TS is demodulated with the aid of synchronization data provided by the burst formatting. After demodulation, the binary information is deciphered, de-interleaved, channel decoded, and speech decoded.

## 10.4 CDMA Digital Cellular Standard (IS-95)

As discussed in Chapter 8, Code Division Multiple Access (CDMA) offers some advantages over TDMA and FDMA. A U.S. digital cellular system based on CDMA which promises increased capacity [Gil91] has been standardized as Interim Standard 95 (IS-95) by the U.S. Telecommunications Industry Association (TIA) [TIA93]. Like IS-54, the IS-95 system is designed to be compatible with the existing U.S. analog cellular system (AMPS) frequency band, hence mobiles and base stations can be economically produced for dual mode operation. Pilot production, CDMA/AMPS, dual mode phones were made available by Qualcomm in 1994.

IS-95 allows each user within a cell to use the same radio channel, and users in adjacent cells also use the same radio channel, since this is a direct sequence spread spectrum CDMA system. CDMA completely eliminates the need for frequency planning within a market. To facilitate graceful transition from AMPS to CDMA, each IS-95 channel occupies 1.25 MHz of spectrum on each one-way link, or 10% of the available cellular spectrum for a U.S. cellular

provider (recall, the U.S. cellular system is allocated 25 MHz and each service provider receives half the spectrum or 12.5 MHz). In practice, AMPS carriers must provide a 270 kHz guard band (typically 9 AMPS channels) on each side of the spectrum dedicated for IS-95. IS-95 is fully compatible with the IS-41 networking standard described in Chapter 9.

Unlike other cellular standards, the user data rate (but not the channel chip rate) changes in real-time, depending on the voice activity and requirements in the network. Also, IS-95 uses a different modulation and spreading technique for the forward and reverse links. On the forward link, the base station simultaneously transmits the user data for all mobiles in the cell by using a different spreading sequence for each mobile. A pilot code is also transmitted simultaneously and at a higher power level, thereby allowing all mobiles to use coherent carrier detection while estimating the channel conditions. On the reverse link, all mobiles respond in an asynchronous fashion and have ideally a constant signal level due to power control applied by the base station.

The speech coder used in the IS-95 system is the Qualcomm 9600 bps Code Excited Linear Predictive (QCELP) coder. The original implementation of this vocoder detects voice activity, and reduces the data rate to 1200 bps during silent periods. Intermediate user data rates of 2400, 4800, and 9600 bps are also used for special purposes. As discussed in Chapter 7 and section 10.4.4, a 14,400 bps coder which uses 13.4 kbps of speech data (QCELP13) was introduced by Qualcomm in 1995.

### 10.4.1 Frequency and Channel Specifications

IS-95 is specified for reverse link operation in the 824 - 849 MHz band and 869 - 894 MHz for the forward link. A forward and reverse channel pair is separated by 45 MHz. Many users share a common channel for transmission. The maximum user data rate is 9.6 kb/s. User data in IS-95 is spread to a channel chip rate of 1.2288 Mchip/s (a total spreading factor of 128) using a combination of techniques. The spreading process is different for the forward and reverse links. On the forward link, the user data stream is encoded using a rate 1/2 convolutional code, interleaved, and spread by one of sixty-four orthogonal spreading sequences (Walsh functions). Each mobile in a given cell is assigned a different spreading sequence, providing perfect separation among the signals from different users, at least for the case where multipath does not exist. To reduce interference between mobiles that use the same spreading sequence in different cells, and to provide the desired wideband spectral characteristics (not all of the Walsh functions yield a wideband power spectrum), all signals in a particular cell are scrambled using a pseudorandom sequence of length $2^{15}$ chips.

Orthogonality among all forward channel users within a cell is preserved because their signals are scrambled synchronously. A pilot channel (code) is provided on the forward link so that each subscriber within the cell can determine

and react to the channel characteristics while employing coherent detection. The pilot channel is transmitted at higher power than the user channels.

On the reverse link, a different spreading strategy is used since each received signal arrives at the base station via a different propagation path. The reverse channel user data stream is first convolutionally encoded with a rate 1/3 code. After interleaving, each block of six encoded symbols is mapped to one of the 64 orthogonal Walsh functions, providing sixty-four-ary orthogonal signaling. A final fourfold spreading, giving a rate of 1.2288 Mchip/s, is achieved by spreading the resulting 307.2 kchip/s stream by user-specific and base-station specific codes having periods of $2^{42} - 1$ chips and $2^{15}$ chips, respectively. The rate 1/3 coding and the mapping onto Walsh functions result in a greater tolerance for interference than would be realized from traditional repetition spreading codes. This added robustness is important on the reverse link, due to the noncoherent detection and the in-cell interference received at the base station.

Another essential element of the reverse link is tight control of each subscriber's transmitter power, to avoid the "near-far" problem that arises from varying received powers of the users. A combination of open-loop and fast, closed-loop power control is used to adjust the transmit power of each in-cell subscriber so that the base station receives each user with the same received power. The commands for the closed-loop power control are sent at a rate of 800 b/s, and these bits are stolen from the speech frames. Without fast power control, the rapid power changes due to fading would degrade the performance of all users in the system.

At both the base station and the subscriber, RAKE receivers are used to resolve and combine multipath components, thereby reducing the degree of fading. As described in Chapter 6, a RAKE receiver exploits the multipath time delays in a channel and combines the delayed replicas of the transmitted signal in order to improve link quality. In IS-95, a three finger RAKE is used at the base station. The IS-95 architecture also provides base station diversity during "soft" handoffs, whereby a mobile making the transition between cells maintains links with both base stations during the transition. The mobile receiver combines the signals from the two base stations in the same manner as it would combine signals associated with different multipath components.

### 10.4.2 Forward CDMA Channel

The forward CDMA channel consists of a pilot channel, a synchronization channel, up to seven paging channels, and up to sixty-three forward traffic channels [Li93]. The pilot channel allows a mobile station to acquire timing for the Forward CDMA channel, provides a phase reference for coherent demodulation, and provides each mobile with a means for signal strength comparisons between base stations for determining when to handoff. The synchronization channel broadcasts synchronization messages to the mobile stations and operates at 1200

bps. The paging channel is used to send control information and paging messages from the base station to the mobiles and operates at 9600, 4800, and 2400 bps. The forward traffic channel (FTC) supports variable user data rates at 9600, 4800, 2400, or 1200 bps.

The forward traffic channel modulation process is described in Figure 10.14 [EIA90]. Data on the forward traffic channel is grouped into 20 ms frames. The user data is first convolutionally coded and then formatted and interleaved to adjust for the actual user data rate, which may vary. Then the signal is spread with a Walsh code and a long PN sequence at a rate of 1.2288 Mcps. Table 10.4 lists the coding and repetition parameters for the forward traffic channel.

**Table 10.4 IS-95 Forward traffic channel modulation parameters summary (does not reflect new 13.4 kbps coder)**

| Parameter | Data Rate (bps) | | | |
|---|---|---|---|---|
| User data rate | 9600 | 4800 | 2400 | 1200 |
| Coding Rate | 1/2 | 1/2 | 1/2 | 1/2 |
| User Data Repetition Period | 1 | 2 | 4 | 8 |
| Baseband Coded Data Rate | 19,200 | 19,200 | 19,200 | 19,200 |
| PN Chips/Coded Data Bit | 64 | 64 | 64 | 64 |
| PN Chip Rate (Mcps) | 1.2288 | 1.2288 | 1.2288 | 1.2288 |
| PN Chips/Bit | 128 | 256 | 512 | 1024 |

The speech data rate applied to the transmitter is variable over the range of 1200 bps to 9600 bps.

#### 10.4.2.1 Convolutional Encoder and Repetition Circuit

Speech coded voice or user data are encoded using a half-rate convolutional encoder with constraint length 9. The encoding process is described by generator vectors $G_0$ and $G_1$ which are 753 (octal) and 561 (octal), respectively.

The speech encoder exploits pauses and gaps in speech, and reduces its output from 9600 bps to 1200 bps during silent periods. In order to keep a constant baseband symbol rate of 19.2 kbps, whenever the user rate is less than 9600 bps, each symbol from the convolution encoder is repeated before block interleaving. If the information rate is 4800 bps, each code symbol is repeated 1 time. If the information rate is 2400 bps or 1200 bps, each code symbol is repeated 3 or 7 times, respectively. The repetition results in a constant coded rate of 19,200 symbols per second for all possible information data rates.

#### 10.4.2.2 Block Interleaver

After convolution coding and repetition, symbols are sent to a 20ms block interleaver, which is a 24 by 16 array.

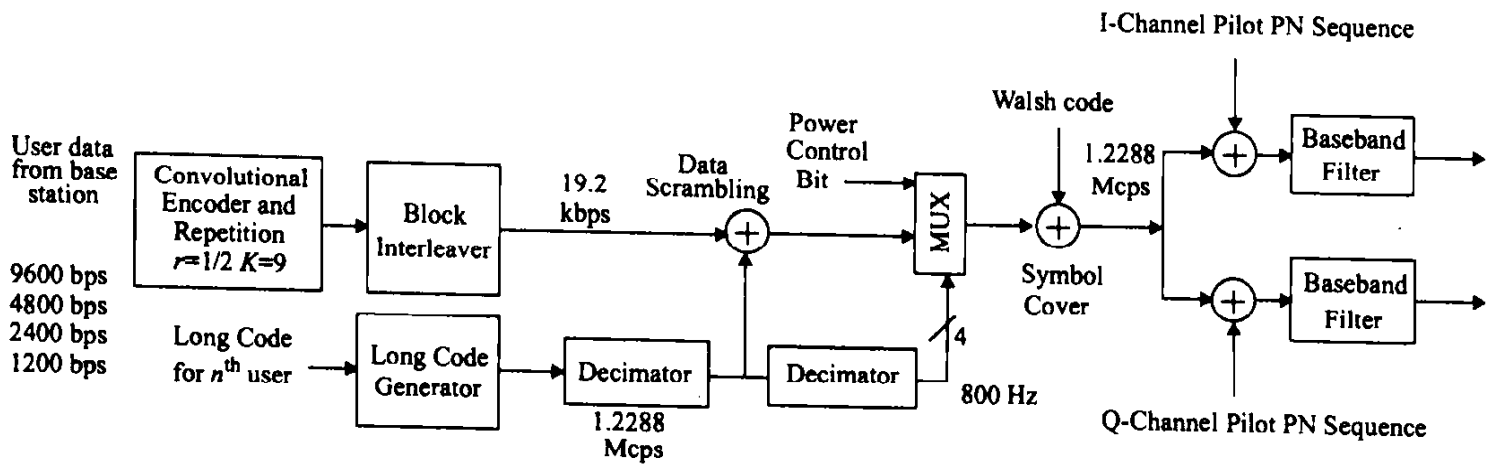



Figure 10.14
Forward CDMA channel modulation process.

### 10.4.2.3 Long PN Sequence

In the forward channel, direct sequence is used for data scrambling. The long PN sequence is uniquely assigned to each user is a periodic long code with period $2^{42}$-1 chips. (This corresponds to repeating approximately once per century). The long code is specified by the following characteristic polynomial [TIA93]

$$p(x) = x^{42} + x^{35} + x^{33} + x^{31} + x^{27} + x^{26} + x^{25} + x^{22} + x^{21} + x^{19} + x^{18} + x^{17} + x^{16} + x^{10} + x^{7} + x^{6} + x^{5} + x^{3} + x^{2} + x^{1} + 1$$

Each PN chip of the long code is generated by the modulo-2 inner product of a 42 bit mask and the 42 bit state vector of the sequence generator. The initial state of the generator is defined to be when the output of the generator becomes '1' after following 41 consecutive '0' outputs, with the binary mask consisting of '1' in the most significant bit (MSB) followed by 41 '0's.

Two types of masks are used in the long code generator: a public mask for the mobile station's electronic serial number (ESN) and a private mask for the mobile station identification number (MIN). All CDMA calls are initiated using the public mask. Transition to the private mask is carried out after authentication is performed. The public long code is specified as follows: $M_{41}$ through $M_{32}$ is set to 1100011000, and $M_{31}$ through $M_0$ is set to a permutation of the mobile station's ESN bits. The permutation is specified as follows [TIA93]:

$$ESN = (E_{31}, E_{30}, E_{29}, E_{28}, E_{27} \ldots\ldots\ldots E_3, E_2, E_1, E_0)$$

$$\text{Permuted ESN} = (E_0, E_{31}, E_{22}, E_{13}, E_4, E_{26}, E_{17}, E_8, E_{30}, E_{21}, E_{12}, E_3, E_{25}, E_{16}, E_7, E_{29}, E_{20}, E_{11}, E_2, E_{24}, E_{15}, E_6, E_{28}, E_{19}, E_{10}, E_1, E_{23}, E_{14}, E_5, E_{27}, E_{18}, E_9)$$

The private long code mask is specified so $M_{41}$ and $M_{40}$ are set to '01', and $M_{39}$ through $M_0$ are set by a private procedure. Figure 10.15 illustrates the long code mask format.

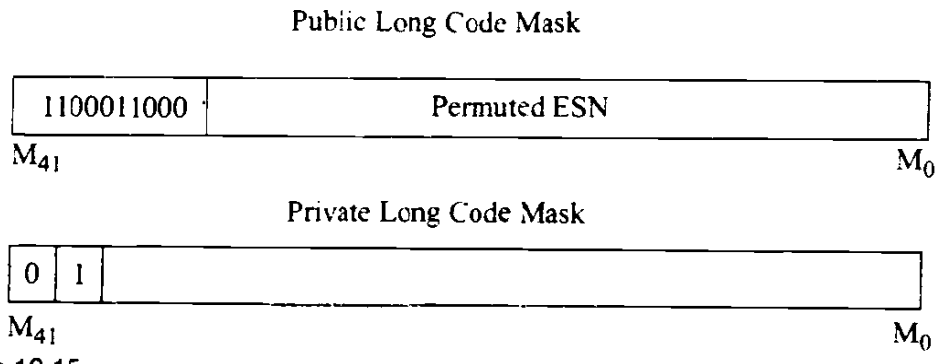



Figure 10.15
Long code mask format for IS-95.

#### 10.4.2.4 Data Scrambler

Data scrambling is performed after the block interleaver. The 1.2288 MHz PN sequence is applied to a decimator, which keeps only the first chip out of every sixty-four consecutive PN chips. The symbol rate from the decimator is 19.2 ksps. The data scrambling is performed by modulo-2 addition of the interleaver output with the decimator output symbol as shown in Figure 10.14.

#### 10.4.2.5 Power Control Subchannel

To minimize the average BER for each user, IS-95 strives to force each user to provide the same power level at the base station receiver. The base station reverse traffic channel receiver estimates and responds to the signal strength (actually, the signal strength and the interference) for a particular mobile station. Since both the signal and interference are continually varying, power control updates are sent by the base station every 1.25 ms. Power control commands are sent to each subscriber unit on the forward control subchannel which instruct the mobile to raise or lower its transmitted power in 1 dB steps. If the received signal is low, a '0' is transmitted over the power control subchannel, thereby instructing the mobile station to increase its mean output power level. If the mobile's power is high, a '1' is transmitted to indicate that the mobile station should decrease its power level. The power control bit corresponds to two modulation symbols on the forward traffic channel. Power control bits are inserted after data scrambling as shown in Figure 10.16.

Power control bits are transmitted by using puncturing techniques [TIA93]. During a 1.25 ms period, twenty-four data symbols are transmitted, and IS-95 specifies sixteen possible power control group positions for the power control bit. Each position corresponds to one of the first sixteen modulation symbols. Twenty-four bits from the long code decimator are used for data scrambling in a period of 1.25 ms. Only the last 4 bits of the 24 bits are used to determine the position of the power control bit. In the example shown in Figure 10.16, the last 4 bits (23, 22, 21, and 20) are '1011' (11 decimal), and the power control bit consequently starts in position eleven.

#### 10.4.2.6 Orthogonal Covering

Orthogonal covering is performed following the data scrambling on the forward link. Each traffic channel transmitted on the forward CDMA channel is spread with a Walsh function at a fixed chip rate of 1.2288 Mcps. The Walsh functions comprise of sixty-four binary sequences, each of length 64, which are completely orthogonal to each other and provide orthogonal channelization for all users on the forward link. A user that is spread using Walsh function $n$ is assigned channel number $n$ ($n$ = 0 to 63). The Walsh sequence repeats every 52.083 μs, which is equal to one coded data symbol. In other words, each data symbol is spread by 64 Walsh chips.

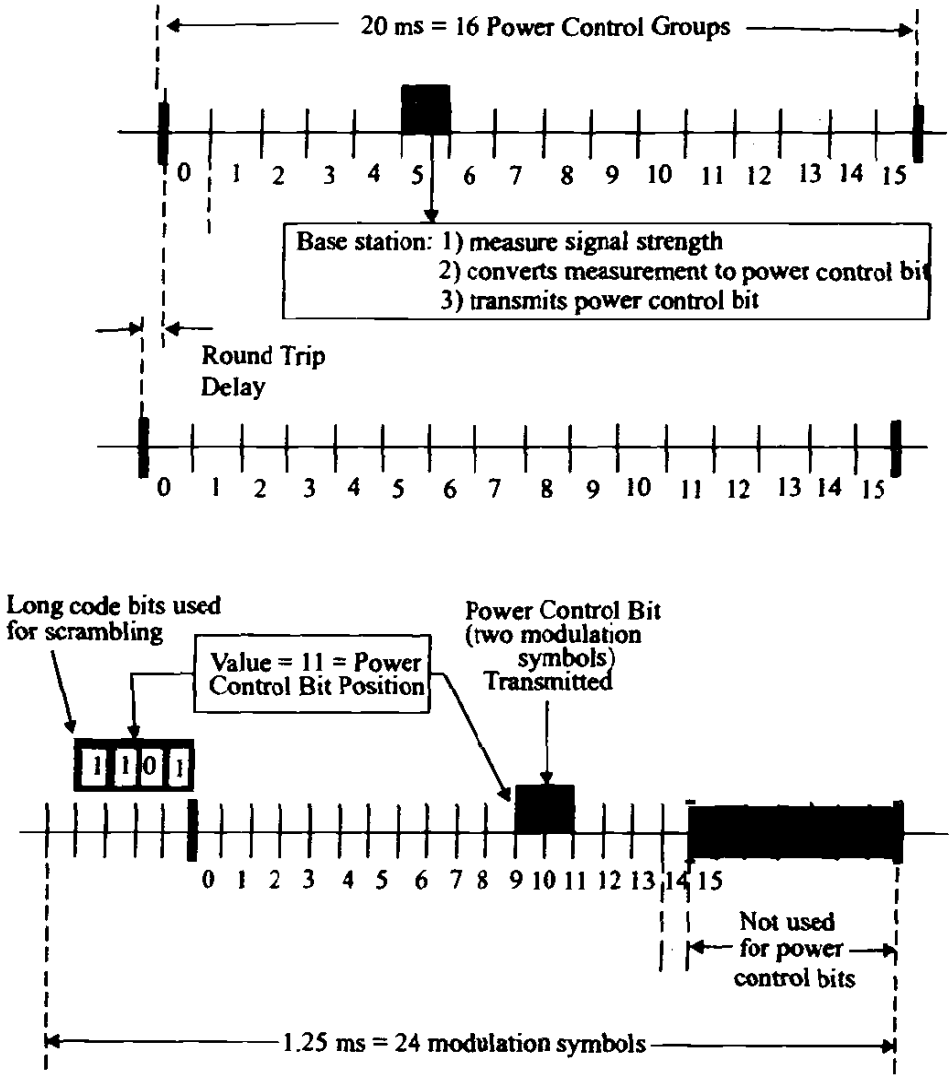



Figure 10.16
**Randomization of power control bit positions in a IS-95 forward traffic channel.**

The 64 by 64 Walsh function matrix (also called a Hadamard matrix) is generated by the following recursive procedure:

$$H_1 = 0 \qquad H_2 = \begin{matrix} 0 & 0 \\ 0 & 1 \end{matrix}$$

$$H_4 = \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{matrix} \qquad H_{2N} = \begin{matrix} H_N & H_N \\ H_N & \overline{H_N} \end{matrix}, \text{ where } N \text{ is a power of 2.}$$

Each row in the 64 by 64 Walsh function matrix corresponds to a channel number. For channel number $n$, the symbols in the transmitter are spread by the sixty-four Walsh chips in the $n$th row of the Walsh function matrix. Channel number 0 is always assigned to the pilot channel. Since Channel 0 represents Walsh code 0, which is the all zeros code, then the pilot channel is nothing more than a "blank" Walsh code and thus consists only of the quadrature PN spreading code. The synchronization channel is vital to the IS-95 system and is assigned channel number 32. If paging channels are present, they are assigned to the lowest code channel numbers. All remaining channels are available for forward traffic channels.

#### 10.4.2.7 Quadrature Modulation

After the orthogonal covering, symbols are spread in quadrature as shown in Figure 10.14. A short binary spreading sequence, with a period of $2^{15}$-1 chips, is used for easy acquisition and synchronization at each mobile receiver and is used for modulation. This short spreading sequence is called the pilot PN sequence, and it is based on the following characteristic polynomials:

$$P_I(x) = x^{15} + x^{13} + x^9 + x^8 + x^7 + x^5 + 1$$

for the in-phase ($I$) modulation and

$$P_Q(x) = x^{15} + x^{12} + x^{11} + x^{10} + x^6 + x^5 + x^4 + x^3 + 1$$

for the quadrature ($Q$) modulation.

Based on the characteristic polynomials, the pilot PN sequences $i(n)$ and $q(n)$ are generated by the following linear recursions:

$$i(n) = i(n\text{-}15) \oplus i(n\text{-}10) \oplus i(n\text{-}8) \oplus i(n\text{-}7) \oplus i(n\text{-}6) \oplus i(n\text{-}2)$$

and,

$$q(n) = q(n\text{-}15) \oplus q(n\text{-}13) \oplus q(n\text{-}11) \oplus q(n\text{-}10) \oplus q(n\text{-}9) \oplus q(n\text{-}5) \oplus q(n\text{-}4) \oplus q(n\text{-}3)$$

where the in-phase and quadrature PN codes are used respectively, and $\oplus$ represents modulo-2 addition. A '0' is inserted in each sequence after the contiguous succession of fourteen '0's to generate pilot PN sequences of length $2^{15}$. The initial state of both $I$ and $Q$ pilot PN sequences is defined as the state in which the output of the pilot PN sequence generator is the first '1' output following fifteen consecutive '0' outputs. The chip rates for the pilot PN sequences are 1.2288 Mcps. The binary $I$ and $Q$ outputs of the quadrature spreading are mapped into phase according to Table 10.5.

### 10.4.3 Reverse CDMA Channel

The reverse traffic channel modulation process is shown in Figure 10.17. User data on the reverse channel are grouped into 20 ms frames. All data transmitted on the reverse channel are convolutionally encoded, block interleaved, modulated by a 64-ary orthogonal modulation, and spread prior to

**Table 10.5 Forward CDMA Channel I and Q Mapping**

| *I* | *Q* | Phase |
|---|---|---|
| 0 | 0 | $\pi/4$ |
| 1 | 0 | $3\pi/4$ |
| 1 | 1 | $-3\pi/4$ |
| 0 | 1 | $-\pi/4$ |

transmission. Table 10.6 shows the modulation parameters for the reverse traffic channel [EIA92]. The speech or user data rate in the reverse channel may be sent at 9600, 4800, 2400, or 1200 bps.

**Table 10.6 Reverse Traffic Channel Modulation Parameters Summary (does not reflect recent 13.4 kbps coder)**

| Parameter | Data Rate (bps) | | | |
|---|---|---|---|---|
| User data rate | 9600 | 4800 | 2400 | 1200 |
| Code Rate | 1/3 | 1/3 | 1/3 | 1/3 |
| TX Duty Cycle (%) | 100.0 | 50.0 | 25.0 | 12.5 |
| Coded Data Rate (sps) | 28,800 | 28,800 | 28,800 | 28,800 |
| Bits per Walsh Symbol | 6 | 6 | 6 | 6 |
| Walsh Symbol Rate | 4800 | 4800 | 4800 | 4800 |
| Walsh Chip Rate (kcps) | 307.2 | 307.2 | 307.2 | 307.2 |
| Walsh Symbol Duration (μs) | 208.33 | 208.33 | 208.33 | 208.33 |
| PN Chips/Code Symbol | 42.67 | 42.67 | 42.67 | 42.67 |
| PN Chips/Walsh Symbol | 256 | 256 | 256 | 256 |
| PN Chips/Walsh Chip | 4 | 4 | 4 | 4 |
| PN Chip Rate (Mcps) | 1.2288 | 1.2288 | 1.2288 | 1.2288 |

The reverse CDMA channels are made up of access channels (AC) and reverse traffic channels (RTC). Both share the same frequency assignment, and each Traffic/Access channel is identified by a distinct user long code. The access channel is used by the mobile to initiate communication with the base station and to respond to paging channel messages. The access channel is a random access channel with each channel user uniquely identified by their long codes. The Reverse CDMA channel may contain a maximum of 32 ACs per supported paging channel. While the RTC operates on a variable data rate, the AC works at a fixed data rate of 4800 bps.

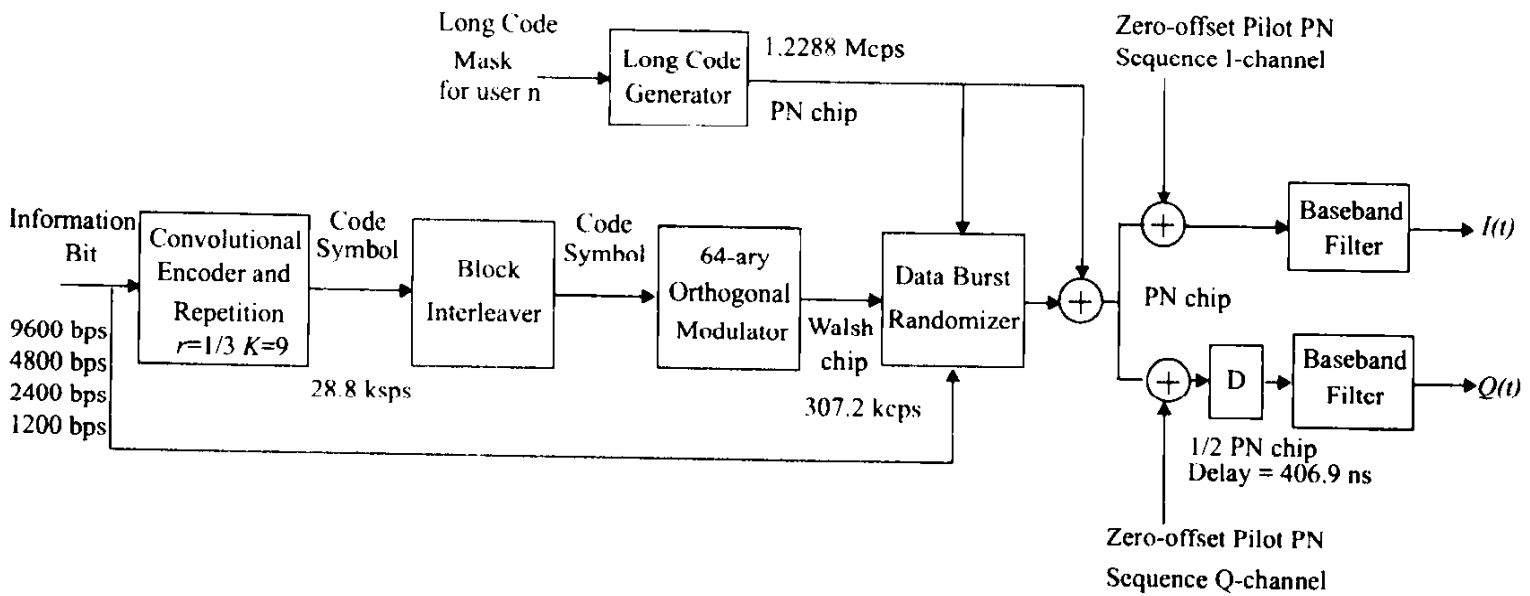



Figure 10.17
Reverse IS-95 channel modulation process for a single user.

#### 10.4.3.1 Convolutional Encoder and Symbol Repetition

The convolutional coder used in the reverse traffic channel is rate 1/3 and constraint length 9. The three generator vectors $g_0$, $g_1$, and $g_2$ are 557 (octal), 663 (octal), and 771 (octal), respectively.

Coded bits after the convolutional encoder are repeated before interleaving when the data rate is less than 9600 bps. This is identical to the method used on the forward channel. After repetition, the symbol rate out of the coder is fixed at 28,800 bps.

#### 10.4.3.2 Block Interleaver

Block interleaving is performed following convolutional encoding and repetition. The block interleaver spans 20 ms, and is an array with 32 rows and 18 columns. Code symbols are written into the matrix by columns and read out by rows.

#### 10.4.3.3 Orthogonal Modulation

A 64-ary orthogonal modulation is used for the reverse CDMA channel. One of sixty-four possible Walsh functions is transmitted for each group of six coded bits. Within a Walsh function, sixty-four Walsh chips are transmitted. The particular Walsh function is selected according to the following formula:

$$\text{Walsh function number} = c_0 + 2c_1 + 4c_2 + 8c_3 + 16c_4 + 32c_5,$$

where $c_5$ represents the last coded bit and $c_0$ represents the first coded bit of each group of six coded symbols that are used to select a Walsh function. Walsh chips are transmitted at a rate of 307.2 kcps as shown in equation (10.1)

$$28.8 \text{ kbps} \times (64 \text{ Walsh chips}) / (6 \text{ coded bits}) = 307.2 \text{ kbps} \qquad (10.1)$$

Note that Walsh functions are used for different purposes on the forward and reverse channels. On the forward channel, Walsh functions are used for spreading to denote a particular user channel, while on the reverse channel, Walsh functions are used for data modulation.

#### 10.4.3.4 Variable Data Rate Transmission

Variable rate data are sent on the reverse CDMA channel. Code symbol repetition introduces redundancy when the data rate is less than 9600 bps. A data randomizer is used to transmit certain bits while turning the transmitter off at other times. When the data rate is 9600 bps, all interleaver output bits are transmitted. When the data rate is 4800 bps, half of the interleaver output bits are transmitted, and the mobile unit does not transmit 50% of the time, and so forth (see Table 10.6). Figure 10.18 illustrates the process under different data rates [EIA92]. Data in each 20 ms frame are divided into sixteen power control groups, each with period 1.25 ms. Some power control groups are gated-on, while others are gated-off. The data burst randomizer ensures that every repeated code symbol is transmitted exactly once. During the gate-off process, the mobile

station reduces its EIRP either by at least 20 dB with respect to the power of the most recent gated-on period, or to the transmitter noise floor, whichever is greater. This reduces interference to other mobile stations operating on the same reverse CDMA channel.

The data burst randomizer generates a masking pattern of '0's and '1's that randomly masks the redundant data generated by the code repetition process. A block of 14 bits taken from the long code determines the masking pattern. The last 14 bits of the long code used for spreading in the second to last power control group of the previous frame are used to determine the random mask for the gating. These 14 bits are denoted as

$$b_0\ b_1\ b_2\ b_3\ b_4\ b_5\ b_6\ b_7\ b_8\ b_9\ b_{10}\ b_{11}\ b_{12}\ b_{13}$$

where $b_0$ represents the earliest bit, and $b_{13}$ represents the latest bit. The data randomizer algorithm is as follows:

- If the user data rate is 9600 bps, transmission occurs on all sixteen power control groups.
- If the user data rate is 4800 bps, transmission occurs on eight power control groups given as

  $$b_0,\ 2 + b_1,\ 4 + b_2,\ 6 + b_3,\ 8 + b_4,\ 10 + b_5,\ 12 + b_6,\ 14 + b_7$$

- If the user data rate is 2400 bps, transmission occurs on four power control groups numbered

  1) $b_0$ if $b_8 = 0$, or $2 + b_1$ if $b_8 = 1$
  2) $4 + b_2$ if $b_9 = 0$, or $6 + b_3$ if $b_9 = 1$
  3) $8 + b_4$ if $b_{10} = 0$, or $10 + b_5$ if $b_{10} = 1$
  4) $12 + b_6$ if $b_{11} = 0$, or $14 + b_7$ if $b_{11} = 1$

- If the user data rate is 1200 bps, transmission occurs on two power control groups numbered:

  1) $b_0$ if ($b_8 = 1$ and $b_{12} = 0$), or $2 + b_1$ if ($b_8 = 1$ and $b_{12} = 0$), or $4 + b_2$ if ($b_9 = 0$ and $b_{12} = 1$), or $6 + b_3$ if ($b_9 = 1$ and $b_{12} = 1$);
  2) $8 + b_4$ if ($b_{10} = 0$ and $b_{13} = 0$), or $10 + b_5$ if ($b_{10} = 1$ and $b_{13} = 0$), or $12 + b_6$ if ($b_{11} = 0$ and $b_{13} = 1$), or $14 + b_7$ if ($b_{11} = 1$ and $b_{13} = 1$).

#### 10.4.3.5 Direct Sequence Spreading

The reverse traffic channel is spread by the long code PN sequence which operates at a rate of 1.2288 Mcps. The long code is generated as described in Section 10.4.2.3 for the forward channel. Each Walsh chip is spread by four long code PN chips.

#### 10.4.3.6 Quadrature Modulation

Prior to transmission, the reverse traffic channel is spread by *I* and *Q* channel pilot PN sequences which are identical to those used in the forward CDMA channel process. These pilot sequences are used for synchronization purpose. The reverse link modulation is offset quadrature phase shift keying (OQPSK). The data spread by the *Q* pilot PN sequence is delayed by half a chip (406.901 ns)

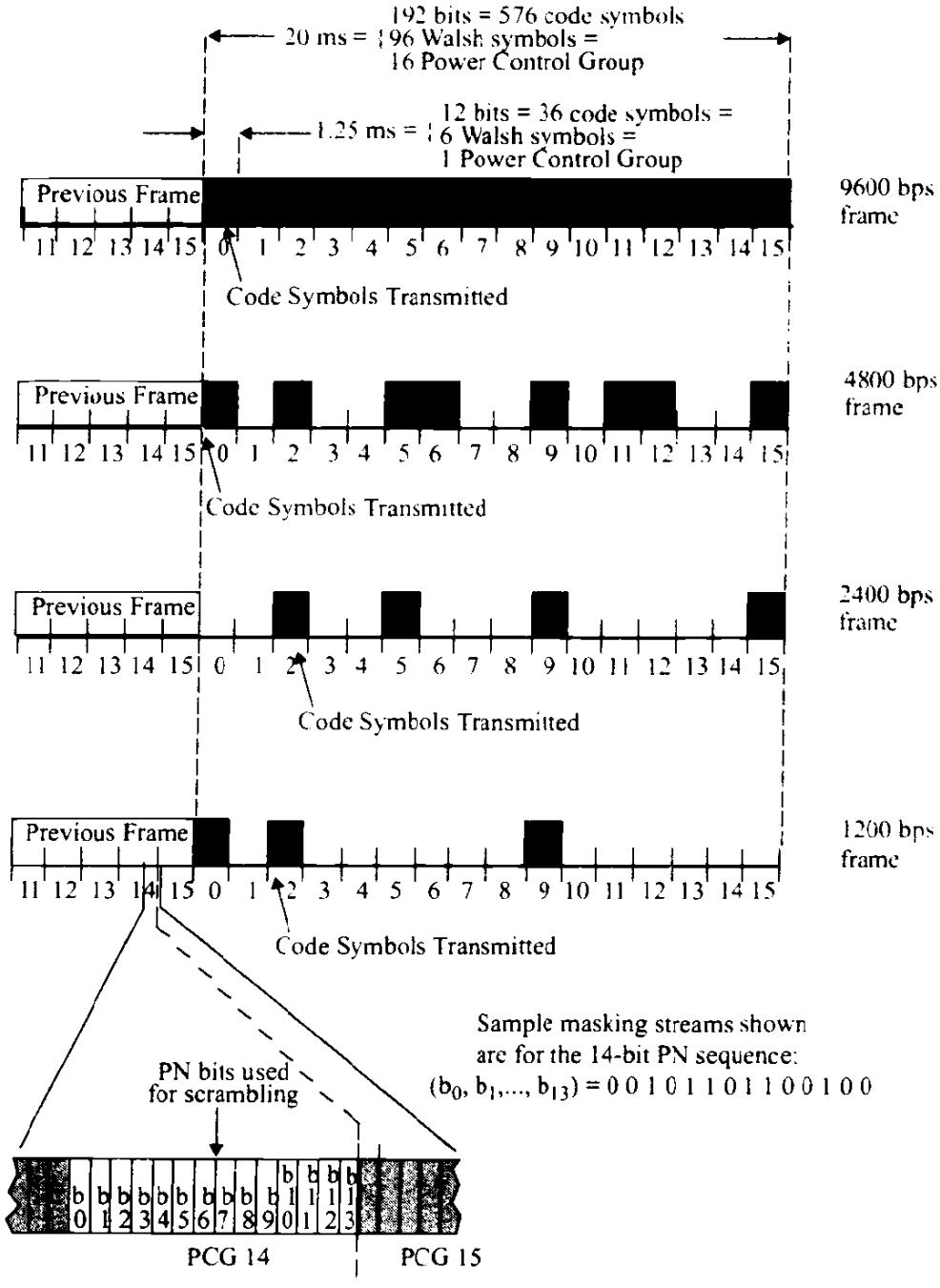



Figure 10.18
Reverse IS-95 channel variable data rate transmission example.

with respect to the data spread by the $I$ pilot PN sequence. This delay is used for improved spectral shaping and synchronization. The binary $I$ and $Q$ data are mapped into the phase according to Table 10.5.

### 10.4.4 IS-95 with 14.4 kbps Speech Coder [ANS95]

To accommodate a higher data rate for better speech quality, the IS-95 air interface structure has been modified to accommodate higher data rate services for PCS. On the reverse link, the convolutional code rate has been changed from rate 1/3 to rate 1/2. On the forward link, the convolutional code rate has been changed from rate 1/2 to rate 3/4 by puncturing two of every six symbols from the original rate 1/2 encoded symbol stream. These changes increase the effective information data rates from 9600, 4800, 2400, and 1200 bps to 14,400, 7200, 3600, and 1800 bps respectively, while keeping the remaining numerology of the air interface structure unchanged.

A variable rate speech coder, QCELP13, has been designed to operate over this higher data rate channel. QCELP13 is a modified version of QCELP. In addition to using the higher data rate for improved quantization of the LPC residual, QCELP13 has several other improvements over the QCELP algorithm including improved spectral quantization, improved voice activity detection, improved pitch prediction, and pitch post-filtering. The QCELP13 algorithm can operate in several models. Mode 0 operates in the same manner as the original QCELP speech coder. QCELP13 codes the speech signal at the highest data rate when active speech is present and at the lowest data rate when idle. Intermediate data rates are used for different modes of speech, such as stationary voiced and unvoiced frames, thereby reducing the average data rate and thus increasing system capacity.

## 10.5 CT2 Standard For Cordless Telephones

CT2 is the second generation of cordless telephones introduced in Great Britain in 1989 [Mor89a]. The CT2 system is designed for use in both domestic and office environments. It is used to provide *telepoint services* which allow a subscriber to use CT2 handsets at a public telepoint (a public telephone booth or a lamp post) to access the PSTN.

### 10.5.1 CT2 Services and Features

CT2 is a digital version of the first generation, analog, cordless telephones. When compared with analog cordless phones, CT2 offers good speech quality, is more resistant to interference, noise, and fading, and like other personal telephones, uses a compact handset with built-in antenna. The digital transmission provides better security. Calls may be made only after entering a PIN, thereby rendering handsets useless to unauthorized users. The battery in a CT2 sub-

scriber unit typically has a talk-time of 3 hours and a standby-time of 40 hours. The CT2 system uses dynamic channel allocation which minimizes system planning and organization within a crowded office or urban environment.

### 10.5.2 The CT2 Standard

The CT2 standard defines how the Cordless Fixed Part (CFP) and the Cordless Portable Part (CPP) communicate through a radio link. The CFP corresponds to a base station and the CPP corresponds to a subscriber unit. The frequencies allocated to CT2 in Europe and Hong Kong are in the 864.10 MHz to 868.10 MHz band. Within this frequency range, forty TDD channels have been assigned, each with 100 kHz bandwidth.

The CT2 standard defines three air interface signaling layers and the speech coding techniques. Layer 1 defines the TDD technique, data multiplexing and link initiation, and handshaking. Layer 2 defines data acknowledgment and error detection as well as link maintenance. Layer 3 defines the protocols used to connect CT2 to the PSTN. Table 10.7 summarizes the CT2 air interface specification.

**Table 10.7 CT2 Radio Specifications Summary**

| Parameter | Specification |
|---|---|
| Frequency | 864.15 — 868.05 MHz |
| Multiple Access | FDMA |
| Duplexing | TDD |
| Number of Channels | 40 |
| Channel Spacing | 100 kHz |
| Number of Channels/Carrier | 1 |
| Modulation Type | 2 level GMSK (BT=0.3) |
| Peak Frequency Deviation Range | 14.4 — 25.2 kHz |
| Channel Data Rate | 72 Kbps |
| Spectral Efficiency | 50 Erlangs/km$^2$/MHz |
| Bandwidth Efficiency | 0.72 bps/Hz |
| Speech Coding | 32 kbps ADPCM (G.721) |
| Control Channel Rate (net) | 1000/2000 bps |
| Max. Effective Radiated Power | 10 mW |
| Power Control | Yes |
| Dynamic Channel Allocation | Yes |
| Receiver Sensitivity | 40 dB V/m or better @ BER of 0.001 |
| Frame Duration | 2 ms |
| Channel Coding | (63,48) cyclic block code |

**Modulation** — All channels use Gaussian filtered binary frequency-shift keying (GFSK) with bit transitions constrained to be phase continuous. The most commonly used filter has a bandwidth-bit period product $BT = 0.3$, and the peak frequency deviation is a maximum of 25.2 kHz under all possible data patterns. The channel transmission rate is 72 kbps.

**Speech Coding** — Speech waveforms are coded using ADPCM with a bit rate of 32 kbps [Det89]. The algorithm used is compliant with CCITT standard G.721.

**Duplexing** — Two-way full duplex conversation is achieved using time division duplex (TDD). A CT2 frame has a 2 ms duration and is divided equally between the forward and reverse link. The 32 kbps digitized speech is transmitted at a 64 kbps rate. Each 2 ms of user speech is transmitted in a 1 ms, with the 1ms gap used for the speech return path. This eliminates the need for paired frequencies or a duplex filter in the subscriber unit. Since each CT2 channel supports 72 kbps of data, the remaining 8 kbps is used for control data (the D subchannel) and burst synchronization (the SYN subchannel). Depending on CT2 situations, the channel bandwidth may be allocated to one or more of the subchannels. The different possible subchannel combinations are called *multiplexes*, and three different multiplexes may be used in CT2 (additional details may be found in [Ste90],[Pad95]).

## 10.6 Digital European Cordless Telephone (DECT)

The Digital European Cordless Telephone (DECT) is a universal cordless telephone standard developed by the European Telecommunications Standards Institute (ETSI) [Och89],[Mul91]. It is the first pan-European standard for cordless telephones and was finalized in July 1992.

### 10.6.1 Features and Characteristics

DECT provides a cordless communications framework for high traffic density, short range telecommunications, and covers a broad range of applications and environments. DECT offers excellent quality and services for voice and data applications [Owe91]. The main function of DECT is to provide local mobility to portable users in an in-building Private Branch Exchange (PBX). The DECT standard supports telepoint services, as well. DECT is configured around an open standard (OSI) which makes it possible to interconnect wide area fixed or mobile networks, such as ISDN or GSM, to a portable subscriber population. DECT provides low power radio access between portable parts and fixed base stations at ranges of up to a few hundred meters.

### 10.6.2 DECT Architecture

The DECT system is based on OSI (Open System Interconnection) principles in a manner similar to ISDN. A control plane (C-plane) and a user plane (U-plane) use the services provided by the lower layers (i.e., the physical layer and the medium access control (MAC) layer). DECT is able to page up to 6000 subscribers without the need to know in which cell they reside (no registration required), and unlike other cellular standards such as AMPS or GSM, DECT is not a total system concept. It is designed for radio local loop or metropolitan area access, but may be used in conjunction with wide area wireless systems such as GSM [Mul91]. DECT uses dynamic channel allocation based on signals received by the portable user and is specifically designed to only support handoffs at pedestrian speeds.

**Physical Layer** — DECT uses a FDMA/TDMA/TDD radio transmission method. Within a TDMA time slot, a dynamic selection of one out of ten carrier frequencies is used. The physical layer specification requires that the channels have a bandwidth which is 1.5 times the channel data rate of 1152 kbps, resulting in a channel bandwidth of 1.728MHz. DECT has twenty-four time slots per frame, and twelve slots are used for communications from the fixed part to the portable (base to handset) and twelve time slots for portable to fixed (handset to base) communications. These twenty-four time slots make up a DECT frame which has a 10 ms duration. In each time slot, 480 bits are allocated for 32 synchronization bits, 388 data bits, and 60 bits of guard time. The DECT TDMA time slot and frame structures are shown in Figure 10.18.

**Medium Access Control (MAC) Layer** — The MAC layer consists of a paging channel and a control channel for the transfer of signaling information to the C-plane. The U-plane is served with channels for the transfer of user information (for ISDN services and frame-relay or frame-switching services). The normal bit rate of the user information channel is 32 kbps. DECT, however, also supports other bit rates. For example, 64 kbps and other multiples of 32 kbps for ISDN and LAN-type applications. The MAC layer also supports handover of calls and a broadcast "beacon" service that enables all idle portable units to find the best fixed radio port to lock onto.

**Data Link Control (DLC) Layer** — The DLC layer is responsible for providing reliable data links to the network layer and divides up the logical and physical channels into time slots for each user. The DLC provides formatting and error protection/correction for each time slot.

**Network Layer** — The network layer is the main signaling layer of DECT and is based on ISDN (layer 3) and GSM protocols. The DECT network layer provides call control and circuit-switched services selected from one of the DLC services, as well as connection-oriented message services and mobility management.

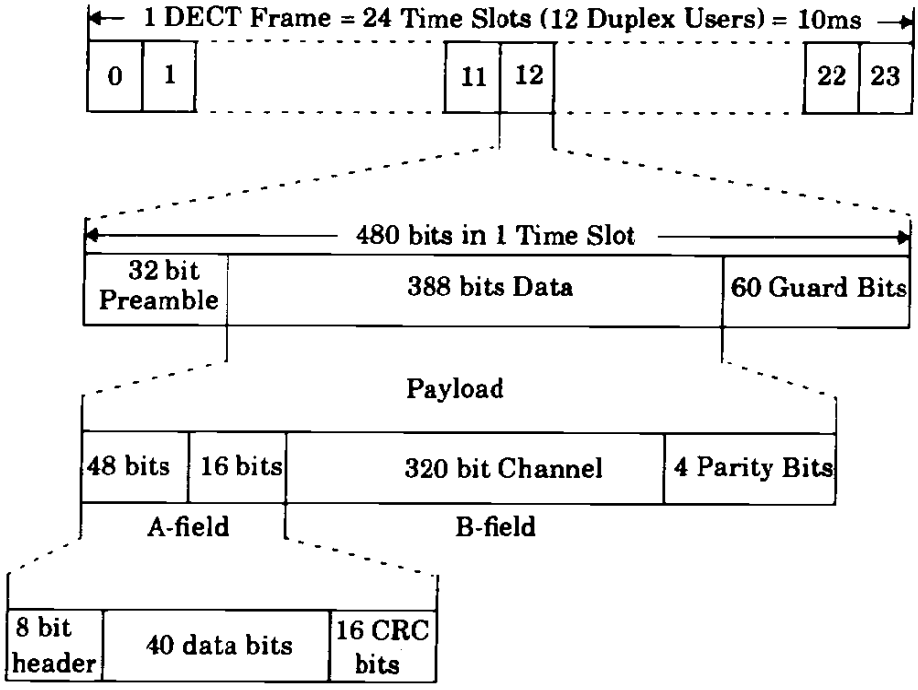



Figure 10.18
DECT TDMA frame structure.

### 10.6.3 DECT Functional Concept

The DECT subsystem is a microcellular or picocellular cordless telephone system that may be integrated with or connected to a Private Automatic Branch Exchange (PABX) or to the Public Switched Telephone Network (PSTN). A DECT system always consists of the following five functional entities as shown in Figure 10.19:

- **Portable Handset (PH)** — This is the mobile handset or the terminal. In addition, cordless terminal adapters (CTAs) may be used to provide fax or video communications.
- **Radio Fixed Part (RFP)** — This supports the physical layer of the DECT common air interface. Every RFP covers one cell in the microcellular system. The radio transmission between RFP and the portable unit uses multi carrier TDMA. A full duplex operation is achieved using time division duplexing (TDD).
- **Cordless Controller (CC or Cluster Controller)** — This handles the MAC, DLC, and network layers for one or a cluster of RFPs and thus forms the central control unit for the DECT equipment. Speech coding is done in the CC using 32 kbps ADPCM.
- **Network-specific Interface Unit** — This supports the call completion

facility in a multi handset environment. The interface recommended by the CCITT is the G.732 based on ISDN protocols.

- **Supplementary Services** — This provides centralized authentication and billing when DECT is used to provide telepoint services, and provides mobility management when DECT is used in multilocation PABX network.

Since the system is limited by $C/I$, the capacity can be increased and the interference from other systems decreased by installing the RFPs in closer proximity. This is illustrated in Figure 10.19.

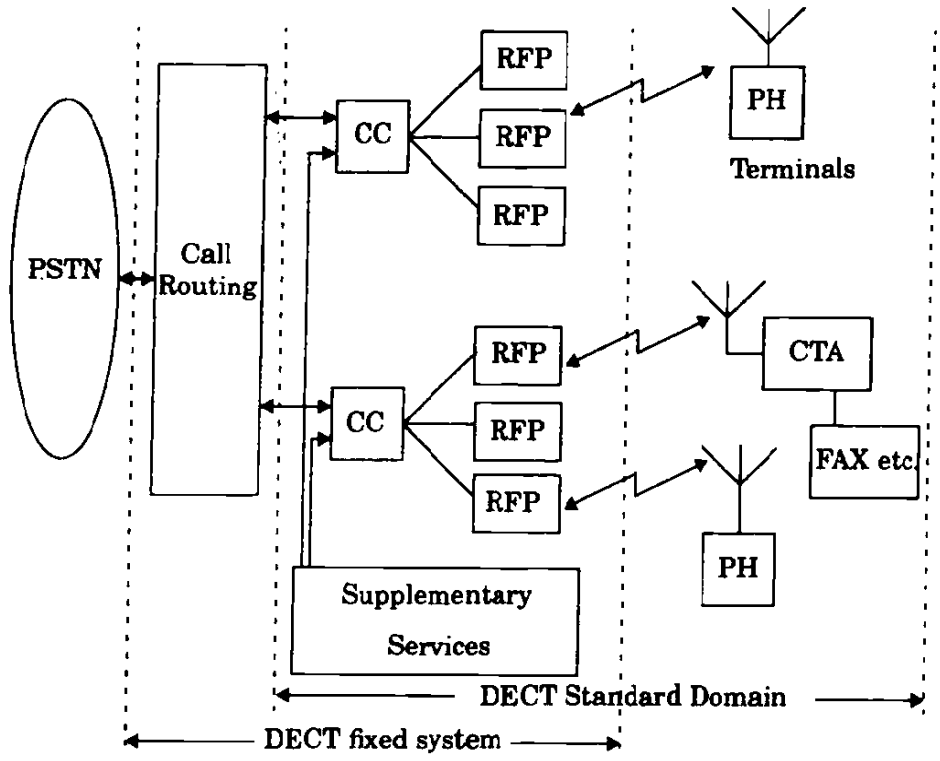



Figure 10.19
DECT functional concept.

### 10.6.4 DECT Radio Link

DECT operates in the 1880 MHz to 1900 MHz band. Within this band, the DECT standard defines ten channels from 1881.792 MHz to 1897.344 MHz with a spacing of 1728 kHz. DECT supports a Multiple Carrier/TDMA/ TDD structure. Each base station provides a frame structure which supports 12 duplex speech channels, and each time slot may occupy any of the DECT channels. Thus, DECT base stations support FHMA on top of the TDMA/TDD structure. If the frequency hopping option is disabled for each DECT base station, a total of 120 channels within the DECT spectrum are provided before frequency reuse is

required. Each time slot may be assigned to a different channel in order to exploit advantages provided by frequency hopping, and to avoid interference from other users in an asynchronous fashion.

**Channel Types** — DECT user data is provided in each B-field time slot (see Figure 10.18). 320 user bits are provided during each time slot yielding a 32 kbps data stream per user. No error correction is provided although 4 parity bits are used for crude error detection.

DECT control information is carried by 64 bits in every time slot of an established call (see Figure 10.18). These bits are assigned to one of the four logical channels depending on the nature of the control information. Thus, the gross control channel data rate is 6.4 kbps per user. DECT relies on error detection and retransmission for accurate delivery of control information. Each 64 bit control word contains 16 cyclic redundancy check (CRC) bits, in addition to the 48 control data bits. The maximum information throughput of the DECT control channel is 4.8 kbps.

**Speech Coding** — Analog speech is digitized into PCM using a 8 kHz sampling rate. The digital speech samples are ADPCM encoded at 32 kbps following CCITT G.721 recommendations.

**Channel Coding** — For speech signals, no channel coding is used since DECT provides frequency hopping for each time slot. Channel coding and interleaving are avoided because the DECT system is meant for use in indoor environments where the tolerable end-to-end system delay is small and the channel may be modeled as "on" or "off" (see Chapter 6). However, the control channels use 16 bit cyclic redundancy check (CRC) codes in each time slot.

**Modulation** — DECT uses a tightly filtered GMSK modulation technique. As discussed in Chapter 5, minimum shift keying (MSK) is a form of FSK where the phase transitions between two symbols are constrained to be continuous. Before the modulation, the signal is filtered using a Gaussian shaping filter.

**Antenna Diversity** — In DECT, spatial diversity at the RFP (base station) receiver is implemented using two antennas. The antenna which provides the best signal for each time slot is selected. This is performed on the basis of a power measurement or alternatively by using an appropriate quality measure (such as interference or BER). Antenna diversity helps solve fading and interference problems. No antenna diversity is used at the subscriber unit.

## 10.7 PACS — Personal Access Communication Systems

PACS is a third generation Personal Communications System originally developed and proposed by Bellcore in 1992 [Cox87], [Cox92]. PACS is able to support voice, data, and video images for indoor and microcell use. PACS is designed to provide coverage within a 500 meter range. The main objective of PACS is to integrate all forms of wireless local loop communications into one sys-

CHAPTER 3

# General Packet Radio Service

The main goal of integrating the *General Packet Radio Service* (GPRS) into the *Global System for Mobile communication* (GSM) is to use the GSM radio resources more efficiently than is possible with the existing GSM phase 2 services. This is realized by allowing a number of logical connections per bearer to co-exist and by multiplexing packet data of these connections onto the GSM physical channels (see Chapter 2).

GPRS has been standardized by the ETSI as part of the GSM Phase 2+ development. The phase 2+ specifications are defining the implementation of packet switching within GSM, which is essentially a circuit-switched (CS) technology.

Packet-switching means that GPRS radio resources are used only when users are actually sending or receiving data. Rather than dedicating a radio channel to a mobile data user for a fixed period of time, the available radio resource can be shared between several users. The actual number of users supported depends on the application being used and how much data is being transferred. Through multiplexing of several logical connections to one or more GSM physical channels, GPRS reaches a flexible use of channel capacity for applications with variable bit rates.

Additionally it brings *Internet Protocol* (IP) capability to the GSM network and enables the access to a wide range of public and private data networks using industry standard data protocols such as *Transport Control Protocol* (TCP)/*Internet Protocol* (IP) and X.25.

The standardization of the GPRS was broadly concluded in 1997. However, details were regularly discussed in the boards of standardization and the GPRS standard was regularly modified. First proposals for a packed-oriented data service in GSM were published in 1991 [Brasche and Walke (1997); Decker (1993, 1995); Walke et al. (1991a,b)].

The first GPRS-based services have been available since 2001 in Europe. Many countries worldwide will introduce GPRS in the next few years. With these new services mobile multimedia applications with net bit rates of up to 117 kbit/s will be offered and established on the market.

## 3.1 Design Approach

The motivation to define a GSM packet-switched data service in 1994 was motivated by the growing number of dedicated packet radio networks. The provision of a circuit-switched channel for bursty traffic was felt to be inefficient. Applications in mind were fleet management, logistics, telematics and mobile offices [Hillebrandt (2002)].

In order to realize economically priced packet services, it was the premise to change the GSM components as little as possible and to develop the new service in consideration of the limits of the existing GSM tele- and bearer services [Brasche (1999); Stuckmann (2002b)].

GPRS integrates a packet-based air interface into the existing circuit-switched GSM network. The GSM infrastructure is not to be replaced. A couple of new network elements have been added (see Section 3.2).

The GPRS specification does not provide an upper limit for the amount of data which can be transmitted per access. GPRS was initially designed for

- the frequent, regular transmission of short data packets up to 500 bytes, and
- the irregular transmission of short to medium-sized data packets up to a few kbyte.

The basic approach to integrate the packet data service into the GSM standard represents the reservation and the logical subdivision of certain GSM channels. The number of channels allocated for GPRS can be dynamically adapted to the load situation in the respective radio cell.

## 3.2 Logical Architecture

The existing GSM network does not provide sufficient functionality to realize a packet data service. Integrating GPRS into a GSM network requires the addition of components which provide the packet-switched service (see Figure 3.1). Hence, the GSM network is extended by two additional nodes:

**Gateway GPRS Support Node (GGSN):** The GGSN serves as the interface towards external *Packet Data Networks* (PDNs) or other *Public Land Mobile Networks* (PLMNs). Here, switching functions are fulfilled, e.g., the evaluation of *Packet Data Protocol* (PDP) addresses and the routing to mobile subscribers via the SGSN.

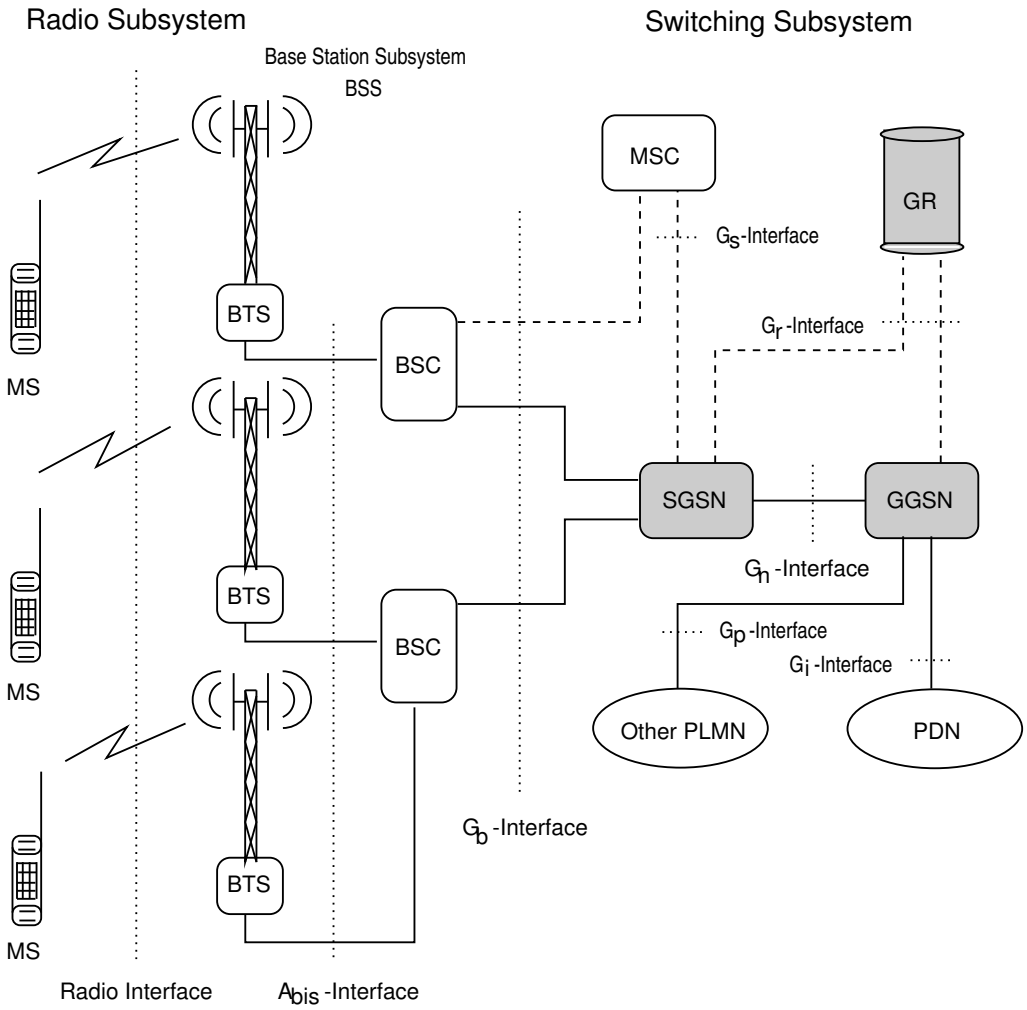



**Figure 3.1:** GPRS logical architecture

**Serving GPRS Support Node (SGSN):** The SGSN represents the GPRS switching center in analogy to the *Mobile-services Switching Center* (MSC). Packet data addresses are evaluated and mapped onto the *International Mobile Subscriber Identity* (IMSI). The SGSN is responsible for the routing inside the packet radio network and for mobility and resource management. Furthermore, it provides authentication and encryption for the GPRS subscribers.

For the communication between the SGSN and the GGSN within one PLMN, the Intra PLMN *IP Version 6* (IPv6) or *IP Version 4* (IPv4) backbone is used. SGSN and GGSN encapsulate and decapsulate, respectively, the packets using a special protocol called the *GPRS Tunneling Protocol* (GTP) that operates on top of standard TCP/IP protocols. The SGSN and GGSN functions may also be combined into one physical node.

## 3.3 Radio Interface

The radio interface $U_m$ is located between MS and BSS. The GSM recommendations combine a *Frequency Division Multiplex* (FDM) and a *Time Division Multiplex* (TDM) scheme (see Chapter 2).

The communication over the GPRS radio interface comprises functions of the *Radio Link Control* (RLC)/*Medium Access Control* (MAC) layer and the physical layer. For the transfer of layer-2 messages over the radio interface the GSM logical channel concept is reused.

### 3.3.1 Channel Concept

The basic approach to integrate the packet data service into the GSM standard represents the reservation and the logical subdivision of certain GSM channels for GPRS. The physical channels dedicated to packet data traffic are called *Packet Data Channels* (PDCHs).

Similar to GSM, logical channels are mapped onto physical channels using a cyclically recurring multiframe structure. Two 26-frame multiframes are assembled to one GPRS 52-frame multiframe. A 52-frame multiframe represents one physical GPRS channel consisting of 12 radio blocks and four `Idle` bursts, each block comprising four radio bursts distributed on time slots with the same *Timeslot Number* (TN) in consecutive TDMA frames (see Figure 3.2 and Figure 3.3).

Since one TDMA frame represents eight physical channels to be shared by GSM and GPRS, there have to be strategies on how to distribute these resources, whether fixed or on-demand (see Figure 3.4). In [3GPP TSG GERAN (2002e)] the principles concerning this task are specified. The specific details of implementation are left to manufacturers and network operators.

#### 3.3.1.1 The Master-slave Concept

One or more PDCHs are run as *Master Packet Data Channels* (MPDCHs) (see Figure 3.3), and provide *Packet Common Control Channels* (PCCCHs) that carry the necessary control and signaling information to initiate a packet data transmission. These PCCCHs have to be run if signaling information is not transmitted via existing *Common Control Channels* (CCCHs).

Furthermore PDCHs operate as slaves and are utilized for transmission of user data over *Packet Data Traffic Channels* (PDTCHs) and associated control over *Packet Associated Control Channels* (PACCHs).

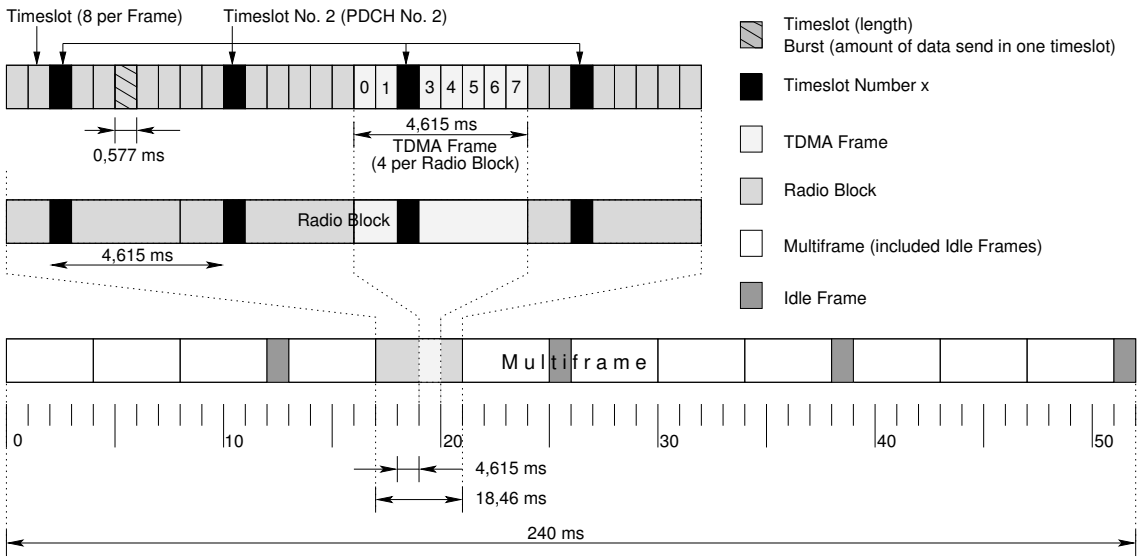



**Figure 3.2:** Duration of a time slot, TDMA frame, radio block and 52-frame

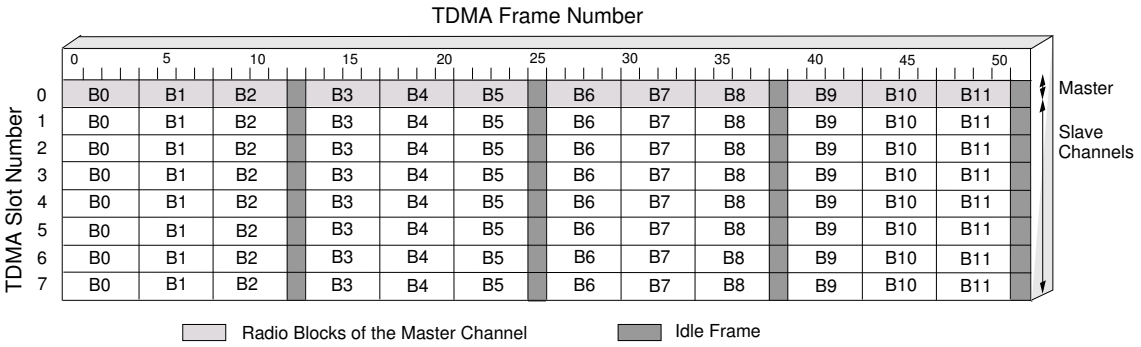



**Figure 3.3:** GPRS channel structure

#### 3.3.1.2 The Capacity-on-demand Concept

GPRS does not require fixed allocated PDCHs. Capacity assignment for packet data transmission can be done according to actual demand. The decision about the number of fixed and on-demand PDCHs is left with the radio network operator.

There are several mechanisms to increase or diminish the number of actually assigned PDCHs on a capacity-on-demand base.

**Load monitoring** The PDCH utilization is supervised by a load monitor instance that should be implemented as part of the MAC functionality.

**Dynamic allocation of PDCHs** Unused channels, whether by GSM or GPRS, may be allocated as PDCHs to increase the GPRS QoS. PDCHs are released in order to meet the obligations of higher priority services.

**Release of PDCHs** Fast release of PDCHs is an important criterion for a pool of radio resources to be dynamically shared by circuit as well as packet switched services. To achieve this goal, there are the following possibilities:

- Channel release is delayed until there are no more active packet data flows on the regarded PDCHs.
- Each user having allocated the PDCHs to be released has to be notified, e.g. by `Packet Resource Reassignment` messages from the network side.

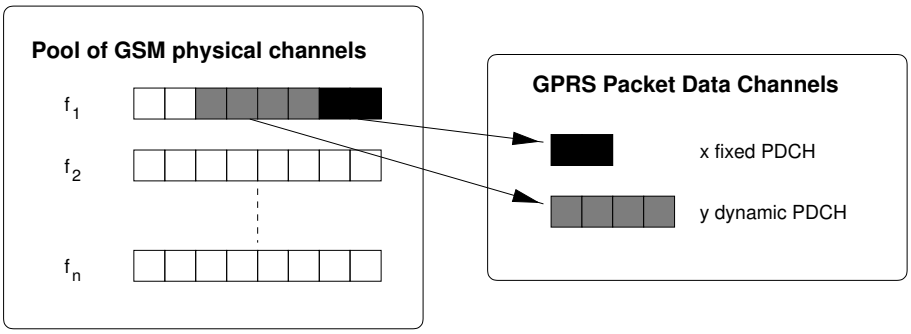



**Figure 3.4:** Assignment of GSM physical channels for GPRS

- The channel release notification is broadcast. There has to be one PDCHs monitored by all MSs to assure the reception of this notification.

In practice, a combination of the above methods can be implemented. In case of an MS not receiving a channel release notification there will be temporary channel interference because of the MS using a channel provided for other services. However, the MS will terminate the packet data flow as soon as it receives an inappropriate response on the downlink, and initiate a retransmission on a different PDCH.

### 3.3.2 Logical Channels

The packet data logical channels are mapped onto the physical channels dedicated to packet data (PDCHs).

Table 3.1 lists the GPRS logical channels and their functions. A detailed description for each channel is presented below.

#### 3.3.2.1 Packet Common Control Channel (PCCCH)

The PCCCH comprises the following logical channels for GPRS common control.

**PRACH:** the *Packet Random Access Channel* (PRACH) is used by the MS in the uplink direction to initiate uplink transfer, e.g., for sending data or a paging response. In other words the channel is used by the MS to initiate a packet transfer or respond to paging messages. On this channel an MS transmits access bursts with long guard times. On receiving access bursts, the BSS assigns a value for the *Timing Advance* (TA) algorithm to each terminal.

**PPCH:** the *Packet Paging Channel* (PPCH) is used in the downlink direction to page an MS prior to downlink packet transfer. The PPCH uses paging groups in order to allow usage of *Discontinuous Reception* (DRX) mode.

**PAGCH:** the *Packet Access Grant Channel* (PAGCH) is used in the establishment phase of the packet transfer to send a resource assignment in the downlink direction to an MS requesting a packet transfer. Additionally, resource assignment for a downlink packet transfer can be sent on a PACCH if the MS is currently involved in a packet transfer.

**PNCH:** the *Packet Notification Channel* (PNCH) is used in the downlink direction to send a *PTM Multicast* (PTM-M) notification to a group of MSs prior to a PTM-M packet transfer. The notification has the form of a resource assignment for the packet transfer. DRX mode will be provided for monitoring the PNCH. Furthermore, a `PTM-`

**Table 3.1:** GPRS logical channels

| Group | Channel | Name | Direction | Function |
|---|---|---|---|---|
| PCCCH | PRACH | Packet Random Access Channel | UL | random access |
| | PPCH | Packet Paging Channel | DL | paging |
| | PAGCH | Packet Access Grant Channel | DL | access grant |
| | PNCH | Packet Notification Channel | DL | multicast |
| PBCCH | PBCCH | Packet Broadcast Control Channel | DL | broadcast |
| PTCH | PDTCH | Packet Data Traffic Channel | UL/DL | data |
| | PACCH | Packet Associated Control Channel | UL/DL | assoc. control |

`M new message` indicator may optionally be sent on all individual paging channels to inform MSs interested in PTM-M when they need to listen to the PNCH.

#### 3.3.2.2 Packet Broadcast Control Channel (PBCCH)

The *Packet Broadcast Control Channel* (PBCCH) broadcasts packet data specific system information. If the PBCCH is not allocated, the *Broadcast Control Channel* (BCCH) transmits the system information to all GPRS terminals in a radio cell.

#### 3.3.2.3 Packet Traffic Channel (PTCH)

The PTCH is allocated to carry user data and associated control information and is subdivided into the following channels.

**PDTCH:** the *Packet Data Traffic Channel* (PDTCH) is a channel allocated for data transfer. It is temporarily dedicated to one MS or in the *Point-To-Multipoint* (PTM) case to a group of MSs. In multislot operation, one MS may use multiple PDTCHs, e.g., more than one PDTCH, simultaneously for individual packet transfer.

**PACCH:** the *Packet Associated Control Channel* (PACCH) is used to convey control information related to a given MS, such as an acknowledgement message. The PACCH also carries resource assignment and reassignment messages comprising the assignment of capacity for PDTCHs and for further occurrences of PACCH. One PACCH is associated with one or several PDTCHs assigned to one MS.

### 3.3.3 Mapping of the Logical Channels

The mapping of the logical channels is defined by the multiframe structure. As described in Section 3.3.1 the multiframe structure for a PDCH consists of 52 TDMA frames, divided into 12 blocks (of four frames each) and four idle frames according to Figure 3.5. The mapping of the logical channels onto the radio block periods is defined by means of the ordered list of blocks (B0, B6, B3, B9, B1, B7, B4, B10, B2, B8, B5, B11). With this list a certain number of channels per PDCH is uniformly distributed over the multiframe. For example one PBCCH and three PAGCHs on the master channel would be allocated to B0, B6, B3 and B9 of the first PDCH.

PDCH that contain PCCCHs are indicated on the BCCH. The number of PDCHs containing PCCCHs is defined by the `BS PCC CHANS` parameter. The following restrictions are defined in [3GPP TSG GERAN (2002d)] concerning PCCCHs.

Only one PDCH—the master channel—contains one or more PBCCHs All other channels do not contain PBCCHs and are called slave channels. The logical channel PAGCH can only be mapped on the master channel and the slave channels that contain PCCCHs.

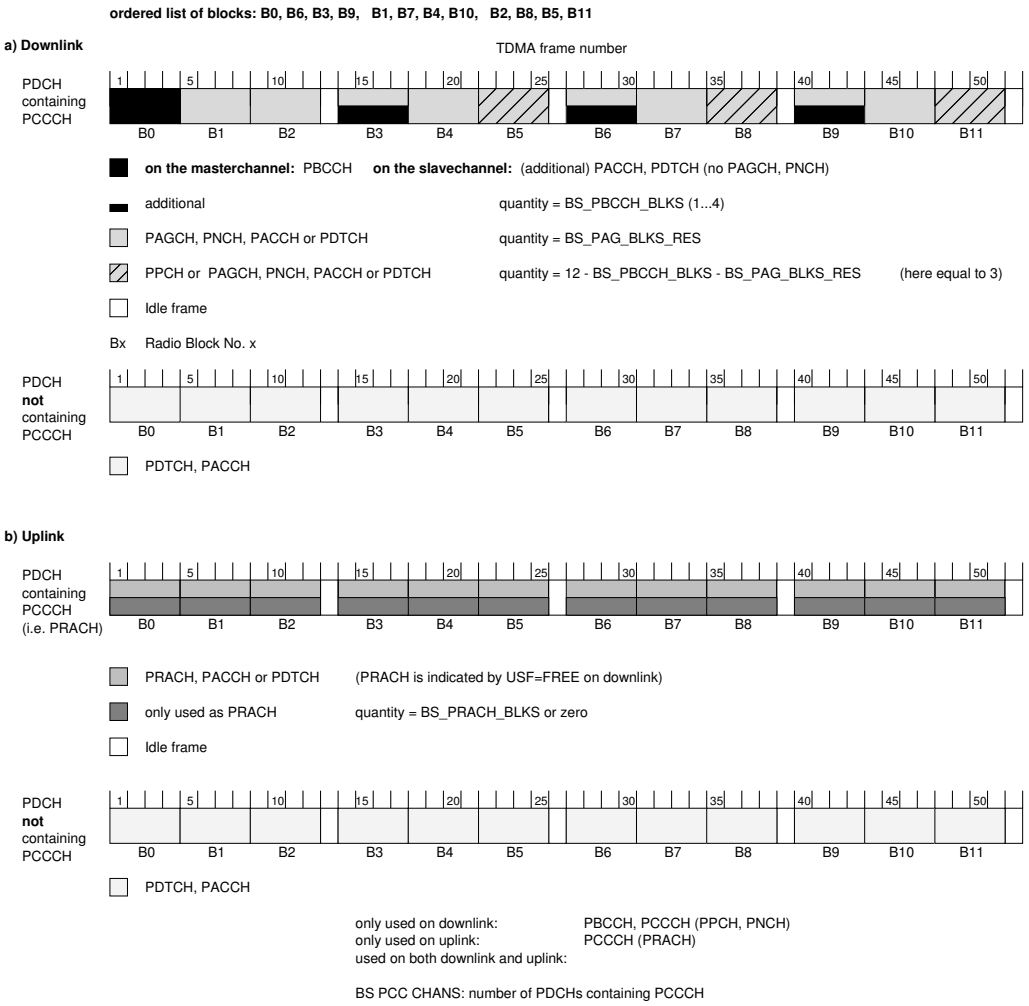



**Figure 3.5:** Mapping of the logical channels onto the radio blocks

The response message to a `Packet Channel Request` has to be sent on the same PDCH that was used for the `Packet Channel Request`.

The PRACH can only be mapped on an uplink PDCH containing PCCCHs. On PRACHs, access bursts are used, while on all other packet data logical channels, radio blocks comprising four normal bursts are used. The only exception are messages on uplink PACCHs, which can comprise four consecutive access bursts to increase robustness. The detailed mapping rules of the single channels as defined in [3GPP TSG GERAN (2002d)] are described in the following (see Figure 3.5).

#### 3.3.3.1 Downlink

**Mapping of PBCCHs:** On the downlink of the master channel the first block in the ordered list of blocks (B0) is always used as a PBCCH. If required up to three more blocks on the same PDCH can be used as additional PBCCHs. The parameter `BS PBCCH BLKS` indicates the total number of PBCCH blocks (from one up to four) on the downlink master channel. Therefore on this PDCH the first `BS PBCCH BLKS` blocks in the ordered list of blocks are reserved for PBCCHs.

Any additional PDCH containing PCCCHs is indicated on the PBCCH. On these PDCHs the `BS PBCCH BLKS` first blocks in the ordered list of blocks are used as

PDTCHs or PACCHs in the downlink.

**Mapping of PAGCHs, PNCHs, PDTCHs and PACCHs:** On any PDCH containing PCCCHs (with or without PBCCH), the next `BS PAG BLKS RES` blocks in the ordered list of blocks are used as PAGCHs, PNCHs, PDTCHs or PACCHs. The `BS PAG BLKS RES` parameter is broadcast on the PBCCH and ranges from 0 up to `12 - BS PBCCH BLKS`.

**Mapping of PPCHs:** The remaining blocks in the ordered list of blocks are used as PPCHs. But they can also be used as PAGCHs, PNCHs, PDTCHs or PACCHs, if no paging messages are waiting for transmission.

#### 3.3.3.2 Uplink

**Mapping of PRACHs:** On a PDCH containing PCCCHs, all blocks in the multiframe can be used as a PRACH—indicated by an Uplink State Flag in the previous downlink message—or as a PDTCH or a PACCH.
Optionally the `BS PRACH BLKS` first blocks in the ordered list of blocks are only used as PRACHs. The `BS PRACH BLKS` parameter is broadcast on the PBCCH. The remaining blocks in the multiframe are used as PRACH, PDTCH or PACCH.

**Mapping of PACCHs and PDTCHs:** On a PDCH that does not contain a PCCCH, i.e., that does not contain a PRACH (`BS PRACH BLKS`=0), all blocks are used as PDTCHs or PACCHs.

## 3.4 GPRS User Plane

The GPRS protocol architecture follows the *International Standards Organization* (ISO)/ *Open Systems Interconnection* (OSI) reference model [ISO/IEC JTC1 (1994)]. This reference model provides a common basis for the coordination of standards development for the purpose of systems interconnection, while allowing existing standards to be placed into perspective within the overall reference model. It divides the functions of a system into seven layers. The communication between the corresponding partner entities of each layer is specified by communication protocols comprising the syntax, semantics, timing and pragmatics of the exchanged data units.

The GPRS user plane, also called transmission plane, comprises all protocols used for user data transmission. In GPRS networks the user plane can be considered separate from the control plane, since communication is ongoing in two phases. While a so-called *Packet Data Protocol* (PDP) context, a logical context between an MS and an external *Packet Data Network* (PDN) based on TCP/IP or X.25, can be set up without actually transmitting user data, the user plane will only be utilized when the MS is actually transmitting or receiving user data. Nevertheless some protocols of the user plane are reused to carry signaling messages (see Section 3.5).

The kernel of the GPRS standard is the *Radio Link Control* (RLC)/*Medium Access Control* (MAC) protocol that operates on top of the GSM channel structure (see Section 3.3.1) and realizes the reliable packet-oriented transmission between *Mobile Station* (MS) and *Base Station Subsystem* (BSS) by statistical multiplexing of several logical connections on the PDCHs available for GPRS in the radio cell. While RLC/MAC contains already an *Automatic Repeat Request* (ARQ) scheme to handle transmission errors on the air interface, the *Logical Link Control* (LLC) protocol provides a reliable ciphered link between MS and SGSN applying flow control and error handling mechanisms known from link layer protocols in fixed networks like *Integrated Services Digital Network* (ISDN). The ARQ scheme in LLC is not designed to handle transmission errors on the air interface,

but for handling frame losses and errors in the fixed network or for link recovery after a cell change.

The *Base Station Subsystem GPRS Protocol* (BSSGP), which is operating on top of a network service such as *Frame Relay* (FR), is responsible for flow control between BSS and SGSN. To support different network layer standards that may be used in GPRS networks the *Sub-Network Dependent Convergence Protocol* (SNDCP) adapts the network layer *Protocol Data Units* (PDUs) to the GPRS logical link. Finally the *GPRS Tunneling Protocol* (GTP) is used as part of the GPRS user plane to realize a transport service for IP datagrams by establishing an IP tunnel through the GPRS *Core Network* (CN) between SGSN and GGSN.

This section will describe all GPRS-specific protocols (see shaded blocks in Figure 3.6) that are involved in user data transmission in a top-down approach.

### 3.4.1 GPRS Tunneling Protocol (GTP)

In GPRS networks the interworking with TCP/IP or X.25 based networks is foreseen [3GPP TSG CN (2002d)] on the $G_i$ interface. Between SGSN and GGSN on the $G_n$ interface, encapsulated data packets are transmitted with the help of the GTP. GTP signaling is used by the *GPRS Mobility Management* (GMM)/*Session Management* (SM) (see Section 3.5) to create, modify and delete tunnels [3GPP TSG CN (2002c)]. GTP uses the *User Datagram Protocol* (UDP) for the transmission in the core network.

Tunnels are used to carry encapsulated PDUs between a given GSN pair for individual mobile stations. The *Tunnel Identifier* (TID), part of the GTP header, indicates to which tunnel a particular datagram belongs. The TID specifies the PDP context (see Section 3.5). Like this, datagrams injected into the GPRS network at the GGSN are tunneled to the correct SGSN area, where the MS is presently located, which means that mobility-related routing is completely supported by the GTP. In other mobile core networks, e.g., networks based on the 3GPP2 specifications, this function is realized by *Mobile IP* [Perkins (1996)].

The GTP comprises two different entities, the GTP-U entity that is responsible for user data tunneling and the GTP-C entity responsible for signaling between SGSN and GGSN (see Section 3.5.2) using path, tunnel, location and mobility management messages.

### 3.4.2 Base Station Subsystem GPRS Protocol (BSSGP)

On the $G_b$ interface, the BSSGP [3GPP TSG GERAN (2002c)] provides a connectionless link between BSS and SGSN. This protocol's main task is the flow control for the downlink transfer of LLC PDUs. No flow control is performed in the uplink direction. The SGSN is expected to accept all data that is sent to it—the buffers and link capacity have to be dimensioned according to this premise to avoid loss of uplink data.

#### 3.4.2.1 Flow Control between SGSN and BSS

Each cell in the coverage area of one BSS is virtually connected to the SGSN by a *BSSGP Virtual Connection* (BVC), identified by a *BSSGP Virtual Connection Identifier* (BVCI). The BSS maintains one queue for each BVCI that contains the LLC PDUs for that particular cell. There are several ways to further split these queues, e.g., into one queue per MS, or into one queue that serves all LLC PDUs belonging to a certain delay or precedence class. The queues are filled with downlink LLC PDUs and emptied by the BSS forwarding these PDUs to the MSs addressed. The SGSN is regularly informed by the BSS about the maximum queue size available for each BVC and MS at a given time, and also about the

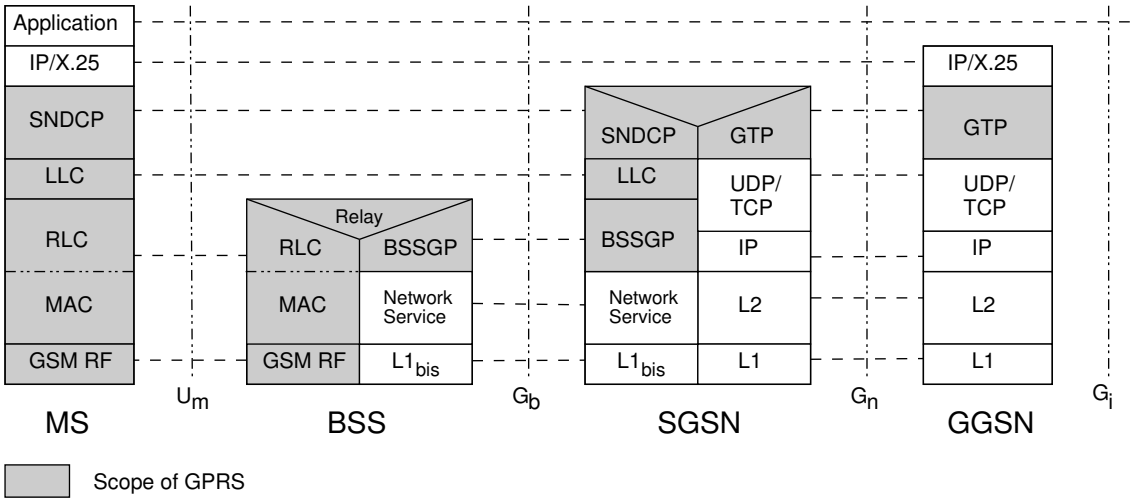



**Figure 3.6:** GPRS user plane

rate at which the available queue size increases. Thus, the SGSN is enabled to estimate the maximum allowable throughput per BVC as well as per MS belonging to the BVC.

Within the SGSN there are queues provided, similar to those in the BSS. Here the downlink LLC PDUs are queued and scheduled as long as the maximum allowable throughput per BVC and per MS is not exceeded. By this mechanism, it is ensured that both the MS and the BSS are able to handle the incoming traffic.

#### 3.4.2.2 BSS Context

The SGSN can provide a BSS with information related to ongoing user data transmission. The information related to one MS is stored in a BSS context. The BSS may contain BSS contexts for several MSs. A BSS context contains a number of BSS *Packet Flow Contexts* (PFCs). Each BSS PFC is identified by a *Packet Flow Identifier* (PFI) assigned by the SGSN. A BSS PFC is shared by one or more activated *Packet Data Protocol* (PDP) contexts with identical or similar negotiated QoS profiles. The data transmission related to PDP contexts that share the same BSS PFC comprises one packet flow.

Three packet flows are predefined, and identified by three reserved PFI values. The BSS does not negotiate BSS PFCs for these pre-defined packet flows with the SGSN. One pre-defined packet flow is used for best-effort service, one is used for SMS, and one is used for signaling. The SGSN can assign the best-effort or SMS PFI to any PDP context. In the SMS case, the BSS handles the packet flow for the PDP context with the same QoS that it handles SMS with.

The combined BSS QoS profile for the PDP contexts that share the same packet flow is called the *Aggregate BSS QoS Profile* (ABQP). The ABQP is considered to be a single parameter with multiple data transfer attributes. It defines the QoS that must be provided by the BSS for a given packet flow between the MS and the SGSN, i.e., for the $U_m$ and $G_b$ interfaces combined. The ABQP is negotiated between the SGSN and the BSS.

A BSS packet flow timer indicates the maximum time for which the BSS may store the BSS PFC. The BSS packet flow timer is started when the BSS PFC is stored in the BSS and when an LLC frame is received from the MS. When the BSS packet flow timer expires the BSS deletes the BSS PFC.

When a PDP context is activated, modified, or deactivated (see Section 3.5.2.3), the SGSN may create, modify, or delete BSS PFCs.

### 3.4.3 Sub-Network Dependent Convergence Protocol (SNDCP)

Network layer protocols are intended to be capable of operating over services derived from a wide variety of subnetworks and data links. GPRS supports several network layer protocols providing protocol transparency for the users of the service. Therefore, all functions related to transfer of network layer PDUs are carried out in a transparent way by the GPRS network entities. SNDCP provides this adaptation of different network layers to the logical link by using *Network layer Service Access Point Identifiers* (N-SAPIs) to map different PDPs onto the services provided by the LLC layer. Additionally SNDCP supports compression of redundant user data and protocol control information. The main functions of the SNDCP layer are [3GPP TSG CN (2002b)]:

- Multiplexing of several PDPs.
- N-PDU buffering.
- Compression and decompression of user data and control information (e.g., TCP/IP header compression).
- Segmentation and reassembly of network PDUs.

The N-PDUs are buffered in the SNDCP layer before they are compressed, segmented and transmitted to the LLC layer. For acknowledged data transfer, the SNDCP entity buffers an N-PDU until successful reception of all segments of the N-PDU have been confirmed. The confirmation is carried out by the underlying layer using a confirm service primitive. N-PDUs buffered at the transmitting side, which which have been acknowledged by the receiver side are cleared. For unacknowledged data transfer, the SNDCP deletes an N-PDU immediately after it has been delivered to the LLC layer. Acknowledged or unacknowledged mode on SNDCP level only means that the SNDCP layer is using the acknowledged or unacknowledged data service from the LLC layer. In the SNDCP layer itself no error handling functions are performed.

The protocol control information compression method is specific for each network layer protocol type. TCP/IP (IPv4) header compression is specified in [Jacobson (1990)]. For data compression [ITU-T (1990)] may be used.

After the SNDCP header is added the compressed N-PDU is segmented if it is longer than the maximum payload size for an LLC frame.

### 3.4.4 Logical Link Control (LLC)

The LLC layer [3GPP TSG CN (2002a)] provides a highly reliable and ciphered link between the MS and the SGSN. It includes functions for:

- Provision of one or more logical connections.
- Sequence control.
- Error detection and correction.
- Flow control.
- Ciphering.

The protocol for the LLC sublayer is based on well-known data link protocols like the *Link Access Procedure on the D-channel* ($\text{LAPD}_\text{m}$) used in GSM and the *High level Data Link Control* (HDLC) protocol standardized by ISO for *Open Systems Interconnection* (OSI).

The key modifications can be summarized as follows:

**Variable frame length** The GPRS protocol architecture allows a variable frame length at the LLC level. Therefore frame delimiters and bit stuffing are not necessary but an additional field is required in the frame header for specification of the frame length.

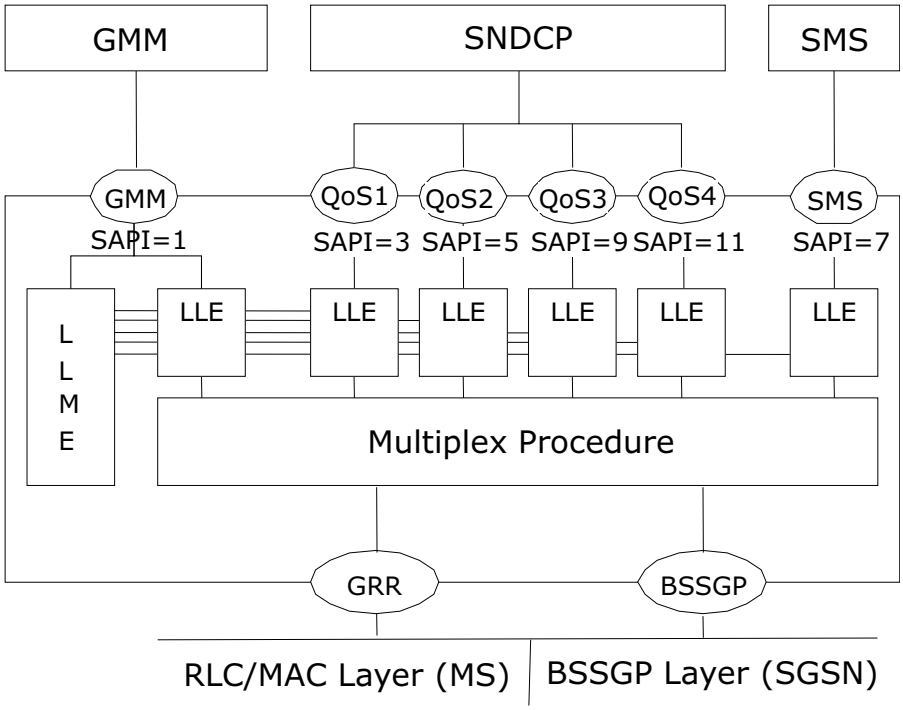



**Figure 3.7:** LLC layer structure

**Prioritized Service Access Points (SAPs)** The priority classes in GPRS are considered in the introduction of new SAPs with priorities.

#### 3.4.4.1 Layer Entities and Service Access Points

The LLC layer provides six SAPs to the upper layer:

- Four SAPs are dedicated to the SNDCP that manages data packet transmission; one SAP exists for each QoS class
- One SAP is dedicated to *GPRS Mobility Management* (GMM) (see Section 3.5)
- One SAP is dedicated to *Short Message Service* (SMS)

The related services are performed by different logical link entities (see Figure 3.7) that are managed by the logical link management entity. Corresponding to four radio priority levels, four *Service Access Point Identifiers* (SAPIs) are required for packet data transmission. SAPIs of the LLC layer identify a point at which LLC services are provided by an LLC entity to a layer-3 entity. With the introduction of service or subscriber differentiation, it is also necessary that flow control and error correction is performed on a priority basis.

#### 3.4.4.2 LLC Frame Structure

The LLC frame format is shown in Figure 3.8. The frame header consists of the address field and the control field and ranges from two to 37 octets. The address field consists of a single octet containing the SAPI (see Section 3.4.4.1) and the identifier of the logical link, the *Data Link Connection Identifier* (DLCI). The control field typically consists of between one and three octets (see Figure 3.9). LLC frames can be classified into information frames (I frames) containing sequence numbers for acknowledged data transmission, unnumbered

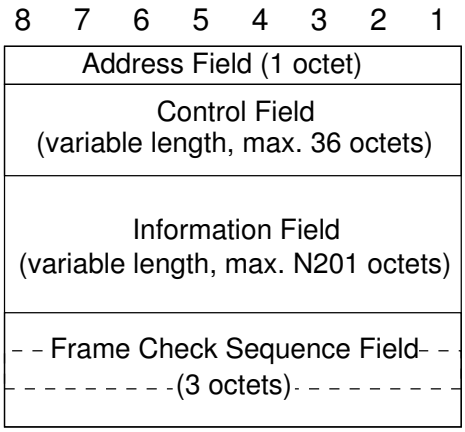



**Figure 3.8:** LLC frame format

Control Field Bits

| | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | Octet |
|---|---|---|---|---|---|---|---|---|---|
| I format (I+S) | 0 | A | X | N(S) | | | | | 1 |
| | N(S) | | | | X | N(R) | | | 2 |
| | N(R) | | | | | | $S_1$ | $S_2$ | 3 |
| S format | 1 | 0 | A | X | X | N(R) | | | 1 |
| | N(R) | | | | | | $S_1$ | $S_2$ | 2 |
| UI format | 1 | 1 | 0 | X | X | N(U) | | | 1 |
| | N(U) | | | | | | E | PM | 2 |
| U format | 1 | 1 | 1 | P/F | $M_4$ | $M_3$ | $M_2$ | $M_1$ | 1 |

| | |
|---|---|
| A | Acknowledgement request bit |
| E | Encryption function bit |
| M | Unnumbered function bit |
| N(R) | Transmitter receive sequence number |
| N(S) | Transmitter send sequence number |
| N(U) | Transmitter unconfirmed sequence number |
| P/F | Poll bit, when issued as a command |
| | Final bit, when issued as a response |
| PM | Protected mode bit |
| S | Supervisory function bit |
| X | Spare bit |

**Figure 3.9:** LLC control field

information frames (UI frames) for unacknowledged data transmission and supervisory frames (S frames) and unnumbered frames (U frames) containing LLC control messages. The *Selective Acknowledgement* (SACK) supervisory frame additionally includes a bitmap field of variable length of up to 32 octets. The information field, if present, follows the control field. The maximum length of the information field depends on the SAPI. The *Frame Check Sequence* (FCS) field consists of a 24-bit *Cyclic Redundancy Check* (CRC) code that is used to detect bit errors in the frame header and information field.

#### 3.4.4.3 LLC Frame Transmission

The LLC layer performs an ARQ protocol based on retransmissions after timeouts or frame loss detections and optionally uses bitmap-based selective acknowledgements.

##### 3.4.4.3.1 LLC Modes

LLC can operate in acknowledged and unacknowledged mode. In unacknowledged mode the network layer PDUs are transmitted in *Asynchronous Disconnected Mode* (ADM) using UI frames (see Figure 3.10). Neither LLC error recovery nor reordering procedures are defined, but transmission and format errors are detected. Duplicate UI frames are discarded.

MSC: UA Transfer

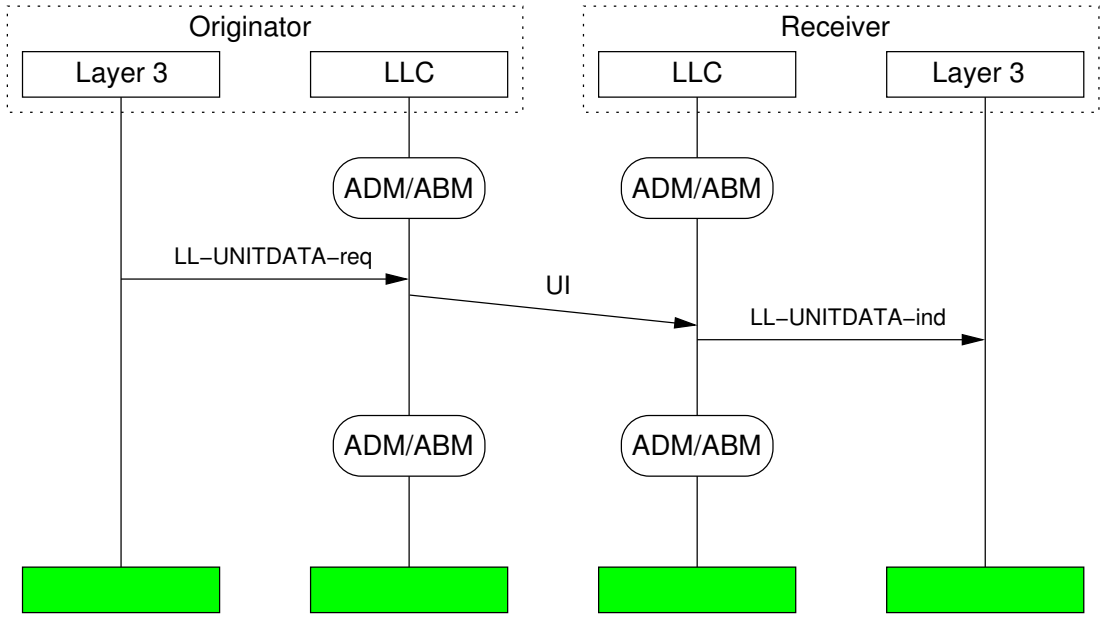



**Figure 3.10:** LLC unacknowledged mode

MSC: ABM Establishment

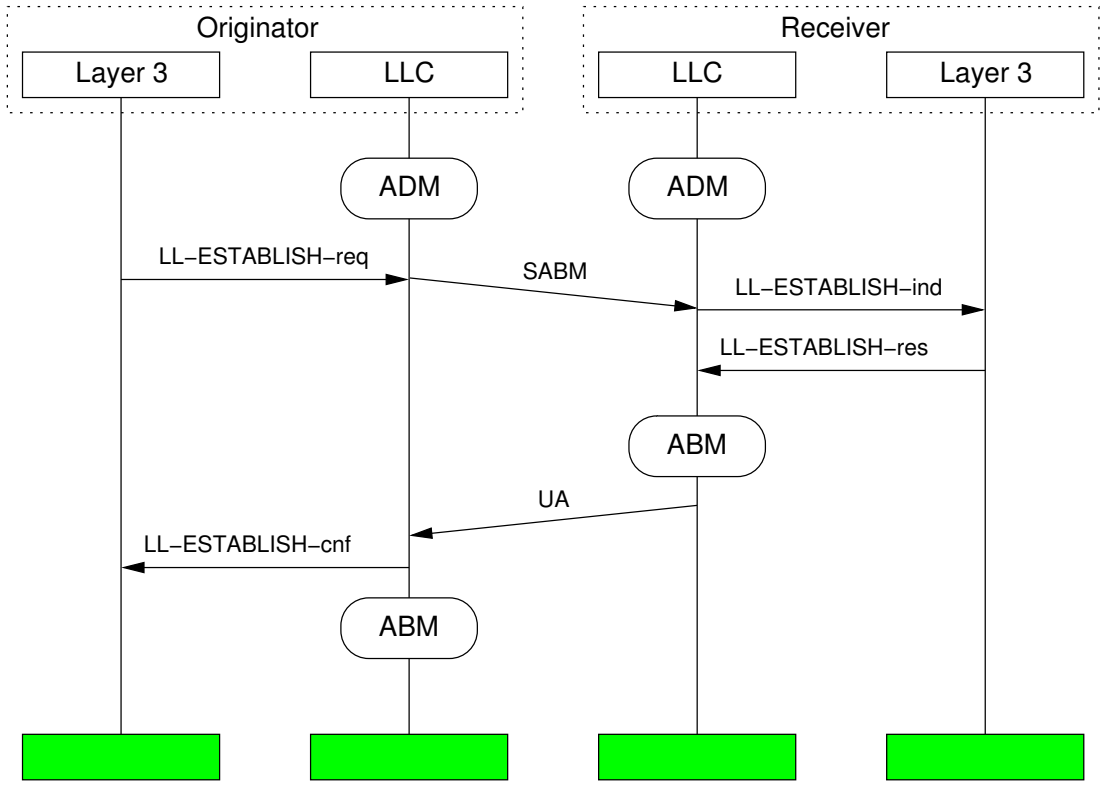



**Figure 3.11:** LLC acknowledged mode

Flow control procedures are not defined in unacknowledged mode. In acknowledged mode network layer information is transmitted within numbered I frames that are acknowledged by the peer entity. Error recovery and reordering procedures based on retransmission of unacknowledged I frames are specified. Acknowledged operation requires that the *Asynchronous Balanced Mode* (ABM) has been initiated in an establishment procedure using the *Set Asynchronous Balance Mode* (SABM) command (see Figure 3.11).

#### 3.4.4.3.2 Flow Control

The flow control between the LLC peer entities is determined by the LLC window size that ranges from 2 to 16 and a modulus of 64 frames.

To perform vertical flow control between LLC and RLC/MAC, i.e. not to overload the RLC/MAC layer with I frames, a maximum buffer size is introduced. The actual buffer size is incremented with the length of the information field of a transmitted I frame and decremented with the size of an acknowledged LLC frame. The value of the buffer size shall never exceed the maximum. If a new I frame to transmit would make the buffer size exceed the maximum, then the LLC entity shall not transmit any new I frames, but still may retransmit I frames as a result of the error recovery procedures.

### 3.4.5 Radio Link Control (RLC) and Medium Access Control (MAC)

The idea of packet-switched services in GSM is to multiplex several users onto one single physical channel in order to use its capacity in a more efficient way. Furthermore, one single MS can use more than one PDCH simultaneously to increase the data rate. The maximum number of PDCHs that can be used in parallel is determined by the multislot capabilities of the MS, which can range in uplink and downlink from multislot capability 1 (one PDCH) up to multislot capability 8 (all eight PDCHs). The *multislot class* specified in the GPRS standard is a combination of the multislot capabilities in the uplink and downlink.

The *Data Link Control* (DLC) layer at the mobile $U_m$ interface is divided in two sublayers: the RLC layer and the MAC layer, see Figure 3.6. The RLC sublayer provides a reliable logical connection between MS and BSS, while the MAC sublayer controls the access to the physical medium, the radio link.

#### 3.4.5.1 Multiplexing Principles

The radio interface consists of asymmetric and independent uplink and downlink channels (see Section 3.3.2). The downlink carries packet data from the network to multiple MSs and does not require contention arbitration. The uplink is shared among multiple MSs and requires contention control procedures.

The access to the uplink is realized with a *Slotted-Aloha* based reservation protocol. A `Packet Channel Request` sent by an MS on the *Packet Random Access Channel* (PRACH) is responded by a `Packet Uplink Assignment` message indicating the uplink resources reserved for the MS.

LLC frames are segmented into RLC data blocks. At the RLC/MAC layer, a selective ARQ protocol between the MS and the network provides retransmission of erroneous RLC data blocks. When a complete LLC frame is successfully transferred across the RLC layer, it is forwarded to the peer LLC entity.

A *Temporary Block Flow* (TBF) is introduced as a virtual connection to support the unidirectional transfer of LLC PDUs on the packet data physical channels. The TBF is the allocated radio resource on one or more PDCHs and comprises a number of RLC/MAC

blocks carrying one or more LLC PDUs. A TBF is temporary and maintained only for the duration of the data transfer, i.e., until there are no more RLC/MAC blocks to be transmitted and, in RLC acknowledged mode, all of the transmitted RLC/MAC blocks have been successfully acknowledged by the receiving entity.

A *Temporary Flow Identity* (TFI) is assigned for each TBF by the network. The MS assumes that the TFI value is unique among the concurrent TBFs in each direction. The same TFI value may be used concurrently for TBFs in opposite directions. An RLC/MAC block associated with a certain TBF comprises the TFI. The TBF is identified by the TFI together with the direction in which the RLC data block is sent and in case of an RLC/MAC control message additionally with the message type.

An *Uplink State Flag* (USF) is included in the header of each RLC/MAC block on a downlink PDCH. The USF may be used on the PDCH to allow multiplexing of radio blocks from different MSs and comprises three bits at the beginning of each radio block that is sent on the downlink. It enables the coding of *eight* different USF states which are used to multiplex the uplink traffic. On PCCCH, one USF value is used to denote PRACH (USF = FREE). The *seven* other USF values are used to reserve the uplink for different MSs.

Three medium access modes are supported:

- Dynamic Allocation (characterized by a USF)
- Extended Dynamic Allocation (supports the multislot functionality)
- Fixed Allocation (no uplink multiplexing during the TBF)

The Extended Dynamic Allocation medium access method extends the Dynamic Allocation medium access method to allow higher uplink throughput. The network allocates to the MS a subset of 1 to $N$ consecutive PDCHs, where $N$ depends on the multislot class of the MS.

In `packet idle mode` the MS monitors the relevant paging subchannels on the PCCCH. In `packet idle mode` no TBF exists and the upper layer may require the transfer of an LLC PDU, which implicitly triggers the establishment of a TBF and the transition to the `packet transfer mode`.

In `packet transfer mode` a TBF provides a physical point-to-point connection on one or more packet data physical channels for the unidirectional transfer of LLC PDUs between the network and the MS. Continuous transfer of one or more LLC PDUs is possible. Two parallel TBFs may be established in opposite directions for each MS.

#### 3.4.5.2 RLC/MAC Block Structure

An RLC/MAC block consists of a MAC Header and an RLC data block or RLC/MAC control block, respectively. In the PLL a *Block Check Sequence* (BCS) is added for error detection. The RLC/MAC block together with the BCS are building one radio block that is transmitted in one PDCH in one radio block period (see Section 3.3.1). The RLC/MAC block structure is shown in Figure 3.12.

The RLC data block consists of RLC header, an RLC data field and spare bits. Each RLC data block may be encoded using any of the available channel coding schemes CS-1, CS-2, CS-3 or CS-4. The size of the RLC data block for each of the channel coding schemes is shown in Table 3.2. Two octets are always used in the RLC header for TFI, *Final Block Indicator* (FBI) and *Block Sequence Number* (BSN). The field for the `Length Indicator` and the `Extension`-Bit needs at least one octet. One additional octet is needed for each additional LLC frame. The remaining space is used for user data.

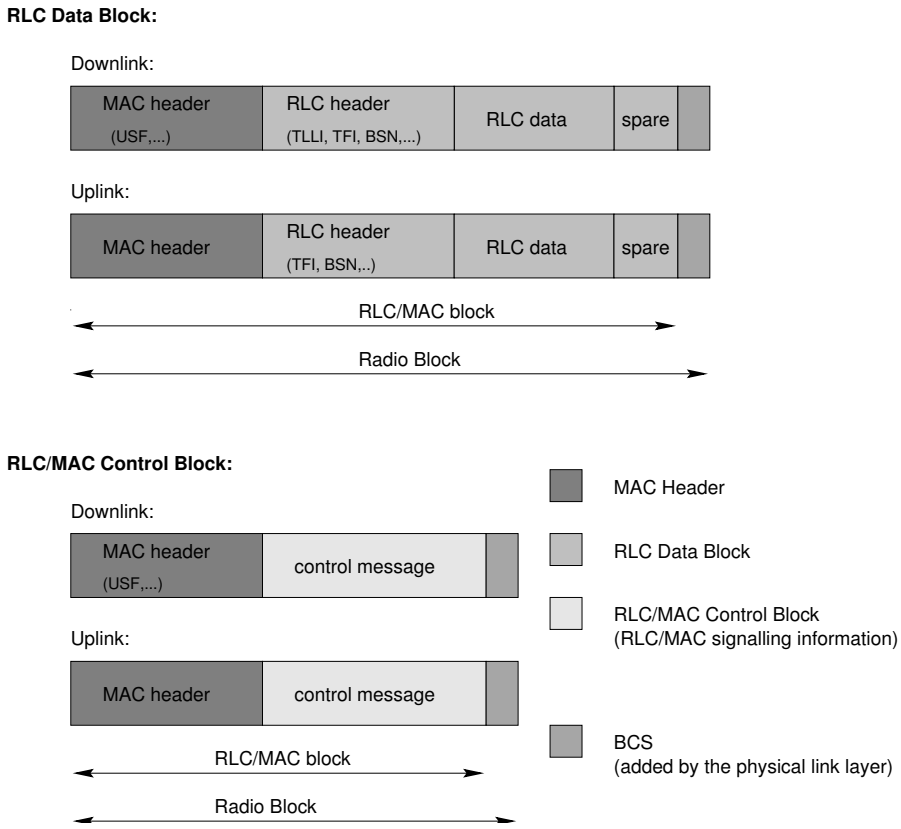



**Figure 3.12:** Structure of the radio block and RLC/MAC block

**Table 3.2:** RLC block size

| Coding Scheme | RLC info field size [byte] | RLC data block size (without spare bits) [byte (bit)] | Spare bits [bit] | RLC/MAC block size [bit] |
|---|---|---|---|---|
| CS-1 | 20 | 22 (176) | 0 | 181 |
| CS-2 | 30 | 32 (256) | 7 | 268 |
| CS-3 | 36 | 38 (304) | 3 | 312 |
| CS-4 | 50 | 52 (416) | 7 | 428 |

The RLC/MAC blocks containing an RLC/MAC control block always consists of 23 bytes, one byte for the MAC header and 22 bytes for the RLC/MAC control block containing the control message.

Each radio block carrying an RLC data block can be coded with one of the four different Channel Coding Schemes CS-1 to CS-4. For a radio block containing an RLC/MAC control block always the coding scheme CS-1 is used. The exception is messages that use the access burst, e.g., the `Packet Channel Request` message. The four different coding schemes are explained in detail in Section 3.4.5.6.

#### 3.4.5.3 RLC Functions

The RLC functions in GPRS provide an interface towards the LLC layer, especially the segmentation and reassembly of LLC PDUs into RLC data blocks depending on the used *Coding Scheme* (CS) (see Table 3.2). *Backward Error Correction* (BEC) is used to enable

the selective retransmission of uncorrectable PDUs. The BCS for error detection in GPRS is provided already by the Physical Link Layer (see Section 3.4.6).

There are two possible modes provided by the RLC/MAC layer: the acknowledged mode, used for reliable data transmission, and the unacknowledged mode, mainly used for real-time services, such as video or voice, where time delay is most critical, whereas bit errors are less important unless they pass the accepted range. For RLC acknowledged mode, a selective ARQ protocol is used between the MS and the BSS, which provides retransmission of erroneous RLC data blocks. As soon as a complete LLC frame is successfully transferred across the RLC layer, it is forwarded to the peer LLC entity.

#### 3.4.5.3.1 Data Transmission in RLC Acknowledged Mode

In this mode, the transmitting side numbers the RLC data blocks with the *Block Sequence Number* (BSN) to supervise the acknowledged transmission of the RLC blocks. With *Acknowledgement* (ACK)/*Negative Acknowledgement* (NACK) messages on uplink and downlink (`Packet Uplink/Downlink ACK/NACK`), the receiving side can request retransmissions of RLC blocks if needed. A bitmap-based selective ARQ scheme is used with a window size of 64 and a modulus of 128.

Since the radio resource management is located in the BSS, the transmission of the acknowledgement messages in both directions is controlled by the BSS. The MS sends the ACK/NACK message in the reserved radio block which is requested from the peer entity by polling. `Packet Uplink ACK/NACK` messages can be transmitted directly by the BSS on the PACCH. In case of a negative acknowledgement, only those blocks listed as erroneous are retransmitted. In the case that the TBF ends before the LLC frame is transmitted completely, the missing part has to be transmitted during the new TBF.

#### 3.4.5.3.2 Data Transmission in RLC Unacknowledged Mode

This mode comprises no retransmissions. The BSN is now only used to reassemble the RLC blocks. The `Packet ACK/NACK` messages are sent to convey the necessary control signaling, as for instance the monitoring of the channel quality. The MS transmits RLC data blocks without receiving a `Packet ACK/NACK` message until the *Window Size* (WS) is reached. Then it starts the timer `T3182` to wait for a `Packet Uplink ACK/NACK` message. On receipt of this message, the timer will be stopped. On expiry of the timer, an abnormal release with cell reselection will be performed.

### 3.4.5.4 MAC Functions

The GPRS MAC layer is responsible for providing efficient multiplexing of data and control signaling on the uplink and the downlink. The multiplexing on the downlink is controlled by the downlink scheduler, which has knowledge of the active MSs in the system and of the downlink traffic. Therefore, an efficient multiplexing on the PDCHs can easily be made. On the uplink, the multiplexing is controlled by channel reservation to individual MSs. This is done by resource requests, which are sent by the MS to the network, which then has to schedule the uplink PDCHs. Additionally, a contention resolution is needed when several MSs attempt simultaneously to access the control channel for resource requests. Collision detection and recovery are included when using the Slotted Aloha reservation protocol for recovery, which means that the MS has to wait a random backoff time before a new access attempt is done.

The MAC procedures include the provision of a TBF, which allows a point-to-point transfer of data in one cell between the BSS and the MS. As a physical connection between

two radio resource entities, the TBF is used to transport the RLC data blocks on the PDCHs. The TBF is maintained only for the duration of the data transfer and is released as soon as all data is sent (RLC send queue is cleared). A TFI is assigned to the TBF by the network to associate the MS with the current TBF.

An *Uplink State Flag* (USF) is used by the network to control the multiplexing of different mobile stations on uplink PDCHs. It is included in the header of each RLC/MAC PDU on a downlink PDCH. The USF indicates which station is allowed to transmit during the next uplink radio block period (see Section 3.4.5.1).

#### 3.4.5.4.1 TBF Setup

Both the network and the MS can initiate the establishment of a *Temporary Block Flow* (TBF) on the PCCCH allocated in the cell. The access is carried out on the PCCCH in either one or two phases, whereas two-phase access is used if the requested RLC mode is unacknowledged mode to ensure a safe establishment or if more than one time slot is requested by the MS. A one-phase access procedure is used if the amount of data to send fits into eight or fewer RLC/MAC blocks using CS-1.

##### 3.4.5.4.1.1 Uplink TBF Setup

The MS enters the packet access procedure by sending a `Packet Channel Request` message on the PRACH and entering the `packet transfer mode` (see Figure 3.13). This `Packet Channel Request` message contains parameters required to indicate the mobile station's request for radio resources and the type of access needed. Access persistence control on PRACH can be steered either by the network (network steered method) or by the mobile station (mobile station steered method) to avoid collision failure.

In one-phase access, the network responds to the `Packet Channel Request` with the `Packet Immediate Assignment` (`Packet Uplink Assignment`), reserving the resources on PDCHs for uplink transfer of a number of radio blocks.

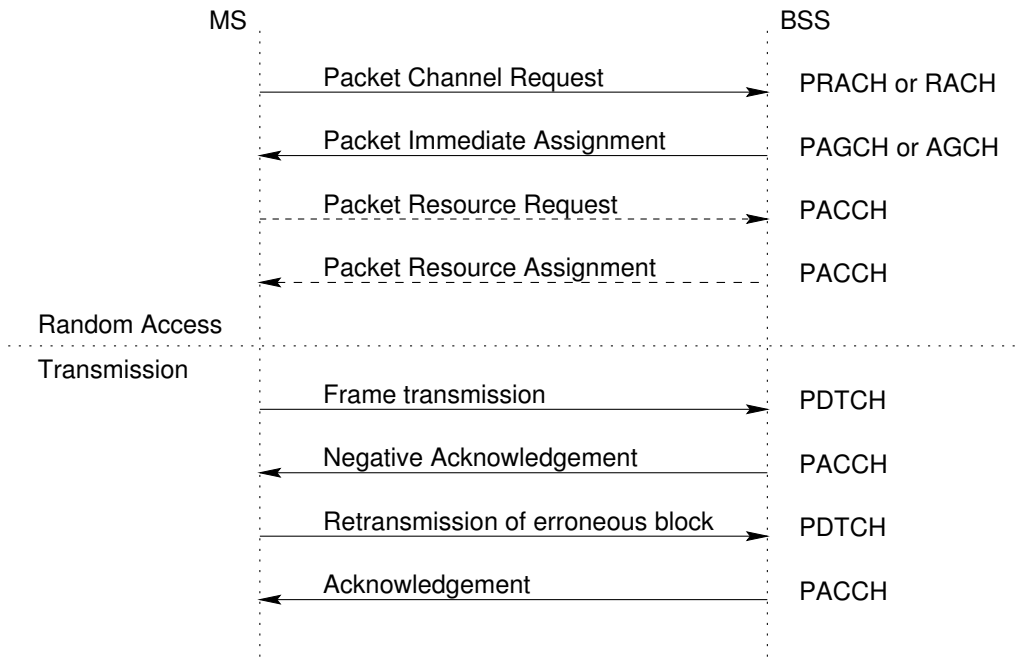



**Figure 3.13:** Uplink TBF establishment and data transmission

During two phase access, the network responds to the `Packet Channel Request` with `Packet Immediate Assignment` (`Packet Uplink Assignment`), which reserves the uplink resources for transmitting the `Packet Resource Request`. The `Packet Resource Request` message carries the complete description of the requested resources for the uplink transfer. Thereafter, the network responds with a `Packet Resource Assignment` message (`Packet Uplink Assignment`), reserving resources for the uplink transfer.

If there is no response to the `Packet Channel Request` within a predefined time period (`T3168`), the MS retries after a random back-off time. On receipt of a `Packet Channel Request`, the network sends a `Packet Uplink Assignment` message on the same PCCCH on which the network has received the `Packet Channel Request` message.

Packet data traffic is bursty in nature. Sometimes, the BSS will receive more channel requests than it can serve within a certain time limit. To avoid the repeat of `Packet Channel Requests` the sender is notified with a `Packet Queueing Notification` that its message is correctly received and will be handled later.

Efficient and flexible utilization of the available spectrum for packet data traffic with one or more PDCHs in a cell can be obtained using a multislot channel reservation scheme. Blocks from one MS can be sent on different PDCHs simultaneously, thus reducing the packet delay for transmission across the air interface. The bandwidth may be varied by allocating one to eight time slots in each TDMA frame depending on the number of available PDCHs, the multislot capabilities of the MS, and the current system load.

#### 3.4.5.4.1.2 Downlink TBF Setup

A BSS initiates a packet transfer by sending a `Packet Paging Request` on the PPCH on the downlink in order to determine the location of an MS (if it is not already known). The MS responds to the paging message by initiating a procedure for page response very similar to the packet access procedure described earlier. If the location is known, the paging procedure is followed by the `Packet Resource Assignment` for downlink frame transfer containing the list of PDCHs to be used. The last procedure is also performed before downlink transfer, if the location of the MS is already known.

#### 3.4.5.4.2 RLC Block Transfer

#### 3.4.5.4.2.1 Uplink

When the MS receives the complete uplink assignment, it begins to monitor the assigned PDCHs for the USF value. If there is already a TBF running, the MS waits for the moment of the TBF starting time, which is specified in the `Packet Uplink Assignment` message. Then the MS starts to use the new parameters. Otherwise, if there is no TBF running, the MS begins to monitor the PDCH for USFs as soon as the starting time expires.

It is possible to set the `RLC-DATA-BLOCKS-GRANTED` information element in the `Packet Uplink Assignment` message to allow the MS to send only a specified number within the TBF.

When transmitting RLC blocks on the uplink, the first three RLC data blocks of the uplink TBF contain a *Temporary Logical Link Identifier* (TLLI) field in the RLC data block header. Each time the MS detects an assigned USF value on an assigned PDCH, one RLC block will be transmitted on the same PDCH in the next block period.

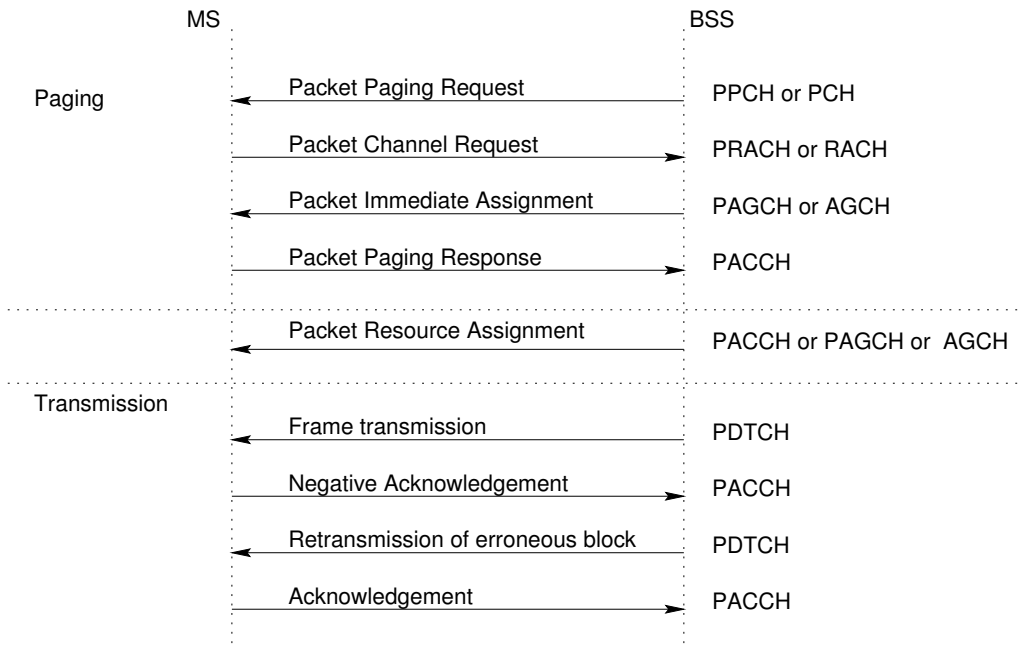



**Figure 3.14:** Downlink TBF establishment and data transmission

#### 3.4.5.4.2.2 Downlink

After the reception of the `Packet Downlink Assignment` message, the MS starts the timer `T3190` to define when to stop waiting for valid data from the network.

Receiving a valid RLC data block without the FBI bit set to 1, the MS resets and restarts timer `T3190` to define when to stop waiting for valid data from the network.

### 3.4.5.5 RLC/MAC Scheduling

RLC/MAC scheduling in the (E)GPRS BSS can be subdivided into three steps: the selection of the traffic class (see Section 2.2.2.2 and Section 3.5.3.3), scheduling of the next TBF inside the selected traffic class and scheduling of the next RLC block of the selected TBF (see Figure 3.4.5.5).

#### 3.4.5.5.1 Traffic Class Scheduling

The MAC scheduler classifies the incoming radio resource requests of established TBFs regarding the application and subscription of the MSs (see Section 3.5.3). For example, the TBF can be classified in one of three subscriber classes, *Gold* service, *Standard* service and *Best-effort* service. In case the TBF belongs to a *Standard* subscriber it might additionally be classified according to the application QoS profile to one of the four standard traffic classes, *Conversational*, *Streaming*, *Interactive*, or *Background*. Within the resulting six traffic class queues which are also called *TBF queues* only the identifiers of the TBFs are registered. The traffic class scheduler only has the information about TBFs which are requesting a data transfer and does not have information on the amount of data to be transmitted for each TBF. The traffic class scheduler selects a traffic class queue to be served by applying a class scheduling algorithm. This algorithm is not specified in the standard and can be optimized by the system designer. The algorithm can be, e.g. a priority algorithm or a bandwidth sharing algorithm.

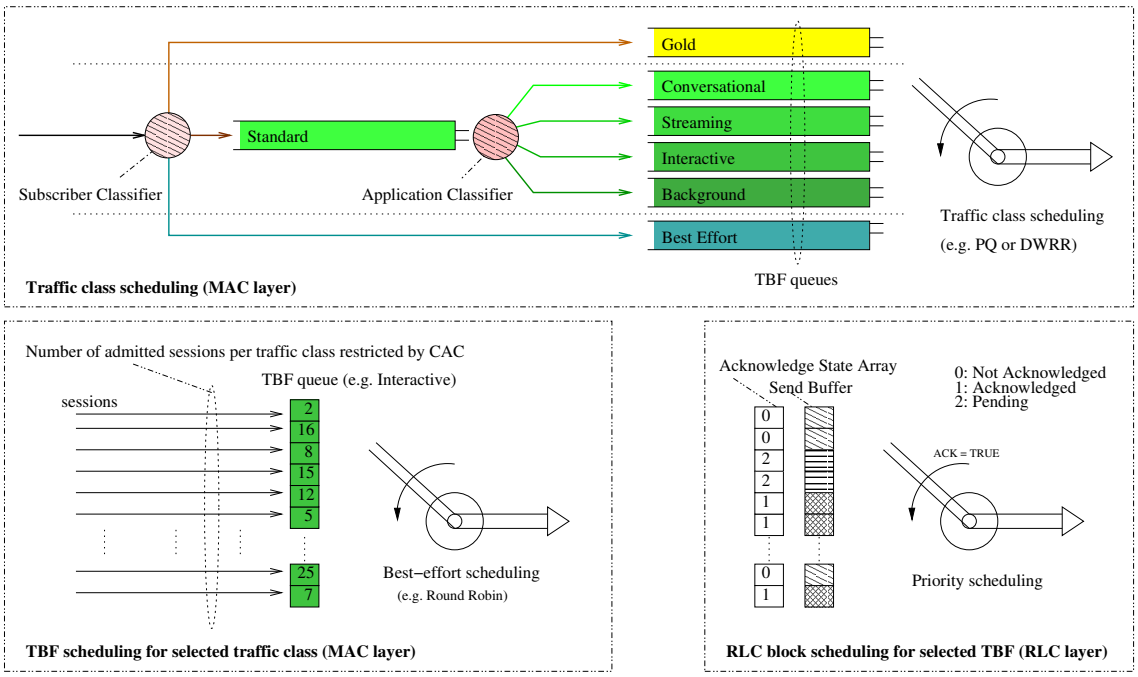



**Figure 3.15:** Scheduling of traffic classes, TBFs and RLC blocks

#### 3.4.5.5.2 TBF Scheduling

Once a traffic class queue containing all TBF identifiers of this traffic class has been selected by the traffic class scheduler, the TBF scheduler selects one TBF of this TBF queue applying the TBF scheduling algorithm. This algorithm is also implementation-specific. As an example a *Round Robin* (RR) algorithm can be applied. The TBF scheduler only has the information that a TBF is established and has neither information on the amount of data to transmit nor if the TBF actually has data available. So the scheduler starts with the first TBF listed in the queue and checks if it has been allocated to the regarded PDCH. If not, the scheduler continues with the following TBF. In case the TBF is able to use the regarded PDCH the related RLC entity is polled for data until it reaches the predefined RR quantum or until it has no more radio blocks to transmit. Then the following TBF of the same class queue is served if the same traffic class is still selected by the traffic class scheduler. Typically the RR quantum is in the order of 1–20 radio blocks.

#### 3.4.5.5.3 RLC Block Scheduler

In the third step, the RLC entity which has been polled for data by the MAC scheduler, checks if there are any data blocks available in the transmit buffer.

In case of *RLC acknowledged mode* the elements in V(B) indicate the acknowledgement status of related RLC data blocks. There are three possible states for each RLC data block:

- `NACK` indicates an RLC block which has not been transmitted yet, which has been negatively acknowledged or which has an expired timer
- `PENDING_ACK` indicates an RLC block which has been sent, but no acknowledgement has been received for this block yet
- `ACK` indicates data which has been sent and has already been acknowledged

The RLC block scheduling algorithm determines the order of transmission of the RLC blocks inside the RLC send buffer of a regarded TBF. The RLC data blocks in the RLC transmit window with the acknowledge state `NACK`, are forwarded to the MAC starting with the oldest one. If no `NACK` data block exists, the oldest RLC data block with the acknowledge state `PENDING_ACK` is retransmitted.

The priority of `NACK` blocks to `PENDING_ACK` blocks inside one RLC entity is specified in the standard [3GPP TSG GERAN (2002d)]. It is also specified that `PENDING_ACK` blocks should be transmitted if a radio block period is scheduled for the regarded TBF and if no `NACK` block exists for this TBF. A decoupled implementation of RLC and MAC leads to the transmission of `PENDING_ACK` blocks, while other TBFs still could have `NACK` blocks to transmit that are more urgent. This gives the motivation to implement an RLC/MAC layer with a MAC TBF scheduler that serves TBFs with `NACK` RLC blocks ahead of TBFs with only `PENDING_ACK` blocks, which is consistent with the GPRS standard.

#### 3.4.5.6 Channel Coding Schemes

The redundancy added to the RLC/MAC blocks can be adapted by the operator to the channel conditions in the network. The GPRS channel *Coding Schemes* (CSs) CS-1 to CS-4 enable code rates from 1/2 to 1 (see Table 3.3). The RLC data block size depends on the number of LLC PDUs concatenated in the radio block. In this table one LLC PDU per radio block is assumed. CS-1 has to be used for RLC/MAC control blocks after the GPRS specification. Although all CSs can be used for RLC/MAC data blocks, only CS-2 is used for RLC/MAC data blocks in operational GPRS networks.

In the first step of the coding procedure a BCS is added for error detection. The second step consists of precoding the USF (except for CS-1), adding four tail bits, convolutional coding for error correction and puncturing to realize the desired code rate. The result is a bit pattern with a length of 456 bit independent of the coding scheme used. For CS-4 no tail bit is added and no convolutional coding (and puncturing) is performed; only the

**Table 3.3:** Coding parameters

| | CS-1 | CS-2 | CS-3 | CS-4 |
|---|---|---|---|---|
| Code rate | 1/2 | 2/3 | 3/4 | 1 |
| GPRS net data rate [kbit/s] | 9.05 | 13.4 | 15.6 | 21.4 |
| RLC data block size [bit] | 176 | 263 | 307 | 423 |
| MAC header size (ex. USF) [bit] | 5 | 5 | 5 | 5 |
| RLC/MAC block size (ex. USF) [bit] | 181 | 268 | 312 | 428 |
| USF [bit] | 3 | 3 | 3 | 3 |
| BCS [bit] | 40 | 16 | 16 | 16 |
| Precoded USF [bit] | 3 | 6 | 6 | 12 |
| Tail bits | 4 | 4 | 4 | 0 |
| Radio block [bit] (before conv. coding) | 228 | 294 | 338 | 456 |
| Convolutional code rate | 1/2 | 1/2 | 1/2 | 1 |
| Coded radio block size [bit] | 456 | 588 | 676 | 456 |
| Puncturing [bit] | 0 | 132 | 220 | 0 |
| Radio block size [bit] | 456 | 456 | 456 | 456 |

USF is precoded [3GPP TSG GERAN (2002e)].

In GPRS a punctured convolutional code that was specified for GSM is used [Stuckmann (2002b)]. The decoding of the USF is simplified whereby for coding CS-2–4 a 12-bit long code word, which is not punctured, is generated for the USF. With CS-2 and CS-3 the USF is precoded with 6 bit before the frame is convolutionally coded, with the first 12 bit not being punctured. When CS-1 is used, the entire frame is coded and the USF has to be decoded as part of the information field. Figure 3.16 shows an example of the coding of an information frame for CS-3.

The GPRS *net data rate* $R_{\text{net}}$ is defined by the data rate of RLC/MAC blocks considering the channel coding overhead and considering the idle frames. Since one 52 multiframe comprises 12 radio block periods, the mean transmission duration of a radio block $D_{\text{RB}}$ can be calculated as:

$$D_{\text{RB}} = \frac{52\ frames \cdot 4.615\ ms}{12\ frames} = 20\ ms \tag{3.1}$$

With X as the number of bits in one RLC/MAC block from Table 3.3 (RLC/MAC block size) the net data rate can be calculated as:

$$R_{\text{net}} = \frac{X\ bit}{D_{\text{RB}}} \tag{3.2}$$

For example:

$$R_{\text{net(CS-1)}} = \frac{181\ bit}{20\ ms} = 9.05\ \frac{kbit}{s} \tag{3.3}$$

The net data rate in GPRS does not equal the data rate that the RLC/ MAC layer offers to the LLC layer. The latter can be calculated based on the number of RLC data block information bytes from Table 3.2.

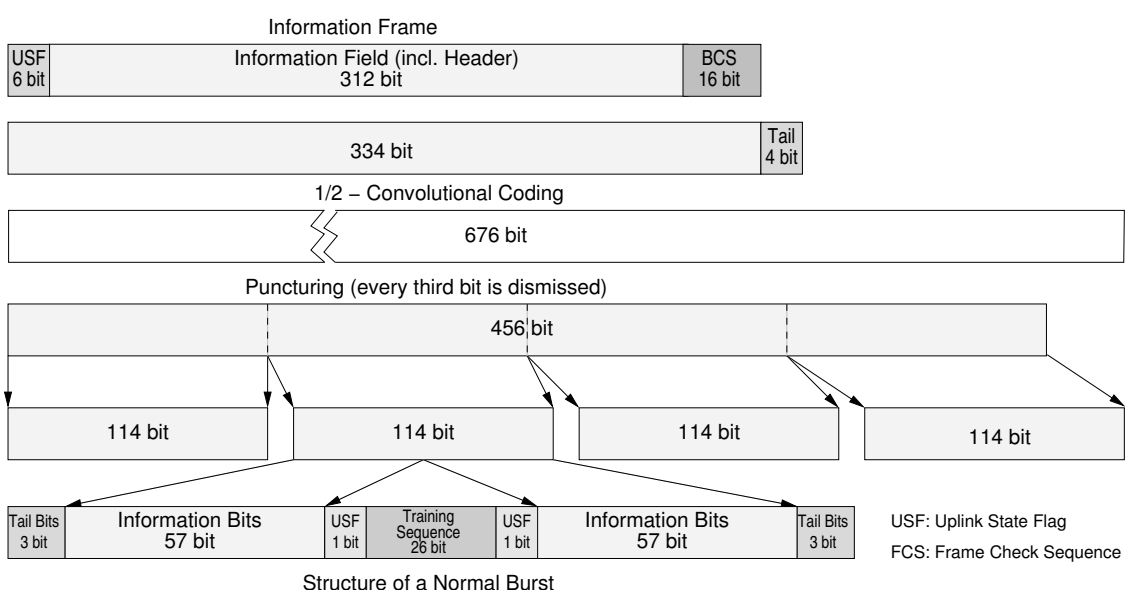



**Figure 3.16:** Coding of an information frame according to CS-3

### 3.4.6 Physical Layer (PL)

The *Physical Layer* (PL) at the air interface is divided into the *Physical Link Layer* (PLL) and the *Radio Frequency* (RF) layer. These sublayers are specified in [3GPP TSG GERAN (2002a,b)]. While in the RF sublayer mainly modulation and demodulation are carried out, the PLL sublayer provides services for data transmission over the radio interface. The PLL sublayer is responsible for the *Forward Error Correction* (FEC) coding, allowing the detection and correction of erroneous transmitted code words and the indication of uncorrectable code words. It is also responsible for the interleaving of one radio block over four bursts in consecutive TDMA frames and provides procedures for synchronization.

## 3.5 GPRS Control Plane

The GPRS protocol architecture is organized into two planes, the user plane, also called transmission plane, and the control plane, also called signaling plane. The user plane is responsible for data transmission, when packet data transfer is actually requested in uplink or downlink. To realize the transfer between the correct network nodes with adequate performance characteristics, the user plane protocols need certain information such as addresses of peer entities, the status of network elements or requested protocol options. This information has to be provided at the beginning of a session and has to be kept up-to-date during an ongoing session, while the mobile user might move around with his terminal or use different services in parallel or consecutively with idle periods in between.

The GPRS control plane realizes these management functions, namely *GPRS Mobility Management* (GMM), *Session Management* (SM) and *Quality of Service* (QoS) management. GMM keeps track of the MS's location and its state, updates databases with this information and supports cell change procedures. The SM manages the logical context between the MS and an external *Packet Data Network* (PDN) based on TCP/IP or X.25, when a new session is set up or changing its performance requirements. In GPRS networks different applications with different performance requirements are supported. QoS management in GPRS networks enables the network to establish QoS contracts with the MS differentiating applications and subscribers according to their QoS requirements. The GPRS QoS support is based on the SM procedures.

In Figure 3.17 the GPRS control plane between MS and SGSN is shown. The signaling layers SM and GMM reuse the layer-2 protocols and the GPRS channel interface of the user plane (see Section 3.4) to transmit signaling messages over the air interface.

### 3.5.1 Mobility Management

Mobility management in GSM and GPRS [3GPP TSG SSA (2002a)] includes functions for location registration with the PLMN and location updating to report the current location of an MS. Additionally it is responsible for identification and authentication of subscribers.

The GMM concept allows combined procedures for GSM and GPRS to reduce the signaling load. The handover procedure is split up into the cell update and the *Routing Area* (RA) update.

In contrast to the GSM system, where the handover decision is made by the network based on measurement reports sent by the MS [3GPP TSG CN (2001)], the cell change in GPRS is not necessarily prepared and triggered by the network, but is mostly decided

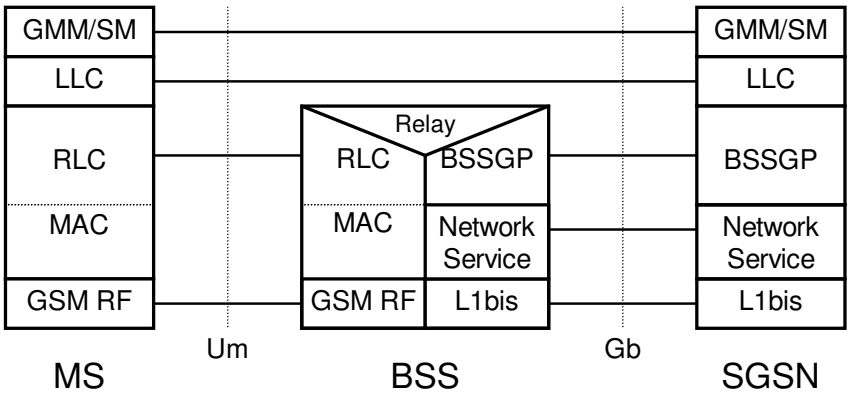



**Figure 3.17:** GPRS control plane (MS - SGSN)

by the MS, when it detects that it has entered a new cell or even a new RA by reading the system information on the broadcast channel.

#### 3.5.1.1 Location Area and Routing Area

The *Location Area* (LA) and *Routing Area* (RA) are different areas comprising a certain number of cells (see Figure 3.18). For example one LA may comprise 30 cells. This hierarchical structure with LAs is chosen for GSM networks to manage the location of an MS. The network initially only knows the LA in which the MS is located. The MS provides this information periodically, or when the MS changes its LA. When it becomes necessary to determine the exact cell location of an MS, the network sends a paging command in all cells of the LA. The MS, which continuously monitors the broadcast channel, responds to this paging command so that a connection or a context can be established. With this procedure the signaling load on the air interface is minimized, since the network need not know about the exact cell location of the MS when it is not active. Additionally, the VLR does not have to be updated by the SGSN, as long as the MS is staying inside one LA.

The RA that does not exist in GSM networks without GPRS support was introduced to speed up the paging procedure and lower the signaling load for paging. The number of cells in an RA is generally less than or equal to that of an LA. For example, one LA with 30 cells may comprise three RAs with 10 cells each. An RA can not span more than one LA and an RA is served by only one SGSN. One SGSN can handle several RAs and the size of an RA can range from a part of a city to an entire province or even a small country. Figure 3.18 shows an example cell scenario, in which it is possible to see how cells could be grouped in a cell planning configuration. Cells 1, 2 and 3 together build one RA. Cell 4 alone builds another distinct RA. Together these two RAs form one LA. A set of LAs are served by one BSC, and many BSCs are served by an SGSN. The SGSNs are interconnected and represent the highest level in the GMM hierarchy. Cell 14 belongs to another SGSN in this example.

After an MS has changed from one cell to another it reads the system information of the new cell. Since the RA is broadcast as part of the system information, the MS can determine whether it has changed its RA. If this is the case, the MS sends an RA update to the SGSN to inform it about the new RA.

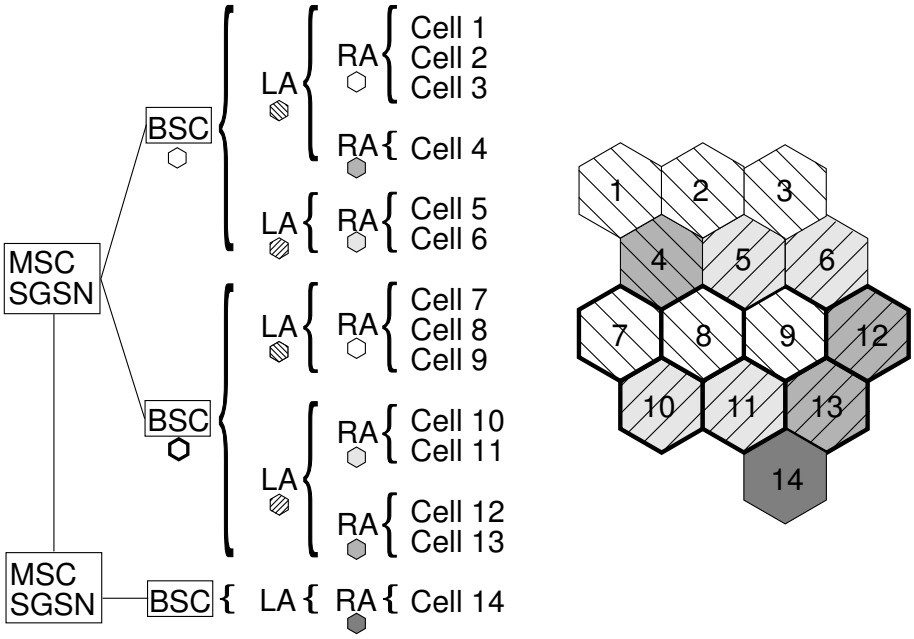



**Figure 3.18:** Example of a cell scenario comprising all cell change possibilities

#### 3.5.1.2 Mobility Management Procedures

The *GPRS Mobility Management* (GMM) procedures comprise access control and authentication functions. To obtain access to GPRS services the MS has to initiate the GPRS attach procedure. During *Mobility Management* (MM) procedures user data can, in general, be transmitted while the signaling is going on. This can lead to loss of data during attach, authentication and RA update procedures. In order to minimize retransmissions the MS and SGSN should not transmit user data during these procedures.

#### 3.5.1.3 States of the Mobility Management

The GMM functionalities are based on three different MM states defined at the MS and the SGSN states, in which each MS can be. These MM states are:

1. The `idle` state.
2. The `standby` state.
3. The `ready` state.

In each of these MM states the MS and the SGSN hold a different set of information about the GPRS terminal. These are denoted MM contexts. Figure 3.19 shows the different states and the related transitions. The MM context in the MS comprises among others the MM state MS identifiers, location and security data and zero or more PDP contexts, while the MM context at the SGSN additionally contains classmarks for the communication with the registers. The following states are defined.

**In the idle state** the MM context is empty, because the subscriber is not attached to the GPRS network. The MS and SGSN do not hold any valid information about the cell and RA in which the MS is located. Therefore no MM procedures are performed. The MS is seen as not reachable and can not be paged.

**In the standby state** the MS is attached to the network. Each of the MS and SGSN have established an MM context. Both the MS and the SGSN know the RA in which the

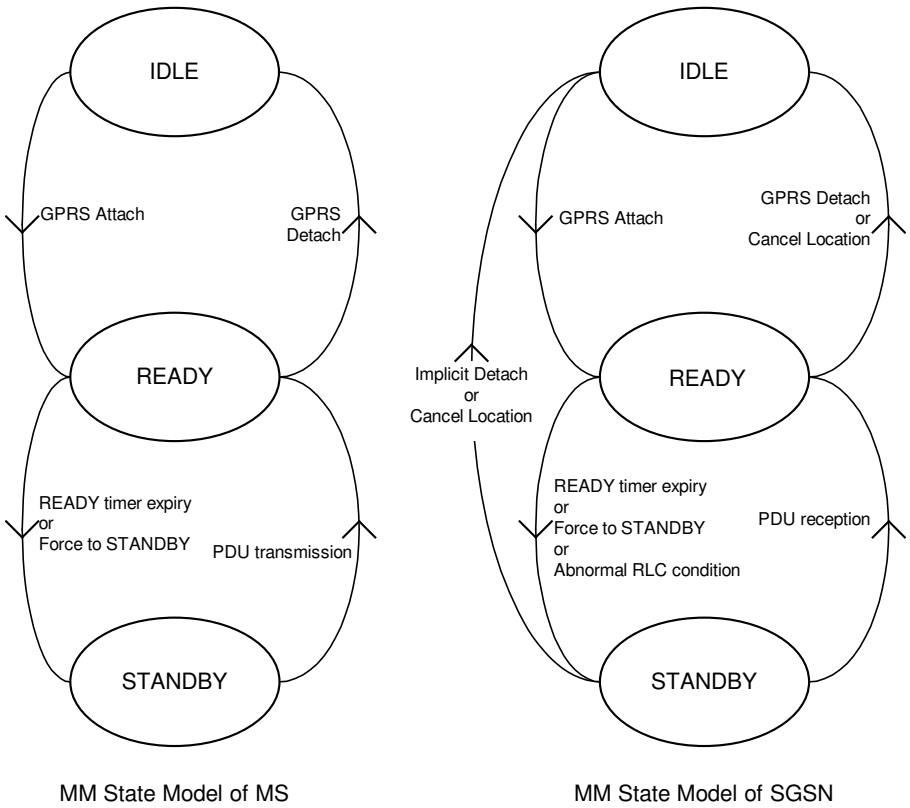



**Figure 3.19:** MM state transitions

MS is located. In this state the SGSN can send paging messages to all the cells of the RA to find the MS, when it messages have to be sent in the downlink direction. The MS may receive paging messages, but can not receive or transmit any data.

**In the ready state** the SGSN knows in which cell the MS is located and can therefore send a continuous stream of data packets on the downlink, even if the MS changes its cell supported by *Location Management* (LM) procedures (see Section 3.5.1.4). The MM context is practically the same as in the case of the `standby` state but extended by the information on the cell in which the MS is located. In this state the SGSN need not initially page the MS to send data. The MS can activate or deactivate PDP contexts. The MM context remains in the `ready` state even if there is no data to send. It returns to the `standby` state after the expiry of the `ready` timer.

#### 3.5.1.4 Location Management Procedures

The *Location Management* (LM) function controls the cell and PLMN selection and provides a mechanism which enables the SGSN to know the RA of the MS in the `standby` and `ready` states [3GPP TSG SSA (2002a)]. If an MS is in the `ready` state, the SGSN knows the location of the MS on cell level.

When an MS enters a new cell, three different procedures can be performed:

1. A cell update procedure
2. An RA update procedure (intra- or inter-SGSN)

In Figure 3.20 all the cell change possibilities are shown. These can be derived from Figure 3.18, in which a cell planning scenario is shown. The first possibility is the intra-

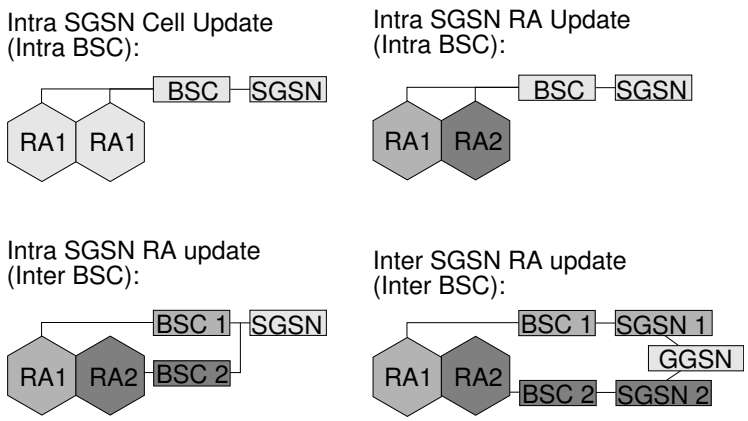



**Figure 3.20:** All cell change possibilities in GPRS

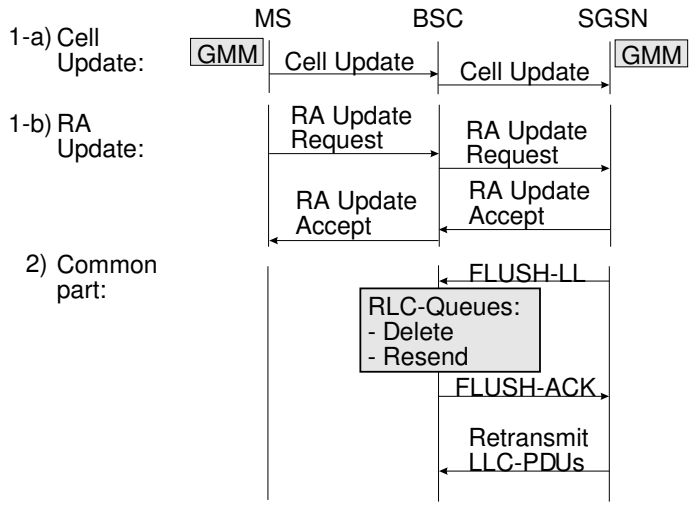



**Figure 3.21:** Message sequence chart of the cell change procedure

SGSN cell update intra-BSC where an MS changes between two cells belonging to the same RA served by one BSC. In this case, the MS performs a cell update procedure. The second possibility is the intra-SGSN RA update intra-BSC. Here the RA changes from one cell to the other. The MS has to perform an RA update procedure. In the third case the responsible BSC changes (intra-SGSN RA update inter-BSC). The fourth and last case is the inter-SGSN RA update where also the responsible SGSN changes.

#### 3.5.1.5 Cell Change Signaling

Figure 3.21 shows a simplified overview of the cell change procedures in GPRS. Basically, two cell change possibilities are foreseen. The first one (1-a) is the cell update, where the MS notifies the SGSN GMM about a cell change by sending an uplink PDU in the new cell it enters. The second one (1-b) is the RA update, where the MS sends an `RA Update Request` message in the new cell it enters and must wait for a response from the GMM. The response shown in Figure 3.21 is an `RA Update Accept` message, but it also might be an `RA Update Reject` message.

Then, there is a common part (2) for both the cell update and the RA update where the GMM decides with the `FLUSH-LL` message to delete or to forward RLC PDUs queued in the

old BSC [3GPP TSG GERAN (2002c)]. This message is acknowledged with the message `FLUSH-ACK`. After this response the SGSN can retransmit the LLC-PDUs immediately, if they have been deleted at the BSC.

### 3.5.2 Session Management

*Session Management* (SM) in GPRS comprises all signaling functions necessary for the access to external *Packet Data Networks* (PDNs). For interworking with external IP networks the IP interworking model is defined in [3GPP TSG CN (2002d)].

#### 3.5.2.1 The IP Interworking Model

Network interworking is required whenever a PLMN is involved in communications with another network to provide end-to-end transport. The PLMN interconnects in a manner consistent with that of a *Packet Data Network* (PDN) defined by the requirements for interworking with *Public Switched Data Networks* (PSDNs) X.75 [ITU-T (1996a)].

GPRS supports interworking with networks based on the Internet Protocol. These interworked networks may be either intranets or the Internet [3GPP TSG CN (2002d)]. When interworking with IP networks, GPRS can operate IPv4 or IPv6. The interworking point with IP networks is at the $G_i$ reference point as shown in Figure 3.22. The GGSN is the access point of the GPRS network. In this case the GPRS network will look like any other IP network or subnetwork.

The $G_i$ reference point is between the GGSN and the external IP network. From the external IP network's point of view, the GGSN is seen as an ordinary IP router. The access to the Internet, an intranet or *Internet Service Provider* (ISP) may involve specific functions such as user authentication, user's authorization, end-to-end encryption between MS and Intranet/ISP and the allocation of a dynamic address belonging to the PLMN/Intranet/ISP addressing space. For this purpose the GPRS PLMN may offer:

- either direct transparent access to the Internet
- or a non-transparent access to an intranet/ISP.

In the transparent case the mobile station is given an address belonging to the operator's addressing space (see Figure 3.22). The address is a public IP address given either at subscription in which case it is a static address or at PDP context activation in which case

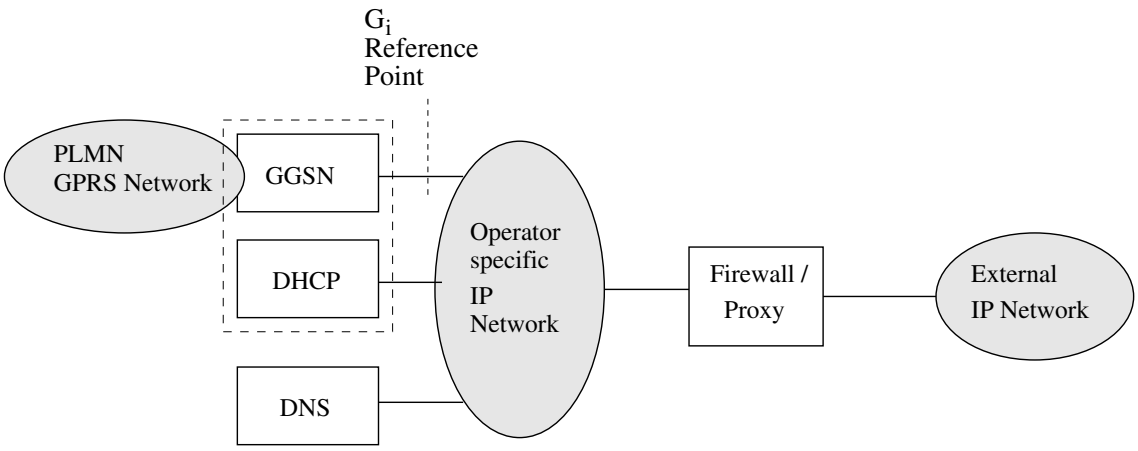



**Figure 3.22:** IP interworking

it is a dynamic address. This address is used for packet forwarding between the Internet and the GGSN and within the GGSN to tunnel user data through the core network.

The transparent case provides at least a basic ISP service. As a consequence of this it may therefore provide a bearer service for a tunnel to a private Intranet. In the non-transparent case the mobile station is given an address belonging to the Intranet/ISP private addressing space, static or dynamic. Packet forwarding within the GGSN and on the Intranet/ISP requires a link between the GGSN and an address allocation server, like *Remote Authentication Dial In User Service* (RADIUS) and *Dynamic Host Configuration Protocol* (DHCP) belonging to the Intranet/ISP. This server responses to authentication requests at a PDP context activation.

#### 3.5.2.2 PDP Context Handling

The main function of *Session Management* (SM) [3GPP TSG CN (2002e)] is to support PDP context handling of the MS.

QoS profile negotiation between MS and the network is performed within the procedures for the activation or modification of PDP contexts. Therefore, SM procedures play an important role in the GPRS mechanisms for QoS support.

#### 3.5.2.3 Session Management Procedures

The purpose of the PDP context activation procedure is to establish a PDP context between the MS and the network for a specific QoS on a specific N-SAPI (see Section 3.4.3). The PDP context activation may be initiated by the MS or the initiation may be requested by the network.

##### 3.5.2.3.1 PDP Context and Traffic Flow Template

In GPRS Release 97/98, there is only one PDP context possible per MS. In GPRS Release 99, each PDP address may be described by one or more PDP contexts in the MS or the network. The first PDP context activated for a PDP address is called the *primary PDP context*, whereas all additional contexts associated with the same PDP address are called *secondary PDP contexts*. When more than one PDP context is associated with a PDP address, there will be a *Traffic Flow Template* (TFT) for each additional context. The TFT will be sent transparently via the SGSN to the GGSN to enable packet classification and policing for downlink data transfer [3GPP TSG SSA (2002a)]. A TFT consists of from one up to eight packet filters, each identified by a unique packet filter identifier. A packet filter has an evaluation precedence index that is unique within all TFTs associated with the PDP contexts that share the same PDP address. The MS manages packet filter identifiers and their evaluation precedence indexes, and creates the packet filter contents. The packet filter specifies attributes for incoming and outgoing IP packets. The IP datagram attributes have to match at least for the packet filter of one evaluation precedence index of the TFT to pass the GGSN.

##### 3.5.2.3.2 PDP Context Information Element

All PDP context information is stored within an appropriate *Information Element* (IE) [3GPP TSG CN (2002c)]. In Figure 3.23 the fields corresponding to QoS Subscribed (QoS Sub), QoS Requested (QoS Req), and QoS Negotiated (QoS Neg), as well as the PDP Address, are emphasized. They play an important role in the implementation of the mechanisms to support QoS.

| Octet \ Bit | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|---|---|
| 1 | Type = 130 (Decimal) | | | | | | | |
| 2–3 | Length | | | | | | | |
| 4 | Res–erved | AA | Res–erved | Order | NSAPI | | | |
| 5 | X | X | X | X | SAPI | | | |
| 6–8 | QoS Sub | | | | | | | |
| 9–11 | QoS Req | | | | | | | |
| 12–14 | QoS Neg | | | | | | | |
| 15–16 | Sequence Number Down (SND) | | | | | | | |
| 17–18 | Sequence Number Up (SNU) | | | | | | | |
| 19 | Send N–PDU Number | | | | | | | |
| 20 | Receive N–PDU Number | | | | | | | |
| 21–22 | Uplink Flow Label Signalling | | | | | | | |
| 23 | Spare 1 1 1 1 | | | | PDP Type Organization | | | |
| 24 | PDP Type Number | | | | | | | |
| 25 | PDP Address Length | | | | | | | |
| 26–m | PDP Address [1..63] | | | | | | | |
| m+1 | GGSN Address for signalling Length | | | | | | | |
| (m+2)–n | GGSN Address for signalling [4..16] | | | | | | | |
| n+1 | APN length | | | | | | | |
| (n+2)–o | APN | | | | | | | |
| o+1 | Spare (sent as 0 0 0 0) | | | | Transaction Identifier | | | |

**Figure 3.23:** PDP context information element

All QoS-related information, whether subscribed, requested, or negotiated, is included in a QoS profile. The associated IE is described in Section 3.5.3.2.

### 3.5.3 Quality of Service Management

As mentioned in Section 3.4.2, Section 3.4.5.5 and Section 3.5.2, there are several logical entities and protocols involved in GPRS QoS management.

PDP contexts as well as QoS profiles are negotiated between MS and SGSN. The BSS is provided with a *Packet Flow Context* (PFC) containing the *Aggregate BSS QoS Profile* (ABQP) (see Section 3.4.2) and is responsible for resource allocation on a TBF base and scheduling of packet data traffic with respect to the relevant QoS profiles negotiated. Moreover, it has to regularly inform the SGSN about the current load conditions in the radio cell. The tasks of the GGSN comprise mapping PDP addresses as well as classification of incoming traffic from external networks on behalf of downlink TFTs. The *GPRS Registers* (GRs) hold the QoS-related subscriber information and deliver it, on demand, to the *GPRS Support Nodes* (GSNs).

In Figure 3.24, part of a GPRS session is schematically outlined, depicting the instances involved, messages exchanged, and parameters negotiated for PDP context, PFC and TFT

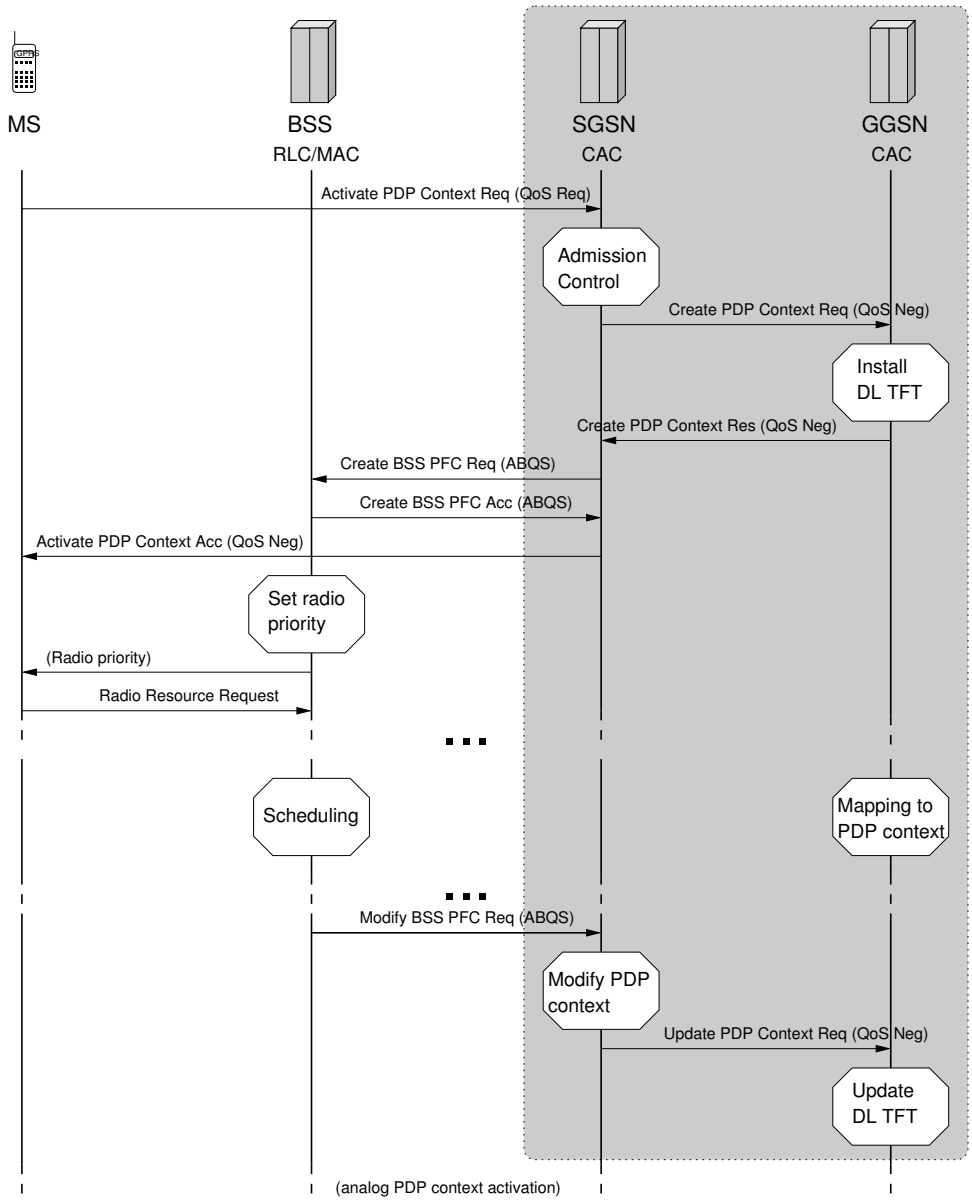



**Figure 3.24:** QoS negotiation and renegotiation procedures (example)

installation and renegotiation.

From a time-scale point of view, the mechanisms for QoS management within the GPRS might be regarded as a three-stage model (see Figure 3.25). On PDP context activation, the QoS parameters are negotiated. As long as the PDP context remains active, these parameters should be guaranteed, unless there is a QoS renegotiation. The QoS profile is considered both for each TBF and for each radio block period. At TBF setup, radio resources, like the number of PDCHs, the associated USFs and TFIs, are assigned according to the negotiated QoS parameters. During the TBF, radio blocks are scheduled at the BSS in competition with other existing TBFs in the radio cell. This scheduling function has to be done considering the QoS profiles of the PDP contexts associated with the TBFs [Baig et al. (2001); Stuckmann (2002a); Stuckmann and Müller (2001)].

A QoS profile can be considered as a single parameter value that is defined by a unique combination of attributes. There are numerous QoS profiles available based on

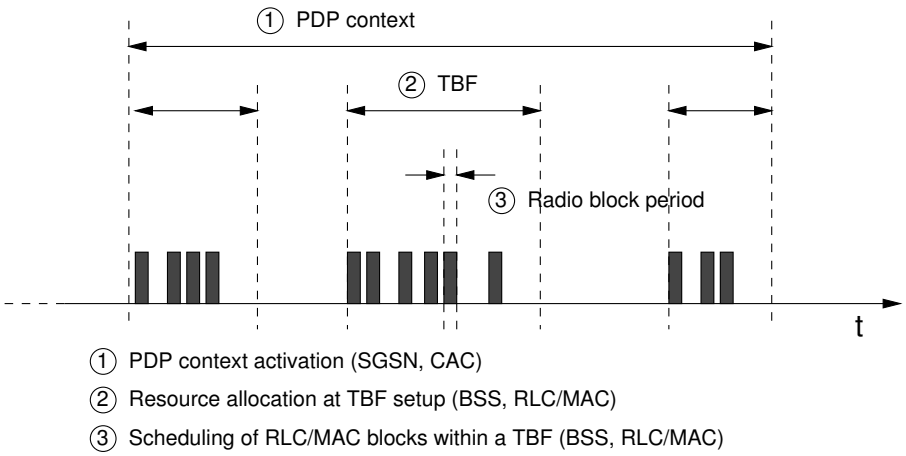



**Figure 3.25:** Three-stage model of QoS management

permutations of these different attributes, but each mobile network operator must choose to support only a limited subset, reflecting their planned range of GPRS subscriptions. In the following subsections, the QoS attributes defined in GPRS Release 97/98, as well as the changes made for Release 99, will be explained.

#### 3.5.3.1 QoS Attributes According to GPRS Release 97/98

A QoS profile defines the QoS within the range of the following service classes [3GPP TSG SSA (2001a, 2002a)].

##### 3.5.3.1.1 Precedence Classes

Under normal circumstances the network should try to meet all profiles' QoS agreements. The precedence specifies the relative importance to keep the conditions even under critical circumstances, e.g., momentarily high network load. The various precedence classes are presented in Table 3.4.

##### 3.5.3.1.2 Delay Classes

The packet delay is defined by the time needed for transmission from one GPRS SAP to another. Delays outside the system, e.g., in transit networks, are not considered. [3GPP TSG SSA (2001a)] determines four delay classes (see Table 3.5). The network operator has to provide for convenient resources on the air interface to be able to serve the number of participants with a certain delay class expected within each cell. Although there is no need for all delay classes to be available, at least *best effort* has to be offered.

**Table 3.4:** Precedence classes

| Precedence class | Identifier | To be served |
|---|---|---|
| 1 | High priority | preferably before classes 2 and 3 |
| 2 | Normal priority | preferably before class 3 |
| 3 | Low priority | without preference |

**Table 3.5:** Delay classes

| Delay class | 128 byte packet | | 1 024 byte packet | |
|---|---|---|---|---|
| | Mean delay [s] | 95 % [s] | Mean delay [s] | 95 % [s] |
| 1 (predictive) | 0.5 | 1.5 | 2 | 7 |
| 2 (predictive) | 5 | 25 | 15 | 75 |
| 3 (predictive) | 50 | 250 | 75 | 375 |
| 4 (best effort) | unspecified | | | |

**Table 3.6:** Reliability class

| Reliability classes | GTP mode | LLC frame mode | LLC data mode | RLC block mode | Traffic type security |
|---|---|---|---|---|---|
| 1 | ACK | ACK | PR | ACK | NRT traffic, error sensitive, loss sensitive |
| 2 | UACK | ACK | PR | ACK | NRT traffic, error sensitive, slightly loss sensitive |
| 3 | UACK | UACK | PR | ACK | NRT traffic, error sensitive, not loss sensitive |
| 4 | UACK | UACK | PR | UACK | RT traffic, error sensitive, not loss sensitive |
| 5 | UACK | UACK | UPR | UACK | RT traffic not error sensitive, not loss sensitive |
| (U)ACK | (Un)acknowledged | | | NRT | Non-Realtime |
| PR/UPR | Protected/Unprotected | | | RT | Realtime |

#### 3.5.3.1.3 Reliability Classes

Data services generally require a low residual *Bit Error Ratio* (BER). Erroneous data is usually useless, while incorrectly received speech only leads to a worse perception. Reliability of data transmission is defined within the scope of the following cases:

- probability of loss of data
- probability of out-of-sequence data delivery
- probability of multiple delivery of data and
- probability of erroneous data.

The reliability classes specify the requirements for each layer's services. The combination of different modes of operation of the GPRS specific protocols GTP, LLC, and RLC, explained in Section 3.4, support the reliability requirements of various applications, e.g., *Real-Time* (RT) or *Non Real-Time* (NRT). The reliability classes are summarized in Table 3.6.

#### 3.5.3.1.4 Peak Throughput Classes

User data throughput is specified within the scope of a set of throughput classes that characterize the expected bandwidth for a requested PDP context. It is defined by the choice of peak and mean throughput class. Peak throughput is measured in $^{\text{byte}}/_{\text{s}}$ at the reference point $G_i$ and in the terminal (see Figure 3.1). Peak throughput specifies the

**Table 3.7:** Peak throughput classes

| Peak throughput class | Peak throughput | |
|---|---|---|
| | [byte/s] | [kbit/s] |
| 1 | up to 1 000 | 8 |
| 2 | up to 2 000 | 16 |
| 3 | up to 4 000 | 32 |
| 4 | up to 8 000 | 64 |
| 5 | up to 16 000 | 128 |
| 6 | up to 32 000 | 256 |
| 7 | up to 64 000 | 512 |
| 8 | up to 128 000 | 1 024 |
| 9 | up to 256 000 | 2 048 |

**Table 3.8:** Mean throughput classes

| Mean throughput class | Mean throughput | |
|---|---|---|
| | [byte/h] | ≈ [bit/s] |
| 1 | 100 | 0.22 |
| 2 | 200 | 0.44 |
| 3 | 500 | 1.11 |
| 4 | 1 000 | 2.2 |
| 5 | 2 000 | 4.4 |
| 6 | 5 000 | 11.1 |
| 7 | 10 000 | 22 |
| 8 | 20 000 | 44 |
| 9 | 50 000 | 111 |
| 10 | 100 000 | 220 |
| 11 | 200 000 | 440 |
| 12 | 500 000 | 1 110 |
| 13 | 1 000 000 | 2 200 |
| 14 | 2 000 000 | 4 400 |
| 15 | 5 000 000 | 11 100 |
| 16 | 10 000 000 | 22 000 |
| 17 | 20 000 000 | 44 000 |
| 18 | 50 000 000 | 111 000 |
| 31 | Best effort | |

maximum rate at which data is transmitted within a certain PDP context. No guarantee is given that this data rate is actually achieved at any time during transmission. Rather, this depends on the resources available and the capabilities of the MS. The operator may limit the user data rate to the peak data rate agreed on, even if there is capacity left for disposal. The peak throughput classes are presented in Table 3.7.

#### 3.5.3.1.5 Mean Throughput Classes

Like peak throughput, mean throughput is also measured in byte/s at the reference points $G_i$ and R. It specifies the average rate data transmitted within the time remaining for a certain PDP context. The operator may limit the user data rate to the mean data rate negotiated, even if excessive capacity is available. If *best effort* has been agreed on as the throughput class, throughput is made available to a MS whenever there are resources

needed and at its disposal. Table 3.8 summarizes the classes of mean throughput.

#### 3.5.3.2 QoS Profile Information Element

All QoS-related information to be exchanged between MS and SGSN is stored in a QoS profile. The appropriate IE is shown in Figure 3.26. It consists of an *Information Element Identifier* (IEI), a length field, five fields that contain the values of the service classes (see Section 3.5.3.1), and three fields filled with spare bits [3GPP TSG CN (2002e)]. IEI and length fields are not part of a PDP context IE (see Section 3.5.2.3.2).

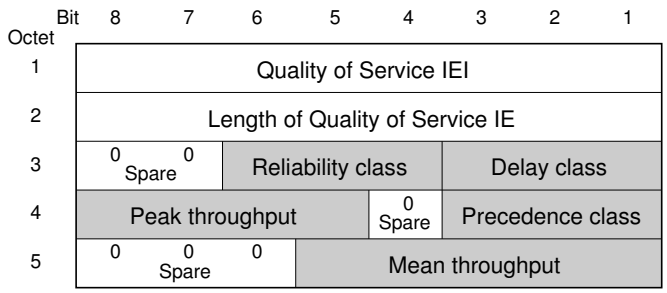



**Figure 3.26:** QoS profile information element

#### 3.5.3.3 QoS in GPRS Release 99

It is evident from the above description that the QoS architecture defined in GPRS Release 97/98 shows some major drawbacks (see also [Gudding (2000); Stuckmann and Müller (2001)]):

1. The BSS is not aware of the negotiated QoS profile. This restricts the ability of the BSS to perform scheduling and resource management on the radio interface.
2. Neither MS nor GGSN can influence the QoS profile, even if they detect congestion in external networks as well as changing radio conditions or varying application requirements.
3. It is only possible to have one QoS profile for every PDP context that is associated with a specific service access. Thus, there can only be one QoS profile utilized by all applications for one PDP address.

GPRS Release 99 defines several further QoS parameters with finer grained properties to meet requirements on different levels of service for applications [3GPP TSG SSA (2002b)]:

- maximum bitrate
- delivery order
- SDU format information
- residual bit error ratio
- transfer delay
- allocation/retention priority
- guaranteed bitrate
- maximum SDU size
- SDU error ratio
- delivery of erroneous SDUs
- traffic handling priority
- source statistics descriptor (‘speech’/‘unknown’)

Additionally, there are four distinct *traffic classes* introduced, with different parameters specifying their QoS requirements (see Table 3.9):

**Table 3.9:** End-user performance expectations for selected services belonging to different traffic classes [3GPP TSG SSA (2001b)]

| Traffic class | Medium | Application | Data rate (kbit/s) | One-way delay |
|---|---|---|---|---|
| Conversational | Audio | Telephony | 4–25 | $< 150\,\mathrm{ms}$ |
| | Data | Telnet | $< 8$ | $< 250\,\mathrm{ms}$ |
| Streaming | Audio | Streaming (HQ) | 32–128 | $< 10\,\mathrm{s}$ |
| | Video | One-way | 32–384 | $< 10\,\mathrm{s}$ |
| | Data | FTP | — | $< 10\,\mathrm{s}$ |
| Interactive | Audio | Voice messaging | 4–13 | $< 1\,\mathrm{s}$ |
| | Data | Web-browsing | — | $< 4\,{}^{\mathrm{s}}/_{\mathrm{page}}$ |

- *conversational*
- *interactive*
- *streaming*
- *background*

Delay-sensitive services belonging to the *conversational* class, for example, do need absolute guarantees in terms of *guaranteed bitrate* and *transfer delay* attributes, while for *background* traffic only bit integrity is necessary.

In GPRS Release 97/98, the BSS cannot use QoS profile information to schedule resources on a continuous data flow, neither on the downlink, since there is no mechanism provided to download the QoS profile from the SGSN, nor on the uplink, because there is no QoS information available from the MS. With the introduction of Release 99, the BSS is not only provided with QoS profiles on a PFC base, but also with the ability to modify the QoS profile associated with a data flow in case of changing load conditions. Likewise, MS and GGSN may initiate QoS profile renegotiation, either because of changing application requirements, or due to congestion or a change in radio link quality.

Release 99 also solves the Release 97/98 problem of having only one PDP context installed per PDP address; thus all different applications running on top of this PDP address with one single QoS profile, independent of their specific requirements on, e.g., delay or reliability. GPRS Release 99 provides the possibility to install multiple PDP contexts per PDP address. Each PDP context is uniquely associated with a TFT which identifies the traffic flow. This makes it possible to assign different QoS profiles to simultaneous traffic flows, so that each application may receive the appropriate QoS requirement.

CHAPTER 4

# Enhanced Data Rates for GSM Evolution

*Enhanced Data rates for GSM Evolution* (EDGE) is a further development of the GSM data services *High-Speed Circuit-Switched Data* (HSCSD) and GPRS and is suitable for circuit- and packet-switched services. The circuit-oriented part is the *Enhanced Circuit-Switched Data* (ECSD). The packet-oriented part is the *Enhanced General Packet Radio Service* (EGPRS). Applying modified modulation and coding schemes EDGE reaches very high raw bit rates of up to 69 kbit/s per GSM physical channel. If a user utilizes all 8 time slots in parallel, the theoretical maximum raw bit rate rises to 554 kbit/s. The maximum bearer bit rate achievable rises to about 384 kbit/s [Furuskär et al. (1999b)]. EDGE was introduced to the ETSI for the first time in 1997 for the evolution of GSM. After a successful feasibility study of the ETSI the standardization process for EDGE was initiated. Although EDGE was introduced for the evolution of GSM, this concept can be applied to increase the data rate in other systems.

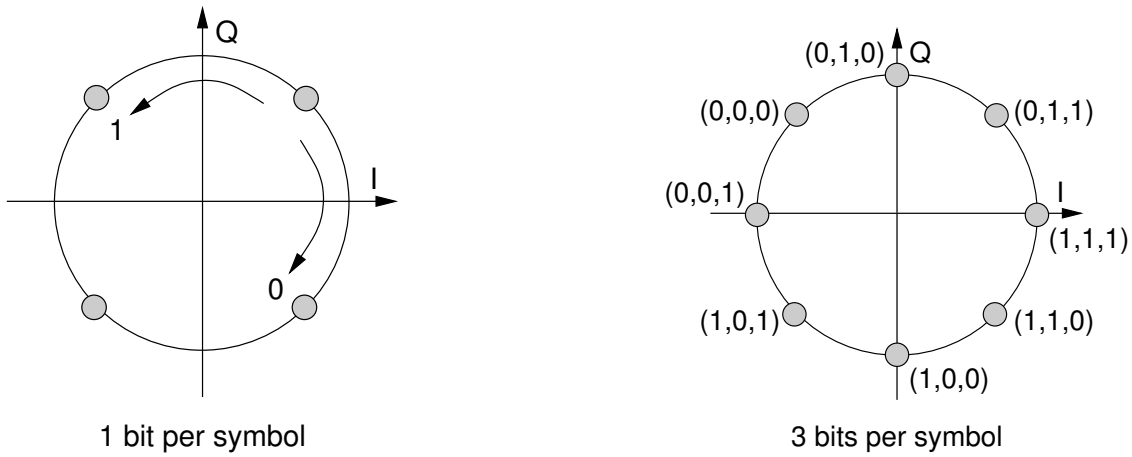
Since the network architecture of the GSM will remain similar for EDGE, the modifications at the air interface are depicted in this chapter. To support higher data rates, a modulation scheme called *8-Phase-Shift-Keying* (8-PSK) is introduced which will not replace the current *Gaussian Minimum Shift Keying* (GMSK) but coexist with it. With 8-PSK it is possible to provide a higher data rate, which is necessary to support bandwidth extensive data applications.

The modifications mostly concern the RLC/MAC layer and the physical layer. Since these protocols are implemented in the MS and the *Base Station* (BS), both have to be modified. In reality, the changes that have to be made comprise a new EDGE transceiver unit and software upgrades to the *Base Station Controller* (BSC), which then can handle standard GSM or GPRS traffic and will automatically switch to EDGE mode when needed.

The core of EDGE is the *Link Quality Control* (LQC) mechanism that allows the adaptation of the *Modulation and Coding Schemes* (MCSs) to a changing radio link quality. Although *Link Adaptation* (LA) is already possible within the GPRS standard, higher CSs are not supported by the actual equipment and will probably only be introduced together with EDGE functionality.

Additionally, a type-II-hybrid ARQ (soft ARQ) scheme is introduced. Soft information is stored during retransmissions to enable *Incremental Redundancy* (IR). The RLC/MAC protocol structure and retransmission mechanism proposed for EGPRS are based on the GPRS standard. An LLC frame is divided into a number of RLC blocks. The average duration of an RLC block is 20 ms, whereas its information content ranges from 176 to 1184 bit, depending on the MCS used. For error detection, the RLC blocks contain a CRC field. Upon reception of an RLC block, the receiver checks the CRC and determines if retransmission of the RLC block is necessary. To give a more detailed introduction, in the following the modulation scheme 8-PSK used for EDGE is described and compared with GMSK currently used for GSM. Additionally the new MCSs are presented. Then the main characteristics of the LQC mechanisms are given in Section 4.3, comprising LA and IR.

Finally the RLC/MAC protocol is adapted to the higher data rates that have to be supported. The RLC parameters *Sequence Number Space* (SNS), *Window Size* (WS) and




(a) GMSK modulation

(b) 8-PSK modulation

**Figure 4.1:** I/Q-diagram of GMSK and 8-PSK modulation

consequentially the *Block Sequence Number* (BSN) field and the *Received Block Bitmap* (RBB) structure are adapted (see Section 3.4.5.3).

## 4.1 8-PSK Modulation versus GMSK Modulation

The modulation scheme that is used in GSM is called *Gaussian Minimum Shift Keying* (GMSK). In GPRS the same modulation type will be used and high bit rates are achieved through multislot operation. To allow higher bit rates within EGPRS, the *8-Phase-Shift-Keying* (8-PSK) modulation scheme is introduced in addition to GMSK [Furuskär et al. (1999b); Sollenberger et al. (1999)]. The differences between GMSK and 8-PSK are illustrated in an I/Q diagram as shown in Figure 4.1.

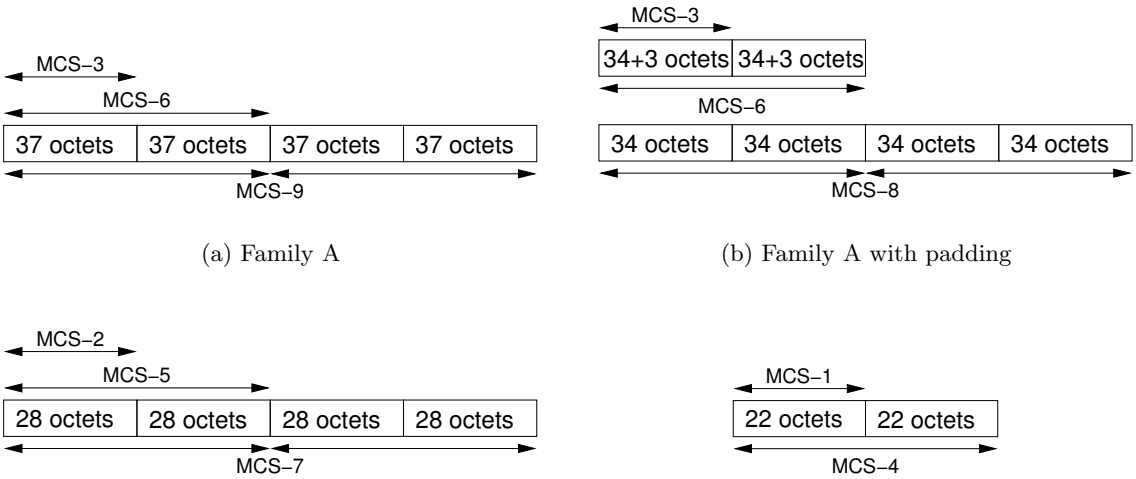
Within GMSK transmitting a 0-bit or a 1-bit is represented by incrementing the phase with $\pm 1/2\,\pi$. Every symbol that is transmitted represents one bit, which means that symbol rate and bit rate are equal. The symbol rate of 271 kbit/s is divided over eight time slots. After burst formatting (see Section 3.4.6) a bit rate of 22.8 kbit/s per time slot is achieved. Via different channel coding schemes this results in different net bit rates.

Within 8-PSK three consecutive bits are mapped onto one symbol in the I/Q-plane. With the same symbol rate as in GMSK of 271 kbit/s, bit rates of up to 813 kbit/s can be achieved. With burst formatting, this results in a bit rate of 69.2 kbit/s per time slot. 8-PSK modulation still has the GSM spectrum mask and leaves the burst duration unchanged, which offers the possibility of using both modulation schemes next to each other. However, 8-PSK modulated data is less robust if the channel conditions are bad. The advantage of the new modulation scheme is support for higher data rates under good channel conditions, and to reuse at the same time the channel structure of the GPRS system. Since the *Bit Error Ratio* (BER) depends significantly on the channel conditions, *Link Quality Control* (LQC) is of major importance for a high throughput in EGPRS.

LA as part of LQC gives the possibility of changing to another modulation scheme or channel coding, when the channel quality changes. Through the adaptation to the radio channel conditions it is possible to offer the highest data rates in good propagation conditions close to the site of the BSs, whereas under lower quality conditions a more

**Table 4.1:** EGPRS modulation and coding schemes and coding parameters

| MCS | CR | HCR | Data per radio block [bit] | Fam. | RLC info [byte] | Data rate [kbit/s] |
|---|---|---|---|---|---|---|
| 9 (8-PSK) | 1.0 | 0.36 | 2×592 | A | 2×74 | 59.2 |
| 8 (8-PSK) | 0.92 | 0.36 | 2×544 | A | 2×68 | 54.4 |
| 7 (8-PSK) | 0.76 | 0.36 | 2×448 | B | 2×56 | 44.8 |
| 6 (8-PSK) | 0.49 | 1/3 | 592 *(544+48)* | A | 74 | 29.6 (27.2) |
| 5 (8-PSK) | 0.37 | 1/3 | 448 | B | 56 | 22.4 |
| 4 (GMSK) | 1.0 | 0.53 | 352 | C | 44 | 17.6 |
| 3 (GMSK) | 0.80 | 0.53 | 296 *(272+24)* | A | 37 | 14.8 (13.6) |
| 2 (GMSK) | 0.66 | 0.53 | 224 | B | 28 | 11.2 |
| 1 (GMSK) | 0.53 | 0.53 | 176 | C | 22 | 8.8 |




(a) Family A

(b) Family A with padding

(c) Family B

(d) Family C

**Figure 4.2:** Basic block sizes and MCS families

robust coding scheme is selected.

## 4.2 Modulation and Coding Schemes

The additional use of 8-PSK modulation enables the introduction of new *Modulation and Coding Schemes* (MCSs). The possible MCSs are depicted in Table 4.1. MCS-1 to MCS-4 use GMSK as the modulation scheme and nearly equal the coding schemes used in GPRS, providing data rates up to 17.6 kbit/s. MCS-5 to MCS-9 use the previously described 8-PSK modulation scheme enhancing the data rate up to 59.2 kbit/s per time slot. The MCSs are classified in three different MCS families consisting of different basic payload sizes. Figure 4.2 gives an overview of how an RLC block is created depending on the MCS family. For instance, MCS-9 comprises four basic units of 37 octets each, whereas MCS-6 combines two basic units and MCS-3 only one. Through transmitting a different number of payload units within a 20 ms radio block duration, different code rates are achieved, resulting in bit rates of 8.8 kbit/s up to 59.2 kbit/s per time slot. According to the link quality, an initial MCS is selected for each RLC block. For retransmissions, it is only

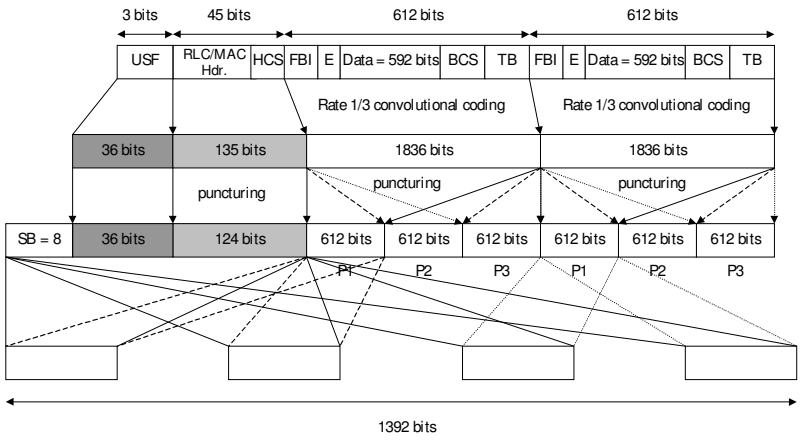



**Figure 4.3:** Coding and puncturing for MCS-9, uncoded 8-PSK, two RLC blocks

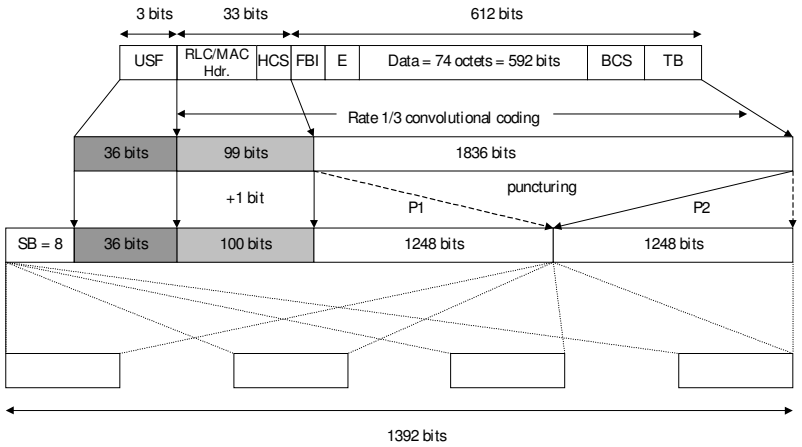



**Figure 4.4:** Coding and puncturing for MCS-6, rate 0.49 8-PSK, one RLC block

allowed to use the same MCS or to switch to another MCS of the same MCS family.

The following figures show the coding and puncturing used in Family A (MCS-9: Figure 4.3, MCS-6: Figure 4.4 and MCS-3: Figure 4.5). They differ in the amount of redundancy used for error correction, which determines accordingly the payload size.

Note that using MCS-7, MCS-8 and MCS-9, two RLC PDUs are transmitted within 20 ms. Both PDUs are interleaved over only two bursts, whereas their common header is interleaved over all four. Using a lower MCS, the RLC PDU (data and header) is interleaved over all four bursts. Figure 4.3 shows a MCS-9 RLC/MAC block consisting of two RLC PDUs of 592 bit (74 octets) payload and separate header containing the FBI field, extension field and the BSN (not pictured) to identify the RLC PDU within a TBF. The two RLC PDUs do not have to be in sequence; for example the first PDU can be a new RLC PDU and the second a retransmission of a previously sent RLC PDU. If there are no further RLC blocks to send, no retransmissions or pending retransmissions requested, the second payload is padded.

Initially, the RLC data block will be encoded with Puncturing Scheme PS-1, whereas PS-2 and PS-3 are used for retransmissions, then enabling the use of IR by using the different punctured information in PS-1, PS-2 and PS-3. The EDGE standard foresees an LA algorithm for the dynamic selection of MCSs to adapt to the quality of the radio link. This includes measurement and reporting the quality of the downlink and demands the

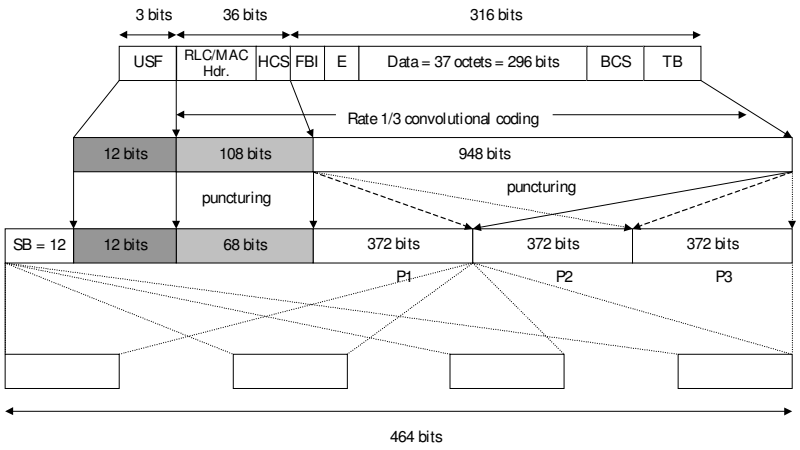



**Figure 4.5:** Coding and puncturing for MCS-3, rate 0.80 GMSK, one RLC block

indication of new modulation schemes (see Figure 4.6).

## 4.3 Link Quality Control

*Link Quality Control* (LQC) is a term used for techniques to adapt the channel coding of the radio link to the varying channel quality. Different modulation and coding schemes are optimal during different situations, depending on the link quality. The LQC used for EDGE is performed through the techniques of

- *Link Adaptation* (LA)
- *Incremental Redundancy* (IR)

LA provides a dynamic switching between coding and modulation schemes, so that the highest throughput (e.g., maximum user bit rate) according to the time-varying link quality (e.g. $C/I$) is achieved. With IR, information is first sent with very little coding. If decoding is successful, this will yield a very high user bit rate or throughput. However, if decoding is unsuccessful, then additional coded bits are sent until decoding is successful. The more coding bits that have to be sent, the less the resulting bit rate and the higher the delay. As a result of the link quality control, a low radio link quality will not cause a dropped data transfer, but only give a reduced bit rate for the user. Furthermore, a tight frequency plan (with a small cluster size, say 3) can be introduced while still providing high data rates for packet data services.

### 4.3.1 Link Adaptation

An LA scheme is proposed to maximize the user bit rate on the link through estimation of the time varying link quality in EGPRS and respectively adapt the most appropriate modulation and coding scheme. Through measurements of the link quality, the predictive algorithm estimates the performance of the currently used scheme and decides whether another MCS would perform better than the currently used one. The algorithm chooses the MCS with the best performance for the currently measured radio link quality and therefore for the expected link quality during the next bursts (it is assumed that the link quality can vary with time and location). During active TBF the most appropriate MCS is chosen on the basis of the channel quality measurements concerning the used MCS family. For compatibility reasons it is not allowed to switch the chosen MCS family during retransmissions, otherwise too much padding would decrease efficiency.

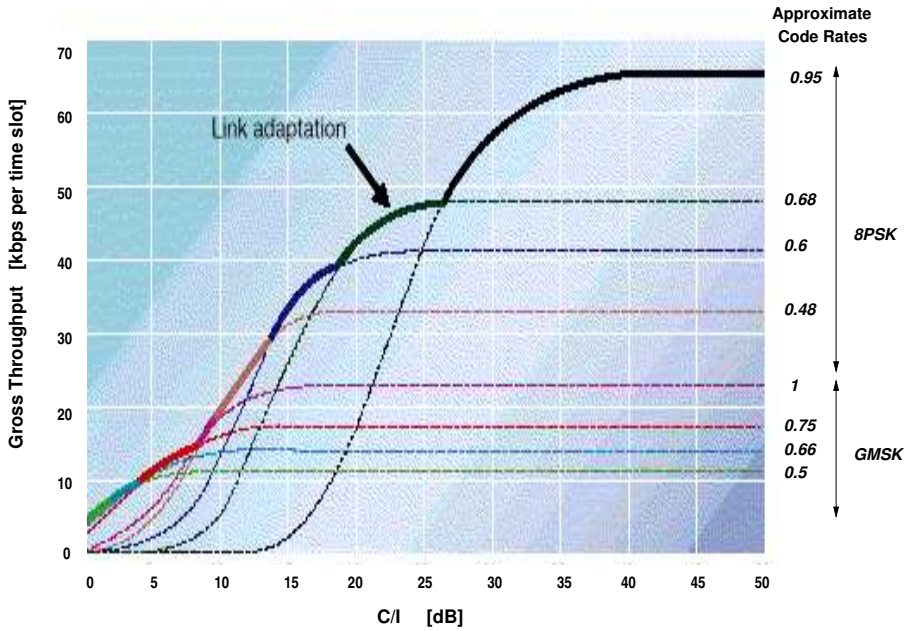



**Figure 4.6:** Link adaptation algorithm [Furuskär and Olofsson (1999)]

In order to perform the most efficient LA, measurements of the quality of the radio channel are the basis of the choice of the MCS. The operation of link quality control relies on measurements of the link quality, depending also on $C/I$ values, frequency errors, time dispersion, interleaving gain due to frequency hopping, velocities and bit error probability.

### 4.3.2 Incremental Redundancy

Operating in RLC acknowledged mode, retransmissions are possible, when requested, or when capacity is available for pending retransmissions (RLC blocks not yet acknowledged). An ARQ scheme ensures an error-free data transmission by retransmitting erroneous blocks. When an error is detected in a block, all information about the block is usually discarded. The basic idea of IR is soft combining, i.e., not to forget what was already received in the erroneous transmission. Instead, the soft bit information from the erroneous blocks is saved and combined with the information received with the next (erroneous) retransmission. That way the *Block Error Ratio* (BLER) is reduced.

With IR mode PS-1 is used at first, and if decoding is unsuccessful, PS-2 will be used, etc. The IR functionality saves soft information of PS-1 and performs a joint decoding with PS-2 after the received retransmission with PS-2. A *Synchronization downlink Burst* (SB) is used to indicate which code rate and which puncturing scheme is used. Information about link quality control is given in [3GPP TSG GERAN (2002d)], [3GPP TSG GERAN (2002e)], and [3GPP TSG GERAN (2002f)]. Through measurements of the link quality every 60 ms (example inter-arrival time of the measurement report), the most suitable MCS family is chosen, based on the mean and variance of the $C/I$. For the following explanation, MCS family A is assumed. All assumptions apply also for the other families. For retransmissions three possible scenarios are drawn. The first one assumes that the link quality remains stable during the retransmissions. Then the RLC data blocks can be retransmitted with the originally used MCS. If the link quality worsens,

an MCS downgrading becomes necessary, otherwise if the link quality improves, an MCS upgrading is possible. In the following, proposals for upgrading and downgrading actions are provided.

**Upgrading** When performing an MCS upgrade, retransmissions can be executed using the old MCS, allowing the use of IR by using different puncturing schemes (PS-1, PS-2, PS-3). That way, the BLER is reduced by using information from the different puncturing schemes. Otherwise no IR is possible.

**Downgrading** If the channel quality worsens, an MCS downgrading is necessary to provide a low BLER through higher protection of the bits. Thus, the retransmissions are carried out using a lower MCS, rejecting the possibility of the use of IR because of the incompatible puncturing of the different MCSs. For instance, when MCS-6 is chosen, when previous RLC blocks were sent with MCS-9, no coding problem occurs since MCS-9 comprises two individual RLC blocks consisting of two basic blocks each. MCS-6 also comprises two basic blocks, therefore only the coding, puncturing and interleaving will change and not the number of RLC blocks distinguished by separate BSNs. Retransmission of a previous MCS-9 RLC block results in a retransmission of two separate RLC blocks in MCS-6.

Assuming MCS-6 and downgrading to MCS-3, the payload data has to be deblocked into two subblocks (also called split blocks) and the header duplicated, creating two MCS-3 RLC blocks with the same BSN. Both MCS-3 blocks are retransmitted consecutively with the split bit set to 1 for the first and 2 for the second MCS-3 block to allow accurate assembly at the receiver side requiring the error-free arrival of both the first and second subblocks. Performing downgrading, all consecutive RLC blocks will be sent using the lower MCS, remaining in this MCS even if changed channel quality allows upgrading. This applies only for retransmissions, whereas the newly segmented LLC frame will be sent with the actual MCS, taking advantage of the better channel quality and the transmission capacity of the higher MCS.

It is not possible to acknowledge the subblocks separately, therefore the receiver has to wait until both subblocks have arrived before they enter the receiver window. The RLC block, consisting of two subblocks, is acknowledged positively if both subblocks have arrived error-free, otherwise both subblocks have to be requested by using the BSN of the former MCS-6 PDU.

## 4.4 Flow Control Modifications for EDGE

To support higher data rates in an efficient way the flow control function in the RLC/MAC protocol has to be modified. The *Window Size* (WS) is increased from 64 in GPRS to the range from 64 to 1024 in EGPRS. Consequentially the *Sequence Number Space* (SNS) is adapted from 128 to 2048 and the *Block Sequence Number* (BSN) is increased from 7 bit to 11 bit. To be able to maintain the selective acknowledgement function for this large SNS the *Received Block Bitmap* (RBB) is segmented into a *First Partial Bitmap* (FPB) and *Next Partial Bitmap* (NPB) controlled by a *Extended Polling Function* in EGPRS.

# 3

# Introduction to WCDMA

Peter Muszynski and Harri Holma

## 3.1 Introduction

This chapter introduces the principles of the WCDMA air interface. Special attention is drawn to those features by which WCDMA differs from GSM and IS-95. The main parameters of the WCDMA physical layer are introduced in Section 3.2. The concept of spreading and despreading is described in Section 3.3, followed by a presentation of the multipath radio channel and Rake receiver in Section 3.4. Other key elements of the WCDMA air interface discussed in this chapter are power control and soft and softer handovers. The need for power control and its implementation are described in Section 3.5, and soft and softer handover in Section 3.6.

## 3.2 Summary of the Main Parameters in WCDMA

We present the main system design parameters of WCDMA in this section and give brief explanations for most of them. Table 3.1 summarizes the main parameters related to the WCDMA air interface. Here we highlight some of the items that characterize WCDMA:

- WCDMA is a wideband Direct-Sequence Code Division Multiple Access (DS-CDMA) system, i.e. user information bits are spread over a wide bandwidth by multiplying the user data with quasi-random bits (called chips) derived from CDMA spreading codes. In order to support very high bit rates (up to 2 Mbps), the use of a variable spreading factor and multicode connections is supported. An example of this arrangement is shown in Figure 3.1.
- The chip rate of 3.84 Mcps leads to a carrier bandwidth of approximately 5 MHz. DS-CDMA systems with a bandwidth of about 1 MHz, such as IS-95, are commonly referred to as narrowband CDMA systems. The inherently wide carrier bandwidth of WCDMA supports high user data rates and also has certain performance benefits, such as increased multipath diversity. Subject to his operating license, the network operator can deploy multiple 5 MHz carriers to increase capacity, possibly in the form of hierarchical cell layers. Figure 3.1 also shows this feature. The actual carrier spacing can be selected on a 200 kHz grid between approximately 4.4 and 5 MHz, depending on interference between the carriers.

*WCDMA for UMTS: HSPA Evolution and LTE, Fifth Edition* Edited by Harri Holma and Antti Toskala

**Table 3.1** Main WCDMA parameters

| | |
|---|---|
| Multiple access method | DS-CDMA |
| Duplexing method | Frequency division duplex/time division duplex |
| Base station synchronization | Asynchronous operation |
| Chip rate | 3.84 Mcps |
| Frame length | 10 ms |
| Service multiplexing | Multiple services with different quality of service requirements multiplexed on one connection |
| Multirate concept | Variable spreading factor and multicode |
| Detection | Coherent using pilot symbols or common pilot |
| Multiuser detection, smart antennas | Supported by the standard, optional in the implementation |

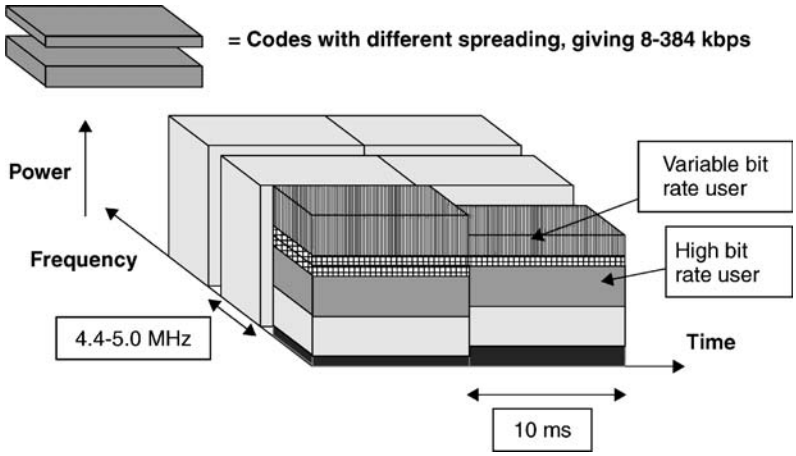



**Figure 3.1** Allocation of bandwidth in WCDMA in the time–frequency–code space

- WCDMA supports highly variable user data rates, in other words, the concept of obtaining Bandwidth on Demand (BoD) is well supported. The user data rate is kept constant during each 10 ms frame. However, the data capacity among the users can change from frame to frame. Figure 3.1 also shows an example of this feature. This fast radio capacity allocation will typically be controlled by the network to achieve optimum throughput for packet data services.
- WCDMA supports two basic modes of operation: Frequency Division Duplex (FDD) and Time Division Duplex (TDD). In the FDD mode, separate 5 MHz carrier frequencies are used for the uplink and downlink respectively, whereas in TDD only one 5 MHz is time-shared between the uplink and downlink. Uplink is the connection from the mobile to the base station, and downlink is that from the base station to the mobile.
- The TDD mode is based heavily on FDD mode concepts and was added in order to leverage the basic WCDMA system also for the unpaired spectrum allocations of the ITU for the IMT-2000 systems. The TDD mode is described in detail in Chapter 18.
- WCDMA supports the operation of asynchronous base stations, so that, unlike in the synchronous IS-95 system, there is no need for a global time reference such as a GPS. Deployment of indoor and micro-base stations is easier when no GPS signal needs to be received.

- WCDMA employs coherent detection on uplink and downlink based on the use of pilot symbols or common pilot. While already used on the downlink in IS-95, the use of coherent detection on the uplink is new for public CDMA systems and will result in an overall increase of coverage and capacity on the uplink.
- The WCDMA air interface has been crafted in such a way that advanced CDMA receiver concepts, such as multiuser detection and smart adaptive antennas, can be deployed by the network operator as a system option to increase capacity and/or coverage. In most second generation systems no provision was made for such receiver concepts and as a result they are either not applicable or can be applied only under severe constraints with limited increases in performance.
- WCDMA is designed to be deployed in conjunction with GSM. Therefore, handovers between GSM and WCDMA are supported in order to be able to leverage the GSM coverage for the introduction of WCDMA.

In the following sections of this chapter we will briefly review the generic principles of CDMA operation. In the subsequent chapters, the above-mentioned aspects specific to the WCDMA standard will be presented and explained in more detail. The basic CDMA principles are also described in [1, 2, 3 and 4].

## 3.3 Spreading and Despreading

Figure 3.2 depicts the basic operations of spreading and despreading for a DS-CDMA system. User data is here assumed to be a BPSK-modulated bit sequence of rate $R$, the user data bits assuming the values of $\pm 1$. The spreading operation, in this example, is the multiplication of each user data bit with a sequence of 8 code bits, called chips. We assume this also for the BPSK spreading modulation. We see that the resulting spread data is at a rate of $8 \times R$ and has the same random (pseudo-noise-like) appearance as the spreading code. In this case we would say that we used a spreading factor of 8. This wideband signal would then be transmitted across a wireless channel to the receiving end.

During despreading, we multiply the spread user data/chip sequence, bit duration by bit duration, with the very same 8 code chips as we used during the spreading of these bits. As shown, the original

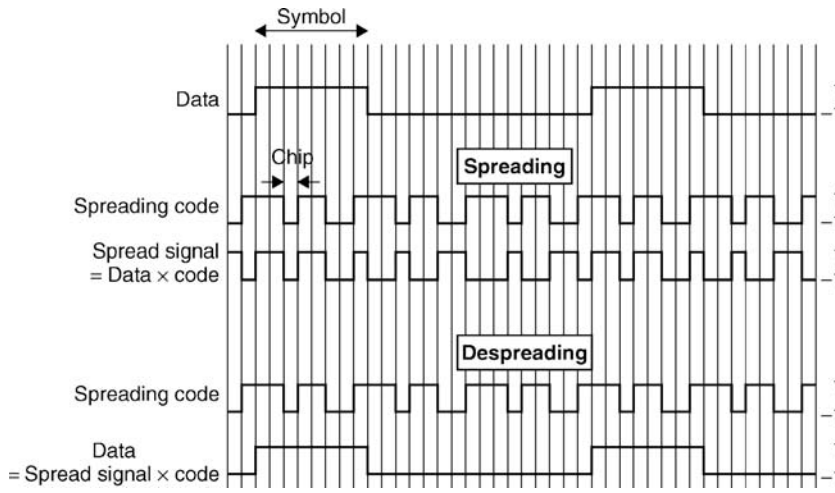



**Figure 3.2** Spreading and despreading in DS-CDMA

user bit sequence has been recovered perfectly, provided we have (as shown in Figure 3.2) also perfect synchronization between the spread user signal and the (de)spreading code.

The increase of the signaling rate by a factor of 8 corresponds to a widening (by a factor of 8) of the occupied spectrum of the spread user data signal. Due to this virtue, CDMA systems are more generally called spread spectrum systems. Despreading restores a bandwidth proportional to $R$ for the signal.

The basic operation of the correlation receiver for CDMA is shown in Figure 3.3. The upper half of the figure shows the reception of the desired own signal. As in Figure 3.2, we see the despreading operation with a perfectly synchronized code. Then, the correlation receiver integrates (i.e. sums) the resulting products (data $\times$ code) for each user bit.

The lower half of Figure 3.3 shows the effect of the despreading operation when applied to the CDMA signal of another user whose signal is assumed to have been spread with a different spreading code. The result of multiplying the interfering signal with the own code and integrating the resulting products leads to interfering signal values lingering around 0.

As can be seen, the amplitude of the own signal increases on average by a factor of 8 relative to that of the user of the other interfering system, i.e. the correlation detection has raised the desired user signal by the spreading factor, here 8, from the interference present in the CDMA system. This effect is termed 'processing gain' and is a fundamental aspect of all CDMA systems, and in general of all spread spectrum systems. Processing gain is what gives CDMA systems the robustness against self-interference that is necessary in order to reuse the available 5 MHz carrier frequencies over geographically close distances. Let's take an example with real WCDMA parameters. Speech service with a bit rate of 12.2 kbps has a processing gain of $25\,\text{dB} = 10 \times \log_{10}$ (3.84e6/12.2e3). After despreading, the signal power needs to be typically a few decibels above the interference and noise power. The required power density over the interference power density after despreading is designated as $E_b/N_0$ in this book, where $E_b$ is the energy, or power density, per user bit and $N_0$ is the interference and noise power density. For speech service $E_b/N_0$ is typically in the order of 5.0 dB, and the required wideband signal-to-interference ratio is therefore 5.0 dB minus the processing gain $= -20.0$ dB. In other words, the signal power can be 20 dB under the interference or thermal noise power, and the WCDMA receiver can still detect the signal. The wideband signal-to-interference ratio

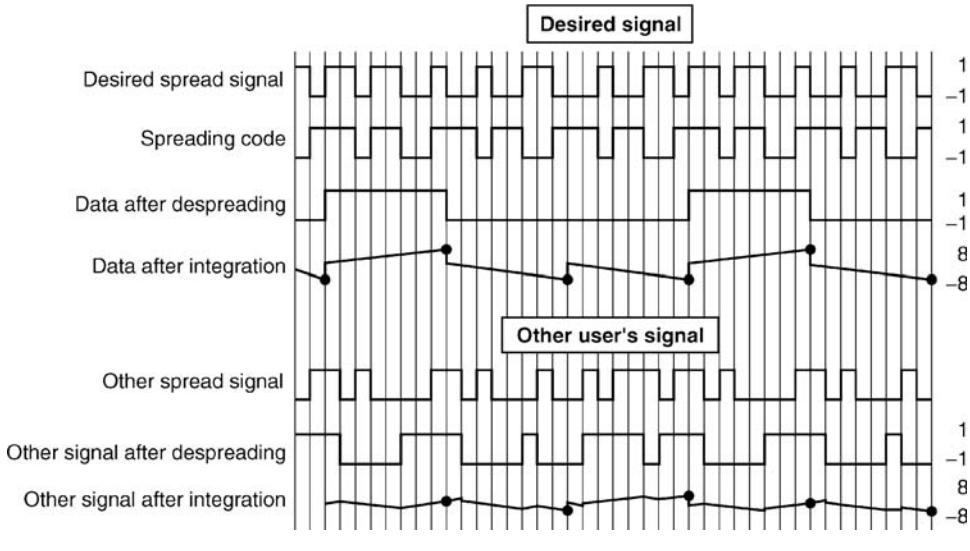



**Figure 3.3** Principle of the CDMA correlation receiver

is also called the carrier-to-interference ratio $C/I$. Due to spreading and despreading, $C/I$ can be lower in WCDMA than, for example, in GSM. A good quality speech connection in GSM requires $C/I = 9 - 12\,\text{dB}$.

Since the wideband signal can be below the thermal noise level, its detection is difficult without knowledge of the spreading sequence. For this reason, spread spectrum systems originated in military applications where the wideband nature of the signal allowed it to be hidden below the omnipresent thermal noise.

Note that within any given channel bandwidth (chip rate) we will have a higher processing gain for lower user data bit rates than for high bit rates. In particular, for user data bit rates of 2 Mbps, the processing gain is less than 2 (= 3.84 Mcps/2 Mbps = 1.92 which corresponds to 2.8 dB) and some of the robustness of the WCDMA waveform against interference is clearly compromised.

Both base stations as well as mobiles for WCDMA essentially use this type of correlation receiver. However, due to multipath propagation (and possibly multiple receive antennas), it is necessary to use multiple correlation receivers in order to recover the energy from all paths and/or antennas. Such a collection of correlation receivers, termed 'fingers', is what comprises the CDMA Rake receiver. We will describe the operation of the CDMA Rake receiver in further detail in the following section, but before doing so, we make some final remarks regarding the transformation of spreading/despreading when used for wireless systems.

It is important to understand that spreading/despreading by itself does not provide any signal enhancement for wireless applications. Indeed, the processing gain comes at the price of an increased transmission bandwidth (by the amount of the processing gain).

All the WCDMA benefits come rather 'through the back door' by the wideband properties of the signals when examined at the system level, rather than the level of an individual radio link:

1. The processing gain together with the wideband nature suggests a frequency reuse of 1 between different cells of a wireless system (i.e. a frequency is reused in every cell/sector). This feature can be used to obtain high spectral efficiency.
2. Having many users share the same wideband carrier for their communications provides interferer diversity, i.e. the multiple access interference from many system users is averaged out, and this again will boost capacity compared to systems where one has to plan for the worst-case interference.
3. However, both the above benefits require the use of tight power control and soft handover to avoid one user's signal blocking the others' communications. Power control and soft handover will be explained later in this chapter.
4. With a wideband signal, the different propagation paths of a wireless radio signal can be resolved at higher accuracy than with signals at a lower bandwidth. This results in a higher diversity content against fading, and thus improved performance.

## 3.4 Multipath Radio Channels and Rake Reception

Radio propagation in the land mobile channel is characterized by multiple reflections, diffractions and attenuation of the signal energy. These are caused by natural obstacles such as buildings, hills, and so on, resulting in so-called multipath propagation. There are two effects resulting from multipath propagation that we are concerned with in this section:

1. The signal energy (pertaining, for example, to a single chip of a CDMA waveform) may arrive at the receiver across clearly distinguishable time instants. The arriving energy is 'smeared' into a certain multipath delay profile, see Figure 3.4, for example. The delay profile extends typically from 1 to 2 μs in urban and suburban areas, although in some cases delays as long as 20 μs or more with significant signal energy have been observed in hilly areas. The chip duration at 3.84 Mcps

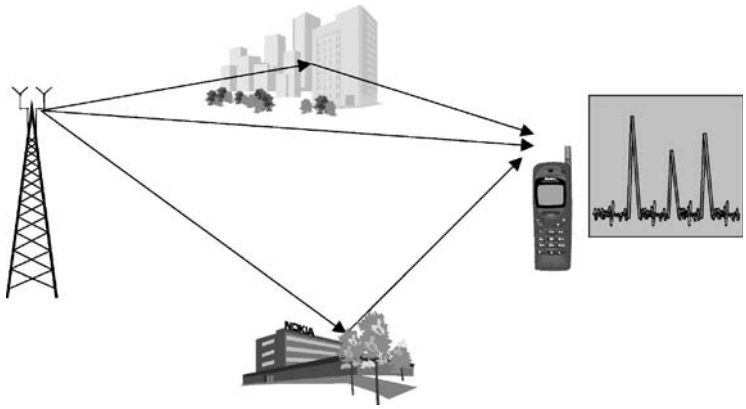


**Figure 3.4** Multipath propagation leads to a multipath delay profile

is 0.26 μs. If the time difference of the multipath components is at least 0.26 μs, the WCDMA receiver can separate those multipath components and combine them coherently to obtain multipath diversity. The 0.26 μs delay can be obtained if the difference in path lengths is at least 78 m (= speed of light ÷ chip rate = $3.0 \cdot 10^8$ ms$^{-1}$ ÷ 3.84 Mcps). With a chip rate of about 1 Mcps, the difference in the path lengths of the multipath components must be about 300 m, which cannot be obtained in small cells. Therefore, it is easy to see that the 5 MHz WCDMA can provide multipath diversity in small cells, which is not possible with IS-95.

2. Also, for a certain time delay position there are usually many paths nearly equal in length along which the radio signal travels. For example, paths with a length difference of half a wavelength (at 2 GHz this is approximately 7 cm) arrive at virtually the same instant when compared to the duration of a single chip, which is 78 m at 3.84 Mcps. As a result, signal cancellation, called fast fading, takes place as the receiver moves across even short distances. Signal cancellation is best understood as a summation of several weighted phasors that describe the phase shift (usually modulo radio wavelength) and attenuation along a certain path at a certain time instant.

Figure 3.5 shows an exemplary fast fading pattern as would be discerned for the arriving signal energy at a particular delay position as the receiver moves. We see that the received signal power can drop considerably (by 20–30 dB) when phase cancellation of multipath reflections occurs. Because of the underlying geometry causing the fading and dispersion phenomena, signal variations due to fast fading occur several orders of magnitude more frequently than changes in the average multipath delay profile. The statistics of the received signal energy for a short-term average are usually well described by the Rayleigh distribution, see, e.g. [5] and [6]. These fading dips make error-free reception of data bits very difficult, and countermeasures are needed in WCDMA. The countermeasures against fading in WCDMA are shown below.

1. The delay dispersive energy is combined by utilizing multiple Rake fingers (correlation receivers) allocated to those delay positions on which significant energy arrives.
2. Fast power control and the inherent diversity reception of the Rake receiver are used to mitigate the problem of fading signal power.
3. Strong coding and interleaving and retransmission protocols are used to add redundancy and time diversity to the signal and thus help the receiver in recovering the user bits across fades.

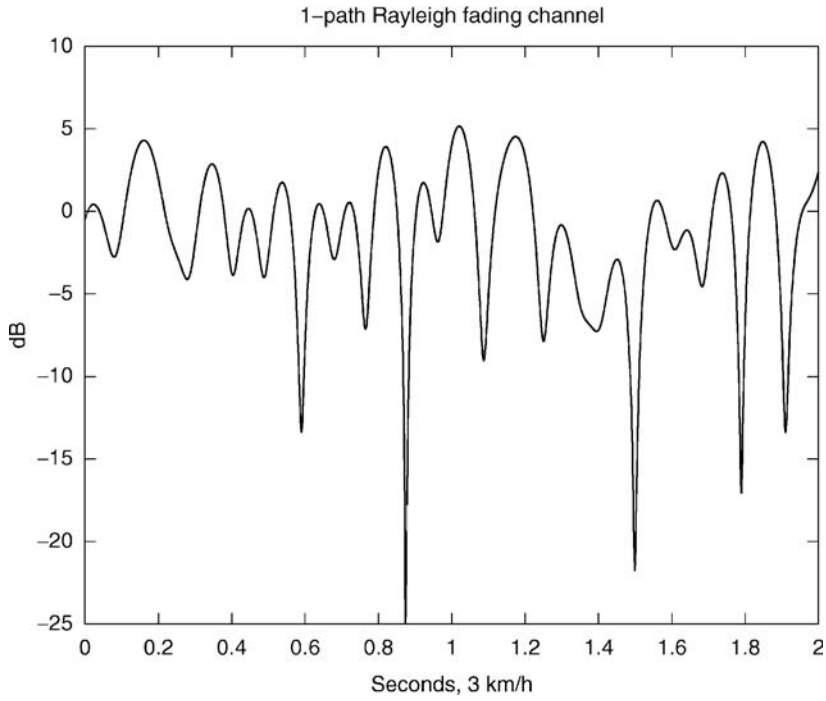



**Figure 3.5** Fast Rayleigh fading as caused by multipath propagation

The dynamics of the radio propagation suggest the following operating principle for the CDMA signal reception:

1. Identify the time delay positions at which significant energy arrives and allocate correlation receivers, i.e. Rake fingers, to those peaks. The granularity for acquiring the multipath delay profile is in the order of one chip duration (typically within the range of $\frac{1}{4}-\frac{1}{2}$ chip duration) with an update rate in the order of some tens of milliseconds.
2. Within each correlation receiver, track the fast-changing phase and amplitude values originating from the fast fading process and remove them. This tracking process has to be very fast, with an update rate in the order of 1 ms or less.
3. Combine the demodulated and phase-adjusted symbols across all active fingers and present them to the decoder for further processing.

Figure 3.6 illustrates points 2 and 3 by depicting modulation symbols (BPSK or QPSK) as well as the instantaneous channel state as weighted complex phasors. To facilitate point 2, WCDMA uses known pilot symbols that are used to sound the channel and provide an estimate of the momentary channel state (value of the weighted phasor) for a particular finger. Then the received symbol is rotated back, so as to undo the phase rotation caused by the channel. Such channel-compensated symbols can then be simply summed together to recover the energy across all delay positions. This processing is also called Maximal Ratio Combining (MRC).

Figure 3.7 shows a block diagram of a Rake receiver with three fingers according to these principles. Digitized input samples are received from the RF front-end circuitry in the form of I and Q branches (i.e. in complex low-pass number format). Code generators and a correlator perform the despreading

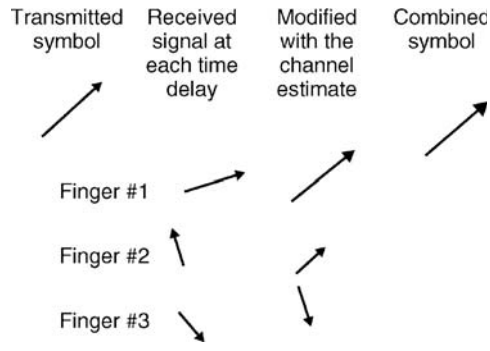



**Figure 3.6** The principle of maximal ratio combining within the CDMA Rake receiver

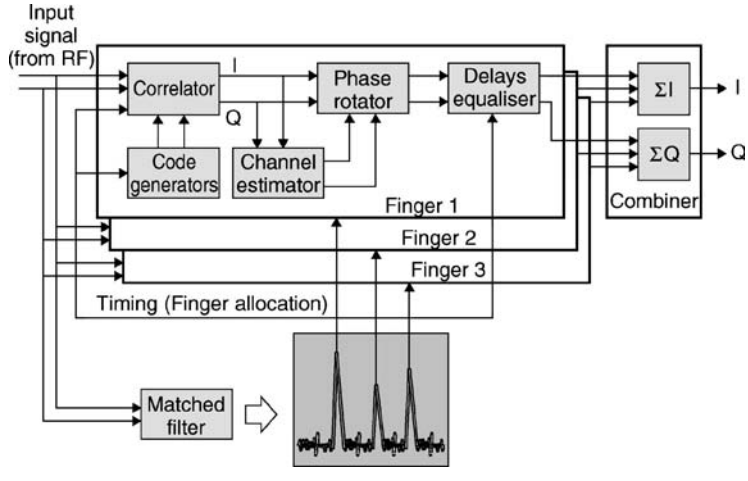



**Figure 3.7** Block diagram of the CDMA Rake receiver

and integration to user data symbols. The channel estimator uses the pilot symbols to estimate the channel state which will then be removed by the phase rotator from the received symbols. The delay is compensated for the difference in the arrival times of the symbols in each finger. The Rake combiner then sums the channel-compensated symbols, thereby providing multipath diversity against fading. Also shown is a matched filter used for determining and updating the current multipath delay profile of the channel. This measured and possibly averaged multipath delay profile is then used to assign the Rake fingers to the largest peaks.

In typical implementations of the Rake receiver, processing at the chip rate (correlator, code generator, matched filter) is done in ASICs, whereas symbol-level processing (channel estimator, phase rotator, combiner) is implemented by a DSP. Although there are several differences between the WCDMA Rake receiver in the mobile and the base station, all the basic principles presented here are the same.

Finally, we note that multiple receive antennas can be accommodated in the same way as multiple paths received from a single antenna: by just adding additional Rake fingers to the antennas, we can then receive all the energy from multiple paths *and* antennas. From the Rake receiver's perspective, there is essentially no difference between these two forms of diversity reception.

## 3.5 Power Control

Tight and fast power control is perhaps the most important aspect in WCDMA, in particular on the uplink. Without it, a single overpowered mobile could block a whole cell. Figure 3.8 depicts the problem and the solution in the form of closed loop transmission power control.

Mobile stations MS1 and MS2 operate within the same frequency, separable at the base station only by their respective spreading codes. It may happen that MS1 at the cell edge suffers a path loss, say, 70 dB above that of MS2 which is near the base station BS. If there were no mechanism for MS1 and MS2 to be power-controlled to the same level at the base station, MS2 could easily overshoot MS1 and thus block a large part of the cell, giving rise to the so-called near–far problem of CDMA. The optimum strategy in the sense of maximizing capacity is to equalize the received power per bit of all mobile stations at all times.

While one can conceive open loop power control mechanisms that attempt to make a rough estimate of path loss by means of a downlink beacon signal, such a method would be far too inaccurate. The prime reason for this is that the fast fading is essentially uncorrelated between uplink and downlink, due to the large frequency separation of the uplink and downlink bands of the WCDMA FDD mode. Open loop power control is, however, used in WCDMA, but only to provide a coarse initial power setting of the mobile station at the beginning of a connection.

The solution to power control in WCDMA is fast closed loop power control, also shown in Figure 3.8. In closed loop power control in the uplink, the base station performs frequent estimates of the received Signal-to-Interference Ratio (SIR) and compares it to a target SIR. If the measured SIR is higher than the target SIR, the base station will command the mobile station to lower the power; if it is too low, it will command the mobile station to increase its power. This measure–command–react cycle is executed at a rate of 1500 times per second (1.5 kHz) for each mobile station and thus operates faster than any significant change of path loss could possibly happen and, indeed, even faster than the speed of fast Rayleigh fading for low to moderate mobile speeds. Thus, closed loop power control will prevent any power imbalance among all the uplink signals received at the base station.

The same closed loop power control technique is also used on the downlink, though here the motivation is different: on the downlink there is no near–far problem due to the one-to-many scenario. All the signals within one cell originate from the one base station to all mobiles. It is, however, desirable to provide a marginal amount of additional power to mobile stations at the cell edge, as they suffer from increased other-cell interference. Also on the downlink a method of enhancing weak signals caused by Rayleigh fading with additional power is needed at low speeds when other error-correcting methods based on interleaving and error correcting codes do not yet work effectively.

Figure 3.9 shows how uplink closed loop power control works on a fading channel at low speed. Closed loop power control commands the mobile station to use a transmit power proportional to the inverse of the received power (or SIR). Provided the mobile station has enough headroom to ramp the

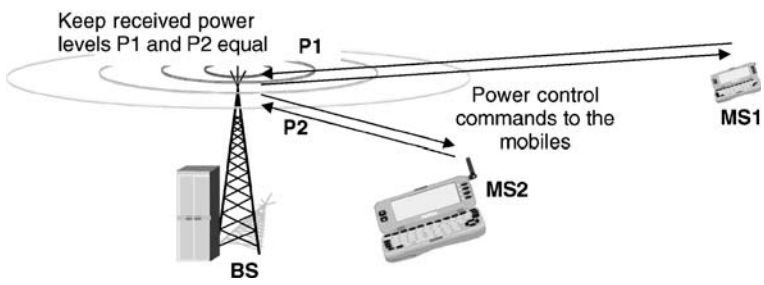



**Figure 3.8** Closed loop power control in CDMA

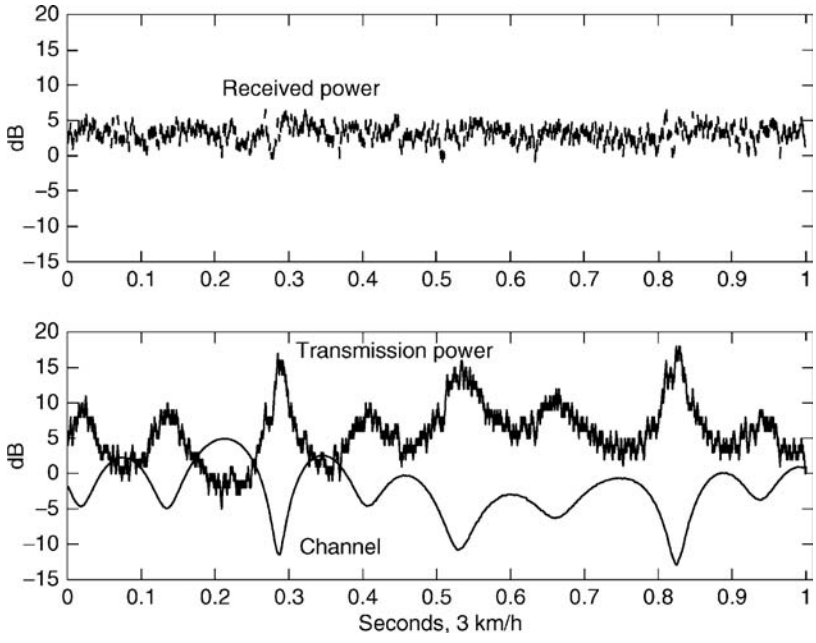



**Figure 3.9** Closed-loop power control compensates a fading channel

power up, only very little residual fading is left and the channel becomes an essentially non-fading channel as seen from the base station receiver.

While this fading removal is highly desirable from the receiver point of view, it comes at the expense of increased average transmit power at the transmitting end. This means that a mobile station in a deep fade, i.e. using a large transmission power, will cause increased interference to other cells. Figure 3.9 illustrates this point. The gain from the fast power control is discussed in more detail in Section 9.2.1.1.

Before leaving the area of closed loop power control, we mention one more related control loop connected with it: outer loop power control. Outer loop power control adjusts the target SIR set point in the base station according to the needs of the individual radio link and aims at a constant quality, usually defined as a certain target bit error rate (BER) or block error rate (BLER). Why should there be a need for changing the target SIR set point? The required SIR (there exists a proportional $E_b/N_0$ requirement) for, say, BLER $= 1\%$ depends on the mobile speed and the multipath profile. Now, if one were to set the target SIR set point for the worst case, i.e. high mobile speeds, one would waste much capacity for those connections at low speeds. Thus, the best strategy is to let the target SIR set point float around the minimum value that just fulfils the required target quality. The target SIR set point will change over time, as shown in the graph in Figure 3.10, as the speed and propagation environment changes. The gain of outer loop power control is discussed in detail in Section 9.2.2.1.

Outer loop control is typically implemented by having the base station tag each uplink user data frame with a frame reliability indicator, such as a CRC check result obtained during decoding of that particular user data frame. Should the frame quality indicator indicate to the Radio Network Controller (RNC) that the transmission quality is decreasing, the RNC in turn will command the base station to increase the target SIR set point by a certain amount. The reason for having outer loop control reside in the RNC is that this function should be performed after a possible soft handover combining. Soft handover will be presented in the next section.

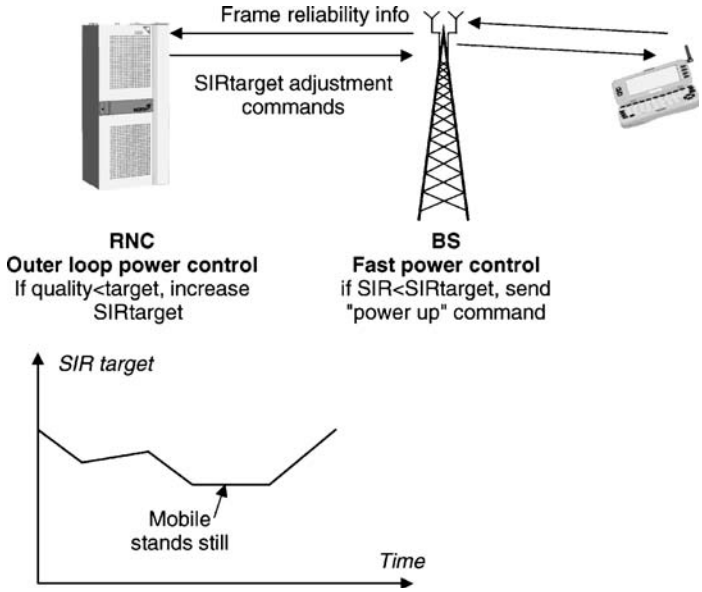



**Figure 3.10** Outer loop power control

## 3.6 Softer and Soft Handovers

During softer handover, a mobile station is in the overlapping cell coverage area of two adjacent sectors of a base station. The communications between mobile station and base station take place concurrently via *two* air interface channels, one for each sector separately. This requires the use of two separate codes in the downlink direction, so that the mobile station can distinguish the signals. The two signals are received in the mobile station by means of Rake processing, very similar to multipath reception, except that the fingers need to generate the respective code for each sector for the appropriate despreading operation. Figure 3.11 shows the softer handover scenario.

In the uplink direction a similar process takes place at the base station: the code channel of the mobile station is received in each sector, then routed to the same baseband Rake receiver and the maximal ratio combined there in the usual way. During softer handover only one power control loop per connection is active. Softer handover typically occurs in about 5–15% of connections.

Figure 3.12 shows the soft handover. During soft handover, a mobile station is in the overlapping cell coverage area of two sectors belonging to different base stations. As in softer handover, the communications between mobile station and base station take place concurrently via two air interface channels from each base station separately. As in softer handover, both channels (signals) are received at the mobile station by maximal ratio combining Rake processing. Seen from the mobile station, there are very few differences between softer and soft handover.

However, in the uplink direction soft handover differs significantly from softer handover: the code channel of the mobile station is received from both base stations, but the received data is then routed to the RNC for combining. This is typically done so that the same frame reliability indicator as provided for outer loop power control is used to select the better frame between the two possible candidates within the RNC. This selection takes place after each interleaving period, i.e. every 10–80 ms. Note that during soft handover two power control loops per connection are active, one for each base station. Power control in soft handover is discussed in Section 9.2.1.3.

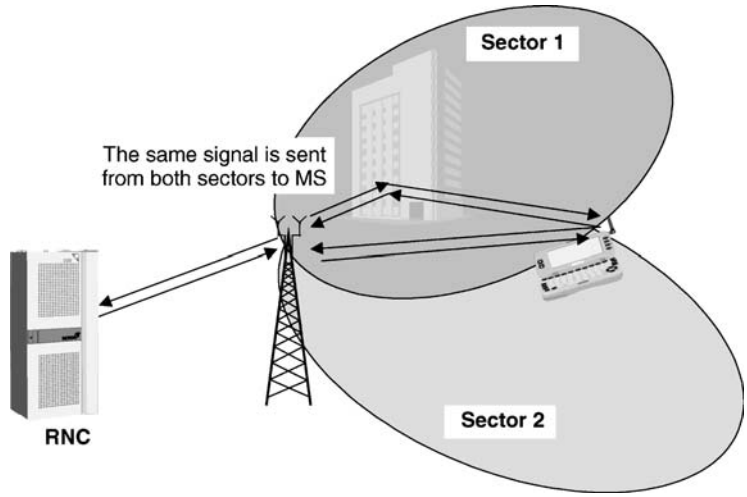



**Figure 3.11** Softer handover

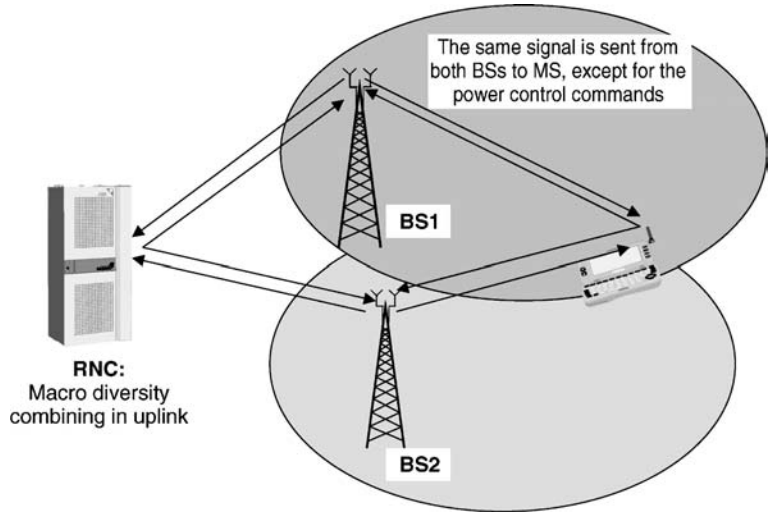



**Figure 3.12** Soft handover

Soft handover occurs in about 20–40% of connections. To cater for soft handover connections, the following additional resources need to be provided by the system and must be considered in the planning phase:

- additional Rake receiver channels in the base stations;
- additional transmission links between base station and RNC;
- additional Rake fingers in the mobile stations.

We also note that soft and softer handover can take place in combination with each other.

Why are these CDMA-specific handover types needed? They are needed for similar reasons as closed loop power control: without soft/softer handover there would be near–far scenarios of a mobile station penetrating from one cell deeply into an adjacent cell without being power-controlled by the latter. Very fast and frequent hard handovers could largely avoid this problem; however, they can be executed only with certain delays during which the near–far problem could develop. So, as with fast power control, soft/softer handovers are an essential interference-mitigating tool in WCDMA. Soft and softer handovers are described in more detail in Section 9.3.

In addition to soft/softer handover, WCDMA provides other handover types:

- Inter-frequency hard handovers that can be used, for example, to hand a mobile over from one WCDMA frequency carrier to another. One application for this is high capacity base stations with several carriers.
- Inter-system hard handovers that take place between the WCDMA FDD system and another system, such as WCDMA TDD or GSM.

## References

[1] Ojanperä, T. and Prasad, R., *Wideband CDMA for Third Generation Mobile Communications*, New York: Artech House, 1998.
[2] Viterbi, A., *Principles of Spread Spectrum Communication*, Reading, MA: Addison-Wesley, 1997.
[3] Cooper, G. and McGillem, C., *Modern Communications and Spread Spectrum*, New York: McGraw-Hill, 1998.
[4] Dixon, R., *Spread Spectrum Systems with Commercial Applications*, Chichester: John Wiley & Sons, 1994.
[5] Jakes, W., *Microwave Mobile Communications*, Chichester: John Wiley & Sons, 1974.
[6] Saunders, S., *Antennas and Propagation for Wireless Communication Systems*, Chichester: John Wiley & Sons, 1999.

# 5

# Radio Access Network Architecture

Fabio Longoni, Atte Länsisalmi and Antti Toskala

## 5.1 Introduction

This chapter gives a wide overview of the Universal Mobile Telephone System (UMTS) architecture, including an introduction to the logical network elements and the interfaces. The UMTS utilizes the same well-known architecture that has been used by all main second-generation systems and even by some first-generation systems. The reference list contains the related 3GPP specifications [1–24].

The UMTS consists of a number of logical network elements that each has a defined functionality. In the standards, network elements are defined at the logical level, but this quite often results in a similar physical implementation, especially since there are a number of open interfaces (for an interface to be 'open', the requirement is that it has been defined to such a detailed level that the equipment at the endpoints can be from two different manufacturers). The network elements can be grouped based on similar functionality, or based on which sub-network they belong to.

Functionally, the network elements are grouped into the Radio Access Network (RAN; UMTS Terrestrial RAN (UTRAN)) that handles all radio-related functionality, and the Core Network (CN), which is responsible for switching and routing calls and data connections to external networks. To complete the system, the User Equipment (UE) that interfaces with the user and the radio interface is defined. The high-level system architecture is shown in Figure 5.1.

From a specification and standardization point of view, both UE and UTRAN consist of completely new protocols, the designs of which are based on the needs of the new WCDMA radio technology. On the contrary, the definition of CN is adopted from GSM. This gives the system with new radio technology a global base of known and rugged CN technology that accelerates and facilitates its introduction, and enables such competitive advantages as global roaming.

Another way to group UMTS network elements is to divide them into sub-networks. The UMTS is modular in the sense that it is possible to have several network elements of the same type. In principle, the minimum requirement for a fully featured and operational network is to have at least one logical network element of each type (note that some features and consequently some network elements are

*WCDMA for UMTS: HSPA Evolution and LTE, Fifth Edition* Edited by Harri Holma and Antti Toskala

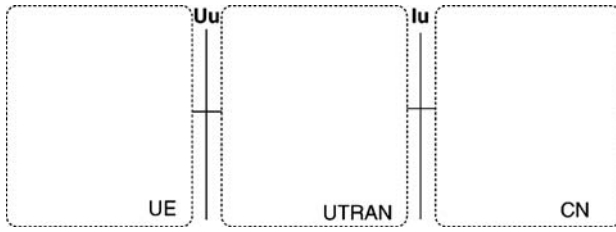



**Figure 5.1** UMTS high-level system architecture

optional). The possibility of having several entities of the same type allows the division of the UMTS into sub-networks that are operational either on their own or together with other sub-networks, and that are distinguished from each other with unique identities. Such a sub-network is called a UMTS Public Land Mobile Network (PLMN). Typically, one PLMN is operated by a single operator, and is connected to other PLMNs as well as to other types of network, such as ISDN, PSTN, the internet, and so on. Figure 5.2 shows elements in a PLMN and, in order to illustrate the connections, also external networks.

The UTRAN architecture is presented in Section 5.2. A short introduction to all the elements is given below. 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, performs authentication algorithms, and stores authentication and encryption keys and some subscription information that is needed at the terminal.

UTRAN also consists of two distinct elements:

1. The Node B converts the data flow between the Iub and Uu interfaces. It also participates in radio resource management. (*Note*: the term 'Node B' from the corresponding 3GPP specifications is used throughout this chapter. The more generic term 'base station' used elsewhere in this book means exactly the same thing.)
2. The Radio Network Controller (RNC) owns and controls the radio resources in its domain (the Node Bs connected to it). The RNC is the service access point for all services that UTRAN provides the CN, e.g. management of connections to the UE.

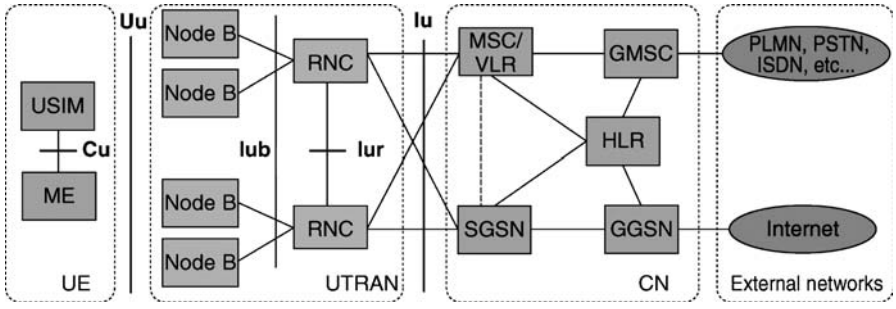



**Figure 5.2** Network elements in a PLMN

The main elements of the GSM CN (there are other entities not shown in Figure 5.2, such as those used to provide IN services) are as follows:

- The Home Location Register (HLR) is a database located in the user's home system that stores the master copy of the user's service profile. The service profile consists of, for example, information on permitted services, forbidden roaming areas, and Supplementary Service information such as the status of call forwarding and the call forwarding number. It is created when a new user subscribes to the system, and remains stored as long as the subscription is active. For the purpose of routing incoming transactions to the UE (e.g. calls or short messages), the HLR also stores the UE location on the level of MSC/VLR and/or SGSN, i.e. on the level of serving system.
- The Mobile Services Switching Center/Visitor Location Register (MSC/VLR) is the switch (MSC) and database (VLR) that serves the UE in its current location for Circuit-Switched (CS) services. The MSC function is used to switch the CS transactions, and the VLR function holds a copy of the visiting user's service profile, as well as more precise information on the UE's location within the serving system. The part of the network that is accessed via the MSC/VLR is often referred to as the CS domain. MSC also has a role in the early UE handling, as discussed in Chapter 7.
- The Gateway MSC (GMSC) is the switch at the point where UMTS PLMN is connected to external CS networks. All incoming and outgoing CS connections go through GMSC.
- The Serving General Packet Radio Service (GPRS) Support Node (SGSN) functionality is similar to that of MSC/VLR but is typically used for Packet-Switched (PS) services. The part of the network that is accessed via the SGSN is often referred to as the PS domain. Similar to MSC, SGSN support is needed for the early UE handling operation, as covered in Chapter 7.
- Gateway GPRS Support Node (GGSN) functionality is close to that of GMSC but is in relation to PS services.

The external networks can be divided into two groups:

1. *CS networks.* These provide circuit-switched connections, like the existing telephony service. ISDN and PSTN are examples of CS networks.
2. *PS networks.* These provide connections for packet data services. The Internet is one example of a PS network.

The UMTS standards are structured so that the internal functionality of the network elements is not specified in detail. Instead, the interfaces between the logical network elements have been defined. The following main open interfaces are specified:

- *Cu interface*. This is the electrical interface between the USIM smartcard and the ME. The interface follows a standard format for smartcards.
- *Uu interface*. This is the WCDMA radio interface, which is the subject of the main part of this book. The Uu is the interface through which the UE accesses the fixed part of the system and, therefore, is probably the most important open interface in UMTS. There are likely to be many more UE manufacturers than manufacturers of fixed network elements.
- *Iu interface*. This connects UTRAN to the CN and is introduced in detail in Section 5.4. Similar to the corresponding interfaces in GSM, A (CS) and Gb (PS), the open Iu interface gives UMTS operators the possibility of acquiring UTRAN and CN from different manufacturers. The permitted competition in this area has been one of the success factors of GSM.
- *Iur interface*. The open Iur interface allows a soft handover between RNCs from different manufacturers and, therefore, complements the open Iu interface. Iur is described in more detail in Section 5.5.1.

- *Iub interface*. The Iub connects a Node B and an RNC. UMTS is the first commercial mobile telephony system where the Controller–Base Station interface is standardized as a fully open interface. Like the other open interfaces, open Iub is expected to motivate further competition between manufacturers in this area. It is likely that new manufacturers concentrating exclusively on Node Bs will enter the market.

## 5.2 UTRAN Architecture

The UTRAN architecture is highlighted in Figure 5.3. UTRAN consists of one or more Radio Network Sub-systems (RNSs). An RNS is a sub-network within UTRAN and consists of one RNC and one or more Node Bs. RNCs may be connected to each other via an Iur interface. RNCs and Node Bs are connected with an Iub Interface. During Release 7, work study on the support of small RNSs was done, meaning use of co-located RNC and Node B functionalities in a flat architecture, and that was found feasible without mandatory specification changes. The flat architecture is further discussed in Chapter 15.

Before entering into a brief description of the UTRAN network elements (in this section) and a more extensive description of UTRAN interfaces (in the following sections), we present the main characteristics of UTRAN that have also been the main requirements for the design of the UTRAN architecture, functions and protocols. These can be summarized in the following points:

- *Support of UTRA* and all the related functionality. In particular, the major impact on the design of UTRAN has been the requirement to support *soft handover* (one terminal connected to the network via two or more active cells) and the WCDMA-specific *Radio Resource Management* algorithms.
- Maximization of the *commonalities in the handling of PS and CS data*, with a unique air interface protocol stack and with use of the same interface for the connection from UTRAN to both the PS and CS domains of the CN.
- Maximization of the *commonalities with GSM*, when possible.
- Use of the *Asynchronous Transfer Mode (ATM) transport* as the main transport mechanism in UTRAN.
- Use of the Internet Protocol (IP)-based transport as the alternative transport mechanism in UTRAN from Release 5 onwards.

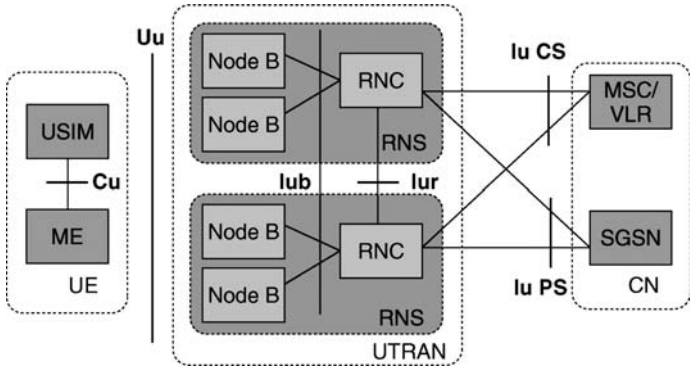



**Figure 5.3** UTRAN architecture

## 5.2.1 The Radio Network Controller (RNC)

The RNC is the network element responsible for the control of the radio resources of UTRAN. It interfaces the CN (normally to one MSC and one SGSN) and also terminates the Radio Resource Control (RRC) protocol that defines the messages and procedures between the mobile and UTRAN. It logically corresponds to the GSM BSC.

### 5.2.1.1 Logical Role of the RNC

The RNC controlling one Node B (i.e. terminating the Iub interface towards the Node B) is indicated as the *Controlling RNC* (CRNC) of the Node B. The CRNC is responsible for the load and congestion control of its own cells, and also executes the admission control and code allocation for new radio links to be established in those cells.

If one mobile–UTRAN connection uses resources from more than one RNS (see Figure 5.4), the RNCs involved have two separate logical roles (with respect to this mobile–UTRAN connection):

- *Serving RNC (SRNC)*. The SRNC for one mobile is the RNC that terminates both the Iu link for the transport of user data and the corresponding RAN application part (RANAP) signaling to/from the CN (this connection is referred to as the RANAP connection). The SRNC also terminates the RRCl Signaling, i.e. the signaling protocol between the UE and UTRAN. It performs the L2 processing of the data to/from the radio interface. Basic Radio Resource Management operations, such as the mapping of Radio Access Bearer (RAB) parameters into air interface transport channel parameters, the handover decision, and outer loop power control, are executed in the SRNC. The SRNC may also (but not always) be the CRNC of some Node B used by the mobile for connection with UTRAN. One UE connected to UTRAN has one and only one SRNC.
- *Drift RNC (DRNC)*. The DRNC is any RNC, other than the SRNC, that controls cells used by the mobile. If needed, the DRNC may perform macrodiversity combining and splitting. The DRNC does not perform L2 processing of the user plane data, but routes the data transparently between the Iub and Iur interfaces, except when the UE is using a common or shared transport channel. One UE may have zero, one or more DRNCs.

Note that one physical RNC normally contains all the CRNC, SRNC and DRNC functionality.

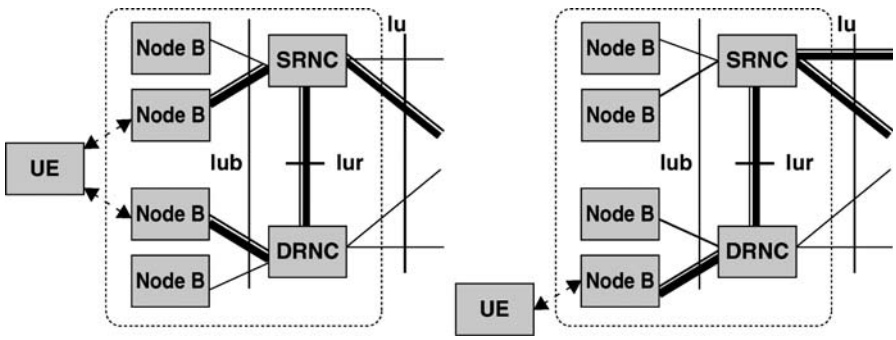



**Figure 5.4** Logical role of the RNC for one UE UTRAN connection. The left-hand scenario shows one UE in inter-RNC soft handover (combining is performed in the SRNC). The right-hand scenario represents one UE using resources from one Node B only, controlled by the DRNC

### 5.2.2 The Node B (Base Station)

The main function of the Node B is to perform the air interface L1 processing (channel coding and interleaving, rate adaptation, spreading, etc.). It also performs some basic Radio Resource Management operation as the inner loop power control. It logically corresponds to the GSM Base Station. The enigmatic term 'Node B' was initially adopted as a temporary term during the standardization process, but then never changed. The logical model of the Node B is described in Section 5.5.2.

## 5.3 General Protocol Model for UTRAN Terrestrial Interfaces

### 5.3.1 General

Protocol structures in UTRAN terrestrial interfaces are designed according to the same general protocol model. This model is shown in Figure 5.5. The structure is based on the principle that the layers and planes are logically independent of each other and, if needed, parts of the protocol structure may be changed in the future while other parts remain intact.

### 5.3.2 Horizontal Layers

The protocol structure consists of two main layers: the Radio Network Layer and the Transport Network Layer. All UTRAN-related issues are visible only in the Radio Network Layer, and the Transport Network Layer represents standard transport technology that is selected to be used for UTRAN but without any UTRAN-specific changes.

### 5.3.3 Vertical Planes

#### 5.3.3.1 Control Plane

The Control Plane is used for all UMTS-specific control signaling. It includes the Application Protocol (i.e. RANAP in Iu, Radio Network System Application Part (RNSAP) in Iur and Node B Application Part (NBAP) in Iub), and the Signaling Bearer to transport the Application Protocol messages.

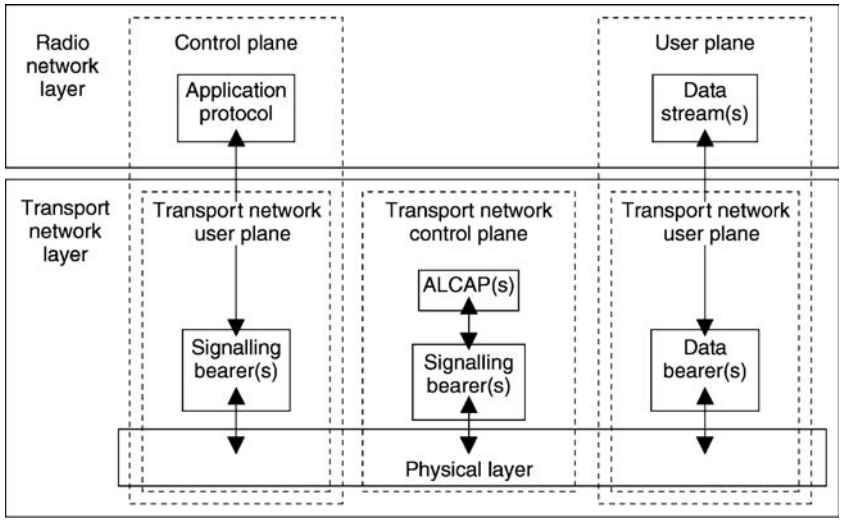



**Figure 5.5** General protocol model for UTRAN terrestrial interfaces

The Application Protocol is used, among other things, for setting up bearers to the UE (i.e. the RAB in Iu and subsequently the Radio Link in Iur and Iub). In the three-plane structure the bearer parameters in the Application Protocol are not directly tied to the User Plane technology, but rather are general bearer parameters.

The Signaling Bearer for the Application Protocol may or may not be of the same type as the Signaling Bearer for the ALCAP. It is always set up by operation and maintenance (O&M) actions.

#### 5.3.3.2 User Plane

All information sent and received by the user, such as the coded voice in a voice call or the packets in an Internet connection, are transported via the User Plane. The User Plane includes the Data Stream(s), and the Data Bearer(s) for the Data Stream(s). Each Data Stream is characterized by one or more frame protocols specified for that interface.

#### 5.3.3.3 Transport Network Control Plane

The Transport Network Control Plane is used for all control signaling within the Transport Layer. It does not include any Radio Network Layer information. It includes the ALCAP protocol that is needed to set up the transport bearers (Data Bearer) for the User Plane. It also includes the Signaling Bearer needed for the ALCAP.

The Transport Network Control Plane is a plane that acts between the Control Plane and the User Plane. The introduction of the Transport Network Control Plane makes it possible for the Application Protocol in the Radio Network Control Plane to be completely independent of the technology selected for the Data Bearer in the User Plane.

When the Transport Network Control Plane is used, the transport bearers for the Data Bearer in the User Plane are set up in the following fashion. First, there is a signaling transaction by the Application Protocol in the Control Plane, which triggers the se-up of the Data Bearer by the ALCAP protocol that is specific for the User Plane technology.

The independence of the Control Plane and the User Plane assumes that an ALCAP signaling transaction takes place. It should be noted that ALCAP might not be used for all types of Data Bearers. If there is no ALCAP signaling transaction, then the Transport Network Control Plane is not needed at all. This is the case when it is enough simply to select the user plane resources, e.g. selecting end-point addresses for IP transport or selecting a preconfigured Data Bearer. It should also be noted that the ALCAP protocols in the Transport Network Control Plane are not used for setting up the Signaling Bearer for the Application Protocol or for the ALCAP during real-time operation.

The Signaling Bearer for the ALCAP may or may not be of the same type as that for the Application Protocol. The UMTS specifications assume that the Signaling Bearer for ALCAP is always set up by O&M actions, and do not specify this in detail.

#### 5.3.3.4 Transport Network User Plane

The Data Bearers in the User Plane and the Signaling Bearers for the Application Protocol also belong to the Transport Network User Plane. As described in the previous section, the Data Bearers in the Transport Network User Plane are directly controlled by the Transport Network Control Plane during real-time operation, but the control actions required to set up the Signaling Bearer(s) for the Application Protocol are considered O&M actions.

## 5.4 Iu, the UTRAN–CN Interface

The Iu interface connects UTRAN to CN. Iu is an open interface that divides the system into radio-specific UTRAN and CN, which handles switching, routing and service control. As can be seen from Figure 5.3, the Iu can have two main different instances, which are Iu CS to connect UTRAN to CS

CN, and Iu PS to connect UTRAN to PS CN. The additional third instance of Iu, the Iu Broadcast (Iu BC, not shown in Figure 5.3), has been defined to support Cell Broadcast Services (see Section 5.4.5). The original design goal in the standardization was to develop only one Iu interface, but then it was realized that fully optimized User Plane transport for CS and PS services can only be achieved if different transport technologies are permitted. Consequently, the Transport Network Control Plane is different. One of the main design guidelines has still been that the Control Plane should be the same for Iu CS and Iu PS, and the differences are minor.

A third instance of the Iu interface, Iu BC, is used to connect UTRAN to the Broadcast domain of the CN. The Iu BC interface is not shown in Figure 5.3.

## *5.4.1 Protocol Structure for Iu CS*

The Iu CS overall protocol structure is depicted in Figure 5.6. The three planes in the Iu interface share a common ATM transport which is used for all planes. The physical layer is the interface to the physical medium: optical fibre, radio link or copper cable. The physical layer implementation can be selected from a variety of standard off-the-shelf transmission technologies, such as SONET, STM1, or E1.

### 5.4.1.1 Iu CS Control Plane Protocol Stack

The Control Plane protocol stack consists of RANAP, on top of Broad Band (BB) Signaling System #7 (SS7) protocols. The applicable layers are the Signaling Connection Control Part (SCCP), the Message Transfer Part (MTP3-b) and Signaling ATM Adaptation Layer for Network to Network

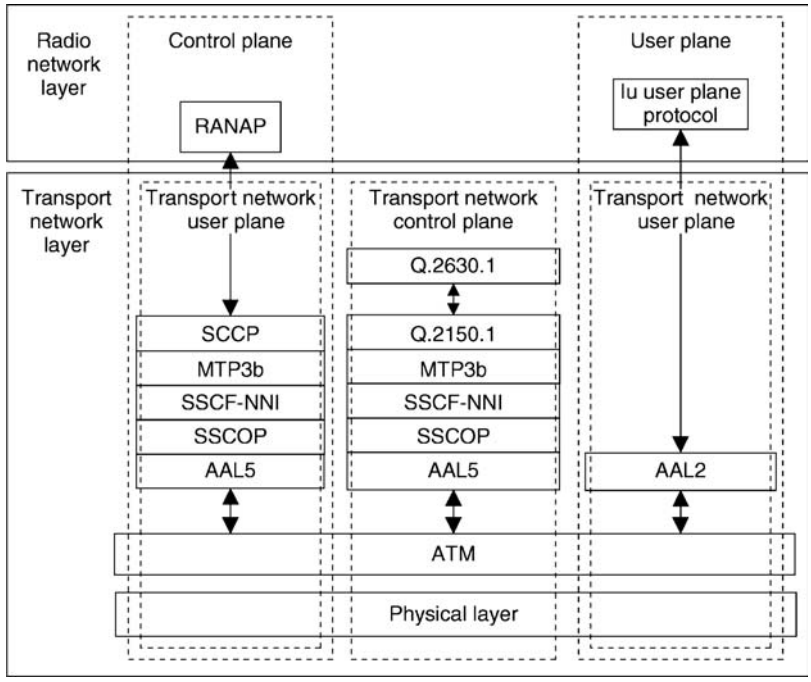



**Figure 5.6** Iu CS protocol structure

Interfaces (SAAL-NNI). SAAL-NNI is further divided into Service Specific Co-ordination Function (SSCF), Service Specific Connection Oriented Protocol (SSCOP) and ATM Adaptation Layer 5 (AAL) layers. SSCF and SSCOP layers are specifically designed for signaling transport in ATM networks, and take care of such functions as signaling connection management. AAL5 is used to segment the data into ATM cells.

#### 5.4.1.2 Iu CS Transport Network Control Plane Protocol Stack

The Transport Network Contro Plane protocol stack consists of the Signaling Protocol for setting up AAL2 connections (Q.2630.1 and adaptation layer Q.2150.1), on top of BB SS7 protocols. The applicable BB SS7 protocols are those described above without the SCCP layer.

#### 5.4.1.3 Iu CS User Plane Protocol Stack

A dedicated AAL2 connection is reserved for each individual CS service. The Iu User Plane Protocol residing directly on top of AAL2 is described in more detail in Section 5.4.4.

## *5.4.2 Protocol Structure for Iu PS*

The Iu PS protocol structure is depicted in Figure 5.7. Again, a common ATM transport is applied for both the User and the Control Planes. Also, the physical layer is as specified for Iu CS.

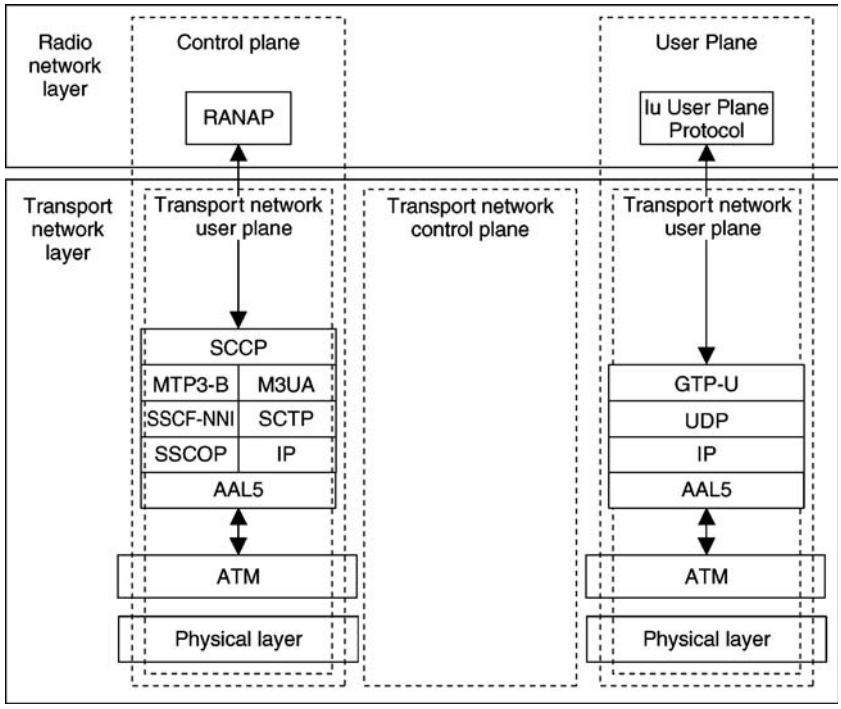



**Figure 5.7** Iu PS protocol structure

#### 5.4.2.1 Iu PS Control Plane Protocol Stack

The Control Plane protocol stack again consists of RANAP, and the same BB SS7-based signaling bearer as described in Section 5.4.1.1. Also as an alternative, an IP-based signaling bearer is specified. The SCCP layer is also used commonly for both. The IP-based signaling bearer consists of SS7 MTP3–User Adaptation Layer (M3UA), Simple Control Transmission Protocol (SCTP), IP, and AAL5 which is common to both alternatives. The SCTP layer is specifically designed for signaling transport in the Internet. Specific adaptation layers are specified for different kinds of signaling protocols, such as M3UA for SS7-based signaling.

#### 5.4.2.2 Iu PS Transport Network Control Plane Protocol Stack

The Transport Network Control Plane is not applied to Iu PS. The setting up of the GTP tunnel requires only an identifier for the tunnel, and the IP addresses for both directions, and these are already included in the RANAP RAB Assignment messages. The same information elements that are used in Iu CS for addressing and identifying the AAL2 signaling are used for the User Plane data in Iu CS.

#### 5.4.2.3 Iu PS User Plane Protocol Stack

In the Iu PS User Plane, multiple packet data flows are multiplexed on one or several AAL5 Pre-defined Virtual Connections. The User Plane part of the GPRS Tunnelling Protocol (GTP-U) is the multiplexing layer that provides identities for individual packet data flow. Each flow uses UDP connectionless transport and IP addressing.

### *5.4.3 RANAP Protocol*

RANAP is the signaling protocol in Iu that contains all the control information specified for the Radio Network Layer. The functionality of RANAP is implemented by various RANAP Elementary Procedures (EPs). Each RANAP function may require the execution of one or more EPs. Each EP consists of either just the request message (class 2 EP), the request and response message pair (class 1 EP), or one request message and one or more response messages (class 3 EP). The following RANAP functions are defined:

- *Relocation*. This function handles both SRNS relocation and hard handover, including intersystem case to/from GSM:
- *SRNS relocation*. The SRNS functionality is relocated from one RNS to another without changing the radio resources and without interrupting the user data flow. The prerequisite for SRNS relocation is that all Radio Links are already in the same DRNC that is the target for the relocation.
- *Inter-RNS hard handover*. This is used to relocate the serving RNS functionality from one RNS to another and to change the radio resources correspondingly by a hard handover in the Uu interface. The prerequisite for hard handover is that the UE is at the border of the source and target cells.
- *RAB management*. This function combines all RAB handling:
  - RAB set-up, including the possibility for queuing the setup;
  - modification of the characteristics of an existing RAB;
  - clearing an existing RAB, including the RAN-initiated case.
- *Iu release*. Releases all resources (Signaling link and U-plane) from a given instance of Iu related to the specified UE. Also includes the RAN-initiated case.
- *Reporting unsuccessfully transmitted data*. This function allows the CN to update its charging records with information from UTRAN if part of the data sent was not successfully sent to the UE.

- *Common ID management*. In this function the permanent identification of the UE is sent from the CN to UTRAN to allow paging coordination from possibly two different CN domains.
- *Paging*. This is used by CN to page an idle UE for a UE terminating service request, such as a voice call. A paging message is sent from the CN to UTRAN with the UE common identification (permanent ID) and the paging area. UTRAN will either use an existing signaling connection, if one exists, to send the page to the UE or broadcast the paging in the requested area.
- *Management of tracing*. The CN may, for O&M purposes, request UTRAN to start recording all activity related to a specific UE–UTRAN connection.
- *UE–CN signaling transfer*. This functionality provides transparent transfer of UE–CN signaling messages that are not interpreted by UTRAN in two cases.
- *Transfer of the first UE message from UTRAN to UE*: this may be, for example, a response to paging, a request of a UE-originated call, or just registration to a new area. It also initiates the signaling connection for the Iu.
- *Direct transfer*: used for carrying all consecutive signaling messages over the Iu signaling connection in both the uplink and downlink directions.
- *Security rode control*. This is used to set the ciphering or integrity checking on or off. When ciphering is on, the signaling and user data connections in the radio interface are encrypted with a secret key algorithm. When integrity checking is on, an integrity checksum, further secured with a secret key, is added to some or all of the radio interface signaling messages. This ensures that the communication partner has not changed, and the content of the information has not been altered.
- *Management of overload*. This is used to control the load over the Iu interface against overload due, for example, to processor overload at the CN or UTRAN. A simple mechanism is applied that allows stepwise reduction of the load and its stepwise resumption, triggered by a timer.
- *Reset*. This is used to reset the CN or the UTRAN side of the Iu interface in error situations. One end of the Iu may indicate to the other end that it is recovering from a restart, and the other end can remove all previously established connections.
- *Location reporting*. This functionality allows the CN to receive information on the location of a given UE. It includes two elementary procedures, one for controlling the location reporting in the RNC and the other to send the actual report to the CN.

### *5.4.4 Iu User Plane Protocol*

The Iu User Plane protocol is in the Radio Network Layer of the Iu User Plane. It has been defined so that it would be, as much as possible, independent of the CN domain that it is used for. The purpose of the User Plane protocol is to carry user data related to RABs over the Iu interface. Each RAB has its own instance of the protocol. The protocol performs either a fully transparent operation, or framing for the user data segments and some basic control signaling to be used for initialization and online control. Based on these cases, the protocol has two modes:

1. *Transparent mode*. In this mode of operation the protocol does not perform any framing or control. It is applied for RABs that do not require such features but that assume fully transparent operation.
2. *Support mode for predefined (service data unit) SDU sizes*. In this mode the User Plane performs framing of the user data into segments of predefined size. The SDU sizes typically correspond to Adaptive Multirate Codec speech frames, or to the frame sizes derived from the data rate of a CS data call. Also, control procedures for initialization and rate control are defined, and a functionality is specified to indicate the quality of the frame based, for example, on a cyclic redundancy check from the radio interface.

## 5.4.5 *Protocol Structure of Iu BC, and the Service Area Broadcast Protocol*

The Iu BC [2] interface connects the RNC in UTRAN with the broadcast domain of the CN, namely with the Cell Broadcast Center. It is used to define the Cell Broadcast information that is transmitted to the mobile user via the Cell Broadcast Service (e.g. name of city/region visualized on the mobile phone display). Note that this should not be confused with the UTRAN or CN information broadcast on the broadcast common control channel. Iu BC is a control-plane-only interface. The protocol structure of Iu BC is shown in Figure 5.8.

### 5.4.5.1 Service Area Broadcast Protocol

The Service Area Broadcast protocol (SABP) [23] provides the capability for the Cell Broadcast Center in the CN to define, modify and remove cell broadcast messages from the RNC. RNC uses them and the NBAP protocol and RRC signaling to transfer such messages to the mobile. The SABP has the following functions:

- *Message handling*. This function is responsible for the broadcast of new messages, amending existing broadcasted messages and stopping the broadcasting of specific messages.

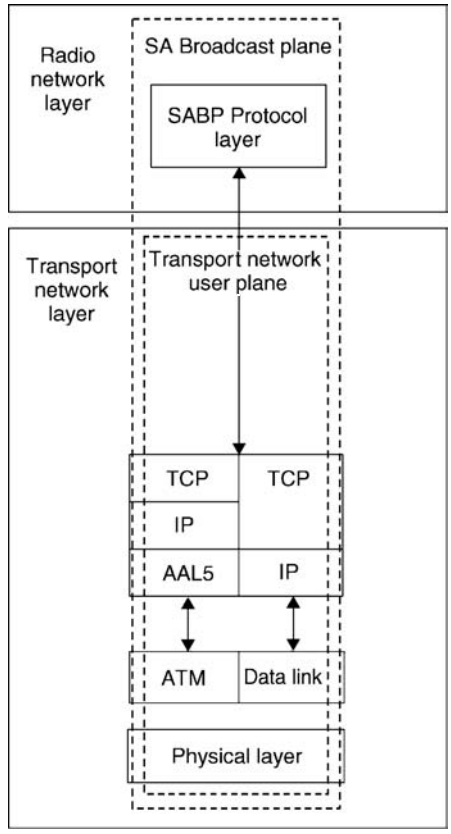



**Figure 5.8** Iu BC protocol structure

- *Load handling*. This function is responsible for determining the loading of the broadcast channels at any particular point in time.
- *Reset*. This function permits the Cell Broadcast Center to end broadcasting in one or more Service Areas.

## 5.5 UTRAN Internal Interfaces

### *5.5.1 RNC–RNC Interface (Iur Interface) and the RNSAP Signaling*

The protocol stack of the RNC-to-RNC interface (Iur interface) is shown in Figure 5.9. Although this interface was initially designed in order to support the inter-RNC soft handover (shown on the left-hand side of Figure 5.4), more features were added during the development of the standard and, currently, the Iur interface provides four distinct functions:

1. Support of basic inter-RNC mobility.
2. Support of dedicated channel traffic.
3. Support of common channel traffic.
4. Support of global resource management.

For this reason, the Iur signaling protocol itself (RNSAP) is divided into four different *modules* (to be intended as groups of procedures). In general, it is possible to implement only part of the four Iur modules between two RNCs, according to the operator's need.

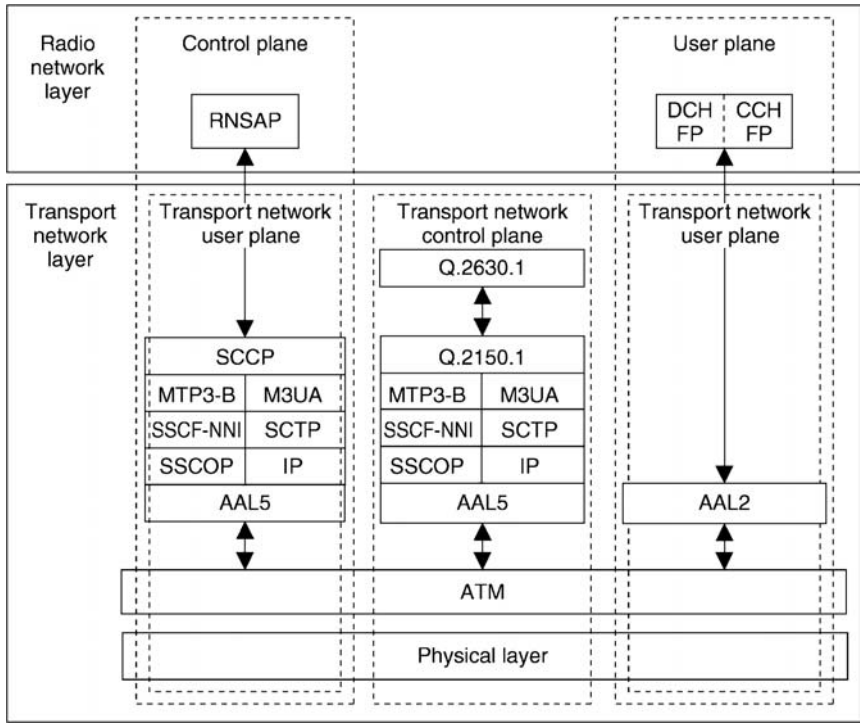



**Figure 5.9** Release 99 protocol stack of the Iur interface. As for the Iu interface, two options are possible for the transport of the RNSAP signaling: the SS7 stack (SCCP and MTP3b) and the new SCTP/IP-based transport. Two User Plane protocols are defined (DCH: dedicated channel; CCH: common channel)

#### 5.5.1.1 Iur1: Support of the Basic Inter-RNC Mobility

This functionality requires the *basic* module of RNSAP signaling as described in [11]. This first brick for the construction of the Iur interfaces provides on its own the functionality needed for the mobility of the user between the two RNCs, but does not support the exchange of any user data traffic. If this module is not implemented, then the Iur interface as such does not exist, and the only way for a user connected to UTRAN via the RNS1 to utilize a cell in RNS2 is to disconnect itself temporarily from UTRAN (release the RRC connection).

The functions offered by the Iur basic module include:

- Support of SRNC relocation.
- Support of inter-RNC cell and UTRAN registration area update.
- Support of inter-RNC packet paging.
- Reporting of protocol errors.

Since this functionality does not involve user data traffic across Iur, the User Plane and the Transport Network Control Plane protocols are not needed.

#### 5.5.1.2 Iur2: Support of Dedicated Channel Traffic

This functionality requires the Dedicated Channel module of RNSAP signaling and allows the dedicated and shared channel traffic between two RNCs. Even if the initial need for this functionality is to support the inter-RNC soft handover state, it also allows the anchoring of the SRNC for all the time the user is utilizing dedicated channels (dedicated resources in the Node B), commonly for as long as the user has an active connection to the CS domain.

This functionality requires also the User Plane Frame Protocol for the dedicated and shared channel, plus the Transport Network Control Plane protocol (Q.2630.1) used for the set-up of the transport connections (AAL2 connections). Each dedicated channel is conveyed over one transport connection, except the coordinated DCH used to obtain unequal error protection in the air interface.

The Frame Protocol for dedicated channels, in short DCH FP [15], defines the structure of the data frames carrying the user data and the control frames used to exchange measurements and control information. For this reason, the Frame Protocol also specifies simple messages and procedures. The user data frames are normally routed transparently through the DRNC; thus, the Iur frame protocol is used also in Iub and referred to as Iur/Iub DCH FP. The user plane procedure for shared channels is described in the Frame Protocol for a common channel in an Iur interface, in short, Iur CCH FP [13].

The functions offered by the Iur DCH module are:

- Establishment, modification and release of the dedicated and shared channel in the DRNC due to handovers in the dedicated channel state.
- Set-up and release of dedicated transport connections across the Iur interface.
- Transfer of DCH Transport Blocks between SRNC and DRNC.
- Management of the radio links in the DRNS, via dedicated measurement report procedures, power setting procedures and compress mode control procedures.

#### 5.5.1.3 Iur3: Support of Common Channel Traffic

This functionality allows the handling of common channel (i.e. RACH, FACH and CPCH) data streams across the Iur interface. It requires the Common Transport Channel module of the RNSAP protocol and the Iur Common Transport Channel Frame Protocol (in short, CCH FP). The Q.2630.1 signaling protocol of the Transport Network Control Plane is also needed if signaled AAL2 connections are used.

If this functionality is not implemented, then every inter-RNC cell update always triggers an SRNC relocation, i.e. the serving RNC is always the RNC controlling the cell used for common or shared channel transport.

The identification of the benefits of this feature caused a long debate in the relevant standardization body. On the one hand, this feature allows the implementation of the total anchor RNC concept, avoiding the SRNC relocation procedure (via the CN); on the other hand, it requires the splitting of the Medium Access Control (MAC) layer functionality into two network elements, generating inefficiency in the utilization of the resources and complexity in the Iur interface. The debate could not reach an agreement; thus, the feature is supported by the standard but is not essential for the operation of the system.

The functions offered by the Iur common transport channel module are:

- Set-up and release of the transport connection across the Iur for common channel data streams.
- Splitting of the MAC layer between the SRNC (MAC-d) and the DRNC (MAC-c). The scheduling for downlink data transmission is performed in the DRNC.
- Flow control between the MAC-d and MAC-c.

#### 5.5.1.4 Iur4: Support of Global Resource Management

This functionality provides signaling to support enhanced radio resource management and O&M features across the Iur interface. It is implemented via the global module of the RNSAP protocol, and does not require any User Plane protocol, since there is no transmission of user data across the Iur interface. The function is considered optional. This function has been introduced in subsequent releases for the support of common radio resource management between RNCs, advanced positioning methods and Iur optimization purposes.

The functions offered by the Iur global resource module are:

- Transfer of cell information and measurements between two RNCs.
- Transfer of positioning parameters between controllers.
- Transfer of Node B timing information between two RNCs.

### *5.5.2 RNC–Node B Interface and the NBAP Signaling*

The protocol stack of the RNC–Node B interface (Iub interface) is shown, with the typical triple plane notation, in Figure 5.10. In order to understand the structure of the interface, it is necessary briefly to introduce the logical model of the Node B, depicted in Figure 5.11. This consists of a common control port (a common signaling link) and a set of traffic termination points each controlled by a dedicated control port (dedicated signaling link). One traffic termination point controls a number of mobiles having dedicated resources in the Node B, and the corresponding traffic is conveyed through dedicated data ports. Common data ports outside the traffic termination points are used to convey RACH, FACH and PCH traffic.

Note that there is no relation between the traffic termination point and the cells, i.e. one traffic termination point can control more than one cell, and one cell can be controlled by more than one traffic termination point.

The Iub interface signaling (NBAP) is divided into two essential components: the common NBAP (C-NBAP), which defines the signaling procedures across the common signaling link, and the dedicated NBAP (D-NBAP), used in the dedicated signaling link.

The User Plane Iub frame protocols define the structures of the frames and the basic in-band control procedures for every type of transport channel (i.e. for every type of data port of the model). The Q.2630.1 signaling is used for the dynamic management of the AAL2 connections used in the User Plane.

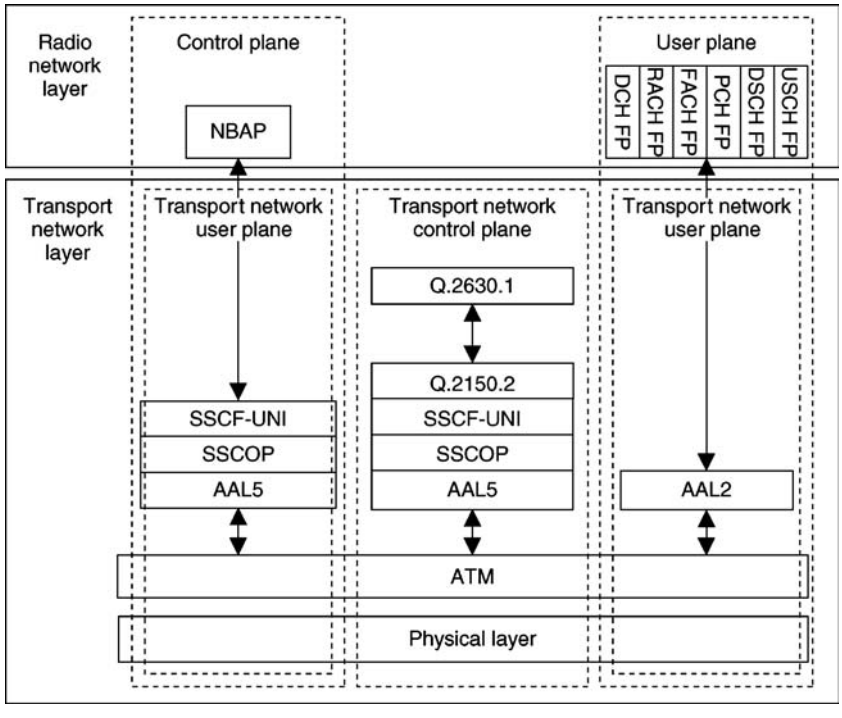



**Figure 5.10** Release 99 protocol stack of the Iub interface. This is similar to the Iur interface protocol, the main difference being that in the Radio Network and Transport Network Control Planes the SS7 stack is replaced by the simpler SAAL-UNI as signaling bearer. Note also that the SCTP/IP option is not present here

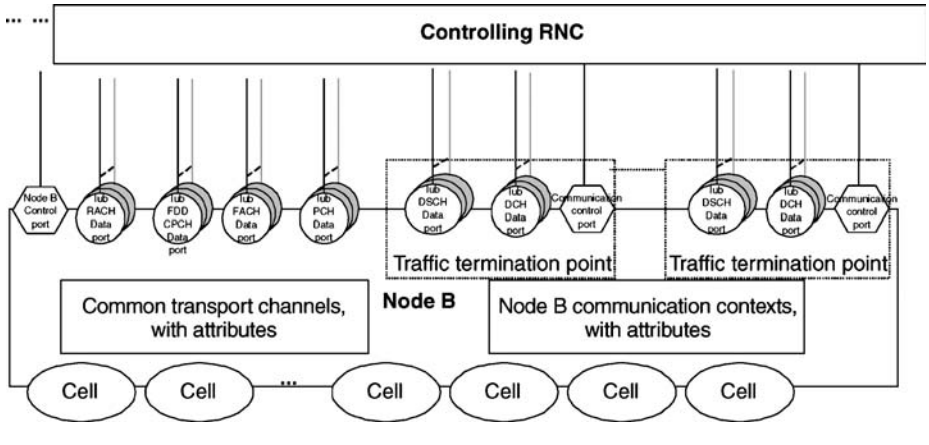



**Figure 5.11** Logical model of the Node B for FDD

#### 5.5.2.1 C-NBAP and the Logical O&M

The C-NBAP procedures are used for the signaling that is not related to one specific UE context already existing in the Node B. In particular, the C-NBAP defines all the procedures for the logical O&M of the Node B, such as configuration and fault management.

The main functions of the C-NBAP are:

- Set-up of the first radio link of one UE, and selection of the traffic termination point.
- Cell configuration.
- Handling of the RACH/FACH/CPCH and PCH channels.
- Initialization and reporting of Cell or Node B specific measurement.
- Location Measurement Unit (LMU) control.
- Fault management.

#### 5.5.2.2 Dedicated NBAP

When the RNC requests the first radio link for one UE via the C-NBAP Radio Link Set-up procedure, the Node B assigns a traffic termination point for the handling of this UE context, and every subsequent signaling related to this mobile is exchanged with D-NBAP procedures across the dedicated control port of the given Traffic Termination Point.

The main functions of the D-NBAP are:

- Addition, release and reconfiguration of radio links for one UE context.
- Handling of dedicated and shared channels.
- Handling of softer combining.
- Initialization and reporting of radio-link-specific measurement.
- Radio-link fault management.

## 5.6 UTRAN Enhancements and Evolution

The Release 99 UTRAN architecture described in Chapter 4 defines the basic set of network elements and interface protocols for the support of the Release 99 WCDMA radio interface. Since then, enhancement of the architecture and related specification were needed in order to support new WCDMA radio interface features, but also as result of the necessity to provide a more efficient, scalable and robust 3GPP system architecture. The four most significant additions to the UTRAN architecture introduced in Release 5 are described in subsequent sections.

### *5.6.1 IP Transport in UTRAN*

ATM is the transport technology used in the first release of the UTRAN. Even before the completion of the specification it was clear that 3GPP could not stay immune from the increasing popularity of IP technology, and a second option for the transport, 'IP transport', was introduced in the specification in Release 5. Accordingly, user plane FP frames can also be conveyed over UDP/IP protocols on Iur/Iub, and over RTP/UDP/IP protocols in an Iu CS interface, in addition to the initially defined option of AAL2/ATM. A second option for the Iub control plane, using SCTP directly below the application part, is also introduced. The protocols to be used to convey IP frames are, in general, left unspecified in order not to limit the use of layer 2 and physical layer interfaces available in the operator networks. Although the IP transport requires small changes in the specification (and almost none in the control plane application parts), the adoption of IP technology is a relevant step for both the operator and the

vendors, changing the way the network itself is managed and, in some cases, the way the network elements are implemented.

### 5.6.2 Iu Flex

The Release 99 architecture presented in Figure 5.3 is characterized by having only one MSC and one SGSN connected to the RNC, i.e. only one Iu PS and Iu CS interface in the RNC. This limitation is overcome in the Release 5 specification with the introduction of the Iu flex (from the word 'flexible') concept, which allows one RNC to have more than one Iu PS and Iu CS interface instances with the core. The main benefits of this feature are to introduce the possibility of load sharing between the CN nodes and to increase the possibility of anchoring the MSC and SGSN in case of SRNS relocation. Iu flex has limited impact in the UTRAN specification, since the CN node to be used is negotiated between the UE and the CN.

### 5.6.3 Stand-Alone SMLC and Iupc Interface

Location-based services are expected to be a very important source of revenue for the mobile operators, and a number of different applications are expected to be available and largely used. Following the example of the GSM BSS, the UTRAN architecture also includes a stand-alone Serving Mobile Location Center (stand-alone SMLC, or, simply, SAS), which is a new network element for the handling of positioning measurements and the calculation of the mobile station position. The SAS is connected to the RNC via the Iupc interface and the Positioning Calculation Application Part (PCAP) is the L3 protocol used for the RNC-SAS signaling. Stand-alone SMLC and Iu PC interface are optional elements, since SMLC functionality can be integrated in the RNC as well; thus, it depends on the individual network implementation whether to use it or not. The first version of the Iu PC supported only Assisted GPS, but then for later versions support for other positioning methods was added. The latest methods being included in Release 7 specifications are Assisted Galileo and Uplink TDOA methods.

### 5.6.4 Interworking between GERAN and UTRAN, and the Iur-g Interface

The Iu interface has also been scheduled to be part of the GSM/EDGE Radio Access Network (GERAN) in GERAN Release 5. This allows reuse of the 3G CN also for the GSM/EDGE radio interface (and frequency band), but also allows more optimized interworking between the two radio technologies. As an effect of this, the RNSAP basic mobility module (described in Section 5.5.1.1) is enhanced to allow also the mobility to and from GERAN cells in the target and the source, and the RNSAP global module (see Section 5.5.1.4) is enhanced in order to allow the GERAN cells' measurements to be exchanged between controllers. The last feature allows a Common Radio Resource Management (CRRM) between UTRAN and GERAN radios. The term Iur-g interface is often used to refer to the above-mentioned set of Iur functionalities that are utilized also by the GERAN.

### 5.6.5 IP-Based RAN Architecture

The increasing role of IP technology in modern telecoms and IT networks has already been mentioned in the previous sections to motivate the introduction of the IP Transport option in UTRAN. We will see in the next section how the need to provide optimized support IP services leads to sensible changes in the architecture of the CN with the introduction of a new subsystem, the IP Multimedia Subsystem (IMS), to form what is now commonly referred to as the All IP CN. Is the introduction

of IP Transport in UTRAN enough to provide the most suitable RAN architecture to be implemented with IP technology, integrated with the always more commonly used IP networks and platforms, and utilized by IP packet services? In 3GPP, work has been done to investigate new architecture alternatives aiming for a more distributed operation from a centralized network structure, with the motivation to achieve a flat architecture similar to the case of LTE with good scalability for handling increased data rates. The developments from Release 99 with features like HSDPA in Release 5 (Chapter 12) or HSUPA in Release 6 (Chapter 13) have taken the first steps in taking the scheduling and retransmissions (MAC layer) closer to the air interface to improve system performance.

The term All IP RAN is nowadays often used to refer to this IP-Optimized RAN architecture concept and implementation, but is currently not yet associated with any 3GPP standard feature. For this reason, this term is sometimes used to refer to a RAN implementation based on the current architecture but using IP Transport. The flat architecture having all radio-related protocols terminated in the Node B site (with RNC functionality co-located with Node B) is discussed further in Chapter 15.

## 5.7 UMTS CN Architecture and Evolution

While the UMTS radio interface, WCDMA, represented a bigger step in the radio access evolution from GSM networks, the UMTS CN did not experience major changes in the 3GPP Release 99 specification. The Release 99 structure was inherited from the GSM CN and, as stated also earlier, both UTRAN- and GERAN-based RANs connect to the same CN.

### *5.7.1 Release 99 CN Elements*

The Release 99 CN has two domains, a CS domain and a PS domain, to cover the need for different traffic types. The division comes from the different requirements of the data, depending on whether it is real time (circuit switched) or non-real time (packet data). We now present the functional split in the CN side; however, it should be understood that several functionalities can be implemented in a single physical entity and all entities do not necessarily exist as separate physical units in real networks. Figure 5.12 illustrates the Release 99 CN structure with both CS and PS domains shown. Figure 5.12

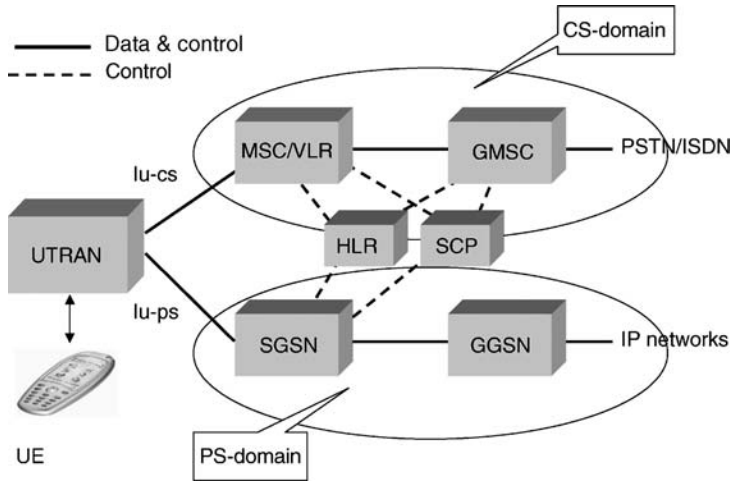



**Figure 5.12** Release 99 UMTS CN structure

also contains registers as well as the Service Control Point (SCP) to indicate the link for providing a particular service to the end user.

The CS domain has the following elements as introduced in Section 5.1:

- MSC, including VLR
- GMSC.

The PS domain has the following elements as introduced in Section 5.1:

- SGSN, which covers similar functions to the MCS for the packet data, including VLR-type functionality.
- GGSN connects PS CN to other networks, for example, to the internet.

In addition to the two domains, the network needs various registers for proper operation:

- HLR with the functionality as covered in Section 5.1.
- Equipment Identity Register contains the information related to the terminal equipment and can be used, for example, to prevent a specific terminal from accessing the network.

### *5.7.2 Release 5 CN and IP Multimedia Subsystem*

The Release 5 CN has many additions compared with Release 99 CNs. Release 4 already included the change in the CN CS domain when the MSC was divided into the MSC server and the Media Gateway (MGW). Also, the GMSC was divided into the GMSC server and the MGW. Release 5 contains the first phase of the IMS, which will enable a standardized approach for IP-based service provision via the PS domain as discussed in Chapter 2. The capabilities of the IMS will be further enhanced in Release 6. The Release 6 IMS will allow the provision of services similar to the CS domain services from the PS domain. The following summarizes the elements in the Release 5-based architecture, added on top of Release 99 and Release 4 architectures. The Release 5 architecture is presented in Figure 5.13, with the simplification that the registers, now part of Home Subscriber Server (HSS), are shown only as independent items without all the connections to the other elements shown.

From a protocols perspective, the key protocol between the terminal and the IMS is the Session Initiation Protocol SIP, which is the basis for IMS-related signaling, with the contents as described in Chapter 2.

The following elements have experienced changes in the CS domain for Release 4:

- The MSC or GMSC server takes care of the control functionality as MSC or GMSC respectively, but the user data goes via the MGW. One MSC/GMSC server can control multiple MGWs, which allows better scalability of the network when, for example, the data rates increase with new data services. In that case, only the number of MGWs needs to be increased.
- The MGW performs the actual switching for user data and network interworking processing, e.g. echo cancellation or speech decoding/encoding.

In the PS domain, the SGSN and GGSN are as in Release 99 with some enhancements, but for the IP-based service delivery the IMS now has the following key elements included:

- Media Resource Function (MRF), which, for example, controls media stream resources or can mix different media streams. The standard defines further the detailed functional split for the MRF.
- Call Session Control Function (CSCF), which acts as the first contact point to the terminal in the IMS (as a proxy). The CSCF covers several functionalities, from handling of the session states to being a contact point for all IMS connections intended for a single user and acting as a firewall towards other operator's networks.

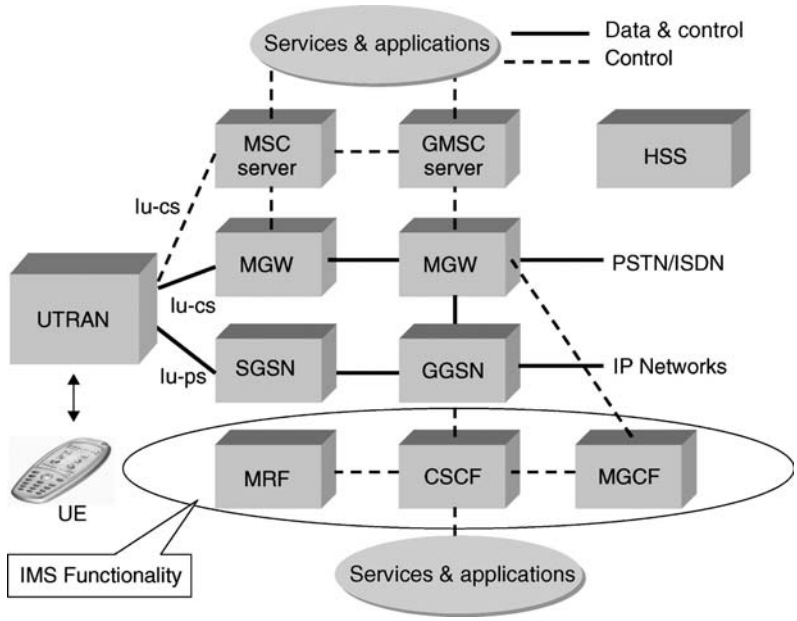



**Figure 5.13** Release 5 UMTS CN architecture

- MGW Control Function, to handle protocol conversions. This may also control a service coming via the CS domain and perform processing in an MGW, e.g. for echo cancellation.

An overview of the different elements and their interfaces can be found in [22] and further details of the CN protocols are given in [25].

## References

[1] 3GPP Technical Specification 25.401 UTRAN Overall Description.
[2] 3GPP Technical Specification 25.410 UTRAN Iu Interface: General Aspects and Principles.
[3] 3GPP Technical Specification 25.411 UTRAN Iu Interface: Layer 1.
[4] 3GPP Technical Specification 25.412 UTRAN Iu Interface: Signalling Transport.
[5] 3GPP Technical Specification 25.413 UTRAN Iu Interface: RANAP Signalling.
[6] 3GPP Technical Specification 25.414 UTRAN Iu Interface: Data Transport and Transport Signalling.
[7] 3GPP Technical Specification 25.415 UTRAN Iu Interface: CN-RAN User Plane Protocol.
[8] 3GPP Technical Specification 25.420 UTRAN Iur Interface: General Aspects and Principles.
[9] 3GPP Technical Specification 25.421 UTRAN Iur Interface: Layer 1.
[10] 3GPP Technical Specification 25.422 UTRAN Iur Interface: Signalling Transport.
[11] 3GPP Technical Specification 25.423 UTRAN Iur Interface: RNSAP Signalling.
[12] 3GPP Technical Specification 25.424 UTRAN Iur Interface: Data Transport and Transport Signalling for CCH Data Streams.
[13] 3GPP Technical Specification 25.425 UTRAN Iur Interface: User Plane Protocols for CCH Data Streams.
[14] 3GPP Technical Specification 25.426 UTRAN Iur and Iub Interface Data Transport and Transport Signalling for DCH Data Streams.
[15] 3GPP Technical Specification 25.427 UTRAN Iur and Iub Interface User Plane Protocols for DCH Data Streams.