3

System Architecture Based on 3GPP SAE

Atte Länsisalmi and Antti Toskala

3.1 System Architecture Evolution in 3GPP

When the evolution of the radio interface started, it soon became clear that the system architecture would also need to be evolved. The general drive towards optimizing the system only for packet switched services is one reason that alone would have set the need for evolution, but some of the radio interface design goals – such as removal of soft handover – opened up new opportunities in the architecture design. Also, since it had been shown by High Speed Packet Access (HSPA) that all radio functionality can be efficiently co-located in the NodeB, the door was left open for discussions of flatter overall architecture.

Discussions for System Architecture Evolution (SAE) then soon followed the radio interface development, and it was agreed to schedule the completion of the work in Release 8. There had been several reasons for starting this work, and there were also many targets. The following lists some of the targets that possibly shaped the outcome the most:

- optimization for packet switched services in general, when there is no longer a need to support the circuit switched mode of operation;
- optimized support for higher throughput required for higher end user bit rates;
- improvement in the response times for activation and bearer set-up;
- improvement in the packet delivery delays;
- overall simplification of the system compared to the existing 3GPP and other cellular systems;
- optimized inter-working with other 3GPP access networks;
- optimized inter-working with other wireless access networks.

Many of the targets implied that a flat architecture would need to be developed. Flat architecture with less involved nodes reduces latencies and improves performance.

LTE for UMTS: Evolution to LTE-Advanced, Second Edition. Edited by Harri Holma and Antti Toskala.

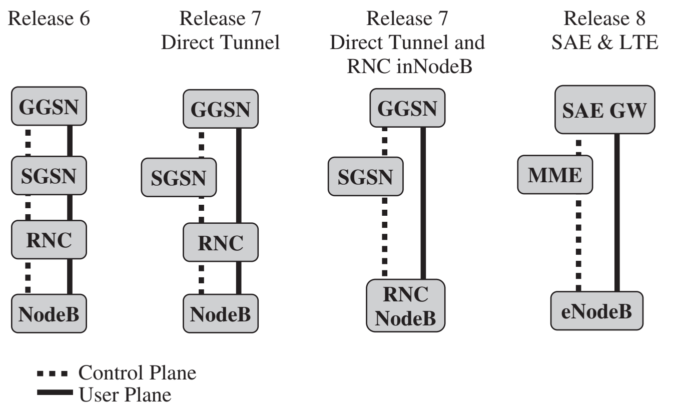

Figure 3.1 3GPP architecture evolution towards flat architecture

Development towards this direction had already started in Release 7 where the Direct Tunnel concept allows User Plane (UP) to bypass the SGSN, and the placement of RNC functions to HSPA NodeB was made possible. Figure 3.1 shows these evolution steps and how this aspect was captured at a high level in SAE architecture.

Some of the targets seem to drive the architecture development in completely different directions. For example, optimized inter-working with several wireless access networks (ANs) indicates the need to introduce a set of new functions and maybe even new interfaces to support specific protocols separately for each one of them. This works against the target of keeping the architecture simple. Therefore, since it is likely that that none of the actual deployments of the architecture would need to support all of the potential inter-working scenarios, the 3GPP architecture specifications were split into two tracks:

- GPRS enhancements for E-UTRAN access [1]: This document describes the architecture and its functions in its native 3GPP environment with E-UTRAN and all the other 3GPP ANs, and defines the inter-working procedures between them. The common nominator for these ANs is the use of GTP (GPRS Tunnelling Protocol) as the network mobility protocol.
- Architecture enhancements for non-3GPP accesses [2]: This document describes the architecture and functions when inter-working with non-3GPP ANs, such as cdma2000® High Rate Packet Data (HRPD), is needed. The mobility functionality in this document is based on IETF protocols, such as MIP (Mobile Internet Protocol) and PMIP (Proxy MIP), and the document also describes E-UTRAN in that protocol environment.

This chapter further describes the 3GPP system architecture in some likely deployment scenarios: basic scenario with only E-UTRAN, legacy 3GPP operator scenario with existing 3GPP ANs and E-UTRAN, and finally E-UTRAN with non-3GPP ANs, where inter-working with cdma2000® is shown as a specific example.

3.2 Basic System Architecture Configuration with only E-UTRAN Access Network

3.2.1 Overview of Basic System Architecture Configuration

Figure 3.2 describes the architecture and network elements in the architecture configuration where only the E-UTRAN AN is involved. The logical nodes and connections shown in this figure represent the basic system architecture configuration. These elements and functions are needed in all cases when E-UTRAN is involved. The other system architecture configurations described in the next sections also include some additional functions.

This figure also shows the division of the architecture into four main high level domains: User Equipment (UE), Evolved UTRAN (E-UTRAN), Evolved Packet Core Network (EPC), and the Services domain.

The high level architectural domains are functionally equivalent to those in the existing 3GPP systems. The new architectural development is limited to Radio Access and Core Networks, the E-UTRAN and the EPC respectively. UE and Services domains remain architecturally intact, but functional evolution has also continued in those areas.

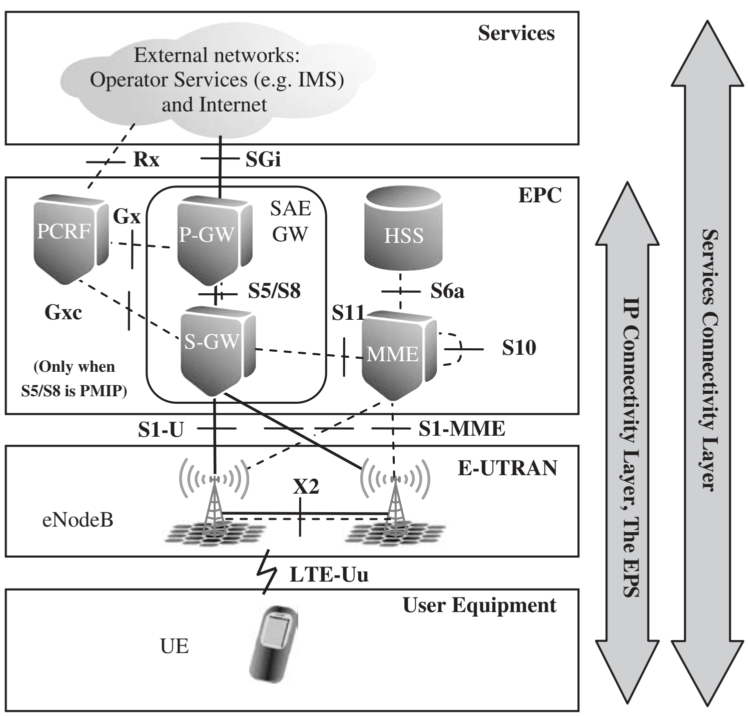

Figure 3.2 System architecture for E-UTRAN only network

UE, E-UTRAN and EPC together represent the Internet Protocol (IP) Connectivity Layer. This part of the system is also called the Evolved Packet System (EPS). The main function of this layer is to provide IP based connectivity, and it is highly optimized for that purpose only. All services will be offered on top of IP, and circuit switched nodes and interfaces seen in earlier 3GPP architectures are not present in E-UTRAN and EPC at all. IP technologies are also dominant in the transport, where everything is designed to be operated on top of IP transport.

The IP Multimedia Sub-System (IMS) [3] is a good example of service machinery that can be used in the Services Connectivity Layer to provide services on top of the IP connectivity provided by the lower layers. For example, to support the voice service, IMS can provide Voice over IP (VoIP) and interconnectivity to legacy circuit switched networks PSTN and ISDN through Media Gateways it controls.

The development in E-UTRAN is concentrated on one node, the evolved Node B (eNodeB). All radio functionality is collapsed there, i.e. the eNodeB is the termination point for all radio related protocols. As a network, E-UTRAN is simply a mesh of eNodeBs connected to neighbouring eNodeBs with the X2 interface.

One of the big architectural changes in the core network area is that the EPC does not contain a circuit switched domain, and no direct connectivity to traditional circuit switched networks such as ISDN or PSTN is needed in this layer. Functionally the EPC is equivalent to the packet switched domain of the existing 3GPP networks. There are, however, significant changes in the arrangement of functions and most nodes and the architecture in this part should be considered to be completely new.

Both Figure 3.1 and Figure 3.2 show an element called SAE GW. As the latter figure indicates, this represents the combination of the two gateways, Serving Gateway (S-GW) and Packet Data Network Gateway (P-GW) defined for the UP handling in EPC. Implementing them together as the SAE GW represents one possible deployment scenario, but the standards define the interface between them, and all operations have also been specified for when they are separate. The same approach is followed in this chapter of the book.

The Basic System Architecture Configuration and its functionality are documented in 3GPP TS 23.401 [1]. This document shows the operation when the S5/S8 interface uses the GTP protocol. However, when the S5/S8 interface uses PMIP, the functionality for these interfaces is slightly different, and the Gxc interface also is needed between the Policy and Charging Resource Function (PCRF) and S-GW. The appropriate places are clearly marked in [1] and the additional functions are described in detail in 3GPP TS 23.402 [2]. In the following sections the functions are described together for all cases that involve E-UTRAN.

3.2.2 *Logical Elements in Basic System Architecture Configuration*

This section introduces the logical network elements for the Basic System Architecture configuration.

3.2.2.1 User Equipment (UE)

UE is the device that the end user uses for communication. Typically it is a hand held device such as a smart phone or a data card such as those used currently in 2G and 3G, or it could be embedded, e.g. to a laptop. UE also contains the Universal Subscriber

Identity Module (USIM) that is a separate module from the rest of the UE, which is often called the Terminal Equipment (TE). USIM is an application placed into a removable smart card called the Universal Integrated Circuit Card (UICC). USIM is used to identify and authenticate the user and to derive security keys for protecting the radio interface transmission.

Functionally the UE is a platform for communication applications, which signal with the network for setting up, maintaining and removing the communication links the end user needs. This includes mobility management functions such as handovers and reporting the terminals location, and in these the UE performs as instructed by the network. Maybe most importantly, the UE provides the user interface to the end user so that applications such as a VoIP client can be used to set up a voice call.

3.2.2.2 E-UTRAN Node B (eNodeB)

The only node in the E-UTRAN is the E-UTRAN Node B (eNodeB). Simply put, the eNodeB is a radio base station that is in control of all radio related functions in the fixed part of the system. Base stations such as eNodeB are typically distributed throughout the networks coverage area, each eNodeB residing near the actual radio antennas.

Functionally eNodeB acts as a layer 2 bridge between UE and the EPC, by being the termination point of all the radio protocols towards the UE, and relaying data between the radio connection and the corresponding IP based connectivity towards the EPC. In this role, the eNodeB performs ciphering/deciphering of the UP data, and also IP header compression/decompression, which means avoiding repeatedly sending the same or sequential data in IP header.

The eNodeB is also responsible for many Control Plane (CP) functions. The eNodeB is responsible for the Radio Resource Management (RRM), i.e. controlling the usage of the radio interface, which includes, for example, allocating resources based on requests, prioritizing and scheduling traffic according to required Quality of Service (QoS), and constant monitoring of the resource usage situation.

In addition, the eNodeB has an important role in Mobility Management (MM). The eNodeB controls and analyses radio signal level measurements carried out by the UE, makes similar measurements itself, and based on those makes decisions to handover UEs between cells. This includes exchanging handover signalling between other eNodeBs and the MME. When a new UE activates under eNodeB and requests connection to the network, the eNodeB is also responsible for routing this request to the MME that previously served that UE, or selecting a new MME, if a route to the previous MME is not available or routing information is absent.

Details of these and other E-UTRAN radio interface functions are described extensively elsewhere in this book. The eNodeB has a central role in many of these functions.

Figure 3.3 shows the connections that eNodeB has to the surrounding logical nodes, and summarizes the main functions in these interfaces. In all the connections the eNodeB may be in a one-to-many or a many-to-many relationship. The eNodeB may be serving multiple UEs at its coverage area, but each UE is connected to only one eNodeB at a time. The eNodeB will need to be connected to those of its neighbouring eNodeBs with which a handover may need to be made.

Both MMEs and S-GWs may be pooled, which means that a set of those nodes is assigned to serve a particular set of eNodeBs. From a single eNodeB perspective this

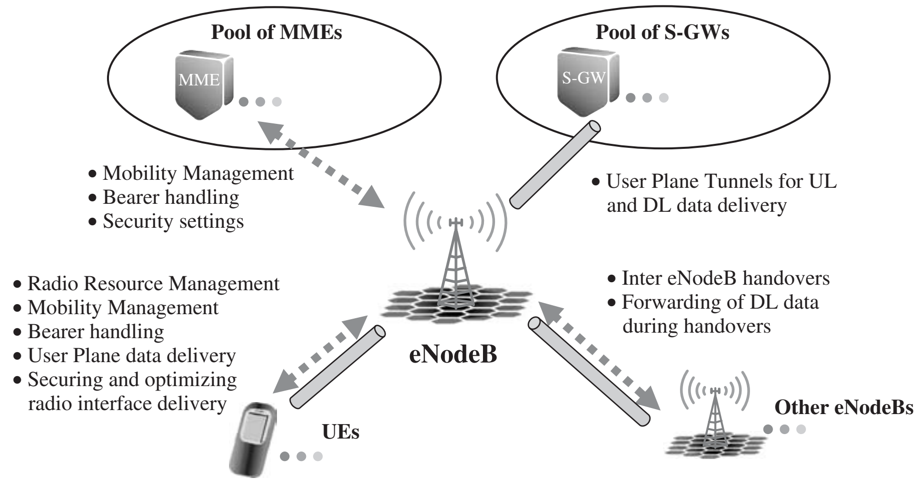

Figure 3.3 eNodeB connections to other logical nodes and main functions

means that it may need to connect to many MMEs and S-GWs. However, each UE will be served by only one MME and S-GW at a time, and the eNodeB has to keep track of this association. This association will never change from a single eNodeB point of view, because MME or S-GW can only change in association with inter-eNodeB handover.

3.2.2.3 Mobility Management Entity (MME)

Mobility Management Entity (MME) is the main control element in the EPC. Typically the MME would be a server in a secure location in the operator's premises. It operates only in the CP, and is not involved in the path of UP data.

In addition to interfaces that terminate to MME in the architecture as shown in Figure 3.2, the MME also has a logically direct CP connection to the UE, and this connection is used as the primary control channel between the UE and the network. The following lists the main MME functions in the basic System Architecture Configuration:

- Authentication and Security: When a UE registers to the network for the first time, the MME initiates the authentication, by performing the following: it finds out the UE's permanent identity either from the previously visited network or the UE itself; requests from the Home Subscription Server (HSS) in UE's home network the authentication vectors which contain the authentication challenge – response parameter pairs; sends the challenge to the UE; and compares the response received from the UE to the one received from the home network. This function is needed to assure that the UE is who it claims to be. The details of EPS-AKA authentication are defined in [4]. The MME may repeat authentication when needed or periodically. The MME will calculate UEs ciphering and integrity protection keys from the master key received in the authentication vector from the home network, and it controls the related settings in E-UTRAN for UP and CP separately. These functions are used to protect the communication from eavesdropping and from alteration by unauthorized third parties respectively. To protect the UE privacy, MME also allocates each UE a temporary identity called the Globally Unique Temporary Identity (GUTI), so that the need to send the permanent UE identity – International Mobile Subscriber Identity (IMSI) – over the radio interface is minimized. The GUTI may be re-allocated, e.g. periodically to prevent unauthorized UE tracking.

- Mobility Management: The MME keeps track of the location of all UEs in its service area. When a UE makes its first registration to the network, the MME will create an entry for the UE, and signal the location to the HSS in the UE's home network. The MME requests the appropriate resources to be set up in the eNodeB, as well as in the S-GW which it selects for the UE. The MME will then keep tracking the UE's location either on the level of eNodeB, if the UE remains connected, i.e. is in active communication, or at the level of Tracking Area (TA), which is a group of eNodeBs in case the UE goes to idle mode, and maintaining a through connected data path is not needed. The MME controls the setting up and releasing of resources based on the UE's activity mode changes. The MME also participates in control signalling for handover of an active mode UE between eNodeBs, S-GWs or MMEs. MME is involved in every eNodeB change, since there is no separate Radio Network Controller to hide most of these events. An idle UE will report its location either periodically, or when it moves to another Tracking Area. If data are received from the external networks for an idle UE, the MME will be notified, and it requests the eNodeBs in the TA that is stored for the UE to page the UE.
- Managing Subscription Profile and Service Connectivity: At the time of a UE registering to the network, the MME will be responsible for retrieving its subscription profile from the home network. The MME will store this information for the duration it is serving the UE. This profile determines what Packet Data Network connections should be allocated to the UE at network attachment. The MME will automatically set up the default bearer, which gives the UE the basic IP connectivity. This includes CP signalling with the eNodeB, and the S-GW. At any point later on, the MME may need to be involved in setting up dedicated bearers for services that benefit from higher treatment. The MME may receive the request to set up a dedicated bearer either from the S-GW if the request originates from the operator service domain, or directly from the UE, if the UE requires a connection for a service that is not known by the operator service domain, and therefore cannot be initiated from there.

Figure 3.4 shows the connections MME has to the surrounding logical nodes, and summarizes the main functions in these interfaces. In principle the MME may be connected to any other MME in the system, but typically the connectivity is limited to one operator network only. The remote connectivity between MMEs may be used when a UE that has travelled far away while powered down registers to a new MME, which then retrieves the UE's permanent identity, the International Mobile Subscriber Identity (IMSI), from the previously visited MME. The inter-MME connection with neighbouring MMEs is used in handovers.

Connectivity to a number of HSSs will also need to be supported. The HSS is located in each user's home network, and a route to that can be found based on the IMSI. Each MME will be configured to control a set of S-GWs and eNodeBs. Both the S-GWs and eNodeBs may also be connected to other MMEs. The MME may serve a number of UEs at the same time, while each UE will only connect to one MME at a time.

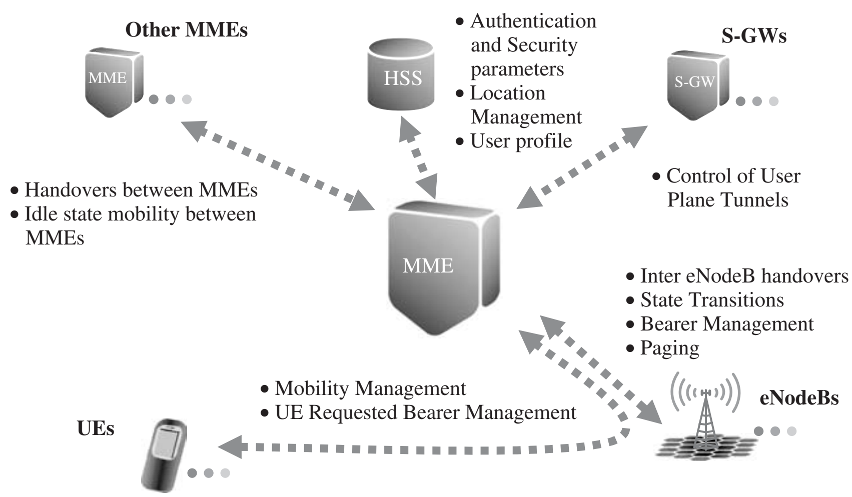

Figure 3.4 MME connections to other logical nodes and main functions

3.2.2.4 Serving Gateway (S-GW)

In the Basic System Architecture configuration, the high level function of S-GW is UP tunnel management and switching. The S-GW is part of the network infrastructure maintained centrally in operation premises.

When the S5/S8 interface is based on GTP, the S-GW will have GTP tunnels on all its UP interfaces. Mapping between IP service flows and GTP tunnels is done in P-GW, and the S-GW does not need to be connected to PCRF. All control is related to the GTP tunnels, and comes from either MME or P-GW. When the S5/S8 interface uses PMIP, the S-GW will perform the mapping between IP service flows in S5/S8 and GTP tunnels in S1-U interfaces, and will connect to PCRF to receive the mapping information.

The S-GW has a very minor role in control functions. It is only responsible for its own resources, and it allocates them based on requests from MME, P-GW or PCRF, which in turn are acting on the need to set up, modify or clear bearers for the UE. If the request was received from P-GW or PCRF, the S-GW will also relay the command on to the MME so that it can control the tunnel to eNodeB. Similarly, when the MME initiated the request, the S-GW will signal on to either the P-GW or the PCRF, depending on whether S5/S8 is based on GTP or PMIP respectively. If the S5/S8 interface is based on PMIP, the data in that interface will be IP flows in one GRE tunnel for each UE, whereas in the GTP based S5/S8 interface each bearer will have its own GTP tunnel. Therefore S-GW supporting PMIP S5/S8 is responsible for bearer binding, i.e. mapping the IP flows in S5/S8 interface to bearers in the S1 interface. This function in S-GW is called Bearer Binding and Event Reporting Function (BBERF). Irrespective of where the bearer signalling started, the BBERF always receives the bearer binding information from PCRF.

During mobility between eNodeBs, the S-GW acts as the local mobility anchor. The MME commands the S-GW to switch the tunnel from one eNodeB to another. The MME may also request the S-GW to provide tunnelling resources for data forwarding, when there is a need to forward data from source eNodeB to target eNodeB during the time UE makes the radio handover. The mobility scenarios also include changing from one S-GW to another, and the MME controls this change accordingly, by removing tunnels in the old S-GW and setting them up in a new S-GW.

For all data flows belonging to a UE in connected mode, the S-GW relays the data between eNodeB and P-GW. However, when a UE is in idle mode, the resources in eNodeB are released, and the data path terminates in the S-GW. If S-GW receives data packets from P-GW on any such tunnel, it will buffer the packets, and request the MME to initiate paging of the UE. Paging will cause the UE to re-connect, and when the tunnels are re-connected, the buffered packets will be sent on. The S-GW will monitor data in the tunnels, and may also collect data needed for accounting and user charging. The S-GW also includes functionality for Lawful Interception, which means the capability to deliver the monitored user's data to authorities for further inspection.

Figure 3.5 shows how the S-GW is connected to other logical nodes, and lists the main functions in these interfaces. All interfaces have to be configured in a one-to-many fashion from the S-GW point of view. One S-GW may be serving only a particular geographical area with a limited set of eNodeBs, and likewise there may be a limited set of MMEs that control that area. The S-GW should be able to connect to any P-GW in the whole network, because P-GW will not change during mobility, while the S-GW may be relocated, when the UE moves. For connections related to one UE, the S-GW will always signal with only one MME, and the UP points to one eNodeB at a time (indirect data forwarding is the exception, see next paragraph). If one UE is allowed to connect to multiple PDNs through different P-GWs, then the S-GW needs to connect to those separately. If the S5/S8 interface is based on PMIP, the S-GW connects to one PCRF for each separate P-GW the UE is using.

Figure 3.5 also shows the indirect data forwarding case where UP data is forwarded between eNodeBs through the S-GWs. There is no specific interface name associated to the interface between S-GWs, since the format is exactly the same as in the S1-U

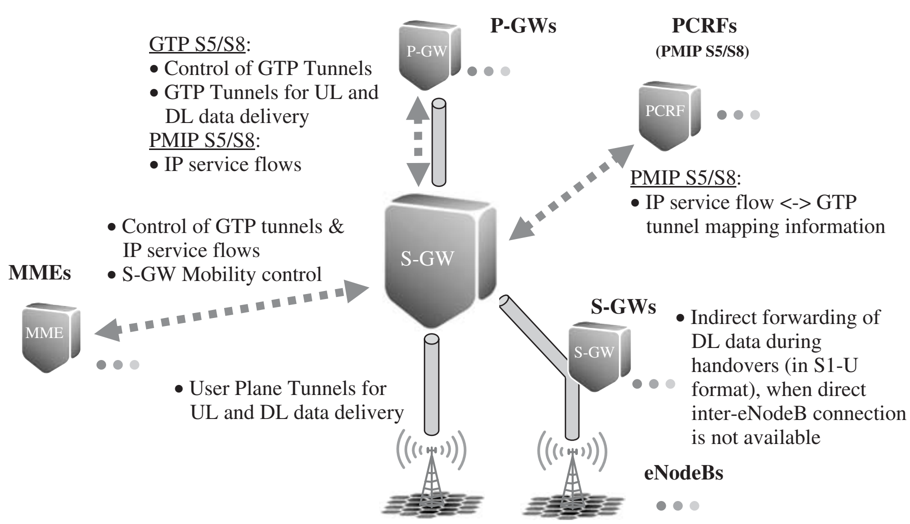

Figure 3.5 S-GW connections to other logical nodes and main functions

interface, and the involved S-GWs may consider that they are communicating directly with an eNodeB. This would be the case if indirect data forwarding takes place via only one S-GW, i.e. both eNodeBs can be connected to the same S-GW.

3.2.2.5 Packet Data Network Gateway (P-GW)

Packet Data Network Gateway (P-GW, also often abbreviated as PDN-GW) is the edge router between the EPS and external packet data networks. It is the highest level mobility anchor in the system, and usually it acts as the IP point of attachment for the UE. It performs traffic gating and filtering functions as required by the service in question. Similarly to the S-GW, the P-GWs are maintained in operator premises in a centralized location.

Typically the P-GW allocates the IP address to the UE, and the UE uses that to communicate with other IP hosts in external networks, e.g. the internet. It is also possible that the external PDN to which the UE is connected allocates the address that is to be used by the UE, and the P-GW tunnels all traffic to that network. The IP address is always allocated when the UE requests a PDN connection, which happens at least when the UE attaches to the network, and it may happen subsequently when a new PDN connectivity is needed. The P-GW performs the required Dynamic Host Configuration Protocol (DHCP) functionality, or queries an external DHCP server, and delivers the address to the UE. Also dynamic auto-configuration is supported by the standards. Only IPv4, only IPv6 or both addresses may be allocated depending on the need, and the UE may signal whether it wants to receive the address(es) in the Attach signalling, or if it wishes to perform address configuration after the link layer is connected.

The P-GW includes the PCEF, which means that it performs gating and filtering functions as required by the policies set for the UE and the service in question, and it collects and reports the related charging information.

The UP traffic between P-GW and external networks is in the form of IP packets that belong to various IP service flows. If the S5/S8 interface towards S-GW is based on GTP, the P-GW performs the mapping between the IP data flows to GTP tunnels, which represent the bearers. The P-GW sets up bearers based on request either through the PCRF or from the S-GW, which relays information from the MME. In the latter case, the P-GW may also need to interact with the PCRF to receive the appropriate policy control information, if that is not configured in the P-GW locally. If the S5/S8 interface is based on PMIP, the P-GW maps all the IP Service flows from external networks that belong to one UE to a single GRE tunnel, and all control information is exchanged with PCRF only. The P-GW also has functionality for monitoring the data flow for accounting purposes, as well as for Lawful Interception.

P-GW is the highest level mobility anchor in the system. When a UE moves from one S-GW to another, the bearers have to be switched in the P-GW. The P-GW will receive an indication to switch the flows from the new S-GW.

Figure 3.6 shows the connections P-GW has to the surrounding logical nodes, and lists the main functions in these interfaces. Each P-GW may be connected to one or more PCRF, S-GW and external network. For a given UE that is associated with the P-GW, there is only one S-GW, but connections to many external networks and respectively to many PCRFs may need to be supported, if connectivity to multiple PDNs is supported through one P-GW.

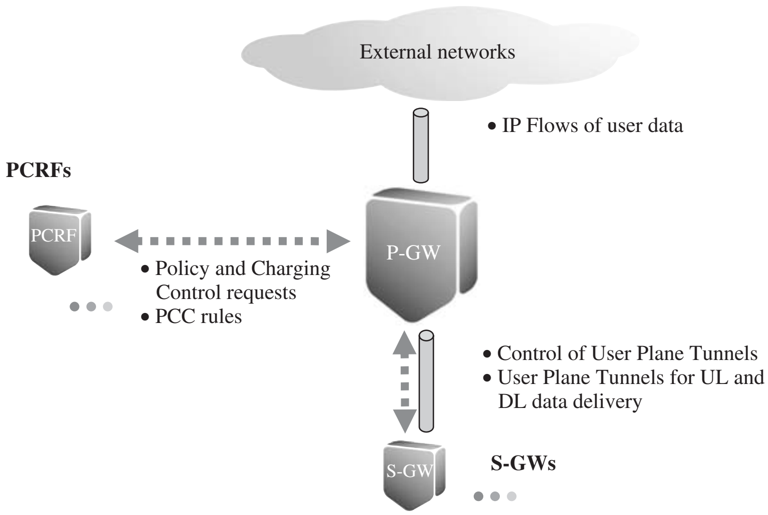

Figure 3.6 P-GW connections to other logical nodes and main functions

3.2.2.6 Policy and Charging Resource Function (PCRF)

Policy and Charging Resource Function (PCRF) is the network element that is responsible for Policy and Charging Control (PCC). It makes decisions on how to handle the services in terms of QoS, and provides information to the PCEF located in the P-GW, and if applicable also to the BBERF located in the S-GW, so that appropriate bearers and policing can be set up. PCRF is part of the PCC framework defined in [5]. PCRF is a server usually located with other CN elements in operator switching centres.

The information the PCRF provides to the PCEF is called the PCC rules. The PCRF will send the PCC rules whenever a new bearer is to be set up. Bearer set-up is required, for example, when the UE initially attaches to the network and the default bearer will be set up, and subsequently when one or more dedicated bearers are set up. The PCRF will be able to provide PCC rules based on request either from the P-GW and also the S-GW in PMIP case, like in the attach case, and also based on request from the Application Function (AF) that resides in the Services Domain. In this scenario the UE has signalled directly with the Services Domain, e.g. with the IMS, and the AF pushes the service QoS information to PCRF, which makes a PCC decision, and pushes the PCC rules to the P-GW, and bearer mapping information to S-GW in PMIP S5/S8 case. The EPC bearers are then set up based on those.

The connections between the PCRF and the other nodes are shown in Figure 3.7. Each PCRF may be associated with one or more AF, P-GW and S-GW. There is only one PCRF associated with each PDN connection that a single UE has.

3.2.2.7 Home Subscription Server (HSS)

Home Subscription Server (HSS) is the subscription data repository for all permanent user data. It also records the location of the user in the level of visited network control node, such as MME. It is a database server maintained centrally in the home operator's premises.

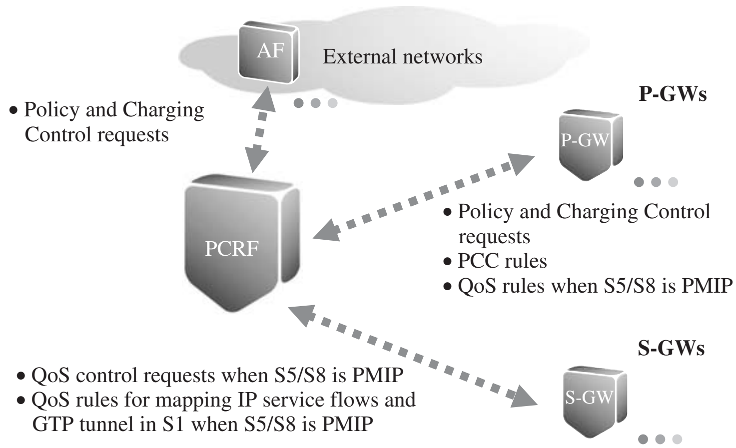

Figure 3.7 PCRF connections to other logical nodes and main functions

The HSS stores the master copy of the subscriber profile, which contains information about the services that are applicable to the user, including information about the allowed PDN connections, and whether roaming to a particular visited network is allowed or not. For supporting mobility between non-3GPP ANs, the HSS also stores the Identities of those P-GWs that are in use. The permanent key, which is used to calculate the authentication vectors that are sent to a visited network for user authentication and deriving subsequent keys for encryption and integrity protection, is stored in the Authentication Center (AuC), which is typically part of the HSS. In all signalling related to these functions, the HSS interacts with the MME. The HSS will need to be able to connect with every MME in the whole network, where its UEs are allowed to move. For each UE, the HSS records will point to one serving MME at a time, and as soon as a new MME reports that it is serving the UE, the HSS will cancel the location from the previous MME.

3.2.2.8 Services Domain

The Services domain may include various sub-systems, which in turn may contain several logical nodes. The following is a categorization of the types of services that will be made available, and a short description of what kind of infrastructure would be needed to provide them:

- IMS based operator services: The IP Multimedia Sub-system (IMS) is service machinery that the operator may use to provide services using the Session Initiation Protocol (SIP). IMS has 3GPP defined architecture of its own, and is described in section 3.6, and more thoroughly, e.g. in [3].
- Non-IMS based operator services: The architecture for non-IMS based operator services is not defined in the standards. The operator may simply place a server into their network, and the UEs connect to that via some agreed protocol that is supported by an application in the UE. A video streaming service provided from a streaming server is one such example.

- Other services not provided by the mobile network operator, e.g. services provided through the internet: This architecture is not addressed by the 3GPP standards, and the architecture depends on the service in question. The typical configuration would be that the UE connects to a server in the internet, e.g. to a web-server for web browsing services, or to a SIP server for internet telephony service (i.e. VoIP).

3.2.3 Self-configuration of S1-MME and X2 Interfaces

In 3GPP Release 8 development it has been agreed to define the support for self-configuration of the S1-MME and X2 interfaces. The basic process is as presented in Figure 3.8, where the eNodeB once turned on (and given that the IP connection exists) will connect to the O&M (based on the known IP address) to obtain then further parameters in terms of which other network elements to connect (and also for eNodeB software download) as well as initial parameters for the operation, such as in which part of the frequency band to operate and what kind of parameters to include for the broadcast channels.

This is expected to include setting the S1-MME connection by first setting up the SCTP association with at least one MME, and once that is connected to continue with application level information exchange to make S1-MME interface operational. Once the link to MME exists, there needs to be then association with S-GW created for UP data transfer.

To enable functionalities such as mobility and inter-cell interference control, the X2 interface configuration follows similar principles to the S1-MME interface. The difference here is that initially the eNodeB will set up the X2 connection for those eNodeBs indicated from the O&M and it may then later adapt more to the environment based on the Automatic Neighbour Relationship (ANR) functionality – as covered in Chapter 7 – to further optimize the X2 connectivity domain based on actual handover needs. The parameters that are exchanged over the X2 interface include:

- global eNodeB ID;
- information of the cell specific parameters such as Physical Cell ID (PCI), uplink/downlink frequency used, bandwidth in use;
- MMEs connected (MME Pool).

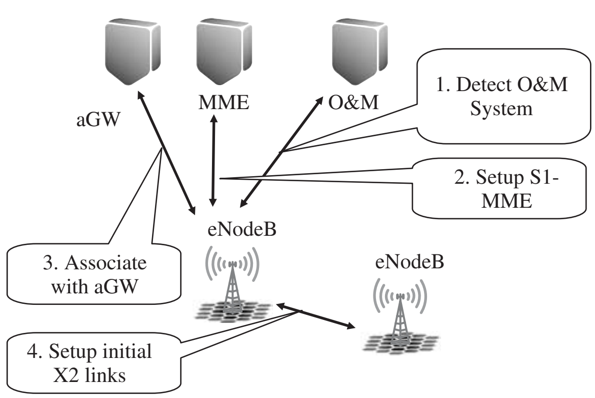

Figure 3.8 eNodeB self-configuration step

For the PCI there is also support for auto-configuration in the Release 8 specifications as covered in Chapter 5, other parameters then coming from the O&M direction with procedures that can be automated to limit the need for on-site configuration by installation personnel.

3.2.4 Interfaces and Protocols in Basic System Architecture Configuration

Figure 3.9 shows the CP protocols related to a UE's connection to a PDN. The interfaces from a single MME are shown in two parts, the one on top showing protocols towards the E-UTRAN and UE, and the bottom one showing protocols towards the gateways. Those protocols that are shown in white background are developed by 3GPP, while the protocols with light grey background are developed in IETF, and represent standard internet technologies that are used for transport in EPS. 3GPP has only defined the specific ways of how these protocols are used.

The topmost layer in the CP is the Non-Access Stratum (NAS), which consists of two separate protocols that are carried on direct signalling transport between the UE and the MME. The content of the NAS layer protocols is not visible to the eNodeB, and the eNodeB is not involved in these transactions by any other means, besides transporting the messages, and providing some additional transport layer indications along with the messages in some cases. The NAS layer protocols are:

- EPS Mobility Management (EMM): The EMM protocol is responsible for handling the UE mobility within the system. It includes functions for attaching to and detaching from

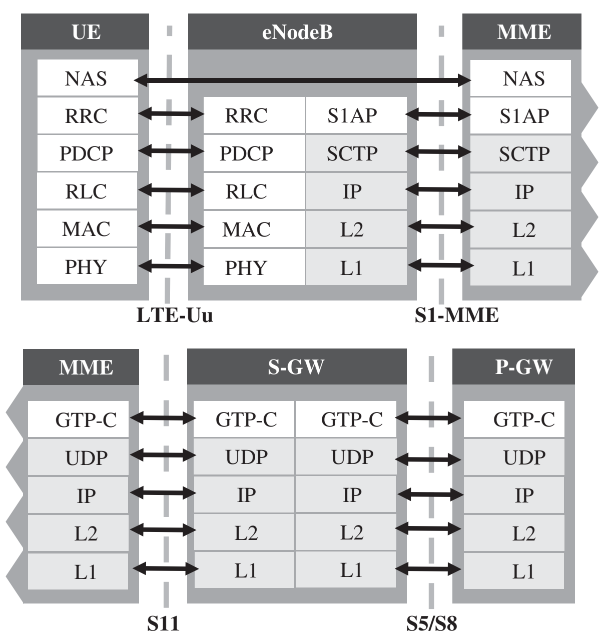

Figure 3.9 Control plane protocol stack in EPS

the network, and performing location updating in between. This is called Tracking Area Updating (TAU), and it happens in idle mode. Note that the handovers in connected mode are handled by the lower layer protocols, but the EMM layer does include functions for re-activating the UE from idle mode. The UE initiated case is called Service Request, while Paging represents the network initiated case. Authentication and protecting the UE identity, i.e. allocating the temporary identity GUTI to the UE are also part of the EMM layer, as well as the control of NAS layer security functions, encryption and integrity protection.
- EPS Session Management (ESM): This protocol may be used to handle the bearer management between the UE and MME, and it is used in addition for E-UTRAN bearer management procedures. Note that the intention is not to use the ESM procedures if the bearer contexts are already available in the network and E-UTRAN procedures can be run immediately. This would be the case, for example, when the UE has already signalled with an operator affiliated Application Function in the network, and the relevant information has been made available through the PCRF.

The radio interface protocols are (only short descriptions are included here, since these functions are described extensively in other sections of this book):

- Radio Resource Control (RRC): This protocol is in control of the radio resource usage. It manages UE's signalling and data connections, and includes functions for handover.
- Packet Data Convergence Protocol (PDCP): The main functions of PDCP are IP header compression (UP), encryption and integrity protection (CP only).
- Radio Link Control (RLC): The RLC protocol is responsible for segmenting and concatenation of the PDCP-PDUs for radio interface transmission. It also performs error correction with the Automatic Repeat Request (ARQ) method.
- Medium Access Control (MAC): The MAC layer is responsible for scheduling the data according to priorities, and multiplexing data to Layer 1 transport blocks. The MAC layer also provides error correction with Hybrid ARQ.
- Physical Layer (PHY): This is the Layer 1 of LTE-Uu radio interface that takes care of OFDMA and SC-FDMA Layer functions.

The S1 interface connects the E-UTRAN to the EPC, and involves the following protocols:

- S1 Application Protocol (S1AP): S1AP handles the UE's CP and UP connections between the E-UTRAN and EPC, including participating in the handover when EPC is involved.
- SCTP/IP signalling transport: The Stream Control Transmission Protocol (SCTP) and Internet Protocol (IP) represent standard IP transport suitable for signalling messages. SCTP provides the reliable transport and sequenced delivery functions. IP itself can be run on a variety of data link and physical layer technologies (L2 and L1), which may be selected based on availability.

In the EPC, there are two alternative protocols for the S5/S8 interface. The following protocols are involved, when GTP is used in S5/S8:

- GPRS Tunnelling Protocol, Control Plane (GTP-C): It manages the UP connections in the EPC. This includes signalling the QoS and other parameters. If GTP is used in the

S5/S8 interface it also manages the GTP-U tunnels. GTP-C also performs the mobility management functions within the EPC, e.g. when the GTP-U tunnels of a UE need to be switched from one node to the other.
- UDP/IP transport. The Unit Data Protocol (UDP) and IP are used as the standard and basic IP transport. UDP is used instead of Transmission Control Protocol (TCP) because the higher layers already provide reliable transport with error recovery and re-transmission. IP packets in EPC may be transported on top of a variety of L2 and L1 technologies. Ethernet and ATM are some examples.

The following protocols are used, when S5/S8 is based on PMIP:

- Proxy Mobile IP (PMIP): PMIP is the alternative protocol for the S5/S8 interface. It takes care of mobility management, but does not include bearer management functions as such. All traffic belonging to a UE's connection to a particular PDN is handled together.
- IP: PMIP runs directly on top of IP, and it is used as the standard IP transport.

Figure 3.10 illustrates the UP protocol structure for UE connecting to P-GW.

The UP shown in Figure 3.10 includes the layers below the end user IP, i.e. these protocols form the Layer 2 used for carrying the end user IP packets. The protocol structure is very similar to the CP. This highlights the fact that the whole system is designed for generic packet data transport, and both CP signalling and UP data are ultimately packet data. Only the volumes are different. Most of the protocols have been introduced already above, with the exception of the following two that follow the selection of protocol suite in S5/S8 interface:

- GPRS Tunnelling Protocol, User Plane (GTP-U): GTP-U is used when S5/S8 is GTP based. GTP-U forms the GTP-U tunnel that is used to send End user IP packets belonging to one EPS bearer. It is used in S1-U interface, and is used in S5/S8 if the CP uses GTP-C.
- Generic Routing Encapsulation (GRE): GRE is used in the S5/S8 interface in conjunction with PMIP. GRE forms an IP in IP tunnel for transporting all data belonging to one UE's connection to a particular PDN. GRE is directly on top of IP, and UDP is not used.

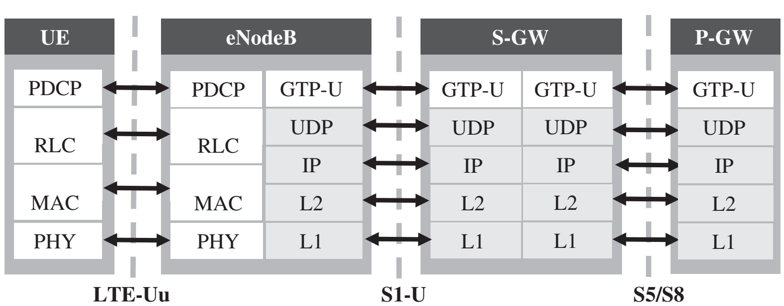

Figure 3.10 User plane protocol stack in EPS

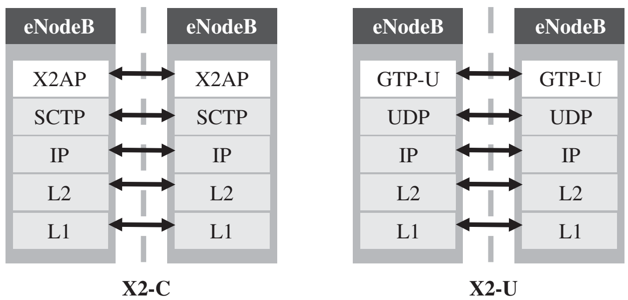

Figure 3.11 Control and user plane protocol stacks for X2 interface

Figure 3.11 illustrates the X2 interface protocol structure, which resembles that of the S1 interface. Only the CP Application Protocol is different. X2 interface is used in mobility between the eNodeBs, and the X2AP includes functions for handover preparation, and overall maintenance of the relation between neighbouring eNodeBs. The UP in the X2 interface is used for forwarding data in a transient state during handover, when the radio interface is already disconnected on the source side, and has not yet resumed on the target side. Data forwarding is done for the DL data, since the UL data can be throttled effectively by the UE.

Table 3.1 summarizes the protocols and interfaces in Basic System Architecture configuration.

Table 3.1 Summary of interfaces and protocols in Basic System Architecture configuration

Interface	Protocols	Specification
LTE-Uu	CP: RRC/PDCP/RLC/MAC/PHY	36.300 [6]
	UP: PDCP/RLC/MAC/PHY	(stage 2)
X2	CP: X2AP/SCTP/IP	36.423 [7]
	UP: GTP-U/UDP/IP	29.274 [8]
S1-MME	S1AP/SCTP/UDP/IP	36.413 [9]
S1-U	GTP-U/UDP/IP	29.274 [8]
S10	GTP-C/UDP/IP	29.274 [8]
S11	GTP-C/UDP/IP	29.274 [8]
S5/S8 (GTP)	GTP/UDP/IP	29.274 [8]
S5/S8 (PMIP)	CP: PMIP/IP	29.275 [10]
	UP: GRE/IP	
SGi	IP (also Diameter & Radius)	29.061 [11]
S6a	Diameter/SCTP/IP	29.272 [12]
Gx	Diameter/SCTP/IP	29.212 [13]
Gxc	Diameter/SCTP/IP	29.212 [13]
Rx	Diameter/SCTP/IP	29.214 [14]
UE – MME	EMM, ESM	24.301 [15]

3.2.5 Roaming in Basic System Architecture Configuration

Roaming is an important functionality, where operators share their networks with each other's subscribers. Typically roaming happens between operators serving different areas, such as different countries, since this does not cause conflicts in the competition between the operators, and the combined larger service area benefits them as well as the subscribers. The words *home* and *visited* are used as prefixes to many other architectural terms to describe where the subscriber originates from and where it roams to respectively.

3GPP SAE specifications define which interfaces can be used between operators, and what additional considerations are needed if an operator boundary is crossed. In addition to the connectivity between the networks, roaming requires that the operators agree on many things at the service level, e.g. what services are available, how they are realized, and how accounting and charging is handled. This agreement is called the *Roaming Agreement*, and it can be made directly between the operators, or through a broker. The 3GPP specifications do not cover these items, and operators using 3GPP technologies discuss roaming related general questions in a private forum called the GSM Association, which has published recommendations to cover these additional requirements.

Roaming defined for SAE follows quite similar principles to the earlier 3GPP architectures. The E-UTRAN is always locally in the visited network, but the data may be routed either to the home network, or can break out to external networks directly from the visited network. This aspect differentiates the two roaming models supported for SAE, which are defined as follows:

- Home Routed model: The P-GW, HSS and PCRF reside in the home operator network, and the S-GW, MME and the radio networks reside in the visited operator network. In this roaming configuration the interface between P-GW and S-GW is called S8, whereas the same interface is called S5 in the non-roaming case when S-GW and P-GW are in the same operator's network. S5 and S8 are technically equivalent. When the S8 interface is based in GTP, the roaming architecture is as shown in Figure 3.2

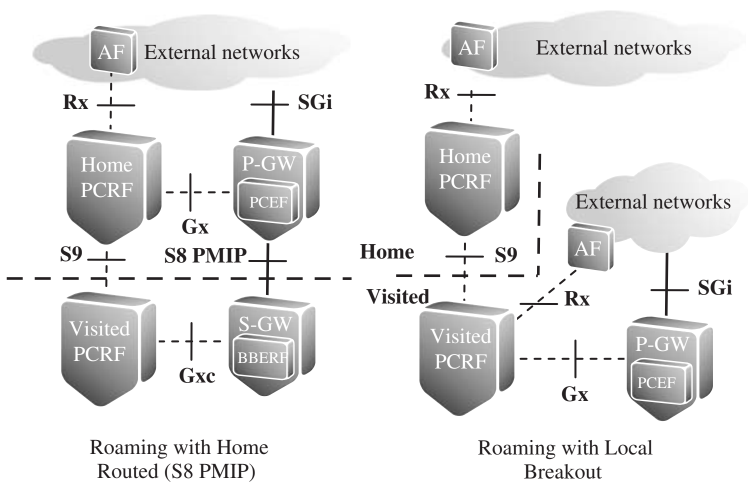

Figure 3.12 Home routed and local breakout roaming

(Gxa does not apply with GTP). When the S8 interface uses PMIP, the PCRF will also be divided into home and visited nodes with the S9 interface between them. This is the scenario shown in Figure 3.12 on the left, and is explained with more detail in section 3.7.1. The Home Routed roaming model applies to legacy 3GPP ANs in the same way, the additional detail being that the SGSN introduced in the next chapter and shown in Figure 3.12 resides in the visited network.

- Local Breakout model: In this model, shown in the right side of Figure 3.12, the P-GW will be located in the visited network, and the HSS is in the home network. If dynamic policy control is used, there will again be two PCRFs involved, one in the home network, and the other in the visited network. Depending on which operator's services are used, the PCRF in that operator's network is also connected to the AF. Also this scenario is explained with more detail in section 3.7.1. With these constraints the Local Breakout model also works with the legacy 3GPP ANs.

3.3 System Architecture with E-UTRAN and Legacy 3GPP Access Networks

3.3.1 Overview of 3GPP Inter-working System Architecture Configuration

Figure 3.13 describes the architecture and network elements in the architecture configuration where all 3GPP defined ANs, E-UTRAN, UTRAN and GERAN, are connected to

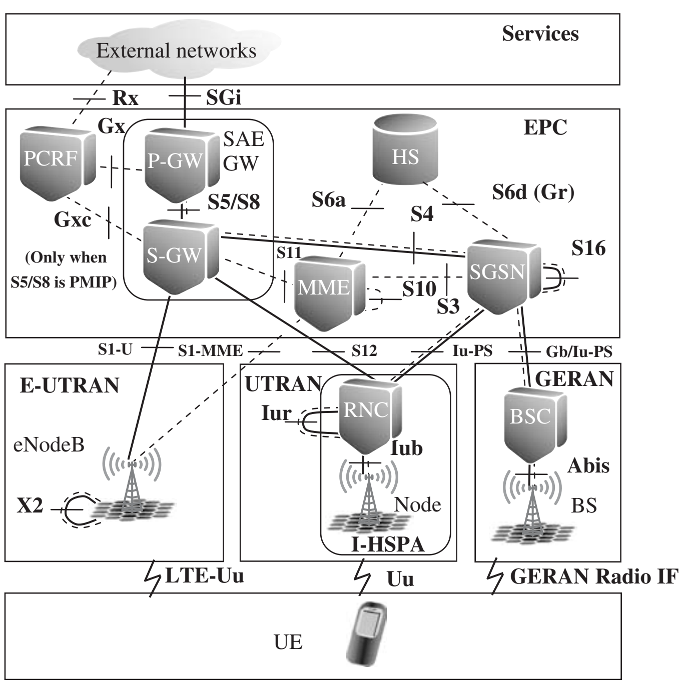

Figure 3.13 System architecture for 3GPP access networks

the EPC. This is called here the 3GPP Inter-working System Architecture Configuration, and it allows optimized inter-working between the mentioned accesses.

Functionally the E-UTRAN, UTRAN and GERAN all provide very similar connectivity services, especially when looking at the situation from the end user point of view, where the only difference may be the different data rates and improved performance, but architecturally these ANs are quite different, and many things are carried out differently. There are, for example, big differences in how the bearers are managed in the EPS compared to the existing networks with UTRAN or GERAN access. However, when UTRAN or GERAN is connected to EPC, they may still operate as before from this perspective, and for this purpose the S-GW simply assumes the role of the Gateway GPRS Support Node (GGSN). Also in optimized inter-working with the E-UTRAN, the GERAN and UTRAN ANs behave almost the same way as they behave when inter-working between themselves. The differences become more visible in the EPC, because what used to be the fixed GGSN is now the S-GW that may be changed along with the SGSN change during UE mobility.

All nodes and functions described in the previous section for the Basic System Architecture Configuration are needed here also. The EPC needs the addition of a few new interfaces and functions to connect and inter-work with UTRAN and GERAN. The corresponding functions will also be required from GERAN and UTRAN. The new interfaces are S3, S4 and S12 as shown in Figure 3.12. The interface from SGSN to HSS can also be updated to Diameter based S6d, but the use of the legacy MAP based Gr is also possible.

Keeping E-UTRAN, i.e. the eNodeB design as focused to, and as optimized for the requirements of the new OFDMA radio interface, and as clean of inter-working functionality as possible, was an important guideline for the inter-working design. Consequently, the eNodeB does not interface directly with the other 3GPP ANs, and the interaction towards the EPC is the same as in other mobility cases that involve EPC. However, optimized inter-working means that the network is in control of mobility events, such as handovers, and provides functionality to hand the communication over with minimum interruption to services. This means that an eNodeB must be able to coordinate UE measuring UTRAN and GERAN cells, and perform handover decisions based on measurement results, and thus E-UTRAN radio interface protocols have been appended to support the corresponding new functions. Similar additions will be required from UTRAN and GERAN to support handover to E-UTRAN.

3.3.2 Additional and Updated Logical Elements in 3GPP Inter-working System Architecture Configuration

3.3.2.1 User Equipment

From the UE point of view, inter-working means that it needs to support the radio technologies in question, and the mobility operations defined for moving between them. The optimized inter-working means that the network controls the usage of radio transmitter and receiver in the UE in a way that only one set of them needs to be operating at the same time. This is called single radio operation, and allows UE implementations where only one pair of physical radio transmitter and receiver is implemented.

The standard does not preclude implementing multiple radio transmitters and receivers, and operating them simultaneously in dual radio operation. However, single radio operation is an important mode, because the different ANs often operate in frequencies that are so close to each other that dual radio operation would cause too much interference within the terminal. That, together with the additional power consumption, will decrease the overall performance.

3.3.2.2 E-UTRAN

The only addition to E-UTRAN eNodeB compared to the Basic System Architecture Configuration is the mobility to and from other 3GPP ANs. From the eNodeB perspective the functions are very similar irrespective of whether the other 3GPP AN is UTRAN or GERAN.

For the purpose of handover from E-UTRAN to UTRAN or GERAN, the neighbouring cells from those networks need to be configured into the eNodeB. The eNodeB may then consider handover for those UEs that indicate corresponding radio capability. The eNodeB requests the UE to measure the signal level of the UTRAN or GERAN cells, and analyses the measurement reports. If the eNodeB decides to start the handover, it signals the need to the MME in the same way that it would signal inter-eNodeB handover when the X2 interface is not available. Subsequently, the eNodeB will receive the information needed for the Handover Command from the target Access System via the MME. The eNodeB will send the Handover Command to the UE without the need for interpreting the content of this information.

In the case of handover from UTRAN or GERAN to E-UTRAN, the eNodeB does not need to make any specific preparations compared to other handovers where the handover preparation request comes through the MME. The eNodeB will allocate the requested resources, and prepare the information for handover command, which it sends to the MME, from where it is delivered to the UE through the other 3GPP Access System that originated the handover.

3.3.2.3 UTRAN

In UTRAN, the radio control functionality is handled by the Radio Network Controller (RNC), and under its control the Node B performs Layer 2 bridging between the Uu and Iub interfaces. UTRAN functionality is described extensively in [16].

UTRAN has evolved from its initial introduction in Release 99 in many ways, including the evolution of architectural aspects. The first such item is Iu flex, where the RNC may be connected to many Serving GPRS Support Nodes (SGSNs) instead of just one. Another such concept is I-HSPA, where the essential set of packet data related RNC functions is included with the Node B, and that connects to Iu-PS as a single node. Figure 3.13 also shows the direct UP connection from RNC to S-GW, which is introduced to 3G CN by the Direct Tunnel concept, where the SGSN is bypassed in UP.

Inter-working with E-UTRAN requires that UTRAN performs the same measurement control and analysis functions as well as the transparent handover information delivery in Handover Command that were described for eNodeB in the earlier section. Also the UTRAN performs similar logic that it already uses with Relocation between RNCs, when the Iur interface is not used.

3.3.2.4 GERAN

GSM EDGE Radio AN (GERAN) is the evolved version of GSM AN, which can also be connected to 3G Core Network. It consists of the Base Station Controller (BSC) and the Base Station (BS), and the radio interface functionalities are divided between them. An overview of GERAN functionality and the whole GSM system can be found in [17].

The GERAN is always connected to the SGSN in both Control and UPs, and this connection is used for all the inter-working functionality. Also the GERAN uses logic similar to that described above for E-UTRAN and UTRAN for inter-working handover.

3.3.2.5 EPC

The EPC has a central role for the inter-working system architecture by anchoring the ANs together. In addition to what has been described earlier, the MME and S-GW will support connectivity and functions for inter-working. Also the SGSN, which supports the UTRAN and GERAN access networks, will need to support these functions, and when these additions are supported, it can be considered to belong to the EPC.

The S-GW is the mobility anchor for all 3GPP access systems. In the basic bearer operations and mobility between SGSNs, it behaves like a GGSN towards the SGSN, and also towards the RNC if UP tunnels are set up in Direct Tunnel fashion bypassing the SGSN. Many of the GGSN functions are actually performed in the P-GW, but this is not visible to the SGSN. The S-GW retains its role as a UP Gateway, which is controlled by either the MME or the SGSN depending on which AN the UE is being served by.

To support the inter-working mobility, the MME will need to signal with the SGSN. These operations are essentially the same as between those two MMEs, and have been described earlier in section 3.2. An additional aspect of the MME is that it may need to combine the change of S-GW and the inter-working mobility with SGSN.

The SGSN maintains its role as the controlling node in core network for both UTRAN and GERAN. These functions are defined in [18]. The SGSN has a role very similar to that of the MME. The SGSN needs to be updated to support for S-GW change during mobility between SGSNs or RNCs, because from the legacy SGSN point of view this case looks like GGSN changing, which is not supported. As discussed earlier, the SGSN may direct the UP to be routed directly between the S-GW and UTRAN RNC, or it may remain involved in the UP handling. From the S-GW point of view this does not really make a difference, since it does not need to know which type of node terminates the far end of the UP tunnel.

3.3.3 Interfaces and Protocols in 3GPP Inter-working System Architecture Configuration

Table 3.2 summarizes the interfaces in the 3GPP Inter-working System Architecture Configuration and the protocols used in them. Interfaces and protocols in legacy 3GPP networks are not listed. Interfaces and protocols listed for Basic System Architecture Configuration are needed in addition to these.

Table 3.2 Summary of additional interfaces and protocols in 3GPP Inter-working System Architecture configuration

Interface	Protocols	Specification
S3	GTP-C/UDP/IP	29.274 [8]
S4	GTP/UDP/IP	29.274 [8]
S12	GTP-U/UDP/IP	29.274 [8]
S16	GTP/UDP/IP	29.274 [8]
S6d	Diameter/SCTP/IP	29.272 [12]

3.3.4 Inter-working with Legacy 3GPP CS Infrastructure

While the EPS is purely a Packet Switched (PS) only system without a specific Circuit Switched (CS) domain with support for VoIP, the legacy 3GPP systems treat CS services such as voice calls with a specific CS infrastructure. IMS VoIP may not be ubiquitously available, and therefore the SAE design includes two special solutions that address inter-working with circuit switched voice. A description of how inter-working between E-UTRAN and the legacy 3GPP CS domain can be arranged is given in Chapter 10 on VoIP. Two specific functions have been defined for that purpose, Circuit Switched Fall Back (CSFB) and Single Radio Voice Call Continuity (SR-VCC).

CSFB [19] is a solution for networks that do not have support for IMS VoIP. Instead, the voice calls are handled by the CS domain, and the UE is handed over there at the time of a voice call. The SGs interface between the MME and MSC Server is used for related control signalling, as shown with more detail in Chapter 10.

SR-VCC [20] is a solution for converting and handing over an IMS VoIP call to a CS voice call in the legacy CS domain. This functionality would be needed when the coverage of an IMS VoIP capable network is smaller than that of the legacy CS networks. SR-VCC allows a UE entering the edge of the VoIP coverage area with an ongoing VoIP call to be handed over to the CS network without interrupting the call. SR-VCC is a one way handover from the PS network with VoIP to the CS network. If E-UTRAN coverage becomes available again, the UE may return there when the call ends and the UE becomes idle. The solution relies on running only one radio at a time, i.e. the UE does not need to communicate simultaneously with both systems. In this solution the MME is connected to the MSC Server in the CS domain via a Sv interface, which is used for control signalling in the SR-VCC handover. The details of the solution are presented in Chapter 10. A summary of additional interfaces and protocols for inter-working with legacy 3GPP CS infrastructure is given in Table 3.3.

Table 3.3 Summary of additional interfaces and protocols for inter-working with legacy 3GPP CS infrastructure

Interface	Protocols	Specification
SGs	SGsAP/SCTP/IP	29.118 [21]
Sv	GTP-C(subset)/UDP/IP	29.280 [22]

3.4 System Architecture with E-UTRAN and Non-3GPP Access Networks

3.4.1 Overview of 3GPP and Non-3GPP Inter-working System Architecture Configuration

Inter-working with non-3GPP ANs was one of the key design goals for SAE, and to support it, a completely separate architecture specification [2] was developed in 3GPP. The non-3GPP Inter-working System Architecture includes a set of solutions in two categories. The first category contains a set of generic and loose inter-working solutions that can be used with any other non-3GPP AN. Mobility solutions defined in this category are also called Handovers without Optimizations, and the same procedures are applicable in both connected and idle mode. The second category includes a specific and tighter inter-working solution with one selected AN, the cdma2000® HRPD. This solution category is also called Handovers with Optimizations, and it specifies separate procedures for connected and idle mode.

The generic non-3GPP Inter-working System Architecture is shown in Figure 3.14. The specific application of the architecture for cdma2000® HRPD inter-working and the required additional interfaces are described with more detail in section 3.5.

Figure 3.14 describes the generic inter-working solution that relies only on loose coupling with generic interfacing means, and without AN level interfaces. Since there are so many different kinds of ANs, they have been categorized to two groups, the trusted and un-trusted non-3GPP ANs, depending on whether it can be safely assumed that 3GPP defined

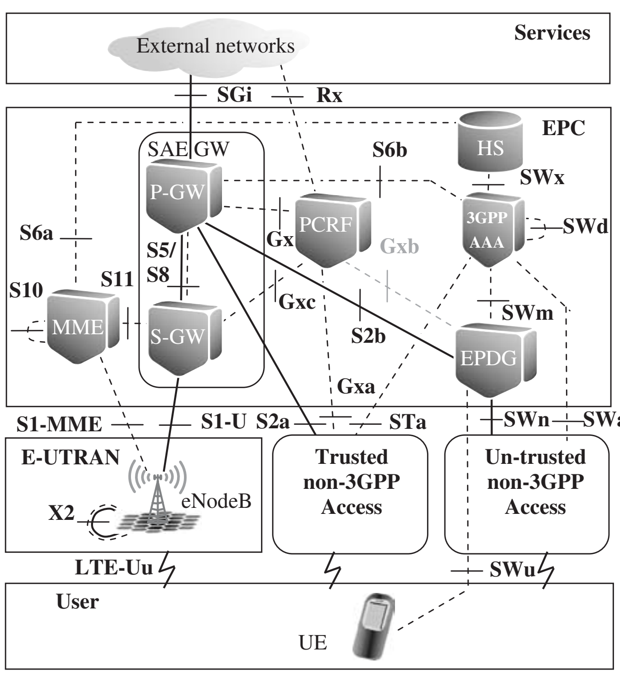

Figure 3.14 System architecture for 3GPP and non-3GPP access networks

authentication can be run by the network, which makes it trusted, or if authentication has to be done in overlay fashion and the AN is un-trusted. The P-GW will maintain the role of mobility anchor, and the non-3GPP ANs are connected to it either via the S2a or the S2b interface, depending on whether the non-3GPP AN functions as a Trusted or Un-trusted non-3GPP AN. Both use network controlled IP layer mobility with the PMIP protocol. For networks that do not support PMIP, Client MIPv4 Foreign Agent mode is available as an option in S2a. In addition to mobility functions, the architecture includes interfaces for authenticating the UE within and through the non-3GPP ANs, and also allows PCC functionality in them via the Gxa and Gxb interfaces. Note that the detailed functions and protocols for Gxb are not specified in Release 8.

In addition to the network controlled mobility solutions, a completely UE centric solution with DSMIPv6 is also included in the inter-working solutions. This scenario is depicted in Figure 3.15.

In this configuration the UE may register in any non-3GPP AN, receive an IP address from there, and register that to the Home Agent in P-GW. This solution addresses the mobility as an overlay function. While the UE is served by one of the 3GPP ANs, the UE is considered to be in home link, and thus the overhead caused by additional MIP headers is avoided.

Another inter-working scenario that brings additional flexibility is called the chained S8 and S2a/S2b scenario. In that scenario the non-3GPP AN is connected to S-GW in the visited Public Land Mobile Network (PLMN) through the S2a or S2b interface, while the P-GW is in the home PLMN. This enables the visited network to offer a roaming subscriber the use of non-3GPP ANs that might not be associated with the home operator at all, even in the case where P-GW is in the home PLMN. This scenario requires that S-GW performs functions that normally belong to P-GW in order to behave as the termination point for the S2a or S2b interfaces. In Release 8, this scenario does not support dynamic policies through the PCC infrastructure, i.e. the Gxc interface will not be used. Also, chaining with GTP based S5/S8 is not supported. All other interfaces related to non-3GPP ANs are used normally as shown in Figure 3.14.

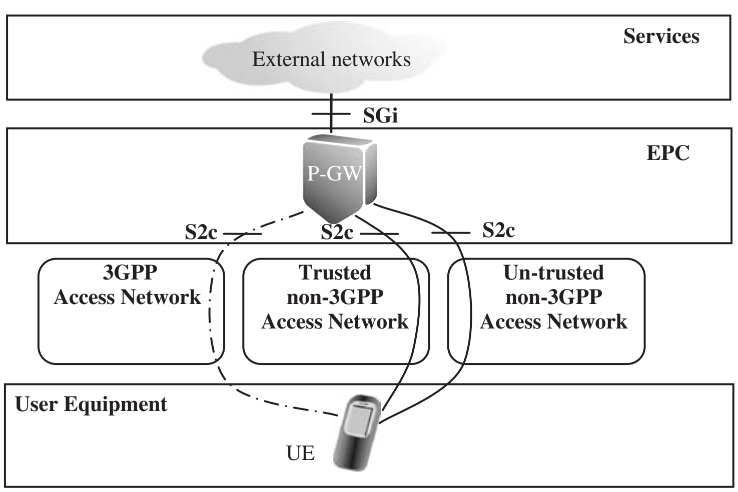

Figure 3.15 Simplified system architecture showing only S2c

3.4.2 Additional and Updated Logical Elements in 3GPP Inter-working System Architecture Configuration

3.4.2.1 User Equipment

Inter-working between the non-3GPP ANs requires that the UE supports the corresponding radio technologies, and the specified mobility procedures. The mobility procedures and required radio capabilities vary depending on whether optimizations are in place or not. The procedures defined for Handovers without Optimizations do not make any assumption about the UE's capability to use the radio transmitters and receivers simultaneously, and both single radio and dual radio configurations can use the procedures. However, the handover gap time is expected to be shorter, if preparing the connections towards the target side can start already while data are still flowing through the source side. This is caused by the fact that Handovers without Optimizations do not have procedures in the network side to assist in handover preparations, and the procedures follow the principle where UE registers to the target network according to the method defined for that network, and then the network switches the flow to the target network. This may be time consuming, since it normally includes procedures such as authentication. Also, the decision to make these handovers is the responsibility of the UE.

The Handovers with Optimizations, i.e. inter-working with cdma2000® HRPD, assume that they do include network control for connected mode, so the handovers are decided by the network, while the idle mode mobility relies on UE decision making, which may use cdma2000® HRPD related information in the LTE-Uu broadcast. Furthermore, the procedures are designed with the assumption that single radio configuration is enough for the UE.

3.4.2.2 Trusted Non-3GPP Access Networks

The term trusted non-3GPP AN refers to networks that can be trusted to run 3GPP defined authentication. 3GPP Release 8 security architecture specification for non-3GPP ANs [23] mandates that the Improved Extensible Authentication Protocol Method for 3rd Generation Authentication and Key Agreement (EAP-AKA') [24] is performed. The related procedures are performed over the STa interface.

The trusted non-3GPP ANs are typically other mobile networks, such as the cdma2000® HRPD. The STa interface supports also delivery of subscription profile information from Authentication, Authorization and Accounting (AAA)/HSS to the AN, and charging information from the AN to AAA Server, which are typical functions needed in mobile networks. It can also be assumed that such ANs may benefit from connecting to the PCC infrastructure, and therefore the Gxc interface may be used to exchange related information with the PCRF.

The trusted non-3GPP AN connects to the P-GW with the S2a interface, with either PMIP or MIPv4 Foreign Agent mode. The switching of UP flows in P-GW is therefore the responsibility of the trusted non-3GPP AN when UE moves into the AN's service area.

3.4.2.3 Un-trusted Non-3GPP Access Networks

To a large extent, the architectural concepts that apply for un-trusted non-3GPP ANs are inherited from the Wireless Local Area Network Inter-Working (WLAN IW) defined originally in Release 6 [25]. The Release 8 functionality for connecting un-trusted non-3GPP ANs to EPC is specified fully in [2] with references to the earlier WLAN IW specifications when applicable.

The main principle is that the AN is not assumed to perform any other functions besides delivery of packets. A secure tunnel is established between UE and a special node called the Enhanced Packet Data Gateway (EPDG) via the SWu interface, and the data delivery takes place through that tunnel. Furthermore, the P-GW has a trust relationship with the EPDG connected to it via the S2b interface, and neither node needs to have secure association with the un-trusted non-3GPP AN itself.

As an optional feature, the un-trusted non-3GPP AN may be connected to the AAA Server with the SWa interface, and this interface may be used to authenticate the UE already in the non-3GPP AN level. This can be done only in addition to authentication and authorization with the EPDG.

3.4.2.4 EPC

The EPC includes quite a few additional functions for the support of non-3GPP ANs, when compared to the previously introduced architecture configurations. The main changes are in the P-GW, PCRF and HSS, and also in S-GW for the chained S8 and S2a/S2b scenario. In addition, completely new elements, such as the EPDG (Evolved Packet Data Gateway) and the AAA are introduced. The AAA infrastructure contains the AAA Server, and it may also contain separate AAA proxies in roaming situations. Figure 3.16 highlights the AAA connections and functions for non-3GPP ANs.

The P-GW is the mobility anchor for the non-3GPP ANs. For PMIP based S2a and S2b interfaces, the P-GW hosts the Local Mobility Anchor (LMA) function in a manner similar to that for the S5/S8 PMIP interfaces. Also the Home Agent (HA) function for the Client MIPv4 Foreign Agent mode in S2a is located in P-GW. The relation between

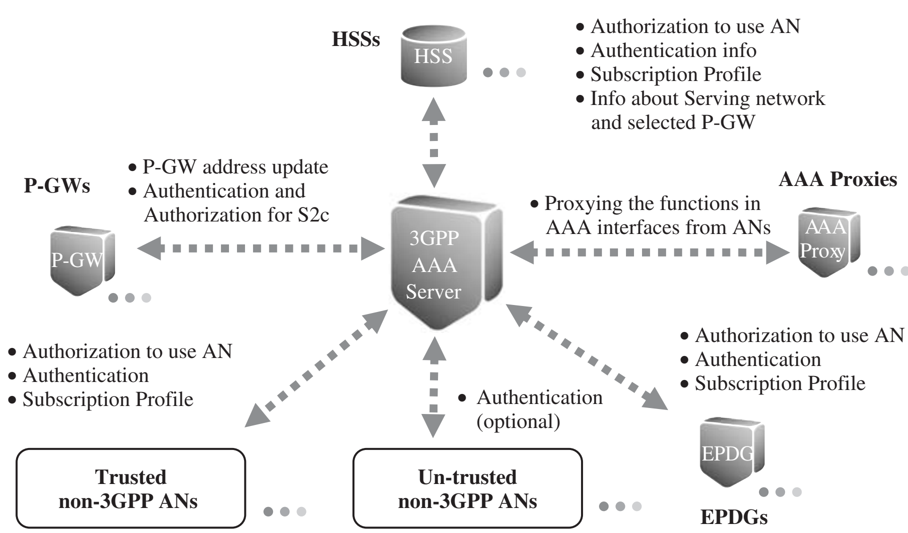

Figure 3.16 3GPP AAA server interfaces and main functions

P-GWs and non-3GPP ANs is many to many. The P-GW will also interface with the AAA Server, which subsequently connects to HSS. This interface is used for reporting the selected P-GW to the HSS so that it is available in mobility between non-3GPP ANs, and to authenticate and authorize users connecting with S2c mode. Each P-GW may connect to more than one AAA server.

The PCRF supports PCC interfaces for non-3GPP ANs. The Gxa is used towards trusted non-3GPP ANs, and Gxb towards un-trusted non-3GPP ANs. Only Gxa is specified in detail level in Release 8. The Gxa interface functions in a fashion similar to that of the Gxc interface. In this case the BBERF function will be in the non-3GPP AN, and it will receive instructions from the PCRF on how to handle the bearer level functions for the IP flows in the S2a interface. The further bearer functions internal to non-3GPP ANs are not addressed in 3GPP specifications.

The EPDG is a dedicated node for controlling the UE and inter-network connection, when an un-trusted non-3GPP AN is connected to EPC. Since the AN is not trusted, the main function is to secure the connection, as defined in [23]. The EPDG establishes an IPsec tunnel to the UE through the un-trusted non-3GPP AN with IKEv2 signalling [26] over the SWu interface. During the same signalling transaction the EAP-AKA authentication is run, and for that the EPDG signals with the AAA Server through the SWm interface. While the SWm interface is logically between UE and the EPDG, the SWn interface represents the interface on a lower layer between the EPDG and the un-trusted non-3GPP AN. The Release 8 specifications do not assume that EPDG would signal with PCRF for any PCC functions, but the architecture already contains the Gxb interface for that purpose.

The 3GPP AAA Server, and possibly a AAA Proxy in the visited network, performs a 3GPP defined set of AAA functions. These functions are a subset of what the standard IETF defined AAA infrastructure includes, and do not necessarily map with the way other networks use AAA infrastructure. The AAA Server acts between the ANs and the HSS, and in doing so it creates a context for the UEs it serves, and may store some of their information for further use. Thus, the 3GPP AAA Server consolidates the signalling from different types of ANs into a single SWx interface towards the HSS, and terminates the access specific interfaces S6b, STa, SWm and SWa. Most importantly the AAA Server performs as the authenticator for the EAP-AKA authentication through the non-3GPP ANs. It checks the authenticity of the user, and informs the AN about the outcome. The authorization to use the AN in question will also be performed during this step. Depending on the AN type in question, the AAA Server may also relay subscription profile information to the AN, which the AN may further use to better serve the UE. When the UE is no longer served by a given non-3GPP AN, the AAA Server participates in removing the UE's association from the HSS. Figure 3.16 summarizes the AAA Server main functions in relation to other nodes.

The HSS performs functions similar to those for the 3GPP ANs. It stores the main copy of the subscription profile as well as the secret security key in the AuC portion of it, and when requested, it provides the profile data and authentication vectors to be used in UEs connecting through non-3GPP ANs. One addition compared to 3GPP ANs is that since the non-3GPP ANs do not interface on the AN level, the selected P-GW needs to be stored in the HSS, and retrieved from there when the UE mobility involves a non-3GPP AN. The variety of different AN types are mostly hidden from the HSS, since the AAA

Server terminates the interfaces that are specific to them, and HSS only sees a single SWx interface. On the other hand, the subscription profile stored in the HSS must reflect the needs of all the different types of ANs that are valid for that operator.

3.4.3 Interfaces and Protocols in Non-3GPP Inter-working System Architecture Configuration

Connecting the non-3GPP ANs to EPC and operating them with it requires additional interfaces to those introduced in earlier sections. Table 3.4 lists the new interfaces.

3.4.3.1 Roaming in Non-3GPP Inter-working System Architecture Configuration

The principles for roaming with non-3GPP accesses are equivalent to those described in section 3.2.4 for 3GPP ANs. Both home routed and local breakout scenarios are supported and the main variations in the architecture relate to the PCC arrangement, which depends on where the services are consumed. This aspect is highlighted more in section 3.7.1.

The additional consideration that non-3GPP ANs bring to roaming is related to the case where the user is roaming to a visited 3GPP network in Home Routed model, and it would be beneficial to use a local non-3GPP AN that is affiliated with the visited network, but there is no association between that network and the home operator. For this scenario, the 3GPP Release 8 includes a so-called *chained case*, where the S-GW may behave as the anchor for the non-3GPP ANs also, i.e. it terminates the S2a or S2b interface, and routes the traffic via the S8 interface to the P-GW in the home network.

Table 3.4 Summary of additional interfaces and protocols in non-3GPP Inter-working System Architecture configuration

Interface	Protocols	Specification
S2a	PMIP/IP, or MIPv4/UDP/IP	29.275 [10]
S2b	PMIP/IP	29.275 [10]
S2c	DSMIPv6, IKEv2	24.303 [27]
S6b	Diameter/SCTP/IP	29.273 [28]
Gxa	Diameter/SCTP/IP	29.212 [13]
Gxb	*Not defined in Release 8*	*N.A.*
STa	Diameter/SCTP/IP	29.273 [28]
SWa	Diameter/SCTP/IP	29.273 [28]
SWd	Diameter/SCTP/IP	29.273 [28]
SWm	Diameter/SCTP/IP	29.273 [28]
SWn	PMIP	29.275 [10]
SWu	IKEv2, MOBIKE	24.302 [29]
SWx	Diameter/SCTP/IP	29.273 [28]
UE – foreign agent in trusted non-3GPP Access	MIPv4	24.304 [30]
UE – Trusted or Un-trusted non-3GPP access	EAP-AKA	24.302 [29]

3.5 Inter-working with cdma2000® Access Networks

3.5.1 Architecture for cdma2000® HRPD Inter-working

The best inter-working performance in terms of handover gap time is achieved by specifying the networks to inter-operate very tightly to exchange critical information. This creates a specific solution that is valid for only the ANs in question. With the limited time and resources available for specification work, the number of such solutions in 3GPP Release 8 could only be limited. A tight inter-working solution also requires changes in the other ANs, and by definition the development of non-3GPP ANs is not within the control of 3GPP. Achieving a well designed solution requires special attention to coordination between the developments in different standardization bodies. With these difficulties at hand, 3GPP Release 8 only includes an optimized inter-working solution with cdma2000® HRPD AN.

Figure 3.17 highlights the architecture for cdma2000® HRPD inter-working. It shows the Evolved HRPD (E-HRPD) network, where a number of modifications have been applied to make it suitable for connecting to the EPC. Due to these modifications it will be called E-HRPD in this chapter to distinguish it from legacy HRPD systems that do not support these functions. The radio interface and the Radio Access Network have been kept as similar as possible, but the HRPD Serving Gateway (HSGW) is a completely new node inheriting many of its functions from S-GW.

The E-HRPD is generally treated as a trusted non-3GPP AN, and it is therefore connected to the EPC via S2a, Gxa and STa interfaces. These interfaces operate as described

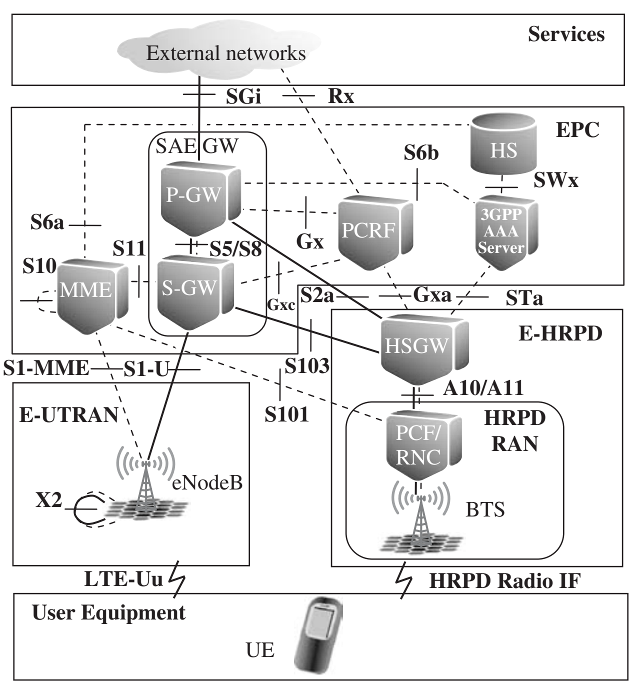

Figure 3.17 System architecture for 3GPP and 2000® HRPD inter-working

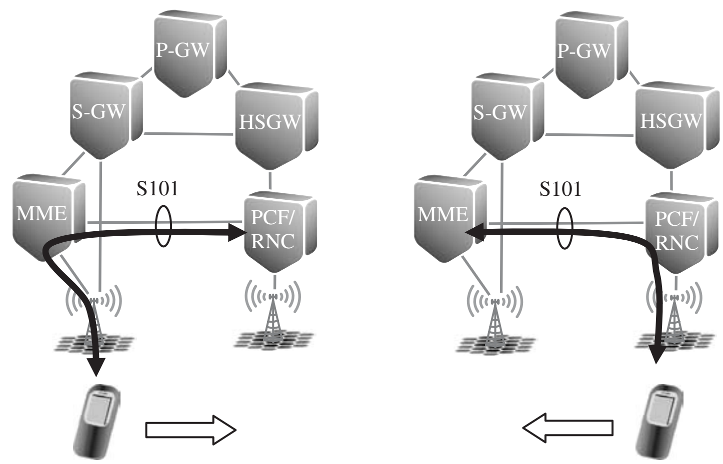

Figure 3.18 Tunnelled pre-registration to eHRPD and to E-UTRAN

earlier. Since the inter-working solution is optimized, and does not rely on UE performing the attach procedure directly to the target network, two new interfaces, S101 and S103, were defined for the CP and UP interactions respectively.

3GPP ANs and the 3GPP2-defined cdma2000® ANs share many things in common, but many things are also different. Both systems use a prepared handover, where the source system signals to the target system to give it essential parameters to be able to serve the terminal there, and the target system gives the source system parameters that can be further given to the terminal to guide it to make the access to the target radio. While there are similarities in these methods, the parameters themselves do not match well at all, and this method could not be used by applying a simple protocol conversion. To ease up on the need to align every information element that would need to be exchanged in handover, it was decided to use a transparent signalling transport method.

Figure 3.18 shows how the transparent tunnel is used in mobility from E-UTRAN to E-HRPD on the left, and the opposite direction is shown on the right. The thick black arrow indicates the signalling which is carried transparently through the source access system and over the S101 interface to the target system. In this method the source access system gives guidance to the UE to register to the target system through the tunnel. This creates the UE context in the target system without the source system having to convert its information to the target system format. This is called pre-registration, and the purpose is to take the time consuming registration/attach function away from the time critical path of handover. The transparent tunnel may also be used to build the bearer context in the target system so that when the time to make the handover is at hand, everything will be ready and waiting at the target side. The actual handover is decided based on radio interface conditions, and this solution requires that both systems are able to handle measurements from the other system. The following inter-working scenarios are supported between E-UTRAN and E-HRPD:

- E-UTRAN → E-HRPD handover: The pre-registration may be performed well before the actual handover takes place, and also all bearers are set up in the E-HRPD side.

The UE remains in a dormant state (equal to idle mode) from the E-HRPD system point of view before handover, and this state may be long lived. When the radio conditions indicate the need for handover, the eNodeB commands the UE to start requesting traffic channel from E-HRPD. This takes place through the transparent tunnel, and once it is completed, the eNodeB commands the UE to make the actual handover. The S103 interface is used only in this handover scenario to forward DL data during the time when the UE is making the switch between the radios.

- E-UTRAN → E-HRPD idle mode mobility: The pre-registration state works as described above for the handover. The UE is in idle mode in the E-UTRAN also, and it moves within the system, selecting the cells on its own, and when it selects an E-HRPD cell, it briefly connects to the E-HRPD network to get the mobility pointers updated to the E-HRPD side.
- E-HRPD → E-UTRAN handover: The E-HRPD AN will request the UE to make tunnelled pre-registration (attach) only at the time the handover is needed, and the UE will immediately proceed to requesting connection directly from the E-UTRAN cell after the registration is complete. The bearers are set up in embedded fashion with the registration and connection request procedures.
- E-HRPD → E-UTRAN idle mode mobility: The idle mode procedure follows the same guidelines as the handover for the tunnelled registration (attach), but the UE accesses the E-UTRAN radio only by reporting its new location (Tracking Area), since there is no need to set up bearers in E-UTRAN for UE in idle mode.

3.5.2 Additional and Updated Logical Elements for cdma2000® HRPD Inter-working

Inter-working with eHRPD in an optimized manner brings a lot of new features in the basic SAE network elements, and introduces few totally new elements in the HRPD side. The UE, eNodeB, MME and S-GW will all be modified to support new functions, and MME and S-GW will also deal with new interfaces. The eHRPD is a new network of its own, and it consists of elements such as Base Station, Radio Network Controller (RNC), Packet Control Function (PCF) and HRPD Serving Gateway (HSGW).

The UE will need to support both radio interfaces. The design of the procedure assumes that UE is capable of single mode operation only. On the other hand, the integration is kept loose enough so that it would be possible to implement terminal with separate chip sets for E-UTRAN and E-HRPD. This means that the UE is not required to make measurements of cells in the other technology in as tightly a timewise controlled manner as is normally seen within a single radio technology. The UE will also need to support the tunnelled signalling operation. The tunnelled signalling itself is the same signalling as the UE would use directly with the other system.

The main new requirement for the eNodeB is that it also needs to be able to control mobility towards the eHRPD access. From the radio interface perspective it does this much in the same manner as with the other 3GPP accesses, by instructing the UE to make measurements of the neighbouring eHRPD cells, and making the handover decision based on this information. On the other hand, the eNodeB does not signal the handover preparation towards the eHRPD, like it would for other handovers in S1 interface. Instead the handover preparation works so that the UE sends traffic channel requests to the eHRPD AN through the transparent tunnel, and the eNodeB is only responsible for marking the uplink messages with appropriate routing information, so that the MME can select the right node in the eHRPD AN, and noting the progress of the handover from the headers of the S1 messages carrying the eHRPD signalling.

The MME implements the new S101 interface towards the eHRPD RAN. For UE originated messages, it needs to be able to route them to the right eHRPD RAN node based on a reference given by the eNodeB. In the reverse direction the messages are identified by the IMSI of the UE, and the basis that the MME can route them to the right S1 signalling connection. The MME does not need to interpret the eHRPD signalling message contents, but the status of the HO progress is indicated along with those messages that require special action from the MME. For example, at a given point during E-UTRAN → E-HRPD handover, the MME will set up the data forwarding tunnels in the S-GW. The MME also needs to memorize the identity of the E-HRPD AN node that a UE has been signalling with, so that if MME change takes place, the MME can update the S101 context in the HRPD AN node.

The S-GW supports the new S103 interface, which is used for forwarding DL data during the time in handover, when the radio link cannot be used. The forwarding function is similar to the function S-GW has for the E-UTRAN handovers. The difference is that S103 is based on a GRE tunnel, and there will be only one tunnel for each UE in handover, so the S-GW needs to map all GTP tunnels from the S1-U interface to a single GRE tunnel in the S103 interface.

The E-HRPD network is a completely new way to use the existing HRPD radio technology with the SAE, by connecting it to the EPC. Compared to the original HRPD, many changes are caused by the inter-working, and connecting to the EPD requires some new functions, e.g. the support of EAP-AKA authentication. The HSGW is taking the S-GW role for E-HRPD access, and performs much like a S-GW towards the P-GW. The HSGW also includes many CP functions. Towards the eHRPD AN, it behaves like the Packet Data Serving Node (PDSN) in a legacy HRPD network. It also signals with the 3GPP AAA Server to authenticate the UE, and to receive its service profile. The CN aspects of the E-HRPD are specified in [31] and the evolved RAN is documented in [32].

3.5.3 Protocols and Interfaces in cdma2000® HRPD Inter-working

The optimized inter-working introduces two new interfaces – S101 and S103 – to the architecture (see Table 3.5). The S2a, Gxc and STa are as described earlier. The following summarizes the new interfaces:

- S101 is a CP interface that in principle forms a signalling tunnel for the eHRPD messages. The CP protocol is S101AP, which is specified in [33]. The S101AP uses the same message structure and coding as the newest version of GTP. The main function is to carry the signalling messages, with the IMSI as a reference and with an additional handover status parameter that is set by either the UE or either one of the networks it signals with. In addition, when the data forwarding tunnel needs to be set up, the address information is also included in S101AP. S101AP also includes a procedure to switch the interface from one MME to another if handover in E-UTRAN causes MME change.
- S103 is a simple GRE tunnel for UP data forwarding in handover. It is only used for DL data in handover from E-UTRAN to E-HRPD. S103 is a UP interface only, and

Table 3.5 Additional interfaces and protocols for inter-working with cdma2000® eHRPD

Interface	Protocols	Specification
S101	S101AP/UDP/IP	29.276 [33]
S103	GRE/IP	29.276 [33]

all control information to set up the GRE tunnel is carried in other interfaces. It is specified with S101AP in [33].

3.5.4 Inter-working with cdma2000® 1xRTT

The cdma2000® 1xRTT is a system supporting CS bearers, and is primarily used for voice calls. In this respect it is functionally equivalent to the legacy 3GPP CS infrastructure such as the MSC and the CS bearer capabilities of GERAN and UTRAN. As described in Chapter 10, the 3GPP standard includes two functions to support inter-working between the E-UTRAN and the legacy CS infrastructure. These are the CSFB [19] and SR-VCC [20]. These functions have been extended to cover inter-working with cdma2000® 1xRTT also, and at a high level they work in the same way as described in Chapter 10. In the 1xRTT case, the interface between MME and the cdma2000® 1xRTT infrastructure is called S102. S102 carries a protocol specified in 3GPP2 for the A21 interface, which is used in cdma2000® systems for voice call continuity (see Table 3.6).

3.6 IMS Architecture

3.6.1 Overview

The IP Multimedia Services Sub-System (IMS) is the preferred service machinery for LTE/SAE. IMS was first introduced in Release 5, and with the well defined inter-working with existing networks and services that have been introduced since, the Rel-8 IMS can now be used to provide services over fixed and wireless accesses alike. The IMS architecture is defined in [36], and the functionality is defined in [37]. A comprehensive description of IMS can also be found in [3]. For the purpose of this book, the functional architecture of IMS is presented in Figure 3.19, and a short description of the main functions follows below.

IMS is an overlay service layer on top of the IP connectivity layer that the EPS provides. Figure 3.19 shows a thick grey line from UE to P-GW that represents the UE's

Table 3.6 Additional interfaces and protocols for inter-working with cdma2000® 1xRTT

Interface	Protocols	Specification
S102	S102 protocol	29.277 [34]
A21	A21 protocol	A.S0008-C [35]

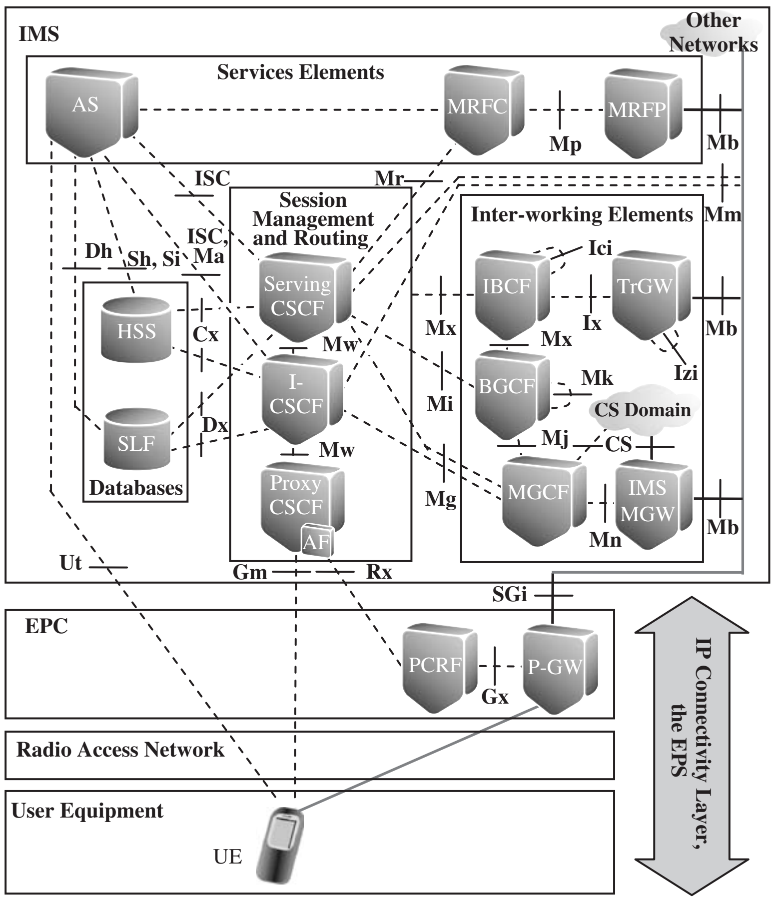

Figure 3.19 IMS architecture

IP connectivity to IMS and other external networks through the RAN and EPC. The signalling interfaces Gm and Ut run on top of this connection, which typically use the default bearer a UE will always have in LTE/SAE. The services may further require that dedicated bearers are set up through EPC, and the service data flows may need to be handled by one of the Inter-working or Services Elements.

In principle the IMS is independent of the connectivity layer, which requires its own registration and session management procedures, but it has also been specifically designed to operate over the 3GPP-defined ANs, and it works seamlessly with the PCC described in section 3.7. IMS uses SIP protocol for registration and for controlling the service sessions. SIP is used both between the terminal and the IMS (Gm Interface) and between various IMS nodes (ISC, Mw, Mg, Mr, Mi, Mj, Mx, Mk and Mm Interfaces). The SIP usage in IMS is defined in [38]. Diameter (Cx, Dx, Dh and Sh Interfaces) and H.248 (Mp) are the other protocols used in IMS.

The UE primarily signals with the CSCFs for the services it wishes to use, and in addition some service specific signalling may be run directly with the Application Servers. The signalling may also be network originated for terminating services. The *Session*

Management and Routing functions are handled by the CSCFs that are in control of the UE's registration in IMS. For that purpose they signal with the *Databases* to get the appropriate information. The CSCFs are also in control of the UE's service sessions in IMS, and for that purpose they may need to signal with one or more of the *Services Elements* to know what kind of connectivity is needed for the service in question, and then with the connectivity layer through the Rx interface to make corresponding requests to bearer resources. Finally, the CSCFs may need to signal with one or more of the *Inter-working Elements* to control the interconnection between networks. Whenever the UP flow is routed through one of the IMS elements, it is done through the Mb interface that connects IMS to IP networks. The following sections introduce the main functions in the functional groups highlighted in Figure 3.19.

Most IMS elements responsible for session management and routing or inter-working are involved in collecting charging information. Rf and Ro interfaces are the main IMS charging interfaces (see section 3.7.1). For simplicity, charging related nodes and interfaces are not shown in Figure 3.19.

3.6.2 Session Management and Routing

The Call State Control Function (CSCF) is the central element in SIP signalling between the UE and the IMS, and it takes care of the UE's registration to the IMS, and service session management. The registration includes authentication. The primary authentication method is IMS-AKA [39], but other methods such as http digest [40] may also be used. CSCF is defined to have three different roles that may reside in the same node, or separate nodes connected through the Mw interface, and all are involved in the UE-related SIP signalling transactions:

- The Serving CSCF (S-CSCF) locates in the user's home network, and it will maintain the user's registration and session state. At registration, it interfaces with the HSS to receive the subscription profile, including authentication information, and it will authenticate the UE. For the service sessions, the S-CSCF signals with the UE through the other CSCFs, and may also interact with the Application Servers (ASs) or the MRFCs for setting up the service session properly. It also carries the main responsibility for controlling the Inter-working Elements. The S-CSCF may also need to interact with MGCF for inter-working with CS networks, or with other multimedia networks for UE requested services.
- The Interrogating CSCF (I-SCSF) is located at the border of the home network, and it is responsible for finding out the UE's registration status, and either assigning a new S-CSCF or routing to the right existing S-CSCF. The request may come from Proxy CSCF (P-CSCF), from other multimedia networks, or from CS networks through the Media Gateway Control Function (MGCF). Also I-CSCF may need to interact with the ASs for service handling. The Ma interface is used for this when Public Service Identity (PSI) is used to identify the service, and the I-CSCF can route the request directly to the proper AS.
- The (P-CSCF) is the closest IMS node the UE interacts with, and it is responsible for all functions related to controlling the IP connectivity layer, i.e. the EPS. For this purpose the P-CSCF contains the Application Function (AF) that is a logical element for the PCC concept, which is described in section 3.7.1. The P-CSCF is typically located in the same network as the EPS, but the Rel-8 includes a so-called Local Breakout

concept that allows P-CSCF to remain in the home network, while PCRF in the visited network may still be used.

In addition to the above-mentioned three CSCF roles, a fourth role, the Emergency CSCF (E-CSCF), has been defined. As the name indicates, the E-CSCF is dedicated to handling the emergency call service in IMS. The E-CSCF connects to the P-CSCF via the Mw interface, and these nodes must always be in the same network. In addition, the E-CSCF is also connected to a Location Retrieval Function (LRF) through the Mi Interface. The LRF can provide the location of the UE, and routing information to route the emergency call appropriately. The E-CSCF and LRF are not shown in Figure 3.19 for simplicity

The CSCFs are connected to each other with the Mw interface, and to other multimedia networks through the Mm interface. Interconnection between CSCFs in different operators' networks may be routed through a common point called the Interconnection Border Control Function (IBCF). See section 3.6.5.

3.6.3 Databases

The Home Subscriber Server (HSS) is the main database used by the IMS. The HSS contains the master copy of subscription data, and it is used in much the same way as with the IP connectivity layer. It provides the location and authentication information based on requests from the I- or S-CSCF, or the AS. The interface between the HSS and the Services Elements will be either Sh or Si depending on the type of services elements. The Sh interface is used in case of SIP or OSA service capability server and the Si when CAMEL based AS is in question.

When there is more than one addressable HSS, another database called the Subscription Locator Function (SLF) may be used to find the right HSS.

3.6.4 Services Elements

The actual service logic is located in the Application Servers (AS). A variety of different services may be provided with different ASs, and the standards do not aim to cover all possible services. Some of the main services are covered in order to facilitate easier inter-working with operators in roaming, and to provide for consistent user experience. One example of a standardized AS is the Telephony Application Server (TAS), which may be used to provide the IMS VoIP service.

The media component of the service can be handled by the Multimedia Resource Function (MRF), which is defined as a separate controller (MRFC) and processor (MRFP). The UP may be routed through MRFP for playing announcements as well as for conferencing and transcoding. For coordination purposes, the MRFC may also be connected to the related AS.

3.6.5 Inter-working Elements

The Inter-working Elements are needed when the IMS interoperates with other networks, such as other IMS networks, or CS networks. The following are the main functions of the standardized inter-working elements:

- The Breakout Gateway Control Function (BGCF) is used when inter-working with CS networks is needed, and it is responsible for selecting where the interconnection will take place. It may select the Media Gateway Control Function (MGCF) if the breakout is to happen in the same network, or it may forward the request to another BGCF in another network. This interaction may be routed through the Interconnection Border Control Function (IBCF).

- The Interconnection Border Control Function (IBCF) is used when interconnection between operators is desired to be routed through defined points, which hide the topology inside the network. The IBCF may be used in interconnection between CSCFs or BGCFs and it is in control of Transition Gateway (TrGW), which is used for the same function in the UP. Note that the IBCF–TrGW interface is not fully specified in Release 8. The IBCFs and the TrGWs in different operators' networks may be interconnected to each other via the Ici and Izi interfaces respectively, and together they comprise the Inter IMS Network to Network Interface (II-NNI).
- The Media Gateway Control Function (MGCF) and IMS-Media Gateway (IMS-MGW) are the CP and UP nodes for inter-working with the CS networks such as the legacy 3GPP networks with CS domain for GERAN or UTRAN, or for PSTN/ISDN. Both incoming and outgoing IMS VoIP calls are supported with the required signalling inter-working and transcoding between different voice coding schemes. The MGCF works in the control of either the CSCF or BGCF.

3.7 PCC and QoS

3.7.1 PCC

Policy and Charging Control (PCC) has a key role in the way users' services are handled in the Release 8 LTE/SAE system. It provides a way to manage the service related connections in a consistent and controlled way. It determines how bearer resources are allocated for a given service, including how the service flows are partitioned to bearers, what QoS characteristics those bearers will have, and finally, what kind of accounting and charging will be applied. If an operator uses only a very simple QoS model, then a static configuration of these parameters may be sufficient, but Release 8 PCC allows the operator to set these parameters dynamically for each service and even each user separately.

The PCC functions are defined in [5] and the PCC signalling transactions as well as the QoS parameter mapping are defined in [41]. Figure 3.20 shows the PCC functions and interfaces in the basic configuration when PCC is applied in one operator's network.

The primary way to set up service flows in Release 8 is one where the UE first signals the request for the service in the Service Layer, and the Application Function (AF) residing in that layer contacts the Policy and Charging Resource Function (PCRF) for appropriate bearer resources. The PCRF is in charge for making the decisions on what PCC to use for the service in question. If subscriber specific policies are used, then the PCRF may enquire subscription related policies from the Subscription Profile Repository (SPR). Further details about SPR structure and, for example, its relation to HSS, and the Sp interface are not specified in Release 8. Based on the decision, the PCRF creates the appropriate PCC rules that determine the handling in the EPS.

If the interface from P-GW to the S-GW is based on GTP, the PCRF pushes the PCC rules to the Policy and Charging Enforcement Function (PCEF) residing in the P-GW, and it alone will be responsible for enforcing the PCC rules, e.g. setting up the corresponding

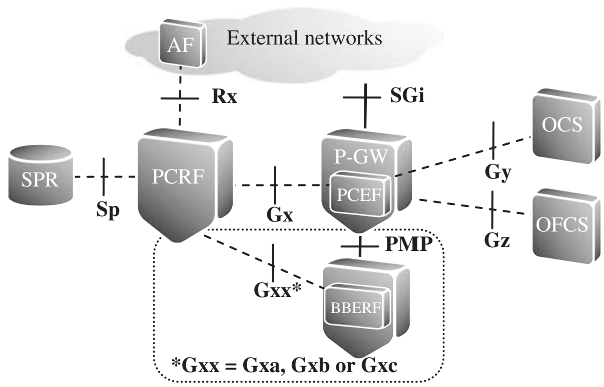

Figure 3.20 Basic PCC functions

dedicated bearers, or modifying the existing bearers so that the new IP service flows can be mapped to them, and by ensuring that only authorized service flows are allowed and QoS limits are not exceeded. In this case the Gxx interface shown in Figure 3.20 does not apply.

If the interface from P-GW towards the AN is based on PMIP, i.e. if it is S5 PMIP, S2a or S2b, there is no means to signal the bearer level information onwards from the P-GW, and the PCRF will create a separate set of QoS rules, and those are first sent to the BBERF, which will handle the mapping between IP service flows and bearers over the AN. Depending on the AN type, the BBERF may reside in S-GW (S5 PMIP), trusted non-3GPP AN, e.g. in HSGW (S2a), or in the EPDG (S2b) for the un-trusted non-3GPP AN (S2b is not supported in Release 8). Also in this case the PCC rules are also sent to PCEF in the P-GW, and it performs the service flow and QoS enforcement.

Release 8 also supports UE initiated bearer activation within the EPS, which is applicable to the case when there is no defined service that both the UE and the serving network could address. In this case the UE signals with the AN and the BBERF requests the service resources from the PCRF. The PCRF makes the PCC decision, and the logic then continues as described above.

The PCC standard [5] defines two charging interfaces, Gy and Gz, which are used for online and offline charging respectively. The Gy interface connects the PCEF to the Online Charging System (OCS), which is used for flow based charging information transfer and control in an online fashion. The Gz interface is used between the P-GW and the Offline Charging System (OFCS), and it is applied when charging records are consolidated in an offline fashion. The charging specifications [42] and [43] further define that the Gy interface is functionally equivalent to the Ro interface that uses Diameter Credit-Control Application as defined in [44]. The Gz interface may be based on either the Rf interface, which relies on the mentioned Diameter Credit-Control Application, or the Ga interface, which uses the 3GPP defined GTP protocol. The Ro and Rf interfaces are also used for charging in IMS, and were originally specified for that purpose.

The PCRF in control of the PCEF/P-GW and the BBERF typically reside in the same operator's network. In the case of roaming, they may reside in different networks, and the S9 interface between PCRFs is used to enable the use of a local PCRF. The S9 interface

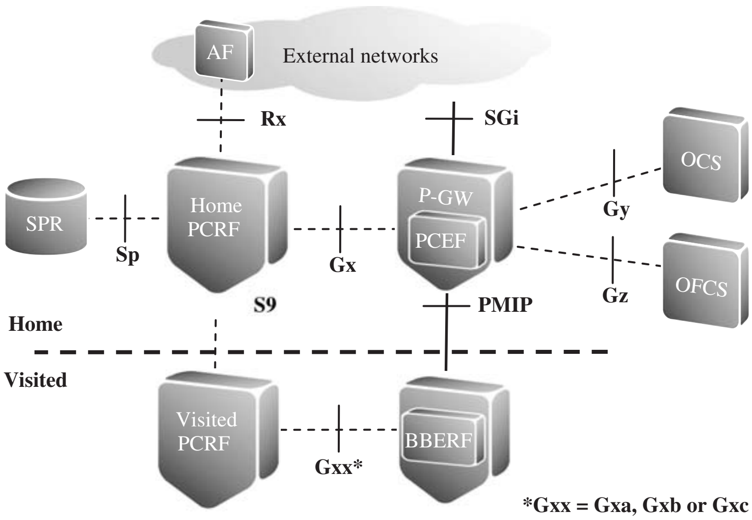

Figure 3.21 PCC functions in roaming with PMIP, home routed model

is defined in [9], and it re-uses the applicable parts from Rx, Gx and Gxx interfaces to convey the information between the PCRFs.

There are two different cases when the S9 interface is used. The first case, which is shown in Figure 3.21, applies when the roaming interface is based on PMIP, and the PCEF and BBERF are in different networks. In this scenario traffic is routed to the home network. In the second case, shown in Figure 3.22, the Local Breakout model is applied, and the P-GW resides in the visited network. The AF and Rx interface will be used from the same network that provides the service in question. The OCS will reside in the home network. As described above, the separate BBERF and Gxx interfaces apply only if PMIP is used from the P-GW in the visited network.

Table 3.7 lists the PCC related interfaces and the protocols, and the standards where they are specified.

3.7.2 QoS

The development of the SAE bearer model and the QoS concept started with the assumption that improvements compared to the existing 3GPP systems with, e.g. UTRAN access, should be made, and the existing model should not be taken for granted. Some potential areas had already been identified. It had not been easy for the operators to use QoS in the legacy 3GPP systems. An extensive set of QoS attributes was available, but it was to some extent disconnected from the application layer, and thus it had not been easy to configure the attributes in the correct way. This problem was emphasized by the fact that the UE was responsible for setting the QoS attributes for a bearer. Also, the bearer model had many layers, each signalling just about the same information. It was therefore agreed that for SAE, only a reduced set of QoS parameters and standardized characteristics would be specified. Also it was decided to turn the bearer set-up logic so that the network resource management is solely network controlled, and the network decides how the parameters

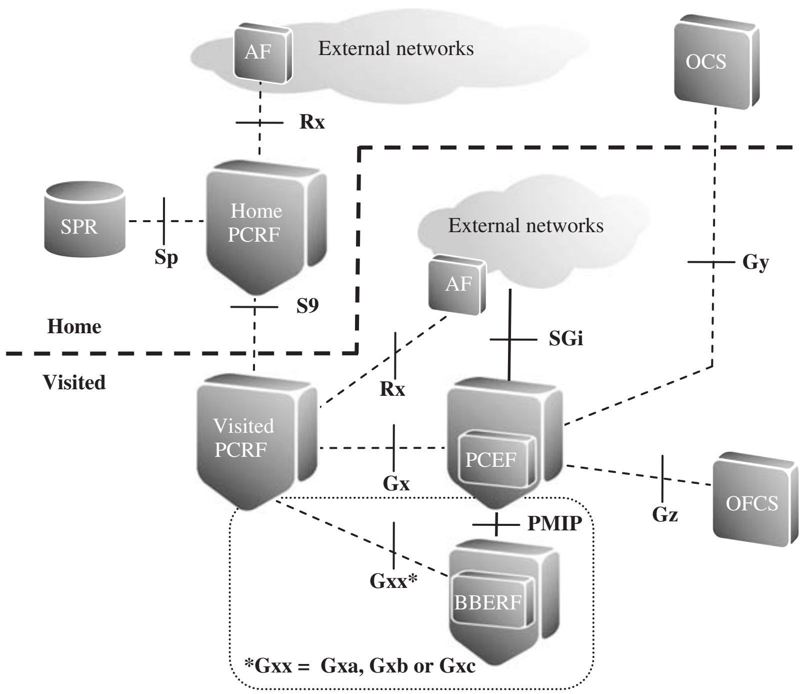

Figure 3.22 PCC functions in roaming, local breakout model

Table 3.7 Summary of PCC interfaces

Interface	Protocols	Specification
Gx	Diameter/SCTP/IP	29.212 [13]
Gxx (Gxa or Gxc)	Diameter/SCTP/IP	29.212 [13]
Rx	Diameter/SCTP/IP	29.214 [14]
S9	Diameter/SCTP/IP	29.215 [45]
Sp	*Not defined in Release 8*	*N.A.*
Gy =		32.240 [42]
Ro	Diameter/SCTP/IP	32.299 [46]
Gz =		32.251 [43]
Rf or	Diameter/SCTP/IP or	32.295 [47] or
Ga	GTP'/UDP or TCP/IP	32.299 [46]

are set, and the main bearer set-up logic consists of only one signalling transaction from the network to the UE and all interim network elements.

The resulting SAE bearer model is shown in Figure 3.23. The bearer model itself is very similar to the GPRS bearer model, but it has fewer layers. EPS supports the always-on concept. Each UE that is registered to the system has at least one bearer called the default bearer available, so that continuous IP connectivity is provided. The default bearer may have quite basic QoS capabilities, but additional bearers may be set up on demand for

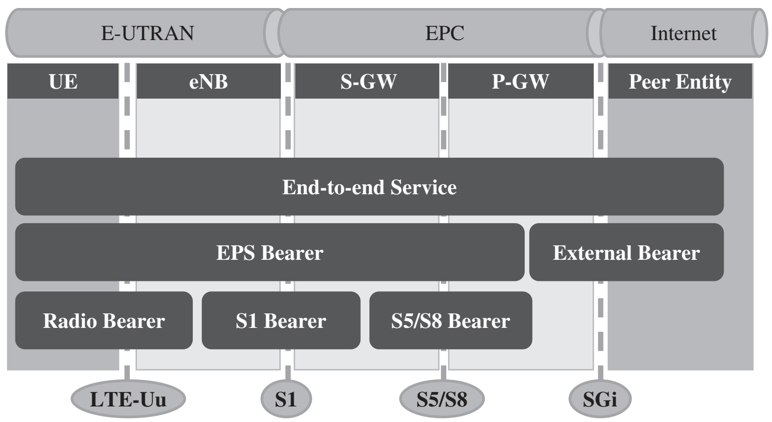

Figure 3.23 SAE Bearer model

services that need more stringent QoS. These are called dedicated bearers. The network may also map several IP flows that have matching QoS characteristics to the same EPS bearer.

The bearer set-up logic works so that the UE first signals on the application layer, on top of the default bearer, to an Application Server (AS) in the operator service cloud, e.g. with IMS, to set up the End-to-end Service. This signalling may include QoS parameters, or simply indication to a known service. The AS will then request the set-up of the corresponding EPS bearer through the PCC infrastructure. There is no separate signalling transaction for the EPS bearer layer, but the EPS bearer is set up together with the signalling for the lower layers, i.e. S5/S8 bearer, S1 Bearer and Radio Bearer. Furthermore, since the eNodeB is responsible for controlling the radio interface transmission in the uplink as well, the UE can operate based on very basic QoS information. The overall goal for network orientation in bearer set-up is to minimize the need for QoS knowledge and configuration in the UE.

Also the QoS parameters were optimized for SAE. Only a limited set of signalled QoS parameters are included in the specifications. They are:

- QoS Class Identifier (QCI): It is an index that identifies a set of locally configured values for three QoS attributes: Priority, Delay and Loss Rate. QCI is signalled instead of the values of these parameters. Ten pre-configured classes have been specified in two categories of bearers, Guaranteed Bit Rate (GBR) and Non-Guaranteed Bit-Rate (Non-GBR) bearers. In addition operators can create their own classes that apply within their network. The standard QCI classes and the values for the parameters within the class are shown in Table 3.8.
- Allocation and Retention Priority (ARP): Indicates the priority of the bearer compared to other bearers. This provides the basis for admission control in bearer set-up, and further in a congestion situation if bearers need to be dropped.
- Maximum Bit Rate (MBR): Identifies the maximum bit rate for the bearer. Note that a Release 8 network is not required to support differentiation between the MBR and GBR, and the MBR value is always set to equal to the GBR.

Table 3.8 QoS parameters for QCI

QCI	Resource type	Priority	Delay budget	Loss rate	Example application
1	GBR	2	100 ms	1e-2	VoIP
2	GBR	4	150 ms	1e-3	Video call
3	GBR	5	300 ms	1e-6	Streaming
4	GBR	3	50 ms	1e-3	Real time gaming
5	Non-GBR	1	100 ms	1e-6	IMS signalling
6	Non-GBR	7	100 ms	1e-3	Interactive gaming
7	Non-GBR	6	300 ms	1e-6	Application with TCP:
8	Non-GBR	8			browsing, email, file
9	Non-GBR	9			download, etc.

- Guaranteed Bit Rate (GBR): Identifies the bit rate that will be guaranteed to the bearer.
- Aggregate Maximum Bit Rate (AMBR): Many IP flows may be mapped to the same bearer, and this parameter indicates the total maximum bit rate a UE may have for all bearers in the same PDN connection.

Table 3.8 shows the QoS parameters that are part of the QCI class, and the nine standardized classes. The QoS parameters are:

- Resource Type: Indicates which classes will have GBR associated to them.
- Priority: Used to define the priority for the packet scheduling of the radio interface.
- Delay Budget: Helps the packet scheduler to maintain sufficient scheduling rate to meet the delay requirements for the bearer.
- Loss Rate: Helps to use appropriate RLC settings, e.g. number of re-transmissions.

References

[1] 3GPP TS 23.401, 'General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (Release 8)'.
[2] 3GPP TS 23.402, 'Architecture enhancements for non-3GPP accesses (Release 8)'.
[3] M. Poikselkä et al., ' The IMS: IP Multimedia Concepts and Services', 2nd edition, Wiley, 2006.
[4] 3GPP TS 33.401, 'Security Architecture (Release 8)'.
[5] 3GPP TS 23.203, 'Policy and charging control architecture (Release 8)'.
[6] 3GPP TS 36.413, 'Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description (Release 8)'.
[7] 3GPP TS 36.423, 'Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 Application Protocol (X2AP) (Release 8)'.
[8] 3GPP TS 29.274, 'Evolved GPRS Tunnelling Protocol (eGTP) for EPS (Release 8)'.
[9] 3GPP TS 36.413, 'Evolved Universal Terrestrial Radio Access (E-UTRA); S1 Application Protocol (S1AP) (Release 8)'.
[10] 3GPP TS 29.275, 'PMIP based Mobility and Tunnelling protocols (Release 8)'.
[11] 3GPP TS 29.061, 'Inter-working between the Public Land Mobile Network (PLMN) supporting packet based services and Packet Data Networks (PDN) (Release 8)'.
[12] 3GPP TS 29.272, 'MME Related Interfaces Based on Diameter Protocol (Release 8)'.
[13] 3GPP TS 29.212, 'Policy and charging control over Gx reference point (Release 8)'.
[14] 3GPP TS 29.214, 'Policy and charging control over Rx reference point (Release 8)'.
[15] 3GPP TS 24.301, 'Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS) (Release 8)'.
[16] H. Holma, A. Toskala, ' WCDMA for UMTS – HSPA Evolution and LTE', 5th edition, Wiley, 2010.

[17] T. Halonen, J. Romero, J. Melero, 'GSM, GPRS and EDGE Performance: Evolution Towards 3G/UMTS', 2nd edition, Wiley, 2003.
[18] 3GPP TS 23.060, 'General Packet Radio Service (GPRS); Service description; Stage 2 (Release 8)'.
[19] 3GPP TS 23.272, 'Circuit Switched (CS) fallback in Evolved Packet System (EPS); Stage 2 (Release 8)'.
[20] 3GPP TS 23.216, 'Single Radio Voice Call Continuity (SRVCC); Stage 2 (Release 8)'.
[21] 3GPP TS 29.118, 'Mobility Management Entity (MME) – Visitor Location Register (VLR) SGs interface specification (Release 8)'.
[22] 3GPP TS 29.280, '3GPP EPS Sv interface (MME to MSC) for SRVCC (Release 8)'.
[23] 3GPP TS 33.402, 'Security aspects of non-3GPP accesses (Release 8)'.
[24] IETF Internet-Draft, draft-arkko-eap-aka-kdf, 'Improved Extensible Authentication Protocol Method for 3rd Generation Authentication and Key Agreement (EAP-AKA)'. (J. Arkko, V. Lehtovirta, P. Eronen) 2008.
[25] 3GPP TS 23.234, '3GPP system to Wireless Local Area Network (WLAN) interworking; System description (Release 7)'.
[26] IETF RFC 4306, 'Internet Key Exchange (IKEv2) Protocol.' C. Kaufman, Editor, 2005.
[27] 3GPP TS 24.303, 'Mobility management based on Dual-Stack Mobile IPv6 (Release 8)'.
[28] 3GPP TS 29.273, 'Evolved Packet System (EPS); 3GPP EPS AAA interfaces (Release 8)'.
[29] 3GPP TS 24.302, 'Access to the Evolved Packet Core (EPC) via non-3GPP access networks (Release 8)'.
[30] 3GPP TS 24.304, 'Mobility management based on Mobile IPv4; User Equipment (UE) – foreign agent interface (Release 8)'.
[31] 3GPP2 Specification X.P0057, 'E-UTRAN – HRPD Connectivity and Interworking: Core Network Aspects (2008)'.
[32] 3GPP2 Specification X.P0022, 'E-UTRAN – HRPD Connectivity and Interworking: Access Network Aspects (E-UTRAN – HRPD IOS) (2008)'.
[33] 3GPP TS 29.276, 'Optimized Handover Procedures and Protocols between EUTRAN Access and cdma2000 HRPD Access (Release 8)'.
[34] 3GPP TS 29.277, 'Optimized Handover Procedures and Protocols between EUTRAN Access and 1xRTT Access (Release 8)'.
[35] 3GPP2 Specification A.S0008-C, 'Interoperability Specification (IOS) for High Rate Packet Data (HRPD) Radio Access Network Interfaces with Session Control in the Access Network (2007)'.
[36] 3GPP TS 23.002, 'Network architecture (Release 8)'.
[37] 3GPP TS 23.228, 'IP Multimedia Subsystem (IMS); Stage 2 (Release 8)'.
[38] 3GPP TS 24.229, 'IP multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3 (Release 8)'.
[39] 3GPP TS 33.203, '3G security; Access security for IP-based services (Release 8)'.
[40] IETF RFC 2617, 'HTTP Authentication: Basic and Digest Access Authentication', J. Franks, P. Hallam-Baker, J. Hostetler, S. Lawrence, P. Leach, A. Luotonen, L. Stewart (1999).
[41] 3GPP TS 29.213, 'Policy and Charging Control signalling flows and QoS parameter mapping (Release 8)'.
[42] 3GPP TS 32.240, 'Telecommunication management; Charging management; Charging architecture and principles (Release 8)'.
[43] 3GPP TS 32.251, 'Telecommunication management; Charging management; Packet Switched (PS) domain charging (Release 8)'.
[44] IETF RFC 4006, 'Diameter Credit-Control Application', H. Hakala, L. Mattila, J.-P. Koskinen, M. Stura, J. Loughney (2005).
[45] 3GPP TS 29.215, 'Policy and Charging Control (PCC) over S9 reference point (Release 8)'.
[46] 3GPP TS 32.299, 'Telecommunication management; Charging management; Diameter charging applications (Release 8)'.
[47] 3GPP TS 32.295, 'Telecommunication management; Charging management; Charging Data Record (CDR) transfer (Release 8)'.

4

Introduction to OFDMA and SC-FDMA and to MIMO in LTE

Antti Toskala and Timo Lunttila

4.1 Introduction

As discussed in Chapter 1, LTE multiple access is different to that of WCDMA. In LTE the downlink multiple access is based on the Orthogonal Frequency Division Multiple Access (OFDMA) and the uplink multiple access is based on the Single Carrier Frequency Division Multiple Access (SC-FDMA). This chapter will introduce the selection background and the basis for both SC-FDMA and OFDMA operation. The basic principles behind the multi-antenna transmission in LTE, using Multiple Input Multiple Output (MIMO) technology, is also introduced. The intention of this chapter is to illustrate the multiple access principles in a descriptive way without too much mathematics. For those interested in the detailed mathematical notation, two selected references are given that provide a mathematical treatment of the different multiple access technologies, covering both OFDMA and SC-FDMA.

4.2 LTE Multiple Access Background

A Single Carrier (SC) transmission means that information is modulated only to one carrier, adjusting the phase or amplitude of the carrier or both. Frequency could also be adjusted, but in LTE this is not affected. The higher the data rate, the higher the symbol rate in a digital system and thus the bandwidth is higher. With the use of simple Quadrature Amplitude Modulation (QAM), with the principles explained, for example in [1], the transmitter adjusts the signal to carry the desired number of bits per modulation symbol. The resulting spectrum waveform is a single carrier spectrum, as shown in Figure 4.1, with the spectrum mask influenced (after filtering) by the pulse shape used.

With the Frequency Division Multiple Access (FDMA) principle, different users would then be using different carriers or sub-carriers, as shown in Figure 4.2, to access the system simultaneously having their data modulation around a different center frequency. Care must be now taken to create the waveform in such a way that there is no excessive

LTE for UMTS: Evolution to LTE-Advanced, Second Edition. Edited by Harri Holma and Antti Toskala.