

# 1

## Basic Concepts

This chapter outlines the basic concepts of mobile communications systems and presents the required background information necessary for a clear understanding of this book. First, an overview of the evolution of mobile communications systems is provided. This encompasses the introduction of first generation analog systems supporting only voice communications to the recent deployment of third generation systems supporting voice and multimedia services. The Global System for Mobile Communication, commonly known as GSM, has been a major breakthrough in the domain of mobile communications. Elements composing a typical GSM network are presented. Another important milestone is the introduction of the General Packet Radio Service (GPRS) allowing the support of packet-based communications in evolved GSM networks. The architecture of a GPRS network is presented. Recently deployed are Universal Mobile Telecommunications Systems (UMTS). These systems support advanced multimedia services requiring high data rates. UMTS services and supporting technologies are also introduced in this chapter. Additionally, the Wireless Application Protocol (WAP) is described. WAP is an enabling technology for developing services such as browsing and multimedia messaging. An overview of latest digital rights management methods is also provided. The last section of this chapter provides pointers to books and reference articles for anybody wishing to further explore the topics covered in this chapter.

### 1.1 Generations of Mobile Communications Networks

In France, in 1956, a very basic mobile telephony network was implemented with vacuum electronic tubes and electron-mechanical logic circuitry. These devices used for wireless communications had to be carried in car boots. In these early days of mobile telephony, service access was far from being ubiquitous and was reserved for a very limited portion of the population. Since the introduction of this experimental network, mobile communications technologies benefited from major breakthroughs commonly categorized in three generations. In the 1980s, *first generation (1G) mobile systems* arrived in Nordic countries. These first generation systems were characterized by analog wireless communications and limited support for user mobility.

Digital communications technology was introduced with *second generation (2G) mobile systems* in the 1990s. Second generation systems are characterized by the provision of better quality voice services available to the mass market. Second generation systems benefited from the cellular concept in which scarce radio resources are used simultaneously by several mobile users without interference. The best known 2G system is the Global System for Mobile Communication (GSM) with the billionth GSM user connected in the first quarter of 2004. Other major 2G systems include cdmaOne (based on CDMA technology), with users in the Americas and Asia, and Japanese Personal Data Cellular (PDC) with the iMode technology for mobile Internet.

Early 2004, first *third generation (3G) mobile systems* have been deployed in several European countries. With 3G systems, various wireless technologies converge with Internet technologies. Third generation services encompass a wide range of multimedia and cost-effective services with support for worldwide user mobility. The migration to 3G systems is facilitated by the introduction of intermediary evolved 2G systems, also known as *2.5G systems*.

## 1.2 Telecommunications Context: Standardization and Regulation

In the telecommunications environment, *Standard Development Organizations (SDOs)* provide the necessary framework for the development of standards. These standards are technical documents<sup>1</sup> defining or identifying the technologies enabling the realization of telecommunication network technologies and services. The prime objective of SDOs is to develop and maintain widely accepted standards allowing the introduction of attractive services over interoperable networks. The actors that are involved in the standardization process are network operators, manufacturers, and third party organizations such as content providers, equipment testers, and regulatory authorities. One of the main objectives of telecommunications regulation authorities is to ensure that the telecommunications environment is organized in a sufficiently competitive environment and that the quality of service offered to subscribers is satisfactory.

In the early days of mobile communications, various regional SDOs developed specifications for network technologies and services independently. This led to the development of heterogeneous networks where interoperability was seldom ensured. The lack of interoperability of first generation mobile systems prevented the expansion of a global international mobile network that would have certainly greatly improved user experience. With second and third generations systems, major SDOs decided to gather their efforts in order to ensure that mobile communication networks will appropriately interoperate in various regions of the world. In 1998, such an effort was initiated by several SDOs including ARIB (Japan), ETSI (Europe), TTA (Korea), TTC (Japan), and T1 (USA). The initiative was named the Third Generation Partnership Project (3GPP). The 3GPP standardization process is presented in Chapter 2.

<sup>1</sup> Technical documents are also known as technical specifications, reports, or recommendations.

## 1.3 Global System for Mobile Communication

Before the introduction of the *Global System for Mobile Communication* (GSM), mobile networks implemented in different countries were usually incompatible. This incompatibility made impracticable the roaming of mobile users across international borders. In order to get around this system incompatibility, the Conférence Européenne des Postes et Télécommunications (CEPT) created the Groupe Spécial Mobile<sup>1</sup> committee in 1982. The main task of the committee was to standardize a pan-European cellular public communication network in the 900 MHz radio band. In 1989, the European Telecommunications Standard Institute (ETSI) took over the responsibility for the maintenance and evolution of GSM specifications. In 2000, this responsibility was transferred to 3GPP. The initiative was so successful that networks compliant with the GSM standard have now been developed worldwide. Variations of the GSM specification have been standardized for the 1800 and 1900 MHz bands and are known as DCS 1800 and PCS 1900, respectively. In March 2004, the GSM association<sup>2</sup> reported a total number of 1046.8 million subscribers distributed over 207 countries.

A GSM network is characterized by digital voice communications and support of low-rate data services. The GSM air interface is based on Time Division Multiple Access (TDMA). With TDMA, a radio band is shared by multiple subscribers by allocating one or more timeslots on given radio carriers to each subscriber. With GSM, the transfer of data can be carried out over circuit-switched connections. For these data communications, bit rates up to 14.4 Kbps can be achieved on single-slot connections. The single-slot configuration is called Circuit Switched Data (CSD). Higher bit rates up to 57.6 Kbps can be attained by allocating more than one slot for a data connection. This multi-slot configuration is called High Speed CSD (HSCSD).

One of the most popular GSM services is the Short Message Service (SMS). This service allows SMS subscribers to exchange short text messages. An in-depth description of this service is provided in Chapter 3. An application-level extension of SMS in the form of the Enhanced Messaging Service (EMS) is presented in Chapter 4.

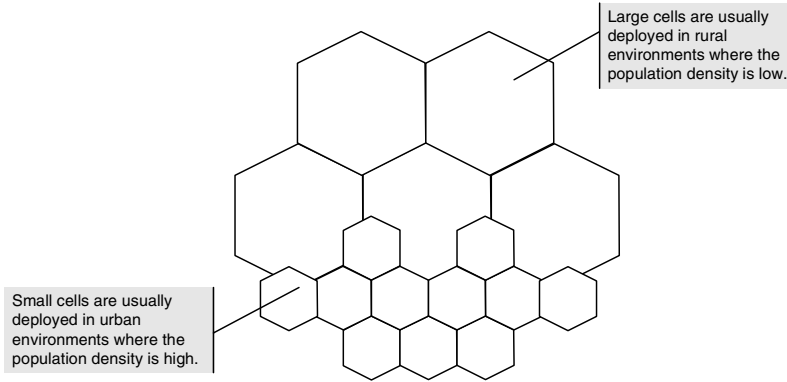
### 1.3.1 Cellular Concept

Radio bands available for wireless communications in mobile networks represent very scarce resources. In order to efficiently use these resources, GSM networks are based on the *cellular concept*. With this concept, the same radio resources (characterized by a frequency band and a timeslot) can be utilized simultaneously by several subscribers without interference if they are separated by a minimum distance. The minimum distance between two subscribers depends on the way radio waves propagate in the environment where the two subscribers are located (e.g., presence of buildings, etc.). In a GSM network, the smaller the cells, the higher is the frequency reuse factor, as shown in Figure 1.1.

In a GSM network, a fixed base station transceiver manages the radio communications for all mobile stations located in a cell. Each geometrical cell in Figure 1.1 represents the radio coverage of one single base station.

<sup>1</sup> The name Groupe Spécial Mobile was later translated to Special Mobile Group (SMG).

<sup>2</sup> <http://www.gsmworld.com>



**Figure 1.1** Cellular concept

### 1.3.2 GSM Architecture

The main elements of the GSM architecture [3GPP-23.002] are shown in Figure 1.2. The GSM network is composed of three subsystems: the *Base Station Subsystem* (BSS), the *Network Subsystem* (NSS), and the *Operation Subsystem* (OSS). The OSS implements functions that allow the administration of the mobile network. For the sake of clarity, elements of the OSS are not represented in the GSM architecture shown in Figure 1.2. Elements of the BSS and NSS are further described in the following sections.

### 1.3.3 Mobile Station

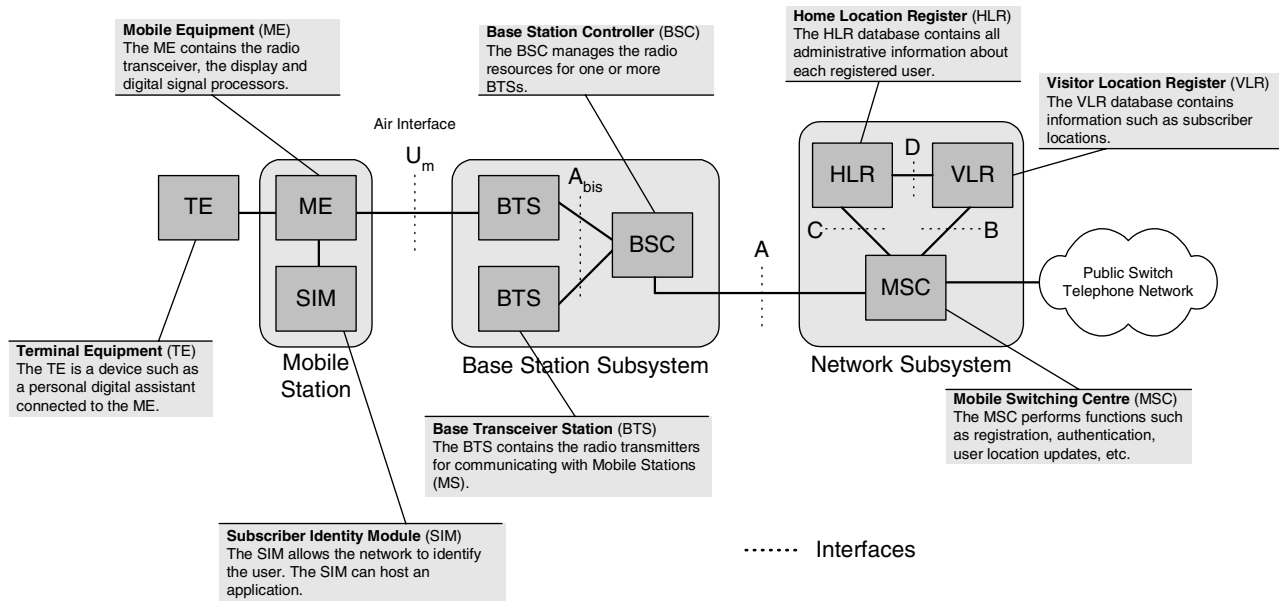
The *Mobile Station* (MS) is a device that transmits and receives radio signals within a cell site. A mobile station can be a basic mobile handset, as shown in Figure 1.3, or a more complex Personal Digital Assistant (PDA). Mobile handset capabilities include voice communications, messaging features, and phone book management. In addition to these basic capabilities, a PDA is usually shipped with an Internet microbrowser and an advanced Personal Information Manager (PIM) for managing contacts and calendaring/scheduling entries. When the user is moving (i.e., while driving), network control of MS connections is switched over from cell site to cell site to support MS mobility. This process is called *handover*.

The mobile station is composed of the *Mobile Equipment* (ME) and the *Subscriber Identity Module* (SIM). The unique *International Mobile Equipment Identity* (IMEI) stored in the ME identifies uniquely the device when attached to the mobile network.

The SIM is usually provided by the network operator to the subscriber in the form of a smart card. The microchip is often taken out of the smart card and directly inserted into a dedicated slot in the mobile equipment. A SIM microchip is shown in Figure 1.4.

Today's mobile stations can be connected to an external device such as a PDA or a personal computer. Such an external device is named a *Terminal Equipment* (TE) in the GSM architecture.

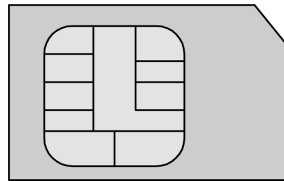
A short message is typically stored in the mobile station. Most handsets have SIM storage capacities. High-end products sometimes complement the SIM storage capacity with



**Figure 1.2** GSM architecture



**Figure 1.3** Mobile station handset – reproduced by permission of Alcatel Business Systems



**Figure 1.4** SIM microchip

additional storage in the mobile equipment itself (e.g., flash memory). It is now common to find handsets shipped with a PIM. The PIM is usually implemented as an ME internal feature and enables elements such as calendar entries, memos, phonebook entries, and of course messages to be stored in the ME. These elements are managed, by the subscriber, with a suitable graphical user interface. These PIM elements remain in the PIM even when the SIM is removed from the mobile handset. Alternatively, simple elements such as short messages and phonebook entries can be directly stored in the SIM. A SIM can contain from 10 short messages to 50 short messages on high-end solutions. Storing elements in the SIM allows messages to be retrieved from any handset simply by inserting the SIM in the desired handset. The benefit of storing messages in the ME is that the ME storage capacity is often significantly larger than the SIM storage capacity.

#### *1.3.4 Base Transceiver Station*

The *Base Transceiver Station* (BTS) implements the air communications interface with all active MSs located under its coverage area (cell site). This includes signal modulation/demodulation, signal equalizing, and error coding. Several BTSs are connected to a single

*Base Station Controller* (BSC). In the United Kingdom, the number of GSM BTSs is estimated around several thousands. Cell radii range from 10 to 200 m for the smallest cells to several kilometers for the largest cells. A BTS is typically capable of handling 20–40 simultaneous communications.

### 1.3.5 Base Station Controller

The BSC supplies a set of functions for managing connections of BTSs under its control. Functions enable operations such as handover, cell site configuration, management of radio resources, and tuning of BTS radio frequency power levels. In addition, the BSC realizes a first concentration of circuits towards the MSC. In a typical GSM network, the BSC controls over 70 BTSs.

### 1.3.6 Mobile Switching Center and Visitor Location Register

The *Mobile Switching Center* (MSC) performs the communications switching functions of the system and is responsible for call set-up, release, and routing. It also provides functions for service billing and for interfacing other networks.

The *Visitor Location Register* (VLR) contains dynamic information about users who are attached to the mobile network including the user's geographical location. The VLR is usually integrated to the MSC.

Through the MSC, the mobile network communicates with other networks such as the Public Switched Telephone Network (PSTN), Integrated Services Digital Network (ISDN), Circuit Switched Public Data Network (CSPDN), and Packet Switched Public Data Network (PSPDN).

### 1.3.7 Home Location Register

The *Home Location Register* (HLR) is a network element containing subscription details for each subscriber. An HLR is typically capable of managing information for hundreds of thousands of subscribers.

In a GSM network, signaling is based on the Signaling System Number 7 (SS7) protocol. The use of SS7 is complemented by the use of the Mobile Application Part (MAP) protocol for mobile specific signaling. In particular, MAP is used for the exchange of location and subscriber information between the HLR and other network elements such as the MSC. For each subscriber, the HLR maintains the mapping between the *International Mobile Subscriber Identity* (IMSI) and the *Mobile Station ISDN Number* (MSISDN).

For security reasons, the IMSI is seldom transmitted over the air interface and is only known within a given GSM network. The IMSI is constructed according to [ITU-E.212] format. Unlike the IMSI, the MSISDN identifies a subscriber outside the GSM network. The MSISDN is constructed according to [ITU-E.164] format (e.g., +33612345678 for a French mobile subscriber).

## 1.4 General Packet Radio Service

In its simplest form, GSM manages voice and data communications over circuit-switched connections. The General Packet Radio Service (GPRS) is an extension of GSM which

allows subscribers to send and receive data over packet-switched connections. The use of GPRS is particularly appropriate for applications with the following characteristics:

- bursty transmission (for which the time between successive transmissions greatly exceeds the average transfer delay);
- frequent transmission of small volumes of data;
- infrequent transmission of large volumes of data.

These applications do not usually need to communicate permanently. Consequently, the continuous reservation of resources for realizing a circuit-switched connection does not represent an efficient way to exploit scarce radio resources. The basic concept behind the GPRS packet-based transmission lies in its ability to allow selected applications to share radio resources by allocating radio resources for transmission only when applications have data to transmit. Once the data have been transmitted by an application, radio resources are released for use by other applications. Scarce radio resources are used more efficiently with this mechanism. GPRS allows more radio resources to be allocated to a packet-based connection than to a circuit-switched connection in GSM. Consequently, a packet-based connection usually achieves higher bit rates (up to 171.2 Kbps) by using a multislot configuration for uplinks and downlinks as shown in Table 1.1. For instance, a mobile station of multislot GPRS class 6 can have a maximum of three slots allocated to the downlink and a maximum of two slots allocated to the uplink. However, for such a mobile station, a maximum of four slots only can be active at a time for both uplink and downlink. The capacity of each slot depends on the channel encoding used. Four channel encoding schemes are available in GPRS with distinct levels of error protection and are typically selected according to the quality of the radio environment. GPRS can offer “always on” connections (sending or receiving data at any time).

**Table 1.1** Multislot GPRS classes

Multislot GPRS class	Downlink slots	Uplink slots	Active slots
1	1	1	2
2	2	1	3
3	2	2	3
4	3	1	4
5	2	2	4
6	3	2	4
7	3	3	4
8	4	1	5
9	3	2	5
10	4	2	5
11	4	3	5
12	4	4	5

### 1.4.1 GPRS Architecture

The main elements composing the GPRS architecture [3GPP-23.060] are shown in Figure 1.5. A GPRS mobile station is categorized according to its capabilities to support simultaneous modes of operation for GSM and GPRS [3GPP-22.060] which are as follows:

- *Class A*: the mobile station supports simultaneous use of GSM and GPRS services (attachment, activation, monitoring, transmission, etc.). A *class A* mobile station may establish or receive calls on the two services simultaneously. The high complexity of designing *class A* devices makes them prohibitively expensive to produce and, therefore, these devices are typically not available for the mass market.
- *Class B*: the mobile station is attached to both GSM and GPRS services. However, the mobile station can only operate in one of the two services at a time.
- *Class C*: the mobile station is attached to either the GSM service or the GPRS service but is not attached to both services at the same time. Prior to establishing or receiving a call on one of the two services, the mobile station has to be explicitly attached to the desired service.

Before a mobile station can access GPRS services, it must execute a *GPRS attachment* procedure to indicate its presence to the network. After its GPRS attachment, the mobile station activates a Packet Data Protocol (PDP) context with the network in order to be able to transmit or receive data. This procedure is called *PDP context activation*.

The GPRS air interface is identical to that of the GSM network (same radio modulation, frequency bands, and frame structure). GPRS is based on an evolved GSM base station subsystem. However, the GPRS core network relies on a GSM network subsystem in which two additional network elements have been integrated: serving and gateway GPRS support nodes. In addition, *Enhanced Data Rate for Global Evolution* (EDGE) can be supported to improve GPRS performances by introducing an enhanced modulation scheme.

### 1.4.2 Serving GPRS Support Node

The *Serving GPRS Support Node* (SGSN) is connected to one or more base station subsystems. It operates as a router for data packets for all mobile stations present in a given geographical area. It also keeps track of the location of mobile stations and performs security functions and access control.

### 1.4.3 Gateway GPRS Support Node

The *Gateway GPRS Support Node* (GGSN) provides the point of attachment between the GPRS domain and other data networks such as the Internet or corporate networks. An *Access Point Name* (APN) is used by the mobile user to establish the connection to the required destination network.

## 1.5 Universal Mobile Telecommunications System

Since 1990, focus has been given to the standardization of third generation mobile systems. The International Telecommunication Union (ITU) has initiated the work on a set of

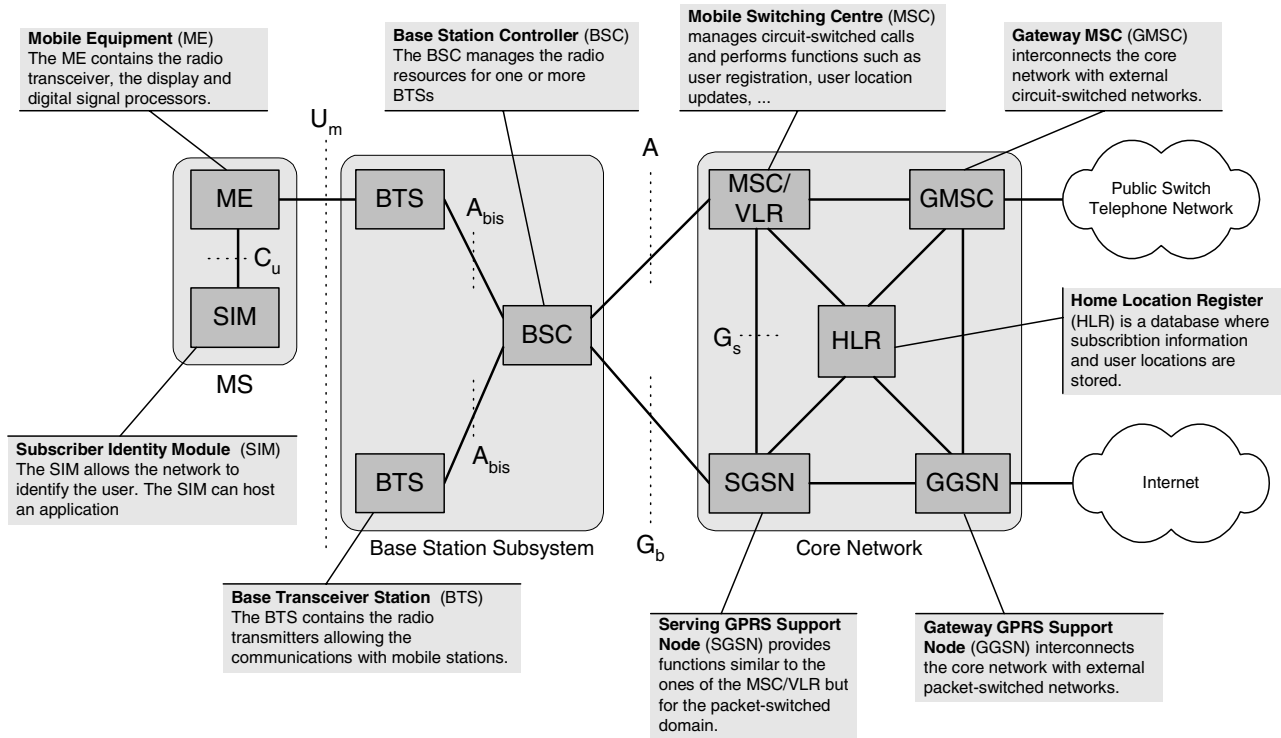


Figure 1.5 GPRS architecture

standards named the International Mobile Telecommunications 2000 (IMT-2000) for the definition of technologies and services for 3G systems. In this family of IMT-2000<sup>1</sup> standards, the Universal Mobile Telecommunications System (UMTS) encompasses the definition of new radio access techniques along with a new service architecture. UMTS aims at providing services such as web browsing, messaging, mobile commerce, videoconferencing, and other services to be developed according to emerging subscribers' needs with the following objectives:

- high transmission rates encompassing circuit-switched and packet-switched connections;
- high spectral efficiency and overall cost improvement;
- definition of common radio interfaces for multiple environments;
- portability of services in various environments (indoor, outdoor, suburban, urban, rural, pedestrian, vehicular, satellite, etc.). This service portability is also known as the *Virtual Home Environment* concept [3GPP-22.121][3GPP-23.127].

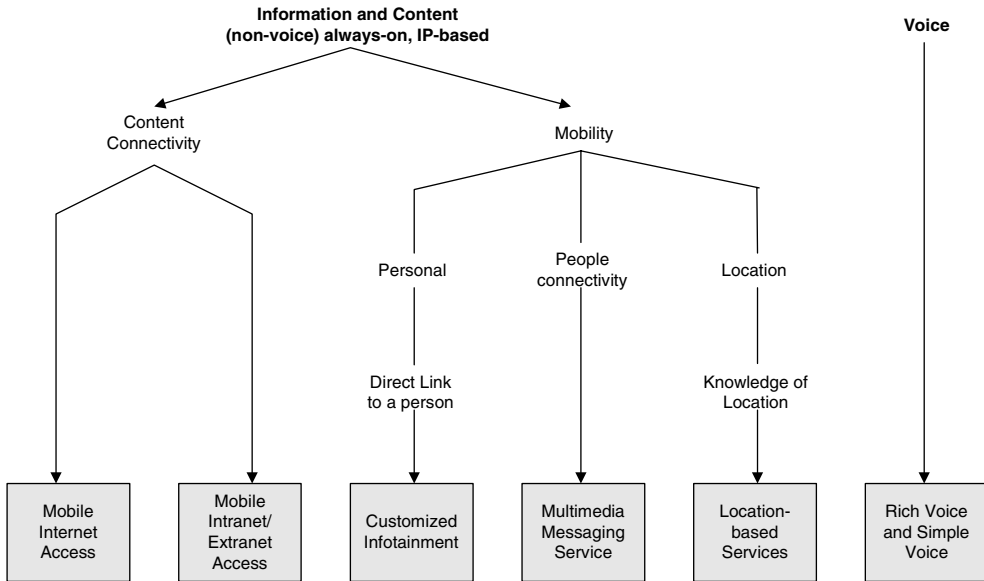
The provisioning of services in heterogeneous environments is enabled with an Open Service Architecture (OSA) [3GPP-22.127][3GPP-23.127]. UMTS extends 2G voice and data capabilities to multimedia capabilities with access to higher bandwidth targeting 384 Kbps for full area coverage and 2 Mb/s for local area coverage. UMTS is expected to become the basis for new mobile telecommunications networks with highly personalized and user-friendly services. UMTS should provide a convergence of communications technologies such as satellite, cellular radio, cordless, and wireless LANs. The network operator NTT Docomo introduced 3G services to the Japanese market in October 2001. Elsewhere, the commercial introduction of UMTS networks and services for the mass market started in 2004.

### 1.5.1 3G Services

Second generation networks provide voice and limited data services. In addition to these 2G services, 3G systems offer multimedia services adapted to the capabilities of multimedia devices and network conditions with a possibility to provide some content specifically formatted according to the subscriber location. The UMTS Forum in [UFR9][UFR13] classifies 3G services into the following six groups as illustrated in Figure 1.6:

- *Mobile Internet access*: a mobile access to the Internet with service quality close to the one offered by fixed Internet Service Providers. This includes full Web access, file transfer, electronic mail, and streaming video and audio.
- *Mobile Intranet/Extranet access*: a secure framework for accessing corporate Local Area Networks (LANs) and Virtual Private Networks (VPNs).
- *Customized infotainment*: a device-independent access to personalized content from mobile portals.
- *Multimedia messaging service*: a means of exchanging messages containing multimedia contents including text, images, and video and audio elements. The multimedia messaging service can be considered as an evolution of SMS where truly multimedia messages can

<sup>1</sup> IMT-2000 was formerly known as Future Public Land Mobile Telecommunications System (FPLMTS).



**Figure 1.6** 3G service categories – source UMTS Forum [UF-Rep-9]

be exchanged between subscribers. An in-depth description of the multimedia messaging service is provided in Chapters 5 and 6.

- *Location-based services*: location-aware services such as vehicle tracking, local advertisements, etc.
- *Rich voice and simple voice*: real-time, two-way voice communications. This includes Voice over IP (VoIP), voice-activated network access, and Internet-initiated voice calls. Mobile videophone and multimedia real-time communications should also be available on high-end multimedia devices.

In the scope of the 3GPP standardization process, the UMTS specification work was divided into two distinct phases. The first phase UMTS, named UMTS Release 99 (also known as Release 3), is a direct evolution from 2G and 2.5G networks (GSM and GPRS networks). The second phase UMTS, also known as UMTS Release 4/5, is a complete revolution introducing new concepts and features.

### 1.5.2 First Phase UMTS

The UMTS architecture [3GPP-23.101] has to meet the requirements of various UMTS services. These requirements range from real-time voice traffic and bursty data access to mixed multimedia traffic. UMTS is intended to offer a true global service availability. To meet this objective, the UMTS architecture includes terrestrial segments complemented by satellite constellations where necessary.

### 1.5.3 First Phase UMTS Architecture

The first phase UMTS architecture is based on evolved GSM and GPRS core networks and a specifically tailored *Universal Terrestrial Radio Access Network* (UTRAN). Two duplexing methods, defining how the received signal is separated from the transmitted signal, have been defined as follows:

- *Universal Terrestrial Radio Access/Time Division Duplex* (UTRA/TDD): this method achieves bi-directional transmission by allowing the use of different time slots over the same radio carrier for the transmission of sent and received signals.
- *Universal Terrestrial Radio Access/Frequency Division Duplex* (UTRA/FDD): this method achieves bi-directional transmission by allowing sent and received signals to be transmitted over two separate and symmetrical radio bands for the two links.

The name *Wideband CDMA* (WCDMA) is also used to identify the two UTRA operating modes (TDD and FDD). Elements composing the first phase UMTS architecture are shown in Figure 1.7.

Elements of the UMTS architecture are grouped into three subsystems: the *User Equipment* (UE), the access network (UTRAN), and the switching and routing infrastructure, also known as the *Core Network* (CN). Elements of the UMTS architecture support both circuit-switched connections and packet-switched connections.

### 1.5.4 User Equipment

The UE, usually provided to the subscriber in the form of a handset, is itself composed of a *Mobile Equipment* (ME) and a *UMTS Subscriber Identity Module* (USIM). The ME contains the radio transceiver, the display, and digital signal processors. The USIM is a 3G application on a *UMTS IC card* (UICC) which holds the subscriber identity, authentication algorithms, and other subscriber-related information. The ME and USIM are interconnected via the Cu electrical interface whereas the UE is connected to the UTRAN via the Uu radio interface. A UE always supports at least one of the operating modes of UTRA: TDD or FDD. In order to allow a smooth transition to UMTS, it is expected that UEs will initially be capable of communicating with legacy systems such as GSM and GPRS. UMTS UEs supporting legacy systems are called multi-mode UEs. The 3GPP classifies multi-mode UEs into the following four categories [3GPP-21.910]:

- *Type 1*: *type 1* user equipment operates in one single mode at a time (GSM or UTRA). It cannot operate in more than one mode at a time. While operating in a given mode, the user equipment does not scan for or monitor any other mode and switching from one mode to another is done manually by the subscriber.
- *Type 2*: while operating in one mode, *type 2* user equipment can scan for and monitor another mode of operation. The user equipment reports to the subscriber on the status of another mode by using the current mode of operation. *Type 2* user equipment does not support simultaneous reception or transmission through different modes. The switching from one mode to another is performed automatically.

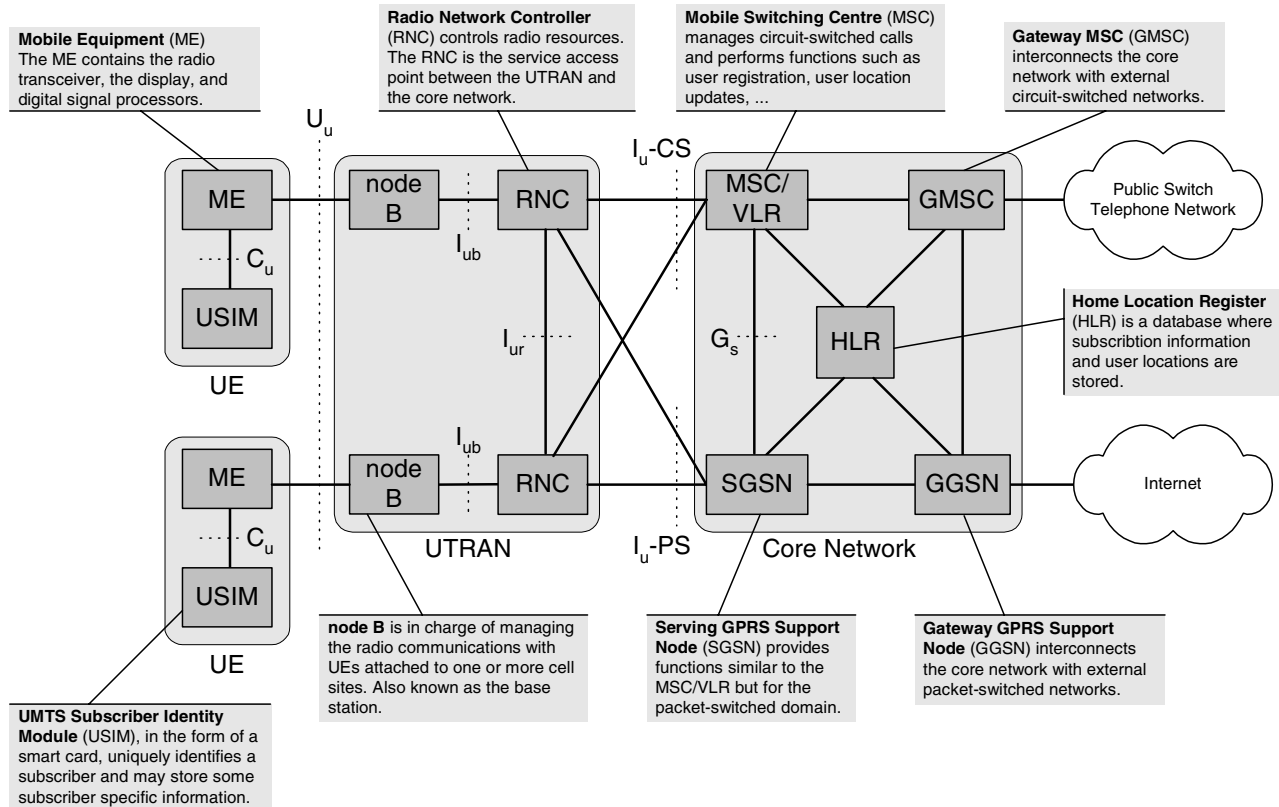


Figure 1.7 UMTS architecture (1st phase)

- *Type 3*: *type 3* user equipment differs from *type 2* user equipment by the fact that the *type 3* UE can receive more than one mode at a time. However, a *type 3* UE cannot emit simultaneously in more than one mode. Switching from one mode to another is performed automatically.
- *Type 4*: *type 4* user equipment can receive and transmit simultaneously in more than one mode. Switching from one mode to another is performed automatically.

### 1.5.5 UTRA Network

The UTRAN is composed of *nodes B* and *Radio Network Controllers* (RNCs). The *node B* is responsible for the transmission of information in one or more cells, to and from UEs. It also participates partly in the system resource management. The *node B* interconnects with the RNC via the Iub interface. The RNC controls resources in the system and interfaces the core network.

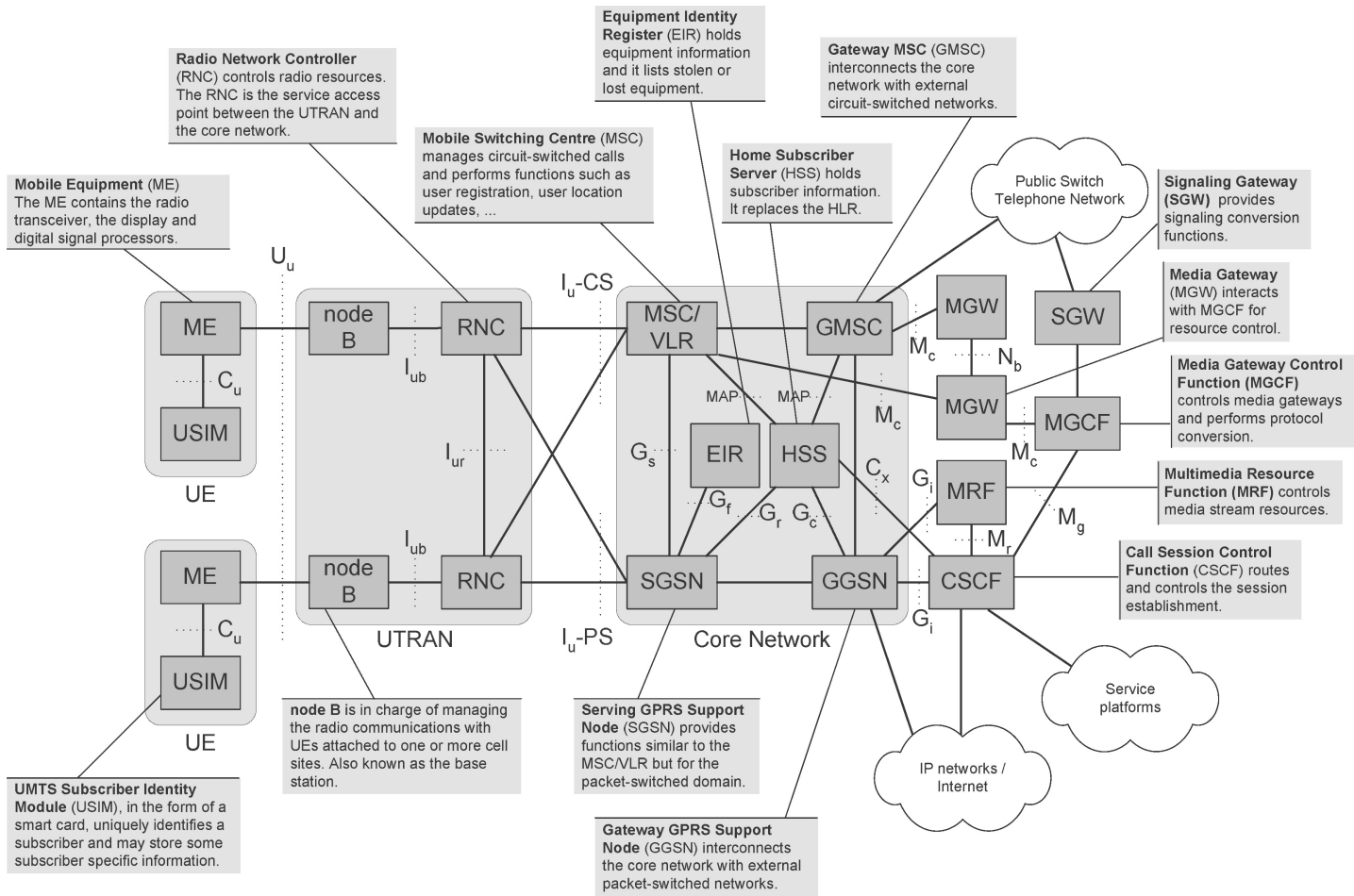
### 1.5.6 First Phase UMTS Core Network

The first phase UMTS core network is based on an evolved GSM network subsystem (circuit-switched domain) and a GPRS core network (packet-switched domain). Consequently, the UMTS core network is composed of the HLR, the MSC/VLR, and the Gateway MSC (to manage circuit-switched connections), and the SGSN and GGSN (to manage packet-based connections).

### 1.5.7 Second Phase UMTS

The initial UMTS architecture presented in this chapter is based on evolved GSM and GPRS core networks (providing support for circuit-switched and packet-switched domains, respectively). The objective of this initial architecture is to allow mobile network operators to rapidly roll out UMTS networks on the basis of existing GSM and GPRS networks. From this first phase UMTS architecture, the next phase is to evolve to an architecture with a core network based on an enhanced packet-switched domain only. The objective is to allow a better convergence with the Internet by using IP-based protocols whenever possible. At the end of 1999, 3GPP started the work on the specification of an “all-IP” architecture. In this architecture, the MSC function is split into a control plane part (MSC server) and a user plane part (media gateway). The core network of the second phase UMTS is interfaced with an IP Multimedia Subsystem, abbreviated IMS [3GPP-22.228][3GPP-23.228]. IMS introduces the capability to support IP-based multimedia services such as Voice over IP (VoIP). In IMS, call control is managed with the Session Initiation Protocol (SIP), published by IETF in [RFC-3261], and all network elements are based on IPv6.

IMS can be seen as an additional layer providing signaling, control, and charging functions for IP-based communications. In this context, service platforms that will initially benefit from IMS are the ones enabling services such as Push-To-Talk (PTT), presence, and location-based services. Figure 1.8 shows the architecture of a second phase UMTS solution.



**Figure 1.8** UMTS architecture (2nd phase)

## 1.6 Wireless Application Protocol

The Wireless Application Protocol (WAP) is the result of a collaborative work between many wireless industry players, carried out in the scope of the WAP Forum. The forum, launched in 1997 by Nokia, Phone.Com (now Openwave), Motorola, and Ericsson produced technical specifications enabling the support of applications over various wireless platforms (GSM, GPRS, UMTS, etc.). For this purpose, the WAP Forum identified and defined a set of protocols and content formats according to the standardization process presented in Chapter 2. In 2002, activities of the WAP Forum were transferred to another standardization organization: the Open Mobile Alliance.

### 1.6.1 Introduction to WAP

The WAP technology is an enabler for building applications (e.g., browsing, messaging, etc.) that run seamlessly over various wireless platforms. The objective of the WAP Forum was to provide a framework for the development of applications with a focus on the following aspects:

- *Interoperability*: applications developed by various parties and hosted on devices, produced by different manufacturers, interoperate in a satisfactory manner.
- *Scalability*: mobile network operators are able to scale services to subscribers' needs.
- *Efficiency*: the framework offers a quality of service suited to the capabilities of underlying wireless networks.
- *Reliability*: the framework represents a stable platform for deploying services.
- *Security*: the framework ensures that user data can be safely transmitted over a serving mobile network, which may not always be the home network. This includes the protection of services and devices and the confidentiality of subscriber data.

In line with these considerations, the WAP technology provides an application model close to the World Wide Web model (also known as the web model). In the web model, content is represented using standard description formats. Additionally, applications known as web browsers retrieve the available content using standard transport protocols. The web model includes the following key elements:

- *Standard naming model*: objects available over the web are uniquely identified by Uniform Resource Identifiers (URIs).
- *Content type*: objects available on the web are typed. Consequently, web browsers can correctly determine the type of a specific content.
- *Standard content format*: web browsers support a number of standard content formats such as the HyperText Markup Language (HTML).
- *Standard protocols*: web browsers also support a number of standard protocols for accessing content on the web. This includes the widely used HyperText Transfer Protocol (HTTP).

The WAP model borrows a lot from the successful web model. However, the web model, as it is, does not efficiently cope with constraints of today's mobile networks and devices. To

cope with these constraints, the WAP model leverages the web model by adding the following improvements:

- The *push technology* allows content to be pushed directly from the server to the mobile device without any prior explicit request from the user.
- The adaptation of content to the capabilities of WAP devices relies on a mechanism known as the *User Agent Profile* (UAProf).
- The support of *advanced telephony features* by applications, such as the handling of calls (establishment and release of calls, placing a call on hold, or redirecting the call to another user, etc.).
- The *External Functionality Interface* (EFI ) allows “plug-in” modules to be added to browsers and applications hosted in WAP devices in order to increase their overall capabilities.
- The *persistent storage* allows users to organize, access, store, and retrieve content from/to remote locations.
- The *Multimedia Messaging Service* (MMS) is a significant added value of the WAP model over the web model. It relies on generic WAP mechanisms such as the push technology and the UAProf to offer a sophisticated multimedia messaging service to mobile users. MMS is further described in Chapters 5 and 6.

The WAP model uses the standard naming model and content types defined in the web model. In addition, the WAP model includes the following:

- *Standard content formats*: browsers in the WAP environment, known as microbrowsers, support a number of standard content formats/languages including the Wireless Markup Language (WML) and the eXtensible HTML (XHTML). WML and XHTML are both applications of the eXtensible Markup Language (XML). See Box 1.1 for a description of markup languages for WAP-enabled devices.
- *Standard protocols*: microbrowsers communicate according to protocols that have been optimized for mobile networks, including the Wireless Session Protocol (WSP) and HTTP from the web model.

The first WAP technical specifications were made public in 1998 and have since evolved to allow the development of more advanced services. The major milestones for WAP technology were reflected in the availability of what the WAP Forum called “specifications suites.” Each specification suite contains a set of WAP technical specifications providing a specific level of features as shown in Table 1.2.

With WAP specification suite 1.x, the WAP device communicates with an application server via a WAP gateway. Communication between the WAP device and the WAP gateway is performed over WSP. In addition, WAP specification suite 2.x allows a better convergence of wireless and Internet technologies by promoting the use of standard protocols from the web model.

**Table 1.2** WAP Forum specification suites

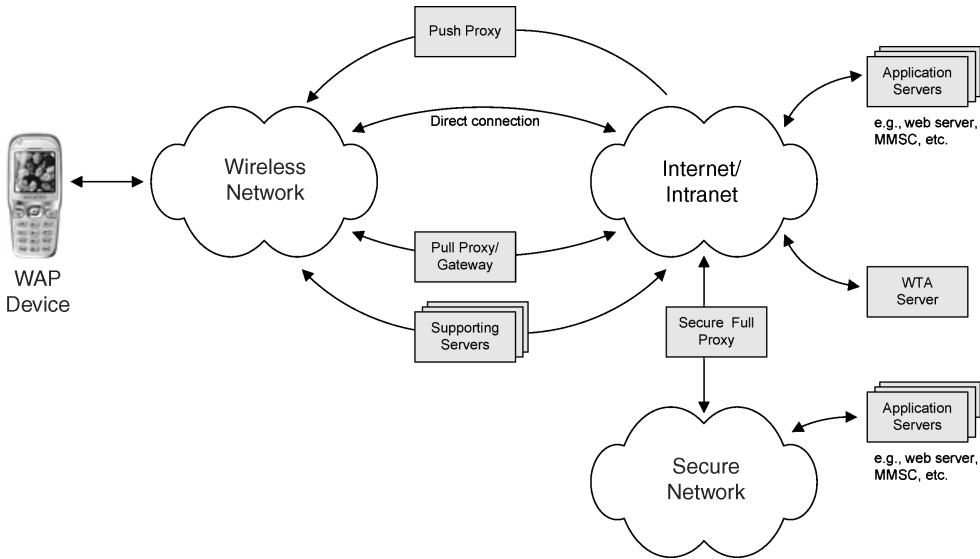
WAP Forum specification suites	Delivery date	Description
WAP 1.0	April 1998	Basic WAP framework Almost no available commercial solutions since the published standards did not allow the design of interoperable solutions
WAP 1.1	June 1999	First commercial solutions supporting: <ul style="list-style-type: none"> <li>- Wireless Application Environment (WAE)</li> <li>- Wireless Session Protocol (WSP)</li> <li>- Wireless Transaction Protocol (WTP)</li> <li>- Wireless Markup Language (WML)</li> <li>- WML script</li> </ul>
WAP 1.2	Nov. 1999	Additional features: <ul style="list-style-type: none"> <li>- Push technology</li> <li>- User Agent Profile (UAProf)</li> <li>- Wireless Telephony Application (WTA)</li> <li>- Wireless Identity Module (WIM)</li> <li>- Public Key Infrastructure (PKI)</li> </ul>
WAP 1.2.1	June 2000	Minor corrections
WAP 2.0	July 2001	Convergence with Internet technologies Additional features: <ul style="list-style-type: none"> <li>- Support of MMS 1.0 (3GPP Release 99)</li> <li>- HTTP, TCP, persistent storage</li> <li>- XHTML, SyncML, client provisioning, etc.</li> </ul>

**Box 1.1** Markup languages for WAP-enabled devices

The HyperText Markup Language (HTML) is the content format commonly used in the World Wide Web. HTML enables a visual presentation of information (text, images, hyperlinks, etc.) on large screens of desktop computers. eXtensible Markup Language (XML) is another markup language that is generic enough to represent the basis for the definition of many other dedicated languages. Several markup languages supported by WAP-enabled devices are derived from XML. This is the case of WML and XHTML. WML has been optimized for rendering information on mobile devices with limited rendering capabilities. The eXtensible HyperText Markup Language (XHTML) is an XML reformulation of HTML. Both WML and XHTML are extensible since the formats allow the addition of new markup tags to meet changing needs.

*1.6.2 WAP Architecture*

Figure 1.9 shows the components of a generic WAP architecture. The WAP device can communicate with remote servers directly or via a number of intermediary proxies and



**Figure 1.9** Generic WAP architecture

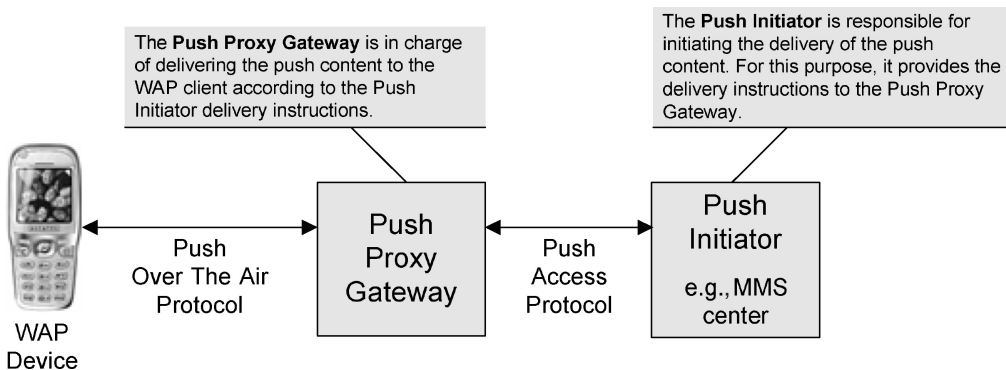
gateways. These proxies/gateways may belong to the mobile network operator or alternatively to service providers. The primary function of proxies/gateways is to optimize the transport of content from servers to WAP devices. Supporting servers, as defined by the WAP Forum, include Public Key Infrastructure (PKI) portals, content adaptation servers, and provisioning servers.

### 1.6.3 Push Technology

In a typical client/server model, a client retrieves the selected information from an application server by explicitly requesting the download of information from the server. This retrieval method is also known as the *pull technology* since the client pulls some data from a server. Web browsing is an example of models based on pull technology.

In contrast, another technology has been introduced in the WAP model and is known as the *push technology*. With push technology, a server is able to push some data to the WAP device with no prior explicit request from the client. In other words, the pull of information is always initiated by the client, whereas the push of information is always initiated by the server.

The push framework, defined by the WAP Forum in [WAP-250], is shown in Figure 1.10. In the push framework, the *push initiator* initiates the push transaction. The push initiator, usually an application server (e.g., web server, MMS center, etc.) transmits the content to be pushed along with XML-formatted delivery instructions to a *Push Proxy Gateway* (PPG). The PPG then delivers the push content to the WAP device according to the delivery instructions. The push initiator interacts with the PPG using the *Push Access Protocol* (PAP). On the other side, the PPG uses the *Push Over The Air* (OTA) protocol (based on WSP or HTTP) to deliver the push content to the WAP device.



**Figure 1.10** The push framework

The PPG may implement network-access-control policies indicating whether push initiators are allowed to push content to WAP devices. The PPG can send back a notification to the Push Initiator to indicate the status of a push request (delivered, canceled, expired, etc.).

Three types of browsing content can be pushed to a WAP microbrowser: Service Indication (SI), Service Loading (SL), and Cache Operation (CO). *Push SI* provides the ability to push content to users to notify them about electronic mail messages awaiting retrieval, news headlines, commercial offers, and so on. In its simplest form, a push SI contains a short message along with an URI. Upon receipt of the push SI, the message is presented to the user who is given the possibility of starting the service (retrieve the content) to which the URI refers. The subscriber may decide to start the service immediately or to postpone it. In contrast to push SI, *push SL* provides the ability to push some content to the WAP device without user explicit request. A push SL contains a URI that refers to the push content. Upon receipt of the push SL, the push content is automatically fetched by the WAP device and is presented to the user. *Push CO* provides a means for invalidating objects stored in the WAP device's cache memory.

In addition to browsing specific push contents, information can also be pushed to other WAP-based applications such as the WTA agent and the provisioning agent. The MMS client embedded in a WAP device also receives application-specific push messages to notify the user about the availability of new messages and for the delivery of reports.

#### 1.6.4 User Agent Profile

The *User Agent Profile* (UAProf) specification was first published in the WAP 1.2 specification suite, improved in WAP 2.0, and further enhanced recently by the Open Mobile Alliance. The objective of this specification is to define a method for describing the capabilities of clients and the preferences of subscribers. In practice, this description (known as a user agent profile) is mainly used for adapting available content to the rendering capabilities of WAP devices. For this purpose, the user agent profile is formatted using a Resource Description Framework (RDF) schema in accordance with Composite Capability/

Preference Profiles (CC/PP). The CC/PP specification defines a high-level framework for exchanging and describing capability, and preference information using RDF. Both RDF and CC/PP specifications have been published by W3C. UAProf, as defined by the WAP Forum and updated by OMA in [OMA-UAProf] (version 2.0), allows the exchange of user agent profiles, also known as *Capability and Preference Information (CPI)*, between the WAP device, intermediate network points, and the origin server (web server or MMS center). These intermediate network points and origin servers can use the CPI to tailor the content of WSP/HTTP responses to the capabilities of receiving WAP devices. The UAProf specification defines a set of *components* that WAP-enabled devices can convey within the CPI. Each component is itself composed of a set of attributes or *properties*. Alternatively, a component can contain a URI pointing to a document describing the capabilities of the client. Such a document is stored on a server known as a *profile repository* (usually managed by device manufacturers or by software companies developing WAP microbrowsers). UAProf is composed of the following components:

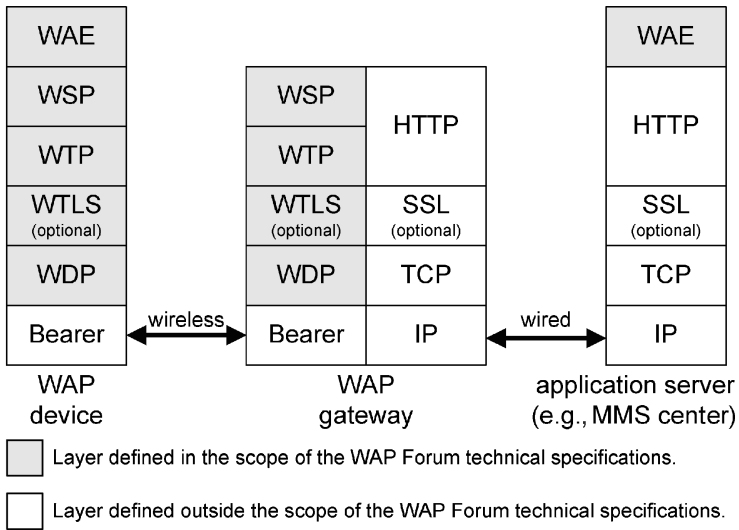
- *Hardware platform*: this component gathers a set of properties indicating the hardware capabilities of a device (screen size, etc.).
- *Software platform*: this component groups a set of properties indicating the software capabilities of a device (operating system, supported image formats, etc.).
- *Browser user agent*: this component gathers properties characterizing the Internet browser capabilities.
- *Network characteristics*: this component informs on network and environment characteristics such as the bearer capacity.
- *WAP characteristics*: this component advertises WAP browsing capabilities of the device. This includes information on the configuration of the WML browser and so on.
- *Push characteristics*: this component indicates push capabilities of the device. This includes the set of supported content types, the maximum message size that can be handled, and whether or not the device can buffer push messages.
- *MMS characteristics*: this component describes the device capabilities for retrieving and rendering multimedia messages (MMS version, maximum message size, supported content types, etc.).

For a configuration involving a WAP device and a gateway communicating with WSP, RDF descriptions can be encoded in binary with the WAP Binary XML (WBXML). In this context, the CPI is provided by the WAP device as part of the WSP session establishment request. The WAP device can also update its CPI at any time during an active WSP session. Note that the WAP gateway may also override a CPI provided by a device.

The use of UAProf in the context of MMS is further explained in Section 5.21.

### 1.6.5 WAP 1.x Legacy Configuration

With the objective of fulfilling the requirements of various services in heterogeneous mobile networks, several network configurations can coexist in the WAP environment. This section and the two following ones present the three most common configurations of the WAP environment: WAP 1.x legacy configuration, WAP HTTP proxy with wireless profiled TCP and HTTP, and HTTP with direct access.



**Figure 1.11** WAP 1.x legacy configuration with WAP gateway

Figure 1.11 shows the protocol stack of the configuration defined in the WAP specification suite 1.x. This configuration is also supported by the WAP specification suite 2.0 in addition to other configurations. In this configuration, the WAP device communicates with a remote server via an intermediary *WAP gateway*. The primary function of the WAP gateway is to optimize the transport of content between the remote server and the WAP device. For this purpose, the content delivered by the remote server is converted into a compact binary form by the WAP gateway prior to the transfer over the wireless link. The WAP gateway converts commands conveyed between datagram-based protocols (WSP, WTP, WTLS, and WDP) and protocols commonly used on the Internet (HTTP, SSL, and TCP).

The *Wireless Application Environment* (WAE) is a general-purpose application environment in which operators and service providers can build applications (e.g., MMS client or MMS center) for a wide variety of wireless platforms.

The *Wireless Session Protocol* (WSP) provides features also available in HTTP (requests and corresponding responses). Additionally, WSP supports long-lived sessions and the possibility to suspend and resume previously established sessions. WSP requests and corresponding responses are encoded in a binary form for transport efficiency.

The *Wireless Transaction Protocol* (WTP) is a lightweight transaction-oriented protocol. WTP improves the reliability over underlying datagram services by ensuring the acknowledgment and retransmission of datagrams. WTP has no explicit connection set-up or connection release. Being a message-oriented protocol, WTP is appropriate for implementing mobile services such as browsing. Optionally, Segmentation And Reassembly (SAR) of packets composing a WTP protocol unit can be supported as described in Section 1.6.8.

The optional *Wireless Transport Layer Security* (WTLS) provides privacy, data integrity, and authentication between applications communicating with the WAP technology. This includes the support of a secure transport service. WTLS provides operations for the establishment and the release of secure connections.

The *Wireless Data Protocol* (WDP) is a general datagram service based on underlying low-level bearers. WDP offers a level of service equivalent to the one offered by the Internet's User Datagram Protocol (UDP).

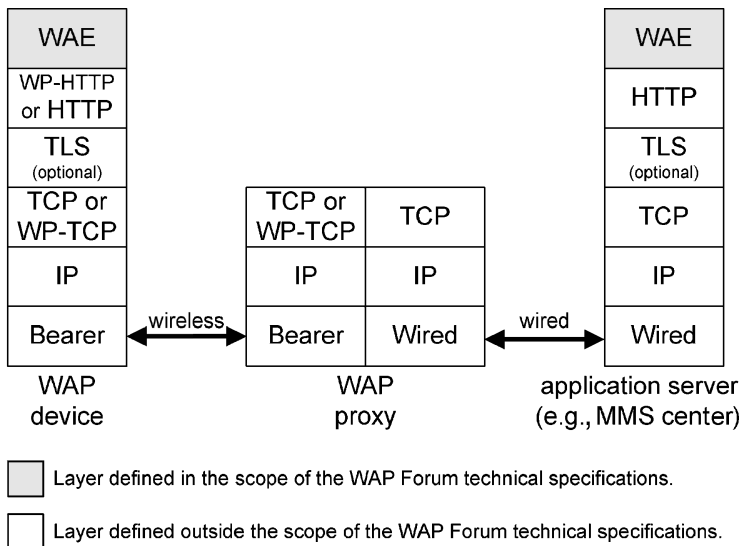
At the bearer level, the connection may be a circuit-switched connection (as found in GSM networks) or a packet-switched connection (as found in GPRS and UMTS networks). Alternatively, the transport of data at the bearer level may be performed over the Short Message Service (e.g., for push messages) or over the Cell Broadcast Service.

### 1.6.6 WAP HTTP Proxy with Wireless Profiled TCP and HTTP

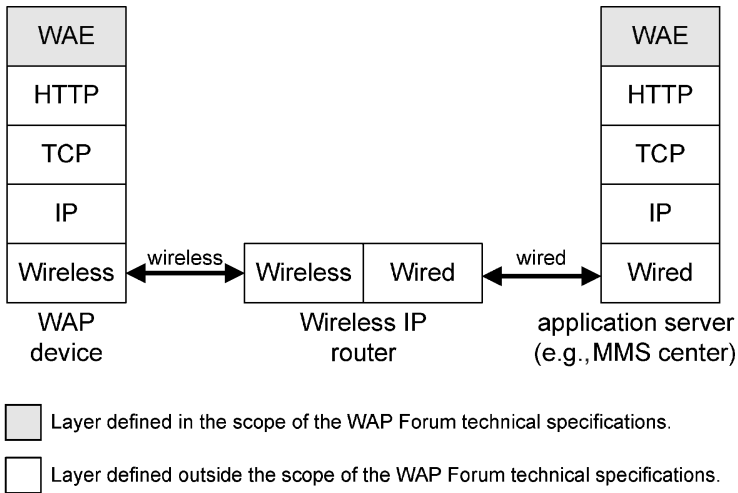
Figure 1.12 shows a configuration in which the WAP device communicates with application servers via an intermediary WAP proxy. The primary role of the proxy is to optimize the transport of content between the fixed Internet and the mobile network. It also acts as a Domain Name Server (DNS) for mobile devices. With this configuration, Internet protocols are preferred against legacy WAP protocols. This is motivated by the need to support IP-based protocols in an end-to-end fashion, from the application server back to the WAP device. The protocol stack of this configuration, defined in the WAP specification suite 2.0, is shown in Figure 1.12.

The *Wireless Profiled HTTP* (WP-HTTP) is an HTTP profile specifically designed for coping with the limitations of wireless environments. This profile is fully interoperable with HTTP/1.1 and supports message compression.

The optional *Transport Layer Security* (TLS) ensures the secure transfer of content for WAP devices involved in the exchange of confidential information.



**Figure 1.12** Configuration with WAP proxy



**Figure 1.13** WAP configuration with direct access

The *Wireless Profiled TCP* (WP-TCP) offers a connection-oriented service. It is adapted to the limitations of wireless environments but remains interoperable with existing Transmission Control Protocol (TCP) implementations.

### 1.6.7 HTTP with Direct Access

Figure 1.13 shows a configuration where the WAP device is directly connected to the application server (via a wireless router that provides a bearer-level connection). The protocol stack shown in this configuration is defined in the WAP specification suite 2.0. A WAP device, compliant with the WAP 2.0 specification suite, may support all configurations by supporting WAP 1.x and WAP 2.0 protocol stacks.

### 1.6.8 WTP Segmentation and Reassembly

In the WAP 1.x legacy configuration, an optional *Segmentation And Reassembly* (SAR) mechanism [WAP-224] allows large transactions to be segmented at the WTP level by the sender and reassembled by the receiver. SAR is specifically used when the size of a transaction (e.g., retrieval of a 50-KB multimedia message) exceeds the WTP Maximum Transmission Unit (MTU). In the context of MMS, SAR is used for transactions including the sending and retrieval of large messages. Note that, in the WAP 1.x configuration, SAR is optional and if it is not supported at the WTP level, then segmentation and reassembly may be supported at an underlying layer (e.g., [RFC-791] for IP, [3GPP-23.040] for SMS, etc.).

With SAR, the WTP transaction is segmented into several packets and packets can be sent by the sender in the form of packet groups. For efficiency, the receiver acknowledges the reception of each single packet group and the sender does not start transmitting packets of a new group if the previous group has not been properly acknowledged by the receiver. A

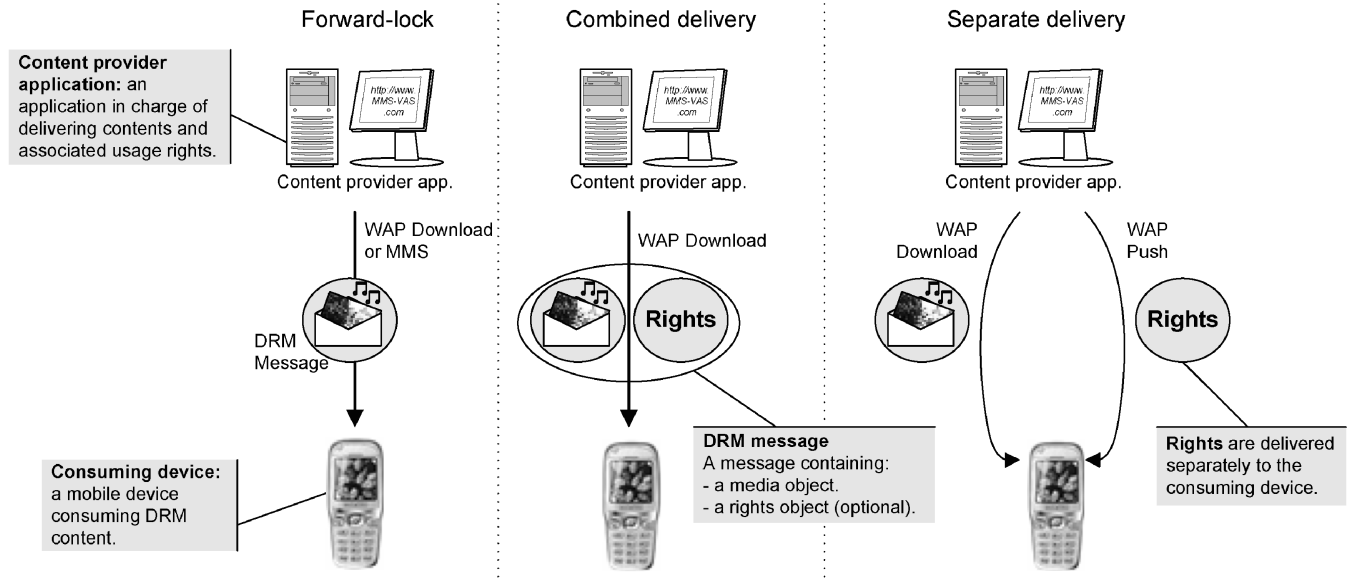


Figure 1.14 OMA digital rights management

group can contain a maximum number of 256 packets. The sender determines the number of packets in a group, preferably according to the characteristics of the network and the ones of the device. The first packet group is sent without knowing the characteristics of the receiver. Therefore, the size of the first packet group should not be too large. SAR allows a selective retransmission of multiple lost packets for a given group. This feature minimizes the number of packets sent over WTP.

### 1.6.9 OMA Digital Rights Management

At the end of 2002, OMA published technical specifications [OMA-DRM] for mechanisms representing the basis for the management of digital rights associated with media objects downloaded via WAP download or MMS. *Digital Rights Management* (DRM) provides a means, for operators and content providers, to control the usage of media objects once they have been downloaded to a mobile device (also known as a 'consuming device' in the DRM context). DRM enables content providers to define usage rules specifying the user's rights regarding the usage of the corresponding media object. For instance, a content provider can grant a user the rights to preview for free and charge for more sophisticated usages. Three main mechanisms are defined in OMA-DRM as shown in Figure 1.14. They differ in the way rights are communicated to the consuming device and are as follows:

- *Combined delivery* consists of delivering the media object along with the associated rights in a single DRM message.
- *Forward lock* is the simplest of the OMA-DRM mechanisms. This is a special case of the combined delivery mechanism in which the DRM message contains only the media object, without the associated rights. For forward lock, the following set of rights applies: the user is not allowed to forward or modify the media object.
- With *separate delivery*, the media object and corresponding rights are conveyed to the consuming device over separate distribution channels. In this context, the media object is converted into a *DRM Content Format* (DCF) [OMA-DRM-CF]. This conversion consists of a symmetric encryption of the original media object, making the converted object unusable, unless the consuming device has the necessary Content Encryption Key (CEK) to convert the object back to its original form. The CEK along with the rights is delivered to the consuming device separately from the associated media object, typically over WAP push.

OMA DRM forward lock is of particular interest to the content-to-person scenario of MMS and is applicable from MMS 1.2, and the support of combined and separate deliveries has also been introduced in MMS 1.3.

The application of OMA DRM in the context of MMS is further explained in Section 5.31.

