# Wireless Multimedia in 3G networks

George Xylomenos and Vasilis Vogkas

xgeorge@aueb.gr and vvogkas@aueb.gr

Mobile Multimedia Laboratory
Department of Informatics
Athens University of Economics and Business
Patision 76, Athens 104 34, Greece

## I. Introduction

This chapter describes the support for IP based multimedia services on *Third Generation* (3G) wireless cellular networks. While earlier cellular networks offered only basic IP connectivity on a best-effort basis, 3G networks provide explicit support for real time multimedia services with guaranteed *Quality of Service* (QoS). As a result, they can offer traditional voice telephony, rich telephony services and multimedia streaming over IP, both inside a 3G network and in conjunction with external networks such as the Internet, the *Public Switched Telephone Network* (PSTN) and the *Integrated Services Digital Network* (ISDN). In fact, 3G networks will deploy for the first time in a large scale technologies such as *IP version 6* (IPv6) and policy based IP QoS, spearheading the introduction of such technologies into the Internet.

The two most important aspects of 3G networks with respect to IP based multimedia services are the *IP Multimedia Subsystem* (IMS) and the *Multimedia Broadcast / Multicast Service* (MBMS). The IMS enables complex IP based multimedia sessions to be created with guaranteed QoS for each media component. Example applications include voice telephony and video conferencing. The IMS interoperates with both traditional telephony services and external IP based multimedia services. The MBMS provides native IP broadcast and multicast support in 3G networks, allowing high bandwidth services to be economically offered to multiple users. Example applications include video streaming via multicast and location based services via broadcast. The MBMS interoperates directly with IP multicasting. While both the IMS and the MBMS are IP based, their standardization is proceeding independently. It is however clear that their combination would allow numerous new services to be provided.

The outline of this chapter is as follows. In Section II we give an overview of cellular networks and their evolution, while in Section III we describe in detail a specific 3G network, the *Universal Mobile Telecommunications System* (UMTS). Section IV presents an introductory description of the features and services provided by the IMS and the MBMS. In Section V we provide the details of the IMS, including service architecture, session setup and control and interworking issues. Section VI describes the MBMS in the same manner as for the IMS. Finally, Section VII discusses the QoS issues for IP based multimedia services, describing the overall QoS concept, the policy based QoS control scheme and its application to IMS sessions. At the end of the chapter, Tables II and III list the acronyms used. An extensive UMTS vocabulary is provided in [1].

## II. Cellular networks

This section provides an overview of cellular network evolution, in terms of both technology and services. *First generation* (1G) systems used analog transmission and provided only circuit switched voice telephony. *Second generation* (2G) systems were fully digital and initially offered voice and circuit switched data services. The increasing popularity of the Internet led to the addition of packet switched data services to 2G networks, turning them into 2.5G networks. Finally, 3G networks are planning to provide all services over packet switching, including voice telephony. This focus on packet switching enables new services to be offered, including IP based multimedia ones.

### A. First generation

All wireless systems face the problem that all transmissions must share the frequency range allocated to the system, thereby limiting the number of simultaneous users. One means to increase the number of users for a given frequency range is the *cellular concept*, illustrated in Figure 1. The area to be covered is divided into hexagonal cells, with a *Base Transceiver Station* (BTS) located at the center of each cell. The BS transmits with just enough power to reach the outer limits of its cell. Each user has a *Mobile Station* (MS), i.e. a cellular telephone. Depending on the cell where the MS is located, it communicates with the network via the corresponding BTS. To avoid interference between transmissions in neighboring cells, the available frequency range is divided into non-overlapping frequency bands and different bands are assigned to adjacent cells. However, the same band can be reused in a non-adjacent cell, thereby increasing the total number of users that can be simultaneously served in the area of coverage.

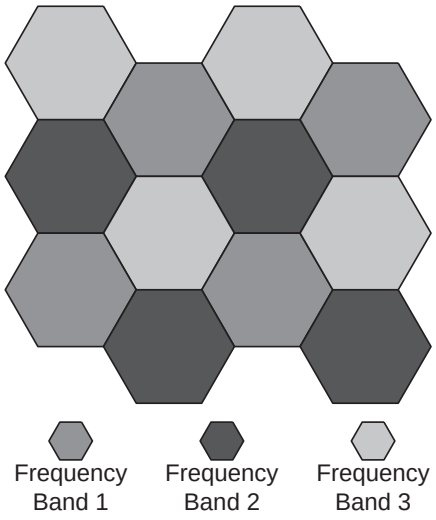


Fig. 1. The cellular concept.

The cellular network infrastructure consists of a BTS in each cell plus a network interconnecting them. The cellular network must also provide connectivity to external networks so that an MS may place calls to or receive calls from fixed telephony networks. When users move to another cell while making a call, a *handover* occurs. That is, the cellular network ensures that the MS transparently switches to the frequency band used by the BTS of its new cell without interrupting the call. When multiple cellular operators exist in the same area, they are allocated different frequency ranges from the local licensing authority. Each operator can deploy a completely different cell structure and frequency band allocation, using its own network to interconnect these cells.

The oldest cellular system, the *Advanced Mobile Phone Service* (AMPS), started operation in the U.S.A. during the 1970s. It is based on the cellular concept described above. Active calls within the same cell share the frequency band via *Frequency Division Multiple Access* (FDMA), that is, each call between an MS and the BTS is allocated a separate part of the frequency band. In addition, the transmissions from the MS to the BTS, called the uplink, and from the BTS to the MS, called the downlink, use separate frequency ranges, in a technique called *Frequency Division Duplexing* (FDD). Note the distinction between FDMA and FDD: the former refers to frequency sharing between different calls, while the latter refers to frequency sharing between the two directions of the same call. The only service offered by AMPS is voice telephony. Voice is transmitted in analog mode, that is, a carrier of the appropriate frequency is modulated by the voice signal. It is also possible to transmit data using an analog modem to modulate the carrier. Since each voice call is allocated a frequency band of 30 KHz, the resulting data rate is limited.

During the 1980s many other analog cellular systems were deployed in various countries, especially in Europe, with each system used in at most a few countries. All these systems have practically vanished due to their low market penetration and the emergence of much more popular 2G systems. In contrast, AMPS is still in operation in the U.S.A., and it actually remains quite popular since it is the only system providing coverage in the entire country.

### *B. Second generation*

The increasing use of digital transmission and switching inside the fixed telephony network, the PSTN, led during the 1980s to the introduction of its fully digital successor, the ISDN. Similarly, 2G digital cellular systems were introduced to replace their analog predecessors. Realizing that their national markets were too small to sustain a 2G cellular system able to compete with AMPS, numerous European manufactures and operators collaborated to create a single pan-European 2G standard. The outcome was the *Global System for Mobile Communications* (GSM), a fully digital cellular system.

While GSM inherits the cellular concept from AMPS and uses FDD to separate the frequencies used for the uplink and downlink transmissions of a call, it differs from AMPS in nearly every other aspect. Voice is digitized and compressed before transmission, therefore each call requires much less bandwidth. Rather than allocating a different frequency range to each call, GSM first divides the frequency band of each cell into smaller bands using FDMA. In each band *Time Division Multiple Access* (TDMA) is used, that is, time is divided into slots of equal duration and each call gets to use a single slot periodically. As a result, multiple calls share the same band by transmitting in different time slots. The advantage of TDMA is that it can support calls that require more bandwidth by allocating them more slots per period.

The original version of GSM, called Phase 1, was purely circuit switched, like the PSTN and ISDN, therefore only circuit switched data services were available. In contrast to the PSTN, and similarly to the ISDN, since GSM is digital there is no need to use a modem; data are transmitted directly in the digital GSM format. Since compressed digitized voice requires lower

bandwidth than uncompressed analog voice, the bandwidth available to GSM circuits was low however; circuit switched data calls only supported speeds of up to 9.6 kbps.

In order for GSM to establish voice calls with the PSTN, an *InterWorking Function* (IWF) is needed at the edge of the GSM network to convert between analog and digital voice and signaling. For data calls, the IWF uses a modem at the PSTN side and relays data from (to) the digital GSM circuit to (from) the analog PSTN circuit. While the ISDN is also digital, the voice and data formats it uses are different than those of GSM, therefore an IWF is also used in this case to convert between the two digital formats.

In parallel with GSM in Europe, other digital 2G systems were deployed in the U.S.A. and other parts of the world. The *Digital Advanced Mobile Phone Service* (D-AMPS) adds digital transmission to the AMPS system. Since D-AMPS needs 10 KHz per call, it can multiplex three digital calls in the frequency band consumed by one analog call, thus tripling system capacity. By using the same frequency allocations as AMPS, D-AMPS can coexist with it, thus allowing gradual migration from analog to digital.

A more radical departure from AMPS is CDMAone which uses the *Code Division Multiple Access* (CDMA) scheme to share the available frequency band between users. With CDMA each call employs the complete frequency range all the time, regardless of the cell that it takes place in. However, each call uses a different code to scramble its data at a very high speed. By the choice of appropriate codes and coding and decoding techniques, CDMA systems have the remarkable property that each call distinguishes its own data, treating all other transmissions as random noise. Each cell is assigned a different primary code and all calls in that cell employ secondary codes derived from the primary one. CDMA can also support calls that require higher bandwidth by allocating them additional secondary codes.

While GSM was the only 2G system in the entire European market, in the U.S.A. three 2G systems competed, D-AMPS, CDMAone and GSM. In addition, while GSM did not have any serious 1G competitors in Europe, AMPS was a well entrenched analog competitor in the U.S.A. As a result, regardless of the technical merits of each 2G system, GSM became the most popular 2G system due to its dominance in Europe, and it was deployed in numerous countries around the world. In the following we will focus on the evolution of GSM.

The increased importance of data services, and especially connectivity to the Internet, has led to the introduction of additional services in the latest version of GSM, called Phase 2+. The *High Speed Circuit Switched Data* (HSCSD) service allows circuit switched calls to use more bandwidth by allocating them more TDMA slots per period. This is roughly equivalent to combining multiple circuits to increase the data rate. As a result, circuit switched data calls can reach speeds of up to 64 kbps. Unfortunately, circuit switching means that these resources are tied up whether the call has any data to transmit or not.

In contrast, the *General Packet Radio Service* (GPRS) offers packet switched data calls. When a GPRS call has data to transmit, the system dynamically allocates it some TDMA slots, but only for the duration of the transmission. As a result, many GPRS calls can dynamically share the available bandwidth without being charged when they are idle. By allocating multiple TDMA slots per period for each transmission, GPRS provides speeds of up to 171 kbps. GPRS packets are transported inside the cellular network using a packet switched backbone, separate from the circuit switched backbone used for voice calls. This packet switching backbone originally provided support for many data protocols, but eventually only IP survived.

### *C. Third generation*

While the success of the various 2G systems has been spectacular, the fierce competition between them and their technological limitations have led to an effort for the standardization of a worldwide digital 3G cellular system. This system, besides using the most advanced technology available, should place emphasis on packet switched data services, especially IP based multimedia. This effort began in the *International Telecommunications Union* (ITU) with the name *International Mobile Telecommunications 2000* (IMT-2000). The IMT-2000 system would replace all 2G systems, providing a way for operators to gradually migrate from 2G to 3G. Even though all parties agreed that this system should be based on CDMA, there was disagreement on the details and on the evolutionary path from 2G to 3G.

The result was the formation of two separate groups, comprised of operators, manufacturers and regulators. The *3G Partnership Project* (3GPP) is designing a system based on *Wideband CDMA* (W-CDMA) for the radio part and GSM for the network backbone, while the *3G Partnership Project 2* (3GPP2) is designing a system based on CDMA2000, an evolution of the CDMAone system [2]. While both projects are moving in the same general direction and are co-ordinated to a large extent, 3GPP is one step ahead in the area of support for IP based multimedia. In the following we will concentrate on the 3GPP system, usually referred to as the *Universal Mobile Telecommunications System* (UMTS).

A UMTS network consists of two parts, the *Radio Access Network* (RAN) and the *Core Network* (CN). The *User Equipment* (UE) connects to the network via the RAN. The RAN is composed of all the network elements providing users with radio access to the UMTS; it is inherently tied to a specific wireless technology. The CN is composed of all the network elements providing UMTS services to the users; it is independent of the wireless technology used. As shown in Figure 2, this separation allows different RAN and CN elements to be combined, thus providing multiple evolutionary paths from 2G to 3G.

For the CN the options are the GSM based circuit switched network and the GPRS based packet switched network, both already provided by 2G networks. For the RAN the options are the *GSM EDGE RAN* (GERAN) and the *Universal Terrestrial*

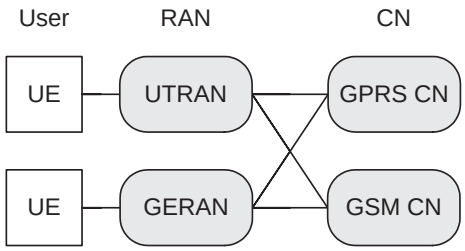


Fig. 2. Components of a UMTS network.

*RAN* (UTRAN). The GERAN is based on the *Enhanced Data rates for GSM Evolution* (EDGE) technology, which reuses the frequency allocations of GSM and provides higher bandwidths by using more advanced modulation and coding schemes [3]. The UTRAN is based on the new W-CDMA technology. An operator may migrate to 3G by first upgrading the CN components to the UMTS specifications and using the GERAN for radio access. The UTRAN can be introduced later, when its high deployment cost is justified.

The UMTS specifications were originally released yearly by the *European Telecommunications Standards Institute* (ETSI), and they were named accordingly, e.g. Release 97. Starting from Release 99 the 3GPP assumed responsibility for UMTS. This release incorporates GSM Phase 2+, which includes GPRS, and the UTRAN. The numbering scheme was changed after Release 99. The next step, Release 4, introduces modifications to the CN and more radio access options. Release 4 also allows IP to be used inside the CN to transport both voice and data. Release 5 introduces the IMS and allows IP to be used inside the RAN to transport both voice and data. Release 6 is currently under development, introducing the MBMS and further extending the IMS [4].

## III. UMTS NETWORKS

This section presents an overview of a UMTS network, based on the 3GPP Release 6 specifications. We first discuss the overall service architecture and its impact on the UE and then separately describe the CN and the RAN in detail. We omit the IMS and MBMS specific functionality, as well as QoS issues, which are covered in later sections.

### A. Services and service capabilities

Unlike 1G networks that provided only a single service, i.e. voice telephony, 2G networks provide multiple services, divided into three types. A *bearer service* is a signal transmission facility inside the network; the signal may be voice. A *teleservice* combines a bearer service with functionality in the terminals to provide a service to the user, such as voice telephony. Finally, a *supplementary service* provides additional facilities for another service, such as call forwarding.

While 3G networks provide all 2G teleservices and supplementary services, for compatibility, the UMTS architecture emphasizes *service capabilities*, that is, parametric bearer services and mechanisms required to implement services. This allows a UMTS network to offer new services based on the same service capabilities, without requiring changes to the network architecture. For this reason, Release 6 specifications only define those capabilities sufficient to implement the required services, without standardizing the services themselves [5]. This emphasis on service capabilities is reflected in the choice of IP for transport: packet switched IP bearers can be used to provide both voice and data services, unlike circuit switched bearers which are more suitable for voice.

Another area where this separation is clear is on the UMTS definition of the UE. Unlike earlier networks where a cellular phone provided both communication and application functionality, i.e. wireless access and voice telephony, the UE in UMTS is split into the *Mobile Terminal* (MT), providing communication, and the *Terminal Equipment* (TE), providing application functionality. A cellular phone can thus either include application functionality, or it may interface to an external device providing application functionality, such as a notebook computer or a personal digital assistant [6].

### B. Core network

The main function of the CN is to provide routing and switching for user and control traffic. The CN is divided into the *circuit switched* (CS) and *packet switched* (PS) domains, as shown in Figure 3. The CS domain is an evolution of GSM, while the PS domain is an evolution of GPRS, modified in both cases so as to handle the new UMTS services. Both domains employ the *Home Subscriber Server* (HSS) that contains all information related to a user, including user preferences, authentication data and location management information.

In the CS domain, calls are handled by two *Mobile services Switching Centers* (MSC): the *Gateway MSC* (GMSC) is located at the user's home network and the *Visitor MSC* (VMSC) is located at the network the user is currently visiting. When the user enters a new network, its VMSC informs the local *Visitor Location Register* (VLR) about the user, and the VLR in turn

informs the appropriate HSS that the user is currently located there. Incoming calls are first directed to the GMSC, which asks the HSS about the user's current location and then directs the call to the appropriate VMSC. Outgoing calls use the reverse path, from the VMSC to the GMSC in the user's home network. For ISDN or PSTN originated or terminated calls, a *Media GateWay* (MGW) is employed by the GMSC for the appropriate translations.

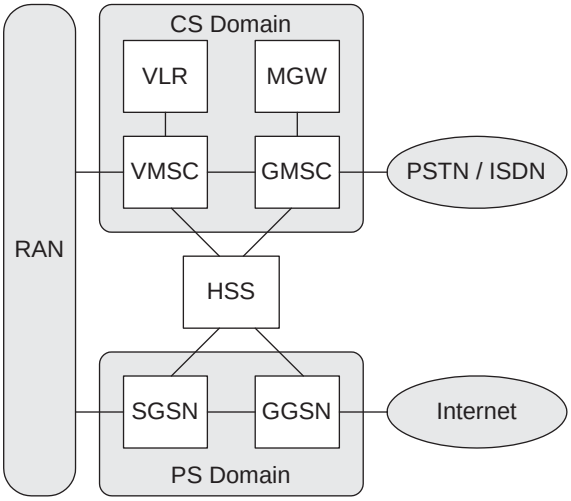


Fig. 3. Core Network architecture.

In the PS domain, calls are similarly handled by two *GPRS Support Nodes* (GSN): the *Gateway GSN* (GGSN) is located at the user's home network and the *Serving GSN* (SGSN) is located at the network the user is currently visiting. While the HSS provides all information related to the user, the GGSN always knows the current SGSN handling the user, therefore there is no need for a VLR in the PS domain. For Internet originated or terminated calls, the GGSN acts as an IP gateway router between the UMTS network and the Internet.

In Release 99, user and control data in both the CS and PS domains were transferred over *Asynchronous Transfer Mode* (ATM) virtual circuits, with *ATM Adaptation Layer 2* (AAL2) used for the CS domain and *ATM Adaptation Layer 5* (AAL5) used for the PS domain. However, starting with Release 4 the CN can employ IP for data transfer in both the CS and PS domains. This reflects the increased importance of IP in UMTS networks.

Since the PS domain is based on GPRS, some additional details on GPRS are provided below. The GPRS was originally designed to efficiently transport any type of packet data. In UMTS only IP is supported, in unicast and multicast modes. In order for a UE to gain IP connectivity, it must first attach to the GPRS. During this procedure the user is authenticated by the HSS and its local SGSN creates a mobility management context to handle the UE. At this point the UE is known to the UMTS network but it is not yet able to communicate. The next step is for the UE to create a *Packet Data Protocol* (PDP) context at the GGSN in its home network. The PDP contains the IP address assigned to the user, which is forwarded to the UE and the SGSN. At this point the user can send and receive IP packets via the SGSN and the GGSN [7].

### *C. Radio access network*

The main function of the RAN is to provide connectivity between the UE and the CN. The first RAN option supported by UMTS is the GERAN, as shown in Figure 4. A GERAN covering a large area is called a *Base Station Subsystem* (BSS). The BSS is divided into smaller regions, with each region controlled by a *Base Station Controller* (BSC). Each BSC region is divided into cells, with each cell served by a BTS. The GERAN is basically an evolution of the GSM radio network with GPRS support. Each circuit switched channel can either be allocated to a single voice call or to multiple packet data calls. For channels allocated to packet data, the system dynamically allocates one or more TDMA slots in each period to each UE that needs to transmit or receive packets, but only for the duration required. Uplink and downlink TDMA slots are allocated separately, thus efficiently supporting asymmetric services, such as file downloads. Depending on the number of TDMA slots allocated to a transmission and the coding scheme used over the channel, GPRS can support data rates between 9 kbps and 171 Kbps.

The EDGE version of GSM offered by the GERAN provides more advanced modulation and coding schemes, supporting data rates of at least 384 kbps for urban environments and 144 kbps for rural environments, without changing the FDMA or TDMA structures of GSM. As a result, the GERAN can be used to support many of the services offered by UMTS networks. However, compatibility with GSM means that the system is fundamentally limited, therefore services requiring higher data rates must resort to a different technology.

This technology is the UTRAN, the second RAN option supported by the UMTS, also shown in Figure 4. In the UTRAN one or more cells are served by a Node-B. Multiple Node-Bs are connected to a *Radio Network Controller* (RNC), and multiple

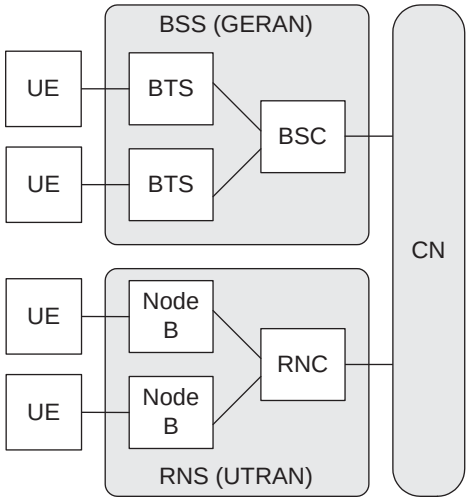


Fig. 4. Radio Access Network architecture.

RNCs form a *Radio Network Subsystem* (RNS). The UTRAN uses W-CDMA to share the available frequency range between multiple simultaneous calls, with each call assigned one or more codes depending on required bandwidth. The W-CDMA system can operate in two modes: in FDD mode uplink and downlink transmissions take place in different frequency bands, while in TDD mode uplink and downlink transmissions use the same frequency band but are multiplexed in time. The TDD mode is more flexible in terms of spectrum allocation but also more complicated in terms of synchronization.

The UTRAN, besides matching the speeds provided by the GERAN, can also support speeds of more than 2048 kbps in small cells or indoor areas. To achieve these high data rates, besides exploiting the flexible bandwidth allocation scheme of W-CDMA, the UMTS network also employs power control. This means that the transmissions between a UE and a Node-B use the minimum amount of power required to achieve the required level of reliability, depending on the distance between them. As a result, the interference among different calls is minimized, and the system provides higher capacity.

Another area where the UTRAN is superior to the GERAN is in the handling of handovers. A handover is performed when a UE making a call moves to a new cell. Since the GERAN uses different frequency bands in each cell, it only supports *hard handover* in which the radio link in the old cell is disconnected before establishing a radio link in the new cell. In contrast, the UTRAN uses the same frequency band everywhere, with cells differentiated only by the primary code used in each cell. The UTRAN therefore supports *soft handover* in which the radio link to the new cell is established before removing the radio link to the old cell. As a result, the user never loses connectivity during a handover.

## IV. Multimedia services

Besides combining multiple media components, multimedia services include at least one continuous, i.e. time sensitive, media component, such as audio or video, and they require all media components to be synchronized with each other. Therefore, while simple IP connectivity allows diverse media to be transmitted over IP in cellular networks, real multimedia services require additional support from the network, at least in the area of QoS provisioning for the continuous media components. Despite many efforts to provide such support, the Internet remains a best effort network, providing no guarantees about end-to-end packet transmission delay or reliability. In contrast, UMTS networks have made significant progress in this direction. One aspect of this progress is the addition of the IMS and MBMS components that will be introduced in this section and analyzed in the two following sections. Another aspect is the provisioning of guaranteed QoS that will be discussed in the last section.

### *A. IP multimedia subsystem*

The *IP Multimedia Subsystem* (IMS) enhances the basic IP connectivity of UMTS by adding network entities that handle multimedia session setup and control and QoS provisioning. Note that the IMS uses the term *session* in place of *call*: a session includes the senders, receivers and data streams participating in an application. The new IMS entities ensure that multimedia sessions will be able to reserve the resources they need in order to perform satisfactorily. A session is able to request different QoS levels for each of its media components and modify these levels during its lifetime.

Following the general philosophy of UMTS, the IMS does not standardize any applications, only the service capabilities required to build various services. As a result, real-time and non real-time multimedia services can be easily integrated over a common IP based transport. These services can directly interwork with all the services available over the Internet. Eventually, even legacy circuit switched services like voice may be replaced inside the UMTS network by real-time IP based packet switched services, making the CS domain obsolete.

Some of the services that can be provided over IMS are voice and video telephony, rich telephone calls, presence services, instant messaging, chat rooms, voice and video conferencing and multiparty gaming. While the possible services are numerous, they are all based on a small set of capabilities [8], [9]:

- Endpoint identities, including telephone numbers and Internet names.
- Media description capabilities, including coding formats and data rates.
- Person-to-person real-time multimedia services, including voice telephony.
- Machine-to-person streaming multimedia services, including TV channels.
- Generic group management, enabling chat rooms and messaging.
- Generic group communication, enabling voice and video conferencing.

These basic services can be controlled by an external *Application Server* (AS) in order to provide actual applications. For example, the IMS does not offer a conferencing or chat room service, but it provides i) point-to-point and point-to-multipoint transmission facilities, ii) group management facilities, and iii) the ability for an external AS to control the group communication. Depending on the functionality of the AS, the application built on top of these capabilities may be a video conference or a chat room.

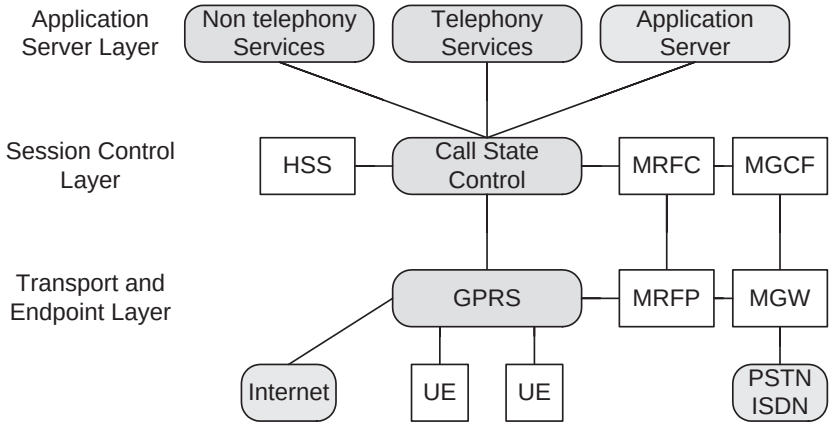


Fig. 5. IMS layered service architecture.

To maximize flexibility, the IMS organizes its functionality in three layers, as shown in Figure 5 (see Section V for the acronyms). The transport and endpoint layer initiates and terminates the signaling needed to setup and control sessions, provides bearer services and support for media conversions. The session control layer provides functionality that allows endpoints to be registered with the network, sessions to be setup between them and media conversions to be controlled. In this layer multiple transport services may be combined in a single session. The application server layer allows sessions to interact with various AS entities. In this layer multiple sessions may be combined in a single application.

### *B. Multimedia broadcast / multicast service*

When packets must be transmitted to many users in a cell, it is more economical to transmit them only once over a common channel received by all users. Since Release 4, UMTS networks have provided a low bandwidth *Cell Broadcast Service* (CBS) that can transmit short messages to all users in a particular region, called the cell broadcast area [10]. CBS messages are unacknowledged, and each one may be directed to a different cell broadcast area, i.e. set of cells. These messages may originate from a number of information providers, which transmit them through a Cell Broadcast Center (CBC). The CBC broadcasts each message periodically, at a frequency and duration arranged between the CBC and the information provider. The frequency normally depends on message content. For example, volatile content such as road traffic reports will probably require more frequent transmissions than weather reports.

The CBS is targeted to text messages, therefore it is unsuitable for multimedia services, due to the high bandwidth these services require. Since Release 99, UMTS networks have also supported IP multicasting, in which IP packets are forwarded to all users belonging to a multicast group. The multicast group is identified by a class D multicast IP address. Unfortunately, as shown in Figure 6, this service is implemented by separately sending packets from the GGSN to each UE. Since multicast packets are sent separately to each receiver in the same cell, no sharing gain is achieved and high bandwidth multimedia services cannot be provided [11].

In order to overcome the limitations of these options, the Release 6 specifications include the *Multimedia Broadcast / Multicast Service* (MBMS) that supports native multicasting and broadcasting over UMTS networks. MBMS is interoperable with IP multicasting, that is, IP multicast packets can be transmitted over MBMS. As shown in Figure 7, the GGSN and SGSN send multicast packets only once to each downstream node. More importantly, these packets are transmitted only once over

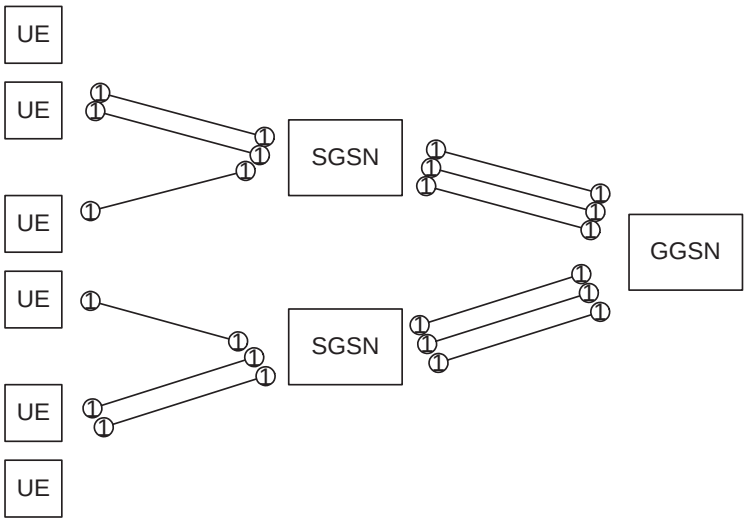


Fig. 6. Multicasting without MBMS.

the wireless link, regardless of the number of receivers in a cell. Note that the services provided by MBMS are unidirectional, that is, data are transmitted only from the GGSN to the UE [12].

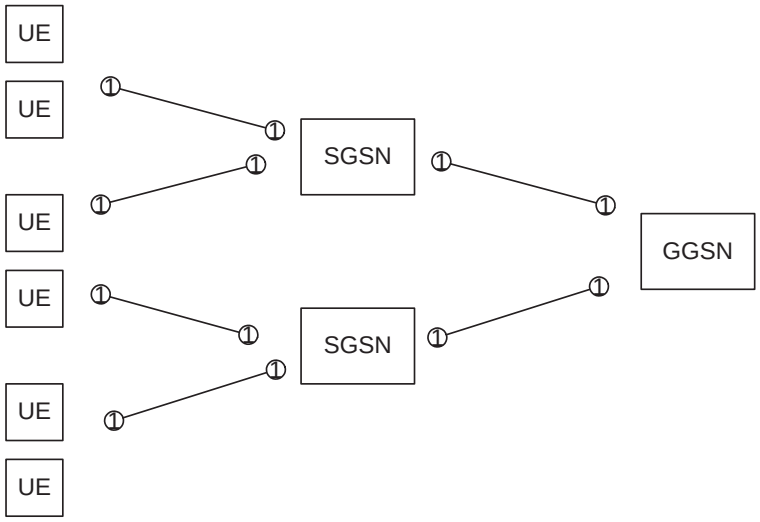


Fig. 7. Multicasting with MBMS.

The services that may be provided over MBMS are classified as follows [13]:

- Streaming: Continuous media flows, such as audio and video, plus supplementary text or images. These services are similar to TV channels, but enhanced with multimedia content.
- File download: Reliable binary data transfers, without any delay constraints. These services are similar to conventional file transfers, but with multiple receivers.
- Carousel: A combination of streaming and file download; static media are sent, but with synchronization constraints. These services are similar to stock quote ticker tapes.

All services can be offered in two modes: in broadcast mode all users that have activated the service receive all data transmitted; in multicast mode only those users that choose to join a multicast group receive the data transmitted to that group. In broadcast mode only the sender can be charged, not the receivers, therefore this mode is suitable for advertising supported services. In multicast mode both the sender and the receivers can be charged, therefore this mode is suitable for either pay per view or advertising supported services. It should also be noted that each service covers a specific area of the network; this allows a provider to transmit different content in each area.

## V. IMS ARCHITECTURE AND IMPLEMENTATION

### A. Service architecture

The relationship of the IMS to the PS and CS domains of a UMTS network is shown in Figure 8. From the figure is should be clear that the IMS is indeed a subsystem of the CN that depends on the PS domain. The actual data transport services are provided by the existing IP based mechanisms offered by GPRS, as enhanced for UMTS networks. What the IMS does is to provide flexible multimedia session management using these IP bearer services. For complete application functionality to be provided, the IMS may have to rely on services provided by an external AS, for example, a conferencing server. The IMS itself only provides session setup and control functions, media processing functions, and media and signaling interworking functions [14]. It also mandates however that all media should be transported using the *Real Time Protocol* (RTP) over UDP/IP, and, most importantly, that the IMS should use exclusively IPv6.

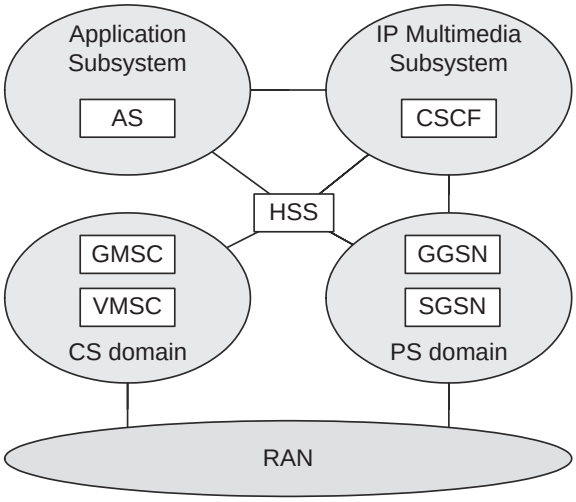


Fig. 8. IMS within a UMTS core network.

The general architecture of the IMS is outlined in Figure 9. Multimedia sessions are setup and controlled via various types of *Call Session Control Functions* (CSCF): the *Proxy CSCF* (P-CSCF) is the local contact point of the UE in the network it is visiting, analogous to the SGSN in GPRS; the *Serving CSCF* (S-CSCF) is controlling the session at the user's home network, analogous to the GGSN in GPRS. Since a network may contain many S-CSCFs for load balancing, an *Interrogating CSCF* (I-CSCF) may be provided at the entry point to an operator's network so as to direct sessions to the appropriate S-CSCF. The I-CSCF and the S-CSCF rely on the HSS for user related information. Networks with multiple HSSs also provide a *Subscription Locator Function* (SLF) that locates the HSS handling a given user.

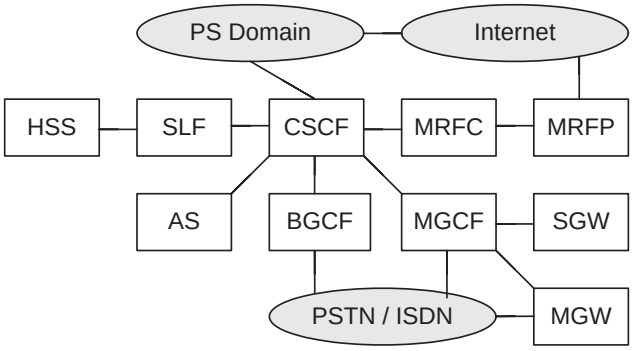


Fig. 9. IMS architecture.

While the IMS does not standardize any applications, it provides a *Media Resource Function Processor* (MRFP) that is able to mix, generate and process media streams under the control of a *Media Resource Function Controller* (MRFC). These entities can be used in conjunction with an appropriate AS to support applications such as voice conferencing, with the MRFP mixing voice streams, or announcement services, with the MRFP generating announcements. The MRFP can also provide transcoding to allow IMS applications to interoperate with other IP based applications employing different encoding schemes. By providing only this basic functionality inside the IMS, many types of applications can be supported by an AS, without

having to transfer the actual media streams to the AS. By separating the MRFP from the MRFC a single control function can oversee many processing functions to achieve scalability.

The IMS also provides *Media GateWay* (MGW) functions to allow IMS sessions to interwork with circuit switched networks, including the PSTN and the ISDN, and even the CS domain of UMTS. The MGW simply transcodes the data streams to and from the format used in the external network. It is controlled by a *Media Gateway Control Function* (MGCF) that also handles the signaling to and from the circuit switched network. For some types of circuit switched networks, the MGCF is supported by a separate *Signaling GateWay* (SGW). Finally, a *Breakout Gateway Control Function* (BGCF) determines where breakout should occur, i.e. where an outgoing session should exit the IMS and enter the circuit switched network.

### *B. Session setup and control*

The various CSCFs provide session setup and control for users accessing the IMS services. All signaling between the UE and the CSCFs, as well as between the CSCFs themselves, is based on the *Session Initiation Protocol* (SIP) [15], with some extensions specific to UMTS [16]. Since SIP is described in detail in a separate chapter, we will omit the SIP signaling details below, concentrating on the concepts. All SIP messages that are used to setup a session or modify its parameters contain the requirements of its media components using the *Session Description Protocol* (SDP) [17].

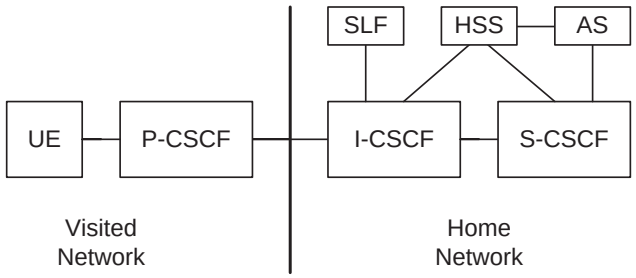


Fig. 10. The various call state control functions.

The CSCF may take on various roles, as shown in Figure 10. In order to understand these roles, we will explain how a UE registers with a SIP server so that it may start originating and terminating IMS sessions. The UE must first attach to the GPRS network and activate at least one PDP context for IPv6, to be used for signaling within the IMS, as explained in the previous section. Having established IP connectivity, the UE must then contact a SIP registry in its home network and register one or more SIP identities. These identities may be either telephone number style or Internet name style SIP identities. All IMS entities that are involved with SIP signaling also have SIP identities.

The UE discovers its P-CSCF either during the initial PDP context activation or by using some other mechanism such as querying the *Domain Name System* (DNS). The UE sends a SIP register message to the P-CSCF, including its SIP identity and the name of its home network. The P-CSCF maps the home network name to an I-CSCF address for that network and forwards the register message there. If the home network contains many HSSs, the SLF is used by the I-CSCF to find the HSS containing the user's information. The I-CSCF contacts this HSS to ask which S-CSCF should handle the user. The HSS checks the user's profile to see which applications the user has enabled, and replies with the address of an S-CSCF that has sufficient resources and capabilities to handle the user. The I-CSCF then forwards the registration to that S-CSCF. The S-CSCF queries the HSS for subscriber information, registers the user, informs the HSS that it is handling the user and returns a SIP confirmation via the I-CSCF and the P-CSCF to the UE. At this point, the user can originate and terminate SIP sessions.

An example of session setup between two UEs is shown in Figure 11 [18]. The caller UE sends a SIP invite message to its P-CSCF, including an SDP description of the media components of the session and the SIP identity of the called UE. The P-CSCF may accept or deny the request, depending on resource availability and the QoS policy of the visited network. If accepted, the invite message will be forwarded to the user's S-CSCF, possibly via the I-CSCF of the user's home network. The S-CSCF will then contact the HSS and accept or deny the request depending on the user's profile. If accepted, the invite message will be forwarded to the S-CSCF handling the called UE at its own home network, again possibly via an I-CSCF. The S-CSCF will contact the appropriate HSS to check whether the call can be accepted, and if so, it will forward the message to the called UE via the corresponding P-CSCF.

After the SIP invite is received by the called UE, its SDP payload will be examined, and the called UE will return a counter proposal in a SIP reply. This reply will follow the reverse path through the CSCFs to reach the caller UE, which will then respond with a final mutually acceptable SDP specification to the called UE. At this point the usual SIP signaling continues, always via the CSCFs, until the session is established. The actual data of the session can then be exchanged directly between the two UEs, without passing any more through the CSCFs. It should be noted that before the actual data exchange begins, each UE must activate secondary PDP contexts for each media component, using the same IP address as for the already established primary PDP context, but different QoS parameters, as appropriate for each media component of the session. This procedure will be described in the last section.

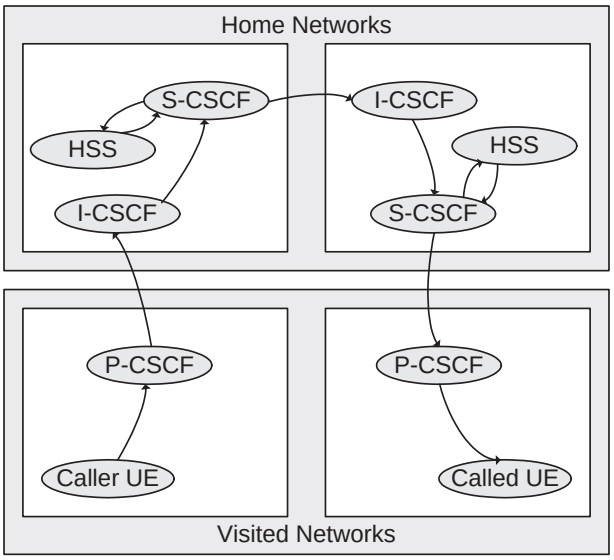


Fig. 11. Session setup example.

In order to allow an external AS to provide services, an AS can install a set of filters in the HSS handling a user. A *filter* describes an event that should cause the AS to be notified. For example, a filter could specify that the AS should be notified when the user sends or receives a SIP invite message with specific parameters in the SDP description. When a UE registers with the IMS, the S-CSCF handling the user receives from the HSS the filters associated with the user and stores them in its database. Then, when SIP messages from or to the user are received by the S-CSCF, they are matched against the filter database and the corresponding ASs are notified by SIP messages. For example, when a UE sends a SIP invite message indicating that it wants to participate in a video conference, the AS controlling the video conference may be notified so as to handle the user.

### *C. Interworking functionality*

Since the GGSN is essentially an IP router, basic connectivity between the IMS and external IP based networks is simple. One complication is that while UMTS networks support both *IP version 4* (IPv4) and IPv6, the IMS uses exclusively IPv6, and most other IP networks use exclusively IPv4. Therefore, IPv4 to IPv6 translation gateways are needed at the edge of an IMS network to convert between the different header formats and addresses. Note that this is an issue for all IPv6 networks, not just IMS.

Another complication is that while SIP is an Internet standard, it has been extended to better handle the requirements of the IMS. As a result, when SIP requests are received from or sent to external IP networks, the S-CSCF will discover that one side does not support the IMS specific extensions. Depending on operator policy, the S-CSCF may either refuse to setup sessions with non IMS conformant SIP endpoints, or accept them and translate between IMS and non IMS specific SIP semantics [19]. If media transcoding is also needed, it may be performed by the MRFP under the control of the MRFC.

Since voice telephony is so important in circuit switched networks, the IMS must interwork as seamlessly as possible with the PSTN, the ISDN and the CS domain of UMTS networks. As explained above, this is achieved by using a MGW to perform the required media translations under the control of a MGCF which also performs the required signaling translations, possibly aided by an SGW. Even though IMS applications can use any codec they desire, each UE and MGW must support at least some standardized codecs so as to provide a minimal guaranteed level of interoperation with other networks. For example, for sessions to or from the PSTN the MGW must convert the RTP packets containing *Adaptive Multi-Rate* (AMR) encoded voice to bit streams containing *Pulse Code Modulation* (PCM) encoded voice, and vice versa. The MGW may also have to generate in band signaling for the PSTN, such as dialing tones. Similarly, the SGW and MGCF must convert the SIP signaling of IMS to PSTN signaling, and vice versa [20].

For sessions originating from the UMTS, the S-CSCF handling the caller UE forwards the SIP invite to the BGCF of the home network. If the BGCF determines that the breakout is to occur locally, it selects the MGCF that will be responsible for the interworking and forwards the SIP invite there. The MGCF will then complete the signaling with the external network in order to setup and control the session. If the breakout is to occur in another network, the BGCF forwards the SIP invite to a BGCF in that network. For sessions originating from external circuit switched networks, the signaling is handled by the MGCF in the home network of the called UE, since the session setup request will arrive there. The MGCF acts then as a SIP proxy server in order to complete the session setup inside the UMTS network.

## VI. MBMS ARCHITECTURE AND IMPLEMENTATION

### A. Service architecture

While the MBMS affects both the CN and the RAN, the emphasis is on the conservation of resources over the air interface, i.e. the wireless link between the RAN and the UE. The MBMS provides various multicast and broadcast *services*, with each service offering some content and covering a possibly different set of cells. The actual transmission of data within a service is a *session*. A service can only have one active session at any given time, but it may use multiple sessions over its lifetime [12]. In order to support MBMS services and sessions, a new functional entity, the *Broadcast / Multicast Service Center* (BM-SC), is added to the PS domain of the CN, as shown in Figure 12. Changes must also be made to the GGSN, SGSN, RAN and UE to provide MBMS support [21].

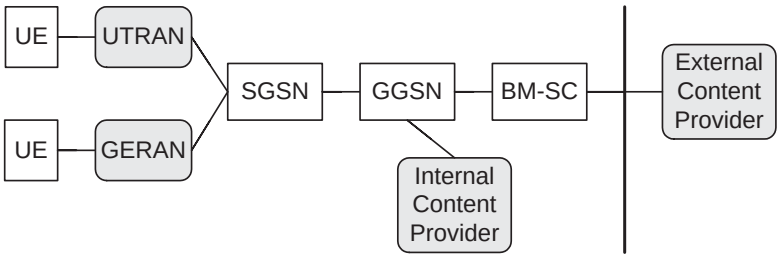


Fig. 12. MBMS architecture.

The BM-SC is the core MBMS component, handling all MBMS services and sessions. With MBMS data can only be transmitted towards the UEs, using IP multicast packets for transport. The content transmitted may originate from either external or internal content providers. External content providers must first be authenticated by the BM-SC and they must then transmit only via the BM-SC. The BM-SC may encode the data to provide additional error resilience. In order for a session to begin, the BM-SC instructs the GGSN to setup MBMS bearers in the required service area at the appropriate QoS level. The session data are always transmitted via the GGSN, which may receive them either directly or indirectly, via the BM-SC, from the content provider.

The data are delivered via the GGSN and SGSN to the RAN. The GGSN is responsible for creating appropriate MBMS bearers, based on instructions from the BM-SC, and routing data to the appropriate SGSNs, i.e. all of them in broadcast mode or only those serving interested UEs in multicast mode. The GGSN generates charging data for each receiver of a multicast session, while the BM-SC generates charging data for the content provider of both multicast and broadcast sessions. The SGSN is responsible for routing the data to the appropriate BSCs or RNCs, depending on the type of RAN used.

The main issue for MBMS multicast data delivery inside the RAN is whether a single point-to-multipoint or multiple point-to-point physical channels should be used over the air interface. Since point-to-multipoint channels are more expensive in terms of resources, each operator must set a threshold: when the number of recipient UEs in the cell exceeds the threshold, a point-to-multipoint channel is established; when the number of recipient UEs drops below the threshold, separate point-to-point channels are established. For MBMS broadcast data delivery, a broadcast physical channel is always used.

Finally, the UE must be modified so as to handle activation and deactivation of broadcast services and joining and leaving of multicast services. The former is a purely local operation, since broadcast data are always transmitted anyway. The latter requires signaling between the UE and the network. Two options are being considered for signaling, either creating an MBMS specific scheme or using the standardized group management protocols of the Internet, i.e. the *Internet Group Management Protocol* (IGMP) for IPv4 or the *Multicast Listener Discovery* (MLD) protocol for IPv6. An MBMS specific solution may be more efficient, but it will require a new protocol to be designed and multiple network elements to be modified. In the following subsections we assume that the IGMP/MLD option is used.

### B. Session setup and control

The MBMS services follow the phases shown in Figure 13. The broadcast service is simpler, as there is no interaction between the network and the users, so we will describe it first. In the *service announcement* phase the users are informed about the service availability in their area. The BM-SC can generate service announcements and distribute them over MBMS using the *Session Announcement Protocol* (SAP) [22]. Announcements can include descriptions of the media that will be included (using SDP), the scheduled time of transmission and other relevant details. The users may also discover the services available in their area by other means, such as web pages or CBS messages. The users may choose to either activate or deactivate reception for each broadcast service locally, i.e. in the UE.

When data are about to be delivered, the BM-SC controlling the service initiates the *session start* phase, during which the GGSN is instructed to setup MBMS bearers. In broadcast mode, this requires the creation of bearers towards all cells in the service area. In the *MBMS notification* phase the users are notified that a session for a specific service is about to begin

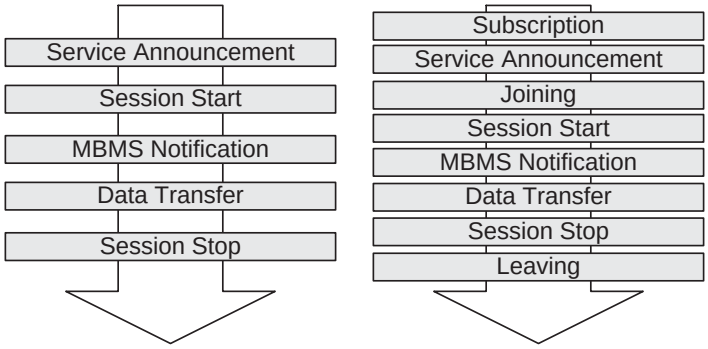


Fig. 13. Multicast and broadcast mode phases.

transmission. In the *data transfer* phase the actual data are transmitted to all cells in the service area and are received by all UEs that have not deactivated the service. Finally, in the *session stop* phase the MBMS bearers of the session are released.

The multicast mode adds three more phases, performed separately by each user that wants to participate in a multicast session, unlike the phases already described that are performed only once per session. In the *subscription* phase the interested users subscribe to specific multicast services via external means, for example, via a web page. This informs the BM-SC that the user has agreed to potentially receive a multicast service. The service announcement phase is the same as in broadcast mode, but the announcement must also include the multicast address that will be used by the service. Either before or after the session start phase, users interested in receiving data enter the *joining* phase, i.e. they send an IGMP/MLD join message to the GGSN indicating the multicast group corresponding to the desired service. These messages are forwarded to the BM-SC which verifies that the user has already subscribed to the service.

During the session start phase, the BM-SC instructs the GGSN to create the appropriate MBMS bearers. In multicast mode bearers are created only towards cells that already contain UEs that have joined the multicast group. During the MBMS notification phase the BTS or Node-B in each cell, depending on the type of RAN used, counts the number of UEs that are interested in this service and establishes either a single point-to-multipoint or multiple point-to-point physical channels to the UEs. In the data transfer phase the actual data are transmitted to all UEs that have joined the service, and in the session stop phase the MBMS bearers are released. Either before or after the session stop phase, users no longer interested in receiving data enter the *leaving* phase, i.e. they send an IGMP/MLD leave message to the GGSN, indicating the multicast group that they wish to leave.

We will now explain how multicast distribution trees are constructed. When a UE wants to join a multicast group corresponding to a service, it sends an IGMP/MLD message to its GGSN using a signaling PDP context. The GGSN checks with the BM-SC if the user has subscribed to the service. If so, the GGSN informs the UE, via the SGSN, that the join is approved. The UE then asks the SGSN to create an MBMS UE context, indicating the multicast service and the UE. In turn, the SGSN asks the GGSN to create a similar MBMS UE context. If these are the first MBMS UE contexts created for a multicast service, the SGSN and the GGSN also create an MBMS bearer context indicating the QoS level for the service, based on information from the BM-SC. The MBMS bearer context also lists the downstream nodes that should receive data. In the session start phase, when the BM-SC instructs the GGSN to create multicast MBMS bearers, by consulting the MBMS UE and bearer contexts the GGSN knows which SGSNs are interested and each SGSN knows which UEs are interested in this multicast service, so bearers are created only towards the appropriate nodes, using the proper QoS level. If some UEs later join or leave the group, the MBMS UE and bearer contexts are updated and the multicast distribution tree is modified accordingly.

### *C. Interworking functionality*

The MBMS is a UMTS specific service, therefore it cannot directly interwork with services provided by external networks. The exception is IP networks, since the MBMS multicast mode is based on IP multicasting. Even in this case however, MBMS multicast groups are controlled for charging purposes and transmissions are strictly unidirectional towards the UEs, therefore the MBMS and IP multicasting service models are quite different. In general, when the content to be delivered via MBMS originates outside the UMTS network, it must be authorized and transmitted by the BM-SC, therefore it must enter the UMTS network in unicast mode and then be replicated as needed.

If an external IP multicast group includes receivers inside the UMTS network, IP multicast routing will terminate at the GGSN; from there on multicast group management will take over, unlike in regular IP networks where multicast routing extends all the way to the local networks, i.e. the cells. However, even though regular multicast routing is not performed inside the UMTS network, the MBMS ensures that data are delivered in the most efficient manner to the UEs, i.e. not as separate unicasts, as was the case before MBMS.

## VII. QUALITY OF SERVICE

### A. Quality of Service architecture

The 3GPP has adopted the layered QoS architecture shown in Figure 14 for UMTS networks. The figure shows a user connected to a UMTS network communicating with a user connected to another network, which could be the PSTN, the ISDN or the Internet. The end-to-end *Quality of Service* (QoS) depends on the QoS offered by the bearer services supported along the path. The Release 6 specifications do not standardize the end-to-end service since it partly lies outside the UMTS network; they only deal with the UMTS bearer service.

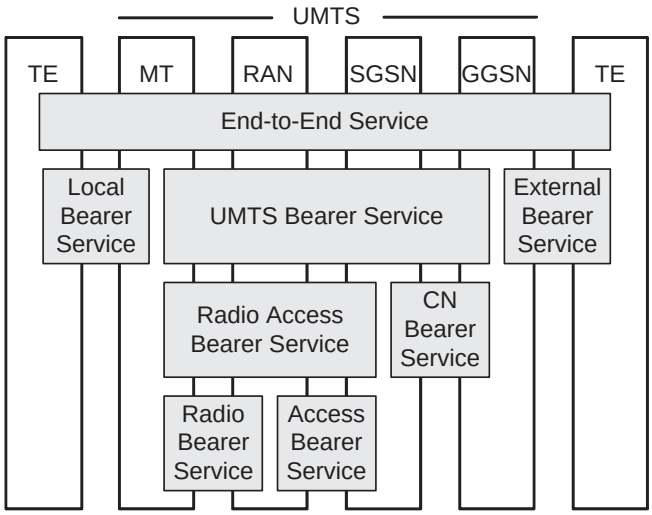


Fig. 14. The UMTS QoS architecture.

The *local bearer service* is offered over a local connection, or internally in UEs integrating both connectivity and terminal functionalities, i.e. both the MT and the TE; it deals with the signaling required to set up an end-to-end service. The *UMTS bearer service*, the actual service provided by the network, consists of two parts: the *radio access bearer service* provided by the RAN and the *core network bearer service* provided by the CN. These services are separated since a UMTS network may include different RANs, i.e. the UTRAN or the GERAN, and different CNs, i.e. the PS and CS domains and the IMS. The radio access bearer service in turn consists of a *radio bearer service* provided over the wireless link and an *access bearer service* provided inside the RAN. Finally, between the edge of the CN and the TE at the other end an *external bearer service* is provided by another network; this could be a PSTN voice call or a best-effort IP session. While these services are not part of the UMTS specifications, some guidelines are proposed to translate between them and UMTS services [23].

Internally, a UMTS network provides four different QoS classes [24]:

- The *conversational* class is suitable for real-time applications with stringent end-to-end delay and delay variation requirements, such as telephony.
- The *streaming* class is suitable for real-time applications with stringent delay variation requirements, such as media streaming.
- The *interactive* class is suitable for request / reply applications that have high reliability but moderate delay requirements, such as web browsing.
- The *background* class is suitable for bulk transfer applications that have high reliability requirements, such as file transfer.

While both the conversational and streaming classes are real-time, only the conversational class requires low end-to-end delay. Similarly, while both the interactive and background classes are non real-time, only the interactive class requires moderate end-to-end delay. Besides its QoS class, a bearer is characterized by a number of additional QoS parameters, summarized in Table I.

All service classes can specify the *maximum bit rate*, *maximum packet size* and *packet error ratio* for the bearer, all of which are self-explanatory. The *residual error ratio* refers to the bit errors that go undetected. The *delivery order* flag is set if the bearer should deliver packets in the order they were sent; this is important for protocols such as TCP. The *delivery of erroneous packets* flag is set if the endpoint wants to inspect these packets; this is useful for protocols performing error recovery. The *allocation / retention priority* indicates how important it is to establish / retain the bearer during resource shortages.

The real-time classes can also specify a *guaranteed bit rate* and a *transfer delay limit*; the latter should cover at least 95% of the packets. The *packet format information* indicates that the bearer will only carry specially formatted packets that can be optimally encoded by the network. The *source statistics descriptor* indicates if the bearer will transfer voice, for which statistical traffic models exist. Finally, the interactive class can specify a *traffic handling priority*, which is used to prioritize

| **Traffic class** | **Conv.** | **Str.** | **Int.** | **Back.** |
|---|---|---|---|---|
| Maximum bit rate | X | X | X | X |
| Maximum packet size | X | X | X | X |
| Packet error ratio | X | X | X | X |
| Residual bit error ratio | X | X | X | X |
| Delivery order | X | X | X | X |
| Delivery of erroneous packets | X | X | X | X |
| Allocation/Retention priority | X | X | X | X |
| Guaranteed bit rate | X | X | | |
| Transfer delay | X | X | | |
| Packet format information | X | X | | |
| Source statistics descriptor | X | X | | |
| Traffic handling priority | | | X | |
| Signaling indication | | | X | |

TABLE I

QOS PARAMETERS PER CLASS.

the interactive bearers, and a *signaling indication* flag that is set if the bearer will be used for signaling. Note that background bearers have lower priority than interactive bearers [23].

When IMS sessions are established, they may involve many media components, as specified by the SDP descriptions carried in the SIP messages. Therefore, in addition to the PDP context used for signaling, each UE must establish secondary PDP contexts with the same IP address but different QoS parameters for each media component. Each PDP context is associated with a filter that shows which IP packets should be handled by it; one PDP context may have no associated filter, serving as the default context. It should also be noted that MBMS sessions may only use the streaming or background QoS classes. This is justified by the fact that they are unidirectional and therefore unsuitable for the conversational and interactive classes. Another MBMS limitation is that the entire multicast or broadcast distribution tree must use the same QoS parameters, which are specified by the BM-SC during session start; these parameters cannot be changed during the session, thus simplifying tree management.

### *B. Policy based Quality of Service*

While the UMTS specifications describe the QoS classes and their parameters in detail, they do not indicate how the corresponding QoS must be provided. It is up to each implementation, i.e. to the manufacturer and operator of a UMTS network, to choose appropriate mechanisms and their parameters to provide the required QoS for the bearers. A clean way to separate the QoS policy from its implementation on various devices is to use a *policy-based* architecture. A UMTS network can optionally implement the policy-based architecture outlined in Figure 15; we will concentrate on this approach for the remainder of this section. In policy-based QoS the operator selects a set of policies to be enforced by the UMTS network, based on the services that should be provided to the users and on interworking requirements with external networks. These policies are translated into actual QoS mechanisms based on the devices used in the UMTS network [25].

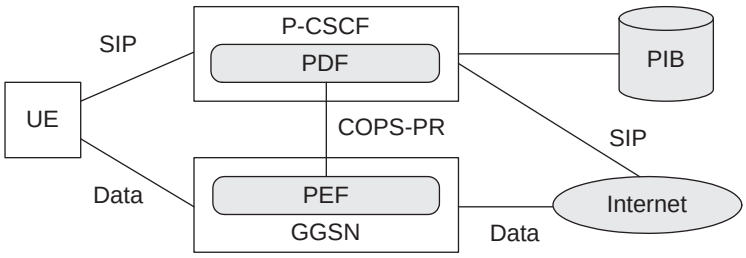


Fig. 15. Policy-based QoS architecture.

This approach splits QoS handling between two entities: the *Policy Decision Function* (PDF) and the *Policy Enforcement Function* (PEF). The PDF intercepts session setup messages specifying a particular QoS level, retrieves the operator's policy rules matching the request from a *Policy Information Base* (PIB) and decides whether the session can be accepted, based on current resource availability. If the session is accepted, the PDF translates the policy rules to appropriate configuration actions and sends them to the PEF, which actually implements the QoS provisioning mechanism. The PEF informs the PDF when the policy has been applied, since the PDF must always be aware of PEF resource availability in order to make decisions.

Even though the PDF and the PEF may be separate entities, it is more economical to combine them with existing UMTS entities [26]. The PDF must intercept session setup signaling, so it can be combined with the P-CSCF serving the UE. The

PEF on the other hand must intercept actual data transmissions, so it can be combined with the GGSN serving the UE. The PEF and the PDF communicate via the *COPS for Policy Provisioning* (COPS-PR) protocol [27], a variant of the *Common Open Policy Service* (COPS) protocol [28]. The COPS-PR messages are transmitted over TCP. The policy rules defined by the operator are stored in a UMTS specific PIB [29].

The model for policy-based QoS provisioning in UMTS assumes that the PEF implements a gate for each bearer service. The gate is basically a QoS specification and a filter that matches the corresponding IP packets based on various header fields. The gate can be opened or closed based on policy decisions made by the PDF. When a SIP session setup request is intercepted by the PDF, the PDF examines the SDP description of its media components and decides whether the session should be established. If so, the PDF sends binding information to the UE; this includes an authorization token uniquely identifying the authorized request and filters matching the approved media components.

When the UE receives the binding information, it can start creating PDP contexts for the various media components of the session. These requests, containing a QoS specification and the binding information returned by the PDF, are sent to the GGSN. The PEF at the GGSN intercepts these messages, extracts the authorization token from the binding information and queries the PDF about the authorized resources. The PDF returns the filters and QoS specifications authorized; if these equal or exceed those in the UE's request, the PDP context is activated, the appropriate filters and QoS specifications are installed by the PEF and the PDF is notified to modify its records. Finally, when the session is established, the PDF intercepting the corresponding SIP message instructs the PEF to open the gate, and the session data may start flowing through the GGSN [26].

### *C. Session setup and control*

To clarify how the policy-based scheme operates, we will present the procedures used for QoS provisioning using as an example a session between two UMTS UEs. There are eight procedures used for QoS control [30]:

- *Authorize QoS resources*. During session setup, when the called UE returns to the caller UE a SIP message containing a modified SDP description, the PDF at the P-CSCF determines if the session can be accepted. If yes, it includes appropriate binding information in the SIP messages sent to both UEs. There is no need to authorize QoS resources earlier, since the called UE may not accept the session.
- *Resource reservation*. When the UE attempts to activate PDP contexts for the media components of the session, it sends a request to the GGSN containing the binding information returned to it by the PDF. The PEF at the GGSN asks the PDF what resources have been authorized for that authorization token. If the resources authorized equal or exceed the requested resources, the PEF installs the packet filters and notifies the PDF that the resources have been reserved.
- *Approval of QoS commit*. When the SIP setup signaling concludes and the session is established, the PDF at the P-CSCF updates its database of available resources and instructs the PEF at the GGSN to open the gate and allow data to flow.
- *Removal of QoS commit*. This procedure reverses the approval procedure, i.e. closes the gate. The filters and reservations are not removed, therefore this procedure is useful when the session is temporarily suspended.
- *Revoke authorization for QoS resources*. This procedure reverses both the reservation and authorization procedures. It is invoked when an established session is normally released using SIP signaling.
- *Indication of PDP context release*. This is similar to the revoke authorization procedure, but it is used when a PDP context is released without previous SIP signaling, i.e. in an abnormal termination.
- *Authorization of PDP context modification*. This procedure is used when the UE wishes to modify the session by requesting additional resources.
- *Indication of PDP context modification*. This procedure is used when a PDP context modification indicates that there is no need for the corresponding resources any more.

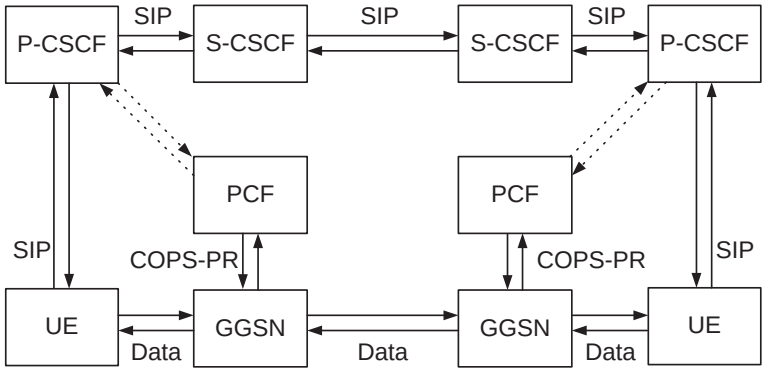


Fig. 16. QoS setup example.

A full session setup example is shown in Figure 16. We assume that both UEs have already registered with their S-CSCFs. When the caller UE sends a SIP invite message, this is routed to the caller UE via the intermediate CSCFs. When the called UE responds with an updated SDP description of the media components, the PDF at the P-CSCF examines the request and decides if it can be authorized. If so, the PDF sends binding information to the called UE and forwards the SIP message to the PDF at the P-CSCF of the caller UE, which repeats the above procedure; if it also approves the request, it also sends binding information to the caller UE. At this point the authorization of resources is complete, but resources have not been reserved yet.

The next step is for both UEs to independently activate appropriate PDP contexts at their GGSNs, using the binding information returned by the PDFs. The included authorization tokens are passed by each PEF at the GGSN to its PDF to determine the amount of resources that have been authorized; if these equal or supersede those requested, each PEF installs the corresponding filters, reports it to the PDF and informs the UE that the PDP contexts have been activated. At this point the reservation of resources is complete but the gates are closed, since the session has not been established.

Eventually the session is established, as indicated by a final SIP message sent by the called UE to the caller UE via the CSCFs. As this message passes through the PCF at each P-CSCF, the PCF instructs the PEF at the GGSN to open the gates, and the PEF confirms that the gates have indeed been opened. When the final SIP message reaches the caller UE, the gates are already open at both GGSNs, therefore session data may start flowing in both directions.

## VIII. Summary

This chapter provided an introduction to the support for IP based multimedia services on 3G wireless cellular networks, focusing on the IP Multimedia Subsystem (IMS) and the Multimedia Broadcast / Multicast Service (MBMS). An overview of cellular networks in general and UMTS networks in particular was first presented to lay the groundwork for the following discussion. Then the features and services of the IMS and the MBMS were introduced, followed by detailed descriptions of both. Finally, the QoS issues for IP based multimedia services were discussed, emphasizing the policy based QoS control scheme of UMTS and its application to the IMS.

## References


[1] 3GPP, “Vocabulary for 3GPP specifications”, TR 21.905, V6.5.0, January 2004.
[2] M. Zeng, A. Annamalai and V.K. Bhargava, “Harmonization of global third-generation mobile systems”, IEEE Communications Magazine, December 2000, pp. 94-104.
[3] A. Furuskär, S. Mazur, F. Müller and H. Olofsson, “EDGE: Enhanced Data Rates for GSM and TDMA/136 Evolution”, IEEE Personal Communications, June 1999, pp. 56-66.
[4] 3GPP, “Evolution of 3GPP system”, TR 21.902, V6.0.0, September 2003.
[5] 3GPP, “Services and service capabilities”, TS 22.105, V6.2.0, June 2003.
[6] 3GPP, “Network architecture”, TS 23.002, V6.3.0, December 2003.
[7] 3GPP “General Packet Radio Service (GPRS); Service description; Stage 2”, TS 23.060, V6.3.0, December 2003.
[8] 3GPP “Service requirements for the Internet Protocol (IP) multimedia core network subsystem; Stage 1”, TS 22.228, V6.5.0, January 2004.
[9] 3GPP “IP Multimedia Subsystem (IMS) group management; Stage 1”, TS 22.250, V6.0.0, December 2002.
[10] 3GPP, “Technical realization of Cell Broadcast Service (CBS)”, TS 23.041, V6.2.0, December 2003.
[11] M. Hauge and Ø. Kure, “Multicast in 3G networks: Employment of existing IP multicast protocols in UMTS”, ACM WoWMoM, September 2002, pp. 96-103.
[12] 3GPP “Multimedia Broadcast / Multicast Service (MBMS); Stage 1”, TS 22.146, V6.3.0, January 2004.
[13] 3GPP “Multimedia Broadcast / Multicast Service (MBMS) user services; Stage 1”, TS 22.246, V6.0.0, January 2004.
[14] 3GPP “IP Multimedia Subsystem (IMS); Stage 2”, TS 23.228, V6.4.1, January 2004.
[15] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley and E. Schooler, “SIP: Session Initiation Protocol”, June 2002, RFC 3261.
[16] M. Garcia-Martin, E. Henrikson and D. Mills, “Private Header (P-Header) extensions to the Session Initiation Protocol (SIP) for the 3rd-Generation Partnership Project (3GPP)”, January 2003, RFC 3455.
[17] M. Handley and V. Jacobson. “SDP: Session Description Protocol”, April 1998, RFC 2327.
[18] K.D. Wong and V.K. Varma, “Supporting real-time IP multimedia services in UMTS”, IEEE Communications Magazine, November 2003, pp. 148-155.
[19] 3GPP “Interworking between the IM CN subsystem and IP networks”, TS 29.162, V1.0.0, March 2002.
[20] 3GPP “Interworking between the IP Multimedia (IM) Core Network (CN) subsystem and Circuit Switched (CS) networks”, TS 29.163, V6.1.0, December 2003.
[21] 3GPP “Multimedia Broadcast / Multicast Service (MBMS) user services; Architecture and functional description”, TS 23.246, V6.1.0, December 2003.
[22] M. Handley, C. Perkins and E. Whelan, “Session Announcement Protocol”, October 2000, RFC 2974.
[23] 3GPP “Quality of Service (QoS) concept and architecture”, TS 23.107, V6.0.0, December 2003.
[24] R. Koodli and M. Puuskari, “Supporting packet-data QoS in next-generation cellular networks”, IEEE Communications Magazine, February 2001, pp. 180-188.
[25] W. Zhuang, Y.S. Gan, K.J. Loh and K.C. Chua, “Policy-based QoS architecture in the IP multimedia subsystem of UMTS”, IEEE Network, May/June 2003, pp. 51-57.
[26] 3GPP “End-to-end Quality of Service (QoS) concept and architecture”, TS 23.207, V5.3.0, March 2002.
[27] K. Chan, J. Seligson, D. Durham, S. Gai, K . McCloghrie, S. Herzog, F. Reichmeyer, R. Yavatkar and A. Smith, “COPS usage for Policy Provisioning (COPS-PR)”, RFC 3084, March 2001.
[28] D. Durham, J. Boyle, R. Cohen, S. Herzog, R. Rajan and A. Sastry, “The COPS (Common Open Policy Service) Protocol”, RFC 2748, January 2000.
[29] 3GPP “Policy control over Go interface”, TS 29.207, V5.7.0, March 2004.
[30] 3GPP “End-to-end Quality of Service (QoS) signaling flows”, TS 29.208, V5.7.0, March 2004.

| | |
|---|---|
| 1G/2G/3G | First/Second/Third Generation |
| 3GPP | 3rd Generation Partnership Project |
| 3GPP2 | 3rd Generation Partnership Project 2 |
| AAL2/5 | ATM Adaptation Layer 2/5 |
| AMPS | Advanced Mobile Phone Service |
| AMR | Adaptive Multi-Rate |
| AS | Application Server |
| ATM | Asynchronous Transfer Mode |
| BGCF | Breakout Gateway Control Function |
| BM-SC | Broadcast / Multicast Service Center |
| BSC | Base Station Controller |
| BSS | Base Station Subsystem |
| BTS | Base Transceiver Station |
| CBC | Cell Broadcast Center |
| CBS | Cell Broadcast Service |
| CDMA | Code Division Multiple Access |
| CN | Core Network |
| COPS | Common Open Policy Service |
| COPS-PR | COPS for Policy Provisioning |
| CS | Circuit Switched |
| D-AMPS | Digital Advanced Mobile Phone Service |
| DNS | Domain Name System |
| EDGE | Enhanced Data Rates for GSM Evolution |
| ETSI | European Telecommunications Standards Institute |
| FDD | Frequency Division Duplexing |
| FDMA | Frequency Division Multiple Access |
| GERAN | GSM EDGE Radio Access Network |
| GGSN | Gateway GPRS Support Node |
| GMSC | Gateway Mobile services Switching Center |
| GPRS | General Packet Radio Service |
| GSM | Global System for Mobile Communications |
| GSN | GPRS Support Node |
| HSCSD | High Speed Circuit Switched Data |
| HSS | Home Subscriber Server |
| I-CSCF | Interrogating Call State Control Function |
| IGMP | Internet Group Management Protocol |
| IMS | IP Multimedia Subsystem |
| IMT-2000 | International Mobile Telecommunications 2000 |
| IPv4/6 | IP version 4/6 |
| ISDN | Integrated Services Digital Network |
| ITU | International Telecommunications Union |
| IWF | InterWorking Function |

TABLE II

GLOSSARY OF ACRONYMS (A TO L).

| | |
|---|---|
| MBMS | Multimedia Broadcast / Multicast Service |
| MGCF | Media Gateway Control Function |
| MGW | Media GateWay |
| MLD | Multicast Listener Discovery |
| MRFC | Multimedia Resource Function Controller |
| MRFP | Multimedia Resource Function Processor |
| MS | Mobile Station |
| MSC | Mobile services Switching Center |
| MT | Mobile Terminal |
| P-CSCF | Proxy Call Session Control Function |
| PCM | Pulse Code Modulation |
| PDF | Policy Control Function |
| PDP | Packet Data Protocol |
| PEF | Policy Enforcement Function |
| PIB | Policy Information Base |
| PS | Packet Switched |
| PSTN | Public Switched Telephone Network |
| QoS | Quality of Service |
| RAN | Radio Access Network |
| RNC | Radio Network Controller |
| RNS | Radio Network Subsystem |
| RTP | Real Time Protocol |
| SIP | Session Initiation Protocol |
| S-CSCF | Serving Call State Control Function |
| SGSN | Serving GPRS Support Node |
| SGW | Signaling GateWay |
| SLF | Subscription Locator Function |
| TDD | Time Division Duplexing |
| TDMA | Time Division Multiple Access |
| TE | Terminal Equipment |
| UE | User Equipment |
| UMTS | Universal Mobile Telecommunications System |
| UTRAN | Universal Terrestrial Radio Access Network |
| VLR | Visitor Location Register |
| VMSC | Visitor Mobile services Switching Center |
| W-CDMA | Wideband Code Division Multiple Access |

TABLE III

GLOSSARY OF ACRONYMS (M TO Z).