URLCC

ITU-R

eMBB

3GPP

5G WIRELESS

ITU-T

mMTC

5G Wireless

A Comprehensive Introduction

Dr. William Stallings

Addison-Wesley

Boston • Columbus • New York • San Francisco • Amsterdam • Cape Town
Dubai • London • Madrid • Milan • Munich • Paris • Montreal • Toronto • Delhi • Mexico City
São Paulo • Sydney • Hong Kong • Seoul • Singapore • Taipei • Tokyo

Library of Congress Control Number: 2021937463

ISBN-13: 978-0-13-676714-5
ISBN-10: 0-13-676714-1

ScoutAutomatedPrintCode

Editor-in-Chief
Mark Taub

Director Product Management
Brett Bartow

Development Editor
Marianne Bartow

Managing Editor
Sandra Schroeder

Technical Reviewers
Toon Norp
Tim Stammers

Senior Project Editor
Lori Lyons

Copy Editor
Catherine D. Wilson

Production Manager
Aswini Kumar/codeMantra

Indexer
Cheryl Ann Lenser

Proofreader
Donna E. Mulder

Editorial Assistant
Cindy Teeters

Cover Designer
Chuti Prasertsith

Compositor
codeMantra

About the Author

Dr. William Stallings has made a unique contribution to understanding the broad sweep of technical developments in computer security, computer networking, and computer architecture. He has authored 20 textbooks, and, counting revised editions, more than 75 books on various aspects of these subjects. His writings have appeared in numerous ACM and IEEE publications, including the *Proceedings of the IEEE* and *ACM Computing Reviews*. He has 13 times received the award for the best computer science textbook of the year from the Text and Academic Authors Association.

In over 30 years in the field, he has been a technical contributor, a technical manager, and an executive with several high-technology firms. He has designed and implemented both TCP/IP-based and OSI-based protocol suites on a variety of computers and operating systems, ranging from microcomputers to mainframes. Currently he is an independent consultant whose clients have included computer and networking manufacturers and customers, software development firms, and leading-edge government research institutions.

He created and maintains the Computer Science Student Resource Site at ComputerScienceStudent.com. This site provides documents and links on a variety of subjects of general interest to computer science students (and professionals). He is a member of the editorial board of *Cryptologia*, a scholarly journal devoted to all aspects of cryptology.

Dr. Stallings holds a PhD from M.I.T. in computer science and a B.S. from Notre Dame in electrical engineering.

About the Technical Reviewers

Toon Norp is a Senior Business Consultant at TNO. He joined TNO (former KPN Research) in 1991, where he has since been working on network aspects of mobile communications. Toon advises European operators, government organizations, and others on strategy and architecture related to mobile networks, M2M/IoT, and 5G. He has been involved in standardization of mobile networks for more than 20 years, and as chairman of the 3GPP SA1 service aspects working group, he was responsible for the requirements specification phase of 5G. Toon has been instrumental in getting several new sectors (e.g., public safety, satellite, media, railway, industry) involved in the 3GPP standardization process. Toon is a member of the 5G-PPP association, a joint initiative between the European ICT industry and the European Commission, a reviewer of European R&D projects, and a regular speaker at conferences. Toon holds a master's degree in electrical engineering from the Eindhoven University of Technology, The Netherlands.

Tim Stammers is a Principal Engineer at Cisco Systems. Tim joined Cisco in 2000, where he has since been working on the architecture and development of products for mobile data services. Prior to that, Tim worked at Alcatel in the area of mobile switching as well as at a number of telecommunications startups.

Tim serves as a technical advisor for cellular core technologies across a number of products and services. Tim was directly involved with Tier 1 operators in the commercial launch of 4G data services and is now focused on 5G, IoT, and multi-access technologies; he has an interest in service adjacencies for private cellular opportunities.

Tim is the author of over 50 U.S. patents in the areas of mobile networking and security.

Tim represents Cisco in 5G-ACIA, an industry group promoting 5G in the area of industrial automation. Tim has provided and reviewed material for 3GPP and participated in the core standards development for ANSI-136.

Tim holds a bachelor of science degree in electrical and electronic engineering from the University of Bristol, United Kingdom.

Chapter 9

Core Network Functionality, QoS, and Network Slicing

Learning Objectives

After studying this chapter, you should be able to:

- List and explain the 5G core network requirements defined by 3GPP
- Explain the relationship among priority, QoS, and policy control
- Present an overview of the concept of tunneling
- Present an overview of the PDU session establishment procedure
- Define 5QI and explain how it is used
- Explain the difference between QoS parameters and QoS characteristics
- Summarize the requirements for network slicing
- Give a functional description of network slice implementation

Chapters 7 and 8 covered the two essential enablers of 5G services provided by core networks: software defined networking (SDN) and network functions virtualization (NFV). With this foundation, this chapter presents an overview of the 5G core network functions and services.

Section 9.1 discusses the requirements for core networks, looking at requirements outlined by ITU-T and by 3GPP. Then, Section 9.2 examines the functional architecture of the core network. First, the section explains the concept of tunneling in 5G networks. Then it discusses the key operational aspect of core networks, which is PDU session establishment. Finally, Section 9.2 examines the central role of the policy control function.

Section 9.3 provides a detailed analysis of quality of service (QoS) and its role in 5G. Section 9.4 discusses network slicing, which is a critical capability for 5G.

9.1 Core Network Requirements

This section examines two sets of core network functional requirements: those defined by ITU-T and those defined by 3GPP.

Network Operational Requirements

As mentioned in Chapter 2, "5G Standards and Specifications," ITU-T Y.3101 (*Requirements of the IMT-2020 Network*, April 2018) defines network operational requirements. The requirements from that document are as follows:

- **Network flexibility and programmability:** The network should support a wide range of devices, users, and applications, with evolving requirements for each. Significant concepts in this regard are network functions virtualization (discussed in Chapter 8, "Network Functions Virtualization"), separation of user and control planes, and network slicing. The latter two concepts are discussed later in this chapter.
- **Fixed mobile convergence:** The focus of this requirement is to enable subscriber access through multi-access networks in seamless, integrated fashion.
- **Enhanced mobility management:** The network should support a wide variety of mobility options.
- **Network capability exposure:** The IMT-2020 network should provide suitable ways (e.g., via application program interfaces [APIs]) to expose network capabilities and relevant information (e.g., information for connectivity, QoS, and mobility) to third parties. This enables third parties to dynamically customize the network capabilities for diverse use cases within the limits set by the IMT-2020 network operator.
- **Identification and authentication:** There should be a unified approach to user and device identification and authentication mechanisms.
- **Security and personal data protection:** The IMT-2020 network must provide effective mechanisms to preserve security and personal data protection for different types of devices, users, and services, including rapid adaptation to dynamic network changes.
- **Efficient signaling:** There are two aspects to this requirement. The signaling mechanisms should be designed to mitigate risks of control and data traffic bottlenecks. Also, the network should provide lightweight signaling protocols and mechanisms to accommodate limited-resource devices.
- **Quality of service control:** The network should support different QoS levels for different services and applications.
- **Network management:** The network should provide a unified network management framework to support interworking of different providers and management of legacy networks.
- **Charging:** The IMT-2020 network needs to support different charging policies and requirements of network operators and service providers, including third parties that may be

involved in a given IMT-2020 network deployment. The charging models to be supported include, but are not limited to, charging based on volume, time, session, and application.

- **Interworking with non-IMT-2020 networks:** IMT-2020 networks should support user-transparent interworking with legacy networks.
- **IMT-2020 network deployment and migration:** The network design should accommodate incremental deployment with migration capabilities for services and related users.

For each of the 12 general requirements listed above, Y.3101 includes a number of specific, more detailed requirements. Figure 9.1, which repeats Figure 2.10, lists these requirements.

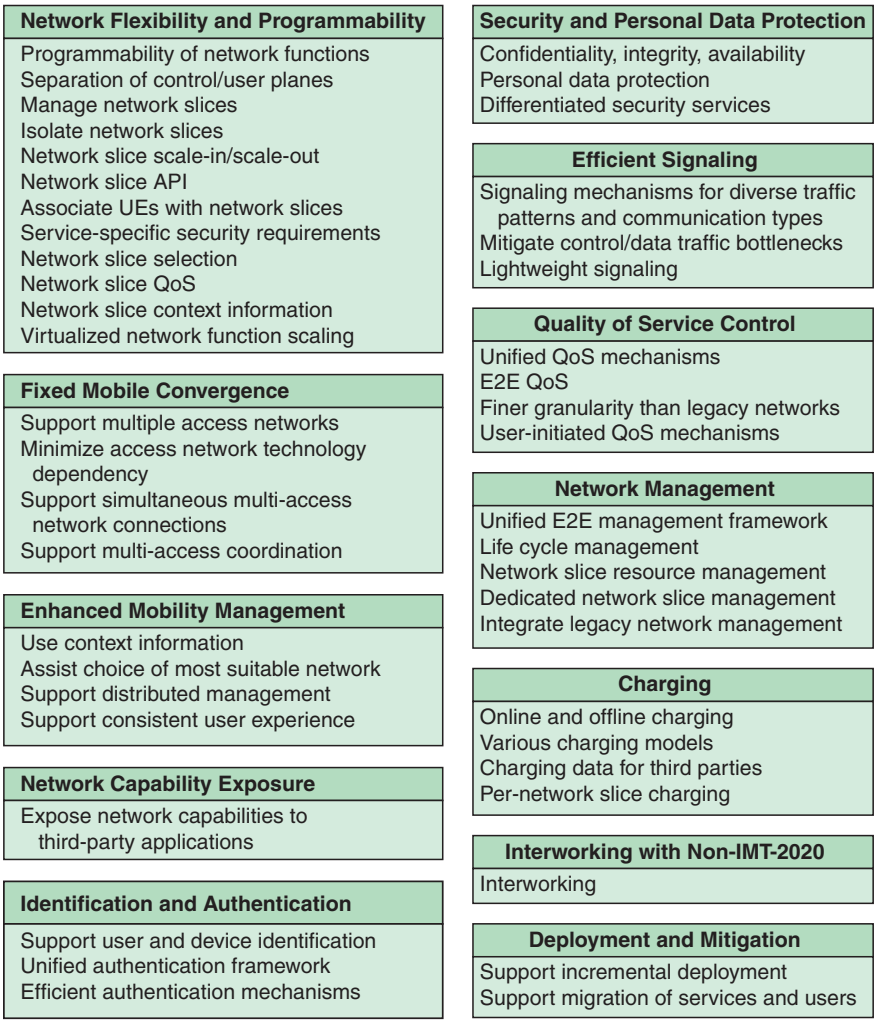


FIGURE 9.1 Requirements from the Network Operation Point of View

Basic Network Requirements

3GPP Technical Specification TS 22.261 (*Technical Specification Group Services and System Aspects, Service requirements for the 5G system, Stage 1 (Release 17)*, December 2020) defines requirements for 34 basic capabilities to be provided by a 5G network. Figure 9.2 lists these requirements (and repeats Figure 2.17). For each capability, TS 22.261 provides a description and elaborates on the requirements for that capability. The remainder of this section provides details on the key capabilities that are new to 5G.

Network slicing	Subscription aspects	Ethernet transport services
Diverse mobility management	Energy efficiency	Non-public networks
Multiple access technologies	Markets requiring minimal service levels	5G LAN-type service
Resource efficiency	Extreme long-range coverage in low-density areas	Positioning services
Efficient user plane	Multi-network connectivity and service delivery across operators	Cyber-physical control applications in vertical domains
Efficient content delivery	3 GPP access network selection	Messaging aspects
Priority, QoS, and policy control	eV2X aspects	Steering of roaming
Dynamic policy control	NG-RAN sharing	Minimization of service interruption
Connectivity models	Unified access control	UAV aspects
Network capability exposure	QoS monitoring	Video, imaging, and audio for professional applications
Context aware network		Critical medical applications
Self backhaul		
Flexible broadcast/multicast service		

eV2X = Enhanced Vehicle-to-Everything
UAV = Unmanned Aerial Vehicle

FIGURE 9.2 3GPP Basic Capability Requirements

Network Slicing

Network slicing enables operators to customize their network for different applications and customers. Slices can differ in services (e.g., priority, policy control, and security), in performance (e.g., latency, availability, reliability, and data rates), in the types and detail of assurance data, and in the type of failure diagnosis offered. Alternatively, a network slice can serve only specific users (e.g., public safety users, corporate customers, or industrial users). A network slice can provide the functionality of a complete network, including radio access network (RAN) and core network functions. Section 9.4 examines network slicing in detail.

Efficiency

TS 22.261 includes the following four capabilities related to efficiency:

- **Resource efficiency:** 5G networks need to be optimized for supporting diverse user equipment (UE) and services. Some of the underlying principles include bulk provisioning, resource efficient access, optimization for UE-originated data transfer, and efficiencies based on reduced needs related to mobility management for stationary UE and UE with restricted range of movement.
- **Efficient user plane:** Cloud-based applications can involve substantial computation that occurs far from the end user device, with substantial or time-sensitive data transfers. Such cases require

low end-to-end latencies and high data rates. 5G optimizes the user plane efficiency for such scenarios by locating applications in a service-hosting environment close to the end user. Video-based services (e.g., live streaming, virtual reality) and personal data storage applications have generated massive growth in mobile broadband traffic. In-network content caching—provided by the operator, third party, or both—can improve the user experience, reduce backhaul resource usage, and make more efficient use of radio resources for such applications.

- **Efficient content delivery:** Video-based services, such as live streaming and virtual reality, can place a considerable burden on the cellular network. To support such services, 5G networks emphasize caching content as much as possible near the end user, such as by using multi-access edge computing. In addition, 5G must support applications that involve a relatively small amount of data but have stringent latency requirements. Efficient delivery of such small packets requires the use of signaling protocols that do not require lengthy procedures and do not involve large amounts of control data.
- **Energy efficiency:** For mobile devices, energy efficiency translates directly into battery usage. Constrained IoT devices are especially of concern; such a device has a small battery, and this not only puts constrains on general power usage but also implies limitations on both the maximum peak power and continuous current drain. Thus, the 5G design must put minimal control signaling burden on such devices.

Diverse Mobility Management

Mobility management refers to a relationship between the mobile station and the RAN that is used to set up, maintain, and release the various physical channels. 5G supports different mobility management methods that minimize signaling overhead and optimizes access for user equipment with different mobility management needs. Devices may be:

- Stationary during their entire usable life (e.g., sensors embedded in infrastructure)
- Stationary during active periods but nomadic between activations (e.g., fixed access)
- Mobile within a constrained and well-defined space (e.g., in a factory)
- Fully mobile

Different applications have varying requirements for the network to mitigate the effects of mobility. Applications such as voice telephony rely on the network to ensure seamless mobility. Applications such as video streaming, on the other hand, have application layer functionality (e.g., buffering) to handle service delivery interruptions during mobility. These applications still require the network to minimize the interruption time.

Although mobility management is primarily a RAN responsibility, because of the much more distributed nature of 5G networks, mobility also has an impact in the core network. With IP traffic offload or service hosting close to the network edge, mobility of a device also implies that the anchor node in the network may need to be updated. Internet peering and service hosting have to follow the device when it is traveling across the network coverage area.

Priority, QoS, and Policy Control

Policy control is a generic term, and in a network, there are many different policies that could be implemented, such as policies related to security, mobility, and use of access technologies. When discussing policies, it is thus important to understand the context. In the context of the discussion of 5G capabilities and the policy control function, the following definitions are useful:

- **Policy:** A set of rules specifying the user plane services and functions available to a particular user, supplied by the network. In particular, a policy specifies the priority to be applied to a given user's traffic and the quality of service (QoS) to be provided to the user.
- **Policy control:** The process by which network resources are controlled to implement a given policy for a given user.
- **Priority:** A value assigned to specific packets transmitted to/from a user that determines the relative importance of transmitting those packets during the upcoming opportunity to use the medium.
- **Quality of service:** The measurable end-to-end performance properties of a network service, which can be guaranteed in advance by a service-level agreement between a user and a service provider, to satisfy specific customer application requirements. These properties may include throughput (bandwidth), transit delay (latency), error rates, security, packet loss, packet jitter, and so on.

Priority is typically included under the category of QoS, but it useful, when discussing policy control, to separate priority from other QoS parameters. The 5G network supports many commercial services and regulatory services (e.g., public safety communication) that need priority treatment. Some of these services share common QoS characteristics, such as latency and packet loss rate, but they may have different priority requirements. Mobile telephony and voice-based services for public safety share common QoS characteristics but may have different priority requirements.

Further, there are situations in which it is desirable to change the priority of a user connection but hold other QoS parameters constant and vice versa. As an example of the former, consider a healthcare patient monitored by wearable and/or implanted sensors, which provide periodic reports sent by a low-priority service. If the sensors detect a health emergency, such as the patient falling down, high-priority messages need to be sent to provide rapid response. Another example is a robot in an automated factory environment: During certain phases of the robot's operation, it may require lower latency or a higher data rate.

Connectivity Models

5G networks support both direct and indirect connectivity models for user equipment (UE). Figure 9.3 illustrates examples of three connection models:

- **Direct 3GPP connection:** An example is a sensor that communicates with an application server or with another device through a 5G network. Figure 9.3 shows a surveillance camera with 5G air interface capability.

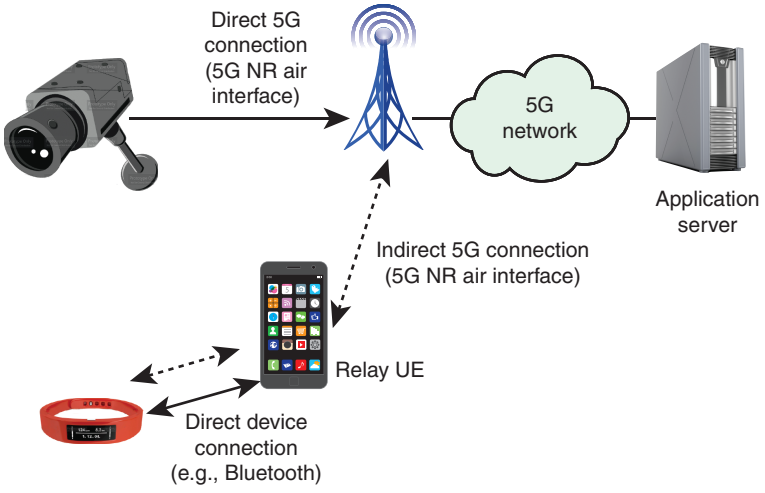


FIGURE 9.3 Connectivity Modes for Devices

- **Indirect 3GPP connection:** An example is a smart wearable that communicates through a smartphone to the 5G network. Figure 9.3 shows a fitness tracker that communicates with the user's smartphone via Bluetooth and through the phone to the 5G network. Another example is an IoT device that communicates using some wireless protocol with an IoT gateway or relay to a 5G network.
- **Direct device connection:** An example is a biometric device that communicates directly with other biometric devices or with a smartphone associated with the same patient. This is illustrated in Figure 9.3 by the connection between the fitness tracker and the smartphone.

Figure 9.3 shows an example of wearable health monitor device. In a remote setting, such as the patient's home, the device communicates indirectly to the application server. This is accomplished by a direct wireless connection to a local device that is capable of connecting to the 5G network. In an environment that provides direct support for multiple IoT devices, such as a hospital, the health monitoring devices can connect directly with the 5G network. In a large hospital setting, there may be thousands of wearable and implanted patient devices, as well as numerous other hospital-related IoT devices. This would require a massive machine type communications (mMTC) capability.

Network Capability Exposure and Context Awareness

In order to allow third parties to access information regarding capabilities provided by the 5G network (e.g., information for connectivity, QoS, and mobility) and to dynamically customize the network capabilities for diverse use cases, the 5G network should provide suitable ways (e.g., via APIs) to expose network capabilities and relevant information to third parties. Network capability exposure enables a third party to customize a dedicated network slice or allows the third party to manage an application in a service-hosting environment.

Applications may also provide the network with context awareness information. For example, radio resource management can be optimized if the network is informed about application characteristics (e.g., expected traffic over time). Other characteristics of a device, such as mobility, speed, and battery status, can be used to optimize allocation of functionality and content in the network.

Flexible Broadcast/Multicast Service

A flow, as implemented in SDN, is a distinguishable stream of related packets that results from a single user activity and requires the same QoS. The term **multicast** refers to a flow that has more than one recipient. If the group of recipients consists of all the potential recipients in some context, such as all the UE on a local network, or all the UE attached to a given virtual network, the term **broadcast** applies.

A high-capacity multicast/broadcast capability is an important requirement for 5G to meet the increasing demand for video services, ad hoc multicast/broadcast streams, software delivery over wireless, group communications, and broadcast/multicast IoT applications. A broadcast/multicast service should allow flexible and dynamic allocation of resources between unicast and multicast services within a network, and it should also allow the deployment of standalone broadcast networks. It should be possible to stream multicast/broadcast content efficiently over wide geographic areas as well as target the distribution of content to very specific geographic areas spanning only a limited number of base stations.

A white paper from 5G PPP (*View on 5G Architecture*, Version 3.0, June 2019) lists the following requirements for core network support of a flexible broadcast/multicast service:

- Enabling multicast and broadcast capabilities should require a small footprint on top of the existing unicast architecture.
- Wherever possible, treat multicast and broadcast as an internal optimization tool inside the network operator's domain.
- Consider terrestrial broadcast as a service offered also to UE without uplink capabilities that can be delivered as a self-containing service by a subset of functions of multicast and broadcast architecture.
- Simplify the system setup procedure to keep the system cost marginal. The goal is to develop an efficient system in terms of architecture/protocol simplicity and resource efficiency. Despite simplified procedures, the architecture should also allow flexible session management.
- Focus on the protocols that allow efficient IP multicast.
- Enable caching capabilities inside the network.

9.2 Core Network Functional Architecture

Chapter 3, "Overview of 5G Use Cases and Architecture," introduces the core network functional architecture. Figure 9.4, which repeats Figure 3.6, shows a service-based representation of the functional architecture defined in TS 23.501 (*Technical Specification Group Services and System Aspects, System Architecture for the 5G System [5GS], Stage 2 [Release 16]*, December 2020).

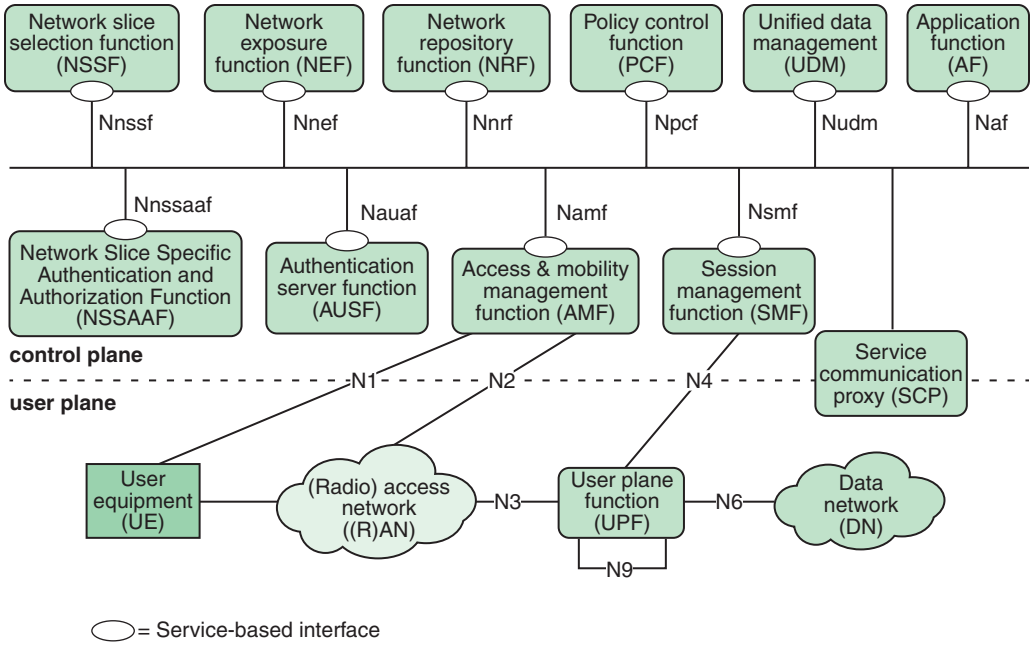


FIGURE 9.4 Non-Roaming 5G System Architecture

Table 9.1 summarizes the functionality of each network function (NF).

TABLE 9.1 Core Network Functions

Function	Description
Application function (AF)	Provides session-related information to the PCF so the SMF can use this information for session management.
Access and mobility management function (AMF)	Includes registration, reachability, and mobility management tasks.
Authentication server function (AUSF)	Performs authentication between UE and the network. The AMF initiates the UE authentication by invoking the AUSF. The AUSF selects an authentication method and performs UE authentication procedures.
Network exposure function (NEF)	Exposes capabilities of network functions and network slices as a service to third parties. In order to expose the capabilities, NEF stores the capability information and provides it upon capability discovery request.
Network repository function (NRF)	Assists the discovery and selection of required network functions (NFs). Each NF instance registers itself when instantiated and updates its status (i.e., activation/deactivation) so that the NRF can maintain information about the available network function instances. In general, each network slice instance has its own NRF, at least logically. In certain cases, such as when the network slice instances are in the same administrative domain, a single NFR instance can be shared by multiple network slice instances.

Function	Description
Network slice selection function (NSSF)	Selects appropriate network slice instances for UE. When UE requests registration with the network, the AMF sends a network slice selection request to the NSSF with preferred network slice selection information. The NSSF responds with a message including the list of appropriate network slice instances for the UE.
Policy control function (PCF)	Controls and manages policy rules, including rules for QoS enforcement, charging, and traffic routing. The PCF enables end-to-end QoS enforcement with QoS parameters (e.g., maximum bit rate, guaranteed bit rate, priority level) at the appropriate granularity (e.g., per UE, per flow, and per PDU session).
Session management function (SMF)	Provides connectivity (i.e., PDU session) for UE as well as control of the user plane for that connectivity (e.g., selection/re-selection of user plane network functions and user path, enforcement of policies including QoS policy and charging policy).
User plane function (UPF)	Performs traffic routing and forwarding, PDU session tunnel management, and QoS enforcement. The PDU session tunnels are used between access network and UPFs, as well as between different UPFs as user plane data transport for PDU sessions.
Unified data management (UDM)	Responsible for access authorization and subscription management. UDM works with the AMF and AUSF as follows: The AMF provides UE authentication, authorization, and mobility management services. The AUSF stores data for authentication of UE, and the UDM stores UE subscription data.

The interworking of these various NFs to implement the various procedures performed by the core network is extraordinarily complex. TS 23.502 (*Technical Specification Group Services and System Aspects, Procedures for the 5G System [5GS], Stage 2 [Release 16]*, December 2020) lists dozens of these procedures. The current version of the document is 603 pages long, suggesting the scale of the implementation task. This section is intended to provide insight into the functional operation of a 5G core network. The first two subsections examine two of the procedures defined in TS 23.502. The final subsection focuses specifically on the key role of the PCF.

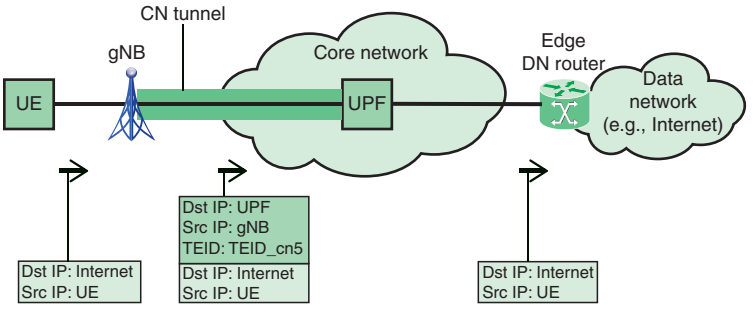
Tunneling

Before discussing the session establishment process, we need to cover the concept of an IP tunnel. Referring back to Figure 3.8, the 5G base stations, designated *gNB*, are connected to the core network and specifically provide user plane and control plane protocol terminations toward the UE. That is, a protocol connection exists between the gNB and two elements of the core network: the AMF and the UPF.

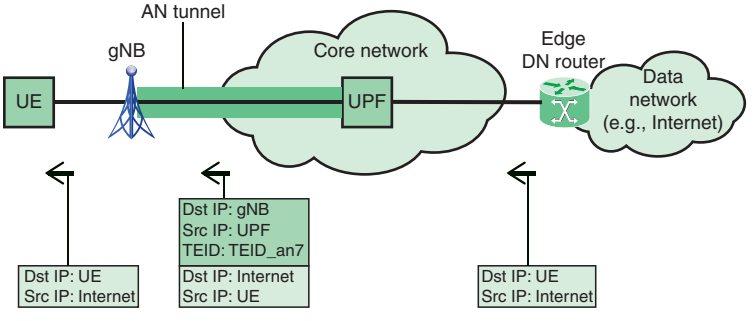
Consider UE that wishes to send an IP packet to an endpoint attached to the Internet. The UE communicates with the Internet in three stages: (1) to/from the radio access network (RAN) that provides a wireless link between the UE and the gNB; (2) using a link, typically a wireline, between the gNB and the core network; and (3) using a link from the core network and the endpoint on the Internet.

Figure 9.5a illustrates the process of transmitting an IP packet from the UE. The UE creates an IP packet. The packet header includes the source IP address of the UE and the destination IP address of an endpoint on the Internet. This packet is sent directly over the RAN to the gNB. However, the gNB does

not have a direct connection to the Internet. Instead, it has a connection to the core network—typically a fiber connection. (Chapter 15, "5G Radio Access Network," discusses other possibilities.) The gNB needs to send this packet to the UPF that is managing this session for the UE. To do this, the gNB encapsulates the entire IP packet from the UE by appending a new IP header with a source IP address of the gNB and a destination IP address of the UPF entity assigned to this session. In this operation, the original IP packet, including its header, from the UE is treated as a data block for the new, outer IP packet. This process is known as **tunneling,** and the path from the gNB to the UPF is referred to as a **tunnel**. In this case, the tunnel is known as a CN tunnel.



(a) Uplink traffic



(b) Downlink traffic

FIGURE 9.5 CN and AN Tunnels

The header for the tunnel packet also includes a tunnel endpoint ID (TEID), in this case labeled TEID_cn5. There will be multiple UEs connected to the gNB, and each will generate one CN tunnel (possibly more) to the same UPF. The core network has to be able to distinguish which tunnel belongs to which UE, and that is the purpose of the TEID.

Once the tunnel packet reaches the UPF, it strips off the outer header and sends the original IP packet to a router on the edge of the DN. The DN then routes that packet to a destination device on the DN. The core network is not concerned with this final step. The core network establishes a session that runs from the UE through the RAN and the CN and terminates at the edge router of the DN.

Figure 9.5a and the preceding discussion somewhat simplify the potential configuration. It may be that the session travels through a UPF located near the gNB to a UPF located near the DN edge router. In that case, an additional tunnel is needed between the two UPFs.

The CN tunnel is unidirectional, providing an uplink for the UE. For bidirectional data exchange, a similar unidirectional tunnel, called an AN tunnel, is needed in the downlink direction, as shown in Figure 9.5b. In this case, it is the UPF that adds the encapsulating IP header with an AN tunnel TEID. The header is stripped off by the gNB before the packet is delivered to the UE.

PDU Session Establishment

This subsection looks at the simple case of PDU session establishment initiated by UE. A PDU session—which may simply be called a session—is an association between the UE and a data network that provides a PDU connectivity service. A PDU connectivity service is a service that provides for the exchange of PDUs between UE and a data network (e.g., the Internet). The objective of the UE's PDU session establishment is to establish a default QoS flow between the UE and the data network (DN). The UE can then use the default QoS flow inside the established PDU session to exchange traffic with the DN. In 5G, QoS flow is the lowest granularity of a traffic flow where QoS and charging can be applied.

Here we return to the session establishment process for a new PDU session, which is briefly examined in Chapter 3, "Overview of 5G Use Cases and Architecture," and illustrated in simplified form in Figure 3.10. Figure 9.6, from TS 23.502, shows the process in greater detail. The procedure assumes that the UE has already registered on the AMF and that the AMF has already retrieved the user subscription data from the UDM. The following steps are involved:

Step 1. From UE to AMF: In order to establish a new PDU session, UE generates a new PDU session ID and sends a message containing a PDU session establishment request to the AMF. The message contains a number of parameters, including a PDU session ID, request type, session management capabilities, protocol configuration option, data network name (DNN), and PDU data network (DN) request container (authorization information). The access network (AN) encapsulates the message sent by the UE in an RP-an message (where the RP-an is a reference point between the AN and AMF) and sends it, together with user location information and access type information, to the AMF.

Step 2. The AMF determines that the message corresponds to a request for a new PDU session based on the fact that the request type indicates "initial request" and that the PDU session ID is not used for any existing PDU session(s) of the UE. The AMF selects an SMF, considering the target data network, network slice instance (NSI), subscription information retrieved from the UDM, and access type information.

Step 3. From AMF to SMF: The AMF sends a Nsmf_PDUSession_CreateSMContext request to the SMF.

Step 4. Based on the data provided by UE, the SMF communicates with the UDM and PCF to get relevant information for PDU session creation and to determine whether the request is valid.

Step 5. From SMF to AMF: If the request is valid, the SMF returns a Nsmf_PDUSession_CreateSMContext response, which includes SM context information.

Step 6. Optional secondary authentication/authorization: If the session requires authentication and authorization, this is performed, as described in a separate part of TS 23.502.

Step 7. The purpose of step 7 is to receive policy and charging control (PCC) rules before selecting the UPF instance. Policy control is the process whereby the PCF indicates to the SMF how to control the QoS flow. Policy control includes QoS control and/or gating control. Gating control is the process of blocking or allowing packets that belong to a service data flow or detected application's traffic to pass through to the UPF. Charging control is the process of applying online charging and/or offline charging, as appropriate. Step 7 has two substeps:

Step 7a. The SMF selects a PCF instance for this session. The following factors may be considered at PCF discovery and selection for a PDU session:

- Local operator policies
- Selected data network name (DNN)
- The network slice instance of the PDU session

Step 7b. The SMF may perform an SM policy association establishment procedure to establish a session management (SM) policy association with the PCF and get the default PCC rules for the PDU session.

Step 8. If the request type in the PDU session establishment request is "initial request," the SMF selects one or more UPFs, as needed. For IP type PDU sessions, the SMF allocates an IP address (prefix) for the PDU session. If the request type is "existing PDU session," the SMF maintains the same IP address (prefix) that has already been allocated to the UE. The selection and reselection of the UPF are performed by the SMF by considering UPF deployment scenarios such as a centrally located UPF and a distributed UPF located close to or at the access network site. The selection of the UPF also enables deployment of the UPF with different capabilities (e.g., UPFs supporting no or a subset of optional functionalities).

Step 9. The SMF may perform an SMF-initiated SM policy association modification procedure in the event that a policy control request trigger condition is met. The policy control request triggers relevant for SMF define the conditions when the SMF shall interact again with the PCF after PDU session establishment. Examples of triggers include the UE moving to another operator's domain, a QoS change, and a routing information change. The PCF may provide updated policies to the SMF.

Step 10. The SMF initiates a session establishment procedure with the selected UPF. This involves the following two substeps:

Step 10a. The SMF sends a session establishment request to the UPF and provides packet detection, enforcement, and reporting rules to be installed on the UPF for this PDU session. If the SMF is configured to request IP address allocation from the UPF, then the SMF indicates to the UPF to perform the IP address/prefix allocation and includes the information required for the UPF to perform the allocation. If CN tunnel information is allocated by the SMF, the CN tunnel information is provided to UPF in this step. The SMF also determines a number of other services related to this session.

Step 10b. The UPF acknowledges by sending a session establishment response.

Step 11. SMF to AMF: The SMF requests the AMF to transfer SM information for the requested PDU session to the UE and AN. The SM message transfer signaling message to the AMF contains the PDU session ID, which allows the AMF to know which AN toward the UE to use. The message contains SM information to be forwarded to the AN by the AMF that includes CN tunnel information, QoS-related information, and other session-related information. The message also contains SM information to be forwarded to the UE by the AMF via the AN, which includes CN tunnel information, QoS-related information, and other session-related information.

Step 12. AMF to AN: The AMF sends session-related information to the AN that is related to the information received from the SMF.

Step 13. AN to UE: The AN allocates an AN tunnel (between the AN gNB and the core network) for the PDU session. The AN sends session-related information to the UE that is related to the information received from the SMF.

Step 14. AN to AMF: The AN sends a response message to the AMF with AN tunnel information and other SM-related information.

Step 15. AMF to SMF: The AMF forwards the SM information received from the AN to the SMF through a context request message.

Step 16. SMF-UPF exchange: The SMF initiates a session modification procedure with the UPF by sending a session modification request message. The SMF provides AN tunnel information to the UPF as well as corresponding forwarding rules. The UPF responds to the SMF with an RP-su session modification response message. If multiple UPFs are used in the PDU session, the UPF terminating the CN tunnel performs this procedure. After this step, the UPF delivers any downlink PDUs to the UE. If this is an authenticated UE, the SMF registers the PDU session with the UDM by providing SM information.

Step 16a. The SMF initiates a session modification procedure with the UPF. The SMF provides AN tunnel information to the UPF as well as the corresponding forwarding rules.

Step 17. SMF to AMF: Nsmf_PDUSession_UpdateSMContext response (cause): The SMF responds to the context request received from the AMF in step 15. After this step, the AMF forwards relevant events that the SMF subscribes to (e.g., location reporting, UE moving into or out of the area of interest).

Step 18. (Conditional) SMF to AMF: If any time after step 5 the PDU session establishment is not successful, the SMF informs the AMF by invoking Nsmf_PDUSession_SMContextStatusNotify (release). The SMF also releases any session(s) created, any PDU session address, if allocated (e.g., IP address), and the association with the PCF, if any. In this case, step 19 is skipped.

Step 19. SMF to UE: In the case of PDU session type IPv6 or IPv4v6, the SMF generates an IPv6 router advertisement and sends it to the UE.

Step 20. If the UE has indicated support for transferring port management information containers, the SMF informs the PCF that a manageable Ethernet port has been detected. The SMF also includes the port number, MAC address, and port management information container.

Step 21. If the PDU session establishment failed after step 4, the SMF unsubscribes the modification of SM subscription data.

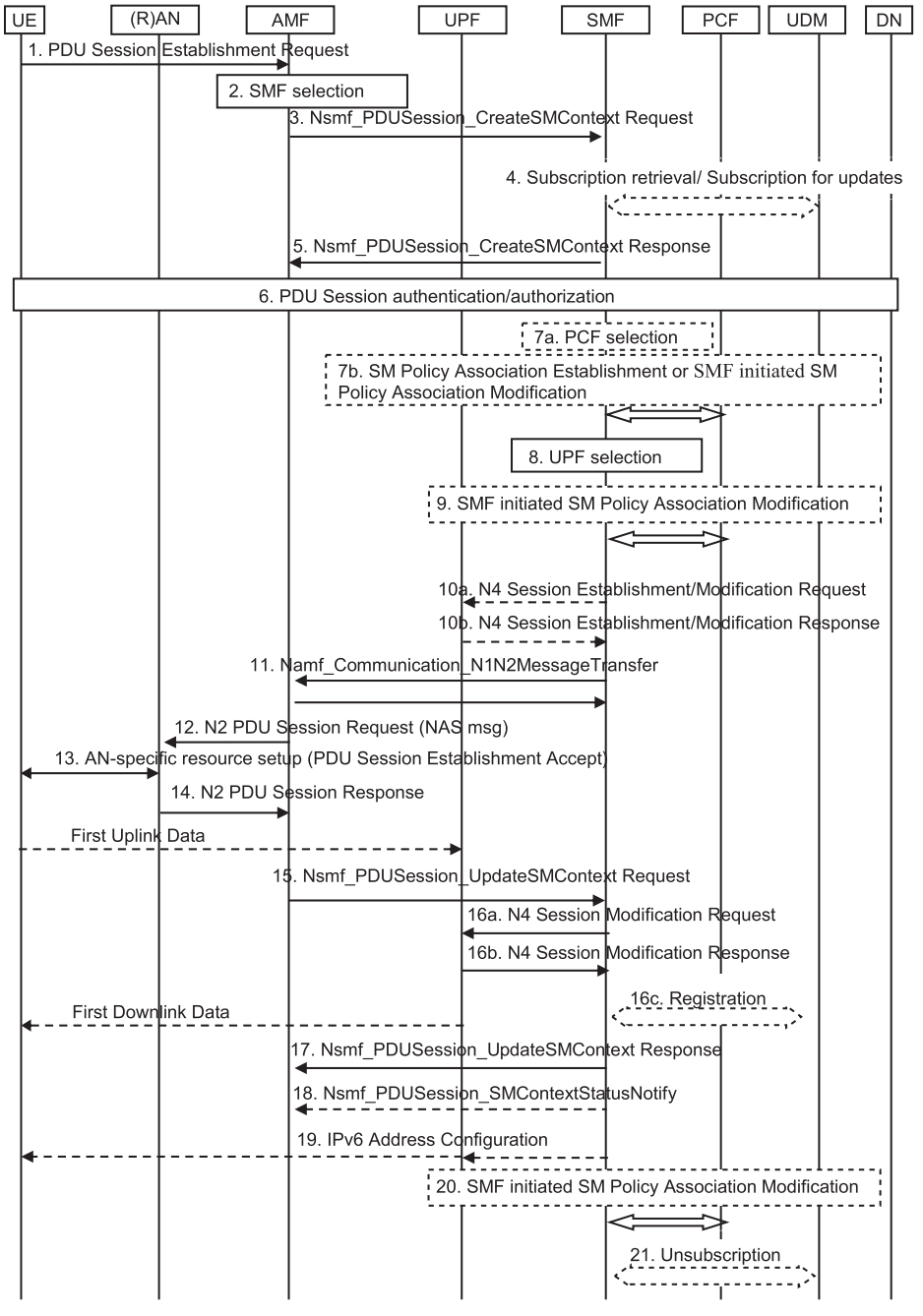


FIGURE 9.6 UE-Requested PDU Session Establishment

Figure 9.7 illustrates the message flows involved in PDU session establishment and gives some idea of the complexity of the operation. However, the preceding 21-step enumeration of tasks is only a summary overview. The specification in TS 23.502 occupies 21 pages, and it includes numerous references to other sections in the same TS as well as other TS documents. Thus, the full specification runs to well over 100 pages—and this is just one of dozens of procedures that must be implemented in the core network.

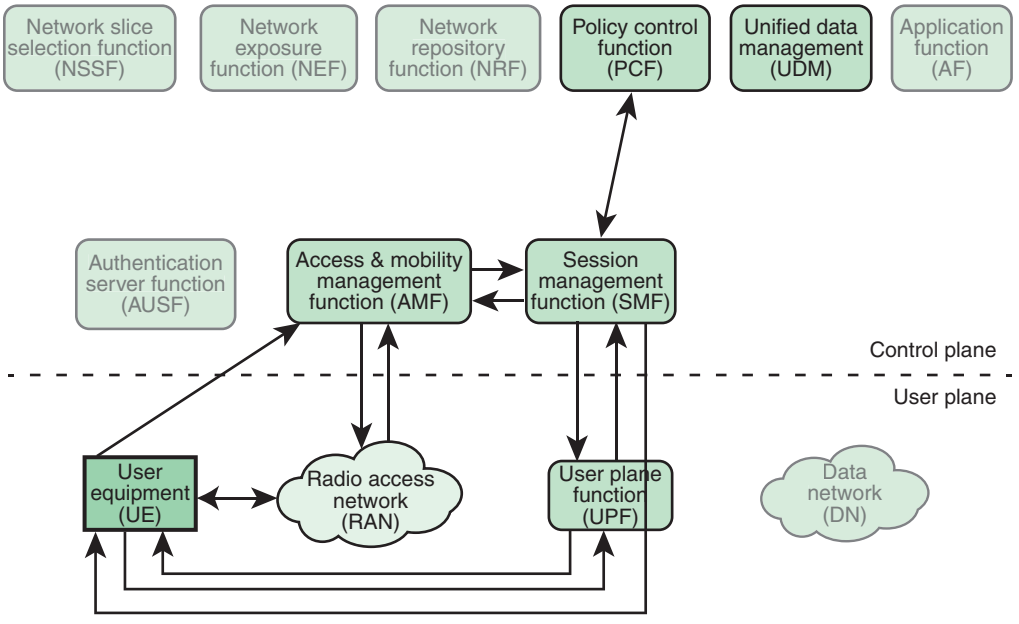


FIGURE 9.7 PDU Session Establishment Message Flow

Policy Control Function

The core network supports a common policy framework, together with network policies that allow UEs to choose the most suitable access network and access-agnostic QoS mechanisms. Thus a key element of the core network functional architecture is the policy control function (PCF).

Before discussing PCF functionality, consider the following related terms:

- **PDU session:** This is a logical connection that carries all the communication between UE and a data network (DN).
- **QoS flow:** This is the lowest level of granularity within the 5G system for defining policy and charging rules. A PDU session may contain multiple QoS flows.
- **Service data flow (SDF):** An SDF provides an end-to-end packet flow between UE and a specific application at the DN. One or more SDFs can be transported in the same QoS flow if they share the same policy and charging rules.

PCF Requirements

The overall requirement on the PCF function is to enable the core network to apply policy and charging control to UE accesses. As mentioned in the preceding subsection, policy control is the process whereby the PCF indicates to the SMF how to control the QoS flow. Policy control includes QoS control and/or gating control. Gating control is the process of blocking or allowing packets that belong to a service data flow or detected application's traffic to pass through to the UPF. Charging control is the process of applying online charging and/or offline charging, as appropriate.

3GPP TS 23.503 (*Technical Specification Group Services and System Aspects*; *Policy and charging control framework for the 5G System [5GS]*; *Stage 2 [Release 16]*, April 2020) breaks down specific PCF requirements into non-session management–related requirements and session management–related requirements. The non-session management–related requirements are as follows:

- **Access- and mobility-related policy control requirements:** These requirements support the AMF, as described later in this chapter.
- **UE policy control requirements:** These requirements provide policy information to the UE.
- **Network status analytics information requirements:** These requirements relate to collecting slice-specific network status analytic information and using that data in policy decisions.
- **Management of packet flow descriptions:** These descriptions provide the capability to create, update, or remove PFDs in the NEF (PFDF) and the distribution from the NEF (PFDF) to the SMF and finally to the UPF.
- **SMF selection management–related policy control requirements:** These requirements provide SMF selection management–related policies to the AMF.
- **Support for non-session management–related network capability exposure:** This support enables an AF to request non-session management–related policy control functionality from the NEF.

The session management–related requirements are as follows:

- **Charging-related requirements:** These requirements provide information to allow for charging control on each SDF.
- **Policy control requirements:** These requirements cover gating control and QoS control requirements:
    - **Gating control:** You apply gating control on a per-SDF basis. For example, session termination triggers gating control for each affected SDF.
    - **QoS control:** The PCF must support QoS control and the SDF, QoS flow, and PDU session levels.

- **Usage monitoring control requirements:** The PCF may use monitoring of both volume and time use to make dynamic policy decisions. It sends the applicable thresholds (of time or volume) to the SMF for monitoring and notification to the PCF. The monitoring is possible for an individual SDF or a group of SDFs or all traffic on a PDU session.
- **Application detection and control requirements:** The application detection and control feature comprises the request to detect the specified application traffic, the report from the SMF to the PCF on the start or stop of application traffic, and the application of the specified enforcement and charging actions.
- **Support for session management–related network capability exposure:** This support enables an AF to request session management–related policy control functionality for capability exposure.
- **Traffic steering control:** This is the capability to activate/deactivate traffic steering policies from the PCF in the SMF.

Interfaces with Other Network Functions

Figure 9.8 shows a reference point representation indicating how the PCF interfaces with other network functions.

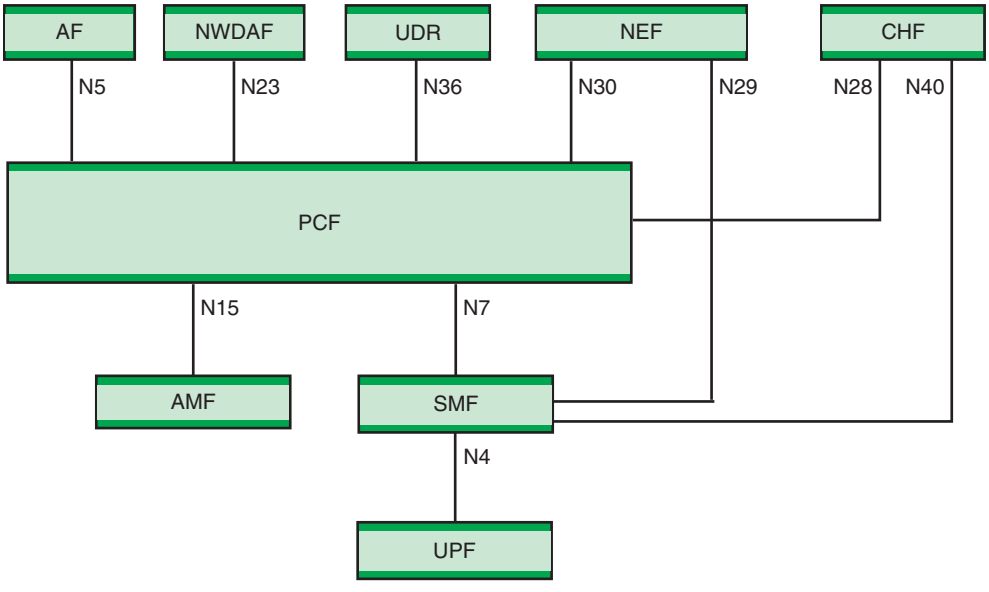


FIGURE 9.8 Overall Non-Roaming Reference Architecture of Policy and Charging Control of Framework for the 5G System (Reference Point Representation)

This figure introduces the following NFs that are not shown in Figure 9.4:

- **Network data analytics function (NWDAF):** Used for data collection and analytics for centralized as well as edge computing resources. It provides network slice–specific data analytics to the PCF and NSSF, which in turn use this data for policy decisions (PCF) and slice selections (NSSF).
- **Unified data repository (UDR):** Serves as a single repository of subscription data, application data, and policy data by integrating with NF consumers (including NEF, AMF, and PCF). It also notifies for the subscription data changes.
- **Charging function (CHF):** Provides an account balance management function, a rating function, and a charging gateway function.

The PCF interacts with other interfaces as follows:

- **PCF–AF interface:** This interface allows for the transport of application-level session information and Ethernet port management information from the AF to the PCF. This includes bandwidth requirements for QoS, identification of application service providers and applications, traffic routing based on applications access, and identification of application traffic for policy and charging control.
- **PCF–SMF interface:** This interface enables the PCF to have dynamic control over the policy and charging behavior at an SMF. The SMF receives control plane information from NFs and user plane information from the UPF. An SMF triggers the PCF to enforce policy decisions when the policy trigger related to session management is met.
- **PCF–SMF–UPF interface:** The PCF and UPF don't communicate directly with each other. They exchange policy actions/enforcements via the SMF. The SMF provisions the policy and threshold rules on the UPF for usage control based on the static/dynamic policy rules configured in the PCF, predefined rules in the SMF, and/or credit control triggers received from the CHF.
- **PCF–AMF interface:** The AMF acts as a single entry point for the UE connection. The PCF provides access and mobility management–related policies for the AMF in order to trigger policy rules on the UE or user sessions.
- **PCF–UDR interface:** The PCF retrieves the policy-/subscription-/application-specific data from the UDR. Policy control–related subscription and application-specific data gets provisioned into the UDR. The UDR can also generate notifications based on the changes in the subscription information, according to the operator's pricing model.
- **PCF–CHF interface:** This interface enables the PCF to access policy control status information related to subscriber spending. The CHF stores the policy counter information against the subscriber pricing plan and notifies the PCF whenever the subscriber breaches the policy thresholds, based on usage consumption. On receiving policy trigger information, the PCF

applies the policy decision by interacting with the SMF (which in turn informs the UPF for the policy enforcement).

- **PCF–NEF interface:** The NEF exposes network function services and resources to the external world. In terms of interaction with the PCF, it exposes the capabilities of network functions for supporting policy and charging.
- **PCF–NWDAF interface:** The PCF collects slice-specific network status analytic information from the NWDAF. The NWDAF provides network data analytics (i.e., load-level information) to the PCF on a network slice level. The PCF is able to use that data in its policy decisions.

9.3 Quality of Service

A wide variety of applications and devices use 5G networks, including cloud computing, big data, the pervasive use of mobile devices on enterprise networks, and the increasing use of video streaming. These factors together contribute to the increasing difficulty in maintaining satisfactory network performance. The key tool in characterizing and measuring the network performance that an enterprise desires to achieve is quality of service (QoS). QoS is the measurable end-to-end performance properties of a network service, which can be guaranteed in advance by a service-level agreement (SLA) between a user and a service provider in order to satisfy specific customer application requirements. QoS enables a network manager to determine whether the network is meeting user needs and to diagnose problem areas that require adjustment to network management and network traffic control.

This section begins by summarizing the QoS capabilities required in a 5G network, as defined by ITU-T. Then, it introduces a QoS architectural framework that provides insight into the scope and complexity of a QoS system. The remainder of the section covers QoS details defined by 3GPP.

QoS Capabilities

QoS capabilities and accompanying SLAs serve two purposes:

- Enable networks to offer different levels of QoS to customers on the basis of customer requirements.
- Allocate network resources efficiently, maximizing effective capacity.

ITU-T Y.3106 (*Quality of Service Functional Requirements for the IMT-2020 Network*, April 2019) defines a QoS life cycle management process that encompasses the entire range of capabilities involved in providing QoS. Figure 9.9, based on a figure in Y.3106, shows the four stages in the QoS management life cycle. As shown, a QoS capability encompasses the interface between the UE and the AN, where the QoS service is visible to the user; the functionality in the AN and in the CN to provide the QoS; and the ability to exchange QoS information and requirements with other networks.

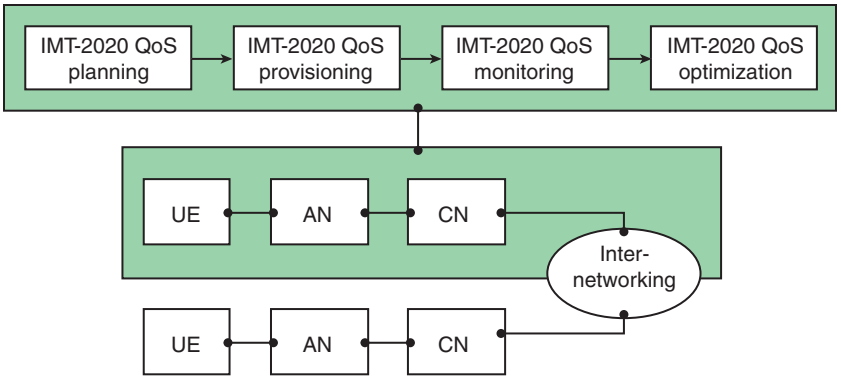


FIGURE 9.9 High-Level QoS Capabilities of the IMT-2020 Network

Table 9.2 defines the four QoS management categories.

TABLE 9.2 QoS Management Categories

Category	Definition
QoS planning	The process of determining the mechanisms and services to be implemented on the network.
QoS provisioning	The process of configuring and maintaining selected network elements based on customer SLAs and observed quality performance.
QoS monitoring	The process of collecting QoS statistics, faults, and warnings. This data is then used for generating analysis reports and making changes and upgrades to the network.
QoS optimization	The process of accessing monitored information, processing the data to determine service and network quality metrics, and initiating corrective actions when any of the quality levels is considered unsatisfactory.

Important requirements for QoS planning include the following:

- Support service-driven QoS planning for the IMT-2020 network.
- Support dynamical modeling of diversified IMT-2020 usage scenarios (e.g., eMBB, MTC, and URLLC).
- Convert service models to traffic models accurately.
- Support an accurate estimate of network coverage, capacity, resources, and network slice requirements.
- Estimate and allocate network resources in a way that efficiently maximizes utilization.
- Support QoS-aware routing to satisfy different service requirements for delay, bandwidth, throughput, load balance, cost, etc.

QoS provisioning requirements are as follows:

- Support E2E QoS for diversified IMT-2020 usage scenarios (eMBB, MMTC, and URLLC).
- Support translation of service-centric SLA to resource-facing network slice descriptions.
- Support efficient E2E QoS provisioning with the capabilities of a global network view, on-demand softwarized network functions, autonomous network slicing management, and orchestration.
- Support unified and access-agnostic (fixed or mobile access) QoS control from a core network (CN) perspective.
- Support proper QoS interworking and mapping among UE, AN, CN, and other data networks (DNs).
- Support a finer level of QoS granularity based on flows to meet different service requirements.
- Support QoS enforcement, which includes flow classification, marking, congestion avoidance, queue shaping, and queue scheduling based on QoS rules.

QoS monitoring requirements are as follows:

- Provide a mechanism for supporting real-time E2E QoS monitoring.
- Provide an interface to applications for QoS monitoring (e.g., to initiate QoS monitoring, request QoS parameters, events, or logging information).
- Respond to an authorized user request to provide real-time QoS monitoring information within a specified time after receiving the request.
- Provide real-time QoS parameters and events information to an authorized application or network entity.
- Support an update or refresh rate for real-time QoS monitoring within a specified time.
- Log the history of QoS events, including, for example, parts of the SLA that are not met and timestamps of the events and event positions (e.g., UE and radio access points associated with the events).
- Support different levels of granularity for QoS monitoring (e.g., per flow or per set of flows).

QoS optimization requirements are as follows:

- Support intelligent QoS anomaly detection based on the analysis of QoS data.
- Support traffic prediction based on the analysis of QoS data.
- Support QoS anomaly prediction based on the analysis of QoS data.
- Support QoS optimization to provide and ensure a desired service performance level during the life cycle of the service.

QoS Architectural Framework

Before we look at the Internet standards that deal with provision of QoS on the Internet and in private internetworks, in this section we consider an overall architectural framework that relates the various elements that go into QoS provision. Such a framework has been developed by ITU-T. Recommendation Y.1291 (*An Architectural Framework for Support of Quality of Service in Packet Networks*, May 2004) provides an overview of the mechanisms and services that comprise a QoS facility.

The Y.1291 framework consists of a set of generic network mechanisms for controlling the network service response to a service request, which can be specific to a network element, or for signaling between network elements, or for controlling and administering traffic across a network. Figure 9.10 shows the relationships among these elements, which are organized into three planes: data, control, and management. This architectural framework is an excellent overview of QoS functions and their relationships and provides a useful basis for summarizing QoS.

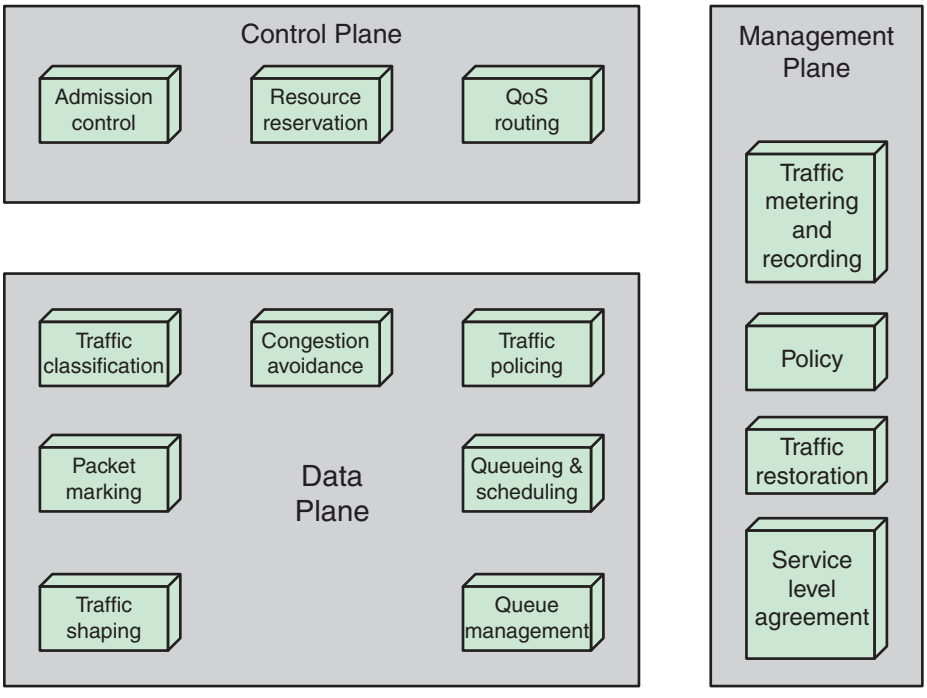


FIGURE 9.10 Architectural Framework for QoS Support

Data Plane

The data plane includes mechanisms that operate directly on flows of data. The following discussion briefly describes each mechanism in turn.

Traffic classification refers to the assignment of packets to a traffic class by the ingress router at the ingress edge of the network. Typically, the classification entity looks at multiple fields of a packet, such as source and destination address, application payload, and QoS markings, and determines the aggregate to which the packet belongs. This classification gives network elements a method to weigh the relative importance of one packet over another in a different class. All traffic assigned to a particular flow or other aggregate can be treated similarly. The flow label in the IPv6 header can be used for traffic classification. Other routers en route perform a classification function as well, but the classification does not change as the packets traverse the network.

Packet marking encompasses two distinct functions. First, packets may be marked by ingress edge nodes of a network to indicate some form of QoS that the packet should receive. Examples include the Differentiated Services (DS) field in IPv4 and IPv6 packets and the Traffic Class field in Multiprotocol Label Switching (MPLS) labels. An ingress edge node can set the values in these fields to indicate a desired level of QoS. Such markings may be used by intermediate nodes to provide differential treatment to incoming packets. Second, packet marking can be used to mark packets as nonconformant, either by the ingress node or intermediate nodes, so that they can be dropped later, if congestion is experienced.

Traffic shaping controls the rate and volume of traffic entering and transiting the network on a per-flow basis. The entity responsible for traffic shaping buffers nonconformant packets until it brings the respective aggregate into compliance with the traffic. The resulting traffic thus is not as bursty as the original and is more predictable.

Congestion avoidance deals with means for keeping the load of the network under its capacity such that it can operate at an acceptable performance level. The specific objectives are to avoid significant queuing delays and, especially, to avoid congestion collapse. In a typical congestion avoidance scheme, senders reduce the amount of traffic entering the network upon an indication that network congestion is occurring (or is about to occur). Unless there is an explicit indication, packet loss or timer expiration is normally regarded as an implicit indication of network congestion.

Traffic policing determines whether the traffic being presented is, on a hop-by-hop basis, compliant with prenegotiated policies or contracts. Nonconformant packets may be dropped, delayed, or labeled as nonconformant.

Queuing and scheduling algorithms, also referred to as queuing discipline algorithms, determine which packet to send next and are used primarily to manage the allocation of transmission capacity among flows.

Queue management algorithms manage the length of packet queues by dropping packets when necessary or appropriate. Active management of queues is concerned primarily with congestion avoidance.

Control Plane

The **control plane** is concerned with creating and managing the pathways through which user data flows. It includes admission control, QoS routing, and resource reservation.

Admission control determines what user traffic may enter the network. This may be in part determined by the QoS requirements of a data flow compared to the current resource commitment in the network. But beyond balancing QoS requests with available capacity to determine whether to accept a request, there are other considerations in admission control. Network managers and service providers must be able to monitor, control, and enforce use of network resources and services based on policies derived from criteria such as the identity of users and applications, traffic/bandwidth requirements, security considerations, and time of day/week.

QoS routing determines a network path that is likely to accommodate the requested QoS of a flow. This contrasts with the philosophy of the traditional routing protocols, which generally look for a least-cost path through the network.

Resource reservation is a mechanism that reserves network resources on demand for delivering desired network performance to a requesting flow.

Management Plane

The management plane is concerned with mechanisms that affect both control plane and data plane mechanisms. The control plane deals with the operation, administration, and management aspects of the network. It includes SLAs, traffic restoration, traffic metering and recording, and policy.

A **service-level agreement (SLA)** typically represents an agreement between a customer and a provider of a service that specifies the level of availability, serviceability, performance, operation, or other attributes of the service.

Traffic metering and recording concerns monitoring the dynamic properties of a traffic stream using performance metrics such as data rate and packet loss rate. It involves observing traffic characteristics at a given network point and collecting and storing the traffic information for analysis and further action. Depending on the conformance level, a meter can invoke necessary treatment (e.g., dropping or shaping) for the packet stream. Section 9.7 discusses the types of metrics that are used in this function.

Traffic restoration refers to the network response to failures. This encompasses a number of protocol layers and techniques.

Policy is a category that refers to a set of rules for administering, managing, and controlling access to network resources. These rules can be specific to the needs of the service provider or reflect an agreement between the customer and service provider, which may include reliability and availability requirements over a period of time and other QoS requirements.

QoS Classification, Marking, and Differentiation

The 3GPP document TS 23.501 uses the following terms:

- **Traffic classification:** Grouping traffic into classes based on user-defined QoS values
- **User plane marking:** Marking packets to indicate to which QoS classification they belong
- **QoS differentiation:** Using a different QoS set of values for different categories of traffic

Recall from Section 9.2 that a PDU session between UE and a DN may contain multiple QoS flows, and each QoS flow may contain one or more service data flows (SDFs). An SDF is associated with a particular application. A QoS flow is where QoS differentiation takes place and where packets are marked to indicate their traffic classification.

Figure 9.11, from TS 23.501, illustrates 5G principles for classification and marking of user plane traffic and mapping of QoS flows. The mapping happens two times. In the core, the UPF maps a QoS flow to a single tunnel. There is a one-to-many relationship between the tunnel on the AN core interface and the data radio bearers on the air interface. A RAN node (gNB) may map multiple QoS flows to one data radio bearer. Incoming application data packets are classified based on the QoS and service requirements of the service data flows of the application. The session management function (SMF) assigns the QoS flow ID (QFI) and derives its QoS profile from the information provided by the PCF.

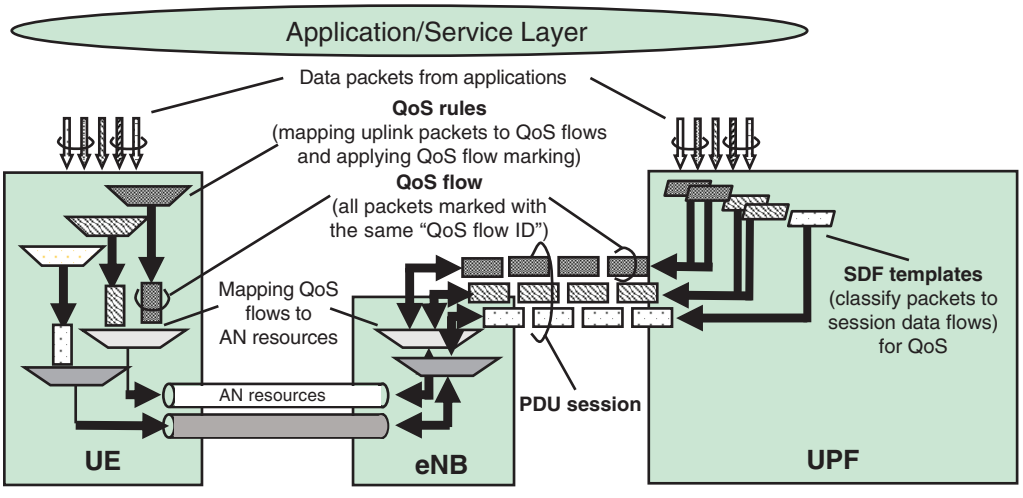


FIGURE 9.11 The Principles for Classification, User Plane Marking, and Differentiation in 5G

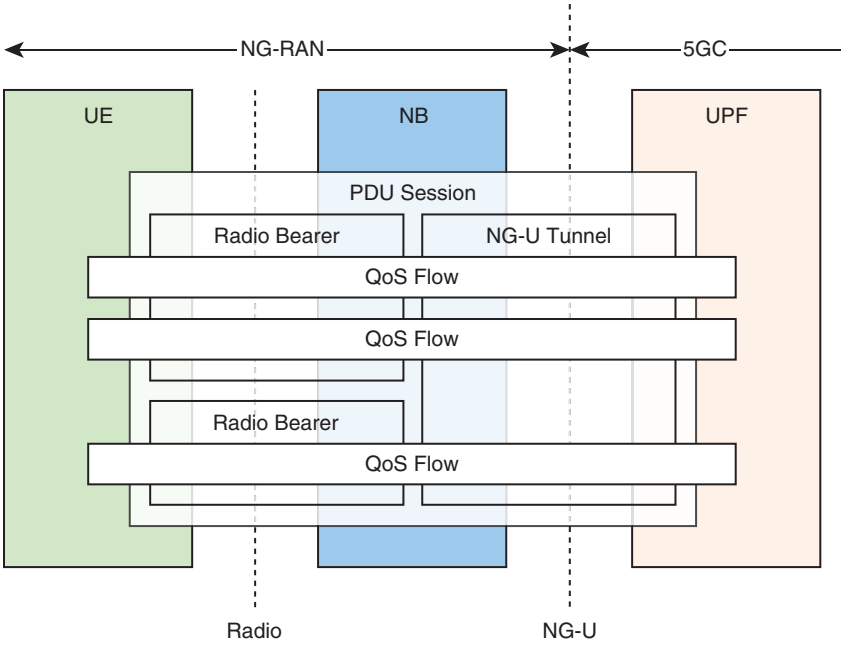
The SMF provides:

- The QFI together with the QoS profile to the AN
- QoS flow marking (i.e., the QFI) and the necessary information to enable classification
- Bandwidth enforcement and marking of user plane traffic to the UPF
- QoS rules enabling classification and marking of user plane traffic to the UE

The QoS capability includes reflective QoS, which is a method to reduce the signaling to the UE for uplink (UL) data classification. Reflective QoS applies the same QoS profile to both UL and downlink (DL), allowing a simple principle for applying the same QoS differentiation to application data in the DL and UL.

3GPP QoS Architecture

Figure 9.12, from TS 38.300 (*Technical Specification Group Radio Access Network*; *NR*; *NR and NG-RAN Overall Description*; *Stage 2 [Release 16]*; September 2020), shows a high-level view of the 3GPP QoS architecture, encompassing both the radio access network (RAN) and the core network (5GC). NG-RAN and 5GC ensure QoS by mapping packets to appropriate QoS flows.



5GC = 5G Core Network
NB = NG-RAN Node
NG-RAN: Next Generation Radio Access Network
NG-U: User Plane Interface Between NG-RAN and 5GC
UE = User Equipment
UPF = User Plane Function

FIGURE 9.12 3GPP QoS Architecture

At the RAN, radio bearer paths may transmit one or more QoS flows, as long as the performance parameters of the radio bearer are sufficient for the flows. Recall that a radio bearer is an information transmission path of defined capacity, delay, bit error rate, and other parameters.

On the core network side, the flows in a PDU session are exchanged through the bidirectional tunnels to the UPF. The core network then performs the functions required to support the QoS for each flow.

QoS Parameters

The QoS model developed in TS 23.501 makes use of a key set of QoS parameters and QoS characteristics, as shown in Figure 9.13. Together, the parameters and characteristics define the requirements associated with a QoS flow.

QoS Parameters
5QI (5G QoS identifier)
ARP (allocation and retention priority)
RQA (reflective QoS attribute)
Notification control
Flow bit rates
Aggregate bit rates
Default values
Maximum packet loss rate
Wireline access network-specific 5G QoS parameters

QoS Characteristics
Resource type
Priority level
Packet delay budget
Packet error rate
Averaging window
Maximum data burst volume

FIGURE 9.13 Elements of 3GPP QoS Model

The following parameters are associated with a QoS flow:

- **5QI (5G QoS identifier):** This is an integer value used as a reference to a set of values assigned to QoS characteristics. Thus, a standardized combination of QoS characteristics can be preconfigured so that the AN and CN are informed of the QoS characteristics for a flow by means of the 5QI.
- **ARP (allocation and retention priority):** The ARP consists of three attributes:
  - **ARP priority level:** Defines the relative importance of a QoS flow. The range of the ARP priority level is 1 to 15, with 1 as the highest level of priority. In cases of congestion, when all QoS requirements cannot be fulfilled for one or more QoS flows, the priority level determines for which QoS flows the QoS requirements are prioritized. In cases where there is no congestion, the priority level determines the resource distribution between QoS flows.
  - **ARP pre-emption capability:** Defines whether a QoS flow may get resources that were already assigned to another QoS flow with a lower ARP priority level. It is set as either enabled or disabled.

- **ARP pre-emption vulnerability:** Defines whether a QoS flow may lose the resources assigned to it in order to admit a QoS flow with a higher ARP priority level. It is set as either enabled or disabled.

- **RQA (reflective QoS attribute):** Reflective QoS means that the UE uses the same QoS parameters on the uplink as obtained from the downlink QoS flow. RQA, when included, indicates that some (not necessarily all) traffic carried on this QoS flow is subject to reflective QoS.
- **Flow bit rates:** There are two categories of flow bit rates. A guaranteed bit rate (GBR) guarantees at least a minimum bit rate capacity for the flow. A non-GBR QoS flow does not guarantee the bit rate. GBR QoS flows include the following parameters:
  - **Guaranteed flow bit rate (GFBR) UL and DL:** Denotes the bit rate that is guaranteed to be provided by the network to the QoS flow over the averaging time window.
  - **Maximum flow bit rate (MFBR) UL and DL:** Limits the bit rate to the highest bit rate that is expected by the QoS flow (e.g., excess traffic may get discarded or delayed by a rate-shaping or policing function at the UE, RAN, or UPF). Bit rates above the GFBR value and up to the MFBR value may be provided with relative priority determined by the priority level of the QoS flows.
- **Notification control:** This parameter indicates whether notifications are requested from the NG-RAN when the GFBR can no longer (or can again) be guaranteed for a QoS flow during the lifetime of the QoS flow. Notification control may be used for a GBR QoS flow if the application traffic is able to adapt to the change in the QoS (e.g., if the AF is capable of triggering rate adaptation).
- **Aggregate bit rates:** Two parameters related to bit rates are associated with each UE:
  - **Per session aggregate maximum bit rate (Session-AMBR):** For each PDU session of a UE, this parameter limits the aggregate bit rate that can be expected to be provided across all non-GBR QoS flows for a specific PDU session. Session-AMBR is measured over an AMBR averaging window, which is a standardized value. Session-AMBR is not applicable to GBR QoS flows.
  - **Per UE aggregate maximum bit rate (UE-AMBR):** This parameter limits the aggregate bit rate that can be expected to be provided across all non-GBR QoS flows of a UE. Each AN sets its UE-AMBR to the sum of the Session-AMBR of all PDU sessions with an active user plane to this AN up to the value of the received UE-AMBR from the AMF. The UE-AMBR is a parameter provided to the AN by the AMF, based on the value of the subscribed UE-AMBR retrieved from UDM or the dynamic serving network UE-AMBR retrieved from the PCF (e.g., for a roaming subscriber). The AMF provides the UE-AMBR provided by the PCF to the AN, if available. The UE-AMBR is measured over an AMBR averaging window, which is a standardized value. The UE-AMBR is not applicable to GBR QoS flows.

- **Default values:** For each PDU session setup, these default values apply to one or more non-GBR flows. The SMF retrieves the subscribed Session-AMBR values as well as the subscribed default values for the 5QI and the ARP and, optionally, the 5QI priority level, from the UDM.
- **Maximum packet loss rate:** This parameter indicates the maximum rate for lost packets of the QoS flow that can be tolerated in the uplink and downlink directions.
- **Wireline access network-specific 5G QoS parameters:** There are additional parameters applicable only to wireline access networks.

Note that the guaranteed flow bit rate (GFBR) and maximum flow bit rate (MFBR) only apply to GBR flows. The per session aggregate maximum bit rate and the per UE aggregate maximum bit rate only apply to non-GBR flows.

QoS Characteristics

Each 5QI has associated with it a set of values of characteristics that describe the packet forwarding treatment that a QoS flow receives edge-to-edge between the UE and the UPF. The following characteristics are associated with a QoS flow:

- **Resource type:** There are three resource types: GBR, delay-critical GBR, and non-GBR. Both GBR types are typically authorized on demand, which requires dynamic policy and charging control. The definitions of PDB (packet delay budget) and PER (packet error rate) are different for GBR and delay-critical GBR resource types, and MDBV (maximum data burst volume) applies only to the delay-critical GBR resource type. A non-GBR QoS flow may be pre-authorized through static policy and charging control.
- **Priority level:** This characteristics indicates a priority in scheduling resources among QoS flows, and it has following characteristics:
  - The lowest numeric value corresponds to the highest priority.
  - The priority level is used to differentiate between QoS flows of the same UE, and it is also used to differentiate between QoS flows from different UEs.
  - When there is congestion such that all QoS requirements cannot be fulfilled for one or more QoS flows, the priority level is used to select which QoS flows the QoS requirements are prioritized such that a QoS flow with priority level value *N* is prioritized over QoS flows with higher priority level values. For example, the priority level serves as a tie-breaker when two packets compete for a given network resource at the same time.
  - In the absence of congestion, the priority level is used to define the resource distribution between QoS flows. In addition, the scheduler may prioritize QoS flows based on other parameters (e.g., resource type, radio condition) in order to optimize application performance and network capacity.

- Every standardized 5QI is associated with a default value for the priority level, as specified in TS 23.501.
- Priority level may also be signaled together with a standardized 5QI to the AN, and if it is received, it is used instead of the default value.
- Priority level may also be signaled together with a preconfigured 5QI to the AN, and if it is received, it is used instead of the preconfigured value.

- **Packet delay budget (PDB):** This characteristic, which defines an upper bound for the time that a packet may be delayed between the UE and the UPF, has the following characteristics:
  - For some 5QI, the value of the PDB is the same in UL and DL.
  - In the case of 3GPP access, the PDB is used to support the configuration of scheduling and link layer functions (e.g., the setting of scheduling priority weights).
  - For GBR QoS flows using the delay-critical resource type, a packet delayed more than PDB is counted as lost if the data burst is not exceeding the MDBV within the period of PDB and the QoS flow is not exceeding the GFBR.
  - For GBR QoS flows with GBR resource type not exceeding GFBR, 98% of the packets do not experience delay exceeding the 5QI's PDB.
  - Services using a GBR QoS flow and sending at a rate smaller than or equal to the GFBR can in general assume that congestion-related packet drops will not occur.
  - Services using non-GBR QoS flows should be prepared to experience congestion-related packet drops and delays. In uncongested scenarios, 98% of the packets should not experience delay exceeding the 5QI's PDB.
  - PDB for non-GBR and GBR resource types denotes a "soft upper bound" in the sense that an "expired" packet (e.g., a link layer SDU that has exceeded the PDB) does not need to be discarded and is not added to the PER.
  - For a delay-critical GBR resource type, packets delayed more than the PDB are added to the PER and can be discarded or delivered depending on local decision.
- **Packet error rate (PER):** This characteristics defines an upper bound for the rate of PDUs (e.g., IP packets) that have been processed by the sender of a link layer protocol but that are not successfully delivered by the corresponding receiver to the upper layer. Equivalently, the PER defines an upper bound for a rate of non-congestion-related packet losses. For GBR QoS flows with delay-critical GBR resource type, a packet that is delayed more than PDB is counted as lost and is included in the PER unless the data burst exceeds the MDBV within the period of PDB or the QoS flow exceeds the GFBR.
- **Averaging window:** This characteristics represents the duration over which the GFBR and MFBR are calculated (e.g., in the RAN, UPF, and UE). Thus, the bit rate is calculated as B/W, where B is the number of bits transmitted during a window of size W seconds. Each GBR QoS

flow is associated with an averaging window. Every standardized 5QI (of GBR and delay-critical GBR resource types) is associated with a default value for the averaging window. The averaging window may also be signaled together with a standardized 5QI to the RAN and UPF, and if it is received, it is used instead of the default value. The averaging window may also be signaled together with a preconfigured 5QI to the RAN, and if it is received, it is used instead of the preconfigured value.

- **Maximum data burst volume (MDBV):** This characteristics denotes the largest amount of data that the 5G-AN is required to serve within a period of 5G-AN PDB (i.e., the 5G-AN part of the PDB). Each GBR QoS flow with a delay-critical resource type is associated with an MDBV. Every standardized 5QI of delay-critical GBR resource type is associated with a default value for the MDBV. The MDBV may also be signaled together with a standardized 5QI to the (R)AN, and if it is received, it is used instead of the default value. The MDBV may also be signaled together with a preconfigured 5QI to the (R)AN, and if it is received, it is used instead of the preconfigured value.

Standardized 5QI-to-QoS Characteristic Mapping

TS 23.501 provides a collection of predefined QoS profiles with associated 5QI values. These standardized 5G QoS identifier (5QI) values correspond to services that are likely to be frequently used in 5G networks and that would thus benefit from optimized signaling through the use of standardized QoS characteristics. From the user perspective, the standardized QoS profiles relieve the user from the necessity of designing a set of QoS characteristic values for services that are common. Nonstandardized 5QIs can be used in an operator network or by agreement between two or more operators.

For each of the 5QI entries, TS 23.501 lists examples of applications. These include a variety of traffic classes, which are as follows:

- **Conversational:** This interactive service provides for bidirectional communication by means of real-time (no store-and-forward) end-to-end information transfer from user to user. Examples of such flows include telephony speech and also VoIP and video conferencing. Sensitivity to delay is high because of the real-time nature of the flows. The time relationship between the stream entities has to be preserved (to maintain the same experience for all flows and all parties involved in the conversation).
- **Streaming:** The streaming class refers to flows in which the user is watching real-time video or listening to real-time audio (or both). The real-time data flow is always aiming at a live (human) destination. Streaming is both a real-time flow and a one-way transport. The delay sensitivity is lower than that of conversational flows because it is expected that the receiving end includes a time-alignment function (e.g., buffering). Because the flow is unidirectional, variations in delay do not adversely affect the user experience as long as the variation is within the alignment function boundaries.

- **Mission critical:** The failure or disruption of this type of service would result in serious damage to the users of the service. Mission-critical communications are secure, reliable, and readily available, and typically time is a vital factor.

The 5QI values fall into three groupings for the three resource types (i.e., GBR, delay-critical GBR, and non-GBR). Table 9.3 shows the values of characteristics for various QoS profiles of the GBR resource type and suggests examples of applications for each 5QI entry.

TABLE 9.3 Standardized 5QI to QoS Characteristics Mapping for GBR Resource Type*

5QI	Default Priority	PDB (ms)	PER	Default MDBV	Default Averaging Window	Examples of Services
1	20	100	10^{-2}	N/A	2 s	Conversational voice
2	40	150	10^{-3}	N/A	2 s	Conversational video
3	30	50	10^{-3}	N/A	2 s	Real-time gaming V2X messages Electricity distribution (medium voltage) Process automation monitoring
4	50	300	10^{-6}	N/A	2 s	Non-conversational video (buffered streaming)
65	7	75	10^{-2}	N/A	2 s	Mission-critical user plane push-to-talk voice
66	20	100	10^{-2}	N/A	2 s	Non-mission-critical user plane push-to-talk voice
67	15	100	10^{-3}	N/A	2 s	Mission-critical user plane video
71	56	150	10^{-6}	N/A	2 s	Live uplink streaming
72	56	300	10^{-4}	N/A	2 s	
73	56	300	10^{-8}	N/A	2 s	
74	56	500	10^{-6}	N/A	2 s	
76	56	500	10^{-4}	N/A	2 s	

**PDB = Packet Delay Budget; PER = Packet Error Rate; MDBV = Maximum Data Burst Volume*

The profiles show support for, among others, the following:

- Smart grid
- Process automation monitoring
- Autonomous vehicles (V2X messaging)
- Mission-critical public safety applications
- Low-latency enhanced mobile broadband (eMBB)

- Augmented reality
- Discrete automation
- Intelligent transport systems

Several observations should be made. The two mission-critical applications in Table 9.3 have significantly higher priority (i.e., lower priority number) than the other applications. Push-to-talk (PTT), also known as press-to-transmit, is a method of having conversations or talking on half-duplex communication lines, including two-way radio, using a momentary button to switch from voice reception mode to transmit mode. For example, the PTT feature on Zoom allows a user to remain muted throughout a Zoom meeting and hold down the spacebar to unmute and talk. Both mission-critical and non-mission-critical PTT have relatively high priority, though the mission-critical priority is higher.

Live uplink streaming is a service in which a user with a radio connection (e.g., reporter, drone) streams a live video feed into the network or to a second party. TS 23.501 defines five different characteristic value combinations for this service (5QI values 71, 72, 73, 74, and 76). TR 26.939 (*Technical Specification Group Services and System Aspects, Guidelines on the Framework for Live Uplink Streaming [FLUS] [Release 16]*, September 2019) defines eight different uses cases for live uplink streaming and provides guidance on the choice of QoS parameters and characteristics for each use case. The five QoS profiles in Table 9.3 support the likely choices to be made for the eight use cases.

Table 9.4 shows the values of characteristics for various QoS profiles of the delay-critical GBR resource type. All of these application examples are assigned relatively high priority.

TABLE 9.4 Standardized 5QI-to-QoS Characteristics Mapping for the Delay-Critical GBR Resource Type*

5QI	Default Priority	PDB (ms)	PER	Default MDBV	Default Averaging Window	Examples of Services
82	19	10	10^{-4}	255 bytes	2 s	Discrete automation
83	22	10	10^{-4}	1354 bytes	2 s	Discrete automation V2X messages (cooperative lane change)
84	24	30	10^{-5}	1354 bytes	2 s	Intelligent transport systems
85	21	5	10^{-5}	255 bytes	2 s	Electricity distribution (high voltage) V2X messages (remote driving)
86	18	5	10^{-4}	1354 bytes	2 s	V2X messages (advanced driving, collision avoidance)

**PDB = Packet Delay Budget; PER = Packet Error Rate; MDBV = Maximum Data Burst Volume*

Table 9.5 shows the values of characteristics for various QoS profiles of the non-GBR resource type. 5QI 70 is non-GBR, intended for mission-critical data, with a priority of 55, a PDB of 200 ms, and a

PER tolerance of at most 10^{-6}. The traffic types intended for 5QI 70 are the same as for 5QIs 6, 8, and 9: buffered streaming video and TCP-based traffic, such as www, email, chat, FTP, P2P, and other file sharing applications. However, 5QI 70 is specifically intended for applications that are mission critical. For this reason, 5QI 70 priority is higher than 6, 8, or 9 priorities (55 versus 60, 80, and 90, respectively).

TABLE 9.5 Standardized 5QI-to-QoS Characteristics Mapping for the Non-GBR Resource Type*

5QI	Default Priority	PDB (ms)	PER	Default MDBV	Default Averaging Window	Examples of Services
5	10	100	10^{-6}	N/A	N/A	IMS (IP multimedia subsystem) signaling
6	60	300	10^{-6}	N/A	N/A	Video (buffered streaming) TCP based (e.g., www, email, chat, FTP, P2P file sharing)
7	70	100	10^{-3}	N/A	N/A	Voice Video (live streaming) Interactive gaming
8	80	300	10^{-6}	N/A	N/A	Video (buffered streaming) TCP based (e.g., www, email, chat, FTP, P2P file sharing)
9	90	300	10^{-6}	N/A	N/A	Video (buffered streaming) TCP based (e.g., www, email, chat, FTP, P2P file sharing)
69	5	60	10^{-6}	N/A	N/A	Mission-critical delay-sensitive signaling
70	55	200	10^{-6}	N/A	N/A	Mission-critical data
79	65	50	10^{-2}	N/A	N/A	V2X messages
80	68	10	10^{-6}	N/A	N/A	Low-latency eMBB applications AR

**PDB = Packet Delay Budget; PER = Packet Error Rate; MDBV = Maximum Data Burst Volume*

Tables 9.3 through 9.5 provide some insight into the distinction between QoS parameters and QoS characteristics. Although there is some overlap, very broadly it can be said that the QoS parameters are used at configuration time to determine the network resources needed for creating a network slice for supporting this set of QoS parameter values. The QoS characteristics are more relevant to dynamic decisions made during the operation of the QoS flow, such as using the priority level as a tie-breaker when two flows compete for a resource.

Another perspective is based on the distinction between an application and a specific instance of an application. 3GPP has determined that the variables defined as characteristics are typically the same for a wide variety of instances of an application, and so it is efficient and useful to the user to provide standardized sets of values. Depending on the context, individual instances of an application may require different sets of values for some of the parameters, especially the flow bit rates (GFBR and MFBR) and the aggregate bit rates (session-AMBR and UE-AMBR).

9.4 Network Slicing

One of the most important features of 5G is network slicing. Network slicing involves virtualization technologies such as SDN and NFV, which enable a 5G network operator to provide customized networks by creating multiple virtual end-to-end networks, referred to as network slices. Each network slice can be defined according to different requirements on functionality, QoS, and specific users. Table 9.6 defines the network slicing terms used in 3GPP documents.

TABLE 9.6 3GPP Network Slicing Terminology

Term	Definition
Network slice	A logical network that provides specific network capabilities and network characteristics.
Network slice instance	A set of network function instances and the required resources (e.g., compute, storage, and networking resources) that form a deployed network slice.
Network slice instance identifier (NSI ID)	An identifier for identifying the core network part of a network slice instance when multiple network slice instances of the same network slice are deployed and there is a need to differentiate between them in the 5GC.
NF instance	An identifiable instance of the NF.
NF set	A group of interchangeable NF instances of the same type, supporting the same services and the same network slice(s). The NF instances in the same NF set may be geographically distributed but have access to the same context data.
Slice/service type (SST)	The expected network slice behavior in terms of features and services.
Slice differentiator (SD)	Optional information that complements the SST(s) to differentiate among multiple network slices of the same SST.
Single network slice selection assistance information (S-NSSAI)	Defines a unique network slice, composed of an SST and an SD.
Network slice selection assistance information (NSSAI)	A vector of up to eight S-NSSAI values that are used to identify and select slice instances associated with a UE.
Network slice selection function (NSSF)	A network function that selects the set of network slice instances to serve the UE. It also determines the allowed and configured NSSAI and, if necessary, maps to the subscribed S-NSSAIs and determines the AMF set to be used (or a candidate list) to serve the UE.
Network slice selection policy (NSSP)	A set of rules (at least one rule), where each rule attributes an application with a certain S-NSSAI. This is used by the UE to associate the matching application with the S-NSSAI.
Network slice-specific authentication and authorization (NSSAA)	A network function that allows a third party to add and remove users to and from a network slice instance.

[LI17] lists the following advantages of slice-based networking compared with traditional networking:

- Network slicing can provide logical networks with better performance than one-size-fits-all networks.
- A network slice can scale up or down as service requirements and the number of users change.
- Network slices can isolate the network resources of one service from the others; the configurations among various slices don't affect each other. Therefore, the reliability and security of each slice can be enhanced.
- A network slice is customized according to QoS requirements, which can optimize the allocation and use of physical network resources.

Network slicing is made possible by the softwarization techniques of network functions virtualization (NFV) and software-defined networking (SDN). NFV implements the network functions (NFs) in a network slice, enabling the isolation of each network slice from all other network slices. Isolation can be achieved by one or more of the following: (1) using a different physical resource, (2) separation by virtualization, which may allow sharing of physical resources, or (3) through sharing a resource with the guidance of a respective policy that defines the access rights for each tenant. Isolation assures QoS and security requirements for that slice, independent of other slices operating on the network from the same or different users.

Once a network slice is defined, SDN operates to monitor and enforce QoS requirements by controlling the behavior of the QoS flow for each slice.

Network Slicing Concepts

Network slicing permits a physical network to be separated into multiple virtual networks (logical segments) that can support different radio access networks or several types of services for certain customer segments, greatly reducing network construction costs by using communication channels more efficiently. In essence, network slicing allows the creation of multiple virtual networks atop a shared physical infrastructure. This virtualized network scenario devotes capacity to certain purposes dynamically, according to need. As needs change, so can the devoted resources. Using common resources such as storage and processors, network slicing permits the creation of slices devoted to logical, self-contained, and partitioned network functions. Network slicing supports the creation of virtual networks to provide a given level of QoS, such as guaranteed delay, throughput, reliability, and/or priority.

A network slice creates a partition of the core network consisting of virtualized network functions and resources running on some of the core network hardware resources. Figure 9.14, based on concepts in the Next Generation Mobile Networks (NGMN) document *5G End-to-End Architecture Framework* (August 2019), illustrates network slicing concepts. It shows a simple core network configuration composed of three types of devices.

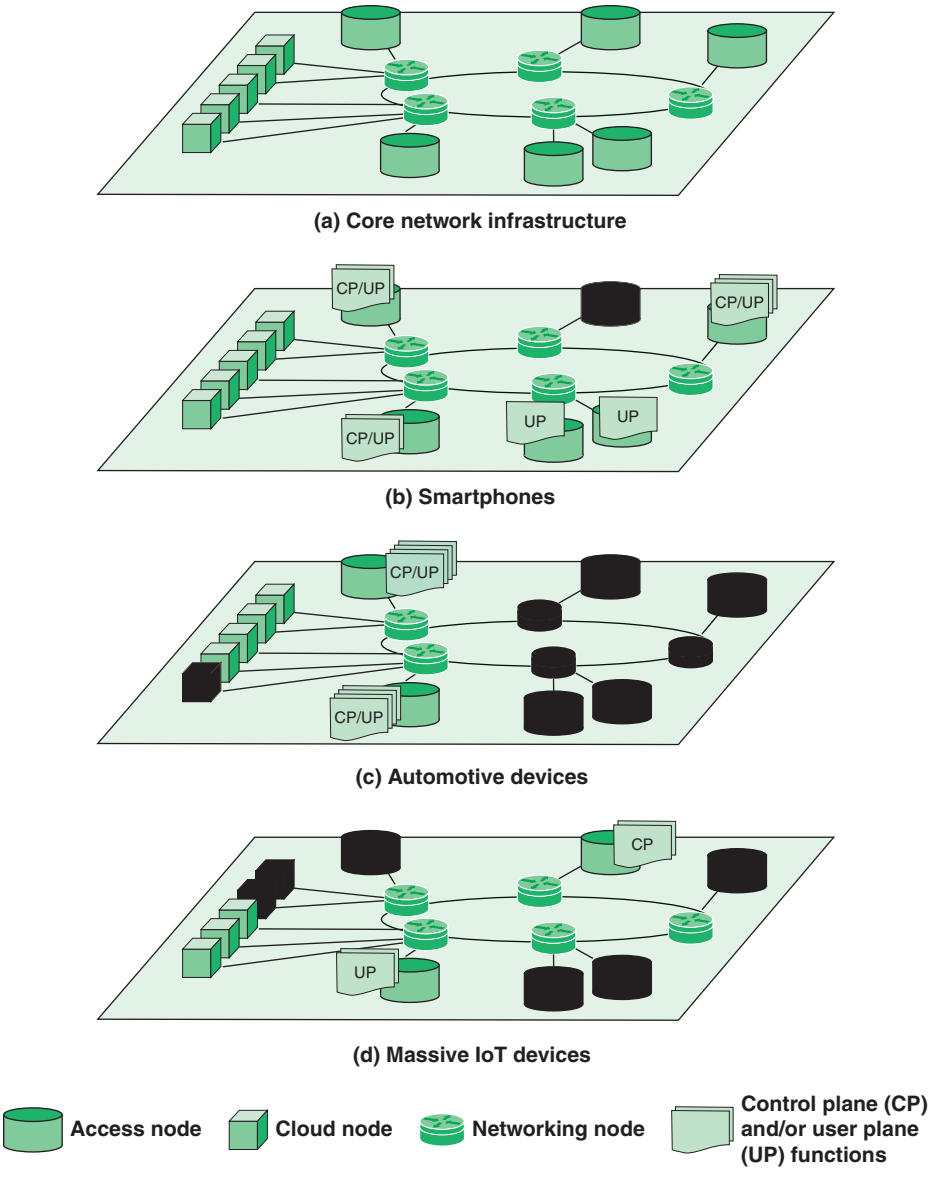


FIGURE 9.14 5G Network Slices Implemented on the Same Infrastructure

The devices are as follows:

- **Cloud nodes:** These nodes provide cloud services, software, and storage resources. There are likely to be one or more central cloud nodes that provide traditional cloud computing service. In addition, cloud-edge nodes provide low-latency and higher-security access to client devices

at the edge of the network. All of these nodes include virtualization system software to support virtual machines and containers. NFV enables effective deployment of cloud resources to the appropriate edge node for a given application and given fixed or mobile user. The combination of SDN and NFV enables the movement of edge resources and services to dynamically accommodate mobile users.

- **Networking nodes:** These nodes are IP routers and other types of switches for implementing a physical path through the network for a 5G connection. SDN provides for flexible and dynamic creation and management of these paths.
- **Access nodes:** These nodes provide an interface to radio access networks (RANs), which in turn provide access to mobile UE. SDN creates paths that use an access node for one or both ends of a connection involving a wireless device.

The remainder of Figure 9.14 illustrates three use cases. The blacked-out core network resources represent resources not used to create the network slice. Cloud nodes that are part of the slice may include the following:

- Control plane functions associated with one or more user plane functions (e.g., a reusable or common framework of control)
- Service- or service category–specific control plane and user plane function pairs (e.g., a user-specific multimedia application session)

The first network slice depicted in Figure 9.14 is for a typical smartphone use case. Such a slice might have fully fledged functions distributed across the network. The second network slice in Figure 9.14 indicates the type of support that may be allocated for automobiles in motion. This use case emphasizes the need for security, reliability, and low latency. A configuration to achieve this would limit core network resources to nearby cloud edge nodes,[1] plus the recruitment of sufficient access nodes to support the use case. The final use case illustrated in Figure 9.14 is for a massive IoT deployment, such as a huge number of sensors. The slice can contain just some specific CP and UP functions with, for example, no mobility functions. The CP and UP functions might include filtering and preliminary data analysis at the edge and big data types of analysis at a more central node. This slice would only need to engage access nodes nearest to the IoT device deployment.

Requirements for Network Slicing

TS 22.261 lists requirements for network slicing in two categories: general requirements and management requirements.

1. Cloud-edge computing, or cloud-edge networking, refers to the deployment of cloud capabilities at the network edge. Chapter 10 explores this topic.

General Requirements

The general requirements for network slicing are as follows:

- Support is needed to provide connectivity to home and roaming users in the same network slice.
- In a shared 5G network configuration, operators can apply all the requirements to their allocated network resources.
- IMS needs to be supported as part of a network slice.
- IMS needs to be supported independent of network slices.

IP Multimedia Subsystem (IMS) is a standards-based architectural framework for delivering multimedia communications services such as voice, video, and text messaging over IP networks [KOUK06]. The IMS specifications were originally developed by 3GPP in the early 2000s to standardize access to multimedia services using cellular networks. The specifications define a complete framework and architecture that enables the convergence of video, voice, data, and mobile network technologies.

Management Requirements

The management requirements for network slicing are as follows:

- The operator should be able to create, modify, and delete a network slice.
- The operator should be able to define and update the set of services and capabilities supported in a network slice.
- The operator should be able to configure the information that associates UE to a network slice.
- The operator should be able to configure the information that associates a service to a network slice.
- The operator should be able to assign UE to a network slice, to move UE from one network slice to another, and to remove UE from a network slice based on subscription, UE capabilities, the access technology being used by the UE, the operator's policies, and services provided by the network slice.
- A mechanism is needed for the VPLMN (visited public land mobile network), as authorized by the HPLMN (home public land mobile network), to assign UE to a network slice with the needed services or to a default network slice.
- A UE should be able to be simultaneously assigned to and access services from more than one network slice of one operator.
- Traffic and services in one network slice should have no impact on traffic and services in other network slices in the same network.
- Creation, modification, and deletion of a network slice should have no or minimal impact on traffic and services in other network slices in the same network.

- A network slice should be able to scale (i.e., adapt its capacity).
- The network operator should be able to define a minimum available capacity for a network slice. Scaling of other network slices on the same network should have no impact on the availability of the minimum capacity for that network slice.
- The network operator should be able to define a maximum capacity for a network slice.
- The network operator should be able to define a priority order between different network slices in the event that multiple network slices compete for resources on the same network.
- The operator should be able to differentiate policy control, functionality, and performance provided in different network slices.

Identifying and Selecting a Network Slice

Single network slice selection assistance information (S-NSSAI) defines a single network slice. An S-NSSAI consists of two elements:

- **Slice/service type (SST):** An identifier that refers to the expected slice behavior in terms of features and services. Standardized SST values provide a way to establish global interoperability for slicing so that 5G networks can support the roaming use case more efficiently for the most commonly used SSTs. Table 9.7 lists the standardized SSTs.
- **Slice differentiator (SD):** Optional information that complements the SST to differentiate among multiple network slices of the same SST.

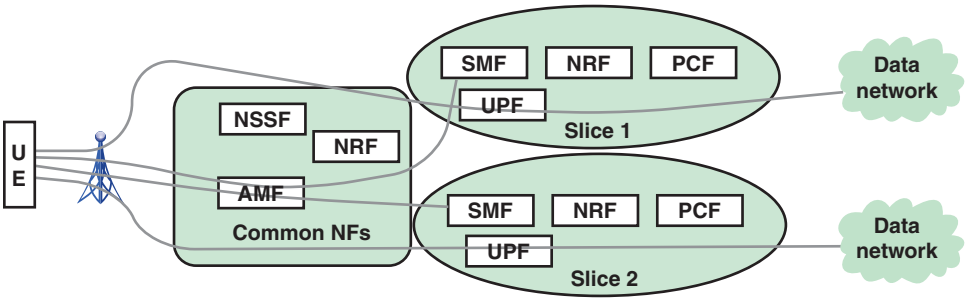
TABLE 9.7 Standardized Slice/Service Type Values

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 and low-latency communications
MIoT	3	Slice suitable for the handling of massive IoT
V2X	4	Slice suitable for the handling of V2X services

UE may be served by up to eight network slices at a time, each identified by an S-NSSAI. The set of S-NSSAIs associated with a UE form a network slice selection assistance information (NSSAI) data object.

Functional Aspects of Network Slicing

Figure 9.15 illustrates how core network functions (NFs) are used to implement network slices. Some NF instances support multiple network slices serving UE, and others are specific to a given slice.



AMF = Access and Mobility Management Function
NF = Network Function
NRF = Network Repository Function
NSSF = Network Slice Selection Function
PCF = Policy Control Function
SMF = Session Management Function
UPF = User Plane Function

FIGURE 9.15 Network Functions That Support Network Slicing

The common NFs are:

- **Access and mobility management function (AMF):** Network slice instance selection is usually triggered as part of the registration procedure by the first AMF that receives the registration request from the UE. When UE accesses the network, the AMF provides functionalities to register and de-register the UE with the network, and it establishes the user context in the network. In the registration procedure, AMF performs (but is not limited to) network slice instance selection, UE authentication, authorization of network access and network services, and network access policy control. In addition, when a session establishment request message is received from UE, the AMF performs discovery and selection of the SMF that is the most appropriate to manage the session.
- **Network slice selection function (NSSF):** The AMF retrieves the slices that are allowed by the user subscription and interacts with the NSSF to select the appropriate network slice instance (e.g., based on allowed S-NSSAIs, 5G network ID, and other parameters). The NSSF responds with a message that includes the list of appropriate network slice instances for the UE. As a result, the registration process may switch to another AMF if needed.
- **Network repository function (NRF):** During the AMF-NSSF interaction, the NSSF may return the identity of one or more NRFs to be used to select NFs and services within the selected network slice instance(s).

The slice-specific NFs are:

- **Session management function (SMF):** The UE sends a message to the AMF, requesting that a PDU session be associated to one S-NSSAI and one data network (DN). The AMF selects the appropriate SMF, which manages the PDU session. The SMF sets up the PDU session for the UE and controls the user plane operation. The SMF selects the UPF and invokes enforcement of QoS and charging policies.
- **User plane function (UPF):** Once a PDU session is established, QoS flows for this PDU session over this network slice pass through the UPF.
- **Policy control function (PCF):** The SMF gets policy information related to session establishment from the PCF.
- **Network repository function (NRF):** The SMF uses the NRF to discover the required NFs for the individual network slice.

Generic Slice Template

3GPP TS 28.531 (*Technical Specification Group Services and System Aspects, Management and orchestration, Provisioning [Release 16]*, December 2020) includes a description of the Generic Network Slice Template (GST) concept, which is specified by the GSM Association (GSMA). The GST provides a standardized list of attributes that can be used to characterize different types of network slices [GSMA20]. A network slice type (NEST) is a GST filled with (ranges of) values. There may be two kinds of NESTs:

- **Standardized NEST (S-NEST):** Attributes are assigned (ranges of) values by standards-developing organization (SDOs), working groups, forums, and so forth, such as 3GPP, GSMA, 5G Automotive Association (5GAA), and 5G Alliance for Connected Industry and Automation (5G-ACIA).
- **Private NEST (P-NEST):** Attributes are assigned (ranges of) values by the network slice providers that are different from those assigned in S-NESTs.

Network slice providers can build their network slice product offerings based on S-NESTs and/or P-NESTs.

GSMA has developed the GST to be a list of attributes sufficient for describing a wide range of NESTs that can be fully constructed by allocating values (or ranges of values) to each relevant attribute in the GST. A network operator can use a NEST to identify the network resources and

functions needed to instantiate network slices. The process to fill in the GST and to create a NEST involves three steps:

Step 1. Study use cases and derive service requirements based on discussions with the slice customers, such as vertical industries or specific enterprises.

Step 2. Convert the service requirements identified in step 1 into technical requirements.

Step 3. Document the technical requirements produced in step 2 using the NEST by filling in the values of each of the attributes of the GST.

The current version of the GST lists 35 attributes, shown in Figure 9.16.

Availability	Network functions owned by network slice customer	Support for non-IP traffic
Area of service	Maximum number of PDU sessions	Supported device velocity
Delay tolerance	Maximum number of UEs	Synchronicity
Deterministic communication	Performance monitoring	UE density
Downlink throughput per network slice	Performance prediction	Uplink throughput per network slice
Downlink maximum throughput per UE	Positioning support	Uplink maximum throughput per UE
Energy efficiency	Radio spectrum	User management openness
Group communication support	Root cause investigation	User data access
Isolation level	Session and service continuity support	V2X communication mode
Maximum supported packet size	Simultaneous use of the network slice	Latency from (last) UFP to application server
Mission-critical support	Slice quality of service parameters	Network Slice Specific Authentication and Authorization (NSSAA) required
MMTel support		
NB-IoT support		

FIGURE 9.16 Generic Network Slice Template Attributes

9.5 SDN and NFV Support for 5G

The 5G Infrastructure Public Private Partnership refers to SDN and NFV as the fundamental pillars that support the wide range of key performance indicators (KPIs) for the new 5G use cases in a cost-efficient way [5GPP20]. With the SDN and NFV framework of 5G, mobile network operators are able to offer new services to consumers, enterprises, verticals, and third-party tenants by addressing their respective requirements.

The European Telecommunications Standards Institute document ETSI GS NFV-EVE 005 (*Network Functions Virtualisation [NFV], Ecosystem, Report on SDN Usage in NFV Architectural Framework*, December 2015) examines the manner in which SDN can be incorporated in the NFVI to provide connectivity services. Figure 9.17, based on a figure in GS NFV-EVE 005, illustrates the placement of SDN controllers to achieve appropriate cooperation between SDN and NFV. The framework

incorporates two controllers: one logically placed at the tenant level and another at the NFVI level. Each controller centralizes the control plane functionalities and provides an abstract view of all the connectivity-related components it manages.

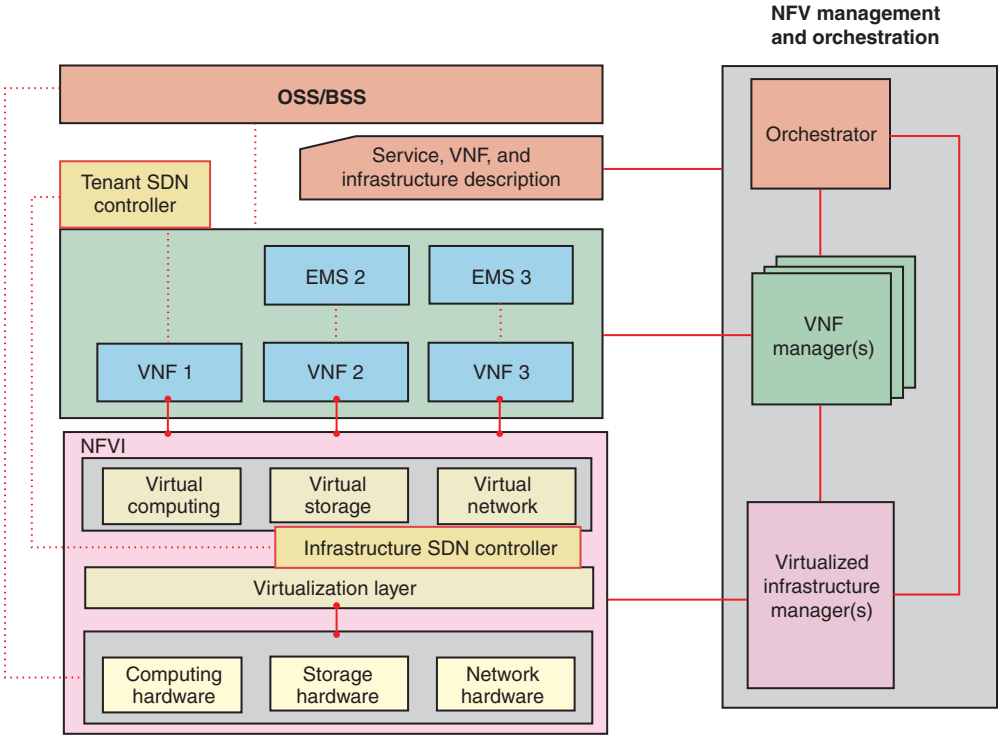


FIGURE 9.17 Integrating SDN Controllers into the Reference NFV Architectural Framework

The controllers are as follows:

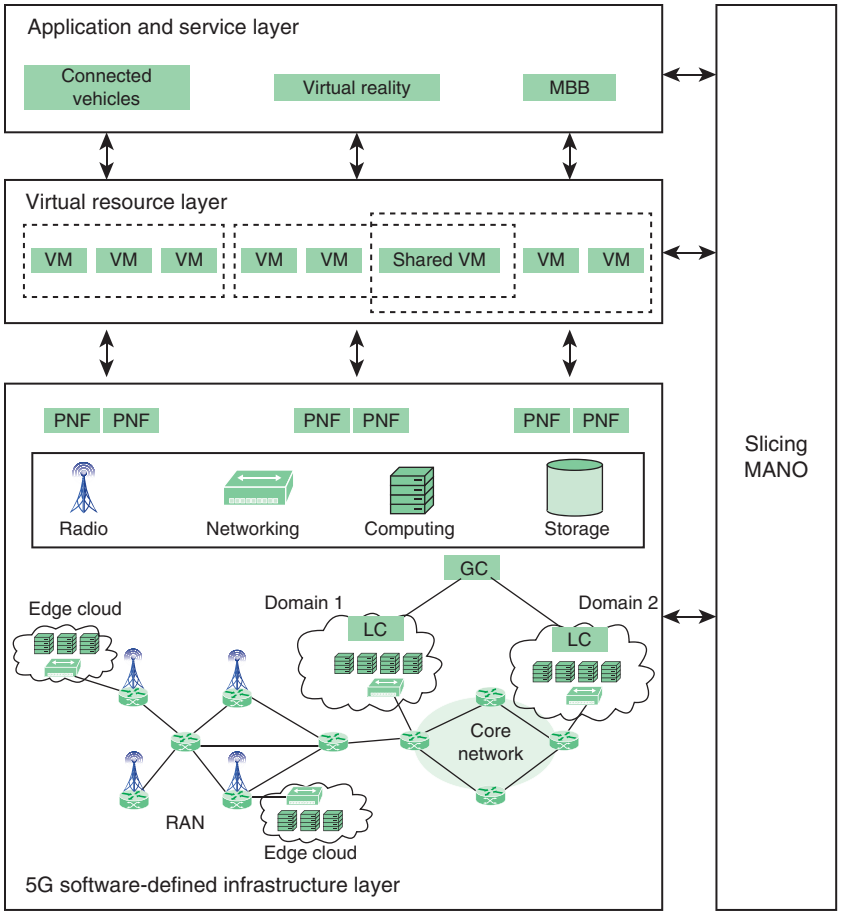
- **Infrastructure SDN controller (IC):** This controller enables communication among VNFs and among their components, including the cases when those VNFs are instantiated in separated PoPs, reachable through a WAN connection. Managed by the VIM, this controller may change infrastructure behavior on demand according to VIM specifications, adapted from tenant requests.

- **Tenant SDN controller (TC):** Instantiated in the tenant domain as one of the VNFs or as part of the NMS, this second controller dynamically manages the pertinent VNFs used to realize the tenant's network service(s). These VNFs are the underlying forwarding plane resources of the TC. The operation and management tasks that the TC carries out are triggered by the applications running on top of it (e.g., the OSS).

[ORDO17] describes the manner in which the architecture of Figure 9.17 supports network slicing. The two controllers manage and control their underlying resources via programmable southbound interfaces, implementing protocols such as OpenFlow. The two controllers provide different levels of abstraction. The IC provides an underlay to support the deployment and connectivity of VNFs, and the TC provides an overlay comprising tenant VNFs that, properly composed, define the network service(s) that such a tenant independently manages on its slice(s). The IC is not aware of the number of slices that utilize the VNFs it connects or the tenant or tenants that operate such slices. For the TC, the network is abstracted in terms of VNFs, without notions of how those VNFs are physically deployed. Despite their different abstraction levels, both controllers have to coordinate and synchronize their actions.

Figure 9.18, from [LI17], provides a more concrete depiction of how SDN and NFV cooperate to support 5G network slices. The framework is constructed as three layers:

- **Application and service layer:** This layer contains a heterogenous collection of service instances. The example of Figure 9.17 shows three instances: connected vehicles, virtual reality, and mobile broadband (MBB). A service instance may serve multiple tenants or users. The service instance informs the slicing MANO module, described later in this section, of its service requirements, which are then mapped to a network slice.
- **Virtual resource layer:** This layer provides all the virtual resources required for network slices, such as radio, computing, storage, and network bandwidth. These resources are provided as virtual network functions residing on virtual machines (VMs). In this example, the three VMs on the left support the connected vehicles application. The next two VMs plus a shared VM support the virtual reality application. The shared VM and the last two VMs support the MBB application.
- **Software-defined infrastructure layer:** This layer is composed of a software-defined infrastructure with software-based control and management encompassing multiple SDN domains and cloud-edge networks. Each SDN domain has a local controller. There is a global SDN controller to coordinate the local controllers.



MANO = Management and Orchestration
MBB = Mobile Broadband
PNF = Physical Network Function
VM = Virtual Machine
GC = SDN Global Controller
LC = SDN Local Controller

FIGURE 9.18 5G Network Slicing Framework

The final component of this architecture is the slicing MANO, which is essentially an enhanced NFV MANO that manages NFV, SDN, and slicing resources. Its tasks include:

- Creating and managing VM instances by using the infrastructure resources
- Mapping network functions to virtual resources and connecting network functions to create service chains
- Managing the life cycle of network slices by interacting with the application and service layer (e.g., automated creation of service-oriented slices, dynamic maintenance by monitoring service requirements and virtual resources)

9.6 Key Terms and Review Questions

Key Terms

admission control	QoS parameters
broadcast	QoS routing
congestion avoidance	queuing
conversational	quality of service (QoS)
core network	resource reservation
mission critical	scheduling
multicast	service-level agreement (SLA)
network slice	service data flow (SDF)
network slicing	streaming
packet marking	traffic classification
PDU session	traffic policing
policy	traffic restoration
policy control	traffic shaping
priority	tunnel
QoS characteristics	tunneling
QoS differentiation	user plane marking
QoS flow	

Review Questions

1. List and define the IMT-2020 network operational requirements.
2. List and define the basic network requirements for 5G specified by 3GPP.
3. What are the differences between priority, QoS, and policy control?
4. What is 5G tunneling?
5. Define the two types of tunnels and explain the purpose of each of them.
6. What are the differences among PDU session, QoS flow, and service data flow?

7. What are the main purposes of QoS capabilities?
8. What are the four stages in the QoS life cycle?
9. Define the common functions used to provide QoS in the data plane.
10. Define the common functions used to provide QoS in the management plane.
11. What are the differences between QoS classification, marking, and differentiation?
12. List and briefly define the QoS parameters in the TS 23.501 model.
13. List and briefly define the QoS characteristics in the TS 23.501 model.
14. What is the difference between QoS parameters and characteristics?
15. List and define the general requirements for network slicing specified by 3GPP.
16. List and define the management requirements for network slicing specified by 3GPP.
17. What NFs support multiple slices for UE?
18. What NFs support a single slice instance?

9.7 References and Documents

References

5GPP20 5G Infrastructure Public Private Partnership. *View on 5G Architecture*. 5G PPP Architecture Working Group white paper, February 2020.

GSMA20 GSM Association. *Generic Network Slice Template Version 3.0.* May 2020.

KOUK06 Koukal, M., and Bestak, R. "Architecture of IP Multimedia Subsystem." *Proceedings ELMAR Symposium*, June 2006.

LI17 Li, X., et al. "Network Slicing for 5G: Challenges and Opportunities." *IEEE Internet Computing*, September/October 2017.

ORDO17 Ordonez-Lucena, J., et al. "Network Slicing for 5G with SDN/NFV: Concepts, Architectures and Challenges." *IEEE Communications Magazine*, May 2017.

Documents

3GPP TR 22.891 *Technical Specification Group Services and System Aspects, Feasibility Study on New Services and Markets Technology Enablers, Stage 1 (Release 14)*. September 2016.

3GPP TR 26.939 *Technical Specification Group Services and System Aspects, Guidelines on the Framework for Live Uplink Streaming (FLUS) (Release 16)*. September 2019.

3GPP TS 22.261 *Technical Specification Group Services and System Aspects, Service requirements for the 5G system, Stage 1 (Release 17)*. December 2020.

3GPP TS 23.501 *Technical Specification Group Services and System Aspects, System architecture for the 5G System (5GS), Stage 2 (Release 16).* December 2020.

3GPP TS 23.502 *Technical Specification Group Services and System Aspects, Procedures for the 5G System (5GS), Stage 2 (Release 16)*. December 2020.

3GPP TS 23.503 *Technical Specification Group Services and System Aspects, Policy and charging control framework for the 5G System (5GS), Stage 2 (Release 16)*. April 2020.

3GPP TS 28.531 *Technical Specification Group Services and System Aspects, Management and orchestration, Provisioning (Release 16)*. December 2020.

3GPP TS 38.300 *Technical Specification Group Radio Access Network, NR, NR and NG-RAN Overall Description, Stage 2 (Release 16).* September 2020.

5G PPP *View on 5G Architecture*. Version 3.0. June 2019.

ETSI GS NFV-EVE 005 *Network Functions Virtualisation (NFV), Ecosystem, Report on SDN Usage in NFV Architectural Framework*. December 2015.

ITU-T Y.1291 *An Architectural Framework for Support of Quality of Service in Packet Networks*. May 2004.

ITU-T Y.3101 *Requirements of the IMT-2020 Network*. April 2018.

ITU-T Y.3106 *Quality of Service Functional Requirements for the IMT-2020 Network*. April 2019.

NGMN *5G End-to-End Architecture Framework*. August 2019.