2

Network Architecture

Sudeep Palat and Philippe Godin

2.1 Introduction

As mentioned in the preceding chapter, LTE has been designed to support only Packet-Switched (PS) services, in contrast to the Circuit-Switched (CS) model of previous cellular systems. It aims to provide seamless Internet Protocol (IP) connectivity between User Equipment (UE) and the Packet Data Network (PDN), without any disruption to the end users' applications during mobility. While the term 'LTE' encompasses the evolution of the radio access through the Evolved-UTRAN[1] (E-UTRAN), it is accompanied by an evolution of the non-radio aspects under the term 'System Architecture Evolution' (SAE) which includes the Evolved Packet Core (EPC) network. Together LTE and SAE comprise the Evolved Packet System (EPS).

EPS uses the concept of *EPS bearers* to route IP traffic from a gateway in the PDN to the UE. A bearer is an IP packet flow with a defined Quality of Service (QoS). The E-UTRAN and EPC together set up and release bearers as required by applications. EPS natively supports voice services over the IP Multimedia Subsystem (IMS) using Voice over IP (VoIP), but LTE also supports interworking with legacy systems for traditional CS voice support.

This chapter presents the overall EPS network architecture, giving an overview of the functions provided by the Core Network (CN) and E-UTRAN. The protocol stack across the different interfaces is then explained, along with an overview of the functions provided by the different protocol layers. Section 2.4 outlines the end-to-end bearer path including QoS aspects, provides details of a typical procedure for establishing a bearer and discusses the inter-working with legacy systems for CS voice services. The remainder of the chapter presents the network interfaces in detail, with particular focus on the E-UTRAN interfaces and associated procedures, including those for the support of user mobility. The network elements and interfaces used solely to support broadcast services are covered in Chapter 13, and the aspects related to UE positioning in Chapter 19.

[1] Universal Terrestrial Radio Access Network.

LTE – The UMTS Long Term Evolution: From Theory to Practice, Second Edition.
Stefania Sesia, Issam Toufik and Matthew Baker.
© 2011 John Wiley & Sons, Ltd. Published 2011 by John Wiley & Sons, Ltd.

2.2 Overall Architectural Overview

EPS provides the user with IP connectivity to a PDN for accessing the Internet, as well as for running services such as VoIP. An EPS bearer is typically associated with a QoS. Multiple bearers can be established for a user in order to provide different QoS streams or connectivity to different PDNs. For example, a user might be engaged in a voice (VoIP) call while at the same time performing web browsing or File Transfer Protocol (FTP) download. A VoIP bearer would provide the necessary QoS for the voice call, while a best-effort bearer would be suitable for the web browsing or FTP session. The network must also provide sufficient security and privacy for the user and protection for the network against fraudulent use.

Release 9 of LTE introduced several additional features. To meet regulatory requirements for commercial voice, services such as support of IMS, emergency calls and UE positioning (see Chapter 19) were introduced. Enhancements to Home cells (HeNBs) were also introduced in Release 9 (see Chapter 24).

All these features are supported by means of several EPS network elements with different roles. Figure 2.1 shows the overall network architecture including the network elements and the standardized interfaces. At a high level, the network is comprised of the CN (i.e. EPC) and the access network (i.e. E-UTRAN). While the CN consists of many logical nodes, the access network is made up of essentially just one node, the evolved NodeB (eNodeB), which connects to the UEs. Each of these network elements is inter-connected by means of interfaces which are standardized in order to allow multivendor interoperability.

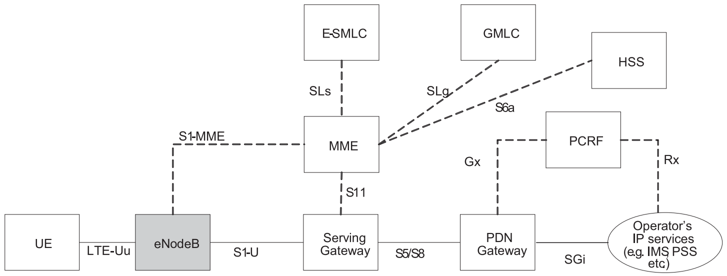

Figure 2.1: The EPS network elements.

The functional split between the EPC and E-UTRAN is shown in Figure 2.2. The EPC and E-UTRAN network elements are described in more detail below.

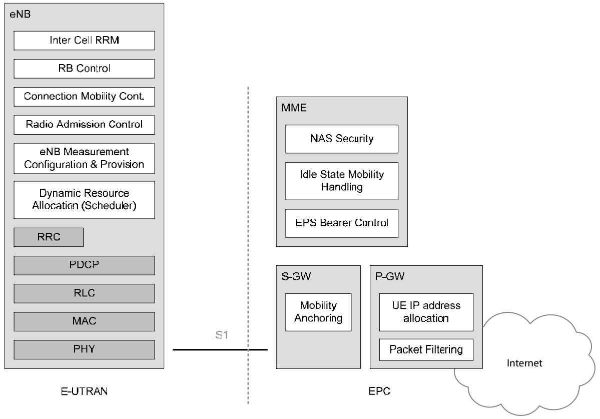

Figure 2.2: Functional split between E-UTRAN and EPC. Reproduced by permission of © 3GPP.

2.2.1 The Core Network

The CN (called the EPC in SAE) is responsible for the overall control of the UE and the establishment of the bearers. The main logical nodes of the EPC are:

- PDN Gateway (P-GW);
- Serving GateWay (S-GW);
- Mobility Management Entity (MME) ;
- Evolved Serving Mobile Location Centre (E-SMLC).

In addition to these nodes, the EPC also includes other logical nodes and functions such as the Gateway Mobile Location Centre (GMLC), the Home Subscriber Server (HSS) and the Policy Control and Charging Rules Function (PCRF). Since the EPS only provides a bearer path of a certain QoS, control of multimedia applications such as VoIP is provided by the IMS which is considered to be outside the EPS itself. When a user is roaming outside his home country network, the user's P-GW, GMLC and IMS domain may be located in either the home network or the visited network. The logical CN nodes (specified in [1]) are shown in Figure 2.1 and discussed in more detail below.

- **PCRF.** The PCRF is responsible for policy control decision-making, as well as for controlling the flow-based charging functionalities in the Policy Control Enforcement Function (PCEF) which resides in the P-GW. The PCRF provides the QoS authorization (QoS class identifier and bit rates) that decides how a certain data flow will be treated in the PCEF and ensures that this is in accordance with the user's subscription profile.

- **GMLC.** The GMLC contains functionalities required to support LoCation Services (LCS). After performing authorization, it sends positioning requests to the MME and receives the final location estimates.

- **Home Subscriber Server (HSS).** The HSS contains users' SAE subscription data such as the EPS-subscribed QoS profile and any access restrictions for roaming (see Section 2.2.3). It also holds information about the PDNs to which the user can connect. This could be in the form of an Access Point Name (APN) (which is a label according to DNS[2] naming conventions describing the access point to the PDN), or a PDN Address (indicating subscribed IP address(es)). In addition, the HSS holds dynamic information such as the identity of the MME to which the user is currently attached or registered. The HSS may also integrate the Authentication Centre (AuC) which generates the vectors for authentication and security keys (see Section 3.2.3.1).

- **P-GW.** The P-GW is responsible for IP address allocation for the UE, as well as QoS enforcement and flow-based charging according to rules from the PCRF. The P-GW is responsible for the filtering of downlink user IP packets into the different QoS-based bearers. This is performed based on Traffic Flow Templates (TFTs) (see Section 2.4). The P-GW performs QoS enforcement for Guaranteed Bit Rate (GBR) bearers. It also serves as the mobility anchor for inter-working with non-3GPP technologies such as CDMA2000 and WiMAX networks (see Section 2.4.2 and Chapter 22 for more information about mobility).

- **S-GW.** All user IP packets are transferred through the S-GW, which serves as the local mobility anchor for the data bearers when the UE moves between eNodeBs. It also retains the information about the bearers when the UE is in idle state (known as EPS Connection Management IDLE (ECM-IDLE), see Section 2.2.1.1) and temporarily buffers downlink data while the MME initiates paging of the UE to re-establish the bearers. In addition, the S-GW performs some administrative functions in the visited network, such as collecting information for charging (e.g. the volume of data sent to or received from the user) and legal interception. It also serves as the mobility anchor for inter-working with other 3GPP technologies such as GPRS[3] and UMTS[4] (see Section 2.4.2 and Chapter 22 for more information about mobility).

- **MME.** The MME is the control node which processes the signalling between the UE and the CN. The protocols running between the UE and the CN are known as the *Non-Access Stratum* (NAS) protocols.

  The main functions supported by the MME are classified as:

  Functions related to bearer management. This includes the establishment, maintenance and release of the bearers, and is handled by the session management layer in the NAS protocol.

  Functions related to connection management. This includes the establishment of the connection and security between the network and UE, and is handled by the connection or mobility management layer in the NAS protocol layer.

[2] Domain Name System.
[3] General Packet Radio Service.
[4] Universal Mobile Telecommunications System.

NAS control procedures are specified in [1] and are discussed in more detail in the following section.

Functions related to inter-working with other networks. This includes handing over of voice calls to legacy networks and is explained in more detail in Section 2.4.2.

- **E-SMLC.** The E-SMLC manages the overall coordination and scheduling of resources required to find the location of a UE that is attached to E-UTRAN. It also calculates the final location based on the estimates it receives, and it estimates the UE speed and the achieved accuracy. The positioning functions and protocols are explained in detail in Chapter 19.

2.2.1.1 Non-Access Stratum (NAS) Procedures

The NAS procedures, especially the connection management procedures, are fundamentally similar to UMTS. The main change from UMTS is that EPS allows concatenation of some procedures so as to enable faster establishment of the connection and the bearers.

The MME creates a *UE context*when a UE is turned on and attaches to the network. It assigns to the UE a unique short temporary identity termed the SAE-Temporary Mobile Subscriber Identity (S-TMSI) which identifies the UE context in the MME. This UE context holds user subscription information downloaded from the HSS. The local storage of subscription data in the MME allows faster execution of procedures such as bearer establishment since it removes the need to consult the HSS every time. In addition, the UE context also holds dynamic information such as the list of bearers that are established and the terminal capabilities.

To reduce the overhead in the E-UTRAN and the processing in the UE, all UE-related information in the access network can be released during long periods of data inactivity. The UE is then in the ECM-IDLE state. The MME retains the UE context and the information about the established bearers during these idle periods.

To allow the network to contact an ECM-IDLE UE, the UE updates the network as to its new location whenever it moves out of its current Tracking Area (TA); this procedure is called a 'Tracking Area Update'. The MME is responsible for keeping track of the user location while the UE is in ECM-IDLE.

When there is a need to deliver downlink data to an ECM-IDLE UE, the MME sends a paging message to all the eNodeBs in its current TA, and the eNodeBs page the UE over the radio interface. On receipt of a paging message, the UE performs a service request procedure which results in moving the UE to the ECM-CONNECTED state. UE-related information is thereby created in the E-UTRAN, and the bearers are re-established. The MME is responsible for the re-establishment of the radio bearers and updating the UE context in the eNodeB. This transition between the UE states is called an 'idle-to-active transition'. To speed up the idle-to-active transition and bearer establishment, EPS supports concatenation of the NAS and AS[5] procedures for bearer activation (see also Section 2.4.1). Some inter-relationship between the NAS and AS protocols is intentionally used to allow procedures to run simultaneously, rather than sequentially as in UMTS. For example, the bearer establishment procedure can be executed by the network without waiting for the completion of the security procedure.

[5] Access Stratum – the protocols which run between the eNodeBs and the UE.

Security functions are the responsibility of the MME for both signalling and user data. When a UE attaches with the network, a mutual authentication of the UE and the network is performed between the UE and the MME/HSS. This authentication function also establishes the security keys which are used for encryption of the bearers, as explained in Section 3.2.3.1. The security architecture for SAE is specified in [2].

The NAS also handles IMS Emergency calls, whereby UEs without regular access to the network (i.e. terminals without a Universal Subscriber Identity Module (USIM) or UEs in limited service mode) are allowed access to the network using an 'Emergency Attach' procedure; this bypasses the security requirements but only allows access to an emergency P-GW.

2.2.2 The Access Network

The access network of LTE, E-UTRAN, simply consists of a network of eNodeBs, as illustrated in Figure 2.3. For normal user traffic (as opposed to broadcast), there is no centralized controller in E-UTRAN; hence the E-UTRAN architecture is said to be flat.

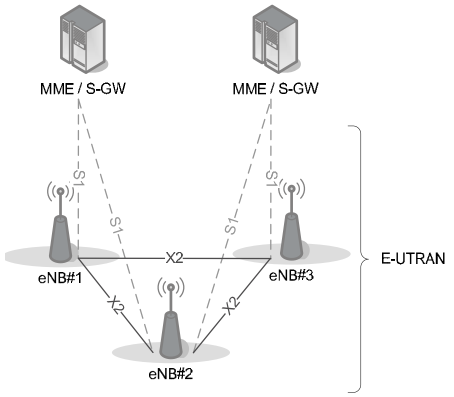

Figure 2.3: Overall E-UTRAN architecture. Reproduced by permission of © 3GPP.

The eNodeBs are normally inter-connected with each other by means of an interface known as *X2*, and to the EPC by means of the *S1* interface – more specifically, to the MME by means of the S1-MME interface and to the S-GW by means of the S1-U interface.

The protocols which run between the eNodeBs and the UE are known as the *Access Stratum* (AS) protocols.

The E-UTRAN is responsible for all radio-related functions, which can be summarized briefly as:

- **Radio Resource Management.** This covers all functions related to the radio bearers, such as radio bearer control, radio admission control, radio mobility control, scheduling and dynamic allocation of resources to UEs in both uplink and downlink.
- **Header Compression.** This helps to ensure efficient use of the radio interface by compressing the IP packet headers which could otherwise represent a significant overhead, especially for small packets such as VoIP (see Section 4.2.2).
- **Security.** All data sent over the radio interface is encrypted (see Sections 3.2.3.1 and 4.2.3).
- **Positioning.** The E-UTRAN provides the necessary measurements and other data to the E-SMLC and assists the E-SMLC in finding the UE position (see Chapter 19).
- **Connectivity to the EPC.** This consists of the signalling towards the MME and the bearer path towards the S-GW.

On the network side, all of these functions reside in the eNodeBs, each of which can be responsible for managing multiple cells. Unlike some of the previous second- and third-generation technologies, LTE integrates the radio controller function into the eNodeB. This allows tight interaction between the different protocol layers of the radio access network, thus reducing latency and improving efficiency. Such distributed control eliminates the need for a high-availability, processing-intensive controller, which in turn has the potential to reduce costs and avoid 'single points of failure'. Furthermore, as LTE does not support soft handover there is no need for a centralized data-combining function in the network.

One consequence of the lack of a centralized controller node is that, as the UE moves, the network must transfer all information related to a UE, i.e. the UE context, together with any buffered data, from one eNodeB to another. As discussed in Section 2.3.1.1, mechanisms are therefore needed to avoid data loss during handover. The operation of the X2 interface for this purpose is explained in more detail in Section 2.6.

An important feature of the S1 interface linking the access network to the CN is known as *S1-flex*. This is a concept whereby multiple CN nodes (MME/S-GWs) can serve a common geographical area, being connected by a mesh network to the set of eNodeBs in that area (see Section 2.5). An eNodeB may thus be served by multiple MME/S-GWs, as is the case for eNodeB#2 in Figure 2.3. The set of MME/S-GW nodes serving a common area is called an *MME/S-GW pool* , and the area covered by such a pool of MME/S-GWs is called a *pool area*. This concept allows UEs in the cell(s) controlled by one eNodeB to be shared between multiple CN nodes, thereby providing a possibility for load sharing and also eliminating single points of failure for the CN nodes. The UE context normally remains with the same MME as long as the UE is located within the pool area.

2.2.3 Roaming Architecture

A network run by one operator in one country is known as a Public Land Mobile Network (PLMN). Roaming, where users are allowed to connect to PLMNs other than those to which they are directly subscribed, is a powerful feature for mobile networks, and LTE/SAE is no exception. A roaming user is connected to the E-UTRAN, MME and S-GW of the visited LTE network. However, LTE/SAE allows the P-GW of either the visited or the home network to be used, as shown in Figure 2.4. Using the home network's P-GW allows the user to access the home operator's services even while in a visited network. A P-GW in the visited network allows a 'local breakout' to the Internet in the visited network.

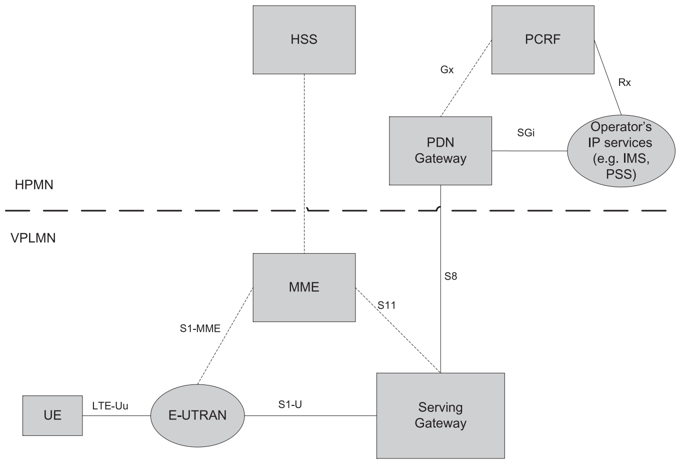

Figure 2.4: Roaming architecture for 3GPP accesses with P-GW in home network.

2.3 Protocol Architecture

We outline here the radio protocol architecture of E-UTRAN.

2.3.1 User Plane

An IP packet for a UE is encapsulated in an EPC-specific protocol and tunnelled between the P-GW and the eNodeB for transmission to the UE. Different tunnelling protocols are used across different interfaces. A 3GPP-specific tunnelling protocol called the GPRS Tunnelling Protocol (GTP) [4] is used over the core network interfaces, S1 and S5/S8.[6]

The E-UTRAN user plane protocol stack, shown greyed in Figure 2.5, consists of the Packet Data Convergence Protocol (PDCP), Radio Link Control (RLC) and Medium Access Control (MAC) sublayers which are terminated in the eNodeB on the network side. The respective roles of each of these layers are explained in detail in Chapter 4.

[6] SAE also provides an option to use Proxy Mobile IP (PMIP) on S5/S8. More details on the MIP-based S5/S8 interface can be found in [3].

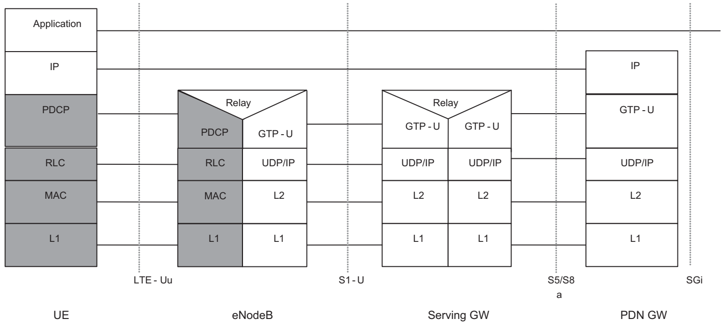

Figure 2.5: The E-UTRAN user plane protocol stack. Reproduced by permission of © 3GPP.

2.3.1.1 Data Handling During Handover

In the absence of any centralized controller node, data buffering during handover due to user mobility in the E-UTRAN must be performed in the eNodeB itself. Data protection during handover is a responsibility of the PDCP layer and is explained in detail in Section 4.2.4.

The RLC and MAC layers both start afresh in a new cell after handover is completed.

2.3.2 Control Plane

The protocol stack for the control plane between the UE and MME is shown in Figure 2.6.

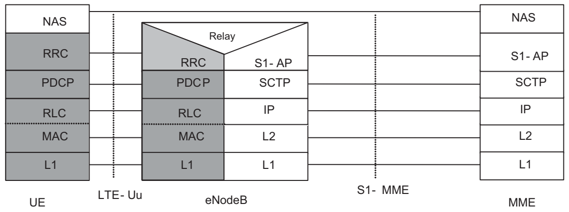

Figure 2.6: Control plane protocol stack. Reproduced by permission of © 3GPP.

The greyed region of the stack indicates the AS protocols. The lower layers perform the same functions as for the user plane with the exception that there is no header compression function for control plane.

The Radio Resource Control (RRC) protocol is known as ‘Layer 3’ in the AS protocol stack. It is the main controlling function in the AS, being responsible for establishing the radio bearers and configuring all the lower layers using RRC signalling between the eNodeB and the UE. These functions are detailed in Section 3.2.

2.4 Quality of Service and EPS Bearers

In a typical case, multiple applications may be running in a UE at the same time, each one having different QoS requirements. For example, a UE can be engaged in a VoIP call while at the same time browsing a web page or downloading an FTP file. VoIP has more stringent requirements for QoS in terms of delay and delay jitter than web browsing and FTP, while the latter requires a much lower packet loss rate. In order to support multiple QoS requirements, different bearers are set up within EPS, each being associated with a QoS.

Broadly, bearers can be classified into two categories based on the nature of the QoS they provide:

- **Minimum Guaranteed Bit Rate (GBR) bearers** which can be used for applications such as VoIP. These have an associated GBR value for which dedicated transmission resources are permanently allocated (e.g. by an admission control function in the eNodeB) at bearer establishment/modification. Bit rates higher than the GBR may be allowed for a GBR bearer if resources are available. In such cases, a Maximum Bit Rate (MBR) parameter, which can also be associated with a GBR bearer, sets an upper limit on the bit rate which can be expected from a GBR bearer.

- **Non-GBR bearers** which do not guarantee any particular bit rate. These can be used for applications such as web browsing or FTP transfer. For these bearers, no bandwidth resources are allocated permanently to the bearer.

In the access network, it is the eNodeB’s responsibility to ensure that the necessary QoS for a bearer over the radio interface is met. Each bearer has an associated Class Identifier (QCI), and an Allocation and Retention Priority (ARP).

Each QCI is characterized by priority, packet delay budget and acceptable packet loss rate. The QCI label for a bearer determines the way it is handled in the eNodeB. Only a dozen such QCIs have been standardized so that vendors can all have the same understanding of the underlying service characteristics and thus provide the corresponding treatment, including queue management, conditioning and policing strategy. This ensures that an LTE operator can expect uniform traffic handling behaviour throughout the network regardless of the manufacturers of the eNodeB equipment. The set of standardized QCIs and their characteristics (from which the PCRF in an EPS can select) is provided in Table 2.1 (from Section 6.1.7 in [5]).

The priority and packet delay budget (and, to some extent, the acceptable packet loss rate) from the QCI label determine the RLC mode configuration (see Section 4.3.1), and how the scheduler in the MAC (see Section 4.4.2.1) handles packets sent over the bearer (e.g. in terms of scheduling policy, queue management policy and rate shaping policy). For example, a packet with a higher priority can be expected to be scheduled before a packet with lower priority. For bearers with a low acceptable loss rate, an Acknowledged Mode (AM) can be

Table 2.1: Standardized QoS Class Identifiers (QCIs) for LTE.

QCI	Resource type	Priority	Packet delay budget (ms)	Packet error loss rate	Example services
1	GBR	2	100	10^{-2}	Conversational voice
2	GBR	4	150	10^{-3}	Conversational video (live streaming)
3	GBR	5	300	10^{-6}	Non-conversational video (buffered streaming)
4	GBR	3	50	10^{-3}	Real time gaming
5	Non-GBR	1	100	10^{-6}	IMS signalling
6	Non-GBR	7	100	10^{-3}	Voice, video (live streaming), interactive gaming
7	Non-GBR	6	300	10^{-6}	Video (buffered streaming)
8	Non-GBR	8	300	10^{-6}	TCP-based (e.g. WWW, e-mail) chat, FTP, p2p file sharing, progressive video, etc.
9	Non-GBR	9	300	10^{-6}	

used within the RLC protocol layer to ensure that packets are delivered successfully across the radio interface (see Section 4.3.1.3).

The ARP of a bearer is used for call admission control – i.e. to decide whether or not the requested bearer should be established in case of radio congestion. It also governs the prioritization of the bearer for pre-emption with respect to a new bearer establishment request. Once successfully established, a bearer's ARP does not have any impact on the bearer-level packet forwarding treatment (e.g. for scheduling and rate control). Such packet forwarding treatment should be solely determined by the other bearer-level QoS parameters such as QCI, GBR and MBR.

An EPS bearer has to cross multiple interfaces as shown in Figure 2.7 – the S5/S8 interface from the P-GW to the S-GW, the S1 interface from the S-GW to the eNodeB, and the radio interface (also known as the LTE-Uu interface) from the eNodeB to the UE. Across each interface, the EPS bearer is mapped onto a lower layer bearer, each with its own bearer identity. Each node must keep track of the binding between the bearer IDs across its different interfaces.

An S5/S8 bearer transports the packets of an EPS bearer between a P-GW and an S-GW. The S-GW stores a one-to-one mapping between an S1 bearer and an S5/S8 bearer. The bearer is identified by the GTP tunnel ID across both interfaces.

An S1 bearer transports the packets of an EPS bearer between an S-GW and an eNodeB. A radio bearer [6] transports the packets of an EPS bearer between a UE and an eNodeB. An E-UTRAN Radio Access Bearer (E-RAB) refers to the concatenation of an S1 bearer and the corresponding radio bearer. An eNodeB stores a one-to-one mapping between a radio bearer ID and an S1 bearer to create the mapping between the two. The overall EPS bearer service architecture is shown in Figure 2.8.

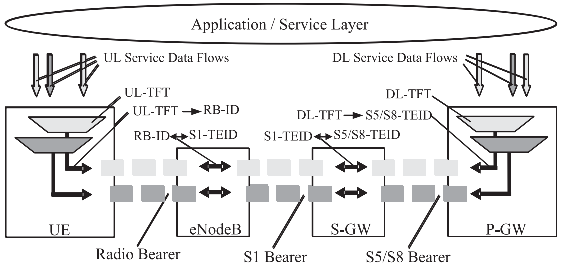

Figure 2.7: LTE/SAE bearers across the different interfaces. Reproduced by permission of © 3GPP.

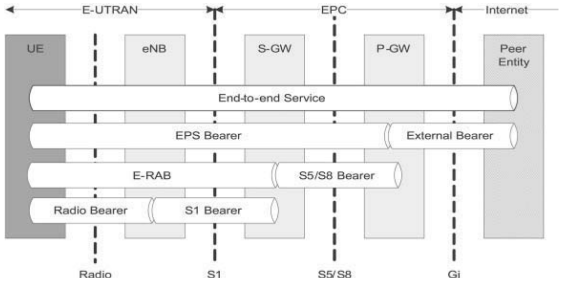

Figure 2.8: The overall EPS bearer service architecture. Reproduced by permission of © 3GPP.

IP packets mapped to the same EPS bearer receive the same bearer-level packet forwarding treatment (e.g. scheduling policy, queue management policy, rate shaping policy, RLC configuration). Providing different bearer-level QoS thus requires that a separate EPS bearer is established for each QoS flow, and user IP packets must be filtered into the different EPS bearers.

Packet filtering into different bearers is based on Traffic Flow Templates (TFTs). The TFTs use IP header information such as source and destination IP addresses and Transmission Control Protocol (TCP) port numbers to filter packets such as VoIP from web browsing traffic so that each can be sent down the respective bearers with appropriate QoS. An UpLink TFT (UL TFT) associated with each bearer in the UE filters IP packets to EPS bearers in the uplink direction. A DownLink TFT (DL TFT) in the P-GW is a similar set of downlink packet filters.

As part of the procedure by which a UE attaches to the network, the UE is assigned an IP address by the P-GW and at least one bearer is established, called the default bearer, and it remains established throughout the lifetime of the PDN connection in order to provide the UE with always-on IP connectivity to that PDN. The initial bearer-level QoS parameter values of the default bearer are assigned by the MME, based on subscription data retrieved from the HSS. The PCEF may change these values in interaction with the PCRF or according to local configuration. Additional bearers called dedicated bearers can also be established at any time during or after completion of the attach procedure. A dedicated bearer can be either GBR or non-GBR (the default bearer always has to be a non-GBR bearer since it is permanently established). The distinction between default and dedicated bearers should be transparent to the access network (e.g. E-UTRAN). Each bearer has an associated QoS, and if more than one bearer is established for a given UE, then each bearer must also be associated with appropriate TFTs. These dedicated bearers could be established by the network, based for example on a trigger from the IMS domain, or they could be requested by the UE. The dedicated bearers for a UE may be provided by one or more P-GWs.

The bearer-level QoS parameter values for dedicated bearers are received by the P-GW from the PCRF and forwarded to the S-GW. The MME only transparently forwards those values received from the S-GW over the S11 reference point to the E-UTRAN.

2.4.1 Bearer Establishment Procedure

This section describes an example of the end-to-end bearer establishment procedure across the network nodes using the functionality described in the previous sections.

A typical bearer establishment flow is shown in Figure 2.9. Each of the messages is described below.

When a bearer is established, the bearers across each of the interfaces discussed above are established.

The PCRF sends a 'PCC[7] Decision Provision' message indicating the required QoS for the bearer to the P-GW. The P-GW uses this QoS policy to assign the bearer-level QoS parameters. The P-GW then sends to the S-GW a 'Create Dedicated Bearer Request' message including the QoS and UL TFT to be used in the UE.

The S-GW forwards the Create Dedicated Bearer Request message (including bearer QoS, UL TFT and S1-bearer ID) to the MME (message 3 in Figure 2.9).

The MME then builds a set of session management configuration information including the UL TFT and the EPS bearer identity, and includes it in the 'Bearer Setup Request' message which it sends to the eNodeB (message 4 in Figure 2.9). The session management configuration is NAS information and is therefore sent transparently by the eNodeB to the UE.

The Bearer Setup Request also provides the QoS of the bearer to the eNodeB; this information is used by the eNodeB for call admission control and also to ensure the necessary QoS by appropriate scheduling of the user's IP packets. The eNodeB maps the EPS bearer QoS to the radio bearer QoS. It then signals a 'RRC Connection Reconfiguration' message (including the radio bearer QoS, session management configuration and EPS radio bearer identity) to the UE to set up the radio bearer (message 5 in Figure 2.9). The RRC Connection Reconfiguration message contains all the configuration parameters for the radio interface.

[7] Policy Control and Charging.

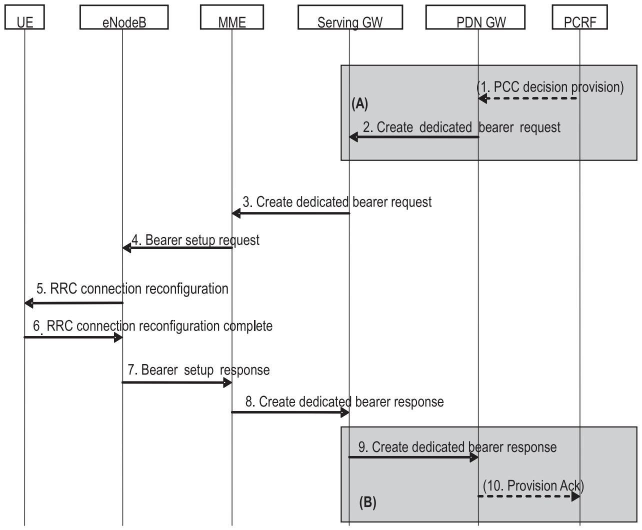
UE
eNodeB
MME
Serving GW
PDN GW
PCRF
(1. PCC decision provision)
(A)
2. Create dedicated bearer request
3. Create dedicated bearer request
4. Bearer setup request
5. RRC connection reconfiguration
6 RRC connection reconfiguration complete
7. Bearer setup response
8. Create dedicated bearer response
9. Create dedicated bearer response
(10. Provision Ack)
(B)

Figure 2.9: An example message flow for an LTE/SAE bearer establishment. Reproduced by permission of © 3GPP.

This is mainly for the configuration of the Layer 2 (PDCP, RLC and MAC) parameters, but also the Layer 1 parameters required for the UE to initialize the protocol stack.

Messages 6 to 10 are the corresponding response messages to confirm that the bearers have been set up correctly.

2.4.2 Inter-Working with other RATs

EPS also supports inter-working and mobility (handover) with networks using other Radio Access Technologies (RATs), notably GSM[8], UMTS, CDMA2000 and WiMAX. The architecture for inter-working with 2G and 3G GPRS/UMTS networks is shown in Figure 2.10. The S-GW acts as the mobility anchor for inter-working with other 3GPP technologies such as GSM and UMTS, while the P-GW serves as an anchor allowing seamless mobility to non-3GPP networks such as CDMA2000 or WiMAX. The P-GW may also support a Proxy Mobile Internet Protocol (PMIP) based interface. While VoIP is the primary mechanism for voice services, LTE also supports inter-working with legacy systems for CS voice services.

[8] Global System for Mobile Communications.

This is controlled by the MME and is based on two procedures outlined in Sections 2.4.2.1 and 2.4.2.2.

More details of the radio interface procedures for inter-working with other RATs are specified in [3] and covered in Sections 2.5.6.2 and 3.2.4.

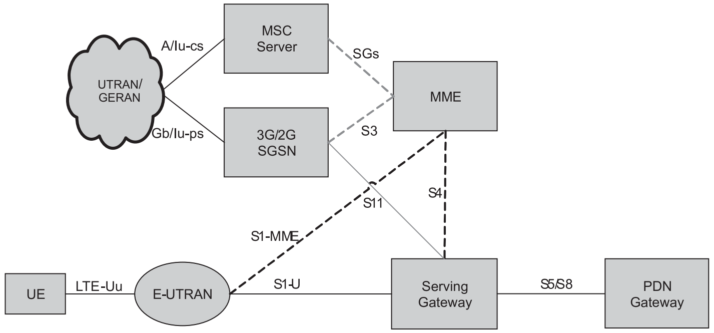

Figure 2.10: Architecture for 3G UMTS interworking.

2.4.2.1 Circuit-Switched Fall Back (CSFB)

LTE natively supports VoIP only using IMS services. However, in case IMS services are not deployed from the start, LTE also supports a Circuit-Switched FallBack (CSFB) mechanism which allows CS voice calls to be handled via legacy RATs for UEs that are camped on LTE.

CSFB allows a UE in LTE to be handed over to a legacy RAT to originate a CS voice call. This is supported by means of an interface, referred to as SGs[9], between the MME and the Mobile Switching Centre (MSC) of the legacy RAT shown in Figure 2.10. This interface allows the UE to attach with the MSC and register for CS services while still in LTE. Moreover it carries paging messages from the MSC for incoming voice calls so that UEs can be paged over LTE. The network may choose a handover, cell change order, or redirection procedure to move the UE to the legacy RAT.

Figure 2.11 shows the message flow for a CSFB call from LTE to UMTS, including paging from the MSC via the SGs interface and MME in the case of UE-terminated calls, and the sending of an Extended Service Request NAS message from the UE to the MME to trigger either a handover or redirection to the target RAT in the case of a UE-originated call. In the latter case, the UE then originates the CS call over the legacy RAT using the procedure defined in the legacy RAT specification. Further details of CSFB can be found in [7].

[9] SGs is an extension of the Gs interface between the Serving GPRS Support Node (SGSN) and the Mobile Switching Centre (MSC)

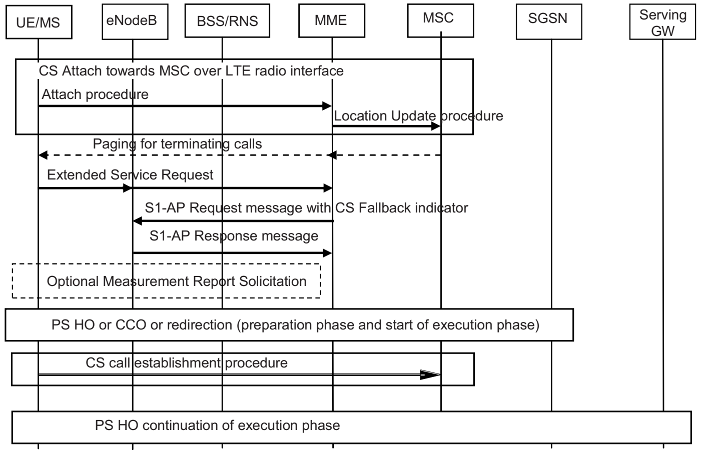
UE/MS
eNodeB
BSS/RNS
MME
MSC
SGSN
Serving GW
CS Attach towards MSC over LTE radio interface
Attach procedure
Location Update procedure
Paging for terminating calls
Extended Service Request
S1-AP Request message with CS Fallback indicator
S1-AP Response message
Optional Measurement Report Solicitation
PS HO or CCO or redirection (preparation phase and start of execution phase)
CS call establishment procedure
PS HO continuation of execution phase

Figure 2.11: Message sequence diagram for CSFB from LTE to UMTS/GERAN.

2.4.2.2 Single Radio Voice Call Continuity (SRVCC)

If ubiquitous coverage of LTE is not available, it is possible that a UE involved in a VoIP call over LTE might then move out of LTE coverage to enter a legacy RAT cell which only offers CS voice services. The Single Radio Voice Call Continuity (SRVCC) procedure is designed for handover of a Packet Switched (PS) VoIP call over LTE to a CS voice call in the legacy RAT, involving the transfer of a PS bearer into a CS bearer.

Figure 2.12 shows an overview of the functions involved in SRVCC. The eNodeB may detect that the UE is moving out of LTE coverage and trigger a handover procedure towards the MME by means of an SRVCC indication. The MME is responsible for the SRVCC procedure and also for the transfer of the PS E-RAB carrying VoIP into a CS bearer. The MSC Server then initiates the session transfer procedure to IMS and coordinates it with the CS handover procedure to the target cell. The handover command provided to the UE to request handover to the legacy RAT also provides the information to set up the CS and PS radio bearers. The UE can continue with the call over the CS domain on completion of the handover. Further details of SRVCC can be found in [8].

2.5 The E-UTRAN Network Interfaces: S1 Interface

The S1 interface connects the eNodeB to the EPC. It is split into two interfaces, one for the control plane and the other for the user plane. The protocol structure for the S1 and the functionality provided over S1 are discussed in more detail below.

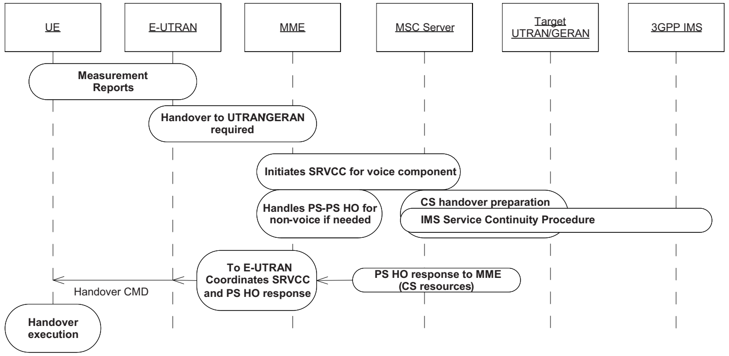
UE
E-UTRAN
MME
MSC Server
Target UTRAN/GERAN
3GPP IMS
Measurement Reports
Handover to UTRANGERAN required
Initiates SRVCC for voice component
Handles PS-PS HO for non-voice if needed
CS handover preparation
IMS Service Continuity Procedure
To E-UTRAN Coordinates SRVCC and PS HO response
PS HO response to MME (CS resources)
Handover CMD
Handover execution

Figure 2.12: The main procedures involved in an SRVCC handover of a PS VoIP call from LTE to CS voice call in UMTS/GERAN.

2.5.1 Protocol Structure over S1

The protocol structure over S1 is based on a full IP transport stack with no dependency on legacy SS7[10] network configuration as used in GSM or UMTS networks. This simplification provides one area of potential savings on operational expenditure with LTE networks.

2.5.1.1 Control Plane

Figure 2.13 shows the protocol structure of the S1 control plane which is based on the Stream Control Transmission Protocol / IP (SCTP/IP) stack.

The SCTP protocol is well known for its advanced features inherited from TCP which ensure the required reliable delivery of the signalling messages. In addition, it makes it possible to benefit from improved features such as the handling of multistreams to implement transport network redundancy easily and avoid head-of-line blocking or multihoming (see 'IETF RFC4960' [9]).

A further simplification in LTE (compared to the UMTS Iu interface, for example) is the direct mapping of the S1-AP (S1 Application Protocol) on top of SCTP which results in a simplified protocol stack with no intermediate connection management protocol. The individual connections are directly handled at the application layer. Multiplexing takes place between S1-AP and SCTP whereby each stream of an SCTP association is multiplexed with the signalling traffic of multiple individual connections.

One further area of flexibility that comes with LTE lies in the lower layer protocols for which full optionality has been left regarding the choice of the IP version and the choice of the data link layer. For example, this enables the operator to start deployment using IP version 4 with the data link tailored to the network deployment scenario.

[10] Signalling System #7 (SS7) is a communications protocol defined by the International Telecommunication Union (ITU) Telecommunication Standardization Sector (ITU-T) with a main purpose of setting up and tearing down telephone calls. Other uses include Short Message Service (SMS), number translation, prepaid billing mechanisms, and many other services.

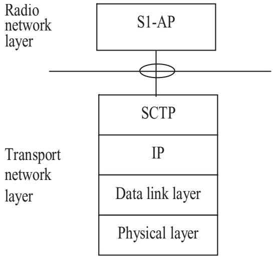

Figure 2.13: S1-MME control plane protocol stack. Reproduced by permission of © 3GPP.

2.5.1.2 User Plane

Figure 2.14 shows the protocol structure of the S1 user plane, which is based on the GTP/ User Datagram Protocol (UDP) IP stack which is already well known from UMTS networks.

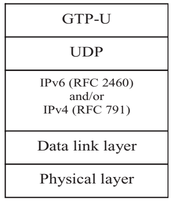

Figure 2.14: S1-U user plane protocol stack. Reproduced by permission of © 3GPP.

One of the advantages of using GTP-User plane (GTP-U) is its inherent facility to identify tunnels and also to facilitate intra-3GPP mobility.

The IP version number and the data link layer have been left fully optional, as for the control plane stack.

A transport bearer is identified by the GTP tunnel endpoints and the IP address (source Tunnelling End ID (TEID), destination TEID, source IP address, destination IP address).

The S-GW sends downlink packets of a given bearer to the eNodeB IP address (received in S1-AP) associated to that particular bearer. Similarly, the eNodeB sends upstream packets of a given bearer to the EPC IP address (received in S1-AP) associated to that particular bearer.

Vendor-specific traffic categories (e.g. real-time traffic) can be mapped onto Differentiated Services (Diffserv) code points (e.g. expedited forwarding) by network O&M (Operation and Maintenance) configuration to manage QoS differentiation between the bearers.

2.5.2 Initiation over S1

The initialization of the S1-MME control plane interface starts with the identification of the MMEs to which the eNodeB must connect, followed by the setting up of the Transport Network Layer (TNL).

With the support of the S1-flex function in LTE, an eNodeB must initiate an S1 interface towards each MME node of the pool area to which it belongs. This list of MME nodes of the pool together with an initial corresponding remote IP address can be directly configured in the eNodeB at deployment (although other means may also be used). The eNodeB then initiates the TNL establishment with that IP address. Only one SCTP association is established between one eNodeB and one MME.

During the establishment of the SCTP association, the two nodes negotiate the maximum number of streams which will be used over that association. However, multiple pairs of streams[11] are typically used in order to avoid the head-of-line blocking issue mentioned above. Among these pairs of streams, one particular pair must be reserved by the two nodes for the signalling of the common procedures (i.e. those which are not specific to one UE). The other streams are used for the sole purpose of the dedicated procedures (i.e. those which are specific to one UE).

Once the TNL has been established, some basic application-level configuration data for the system operation is automatically exchanged between the eNodeB and the MME through an 'S1 SETUP' procedure initiated by the eNodeB. This procedure is one case of a Self-Optimizing Network process and is explained in detail in Section 25.3.1.

Once the S1 SETUP procedure has been completed, the S1 interface is operational.

2.5.3 Context Management over S1

Within each pool area, a UE is associated to one particular MME for all its communications during its stay in this pool area. This creates a context in this MME for the UE. This particular MME is selected by the NAS Node Selection Function (NNSF) in the first eNodeB from which the UE entered the pool.

Whenever the UE becomes active (i.e. makes a transition from idle to active mode) under the coverage of a particular eNodeB in the pool area, the MME provides the UE context information to this eNodeB using the 'INITIAL CONTEXT SETUP REQUEST' message (see Figure 2.15). This enables the eNodeB in turn to create a context and manage the UE while it is in active mode.

Even though the setup of bearers is otherwise relevant to a dedicated 'Bearer Management' procedure described below, the creation of the eNodeB context by the INITIAL CONTEXT SETUP procedure also includes the creation of one or several bearers including the default bearers.

At the next transition back to idle mode following a 'UE CONTEXT RELEASE' message sent from the MME, the eNodeB context is erased and only the MME context remains.

[11] Note that a stream is unidirectional and therefore pairs must be used.

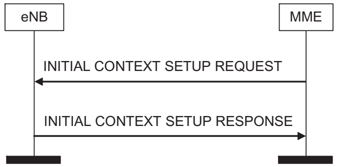

Figure 2.15: Initial context setup procedure. Reproduced by permission of © 3GPP.

2.5.4 Bearer Management over S1

LTE uses independent dedicated procedures respectively covering the setup, modification and release of bearers. For each bearer requested to be set up, the transport layer address and the tunnel endpoint are provided to the eNodeB in the 'BEARER SETUP REQUEST' message to indicate the termination of the bearer in the S-GW where uplink user plane data must be sent. Conversely, the eNodeB indicates in the 'BEARER SETUP RESPONSE' message the termination of the bearer in the eNodeB where the downlink user plane data must be sent.

For each bearer, the QoS parameters (see Section 2.4) requested for the bearer are also indicated. Independently of the standardized QCI values, it is also still possible to use extra proprietary labels for the fast introduction of new services if vendors and operators agree upon them.

2.5.5 Paging over S1

As mentioned in Section 2.5.3, in order to re-establish a connection towards a UE in idle mode, the MME distributes a 'PAGING REQUEST' message to the relevant eNodeBs based on the TAs where the UE is expected to be located. When receiving the paging request, the eNodeB sends a page over the radio interface in the cells which are contained within one of the TAs provided in that message.

The UE is normally paged using its S-TMSI. The 'PAGING REQUEST' message also contains a UE identity index value in order for the eNodeB to calculate the paging occasions at which the UE will switch on its receiver to listen for paging messages (see Section 3.4).

In Release 10, paging differentiation is introduced over the S1 interface to handle Multimedia Priority Service (MPS)[12] users. In case of MME or RAN overload, it is necessary to page a UE with higher priority during the establishment of a mobile-terminated MPS call. In case of MME overload, the MME can itself discriminate between the paging messages and discard the lower priority ones. In case of RAN overload in some cells, the eNodeB can perform this discrimination based on a new Paging Priority Indicator sent by the MME. The MME can signal up to eight such priority values to the eNodeB. In case of an IMS MPS call, the terminating UE will further set up an RRC connection with the same eNodeB that will also get automatically prioritized. In case of a CS fallback call, the eNodeB will instead

[12] MPS allows the delivery of calls or complete sessions of a high priority nature, in case for example of public safety or national security purposes, from mobile to mobile, mobile to fixed, and fixed to mobile networks during network congestion conditions.

signal to the UE that it must set the cause value ‘high priority terminating call’ when trying to establish the UMTS RRC Connection.

2.5.6 Mobility over S1

LTE/SAE supports mobility within LTE/SAE, and also to other systems using both 3GPP and non-3GPP technologies. The mobility procedures over the radio interface are defined in Section 3.2. These mobility procedures also involve the network interfaces. The sections below discuss the procedures over S1 to support mobility. The mobility performance requirements from the UE point of view are outlined in Chapter 22.

2.5.6.1 Intra-LTE Mobility

There are two types of handover procedure in LTE for UEs in active mode: the S1-handover procedure and the X2-handover procedure.

For intra-LTE mobility, the X2-handover procedure is normally used for the inter-eNodeB handover (described in Section 2.6.3). However, when there is no X2 interface between the two eNodeBs, or if the source eNodeB has been configured to initiate handover towards a particular target eNodeB via the S1 interface, then an S1-handover will be triggered.

The S1-handover procedure has been designed in a very similar way to the UMTS Serving Radio Network Subsystem (SRNS) relocation procedure and is shown in Figure 2.16: it consists of a preparation phase involving the core network, where the resources are first prepared at the target side (steps 2 to 8), followed by an execution phase (steps 8 to 12) and a completion phase (after step 13).

Compared to UMTS, the main difference is the introduction of the ‘STATUS TRANSFER’ message sent by the source eNodeB (steps 10 and 11). This message has been added in order to carry some PDCP status information that is needed at the target eNodeB in cases when PDCP status preservation applies for the S1-handover (see Section 4.2.4); this is in alignment with the information which is sent within the X2 ‘STATUS TRANSFER’ message used for the X2-handover (see below). As a result of this alignment, the handling of the handover by the target eNodeB as seen from the UE is exactly the same, regardless of the type of handover (S1 or X2) the network had decided to use; indeed, the UE is unaware of which type of handover is used by the network.

The ‘Status Transfer’ procedure is assumed to be triggered in parallel with the start of data forwarding after the source eNodeB has received the ‘HANDOVER COMMAND’ message from the source MME. This data forwarding can be either direct or indirect, depending on the availability of a direct path for the user plane data between the source eNodeB and the target eNodeB.

The ‘HANDOVER NOTIFY’ message (step 13), which is sent later by the target eNodeB when the arrival of the UE at the target side is confirmed, is forwarded by the MME to trigger the update of the path switch in the S-GW towards the target eNodeB. In contrast to the X2-handover, the message is not acknowledged and the resources at the source side are released later upon reception of a ‘RELEASE RESOURCE’ message directly triggered from the source MME (step 17 in Figure 2.16).

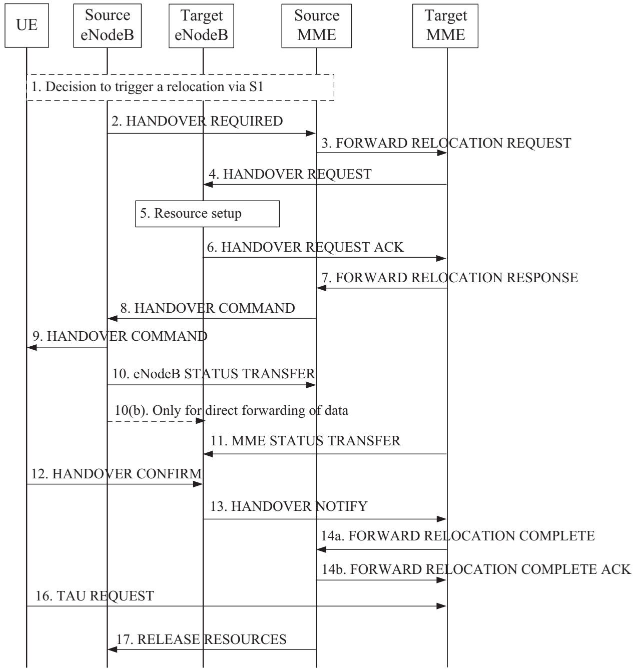

Figure 2.16: S1-based handover procedure. Reproduced by permission of © 3GPP.

2.5.6.2 Inter-RAT Mobility

One key element of the design of LTE is the need to co-exist with other Radio Access Technologies (RATs).

For mobility from LTE towards UMTS, the handover process can reuse the S1-handover procedures described above, with the exception of the ‘STATUS TRANSFERŠ message which is not needed at steps 10 and 11 since no PDCP context is continued.

For mobility towards CDMA2000, dedicated uplink and downlink procedures have been introduced in LTE. They essentially aim at tunnelling the CDMA2000 signalling between the UE and the CDMA2000 system over the S1 interface, without being interpreted by the eNodeB on the way. An ‘UPLINK S1 CDMA2000 TUNNELLING’ message is sent from the eNodeB to the MME; this also includes the RAT type in order to identify which CDMA2000

RAT the tunnelled CDMA2000 message is associated with in order for the message to be routed to the correct node within the CDMA2000 system.

2.5.6.3 Mobility towards Home eNodeBs

Mobility towards HeNBs involves additional functions from the source LTE RAN node and the MME. In addition to the E-UTRAN Cell Global Identifier (ECGI), the source RAN node should include the Closed Subscriber Group Identity (CSG ID) and the access mode of the target HeNB in the 'HANDOVER REQUIRED' message to the MME so that the MME can perform the access control to that HeNB. If the target HeNB operates in closed access mode (see Chapter 24) and the MME fails the access control, the MME will reject the handover by sending back a 'HANDOVER PREPARATION FAILURE' message. Otherwise the MME will accept and continue the handover while indicating to the target HeNB whether the UE is a 'CSG member' if the HeNB is operating in hybrid mode. A detailed description of mobility towards the HeNB and the associated call flow is provided in Chapter 24.

2.5.7 Load Management over S1

Three types of load management procedures apply over S1: a normal 'load balancing' procedure to distribute the traffic, an 'overload' procedure to overcome a sudden peak in the loading and a 'load rebalancing' procedure to partially/fully offload an MME.

The MME load balancing procedure aims to distribute the traffic to the MMEs in the pool evenly according to their respective capacities. To achieve that goal, the procedure relies on the normal NNSF present in each eNodeB as part of the S1-flex function. Provided that suitable weight factors corresponding to the capacity of each MME node are available in the eNodeBs beforehand, a weighted NNSF done by every eNodeB in the network normally achieves a statistically balanced distribution of load among the MME nodes without further action. However, specific actions are still required for some particular scenarios:

- If a new MME node is introduced (or removed), it may be necessary temporarily to increase (or decrease) the weight factor normally corresponding to the capacity of this node in order to make it catch more (or less) traffic at the beginning until it reaches an adequate level of load.

- In case of an unexpected peak in the loading, an 'OVERLOAD' message can be sent over the S1 interface by the overloaded MME. When received by an eNodeB, this message calls for a temporary restriction of a certain type of traffic. An MME can adjust the reduction of traffic it desires by defining the number of eNodeBs to which it sends the 'OVERLOAD' message and by defining the types of traffic subject to restriction.Two new rejection types are introduced in Release 10 to combat CN Overload:

    - 'reject low priority access', which can be used by the MME to reduce access of some low-priority devices or applications such as Machine-Type Communication (MTC) devices (see Section 31.4);
    - 'permit high priority sessions', to allow access only to high-priority users and mobile-terminated services.

- Finally, if the MME wants to force rapidly the offload of part or all of its UEs, it will use the rebalancing function. This function forces the UEs to reattach to another MME by using a specific 'cause value' in the 'UE Release Command S1' message. In a first step it applies to idle mode UEs and in a second step it may also apply to UEs in connected mode (if the full MME offload is desired, e.g. for maintenance reasons).

2.5.8 Trace Function

In order to trace the activity of a UE in connected mode, two types of trace session can be started in the eNodeB:

- Signalling-Based Trace. This is triggered by the MME and is uniquely identified by a trace identity. Only one trace session can be activated at a time for one UE. The MME indicates to the eNodeB the interfaces to trace (e.g. S1, X2, Uu) and the associated trace depth. The trace depth represents the granularity of the signalling to be traced from the high-level messages down to the detailed ASN.1[13] and is comprised of three levels: minimum, medium and maximum. The MME also indicates the IP address of a Trace Collection Entity where the eNodeB must send the resulting trace record file. If an X2 handover preparation has started at the time when the eNodeB receives the order to trace, the eNodeB will signal back a TRACE FAILURE INDICATION message to the MME, and it is then up to the MME to take appropriate action based on the indicated failure reason. Signalling-based traces are propagated at X2 and S1 handover.

- Management-Based Trace. This is triggered in the eNodeB when the conditions required for tracing set by O&M are met. The eNodeB then allocates a trace identity that it sends to the MME in a CELL TRAFFIC TRACE message over S1, together with the Trace Collection Entity identity that shall be used by the MME for the trace record file (in order to assemble the trace correctly in the Trace Collection Entity). Management-based traces are propagated at X2 and S1 handover.

In Release 10, the trace function supports the Minimization of Drive Tests (MDT) feature, which is explained in Section 31.3.

2.5.9 Delivery of Warning Messages

Two types of warning message may need to be delivered with the utmost urgency over a cellular system, namely Earthquake and Tsunami Warning System (ETWS)) messages and Commercial Mobile Alert System (CMAS) messages (see Section 13.7). The delivery of ETWS messages is already supported since Release 8 via the S1 Write-Replace Warning procedure which makes it possible to carry either primary or secondary notifications over S1 for the eNodeB to broadcast over the radio. The Write-Replace Warning procedure also includes a Warning Area List where the warning message needs to be broadcast. It can be a list of cells, tracking areas or emergency area identities. The procedure also contains information on how the broadcast is to be performed (for example, the number of broadcasts requested).

[13] Abstract Syntax Notation One

In contrast to ETWS, the delivery of CMAS messages is only supported from Release 9 onwards. One difference between the two public warning systems is that in ETWS the eNodeB can only broadcast one message at a time, whereas CMAS allows the broadcast of multiple concurrent warning messages over the radio. Therefore an ongoing ETWS broadcast needs to be overwritten if a new ETWS warning has to be delivered immediately in the same cell. With CMAS, a new Kill procedure has also been added to allow easy cancellation of an ongoing broadcast when needed. This Kill procedure includes the identity of the message to be stopped and the Warning Area where it is to be stopped.

2.6 The E-UTRAN Network Interfaces: X2 Interface

The X2 interface is used to inter-connect eNodeBs. The protocol structure for the X2 interface and the functionality provided over X2 are discussed below.

2.6.1 Protocol Structure over X2

The control plane and user plane protocol stacks over the X2 interface are the same as over the S1 interface, as shown in Figures 2.17 and 2.18 respectively (with the exception that in Figure 2.17 the X2-AP (X2 Application Protocol) is substituted for the S1-AP). This also means again that the choice of the IP version and the data link layer are fully optional. The use of the same protocol structure over both interfaces provides advantages such as simplifying the data forwarding operation.

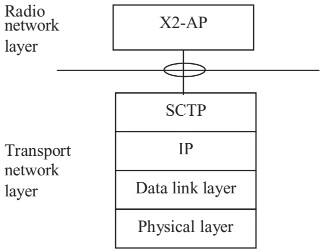

Figure 2.17: X2 signalling bearer protocol stack. Reproduced by permission of © 3GPP.

2.6.2 Initiation over X2

The X2 interface may be established between one eNodeB and some of its neighbour eNodeBs in order to exchange signalling information when needed. However, a full mesh is not mandated in an E-UTRAN network. Two types of information may typically need to be exchanged over X2 to drive the establishment of an X2 interface between two eNodeBs: load or interference related information (see Section 2.6.4) and handover related information (see mobility in Section 2.6.3).

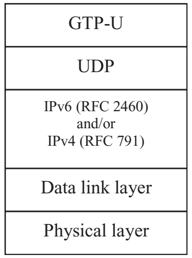

Figure 2.18: Transport network layer for data streams over X2. Reproduced by permission of © 3GPP.

Because these two types of information are fully independent of one another, it is possible that an X2 interface may be present between two eNodeBs for the purpose of exchanging load or interference information, even though the X2-handover procedure is not used to hand over UEs between those eNodeBs.[14]

The initialization of the X2 interface starts with the identification of a suitable neighbour followed by the setting up of the TNL.

The identification of a suitable neighbour may be done by configuration, or alternatively by a self-optimizing process known as the Automatic Neighbour Relation Function (ANRF).[15] This is described in more detail in Section 25.2.

Once a suitable neighbour has been identified, the initiating eNodeB can further set up the TNL using the transport layer address of this neighbour – either as retrieved from the network or locally configured. The automatic retrieval of the X2 IP address(es) via the network and the eNodeB Configuration Transfer procedure are described in details in Section 25.3.2.

Once the TNL has been set up, the initiating eNodeB must trigger the X2 setup procedure. This procedure enables an automatic exchange of application level configuration data relevant to the X2 interface, similar to the S1 setup procedure already described in Section 2.5.2. For example, each eNodeB reports within the 'X2 SETUP REQUESTŠ message to a neighbour eNodeB information about each cell it manages, such as the cell's physical identity, the frequency band, the tracking area identity and/or the associated PLMNs.

This automatic exchange of application-level configuration data within the X2 setup procedure is also the core of two additional SON features: automatic self-configuration of the Physical Cell Identities (PCIs) and RACH self-optimization. These features both aim to avoid conflicts between cells controlled by neighbouring eNodeBs; they are explained in detail in Sections 25.4 and 25.7 respectively.

Once the X2 setup procedure has been completed, the X2 interface is operational.

[14] In such a case, the S1-handover procedure is used instead.

[15] Under this function the UEs are requested to detect neighbour eNodeBs by reading the Cell Global Identity (CGI) contained in the broadcast information.

2.6.3 Mobility over X2

Handover via the X2 interface is triggered by default unless there is no X2 interface established or the source eNodeB is configured to use the S1-handover instead.

The X2-handover procedure is illustrated in Figure 2.19. Like the S1-handover, it is also composed of a preparation phase (steps 4 to 6), an execution phase (steps 7 to 9) and a completion phase (after step 9).

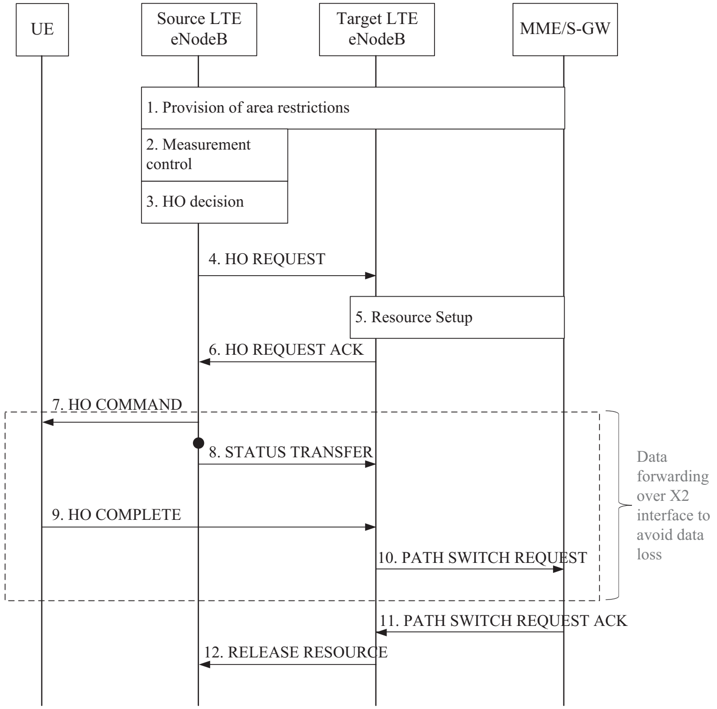

Figure 2.19: X2-based handover procedure.

The key features of the X2-handover for intra-LTE handover are:

- The handover is directly performed between two eNodeBs. This makes the preparation phase quick.

- Data forwarding may be operated per bearer in order to minimize data loss.
- The MME is only informed at the end of the handover procedure once the handover is successful, in order to trigger the path switch.
- The release of resources at the source side is directly triggered from the target eNodeB.

For those bearers for which in-sequence delivery of packets is required, the 'STATUS TRANSFER' message (step 8) provides the Sequence Number (SN) and the Hyper Frame Number (HFN) which the target eNodeB should assign to the first packet with no sequence number yet assigned that it must deliver. This first packet can either be one received over the target S1 path or one received over X2 if data forwarding over X2 is used (see below). When it sends the 'STATUS TRANSFER' message, the source eNodeB freezes its transmitter/receiver status – i.e. it stops assigning PDCP SNs to downlink packets and stops delivering uplink packets to the EPC.

Mobility over X2 can be categorized according to its resilience to packet loss: the handover can be said 'seamless' if it minimizes the interruption time during the move of the UE, or 'lossless' if it tolerates no loss of packets at all. These two modes use data forwarding of user plane downlink packets. The source eNodeB may decide to operate one of these two modes on a per-EPS-bearer basis, based on the QoS received over S1 for this bearer (see Section 2.5.4) and the service in question. These two modes are described in more detail below.

2.6.3.1 Seamless Handover

If, for a given bearer, the source eNodeB selects the seamless handover mode, it proposes to the target eNodeB in the 'HANDOVER REQUEST' message to establish a GTP tunnel to operate the downlink data forwarding. If the target eNodeB accepts, it indicates in the 'HANDOVER REQUEST ACK' message the tunnel endpoint where the forwarded data is expected to be received. This tunnel endpoint may be different from the one set up as the termination point of the new bearer established over the target S1.

Upon reception of the 'HANDOVER REQUEST ACK' message, the source eNodeB can start forwarding the data freshly arriving over the source S1 path towards the indicated tunnel endpoint in parallel with sending the handover trigger to the UE over the radio interface. This forwarded data is thus available at the target eNodeB to be delivered to the UE as early as possible.

When forwarding is in operation and in-sequence delivery of packets is required, the target eNodeB is assumed to deliver first the packets forwarded over X2 before delivering the first ones received over the target S1 path once the S1 path switch has been performed. The end of the forwarding is signalled over X2 to the target eNodeB by the reception of some 'special GTP packets' which the S-GW has inserted over the source S1 path just before switching this S1 path; these are then forwarded by the source eNodeB over X2 like any other regular packets.

2.6.3.2 Lossless Handover

If the source eNodeB selects the lossless mode for a given bearer, it will additionally forward over X2 those user plane downlink packets which it has PDCP processed but are still buffered locally because they have not yet been delivered and acknowledged by the UE. These packets are forwarded together with their assigned PDCP SN included in a GTP extension header field. They are sent over X2 prior to the freshly arriving packets from the source S1 path. The same mechanisms described above for the seamless handover are used for the GTP tunnel establishment. The end of forwarding is also handled in the same way, since in-sequence packet delivery applies to lossless handovers. In addition, the target eNodeB must ensure that all the packets – including the ones received with sequence number over X2 – are delivered in sequence at the target side. Further details of seamless and lossless handover are described in Section 4.2.

Selective retransmission. A new feature in LTE compared to previous systems is the optimization of the radio interface usage by selective retransmission. When lossless handover is operated, the target eNodeB may, however, not deliver over the radio interface some of the forwarded downlink packets received over X2 if it is informed by the UE that those packets have already been received at the source side (see Section 4.2.6). This is called downlink selective retransmission.

Similarly in the uplink, the target eNodeB may desire that the UE does not retransmit packets already received earlier at the source side by the source eNodeB, for example to avoid wasting radio resources. To operate this uplink selective retransmission scheme for one bearer, it is necessary that the source eNodeB forwards to the target eNodeB, over another new GTP tunnel, those user plane uplink packets which it has received out of sequence. The target eNodeB must first request the source eNodeB to establish this new forwarding tunnel by including in the 'HANDOVER REQUEST ACK' message a GTP tunnel endpoint where it expects the forwarded uplink packets to be received. The source eNodeB must, if possible, then indicate in the 'STATUS TRANSFER' message for this bearer the list of SNs corresponding to the forwarded packets which are to be expected. This list helps the target eNodeB to inform the UE earlier of the packets not to be retransmitted, making the overall uplink selective retransmission scheme faster (see also Section 4.2.6).

2.6.3.3 Multiple Preparation

'Multiple preparation' is another new feature of the LTE handover procedure. This feature enables the source eNodeB to trigger the handover preparation procedure towards multiple candidate target eNodeBs. Even though only one of the candidates is indicated as target to the UE, this makes recovery faster in case the UE fails on this target and connects to one of the other prepared candidate eNodeBs. The source eNodeB receives only one 'RELEASE RESOURCE' message from the final selected eNodeB.

Regardless of whether multiple or single preparation is used, the handover can be cancelled during or after the preparation phase. If the multiple preparation feature is operated, it is recommended that upon reception of the 'RELEASE RESOURCE' message the source eNodeB triggers a 'cancel' procedure towards each of the non-selected prepared eNodeBs.

2.6.3.4 Mobility Robustness Handling

In order to detect and report the cases where the mobility is unsuccessful and results in connection failures, specific messages are available over the X2 interface from Release 9 onwards to report handovers that are triggered too late or too early or to an inappropriate cell. These scenarios are explained in detail in Section 25.6.

2.6.3.5 Mobility towards Home eNodeBs via X2

In Release 10, in order to save backhaul bandwidth reduce delays, mobility between two HeNBs does not necessarily need to use S1 handover and transit via the MME but can directly use the X2 handover. This optimization is described in detail in Section 24.2.3.

2.6.4 Load and Interference Management Over X2

The exchange of load information between eNodeBs is of key importance in the flat architecture used in LTE, as there is no central Radio Resource Management (RRM) node as was the case, for example, in UMTS with the Radio Network Controller (RNC).

The exchange of load information falls into two categories depending on the purpose it serves:

- **Load balancing.** If the exchange of load information is for the purpose of load balancing, the frequency of exchange is rather low (in the order of seconds). The objective of load balancing is to counteract local traffic load imbalance between neighbouring cells with the aim of improving the overall system capacity. The mechanisms for this are explained in detail in Section 25.5.

  In Release 10, partial reporting is allowed per cell and per measurement. Therefore, if a serving eNodeB does not support some measurements, it will still report the other measurements that it does support. For each unsupported measurement, the serving eNodeB can indicate if the lack of support is permanent or temporary.

- **Interference coordination.** If the exchange of load information is to optimize RRM processes such as interference coordination, the frequency of exchange is rather high (in the order of tens of milliseconds). A special X2 'LOAD INDICATION' message is provided over the X2 interface for the exchange of load information related to interference management. For uplink interference management, two indicators can be provided within the 'LOAD INDICATION' message: a 'High Interference Indicator' and an 'Overload Indicator'. The usage of these indicators is explained in detail in Section 12.5.

The Load Indication procedure allows an eNodeB to signal to its neighbour eNodeBs new interference coordination intentions when applicable. This can either be frequency-domain interference management, as explained in Sections 12.5.1 and 12.5.2, or time-domain interference management, as explained in Section 31.2.3.

2.6.5 UE Historical Information Over X2

The provision of UE historical information is part of the X2-handover procedure and is designed to support self-optimization of the network.

Generally, the UE historical information consists of some RRM information which is passed from the source eNodeB to the target eNodeB within the 'HANDOVER REQUEST' message to assist the RRM management of a UE. The information can be partitioned into two types:

- UE RRM-related information, passed over X2 within the RRC transparent container;

- Cell RRM-related information, passed over X2 directly as an information element of the ‘X2 AP HANDOVER REQUEST’ message itself.

An example of such UE historical information is the list of the last few cells visited by the UE, together with the time spent in each one. This information is propagated from one eNodeB to another and can be used to determine the occurrence of ping-pong between two or three cells for instance. The length of the history information can be configured for more flexibility.

2.7 Summary

The EPS provides UEs with IP connectivity to the packet data network. In this chapter we have seen an overview of the EPS network architecture, including the functionalities provided by the E-UTRAN access network and the evolved packet core network..

It can be seen that the concept of EPS bearers, together with their associated quality of service attributes, provide a powerful tool for the provision of a variety of simultaneous services to the end user. Depending on the nature of the application, the EPS can supply the UE with multiple data flows with different QoSs. A UE can thus be engaged in a VoIP call which requires guaranteed delay and bit rate at the same time as browsing the web with a best effort QoS.

From the perspective of the network operator, the LTE system breaks new ground in terms of its degree of support for self-optimization and self-configuration of the network via the X2, S1 and Uu interfaces; these aspects are described in more detail in Chapter 25.

References[16]

[1] 3GPP Technical Specification 24.301, ‘Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3’, www.3gpp.org.

[2] 3GPP Technical Specification 33.401, ‘System Architecture Evolution (SAE): Security Architecture’, www.3gpp.org.

[3] 3GPP Technical Specification 23.402, ‘Architecture enhancements for non-3GPP accesses’, www.3gpp.org.

[4] 3GPP Technical Specification 29.060, ‘General Packet Radio Service (GPRS); GPRS Tunnelling Protocol (GTP) across the Gn and Gp interface’, www.3gpp.org.

[5] 3GPP Technical Specification 23.203, ‘Policy and charging control architecture’, www.3gpp.org.

[6] 3GPP Technical Specification 36.300, ‘Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2’, www.3gpp.org.

[7] 3GPP Technical Specification 23.272, ‘ Circuit Switched (CS) fallback in Evolved Packet System (EPS); Stage 2’, www.3gpp.org.

[8] 3GPP Technical Specification 23.272, ‘Single Radio Voice Call Continuity (SRVCC); Stage 2’, www.3gpp.org.

[9] Request for Comments 4960 The Internet Engineering Task Force (IETF), Network Working Group, ‘Stream Control Transmission Protocol’, http://www.ietf.org.

[16] All web sites confirmed 1st March 2011.