# 5G Slicing and Fronthaul

Aug-2018

## AGENDA

1. 5G Overview and Use-cases
2. Handling of Services in 4G
3. Handling of Services in 5G
4. Fronthaul - RAN-Split Options
5. Fronthaul - Transport
6. Summary

# 5G Use Cases from 3GPP

## IMT-2020 Requirements

- Downlink Peak Data Rate 20 Gbps
- Uplink Peak Data Rate 10 Gbps
- Downlink Peak Spectral Efficiency 30 bits/s/Hz
- Uplink Peak Spectral Efficiency 15 bits/s/Hz

Throughout > 10 Gbps per user

3D Video , UHD

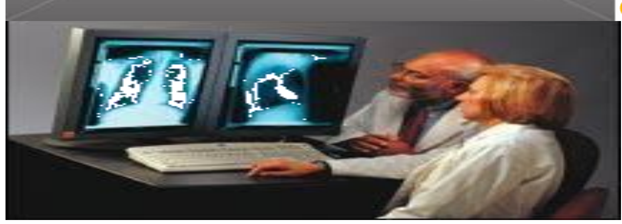
High spectral efficiency

Work and Play in the cloud

**Communication Entertainment , Internet**

eMBB & Gigabyte / second





Mission critical & URLCC

**Automotive & Medical**

- Userplane Latency < 1 ms
- Controlplane Latency < 20 ms

mMTC

**Smart City, Retail, Manufacturing**

- Connection Density 1,000,000 per sq KM
- Battery life > 10 Years

| 5G Technology compared to 4G will provide following additional features: | | |
|---|---|---|
| Ultra-high throughput speed (20Gbps/UE) - New RAT, mmWave, massive MIMO, massive Aggregation | Ultra-Reliable and Low Latency Communications(E2E few msec) - Tactile Internet, autonomous driving, remote controlled machine | Massive connectivity (hundreds of millions of IoT devices) |

# 5G Use Cases & Requirements

## Ultra Reliable Low Latency Communication (uRLLC)

Traffic Safety & Control
Remote Healthcare
Industrial Application & Control
Remote Manufacturing
Self Driving Car

**Features**

- Ultra Reliable
- Very Low Latency
- Very High Availability

**IMT-2020 Requirements**

- Userplane Latency < 1 ms
- Controlplane Latency < 20 ms

## Enhanced Mobile BroadBand (eMBB)

Media Everywhere
UHD Broadcast
3D Video, 4K,8K UHD
Gigabytes in a Second
AR/VR
High Speed Train

**Features**

- Very High Bandwidth
- Widespread Coverage

**IMT-2020 Requirements**

- Downlink Peak Data Rate 20 Gbps
- Uplink Peak Data Rate 10 Gbps
- Downlink Peak Spectral Efficiency 30 bits/s/Hz
- Uplink Peak Spectral Efficiency 15 bits/s/Hz

## Massive MTC (mMTC)

Tracking
Fleet Management
Smart Meter
Smart City
Smart Home/Building

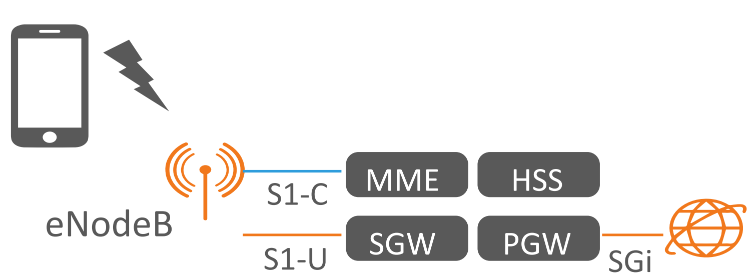
**Features**

- Massive numbers
- Small Data Volumes
- Low Cost
- High Battery Life

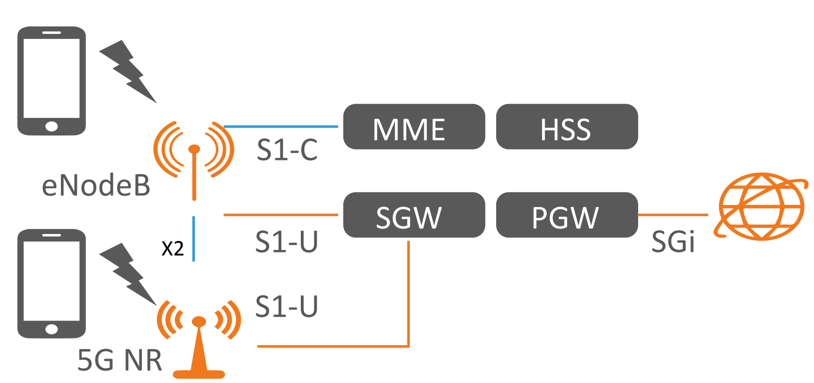
**IMT-2020 Requirements**

- Connection Density 1,000,000 per sq KM
- Battery life > 10 Years

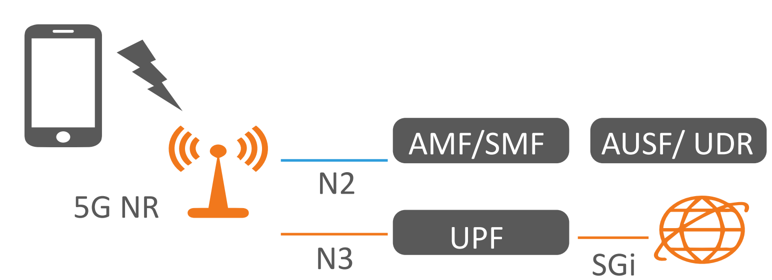
## 4G Interfaces




## Non Stand Alone Model




## Stand Alone Model



# 4G to 5G Network

Aricent
.ALTRAN GROUP

## 4G network

- Defined primarily to provide broadband data only
- Voice is considered to be sent over the same data-pipe (though with different QOS properties)
- **Same Air-interface or RAN network** to handle requirements of different services or verticals
- For NB-IOT/Cat-M1 support, separate set of devices and core nodes (C-SGN) required to handle the requirement

## 5G network

- **Horizontal** - Apart from support for higher broadband, 5G air-interface and network is designed to extend to handle other service attributes like
  - Support for massive or large number of devices)
  - Support for critical (low-latency) applications
  - Support for ultra reliable (very high reliability over AI) or tactile (AR/VR and related) scenarios
- **Vertical** - In addition 5G also focusses to address "verticals" specific requirements like automotive, industrial, healthcare, logistics, gaming /entertainment etc

| Automotive | Aerospace | Railway | Medical | Factory | Utilities |
|---|---|---|---|---|---|
| Autonomous Vehicle<br>Online Predictive Maintenance<br>Infotainment | Online Predictive Maintenance<br>UAV-based surveillance<br>Smart airport<br>Fleet Management | Online Predictive Maintenance<br>Integrated Railway Signalling<br>Intelligent information systems | Remote surgery<br>Predictive Maintenance<br>Asset tracking | Task automation<br>Predictive Maintenance<br>Mission-critical control<br>Smart metering | Massive sensor information collection<br>Mission-critical sensor information collection |

# Example of 5G Use-case

| Use-case | Parameters | Realized By |
|---|---|---|
| Stadium/Events – Crowdsourced video | Latency ~5s<br>Data-rate ~Mbps/user<br>Reliability >99.9% and High Connection density | eMBB |
| Stadium/Events – UHD Broadcasting | Latency ~10ms<br>Data-rate ~Gbps<br>Reliability >99.999% | eMBB |
| Stadium/Events – Power Management | Latency ~100ms<br>Data-rate ~kbps<br>Reliability >99.999% | mMTC |
| Stadium/Events – First Responders | Latency ~1ms<br>Data-rate ~Mbps<br>Reliability >99.999% | URLLC |
| Mission Critical Applications (UAV/drones) | Latency 10-150ms<br>Very high reliability >99.999%<br>Data-rate ~Mbps<br>High Mobility | eMBB |
| Augmented Reality (Head Mounted display) | Latency ~10 ms<br>High requirements on data rate ~Gbps - Mbps<br>Reliability 99.9% | eMBB |
| Machine-type communications | High energy-efficiency (10 years battery life)<br>Latency ~100ms<br>Low data-rate ~kbps and low packet size, $10^6$ devices/sqkm | mMTC |

Every use-case will have different set of requirements and same fronthaul/backhaul network and RAN nodes should handle all the requirements

- ✓ Latency, data-rate, reliability, availability, packet size, security
- ✓ mobility, connection density, traffic density, coverage, radio conditions etc

# Example of 5G Use-case (Industry 4.0)

| Use-case | Parameters | Realized By |
|---|---|---|
| Autonomous Guided vehicles (AGV) | High requirements on latency, reliability<br>Data-rate ~Mbps-kbps<br>Seamless mobility support | URLLC |
| Motion Control Applications (Industrial Robot) | Highest requirements on latency <1ms<br>Low data-rate ~Mbps – Kbps and low packet size<br>Very high reliability >99.9999% | URLLC |
| Augmented Reality (Head Mounted display) | Latency ~10 ms<br>High requirements on data-rate ~Gbps - Mbps<br>Reliability 99.9% | eMBB |
| Mobile Human-Machine-Interfaces (Terminal) | Highest requirements on latency ~4 ms<br>Data-rate – Mbps - kbps<br>Reliability >99.9999% and Seamless mobility support | eMBB/URLLC |
| Process Monitoring (Wireless Sensor Networks) | High energy-efficiency<br>Latency ~100ms<br>Low data-rate ~kbps)and low packet size | mMTC |
| Video Surveillance | Latency ~10 ms<br>Data-rate ~Mbps<br>Reliability 99.9% | eMBB |
| Channel models<br>(No pre-defined channel model to characterize the channel.) | Challenging RF environment (blocking and reflections due to moving objects – AGV, cranes etc). Spatial multiplexing increases capacity and Spatial diversity provides reliability & mitigate RF blockage | CoMP (with multiple antennas) |
| Others | Support for multiple services with different SLAs/QoS), Private and public network support, and support for licenses / unlicensed bands | Edge Cloud & Network Slicing |
| | Microsecond time synchronization, Handling Ethernet switch function | TSN |

✓ Need to support different service requirements within network

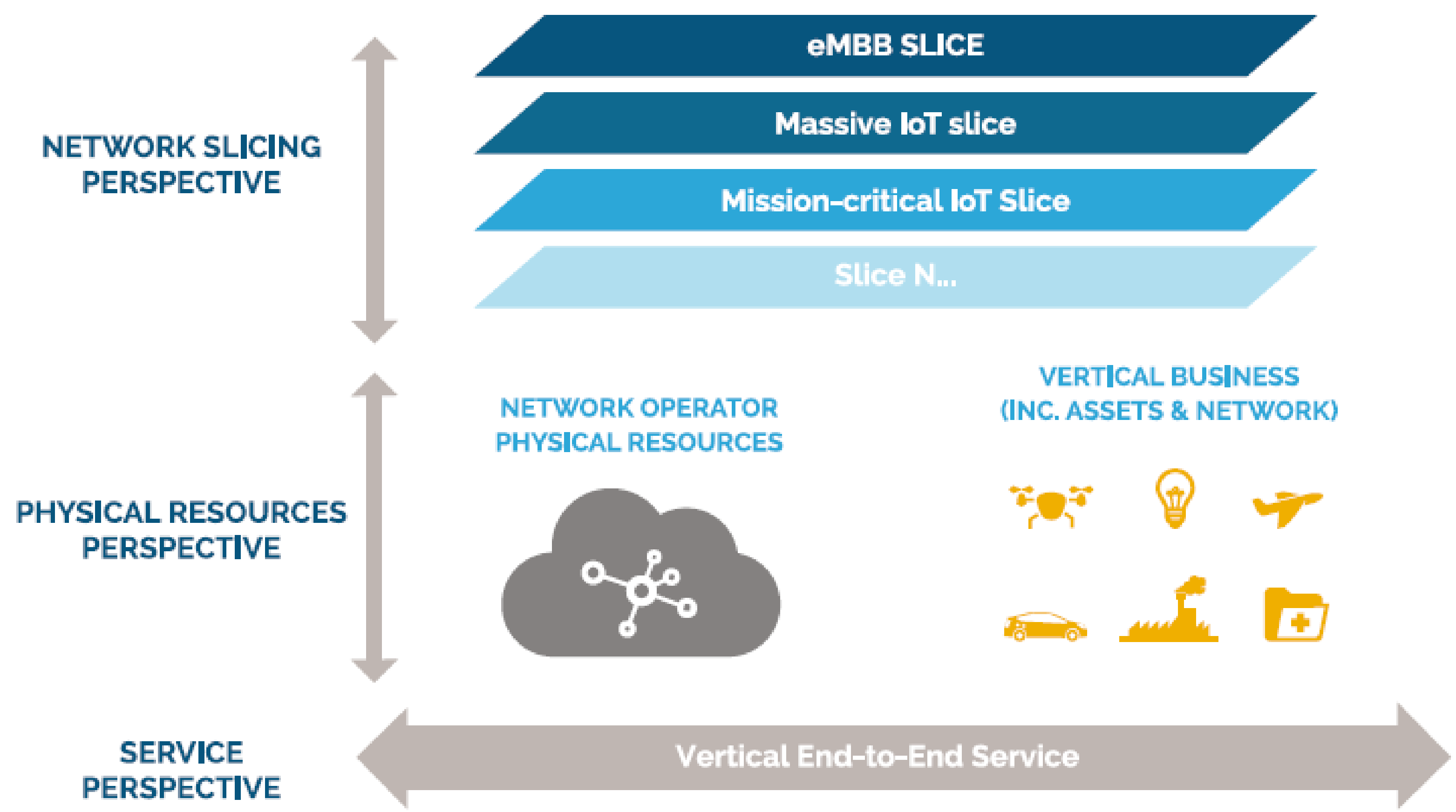
# What is Network Slicing ?

5G they need to support different type of services eMBB, eMTC, URLLC wherein

- Support different service requirements for each vertical in terms of mobility, security, policy control, latency, availability, bandwidth, datarate etc
- The network need to ensure the different requirements are met using same network and not have dedicated network for each type of service – and this is achieved through Network slicing.
- Network slicing allows multiple logical networks (slices) to be created over a single physical network, with each slice built to address a set of service parameters.

Network slices provide separation of traffic:

- from different users
- from different mobile operators
- from different types of traffic (QoS requirements)
- from different vendors (offering differentiated services for service type)
- Network slices can use isolated resources and/or shared resources



# Network Slicing Impacts end-to-end Network

- A network slice is an end-to-end logical subnet and requires coordination of

**Core network (control plane and user plane)**

- Split the control and user-plane so that user-plane could move closer to radio network to reduce overall latency and processing

**Transport network (backhaul and fronthaul)**

- Support the latency, datarate requirements with optimal utilization of the overall pipe bandwidth

**RAN node**

- handle different set of radio processing requirements (eMBB, URLLC, mMTC) for different slices
- Efficient RB allocation ensuring the key objectives such as spectrum efficiency, system capacity, and network quality are not impacted

**MEC (Multi-access Edge Computing) node**

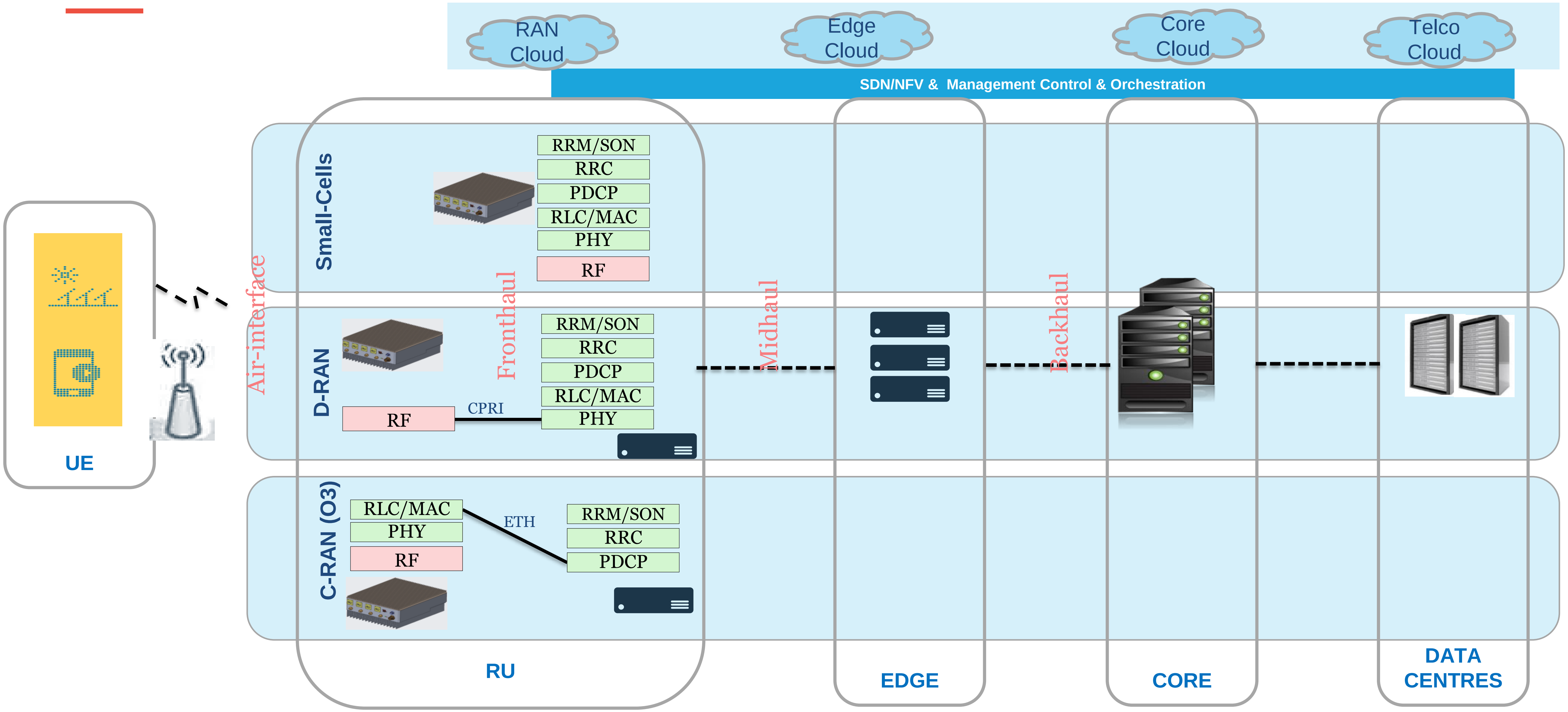
- Push applications & processing to the network edge to reduce network congestion and improve application performance
- Brings real-time, high-bandwidth, low-latency access to radio network information

**Network Softwarization**

- Modular and programmable network architecture realized using concept of SDN and NFV

# Handling of Services in 4G

# (Typical) 4G RAN Network




- ✓ High-level view of 4G Network architecture (details of switches / other nodes not shown) for RAN deployment options
- ✓ Mix of small-cells, D-RAN and C-RAN (Option-2) based deployment
- ✓ CPRI (optical) is typically used for fronthaul in D-RAN/C-RAN scenario
- ✓ Same RAN and Core NW architecture to handle different service requirements

# 4G QoS

## Static bearer based QoS

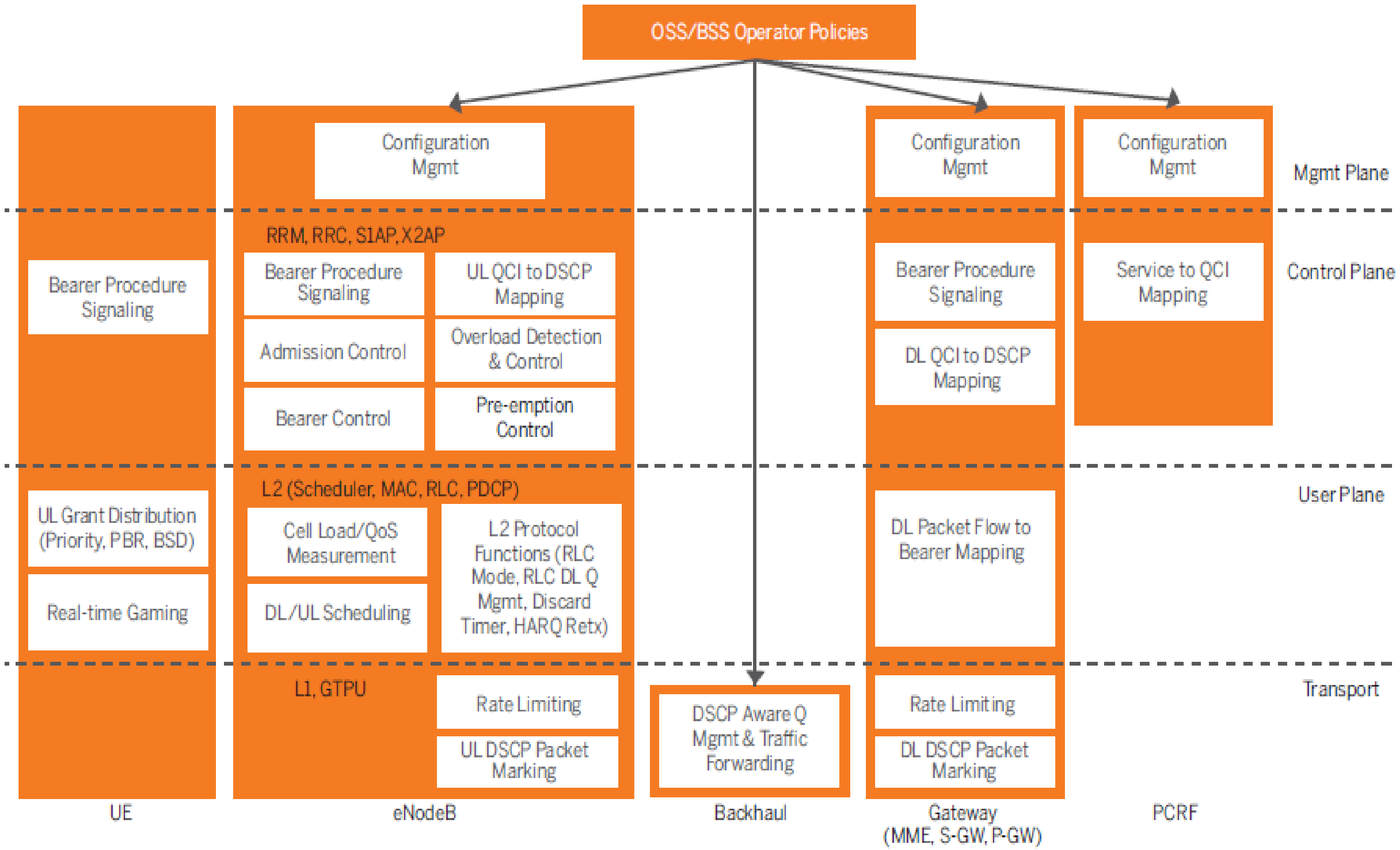
- **IP QOS**
  - DSCP
- **Bearers**
  - Dedicated (GBR or non-GBR), Default (non-GBR)
  - GBR bearer (GBR, MBR)
  - non-GBR bearer (APN-AMBR, UE-AMBR)
- **QCI**
  - Multiple levels for different traffic classes
  - Priority, PDB, PLER
- **Packet Scheduler**
  - ARP
  - Scheduling over air-interface (proprietary) QOS, PFS, SPS

| QCI | Resource Type | Priority Level | Packet Delay Budget (NOTE 13) | Packet Error Loss Rate (NOTE 2) | Example Services |
|---|---|---|---|---|---|
| 1 (NOTE 3) | | 2 | 100 ms (NOTE 1, NOTE 11) | $10^{-2}$ | Conversational Voice |
| 2 (NOTE 3) | GBR | 4 | 150 ms (NOTE 1, NOTE 11) | $10^{-3}$ | Conversational Video (Live Streaming) |
| 3 (NOTE 3), NOTE 14 | | 3 | 50 ms (NOTE 1, NOTE 11) | $10^{-3}$ | Real Time Gaming, V2X messages |
| 4 (NOTE 3) | | 5 | 300 ms (NOTE 1, NOTE 11) | $10^{-6}$ | Non-Conversational Video (Buffered Streaming) |
| 65 (NOTE 3, NOTE 9, NOTE 12) | | 0.7 | 75 ms (NOTE 7, NOTE 8) | $10^{-2}$ | Mission Critical user plane Push To Talk voice (e.g., MCPTT) |
| 66 (NOTE 3, NOTE 12) | | 2 | 100 ms (NOTE 1, NOTE 10) | $10^{-2}$ | Non-Mission-Critical user plane Push To Talk voice |
| 75 (NOTE 14) | | 2.5 | 50 ms (NOTE 1) | $10^{-2}$ | V2X messages |
| 5 (NOTE 3) | | 1 | 100 ms (NOTE 1, NOTE 10) | $10^{-6}$ | IMS Signalling |
| 6 (NOTE 4) | | 6 | 300 ms (NOTE 1, NOTE 10) | $10^{-6}$ | Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.) |
| 7 (NOTE 3) | Non-GBR | 7 | 100 ms (NOTE 1, NOTE 10) | $10^{-3}$ | Voice, Video (Live Streaming) Interactive Gaming |
| 8 (NOTE 5) | | 8 | 300 ms (NOTE 1) | $10^{-6}$ | Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.) |
| 9 (NOTE 6) | | 9 | | | |
| 69 (NOTE 3, NOTE 9, NOTE 12) | | 0.5 | 60 ms (NOTE 7, NOTE 8) | $10^{-6}$ | Mission Critical delay sensitive signalling (e.g., MC-PTT signalling) |
| 70 (NOTE 4, NOTE 12) | | 5.5 | 200 ms (NOTE 7, NOTE 10) | $10^{-6}$ | Mission Critical Data (e.g. example services are the same as QCI 6/8/9) |
| 79 (NOTE 14) | | 6.5 | 50 ms (NOTE 1, NOTE 10) | $10^{-2}$ | V2X messages |

Ref:23.203

Standardized QCI characteristics

# Handling of 4G QoS across nodes




Ref: (Aricent) Developing a QOS Aware Framework for LTE

- Bearer based QoS framework and class based operation, with limited number of QoS levels
- Control of packet flows that can be mapped to dedicated bearer (and QoS level associated) through policies provisioned in core
- Handling of QoS enforcement towards air-interface handled in RAN
- Incapable of providing service specific requirement handling and adapting per user sessions

- ✓ Priority: used to differentiate between bearers with same UE, as well as between bearers with different UE (1 = highest)
- ✓ Packet Delay Budget (PDB): defines the upper limit on the delay experienced by a packet, associated with a bearer, between UE and Gateway
- ✓ Packet Error Loss Rate (PLER): defines the upper limit on the rate of non-congestion-related packet losses. The PLER can be used for configuration of appropriate link layer protocol like RLC mode, max HARQ retransmissions, etc.
  - Typically, RLC AM mode used for PLER below 10-3, RLC UM for PLER above 10-3, RM UM mode

# Other Options in 4G

### RAN/Network Sharing

- MORAN (RAN sharing with separate spectrum)
- MOCN (RAN and spectrum sharing)
- GWCN (sharing of Gateway also)
- RAN scheduler to allocate the RBs across operators – static / semi-static

### DECOR (Dedicated Core Network)

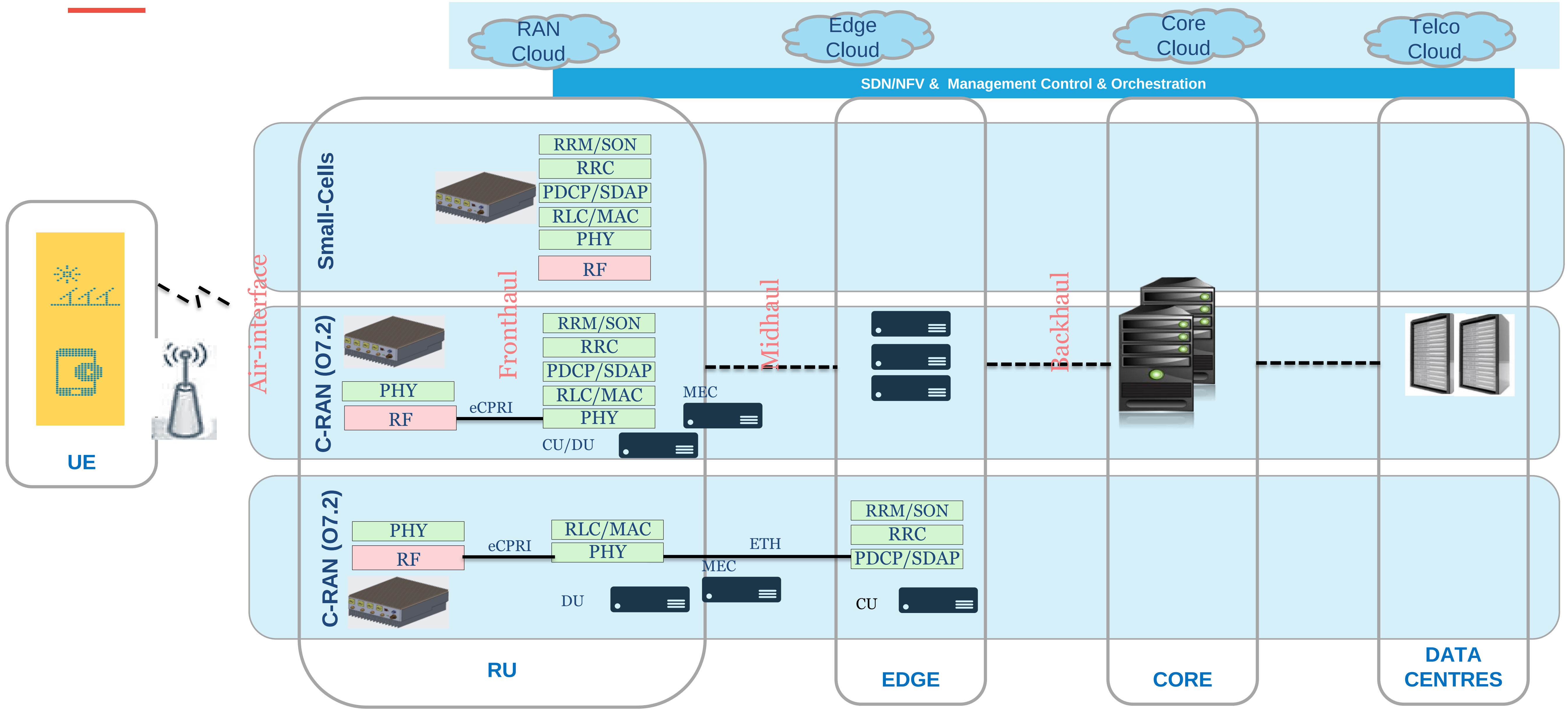
- Intent of network slicing realized by connecting UE to a customized core network
- eNodeB selects CN based on information received from UE
  - On receiving the NAS Attach Request, CN selects the MME based on the user's subscription information and redirects UE to select MME
- No changes needed to UE

### eDECOR

- Using the last selected Dedicated Core Network (DCN) based on information from UE
- UE provides the DCN selected last time in RRC Connection Setup Complete
  - Based on DCN Id, eNodeB routes the Initial UE Message to the correct MME in one go
  - This does not require a reroute from MME and hence, reduces the signaling and latency
- Changes needed in network and in UE as well

# Handling of Services in 5G

# 5G RAN Network




- ✓ Cloud-RAN based architecture shall meet 5G requirements of ultra-low latency and massive data throughput
- ✓ Multiple RAN Deployment options based on service requirements
  - ex: for low-latency scenarios, CU/DU can be collapsed into single node and MEC could be part of the topology
  - For eMBB case, CU and DU could be distributed across and it may be possible to have MEC entities distributed across the cloud
- ✓ Mix of small-cells, C-RAN (Option 7.2) based deployment based on 5G
- ✓ eCPRI is used for fronthaul and multiple options exist for fronthaul (OTN/PON, Ethernet)

# 5G QoS

## Flow based QoS

- 5QI
  - Common
    - Resource Type
    - Priority level
    - PDB, PER
  - GBR
    - Averaging Window *
    - Max Data Bursty Volume **
- Flows
  - Dedicated (GBR or non-GBR), Default (non-GBR)
  - GBR or delay-critical GBR (GFBR, MFBR, Max Packet Loss Rate, Notification Control)
  - non-GBR (Reflective QoS, UE-AMBR, Session-AMBR)
- Packet Scheduler
  - ARP
  - Scheduling over air-interface (proprietary)
  - Scheduling over air-interface (proprietary) QOS, PFS, SPS

* Averaging window application for GBR
** MDBV applicable for delay-critical GBR

- Flow-based QoS mechanism
  - GBR flow
  - Delay-critical GBR flow
  - Non-GBR flow
- Reflective QoS - Network decides on the QoS to apply on DL traffic, and UE reflects the DL QoS to associated UL traffic
- Notification control - indicates whether notifications are requested from the RAN when the GFBR can no longer (or again) be fulfilled for a QoS Flow
  - If NG-RAN determines that the GFBR cannot be fulfilled, RAN shall send a notification towards Core Network
- Flow bit-rates for GBR
- Per session aggregate bit-rates for non-GBR

# 5G QI to QoS mapping

| 5QI Value | Resource Type | Default Priority Level | Packet Delay Budget | Packet Error Rate | Default Maximum Data Burst Volume (NOTE 2) | Default Averaging Window | Example Services |
|---|---|---|---|---|---|---|---|
| 1 | GBR NOTE 1 | 20 | 100 ms | $10^{-2}$ | N/A | 2000 ms | Conversational Voice |
| 2 | | 40 | 150 ms | $10^{-3}$ | N/A | 2000 ms | Conversational Video (Live Streaming) |
| 3 | | 30 | 50 ms | $10^{-3}$ | N/A | 2000 ms | Real Time Gaming, V2X messages Electricity distribution – medium voltage, Process automation - monitoring |
| 4 | | 50 | 300 ms | $10^{-6}$ | N/A | 2000 ms | Non-Conversational Video (Buffered Streaming) |
| 65 | | 7 | 75 ms | $10^{-2}$ | N/A | 2000 ms | Mission Critical user plane Push To Talk voice (e.g., MCPTT) |
| 66 | | 20 | 100 ms | $10^{-2}$ | N/A | 2000 ms | Non-Mission-Critical user plane Push To Talk voice |
| 67 | | 15 | 100 ms | $10^{-3}$ | N/A | 2000 ms | Mission Critical Video user plane |
| 75 | | 25 | 50 ms | $10^{-2}$ | N/A | 2000 ms | V2X messages |
| 5 | Non-GBR NOTE 1 | 10 | 100 ms | $10^{-6}$ | N/A | N/A | IMS Signalling |
| 6 | | 60 | 300 ms | $10^{-6}$ | N/A | N/A | Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.) |

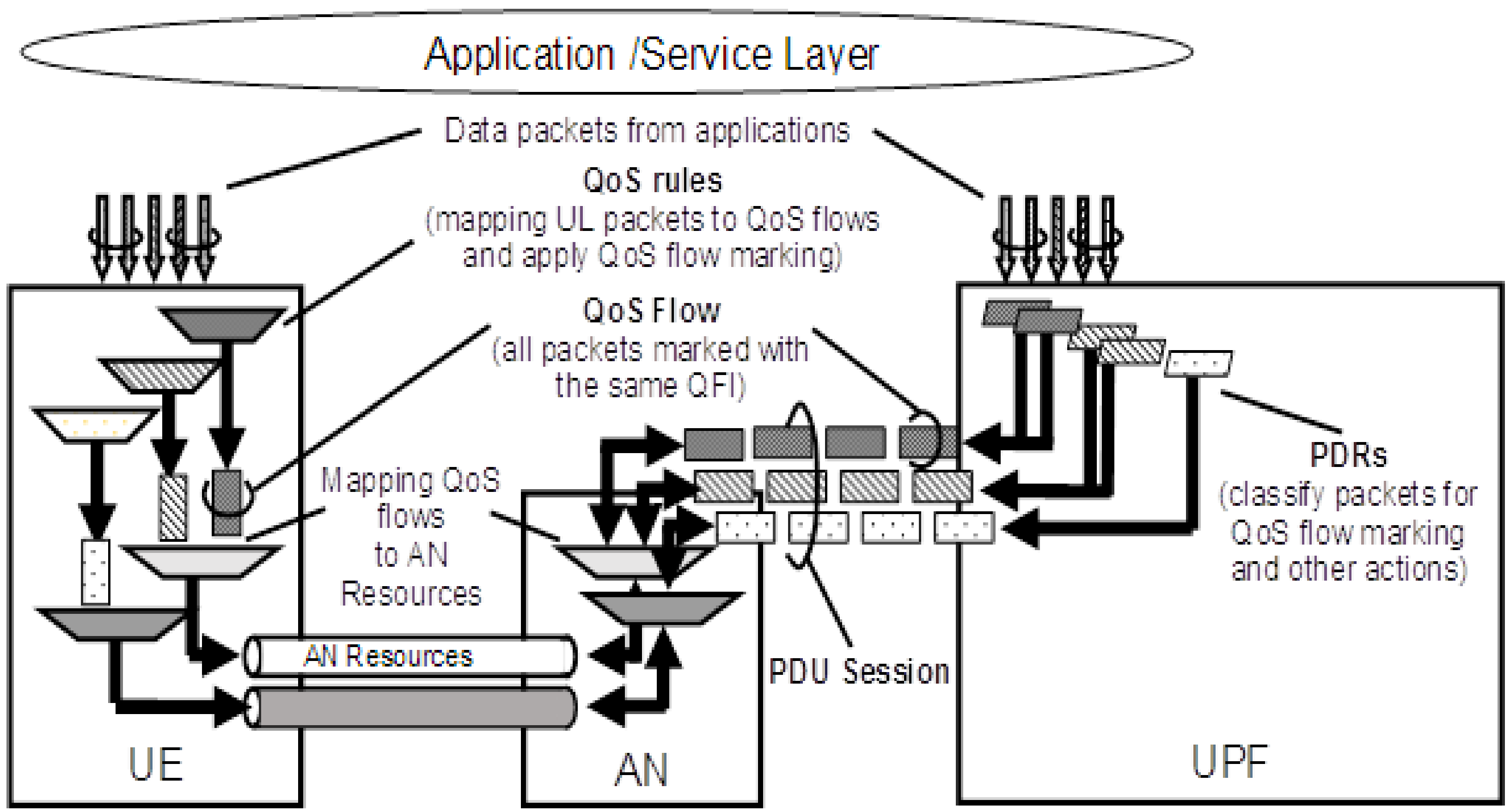
| 5QI Value | Resource Type | Default Priority Level | Packet Delay Budget | Packet Error Rate | Default Maximum Data Burst Volume (NOTE 2) | Default Averaging Window | Example Services |
|---|---|---|---|---|---|---|---|
| 7 | | 70 | 100 ms | $10^{-3}$ | N/A | N/A | Voice, Video (Live Streaming) Interactive Gaming |
| 8 | | 80 | 300 ms | $10^{-6}$ | N/A | N/A | Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.) |
| 9 | | 90 | | | | | |
| 69 | | 5 | 60 ms | $10^{-6}$ | N/A | N/A | Mission Critical delay sensitive signalling (e.g., MC-PTT signalling) |
| 70 | | 55 | 200 ms | $10^{-6}$ | N/A | N/A | Mission Critical Data (e.g. example services are the same as QCI 6/8/9) |
| 79 | | 65 | 50 ms | $10^{-2}$ | N/A | N/A | V2X messages |
| 80 | | 68 | 10 ms | $10^{-6}$ | N/A | N/A | Low Latency eMBB applications Augmented Reality |
| 81 | Delay Critical GBR | 11 | 5 ms | $10^{-5}$ | 160 B | 2000 ms | Remote control (see TS 22.261 [2]) |
| 82 | | 12 | 10 ms NOTE 5 | $10^{-5}$ | 320 B | 2000 ms | Intelligent transport systems |
| 83 | | 13 | 20 ms | $10^{-5}$ | 640 B | 2000 ms | Intelligent Transport Systems |
| 84 | | 19 | 10 ms | $10^{-4}$ | 255 B | 2000 ms | Discrete Automation |
| 85 | | 22 | 10 ms | $10^{-4}$ | 1358 B NOTE 3 | 2000 ms | Discrete Automation |

Standardized 5QI to QoS characteristics mapping

Ref:23.501

- ✓ Priority: Used to differentiate between QoS Flows of same UE, and also differentiate QoS Flows from different UEs (1 = highest)
- ✓ Packet Delay Budget (PDB): upper bound for the time that a packet may be delayed between the UE and the UPF
- ✓ Packet Error Loss Rate (PLER): upper bound for the rate of PDUs (IP packets) that have been processed by RLC but that are not successfully delivered by the corresponding receiver to PDCP. Thus, the PER defines an upper bound for a rate of non-congestion related packet losses. The PLER can be used for configuration of appropriate link layer protocol like RLC mode, max HARQ retransmissions, etc.

# 5G QoS flow and mapping



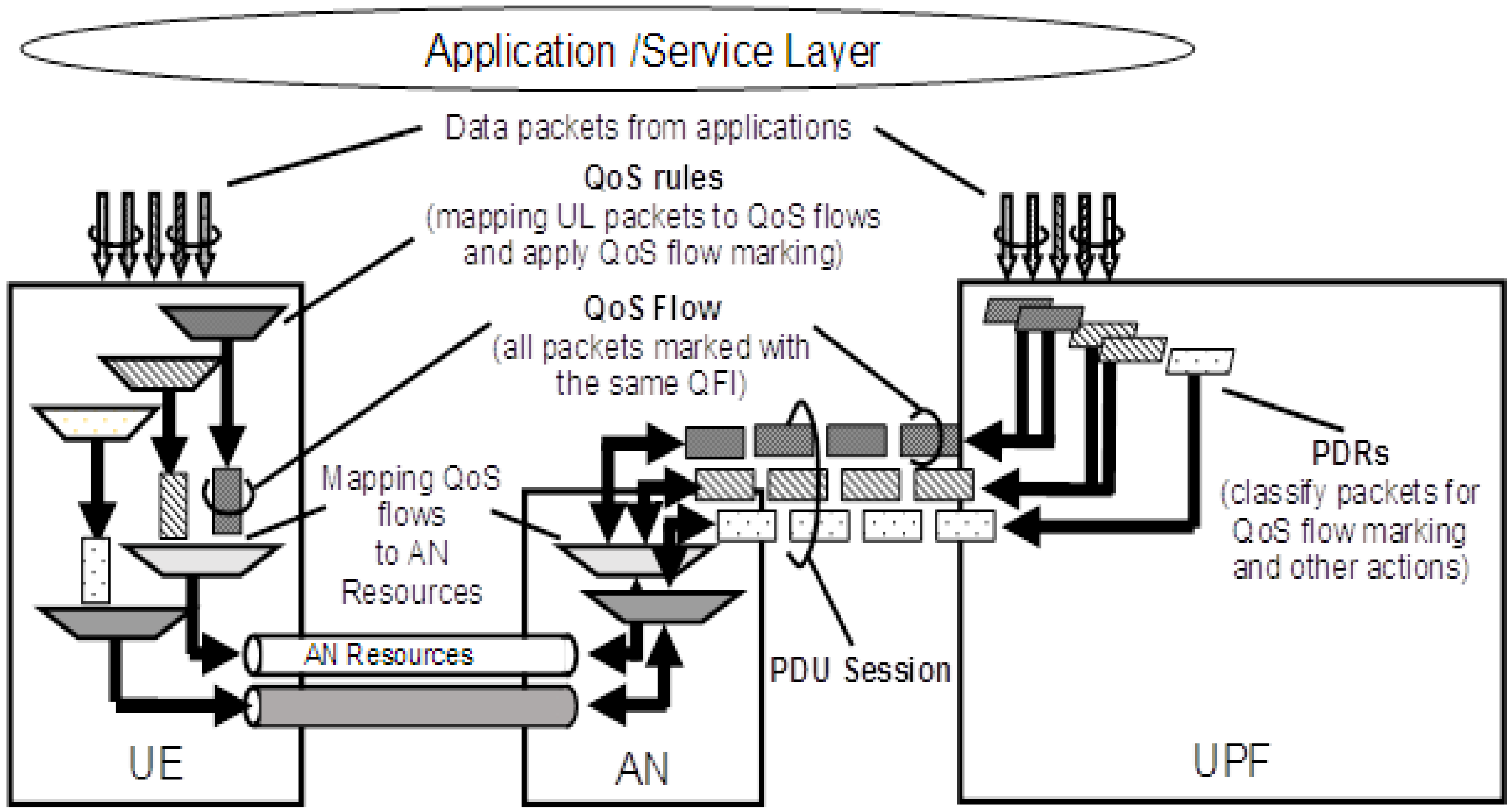

PDR – Packet detection rule

5G QoS flows and mapping

Ref:23.501

- RAN map a guaranteed bit rate (GBR), or multiple GBR flows to the same DRB. The mapping of an E2E session to QoS flows, and DRBs can be updated dynamically
- Different DRBs may be established for QoS flows requiring different packet forwarding treatment (ex: different requirements such as latency budget, packet loss rate tolerance, GBR).
- On the terminal side, the concept of reflective QoS eliminates the need to use dedicated flow filters signaled by the network to match traffic to QoS flows ie terminal derives the mapping of uplink traffic to QoS flows by correlating the corresponding downlink traffic and its attributes
- MAC-level scheduler aims at fulfilling the requirements for the users DRBs, as well as to prioritize accordingly if the system reaches congestion where requirements for all users cannot be simultaneously fulfilled.

# 5G QoS flow – DL traffic



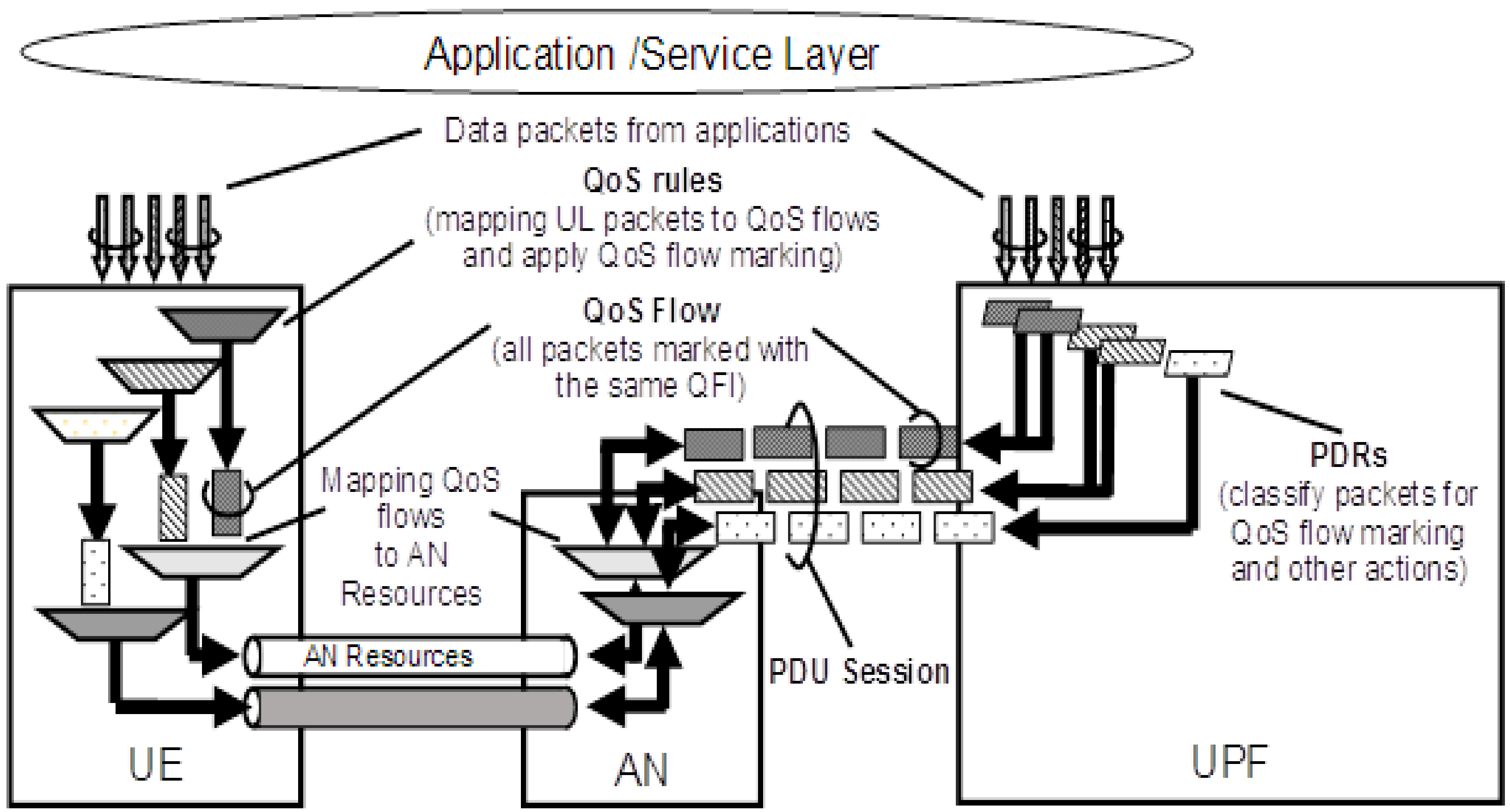

PDR – Packet detection rule

Ref:23.501

5G QoS flows and mapping

1) Incoming data packets are classified by the UPF based on the Packet Filter Sets of the DL PDRs in the order of their precedence
2) UPF performs Session-AMBR enforcement and counting of packets for charging
3) UPF maps User Plane traffic to QoS Flows based on the PDRs.
4) UPF conveys the classification of the User Plane traffic belonging to a QoS Flow using QFI
5) UPF performs transport level packet marking in DL (ex: DSCP) in outer IP header
6) gNB binds QoS Flows to AN resources (i.e. Data Radio Bearers)
    - There is no strict 1:1 relation between QoS Flows and AN resources
    - It is up to the AN to establish the necessary AN resources that QoS Flows can be mapped to, and to release them.
    - The RAN shall indicate to the SMF when the AN resources onto which a QoS flow is mapped are released

# 5G QoS flow – UL traffic




PDR – Packet detection rule

5G QoS flows and mapping

Ref:23.501

1) UE uses the stored QoS rules to determine mapping between UL User Plane traffic and QoS Flows
2) UE marks the UL PDU with the QFI of the QoS rule containing the matching Packet Filter and transmits the UL PDUs using the corresponding access specific resource for the QoS Flow based on the mapping provided by (R)AN
3) (R)AN performs transport level packet marking in the UL, transport level packet marking may be based on the 5QI and ARP of the associated QoS Flow
4) When applicable, (R)AN uses the transport level packet marking value (ex: DSCP) if it is provided by the SMF during the PDU Session Establishment/Modification.
5) UPF verifies whether QFIs in the UL PDUs are aligned with the QoS Rules provided to the UE or implicitly derived by the UE in the case of Reflective QoS).
6) UPF and UE perform Session-AMBR enforcement and the UPF performs counting of packets for charging

# 5G Scheduler (L2 slices)

- MAC scheduler allocates the RBs on per-user basis dynamically for both DL and UL. The key objective of the scheduler is to fulfill QoS service targets for all the DRBs of the served UEs.

- Different services shall map to different set of functionally processing within L2 SW (ie different L2 slices)

- This shall also map to different handling in time-domain scheduling and frequency-domain scheduling of the resources (and maps to Physical layer slices)
  - Smallest time-domain scheduling resolution for the MAC scheduler is mini-slot, but it is also possible to schedule users on slot resolution, or on resolution of multiple slots.
  - So dynamic scheduling with different TTIs is to be considered and this enables the MAC scheduler to match the radio resource allocations more efficiently for different users in coherence with the radio condition, QoS requirements, and cell load conditions
  - Minimum frequency-domain scheduling resolution is 1 physical resource block of 12 subcarriers corresponding to 180 kHz for 15 kHz SCS, 360 kHz for 30 kHz etc

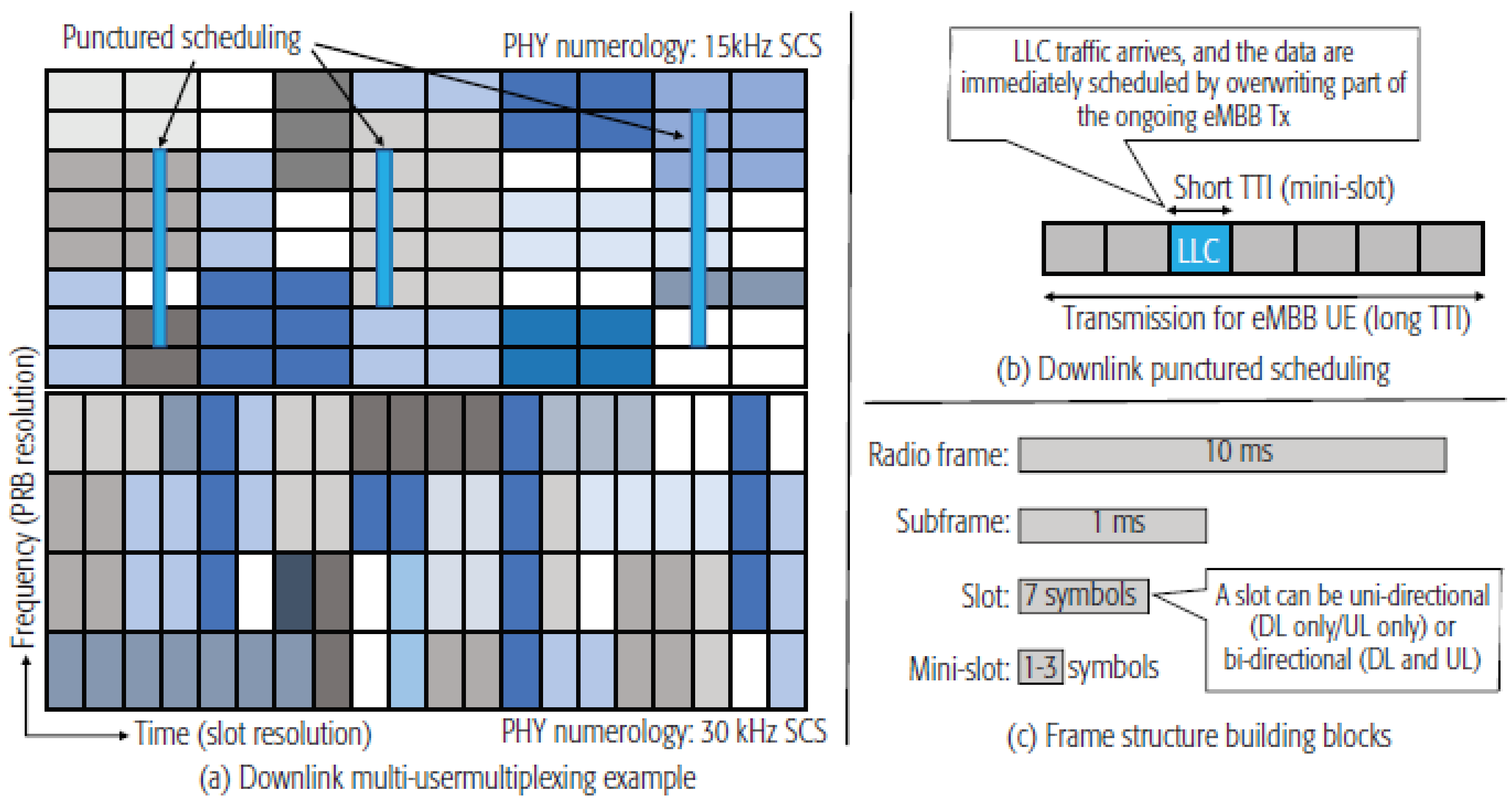
- Efficiency of resource usage and allocation dependent on MAC scheduler design (which is proprietary)

| | | |
|---|---|---|
| eMBB | 5G Upper | -Ciphering<br>-RoHC<br>-ARQ (per segment)<br>-Packet Re-ordering<br>-Duplication detection<br>-Segmentation / Concatenation |
| | 5G Lower | -MAC Multiplexing<br>-Dynamic Beam Management<br>-Dynamic AIV resource mapping<br>-Scheduling (UE)<br>-Logical Channel Prioritization<br>-DRX / DTX<br>-HARQ<br>-RACH |
| URLLC | 5G Upper | -No Ciphering<br>-Optional RoHC<br>-Re-transmission (PDCP level)<br>-Semi-static AIV to resource mapping |
| | 5G Lower | -Fixed size LCP<br>-Optional MAC Multiplexing<br>-Flow-based QoS BSR<br>-Prioritized RACH, Scheduling(UE)<br>-Chase-and-combining HARQ |
| mMTC | 5G Upper | -Optional RoHC<br>-Group-based Scheduling<br>-Semi-static AIV to resource mapping |
| | 5G Lower | -MAC Multiplexing<br>-Scheduling (UE)<br>-Fixed Size LCP<br>-HARQ optimized for coverage<br>-Group-based RACH |

Ref: Service-Tailored User-Plane Design Framework and Architecture Considerations in 5G Radio Access Networks

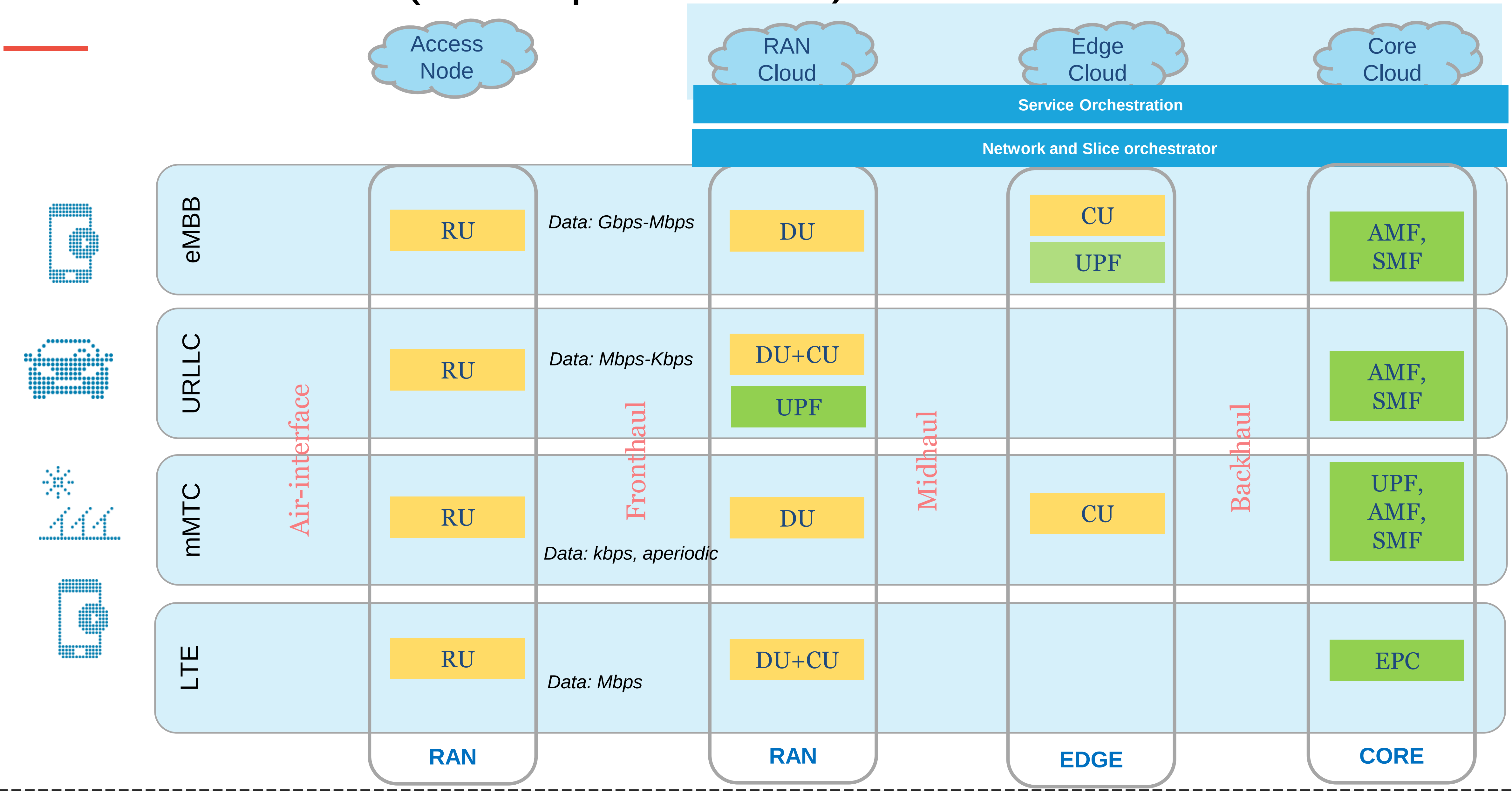
# 5G Scheduler (L2 Slices)

- Schedulers used in 4G can be used for eMBB
- URLLC – priority handling is needed and need to be scheduled immediately to meet latency requirements
  - MAC scheduler immediately transmits it to the designated device by overwriting part of a current scheduled transmission, using mini-slot transmission
  - URLLC payload is transmitted immediately without waiting for current scheduled transmissions to be completed and without the need for pre-reserving radio resources for LLC traffic
  - Option of pre-allocated RBs or contention based handling for UL URLLC traffic




Ref: Agile 5G Scheduler for Improved E2E Performance and Flexibility for Different Network

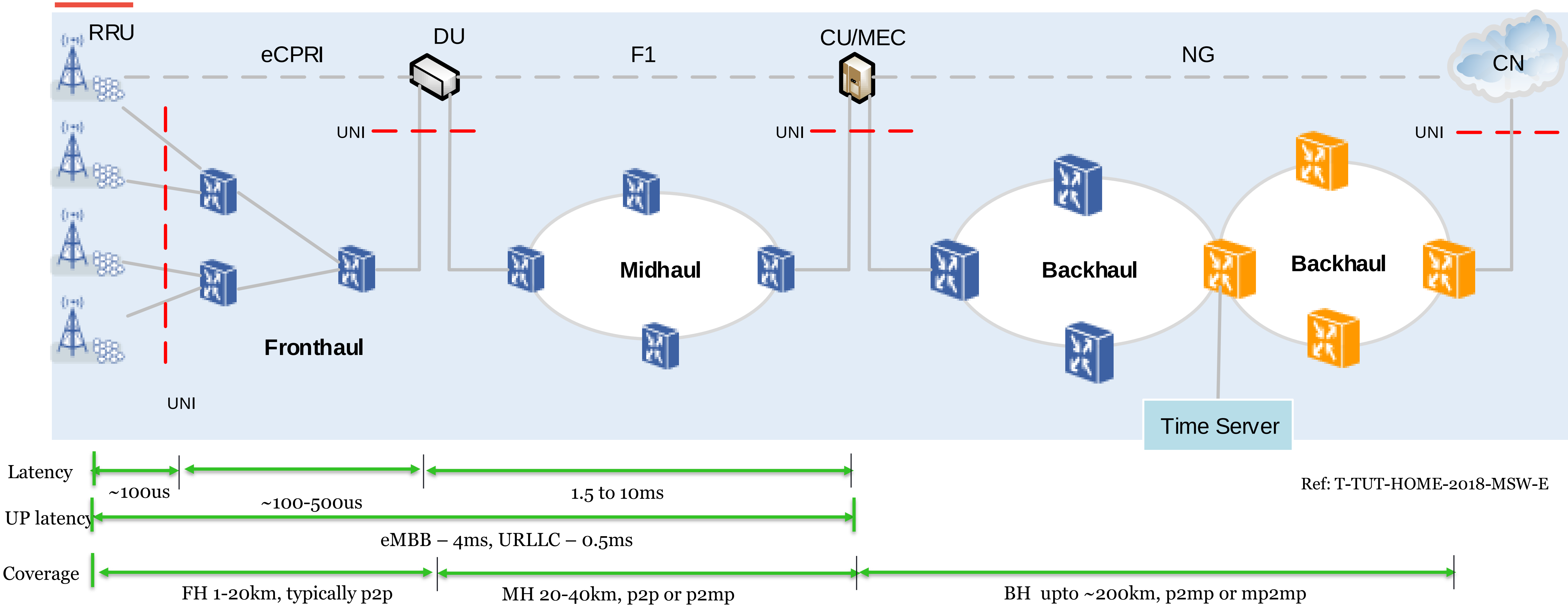
# 5G RAN Network (Slice specific view)




- Network slicing is key concept supported in 5G to be able to address service specific QOS handling (as service slices) and the handling is distributed across all network nodes
- For scenarios requiring low-latency, the UPF functionality of 5G NGC can also move to RAN
- Fronthaul bandwidth shall vary per slice both from data-rate, periodicity/burstiness, and the fronthaul should be able to support statistical multiplexing for efficient utilization
- Fronthaul, Midhaul and Backhaul latency requirements determined based on deployment option (for different slices)

# Fronthaul in 5G

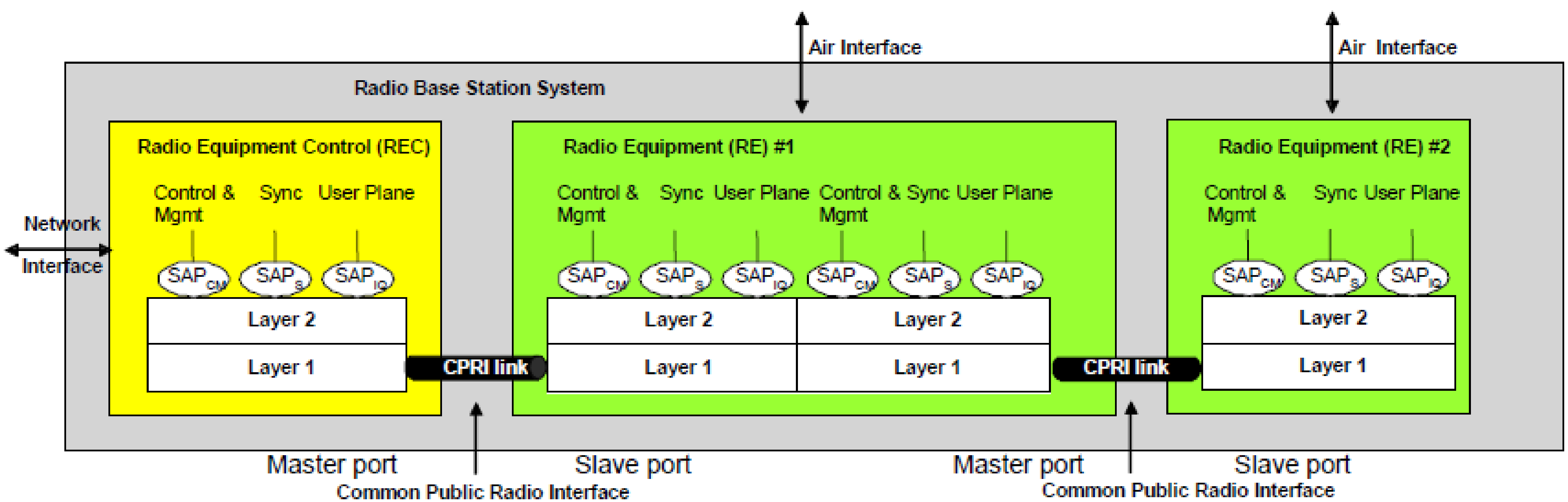
# Typical 5G Transport Architecture




- ✓ Fronthaul – network between RU and DU
- ✓ Midhaul - network between DU and CU (F interface)
- ✓ Backhaul - network between CU and 5G NGC (and EPC)
- ✓ In case CU and DU are co-located then there shall be no midhaul. For small-cell case, there shall be backhaul only (as RU, DU and CU are collapsed into single node)

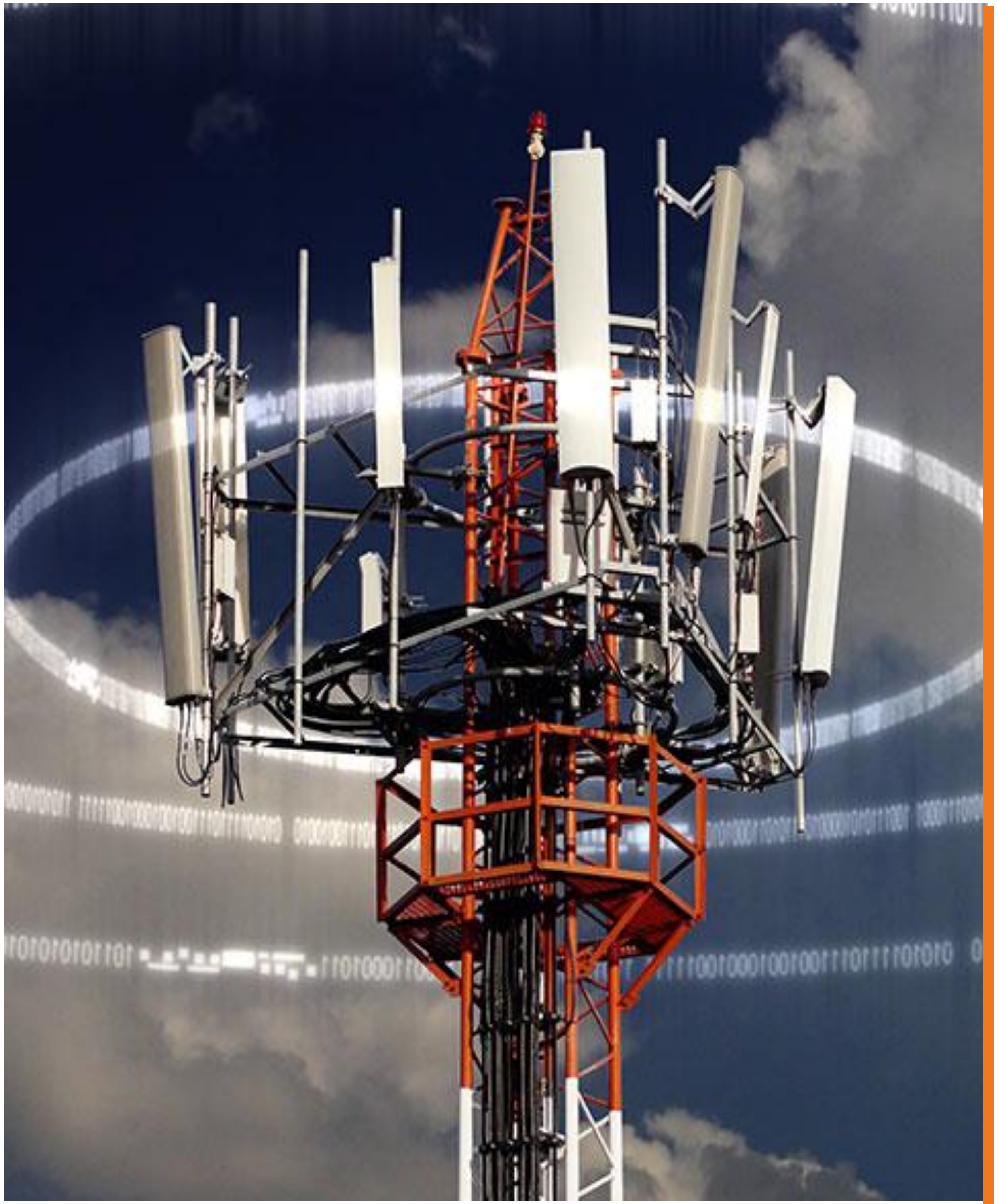
# 4G – CPRI Fronthaul

- CPRI is defined as fronthaul between CU/DU and RU for 4G network, and defines how to exchange radio signal data in form of IQ data between RE and REC
- The IQ data of different antenna carriers (AxCs) are multiplexed by **TDM scheme** onto an electrical or optical transmission line, and link is always "ON" with **Constant bit-rate data**
    - CPRI v7.0 bit rates range from 614 Mbit/s (Rate 1) up to 24.330 Gbit/s (Rate 10)
- Provide time and synchronization information for the Radio Air Interface
- Specified for point-to-point topology and is more antenna dependent (rather than traffic dependent)
- Maximum latency assuming no intermediate nodes
- Multipoint topologies supported but networking management left to application layer




Ref: CPRI specification

# 4G Fronthaul Limitation

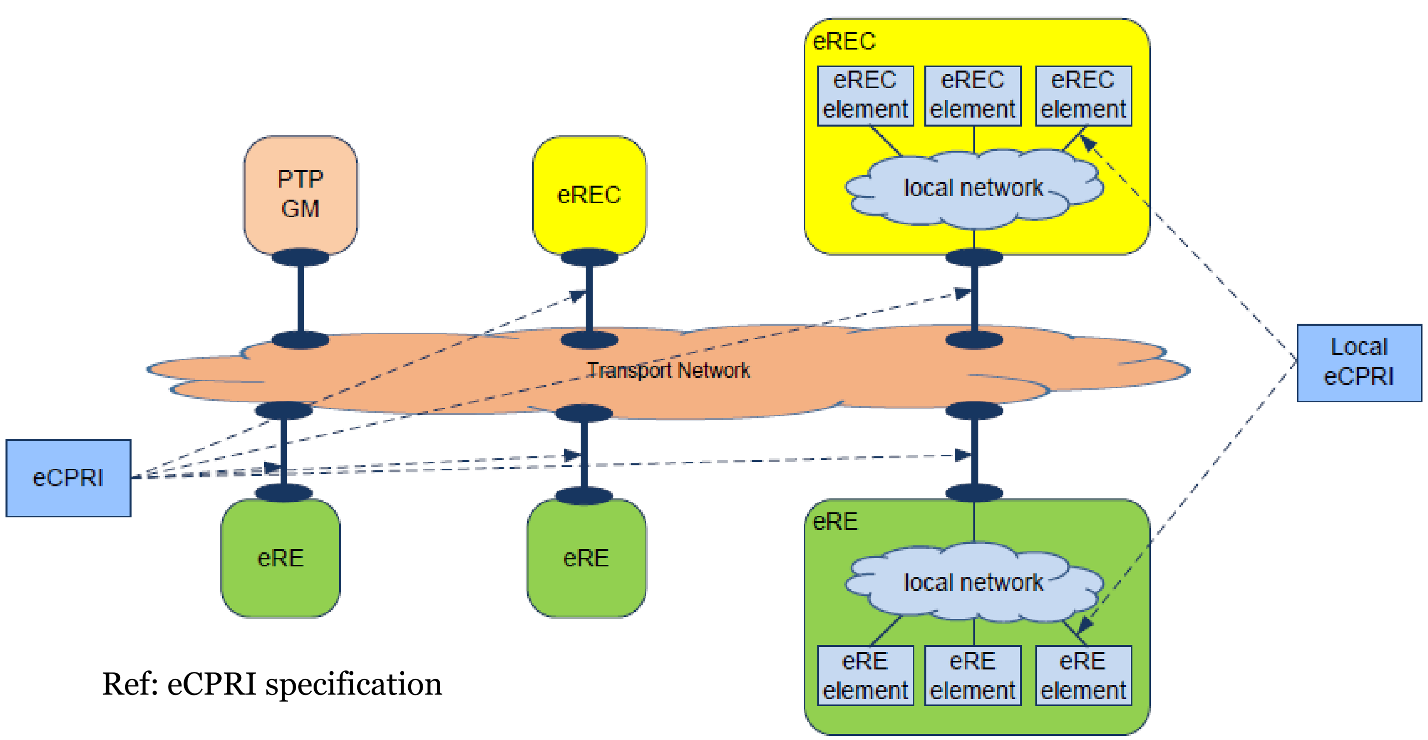


- CPRI is pre-dominantly used in 4G fronthaul. Max data rate supported in CPRI v7.0 is 24.33 Gbps (rate 10)
- For typical LTE scenario of 20MHz, 2x2 DL MIMO, the fronthaul data rate is ~1.96Gbps
- Consider 5G eMBB scenario - *100MHz and 256QAM UL/DL, 8 MIMO layers and 32Tx/Rx, the data rate requirement will be 157.3Gbps*
  - As the number of antenna elements increase, the data rate requirement will further increase
  - *This is much higher than what can be supported in CPRI.* Even with compression, CPRI cannot meet this 5G requirements
  - CPRI cannot handle the high fronthaul bandwidth, even if high-compression is considered (over CPRI)

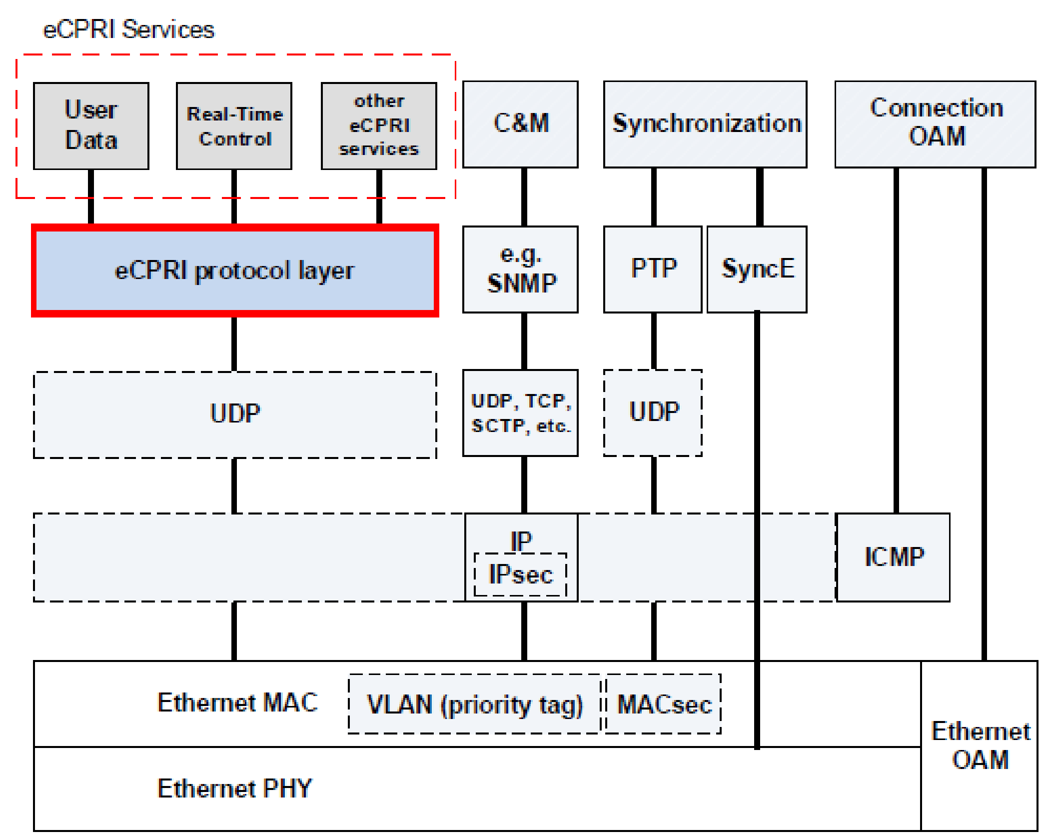
- Option is to consider split point so as to reduce the fronthaul requirements and also define new specification
- eCPRI specification is defined to handle 5G fronthaul datarates
- Flexible deployment split options defined for 5G that provides
  - Flexibility to have multiple deployment split options that can reduce fronthaul bandwidth requirement and also allow flexible and scalable HW implementations
  - Allows for coordination for performance features, load management, real-time performance optimization, and enables NFV/SDN
  - Configurable functional splits enables adaptation to various use cases, such as variable latency on transport

# 5G – eCPRI Fronthaul

- eCPRI is used as fronthaul between CU/DU and RU for 5G network, and defines a protocol for the transfer of user plane information between eREC and eRE via **packet based** fronthaul transport network
- Enables flexible functional decomposition while limiting the complexity of the eRE
  - Supports for Ethernet interface types – 10G, 25G, 40G and 100G
- More traffic dependent rather than antenna dependent
- Typically data is transmitted intermittently rather than having a steady data stream ie link is asynchronous & “bursty” in nature, and Ethernet can handle this with support of statistical multiplexing
- Interface is future proof allowing new feature introductions by SW updates in the radio network
- However, Synchronization of nodes is critical with this interface



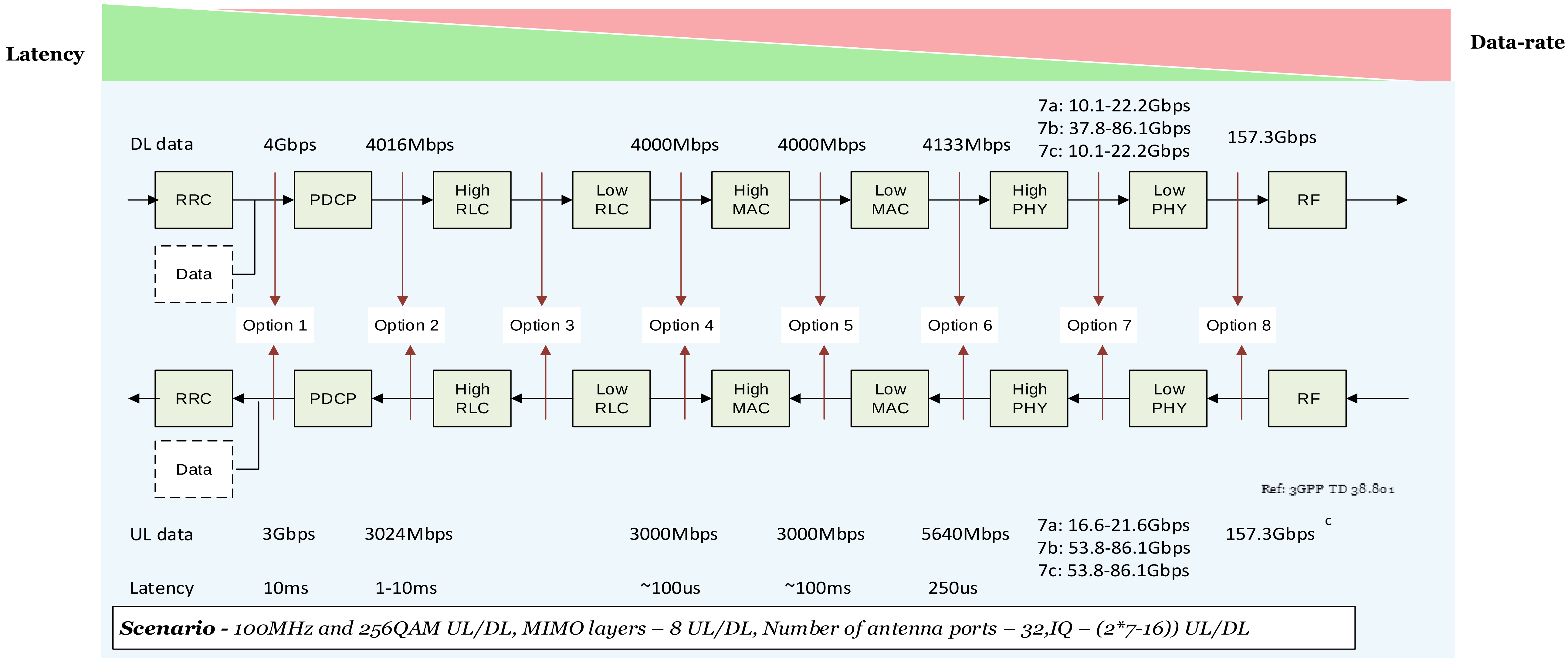

Ref: eCPRI specification




Ref: eCPRI specification

IEEE1914.3 is also alternative option for 5G fronthaul

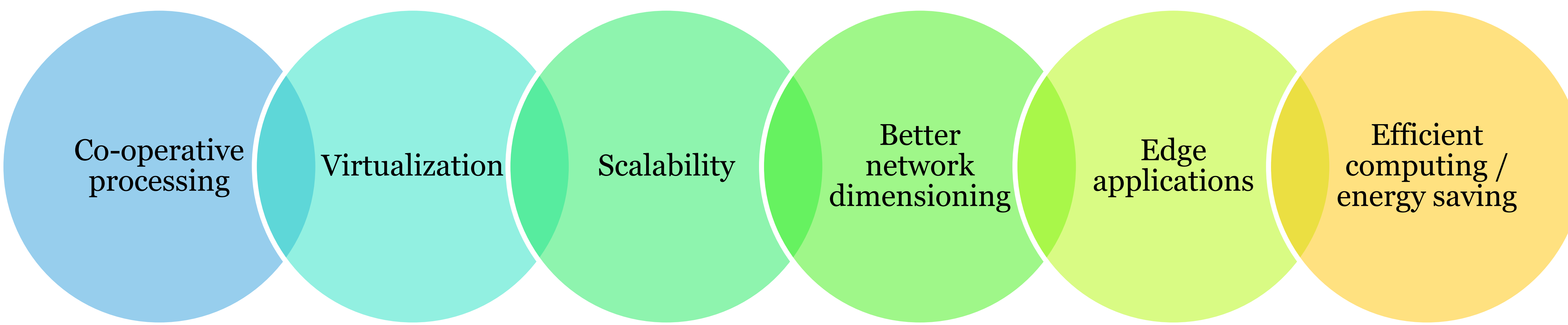
# 5G RAN Split Options




- ✓ Latency requirement becomes stringent and data-rates also increase as we move to option-7/option-8
- ✓ Compression of data reduces the data-rate requirement in Option-7
- ✓ Centralized scheduler possible for options-6 onwards leading to better realization of high-gain coordinated algorithms (like joint scheduling, joint reception, and joint transmission options as part of 5G CoMP)
- ✓ Provides scalable and virtualized architecture options based on CU/DU and RU architecture
- ✓ Typically high-layer split is used for eMBB (>6GHz) and lower-layer split for eMBB (sub-6GHz)

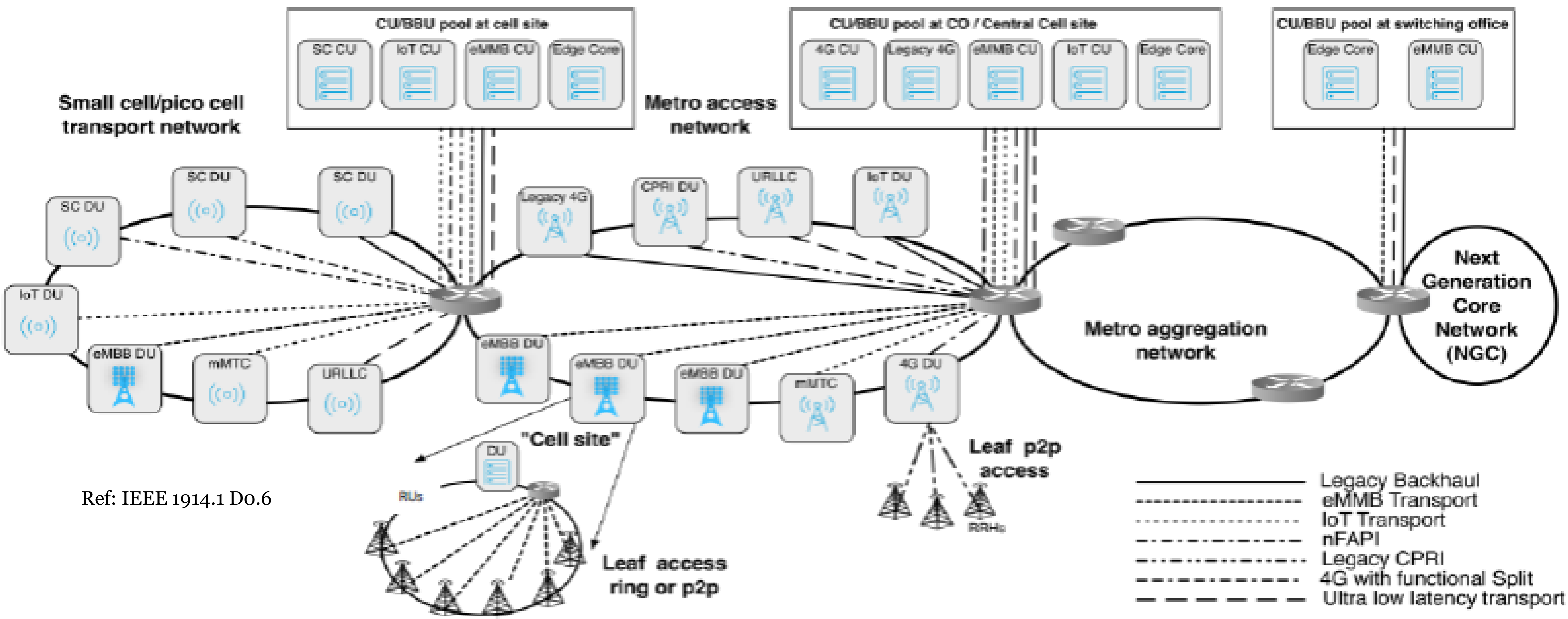
# RAN Split Benefits

- RAN split in deployment depends on number of factors like
  - Radio network deployment scenarios, and user-cases and service requirements (access to edge or centralized services and necessary response times)
  - Specific user density and load demand per given geographical area (which may influence the level of RAN coordination)
  - Transport network capabilities and able to function with transport networks with different performance levels, from ideal to non-ideal
  - Typically high-layer splits enables mmWave eMBB and low-latency scenarios and low-layer split enables high spectral efficiency (sub-6)




  - By using a mix of split points, transport capacity could be minimized and network performance maximized – however dynamic reallocation of functions across split points is complex and can be considered as possible evolution

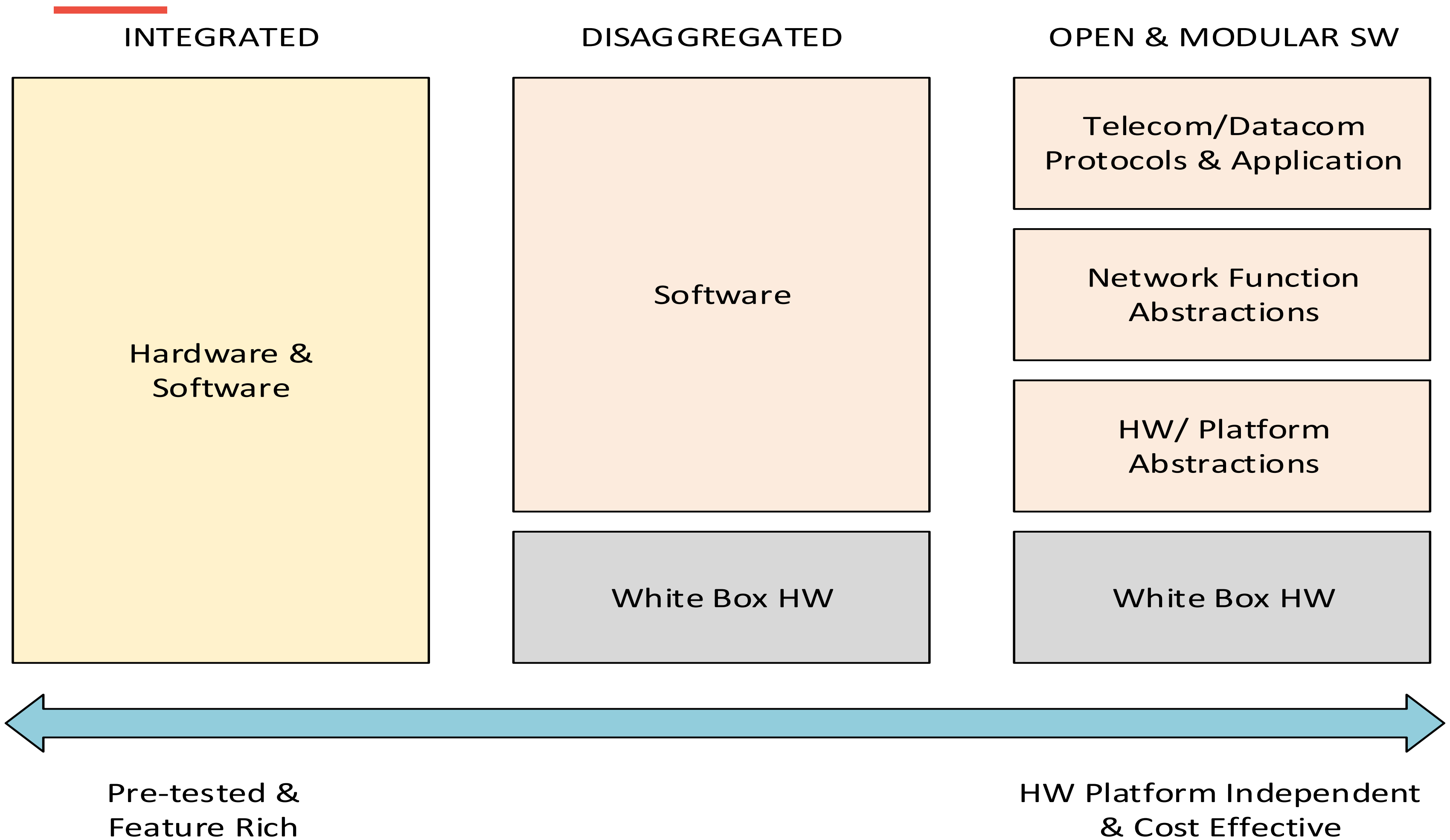
# 5G Network Architecture




Ref: IEEE 1914.1 Do.6

- ✓ Mix of small-cells, CRAN and DRAN base-stations, and mix of LTE eNodeB and gNB nodes
- ✓ Network deployment and MEC placement is dependent on services addressed
- ✓ Fronthaul/backhaul based on deployment and use-cases (date-rates/service requirements)
- ✓ Tree topology is preferred to Ring topology for 5G
    - Transmission latency is accumulated node by node on ring networks. To decrease latency, the processing latency of each node must be greatly reduced, and congestion must not occur.
    - In tree networks, only the latency accumulation between the source and sink nodes needs to be considered, improving the latency tolerance of the entire network

# Evolution towards “COTS” Hardware




OPEN Compute Project
OPEN19 Foundation

- Reference HW solution being proposed in open-source computing hardware can be used for 5G Network solution till DU
- RU may still be proprietary HW
- Open19, OCP and also Intel Rack Sack Architecture are some of the options that can be explored

✓ CU/DU Hardware
  - 5G is accelerating the adoption of **commodity HW**, disaggregated solution within Telco’s.
  - Limited SOC options available for 5G PHY (Ceva). CPU based (x86 or ARM) solution with FPGA used
  - ASIC based solutions for baseband will come into picture once the solutions are verified

✓ RU Hardware
  - 5G RU designs will be **“inherently intelligent”.** Part of PHY runs in RU and also handling for digital beam-forming functionality
  - This will also have challenge wrt some of the key considerations of RU design like size, weight, and power

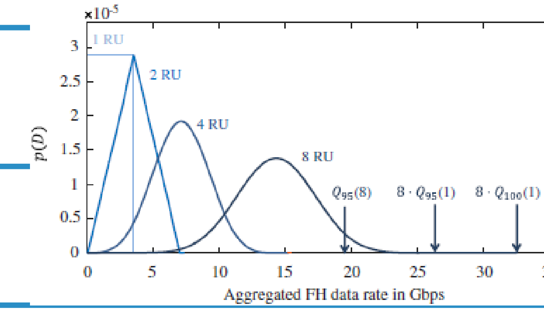
# Fronthaul Transport

# 5G Fronthaul Transport Requirements

Handling of very high data-rate requirements

Handling of traffic and for different service types (or slices) with varied priorities (include packet prioritization) and quality of service over a unified network

Flexibility to scale the bandwidth based on user plane traffic




Statistical multiplexing for aggregating traffic from multiple sites

Should be traffic dependent and NOT be antenna dependent

Support for multiple network architectures

Meet stringent Synchronization and Timing requirements for 5G

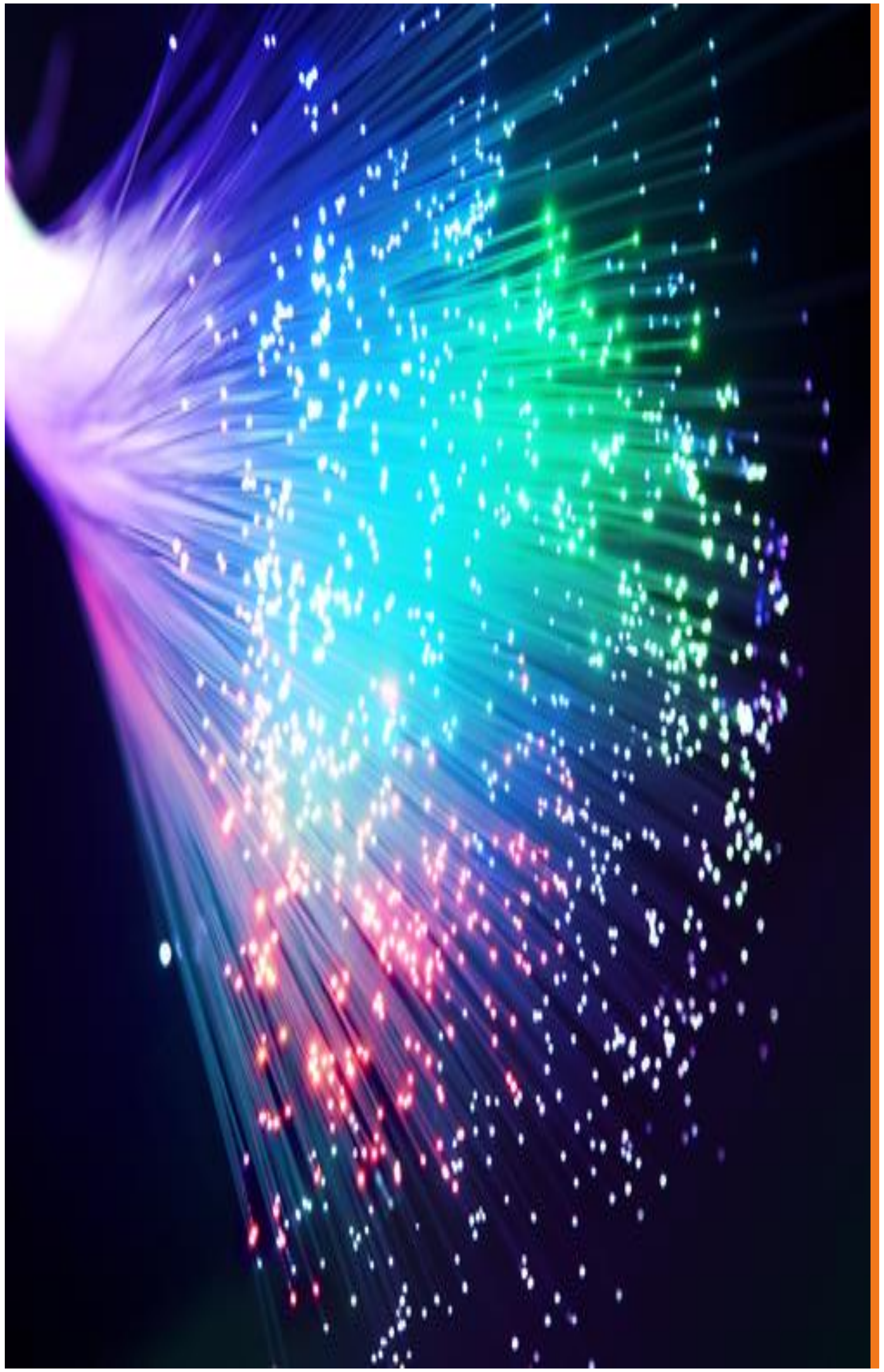
Cost / Performance Trade-off by selecting proper FH/BH suitable for the network deployment

- ✓ Mix of transport technologies – optical, packet, microwave
- ✓ Ethernet based solutions provide - Reuse of existing infrastructure, flexibility, statistical multiplexing, flexible routing
- SDN is a key enabler for converged FH/BH networks in 5G to virtualize the transport network to support slicing and allow a flexible deployment of virtual functions in different places of the network
- GNSS, 1588, PTP are some of the synchronization options

*p(D) = offered traffic can be transported without queueing with a probability*

Ref: 5G transport network requirements for the next generation fronthaul interface

# Transport methodologies - Optical



PON

- uses multiple different wavelengths (WDM) over a physical point-to-multipoint fiber infrastructure that contains no active components
- use of different wavelengths allows for traffic separation within the same physical fiber ie provides logical point-to-point connections over a physical point-to-multipoint network topology

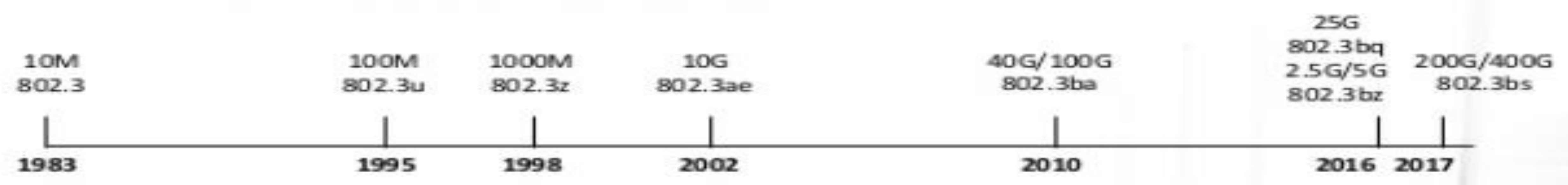
Active/Passive WDM

- Passive WDM – has no active components and CWDM optical modules directly installed in the RRH and BBU, and so can be placed directly on cell-site
- Active WDM – requires external transponders and so requires power supply, and can fit more wavelengths in single fiber
- Limited BW (Nx10G), and need to have different WDM colours for different clients
- Cannot meet 5G requirements

G.709 Optical Transport Network (OTN)

- method for transparent transport of services over optical wavelengths in WDM systems
- Standards based multiplexing (Static)
- time multiplexing of signals on same wavelength - can save fiber
- Mature and robust OAM

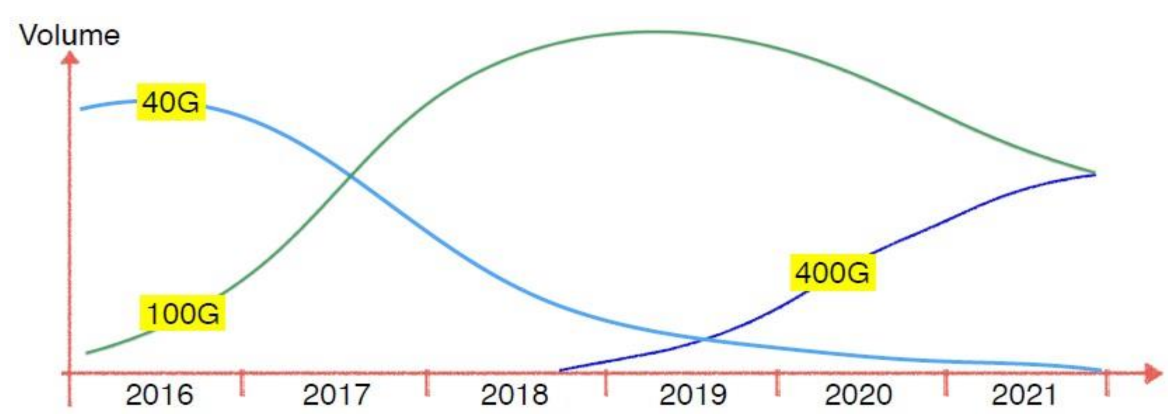
# Is the Future Transport Ethernet 400G?




## Scaling up with OTN

- OTN natively defines how to frame a number of different protocols into OTN frames, it is more suitable than Ethernet for transport of legacy services.
- Carrier Ethernet, it now offers the same level of OAM functionality as OTN.
- OTN with static multiplexing capability matched by Ethernet with IHON
- Support for FEC for both OTN & 100G & above optics
- Support Timing & Synchronization via TSN, IEEE1588

Ethernet is currently a beneficial choice for mobile transport, access and metro networks while for the time being OTN is defined for high bitrate long-haul transport, but this could eventually change once IEEE chooses to define long-haul Ethernet interfaces.




Note: 400G port numbers includes both 8x50G and 4x100G implementations

| Lane Speed | 10Gbps | 25Gbps | 50Gbps | 100Gbps | |
|---|---|---|---|---|---|
| 1X | 10G | 25G | 50G | 100G | Server Interface |
| 2X | — | 50G | 100G | 200G | |
| 4X | 40G | 100G | 200G | 400G | Leaf-Spine Interface |
| 8X | — | — | 400G | 800G | |
| Availability | 2010 | 2015 | 2018 | 2020 | |

Andy Bechtolsheim 5 Years 3 Years 2 Years

Ref: https://www.nextplatform.com/2018/02/21/road-400g-ethernet-paved-bechtolsheims-intentions/

# Transport methodologies - Wireless

Unlicensed TVWS (600– 800 MHz):

- The unused TV spectrum, a suitable option for sparsely populated areas ,
- Allows NLOS connectivity (1-5km) , but the data rates be may not be sufficient for backhauling 5G networks.

The licensed sub-6 GHz spectrum (800 MHz – 6 GHz)

- Is suitable for backhauling SBSs in rural as well as urban areas (1.5-2.5km
- Suffers from severe interference, spectrum is costly and provides narrow BW's
- Not recommended for backhauling 5G small cells in dense urban areas.

Microwave (6– 60 GHz):

- Common choice for backhauling SBSs in urban and rural area.
- Can provide 1 Gbps + data rates for real-time and non-real-time services
- Spectrum is costly and requires antenna alignment for LOS communication to achieve high directivity gain.

Unlicensed millimeter wave (60– 90 GHz):

- Promising candidate for backhauling 5G small cells in dense urban areas.
- Suffer from high propagation loss. However, this may work for densely deployed small cells, in turn for increased capability of frequency reuse.
- Extremely high frequency signals enable the installation of a large number of antenna elements within a small area which enable high antenna array gain.
- The wider channel bandwidth offered by mmWave spectrum and high directivity can provide multi-Gbps data rates. Requires antenna alignment and LOS connectivity, there is no additional spectrum cost.

# Summary

- Network slicing impacts e2e architecture
- Efficiency of MAC Scheduler  efficient allocation of RBs to various slices in UL/DL is critical
- Network architecture and CU/DU split dependent on network slices and services offered
- Transport used for Fronthaul should take care of dynamicity of traffic handling of different slices
- Choice of Fronthaul methodology is based on network engineering and cost/roadmap considerations
- Usage of different “non-ideal” fronthaul methods cannot be offered for low-latency scenarios
- Network architecture moving towards “open” platforms

# Thank You

**Headquarters**

3979 Freedom Circle
Santa Clara, CA  95054

www.aricent.com