1

Introduction to LTE/SAE

1.1 Role of 3GPP

The 3rd Generation Partnership Project (in short 3GPP) is a joint international standardization initiative between North American (Alliance for Telecommunications Industry Solutions (ATIS)), European (European Telecommunications Standards Institute (ETSI)), and Asian organizations (Association of Radio Industries and Businesses (ARIB) and Telecommunication Technology Committee (TTC) in Japan, Telecommunications Technology Association (TTA) in Korea and China Communication Standards Association (CCSA) in China) that was originally established in December 1998. The participating organizations are also called organizational partners. Scope of 3GPP was to specify a new worldwide mobile radio system (the Global System for Mobile Communications (GSM) was a European initiative while Code Division Multiple Access (CDMA) was initiated in North America, both are not compatible with each other) based on the evolved GSM techniques General Packet Radio Service (GPRS)/EDGE. This activity has led to the standardization of the third-generation Universal Mobile Telecommunications System (UMTS), which consists of Wideband Code Division Multiple Access (WCDMA) as radio technology and a core network supporting both circuit-based voice calls and packet-based data services. UMTS was meant as a universal standard that allows subscribers to use their UMTS-capable mobile phones and subscriptions worldwide through roaming (for an explanation of the term "roaming," see Section 1.13.1) agreements between mobile operators. UMTS is a big success story with around 1.4 billion WCDMA subscriptions deployed until now.

But 3GPP did not stop work after UMTS, in the following years enhancements of UMTS like High-Speed Packet Access (HSPA)/HSPA+, new services such as Multicast/Broadcast delivery, Location services, and the IP Multimedia Subsystem (IMS) were introduced. Long-Term Evolution (LTE) with a new Orthogonal Frequency-Division Multiplexing (OFDM)-based radio technology and an All-IP core network architecture is the newest development of 3GPP.

3GPP is organized in different working groups (see Figure 1.1) that are responsible for different parts of the 3GPP system. The Radio Access Network (RAN) groups define the radio parts of the UMTS/LTE system, i.e. the physical layer, and radio protocols. The GSM/EDGE Radio Access Networks (GERAN) groups work specifically on the maintenance and development of GSM/EDGE access technologies. The System Architecture (SA) and Core/Terminal

LTE for Public Safety, First Edition. Rainer Liebhart, Devaki Chandramouli, Curt Wong and Jürgen Merkel.
© 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.

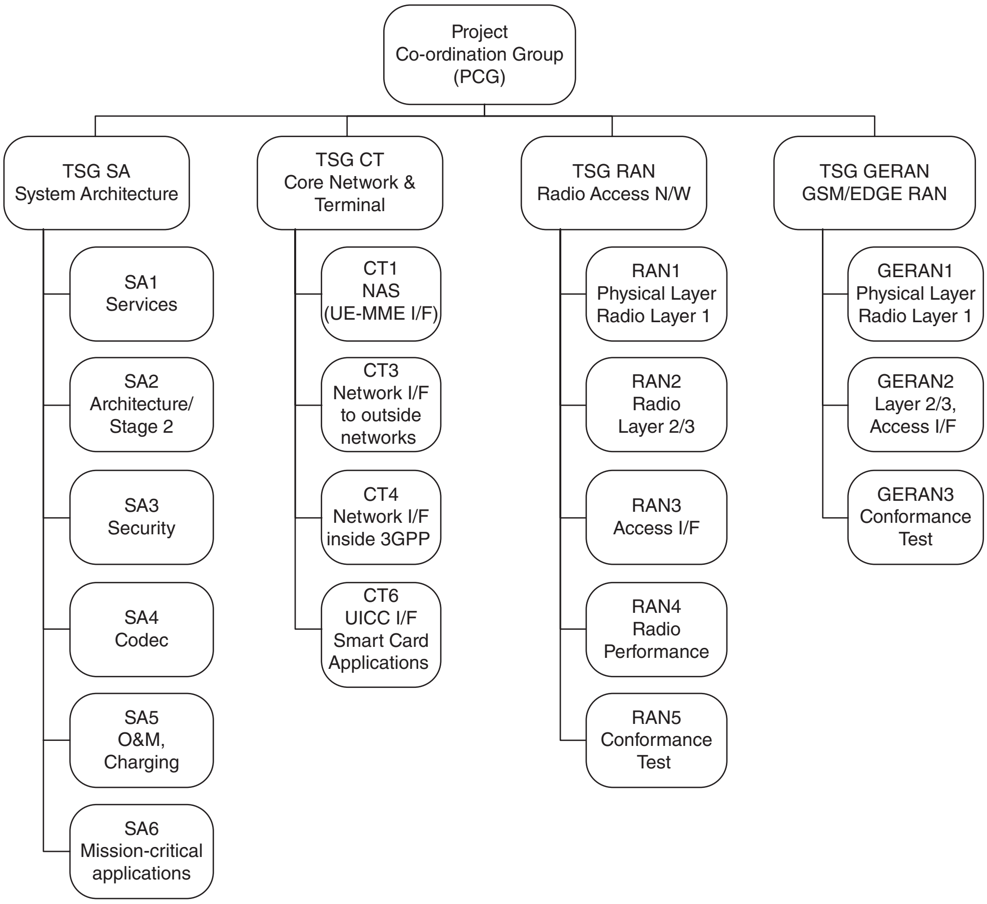


Figure 1.1 3GPP organizational structure

(CT) groups specify all parts of the overall system (e.g., architecture, security, charging) and all non-radio protocols (between the mobile device and network, within the network and between networks). A new working group SA6 will be operational from January 2015 onwards to standardize a Mission Critical Push To Talk (MCPTT) application in 3GPP. For details regarding MCPTT, please refer to Chapter 5.

3GPP follows a phased approach; working output is delivered as a set of Technical Specifications (TS) in so-called System Releases. Technical Specifications contain normative requirements that have to be implemented by chipset, device, and network equipment vendors. Interim results of ongoing work in 3GPP are usually captured in non-normative Technical Reports (TR). Test specifications are also created by 3GPP (mainly test cases for User Equipment (UE) to network communication). It has to be noted that 3GPP defines only functions and protocols, how these functions are implemented in concrete network nodes or whether some functions are implemented in the same node is up to the network vendor. One basic design principle in 3GPP's standardization process is backward compatibility of new features with existing ones. This ensures that new features can be introduced in one network without the need to upgrade all interconnected networks or all other nodes within this network at the same time.

Table 1.1 3GPP milestones up to Release 12

Release	Date	Main content
Phases 1 and 2	1992 and 1995	Basic GSM functions
Release 96, 97, 98, 99	1996, 1997, 1998, 1999	GPRS, HSCSD, EDGE, UMTS
Release 4	2001	MSC server split architecture
Release 5	2002	HSDPA, IMS
Release 6	2004	HSUPA, MBMS, Push to Talk over Cellular (PoC)
Release 7	2007	HSPA, EDGE evolution
Release 8	2008	LTE/SAE
Release 9	2009	LTE/SAE enhancements, Public Warning System (PWS), IMS emergency sessions over LTE/HSPA
Release 10	2011	LTE Advanced Local IP Access (LIPA) Selective IP Traffic Offload (SIPTO)
Release 11	2012	Heterogeneous Network (HetNet) Support Coordinated Multipoint Operation (CoMP)
Release 12	2014	Public Safety Machine type communication HSPA/LTE carrier aggregation

Table 1.1 provides a brief overview of the official release dates and milestones of the 3GPP releases up to Release 12.

For more information on the history and structure of 3GPP, visit the official 3GPP site at http://www.3gpp.org/about-3gpp/about-3gpp.

1.2 History of LTE

Main drivers for evolution of mobile networks are usually higher bandwidth on the air interface and better spectral efficiency (i.e., the information rate transmitted over a given bandwidth). After improving these key factors for WCDMA over several years, which led to the specification of HSPA and its evolution HSPA+, the 3GPP standardization forum, in 2004, started evaluating a new radio technology as successor for WCDMA. Objectives for starting this work were higher peak data rates (>100 Mbit/s in Downlink and >50 Mbit/s in Uplink) and lower latency besides other improvements. This work is formed under the name LTE. As the search for a more appealing name/acronym had no result, LTE is now used as the radio interface name in most official publications. Inside 3GPP the newly developed radio access network is called Evolved UMTS Radio Access Network (E-UTRAN) to indicate the evolution path from GERAN (GSM/GPRS/EDGE) second-generation networks (2G) to UTRAN (WCDMA/HSPA) third-generation networks (3G), and finally to E-UTRAN (LTE) fourth-generation networks (4G). It has to be noted that LTE initially did not fulfill International Telecommunication Union's (ITU) International Mobile Telecommunication (IMT) Advanced requirements, namely, the 1 Gbit/s peak data rate, which are officially the criteria for a network technology to be called 4G. Only LTE Advanced (LTE-A) will be able to support such data rates. So, strictly speaking calling LTE 4G was not correct but due to marketing battles in the United States, 4G was anchored as synonym for LTE and the ITU took the decision to allow any technology that provides an evolution path toward IMT Advanced to be called 4G. In parallel to the work on a new radio interface, 3GPP initiated a study to evolve the 2G/3G packet core network to which GERAN and UTRAN are connected (known as GPRS core) in order to cope with the new demands of LTE. This core network study was called System Architecture Evolution (SAE) and it was documented in the Technical Report TR 23.882 [1]. The final outcome of this work was a new packet core design in 3GPP's Release 8 documented in Technical Specifications TS 23.401 [2] and TS 23.402 [3], called the Evolved Packet Core (EPC); GPRS-specific parts are documented in TS 23.060 [4]. 3GPP Release 8 was officially completed in March 2009. EPC allows connecting LTE, GERAN/UTRAN, non-3GPP access systems such as Wireless Local Area Network (WLAN), WiMAX, and CDMA and also 3GPP-compliant small Femto Access Points (GERAN/UMTS/LTE radio stations connected via consumer links such as DSL or TV cable to the EPC) installed at homes, offices, and smaller campus areas. Special emphasis was put on optimized handover procedures between LTE and CDMA2000 eHRPD (Evolved High-Rate Packet Data) access due to requirements from CDMA network operators in the United States and Japan, which introduced LTE very early (starting in 2010). The EPC together with the connected radio access systems and the UEs is called Evolved Packet System (EPS) – the term LTE/SAE is also being used in this context. In contrast to the 2G and 3G systems, EPS no longer contains a Circuit-Switched (CS) part offering classical telephony network connectivity (e.g., EPS lacks optimized and dedicated radio bearers for CS voice calls) but only contains the Packet-Switched (PS) part providing data connectivity (for a description of CS and PS, please refer to Section 1.13.2). For that reason, supporting voice in EPS and providing a smooth migration story from 2G/3G voice and SMS (Short Message Service) to voice and SMS in EPS is extremely important for the acceptance of the new system. A lot of effort was spent in specifying solutions for the above-mentioned issues during Releases 8 and 9 (and continuing in later releases). This resulted in features called CS Fallback, SMS over SGs, SMS in MME (Mobility Management Entity), IMS Centralized Services, Session Continuity, and Single Radio Voice Call Continuity (SRVCC) among others.

As already mentioned EPC is an evolution of the 3GPP system architecture that finally realized the vision of an all-IP network. EPC in conjunction with the IMS delivers various services such as VoIP (Voice over Internet Protocol), Short Message Service (SMS), Video call, Picture share, Instant Messaging, and Presence. EPC and IMS support mobility with the existing 2G/3G wireless networks as well as fixed networks to facilitate smooth migration, interworking, and service continuity across all these networks. Nevertheless, as a pure IP-based network the main application for LTE/SAE will be the "Internet" with its rapidly increasing demands for more bandwidth and lower latency coming from P2P services or applications such as online gaming, Mobile TV, and Machine-to-Machine network deployments like smart traffic control or smart grid.

LTE/SAE was designed to cope with the challenges of the growing broadband mobile market: more data per user, "always-on" with high expectation on quality and reliability, connected everywhere, network efficiency, more devices that are not operated by humans, and the need to connect with an all-IP world.

As of today more than 300 operators in more than 100 countries have commercially launched LTE services. Nearly 160 million LTE subscriptions were issued worldwide, while nearly 1900

LTE capable devices were launched in the meantime. Without doubts LTE is the de facto standard in mobile broadband communication around the globe that will deliver broadband multimedia services to hundreds of million subscribers in the near future. These services, whether provided in the Internet or mobile networks, will be accessible with standardized devices (Smartphones, tablets, laptops) in nearly all countries of the world with an end user quality comparable to fixed broadband networks.

1.3 Drivers for LTE

As outlined in the previous section, main drivers to start work on a new radio technology and core architecture such as LTE/SAE are the need for higher data rates and a significant reduction of control plane latency and round trip delay to support future broadband, high quality services. After work on LTE started, 3G standards also progressed further, currently providing peak data rates well above 300 Mbit/s. Higher data rates are a must when recognizing the tremendous increase of mobile data traffic: forecasts indicated that mobile data traffic will increase 18-fold between 2011 and 2016 (an increase three times faster than fixed IP traffic) and will account for around 60% of the total IP traffic by 2016. As the usage of Internet services such as Email, browsing, chatting, or community applications increases dramatically in mobile networks, the limitations of 2G/3G radio and packet core networks has become apparent nowadays. Reduced control plane latency is a need to provide a high-quality (minimum delays when connecting to the network or during handover) “always-on” experience to the end user. Latency of control plane messages and big round trip delays were seen as drawbacks of existing legacy 2G/3G systems by many operators. Another driver for a flat, pure IP-based and simplified architecture with less radio and core network nodes was certainly the desire to decrease overall costs (OPEX and CAPEX). Finally, a purely IP-based architecture provides the possibility for introducing a PS optimized system while, for example, the 3G system had to make some compromises in support of packet-based services as 3G radio bearers have to be optimized for CS voice calls.

Another important design principle of EPC was backward compatibility and the capabilities to connect to non-3GPP access systems such as WLAN, WiMAX, or CDMA2000 systems. As such EPC and in general EPS are providing inherent mechanisms to support mobility for devices when changing radio access between 2G, 3G, and LTE, either based on GPRS Tunneling Protocol (GTP) or Proxy Mobile IP (PMIP) mobility protocols. For non-3GPP access systems such as WLAN or WiMAX, LTE/SAE is supporting mobility by reuse of generic mobility protocols defined in IETF, namely, PMIPv6 as specified in IETF RFC 5213 [5] and DSMIPv6 (Dual-Stack Mobile IP) as specified in IETF RFC 5555 [6]. For CDMA2000 eHRPD handover to LTE was improved by introducing special control plane and user plane interfaces to the EPC in order to deliver information from one access system to the other before the actual handover takes place to speed up the overall handover process. This harmonized core network architecture supporting 2G/3G, LTE, and non-3GPP access systems was another important objective when designing EPC.

While cost reduction is the main driver for network operators to introduce LTE/SAE, delay optimizations (minimized latency and round-trip delay leads to high-TCP traffic throughput, low-UDP/RTP traffic jitter to high-quality real-time services) and fast service availability caused by low bearer setup times are the benefits for the end user. These benefits increase the acceptance of the new technology and will pay back the investments of operators, terminal, and infrastructure vendors.

LTE/SAE is a broadband, standardized, wireless and packet-based system and constantly evolving to meet the needs of industry demands through 3GPP standardization activities. This allows public safety personnel to take advantage of an advanced wireless packet system and the vast support of innovative new applications for real-time information sharing and collaboration during emergencies and day-to-day operations. It improves situational awareness and enhances the safety of first responders and the public in general. LTE/SAE-based public safety networks will allow us to use any kind of multimedia services, voice, video, text, picture/file sharing, location-based services, in any situation (mobile and stationary) with high quality and in a reliable/resilient way. Built-in features such as network sharing and broadcast message delivery can be easily reused in public safety network deployments. Adopting IMS to provide mission critical push to talk and any kind of multimedia services over LTE is a further step toward a fully standardized, interoperable LTE public safety network that allows easy interconnection with other mobile and fixed networks, providing economy of scale on a high level.

1.4 EPS compared to GPRS and UMTS

When comparing E-UTRAN with UTRAN or GERAN the obvious change (besides use of the radio technology OFDM) is that E-UTRAN knows only one network element, the so-called Evolved NodeB (eNodeB or eNB) while in UTRAN NodeB and Radio Network Controller (RNC) and in GERAN Base Transmitter Station (BTS) and Base Station Controller (BSC) exist. Main reasons for this simplified E-UTRAN architecture were reducing complexity, latency, and costs while increasing data throughput.

On the core network side the obvious difference between the GPRS core (i.e., the 2G/3G core network with Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN) functions) and EPC is the strict and built-in separation of control and user plane in EPC. While this is possible in GPRS as well by using the so-called Direct Tunnel feature, that is, by which the user plane traffic is directly tunneled from RNC to GGSN bypassing the SGSN, the EPC was designed in this way from the very beginning. The main reason for this separation of user data and control (signaling) traffic was the ability to dimension the infrastructure for user traffic differently from the infrastructure parts handling the control of the user traffic, making it easy to adapt to growing user traffic needs. There are two functional elements in the user plane, the Serving Gateway (S-GW) and Packet Data Network Gateway (P-GW or PDN-GW), and one additional element in the control plane, the MME. As a consequence, in minimum two (eNodeB and combined S-GW/P-GW) and in maximum three nodes (eNodeB, S-GW, and P-GW) are in the EPS user plane path when LTE/SAE is deployed (1–2 less than in 2G/3G). This comes with a higher degree of simplicity, higher throughput, and less latency. Besides support of legacy 2G/3G access systems and LTE, the EPC also supports non-3GPP access systems that can be trusted or non-trusted from the EPC point of view. Support means that EPC provides means to authenticate, authorize, and charge subscribers using these non-3GPP access systems; the user plane is securely routed to and through the EPC toward a Packet Data Network (PDN) like the Internet. Last but not the least, mobility between 3GPP and non-3GPP access systems is enabled with the mobility anchor located in the EPC (P-GW). This kind of enhanced support of interworking with non-3GPP access systems is not supported by classical 2G GPRS.

In 2G/3G, the UE can be attached to the network without having any Packet Data Protocol (PDP) context established (i.e., no IP address is assigned to the UE). This concept was changed with the introduction of LTE to provide "always-on" connectivity to the UE anytime it is registered with the network. In LTE, a default bearer is established when the UE performs initial attach (explained later in this chapter) and when the last bearer is deactivated, the UE is detached. Thus, by default the UE is assumed to have at least one bearer context with which it can send and receive data when it is attached to an LTE network.

One technical detail that is not obvious is the concept of network-initiated bearer establishment. It was introduced to 2G/3G rather late and it was more an exception than the usual procedure. Until then, a bearer/context establishment request was always initiated by the UE. However, in LTE/SAE network-initiated bearer establishment is the main mechanism to establish dedicated bearers. Nevertheless, a UE-initiated bearer resource request procedure still exists in LTE/SAE but is considered as an exception.

To reduce the number of signaling messages between UE and core network the concept of a Tracking Area (TA) list allocated to the UE was introduced. Each TA of the list consists of one or more cells. Different TA lists allocated to different UEs in one area reduce the probability of simultaneous TA updates when a huge number of UEs are moving from one TA to another at the same time (e.g., when the users are riding the same train). Furthermore, a TA list allocated to one UE naturally decreases the need of the UE to perform TA updates but it comes with the drawback of bigger paging areas (the concepts of TA updates and paging are described later in this chapter).

1.5 Spectrum Considerations

To make LTE a worldwide success story, the technology must be flexible, adopt, and adapt to spectrum requirements in different countries around the globe. For that purpose, LTE was designed such that it can be deployed in a large variety of frequency bands, bands that might be allocated for a mobile broadband system by the national authorities in a specific country. Table 1.2 is taken from 3GPP TS 36.104 [7] and shows the supported Frequency Division Duplex (FDD) and Time Division Duplex (TDD) frequency bands for LTE.

More frequencies might be adopted for LTE once a commercial need arises in a specific country. It has to be noted that some bands are currently occupied by other technologies but LTE can coexist with these.

The process to allocate LTE spectrum for Public Safety deployments is ongoing. As it has not only technological but also economic and commercial consequences to allocate a certain portion of spectrum to a specific technology like LTE, spectrum-related discussions are ongoing in different regions of the world. However, some countries already took decisions which spectrum to allocate for Public Safety networks based on LTE. In the United States, the middle class tax relief and job creation act of 2012 reallocated the 700 MHz D-Block spectrum to public safety and included additionally $7 billion in federal funding for a nationwide LTE network for first responders.

The Australian Public-Safety Communications Officials (APCO) public safety bulletin No. 16 advised to develop a national interoperable public safety mobile broadband network based on LTE technology in the 4.9 GHz band. The Australian Communications and Media Authority

Table 1.2 LTE FDD/TDD frequency bands

Band	Uplink (UL) frequency band BS receive, UE transmit (MHz)		Downlink (DL) frequency band BS transmit, UE receive (MHz)		Duplex mode
1	1920	1980	2110	2170	FDD
2	1850	1910	1930	1990	FDD
3	1710	1785	1805	1880	FDD
4	1710	1755	2110	2155	FDD
5	824	849	869	894	FDD
6	830	840	875	885	FDD
7	2500	2570	2620	2690	FDD
8	880	915	925	960	FDD
9	1749.9	1784.9	1844.9	1879.9	FDD
10	1710	1770	2110	2170	FDD
11	1427.9	1447.9	1475.9	1495.9	FDD
12	699	716	729	746	FDD
13	777	787	746	756	FDD
14	788	798	758	768	FDD
15	Reserved		Reserved		FDD
16	Reserved		Reserved		FDD
17	704	716	734	746	FDD
18	815	830	860	875	FDD
19	830	845	875	890	FDD
20	832	862	791	821	
21	1447.9	1462.9	1495.9	1510.9	FDD
22	3410	3490	3510	3590	FDD
23	2000	2020	2180	2200	FDD
24	1626.5	1660.5	1525	1559	FDD
25	1850	1915	1930	1995	FDD
26	814	849	859	894	FDD
27	807	824	852	869	FDD
28	703	748	758	803	FDD
29	N/A	N/A	717	728	FDD[2]
30	2305	2315	2350	2360	FDD
31	452.5	457.5	462.5	467.5	FDD
…					
33	1900	1920	1900	1920	TDD
34	2010	2025	2010	2025	TDD
35	1850	1910	1850	1910	TDD
36	1930	1990	1930	1990	TDD
37	1910	1930	1910	1930	TDD
38	2570	2620	2570	2620	TDD
39	1880	1920	1880	1920	TDD
40	2300	2400	2300	2400	TDD
41	2496	2690	2496	2690	TDD
42	3400	3600	3400	3600	TDD
43	3600	3800	3600	3800	TDD
44	703	803	703	803	TDD

Note 1: Band 6 is not applicable.

Note 2: Restricted to E-UTRA operation when carrier aggregation is configured.

(ACMA) allocated an additional 60 MHz of spectrum across a number of bands (e.g., 800 MHz spectrum) to facilitate the deployment of high-speed, nationally interoperable mobile broadband networks for use by Australia's public safety authorities.

The European community is planning to recommend spectrum for public safety use in the 400 or 700 MHz bands. European spectrum community officials stated that the World Radio Conference (WRC) to be held in November 2015 is the best opportunity to get a dedicated spectrum allocation for Public Safety LTE deployments.

In Europe, many Asia-Pacific countries and parts of South America, the 400 MHz band is currently used by public safety agencies for their Terrestrial Trunked Radio (TETRA) and TETRAPOL systems. Using the same frequency band for LTE-based public safety networks around the world will allow minimizing investment costs by reuse of existing sites and assets.

In summary, Map 1.1 shows example bands that may be used for Public Safety. These bands are listed in ITU-R Res 646 and are recommended for Public Safety use. Which bands will be actually used in which region depends on national regulation at the end.

1.6 Network Architecture

1.6.1 Radio Access Network and Core Network

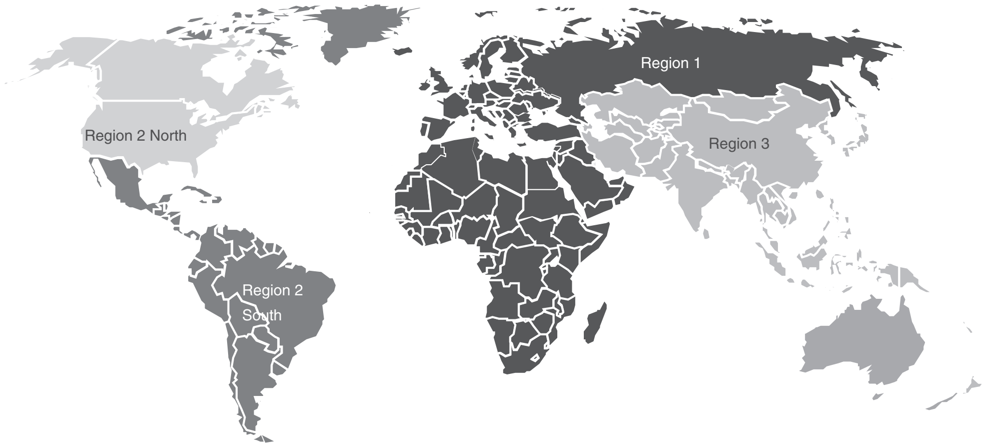
A mobile network (also called Public Land Mobile Network – PLMN) is usually separated into a Radio Access Network (RAN) and a Core Network (CN). The functional elements are specified by 3GPP. Functional elements relevant for LTE are described in later sections and chapters of this book (see e.g., section 1.6.5).

The RAN consists of all functions that are necessary to establish, maintain, and teardown connections between a user device (also called UE) and the network via the air interface. The RAN consists of radio base stations with their antennas that are spread over the geographical serving areas of the whole country.

The CN consists of all functions that are necessary to authenticate and authorize the user, setup voice calls or data connections, support mobility, charging, and lawful interception. The CN provides also interfaces to other mobile or fixed networks and to data networks such as the Internet or a company Intranet. Subscriber data are stored in a central register called Home Subscriber Server (HSS) that is part of the CN.

1.6.2 Architecture Principles

LTE/SAE has evolved from the 2G/3G PS domain, and especially EPC has its roots in the GPRS core network. Separation of control and user plane functions was a key in the design of EPC, thus the EPC basically consists of three functional elements. One is the MME that resides in the control plane of EPC. The MME can be seen as an evolution of the SGSN control plane function in GPRS. The Serving Gateway (S-GW) correlates with the SGSN user plane function in GPRS. All user plane packets in Uplink (UL) and Downlink (DL) are traversing the S-GW and the S-GW also acts as a local mobility anchor that is able to buffer downlink packets during handover. The P-GW finally is the global IP mobility anchor point comparable to



Map 1.1 Potential frequency bands for Public Safety. Numbers indicate MHz. Region 1 (mainly Africa, Europe, and Russia): 380–385/390–395. Region 2 North (United States and Canada): 3GPP band 14, 758–768/788–798. Region 2 South (Latin America): 746–806, 806.869, 4940–4990. Region 3 (mainly Asia): 406.1–430, 440–470, 806–824/851–869, 4940–4990, 5850–5925. Australia and other countries plan for the APT700 band (APT = Asia-Pacific Telecommunity), which is a segmentation of the 698–806 MHz band

the GGSN in GPRS. It allocates IP addresses to UEs and provides the interface toward Packet Data Networks (PDNs) such as the Internet or the mobile operator's service domain. The P-GW also contains the Policy Enforcement Function (PCEF) for the detection of service data flows, policy enforcement (e.g., discarding of packets), and charging (see section 1.6.12). All these network elements are logical functions, that is, in real implementations two or more functions (e.g., S-GW and P-GW) can reside on the same physical hardware platform. While the MME is selected by the eNodeB for a new session, the MME itself selects S-GW and P-GW by constructing special domain names and resolving these names by means of operator's Domain Name System (DNS) infrastructure. In the following sections, we will describe the EPS architecture variants for non-roaming and roaming cases and the architectures for interworking to 2G/3G. The functional description of the network elements can be found in Section 1.6.5.

1.6.3 Non-roaming Architecture

Figure 1.2 gives an overview of the logical LTE/SAE architecture in the non-roaming case, that is, the UE is served by its Home Public Land Mobile Network (HPLMN). In this and the following architecture figures, control plane interfaces are indicated with dotted lines while user plane interfaces are using full solid lines.

An overview of the main EPS reference points, their roles, and the underlying protocols can be found in the Appendix.

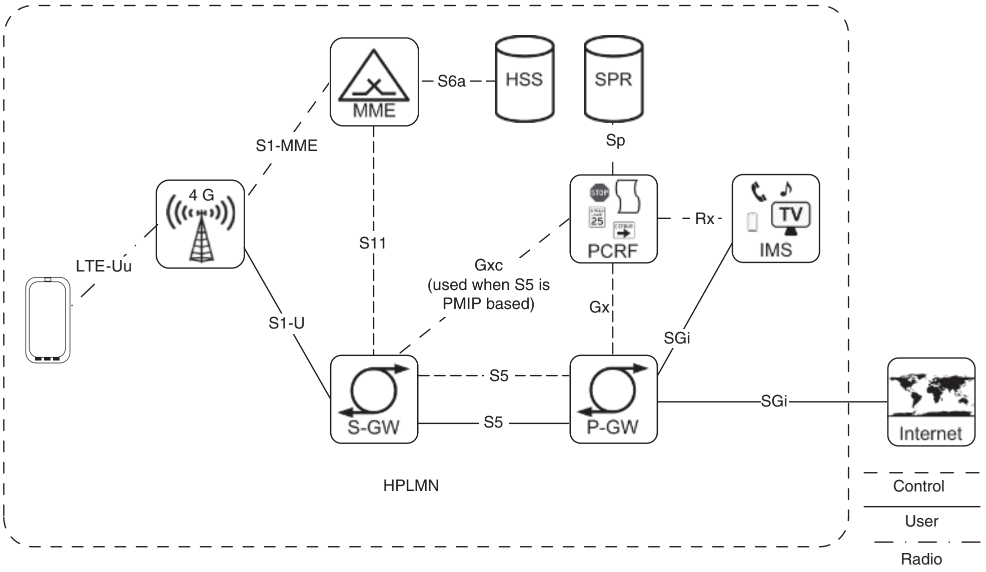


Figure 1.2 LTE/SAE non-roaming architecture

1.6.4 Roaming Architectures

1.6.4.1 Home Routed Roaming Architecture

The roaming architecture, that is, the UE is served by a VPLMN, is not much different from the non-roaming architecture. Main difference between non-roaming and roaming architecture is that in the roaming case S-GW is located in the VPLMN while P-GW is usually located in the HPLMN as most of the traffic is home routed traffic. Only if local breakout is used, the P-GW can be located in the VPLMN, but this scenario requires special arrangements between operators (e.g., usage of special Access Point Names (APN), providing charging information from VPLMN to HPLMN) and the user must be subscribed to this service. Thus, home routed traffic is the dominant usage scenario for PS services until now and this may be true also in the near future. The roaming architecture in the home routed scenario is shown in Figure 1.3.

As can be seen the S5 interface is replaced by S8 in roaming cases and in fact both are providing nearly identical functionality. This is similar to usage of Gn/Gp interfaces (see TS 29.060 [8]) in 2G/3G. S8 is either based on GTP or PMIP like S5. However, using PMIP for S8 requires either appropriate roaming agreements between operators or the VPLMN needs an interworking function to translate PMIP into GTP at the network border.

1.6.4.2 Local Breakout Roaming Architecture

Finally, we show the roaming architecture for local breakout to use services accessible via the VPLMN. The key point in this scenario is that the P-GW is located in the VPLMN and

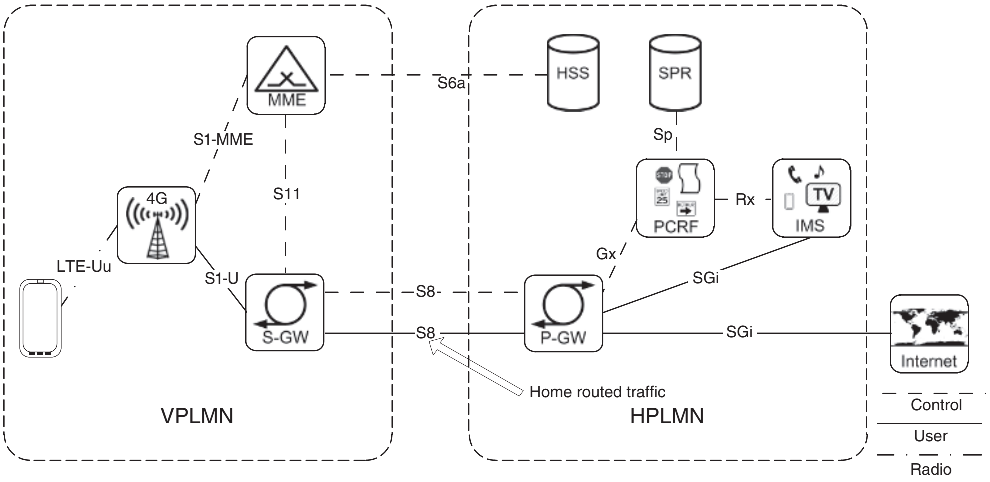


Figure 1.3 Roaming with home routed traffic

a Visited PCRF (V-PCRF) exists in the VPLMN as well to terminate the Gx interface (Gx is not an inter-operator interface). To receive QoS rules from HPLMN where the subscription data are stored (e.g., the subscribed maximum bit rate), a new interface S9 was introduced that connects the Home PCRF (H-PCRF) in HPLMN with the V-PCRF in VPLMN.

One possible solution to obtain VPLMN services requires the use of specially constructed APNs that are configured in the UE and used in the VPLMN to resolve them to a P-GW address in the VPLMN. Another possible solution is for UEs to use well-known standardized APNs like the IMS APN (see also Section 1.7) defined by Global System for Mobile Communications Association (GSMA) to get connectivity to a P-GW in the VPLMN.

As can be seen from Figure 1.4, the user can potentially use operator services in the HPLMN and VPLMN when local breakout is deployed.

1.6.5 Description of Functional Entities

1.6.5.1 User Equipment, Mobile Equipment, and the Universal SIM

In 3GPP terminology the UE is a device used by the end user for communication with the network. It is typically a smartphone, tablet, or modem equipped with LTE radio and is most often multiradio capable, that is, equipped also with other radios such as WLAN, cdma2000® 1xRTT, UTRAN, and GERAN. A UE consists of a Mobile Equipment (ME) part, which is the phone hardware that might be constituted of display, keypad, battery, and all electronics necessary to access and communicate with the network and providing the interface to the user. The only other piece of hardware required to form a UE is a Universal Integrated Circuit Card (UICC), which according to the definition in 3GPP TS 21.905 [9] is a “physically secure device, an IC card (or “smart card”), that can be inserted and removed from the terminal. It may contain one or more applications. One of the applications may be a USIM.”

The Universal Subscriber Identity Module (USIM) contains all data and algorithms necessary to authenticate the UE and verifying the authenticity of the network (only in case of

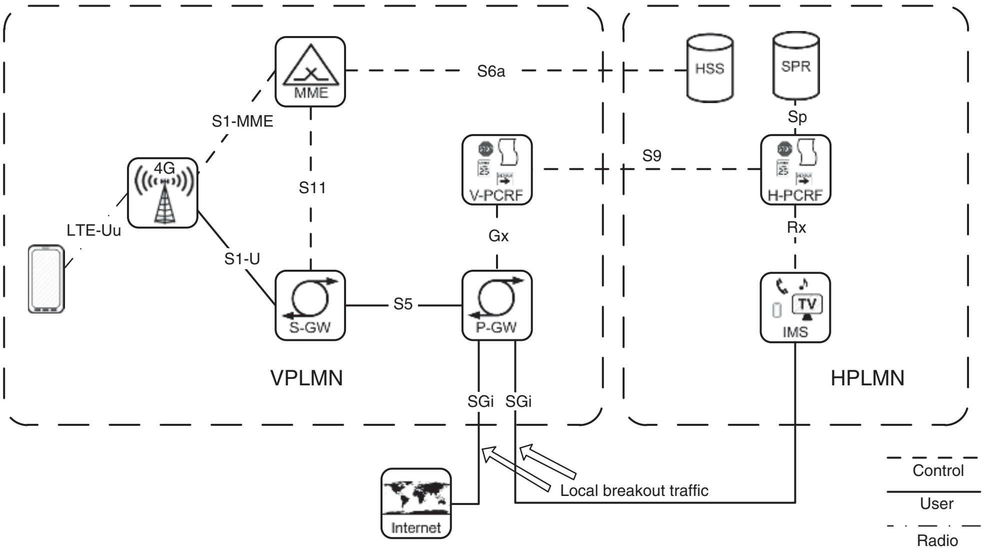


Figure 1.4 Roaming with local breakout

UMTS and LTE). Frequently, the terms UICC and USIM are referred to as Subscriber Identity Module (SIM), as this is the term used in the predecessor organization of 3GPP, the ETSI Special Mobile Group (ETSI SMG).

In addition to the USIM application, the UICC can also contain other applications like the IMS – SIM (ISIM) that provides all necessary data for the UE to access the IP Multimedia Services domain.

A typical UE must go through a series of steps to request and manage a service. When it powers on it needs to perform the following steps:

1. Scanning for a LTE cell, synchronize with the network, and listen for system information over the broadcast channels.
2. Establish a signaling connection in order to communicate with the network.
3. Register with the network.
4. Establish a data connection to be "always-on."
5. Respond to authentication requests when initiated by the network.
6. Receive an IP address for IP connectivity.

Afterwards the UE can request for specific resources that are needed to run one or more applications.

1.6.5.2 E-UTRAN Node B (eNodeB/eNB)

Main part of the E-UTRAN architecture is the eNodeB. The name is derived from the NodeB in UMTS and just extended with the letter "e" that stands for "evolved." It is the base station that is in control of all radio-related functions. It is typically deployed throughout the network coverage area, each eNodeB resides near the actual radio antennas. As the eNodeB is the only E-UTRAN node it consists of functions that reside in UMTS in the NodeB and partly in the RNC (other RNC functions were moved to the MME).

It terminates the radio protocols from the UE and relays data between UE and EPC. It supports the following main functions:

- Ciphering/deciphering of user plane traffic, IP header compression, and decompression.
- Radio Resource Management (RRM) functions including radio bearer control and admission control.
- Scheduling traffic according to the assigned Quality of Service (QoS), constant monitoring of the resource usage situation, and radio resource allocation in both uplink and downlink directions toward the UE.
- Mobility management functions: controls and analyzes the radio signal level measurements carried out by the UE, performs measurements, and takes handover decisions for UE(s) based on current radio link quality.
- Select a MME and route the data traffic to the S-GW.

The eNodeBs are interconnected by the X2 interface, connected to the MME by the S1-MME (control plane) interface, and connected to the S-GW by the S1-U (user plane) interface. Note that an eNodeB can connect to multiple MME(s) and/or multiple S-GWs for load balancing and network sharing purposes.

1.6.5.3 Mobility Management Entity (MME)

The MME is the main control element in the EPC. It is derived from the control plane part of the 2G/3G SGSN, enhanced with some functions that were inherited from the 3G RNC. Typically a MME will be a server in a secure location in operator's premises. The MME is only part of the control plane path. User plane data bypass the MME. Thus, the MME has no charging functions, except when it supports the feature "SMS in MME" (see Section 1.8.2).

Besides physical connections to the eNodeB, the MME has also a logical connection with the UE. This logical connection is referred to as Non Access Stratum (NAS). A description of NAS and Access Stratum (AS) can be found in Section 1.13.3. The MME connects also to the user's home HSS to authenticate and authorize the UE and retrieve subscription data. In a nutshell, MME supports the following functions:

- Control plane traffic handling (termination of signaling from the UE).
- Session and mobility management, for example, idle mode mobility and handover control.
- Paging of UE(s) that are in idle mode.
- TA list management.
- Selection of P-GW and S-GW.
- MME selection during handover.
- Coordinates inter-S-GW and inter-MME relocations.
- Authentication and authorization of the UE.
- Bearer management including dedicated bearer establishment.
- Lawful interception of signaling traffic.

1.6.5.4 Serving Gateway (S-GW)

Main function of the Serving Gateway is routing user plane packets between E-UTRAN and P-GW. The S-GW is derived from the user plane part of the 2G/3G SGSN. It is part of the network infrastructure and can be deployed centrally or decentrally in the network. It acts as the local anchor point for inter-eNodeB mobility. If a bearer was established for a certain UE and the UE is in idle mode, that is, there is no signaling connection between UE and network, the S-GW will buffer incoming data packets and request the MME to page the UE. The S-GW is connected to one or more P-GW(s). For each UE and each bearer, a tunnel between S-GW and P-GW is established. The S-GW supports the following functions:

- User plane anchor for mobility between 2G/3G and LTE and for inter-eNodeB handover.
- Lawful Interception (LI).
- Packet buffering and initiation of paging.
- Packet routing and forwarding.
- Transport level packet marking.
- Generation of charging events for interoperator accounting in case of roaming.

1.6.5.5 Packet Data Network Gateway (P-GW)

The P-GW is the first IP hop router from UE point of view and the edge router between the EPS and external PDNs. It is the central mobility anchor and acts as the IP point of attachment for the UE. The UE may be connected to multiple PDNs at the same time through the same or different P-GWs (connecting to different P-GWs is in practice rather unusual). During handover the P-GW is not changed. It is anchoring the user plane for inter-S-GW mobility by maintaining tunnels per UE and bearer toward the S-GW. It is also part of the network infrastructure maintained centrally in operator premises.

One important function of the P-GW is the allocation of an IP address to the UE per PDN connection. On the basis of the particular PDN this can be a private or public IPv4 address or an IPv6 prefix.

On the basis of dynamic or static policy rules the P-GW performs enforcement functions such as shaping, gating, filtering, and packet marking. More precisely, enforcement is done in the Policy Enforcement Function (PCEF), which is an integral part of the P-GW. It ensures that the downlink data rate does not exceed the allowed (i.e., subscribed) maximum bit rate for a particular user and that data packets received in downlink direction are using the correct bearer/tunnel (this is called bearer binding). Following are the main functions supported by the P-GW:

- Edge router to other networks.
- Allocation of IP addresses to the UE.
- Central mobility anchor.
- Policy and Charging Enforcement, bearer binding.
- Packet Filtering (optionally Deep Packet Inspection).
- Accounting per UE and per bearer.
- Lawful Interception.

1.6.5.6 Policy and Charging Rules Function (PCRF)

The Policy and Charging Rules Function (PCRF) is responsible for dynamic policy and charging control (PCC). For more details on PCC, see Section 1.6.7. It is usually located in the operator premises along with other core network elements, for example, close to the P-GW.

The PCRF translates session data coming from the application layer (e.g., information on the codec used for a multimedia session) into access specific parameters. On the basis of these parameters it generates so-called PCC rules that specify which kind of QoS is applicable for which IP flows. These rules are installed and later on executed in the Policy Control Enforcement Function (PCEF) that is part of the P-GW. These rules also specify whether the P-GW should grant resource requests and whether it is allowed to process packets for a given IP flow. The PCRF takes subscription information (e.g., the maximum allowed bit rate) stored in the Subscriber Profile Repository (SPR) into account to make its decisions. The SPR is usually part of the HSS. In the 3GPP architecture the SPR has Sp interface to the PCRF, but this interface was never specified. If GTP is used between S-GW and P-GW, QoS rules are provided from PCRF to P-GW as P-GW can send QoS parameters further down to the RAN via GTP. If an operator decides to use PMIP between S-GW and P-GW, bearers are terminated in the S-GW as PMIP has no bearer concept implemented. In this case, QoS rules are provided from the PCRF to the S-GW while policy and charging rules are still provided to the P-GW.

1.6.5.7 Home Subscriber Server (HSS) and Subscriber Profile Repository (SPR)

The HSS is the central database for all subscriber related data in the network. It contains subscriber-specific data such as International Mobile Subscriber Identity (IMSI), Mobile Station International ISDN Number (MSISDN), subscribed APN, priority indication, and subscribed supplementary voice services (e.g., call forwarding). It is centrally located in the operator premises. The HSS contains logically the Home Location Register (HLR) function that is the central register in legacy 2G/3G networks. Originally the HSS function was introduced with 3GPP Release 5 to host IMS subscription data like the IMPI and IMPU(s).

Although the SPR is an extra functional entity defined by 3GPP, it is usually (in real-life implementations) collocated with the HSS. The SPR contains subscription information that is used by the PCRF to make policy decisions and generate appropriate PCC rules. Such data are, for example, the maximum allowed bit rates for a user, the subscribed guaranteed bandwidth, whether user has a prepaid or postpaid contract and whether user receives preferred treatment based on his status (“bronze,” “silver,” and “gold” users).

The HSS is involved in user authentication and authorization and other security-related functions, generating and storing keys, parameters for ciphering, mutual authentication, and message integrity checking. The HSS also records user’s physical location (the addresses of MME, SGSN, and MSC serving the user). More than one HSS may be present in the network, depending on the number of subscribers and capacity of the hardware platform. In this case, identifiers such as MSISDN or IMSI can be used to select the correct HSS.

When the MME authenticates the user and authorizes the user request to access network resources (e.g., authorizing the APN provided by the UE), it needs access to the subscription data stored in the HSS. The HSS provides also the security key $\mathbf{K}_{\mathrm{ASME}}$ from which the MME derives the security keys to cipher and integrity protect NAS messages exchanged between UE and MME. For location management purposes, the HSS stores the addresses of MME serving a particular UE.

1.6.5.8 Authentication, Authorization, and Accounting (AAA) Function

The Authentication, Authorization, and Accounting (AAA) server (not shown in the various architecture figures) is either used to authenticate and authorize users who are accessing the EPC through non-3GPP access systems or, optionally, by the P-GW to authorize usage of the provided APN and to allocate an IP address to the UE. The P-GW uses RADIUS or DIAMETER according to 3GPP TS 29.061 [10] to access the AAA server via the SGi interface. The operator's AAA server can also work as an AAA proxy and interwork with AAA servers of other PLMN(s) or in company networks. It is centrally located in the operator premises and has a direct interface to the HSS.

1.6.6 Session Management

1.6.6.1 Quality of Service and EPS Bearers

The EPS provides IP connectivity between a UE and a packet data network external to the PLMN. This is referred to as PDN connectivity service. An EPS bearer uniquely identifies traffic flows that receive a common QoS treatment. It is the level of granularity for bearer level QoS control in the EPC/E-UTRAN. All traffic mapped to the same EPS bearer receives the same bearer level packet forwarding treatment. Providing different bearer level packet forwarding treatment requires separate EPS bearers.

An EPS bearer is referred to as a GBR bearer, if dedicated network resources related to a Guaranteed Bit Rate (GBR) are permanently allocated once the bearer is established or modified. Otherwise, an EPS bearer is referred to as a non-GBR bearer.

Each EPS bearer is associated with a QoS profile including the following data:

- QoS Class Identifier (QCI): A scalar pointing in the P-GW and eNodeB to node-specific parameters that control the bearer level packet forwarding treatment in this node.
- Allocation and Retention Priority (ARP): Contains information about the priority level, the pre-emption capability, and the pre-emption vulnerability. The primary purpose of the ARP is to decide whether a bearer establishment or modification request can be accepted or needs to be rejected due to resource limitations.
- GBR: The bit rate that can be expected to be provided by a GBR bearer.
- Maximum Bit Rate (MBR): Limits the bit rate that can be expected to be provided by a GBR bearer.

Following QoS parameters are applied to an aggregated set of EPS bearers and are part of user's subscription data:

- APN Aggregate Maximum Bit Rate (APN-AMBR): Limits the aggregate bit rate that can be expected to be provided across all non-GBR bearers and across all PDN connections associated with the APN.

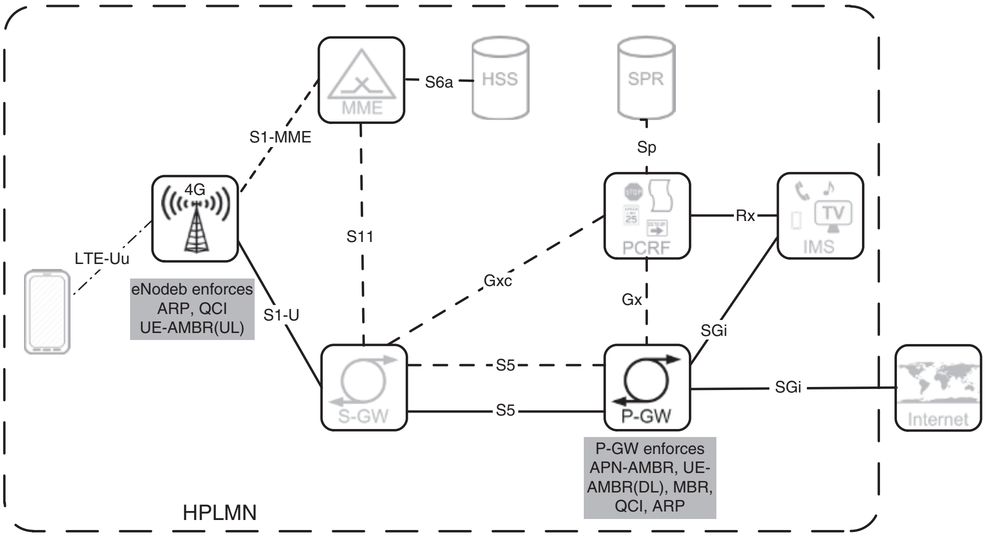


Figure 1.5 QoS enforcement in EPS

- UE Aggregate Maximum Bit Rate (UE-AMBR): Limits the aggregate bit rate that can be expected to be provided across all non-GBR bearers of a UE.

The UE routes uplink packets to the different EPS bearers based on uplink packet filters assigned to the bearers while the P-GW routes downlink packets to the different EPS bearers based on downlink packet filters assigned to the bearers in the PDN connection.

Figure 1.5 shows the nodes where QoS parameters are enforced in the EPS system.

1.6.6.2 Session and Bearer Management

With Session and Bearer Management procedures, EPS bearers for a particular UE are established and maintained. The default EPS bearer context is activated during the EPS Attach procedure. Upon successful attach, the UE can request setting up connections to additional PDNs. For each additional PDN connection, the MME activates a separate default EPS bearer. A default EPS bearer context remains activated throughout the lifetime of the PDN connection. Each PDN connection is characterized by the IP address that is assigned to this connection.

A dedicated EPS bearer is always linked to a default EPS bearer and inherits from it characteristics like the IP address, that is, UL traffic on a dedicated bearer uses the same source IP address as traffic on the default bearer. A dedicated bearer is used when additional EPS bearer resources with a specific QoS between UE and PDN are required. The distinction between default and dedicated bearers is transparent to the eNodeB. As an example: In case of VoLTE (see Section 1.8.1) voice and video streams use dedicated bearers associated with the IMS PDN connection while IMS/SIP signaling can use the default EPS bearer. A dedicated bearer is established via the dedicated bearer context activation procedure. It can either be part of the attach procedure or initiated together with the default EPS bearer context activation procedure.

The procedure is usually initiated by the network, but may be also requested by the UE. If the UE requests additional EPS bearer resources, the network decides whether to fulfill such a request by activating a new dedicated bearer or modifying an existing dedicated or default bearer.

By using the PCC framework the network can initiate the activation of dedicated EPS bearers together with the activation of the default EPS bearer or at any time later, as long as the default EPS bearer remains activated.

Default and dedicated EPS bearers can be modified. Dedicated EPS bearers can be released without affecting the default EPS bearer. If the default EPS bearer is released, all dedicated bearers linked to it are also released.

Readers interested in detailed call flows for UE-requested PDN connectivity request and network-initiated dedicated bearer activation procedure may refer to the Appendix.

1.6.6.3 IP Address Allocation

The network can assign three types of IP addresses to the UE once it connects to the network: either a private or public IPv4 address or an IPv6 prefix or both. The IP address can be allocated by the HPLMN, VPLMN (in case of local breakout), or potentially by an external service provider (e.g., a company network).

The type of IP address allocated to the UE depends on the PDN connection and on UE capabilities. The HSS stores one or more PDN types per APN in the subscription data. During Attach or UE-requested PDN connectivity procedure, the MME takes the requested and the subscribed PDN types into account before finally determining the PDN type for the connection.

DHCPv4 or 3GPP-specific signaling during PDN connection establishment is used to deliver an IPv4 address to the UE. IPv6 stateless address auto-configuration is used to assign a 64-bit globally unique prefix to the UE. If a shorter prefix has to be allocated for a PDN connection, it is delivered using DHCPv6 prefix delegation after the UE has first received the 64-bit prefix with stateless auto-configuration. It has to be noted that neither DHCPv4 nor DHCPv6 prefix delegation is currently widely used in live networks. In most of the cases, an AAA server assigns an IP address for a PDN connection and provides it to the P-GW via SGi signaling (RADIUS or DIAMETER).

1.6.7 Policy and Charging Control

The PCC system allows operators to dynamically maintain IP bearers with a certain QoS and provide means for online and offline charging of single-service data flows (for the charging aspects please refer to Section 1.6.12).

PCC helps to enforce the service data flows that are transmitted over a certain bearer. A service data flow is an aggregate of IP packet flows defined by 5-tuple filters (note that one service data flow can consist of several IP packet flows). A bearer can be seen as a transmission channel from the UE via eNodeB toward GGSN/P-GW with a certain capacity, delay, and bit error rate provided to all service data flows transported within it. Enforcement of QoS is performed hop-by-hop, on the radio link between UE and eNodeB, on the transport link between eNodeB and GGSN/P-GW, and finally beyond the GGSN/P-GW. However, support for QoS beyond GGSN/P-GW is out of scope of 3GPP and is up to configuration.

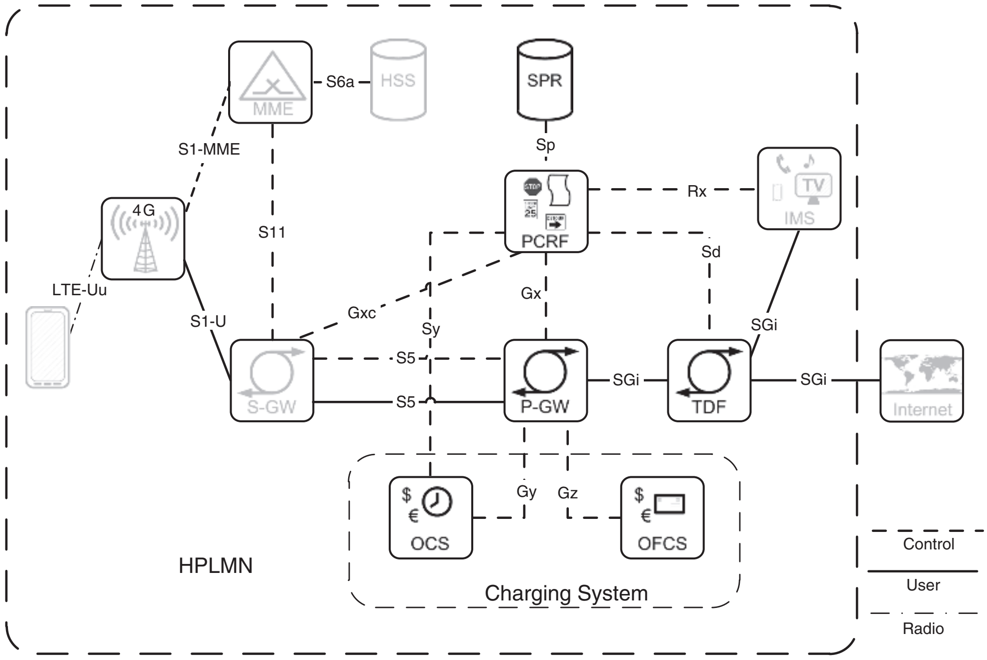


Figure 1.6 PCC architecture

Figure 1.6 gives a simplified overview of the PCC architecture. For details, see 3GPP TS 23.203 [11].

The central control function in the PCC architecture is the Policy and Charging Rules Function (PCRF). Basic idea is that the PCRF obtains information about new and ongoing media sessions from the application layer (called Application Function (AF)) and translates this into so-called policy rules that are enforced at the PCEF. AF and PCRF communicate via the DIAMETER-based Rx interface with each other (see 3GPP TS 29.214 [12]). In LTE networks the PCEF is implemented in the P-GW. The PCEF is responsible for the binding of service data flows to bearers, that is, to decide which flow goes into which bearer. IP flows in downlink and uplink that are traveling through the P-GW are then handled according to these policy rules. Policy rules are installed by the PCRF in the PCEF via the Gx interface, which is also based on DIAMETER. The PCC rule determines the authorized QoS, which is also signaled to the eNodeB, applicability of online/prepaid or offline/postpaid charging for this particular bearer, and addresses of respective online or offline charging servers. Policy rules can enforce traffic from/to certain destinations on a bearer, thus blocking traffic flows from/to other destinations. QoS-relevant subscription data (e.g., subscribed maximum bandwidth per user) are stored in the SPR, which is usually part of the HSS, but can be a stand-alone function. The PCRF has access to the SPR via the non-standardized Sp interface. The P-GW interacts with the Online and Offline Charging Systems (OCS/OFCS) via the DIAMETER-based Gy/Gz interfaces. Alternatively, Gz may use the GTP' protocol as well, which is a flavor of GTP, and P-GW may even send offline charging records directly to a billing system without passing through an OFCS. The OCS allows deploying prepaid charging schemes based on time or volume usage. The Sy interface allows the OCS to change policy rules based on the current usage of resources, for example, to downgrade the available bit rate for a user who has spent more than 4 GB data volume in a month. The Traffic Detection Function (TDF) is the 3GPP terminus for a Deep Packet Inspection (DPI) node and allows detecting applications based on different criteria like IP 5-tuples or the characteristics of application-specific messages. The PCRF can instruct the TDF via the Sd interface to start detection of applications for certain users and the TDF can inform the PCRF that an application was detected. On the basis of this information, policy rules can be installed at the PCEF (e.g., to block the service or to apply different charging rules). The TDF can be a stand-alone function or co-located with the P-GW.

1.6.8 Interfaces and Protocols in EPS

1.6.8.1 Control Plane

The control plane consists of all protocols used for signaling between UE and network or between two network entities.

UE – eNodeB – MME

Figure 1.7 shows the control plane protocol stack between UE, eNodeB, and MME.

A description for the 3GPP defined protocols (white background in the figure) and corresponding reference specifications are provided here. Other protocols based on IETF standards are not described.

Non Access Stratum (NAS)

NAS is the highest layer of the control plane between UE and MME at the radio interface. Main functions of the NAS protocol are the support of mobility of the UE and the support of session management procedures to establish and maintain IP connectivity between UE and P-GW. For further details, refer to 3GPP TS 24.008 [13], 3GPP TS 24.301 [14], and 3GPP TS 23.122 [15]. See also Section 1.13.3 for a comparison of NAS and AS.

Radio Resource Control (RRC)

RRC is the primary control protocol of the LTE–Uu interface, which is the interface between UE and eNB. It is responsible for a wide variety of system functions, including system information over the broadcast channel, radio configuration (set up, maintenance, and teardown of radio resources), measurements, and mobility. For further details, please refer to 3GPP TS 36.300 [16].

Packet Data Convergence Protocol (PDCP)

PDCP is responsible for ensuring the integrity of packets sent over the air interface. In addition, the PDCP layer compresses the IP header of a data packet. For further details, please refer to 3GPP TS 36.300 [16].

Radio Link Control (RLC)

RLC provides a logical link control mechanism over the air interface. Main task of RLC is the segmentation of PDCP packets in smaller parts that can be transmitted over the air. For further details, please refer to 3GPP TS 36.300 [16].

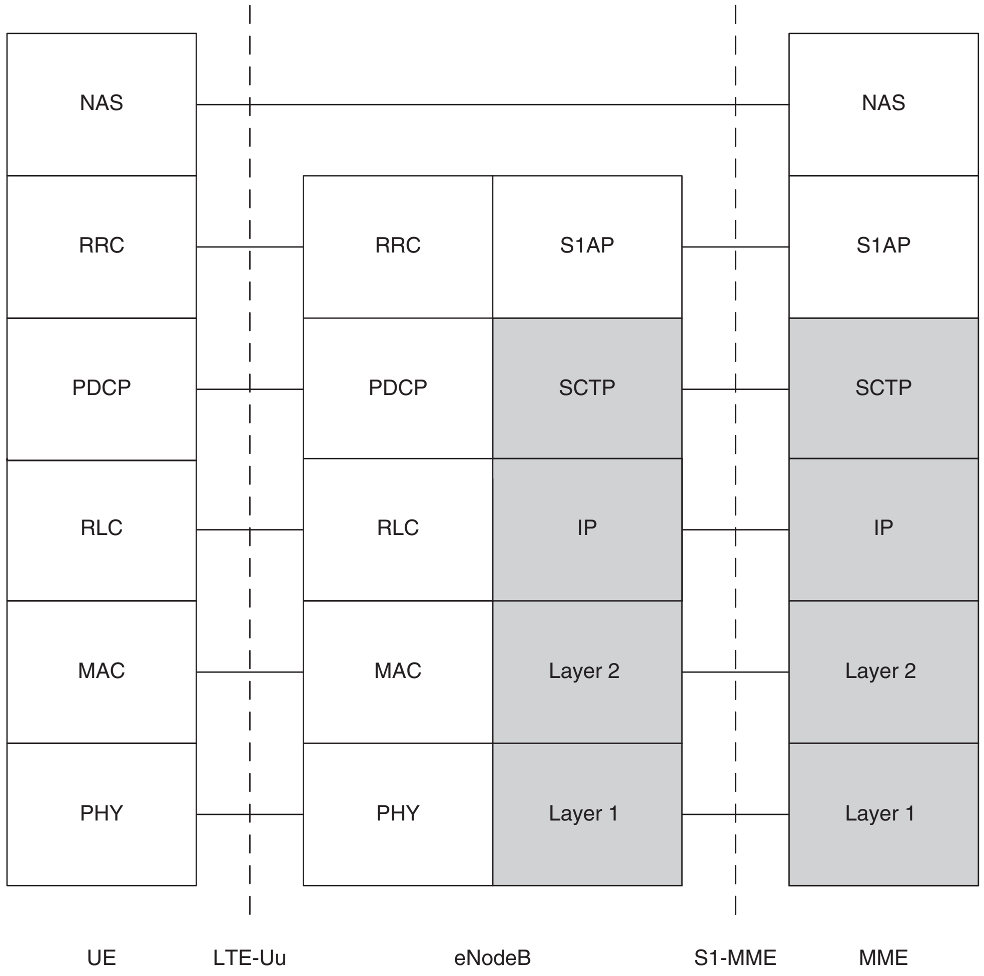


Figure 1.7 Control plane stack between UE, eNodeB, and MME

Medium Access Control (MAC)
The MAC layer selects a transport channel to transmit data and manages the mapping of logical channels to transport channels. In addition, the MAC layer is responsible to multiplex data to common and shared channels. For further details, please refer to 3GPP TS 36.300 [16].

Physical Layer (PHY)
This is the layer 1 of the LTE−Uu interface. Physical channels differentiate the source of the transmission as well as its destination. A particular message may be destined for a specific UE (a handover command) or may be intended for all active UE(s) (system broadcast messages). For further details, please refer to 3GPP TS 36.300 [16].

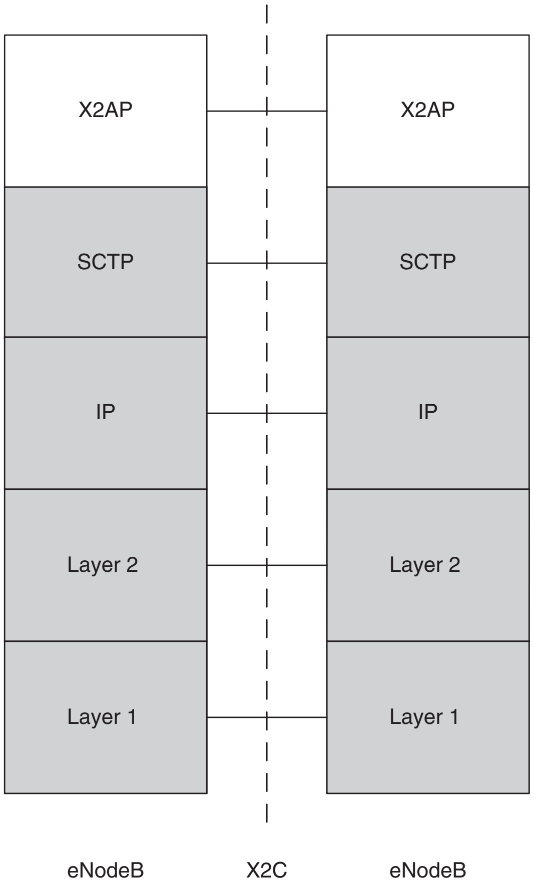


Figure 1.8 Control plane stack between eNB and eNB

S1-Application Protocol (S1AP)

S1AP provides the signaling service between E-UTRAN (eNB) and the EPC (MME). S1AP services are divided into two groups:

- Non-UE-associated services: They are related to the whole S1 interface between the eNB and MME utilizing a generic signaling connection.
- UE-associated services: They are related to one UE. S1AP functions that provide these services are associated with a UE-associated signaling connection that is maintained for the UE.

For further details, please refer to 3GPP TS 36.413 [17].

eNodeB–eNodeB

Figure 1.8 shows the control plane protocol stack between two eNB(s) (X2 control reference point).

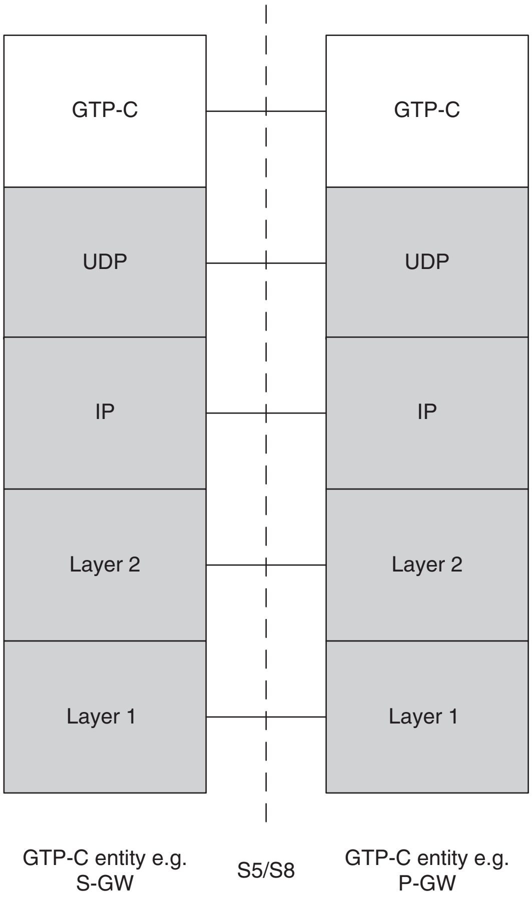


Figure 1.9 GTP-C control plane protocol stack

X2-Application Protocol (X2AP)
X2AP is used for mobility management procedures between two eNB(s) including handover preparation. In general, it is used to maintain the relationship between two eNBs. For further details, please refer to 3GPP TS 36.423 [18].

MME–S-GW–P-GW
Figure 1.9 illustrates the control plane protocol stack between any two network elements that support the GTP-C protocol. This can apply for MME–S-GW (S11 interface), MME–MME (S10 interface), and S-GW–P-GW (S5/S8-GTP-C-based interfaces).

Short description for GTP-C (3GPP defined protocol) is provided here. Other protocols are based on standard IETF techniques.

GPRS Tunneling Protocol for the Control Plane (GTP-C)
Main purpose of GTP-C is to establish and maintain bearer tunnels (GTP-U tunnels) associated with bearer contexts in LTE/SAE for user sessions that require a certain QoS. For that purpose,

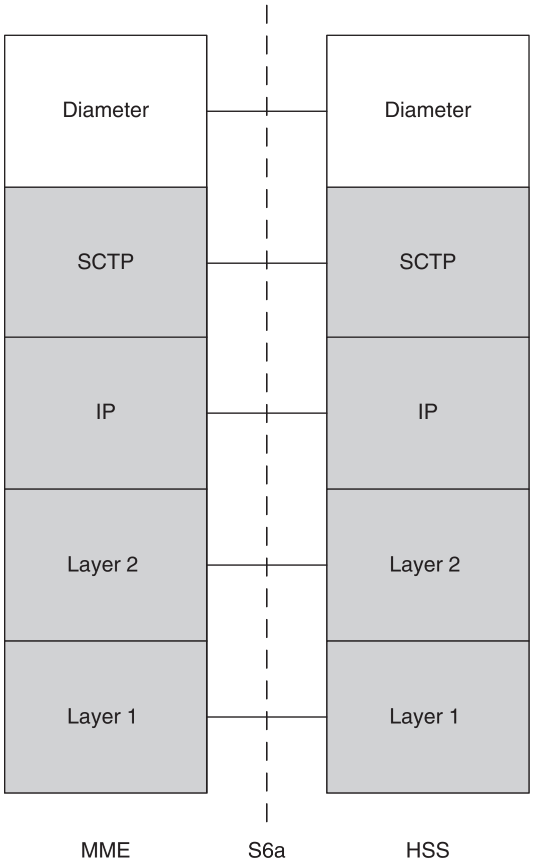


Figure 1.10 Control plane stack between MME and HSS

GTP-C enables the exchange of tunnel identifiers and tunnel addresses (IP addresses, port numbers) between the two involved entities. Such tunnels exist between radio and core network and within the core network. In EPC, the GTP tunnel is established between eNB and S-GW and between S-GW and P-GW. Together with the bearer on the radio interface between UE and eNB, this results in an end-to-end bearer with a certain QoS spanning from the UE toward the P-GW. Besides that GTP-C is also used to transport mobility management messages in case of relocation/handover. For further details refer to 3GPP TS 29.274 [19].

MME – HSS

Figure 1.10 shows the control plane protocol stack between MME and HSS.

S6a DIAMETER Application

Main function of S6a is to support transferring of subscription and authentication data for user authentication and authorization between MME and HSS. DIAMETER is defined in RFC 3588 [20] and the S6a DIAMETER application in 3GPP TS 29.272 [21].

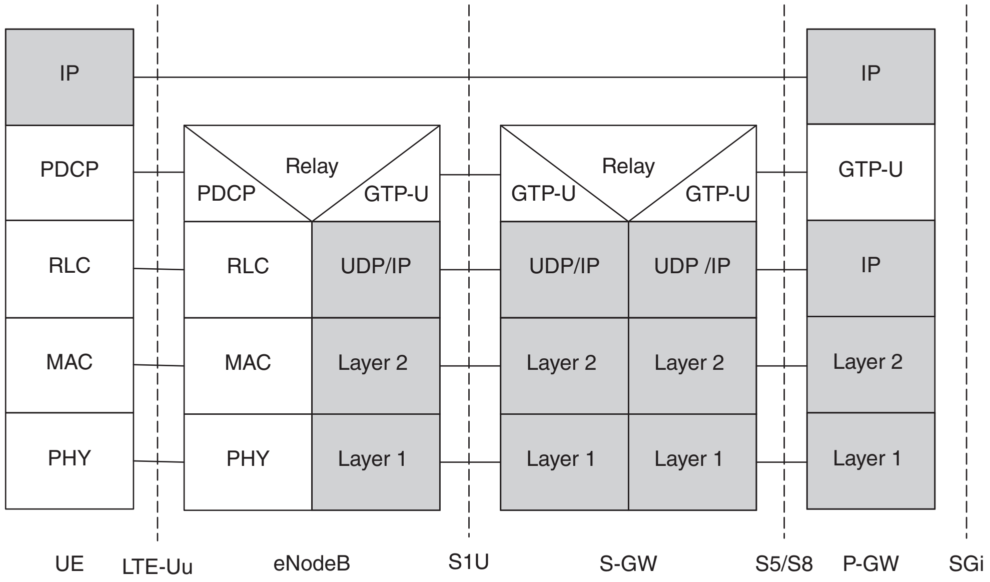


Figure 1.11 User plane stack between UE and P-GW

1.6.8.2 User Plane

Figure 1.11 shows the end-to-end user plane stack for a UE connecting toward a P-GW in EPS.

The protocol stack is similar to the control plane stack. The only new protocol being introduced here is GTP-U.

GPRS Tunneling Protocol for the User Plane (GTP-U)
This protocol tunnels user data between eNodeB and eNodeB, eNodeB and S-GW, as well as between S-GW and P-GW. GTP-U encapsulates all user plane packets. For further details on GTP-U, please refer to 3GPP TS 29.281 [22].

1.6.8.3 Summary of Reference Points and Protocols

Table 1.3 summarizes the reference points and protocols used for LTE/SAE and PCC.

1.6.9 Mobility Management

In a broader sense “mobility management” includes all procedures used to support registration and mobility of a UE, such as the following:

- Selecting a network.
- Attaching to and detaching from the network.

Table 1.3 Reference points and protocols for LTE/SAE and PCC

Reference point	Protocols	Specifications
LTE-Uu	CP:PHY/MAC/RLC/PDCP/RRC UP:PHY/MAC/RLC/PDCP	TS 36.300 [16]
UE – MME	NAS (EMM, ESM)	TS 24.301 [14]
X2	CP: X2AP	TS 36.423 [18]
	UP: GTP-U	TS 29.274 [19]
S1-MME	S1-AP	TS 36.413 [17]
S1-U	GTPv1-U	TS 29.281 [22]
S5	GTPv2-C/GTPv1-U	TS 29.274 [19]/TS 29.281 [22]
	PMIPv6	TS 29.275 [23]
S8	GTPv2-C/GTPv1-U	TS 29.274 [19]/TS 29.281 [22]
	PMIPv6	TS 29.275 [23]
S6a	DIAMETER	TS 29.272[21]
S9	DIAMETER	TS 29.215 [24]
S11	GTPv2-C	TS 29.274 [19]
Sp	Not specified in 3GPP	Not specified in 3GPP
Gx	DIAMETER	TS 29.212 [25]
Gxa	DIAMETER	TS 29.212 [25]
Gxb	Not specified in 3GPP	Not specified in 3GPP
Gxc	DIAMETER	TS 29.212 [25]
SGi	IPv4, IPv6, RADIUS, DIAMETER, DHCP	TS 29.061 [10]
Rx	DIAMETER	TS 29.214 [12]

- Maintaining the connection to the network while the UE is moving (referred to as “handover”).
- Keeping the network informed about the present location of the UE in order to be reachable for paging even after the connection to the network has been released.
- Re-establishing the connection between UE and network when the UE needs to send uplink signaling or user data or when the UE was paged by the network because the network wants to send downlink signaling or user data.

Furthermore, certain security-related tasks are usually performed as part of the mobility management procedures: authentication, confidentiality protection of subscriber’s identity, and confidentiality and/or integrity protection of signaling messages and user data.

Readers interested in detailed call flows for Attach, Detach, TA Update, Paging, Service Request, and Handover procedures in addition to the overview provided in this section may refer to the Appendix.

1.6.9.1 Attach

Each UE needs to register with the network to receive EPS services. This registration is called “network attachment.” Always-on IP connectivity for a UE is enabled in EPS by establishing a so-called default EPS bearer during network attachment. The UE may request an IP address during the Attach procedure. The Attach procedure is triggered by the UE by sending an Attach Request message to the network. This message terminates at the MME. The Attach Request message is encapsulated in a RRC message toward the eNB and in a S1-MME control message to the MME. The MME that terminates the Attach Request message is called "serving MME." Its address is stored in the HSS.

One purpose of the Attach procedure is to authenticate the UE and activate integrity protection and ciphering of NAS messages exchanged between UE and MME. Necessary information for the authentication of a particular UE is obtained from the HSS. This information consists of so-called authentication vectors that are generated in the HSS from the keys stored in the subscription data.

If dynamic policy control is applied, the P-GW obtains PCC rules for the UE from the PCRF during the Attach procedure. If dynamic policy control is not deployed, the P-GW may apply local QoS policies. This could result in the establishment of a number of dedicated bearers for the UE in association or combination with the default bearer.

Finally, the UE is registered with the network to receive EPS services. It can request for specific resources to run a certain application and perform handover when radio coverage conditions change.

1.6.9.2 Detach

The Detach procedure allows the UE to inform the network that it does not want to access the network any longer and allows the network to inform the UE that it does not have access to the network any longer. Detaching a UE implies that all PDN connections and associated bearers are released.

The UE is detached either explicitly or implicitly. In case of an explicit detach, the network or UE explicitly requests detach and signal with each other. In case of an implicit detach, the network detaches the UE without notifying the UE. This is typically done when the network presumes that it is not able to communicate with the UE, for example, due to lack of radio coverage.

1.6.9.3 Tracking Area Update

Once the UE is successfully attached, it needs to keep the network informed about its current location in order to be reachable for downlink signaling and user data (incoming voice call or SMS) even when the UE is in "idle mode," that is, when the signaling connection between UE and network has been released. Moving in idle mode and keeping the network informed about its current location is in general referred to as "idle mode mobility." Although normally not active in idle mode the UE periodically wakes up and monitors the broadcast channel in order to receive information about incoming signaling or data traffic.

E-UTRAN cells are combined to Tracking Areas. A UE camping on an E-UTRAN cell in idle mode is listening to the system information broadcast in this cell, which includes the identity of the TA the cell belongs to. When the UE moves to another cell and the received TA identity indicates that the new cell belongs to a TA to which the UE is currently not registered (i.e., the TA is not in the list of registered TAs stored in the UE), the UE initiates a TA updating procedure.

In the simplest case, if the new cell and the new TA are served by the same MME to which the UE is already registered, only two NAS messages, TA Update Request and TA Update Accept, need to be exchanged between UE and MME. Otherwise, if the new cell is served by a new MME, or if the MME decides to change the S-GW, further network entities (HSS, old MME or old SGSN, old S-GW, P-GW) need to be involved in the procedure.

The TA or list of TAs allocated by the MME during the TA updating procedure or attach procedure can be used by the MME to page the UE in a certain area, when the signaling connection to the UE has been released and the network needs to send downlink signaling or user data. The MME can page the UE first in the last known TA, if the UE does not respond, paging can be performed in a wider area (e.g., in neighboring TAs) before the MME pages the UE in all TAs of the allocated TA list. The MME knows the location of a UE only up to the granularity of a TA, and it has no knowledge in which particular cell of a TA the UE is currently camping on.

1.6.9.4 Paging

The paging procedure is normally initiated by the network to request the establishment of a NAS signaling connection toward an UE that is in idle mode. In addition, the network can also initiate paging to inform the UE in RRC-IDLE or RRC-CONNECTED mode about system information change, inform the UE about an impending warning notification (e.g., for CMAS and ETWS as described in chapter 2). It can also be initiated by the network to request the UE to perform re-attach when the network has lost UE context due to a failure situation.

The trigger for paging a UE is usually a mobile-terminated transaction destined for the UE such as downlink data, an incoming SMS, or voice call.

It is up to the network (i.e., the MME) to decide how and with which priority to page the UE. The paging could be, for example, started in the last known cell and then extended to a wider area such as TA and TA list. Priority of paging can be determined by the bearer QCI and ARP.

When the UE is in RRC-IDLE state, the UE periodically monitors for a paging message from the network. The monitoring frequency in the UE is determined by the “Discontinuous Reception (DRX) cycle in idle mode.” This idle mode DRX value can be configured via the Broadcast Control Channel (BCCH) and the NAS layer. If the UE has to monitor paging periodically, this consumes some power (i.e., battery life) in the UE. Thus, in Release-12, 3GPP introduced a new mode called “Power Saving Mode (PSM)” in order to provide additional means to save battery life for devices that normally require only infrequently mobile-originated transactions. When UE is in PSM, idle mode procedure does not apply, that is, UE does not listen to paging nor perform measurements. Thus, the UE is not reachable for mobile-terminated transactions. This is meant mainly for devices (e.g., with frequency such as twice a day transmission, 8 bytes a day transmission) for which mobile-terminated transactions are very infrequent or some delay in mobile-terminated transactions is acceptable without impacting the end user experience.

1.6.9.5 Service Request

The purpose of the Service Request (SR) procedure is to move the UE from ECM-IDLE to ECM-CONNECTED state and establish EPS bearers when user or signaling data have to be transmitted. Other purposes are to trigger MO/MT CS Fallback (CSFB), explained later in this chapter, and Proximity Services procedures (see Chapter 4).

If the UE has a pending mobile-originated transaction when it is camping in E-UTRAN, then the UE initiates a Service Request procedure toward the network. If the network has a pending mobile-terminated transaction, the network initiates a paging procedure first. Once the paging message has been successfully processed by the UE, it responds with a Service Request message toward the network to establish the necessary bearers.

The NAS Service Request message is used for fast re-establishment of the NAS signaling connection and the user plane bearers. In order to meet the strict performance requirements for EPS when it is sent as a response to a paging message, the NAS Service Request message transmitted over the air was carefully designed to fit within a single-radio transport block so that the message need not be segmented over the air interface.

When new features caused the addition of information elements, this requirement could no longer be met. For this reason, a new Extended Service Request (ESR) message was defined. ESR is used in special cases (e.g., to trigger a CS Fallback procedure) when additional parameters need to be sent. ESR follows the layout of regular EMM messages.

Thus, the Service Request procedure can be initiated using a regular Service Request or an Extended Service Request. The detailed conditions for when to use Service Request or Extended Service Request are specified in 3GPP TS 24.301 [14].

Successful completion of a Service Request procedure is determined by the UE either based on indication of successful establishment of radio bearers (when the UE continues to remain in E-UTRAN) or based on successful intersystem change (in case of CS Fallback).

1.6.10 Intra E-UTRAN Handover

1.6.10.1 General

“Handover” refers to situations where the UE is moving from one cell to another while it maintains a signaling connection with the network. In case of intra E-UTRAN handover, the UE moves from one LTE cell to another one. Inter-RAT handover refers to the case where source and target cell belong to different radio technologies, for example, UE moves from an E-UTRAN to a UTRAN cell.

In case of intra E-UTRAN handover, the UE can be handed over from the currently serving eNB (“source eNB”) to a new one (“target eNB”). This handover procedure considers also existing data connections, that is, data connections are moved from the source to the target eNB. The handover process can even lead to a change of the serving S-GW and/or MME once the UE has moved into a service area that is no longer served by the old S-GW and/or MME. Only the P-GW as mobility anchor point is never changed during handover.

Two different signaling procedures have been defined for handover: X2-based and S1-based handover, named after the interface used for the exchange of signaling messages during the handover preparation.

1.6.10.2 X2-Based Handover

If both eNBs are connected to the same MME, source eNB and target eNB can exchange the S1AP signaling for the handover preparation and execution directly via the X2 interface (for details, see 3GPP TS 36.300 [16]).

The MME is involved in the procedure only during the handover completion phase when the UE has already been handed over to the target eNodeB. During the handover, the target eNB sends an S1AP message (Path Switch Request) to the MME indicating that the S1 interface needs to be switched from the source to the target eNB. Through this message MME is also aware that the UE has moved into a new cell. The MME updates the S-GW with the new address information for downlink user data and confirms the relocation of the S1 interface toward the target eNodeB. If the MME wants to use a different S-GW, it can allocate resources on the new S-GW and provide the target eNB with the necessary address information for sending uplink user data.

During the handover execution, downlink packets are forwarded by the source eNodeB to the target eNodeB. Once the S-GW receives the command to switch the user plane to the target eNodeB, it sends one or several “end marker” packets toward the source eNodeB to assist the target eNodeB with the reordering of packets.

After completion of the handover, the UE receives the system information broadcast in the target cell. If the target cell belongs to a TA to which the UE is not registered, the UE has to initiate a TA update procedure.

1.6.10.3 S1-Based Handover

S1-based handover is used in all cases when X2-based handover cannot be used, for example, when source eNB and target eNB are served by different MMEs (indicated by the TA identity of the target cell). During the handover preparation, the signaling information is exchanged between source eNB and target eNB via the involved MME(s), that is, messages are sent via the S1 interfaces, and possibly via the S10 interface between the MMEs. Independent of a possible MME change also the S-GW can be changed during S1-based handover.

In case of MME change, the source MME transfers context information about the UE to the target MME, including the EPS security context, and mobility management and session management information. For proper routing of uplink data packets, the MME provides the target eNB with the S-GW address.

During the handover execution, downlink packets are forwarded by the source eNB to the target eNB either directly or indirectly, that is, via the target S-GW.

After completion of the handover, the UE receives the system information broadcast in the target cell. If the target cell of the handover belongs to a TA to which the UE is not registered, the UE initiates a TA updating procedure.

1.6.11 Security

Communication via the radio interface is vulnerable to various security attacks such as eavesdropping, “man-in-the-middle” attacks, or subscriber tracking.

To protect the subscriber against eavesdropping, EPS supports encryption of signaling and user data.

In a "man-in-the-middle" attack, a "false base station" impersonates a "real" base station toward the UE and a UE toward the network, and relays, possibly modifies, the signaling between UE and network. To avoid "man-in-the-middle" attacks, the EPS supports integrity protection of signaling data.

The use of ciphering in a network is optional for the operator, whereas the use of integrity protection is mandatory. Without activation of integrity protection, the UE will not be able to successfully attach to the network and receive services. As an exception to this rule, the network can activate a so-called null integrity protection algorithm (see 3GPP TS 33.401 [26]) during an emergency attach for an unauthenticated, (U)SIM-less, emergency call. The null integrity protection algorithm is effectively providing no protection.

Before ciphering and integrity protection can be activated, UE and network need to establish an EPS security context. This security context is either created during an authentication procedure or derived from a UMTS security context during intersystem change from GERAN/UTRAN to E-UTRAN.

Once the EPS security context is used by the MME by means of a security mode control procedure, the UE will send all NAS messages integrity protected, including the initial NAS messages for subsequent network access and for re-attaching when the UE has temporarily detached from the network. Integrity protection in downlink direction and ciphering of NAS messages is started by the MME after successful authentication of the UE or when the network has verified the integrity protection of NAS messages sent by the UE by means of an already available EPS security context. Each time the UE accesses the network and establishes a new signaling connection, the MME needs to restart integrity protection and ciphering of NAS messages.

Security on AS level, including the ciphering of user data, is controlled separately by AS signaling procedures. Each time the UE accesses the network and establishes a new signaling connection, the MME needs to restart AS security.

A further aspect of subscriber confidentiality is the protection against location tracking of subscribers by third parties. In certain situations, signaling messages containing a subscriber identity need to be sent unciphered. To protect the subscriber against tracking, UE and network are using a temporary user identity whenever possible. During the initial attach procedure, the UE may need to identify itself toward the network with its permanent subscriber identity, the IMSI, but for subsequent accesses the UE will use the Globally Unique Temporary Identity (GUTI), which is assigned by the MME during attach procedure or the UE will use the S-TMSI which is a part of the GUTI. The GUTI is regularly reallocated by the MME via security-protected signaling.

In the downlink direction, the network is normally using the S-TMSI to page the subscriber.

An additional permanent identity related to the UE, the International Mobile Equipment Identity (IMEI), can be retrieved by the MME only via integrity-protected signaling. Thus, the IMEI can only be requested by an authorized network entity and is usually transmitted in ciphered form.

In addition to security on a per user basis, also security associations for links such as S1 (eNB to MME and S-GW) and X2 (eNB to eNB) providing integrity and confidentiality protection for all control and user plane traffic going via these interfaces exist. Figure 1.12 provides an overview of the overall EPS security architecture.

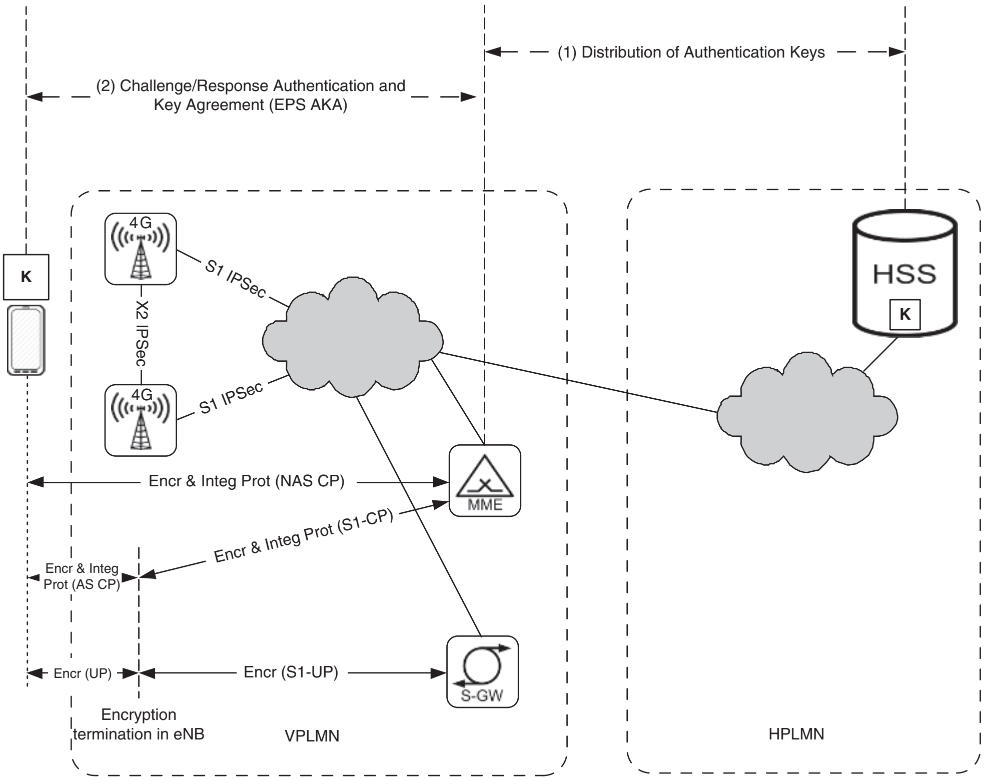

(1) Distribution of Authentication Keys
(2) Challenge/Response Authentication and Key Agreement (EPS AKA)
K
4G
S1 IPSec
X2 IPSec
4G
S1 IPSec
HSS
K
MME
Encr & Integ Prot (NAS CP)
Encr & Integ Prot (S1-CP)
Encr & Integ Prot (AS CP)
Encr (UP)
Encr (S1-UP)
S-GW
Encryption termination in eNB
VPLMN
HPLMN

Figure 1.12 EPS security architecture

Following are the list of per-user security associations.

1. RRC signaling between UE and eNB is integrity and confidentiality protected.
2. User plane between UE and eNB is confidentiality protected.
3. NAS signaling between UE and MME is integrity and confidentiality protected.

Following are the list of security associations that are user independent.

1. S1-C (signaling) between eNB and MME is integrity and confidentiality protected.
2. S1-U (user data) between eNB and S-GW is integrity and confidentiality protected.
3. X2 (signaling and user data) between two eNBs is integrity and confidentiality protected.
4. Various signaling protocols for core network internal interfaces are integrity and confidentiality protected.

For more information related to security aspects, we refer the reader to 3GPP TS 33.401 [26] and 3GPP TS 33.402 [27].

1.6.12 Charging

Logical charging functions in the EPC network are responsible to collect data for charging and billing purposes. These charging functions are specified in 3GPP TS 32.240 [28], together with the reference points that are used to transfer charging events and consolidate charging information between those functions. The mapping of the logical charging functions to the EPC architecture is described in 3GPP TS 32.251 [29].

1.6.12.1 Charging Principles

EPC nodes provide functions that implement Online Charging and/or Offline Charging mechanisms on bearer level, in particular for the PDP context/IP-CAN bearer, and on Service Data Flow (SDF) level. Online Charging performs traffic supervision in real time, while Offline Charging records the usage of resources (e.g., number of sent/received packages or volume or the duration of voice calls). The collected resource usage data can be used for billing of individual subscribers, inter-operator accounting, or general statistical purposes.

Offline Charging is a process where charging information for network resource usage is collected concurrently. The information is passed from the Charging Trigger Function (CTF) to the billing system. Offline charging does not affect the rendered service in real time.

Online Charging, on the other hand, is a process where charging information for network resource usage is not only collected concurrently during their usage, but also the authorization for using these network resources must be obtained before the actual usage. This authorization is granted by the Online Charging System (OCS) upon request from network functions such as GGSN or P-GW/PCEF. More details of the OCS architecture can be found in 3GPP TS 32.296 [30].

Offline and Online Charging can be performed simultaneously and independently for the same chargeable event, for example, a voice call or packet data transfer.

1.6.12.2 EPC Charging

In case of EPC charging, the following functional entities provide charging data:

- The SGSN, S-GW, and ePDG (see chapter 1.10) record user's access to VPLMN resources. In addition, the SGSN records user's mobility management activities, SMS, and Multimedia Broadcast/Multicast Service (MBMS) usage.
- The P-GW and GGSN record user's access to external networks such as the Internet.
- The MME record SMS usage when the "SMS in MME" feature is supported.

The functional entities in EPC differ with regards to the charging information they are able to supervise and report. So-called Charging Characteristic profiles stored in the HSS specify under which conditions a certain charging event is generated. Three default profiles exist: one for the non-roaming case, one for the local breakout roaming case, and one for the home routed roaming case. Besides these profiles stored in HSS, SGSN/S-GW and GGSN/P-GW/TDF can also use locally configured profiles.

1.6.12.3 Offline Charging

If a subscriber is using EPS network resources, the corresponding charging information is collected by the EPC functions serving the subscriber. SGSN and S-GW capture information regarding the usage of radio network resources as well as data pertaining to mobility management, while GGSN, ePDG, and P-GW collect charging information that relate to the utilization of external data network resources. The subscriber’s usage of resources in the EPC network itself is recorded by each of them. Access to EPC network resources is provided by PDP contexts/IP-CAN bearers that offer the user a logical connection to services with a certain QoS. An APN identifies the service to which access is provided and network resources are consumed. Examples of such services are IMS or Internet access.

The EPC entities collect information pertaining to these PDP context/IP-CAN bearers and the resource utilization that comes along with their usage. EPC charging comprises collecting information about transferred data volume separated for downlink and uplink traffic and categorized by the provided QoS and used protocols, the duration of this usage (i.e., how long the PDP context/IP-CAN bearer is activated), destination, and source address, the APN as well as the location of the UE, that is, the network to which the UE is currently attached to. Regarding the GGSN and P-GW, the accuracy of this location information is limited to the SGSN address, whereas in SGSN and S-GW also the E-UTRAN cell identity is available. The MME plays a role for charging only in case of SMS offline charging when “SMS in MME” feature is supported.

User’s activities are recorded in charging events by the Charging Data Function (CDF) and finally formatted into so-called Charging Data Records (CDR) generated in the entities serving the user. The CDR is transferred to a Charging Gateway Function (CGF) for further processing and from there to the Billing Domain (BD).

For EPC charging the creation of a CDR is triggered by a charging event during PDP context/IP-CAN bearer activation (e.g., in P-GW). Thus, EPS bearer context activation is, for example, a chargeable event. At the same time, volume counters for this context are initialized counting the transferred data volume in uplink and downlink. Upon occurrence of certain chargeable events, such as change of QoS or radio access type, these volume counters are captured together with a timestamp and the applied QoS in the collecting node. Other charging triggers, such as deactivation of the PDP context/IP-CAN bearer or operator-defined limits for time or volume lead finally to the closure of the CDR. If the IP-CAN bearer remains active, a new charging record is created. Besides data volume or elapsed time, other data such as user’s IP address, protocol type, and APN are also stored in a CDR.

Common information to all CDRs is the IMSI, optionally the MSISDN and the IMEI, identifying the subscriber and the UE.

The CGF receives charging records over the Ga reference point, if it is not integrated in the node sending the records. The CGF can also be a separate entity or part of the BD . The BD is responsible to create the final bill toward the subscriber (e.g., on a monthly basis). In all these cases, charging records are transferred from CDF to CGF via the GTP protocol. The transfer of the CDR files from the CGF to the BD uses the Bp reference point.

1.6.12.4 Online Charging

In contrast to offline charging, the solutions for online charging on the SGSN/S-GW and the GGSN/ePDG/P-GW differ significantly.

The SGSN uses legacy non-IP techniques for online charging while EPC online charging at the GGSN, ePDG, and P-GW uses DIAMETER. Online charging may apply on a PDP context/IP-CAN bearer level or on individual service flow level. EPC online charging at the P-GW and TDF is utilizing the Ro interface and the associated DIAMETER Credit-Control application toward the OCS.

The P-GW (i.e., the PCEF as part of the P-GW) collects charging information on the IP-CAN bearer level for each user separated for uplink and downlink. The defined chargeable events correspond to the ones for offline charging for debiting, that is, start and stop of a PDP context/IP-CAN bearer, reaching of time or volume limits, as well as QoS or tariff time changes. If such events occur, the OCS is informed by the P-GW/PCEF and need to authorize the event beforehand, for example, that a new bearer can be established.

Upon establishment of a PDP context/IP-CAN bearer the P-GW/PCEF requests authorization of this event from the OCS. The OCS either authorizes the request when the answer contains a certain volume and/or time quota or denies the context establishment. If authorization is confirmed, a volume counter and/or time measurement is granted in the P-GW/PCEF that debits the counters based on the traffic transmitted via the PDP context/IP-CAN bearer.

If for an established PDP context/IP-CAN bearer the supplied quota is used, the OCS is requested to perform a re-authorization. In this case the used volume count is reported by the P-GW/PCEF to the OCS. When a change of charging conditions has occurred, this information is used to determine how much of the quota has been consumed for the old QoS and during the tariff time.

Generally, the OCS replies to the P-GW/PCEF with quota and instructions how the P-GW/PCEF shall further proceed, for example, continue or terminate an ongoing session.

1.6.12.5 Flow-Based Bearer Charging

To allow for a service-based charging below bearer level, Flow-Based Charging (FBC) has been introduced with 3GPP Release 6 (see 3GPP TS 23.203 [4]). It adds new functionalities to the EPC network and especially extends the capabilities of the P-GW in such a way, that it is now able to identify different Service Data Flows (SDF) within a single PDP context/IP-CAN bearer.

From the point of view of charging information collection, FBC can be considered as an extension of the classical EPC charging by being able to sub-categorize the total data volume of a PDP context/IP-CAN bearer by the different flows the context/bearer contains. For each of these SDFs, an own uplink and downlink volume counter is required.

With FBC it is possible to employ the same charging models for both offline and online charging and irrespective of whether the subscriber is prepaid or postpaid. FBC provides a high degree of flexibility regarding which SDF to consider and how to recognize a SDF by means of charging rules.

The overall FBC concept is realized by three functional elements: the Policy and Charging Rules Function (PCRF), the Application Function (AF), and the Policy and Charging Enforcement Function (PCEF). The overall PCC architecture is described in Section 1.6.7.

The PCRF provides Charging Rules via the Gx reference point to the PCEF and decides which rules have to be applied based on data received from the PCEF, such as user- and bearer-related information, as well as from the AF, which provides session- and media-related information. Such rules contain information how packets belonging to a flow can be identified and how they shall be treated, in particular regarding the applicable QoS. Charging rules need to be specified for each particular service.

The AF provides services to the user for which PDP context/IP-CAN bearer resources are required. The AF can provide additional information to the PCRF for charging rule selection or generation, such as an application identifier, specific user information, and packet filters allowing for the identification of packets belonging to the respective service data flows.

The PCEF is responsible for identifying the user data traffic based on the charging rules received from the PCRF. It is used for both online and offline charging. In case of online charging, the PCEF has also to take care of Credit Control, that is, maintaining the assigned quota as well as communication with the OCS. The PCEF can be located in the P-GW or in case of untrusted WLAN access in the ePDG.

In case of offline charging, charging information collected by the PCEF is written into a PGW-CDR. When flow-based offline charging is activated in a P-GW, classical IP-CAN bearer charging is not done anymore.

In case of online charging, the PCEF creates charging events to request authorization regarding the resource usage being caused by a SDF. On IP-CAN bearer activation a FBC Credit-Control session is started. The OCS replies with a charging event response either granting an initial quota or denying the service request. Following requests toward the OCS may be triggered when the subscriber starts using a new service, that is, when the PCEF encounters one or more new SDFs or when the granted quota is used up. The charging session is closed when the IP-CAN bearer is deactivated or the OCS indicates session termination as a result of the fact that the subscriber has consumed his or her credit.

1.6.12.6 Charging Rules

FBC uses charging rules to identify and charge SDFs. A SDF consists of one or more IP flows that shall be treated and charged as a whole.

A SDF is identified by means of SDF filters. Such a filter may be an IP 5-tuple containing of destination/source IP addresses, destination/source port numbers and protocol (TCP, UDP, SCTP), possibly with wildcards, and also other filters for application protocol or content recognition and form part of a charging rule. The wildcard represents the so-called default SDF, which comprises all traffic. If it is the only defined SDF in a PCEF, the resulting behavior of FBC corresponds to the classical EPC charging on bearer level. Otherwise, the default SDF captures all traffic that does not match at least one of the more specific flow filters.

Charging rules in the PCEF are applied by matching received packets with the SDF filters that are part of these charging rules. Pre-defined charging rules are completely configured, that is, they contain all necessary information to be applied. They may already be installed on the PCEF or provided by the PCRF via the Gx reference point when needed. In contrast, dynamic charging rules are newly generated or completed dynamically using application-specific criteria to identify the SDF. This information is provided by the AF to the PCRF on request over Rx.

Besides SDF filters, charging rules may comprise further information that, for instance, define how the charging process shall take place, that is, whether time- and/or volume-based resource usage information shall be provided, which precedence the charging rule has in case of a rule overlap, whether online or offline charging has to be applied, and identifiers for charging correlation and service identification.

Charging rule provision takes place over the Gx reference point. The provision of rules can either occur following a charging rule request by the PCEF (caused, e.g., by bearer establishment, bearer modification, or QoS change) or unsolicited by the PCRF as the result of new information received from an AF or via notification received from the OCS.

The Rx reference point plays an important role for dynamic charging rule generation and completion and is used for exchanging information between the PCRF and the AF. When the AF becomes aware of new media used by an application, it provides this information, which in particular relates to the different flows and how they can be identified, to the PCRF, which in turn generates or completes dynamic charging rules by adding SDF filters. This procedure may also be initiated by a PCEF requesting charging rules and a PCRF recognizing that specific information is required for dynamic rule provisioning. The PCRF then contacts the corresponding AF to acquire the needed information.

1.6.12.7 MBMS Charging

For MBMS charging, the BM-SC contains an integrated CTF that generates charging events for mobile subscribers receiving services through the MBMS user service and/or for content providers delivering content through the MBMS bearer service. Transactions involving the content provider are recorded per subscriber. Online charging at the BM-SC utilizes the Ro reference point while offline charging at the BM-SC utilizes the Rf reference point.

The MBMS GW collects charging information for each MBMS bearer service that is activated. Following information is reported by the MBMS GW:

- Start of MBMS bearer context: A CDR for the MBMS bearer context is created and the data volume is captured for the context.
- MBMS bearer context termination in the MBMS GW.
- Expiry of an operator configured time or data volume limit per MBMS bearer context. This event closes the MBMS bearer context CDR and a new CDR is opened, if the context is still active.
- Change of charging condition, for example, tariff time change. In this case the current volume count is captured and a new volume count is started.
- Expiry of an operator configured change of charging condition limit per MBMS bearer context closes the MBMS bearer context CDR. A new CDR is opened, if the MBMS bearer context is still active.

More details on MBMS charging can be found in 3GPP TS 32.273 [31].

1.7 IP Multimedia Subsystem

3GPP Rel-5 in 2001 was the first release where the basic architecture and procedures of the IP Multimedia Subsystem (IMS) were specified. While mobile operators were initially reluctant to commercially deploy IMS in their networks to replace existing voice and SMS, with the introduction of LTE as a pure packet-based system without a CS voice component and the general VoLTE profile specified by GSMA in IR.92, IMS is becoming more and more important for mobile operators. Early mobile deployments of IMS were mainly focused on non-voice

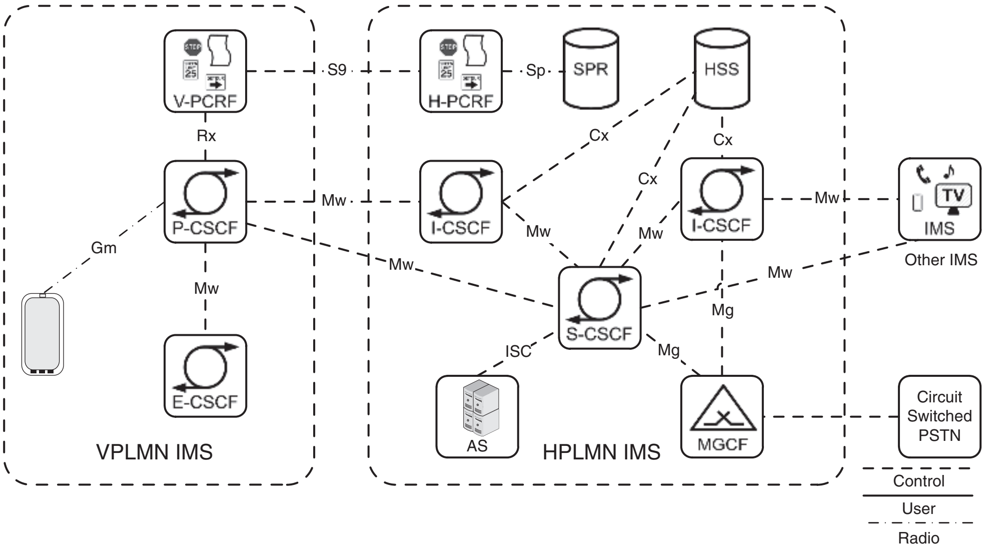


Figure 1.13 IMS roaming architecture

applications (messaging and presence) but nowadays (end of 2014) commercial mobile IMS deployments are planned or even rolled out introducing VoIP and other multimedia services on top of IMS.

In a nutshell IMS can be seen as a general purpose platform for the establishment, modification, and teardown of multimedia sessions between end points to exchange content such as voice, video, messages, and files with the necessary QoS. In principle, IMS can work on any packet-based network providing suitable data rates. One design goal was to provide more flexible communication service creation in an “Internet style” while, on the other hand, being able to handle the limitations of mobile communication systems. For that purpose, the Session Initiation Protocol (SIP) as specified by IETF RFC 3261 [32] was chosen to maintain sessions between UEs or between UEs and servers. The Session Description Protocol (SDP), specified in IETF RFC 4566 [33], transported within SIP message bodies allows end points to exchange information on multimedia content like the used codec. IMS works on the application layer and uses EPS capabilities to setup certain IP bearers with specific QoS between the IMS-capable UE and P-GW. Interworking between IMS and traditional networks such as the 2G/3G CS domain or the fixed public telephone network is enabled by the Media Gateway Control Function/Media Gateway (MGCF/MGW) where the MGCF interworks between SIP and CS protocols like ISDN User Part (ISUP) and the MGW interworks between different transport planes (e.g., between different codecs).

The basic IMS architecture is shown in Figure 1.13.

The main components of IMS are the so-called Call Session Control Functions (CSCF) that can be seen as SIP proxies. Four types of CSCF exist: P-CSCF, S-CSCF, I-CSCF, and E-CSCF.

The Proxy CSCF (P-CSCF) is the outermost SIP entity toward the served UE.

In case of “IMS roaming” the P-CSCF is located in the visited network (VPLMN). To select a proper P-CSCF in the VPLMN, GSMA has specified the well-known “IMS APN” which is resolved by the visited MME into a local P-GW address. The local P-GW selects a P-CSCF in the VPLMN and provides the address to the UE (note that in general P-GW and P-CSCF are always located in the same network, either both in HPLMN or both in VPLMN). Otherwise, the P-CSCF is located in the home network. When a UE registers, it is assigned a P-CSCF as entry point toward the IMS network. The P-CSCF stores information related to that served UE while it remains registered and forwards any SIP message to or from the served UE on the Gm interface. The P-CSCF provides integrity and confidentiality for SIP messages, compression and generation of charging records. It provides also an interface to the PCC and Lawful Interception infrastructure. In PCC terms, the P-CSCF acts as an Application Function (AF) toward the PCRF. Registration in the IMS requires a special application on the UICC, the so-called ISIM (IMS SIM). The ISIM holds information about subscriber identities and keys used in IMS.

The Serving CSCF (S-CSCF) is the central IMS entity located in the home network of the IMS subscriber. It acts as SIP registrar for the UEs, that is, the UE registers its contact address (IP address) at the S-CSCF when it connects to IMS. The S-CSCF retrieves the IMS user profile from the HSS. The user profile is used to authenticate and authorize a user and to execute services the user is subscribed to. Services are usually not executed by the S-CSCF itself but via special Application Servers (AS) that are connected to the S-CSCF through the ISC interface. Examples of Application Servers are Telephony Application Server (TAS) for value-added voice services (e.g., call forwarding) and conferencing, messaging, and announcement servers. Besides routing and address translation, the S-CSCF generates also charging records and interconnects with LI entities.

The Interrogating CSCF (I-CSCF) is the first contact point for SIP messages at the network boundaries. When receiving a SIP message destined for a user the I-CSCF selects an S-CSCF capable to serve that user. The I-CSCF does not store user-related data and it does not stay in the path for subsequent SIP messages. The I-CSCF is used when a UE registers to the network and when a session setup destined for a served user is received in the home IMS.

The Emergency CSCF (E-CSCF) handles emergency calls. Its main purpose is to retrieve location information about the caller and to forward the emergency call to a Public Safety Answering Point (PSAP) or emergency center.

The HSS is the central database located in the home network of a user, containing subscription and location-related information. Examples of these data are user identities (which are, e.g., SIP URIs), security keys used for authentication and authorization on IMS level, address of the S-CSCF serving the user, and the user profile containing a list of subscribed services.

The Media Gateway Control Function (MGCF) converts SIP messages into messages used in circuit-switched networks such as the Public Switched Telephone Network (PSTN) or CS domain of a mobile network. In addition, the MGCF controls Media Gateways (MGW). A MGW converts between the transport layer in a CS network such as TDM (Time Division Multiplex) and RTP (Real Time Transport Protocol) as used in IMS. The MGCF is responsible to allocate and maintain resources on a selected MGW and to provide necessary information received in signaling messages from its peers to the MGW to allow for transport conversion.

1.7.1 Summary of Reference Points and Protocols

Table 1.4 summarizes the reference points and protocols used for IMS.

Table 1.4 Reference points and protocols for IMS

Reference point	Protocols	Specifications
Gm	SIP	TS 24.229 [34]
Mw	SIP	TS 24.229 [34]
Cx	DIAMETER	TS 29.229 [35]
Mg	SIP	TS 24.229 [34], TS 29.163 [36]
ISC	SIP	TS 24.229 [34]

1.8 Voice and SMS in LTE

1.8.1 Voice

LTE/SAE is a pure packet-based system without an in-built circuit-switched voice component such as GSM and UMTS. Thus, the preferred way to provide voice and other multimedia services in LTE/SAE is via the IMS. For a better worldwide acceptability and interoperability of IMS over LTE, GSMA defined an IMS profile for voice and SMS in their IR.92 recommendation. For example, GSMA IR.92 has specified a well-known APN, the so-called IMS APN, to enable IMS roaming. This profile is referred to as "Voice over LTE (VoLTE)."

The LTE/SAE network indicates to the UE that VoIP in LTE is supported; more precisely that the TA(s) the UE is currently camping on provide sufficient QoS and coverage for VoIP. Therefore, an UE that has been provisioned as IMS VoIP capable can start an IMS voice session at its current location based on the received network indication.

However, during the early stages of LTE roll out, actual LTE coverage is spotty. A UE can start an IMS voice session in LTE/SAE but need to continue the voice call when it moves out of LTE coverage. To allow continuation of voice calls in 2G/3G CS domains (usually providing country wide coverage), the so-called Single Radio Voice Call Continuity (SRVCC) feature was introduced (see 3GPP TS 23.216 [37]). It extends the voice services coverage area and provides a good voice experience in the early stages of LTE. SRVCC requires bearer level handover between EPC and Circuit-Switched Core and switching of access legs between IMS and CS domain. The non-voice components (e.g., video streaming, file transfer) can be moved from LTE to 2G/3G PS via general handover procedures. Although 2G/3G CS coverage may allow voice calls all over a country, it is beneficial to start voice and multimedia calls in LTE/SAE whenever possible and even return to LTE/SAE as soon as possible. This is because a broadband network such as LTE/SAE together with a flexible service platform like IMS at the application layer can provide much more value-added services to the end user than the legacy 2G/3G CS systems. Benefit for the operator is that the user can be charged for these value-added services in addition to the normal voice service.

Some LTE/SAE networks may not support IMS-based voice service. In this case, voice service can only be provided over the CS domain of the existing 2G/3G networks. In order to accomplish that, the CS Fallback (CSFB) feature was defined by 3GPP (see TS 23.272 [38]). CSFB allows the UE to switch from LTE to 2G/3G network in a controlled manner to setup a CS voice call. Data connections can be handed over to the 2G/3G PS domain or (if not possible because of lack of resources) suspended in LTE/SAE. After the voice call has ended, the UE returns to LTE and data connections are resumed.

1.8.2 Short Message Service

SMS can be provided in LTE natively over IP (called "SMS over IP") via IMS (see 3GPP TS 23.204 [39]). There is no special requirement being placed on LTE/SAE for this feature, it is just used to provide IP connectivity. The precondition is that the UE must have successfully performed an IMS registration, is supporting the "SMS over IP" feature, and has been configured to use the feature. If this is the case, a SMS is encoded by the network or UE in a special SIP message and send to the peer.

For the scenario in which the LTE operator also runs a 2G or 3G network, SMS over LTE/SAE reusing CSFB mechanisms was specified. In this case, both UE and EPC must support the SMS-specific procedures as defined for CSFB. Unlike CSFB voice, SMS over EPC does not require the UE to tune to 2G/3G radio for SMS. The EPC network provides a kind of tunnel between the UE and the CS Core for SMS delivery. The SMS is forwarded between MME and MSC. The MSC is connected to the legacy SMS infrastructure (Short Message Service Center) in the usual way. The UE is attached in both the EPS and CS networks for SMS delivery.

A SMS can also directly be delivered by the SMS infrastructure to the MME and vice versa via a DIAMETER-based interface, bypassing the MSC. This was introduced to support SMS delivery to PS only UEs. PS only means that the UE does not have any CS subscription data in the HSS. This feature is known as "SMS in MME." SMS in MME was introduced mainly to address requirements from operators who do not deploy a CS core. SMS over IP (i.e., SMS over IMS) could be one solution for these operators. However, SMS over IP requires an IMS/SIP client in the UE, which is too heavy-weight for some types of devices such as smart meters or dongles. Furthermore, inbound roamers whose home operators do not support IMS cannot be offered SMS over IP in the VPLMN.

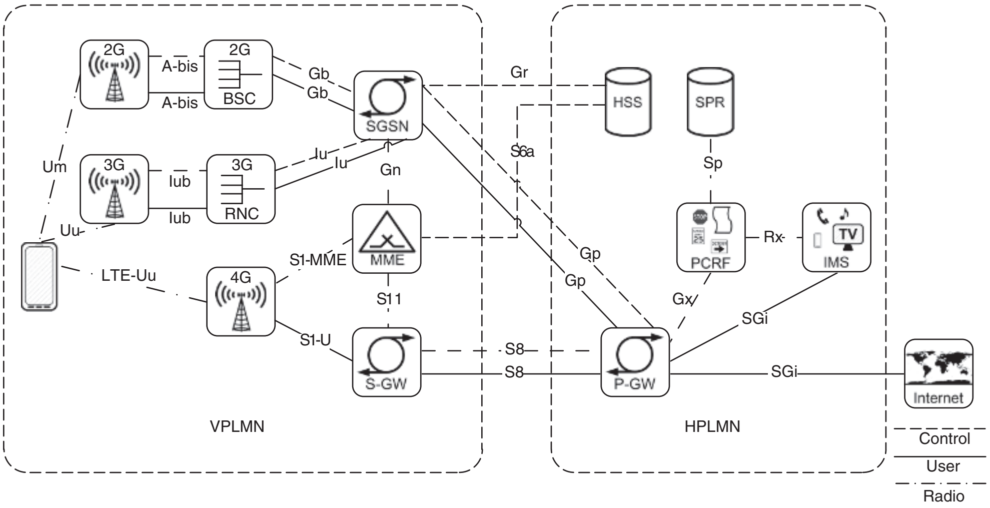


Figure 1.14 Interworking with Gn/Gp-SGSN

1.9 Interworking with 2G/3G Networks

1.9.1 Overview

3GPP has specified one way of interworking between LTE/SAE and existing 2G/3G networks by upgrading 2G/3G SGSN nodes to so-called S4-SGSN nodes. A S4-SGSN acts toward the EPC like a MME, that is, it has control plane interfaces to MME (S3 interface) and S-GW (S4 interface) and user plane interfaces to S-GW (S4), while connecting to 2G/3G radio access networks via the existing GPRS and UMTS interfaces. The S4 interface with the S-GW is used to manage bearers and possibly forward user plane traffic.

1.9.2 Interworking with Legacy Networks

If a mobile operator does not want to upgrade an existing SGSN (the so-called Gn/Gp-SGSN named after the interface between SGSN and GGSN in 2G/3G networks, see 3GPP TS 29.060 [8]) to a S4-SGSN, it is still possible to interwork these legacy SGSNs with the EPC. Such scenarios are important for a smooth introduction of LTE/SAE while leaving existing network elements untouched. The basic idea is to provide Gn/Gp interfaces at MME and P-GW, that is, from Gn/Gp-SGSN point of view the MME behaves like an SGSN and the P-GW behaves like a GGSN. Consequently, MME and P-GW must implement the protocols supported by Gn/Gp-SGSN, and additionally some changes in the mobility and session management procedures are required. It should be noted that this kind of interworking is only possible if GTP-based S5/S8 interfaces are used, as for PMIP-based S5/S8 handover between Gn/Gp-SGSN and MME/S-GW is not supported.

Figure 1.14 shows the Gn/Gp interworking architecture in the roaming case. The user plane is running from GERAN/UTRAN via the Gn/Gp-SGSN in the VPLMN to the P-GW in the HPLMN. The legacy Gr interface toward the HSS is MAP (Mobile Application Part) based.

The non-roaming architecture is similar to the roaming architecture with the difference that S5 is used between S-GW and P-GW, and Gn between Gn/Gp-SGSN and P-GW. If the Gn/Gp-SGSN supports Direct Tunnel (as defined from 3GPP Release 7 onwards) a direct user plane connection between UTRAN and P-GW is possible.

1.9.3 Functional Description

1.9.3.1 UE Aspects

In order to interwork with legacy 3GPP access technologies, an UE needs to be multiradio capable. It needs to support idle mode and connected mode mobility between E-UTRAN, UTRAN, and GERAN. There are two modes of interworking defined: single-radio operation and dual-radio operation.

In case of single-radio operation, the network controls the usage of radio transmitter and receiver in the UE in a way such that only one radio is operating at any time. This is optimized interworking and allows UE implementations where only one pair of physical radio transmitter and receiver is implemented. With dual-radio operation, multiple radio transmitters and receivers are operating simultaneously. Single-radio operation is an important mode because different access networks operate in different frequencies that might be close to each

Table 1.5 Reference points and protocols for interworking with 2G/3G networks

Reference point	Protocols	Specifications
Gn	GTP (v0 and v1)	TS 29.060 [8]
Gp	GTP(v0 and v1)	TS 29.060 [8]
S4	GTPv2-C	TS 29.274 [19]

other. So, dual-radio operation can cause high interference within the UE. Furthermore, it can consume additional power and reduce the overall performance and is more costly in terms of implementation.

1.9.3.2 E-UTRAN Aspects

The main additional function of the eNB is the support of mobility to and from UTRAN and GERAN. From eNB perspective, it needs to provide similar functionality for mobility to UTRAN and GERAN, for example, neighboring GERAN and UTRAN cells from the same network need to be configured in the eNB. Handover to and from UTRAN/GERAN is performed via the MME.

1.9.3.3 EPC Aspects

The S-GW acts as a mobility anchor for all 3GPP access systems. It functions as a GGSN toward SGSN. Although GGSN functions are mainly performed by the P-GW, this is not visible to the SGSN. The S-GW is controlled by the MME or SGSN, depending on the access network where the UE camps on (E-UTRAN or GERAN/UTRAN).

In order to support interworking, the MME needs to support signaling procedures with the SGSN. This is similar to the handover procedure supported by MME when MME relocation occurs. For interworking with legacy Gn/Gp-SGSN, the MME behaves like an SGSN.

1.9.3.4 Summary of Reference Points and Protocols

Table 1.5 summarizes the additional reference points and protocols used for interworking with 2G/3G networks.

1.10 Interworking with Non-3GPP Access Networks

Non-3GPP access networks refer to networks not using 3GPP access technologies, that is, not using GERAN, UTRAN, and E-UTRAN access. A typical and probably the most important example for a non-3GPP access network is a WLAN at a public hotspot, at a campus, or at home. Such a WLAN can, for example, use IEEE 802.11b/g/n radio technology. Interworking with non-3GPP access networks means providing access to the EPC and its services via a non-3GPP access technology and providing mobility between 3GPP and non-3GPP access (e.g., handover of connections from E-UTRAN to WLAN and vice versa). Interworking to non-3GPP access networks is described in detail in 3GPP TS 23.402 [3].

Basic principle of the non-3GPP access interworking architecture is that the P-GW is the IP mobility anchor point, that is, the P-GW is considered the "point of attachment" to external IP networks. In case of mobility between 3GPP and non-3GPP networks, the P-GW does not change.

From 3GPP perspective, non-3GPP access networks can be considered as either trusted or untrusted. It is operator's decision whether an access network is seen as trusted or untrusted. It does not depend on the access technology, but on operator policies and the business relationship between network operator and provider of a non-3GPP access network like a WLAN hotspot. The particular business relationship depends especially on the level of security provided by the access network and whether this is sufficient to allow access to the EPC. A non-3GPP access network can be seen as trusted for one operator while untrusted for another one.

Trusted non-3GPP access networks can be directly connected to the 3GPP core network. If an untrusted non-3GPP access network is used, the UE is connected to a kind of security gateway called Evolved Packet Data Gateway (ePDG) via an IPSec tunnel using the SWu reference point. The ePDG is located in the EPC.

Figure 1.15 gives an overview of the architecture with network-based mobility. The S2a interface connects mobile terminals to the core network over trusted non-3GPP access networks and the S2b interface is used for untrusted access networks. The functionality of S2a and S2b is quite similar and both can be implemented using either GTP or PMIP core network signaling. According to the specification, S2a can also be used in MIP Foreign Agent mode, however, it is not expected that this alternative will be widely deployed. When PMIP is used over S2a or S2b the trusted non-3GPP access network and the ePDG provide the Mobile Access Gateway (MAG) functionality required for PMIP, while the P-GW includes the Local Mobility Anchor (LMA) functionality. The interface between PCRF and ePDG is not specified. In roaming scenarios, that is, when the non-3GPP access network is connected to a VPLMN, the ePDG is in the VPLMN. In roaming scenarios, the 3GPP AAA server is located in the HPLMN and a 3GPP AAA proxy is located in the VPLMN.

Figure 1.16 shows the architecture when client-based (i.e., UE based) mobility based on Dual-Stack Mobile IP (DSMIP) is used. For simplicity not all interfaces are shown. Except for S2a, S2b, and S2c, interfaces in Figures 1.15 and 1.16 are the same. DSMIP between UE and LMA located in the P-GW runs over the S2c interface. In order to avoid user plane tunneling overhead over 3GPP access networks, the 3GPP access network is always the home link in terms of Mobile IP. Therefore, only S2c signaling is used over 3GPP access. The UE provides the DSMIP client function while the P-GW includes the DSMIP Home Agent function (LMA). In roaming scenarios, the 3GPP AAA server is again located in the HPLMN and a 3GPP AAA proxy is located in the VPLMN.

The basic non-3GPP interworking specification in 3GPP TS 23.402 [3] creates a general framework how to access EPC from a non-3GPP access network without putting any requirements on the WLAN. During the past couple of years, more and more 3GPP operators deployed WLANs and it has been recognized that without additional specifications these WLANs cannot be used as trusted non-3GPP access networks. The main technical issues are that in current WLANs the UE has no means to send handover and APN-related information to the network, and the support of multiple PDN connections is missing.

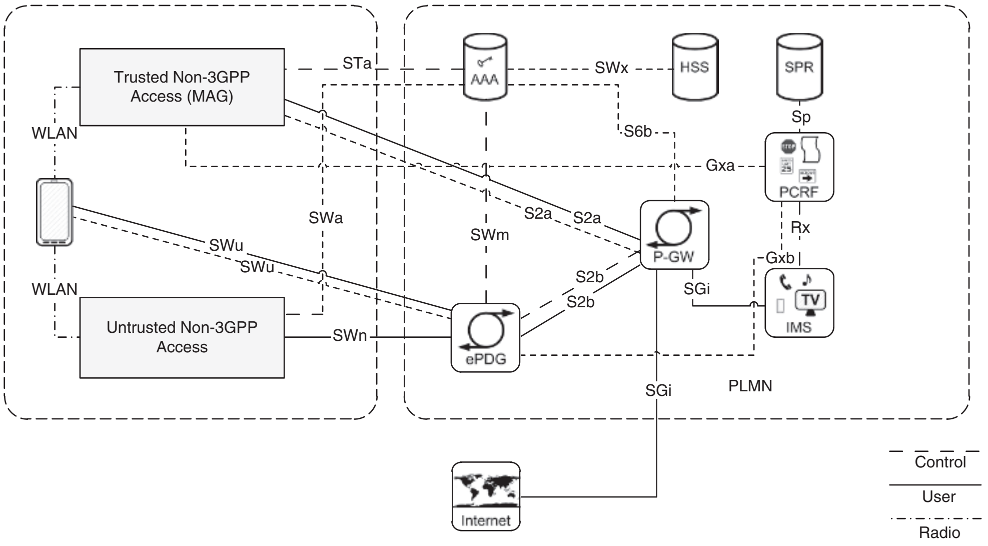


Figure 1.15 Non-3GPP interworking architecture with network-based mobility

In order to enable easy deployment of Trusted WLAN Access Networks (TWAN) in Release 11, a solution that can work without UE impacts was developed. This solution does not create any architecture level changes, just requires some enhancements to the S2a, STa, and SWx

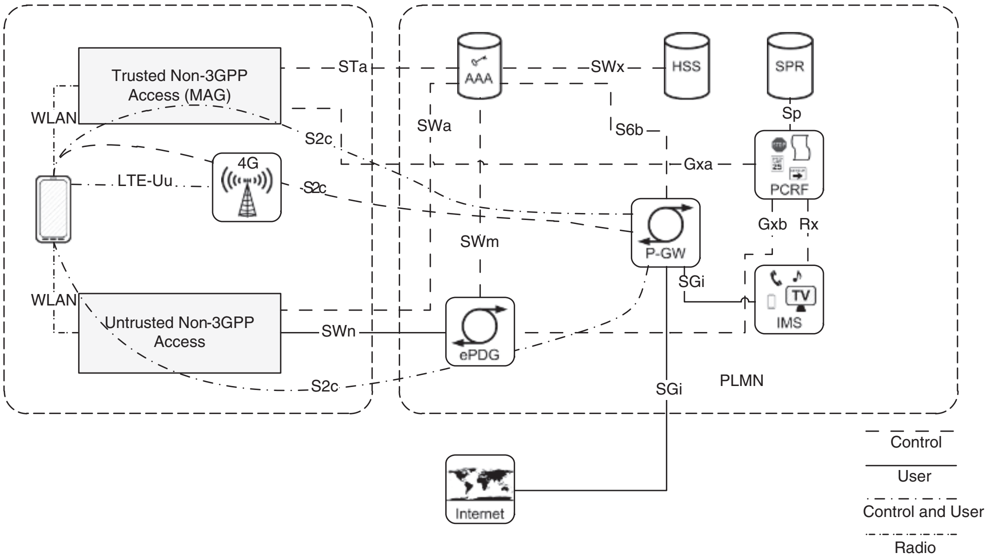


Figure 1.16 Non-3GPP interworking architecture with client-based mobility

interfaces. The AAA interfaces (STa and SWx) are enhanced to carry trusted WLAN authorization related parameters and additional subscriber data for trusted WLAN access. In addition, also GTPv2 support for Trusted WLAN Access over S2a was specified since it was recognized that for operators having only GTP-based interfaces, the deployment of GTP-based S2a might be easier than deploying PMIP-based S2a.

Owing to the lack of handover and APN indication, the solution in Release 11 does not support the following features:

- Handover between TWAN and 3GPP access with IP address preservation.
- Connectivity to a non-default APN (as it is not signaled by the UE).
- UE-initiated connectivity to additional PDNs.

Moreover, simultaneous access to EPC and non-seamless (without preserving UE's IP address) WLAN offload is not supported.

To overcome these limitations, 3GPP specified a new protocol between UE and an entity in the Trusted WLAN Access Network called Trusted WLAN Access Gateway (TWAG) in Release 12. The protocol is called WLAN Control Protocol (WLCP). WLCP signaling is transported over UDP/IP and enables management of PDN connectivity over TWAN. It provides the following functions:

- Establishment, termination, and handover of PDN connections.
- Request the release of a PDN connection by the UE or notify the UE of the connection release.
- IP address assignment.

1.10.1 Summary of Reference Points and Protocols

Table 1.6 summarizes the additional reference points and protocols used for interworking with non-3GPP access networks.

Table 1.6 Reference points and protocols for interworking with non-3GPP access networks

Reference point	Protocols	Specifications
SWa	DIAMETER	TS 29.273 [40]
SWm	DIAMETER	TS 29.273 [40]
SWu	IPSec/IKEv2	TS 24.302 [41]
SWw	WLCP	TS 24.244 [42]
	EAP	IETF RFC 3748 [43]
	EAP-AKA′	IETF RFC 5448 [44]
SWx	DIAMETER	TS 29.273 [40]
STa	DIAMETER	TS 29.273 [40]
S2a/S2b	GTPv2-C/GTPv1-U	TS 29.274 [19]/TS 29.281 [22]
	PMIPv6	TS 29.275 [23]
S2c	DSMIPv6	TS 24.303 [45]

1.11 Network Sharing

Various mechanisms exist for operators to share network deployment costs and increase country wide radio coverage. Increased coverage is an important use case in the beginning of Public Safety network rollouts. These mechanisms allow for sharing of equipment sites, radio network elements, spectrum/frequencies, and core network nodes.

The network sharing solution described in this section is a feature for a shared radio spectrum scenario (i.e., different operators share the same spectrum) where a single cell is broadcasting multiple PLMN ID(s). This feature was originally specified as an option for UMTS in Release 6 and was adopted for EPS in Release 8. In Releases 10 and 11 it was also specified for GERAN access. The feature is defined in 3GPP TS 23.251 [46] and is sometimes referred to as Multi Operator Core Network (MOCN) and/or Gateway Core Network (GWCN) depending on the core network configuration. In the MOCN configuration, operators share the radio access network (eNB) but operate separated core networks (MME, S-GW, P-GW, and HSS) while in the GWCN configuration also the MME is shared (but not S-GW, P-GW, and HSS). It is obvious that sharing RAN nodes provides much more benefits to the operators in terms of cost reduction than just sharing a few CN nodes.

In both configurations – MOCN and GWCN – the eNBs in the radio access network are shared in the same way and the UE behavior is also the same. While in the MOCN configuration the shared eNB connects to core networks of different operators, GWCN allows the MME to be shared and the MME connects to GW(s) and HSS of different operators. Sharing of an eNB or MME means basically that several operators (at least two) can use the same HW and SW resources but can configure parts of the shared node individually (depending on the functionality provided by the eNB/MME vendor). Figure 1.17 shows the MOCN and GWCN configurations.

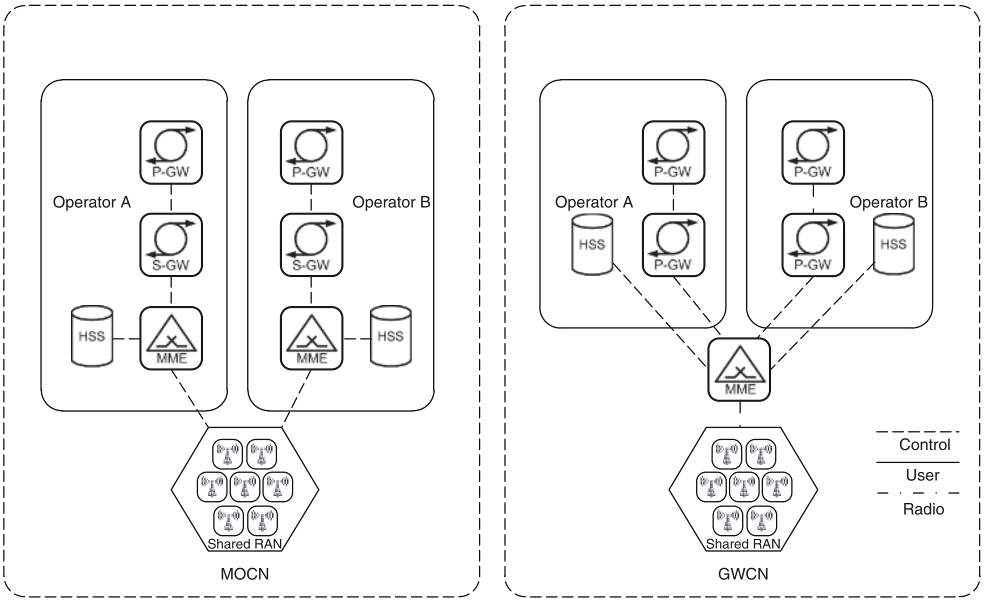


Figure 1.17 MOCN and GWCN configurations

1.11.1 UE-Based Network Selection

Each LTE cell in a shared area broadcasts multiple PLMN Identities (in maximum five) in a list where the first listed identity is the so-called primary PLMN. The broadcasted TA Code is common to all PLMNs. A UE decodes the broadcast system information and takes the information concerning all available core network operators into account in network and cell (re-)selection procedures.

When a UE performs an initial access to a shared network it selects one of the advertised networks (usually its home PLMN) and indicates the selected PLMN identity to the eNB.

1.11.2 RAN-Based Network Selection

The UE informs the eNB about the network identity of the chosen core network. On the basis of this information the eNB routes UE's initial access request to one of the selected operator's MMEs.

Once the UE gets admitted, the MME provides a temporary identity to the UE, which contains sufficient information to enable the eNB to direct subsequent messages to the same MME.

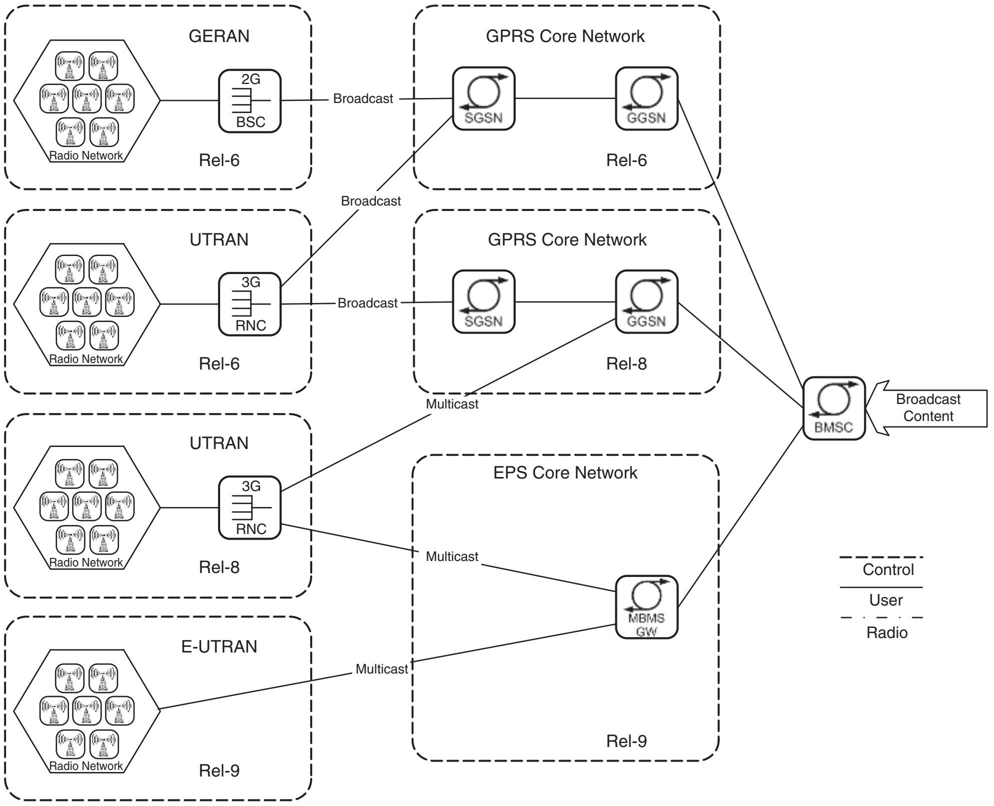


Figure 1.18 Evolution of MBMS user plane

1.12 Multimedia Broadcast Multicast Service

1.12.1 Principles

Multimedia Broadcast Multicast Service (MBMS) is a unidirectional point-to-multipoint service that allows simultaneous transmission of data from a single source (the content provider) to a group of users located in a specific area. MBMS provides means to send data to a potential huge number of users in an efficient manner. For that purpose MBMS uses radio multicast channels on the air interface and IP multicast techniques in the core network. Services that may use MBMS transport capabilities can be divided into two types:

- Streaming services with a continuous data flow
- Download and play services.

Service examples include the following:

- Video distribution, Mobile TV, and Mobile Gaming via streaming or download
- Traffic announcements
- Content distribution such as downloading files, HTML pages, video, audio, or a combination of those and software updates to the device.

MBMS works in broadcast and multicast mode but only broadcast mode is supported in LTE. In broadcast mode, a data stream is transmitted from a single source to multiple UEs in the associated broadcast service area. In multicast mode (only relevant in case of 2G/3G), a data stream is transmitted from a single source to UE(s) that belong to a multicast group in the service area. In multicast mode, only users that are subscribed to the specific multicast service and have joined the multicast group associated with this service can receive data. In broadcast mode, users are not required to join or activate the service in order to receive the data.

MBMS was first specified for GPRS/UMTS in Release 6. To support flat architectures and bypass the SGSN, Rel-8 introduced IP multicast as an option for the distribution of MBMS payloads within the backbone network between GGSN and RNC. Each RNC wishing to receive MBMS data needs to join a corresponding multicast group. The support for MBMS in LTE/SAE was not included in Release 8 because of low interest in the industry.

Release 9 showed increasing interest on MBMS for LTE/SAE and specification work started (see 3GPP TS 23.246 [47]). However, the target design for Release 9 EPS functionality was limited to enable Mobile TV and scheduled file downloads. Therefore, MBMS for EPS (called Evolved MBMS or shortly eMBMS) supports only MBMS broadcast mode.

Figure 1.18 shows the evolution of MBMS from Release 6 GPRS/UMTS to Release 8 UTRAN and Release 9 E-UTRAN.

The Release 9 MBMS broadcast mode function in E-UTRAN differs from GERAN/UTRAN in that E-UTRAN does not support counting of active users in a cell. As a consequence, data are broadcasted to predefined areas regardless whether there are any UEs in this area.

In E-UTRAN IP multicast is the only way for eNodeB(s) to receive MBMS data streams (see Figure 1.18). In UTRAN the RNC may accept or reject IP multicast distribution and the SGSN can establish normal MBMS point-to-point connections to all related RNC(s).

Evolved MBMS refers to the MBMS feature for EPS (specification started in 3GPP Release 9). New functional elements were introduced with eMBMS: The MBMS Gateway (MBMS

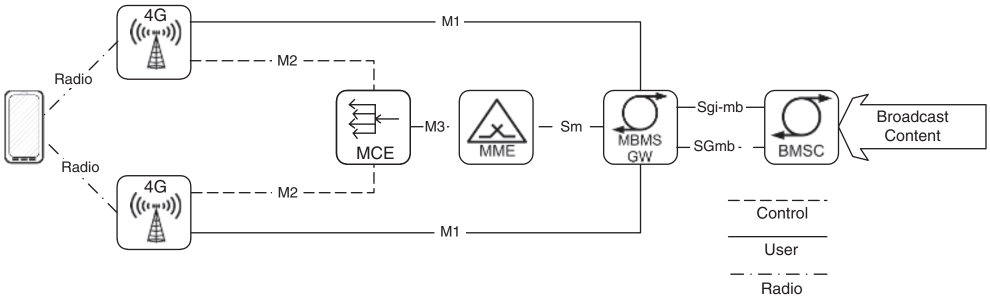


Figure 1.19 Evolved MBMS architecture

GW) replacing the 2G/3G GGSN and the Multicell Coordination Entity (MCE) used in E-UTRAN for uniform MBMS radio resource allocation and control of a group of cells.

Figure 1.19 depicts the eMBMS architecture including these new functional entities.

1.12.2 Description of Functional Entities

1.12.2.1 Broadcast Multicast Service Center (BM-SC)

The Broadcast Multicast Service Center (BM-SC) includes functions for MBMS user service provisioning and delivery. It is the entry point for the content provider, used to authorize and initiate MBMS bearer services within the PLMN via SGmb interface and to schedule and deliver data transmissions via SGi-mb. The BM-SC authenticates, authorizes, and charges access requests from the content provider. The interface between BM-SC and content provider is not specified by 3GPP (except for Public Safety group calls, see chapter 5).

1.12.2.2 MBMS Gateway (MBMS GW)

The MBMS Gateway delivers packets to eNodeBs in configured MBMS service areas via M1 interface and provides MBMS session control signaling (session start/stop/update) toward E-UTRAN (to the MCE) via the MME on Sm interface. The MBMS GW consists of a control and user plane part (MBMS CP and MBMS UP). The MBMS GW may be stand-alone or colocated with other network elements such as the BM-SC, S-GW, or P-GW. The Sn interface between MBMS GW and S4-SGSN is not shown in Figure 1.19. It provides control plane signaling similar to Sm and is used to forward MBMS data in point-to-point mode using GTP.

1.12.2.3 Multicell/Multicast Coordination Entity (MCE)

The Multicell/multicast Coordination Entity (MCE) provides functions for MBMS admission control and MBMS radio resource allocation. It interfaces with the eNodeBs in E-UTRAN via M2 interface. The MCE can be a stand-alone entity or colocated with an eNodeB, in which case only cells controlled by that particular eNodeB can form a MBSFN area (see next sections). In principle the MCE can also be colocated with the MBMS GW.

Table 1.7 Reference points and protocols for MBMS

Reference point	Protocols	Specifications
M1	GTPv1-U	TS 29.281 [22], TS 36.445 [48]
M2	M2AP	TS 36.443 [49]
M3	M3AP	TS 36.444 [50]
Mz (only for GPRS/UMTS)	DIAMETER	TS 29.061 [10]
Sm	GTPv2-C	TS 29.274 [19]
SGmb	DIAMETER	TS 29.061 [10]
SGi-mb	IP unicast or multicast	TS 29.061 [10]

1.12.2.4 MME supporting MBMS

The MME is enhanced for MBMS to support MBMS session control signaling via M3 interface to the MCE. Not shown in the Figure 1.19 is the Mz interface between a BM-SC in HPLMN and a BM-SC in VPLMN (roaming case). Mz is currently supported for GPRS and UMTS only, but not for LTE/SAE.

1.12.2.5 UE supporting MBMS

The MBMS-capable UE needs to support additional functions for the activation and deactivation of the MBMS bearer service and special MBMS security functions (e.g., support of key distribution via the MICKEY protocol).

1.12.2.6 Summary of Reference Points and Protocols

Table 1.7 summarizes the reference points used for MBMS.

1.12.3 MBMS Enhancements

In Release 10, MBMS was enhanced so that the network is capable to manage individual MBMS services depending on the number of users interested in a service. This enables prioritization of different MBMS services depending on their relative priority when there is resource shortage. In addition, a MBMS counting function was introduced to allow counting UE(s) in connected mode, either receiving a particular MBMS service or just interested in receiving a service. Note that only Release 10 devices in connected mode are counted. Release 10 devices in idle mode and Release 9 or older devices are not counted. The MBMS counting function is controlled by the MCE and allows the MCE to enable or disable MBSFN transmission for the service. In support of these new MCE functions, new Release 10 M2 interface procedures were introduced. These procedures support, for example, suspend and resume of a MBMS service, send a MBMS counting request, and obtain MBMS counting results. The prioritization of different MBMS services is also done by the MCE because it is responsible for controlling the allocation of radio resources for MBSFN transmission. So the MCE can pre-empt radio resources used by an ongoing MBMS service according to the Allocation and Retention Priority (ARP) of different MBMS radio bearers.

In LTE Release 11, MBMS was again enhanced to ensure MBMS service continuity in a multicarrier network deployment. MBMS services may be deployed on different carrier frequencies over different geographic areas. Release 11 enhancements allow the network to signal assistance information to MBMS-capable devices that provide information related to the actual MBMS deployment such as carrier frequencies and service area identities. In Release 11, a MBMS-capable device can indicate its interest in MBMS services by indicating the carrier frequencies associated with the MBMS services of interest and the priority between MBMS and unicast service. The network uses this indication for mobility management decisions so that the device is always able to use its receiver at the appropriate carrier frequency layer, thus ensuring continuity of MBMS services. In idle mode, a MBMS-capable device can prioritize a particular carrier frequency during cell reselection depending on the availability of MBMS services on that carrier frequency. To ensure MBMS service continuity in connected mode, the MBMS interest indication received from the device is signaled to the target cell as part of the handover preparation procedure.

1.12.4 MBSFN and MBMS Radio Channels

For the MBMS broadcast mode, E-UTRAN supports the so-called Multimedia Broadcast Single Frequency Network (MBSFN) feature where cells of an MBSFN area are synchronized and produce identical transmissions. MBSFN areas can be predefined. The MCE is in charge of uniform radio resource allocation and synchronized data delivery. The resulting signal will appear to a UE as just one transmission over a time-dispersive radio channel. Multiple cells can belong to a MBSFN area and every cell can be part of up to eight areas. Up to 256 different areas can be defined. It is also possible that certain cells in or at the edge of a MBSFN area do not support MBMS transmission, thus do not belong to the MBSFN area, but transmit other data with low power not to interfere with the MBMS signal. Such cells are referred to as "reserved cells." Figure 1.20 shows a MBMS Single Frequency Network configuration.

MBMS in E-UTRAN requires also new logical, transport, and physical channels. Two logical channels are related to MBMS.

The Multicast Traffic Channel (MTCH) carries data of a certain MBMS service. MBMS services in a MBSFN area may use multiple MTCH. As there is no feedback in the uplink, MTCH uses unacknowledged mode for data transmission. The Multicast Control Channel (MCCH) provides control information to receive MBMS services. There is one MCCH per MBSFN area. One or several MTCH and one MCCH are multiplexed at the MAC layer onto the Multicast Channel (MCH), which is multiplexed to the Physical Multicast Channel (PMCH). The MCCH provides information like the subframe allocation and modulation/coding scheme of each MCH. MCCH can also be used in unacknowledged mode. A notification mechanism is used to announce MCCH changes to the UE. Changes to the MCCH that are not announced can be detected by monitoring the MCCH at each modification period.

The transport format is determined by the MCE and signaled via the MCCH to the UE. During one MCH Scheduling Period (MSP), the different MTCHs and optionally the one MCCH are multiplexed on the MCH. The MCH Scheduling Information (MSI) is provided by the eNodeB at the beginning of the MSP to indicate which subframes are used by each MTCH during the MSP.

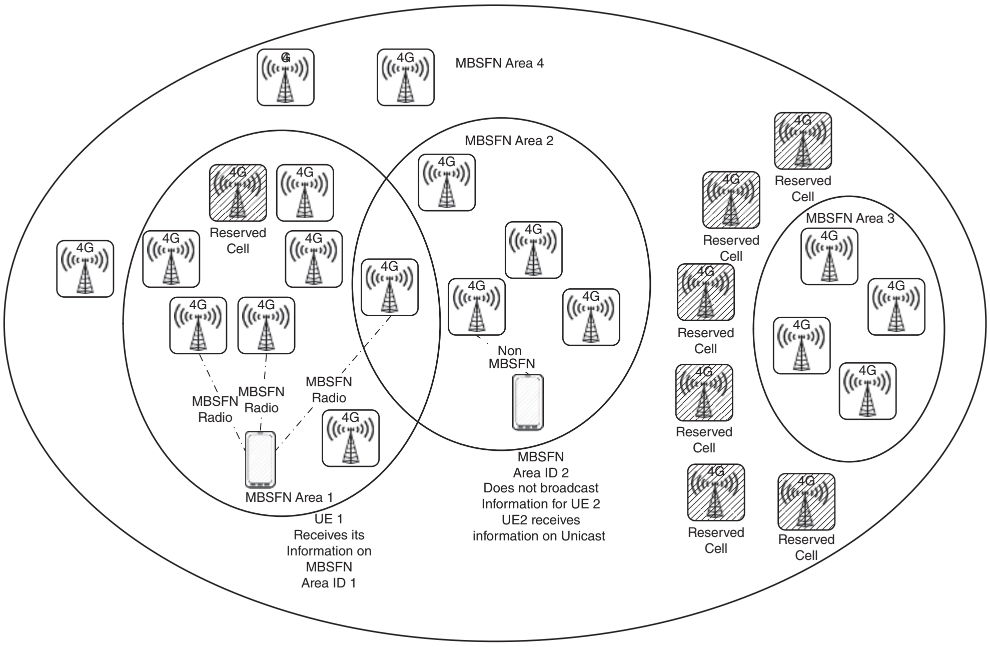

MBSFN Area 4
4G
MBSFN Area 2
MBSFN Area 1
MBSFN Area 3
Reserved Cell
MBSFN Radio
MBSFN Radio
MBSFN Radio
Non MBSFN
UE 1 Receives its Information on MBSFN Area ID 1
MBSFN Area ID 2 Does not broadcast Information for UE 2 UE2 receives information on Unicast

Figure 1.20 MBMS single frequency network

System information broadcast messages carry information on common and shared channels in E-UTRAN and provide also information related to MBMS transmission. They indicate which radio frames contain subframes that can be used for MBMS.

The role of the Broadcast Control Channel (BCCH) is to indicate MCCH-related resource information to the UE (for each MCCH in the cell independently). This information could be the scheduling of the MCCH for multicell transmission on the MCH, the MCCH modification period, the repetition period radio frame offset, and the subframe allocation.

1.13 Terms and Definitions

1.13.1 Roaming

By roaming a subscriber having a contract with network operator A can use network resources of another operator B. Operator A’s network is called the Home PLMN (HPLMN) of the subscriber while operator B’s network is called the Visited PLMN (VPLMN). Usually operator B’s network is deployed in a country different to operator A’s network. However, it is also possible that both networks operate in the same country. In the latter case we speak of “national roaming,” otherwise we speak of “international roaming.” National roaming allows an operator to increase radio coverage by using another operator’s network while the latter operator benefits of additional roaming fees. As an example, Public Safety networks based on the LTE standard can increase their nationwide coverage by signing national roaming agreements with commercial LTE network operators in this country (another way to accomplish this is by sharing radio base stations as described in Section 1.11).

Roaming requires proper agreements between operators (so-called roaming agreements). These agreements clarify how operator A's subscriber can use operator B's network, which services the subscriber can use, and what roaming fees have to be paid. As a prerequisite for roaming, a UE that roams into a foreign network (a roaming-in UE) must support the radio technology and frequencies of this network and the user must have a subscription for the provided radio technology. As an example: If a UE is LTE capable and roams into a foreign LTE network, it cannot register with this network as long as the user has no valid LTE subscription data in his home network. Registration in a foreign network requires that the UE is provisioned with a list of possible/preferred roaming networks. This provisioning process is done by the home operator. On the basis of the list of roaming partners the UE can select the preferred network in a certain country. During the registration process, the HPLMN has to provide authentication and subscription data to the VPLMN via a signaling connection. Although data traffic can be routed by the VPLMN directly to the desired destination, it is common practice that user data are routed from the VPLMN to the HPLMN before traveling toward their final destination (e.g., to a Web server in the Internet). This allows the home operator to apply individual charging rules to the subscriber. For use of visited network resources the subscriber has to pay additional roaming fees on top of the usual service fees. Thus, the VPLMN has to report resource usage of HPLMN's subscribers to the home network (e.g., on a monthly basis).

1.13.2 Circuit-Switched and Packet-Switched Networks

In CS networks a connection (e.g., a voice call) can use a dedicated transmission channel with a constant data rate. This transmission channel can only be used by this particular connection, irrespective of whether data are transmitted or not. Examples for CS networks are the PSTN and networks based on the GSM standard.

In PS networks data that have to be transmitted are separated into data packets and each data packet is transmitted independently from the source to the destination. In principle different data packets can be routed via different paths. Transmission can be done in a connection oriented or connection less manner. Examples for PS networks are IP networks, for example, the Internet. While networks based on the UMTS standard have both a CS and PS component (called CS and PS domains), the new LTE standard consists only of a PS domain.

1.13.3 Access Stratum and Non-Access Stratum

The Latin word "Stratum" means layer and was chosen to avoid confusion with other layers such as the Open Systems Interconnection (OSI) layers.

The Access Stratum (AS) layer consists of all functions that are directly related to the radio access network and the control of connections between end user device and radio network. Protocols on AS layer run between the device and the radio base station in order to establish and maintain radio channels.

The NAS layer, on the other hand, is on top of the AS layer and consists of functions that are related to call control, mobility, and session management. Protocols on the NAS layer are exchanged between the device and the core network, that is, the NAS layer is transparent to

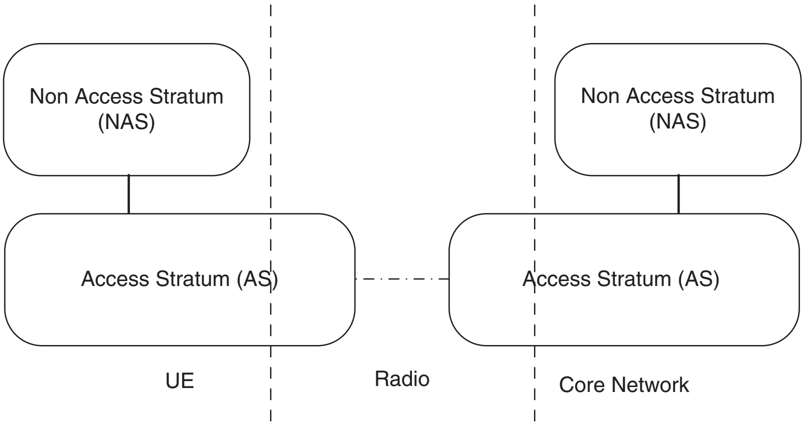


Figure 1.21 AS and NAS layer

the radio access network. The NAS layer and AS layer in the device and core network are able to communicate with each other. This allows the NAS layer to trigger establishment or termination of radio bearers that are used to exchange signaling or user data.

Figure 1.21 provides a simplified overview of the two layers.

References

[1] 3GPP TR 23.882: "3GPP System Architecture Evolution: Report on Technical Options and Conclusions".
[2] 3GPP TS 23.401: "GPRS Enhancements for E-UTRAN Access".
[3] 3GPP TS 23.402: "Architecture Enhancements for Non-3GPP Accesses".
[4] 3GPP TS 23.060: "General Packet Radio Service (GPRS); Service Description".
[5] IETF RFC 5213: "Proxy Mobile IPv6".
[6] IETF RFC 5555: "Mobile IPv6 Support for Dual Stack Hosts and Routers".
[7] 3GPP TS 36.104: "Base Station (BS) Radio Transmission and Reception".
[8] 3GPP TS 29.060: "General Packet Radio Service (GPRS); GPRS Tunnelling Protocol (GTP) across the Gn and Gp Interface".
[9] 3GPP TS 21.905: "Technical Specification Group Services and System Aspects; Vocabulary for 3GPP Specifications".
[10] 3GPP TS 29.061: "Interworking between the Public Land Mobile Network (PLMN) Supporting Packet-Based Services and Packet Data Networks (PDN)".
[11] 3GPP TS 23.203: "Policy and Charging Control Architecture".
[12] 3GPP TS 29.214: "Policy and Charging Control over Rx Reference Point".
[13] 3GPP TS 24.008: "Mobile Radio Interface Layer 3 Specification; Core Network Protocols".
[14] 3GPP TS 24.301: "Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS)".
[15] 3GPP TS 23.122: "Non-Access-Stratum (NAS) Functions Related to Mobile Station (MS) in Idle Mode".
[16] 3GPP TS 36.300: "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRAN); Overall Description".
[17] 3GPP TS 36.413: "S1 Application Protocol (S1AP)".
[18] 3GPP TS 36.423: "X2 application protocol (X2AP)".
[19] 3GPP TS 29.274: "Evolved General Packet Radio Service (GPRS) Tunnelling Protocol for Control Plane (GTPv2-C)".
[20] IETF RFC 3588: "Diameter Base Protocol".
[21] 3GPP TS 29.272: "Evolved Packet System (EPS); Mobility Management Entity (MME) and Serving GPRS Support Node (SGSN) Related Interfaces Based on Diameter Protocol".

[22] 3GPP TS 29.281: “General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U)”.
[23] 3GPP TS 29.275: “Proxy Mobile IPv6 (PMIPv6) Based Mobility and Tunnelling protocols; Stage 3”.
[24] 3GPP TS 29.215: “Policy and Charging Control (PCC) Over S9 Reference Point; Stage 3”.
[25] 3GPP TS 29.212: “Policy and Charging Control (PCC); Reference Points”.
[26] 3GPP TS 33.401: “3GPP System Architecture Evolution (SAE); Security Architecture”.
[27] 3GPP TS 33.402: “3GPP System Architecture Evolution (SAE); Security Aspects of Non-3GPP Accesses”.
[28] 3GPP TS 32.240: “Charging Architecture and Principles”.
[29] 3GPP TS 32.251: “Packet Switched (PS) Domain Charging”.
[30] 3GPP TS 32.296: “Online Charging System (OCS) Applications and Interfaces”.
[31] 3GPP TS 32.273: “Multimedia Broadcast and Multicast Service (MBMS) Charging”.
[32] IETF RFC 3261: “SIP: Session Initiation Protocol”.
[33] IETF RFC 4566: “Session Description Protocol”.
[34] 3GPP TS 24.229: “IP Multimedia Call Control Protocol Based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3”.
[35] 3GPP TS 29.229: “Cx and Dx Interfaces Based on the Diameter Protocol; Protocol Details”.
[36] 3GPP TS 29.163: “Interworking between the IP Multimedia (IM) Core Network (CN) Subsystem and Circuit Switched (CS) Networks”.
[37] 3GPP TS 23.216: “Single Radio Voice Call Continuity (SRVCC)”.
[38] 3GPP TS 23.272: “Circuit Switched (CS) fallback in Evolved Packet System (EPS)”.
[39] 3GPP TS 23.204: “Support of Short Message Service (SMS) over Generic 3GPP Internet Protocol (IP) Access”.
[40] 3GPP TS 29.273: “Evolved Packet System (EPS); 3GPP EPS AAA Interfaces”.
[41] 3GPP TS 24.302: “Access to the 3GPP Evolved Packet Core (EPC) via Non-3GPP Access Networks; Stage 3”.
[42] 3GPP TS 24.244: “IP Multimedia Call Control Protocol Based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3”.
[43] IETF RFC 3748: “Extensible Authentication Protocol (EAP)”.
[44] IETF RFC 5448: “Improved Extensible Authentication Protocol Method for 3rd Generation Authentication and Key Agreement (EAP-AKA′)”.
[45] 3GPP TS 24.303: “Mobility Management Based on Dual-Stack Mobile IPv6; Stage 3”.
[46] 3GPP TS 23.251: “Network Sharing; Architecture and Functional Description”.
[47] 3GPP TS 23.246: “Multimedia Broadcast/Multicast Service (MBMS); Architecture and Functional Description”.
[48] 3GPP TS 36.445: “Evolved Universal Terrestrial Radio Access Network (E-UTRAN); M1 data transport”.
[49] 3GPP TS 36.443: “Evolved Universal Terrestrial Radio Access Network (E-UTRAN); M2 Application Protocol (M2AP)”.
[50] 3GPP TS 36.444: “Evolved Universal Terrestrial Radio Access Network (E-UTRAN); M3 Application Protocol (M3AP)”.