3

Cellular Architecture

For large portions of the world's population, cellular networks represent the only persistent digital connection with the outside world. The combination of device portability, reasonable cost and nearly ubiquitous coverage across an increasing percentage of the globe make the services provided by these networks generally more accessible than those provided by the Internet. While the services have generally been limited in their breadth when compared to the Internet, the increasingly interconnected nature of cellular systems and the Internet will create significant new security issues. In order to understand the implications of this increased connectivity, it is critical that engineers and scientists working in Internet domain are aware of the architecture of cellular systems.

In this chapter, we provide an overview of the architecture of GSM cellular networks. We begin by exploring the history of cellular networks so as to understand the design decisions behind current systems. We then present the core network elements supporting both voice and data communications. These discussions include the reuse of both mechanism and design philosophy throughout the network. Our focus then turns to the network and protocols connecting such elements - the *Signaling System Number 7* (SS7) network. The SS7 architecture is directly compared against the more familiar Internet protocol stack so that readers can easily contrast design decisions. After exploring the wireless portion of the network, we then discuss common network operations including registration and making calls. We finish our discussion be examining current security mechanisms protecting both the core and wireless portions of such network.

A complete treatment of GSM networks is simply not possible outside of reading several thousand pages of standards documents. Accordingly, we intend this chapter to should serve as a jumping-off point for researchers. With a firm understanding of the architecture presented herein, those interested in further exploring additional specific details of both current and next generation networks should consult the documentation available through the *Third Generation Partnership Plan* (3GPP) website [179].

3.1 History of Cellular Telephony

The first analog cellular telephony systems were introduced in the early 1980's. The systems, typified by the *Advanced Mobile Phone System* (AMPS) and the derivative *Total Access Communication System* (TACS) allowed users to receive telephone calls while roaming between systems. System bandwidth was segmented using *Frequency-Division Multiple Access* (FDMA). Bandwidth was broken into many carriers, each of which was capable of supporting a single simplex voice channel. When a call was active, one frequency was used in each direction to allow full duplex communication. Due to propagation characteristics, spatial reuse of frequencies was possible, thus allowing national coverage. These systems were limited in capacity due to inefficiencies of analog voice transport. They were also limited in terms of services due to the use of analog signaling. While largely replaced by more efficient systems, analog cellular networks are still in use across the globe. Products including the OnStar automotive security system [74] and home security systems by ADT [28] and GE Security [73] still rely upon AMPS to communicate with monitoring stations in many areas. Regulatory changes by the FCC permitting cellular providers to cease support for AMPS in early 2008 [68], however, will effectively bring an end to analog cellular systems in the United States.

To overcome these limitations, digital systems, called *Second Generation* (2G) systems were introduced in the early 1990s, most of which are still in use. The first systems combine FDMA with *Time-Division Multiple Access* (TDMA) - each carrier is now also divided into time-slots organized into repeating frames. With these systems, each voice call is assigned a time-slot within a frame. Thus each carrier is capable of supporting multiple calls. Each time slot carries one sample of digitized voice. Because digital transmission allows for redundancy and error correction, transmission at lower power is possible. This enables more aggressive frequency re-use, thus further increasing capacity. The GSM system is the prime example of a TDMA cellular system. GSM originated in Europe and is used in most parts of the world. In the US, a similar system based on the IS-136 standard was introduced. However, most IS-136 will be phased out and fully replaced by GSM by early 2008.

An additional important aspect of these systems was the introduction of digital control channels. These channels allow for a great amount of information exchange between the network and mobile device, thus enabling better security solutions and a richer set of services, such as text messaging.

In parallel with the development of TDMA systems in the US, a second digital system based on *Code-Division Multiple Access* (CDMA) was developed and deployed. While these systems provide seemingly similar services to TDMA systems, they use vastly different wireless technology. These systems are based on the IS-95 series of standards. We explore the differences between FDMA, TDMA and CDMA systems in greater depth in Section 3.5.1.

As cellular voice services matured, the focus of service providers moved to providing mobile data services. With the addition of these services, the

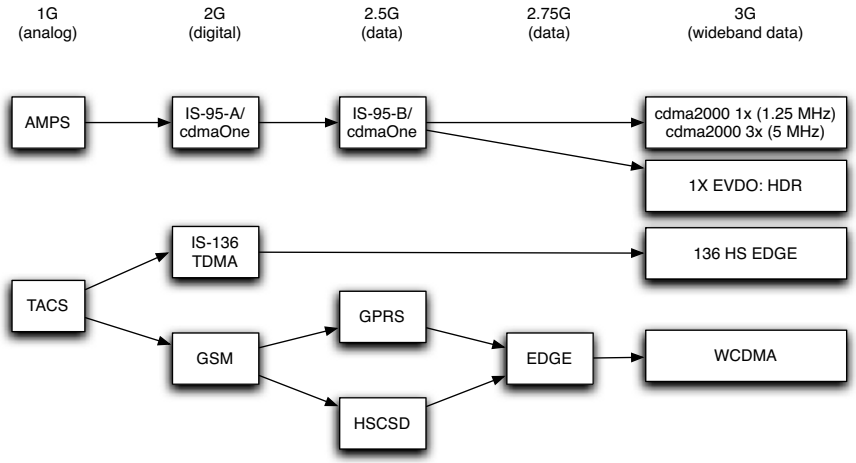

Fig. 3.1. The evolutionary paths of the major telecommunications standards.

so-called "2.5G" networks were deployed. GSM systems were augmented with the *General Packet Radio Service* (GPRS). GPRS provides packet services at data rates of tens of kilobits per second. New modulation techniques called *Enhanced Data Rates for GSM Evolution* (EDGE), often referred to as "2.75G" systems, were introduced to increase these rates to over 200 kbps. CDMA systems were likewise augmented with data services as part of the IS-95B standard.

While theoretically high data rates are possible in 2.5 and 2.75G networks, users typically experience bandwidth comparable to wired dial-up access. *Third Generation* (3G) networks attempt to address this issue through the eventual use of new spectrum and more efficient encoding techniques. However, much of the old core infrastructure will be used at least initially to run such networks. GSM, for instance, is being evolved to the *Universal Mobile Telecommunication System* (UMTS), and will use *Wideband CDMA* (WCDMA) over the air. UMTS promises increased voice capacity and multimedia services with data rates of up to tens of Mbps. IS-95-based systems continue to evolve using narrowband CDMA as part of the CDMA2000 standards. Packet data services at nominal rates of over 10 Mbps are provided using 1xEVDO.

The first, second and early third generation voice systems all re-use the signaling infrastructure of the wired telephone networks. The emerging 3G networks are migrating to Internet-based signaling and services.

We discuss these systems and architecture in more detail below.

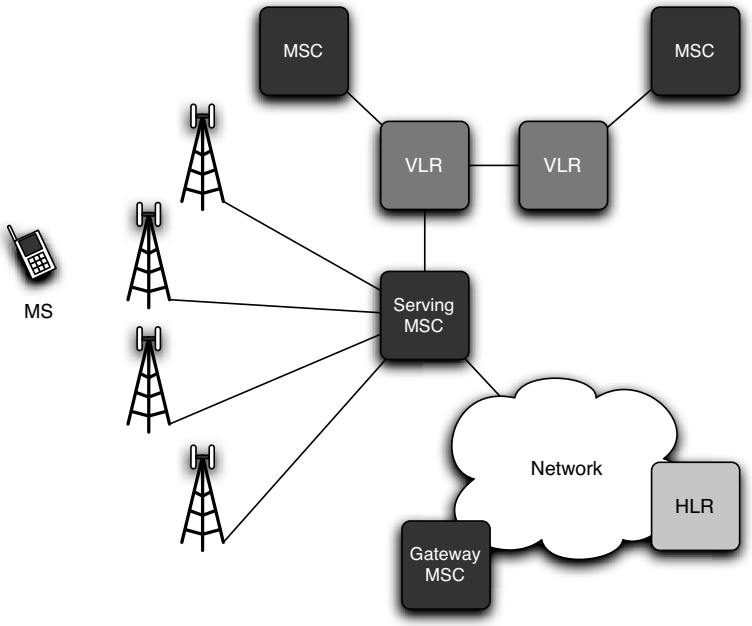

Fig. 3.2. The major components in an SS7 network.

3.2 Cellular Voice Networks

3.2.1 Voice Network Elements

The mobile telecommunication network evolved from the fixed wireline telephone network. These networks use highly intelligent switches made up of switch fabrics and highly reliable processing elements. These processors run programs to enact functions such as routing, resource reservation, digit analysis (for activating services), etc. Separate processors, often called adjunct processor or *Network Control Points* (NCPs), store and execute additional programs for services such as 800-number, credit card calls, etc. To build the mobile telecommunications network, these switches and adjunct processors are loaded with additional software to provide the intelligence to perform mobility management.

Mobility management must address two main issues - connections over which communication takes place must be established between endpoints and services must be available to mobile users. In a wireline network these items are straight-forward. Switches store routing tables by which a dialed telephone number may be routed to its destination. Telephone switches that serve fixed phones have access to databases and processors to provide services to their subscribers. In a mobile network, users must first be located before a connection can be completed. A mobile phone number is simply a logical identifier

for a phone; it has no strong geographical meaning. This requires the network to track the location of devices and to locate them when a call arrives. Because a user may receive a call through many different switches depending on their current location, service profiles and software must be accessible to all switches in the network.

To achieve these goals the mobile voice network introduces several network elements. A simplified network is shown in Figure 3.2.

3.2.2 Home Location Register

The heart of the mobile network is the *Home Location Register* (HLR). Essentially a massive database, the HLR addresses the issues created by mobility in a phone network by storing permanent copies of user profiles, which contains information including the address of the switch currently providing service to a *Mobile Station* (MS) (a.k.a. a phone). All requests involving the user, from incoming calls to the network determining whether a user is eligible to receive certain services (e.g., call-forwarding), are handled by the HLR. Table 3.1 summarizes the mandatory information stored in an HLR according to the standards documents [15].

One of the most important duties of the HLR is authentication. While standards documents mention the presence of an *Authentication Center* (AuC), the functions of these elements are absorbed by nearly every commercially available HLR. As is standard practice in the field, we therefore refer only to the HLR when such operations are performed. In order to perform its authentication duties, an HLR assigns a unique identifier, known as the *International Mobile Subscriber Identity* (IMSI), and a unique cryptographic key, K_i, to each user in the network. In order to determine whether or not a device should be granted access to the network, the HLR creates a challenge that can only be correctly responded to by the device with the correct K_i on its *Subscriber Identity Module* (SIM) card. The protocol used to actually perform such authentication is discussed in greater detail in Section 3.7.

Device level authentication is also possible in GSM networks using an *Equipment Identity Register* (EIR) – functionality that is also typically absorbed by the HLR. In addition to an IMSI, which is stored in the SIM, each device carries a unique identifier known as the *International Mobile Equipment Identity* (IMEI). Should a device be stolen or be known to be causing harm to the network, the HLR can simply add the IMEI to a blacklist and prevent the device from attaching to the network.

Users are assigned to specific HLRs based on their phone number, allowing queries to be efficiently routed. The number of HLRs in a network, however, is highly variable. Limitations on processing power and the number of concurrent database lookups forced early cellular providers to use many HLRs throughout the network. The advantages to this approach were numerous. For instance, failure in one HLR meant that the majority of the network would remain unaffected. However, the cost of administering and maintaining so

Table 3.1. Mandatory data stored in the HLR and/or VLR.

Data	HLR	VLR	Description
IMSI	✓	✓	Permanent and unique identifier assigned to each user. Different than the user's phone number (MSISDN).
IMEI		✓	Unique identifier for an MS.
TMSI/P-TMSI		✓	Temporary identifier used to preserve the privacy of a user's identity over the air.
NAM	✓		Indicates whether a client is registered to receive voice, data or both services.
MSISDN	✓	✓	Voice phone number for an MS.
RAND/SRES and K_C	✓		Random number, signed response and session key in authentication triplets.
K_i	✓		Encryption key shared between HLR and a specific MS.
VLR Number	✓		Identifies the VLR currently serving an MS.
MSC Number	✓		Identifies the MSC currently serving an MS.
SGSN Number	✓	✓	Identifies the SGSN currently serving an MS.
GGSN Number	✓	✓	Identifies the GGSN currently serving an MS.
Roaming Restricted	✓		Notes that a feature is not supported in an SGSN or MSC and can be used to prevent a device from associating with an LA.
Provision of teleservice	✓	✓	Identifies which services (e.g., voice, SMS, data) an MS is able to receive.
Transfer of SM	✓		Indicates whether a text message should be sent via the voice or data network.
MNRR	✓		Specifies the reason why an MS is not reachable (e.g., not GPRS or IMSI attached).
MS purged for non-GPRS flag	✓		MS information has been removed from the VLR.
MS purged for GPRS flag	✓		MS information has been removed from the SGSN.
GGSN-list	✓		Notes the GGSNs (including their number and IP address) to be contacted if a device is not GPRS attached.

many databases in the face of rising computing power has led to a significant centralization of network resources. A number of large providers have now or are in the process of migrating to a single, network-wide HLR.

If an HLR were to fail or be compromised, especially given the move toward centralization, the corresponding cellular network would simply cease to operate. A successful attack on such a device, while difficult, would also give an adversary access to information pertaining to every subscriber in the network. Recognizing this, HLRs are the best provisioned and physically protected elements within the cellular network.

3.2.3 Mobile Switching Center/Visiting Location Register

Mobile Switching Centers (MSCs) act as telephony switches and deliver circuit-switched traffic in a GSM network. Expressing their role so succinctly, however, fails to capture the magnitude of functions tasked to these devices. MSCs can act as gateways between a cellular network and *Public Switched Telephone Network* (PSTN). They connect the wireless portion of the network with core elements and have the most specific information about the location of users under their service. MSCs also facilitate mobility by assisting devices performing "handoffs" between base stations and assist in the billing process.

Managing such tasks requires more than simply switching. Because they are responsible for performing all of the above tasks, MSCs must be aware of context on a per-user basis. Such information can be retrieved from user profiles in the HLR; however, constant lookups in the HLR can be expensive for a number of reasons. Limited bandwidth network links (see Section 3.4), latency concerns and the impact of heightened load on the HLR all prevent such frequent lookups from occurring. To address these issues, temporary copies of user profiles are stored in a nearby database known as the *Visitor Location Register* (VLR). As shown in Table 3.1, the VLR contains many, but not all, of the same user profile data as the HLR. Most notably, the VLR does not have access to a user's K_i. This consideration is critical as phones can receive service from a network operated by a parties other than their provider.

The configuration between MSCs and VLRs varies from network to network. As shown in Figure 3.2, a single VLR may provide service to multiple MSCs. In other systems, the relationship between VLRs and MSCs is one-to-one, and such devices are co-located (and even referred to as MSC/VLRs or simply as MSCs). While the latter arrangement is more common in current networks, we distinguish between the functions of each wherever possible.

3.2.4 Base Station Subsystem

The link connecting wireless devices to a cellular network is provided by *Base Station Subsystems* (BSSs), which are composed of two components. *Base Transceiver Stations* (BTSs) are simply the radios used to transmit messages between mobile devices and the network. Most BTSs are composed of multiple (i.e., three) directional antennas that divide a each cell into smaller sectors (see Section 5.2.3 for more information). *Base Station Controllers* (BSCs) provide intelligence to the radios and are responsible for functions including scheduling and encryption. Deployment is often vendor specific - a single BSC can be combined with a single BTS, or a single BSC can service a large number of BTSs. In general, however, no distinction is made and the unit is generally collectively recognized as the BSS or base station.

Base stations are often arranged into groups of called *Location Areas* (LAs). Such groups often correspond to geographic regions. For instance, New

York City could potentially be divided into five LAs, each corresponding to one of five boroughs. [1] The advantage to dividing towers into such collections are numerous. Mobile devices not currently in a voice or data call with the network can move between towers within an LA without re-registering with the network. Only moves across LA boundaries cause inactive devices to notify the network of their new position. In so doing, signaling across the constrained SS7 and wired channels is greatly reduced.

Users actively using the network are required to perform a handoff for a number of reasons. As the number of users in a cell approaches capacity, the network may direct devices to the resources of a tower providing overlapping coverage. Similarly, if conditions on the wireless network degrade (e.g., increased noise), devices may also change towers. The third, and most common cause of handoffs, results from user mobility. When such an occasion arises, the transfer of device service between two base stations can occur in one of two ways. *Hard Handoffs* require a mobile device to drop its connection with its previous tower before attempting to tune into a neighboring cell. Such a transition should technically be instantaneous, but is often noticeable to users as seemingly unexplained gaps in conversation or prematurely terminated calls. *Soft Handoffs* allow a device to tune in to two or more base stations simultaneously, ensuring a more smooth transfer between cells. Logically, because soft handoffs allow the use of resources in multiple cells concurrently, they can reduce the overall capacity of the network. GSM networks implement hard handoffs between base stations; however, the majority of next generation systems will employ soft handoffs to improve overall quality.

3.3 Cellular Data Networks

3.3.1 Data Network Elements

The data communications system for GSM is called the *General Packet Radio Service* (GPRS). A simplified architecture is shown in Figure 3.3. Recognizing that many of the functions needed to provide cellular data service were already provided by the elements supporting voice in the network core, significant portions of the infrastructure are reused. In particular, signaling is done using SS7 and the HLR is used to perform authentication and store user profiles.

Supporting the new data networking capabilities also required the inclusion of a number of new core elements. In anticipation of high levels of data traffic, providers have deployed new higher bandwidth links within their networks. These links also connect two new network elements responsible for handling data – the *Gateway GPRS Support Node* (GGSN) and the *Serving GPRS Support Node* (SGSN). We examine these new elements in greater detail.

[1] Given its size and population density, such a division is too coarse. Accordingly, Location Areas in New York City are much smaller than this example.

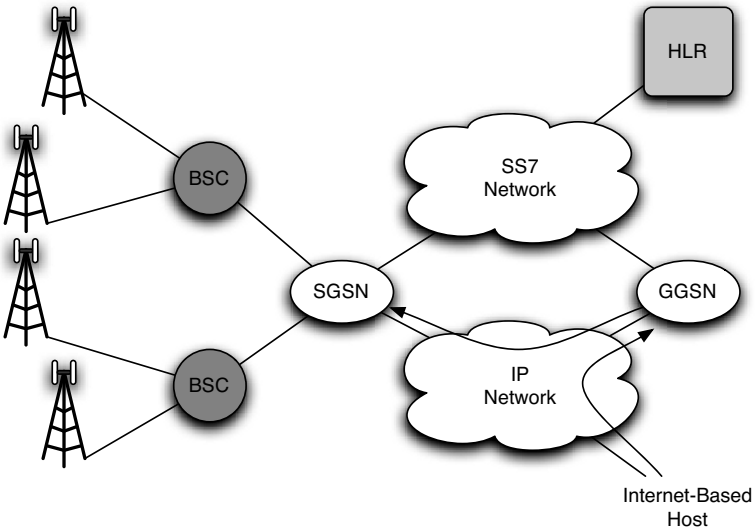

Fig. 3.3. The network architecture of GPRS/EDGE enabled cellular systems.

3.3.2 Gateway GPRS Support Node

Packets crossing the border between a cellular network and the Internet interact with a gateway between the two networks. This node, known as the *Gateway GPRS Support Node* (GGSN), is responsible for more than simply forwarding such packets to their ultimate destination. To provide support for multiple networking protocols, the GGSN can tunnel both IP and X.25 packets to a receiver. While the latter is rarely used, support is provided to ensure that older networks can support data services. GGSNs also provide support for operations more commonly associated with telecommunications networks. Granting and enforcing *Quality of Service* (QoS) markings on specific flows can be managed from the GGSN. The GGSN can also assist in the billing process by recording the amount of bandwidth used by each customer.

The most important function of the GGSN, however, is address and mobility management. As devices register with the network, the GGSN acts much like a DHCP server and assigns addresses. Both public (static) and private (dynamic) IP addresses are currently available from providers. After assigning an address, the GGSN then maintains a listing of the mobile device's current SGSN. Upon the arrival of incoming packets, the GGSN then performs a lookup on the targeted device, determines its SGSN and then tunnels the request.

Table 3.2. Mandatory data stored in the SGSN and/or GGSN.

Data	SGSN	GGSN	Description
IMSI	✓	✓	Permanent and unique identifier assigned to each user. Different than the user's phone number (MSISDN).
NAM	✓		Indicates whether a client is registered to receive voice, data or both services.
MSISDN	✓		Voice phone number for an MS.
P-TMSI	✓		Temporary identifier used to preserve the privacy of a user's identity over the air.
TLLI	✓		A signaling address between the SGSN and a specific MS.
IMEI	✓		Unique identifier for an MS.
RAND/SRES and K_C	✓		Random number, signed response and session key in authentication triplets.
Ciphering Key Seq. Number	✓		Identifier for the currently used K_C.
Ciphering Algorithm	✓		Encryption algorithm in use between the SGSN and MS.
RAI	✓		Identity of current Routing Area for an MS.
Cell Global Identification	✓		A concatenation of the LAI and the Cell Identity.
RA not allowed flag	✓		Applied to restrict service based on lack of roaming agreement or unsupported feature.
Roaming Restricted in the SGSN	✓		Roaming restricted in an SGSN because of an unsupported feature.
MNRG	✓	✓	MS not reachable because it is not GPRS attached.
MM State	✓	✓	The current mobility management state of an MS.
PDP Type	✓	✓	Indicates the protocol used by an MS for data communication (e.g., IP, X.25).
PDP Address	✓	✓	Lists the IP address of an MS.
NSAPI	✓	✓	Identifies a PDP context associated with an address.
SGSN address		✓	The IP address of the SGSN currently serving an MS.
QoS Negotiated	✓	✓	Notes the quality of service negotiated between an MS, its SGSN and the corresponding GGSN.
DRX Parameters	✓		Indicates that the MS is not constantly monitoring paging requests and that it should only be paged during certain times.
Classmark	✓		Specifies the classes of content (e.g., WAP, J2ME) an MS can support.

3.3.3 Serving GPRS Support Node

Much like MSCs, *Serving GPRS Support Nodes* (SGSNs) are responsible for more than simply moving packets toward their ultimate destination. With the assistance of a location register, the SGSN stores user profile information

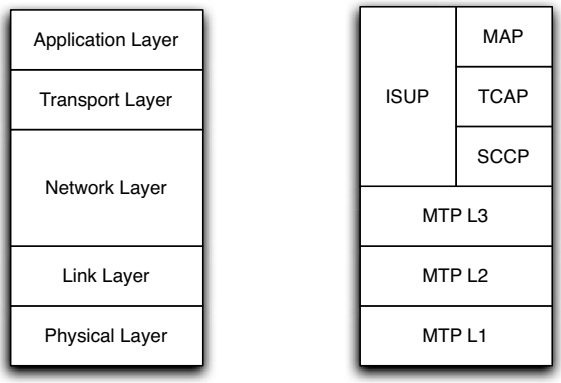

Fig. 3.4. The SS7 protocol stack with the Internet protocol stack as a reference for functionality.

locally. As is the case in MSCs, such profile information is valuable in assisting handoffs, performing authentication and in the billing process. Table 3.2, which is specified in standards documents [15], provides a list of the values stored by both GGSNs and SGSNs.

Because an HLR's knowledge of a user's current location is limited to their current SGSN, these nodes are responsible for tracking such information at a finer granularity. If a specific client is actively receiving data traffic, for instance, the SGSN records the cell or sector in which they are located. As such users move between cells, the SGSN is updated. To reduce the amount of signaling in the network, inactive but registered users (i.e., those whose devices are on but not currently exchanging packets with the network) generally do not alert the network of movement between cells or sectors. Like in the voice network, this is made possible by grouping multiple towers into sets. These *Routing Areas* (RAs) are typically smaller than their voice network Location Area counterparts.

Because many of the mobility management and authentication functions in GPRS are the same as those in pure GSM systems, many of these operations can be performed in parallel. For instance, when a device attempts to register with the voice portion of the network (see Section 3.6), the network can also register the device with the data elements automatically. To minimize the resources dedicate to locating a device, the MSC can defer the process of locating a user (known as paging) to that user's SGSN. Such optimizations are discussed in Chapter 6.

3.4 Signaling Network and Protocols

With an understanding of the core elements found in modern cellular networks, we now discuss the communications between them. We begin with an

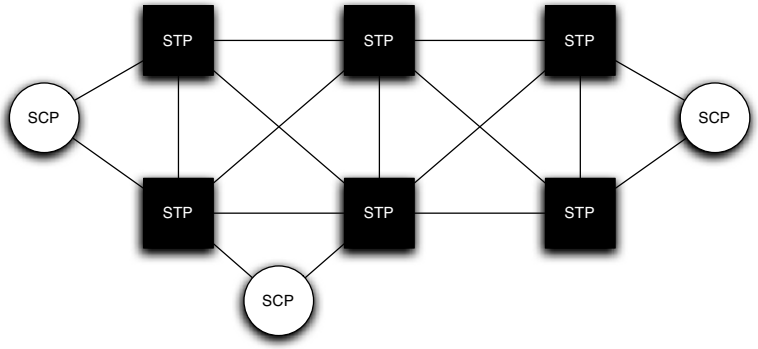

Fig. 3.5. The Common Channel Signaling network.

overview of the signaling network itself. Our focus the turns to the SS7 stack and the role played by each protocol layer in the network. Because we assume that most readers have some familiarity with the IP networking, we compare the functionality of each of the layers of the SS7 protocol stack with the Internet network model. Figure 3.4 provides a comparison of these protocol stacks.

3.4.1 Common Channel Signaling Network

The signaling messages exchanged between the switches, HLRs and VLRs are carried over the *Common Channel Signaling* (CCS) network and are part of the Signaling System No. 7 protocol suite. A simplified view of the CCS network is shown in Figure 3.5. The CCS network is built using special purpose highly redundant packet switches called *Signaling Transfer Points* (STPs). Any element that generates a signaling message (e.g., a HLR, VLR, or switch), is called a *Signaling Control Point* (SCP). All SCPs are connected to at least two STPs. STPs are connected in a quad arrangement. Using this configuration, no single failure in the CCS network will isolate a network element. Due to the cost of the STPs the CCS network is very tightly engineered.

3.4.2 Message Transfer Part

Forming the foundation of the SS7 protocol stack is the *Message Transfer Part* (MTP). MTP is tasked with the reliable delivery of signaling messages, including responding to link outages. In order to support this functionality, MTP is split into three distinct partitions.

Message Transfer Part Level 1 (MTP1) corresponds to the physical layer in the Internet model. All links are bidirectional and support bandwidths as high as 56 KBps in ANSI standard networks and 64 KBps elsewhere. Up to four physical links between two nodes can be combined to create an aggregate

rate of 1.544 Mbps. In order to meet the real-time requirements of telephony, no link can operate at a bit rate of less than 4.8 kbps.

Message Transfer Part Level 2 (MTP2) provides link layer functionality for the SS7 protocol stack. Accordingly, communications between two directly connected network nodes are handled by MTP2. However, MTP2 provides far more functionality than simple point to point addressing. The first such functionality is reliable message delivery. This is in direct contrast to the Internet model, which uses end-to-end reliability services. Delivery guarantees are provided by the Go-Back-N algorithm, which relies upon negative acknowledgments to cause retransmission of packets within a window. Secondly, MTP strictly monitors the error rate of all links. Should the number of errors detected on the link surpass a threshold, MTP2 alerts higher level protocols and the link is shut down. The proper functioning of links is of paramount importance – filler packets are continuously transmitted on all links in the SS7 networks so that error detection can be constantly executed. Finally, MTP2 offers explicit flow control mechanisms. Should the congestion condition on a link exist for a number of seconds, the link is shut down. These functions are handled at such a low layer in order to maintain the real-time requirements of the network.

Many of the responsibilities traditionally assigned to Network layer protocols are provided by *Message Transfer Part Level 3* (MTP3). Accordingly, MTP3 is responsible for routing packets between sources and destinations. Whenever possible, each STP will attempt to balance traffic sent across each link. Like the Internet model, each packet may therefore take slightly different paths through the network. Messages that must be kept strictly in sequence, however, can be flagged to use the same physical link. MTP3 also responds to link outages reported by MTP2. Whether from from processor failure, high link errors or congestion, MTP3 will reconfigure routes around unavailable neighbors to ensure that traffic is delivered. The combination of redundant links and the ability to react quickly to network failure allow such networks to maintain their extremely high levels of availability.

3.4.3 Signaling Connection Control Part

The routing functionality provided MTP3 is somewhat limited. The *Signaling Connection Control Part* (SCCP) addresses these issues by providing the remaining functions common to Network layer protocols. Whereas MTP3 messages can only address nodes in the network, SCCP allows specific functions to become the destination of a request. For instance, support for special global numbers, such as 800 numbers, are supported by SCCP. Requests can be delivered using one of five classes of service. Classes 0 and 1 are connectionless and differ only by the ability to request that all packets be sent on the same physical link in the case of the latter. Classes 2 and 3 are both connection-oriented and require that all packets be delivered over the same links. They differ, however, in that Class 3 SCCP messages provide flow control. Class 4

messages are the same as Class 3, but allows for messages that can not be properly reassembled to be retransmitted.

In conjunction with the three levels of MTP, SCCP forms what is referred to as the *Network Services Part* (NSP).

3.4.4 Transaction Capabilities Application Part

For services control, including mobility management, a transaction-oriented protocol called the *Transactions Capabilities Application Part* (TCAP) is used. TCAP provides a framework through which nodes throughout the network can request the execution of remote procedures. For instance, *Intelligent Network* (IN) functions such as toll free calling and automatic call blocking are invoked with TCAP messages. TCAP messages also provide transaction identifiers, which are functionally similar to port numbers in transport layer protocols.

3.4.5 Mobile Application Part

The majority of the network-supported procedures discussed in this book use the *Mobile Application Part* (MAP). MAP provides application layer functionality to SS7 networks. Services visible to the user, including call handling, text messaging and location-based services are all carried by MAP messages. Less visible services, such as mobility management (both within and outside their home network), service profile downloads between HLRs and VLRs, and authentication procedures are all conducted using MAP.

Because MAP messages contain so much critical information, they must be protected. We briefly discuss MAPsec in Section 3.7 and its associated security weaknesses in Chapter 4.

3.4.6 ISDN User Part

For connection control, the *Integrated Services Digital Network* (ISDN) User Part (ISUP) is used. ISUP carries information so that calls may be routed and resources reserved along the path. It is used for both fixed and mobile networks. In fixed networks, routing is done based on the dialed number. In mobile networks, the first portion of the connection is established using the dialed number; from the gateway MSC to the serving MSC, the call is routed based on the temporary routing number. ISUP messages are routing hop-by-hop through each switch through which the connect will pass.

3.5 Wireless Network

In the following subsections we discuss various facets of the wireless portion of the network including access techniques, frequency issues, voice coding, and provide a brief summary of procedures.

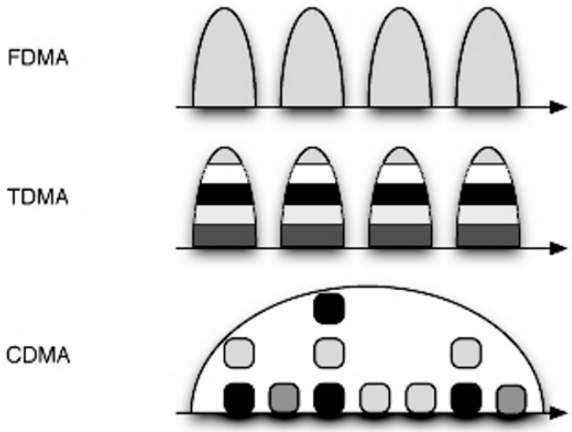

Fig. 3.6. Representations of the three main methods of spectrum access: FDMA, TDMA and CDMA.

3.5.1 Wireless Access Techniques

Dividing wireless spectrum into a medium capable of supporting many users can be achieved through a number of approaches. While many other methods for wireless access exist, the frequency-, time- and code-division multiple access represent the three technologies currently use by cellular telecommunications networks. Each of these general classes has a direct impact on voice quality, interference and the maximum number of users supportable in a cell. Figure 3.6 offers a comparison of these three methods.

Frequency-Division Multiple Access

Frequency-Division Multiple Access (FDMA) is the most basic means of providing concurrent wireless access to multiple parties. Users in an FDMA system each receive exclusive access to independent frequencies, which are referred to as *carriers*. Such separation simplifies a number of systems issues. For instance, hosts do not need to be synchronized in time with each other. Additionally, hardware support is vastly simplified as devices only need to be able to tune to specific frequencies.

For their simplicity, FDMA systems suffer from a number of notable disadvantages. First, in order to prevent two frequencies from interfering, calls in a FDMA must be separated by guard bands – unused spectrum in which interference between two frequencies is likely to be high. The size of guard bands is highly dependent upon the carrier bandwidth allocated in the system. For instance, AMPS used 30 kHz carriers with kHz guard bands. The absence of such buffers could significantly improve capacity of such systems. Moreover, as the number of users in the system increases, so too does the amount of interference. Pure FDMA systems also suffer from a fixed, relatively low bit rate

per carrier, thereby limiting the ability to transmit data in such systems. The security impact of these low bandwidth carriers will be discussed in greater detail in Chapter 5.

Because of their simplicity, FDMA wireless access was used in analog (1G) cellular networks (e.g., AMPS, TACS). However, because more efficient means of spectrum use are now practical, pure FDMA cellular systems are no longer being deployed.

Time-Division Multiple Access

While FDMA systems reserve a small piece of spectrum per caller, the utilization of this bandwidth is relatively inefficient. Periods of silence and short-term limits on the dynamic range of the human voice (see Section 3.5.3) create the potential for individual frequencies to be shared by multiple users. Recognizing this, *Time-Division Multiple Access* (TDMA) systems provide service to users by partitioning a frequency into evenly sized timeslots. Devices listen to a single timeslot, each of which are serviced in a round-robin fashion. In GSM, for instance, each frequency on the air interface is divided into eight timeslots, each of which are serviced every 4.615 msec. As the device samples its assigned timeslot across multiple iterations, a virtual channel is created.

TDMA systems offer a number of advantages over their FDMA counterparts. Because multiple users can share a frequency band, the number of concurrently supportable by such a system is significantly increased. Moreover, the guard bands need to protect users in FDMA systems are no longer required between channels[2], although they do require guard times. Devices listening to a single timeslot, as opposed to constantly monitoring a frequency, also dramatically increase the lifetime of their batteries. Finally, by allowing mobile phones to potentially listen to multiple timeslots, TDMA systems also allow flexible bit-rates for data communications.

These improvements over FDMA come at the cost of complexity. Because access to wireless resources is based on time-division, all devices in the network must be tightly time synchronized. To protect against clock drift, each timeslot must be buffered between guard-time so as to reduce the probability that two devices will accidentally overhear each other. Devices must also frequently resynchronize with the network in order to maintain their ability to operate between the guard-time buffers. Finally, because multiple devices are transmitting on the same frequency, multipath distortion (i.e., reflections of signal received later than the intended timeslot) can significantly impact call quality.

While these costs are significant, TDMA cellular networks far outperform FDMA systems. Examples of deployed networks relying on this technology

[2] In practice, guard bands are still used in real TDMA networks such as GSM as networks use TDMA over multiple frequencies to increase the number of supportable users.

include IS-136 and GSM, the most widely deployed cellular networking technology in the world [79].

Code-Division Multiple Access

Code-Division Multiple Access (CDMA) systems avoid the issue of separating users by time and space by allowing them to transmit simultaneously on the same frequency. To achieve this, each user in a cell receives a unique code (i.e., mask) to spread the spectrum they transmit in to each bit of their data. The size of this code is many times larger than each bit.

Wireless access via CDMA has a number of advantages over the two previously discussed methods. The process of coding transmission not only allows users to a share frequency range, but also makes na ive of confidentiality for the system. Unlike FDMA and TDMA systems in which an adversary can simply scan frequencies or timeslots to intercept traffic, locating a specific signal from within the combined traffic of the network is computationally infeasible as the chip rate becomes large without knowledge of a specific code. However, most codes are available as part of the specification, thereby reducing the search space of an attacker. The additive nature of the coding also places no theoretical limit on the number of users that can be supported in a single area; rather, the number of users to be supported can be increased at the expense of the quality of service received. Moreover, the spreading of signal across a wide band of frequencies reduces the impact of multipath distortion and provides improved quality for voice.

Systems using CDMA face a number of challenges. Because of the additive nature of the coding, the network must keep strict control over the transmission power used by each of the nodes in a cell. As devices move towards and away from the tower and pass behind occlusions, the effort needed to coordinate power control becomes significant. Devices can also potentially self-jam if the pseudo-noise sequences used to spread their signal across the spectrum are not exactly orthogonal.

In spite of these difficulties, CDMA schemes represent the most advanced wireless access systems in the world. Originally used by the military because of their inherent confidentiality and robustness to jamming, CDMA systems are increasingly being used in the civilian arena. While the use of CDMA radios is largely limited to cellular networks in the United States (e.g., IS-95), all major third generation networks will use CDMA-based radios as their wireless access method.

3.5.2 Frequency Issues

Frequency Assignment

In spite of being the most widely deployed cellular networking technology, a GSM-capable mobile device may not be able to interact with every GSM

network it encounters. Even in the presence of nearly universal roaming agreements, one issue still impedes access to telephony services through any GSM network on the globe – frequency assignment. Because each nation is able to regulate how the wireless spectrum is divided and used within their borders, consistent use of spectrum across the world does not exist.

In response, GSM systems exist in one of a number of possible frequency bands. The first networks, deployed largely in Europe, operated within the 900 MHz band and are referred to as GSM-900. These systems, which offer a total of 124 bidirectional 200 kHz carriers, now represent the vast majority of deployed GSM networks in the world, and are deployed on every inhabited continent. In response to rising demand on these networks, most GSM-900 systems now also include an additional 50 carriers and are referred to as Extended-GSM networks (EGSM). Because this portion of the spectrum was already dedicated for other purposes in many locations, other GSM systems transmit at twice the original frequency, or 1800 MHz, and offer up to 374 carriers. Such networks, known as GSM-1800, can be found throughout the United Kingdom, Brazil and parts of Southeast Asia [80].

Similar problems of spectrum allocation exist in the Americas. Because the 900 MHz frequency range used in GSM-900 systems was already allocated to other systems (e.g., IS-54), new spectrum had to be dedicated to support GSM. In less densely populated areas, providers deployed GSM-850. Largely similar to GSM-900 (i.e., supports up to 123 carriers), these networks take advantage of these lower frequency waves to expand their coverage areas. Because higher frequency transmissions allow for increased capacity (up to 298 carriers), GSM-1900 has been deployed in larger metropolitan areas. Such deployment strategies are not compulsory, but generally guide the specific technology deployed when creating new coverage areas.

With all of these possibilities, a logical question arises: Given a mobile device with a single radio, in which frequency band should that device listen in order to receive uninterrupted service? In the ideal case, a provider's network would offer widespread coverage using a single band. In reality, however, neither of these conditions holds. Mobile users may require the use of other networks operating in a different frequency band to maintain connectivity. This case is especially common as users cross national borders. Moreover, a single provider may deploy base stations operating in different frequency bands throughout their network. Accordingly, devices must be able to tune into multiple frequencies in order to ensure service. The majority of modern mobile devices therefore come with so-called *tri-band* and *quad-band* capabilities, which allow them to operate in the presence of most (three) and all GSM frequency bands, respectively.

Frequency Reuse

In spite of the seemingly large wireless capacity available to GSM networks, no single cell has access to the full complement of carriers described above. For

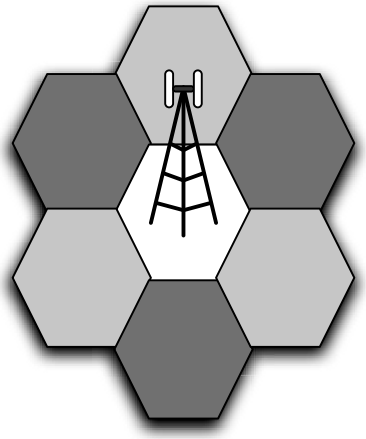

Fig. 3.7. A series of cells in a network with a frequency reuse factor, F_u, of three. Note that no cell borders another cell operating on the same frequency.

instance, if multiple providers offer service to a single geographic area, the spectrum must be divided between them. Moreover, no two bordering cells can use the same carriers, because calls occurring on the same carriers in two neighboring cells will create significant *co-channel interference*. Accordingly, spectrum reuse across cells must be carefully planned.

We use Figure 3.7 and a GSM-900 network to step through a simple example. Accordingly, a total of 124 carriers will be available for all GSM activity in this region. As is done in many real networks are planned, cells are assumed to be hexagonal in shape. For the purposes of this example, we assume that two providers offer competing service within this geographic region. Assuming equal division of resources between the two providers, this immediately halves the number of available carriers. In order to prevent calls in two cells from interfering, a frequency reuse factor, F_u, of three is set. Accordingly, the number of carriers that any one cell can use is further divided by F_u. If cells are further divided into sectors, the number of carriers is again divided by three. The maximum number of TDMA channels that can be provided in each sector of such a network is therefore:

$$\begin{aligned} Channels &= \lfloor \frac{124 \text{ carriers}}{2 \text{ providers}} \times \frac{1}{3(F_u)} \times \frac{1 \text{ cell}}{3 \text{ sectors}} \rfloor \times \frac{8 \text{ channels}}{1 \text{ carrier}} \\ &= 48 \text{ channels/sector/provider} \end{aligned}$$

Given such a setup, it may be possible for a provider to support as many as 46 concurrent voice and data calls in an area (the other two channels would be used for network signaling). The implications of such an arrangement are,

in many ways, more complex. Some cells cover an area of up to 10 square miles, meaning that a large number of users may be forced to compete for a relatively small number of total channels. In more densely populated areas, much smaller micro- and pico-cells may instead be used to combat this problem. However, the relationship between cell/sector size and F_u is inversely proportional. Accordingly, the number of cells that must be between two cells using the same frequency must increase. Network planning must therefore be done in an extremely careful fashion.

Frequency Hopping

While carefully planning frequency reuse throughout a coverage area can significantly reduce interference, GSM networks adopt additional techniques in order to lessen the impact of unrelated transmissions. Because frequency reuse generally protects devices from each other, the most significant source of interference is often a device's own signals. As mentioned in Section 3.5.1, signals reflecting off of nearby objects (e.g., buildings) can considerably degrade the quality of a connection. Much like speaking loudly in a room without sound dampening materials, unaccounted-for multipath distortion can literally make coherent communication impossible.

GSM networks deal with multipath distortion by regularly changing the frequency used within a cell. This technique, known as *frequency hopping*, allows devices simply to tune into a new frequency and timeslot. Instead of receiving signals that may be consistently in the opposite phase of the transmissions sent by the tower, this technique randomizes the effects of interference. Because random noise can easily be removed using error correcting codes, the overall quality of calls in the network is drastically improved. Note that this frequency hopping occurs at a much flower rate than spread spectrum systems that use frequency hopping explicitly (e.g., Bluetooth).

Frequency hopping in GSM can operate under one of two modes – cyclic and pseudorandom [22, 18]. In order to determine the mode used in a sector, devices simply listen for the broadcast of the *Hopping Sequence Number* (HSN) on the control channels. If HSN is set to zero, devices simply use the next highest numbered carrier above their current position at each frequency change. After reaching the highest available carrier, devices simply wrap around to their provider's lowest available carrier. If the HSN is set to a number between 1 and 63, devices change carriers according to tables of pseudo-randomly generated sequences. While cyclic frequency hopping is easier to implement, the use of pseudo-random sequences better randomizes the multipath distortion observed by mobile devices.

Future Frequency Management Issues

The most critical problem facing cellular providers in the future, according to most members of the industry, is spectrum allocation. While current narrowband (200 kHz/carrier) GSM networks deliver voice and data services, higher

bandwidth services and improved connection to larger IP networks (i.e., the Internet) will only be possible through the allocation of new spectrum. Unfortunately, procuring additional spectrum for so-called wide-band services is a timely and expensive process. Most large portions of the usable spectrum have already been allocated. Accordingly, buyers must often wait for technology that previously had exclusive access to a frequency band to become obsolete or unused. Because such events are fairly infrequent, competition for new spectrum is often intense. Wireless providers have, in the past, proposed bids for tens of billions of dollars for access to these new bands.

Spectrum allocation will continue to be a major issue facing all wireless providers in the foreseeable future. At the time of this writing, the *Ultra-High Frequency* (UHF) portion of the spectrum once used for analog television (the 700 MHz band) has only recently been divided by public auction. Unlike previous auctions, however, a number of non-traditional parties expressed interest in acquiring a portion of this band. Google, for instance, proposed an initial offer of US$4.6 Billion [67]. Because it is unlikely that another large portion of the wireless spectrum will become available again during the next decade, bidding grew to over $19 Billion in total [70]. The implications of this auction will take time to become obvious.

3.5.3 Voice Encoding

Modern cellular networks provide high-quality voice communications via digital encoding. Until the mid 1990's, however, only analog service was available. Analog phone systems suffer from a number of significant limitations. As mentioned in Section 3.1, systems such as AMPS operate by transmitting each call on distinct frequencies. The maximum number of users supported in each cell is therefore directly dependent upon wireless spectrum allocation – an expensive and infrequently conducted process. Because noise in analog signals is added at each hop along its traversal between source and destination, the quality of voice telephony is such systems was low. This inability to transmit high-fidelity signals was one of the chief limiting factors in using encryption in early mobile phone systems.

Digital cellular systems offer significant improvements in spectrum efficiency, bandwidth and voice quality. Whether in a TDMA (e.g., IS-136, GSM) or CDMA (e.g., IS-95) system, modern systems no longer require a single frequency per user in an area. Instead, devices in these systems sample, digitize and compress speech. While we focus on the specific mechanisms available to GSM networks, similar techniques are used across all systems.

Analog signals are encoded digitally through *Pulse Code Modulation* (PCM). Instead of capturing the entire analog signal itself, PCM uses regular sampling and records a binary representation of the magnitude of the sound wave. Each sample is encoded as an integer, and a ceiling function is used to remove ambiguity. The G.711 encoder, which takes either 14-bit or 13-bit signed integers depending on the compression algorithm applied (μ-law

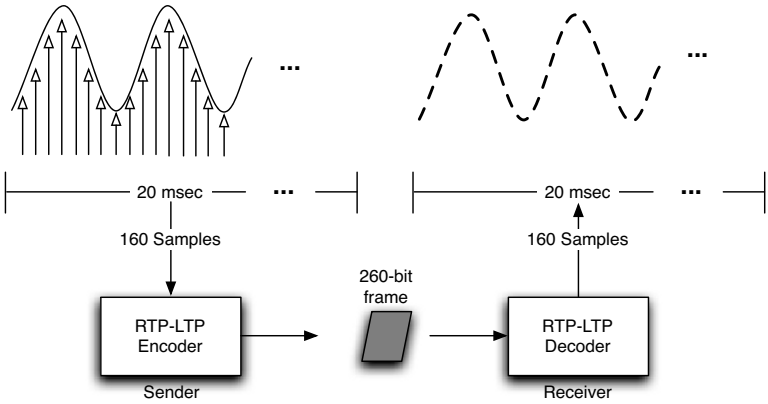

Fig. 3.8. A high level overview of voice encoding. Voice is sampled (white arrows) at a rate of 8,000 times per second (8 kHz) using PCM. Groups of 160 samples (20 msec) are then sent to the RPE-LTP encoder, which generates a 260 bit frame. When this frame is delivered to the receiving client, the reverse process occurs and the voice signals are reconstructed.

and A-law, respectively), then converts each of the inputs into 8-bit samples [93]. At a sampling rate of 8,000 times per second (8 kHz), the resulting voice stream is encoded at 64 kbps. As a comparison, CD quality audio is typically recorded at 1411.2 kbps [88].

While music requires a high bitrate encoding to capture its dynamic and potentially complex nature, because of physical limitations governing the speed in which the mouth can meaningfully change sounds, much of the information included in the 64 kbps stream is redundant. Accordingly, speech encoding can be made more efficient by removing unneeded samples. To achieve these ends, phones apply the Regular Pulse Excitation - Long Term Prediction (RPE-LTP) algorithm, which reduces the average bit rate of the stream to 13 kbps [8, 10]. RPE-LTP takes as input either the unmodified A- or μ-law code or can convert the above 8-bit data to the 13-bit uniform PCM format [94]. The encoder then takes batches of 160 samples (20 msec) and outputs 260-bit encoded blocks. When received by the phone at the other end of a connection, each 260-bit block are used to reconstruct the 160 speech samples. This process is summarized in Figure 3.8.

Reducing the bandwidth required to transmit voice between two parties not only improves spectrum utilization, but also saves power. As mobile phones have traditionally been highly resource constrained, such an improvement significantly improves the lifetime of such devices. Applying additional context-specific information can provide further power savings. For instance, during an average phone call, each user is likely to speak for approximately 50% of the conversation. Accordingly, there is no need for a device to transmit

when its user is silent. *Discontinuous Transmission* (DTX) mode addresses this condition by allowing a device's transmitter to be turned off when a user is not producing "useful" information [7]. In order to determine whether or not a user is producing useful information, DTX relies upon *Voice Activity Detection* (VAD) [26]. VAD analyzes the 20 msec frames generated by the RPE-LTP algorithm to determine whether or not speech is present. If only background noise is detected, the transmitting phone does not send a frame.

While saving a notable amount of power, pure DTX mode is often criticized for adding a decidedly mechanical characteristic to a conversation. Much like two individuals communicating via walkie-talkies, the detection of silence by VAD often makes a voice seem choppy and unnatural given the small detection window. The receiving side is also left listening to silence, which is also frequently confused with call disconnection. These problems are addressed through the addition of "comfort noise" [6]. When VAD determines that voice signals are no longer being sent, the phone transmits a final frame containing a *Silence Descriptor* (SID). This frame includes information to assist the receiving side to generate noise similar to that present on the transmitting side during periods of silence.

There are a number of reasons that individual frames of encoded voice are not received by a destination device. The air interface is inherently noisy and, while error correcting encoding can improve proper reconstruction, some frames can simply not be recovered. The network may also purposefully drop packets in a process known as frame stealing. Frame stealing allows a provider to conduct signaling to users already engaged in a call in-band. For instance, if the user needs to be alerted of an incoming call-waiting request, the network drops a frame of the current conversation to notify the user. For single frames dropped due to noise or for non-disruptive signaling requests (i.e., not call-waiting), devices can either repeat the previous frame or extrapolate the missing values from previous successful frames [9]. The loss of multiple frames is handled by temporarily muting the call.

These techniques allow for wireless providers to offer high-quality voice communications at greatly reduced costs when compared to previous analog systems. Digital encoding reduces the average bandwidth needed to transmit voice by a factor of four over analog methods. In combination with techniques such as DTX, these systems can provide power-efficient yet natural sounding voice between two parties.

3.5.4 Summary of Procedures

To complete registration, location update or call procedures over the air interface, elaborate signaling exchanges are performed over well structured control channels. The details of these control channels are discussed in later chapters. Here we discuss one subtle, but significant service provided on the air interface that provides some measure of privacy. We present it here because this eases the understanding of the discussion in latter sections of this chapter.

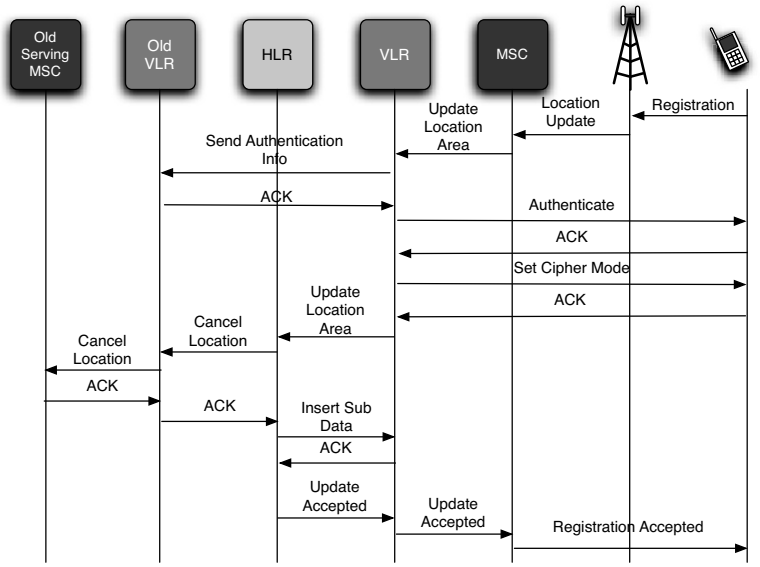

Fig. 3.9. The message flow for a mobile device registration.

To protect the privacy of a mobile device, during each registration and location update, the mobile device is given a pseudonym called a *Temporary Mobile Subscriber Identifier* (TMSI) that is sent over an encrypted channel. This TMSI is the identifier used by the mobile device when communicating over the air. In this way, an eavesdropper cannot track the location of a device. The TMSI used as the index into the subscriber records in the access portion of the network, i.e., by the VLRs and MSCs.

3.6 Registration and Call Setup Procedures

We now explain the operation of the mobile network through an example. Please refer to Figures 3.9 and 3.10 during the following discussion.

At a high level, the user registration with the network occurs as follows: The registration message (or location update message), is sent to the serving MSC/VLR. If the device has already registered previously, it uses its TMSI as its identifier. If the MSC/VLR do not have a record for the mobile device, the message is routed to the HLR of the user based on its mobile telephone number. After authentication procedures are performed, the details of which are discussed in Section 3.8, the HLR enters the current VLR serving the mobile device into its database. A copy of the user's service profile is then sent from the HLR to the VLR. At this time, the HLR points to the VLR, and the MSC/VLR has the ability to provide services to the mobile device.

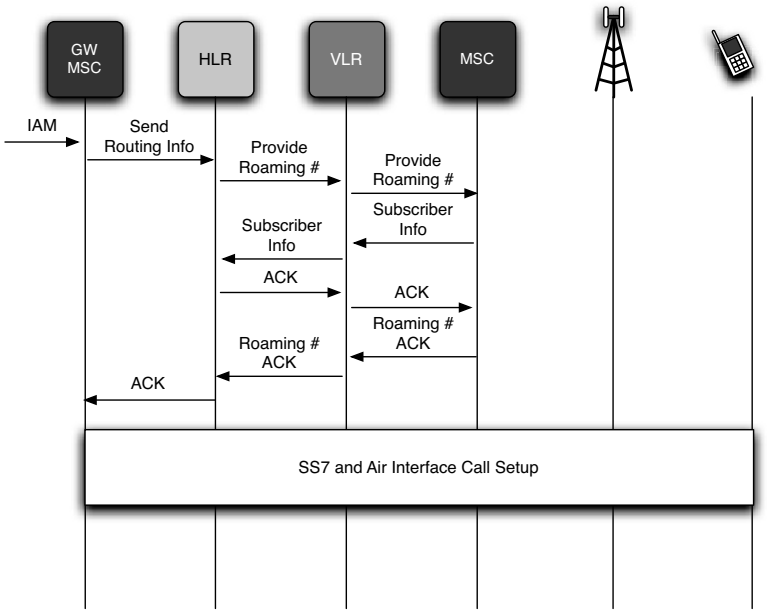

Fig. 3.10. The message flow for call setup to a mobile phone.

When a mobile device is powered on, or if it moves, it must register with the network using the procedures described above so that it can be located if a call for it is placed. While the HLR records the MSC/VLR currently serving the user, it is not aware of a user's current LA. Location updates to the HLR are only made when a mobile device's MSC/VLR changes or periodically to alert the network that it is still alive. A mobile device's serving MSC/VLR stores its current LA. A summary of the hierarchy of location information is shown in Figure 3.11.

Having registered with the network, a mobile device can then receive incoming voice calls. Such calls can originate from mobile devices within and outside a provider's network, from the PSTN or an external data network such as the Internet. Regardless, call setup begins by the transmission of an *Initial Address Message* (IAM). The gateway MSC receiving this message determines the HLR corresponding to the targeted phone number and then sends a request for that device's current location. Upon receiving this request, the HLR performs a lookup on the device and then queries the MSC/VLR listed as currently serving the device. If the MSC/VLR responds with the correct user profile information and the HLR acknowledges receiving this information, the MSC/VLR forwards a temporary local phone number by which a targeted device can be addressed. Having received a temporary routing address known as the *Mobile Station Routing Number* (MSRN), which lasts only as long as it

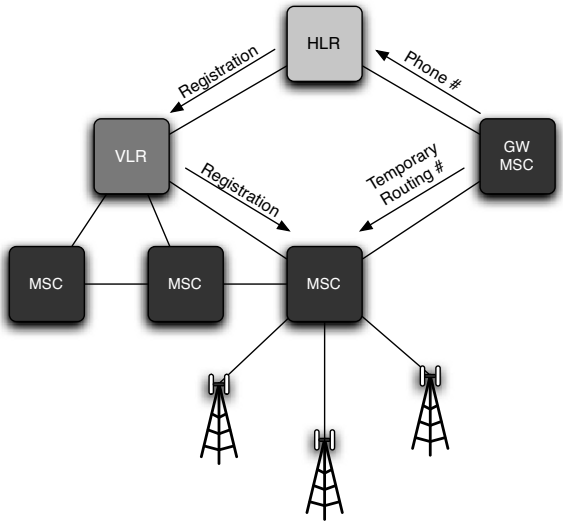

Fig. 3.11. The hierarchy of location information in a cellular network.

takes for the call to be established, the gateway MSC can route the call itself toward the serving MSC/VLR and ultimately the user. Chapter 5 provides the remaining details of how calls are delivered between the MSC/VLR and the mobile device.

Those readers familiar with Mobile IPv6 [101] will notice a number of high-level similarities between it and the SS7 network. HLR's, for instance, provide forwarding service much like a *Home Agent* (HA). The MSC/VLR implement much of the functionality associated with a *Foreign Agent* (FA). While the comparison is not strictly one-to-one, this abstraction may be helpful to readers.

3.7 Core Network Security

Recently, efforts to provide *Network Domain Security* (NDS) have been made to secure the core of the mobile telecommunication network. The aim of these efforts is to provide authentication of signaling messages between network nodes and networks, and to guarantee message integrity. Message confidentially can also be provided. The instantiation of NDS is a protocol targeted at protecting MAP called MAPsec [20, 24]. Thus, security is provided at the application layer.

MAPsec allows security associations to be established internally between network nodes and, more commonly, between networks. Shared cryptographic keys may be distributed manually or automatically using protocols such as IKE [84]. When activated, MAPsec can be applied to all or only a subset of

Table 3.3. MAPsec Protection Levels

Level	Invoke	Result	Error
1	PM1 (Integ.)	PM0 (No Protection)	PM0 (No Protection)
2	PM1 (Integ.)	PM1 (Integ.)	PM0 (No Protection)
3	PM1 (Integ.)	PM2 (Integ. + Conf.)	PM0 (No Protection)
4	(Integ. + Conf.)	PM1 (Integ.)	PM0 (No Protection)
5	(Integ. + Conf.)	PM2 (Integ. + Conf.)	PM0 (No Protection)
6	(Integ. + Conf.)	PM0 (No Protection)	PM0 (No Protection)

messages passing through the network. When MAPsec is used it may provide integrity only or encryption and integrity. Integrity is ensured by generating a CBC MAC using AES over the security header. Encryption is provided using AES counter mode.

MAPsec provides six levels of protection, known as *Protection Modes* (PMs) as shown in Table 3.3. The PM applied to a particular message of set of messages depends on the information communicated. For example, PM3 is suggested for use on messages related to authentication. The invocation of authentication does not contain any secret information, but the reply from the HLR includes the various keys used for encryption and integrity, and must therefore be kept confidential. Handover requests use PM4 because their invocation contains the TMSI. This must be kept confidential so an eavesdropper on the signaling network cannot correlate the mobility of a user; however, only the integrity of the response is protected.

MAPsec is not widely available. In fact, at a panel on telecommunications security at the 2007 USENIX Security Symposium, it was revealed that only a single deployment of MAPsec has ever been fielded. MAPsec's widespread incorporation into live networks was cut short by the substantial degradation it imposed on network performance. Accordingly, many vendor products simply do not include MAPsec capabilities and no known network current runs the protocol [47]. Performance issues aside, MAPsec can assist in combating some message insertion and modifications. However, it will not protect against propagation of attacks if they are launched from a legitimate network or node and a series of other attacks [110].

3.8 Air Interface Security

On the air interface of GSM, there are three main algorithms. The A3 algorithm is used for authentication, A8 for generating a cipher key, and A5 for ciphering data. All security operations are based on a 128 bit key, K_i, shared between the mobile device (actually, the SIM card in the mobile device) and the HLR.

A flow highlighting the authentication procedures is shown in Figure 3.12. When a mobile device first registers with a network, the VLR retrieves a set

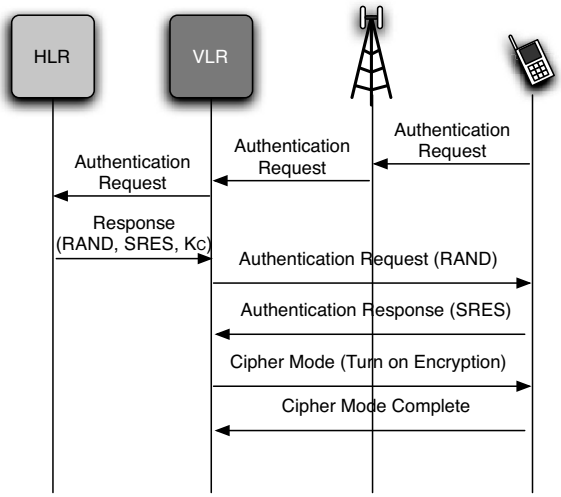

Fig. 3.12. The challenge/response authentication protocol for GSM.

of five triplets from the HLR: RAND, SRES, and K_c. RAND is a random value used by the VLR to challenge the mobile device. SRES is the expected response to the challenge. K_c is the cipher key to be used for communication. When the VLR receives the set of triplets, it challenges the mobile device with RAND. Using the A3 algorithm with K_i and RAND as input, the mobile device computes and returns a 32-bit SRES to the VLR which compares it to the expected value. If it matches, the mobile device is authenticated. In essence, the HLR authenticates the mobile device. The mobile device uses the A8 algorithm with K_i and RAND as input to generate the 64-bit cipher key, K_c used to protect subsequent communication.

The base GSM security solution has many weaknesses. First, the algorithms are supposed to be secret, but in fact have been found to be typically COMP128. As we will discuss in Chapter 4, COMP128 can be broken using a chosen challenge attack. Second, while the cipher key is specified as 64 bits, it is typically 54 bits padded with zeros and can be broken with a brute force attack. Third, the mobile device does not authenticate the network. This has led to false base station attacks in which an adversary can perform chosen challenge attacks and learn the identity of mobile devices by sending it messages. Finally, the use of encryption in these systems was terminated at the base station. Because wireless backhaul (e.g., microwave links) is often used to connect base stations to central offices, eavesdroppers could learn challenges and responses, mobile device identifiers, and most critically, cryptographic keys.

Many of these issues were address with the introduction of UMTS. The procedures in terms of message flows for UMTS are similar to those of GSM, but different information is carried in the messages. UMTS uses five algo-

rithms, F1-F5. The algorithms use a random number, RAND, some local material, a sequence number and a key, K, shared with the mobile, as input at the HLR. F1 outputs a MAC, F2 outputs a signed response (XRES), F3 outputs a cipher key (CK), F4 outputs an integrity key (IK) and F5 outputs an authentication key (AK). The sequence number, AK, local material and MAC are used to generate an authentication token, AUTN. Upon registration, the HLR sends the VLR a set of 5-tuples: RAND, XRES, CK, IK, and AUTN.

The mobile device is sent RAND and AUTN and has local copies of the local material and K. Using these items it can compute XRES, MAC, CK, IK, and AUTN. The AUTN is an authentication token for the network. If the AUTN computed by the device matches the AUTN it received, the device has authenticated the network. If so, it returns the XRES to the VLR for comparison. If the XRES supplied by the mobile device matches the XRES provided by the HLR, the network has authenticated the mobile device. In this way, mutual authentication is performed. CK is used by the mobile device to encrypt all communications. IK is used by the network to sign signaling messages. It is used because the network may command the mobile device to send information in the clear, i.e., disable ciphering. To prevent false base stations from performing this operation, signaling messages are signed with IK.

UMTS uses the public KASUMI algorithm as the basis of its algorithms. Also, encryption is carried back into the wired network to prevent eavesdropping. Combined, these improvements overcome most of the limitations of GSM air interface security from the perspective of stealing service and cloning.

3.9 Summary

In this chapter, we presented an overview of the architecture of a GSM network. We began by offering a historical perspective and providing an intuition behind some of the design decisions made in currently deployed networks. For instance, by understanding the weaknesses of analog cellular systems such as AMPS, it is possible to see why security appeared to be a largely "solved" problem in digital networks from the provider perspective. We then explored the nodes, protocols and physical connections needed to support both voice and data services in GSM networks. Issues ranging from the details of a user's profile and device registration to spectrum allocation offered insight into the many complex technical and political challenges facing real systems. We concluded by discussing the mechanisms used to provide security for both the core and wireless portions of these networks.

Our scope thus far has been intentionally broad. While the information covered in this chapter certainly covered some topics in depth, it is simply not possible to cover every aspect of these complex systems in a single book. However, readers with a good grasp of the information in this chapter will

be well equipped to begin tackling the security issues facing current and next generation networks. The vulnerabilities discussed throughout the remainder of this book can, in fact, are all discoverable using this material. We encourage readers with curiosity in one of the many aspects of the network that we have not discussed in explicit detail to begin their own research in the standards documents.