1

The Internet is Calling – Today's Network Ecosystems and Their Evolution

1.1 Introduction

In this chapter we briefly discuss, at a high level of abstraction, the different kinds of telecommunications networks, the technologies in common use in them today, and then explore why newer technologies continue to remain attractive to users, application developers, and service providers.

Time restrictions, and the need to provide a book that can be carried without the aid of a wheelbarrow, mean we cannot start at the 'very beginning' (since every story starts somewhere after the 'Big Bang'). But we will plunge into the history of telecommunications at a point that will provide readers with some background for the chapters that follow. We have included the key wireline and wireless network technologies in use today and have tried to give a flavor of each, emphasizing what makes them unique and picking out those aspects that are most relevant to the book. This chapter is not intended to be a thorough or complete tutorial and so interested readers are referred to e.g. [Faynberg 1996, Miller 2002, WAP] for more complete discussions of individual topics covered in this chapter.

So what is a network? At the highest level, any communications network may be visualized by the reader as being conformant to the abstract diagram in Figure 1.1. First, there is an access technology – this gives the user access to the network itself, and the services it hosts. Second, there is the core network infrastructure. Multiple access networks may 'hook into' the same core – this typically happens as networks evolve, and a strong need is felt to share services across them either for feature parity (within limits of reason, of course), or for reuse of deployed core network infrastructure and services across different terminal types or for other different reasons altogether. Third, there is a services layer that spans the core network as an overlay. This layer provides the real intelligence and value-add to the core that performs switching and related functions. The services layer also contributes directly to the end-user experience. Finally, we have gateways to other networks, for, as we will see in the main body of the chapter, no network can afford to be an island, and that interconnectedness increases value.

In this chapter, we first explore traditional wired telecommunications networks as this provides a natural lead into the discussions of wireless, WAP and other network technologies.

Parlay/OSA: From Standards to Reality Musa Unmehopa, Kumar Vemuri, Andy Bennett

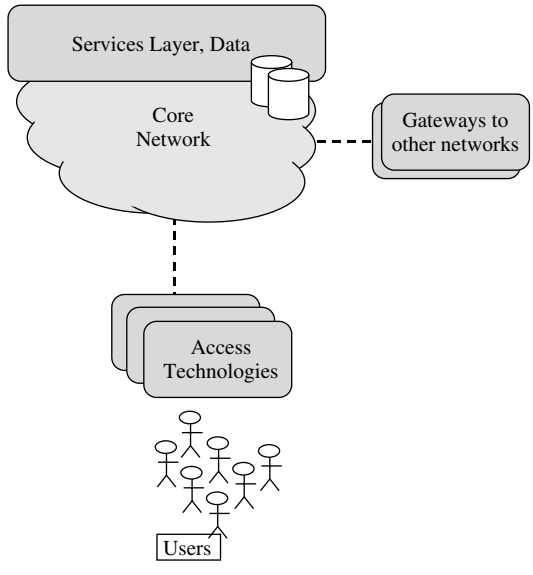

Figure 1.1 Overall logical network reference model

1.2 Traditional Telephony and Intelligent Networks

Since the dawn of time, man has been a gregarious creature with the need to communicate with others of his species. Communication has enabled us to transmit beliefs, traditions and inventions down through the generations and so escape the limits of evolution. People talk, gesture, whisper or find any means available to communicate ideas, feelings, warnings and secrets, and do so for a large part of their waking hours (and sometimes when sleeping too). When we have found a need to communicate over distance we have solved it, using sound (drums), light (beacons along the Great Wall of China) or electricity (the telegraph and the telephone).

However, it was the last of these, the telephone, that fundamentally changed the way people communicate. There was magical quality to hearing a person's voice over hills and valleys, oceans and seas that separated two people talking to each other. The world hasn't been the same since telephony took off in a big way in the third and fourth decades of the last century and things are only getting better, as newer capabilities to share text, documents, video or other media become more widely available.

To set up calls between two parties interested in communicating, one needs an element called a switch. Since not all phones are connected directly to all other phones in the world (the 'two cans and string' model is not very scalable), lines could be connected to a switch, which could link them together whenever the parties tied to those lines wanted to talk to each other. As the number of phones grew, so did the number of switches, and switches had to be connected together as well, to permit users connected to one switch to talk to users connected to others. This led to the birth of telecommunications networks.

Telecommunications networks started out as interconnected networks of switches that permitted users to make and receive simple voice calls (this was referred to as POTS – the Plain Old Telephone Service). A network is an ecosystem, defined in Webster's dictionary as 'the complex of a community of organisms and its environment functioning as an ecological unit'. Here, the network equipment (switches etc.), the user equipment, and interactions between these define the complex environment of interest. Just as in real world ecosystems, changes need to be carefully handled in the network.

For two parties to communicate using the network as a medium, network elements between the calling and the called party (sometimes termed the caller and the callee), need to somehow propagate first the desire to set up the communication path, and then the content of the communication itself, between the two entities. Note that the path is both physical in terms of the trunks tied up to carry the conversation, and logical in that the conversation between the two communicating parties is carried across it. The path between the caller and the callee is not permanent, but needs to be maintained for the duration of their conversation. Thus, the network elements supporting this interaction need to track some state associated with this call. This process is referred to as 'call processing'.

Every phone call between two parties engaging in a conversation was represented in the network in terms of a call model or a state machine on each switching element between the source and the destination[1]. In other words, each switch in a call path would execute a call model or state machine as call processing progressed and the two parties trying to communicate were connected.

Gradually, the need for services became more pronounced, due both to end-users maturing in their use of telephony related technology and in operators' desire to stabilize and expand their subscriber bases. To meet this need, additional features were introduced within call models[2] [Dobrowolski 2001] whereby special code was executed within the context of individual states in the call processing state machine, and new capabilities were provided to the parties involved in the call. These features, since they executed on the switching elements themselves, were typically referred to as 'switch side features'.

But there were issues with this architecture. For one thing, as each new feature was introduced, all switches that needed to support that feature had to be upgraded with this capability. Since there were many switches in the fabric, this kind of an upgrade was not easy to carry out transparently. Also, with the greater proliferation of telecommunications, people began to rely rather heavily on the network, and service outages (both planned outages for the service retrofits, as well as unplanned outages due to the vulnerability of having to update the entire complex fabric) became unacceptable.

Another problem was the degree of difficulty involved in making additions to existing switching logic, and then, testing these additions to ensure that new features did not interact with each other, and with the already deployed features in strange and undesirable ways. This was a far from trivial thing to do. Also, the software architecture of switches rarely allowed a sufficient degree of functional separation of services from the 'normal' logic flow and data structures, etc. relating to call processing. This led to serious issues with switch performance, as well as compromises in the design of the new feature itself.

To alleviate some of these concerns, and to provide a more flexible environment that could change more rapidly with the times as new feature capabilities were added, the Intelligent Network (IN) paradigm was introduced.

In the new (IN) model, switches are no longer merely simple executors of state machines. Service logic is separated from basic call control logic. Service features that were heretofore limited to being hosted on, and executed by, switches, are extracted from the switching elements and collocated into a separate physical element dedicated to execute the enhanced service logic for the newly introduced feature. Such a physical element is called a Service Control Point or SCP. The call model state machines at switches were enhanced to support the capability to query this SCP element (using a well-defined message set) and receive instructions that could be factored into their call processing operations.

Thus, switches now perform two functions – one (called the Call Control Function or CCF in the IN Distributed Functional Plane or DFP) that deals with the execution of the Basic Call State

[1] For readers unfamiliar with call processing and state machines, a very gentle and accessible introduction to these topics is provided in Appendix A [Parlay@Wiley].

[2] The concept of call models is only very briefly introduced here. Later chapters will expand upon this concept as they present more details relating to Call Control in the Parlay space.

Model (BCSM) that implements the call processing logic, and two (called the Service Switching Function or SSF in the IN DFP) that is concerned with the ability of the switch to interact with the SCP, request instructions, receive responses and so on. IN-capable switches are also referred to as SSPs or Service Switching Points. In Wireless Networks, these are sometimes also called MSCs or Mobile Switching Centers. But more about Wireless Networks later.

The SCP itself was a physical manifestation of two logical elements from the IN DFP – the SCF or the Service Control Function, and the SDF or Service Data Function. The former of these refers to the service logic that executes a relevant feature at the particular point in the call at which the switch sought SCP assistance, using the data that the switch provided in the request message to generate a suitable response. The latter refers to the capability whereby subscriber data or other data pertaining to numbers, translations etc., are hosted in a large database and made accessible to the service logic for use as appropriate as features execute[3].

With the proliferation of IN, one can still build in switch side features, but one has added flexibility in deploying new features and capabilities to better the end-user experience, through use of SCPs. Introducing new IN-based services in the network no longer necessitates updating all switching elements in the fabric. Rather, only the physically separated SCPs needed retrofitting. Basic service for call connectivity remains unaffected throughout such an update.

These concepts are illustrated in Figure 1.2. The reader is also referred to [Faynberg 1996, Chapter 5] for more details.

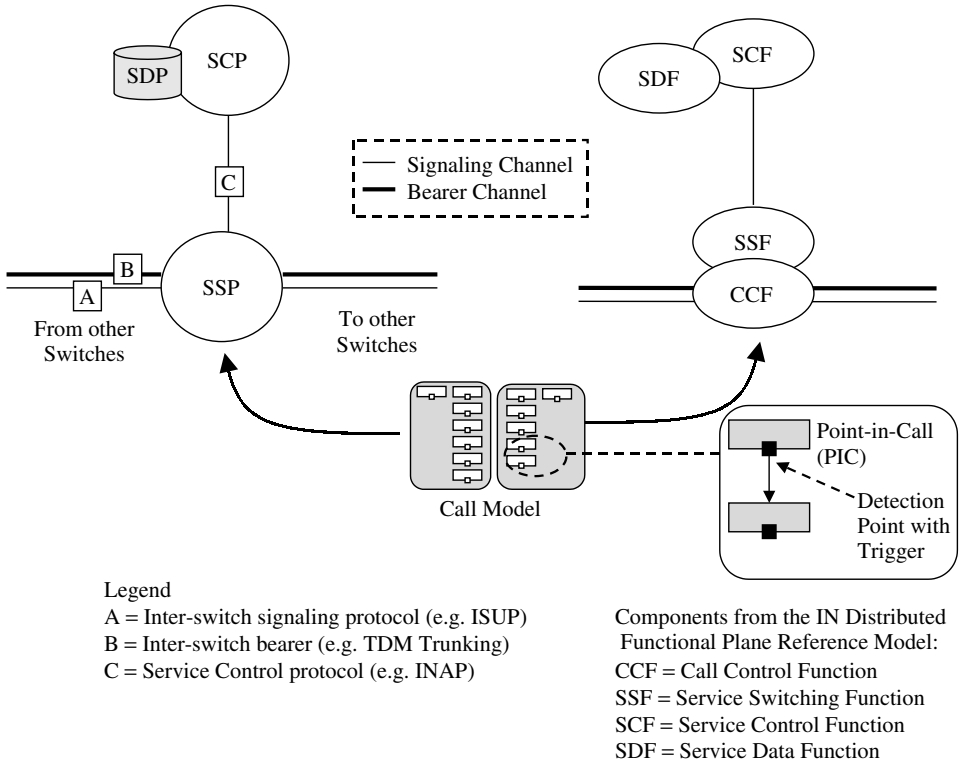

Figure 1.2 Switching components and IN call models

[3] This latter function may also be supported by a dedicated physical element standalone, in which case it would be called a 'Service Data Point' or SDP.

1.3 Signaling

When any two elements communicate they need to use a medium (such as a wire connecting them, for example), and a language they both understand, called a protocol. Communication may involve the transmission of data ('Watson, come here, I need you') or information pertaining to the data transmission (end-user Alexander G. Bell wants to connect a call to end-user Dr. Watson). The former is sometimes referred to as bearer (or payload) information, while the latter is called signaling.

Several signaling protocols are in existence today. Different networks use different signaling protocols. Different protocols are used between different types of network elements, and between the same network elements when they are involved in performing different functions.

Good signaling protocols are designed to be flexible and extensible for the addition of new parameters, messages or functionality, efficient in the number of bits of information that need to be transmitted between two communicating elements to share state or other information, and easy to process with minimal overhead.

Another characteristic of such well-designed protocols is that they are layered, such that each layer provides specific functional capabilities to the protocol as a whole, and the upper layers build on capabilities offered by the lower layers. Accessing these lower layer capabilities takes place through connect points called Service Access Points (SAPs), so they can be used in performing the tasks of the upper layer. The data unit supported by the protocol, and more specifically at each layer, is referred to as a Protocol Data Unit (PDU).

A layered architecture that is widely used in the design and operation of protocol stacks, called the Open Services Interconnect (OSI) data model, was developed by the ITU. This model, as shown in Figure 1.3, is composed of seven layers, and most protocols in use today adhere closely to it. The OSI model is discussed in greater detail in [Tanenbaum 2003].

A complete discussion of the design of good signaling protocols merits a book in itself. The interested reader is referred to [Holzmann 1991].

Signaling can be of different types, depending on where and how it is used. It can be classified in different ways, and In what follows, we study some of these ways.

One way of classifying signaling considers whether the signaling stream touches any end-user equipment (e.g. the phone on your desktop). The signaling link between end-user equipment and the network element (such as a switch) is commonly referred to as UNI or the User-to-Network Interface. Signaling links between network elements are referred to as NNI or Network-to-Network Interface.

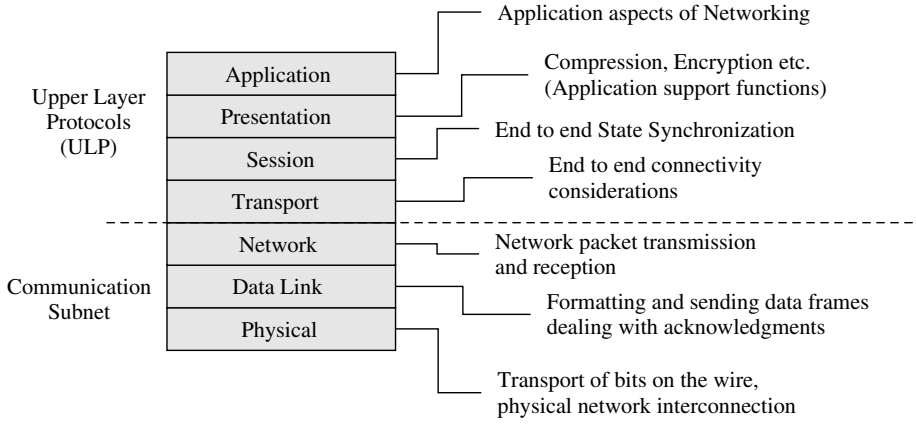

Figure 1.3 The Open Systems Interconnection (OSI) model layered protocol reference architecture

Another categorization considers the role particular signaling protocols play in the overall call flow. For example, user equipment to switch signaling, or switch to switch signaling during call setup, is referred to as call control signaling, while the communication that takes place between SSP and SCP elements is called service control signaling. Typically, different protocols are used in networks to fulfill each of these roles.

There are several other ways of categorizing signaling protocols, such as in-band vs out-of-band, etc. However, we do not study those distinctions for they are best left to books dedicated to signaling. The categorizations we cover above suffice for the purposes of the concepts we intend to develop later in this book, i.e. UNI and NNI, and Call Control and Service Control.

1.3.1 Signaling and Standards Bodies

As signaling pertains to communication among disparate elements in a complex networked environment, some form of agreement on the definition of these protocols is desired. Enter standards. Some standards are developed in bodies focused almost exclusively on data networks, while others are focused on voice communications, and some support working groups (WGs) fall into the gray area in between the two. In this section, we briefly look at some standards bodies of interest to this discussion in an attempt to give the reader a better feel for where and how the various signaling protocols are developed.

1.3.1.1 Telecommunications-oriented Standards Bodies

ITU – The ITU (International Telecommunication Union) is a specialized agency of the United Nations. With telecommunications networks spanning the globe, there is a need for standardization and regulation of such networks on the same scale, that is, globally. The mission of the ITU is to ensure efficient and smooth development and operation of telecommunications technology worldwide, and the general availability of this technology to the global population. As these globe-spanning networks were made up of an enormous mixture of national networks, interconnected by countless, often very specific signaling protocols, the Open Services Interconnect (OSI) data model, referred to earlier in this chapter, was developed by the ITU as a reference model for communications networks and their protocols. The development by the ITU of ISUP as the signaling standard for bearer traffic and INAP as the signaling standard for service control served as a major catalyst for the global proliferation of digital circuit switched telephony networks. With IP networks reaching the same ubiquity, the ITU developed H323 as the international standard for session oriented communication over the Internet [ITU].

3GPP™ – the 3rd Generation Partnership Project (3GPP) is a partnership of regional standards bodies that defines the standards for GSM-based wireless networks and for their evolution into a third-generation UMTS architecture. 3GPP provides several technical specifications aimed at addressing specific interfaces, services and network elements from within its reference architecture [3GPP].

3GPP2 – the 3rd Generation Partnership Project2 (3GPP2) is a partnership of regional standards bodies that defines the standards for CDMA-based wireless networks just as 3GPP performs similar functions for GSM technologies. Given the large overlap in technical directions and architecture between 3GPP and 3GPP2, the latter has agreed to reuse the specifications issued by the former body wherever applicable. In addition, most recently a harmonized reference architecture (called IMS or the IP Multimedia Subsystem) that melds both the 3GPP and the 3GPP2 models has been adopted to further drive convergence in the work being done in these two bodies [3GPP2].

1.3.1.2 Data Network-oriented Standards Bodies

IETF[4] – The Internet Engineering Task Force (IETF) is an organization that hosts numerous working groups dedicated to developing protocols and standards that govern network element

[4] There is also a research wing that parallels the work done by the IETF, called the IRTF (Internet Research Task Force) and also run by the same body, the ISOC. This body does more of the 'forward looking' work,

communications within the Internet, and other Internet Protocol (IP)-based networks. In fact, IP was itself designed by this body [IETF].

Among the numerous IETF WGs, the following are of immediate interest and relevance to our current discussion[5]. A brief summary of the work carried out in each of these groups is provided below:

1. *Iptel* – The Iptel working group designs standards for use in supporting telephony over the Internet Protocol, specifically (inter-domain) routing of voice calls over the Internet.
2. *SIP* – This WG is focused on developing the base Session Initiation Protocol and extensions that enable it to be efficiently used in setting up and tearing down multimedia sessions. This WG was spawned off earlier work accomplished under the charter of the MMUSIC (Multiparty Multimedia Session Control) WG of the IETF.
3. *SIPPING* – Session Initiation Protocol Project INvestiGation is dedicated to studying the applications of SIP and non-base-protocol extensions in support of SIP applications.
4. *PINT* – The PSTN/Internet Interworking WG deals with scenarios where an end-user connected to an IP network such as the Internet can request services from an SCP in the PSTN network. Examples of such services include Click-To-Dial (CTD), where a user clicks on a link or submits an HTML form and causes a call to be set up between herself and a customer service representative representing a business.
5. *SPIRITS* – The Services in PSTN/IN Requesting InTernet Services WG addresses scenarios that are an exact converse of PINT scenarios. So the focus here is on services in the PSTN/IN that require IP-host based feature assist capabilities. Internet Call Waiting (ICW) is an example of such a service. If a user is connected to the Internet via his phone line through a modem, incoming call notifications can be piped to the user via that Internet connection even though his phone line is busy at the time. Both PINT and SPIRITS recommend the use of SIP as a signaling protocol.
6. *Sigtran* – The Signaling Transport WG has produced several protocols including SCTP (Stream Control Transmission Protocol) and others that define the lower layers of a protocol stack to enable the transparent transport of SS7-based protocol payloads over IP. The intent here is to promote seamless convergence where possible through use of upper layer protocols (OSI layers four and above) across network types.
7. *Megaco* – The Media Gateway Control WG developed, in concert with the ITU, the Megaco protocol (also referred to as H.248.1) that defines the communications between Media Gateway Controllers and Media Gateways. (See Section 1.4.2 on 'Converged Networks' in this chapter for more details.)

1.3.2 Some Examples of Signaling Protocols

In traditional telephony networks (also called the Public Switched Telephone Network (PSTN) in wired contexts or Public Land Mobile Network (PLMN) in wireless contexts), switches communicate with each other over SS7 (Signaling System #7)-based signaling protocols [Russell 2002].

For example, switches in the PSTN utilize different protocols for user equipment to switch signaling (e.g. Ear & Mouth (E&M) Protocol, Telephone User Part (TUP)), switch-to-switch signaling (e.g. ISDN User Part (ISUP)), and switch to SCP signaling (e.g. IN Application Protocol (INAP)). As explained previously, the first two of these are typically called call processing signaling, while the last of these is referred to as service control signaling. All these are SS7-based.

---

and has made significant contributions to protocols in the area of AAA, SPAM-filtering etc. [IRTF]. The AAA work has since been absorbed into the IETF AAA WG.

[5] In later chapters, work being done in other IETF WGs may also be introduced as appropriate. This listing is merely intended to give the reader a taste for the kind of work the IETF undertakes.

The Internet, the largest, most widely prevalent, almost ubiquitous network today utilizes the Internet Protocol or IP as the basis for communication between computers. Various application level protocols ride atop IP to provide a range of functional capabilities between communicating applications on computers connected to this network. Examples of these protocols include Simple Mail Transfer Protocol (SMTP) for email, File Transfer Protocol (FTP) for file transfers, Hyper Text Transfer Protocol (HTTP) for Browser to Web Server interactions, etc. Most IP-based protocols use either Transmission Control Protocol (TCP) or User Datagram Protocol (UDP) as the layer-4 protocol of choice. The reader is referred to [Comer 1999, Comer 2000] for more details on IP.

Both the SS7 and the IP protocol suites are compliant with the OSI model.

1.4 A Foray into Other Network and Service Architectures

In this section, we discuss some other network and service architectures of interest. A good understanding of some of these will be useful in later chapters as we address how Parlay and OSA technologies relate to them. Others are introduced to give the reader an appreciation of how network evolution takes place, and the different generations of related technologies as new standards are defined.

1.4.1 Voice over the Internet Protocol (VoIP)

Metcalfe's law states, 'The usefulness, or utility, of a network equals the square of the number of users'. The very ubiquity of the Internet, low barriers to the entry of new endpoints, and the overwhelmingly large number of users, along with its ability to carry various protocols that perform different functions leads to increased value per Metcalfe's law, and a positive feedback loop that continually contributes ever more to its growth.

This, combined with a widespread interest in utilizing the Internet for voice communications, has led to voice becoming one of the most widely transmitted payloads on the Internet today[6]. The use of IP to transport voice is referred to as Voice over IP (VoIP). The Internet's inherent ability to transport data, pictures and other visual media such as video in concert, and potentially interleaved with, voice, leads to a truly powerful multi-media user experience.

As with PSTN/PLMN networks, voice-, or more generally multimedia- session setup requires some session setup, processing and teardown signaling, in addition to the ability to transport bearer information over IP. This support signaling can be provided using various protocols. H.323 (developed in the ITU [H.323 2003]) and SIP (Session Initiation Protocol [RFC 3261], defined by the IETF) are popular IP-based protocols for this today. The former is still widely used, while the latter, widely acknowledged to be the protocol of choice for the future, continues to gain in popularity.

IP-based telephony does not have as clear-cut a partitioning between service-related and call-related signaling. The traditional telephone network supports almost all the user services needed, with a minimal set actually supported by the user handset in a manner independent of the network. In contrast, the IP-based telephony model supports a near equal, if not skewed in favor of the handset, distribution of services between the network and the handset domains. This means end-users have greater flexibility in the kinds of services they can access (since this is handset dependent), but also means user reliance on the terminal is greater than in the PSTN world (e.g. user data are hosted more on the user terminal than on the network, so if the user has to initiate a session from a

[6] In fact, with the prevalence of high-speed Internet connections such as those using cable or DSL lines, VoIP technology is now really taking off in a big way. VoIP service companies like Skype [Skype] and Vonage [Vonage] are seeing a sharp uptake in their subscriptions over a high-speed Internet access infrastructure. What seems really interesting with some of these services is that one gets a real phone number assigned to the 'always on' high-speed Internet connection, and this number is not 'geographically bound' – the user can get a local phone number in the New York area, while residing in London, and can use this transparently, without the caller knowing his or her actual physical location. Judicious choice of phone number can cut down on long-distance bills, especially if one makes more calls within a particular area code.

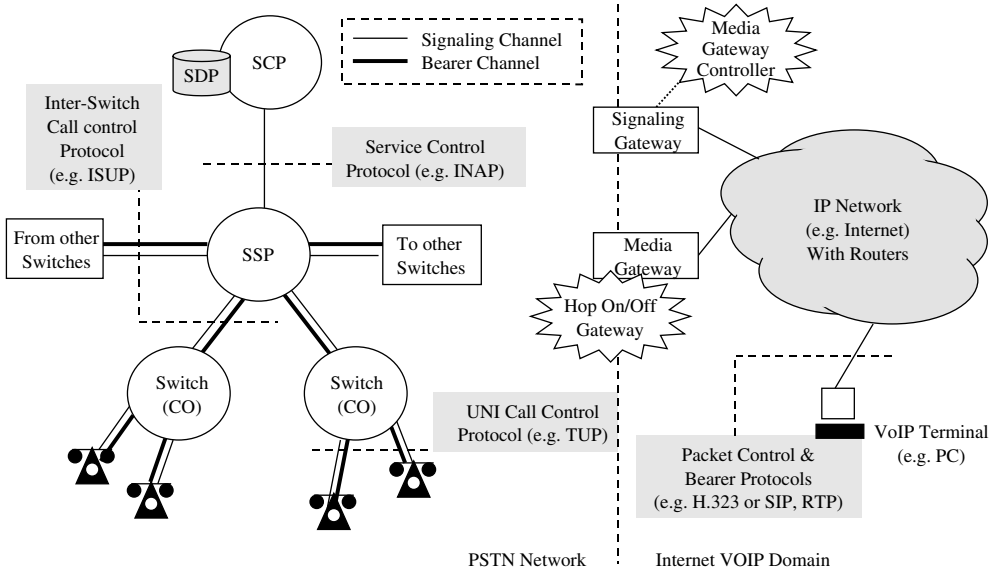

Figure 1.4 The PSTN and IP networks today, and VoIP

different terminal, the experience may be less pleasant than if the data were stored on the network and available to him transparently).

Lately, some protocols, even in the IP-domain have started adding mechanisms to provide support for service control related signaling. But rather than define new protocols aimed specifically at service control, in most cases they have relied on extensions to the base signaling protocol to fulfill these needs as well. However, it is still possible to draw rough analogies between the traditional IN architecture in the PSTN and the IP architecture for multimedia call setup (Figure 1.4). Some of these details for a specific protocol (namely SIP) will be covered in later chapters.

1.4.2 Converged Networks[7]

Isolated networks of users who cannot communicate with users of other networks still feel isolated, though not necessarily alone. Ubiquity, ease of network access, and interconnectedness, contribute towards a feeling of community. The PSTN is useful because it permits a user at any phone connected to it to call another user at any other phone. Connectivity contributes to value. Recall Metcalfe's law.

Definition: Gateway

Where network elements belonging to different networks, and using different protocols, but providing similar functions within their own network contexts need to communicate, an element called a gateway, that speaks both protocols, is used to mediate between these two elements. The gateway element, in its simplest form, functions as a protocol translator, and enables the

[7] Networks today include Cable and DSL access technologies as well. These are used, for instance, to support high-speed broadband Internet Access. In late 2004, in the US, broadband Internet access, for the first time in networking history, surpassed dialup access to the network. For the sake of simplicity, these Cable and DSL aspects are not depicted nor explained in any detail in this chapter.

two elements, one from each network, to talk to each other. Gateways form the basis for most convergence in networks today.

When viewed from the perspective of a signaling flow, a gateway through which a flow enters a given network is typically referred to as a 'hop on' gateway, while one through which it exits is called a 'hop off gateway' (Figure 1.4).

As VoIP took off, service providers gradually came to view the Internet, or other managed IP networks, as a means to offload some of the voice traffic to a more cost efficient, less resource constrained environment that supported more optimized routing (and tied up less resources during session setup). In addition, Internet users wanted the ability to call telephones connected to the traditional PSTN or PLMN. In order to support these and other similar needs, convergent architectures came into being.

Convergence may be achieved in a variety of areas. Convergence in terms of signaling transformations, as call processing signaling transits an IP to PSTN or PSTN to IP network boundary, may be carried out at network elements called Signaling Gateways.

Bearer stream transformation as it transits a network boundary of the kind described above is carried out at network elements called Media Gateways (MG), and is commonly referred to as transcoding[8].

If the signaling stream controls a media or bearer stream associated with it, the Signaling Gateway is also referred to as a Media Gateway Controller (MGC), for not only is the signaling transformed as the network boundary is crossed, but the associated media characteristics are also controlled as this transformation takes place, through interaction with the Media Gateway element where the bearer stream is being transcoded. Megaco ([H.248.1], jointly developed by the ITU with the IETF) is the IP-based protocol of choice for MGC to/from MG communication.

Convergence can also be achieved in the services domain. This is of great interest to service providers and also to end-users. If services originally developed for one network could be transparently used in another, then this offers great benefits. For one thing, it saves money while promoting complete and immediate feature parity. And the immediate availability of all existing services in a new network context does wonders for the end-user experience and in meeting user expectations.

The specific signaling protocols supported both for call/session control and for service control may vary based on the specific domains being inter-worked. The degree or ease of inter-working may also vary depending on how closely the call/session and service state models align between the two types of networks in question.

IP-based telephony might want to reuse IN elements in support of providing deployed features to VoIP users. This could be supported by carrying the IN service control protocols over IP for example. This forms the basis of the Sigtran work in the IETF.

IN SCPs could be enhanced to interact with IP-based application servers for new feature logic that is shared between the IN and IP domains. Work in this area has been done in the IETF PINT [RFC 2848] and SPIRITS working groups [RFC 3910] to support end-user capabilities such as Click-To-Dial (CTD) or Internet Call Waiting (ICW). For more on PINT and SPIRITS, the reader is referred to [Kozik 2000, Gurbani 2003].

These are but two examples of service reuse. Several other elegant models [Vemuri 2000] have been discussed in the literature. For more information on services for converged networks, the reader is referred to [Faynberg 2000].

[8] Media is typically encoded in some bit format for transport across the network. This encoding is done using software called a codec. Since voice is encoded using one set of codecs in the Internet domain, and another different set of codecs in the Telecommunications domain, the operation of switching the encoding scheme or codecs at the media gateways is referred to as 'transcoding'.

1.4.3 Internet Access via the PSTN

It will be useful for the reader to have a basic understanding of how one may access the Internet via the PSTN – the so-called 'dial-up' Internet access. In this section, we provide a high-level summary description of the same. Internet access has also evolved with time. Initially, this was achieved through the use of modems that enabled users to dial in via analog phone lines – this was somewhat slower, and offered speeds in the 56–128 Kbps range depending on the specifics of the modems in use. More recently, broadband Internet access has seen widespread growth where the use of cable modems and digital subscriber lines (DSL) enables subscribers to achieve speeds in the megabits per second range. Again, in order to keep the discussion simple, and since we merely aim to give the reader a flavor for how the basic technology works, we explain only the dial-up access.

An ISP (Internet Service Provider) typically supports modem pools at several geographic locations called POPs (Points of Presence). When the user dials the phone number associated with a POP (and if there is one in your area, you may not have to pay long-distance charges) by running the appropriate dialer software on her computer that talks to her local modem, the modem from the pool answers the phone. Once the basic call is set up, the PPP protocol (Point to Point protocol, developed by, you guessed it, the IETF) runs across the link and sets up the data connection.

The modem pool is collocated with a NAS or Network Access Server, which also functions as the AAA client. This element interacts with a AAA server hosted within the service provider network (typically over protocols such as TACACS [RFC 1492], or RADIUS [RFC 2138] or, more recently, DIAMETER [RFC 3588]) to perform end-user authentication, authorization and accounting procedures, and sets up filters for IP traffic that transit the user connection (PPP link) and IP network for that session. An IP address may be assigned to the computer for the duration of the session (for packets to flow back to it), using either a statically assigned pre-configured ISP-owned address (relatively rare), or a dynamic address obtained through protocols such as DHCP [RFC 2131] or IPCP [RFC 1332].

Once the session is established, the end-user can transmit and receive data from her computer over this link. Once done, the user simply hangs up, and the IP address (if dynamically assigned), becomes free for reuse for other user sessions, as does the port on the modem pool. Figure 1.5 illustrates dial-up access to the Internet.

1.5 Wireless Networks and Generations of Technology

We started this chapter by looking at very abstract network architectures and introduced call/session control signaling and service control signaling, after which we explored how wireline networks have evolved specifically in terms of their protocols and interfaces and the network elements processing these. Now it is time to look at wireless networks.

So far, we have used the term 'wireless network' in a rather generic sense to refer to networks of mobile terminals built to support cellular technology. In this section, we examine these kinds of networks in a little more detail. We shall briefly introduce the concept of cellular communication and then describe wireless networks in terms of their network elements, their signaling protocols, and the service data they store for subscriber services. We shall then look at how the circuit switched core of wireless telephony networks has been expanded with a packet switched domain to support mobile access to services residing in data networks. After that, we will describe how third generation wireless networks are evolving from current mobile communication systems by introducing a new radio access technology and by further evolving the core network.

Generally speaking, wireless and cellular networks are not strictly the same. Every cellular network is a wireless network, but not all wireless networks need necessarily operate using cellular technology. WiFi (Wireless Fidelity or IEEE 802.11b wireless networking) is an example of localized wireless networking that does not operate on cellular technology. For the purposes of this section, we shall use the term 'wireless' to mean cellular in a generic sense.

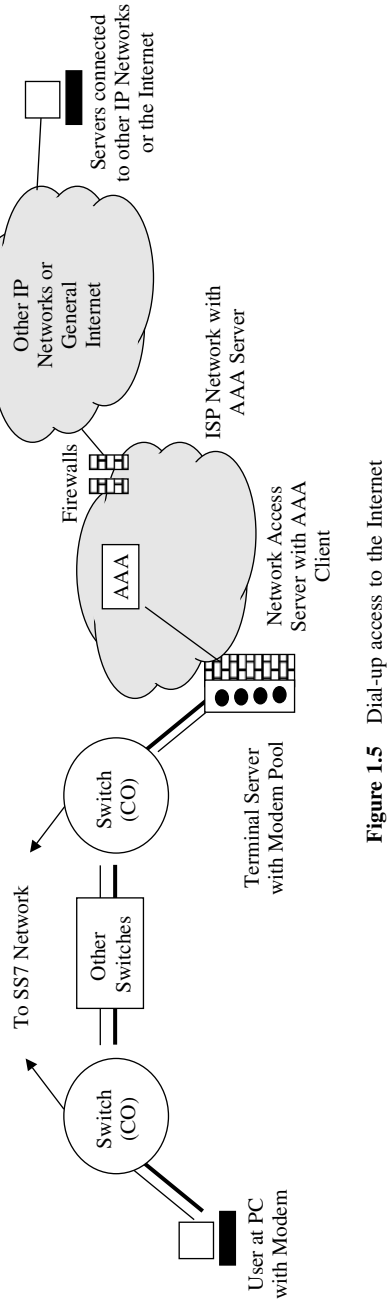

Figure 1.5 Dial-up access to the Internet

1.5.1 Cellular Communication

Wireless (or, to be more precise, cellular) networks are assigned a certain frequency band to use for setting up radio links to the mobile devices of their subscribers to complete the communication path. When engaged in a phone call, the mobile device is allocated a certain frequency in the available spectrum. As radio waves form a shared medium, for the duration of the call, this frequency (or a timeslot in a specific frequency, depending on the details of the wireless radio technology in use) is uniquely assigned to that specific mobile device, in order to avoid interference. This means that the capacity of a mobile network is confined by the number of unique frequencies one can assign within the available spectrum. Cellular systems address this issue of limited capacity by dividing the coverage region of a network in largely non-overlapping areas, called cells. Frequencies are then reused in non-neighboring cells, to increase the overall network capacity.

Cell sizes may vary depending on the area they cover, the technology in use, and in the frequency spectrum utilized by the technology in question. For example, in rural environments a typical cell size may be larger than in the city, as the total number of concurrent mobile phone calls can be safely expected to be lower and hence less reuse of frequencies is required. Also, cells may vary in shape. In dense urban areas the cells may be evenly shaped and arranged like roof tiles or the scales of a fish collectively to cover an entire downtown area. Along major highways or subway and train lines the cells may be stretched in length to offer travelers and commuters continuous radio coverage, whereas on either side of the route coverage may drop quickly.

Although one tries to achieve ubiquitous coverage with cellular technology, sometimes, in the interiors of large buildings or such hard to reach places (for the radio signal), no coverage may be available to make or receive cell-phone calls. Such areas are termed ‘urban canyons’.

1.5.2 Wireless Networks and their Elements

Wireless networks, irrespective of the specific technology deployed, all share a similar network architecture. This wireless network architecture is depicted in Figure 1.6, in which we also recognize of course the overall logical network reference model introduced in Figure 1.1, with the separation of access, core, and services.

One of the most successful wireless network technologies, in terms of global deployment, is GSM. We will draw on GSM to introduce and further define wireless networks, and their network elements and signaling protocols. Interested readers are referred to [Mouly 1992] for an excellent

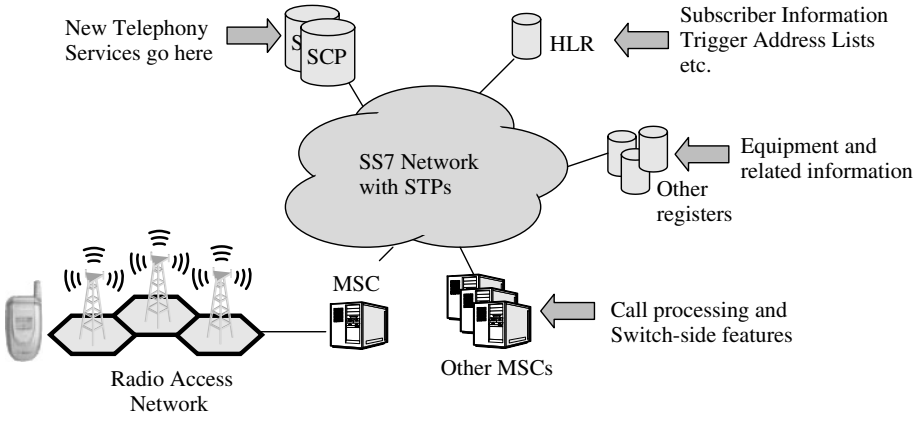

Figure 1.6 Sample wireless network architecture

coverage of GSM technology, in its full breadth and depth. Wireless networks based on CDMA technology will be covered later in the chapter.

GSM networks reuse much of the PSTN and are SS7 networks at their core. Inter-switch signaling is based on the same SS7 protocols deployed in PSTN networks, allowing for large scale reuse whilst smoothly facilitating fixed-to-mobile and mobile-to-fixed calls.

Mobile devices communicate with the network via a radio link to a Base Station System (BSS), consisting of a Base Station Transceiver (BST), or the 'antenna', and a Base Station Controller (BSC). The BSC communicates with the Mobile Switching Center (MSC), connecting the radio part of the network with the SS7 core of the network. The MSC, which is the telephony exchange in GSM networks, performs the basic call processing procedures and interconnects with other MSCs or with the PSTN or ISDN exchanges for network connectivity, via the SS7 core. So mobile telephony systems like GSM only make use of radio resources at the edge of the network, when completing the last step of the communication path to the mobile device.

An MSC differs from a PSTN switch in that it serves mobile devices rather than fixed phones. As mobile subscribers have the freedom of picking up their phone and moving about, contrary to fixed phones and a PTSN switch, there is no static relationship between a mobile device and a specific MSC. Depending on the location, a mobile device is served by a given MSC, which is referred to as the serving MSC. As the serving MSC may alter when the subscriber is changing location, all information pertaining to mobile subscribers is located in a centralized database, called the Home Location Register (HLR) – this is a fundamental difference with the PSTN: as terminal mobility is supported, there is a registry like the HLR that is maintained in wireless architectures. An MSC may query the HLR to obtain service subscription profiles for a given subscriber, or routing information required to locate the subscriber in the network in order to complete an incoming call destined for that subscriber. Originating services are also deployed at the HLR. If for instance the subscriber is not allowed to receive calls while she is registered with another network in a foreign country[9], this information will be stored in the service subscription data of the subscriber. So in the case where the HLR will be queried for routing information with the intention of terminating a call to the subscriber, the HLR will return a decision not to allow further processing of the call and the attempt will be rejected (or barred).

For the purpose of minimizing the need to perform database queries to the centralized HLR database, a temporary local copy of the subscriber data is stored in the Visitor Location Register (VLR) associated with the serving MSC. The VLR record includes information required to page the mobile device and perform call setup procedures. Information relating to so-called terminating services is stored in the VLR as well. An example of a terminating service is 'Call Forwarding on Not Reachable', e.g. when a terminal is switched off. This is a terminating service as only after paging a mobile device, is the not-reachable status for the device established. It is the serving MSC, using service subscription information from the VLR record, that will perform the service logic involved with terminating services, without having to interrogate the HLR.

As service subscription data may change over time, the data stored in the VLR need to be maintained in synchronization with the data kept in the HLR record. Whenever changes occur in the HLR record, the VLR record gets updated. Also, as subscribers move around, they may cross MSC boundaries. As VLR records are associated with the serving MSC, such a crossover (or inter-MSC hand-off) will result in the creation of a new VLR record and the deletion of the old one. The signaling protocol for HLR to VLR communication is the MAP protocol (Mobile Application Part), which is an SS7 based protocol.

As is the case in PSTN networks, IN-based services can be applied in GSM networks as well. In this case, the IN system is referred to as CAMEL (Customized Application for Mobile Enhanced Logic). The service control protocol between MSC and SCP is the CAP protocol (CAMEL

[9] Such a service may serve to protect the subscriber for incurring the additional costs associated with receiving incoming calls when roaming, or it may be applied by the operator for subscribers who have overdrawn their user account.

Application Part), which, and this will not be a surprise by now, is SS7 based[10]. Similar considerations also apply to CDMA architectures, which we will see later on.

In addition to the HLR and the VLR, the SCP now introduces a third location for service data pertaining to the mobile subscriber, and a third location for service logic execution. A MAP (Mobile Application Part) interface is introduced between the SCP and the HLR to ensure service data does not conflict and undesired feature interactions are avoided. Also, the trigger address lists for the CAMEL services of a given subscriber are stored in the HLR.

1.5.3 Evolution of 2nd Generation Wireless Systems

Wireless networks as introduced above are referred to as second generation wireless networks, as they embody the progression from analog technology (the first generation) to digital communication. The second generation GSM network is a circuit switched communication system, seeing that a fixed route through the network is established between the parties, for the entire duration of a call. With the advent of packet switched technologies, and the type of always-on, IP-based services that are facilitated by these technologies, GSM networks evolved by adding a packet domain to the circuit switched core network. The packet domain is used to transport packet data efficiently across the GSM network, from a mobile device to external packet networks. This new GSM bearer service is called General Packet Radio Service, or GPRS.

The first order of business in realizing the packet domain is the introduction of packet switches required to route packet streams. These packet switches are called Serving GPRS Support Nodes, or SGSNs, and their main function is to route the packets to the mobile device and vice versa. As with MSCs, a notion of serving SGSN applies and a VLR record is associated with the serving SGSN. The HLR continues to be the centralized place where subscriber data and service profiles are stored.

In order for the SGSN to transport the data packets to external packet data networks, a Gateway GPRS Support Node (GGSN) is introduced. One of the functions of a GGSN is to perform the translation of GPRS data packets into the data protocol in use within the external packet network. Similarly, an address scheme conversion is required in order to deliver packets originated in an external packet network to a mobile device in the GSM network.

Within the GSM network, a GPRS backbone network is in place between the SGSNs and the GGSNs to carry the data packets. As there may be several external packet networks, e.g. IP or X.25, packet gateways (GGSNs) are required for each such external network. However on the GPRS backbone all packets look alike, as external packets are encapsulated and tunneled across the backbone[11]. A specific session that may exist within a tunnel on the GPRS backbone, established between a GPRS-capable mobile device and a specific address in a given external packet network, is called a Packet Data Protocol Context, or PDP Context. With a PDP Context, a GPRS-capable mobile device in the GSM network is now addressable by entities in the external packet network, and payload packets can be exchanged to and fro.

CAMEL capabilities are in place to allow for IN-based service control of PDP Contexts. To support such service control, a CAP interface exists between the SGSN (or the gprsSSF to be exact) and the SCP.

GPRS itself evolves to EDGE (Enhanced Data-rates for GPRS Evolution), which is sometimes informally called 2.75G. This evolved form of GPRS technology results in increased data rates without any changes to the underlying core network. EDGE is not further discussed in this book.

[10] The reader should note that 3GPP Release 4 is also tending towards including support for TCAP/IP type scenarios as the network continues to evolve. Such work has been in progress for a while in other standards bodies like the IETF for a few years now, where the underlying transport mechanisms for carrying SS7 protocols were being developed. The interested reader is referred to [Sigtran] for more details.

[11] The signaling protocol used on the GPRS backbone is called the GPRS Tunneling Protocol, or GTP. For the remainder of the material addressed in this book, GTP is not important.

1.5.4 Third Generation Wireless Systems

Two developments characterize the dawning of the third generation in wireless networks. The first improvement is the launch of a new radio technology introducing higher data rates, advances in DSP technology and more efficient use of radio spectrum. The second advancement is the establishment of an all-IP core network.

The radio technologies in use in second generation wireless networks are based on frequency division multiplexing, where each connection uses its own dedicated radio frequency, or time division multiplexing, where each connection uses a dedicated frequency only part of the time, in fixed time slots. There are generally two drawbacks with FDMA (Frequency Division Multiple Access) and TDMA (Time Division Multiple Access), and those are that adjacent or nearby frequencies interfere with each other and the fact that each frequency can only be used for one connection (in any given time slot).

Determined to condemn such drawbacks to history, spread spectrum technology emerged that allows multiple mobile devices to use the same time slots and frequencies at the same time. Interference is avoided by cutting up the speech payload of all active mobile devices into tiny fragments and transmitting all of them simultaneously over the radio link. A unique code is assigned to the speech segments of each individual connection. So even though all communication data are shared over the airwaves, any given mobile device will be able to distinguish and identify the speech payload destined for it, by means of the unique code for its speech connection. This radio technology is termed CDMA (Code Division Multiple Access).

The second advancement in third generation wireless networks is the establishment of an all-IP core network. With the evolution of 2G (second generation) networks we have seen that GPRS adds the possibility for a mobile device to connect to external packet networks through the GSM network, and obtain services residing and executed in those external networks. To facilitate the exploitation of increased efficiency and enriched service capabilities made possible by IP technology, wireless networks need to advance beyond the capability of offering access to external packet networks. This trend is visible as an evolution of the nucleus of wireless networks into an IP core. In 3GPP this core is called the IP Multimedia Subsystem (IMS). The principal objective of IMS is the realization of an integrated voice and data network infrastructure, capable of delivering multimedia capabilities, be it real-time or otherwise. IMS is gaining wider industry acceptance and is likely to see widespread deployment by the time this book is published. Later sections in this chapter introduce IMS in a bit more detail.

1.5.5 CDMA Network Evolution

Broadly speaking, CDMA networks evolve along similar lines, though the details are somewhat different (and a discussion of these finer points merits a book in itself). A high-level summary view is presented here. The 2G CDMA networks evolve forward to support CDMA 1X-RTT (Radio Transmission Technology) – a technology that provides for more efficient over the air interfaces and higher bandwidths. An overlay network, called CDMA 1X-EVDO (Evolution for Data Optimized, sometimes also called Data Only), may also be deployed to support packet traffic as the evolution continues forward. EVDO was not designed to support voice[12], and so CDMA 1X-RTT evolves forward into an integrated packet infrastructure with CDMA 1X-EVDV (Evolution for Data and Voice). CDMA evolution, with support for CDMA 1X-EVDO is depicted in Figure 1.7. The reader interested in learning more is referred to [Viterbi 1995].

CDMA 1X-EVDO supports nodes such as the PCF (Packet Control Function), and the PDSN (Packet Data Serving Node), and these roughly translate, at the highest layer of abstraction, to elements similar to the SGSN and GGSN from GPRS networks. While the GPRS networks support

[12] Strictly speaking, EVDO evolves towards EVDO Rev A also sometimes referred to as DOrA (read 'Dora') that can in fact support VoIP. EVDV, the next phase of the evolution, provides for higher bandwidth and increased data rates over and above EVDO.

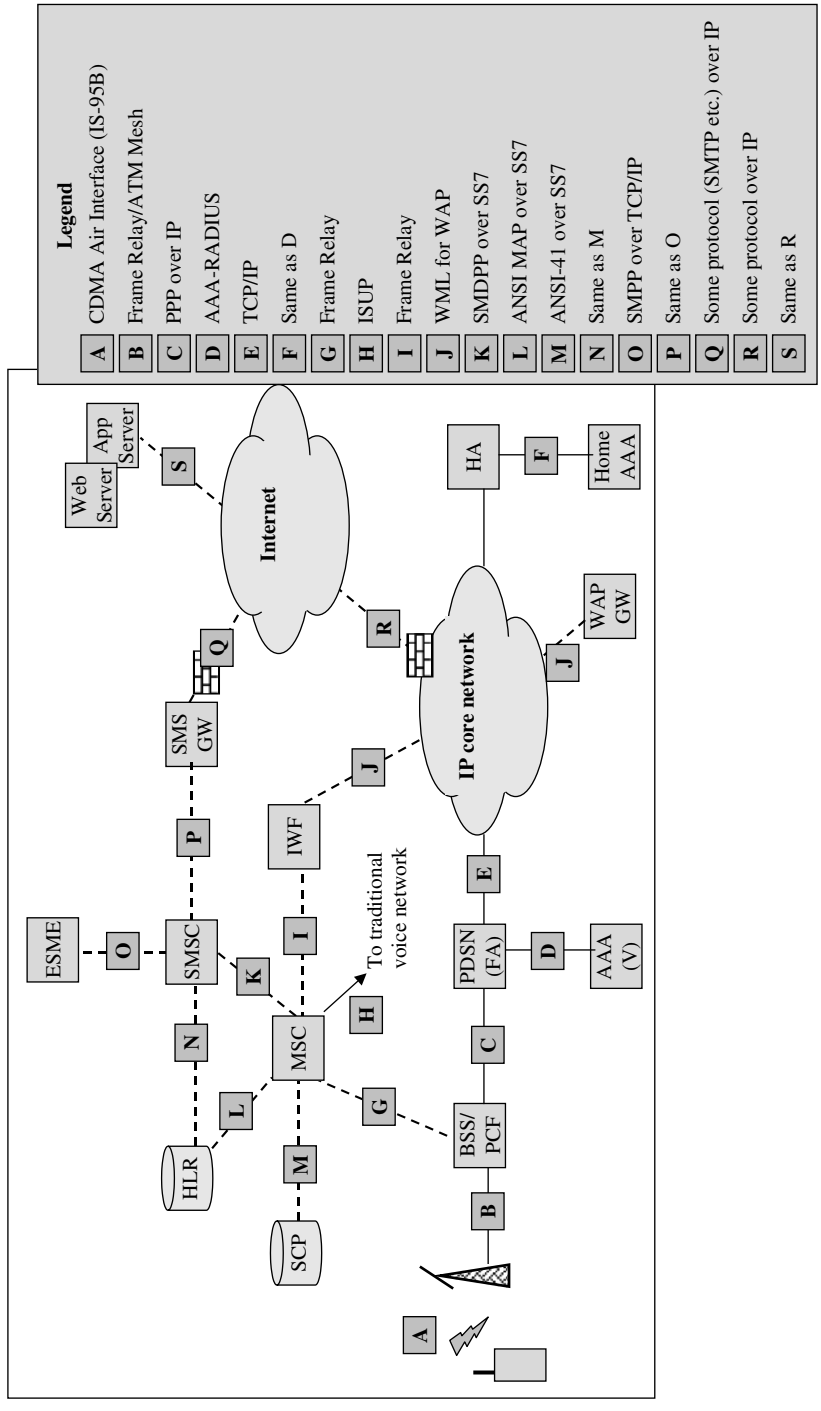

Figure 1.7 An evolved CDMA network with 1X-EVDO

the establishment of PDP contexts for handling end-user sessions, CDMA makes use of PPP (recall this was used in 'dial-up' scenarios). Also, CDMA 1X-EVDO utilizes Mobile IP (designed by the IETF Mobile IP WG) for mobility management.

The astute reader can conclude from the above sections that conceptually GSM and CDMA networks operate on the same principles – signaling and radio protocols are different, but at a high level, they are very similar indeed. So the reader may draw a high level generic model of mobile networks in her mind. Both networks can be viewed as more detailed instances of the generic wireless network architecture shown in Figure 1.6. The reader will see, however, that understanding of some of the differences will help in later chapters as we study Parlay/OSA service capabilities and mappings of service capability APIs to underlying networking technology details. For now though, we continue to focus on the similarities by recognizing that whilst the radio access network technology between 3GPP and 3GPP2 networks differ, the IP core networks of both are harmonized. Both organizations partner in the development and standardization of the IMS.

1.6 The IP Multimedia Subsystem (IMS)

So far, we have looked at various network ecosystems in place today, and have studied the evolution of cellular networks, from the current 2G incarnations to the future 3G evolved forms. At times, a reference was made to an all-IP manifestation of these 3G networks, called the IMS or IP Multimedia Subsystem in 3GPP. Mention was also made that a similar evolved architecture is supported by 3GPP2 as it describes CDMA evolution into an all-IP environment, and that in the latter case, it also goes by the same appellation in addition to sometimes being called the Multi-Media Domain or MMD. We shall use IMS to refer to both.

The IMS architecture is poised to enable the dream of anywhere, anytime communication. What this means is that IMS will enable every networked device, and the people using them, to communicate with any other device, over any network – be it wireline or wireless - with any service, in any media. IMS creates a common core network that can span both wireless and wireline networks, thereby providing seamless service control and delivery across these two types of networks.

1.6.1 A Standards View

In these sections, we study IMS in some detail. 3GPP defines most of the architecture, requirements, and call flows for the IMS in documents such as [3GPP 2002a, 3GPP 2004a, 3GPP 2004b, 3GPP 2005a, 3GPP 2005b] among others[13], and 3GPP2 also utilizes these documents as a basis for its own standards (this enables quicker convergence and reuse) which include the following documents [3GPP2 2003a, 3GPP2 2003b, 3GPP2 2003c, 3GPP2 2003d, 3GPP2 2003e, 3GPP2 2003f, 3GPP2 2003g, 3GPP2 2003h, 3GPP2 2003i, 3GPP2 2003j]. These documents, just like the 3GPP documents previously indicated, contain overviews of the IMS architecture, descriptions of reference points, reference point operational descriptions, and finally, protocol mappings and functional call flows in support of particular required capabilities. Last but not least, the OMA or Open Mobile Alliance™ [OMA] also talks about how the IMS architecture can be supported, albeit more from a services perspective, in a manner that promotes seamless access and use of IMS capabilities in both 3GPP and 3GPP2 contexts. The last of these (i.e. the OMA documents on IMS) are covered by an OMA Enabler Release, called 'IMSinOMA' [IMSinOMA 2005]. [Brenner 2005] provides an introduction into OMA and some of its activities.

In what follows, we shall explore the IMS architecture. This is admittedly a simplified view of IMS – for a more comprehensive treatment, the reader is referred to the standards documents

[13] The interested reader who reads through one or more of these standards documents will soon see how the documents reference one another, and how one quickly gets drawn in, with greater understanding, into more and more other standards documents that explain more of the esoteric details. For help in locating standards documents, the reader is referred to Appendix B, which is included as advanced reading in [Parlay@Wiley].

listed above, which total several hundred pages together. But the lightweight treatment of IMS concepts here shall suffice for most readers to provide a clear view of this all-IP architecture, and its relevance and relationship to Parlay and OSA technologies.

1.6.2 Simplified View of the IMS Architecture

Figure 1.8 depicts a simplified view of the IMS architecture – a view that covers the most important service-related aspects and is sufficient for our purposes. As was alluded to earlier in this chapter, the HLR or Home Location Registry is the centralized repository for service subscription data and service profiles. With the evolution of cellular networks into their all-IP 3G form, the HLR element evolves forward into the HSS or Home Subscriber Server. This HSS element stores information pertaining to subscribers and their subscribed services, among other things, in the 3G environment, and is accessible to call control elements (called CSCFs or Call Session Control Functions), application servers (analogous to SCPs from the traditional IN model), and other authorized network entities that require this information in processing end-user requests.

The IMS supports SIP as the protocol of choice for all signaling, for call control and for most service control interactions between the CSCFs and application servers. Generic SIP (as defined in [RFC 3261]) is used as is, for the most part. The one exception to this is the reference point between the CSCF and the AS, called the ISC (IMS Service Control) reference point. Along this interface, the ISC protocol (SIP with some special private header extensions) is used. Interfaces to the HSS component are normally implemented using the DIAMETER protocol. The main reference points of interest along with the associated protocols are indicated in Figure 1.8.

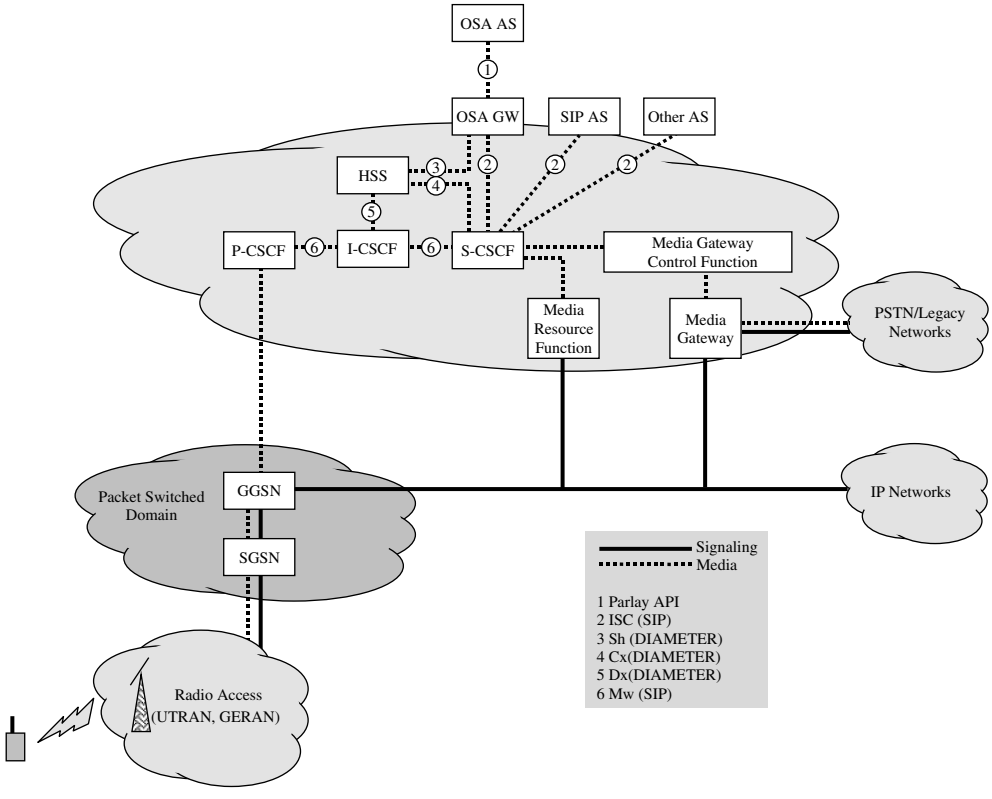

Figure 1.8 IMS network architecture

1.6.2.1 Application Servers in the IMS Architecture

The IMS architecture defines a service layer that supports different kinds of application servers, most prominent among them being the SIP AS and the OSA Gateway.[14]

- The SIP AS supports SIP-based applications and receives SIP (or ISC) messages from the network and responds with messages in the same protocol to enable further processing of user requests at the CSCF or to otherwise be able to provide an enhanced end-user experience[15]. SIP ASs may use any SIP-based technology (e.g. SIP CGI [RFC 3050], SIP CPL [RFC 3880], SIP Servlets [JSR 116], etc.) as they support the value-added application logic.
- The OSA Gateway is the Service Mediation Gateway or SMG that is referred to in later chapters in the book. This is a gateway component that implements the standards defined by the Parlay and OSA specifications. Since 3GPP defines IMS and 3GPP defines OSA, it is logical that references are made to OSA (and not Parlay) in this context. However, as we will see in Chapter 4, the two technologies are similar to the point of being virtually indistinguishable.

The OSA gateway serves as a gateway element (as the name suggests), enabling different OSA-compliant applications that are themselves hosted on application servers (called OSA ASs), access to network capabilities via the OSA-defined SCF APIs. Since most of the book is dedicated to the topic of the OSA Gateway (or Service Mediation Gateway, SMG, as we will call it), we do not discuss that in any more detail here.

1.6.2.2 The Different Types of CSCFs

The IMS architecture classifies CSCFs into three types based on their location and the logical function they perform in call flows. These are as below:

- Proxy-CSCF or P-CSCF: The Proxy CSCF is the contact point into the IMS for an end-user's terminal. The P-CSCF may reside in a visited network, in case the user is roaming. In case the IMS network is realized as an overlay on top of a GPRS network, the P-CSCF is the first point of contact after the GGSN that routes to the user's home IMS network.
- Interrogating-CSCF or I-CSCF: The Interrogating CSCF is the contact point into the user's home IMS network from other networks. Its job is to locate the right S-CSCF for the user after querying the HSS, and then to forward the SIP Registration request to the S-CSCF. Once registration is completed, and the S-CSCF is known, the I-CSCF is no longer involved, and SIP Invite messages are forwarded directly from the P-CSCF to the S-CSCF for outgoing calls and vice versa for incoming calls. There is one notable exception. If for some reason the network operator wishes to keep their network configuration hidden[16], the I-CSCF remains in the path between the P-CSCF and the S-CSCF. In this case, the I-CSCF performs the function of a Topology Hiding Inter-network Gateway (THIG).
- Serving CSCF or S-CSCF: The Serving CSCF serves as the SIP session control point for the end-user's terminal device and, like the I-CSCF, always resides in the user's home IMS network.

[14] The OSA Gateway will be explained in later chapters in its full breadth and depth. For the moment, we shall just focus on its position in the overall IMS architecture, and its relation to the HSS and S-CSCF.

[15] Unlike the SCPs in traditional IN domains that are limited to providing (rather critical) services to call processing, ASs in IMS, which can support capabilities in areas other than just call control (think Presence for example), may be able to provide enhanced end-user experiences even outside immediate call control contexts. Hence the 'or' in this statement.

[16] An example for one such reason could be to hide capacity information like the exact number of S-CSCFs from other networks for competitive motivations.

The S-CSCF maintains state information required for the support of services, however, the S-CSCF does not contain service logic itself. For service logic execution, the S-CSCF refers to application servers (ASs) using the SIP-based IMS Service Control (ISC) interface. All calls and sessions go through the S-CSCF and the S-CSCF controls all services, irrespective whether the end-user is roaming or not, thus ensuring continuous and consistent end-user experience.

1.6.3 Service Control in IMS

Service control in IMS takes place entirely on SIP ASs, as we have seen that the various CSCFs do not contain any service logic themselves. Determining the sequence and invocation of applications running on these SIP ASs for a given call may be done in two places: the S-CSCF (service filtering) and SCIM (service brokering). The procedures for service filtering have been standardized in detail [3GPP 2005a], whereas the mechanisms for service brokering are largely under-standardized. Both mechanisms will be explained below.

1.6.3.1 Filter Criteria

For execution of service logic, ASs are involved by the S-CSCF through the ISC interface. So, like in Intelligent Networking, we see a separation of call or session processing logic and service control logic. Unlike IN however, delegation of service logic execution to ASs is not based on a call model or state machine with detection points. Rather, the S-CSCF may decide to forward a certain SIP message to a particular AS based on a variety of criteria. These criteria include:

- the type of SIP message received (e.g. an INVITE or a REGISTER message);
- whether or not some specific header element is present in the SIP message;
- the content of certain header elements;
- whether the SIP message pertains to an incoming or outgoing call.

These criteria are referred to as filter criteria. Based on the filter criteria, the S-CSCF decides to forward certain SIP messages to a specific AS for service logic execution, whereas other SIP messages are processed by the S-CSCF itself for call or session processing. We can distinguish two types of filter criteria.

1. Initial Filter Criteria (iFC) are part of the subscription and service profile of the end-user and are stored in the HSS. The iFC are downloaded from the HSS into the S-CSCF over the Cx interface, upon registration of the end-user device in the network.
2. Subsequent Filter Criteria (sFC) are determined by the AS, once it has been involved in service control by the S-CSCF as a result of the iFC. SFC are determined dynamically based on service logic execution and signaled back to the S-CSCF over the ISC interface.

The iFC are specific for a given Application Server. Hence if the end-user is subscribed to more than one service, multiple iFCs can be part of the subscriber profile. As part of the iFC a priority is defined which allows the S-CSCF to determine the order in which to contact the various Application Servers. Default behavior is also part of the iFC definition in case the AS in question cannot be contacted.

1.6.3.2 The Service Capability Interaction Manager

The Service Capability Interaction Manager (SCIM) is defined as part of the Application Servers that provide service control in IMS networks. As any aspect within an Application Server is left unspecified in 3GPP, providing it handles any SIP exchange appropriately according to ISC

definitions, the SCIM is left unspecified as well. The role of the SCIM is that of service broker in more complex service interaction scenarios than can be supported through the service filtering mechanism; for example, feature interaction management that provides intricate intelligence, such as the ability to blend Presence and other network information, or more complex and dynamic application sequencing scenarios. This added complexity may provide another reason why SCIM is not fully standardized.

Whereas the service filtering mechanism can be used to manage application interaction and straightforward sequencing, the SCIM may provide a more enhanced end-user experience by blending applications with each other and with context-sensitive information like Presence and Location, and Policy functions. In addition, the SCIM may incorporate multi-session awareness, with a session context that can comprise multiple sub-sessions for example, for voice, video and data streams.

The SCIM is mentioned here, though underspecified in standards and hence mostly proprietary, because the Parlay Gateway, in its capacity as Service Mediation Gateway, can be deployed to fulfill the role and function of the SCIM in IMS networks.

1.7 Related Technologies

Now that we have familiarized ourselves with various networks and their service architectures, a number of related technologies are introduced here as they provide service capabilities in the network that Parlay can provide access to. Later chapters in the book will elaborate on the programmatic interfaces defined by Parlay to make use of these capabilities when building end-user applications.

1.7.1 WAP Technology

WAP or the Wireless Application Protocol, defined initially by the WAP Forum™[17], really came to the fore around 1997, and was the precursor to some of the more exciting 'data to mobile handset' applications of which we will see more as network evolution to 3G continues. A very high-level, simplified view of WAP operation is presented here. As was previously stated in the section that discussed Internet access technology, for WAP as well, multiple alternative access paths to WAP-based services are afforded by the networks of today. In particular, GPRS and EDGE networks (and their CDMA equivalents) provide for the notion of data-session supporting nodes such as SGSN/GGSNs and PCF/PDSNs for WAP sessions. To keep explanations simple, and since we are using the PSTN network and its evolution to drive our discussions, we shall focus on the circuit-switched based data access path for WAP in this section. The interested reader is referred to [WAP] for more details.

WAP enables the user to browse the Internet from her mobile terminal. WAP, in concept, is independent of the radio access or core network technology in use, and can be deployed just as effectively in CDMA and GSM networks. There are several million users of WAP today. Since the screen-size of wireless handsets is usually small, and other limitations exist (such as the thin pipe to the handset over the air interface, etc.), an element called the WAP gateway is introduced into service provider networks where WAP is deployed, to perform conversions of accessed web page contents for suitable rendering on handsets, and for transport over the air.

The WAP standards define a complete protocol stack for use between the handset and the WAP gateway including layers for session control (WSP), security (WTLS), etc., as well as content encoding related aspects. The latter includes a WAP binary format for over the air transmission of accessed

[17] This body has since been subsumed under the OMA [OMA]. As Andrew Tanenbaum once remarked, 'The one good thing about standards is that there are so many to choose from.' Lately, market forces have caused a kind of consolidation of some of these distinct bodies, thereby contributing greater stability, and enabling vendors to make more judicious choices of which protocols to implement in their products.

data, and an encoding format called WML or the Wireless Markup Language – derived from HTML, which most web pages are written in today – for easy rendering to handset screens and so on.

Figure 1.9 provides a view of the network infrastructure needed to support WAP (recall that our focus here is primarily on network technologies). The digital switch or MSC is provided with a connection to an Interworking Function (IWF). All WAP data calls (or dialed calls where the destination is a WAP service) transit this link. The IWF connects on its other interface to a wireless service provider hosted IP network (LAN or WAN), to which a WAP gateway is connected. Some kind of simple handshake takes place between the IWF (representing the user device) and the WAP gateway as this connection is set up, and user credentials such as the subscriber phone number and other information are exchanged across this interface at that time.

The WAP gateway maintains the association between the user identity, and the IP address assigned to this connection, and then works to forward on user requests for web content to web or WAP servers (also called Origin Servers) either within, or outside, the service provider network. The WAP gateway then performs the required conversions on the data returned, and forwards it on along the same path, but in the reverse direction, back to the handset.

The reader should note that here data are being carried over the circuit call established between the handset and the digital switch. When the session ends, the user simply disconnects the call.

1.7.2 Location Based Services

Of late, there has been an upsurge in location technology and its use particularly in mobile networks, but sadly, the uptake here in terms of real-world networks has been somewhat sluggish. Location has been used in wired networks for many years now. The E-911 system in the US, has for example, relied on reverse directory lookups in databases to advise emergency operators and dispatchers of the location and routing information from the nearest police/fire station or hospital, so as to better assist people in distress in more timely a fashion. But use of location technology in networks with wireless handsets is somehow more appealing, primarily due to the mobility of the terminals in question.

In their simplest form, location-based services may be classified into two types. One is where the user himself provides his location while requesting location-specific information from a server. An example of this is where Bob enters his zip code into an HTML form to obtain local weather information, and possibly a Doppler radar image of his vicinity. A second, more enhanced service experience could result if Bob simply asked for location specific information, and his location were transparently obtained by the server in question (factoring in his preferences and privacy

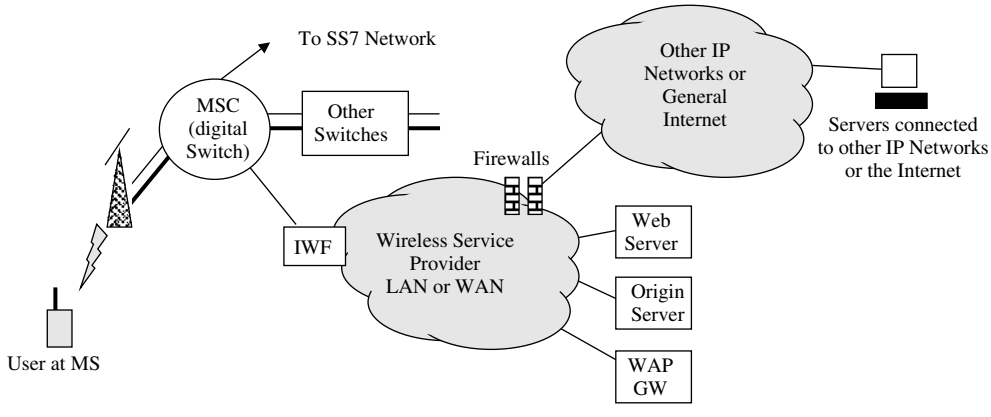

Figure 1.9 The WAP access model

permissions of course), and he were provided with context sensitive information without having to provide his location explicitly to the service.

The latter could be achieved in several different ways in cellular networks today. Recently, phones are coming equipped with GPS receivers, thereby enabling them to provide a fairly accurate location fix that can then be passed on (again, with end-user permission) to network services that require it. Alternatively, network elements such as MPCs (Mobile Positioning Centers) in CDMA networks, and GMLCs (Gateway Mobile Location Centers) in GSM networks, which talk to other Position Determining Equipment (PDE) in the network, are able to obtain location information (Figure 1.10). Such location information could consist of the cell-ID and cell-sector the handset is currently in, or even the latitude and longitude (sometimes even altitude) co-ordinates (sometimes called lat/long or X/Y/Z) determined using various algorithms and triangulation mechanisms such as AFLT, EFLT or Network Assisted GPS (this uses network information in concert with GPS information to locate more accurately a handset).

These MPC and GMLC servers can be made accessible to applications either directly, over protocols such as MLP (Mobile Location Protocol, an XML-based protocol defined by the Location Interoperability Forum or LIF, now subsumed by the OMA standards body), or indirectly, through OSA/Parlay capable service mediation gateway elements via the User Location interfaces supported by such gateways. Regardless of what mechanisms are used, once this information is obtained by the location-based services, specific context sensitive content can be served to users more transparently. Furthermore, this technology may be used very effectively to respond to emergency calls made from cell-phones where the caller either does not know, or is otherwise unable to specify, his or her location.

1.7.3 Short Message Service and Multi-media Messaging

Messaging capabilities are an intrinsic part of the wireless networks of today. Its most well known exponent is the Short Message Service (SMS). Given the popularity of SMS and the resulting high volumes and thus revenues this service generates, it is interesting to consider that the success of the service was really a fluke. In early GSM deployments, part of the available network capacity remained unused. Taking advantage of the characteristics of digital technology available in second generation networks, SMS was introduced as a low-bandwidth, packet-based message exchange mechanism, mostly bundled by equipment vendors at a discount with GSM voice service as part of a package deal. Adding SMS messaging services to the more sparsely used frequency bands in the network allowed network operators to make more use of their available bandwidth, and potentially increase average revenue per user. So, born as a capacity optimization feature, a killer application has emerged blinking into the daylight.

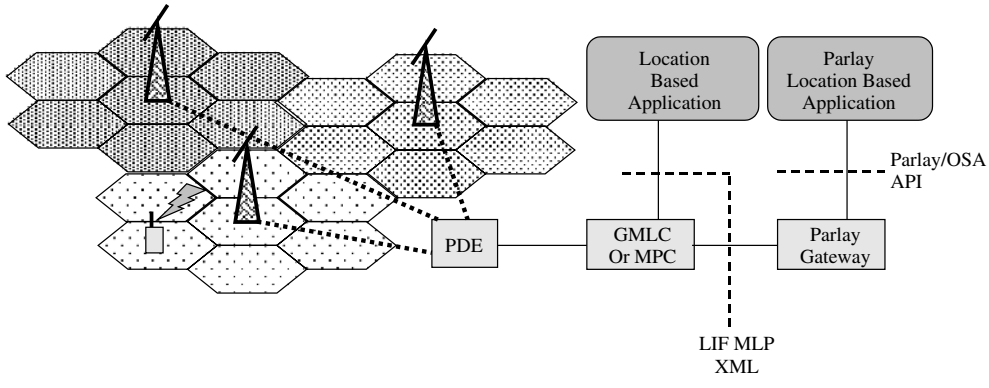

Figure 1.10 Logical architecture schematic of location-based services

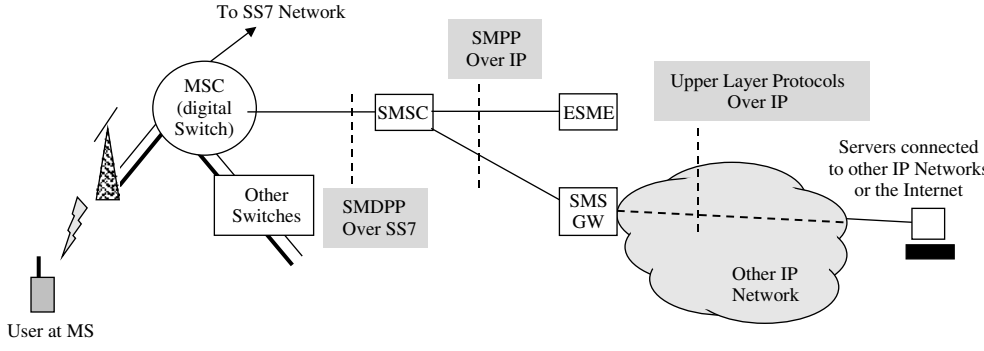

Figure 1.11 SMS network view

The Short Message Service is a store-and-forward message delivery technology, where a Short Message Service Center is introduced in the GSM network as the message store (Figure 1.11). SMS uses the wireless network for message transport and delivery. Because of the store-and-forward nature, SMS basically consists of two point-to-point services, from the originator to the SMSC (Mobile-originated short message, or MO-SM), and from the SMSC to the destination (Mobile-terminated short message, or MT-SM).

The SMSC uses the HLR to locate the destination party for an SMS message. The HLR is also used for supplementary services applicable to the SMS bearer, such as for example the barring of incoming SMS messages. Short messages are short, as they are transmitted out-of-band, over a low bandwidth medium. The messages are limited to 160 alphanumeric characters, although messages may be concatenated. The signaling protocol between the SMSC and the HLR (e.g. to obtain the location of the destination for the short message) is the SS7-based MAP protocol.

Given its enormous popularity, the basic SMS service has been enhanced in many ways. Simple examples include the ability to concatenate the short messages, and the addition of point-to-broadcast to the basic point-to-point capabilities. Enhanced Messaging Service (EMS) adds the capability to send formatted text messages (including bold and italic fonts), simple pictures and animations, and ring tones and logos.

The latest step in this process of enhancing the SMS capabilities and building on the success of the service is the Multimedia Messaging Service (MMS). MMS messages may be used to stream audio or video to the mobile device, or to exchange photos and download games.

In order to support MMS in the network, the basic SMSC does no longer suffice. An MMS Relay/Server is introduced to support MMS capabilities. The basic functionality of the MMS Relay/Server is still the storing and forwarding of messages, MMS messages in this case, but given the much richer content involved, interfaces are introduced to value added service applications, content stores, and external networks.

1.8 Summary

In this chapter, we have covered, with a broad brush, many of the networking technologies in use today. The intent here is to provide the reader with a background and a little more appreciation of the complexity involved in network architectures, and also to introduce, albeit at a high level, the kinds of interfaces in existence, and the reference points where programmatic interfaces could be introduced (this latter point will become more apparent in later chapters). This chapter serves as the basis for the discussions in the rest of the book – we scatter some magic idea seeds here, and these grow into a forest of beanstalks in the pages to come. Next, we look at some marketing, business, and technology drivers for change.