

# Chapter 4

## 5G System Architecture



**Coauthored with: John Kaippallimalil and Amanda Xiang**

This chapter provides a concise description of the 3GPP standards for the 5G System (5GS) and highlights the major enhancements over 4G. Network slicing, virtualization, and edge computing with enhanced connectivity, session/mobility management are some key enhancements. They cater to the requirements of a diverse set of services requiring low latency, high reliability, or a massive number of connections for short, intermittent periods over the same network.

It starts an overview of the end-to-end 5GS. It then describes the service-based architecture (SBA) that organizes core network functions into a set of service-oriented functions and inherent support for virtualization. Network slicing for supporting the broad range of services over the same network is described. An overview of connection and session management including new service and session continuity modes for mobility is given. It then outlines how the 5GC interworks with a 4G Evolved Packet Core (EPC). A detailed description of the control plane (CP) and user plane (UP) protocols in 5GC is then given. Support for virtualized deployments and edge computing is then elucidated. The chapter ends with a discussion on policy and charging for roaming and non-roaming cases.

### 4.1 5G System Architecture

As in the 4G (LTE/EPC) and previous generations, the 3GPP 5G system defines the architecture for communication between a User Equipment (UE) and an end point, such as an Application Server (AS) in the Data Network (DN), or another UE. The interaction between the UE and the Data Network is via the Access Network and Core Network as defined by 3GPP Standards. Figure 4.1 depicts a simple

---

Huawei Technologies, USA  
Plano, TX

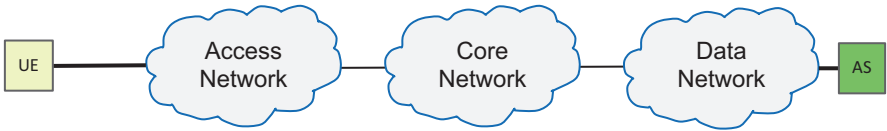


Fig. 4.1 End-to-end architecture

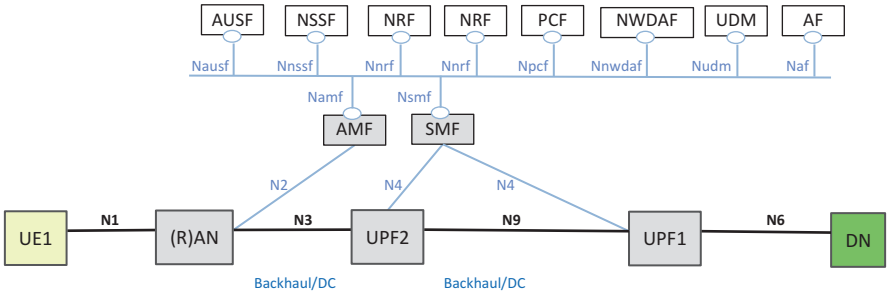


Fig. 4.2 5G System architecture

representation of an end-to-end architecture. In this chapter we focus on describing the 5G Core as defined by 3GPP 5G standards for PLMN [1–3]. The Access Network in 3GPP is referred to as Radio Access Network (RAN).

At a very high level the Core and RAN consist of several Network Functions which are associated with Control Plane and User Plane functionalities. The actual data (also refer it as user data) is normally transported via a path in the User Plane, while the Control Plane is used to establish the path in the User Plane. The Short Message Service (SMS) is an exception in which the data (short message) is communicated via the Control Plane.

The 5G System architecture (5GS) is represented in two ways in the 3GPP standards, one is a service-based representation in which the control plane network functions access each other’s services, and the other is a reference point representation in which the interaction between the network functions is shown with point-to-point reference points. In this chapter we use the service-based representation since the 5G architecture is defined as service-based architecture. The 3GPP 5GS service-based non-roaming reference architecture is shown in Fig. 4.2. In Release-15 specifications the Service-based interfaces are defined within the Control Plane only. In 3GPP terminology, “a network function can be implemented either as a network element on a dedicated hardware, as a software instance running on a dedicated hardware, or as a virtualized function instantiated on an appropriate platform, e.g. on a cloud infrastructure.”

The EPC in Release-14 was enhanced with an optional feature that allowed separation of control plane and user plane. In this feature, the Serving Gateway (SGW) and Packet Gateway (PGW) are divided into distinct control plane and user plane functions (e.g., SGW-C and SGW-U). This optional feature provided more flexibility and efficiency in network deployment—See [4] for details. In 5G architecture, the separation of control plane and user plane is an inherent capability. The Session

Management Function (SMF) handles the control plane functionality for setup and management of sessions while the actual user data is routed through the User Plane Function (UPF). The UPF selection (or re-selection) is handled by SMF. The deployment options allow for centrally located UPF and/or distributed UPF located close to or at the Access Network.

In EPC, the mobility management functionality and session management functionality are handled by Mobility Management Entity (MME). In 5GC, these functionalities are handled by separate entities. The Access and Mobility Management function (AMF) handles the mobility management and procedures. AMF is termination point for control plane connection from (Radio) Access Network ((R)AN) and UE. The connection between UE and AMF (which traversed through RAN) is referred to as Non-Access Stratum (NAS). The Session Management Function (SMF) handles the session management procedures. The separation of the mobility and session management functionalities allows for one AMF to support different Access Networks (3GPP and non-3GPP), while SMF can be tailored for specific Accesses.

Figure 4.3 shows a Roaming architecture with local breakout at the Visited PLMN (VPLMN). In this scenario the Unified Data Management (UDM), which includes the subscription information, and Authentication Server Function (AUSF), which includes authentication/authorization data, are located in the Home PLMN (HPLMN). There are Security Edge Protection Proxies (SEPP) that protect the communication between the Home and Visited PLMNS. UE communicates to Data Network (DN) via the User Plane Functions (UPF) in the VPLMN. The AMF and the Session Management Function (SMF) which handle the mobility and the session management for the UE are located in the VPLMN as well.

4.2 5G Core (5GC) Service-Based Architecture

A major change in the 5G Core (5GC) architecture compared to EPC and the previous generations is the introduction of the service-based architecture. In EPC architecture the control plane functions communicate with each other via the direct interfaces (or reference points) with a standardized set of messages. In the service-based architecture, the Network Functions (NF), using a common framework,

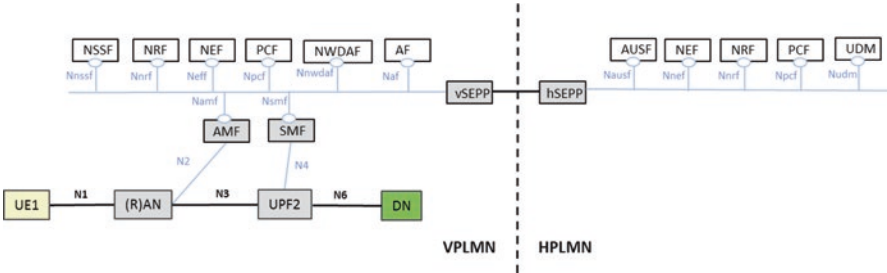


Fig. 4.3 Roaming 5G System architecture—Local breakout scenario

expose their services for use by other network functions. In the 5GC architecture model the interfaces between the networks functions are referred to as Service-Based Interfaces (SBI). The Service Framework defines the interaction between the NFs over SBI using a Producer-Consumer model. As such a service offered by a NF (Producer) could be used by another NF (Consumer) that is authorized to use the service. The services are generally referred to as “NF Service” in 3GPP specifications.

The interaction between the NFs may be a “Request-response” or a “Subscribe-Notify” mechanism. In the “Request-response” model a NF (consumer) request another NF (producer) to provide a service and/or perform a certain action. See Fig. 4.4. In “Subscribe-Notify” model a NF (consumer) subscribes to the services offered by another NF (producer) which notifies the subscriber of the result (Fig. 4.5).

As can be seen in Fig. 4.3, in the 5G System Architecture, each network function has an associated service-based interface designation. For example, “Namf” designates the services exhibited by the Access and Mobility Management function (AMF). 3GPP specifications define a set of Services that are offered/supported by each Network Function. For example, the NF services specified for AMF are shown in Table 4.1. The details for Service descriptions are described in [2].

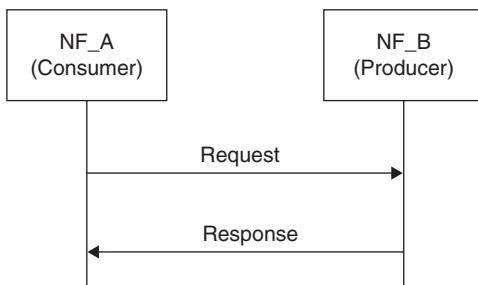
There are three main procedures associated with the Service Framework as defined in 3GPP—see [1, 5] for details:

**NF service registration and de-registration:** to make the Network Repository Function (NRF) aware of the available NF instances and supported services.

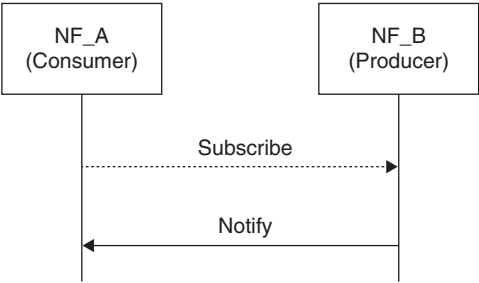
**NF service discovery:** enables a NF (Consumer) to discover NF instance(s) (Producer) that provide the expected NF service(s). A NF typically performs a Services Discovery procedure with NRF for NF and NF service discovery.

**NF service authorization:** to ensure the NF Service Consumer is authorized to access the NF service provided by the NF Service Provider (Producer).

**Fig. 4.4** “Request-response” NF Service illustration



**Fig. 4.5** “Subscribe-Notify” NF Service illustration 1



**Table 4.1** Namf services

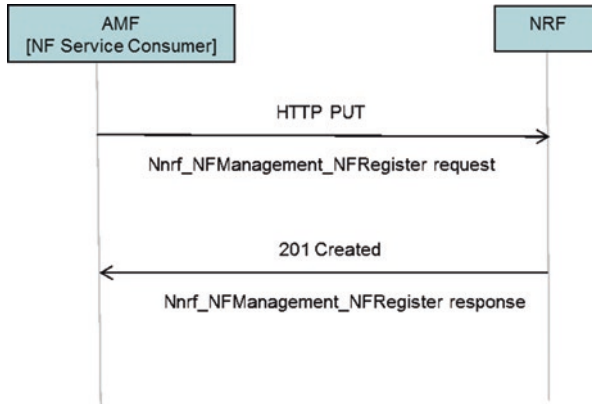
Service name	Description
Namf_communication	Enables an NF consumer to communicate with the UE and/or the AN through the AMF This service enables SMF to request EBI allocation to support interworking with EPS
Namf_EventExposure	Enables other NF consumers to subscribe or get notified of the mobility-related events and statistics
Namf_MT	Enables an NF consumer to make sure UE is reachable
Namf_Location	Enables an NF consumer to request location information for a target UE

**4.2.1 Example of NF Service Registration**

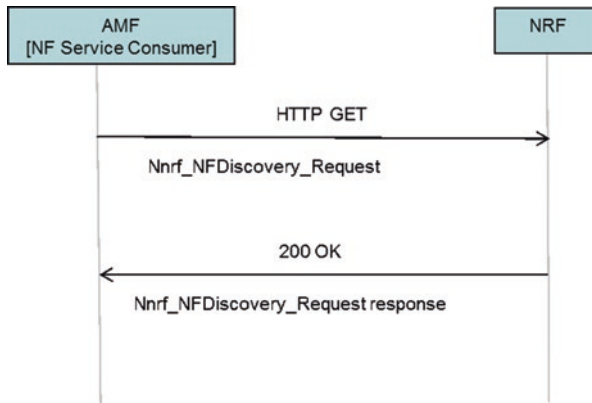
In this example (Fig. 4.6) AMF as the NF service consumer sends a HTTP PUT request to NRF with the resource URI representing the NF Instance. The request contains Nnrf\_NFManagement\_NFRegister request message (the NF profile of NF service consumer) to NRF to inform the NRF of its NF profile. The NF profile of NF service consumer includes information such as NF type, FQDN or IP address of NF, and Names of supported services. The NRF authorizes the request and upon success stores the NF profile of NF service consumer and marks the NF service consumer available. The NRF acknowledge the success of AMF Registration by returning a HTTP 201Created response containing the Nnrf\_NFManagement\_NF Register response (including the NF profile). See 3GPP TS 23.501 [1] and 3GPP TS 29.510 [5] for details.

**4.2.2 Example of NF Service Discovery**

In this example (Fig. 4.7) the AMF as NF service consumer intends to discover NF instances or services available in the network for a targeted NF type. The AMF sends HTTP GET request to NRF in the same PLMN by invoking Nnrf\_NFDiscovery\_Request. This request contains Expected NF service Name, NF Type of the expected NF instance, and NF type of the NF consumer and may also include



**Fig. 4.6** Nnrf\_NF Registration procedure



**Fig. 4.7** Nnrf\_NF service Discovery

other information/parameters such as Subscription Permanent Identifier (SUPI) and AMF Region ID. The NRF authorizes the request, and if allowed the NRF determines the discovered NF instance(s) or NF service instance(s) and provides the search results to the NF service consumer (e.g., AMF) in a HTTP 200 OK. See 3GPP TS 23.501 [1] and 3GPP TS 29.510 [5] for details.

### 4.3 Network Slicing

From the 3GPP point of view, a 5G network slice is viewed as a logical network with specific functions/elements dedicated for a particular use case, service type, traffic type, or other business arrangements with agreed-upon Service-level

Agreement (SLA). It is important to note that 3GPP only defines network slicing for 3GPP defined system architecture and does not address transport network slicing or resource slicing of components.

The most commonly discussed slice types in industry are enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communications (URLLC), and massive IoT (mIoT). However, there could be many more network slices. In 4G systems (EPS) there is an optional feature called eDeco to support Dedicated Core Networks (DCNs) to allow selection of the core networks based on UE’s subscription and usage type. The network slicing in 5GS is a more complete solution that provides capabilities for composing multiple dedicated end-to-end networks as slices.

An end-to-end Network Slice includes the Core Network Control Plane and User Plane Network Functions as well as the Access Network (AN). The Access Network could be the Next Generation (NG) Radio Access Network described in 3GPP TS 38.300 [6], or the non-3GPP Access Network with the Non-3GPP InterWorking Function (N3IWF). To emphasize that there could be multiple instances of a network slice, the 3GPP 5GS specifications define the term “Network Slice instance” as set of Network Function instances and resources (e.g., compute, storage, and networking resources) which form a Network Slice.

In 5GS, the Network Slice Selection Assistance Information (NSSAI) is a collection of identifications for network slices. A network slice is identified by a term referred to as Single-NSSAI (S-NSSAI). The S-NSSAI signaled by the UE to the network assists the network in selecting a particular Network Slice instance. An S-NSSAI comprises a Slice/Service type (SST) and an optional Slice Differentiator (SD) which may be used to differentiate among multiple Network Slices of the same Slice/Service type.

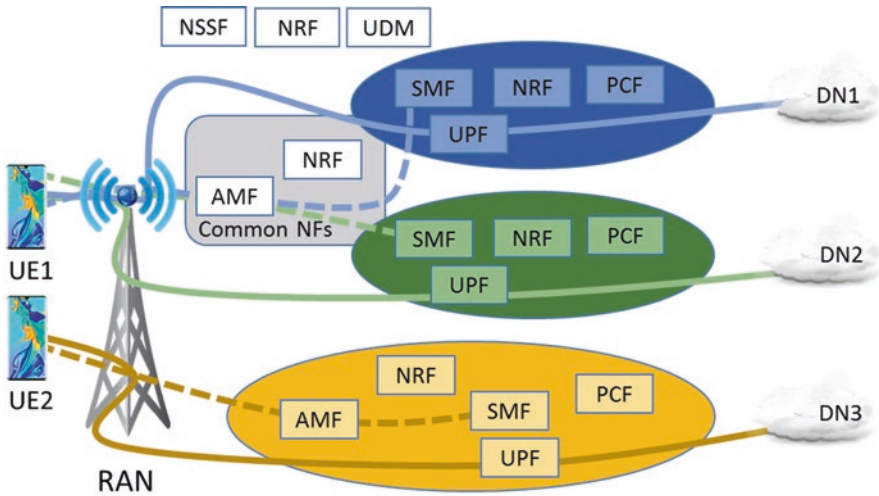
An S-NSSAI can have standard values or nonstandard values. The S-NSSAI with standard value means that it comprises an SST with a standardized SST value. An S-NSSAI with a nonstandard value identifies a single Network Slice within the PLMN with which it is associated.

3GPP has defined some standardized SST values in TS 23.501 [1]. These SST values are to reflect the most commonly used Slice/Service Types and will assist with global interoperability for slicing. The support of all standardized SST values is not required in a PLMN (Table 4.2).

Figure 4.8 shows an example of three Network Slices in 5GS. For Slice 1 and Slice 2, the Access and Mobility Management Function (AMF) instance that is

**Table 4.2** Standardized SST values

Slice/service type	SST value	Characteristics
eMBB	1	Slice suitable for the handling of 5G enhanced mobile broadband
URLLC	2	Slice suitable for the handling of ultra-reliable low latency communications
MIoT	3	Slice suitable for the handling of massive IoT



**Fig. 4.8** Example of Network Slices in 5GS

serving the UE1 and UE2 is common (or logically belongs) to all the Network Slice instances that are serving them. The UE in Slice 3 is served by another AMF. Other network functions, such as the Session Management Function (SMF) or the User Plane Function (UPF) may be specific to each Network Slice.

The Network Slice instance selection for a UE is normally triggered as part of the registration procedure by the first AMF that receives the registration request from the UE. The AMF retrieves the slices that are allowed by the user subscription and may interact with the Network Slice Selection Function (NSSF) to select the appropriate Network Slice instance (e.g., based on Allowed S-NSSAIs, PLMN ID). The NSSF contains the Operators' policies for slice selection. Alternatively, the slice selection policies may be configured in the AMF.

The data connection between the UE and Data Network (DN) is referred to as PDU session in 5GS. In 3GPP Release-15 a PDU Session is associated to one S-NSSAI and one DNN (Data Network Name). The establishment of a PDU session is triggered when the AMF receives a Session Management message from UE. The AMF discovers candidate Session Management Functions (SMF) using multiple parameters (including the S-NSSAI provided in the UE request) and selects the appropriate SMF. The selection of the User Plane Function (UPF) is performed by the SMF. The Network Repository Function (NRF) is used for the discovery of the required Network Functions using the selected Network Slice instance—the detailed procedures are specified in 3GPP TS 23.502 [2]. The data transmission can take place after a PDU session to a Data Network is established in a Network Slice. The S-NSSAI associated with a PDU Session is provided to the (R)AN, and to the policy and charging entities, to apply slice specific policies.

For roaming scenarios, S-NSSAI values applicable in the Visited PLMN (VPLMN) are used to discover a SMF instance in the VPLMN and in Home-Routed deployments S-NSSAI values applicable in the Home PLMN (HPLMN) are also used to discover a SMF instance in the HPLMN.



## 4.4 Registration, Connection, and Session Management

This section provides an overview of the high-level features for registration, connection, and session management in 5GS.

### 4.4.1 Registration Management

A user (UE) registers periodically with the network to remain reachable, in case of mobility or to update its capabilities. During initial registration, the UE is authenticated and access authorization information based on subscription profile in UDM (Unified Data Management) is configured on AMF, and the identifier of the serving AMF is stored in UDM. When this registration process is complete, the state for the UE is 5GMM-REGISTERED (at the UE and AMF). In the 5GMM-REGISTERED state, the UE can perform periodic registration updates to notify that it is active, and mobility registration update if the serving cell is not in the list of TAI (Tracking Area Identifier) that was provisioned during registration. The UE/AMF state machine will transition to 5GMM-DEREGISTERED when the timers expire (and the UE has not performed periodic registration) or if the UE or network explicitly deregisters. When a UE is served by 3GPP and non-3GPP accesses of the same PLMN, the AMF associates multiple access specific registration contexts with the same 5G-GUTI (Globally Unique Temporary Identifier).

### 4.4.2 Connection Management

For signaling between the UE and AMF, NAS (Non-Access Stratum) connection management procedures are used. When the UE is in 5GMM-REGISTERED state and has no NAS connection established with the UE (i.e., in 5GMM-IDLE state), the UE will respond to paging (unless in MICO (Mobile Initiated Connections Only) mode) by performing a Service Request procedure and enter into 5GMM-CONNECTED mode. The UE will perform a Service Request procedure also and enter into 5GMM-CONNECTED mode if it has signaling or user data to send. The AMF enters into 5GMM-CONNECTED mode for that UE when the N2 connection (between Access Network and AMF) is established. Figure 4.9 shows the transitions between 5GMM-IDLE and 5GMM-CONNECTED for both the UE and the AMF.

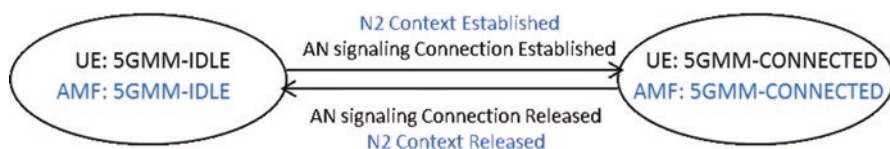


Fig. 4.9 Connection Management State Transitions

The UE can go from 5GMM-CONNECTED to 5GMM-IDLE when the Access Network (AN) signaling connection is released on inactivity (RRC Idle state). The AMF enters 5GMM-IDLE for the UE when the NGAP signaling connection (N2 context) and N3 user plane connection are released. When the UE is in RRC Inactive state and 5GMM-CONNECTED, the UE reachability and paging is managed by the RAN. The AMF provides assistance by configuring UE specific DRX (Discontinuous Reception) values, registration area, periodic registration update timer value and MICO mode indication. The UE monitors for paging with the 5G S-TMSI (Temporary Mobile Subscriber Identity) and RAN identifier.

*Note:* TS 23.501 [1] uses states RM-REGISTERED and RM-DEREGISTERED, while TS 24.501 [7] uses 5GMM-REGISTERED and 5GMM-DEREGISTERED for the same set of states. Similarly, 23.501 uses CM-IDLE and CM-CONNECTED, while TS 24.501 [7] uses 5GMM-IDLE and 5GMM-CONNECTED to refer to the same set of states.

4.4.3 Registration Call Flow

Figure 4.10 provides an overview<sup>1</sup> of the registration management procedure call flow. Initially, the UE and network states are RRC-Idle, 5GMM-IDLE, and 5GMM-DEREGISTERED. An RRC (Radio Resource Connection) layer is needed between

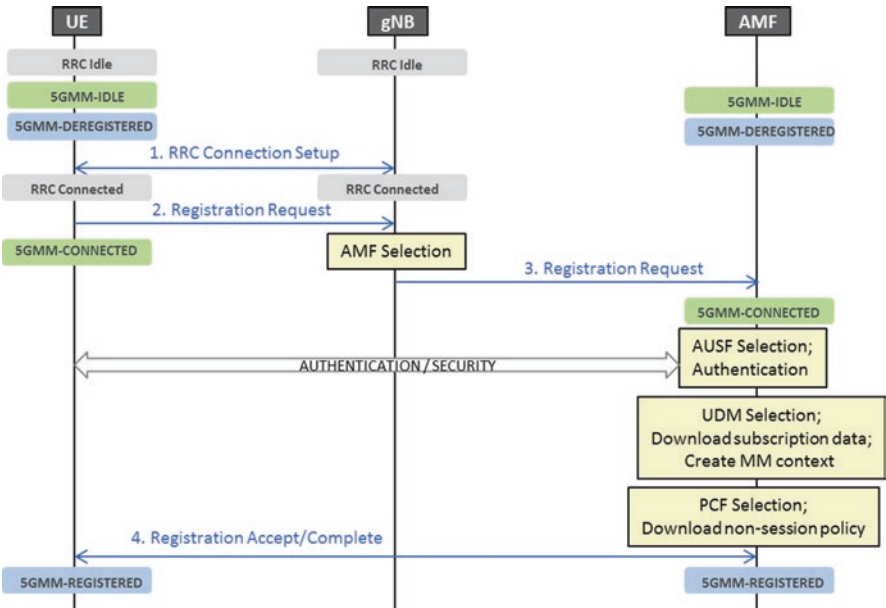


Fig. 4.10 Registration Management

<sup>1</sup> Complete call flow procedure and parameter details are found in 3GPP TS 23.502, 4.2.2 [2].

the UE and gNB to exchange messages. Following the radio link establishment, an RRC connection is established between the UE and gNB (1) (gNB is short term for next generation NodeB).

Once the RRC connection is set up, the UE is ready to start registering itself with the network. The Registration Request (2) includes registration type (initial, mobility registration or periodic registration, emergency registration), UE subscriber and network identifiers (SUCI/SUPI/5G-GUTI), security parameters, requested NSSAI (see details in Sect. 4.3), UE capability and PDU session information (status, sessions to be re-activated, follow-on request and MICO mode preference). The subsequent steps here are for an initial registration. The gNB uses SUPI (Subscription Permanent Identifier) and NSSAI to select an AMF and forward the Registration Request to the selected AMF (2).

The AMF selects an AUSF (Authentication Server Function) and initiates authentication based on SUPI or Subscription Concealed Identifier (SUCI). If it is an emergency registration, authentication is skipped. The AMF initiates NAS security functions and upon completion, the AMF initiates the NGAP procedure (for a logical, per UE association between a 5G access node and AMF). Following the security procedures, the UE is authenticated and gNB/access node stores the security context and uses it to protect messages exchanged with the UE (details are described in TS 33.501).

The AMF selects a UDM (Unified Data Management) and in turn a UDR (Unified Data Repository) instance based on the SUPI and retrieve access and mobility subscription data, SMF selection subscription data. The AMF then creates a MM Context based on the subscription data obtained. The AMF then selects a PCF (Policy Control Function) and requests non-session policy (access and mobility related, further described in the Sect. 4.10).

The AMF sends a Registration Accept to the UE with 5G-GUTI, Registration Area, Mobility restrictions, allowed NSSAI, Periodic Registration Update Timer, Local Area Data Network (LADN), MICO mode information and other session information. If the 5G-GUTI is new, the UE replies with a Registration Complete. The state in the UE and network for this UE is 5GMM-REGISTERED. Before the Periodic Registration Update Timer Expires, the UE can send a Registration Request (type: periodic update) to remain in 5GMM-REGISTERED state (the timer is also reset when the UE/network enter the 5GMM-CONNECTED state).

The UE can request the use of MICO (Mobile Initiated Connection Only) mode during regular (non-emergency) registration in a 3GPP access (gNB). In MICO mode, all NAS timers are stopped (except for periodic registration update timer and a few others) and the UE cannot be paged.

#### ***4.4.4 PDU Session Establishment Call Flow***

The flow description in Fig. 4.11 below provides an overview of how a PDU session can be established following registration of the UE.

The UE initiates session establishment by sending to the AMF a PDU Session Establishment Request (1) with S-NSSAI request type (initial request, existing

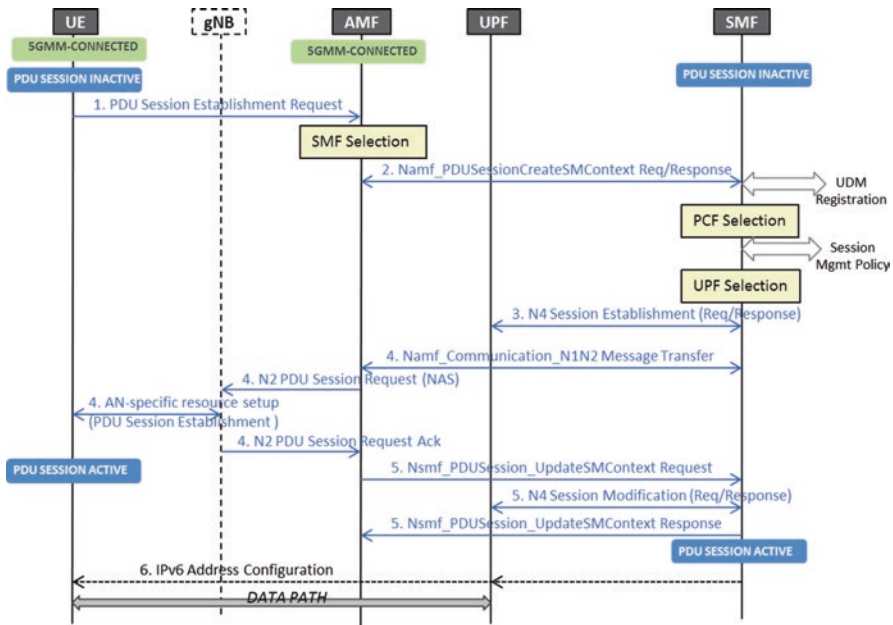


Fig. 4.11 PDU Session Establishment

PDU, emergency request), old PDU session id if one exists, and the N1 SM container (PDU Session Establishment Request).

When the AMF receives this request in (1), it determines the request type—in this case, an initial request—and selects an SMF based on S-NSSAI, subscription permissions on Local Breakout (LBO) roaming, selected Data Network Name (DNN) and per DNN local breakout /S-NSSAI permissions and local operator policies, load and access technology used by the UE. Since this is an initial request, the AMF initiates a create session context messaging request and response sequence (2) with the newly selected SMF. The AMF forwards the SM Container (PDU Session Establishment Request) and parameters along with the GUAMI (Globally Unique AMF Identifier). For an initial request, subscription parameters from UDM include authorized PDU type, authorized SSC modes (described further in Sect. 4.5), default 5QI (5G QoS Identifier), ARP (Allocation and Retention Priority), and subscribed session AMBR (Aggregate Maximum Bit Rate). If the UE request is compliant with subscription policies, the SMF responds back to the AMF (2, response).

The SMF then selects a PCF for dynamic policy, or may apply local policy if dynamic PCC (Policy Control and Charging) is not deployed. Session management related policy control includes gating, charging control, QoS, and per application policy (described further in Sect. 4.10).

Following dynamic policy installation, the SMF selects one or more UPFs (User Plane Functions). UPF selection includes deployment considerations (example central anchor and intermediate UPF close to access node), roaming consideration,

configuration information from OA&M (such as capacity, location, supported capabilities), and dynamic conditions (such as UPF load). The SMF initiates the N4 Session Establishment request/response sequence (3) with the UPF (or UPFs) with packet detection, enforcement and reporting rules, CN tunnel information and user plane inactivity timer for deactivation. The SMF also allocates IPv4 /IPv6 addresses and installs forwarding rules. For an unstructured PDU session, neither MAC nor IP address is assigned by SMF.

Following UPF session establishment, the SMF configures session information at the gNB and UE (sequence of messages 4 in Figure SM). The SMF sends N2 SM Information (PDU Session ID, QFI, QoS Profile, CN Tunnel info, S-NSSAI, AMBR, PDU session type) and N1 SM Container (PDU Session Establishment Accept with QoS, slice and session parameters) in an N1N2Transfer Message to the AMF. The AMF sends the NAS message PDU Session Establishment Accept targeted to the UE, and N2 SM message to the gNB. The responses from the UE and gNB to the AMF contains the list of rejected QFIs and established AN Tunnel information.

The AMF sends Nsmf\_PDUSession\_UpdateSMContext Request (5) to the SMF with rejected QFI and AN tunnel information. The SMF performs an N4 session modification sequence (5) to update the AN tunnel and QoS information. At this point, the PDU session setup is complete.

IPv6 address Router advertisement (6) is initiated by the SMF and forwarded by the UPF for dynamic IP addresses (static IP addresses may be sent in earlier signaling). The interface identifier sent earlier is used to derive the complete IPv6 address. IP data traffic can be sent at this time.

#### **4.4.5 Service Request**

The Service Request procedure allows a UE to transition from a 5GMM-IDLE to 5GMM-CONNECTED state. For example, when the UE is in 5GMM-IDLE state (and not in MICO mode), the network may page the UE to indicate that it has downstream data (which is temporarily buffered at the UPF). Once the service request procedure is executed, the UE and network transition to 5GMM-CONNECTED state for that UE and control and data plane paths are established. In the case above, downstream data that has been buffered at the UPF can then be delivered to the UE.

Service requests can be UE triggered or network triggered. In a UE-triggered service request, the UE in 5GMM-IDLE state requests the establishment of a secure connection to the AMF. The UE triggering service request maybe as a result of receiving a paging request from the network, or when the UE wants to send uplink signaling messages or data. After receiving the service request, the network initiates procedures to set up the control plane and the user plane. The service request procedure may support independent activation of the UP connections for existing PDU sessions.

Network-triggered service request procedure is used when the network needs to signal (N1 signaling) or deliver mobile terminating user data to a UE (e.g., SMS). The network-triggered service request may be invoked in 5GMM-IDLE or 5GMM-CONNECTED state.

#### **4.4.6 Other Procedures**

There are a number of procedures that 5GS provides to support various session management capabilities. For session management, in addition to PDU Session Establishment (described above), they include PDU Session Modification and Session Release.

UE connection, registration, and mobility procedures include all the registration, service request procedures, UE configuration, AN release, and N2 signaling procedures. SMF and UPF procedures are used to set up and manage PDU session state in UPF (setup, modify, delete, reporting, charging). User profile management procedures are used to notify subscriber data updates, session management subscription notifications and purge of subscriber data in AMF.

Details of these and other procedures can be found in 3GPP TS 23.502, section 4 (System Procedures) [2].

### **4.5 Session and Service Continuity in 5GC**

5G support for a range of services from IoT to critical communications introduces various requirements on packet data and user plane including degrees of mobility and session continuity for connections with varying levels of latency, bandwidth, and reliability. Thus, the connectivity service (PDU sessions) modes support classical session continuity with a central anchor as in 4G systems, or newer forms where a PDU session can be retained until after establishing another PDU session to the same data network (DN), or where a PDU session is released prior to a new PDU session establishment to that DN. Handling of IP addresses for these PDU sessions, subnets and the layout of IP networks and gateways are important to consider when discussing how session and service continuity is handled. An outline of how the 5G control plane and user plane entities behave to support the modes of session and service continuity is discussed below.

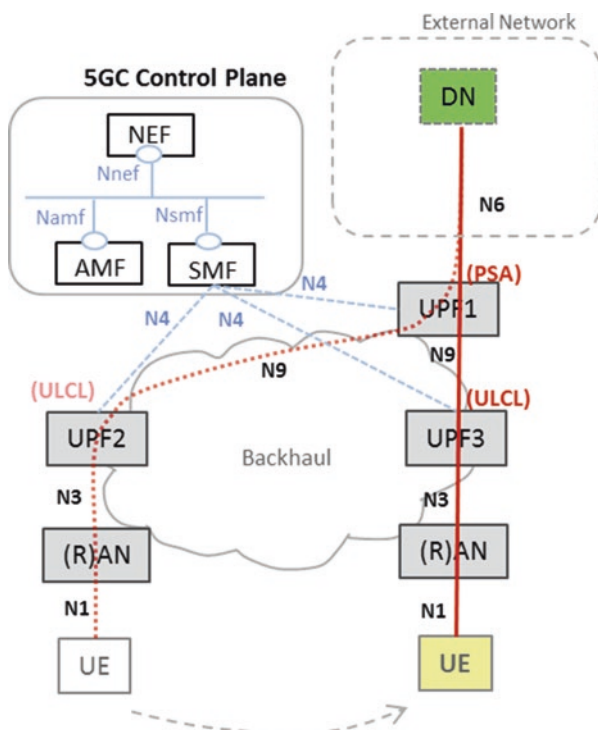
In 5GC, PDU sessions and the IP addresses that identify it are anchored at a UPF PSA (PDU Session Anchor). There are three session continuity modes supported by the 5G system to satisfy the continuity requirements of different applications: SSC Mode 1, Mode 2, and Mode 3. PDU sessions established with SSC Mode 1 maintain the same session anchor for the length of the session. Thus, with SSC mode 1 the network selects a centrally located UPF PSA so that it is able to serve the PDU ses-

sion even when the UE moves across radio access locations. This is shown in Fig. 4.12.

Figure 4.12 also shows an outline of a handover in SSC Mode 1. The UE initially has a PDU session to DN (external data network) with UPF2 as ULCL (Uplink Classifier) and UPF 1 as the PSA. As the UE moves to new radio networks and based on location information from the AMF, the SMF decides to relocate the ULCL from UPF2 to UPF3. The PDU session continues to be anchored at UPF1 since it is SSC Mode 1.

With SSC Mode 2 and Mode 3, there is the option to relocate the PDU session anchor as the PDU session can be released and a new one established to continue the connection. In the case of SSC Mode 2, the UPF anchor (PSA) is released and a new anchor is selected and programmed by the SMF. In the case of SSC Mode 3, the UPF anchor (PSA) is released only after a new anchor is programmed. Applications that need continuous availability of a connection path may select SSC Mode 3, while applications that can tolerate a break in connectivity before a new path can select SSC Mode 2.

In addition to the PDU session management, the 5GC also needs to manage IP addresses. When the SMF selects a new anchor UPF for the PDU session, the IP address assigned to that PDU session also needed to be re-assigned to a topologically



**Fig. 4.12** Handover with Central PDU Anchor







to another radio access, the SMF reprograms the forwarding information at UPF<sub>c</sub> (add forwarding info) and UPF<sub>b</sub> (remove forwarding info). The SMF also provides events to AF (via NEF (Network Exposure Function), if AF cannot directly subscribe to UE information from SMF). The AF can use events received on UE session relocation to provision or migrate the application session from AS1 to AS2.

For session continuity in 5GC as the UE moves, the SMF reprograms UPF forwarding state of the PDU sessions as described in this section. In addition to the change in forwarding path, transport of packets with minimal loss and re-ordering is needed. The UPFs buffer packets and use end-markers to avoid re-ordering during the transition from one UPF to another. End-to-end congestion and flow control mechanisms in TCP and QUIC are also managed to prevent congestion collapse during the handover—especially with short round trip times and low latency flows.

## 4.6 Interworking with EPC

The network deployments may comprise 5GC and EPC as well as coexistence of UEs supporting 5G and 4G within the same network (i.e., within one PLMN). The UEs that supports 5G may also be supporting EPC NAS (Non-Access Stratum) procedures to provide service in legacy 4G networks when roaming to those networks. 3GPP standards have defined architecture options to support interworking and migration from EPC to 5GC. Figure 4.14 shows an example of a non-roaming architecture for interworking between EPC and 5GC. For migration scenarios, it is generally assumed that the subscriber database is common between 5GC and EPC. That is, the UDM in 5GC and the HSS in EPC is a common database. Optionally it could be further assumed that PCRF (Policy and Charging Rules Function) in EPC and PCF in 5G, PGW Control plane function (PGW-C) in 4G and SMF in 5G, PGW User plane function (PGW-U) in 4G and UPF in 5G are collocated, respectively, and dedicated for interworking between EPC and 5GC. These are referred to as HSS + UDM, PCF + PCRF, SMF + PGW-C, and UPF + PGW-U. For User Plane management and connectivity when interworking with EPC, the SMF + PGW-C provides information over N4 to the UPF + PGW-U related to the handling of traffic over S5-U.

An optional interface, N26, is defined to provide an inter Core Networks interface between EPC and 5GC by interconnecting MME and AMF. Since N26 is optional, the networks may provide interworking without N26 interface as well. Interworking procedures with N26 provide IP address continuity on inter-system mobility to UEs that support single registration mode as well as both 5GC NAS and EPC NAS (When the N26 interface is used for interworking, the UE operates in single-registration mode). Networks that support interworking procedures without N26 provide IP address continuity on inter-system mobility to UEs operating in both single-registration mode and dual-registration mode. The N26 interface enables the exchange of mobility and session management states (MM and SM) between the source and target network. In these interworking scenarios the MM state for the UE



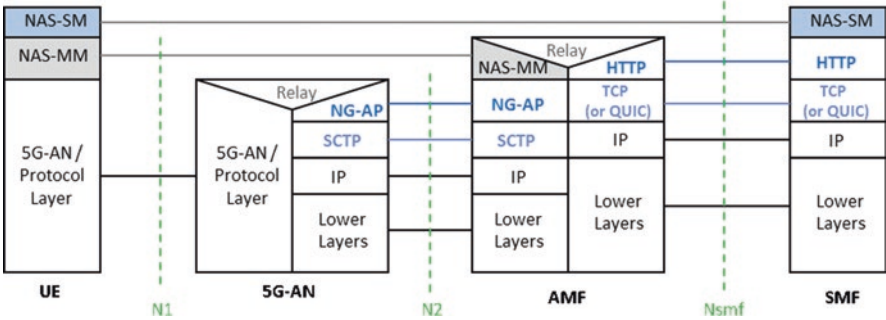


Fig. 4.15 Control Plane Protocol Stack

The set of protocols on N1 interface across UE–5GAN depends on the access network. In the case of NG-RAN, the radio protocol between the UE and NG-RAN (eNB, gNB) is specified in TS 36.300 and TS 38.300 [6]. For non-3GPP accesses, EAP-5G/IKEv2 and IP are used over the non-3GPP radio (WLAN) for establishing the IPsec SA (Security Association) and the NAS is sent over the established IPsec connection.

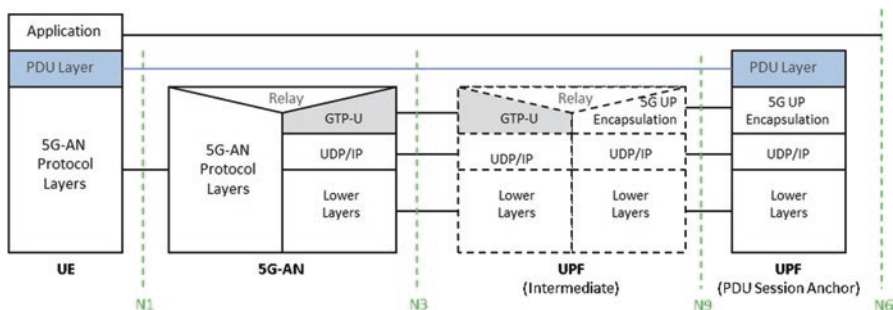
The N2 interface between 5G-AN and AMF has an SCTP/IP transport connection over which the NG-AP protocol runs. The control plane interface between 5G-AN and 5GC supports connection of different kinds of AN (3GPP RAN, N3IWF) with the same protocol. The AN–SMF protocols where N2 SM (Session Management) messages are relayed via AMF uses NG-AP between 5G-AN and AMF. In addition, the NAS protocol from the UE side is also relayed over NG-AP.

The NAS MM layer between UE and AMF is used for Registration Management (RM) and Connection Management (CM) as well as for relaying Session Management (SM) messages. The NAS-SM layer carries session management messages between UE–SMF. 5G NAS protocols are defined in TS 24.501 [7].

The Nsmf interface and protocol between AMF and SMF is based on Service-Based Architecture (SBA) using HTTPS protocol over a transport layer of TCP. It is expected that as QUIC transport is standardized and stable in IETF, 3GPP will adopt it as the SBA signaling transport layer. Functions in the core network interact with each other over in the SBA architecture using 1:N (many) bus over HTTPS. Service discovery in SBA is via the NRF (Nnsf interface) that resolves the destination service function based on a number of request criteria. Control plane signaling between SMF–UPF however uses the N4 interface and extensions to PFCP (Packet Flow Control Protocol) defined in TS 29.244 [8].

4.7.2 User Plane Protocol Stack

The User Plane (UP) protocol stack consists of the protocol stack for a PDU session and the user plane for an untrusted, non-3GPP access (Fig. 4.16).



**Fig. 4.16** User Plane Protocol Stack

User plane protocol stack is primarily to transport PDU sessions across from the UE to the UPF (PDU session anchor). The PDU layer can be IPv4, IPv6, or IPv4v6. For IPv4 and IPv6, the SMF is responsible for allocating and managing the IP address. When the PDU session type is unstructured, Ethernet is used. In the case of Ethernet, the SMF does not allocate MAC or IP addresses. Since the 3GPP access is an NBMA (Non-broadcast Multiple Access) type network, tunneling is used to carry packets to the PDU session anchor. GTP-U (GPRS Tunneling Protocol—User Plane) multiplexes user data over N3 between 5G-AN—intermediate UPF (e.g., UPF that performs uplink classification) and over N9 towards a PDU session anchor UPF. Unlike 4G where the GTP-U tunnels correspond to a bearer (and a UE may have multiple bearers), all UE packets of a UE across an N3 or N9 interface are transported over a single GTP-U connection. QFI/ flow marking associated with a QoS flow is signaled explicitly in this connection to indicate the level of QoS in the IP transport layer.

## 4.8 Support for Virtualized Deployments

The 5GS embraces Network Function virtualization (NFV) and cloudification for its architecture design. The 5GS supports different virtualization deployment scenarios, such as

- A Network Function instance can be deployed as fully distributed, fully redundant, stateless, and fully scalable NF instance that provides the services from several locations and several execution instances in each location.
- A Network Function instance can also be deployed such that several network function instances that are present within a NF set are fully distributed, fully redundant, stateless and scalable as a set of NF instances.

The network slicing feature supported by 5GS is also enabled by virtualization. Network function instances can be created, instantiated and isolated with each other in the virtualized environment, into different network slices in order to serve different services.

In order to manage the life cycle of the virtualized 5GS functions and its instances as well as the virtual resource of network slice, 5G OAM provides means to integrate with virtualized network function management and orchestration capability, as well as providing standardized life cycle management interfaces with other virtualized function management and orchestration system which are defined by other standards, such as ETSI ISG NFV, and other open source project, such as ONAP.

Figure 4.17 illustrates the integration of 5G OAM system with ETSI NFV Management and orchestration (MANO) system [9].

In this illustration, 5G management system provides the 5G service and functionality management, such as NM plays one of the roles of OSS/BSS and provides the functions for the management of mobile network which includes virtualized network functions; and EM/DM is responsible for FCAPS management functionality for a VNF on an application level and physical NE on a domain and element level. ETSI ISG NFV MANO provides the virtualized resource and life cycle management for those virtualized 5G functions and the network service and network slice composed by those functions.

## 4.9 Support for Edge Computing

The 5GS architecture was designed taking the Edge Computing requirements in mind from the start. As such, edge computing is considered a key technology for efficient routing to the application servers as well as addressing the low latency requirements.

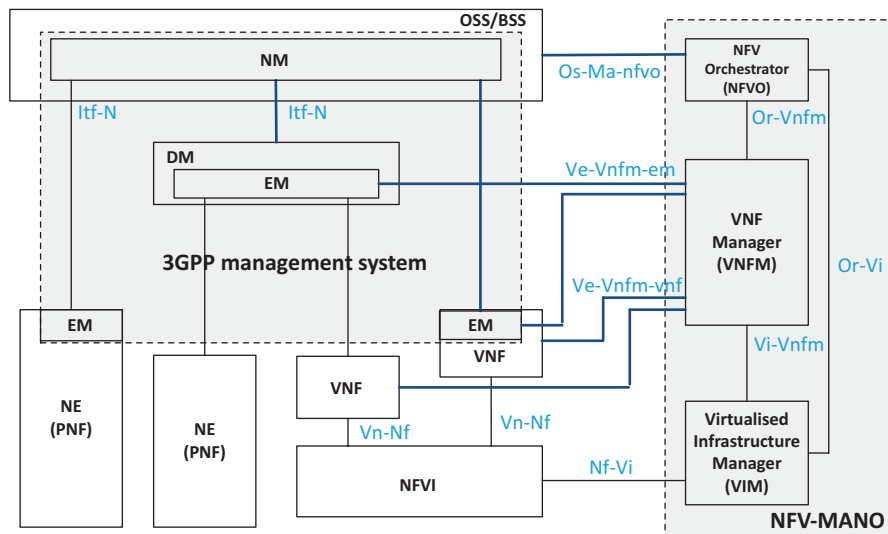


Fig. 4.17 Integration of 5G OAM system with ETSI NFV MANO system

In the context of 3GPP, the Edge Computing refers to the scenarios that the services need to be hosted close to the access network of the UE (e.g., at or close to the RAN). As described earlier, in 5GS the routing of the data (or user) traffic is done via UPF interface to a Data Network. The 5G core network supports the capability to select a UPF that allows routing of traffic to a local Data Network that is close to the UE's access network. This includes the local breakout scenarios for roaming UEs and non-roaming scenarios.

The decision for selection (or reselection) of UPF for local routing may be based on the information from an Edge Computing Application Function (AF) and/or to other criteria such as the subscription, location, and policies. Depending on the operator's policy and arrangements with the third parties, an AF may access the 5G core directly or indirectly via the Network Exposure Function (NEF). For example, an external AF at the edge data center could influence the routing of the traffic by altering the SMF routing decisions via its interaction with the Policy Control Function (PCF).

The 3GPP TS 23.501 clause 5.13 [1] defines several enablers that can support Edge Computing:

- User plane selection and re-selection for UPF to route the user traffic to the local Data Network
- Local Routing and Traffic Steering to the applications in the local Data Network
- Session and service continuity to enable UE and application mobility
- Application Function influencing UPF (re)selection and traffic routing via PCF or NEF
- Network capability exposure between 5G Core Network and Application Function to provide information to each other via NEF
- QoS and Charging procedures in PCF for the traffic routed to the local Data Network
- Support of Local Area Data Network (LADN)

## 4.10 Policy and Charging Control in 5G System

The Policy and Charging Control Function (PCF) is responsible for flow-based offline and online charging control, policy for authorization, mobility, QoS, session management and UE access and PDU session selection. 5G support for URLLC and millions of IoT connections in addition to mobile broadband requires extensive policy and charging capabilities. The policy and charging control system in 5GS supports the ability to manage session QoS and charging, access management (non-session), network and subscription policy in real time. The system also supports service capabilities exposure for applications in edge-networks and policies based on network analytics and load information. There is support for online and offline charging based on the requirements of the operator and application. The above policy control can be broadly classified as session management policy and non-session

management policy. This is described in further detail followed by the functions and architecture for non-roaming and roaming with local breakout or home routed sessions. Non-session management related policy control includes access and mobility management policy, access and PDU selection policy. Access and mobility policies are installed by the PCF when the UE initially registers with AMF. AMF provides the PCF with the SUPI (Subscription Permanent Identifier), user location information, time zone, serving radio access of the UE, and parameters received from the UDM including service area restrictions, RSFP (RAT Frequency Selection Priority) index and GPSI (Generic Public Subscriber Identifier). The PCF makes policy decisions and provides UE access network discovery and selection policy, URSP (UE Route Selection Policy) and revised access and mobility policy (RFSP and service area restrictions). These policies can be modified as a result of changes and notified by the PCF, or re-evaluated if triggered by the AMF.

Session management related policy control includes gating control (to discard packets that do not match any policy), charging control and QoS for PDU session as well as SDF (Service Data Flow) and per application policy (service data flow and application policy is only set up as needed). QoS control policies may be service based, subscription based, or predefined policies that can be applied at the PDU session level or per SDF. SDF filters include flow, precedence, provider identifier, charging key, charging method (online, offline), and measurement method (volume, time, event, or combinations of these). Gating policies allow gate open/closed for SDF, 5QI, reflective QoS, GBR (Guaranteed Bit Rate), and AMBR (Aggregate Maximum Bit Rate) parameters.

The session-based policies also include usage monitoring control, application detection, service capability exposure, and traffic steering. Usage monitoring policies may be applied to a PDU session, service data flow or have rules for PDU session monitoring that excludes specific data flows.

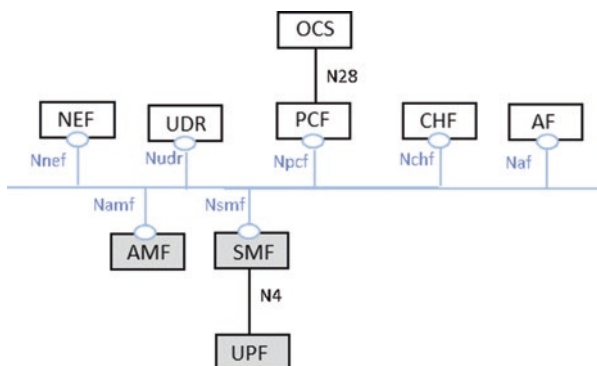
The PCF may subscribe to Network Data Analytic Function (NWDAF), or initiate a request-response sequence for load level information of a network slice instance to make the policy decisions.

Figure 4.18 shows a service-based representation of the policy and charging architecture for the non-roaming scenario. The PCF and AMF interact over the Npcf/Namf interfaces to create an AM Policy association and provision non-session policy for access and mobility management. The PCF and SMF interact over the Npcf/Nsmf interfaces to create an SM Policy association and provision policy for sessions and charging control. The interactions between SMF and CHF enable offline charging, while online charging is handled at the OCS (N28 interface).

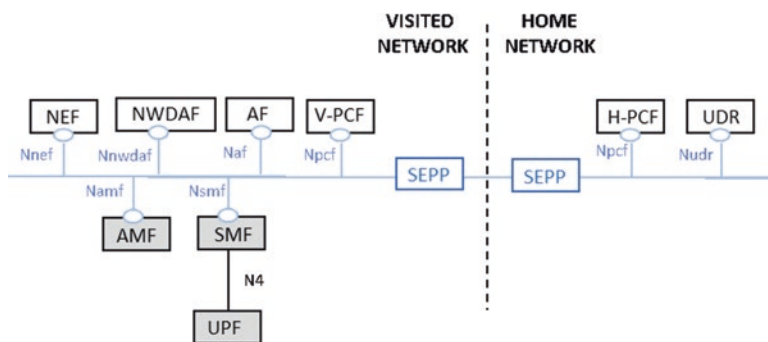
The PCF may access policy control related subscription information at the UDR (Npcf/Nudr interfaces). The UDR may notify the PCF on subscription changes. The PCF may also subscribe to the UDR for AF requests targeting a DNN and S-NSSAI or a group of UEs identified by an internal group identifier.

The PCF and AF interact to transport application level session information to the AF including IP filter information (or Ethernet packet filter information) for policy control or differentiated charging, media bandwidth for QoS control and for sponsored data connectivity. The Npcf and Naf enable the AF to subscribe to notifica-





**Fig. 4.18** Non-Roaming Policy and Charging Control Architecture



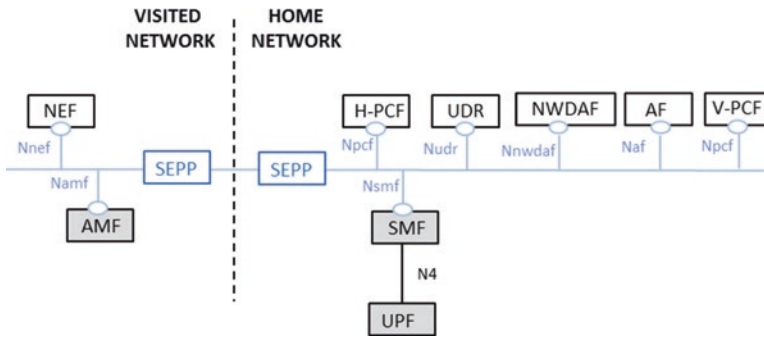
**Fig. 4.19** LBO Roaming Policy and Charging Control Architecture

tions of PDU session events. The PCF and AF may also interact over Rx (Diameter) interface for IMS based and Mission Critical Push to Talk (MCPTT) services. For deployments where the AF belongs to a third party, AF requests to the PCF and SMF may not be allowed to interact directly. In this case, these requests and notifications are handled by the NEF.

Figure 4.19 shows a home routed LBO (Local Breakout) roaming architecture. In the LBO roaming architecture, the H-PCF and UDR are in the home PLMN while other functions are in the visited PLMN. In this case the V-PCF (PCF in the visited PLMN) uses locally configured policies according to the roaming agreement. The V-PCF obtains UE access selection and PDU session selection information from the H-PCF using either the Npcf or N24 interface. There is no offline charging support in this scenario. It may be noted that an SEPP (Security Edge Protection Proxy) may be used for filtering and policing messages on the inter-PLMN control plane interface.

In a home routed roaming architecture, the access management functions are in the visited PLMN while session and policy management are in the home PLMN (Fig. 4.20).





**Fig. 4.20** Home Routed Roaming Policy and Charging Control Architecture

## 4.11 Summary

The 3GPP standards that define the 5G System (5GS) architecture provide enhanced connection, session, and mobility management services with major enhancements from 4G to support network slicing, virtualization, and edge computing. These capabilities are designed to provide support for a range of services requiring low latency, high reliability, high bandwidth, or a massive amount of connectivity over the same network.

As described in this section, the 3GPP standards define the 5GS architecture which consist of network functions and interfaces between them. In the 5GS service-based architecture the service-based interfaces (SBI) are defined within the control plane to allow the network functions to access each other's services using a common framework. The Service Framework defines the interaction between the NFs over SBI using a Producer-Consumer model. The 5GS embraces Network Function virtualization and cloudification for its architecture design. The 5GS supports different virtualization deployment scenarios where for example a NF instance can be deployed as fully distributed, fully redundant, stateless, and fully scalable NF instance.

A major feature of 5GS is network slicing. In 3GPP view network slice is a logical network with specific functions/elements dedicated for a particular use case, service type, traffic type, or other business arrangements. The most commonly discussed slice types in industry are enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communications (URLLC), and massive IoT (mIoT). In this chapter we provided an overview of network slicing as defined in 3GPP 5G core standards.

The 5GS architecture has been designed taking the Edge Computing requirements in mind. In the context of 3GPP, the Edge Computing refers to the scenarios that the services need to be hosted close to the access network of the UE. The 5G core network supports the capability to select a User Plane Function (UPF) that allows routing of traffic to a local Data Network that is close to the UE's access network.

An overview of 5G architecture, network functions, and new capabilities (e.g., Network Slicing), as well as high-level features for registration, connection management, and session management are provided in this chapter. In comparison to 4G/EPS, the 5G System defines more capabilities for degrees of mobility and session continuity for connections with varying levels of latency, bandwidth and reliability. Thus, in 5GS the connectivity service (PDU sessions) modes support classical session continuity with a central anchor as in 4G systems, or newer forms where a PDU session can be retained until after establishing another PDU session to the same data network (DN), or where a PDU session is released prior to a new PDU session establishment to that DN.

Since 5G supports services requiring low latency, high reliability, high bandwidth, or massive connectivity (i.e., URLLC, eMBB and mMTC) over the same network, performance capabilities have to be carefully engineered and evaluated. The next section provides a comprehensive view on the performance evaluation methodologies, metrics and system level simulations.

## References

1. 3GPP TS 23.501, Technical Specification Group Services and System Aspects; System Architecture for the 5G System; Stage 2
2. 3GPP TS 23.502, Technical Specification Group Services and System Aspects; Procedures for the 5G System; Stage 2
3. 3GPP TS 23.503, Technical Specification Group Services and System Aspects; Policy and Charging Control Framework for the 5G System; Stage 2
4. 3GPP TS 23.214, Technical Specification Group Services and System Aspects; Architecture enhancements for control and user plane separation of EPC nodes; Stage 2
5. 3GPP TS 29.510, Technical Specification Group Core Network and Terminals; 5G System; Network Function Repository Services; Stage 3
6. 3GPP TS 38.300, Technical Specification Group Radio Access Network; NR; NR and NG-RAN Overall Description; Stage 2
7. 3GPP TS 24.501, Technical Specification Group Core Network and Terminals; Non-Access-Stratum (NAS) protocol for 5G System (5GS); Stage 3
8. 3GPP TS 29.244, Technical Specification Group Core Network and Terminals; Interface between the Control Plane and the User Plan Nodes; Stage 3
9. 3GPP TS 28.500, Technical Specification Group Services and System Aspects; Telecommunication management; Management concept, architecture and requirements for mobile networks that include virtualized network functions