Session Mangement

DD Owned by David Cherney DISH •••
Last updated: Aug 01, 2023 • Add Workflow

Contents

Session management refers to the management of PDU sessions. Recall that a PDU session has the following structure.

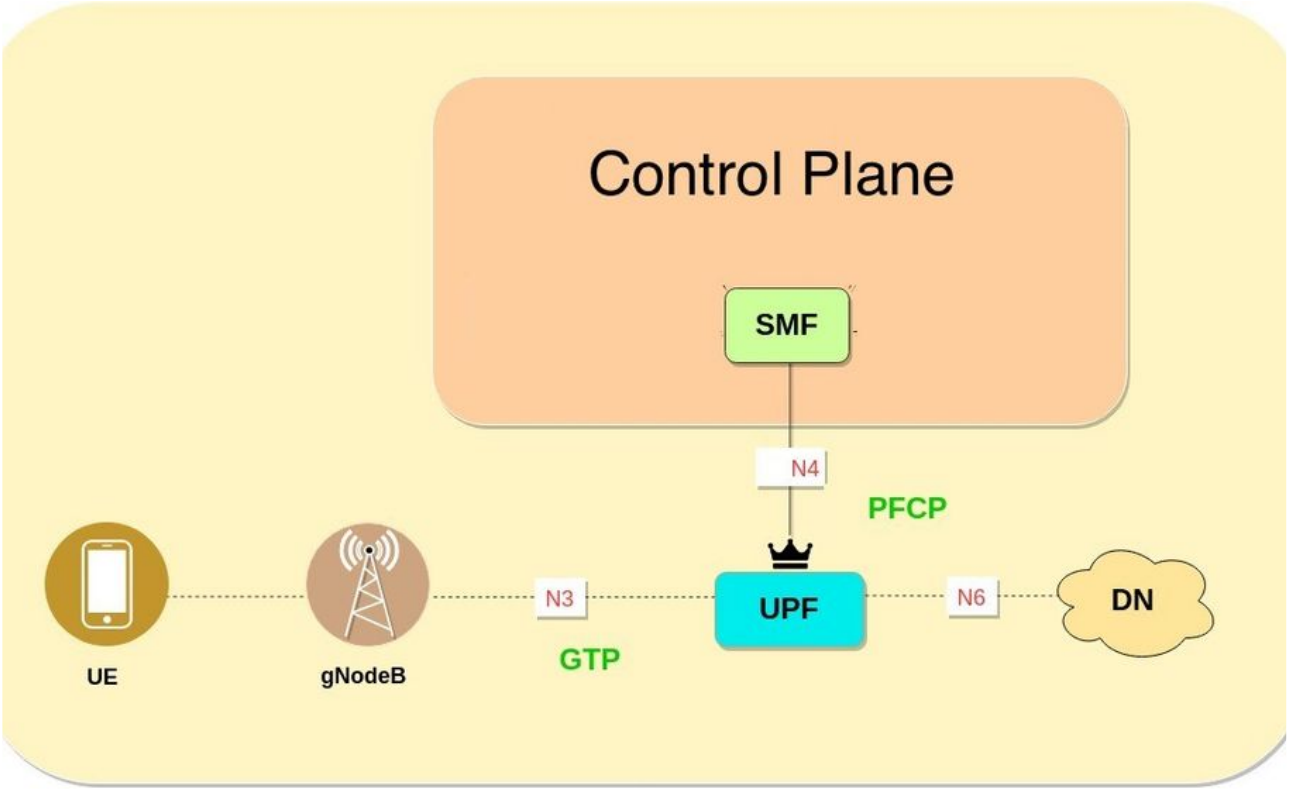
- Grouping together of
  - packets into service data flows (SDFs),
  - SDFs into QoS flows,
  - QoS flows into PDU sessions,
- Differentiated treatment of packets to
  - ensure QoS
  - measure usage at the SDF level
    - charging
    - informing session management decisions

The UPF is the core NF involved in the structure above. The UPF is able to do its work when it has been given the right tools. Those tools are called packet detection rules, PDR. All of the functioning above and more is done in the UPF with PDRs.

Management of PDU sessions is handled by the SMF

The role of the SMF in the *establishment* of a PDU session is to choose a UPF instance to serve as PDU session anchor (PSA), to establish a persistent connection to that UPF called a PFCP session (packet forwarding control protocol), and to give the chosen UPF a set of PDRs through that connection.

The role of the SMF in *running* PDU sessions is to create, update, and delete PDRs in the UPF.



These roles, creating and running, constitute session management.

Management vs Control

The reasons for changing the PDRs in a UPF will be discussed under a separate section named Policy and Charging Control, PCC. Policy control over sessions constitutes a level or control above session management; Policy control is not in the scope of this page. In the rest of this page, we will discuss the ways that PDRs inserted by the SMF into the UPF allow session management .

QoS Systems

Mechanisms within the 5G core for responding to (PDU) service requests with control of flow of packets are divided into 3 kinds for 3 planes; control plane, data plane, and management plane.

Data Plane Mechanisms

These mechanisms act directly on packets.

- traffic classification: a classification entity inspects packets (source, destination, payload, QoS markings) and determines the class to which a packet belongs so that different classes can be treated differently
- packet marking: IPv4 already has ToS (type of service) or DiffServ (differentiated services) field in the IPv4 header, and IPv6 has Traffic Class field in the IPv6 header. These can be used to generate GTP header markings to be used in the 5GC upon ingress. Low priority or non-conforming packets can be dropped when congestion needs to be reduced.
- traffic shaping: an entity can buffer where needed to bring the QoS for the highest priority flows up to SLA levels; high priority traffic can be made more predictable and less bursty with this mechanism.
- congestion avoidance: packet loss and “timer expiration” are typically considered indicators of network congestion. To avoid congestion collapse or significant queueing delays, senders can be made to reduce the amount of traffic entering a network upon such indicators.
- traffic policing: intentional dropping or delaying of some packets.
- queueing: determining which packages to send next
- active queueing: determining which packages to send next in order to avoid congestion.

Control Plane Mechanisms

The control plane is where pathways for user data flows are created and maintained; these mechanisms act on those pathways.

- admission control: A traffic flow can be denied based on identity of users or applications, bandwidth requirements of the flow relative to current commitments, security considerations, or time of the day/week etc.
- QoS routing: instead of looking for the least cost path through the CN, look for paths that will meet QoS.
- resource reservation: Allow only certain data flows through certain resources. e.g. a reserved system of UPFs for a URLLC application like remote surgery requiring <1ms latency might provide the possible lowest possible latency.

Management Plane Mechanisms

The management plane is the intersection of the control and data planes; these mechanisms act on both pathways and packets.

- traffic metering: Metrics like data rate and traffic loss rate at a network point can be used to invoke treatment when necessary, including dropping and shaping.
- traffic restoration: When a network component fails, the network can respond to restore traffic flow.
- policy: Access to network resources can be set to meet needs for periods of time.

The diagram below visually separates these mechanisms for controlling core network traffic to ensure service data flows meet SLAs.

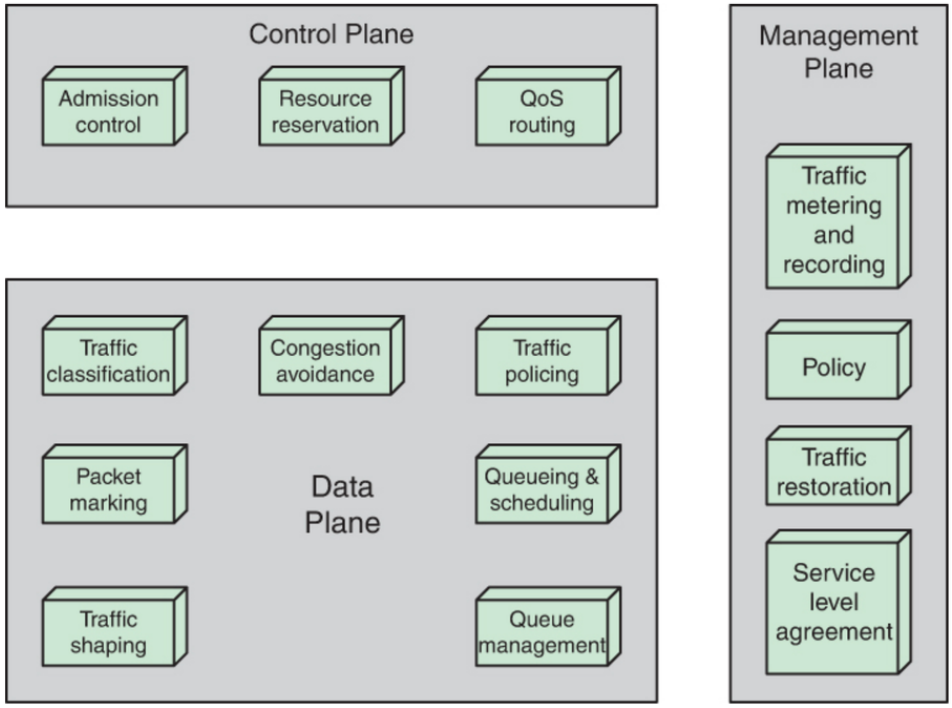


FIGURE 9.10 Architectural Framework for QoS Support

RAN Mechanisms

Radio resources are often the most scarce resource in the transmission of user plane data. In 5G, radio channels that carry user plane data are called data radio bearers (DRB) while channels that carry control plane data are called signaling radio bearers (SRBs). A finite data rate can be sustained across each DRB. That data rate is much lower than the rate that can be sustained across fiber optic backhaul.

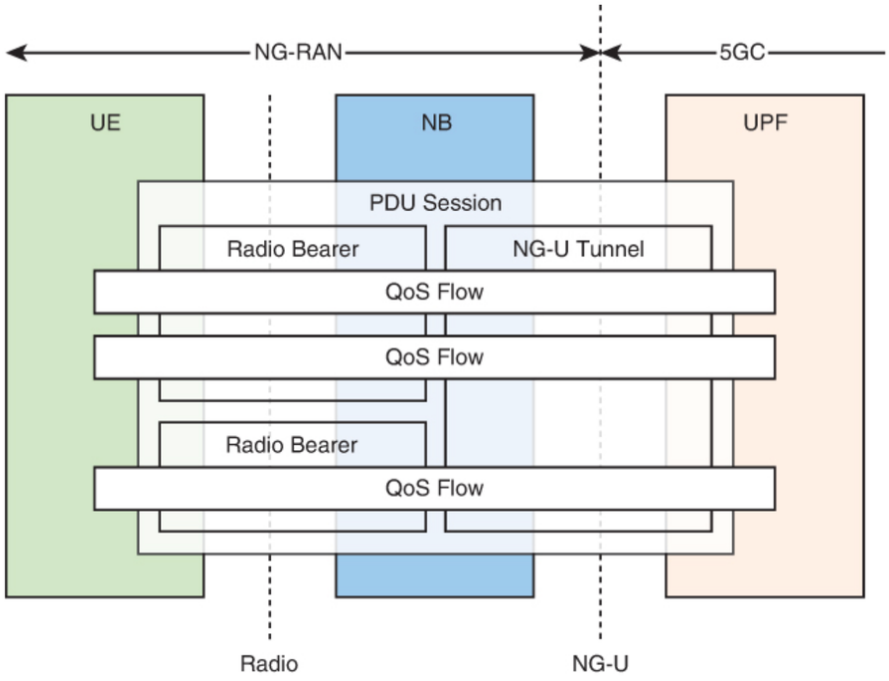
The flows within a PDU session need to be assigned both

- DRBs between UE and base station (labeled NB in the diagram below)

- a tunnel between the base station and the UPF (labeled NG-U below for next generation user plane)

Lessons learned from 4G and and intended diversity of uses of 5G (beyond phones) have motivated a flexible way to put flows into DRBs that is constrained enough to support diverse QoS demands.

- 5G introduces the possibility of putting two QoS flows into a single DRB to conserve radio resources.
- If one DRB is not enough to carry the QoS flows with the required QoS, then another DRB can be utilized simultaneously.
- The SDF flows within a QoS flow are never split between two DRBs so that the assignment of QoS flows to DRBs is always a function (by virtue of being a many to one assignment).



No matter how the QoS flows are divided between data radio bearers, all QoS flows are collected into a single NG-Tunnel.

We now turn to describing the features that describe a QoS flow. Before starting it will help to make a 3-way distinction between QoS flows; this distinction centers on whether or not a QoS flow makes guarantees about bit rate. ‘GBR’ is short for ‘guaranteed bit rate’.

1. non-GBR QoS flow
2. GBR QoS flow
3. delay-critical GBR QoS flow

The distinction will be revisited below, but will be referenced before it is presented in detail, thus the need for this paragraph.

QoS Values (Parameters and Characteristics)

A set of QoS values define the requirements for a single QoS flow.

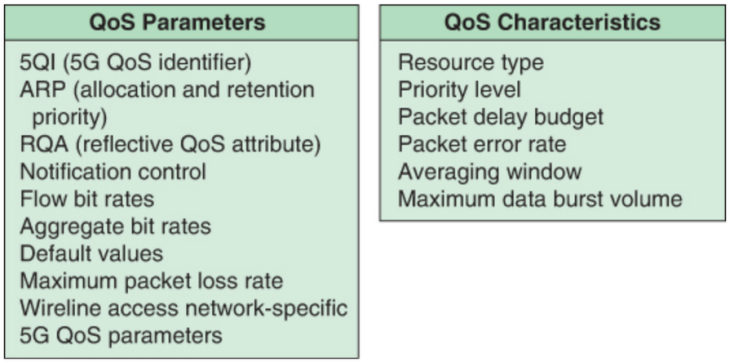


FIGURE 9.13 Elements of 3GPP QoS Model

QoS Parameters

- **5G QoF identifier (5QI)** is a integer code that refers to a set of QoS Characteristics. (See the table below after reading this section.)
- **Allocation and Retention Policy (ARP)** defines the relative priority of a QoS flow relative to other QoS flows with the following three values
  - **ARP priority value** is an integer between 1 and 15 (inclusive). QoS flows with lower ARP priority value are given higher priority.
  - **ARP preemption capacity** is boolean valued and determines if a QoS flow may get resources that were already allocated to another QoS flow with a lower priority number; it is a right to steal from other flows.
  - **ARP preemption vulnerability** is boolean value and determines if a QoS flow may lose resources allocated to so that a QoS flow of lower ARP priority value may have the resources to meet its QoS requirements.
- **Reflective QoS attribute (RQA)** is boolean valued and determines if the QoS parameters are the same for uplink and downlink.
- **Flow bit rates** are the rates of transfer across the core network required by the QoS flow. It comes in two parts, a minimum and a maximum. When the flow rate is between the minimum and maximum then the priority ARP priority level (above) is respected. If a QoS's flow rate is outside this range then the priority levels may not be respected.
  - **Guaranteed flow bit rate (GFBR)** is the minimum rate to be guaranteed by the CN (on average over a specified averaging time window). The value is specified separately for uplink and downlink.
    - If this field is empty then the the flow is classified as a **non-GBR QoS flow**.
    - If the field is non-empty then the QoS flow is classified as
      - a **guaranteed bit rate QoS flow (GBR QoS flow)** if it does not have a maximum data burst volume (presented below)
      - a **delay critical guaranteed bit rate QoS flow (GBR QoS flow)** if it does have a maximum data burst volume
    - Note that the acronyms GFBR and GBR are not redundant since
      - `GFBR` is a key that gets an associated numerical value,
      - 'GBR', 'non-GBR' and 'delay-critical GBR' are
        - adjectives describing QoS flows and are
        - values that the key `resource type` (described below) can take.
  - **Maximum flow bit rate (MFBR)** is the maximum rate that will be required by the QoS flow; QoS guarantees do not apply when flow rate is above this value. That is, if a rate higher than this is generated then priority levels are not enforced and excess packets may be dropped. MFBR is specified separately for uplink and downlink SDF flows.
- **Notification control** is a boolean determining if notifications are requested for moments when the RAN no longer can (or can again) support the GFBR.
- **Aggregate bit rates** are maximum data flow rates.
  - **Per session aggregate maximum bit rate (session-AMBR)** describes a PDU session for the UE; it is the maximum sum of rates of non-GBR QoS flows in the PDU. (This values does not describe GRB QoS flows)
  - **Per UE aggregate maximum bit rate (UE-AMBR)** describes a UE; it is the maximum sum of rates of non-GBR QoS flows to or from the UE. This value is provided to the RAN by either

        - the AMF based on a value retrieved from the UDM (the sum of session-AMBRs for the UE) for non-roaming UE, or
        - the PCF based on dynamic service for roaming UE.
- **Default values** are applied to non-GBR QoS flows and are determined by the 5QI and the ARP; the SMF retrieves these defaults from the UDM.
- **Maximum packet loss rate** is the tolerated loss of packets per unit time for the QoS flow. This value is specified separately for uplink and downlink.
- **Wireline access network-specific 5G QoS parameters** are additional parameters applicable only to wireline networks.

QoS Characteristics

You can think of `QoS_characteristics` as a QoS parameter whose value is a dictionary, and that dictionary has the following list of values. (i.e. `QoS_characteristics: {ResourceType: GBR, … }`).

- **Resource type** has three possible values.
    - delay-critical GBR
        - this allows setting a maximum data burst volume, defined below, which only applies to QoS flows of this type
    - GBR
    - non-GBR
- **Priority level** allows for distinction between QoS flows whether those flows are with the same or with different UEs. (Note that priority level is not the same as APR priority value, introduced above under values.) QoS flows with the lowest priority level get higher priority. Every 5QI is associated with a default value for the priority level, but this default can be overridden.
- **Packet delay budget (PDB)** is the time that a packet may be delayed between the UE and (terminal?) UPF. This value can be different for uplink and downlink.
    - For delay-critical GBR QoS flows, a packet is counted as lost if the packed is delayed by more than the PDB. This policy helps to account for unacceptable lag in delay critical applications.
        - This counting is overridden if either (aka the additional counting applies if both of the following are false)
            - the packet is delivered as part of a burst that exceeds the maximum data burst volume of QoS flow (defined below) over a time interval of width equal to the packet delay budget (PDB) (so the problematic behavior is already identified and need not be accounted for here)
            - the QoS flow rate is above the GFBR (so the problematic behavior is already identified and need not be accounted for here)
    - For GBR QoS flows one can expect 98% of the packets to have delay time less than the PDB and that packets will not be dropped.
        - This only applies when the bit rate is less than the GFBR.
    - For non-GBR QoS flows one can expect 98% of packets to have delay times less than the PDB, and that packets might sometimes be dropped during congestion.
- **Packet error rate (PER)** is the upper bound for (non-congestion related) packet loss rate.
    - For delay-critical GBR QoS flows, a packet that is delayed more than the PDB is counted as lost, so the PER is an upper bound on packets with excessive delay.
- **Averaging window** is the duration W over which the GFBR and MFBR are calculated from the bits transmitted B as $\Delta B/W$.
    - Every standard 5QI with resource type either delay-critical GBR or GBR has a default averaging window.
- **Maximum data burst volume (MDBV)** is the largest amount of data that the QoS will need to serve within a window of time of width PDB.
    - Every QoS flow with resource type delay-critical GBR must have a MDBV (so that packets delayed during bursts that exceed the MDBV do not count toward packets dropped)
    - Every standard 5QI with resource type delay-critical GBR has a default MDBV.

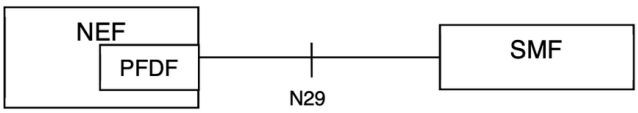
Table 5.7.4-1: Standardized 5QI to QoS characteristics mapping

5QI Value & QFI	Resource Type	Priority Level	Packet Delay Budget	Packet Error Rate	Example Services
1	GBR	20	100 ms	10^{-2}	Conversational Voice
2		40	150 ms	10^{-3}	Conversational Video (Live Streaming)
3		30	50 ms	10^{-3}	Real Time Gaming, V2X messages
4		50	300 ms	10^{-6}	Non-Conversational Video (Buffered Streaming)
65		7	75 ms	10^{-2}	Mission Critical user plane Push To Talk voice (e.g., MCPTT)
66		20	100 ms	10^{-2}	Non-Mission-Critical user plane Push To Talk voice
75		25	50 ms	10^{-2}	V2X messages
5	Non-GBR	10	100 ms	10^{-6}	IMS Signalling
6		60	300 ms	10^{-6}	Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)
7		70	100 ms	10^{-3}	Voice, Video (Live Streaming) Interactive Gaming
8		80	300 ms	10^{-6}	Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file
9		90			sharing, progressive video, etc.)
69		5	60 ms	10^{-6}	Mission Critical delay sensitive signalling (e.g., MC-PTT signalling)
70		55	200 ms	10^{-6}	Mission Critical Data (e.g. example services are the same as QCI 6/8/9)
79		65	50 ms	10^{-2}	V2X messages

Sample of Official 5QI to QoS Characteristics Maps

AF Control of Sessions

The NEF is able to hand new session rules to the SMF for an SDF through a module in the NEF called the packet flow descriptor function (PFDF). A packet flow descriptor (PFD) is a set of packet detection rules (PDRs) that are from an application function (AF) and are intended to be applied to define or redefine an SDF.



Here is the idea; an SDF is a flow between a UE and an application and an SDF template dictates which packets are considered part of that flow. The SMF holds that SDF template and uses it to create, update, and delete packet detection filters in the UPF for that SDF. If there is an application function (AF) for that application in the core then that AF can access the NEF. In particular, the AF can use the NEF as a communication proxy with the SMF and change packet flow descriptors for the SDF. This is dones by sending a PFD. In this way the AF can have control of which packets are considered part of the SDF for the application.