DetNet
Internet-Draft
Intended status: Informational
Expires: 3 April 2026

B. Varga
J. Sachs
Ericsson
F. Duerr
University of Stuttgart
S. Mostafavi
KTH Royal Institute of Technology

30 September 2025

# Latency analysis of mobile transmission
# draft-varga-detnet-mobile-latency-analysis-03

## Abstract

Dependable time-critical communication over a mobile network has its own challenges. This document focuses on a comprehensive analysis of mobile systems latency in order to incorporate its specifics in developments of latency specific network functions. The analysis provides valuable insights for the development of wireless-friendly methods ensuring bounded latency as well as future approaches using data-driven latency characterization.

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 3 April 2026.

## Copyright Notice

Table of Contents

1. Introduction . . . . . 2
2. Comparison between a wired and a mobile virtual DetNet router . . . . . 6
3. Mobile Transmission Latency Breakdown . . . . . 7
3.1. Mobile communication targets . . . . . 7
3.2. Transmission Latency Breakdown . . . . . 8
3.3. QoS architecture within the mobile network . . . . . 9
3.4. Latency contributions in different layers of radio protocols . . . . . 11
3.5. Latency Analysis . . . . . 14
3.5.1. Processing delays in gNB and UE . . . . . 15
3.5.2. Traffic handling and queuing . . . . . 15
3.5.3. Data transmission over the radio interface . . . . . 15
3.5.4. Wireless transmission reliability . . . . . 16
4. Example: Observed characteristics in real network . . . . . 18
5. Scheduling related future work . . . . . 20
6. Summary . . . . . 21
7. Acknowledgements . . . . . 22
8. References . . . . . 22
Authors' Addresses . . . . . 24

## 1. Introduction

Digital transformation of industries and society is resulting in the emergence of a larger family of time-critical services with unique requirements distinct from traditional Internet applications. Such time-critical communication has in the past been mainly prevalent to wired communication system, which is limited to local and isolated network domains. Wireless communication provides flexibility and simplicity, but with inherently stochastic components that lead to packet delay distributions metrics exceeding significantly those found in wired counterparts. These deviations of stochastic characteristics have to be addressed in Deterministic Networking (DetNet) [RFC8655].

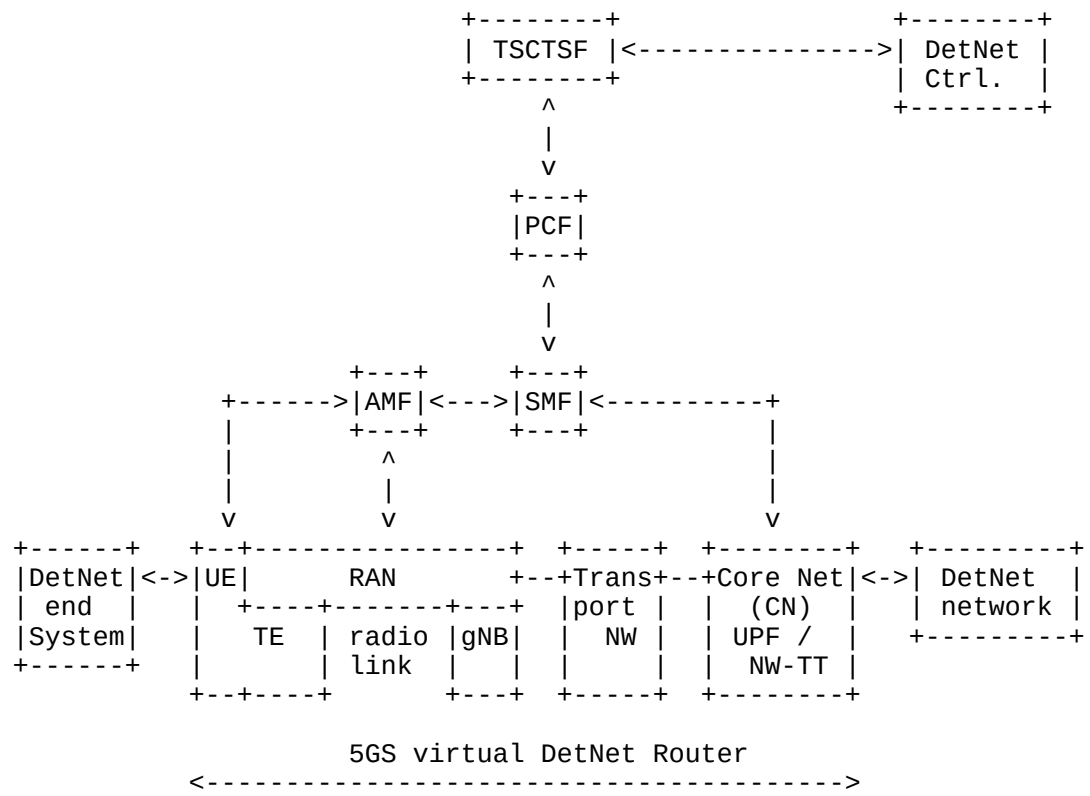
The 5G mobile communication system is specified in the Third Generation Partnership Project (3GPP) and it supports communication with unprecedented reliability and very low latency through the Ultra-Reliable Low Latency Communications (URLLC) enhancements introduced in Release 16. URLLC features targeted reliability,

latency and QoS (e.g., automatic repetitions, antenna techniques, robust physical channels, Orthogonal Frequency Division Multiplex (OFDM) numerology, mini-slots, grant-free access, pre-emption, 5G QoS identifier (5QI) values for multiple time-critical services, QoS monitoring). Providing synchronization and exposure functionality are covered as well.

DetNet support started in Release 18 based on the concept developed for Time-sensitive Networking (TSN) in former releases. The 5G system is represented in the end-to-end architecture as a set of virtual DetNet routers. The 5G network comprises a 5G core network and a Radio Access Network (RAN). A User Plane Function (UPF) of the 5G core network acts as a gateway towards the DetNet network. The RAN can span over a wider geographical area to provide wireless connectivity to one or more User Equipment (UEs).

Note: In general bridging/routing service is out-of-scope for 3GPP specifications, therefore in real network scenarios bridging and routing are for example implemented by additional (external) functions located mainly within or next to the UPF.

Note2: According to TSC (Time Sensitive Communication) concept of 3GPP the whole 5GS is presented towards the wired DetNet network as a virtual DetNet router. Using DetNet technology between the 5GS components is not precluded but out-of-scope in this document.




SMF: Sessions Management Function
AMF: Access & Mobility Management Function
PCF: Policy Control Function
TSCTSF: Time Sensitive Communication and Time Synchronization Function
UE, TE: User, Terminal Equipment
UPF: User Plane Function
RAN: Radio Access Network
gNB: Base Station
NW-TT: Network-side TSC Translator

Figure 1: Internal components of the 5G system acting as a virtual DetNet router

Figure 1 shows the interconnection of the DetNet nodes and the 5G network. The Time-Sensitive Communication and Time Synchronization Function (TSCTSF) connects the DetNet Controller and the 5G control plane. The TSCTSF collects information from the 5G system, and it reports to the DetNet Controller. The DetNet Controller configures the 5G as a virtual DetNet router through the TSCTSF, which maps parameters and sets the configuration via the 5G control plane. Data plane connectivity at the UPF is achieved via the TSC Translators (TT) on the network-side at the UPF (NW-TT). Using a TT function on the device side (DS-TT) is optional (e.g., if time synchronization has to be provided).

The interaction between the DetNet controller and the 5G system is specified in [M23.501] e.g., for the support of periodic deterministic communication. It describes how the a-priori traffic pattern characteristics in the downlink and the uplink direction could be provided from an external network into the 5GS and used by the NG-RAN to optimize resource utilization and to lower the latency and latency variation. The TSC Assistance Information (TSCAI) feature is described as a method how the QoS flow traffic characteristics could be transferred within the 5GS. The TSCAI feature can be helpful e.g., in scenarios where there is an offset between the traffic burst sending times and the reserved resources on the air interface, a mismatch between the periodicity of traffic and scheduled resources, a clock drift of 5GS with a reference to the clock of an external network, or in a combination of these cases. Note: 3GPP systems do not support directly the MPLS data plane of DetNet due to the lack of support for MPLS. DetNet IP data plane is supported via the IP PDU session.

Wireless system and its external interfaces are by nature distributed and with dynamic variations due to radio propagation. The radio transmission suffers from interferences, reflections, scattering and diffraction that affect the reliability of data communications which results in high variable forwarding latency, see a deeper review in [NR-5G].

There are multiple extension directions to overcome the limitations inherited by wireless systems, especially 3GPP ones. The common characteristics of them are that they provide a wireless-friendly toolset to achieve the required latency distribution between the endpoints. The latency analysis described here is intended to help the developers of such wireless-friendly toolsets and provide motivation for new approaches as well. Such new approaches can be based on the predictability of the system, for example via usage of data-driven latency characterization, where network entities have the ability to estimate the evolution of a system metric or state in the future.

Note: this document was written in order to support DetNet WG related discussions but it can be interesting for non-DetNet discussions as well.

## 2. Comparison between a wired and a mobile virtual DetNet router

The same 5G network can form multiple virtual routers, each of which is realized via the UPF instance in the 5G core network. An UPF configured for DetNet support and all UEs connected to that UPF with IP PDU sessions jointly form the virtual 5G DetNet router and its ports. There exist significant differences in the characteristics of such a virtual and a legacy wired DetNet router [D6G-D2.1]:

* Physical distance of ports: In a wired router the physical distance between ingress and egress ports is in the order of a decimeter. In a virtual 5G router the distance is between the UPF and the UE, or between two UEs and can be up to 100’s of meters or even kilometers. This can remarkably impact on network topology. For example, in an industrial wired DetNet network connecting two end points may require many 10’s of hops to be traversed for E2E connectivity. With a 5G virtual router only few hops are needed (e.g., 1-2 (or up to a few) hops to reach the 5G ingress (UE or UPF) and 1-2 (or up to a few) hops to reach the end node from the 5G egress (UE or UPF).

* Number of ports: The number of router ports in a wired router is decided at the design and production of the router; router ports are at fixed locations (in the chassis of the router). In the virtual 5G router, the number of ports depends on the number of UEs connected to the UPF that outlines the virtual router. If a new UE connects to the UPF the number of ports owned by the virtual router increases. This new UE may require interaction(s) with the DetNet Controller (e.g., reporting latency to/from the new port, updating router configurations).

* Latency characteristics: The latency performance of a wired router is in the single-digit microsecond range, with a PDV in the range of some 100’s of nanoseconds. In a wireless router the typical latency values are in the range of milliseconds (without specific configurations for low latency bounds they can reach up to some 10’s of milliseconds). Even by using URLLC and proper DetNet configuration the PD and PDV of a 5G virtual router is substantially larger than for a wired router.

* Dynamicity of characteristics: Characteristics of a (wired) DetNet router are mostly determined at design and production time. A wired router that is tested in a lab prior to normal deployment can be expected to behave in the same way during operations as

during the lab test. In contrast, for a wireless system - and a virtual 5G DetNet router - the performance depends on the radio environment and deployment. So, the characteristics of the virtual 5G router are determined during the operation phase. With a well-planned and deployed RAN, the general 5GS performance is expected to perform according to requirements, but in case of major changes in the radio environment (e.g., walls or large blocking installations being added) changes in the performance might occur.

While the 5GS has been specified to be compatible to DetNet by its external interfaces, the differences in characteristics of the virtual 5G router and a wired DetNet router requires the development of new wireless-friendly solutions, those are able to efficiently ensure bounded latency in mixed (wired and wireless) DetNet scenarios.

## 3. Mobile Transmission Latency Breakdown

### 3.1. Mobile communication targets

In traditional mobile communication networks, the primary key performance indicators of interest have been the achievable data rate and spectral efficiency. In 5G, latency has been added as a further key performance indicator by URLLC. The ambition of 5G URLLC was to provide low-latency communication while providing high reliability for maintaining the latency below a specified latency bound. For example, the objective for the 5G standard is to guarantee that a RAN latency of 1 ms can be achieved with 99.999% probability.

A solution for reliable wireless transmission with high spectral efficiency is to apply Hybrid Automatic Repeat Request (HARQ) retransmissions to recover from unsuccessful transmissions. However, HARQ leads to an increase in latency due to multiple transmissions causing a notable disturbance in the packet delay distribution. URLLC has introduced two major sets of tools: (i) reducing the radio transmission structure for lower latencies (e.g., processing delays, channel access delays), and (ii) providing higher robustness in the transmission to achieve the same latency reliability with fewer transmission attempts, at the costs of reduced spectral efficiency due to extremely conservative transmission modes.

To give an example, an uplink transmission in a millimeter wave carrier can be made in two different configurations [FGS15]:

* Normal 5G New Radio (NR) configuration with up to 3 retransmissions for reliability with packet delay from ~500 us to 2.8 ms, with low resource usage,

* 5G URLLC NR configuration with single-transmission reliability with packet delay from ~500 us to 900 us, involving high resource usage.

Furthermore, very low latencies enabled by URLLC require a thorough network deployment plan (e.g., location and density of base station antennas) to ensure that the capabilities are available throughout the service area. More relaxed latencies are less sensitive to the radio network design.

5G URLLC is the main enabler to support time-critical communication standards that have been defined for fixed networks, like IEEE 802.1 TSN and IETF DetNet.

## 3.2. Transmission Latency Breakdown

Generally, the latency contributions in a 5G network are dominated by the RAN [D6G-D3.1]. The transport network only plays a role if a UPF is far away from the gNB; the amount of packet processing at the UPF (and related processing times) is limited in comparison to RAN.

In the 5G RAN the main latency contributors are:

1. Time-domain reliability based on HARQ

2. Mobility with handover interruptions

3. Time-division duplex structure

4. Congestion due to resource sharing and queuing

HARQ allows for a better utilization of the resources while being robust for a defined loss bound. Retransmissions inherently contribute to the latency of the packet with defined probability of retransmission(s). HARQ should be used as reliability tool, in case that it is permitted by the latency bound; it is a tool that combines high reliability with spectral efficiency (at the cost of increased PDV). Reliability can be achieved without HARQ, by using more robust transmission modes. If a (low) latency bound is provided with 99.999% reliability by a robust single transmission, then the large majority of (i.e., 99.99%) of the packets are over-protected with too high resource allocations in order to ensure that also the worst-case packets mostly achieve the latency bound.

Mobility is ensured by handover, where a UE switches connection from one base station to another, which can lead to handover interruption times. There are tools to minimize this impact, e.g., L3 make-before-break handover where the resources are allocated and ready

before performing the handover, L1 or L2 mobility with multiple transmission-reception point (multi-TRP), multi-connectivity. These options are dependent on deployment and spectrum.

Time Division Duplexing (TDD) pattern is sometimes prescribed by national regulation and subject to harmonization of multiple networks. This can place restrictions on applicable configurations. Each TDD pattern introduces at least PDV at transmission time interval (TTI) level since packets need to wait for their time slots to be transmitted.

When the network is undergoing congestion at high loads, the opportunities for transmission are restricted and, consequently, additional delay is experienced by the packet. Possible solutions are to apply prioritization, resource partitioning, admission control, traffic policing, reservations, or preconfigured access. In most cases there are implications for the implementation, as well as utilization inefficiencies.

### 3.3. QoS architecture within the mobile network

The packet delay of individual packet is strongly dependent on how the packet is handled within the mobile network. Different packets are treated differently according to the service requirements they are associated with. This allows to provide latency-optimized treatment for dependable time-critical services by applying the Quality of Service (QoS) mechanisms of the mobile network. The handling of QoS for traffic passing through the 5G network is defined in the 5G QoS framework [M23.501][FGAQoS], as summarized in Figure 2. The end-to-end traffic flows passing through the 5G network, denoted as service data flows, are mapped at the ingress to the 5G system at the UE and UPF to QoS flows via traffic filter rules. The QoS flow is the finest level of granularity for specifying the service specific traffic treatment in the 5G system. Each QoS flow can have different traffic forwarding treatment configured in the network, according to the defined QoS requirements.

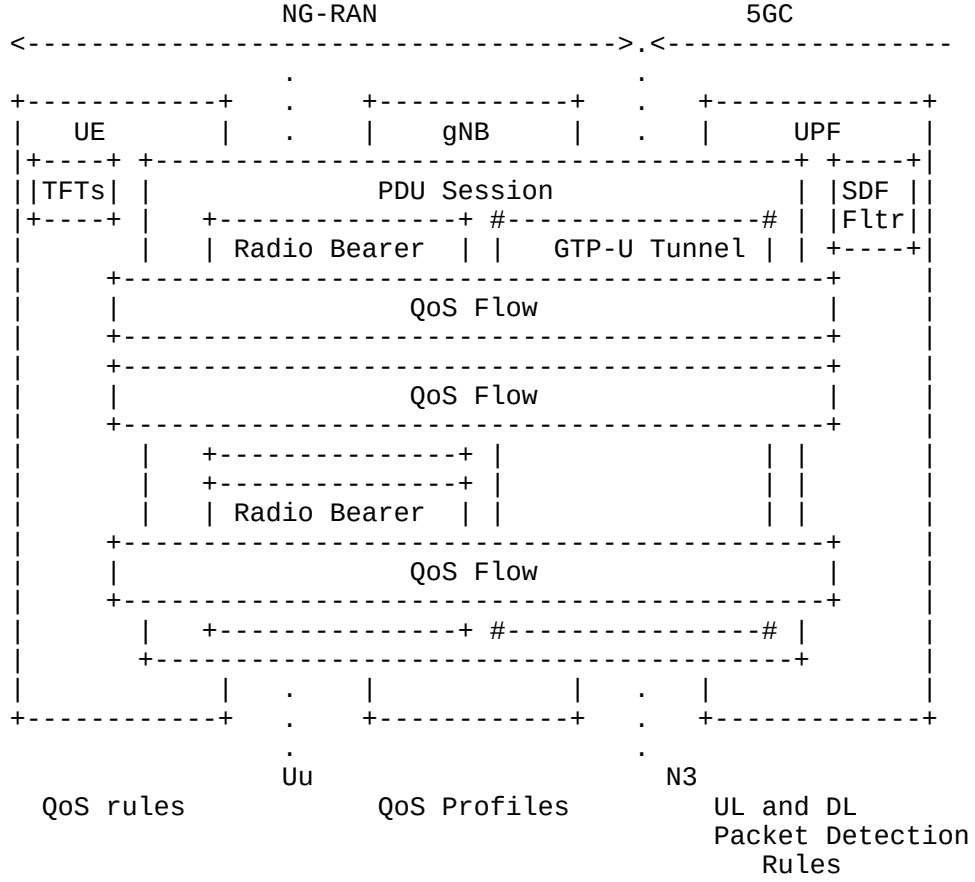
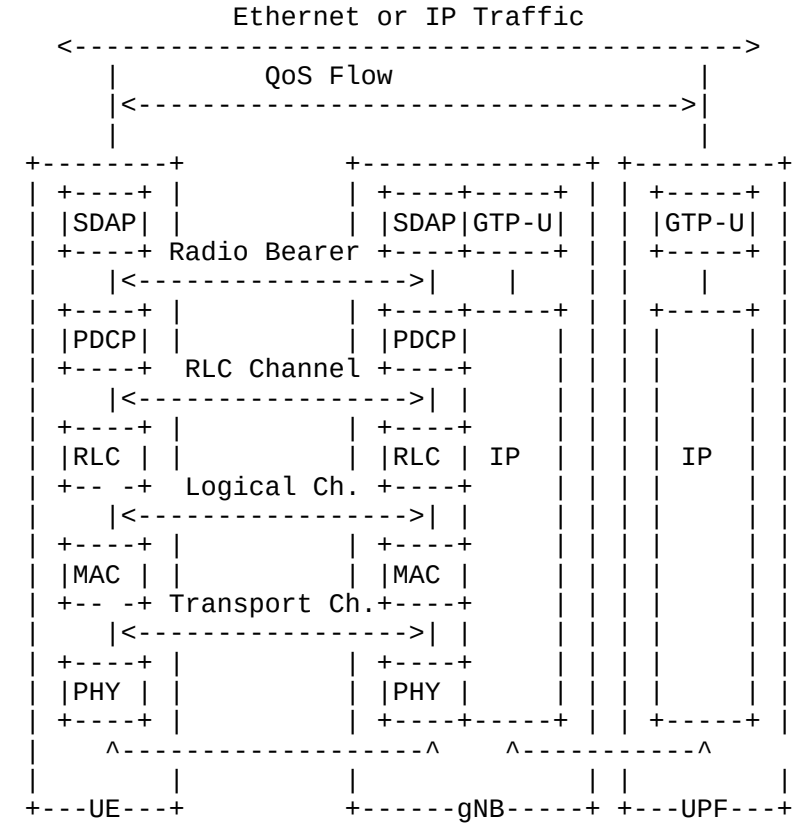



Figure 2: 5G QoS architecture

The QoS flow is transported through the 5G core network via a GTP-U tunnel between the UPF and the gNB over a transport network. In large networks, the UPF can be placed flexibly in the network topology; this allows the UPF to be placed close to the device (UE) and its application and thereby enabling the shortest possible transport connection and reducing latency [ETR20]. In local deployments (e.g., industrial scenarios) a UPF is typically very close to the gNB and can be even located in the same rack. In the RAN, the QoS flow is transported via a radio bearer over the radio interface between the user equipment (UE) and the gNB.

### 3.4. Latency contributions in different layers of radio protocols

The Service Data Adaption (SDAP) layer maps the QoS flows to Data Radio Bearers (DRBs) and marks the packets with the QoS flow identifier. DRBs can be configured to be either in acknowledged mode (AM) or unacknowledged mode (UM) (see Figure 4); for an acknowledged mode DRB lossless data forwarding at handover is enabled for the Packet Data Convergence Protocol (PDCP) layer and Radio Link Control (RLC) operates in acknowledged mode. The latency impact of SDAP on data transfer is negligible.




SDAP: ServiceData Adaptation Protocol
PDCP: Packet Data Convergence Protocol
RLC: Radio Link Control
MAC: Medium Access Control
PHY: Physical Layer

Figure 3: 5G protocol stack for user plane with focus on RAN

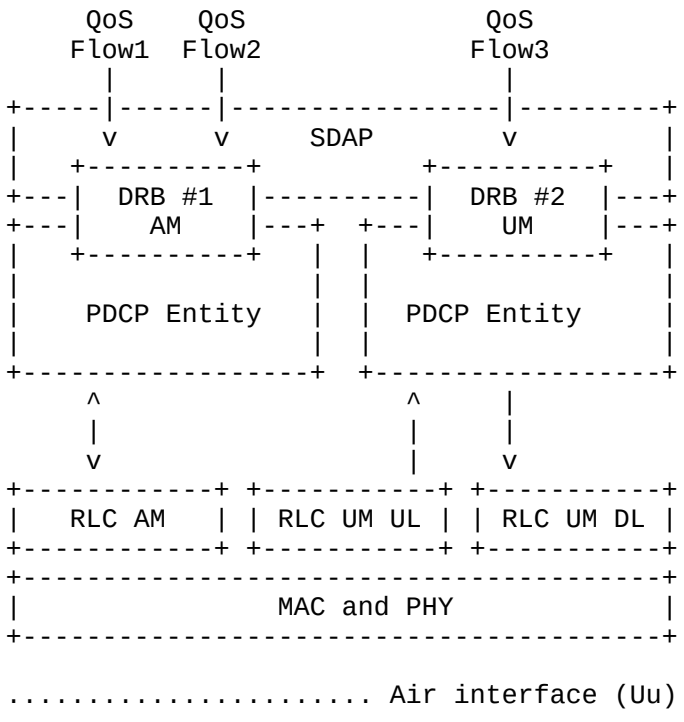



Figure 4: 3GPP 5G protocol stack and data flow

At the next layer, the PDCP (Packet Data Convergence Protocol) layer provides ciphering for encryption of user plane data and optionally also integrity protection and verification via a message authentication code that is calculated for each data Protocol Data Unit (PDU). PDCP assigns a sequence number for each data PDU and forwards it to the underlying RLC layer. PDCP can also perform header compression and decompression over the radio link for the IP headers or Ethernet headers of the end-to-end data flow.

For acknowledged mode DRBs a copy of each PDCP PDU is stored in a local buffer. At changes of the RLC entity, due to either handover or (re-)configuration of dual connectivity or carrier aggregation, a lossless continuation of data transfer is ensured by forwarding not-yet-acknowledged PDCP PDUs to the new RLC entity.

As the underlying protocol layers can lead to packet re-ordering, the PDCP performs packet re-ordering to ensure in-order transmission of data over the DRB. For this, the receiver holds back the received packets until all earlier packets of the DRB have been received and are delivered first. A reordering timer determines how long packets

are held back before delivery. In-order delivery leads to head-of-line blocking, which means that a long packet delay of one PDU (e.g., due to a larger number of retransmissions) affects also earlier packets. The impact of this head-of-line blocking is controlled via the reordering timer, which may reduce head-of-line-induced latencies at an increased risk of sending packets out of order. It is possible to configure the PDCP also for explicit out-of-order delivery, in which case no packet delay propagation within a group of PDUs appears.

The PDCP can be configured for Service Data Unit (SDU) discard, which enables to set a maximum lifetime on a packet in the radio transmission. If a configured SDU discard timer expires, the PDCP sender removes the packet from its buffer and requests the lower layer to purge the related data. SDU discard can be considered as a latency-based active queue management scheme.

The PDCP allows to aggregate multiple radio links over different frequency carriers, based on the NR functionality of carrier aggregation or dual-connectivity. The PDCP connection uses, in this case, multiple RLC entities; this can be used to aggregate the capacity of multiple radio links for the data radio bearer, but it can also be used to provide redundant transmission. For redundant transmission the PDCP entity duplicates PDCP PDUs and transmits them via multiple links; at the PDCP receiver, duplicates are then filtered out.

The PDCP uses one or more RLC channels, via one or more RLC instances. RLC provides reliable data transmission over the radio link via its acknowledged mode (AM); it can also be configured to apply the unacknowledged mode (UM) In AM mode, a selective-repeat ARQ protocol is used, in which correct reception of packets is ensured by detecting packet errors or losses and triggering retransmissions as needed. RLC transmitter and receiver entities maintain a sliding-window buffer, and the receiver entity updates the transmitter entity via status reports about correctly received or missing PDUs. The RLC receiver forwards correctly received PDUs to the PDCP receiving entity, which may comprise packets being delivered out-of-sequence. Reordering for in-sequence delivery is then performed in PDCP. RLC applies segmentation of SDUs towards the Medium Access Control (MAC) layer, so that the MAC protocol can multiplex RLC PDUs into the transport blocks sent by MAC to the physical layer.

From a packet delay perspective, minor latency contributions are made by packet processing. The larger possible latency contribution in acknowledged mode comes from the ARQ operation. A packet is maintained in the receiver buffer until it is successfully transmitted. For this, several RLC retransmissions can be used,

where the maximum number of retransmissions is configurable. An RLC retransmission takes in the order of some tens of milliseconds, so that it can lead to some increased delay of packets that are not correctly transmitted in the first RLC transmission attempt. The need for RLC retransmission depends strongly on the configuration of the reliability that is configured for the lower MAC/PHY layers. For time critical low latency communication, typically the MAC/PHY is configured very reliably so that RLC retransmissions are not necessary. This trade-off we discuss more.

MAC entities are responsible for scheduling the radio resources for all bearers in UEs and gNB in both uplink and downlink directions. The RLC data segments received from multiple logical channels are concatenated along with MAC headers, padded if required, and then encoded to fit inside the scheduled Transport Block (TB) to be transmitted through the radio physical layer [NR-5G]. After the successful reception of the TB, the counterpart MAC entity decodes the TB and demultiplexes to the logical channels. Furthermore, the HARQ process of the MAC layer is responsible for handling most of the radio link errors. HARQ combines ARQ with Forward Error Correction (FEC) to efficiently enhance the reliability of communication in wireless channels. Via fast feedback the receiving MAC provides positive (ACK) or negative acknowledgments (NACK) back to the transmitter about successful TB decoding. One of the key functions of the MAC entity at gNB is to perform radio resource allocation for both Uplink (UL) and Downlink (DL) directions every TTI. The exact resource allocation process, considering factors such as Channel State Indicator (CSI), QoS requirements, and buffer occupancy, is beyond the scope of this document. However, it is important to note that the scheduler plays a crucial role in ensuring that the TB size (TBS) aligns with the chosen Modulation and Coding Scheme (MCS) and the number of Physical Resource Blocks (PRBs) allocated for the transmission. In addition to the above functions, the MAC also manages random access control during the initial access of UEs.

### 3.5. Latency Analysis

The latency analysis focuses on the following areas as contributors to the latency:

* Processing delays at gNB and UE

* Traffic handling / queuing

* Data transmission over the radio interface

* Reliability mechanisms (like HARQ)

In addition, further delays may be incurred due to mobility of devices or activating devices from power-saving idle states.

### 3.5.1. Processing delays in gNB and UE

For RAN processing in both UE and gNB, the most processing-intensive functions are found in the physical layer. They comprise, e.g., channel equalization, channel encoding and decoding, Multiple-input Multiple-output (MIMO) processing. As part of the 5G standardization for URLLC, different UE capabilities with regards to processing times have been defined. For UEs that support faster processing (i.e. "UE capability 2"), this allows the scheduler in the gNB to accelerate certain radio transmission procedures that depend on UE processing times.

### 3.5.2. Traffic handling and queuing

In practical network situations a 5G network provides connectivity for a large number of UEs and a potentially even larger number of traffic flows. The gNB scheduler allocates the radio resources to all UEs and traffic flows in a radio cell for both uplink and downlink. In case that more traffic packets arrive at the wireless 5G transmitter than can be served in the next transmission time interval, which is the scheduling period for which radio resources are allocated, queuing occurs as not all traffic can be handled instantaneously. The queuing of packets thus can introduce additional packet delays.

To ensure that time-critical traffic flows are not impacted by large queuing delays, traffic prioritization is defined. 5G applies a QoS framework, where different traffic flows are separated (into so-called QoS flows), and traffic handling and prioritization is performed between those flows. By appropriate prioritization in the scheduler, the impact of queuing can be minimized for time-critical traffic flows. For this to work, it is also important that the total number and aggregate traffic of time-critical traffic flows, that should obtain priority in scheduling decisions, stays below some threshold fraction of the total 5G network capacity. To this end, admission control is applied when admitting new traffic flows.

### 3.5.3. Data transmission over the radio interface

The data transmission over the radio interface is significantly impacted by the radio interface design and the frame structure. A radio slot consists of 14 Orthogonal Frequency Division Multiplexing (OFDM) symbols, where a flexible numerology with different options of sub-carrier spacing can be applied, which leads to different slot durations [SWD18][LSW19]. The common slot lengths in deployed 5G

networks have a length of 0.5 ms (based on 30 kHz sub-carrier spacing) in frequency bands up to 6 GHz, and a length of 0.125 ms (based on 120 kHz sub-carrier spacing). The transmission of user data is scheduled by the scheduler per slot. 5G can be deployed in a wide range of spectrum bands; multiple spectrum bands can be combined by a 5G network. This includes frequency bands from 450 MHz up to 2.6 GHz which are based on frequency division duplex (FDD), which means that uplink and downlink transmission is ongoing simultaneously on different spectrum carriers. But above 2 GHz typically time-division duplex (TDD) is applied, where the same spectrum carrier is alternatingly used for uplink and downlink transmission. The majority of 5G network deployments are based on TDD spectrum allocations.

In principle, the 5G standard allows a very flexible configuration of TDD patterns. In practice, there are constraints due to coexistence: if two networks use different TDD patterns, this can cause interference between these two networks. For local 5G network deployments the choice of TDD pattern is more flexible, in particular when indoors, since such networks are more isolated from other networks and coexistence is easier. In today’s (public) 5G networks only a set of TDD patterns is used, which are often even with a larger portion of radio resources being allocated to downlink, as most data in public networks is downloaded to devices. From a latency perspective the TDD pattern has a large impact on the transmission latency, as it restricts at what time instances the scheduler can allocate downlink or uplink resources for the transmission of user data or control information (like HARQ feedback).

Other latency-related improvements of the radio transmission include pre-configured transmission opportunities for time-critical devices; this can significantly reduce the time for a UE to obtain access to the radio channel by avoiding an initial request procedure to the gNB [SWD18][LSW19].

#### 3.5.4. Wireless transmission reliability

A new paradigm has been introduced with the 5G standard to address time-critical communications, for which features for URLLC have been standardized. Those include shortened transmission procedures and very robust transmission modes for data and control channels, to significantly reduce the probability of unsuccessful radio transmissions. In addition, a very effective way to provide reliability in a time-varying wireless transmission context is the application of ARQ.

By identifying packet losses and recovering them by retransmissions a reliable transmission over 5G can be provided. Thereby a two-level ARQ mechanism has proven to be very effective [LLM09]. A stop-and-wait Hybrid ARQ mechanism with multiple parallel HARQ processes is implemented in the MAC layer tightly coupled with the physical layer. Fast HARQ feedback (i.e., acknowledgement of negative acknowledgement of successful transmission, ACK or NACK) is enabled via physical channels and allows for fast error recovery. In addition, HARQ is integrated with channel coding by allowing to provide incremental redundancy in the retransmission. This provides a very spectral efficient recovery of transmission errors.

Moreover, a sliding window ARQ mechanisms is provided by the RLC layer. It operates with full ARQ status reports about missing and correctly received RLC PDUs, which are transmitted as RLC control messages including a cyclic redundancy check and normal transmission over the lower MAC/PHY layers. While the majority of transmission errors are recovered by the MAC HARQ, there is a risk of residual HARQ errors, for example due to failure of the binary HARQ feedback, where HARQ NACK may be erroneously misinterpreted as ACK and lead to a packet failure. It is not spectrally efficient to protect such small HARQ signals with very high reliability. The RLC ARQ protocol is well capable at recovering such HARQ failures to provide very high reliability of data transmission. However, the retransmission round-trip time (RTT) of RLC ARQ is significantly larger than the HARQ RTT. For mobile broadband services the benefit of this coordinated two-layer ARQ has been acknowledged as an efficient solution.

As shown in Figure 5, by expanding the service range of 5G to a wider set of critical communication services the focus of latency performance has shifted away from the best-effort latency performance, e.g. expressed as mean packet delay, and which is a relevant latency metric for typical mobile broadband (MBB) applications. For time-critical services, the latency bound comes into focus. To this end, the concept of reliability has been defined in the 5G standardization, which expresses the probability that a packet can be transmitted in a defined maximum delay. Latency performance is thus expressed by a pair of metrics: the latency bound and the reliability with which this bound can be provided.

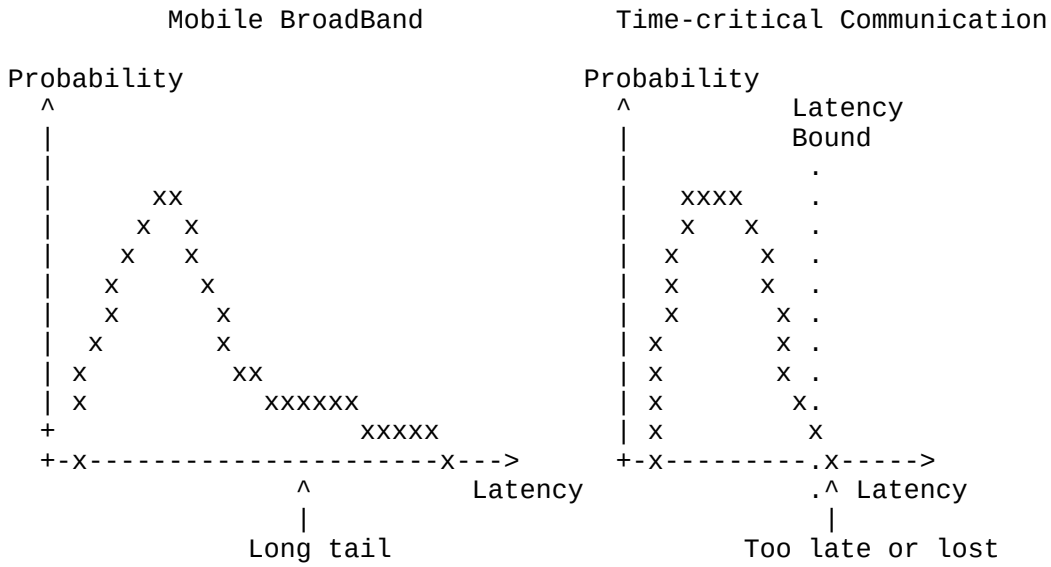



Figure 5: Time-critical communication with URLLC: from best effort to bounded latency performance

The focus of 5G standardization so far was on the latency bound and the reliability for time-critical services. For the integration of 5G with dependable end-to-end communication, e.g., based on TSN or DetNet, packet delay variation may also be of importance. Independent from the latency bound that is provided by 5G, it is clear from the description above that 5G introduces a large PDV; the relative PDV is significantly larger than the one found e.g., in wired nodes.

## 4. Example: Observed characteristics in real network

This section contains real-world observations on packet delay distribution from a 5G system. Through an empirical analysis framework developed for 5G networks as described in [EDAF24], the internal mechanisms in 5G contributing to the packet delay distribution were investigated.

| | Packet Delay | | |
|---|---|---|---|
| | 10 ms | 15 ms | 20 ms |
| Cumulative probability | 99% | 99.99% | 99.999% |
| HARQ Re-transmission | 0.01% | 15% | 45% |
| RAN Transmission | 27% | 25% | 10% |
| RAN Segmentation | 43% | 40% | 30% |
| RAN Queuing Delay | 29.99% | 20% | 15% |

Figure 6: Measurement on internal mechanisms in 5G contributing to the packet delay distribution

In an experiment conducted on OpenAirInterface 5G network [OAI5G] , a traffic generator was deployed on a static UE with ideal coverage to push packets every 10 ms on uplink direction. The end-to-end uplink delay of each packet on a live 5G network was measured and decomposed. Figure 6 on second row displays the cumulative distribution of the packet delays which also indicates the packet's delay violation probability for different delay targets. For instance, it can be observed that 15 ms target was violated with probability of 10e-2 while 20 ms target was violated with probability of 10e-3. Such insight can be useful to incorporate when it comes to determining end-to-end schedules as the violation probability indicates the ratio of packets that will arrive later than the determined window.

In addition, we measured the contribution percentage of 4 distinct delay components to the packet delay violations: HARQ retransmissions, RAN transmission, RAN segmentation, and RAN queuing. Each of these processes contributes on a different level to the delay violations, which is reported in percentage in Figure 6. For instance, 15 ms target delay with violation probability of 10e-2 has 20% contribution from queuing delay, 40% from segmentation delay, 25% from RAN transmission, and 15% from HARQ re-transmissions.

Regarding larger delay targets, where violations are less likely, contribution of HARQ retransmissions starts to dominate, accounting for up to 50% of the e2e delay. This trend was further evident in all experiments, underscoring that the primary contributor to the extended tail in packet delay is the infrequent yet impactful HARQ re-transmissions.

## 5. Scheduling related future work

[D6G-D3.4] describes the concepts and algorithms for optimizing and dynamically adapting end-to-end schedules with wired and wireless network elements to enable deterministic end-to-end guarantees in dynamic environments including mobility, dynamic packet delay, and dynamic stream sets. Moreover, in [D6G-D3.4] several algorithms for calculating and adapting robust end-to-end schedules for scheduled traffic are proposed. These algorithms feature the maximization of robustness, fast adaptation through highly optimized algorithms, graceful degradation of the quality of service under increasing packet delay, and schedules optimized for the smooth handover of mobile stations.

The packet delay characteristics of mobile transmissions has to be considered by the packet scheduling performed at routers to provide reliable end-to-end delay guarantees. Although scheduling is only concerned with providing bounds on queuing delay, the node internal forwarding delay is another integral part of end-to-end delay and must be considered when calculating scheduling parameters or analyzing an end-to-end schedule. The node internal forwarding delay of mobile virtual DetNet routers causes a packet delay that is stochastic and heavy-tailed, i.e., larger delay values are more likely compared to exponentially bounded tails and packet delay variation is relatively large. These properties will lead to the following problems for end-to-end scheduling.

In case of clock-driven scheduling scenarios, similar to scheduled traffic (time-aware shaper) [IEEE8021Qbv] of TSN, the end-to-end scheduling requires the calculation of per-hop time-tables [SOL23] to control packet forwarding:

* Bad reliability-efficiency trade-off: due to the large packet delay variation, larger time windows have to be allocated to flows to isolate flows in time and reliably guarantee delay bounds. With non-work-conserving scheduling (i.e., exclusively allocated time windows) this reduces the number of admitted flows or bandwidth that can be utilized.

* Higher complexity of scheduling problem formulation and solution: stochastic packet delay must be considered in the formulation for calculating time-tables (e.g., Integer Linear Programming, Constrained Programming). This might also increase the time to calculate a feasible schedule or make decisions for admission control.

In case of other (non-clock-driven) scheduling mechanisms, e.g., using static or dynamic priorities or hop-by-hop traffic shaping like the TSN Credit-Based Shaper [IEEE8021Qav], Asynchronous Traffic Shaper [IEEE8021Qcr], etc.:

* Higher complexity of end-to-end delay analysis: stochastic delay with large delay variation needs to be considered in the analysis methodology, e.g., in definition of arrival curves in network calculus [MSL18], to derive tight delay bounds.

In future work, a detailed analysis for each individual scheduling approach is required to analyze the specific impact of the packet delay characteristic onto end-to-end delay bounds, end-to-end delay variation, reliability-efficiency trade-off, runtime of schedule synthesis and analysis, and other KPIs.

Dependable, time-critical communication is poised to become a key technology enabler for future mobile (6G) networks. [D6G-D1.4] provides a comprehensive overview about a system architecture tailored to dependable, time-critical communication, and offers a detailed description of the architecture’s deployment for realizing several time-critical use cases. Furthermore, [D6G-D1.4] describes how to enable new forms of performance predictions for dependable end-to-end traffic management when integrating 6G into time-sensitive or deterministic networks (TSN/DetNet).

## 6. Summary

Wireless communication provides flexibility and simplicity, but with inherently stochastic components that lead to packet delay distributions metrics exceeding significantly those found in wired counterparts. These deviations of stochastic characteristics make traditional approaches to planning and configuration of end-to-end time-critical communication networks such as Time-sensitive Networking (TSN) or Deterministic Networking (DetNet), fall short in their performance regarding service performance, scalability, and efficiency.

Some traffic shaping mechanisms, like time-scheduled transmission (i.e., IEEE 802.1Qbv), expect very deterministic latency behavior in every node on the transmissions path. The latency distribution of a

5G system makes it impracticable to implement some legacy time-schedule configurations. Therefore, to ensure wide integration and interworking with wired deterministic technology such as TSN and DetNet, it is desirable to develop wireless-friendly solutions to ensure the end-to-end latency bounds of deterministic applications.

## 7. Acknowledgements

Authors extend their appreciation to James Gross, Gourav Prateek Sharma, Janos Farkas, Marilet De Andrade Jardim, Gyorgy Miklos, and Damir Hamidovic for their insightful comments and productive discussion that helped to improve the document.

## 8. References

[D6G-D1.4] DETERMINISTIC6G Project, "D1.4: Final report - A Dependable Network Architecture for 6G", 2025, <https://deterministic6g.eu/index.php/library-m/deliverables/>.

[D6G-D2.1] DETERMINISTIC6G Project, "D2.1: First report on 6G-centric Enablers", 2023, <https://deterministic6g.eu/index.php/library-m/deliverables/>.

[D6G-D3.1] DETERMINISTIC6G Project, "D3.1: Report on 6G Convergence Enablers Towards Deterministic Communication Standards", 2023, <https://deterministic6g.eu/index.php/library-m/deliverables/>.

[D6G-D3.4] DETERMINISTIC6G Project, "D3.4: Report on Optimized Deterministic End-toEnd Schedules for Dynamic Systems", 2024, <https://deterministic6g.eu/index.php/library-m/deliverables/>.

[EDAF24] Mostafavi, S., Tillner, M., Sharma, G., and J. Gross, "An End-to-End Delay Analytics Framework for 5G-and-Beyond Networks", arXiv preprint arXiv:2401.09856, 2024.

[ETR20] Alriksson, F., Bostroem, L., Sachs, J., P. E. Wang, Y., and A. Zaidi, "Critical IoT connectivity Ideal for Time-Critical Communications", Ericsson Technology Review, DOI 10.23919/ETR.2020.9905508, 2020.

[FGAQoS] 5G-ACIA, "5G QoS for Industrial Automation", 2021, <https://5g-acia.org/whitepapers/5g-quality-of-service-for-industrial-automation/>.

[FGS15] 5G-SMART, "D1.5: Evaluation of radio network deployment options", 2021, <https://5gsmart.eu/deliverables/>.

[IEEE8021Q] IEEE, "IEEE Standard for Local and Metropolitan Area Networks -- Bridges and Bridged Networks", DOI 10.1109/IEEESTD.2018.8403927, July 2018, <https://ieeexplore.ieee.org/document/8403927>.

[IEEE8021Qav] IEEE, "IEEE Standard for Local and Metropolitan Area Networks -- Amendment 12: Forwarding and Queuing Enhancements for Time-Sensitive Streams", DOI 10.1109/IEEESTD.2010.8684664, July 2018, <https://ieeexplore.ieee.org/document/8684664>.

[IEEE8021Qbv] IEEE, "IEEE Standard for Local and Metropolitan Area Networks -- Amendment 25: Enhancements for Scheduled Traffic", DOI 10.1109/IEEESTD.2016.8613095, 2015, <https://ieeexplore.ieee.org/document/8613095>.

[IEEE8021Qcr] IEEE, "IEEE Standard for Local and Metropolitan Area Networks -- Amendment 34: Asynchronous Traffic Shaping", DOI 10.1109/IEEESTD.2020.9253013, November 2020, <https://ieeexplore.ieee.org/document/9253013>.

[LLM09] Larmo, A., Lindstroem, M., Meyer, M., Pelletier, G., Torsner, J., and H. Wiemann, "The LTE link-layer design", IEEE Communications Magazine, vol. 47, no. 4, pp. 52-59, DOI 10.1109/MCOM.2009.4907407, 2009.

[LSW19] Liberg, O., Sundberg, M., P. E. Wang, Y., Bergman, J., Sachs, J., and G. Wikstroem, "Cellular Internet of Things - From Massive Deployments to Critical 5G Applications", Academic Press, second edition, ISBN: 9780081029022, 2019.

[M23.501] 3GPP 23.501, "System architecture for the 5G System (5GS)", <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144>.

[MSL18] Mohammadpour, E., Stai, E., Mohiuddin, M., and J. Y. Le Boudec, "Latency and Backlog Bounds in Time-Sensitive Networking with Credit Based Shapers and Asynchronous Traffic Shaping", DOI 10.1109/ITC30.2018.10053, 2018, <https://doi.org/10.1109/ITC30.2018.10053>.

[NR-5G] Dahlman, E., Parkvall, S., and J. Skold, "5G NR - The next generation wireless access technology", Academic Press , 2021.

[OAI5G] Kaltenberger, F., Silva, A.P., Gosain, A., Wang, L., and T.T. Nguyen, "OpenAirInterface: Democratizing innovation in the 5G Era", Computer Networks 176,p.107284, 2020.

[RFC8655] Finn, N., Thubert, P., Varga, B., and J. Farkas, "Deterministic Networking Architecture", RFC 8655, DOI 10.17487/RFC8655, October 2019, <https://www.rfc-editor.org/info/rfc8655>.

[SOL23] Stueber, T., Osswald, L., Lindner, S., and M. Menth, "LA Survey of Scheduling Algorithms for the Time-Aware Shaper in Time-Sensitive Networking (TSN)", DOI 10.1109/ACCESS.2023.3286370, 2023, <https://doi.org/10.1109/ACCESS.2023.3286370>.

[SWD18] Sachs, J., Wikstroem, G., Dudda, T., Baldemair, R., and K. Kittichokechai, "5G radio network design for ultra-reliable low-latency communication", IEEE Network vol. 32, pp. 24-31, 2018.

## Authors' Addresses

Balazs Varga
Ericsson
Magyar Tudosok krt. 11
1117 Budapest
Hungary
Email: balazs.a.varga@ericsson.com

Joachim Sachs
Ericsson
Germany
Email: joachim.sachs@ericsson.com

Frank Duerr
University of Stuttgart
Universitaetsstr. 38
70569 Stuttgart
Germany
Email: frank.duerr@ipvs.uni-stuttgart.de

Samie Mostafavi
KTH Royal Institute of Technology
Stockholm
Sweden
Email: ssmos@kth.se