SVEUČILIŠTE U ZAGREBU
Fakultet elektrotehnike i računarstva

Master Programme

Computing

Advanced Architectures of Telecommunication Networks

Ac. year 2022/2023

Lecture 11: Resource Management and Quality of Service

prof. dr. sc. Lea Skorin-Kapov

Outline

- Quality of Service
- Policy control and charging
- Network slicing

Quality of Service

What is Quality of Service (QoS)?

- IETF defines QoS as:

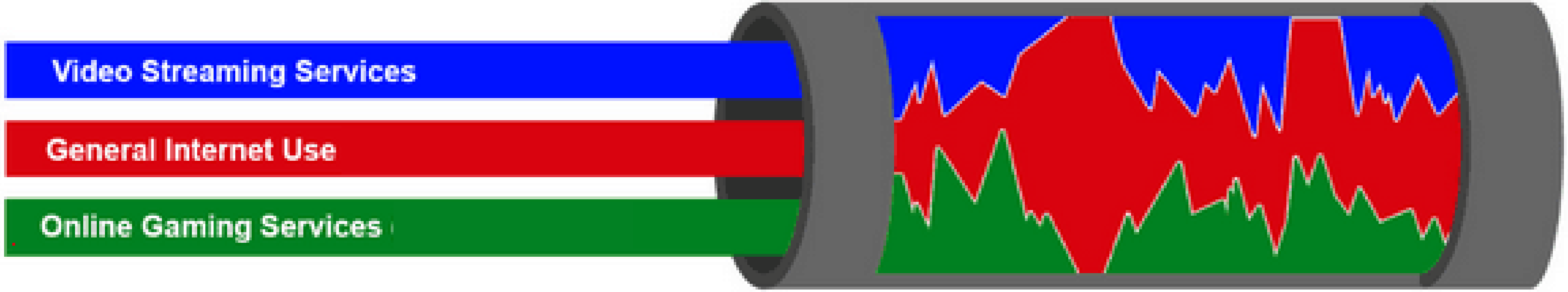
  A set of service requirements to be met by the network while transporting a flow.

- ITU-T defines QoS as:

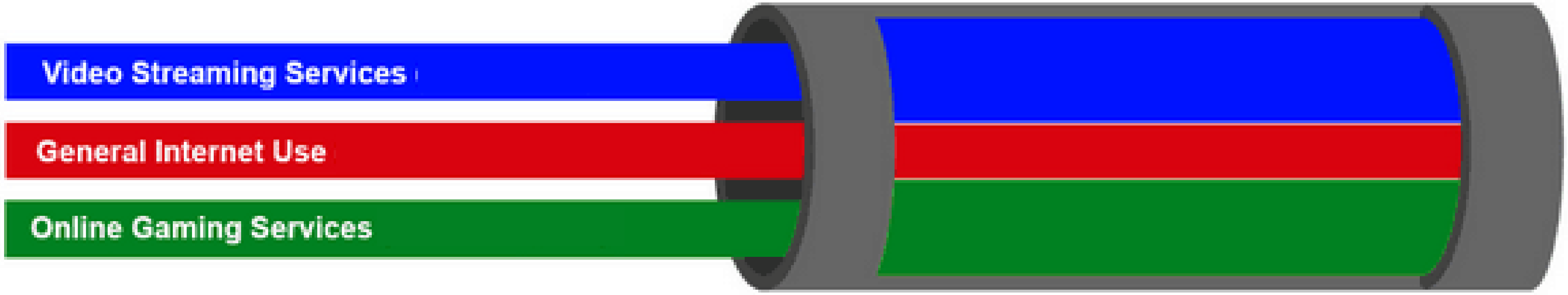
  The totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service.

- In practice, QoS focuses on quantitative network parameters (e.g., packet loss rate, throughput, delay, jitter and/or bandwidth)

Example throughput without QoS guarantees



Example throughput with QoS guarantees (guaranteed bitrate)

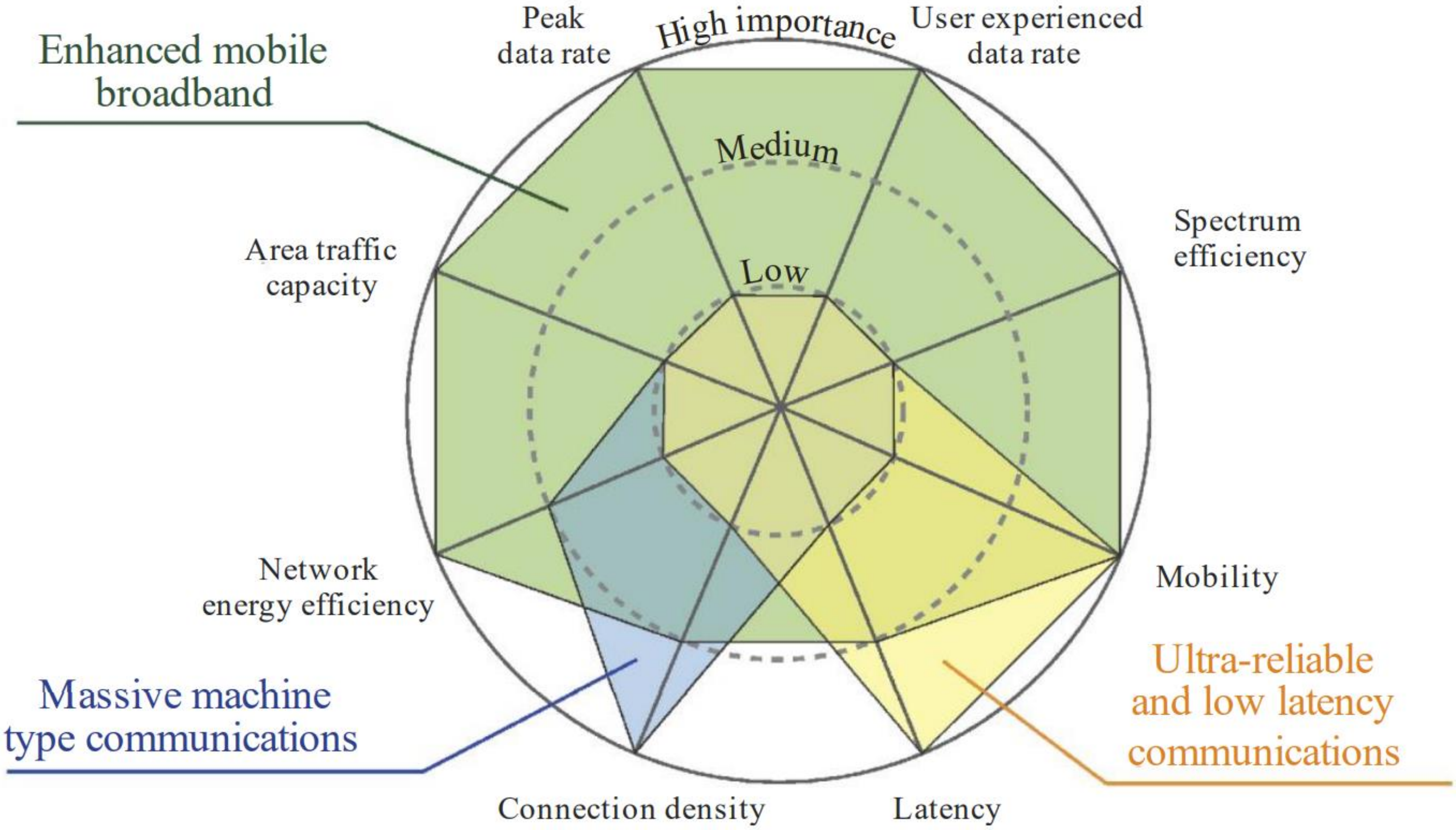


https://www.techplayon.com/5g-nr-qos-architecture-qos-attribute-and-qos-flow/

Why do we need QoS mechanisms in the network?

- E2E services: refer to the network or application services between a UE and the external Data Network (e.g., the Internet).
- Different E2E services require differentiated QoS treatments.
- Operators want the ability to provide a differentiated packet forwarding treatment of data which may belong to different users, different applications or even different services or media within the same applications.
    - For example: provide low latency for a voice flow; allocate high bandwidth to a video streaming flow; provide high reliability for sensor data

5G Use Cases (reminder – Lecture 7!)

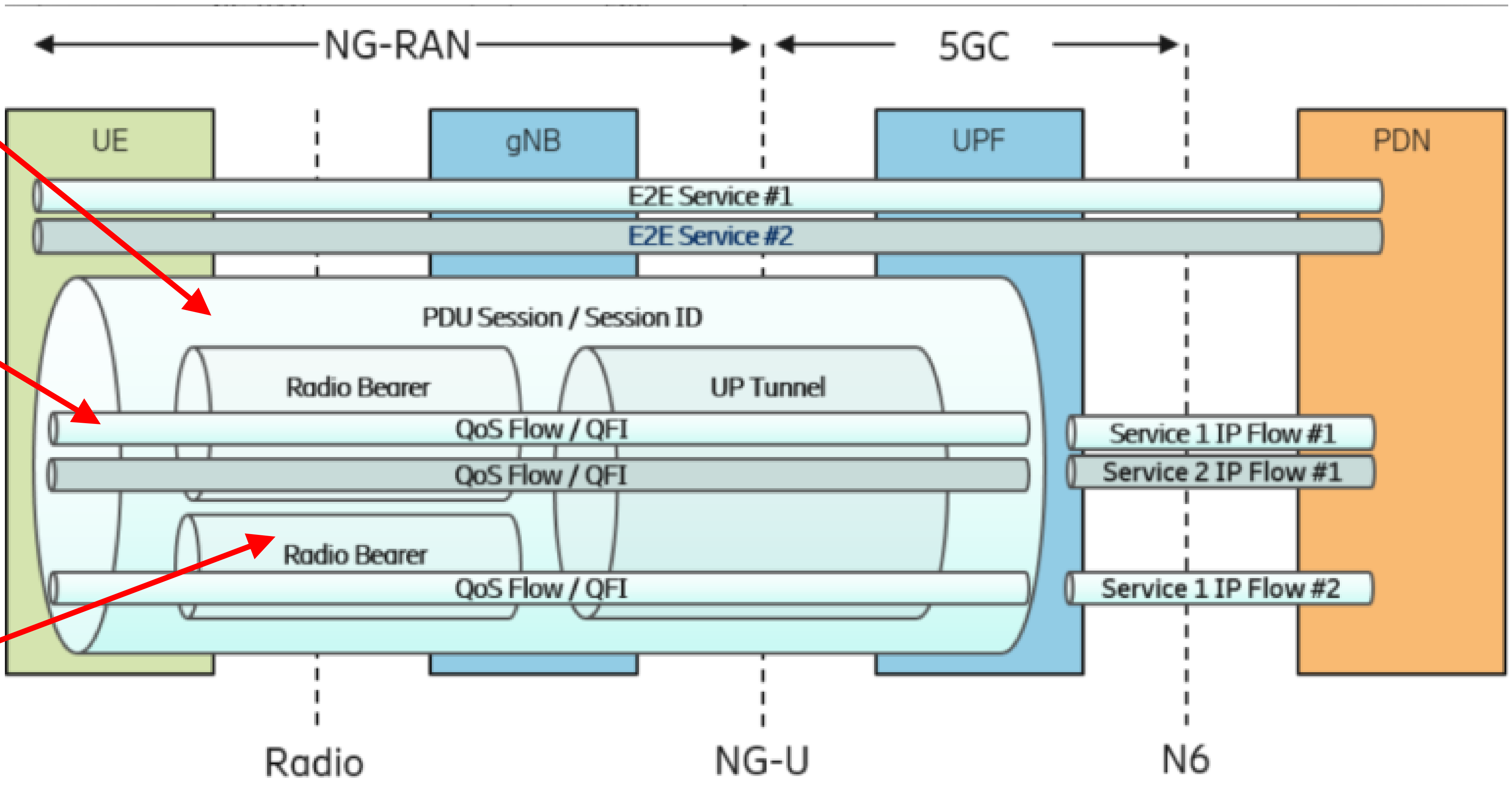


5G QoS framework: flow-based

PDU session: can consist of multiple QoS flows

QoS Flow: has associated certain QoS paramaters; is identified with a QoS Flow ID

Radio Bearer: a bearer service is a link between two points, which is defined by a certain set of (QoS) characteristics; Radio bearers are channels offered by Layer 2 to higher layers for the transfer of either user or control data between the UE and the NG-RAN



https://5ghub.us/quality-of-service-qos-in-5g-networks/

5G QoS framework: flow-based

- **Each E2E service can have one or multiple IP flows**. An SDF (Service Data Flow) is one IP flow or group of IP flows of UE traffic classified by the type of service that is used.
- One or more SDFs can be transported in the same 5G QoS flow if they share the same QoS treatment (*see example on next slide – Whatsapp and Skype video*).
- **Each uplink and downlink packet is mapped to a QoS flow**. This QoS flow provides the forwarding treatment between the UE and the UPF throughout the lifetime of the PDU session.
- **One PDU session can carry one or several QoS flows**
- A QoS Flow is identified by a QoS Flow ID (QFI) in a PDU session
- **Data packets marked with the same QFI receives the same traffic forwarding treatment** (e.g., scheduling, admission threshold).
- A radio bearer can carry one or several QoS flows. Each PDU session has a unique set of radio bearers and the gNB decides over which radio bearer a QoS flow is sent.

Example

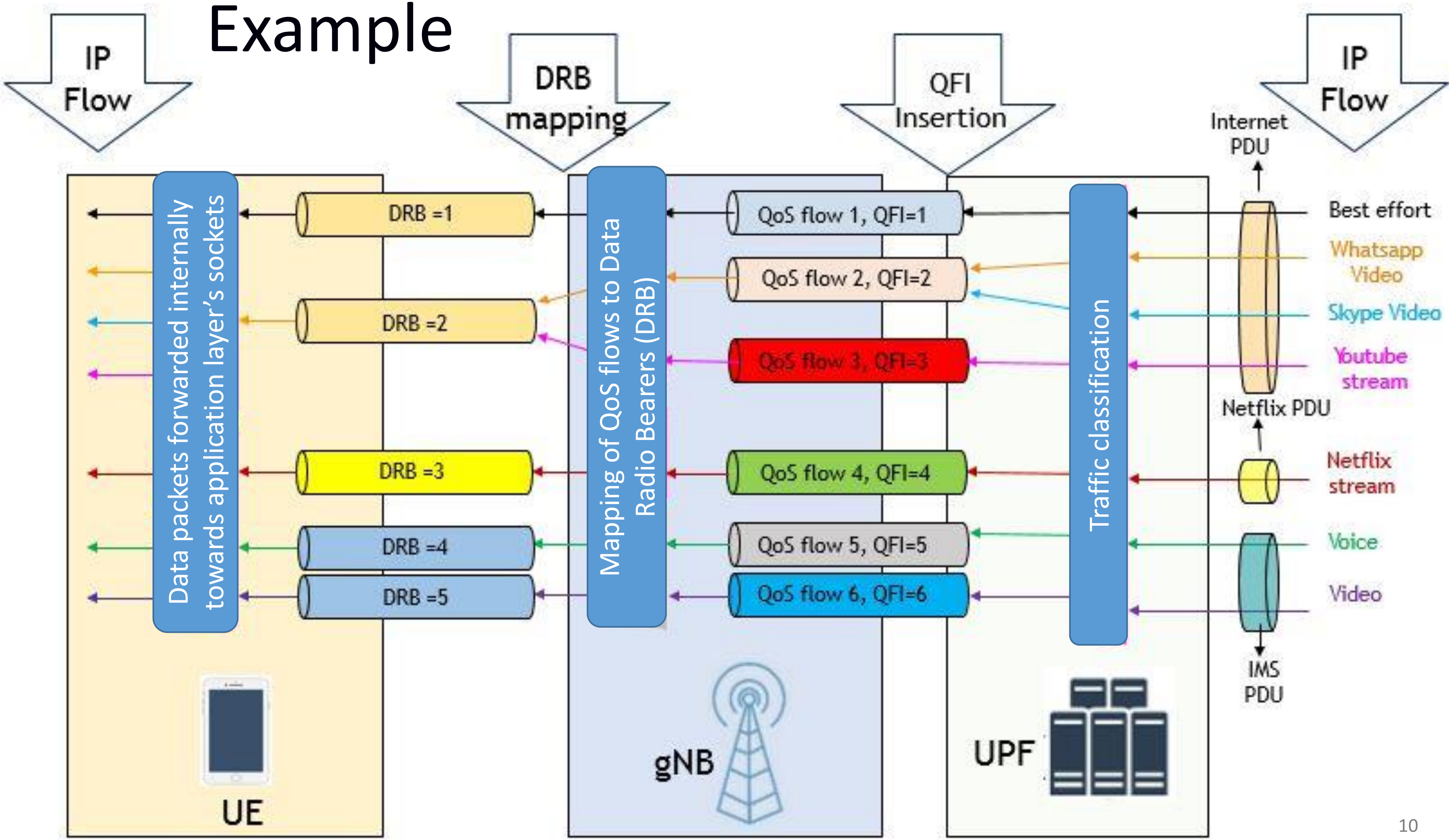


figure adapted from https://www.techplayon.com/5g-nr-qos-architecture-qos-attribute-and-qos-flow/

QoS Profiles

- **A QoS flow has either a:**
  - **Guaranteed Bit Rate (GBR)** (e.g., conversational voice and video; real-time gaming; delay sensitive signaling), or
  - **Non-Guaranteed Bit Rate (non-GBR)** (e.g., TCP-based www, email, ftp; buffered video streaming)
- Each QoS flow is characterized by a set of parameters that are specified in a **QoS profile**:

QoS parameter	Description
5G QoS Identifier	a scalar that is used as a reference to QoS characteristics
Allocation and Retention Priority	whether a service data flow may get resources that were already assigned to another service data flow with a lower priority level; whether a service data flow may lose resources assigned to it in order to admit a service data flow with higher priority level
Flow Bit Rates	If it is a **Guaranteed Bit Rate** flow, then specification of guaranteed and max uplink and downlink bitrates
Max Packet Loss Rate	maximum tolerated uplink and downlink packet loss rate

5G QoS Identifiers (5QI)

A 5QI is **a pointer to a set of QoS characteristics such as priority level, packet delay or packet error rate, etc.**

some examples

5QI	Resource Type	Priority	Delay Budget	Packet error rate	Example Services
1	GBR	20	100 ms	10^{-2}	Conversational voice
2	GBR	40	150 ms	10^{-3}	Conversational (streaming) video
3	GBR	30	50 ms	10^{-3}	Real time gaming
4	GBR	50	300 ms	10^{-6}	Non-Conversational Video
65	GBR	7	75 ms	10^{-2}	Mission Critical user plane Push To Talk voic
66	GBR	20	100 ms	10^{-2}	Non-Mission-Critical user plane Push To Talk
75	GBR	25	50 ms	10^{-2}	Vehicle to everything
5	non-GBR	10	100 ms	10^{-6}	IMS signaling
6	non-GBR	60	300 ms	10^{-6}	Buffered video, TCP-based (www, email…)
7	non-GBR	70	100 ms	10^{-3}	Voice, streaming video, gaming
8	non-GBR	80	300 ms	10^{-6}	Buffered video, TCP-based (www, email…)
9	non-GBR	90	300 ms	10^{-6}	Buffered video, TCP-based (www, email…)
69	non-GBR	5	60 ms	10^{-6}	Mission Critical delay sensitive signalling
70	non-GBR	55	200 ms	10^{-6}	Mission Critical Data
79	non-GBR	65	50 ms	10^{-2}	Vehicle to everything
80	non-GBR	66	10 ms	10^{-6}	Low latency eMBB, augmented reality

Policy control and charging

What are “policies” ?

- **Policies**: rules for how users, data sessions, and data flows shall be controlled or managed, including what services are allowed, how charging shall be done, what quality-of-service applies etc.

- Policies can be applied with different levels of granularity, for example:
  - Policy rules that apply to all users in the network
  - Policy rules that apply to all services for a specific user
  - Policy rules that apply to specific data sessions or data flows for a given user

Policy Control Function (PCF)

The PCF is the main network function in the 5G Core that is responsible for handling policy information

PCF

Policy control related to data sessions

- charging (e.g., per flow or per service), authorized QoS for IP flows; rely on the fact that IP flows can be classified in the UPF using unique packet filters
- rules indicating to the device which data session, slice, or SSC mode to use for a certain application

Policy control NOT related to data sessions

- how a user can access the network (e.g., restricting geographic area where a user can attach – list of allowed or not allowed tracking area IDs)
- which radio access technologies a user can utilize

Charging models

- **Offline charging:** balance deduction happens **after** consumption of resources; charging-related data is collected concurrently as the resources are being used; sent to billing domain and processed after usage of network resources is complete
  - **cannot affect the user data session in real time**

- **Online charging**: balance deduction happens **before** consumption of resources, network resource usage must be authorized → a subscriber must have a pre-paid account
  - **can affect the user data session in real time**

The UPF needs to report about data usage to the SMF so that charging can be performed

Network slicing

Foundations for network slicing

- Move from monolithic core network elements to a service-based architecture
- Virtualized network functions can be provisioned and deployed dynamically
- Virtualization and SDN enable building logical networks on top of a common and shared network infrastructure

Network Slice: a logically separated, self-contained, independent and secured part of the network, targeting different services with different requirements on speed, latency, and reliability.

Network slicing: basic concepts

Goal: separate traffic into multiple logical (virtual) networks that all execute on and share a common physical infrastructure

- Allows a network operator to provide dedicated virtual networks with functionality specific to the service or customer over a common network infrastructure
  - The “customer” is not directly the end-user, but a business entity that has requested specific services from the network operator, e.g., an enterprise, another service provider or the network operator
- Each virtual network (slice) comprises an independent set of logical network functions that support the requirements of the particular use case

Different service requirements

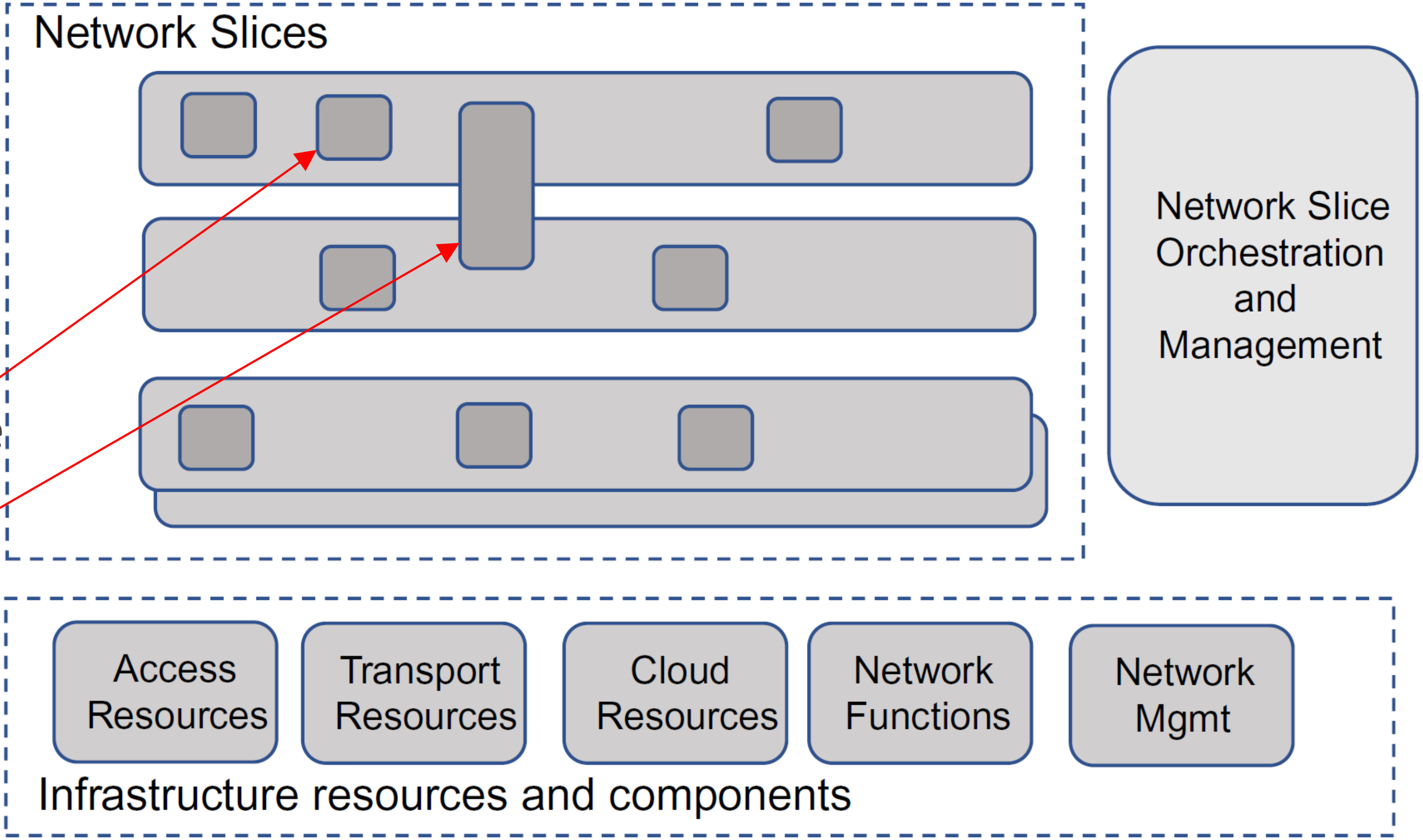
- Depending on service type (e.g., eMBB, URLLC, mIoT), Network Slices can be set up to meet different requirements in terms of:
  - traffic capacity per geographical area
  - coverage area
  - end-to-end latency
  - mobility
  - service availability and reliability
  - priority
  - security
  - charging
  - etc.

Examples:

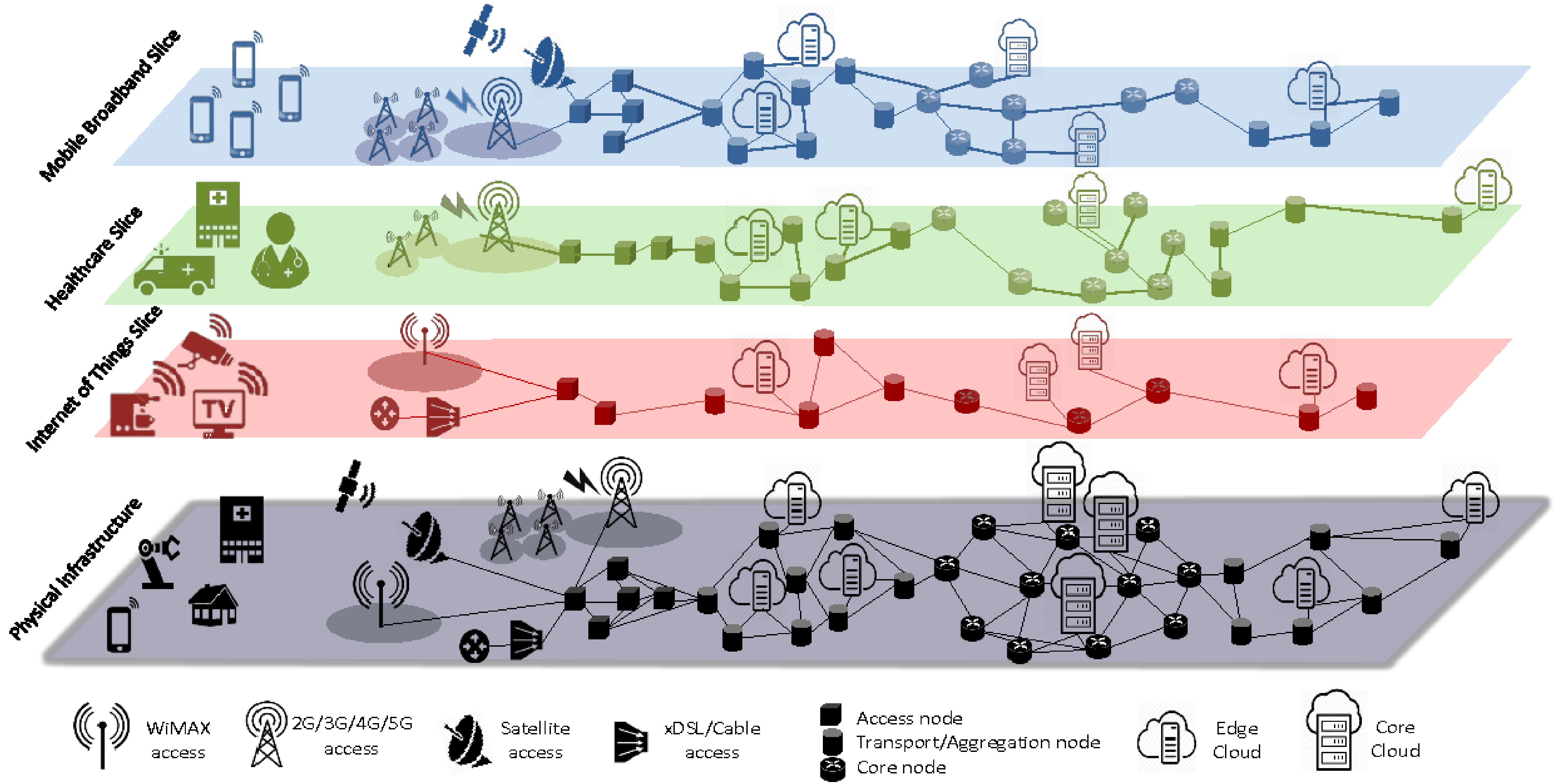
- **Critical IoT slice**: low latency, high bandwidth and ultra-reliability
- **Massive IoT slice**: higher latency and lower bandwidth

Network Slices: resources

- Network Slice: consists of **radio network + core network resources**
- The used physical or virtual infrastructure resources may be dedicated to the Network Slice or shared with other Network Slices



Network slicing: building of logical networks



Ordonez-Lucena, Jose et al. "Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges." IEEE Communications Magazine 55 (2017): 80-87.

Network Slicing: benefits

- better customer experience (will be the same as if using a physically separate dedicated network)
- shorter time-to-market and time-to-customer
- efficient usage and management of network resources
- increased automation (quick creation and updates of slices)
- flexibility and agility
- reduced risks (e.g., if a cyber attack breaches one slice, the attack could be contained to prevent spreading beyond that slice)

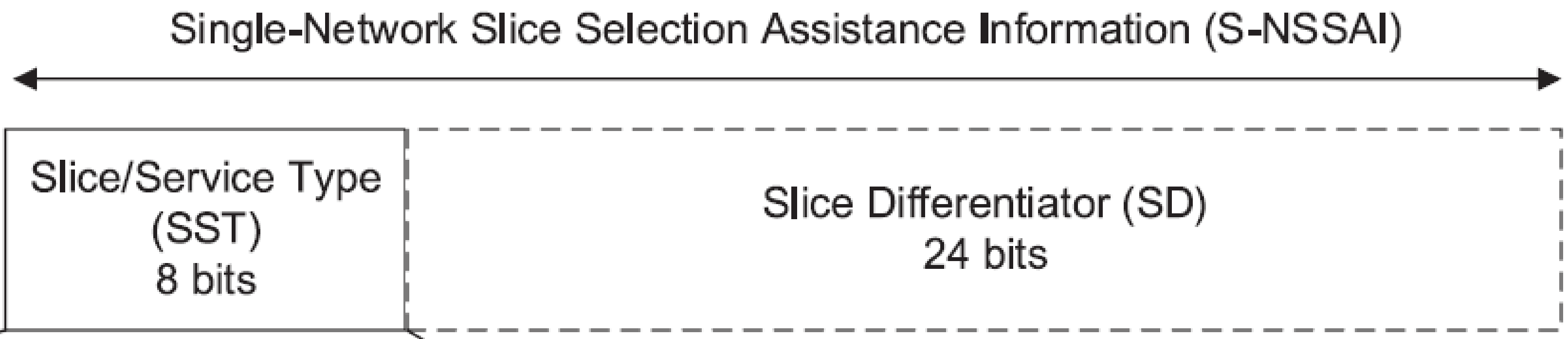
Management and orchestration of Network Slices

- Customers provide requirements using APIs
- **Network Slice preparation**:
  - if an existing Network Slice "template" already exists, it can used. Otherwise, a new one is designed based on customer requirements
  - the Network Slice requirements (template) are validated and uploaded to the production system; preparation of the network environment
- **Creation** of a Network Slice Instance (NSI): allocation and configuration of resources
  - if an existing NSI is being used, then it is scaled to meet the requirments of the new customer; otherwise a new NSI is created
- **Operation**: activation, supervision, performance reporting, resource capacity planning, modification, and de-activation of an NSI.

Connecting UEs to Network Slices

- A specific Network Slice is identified by a parameter called **S-NSSAI** (Single Network Slice Selection Assistance Information), consisting of two sub-parameters:
  - Slice/Service Type (SST)
  - Slice Differentiator (SD) - used to differentiate between multiple slices of the same type
- The UE requests to connect to a certain NSSAI during the initial registration procedure when it connects to the network
- The 5G Core architecture allows one single device to connect to more than one slice simultaneously
  - one or more S-NSSAIs can be provided in an NSSAI

Network Slice identifier

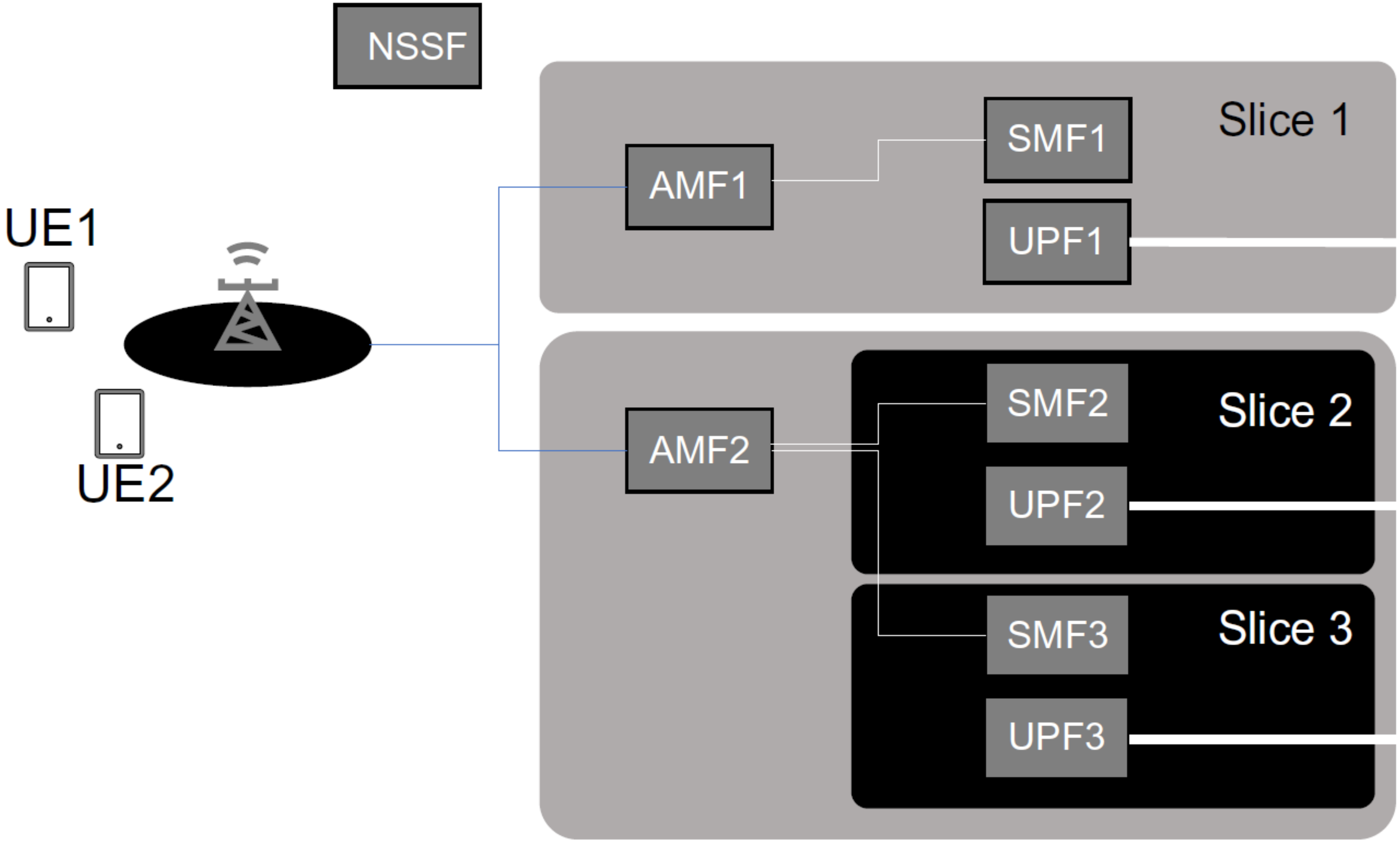


0		Reserved
1	eMBB	Network Slice suitable for handling of 5G enhanced Mobile Broadband
2	URLLC	Network Slice suitable for handling of ultra-reliable low latency communications
3	MIoT	Network Slice suitable for handling of massive IoT
4	V2X	Network Slice suitable for handling of V2X services
5–127		Reserved
128–255		Operator-specific

optional: can be used to differentiate between Network Slices used for different customers

Example: connecting UEs to Network Slices

- the radio network serving the device will use one or more S-NSSAI values requested by the device to select the AMF(s)
- UE1 connected to Slice 1
- UE2 simultaneously connected to Slice 2 and Slice 3



Network Slice availability

NSSF (Network Slice Selection function): selects the optimal network slice available for the service requested by the user

- A Network Slice may be available in the whole mobile network or in one or more Tracking Areas.
- Availability: refers to where the S-NSSAIs are supported.
- O&M (Operations and Management) configure the NSSF and also the 5G-RAN with info on where Network Slices are available. This info is spread using signaling interfaces shown in the figure