DETERMINISTIC6G

# First report on 6G centric enablers

D2.1

The DETERMINISTIC6G project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 1010965604.

# First report on 6G centric enablers

| | |
|---|---|
| Grant agreement number: | 101096504 |
| Project title: | Deterministic E2E communication with 6G |
| Project acronym: | DETERMINISTIC6G |
| Project website: | Deterministic6g.eu |
| Programme: | EU JU SNS Phase 1 |

| | |
|---|---|
| Deliverable type: | Report |
| Deliverable reference number: | D2.1 |
| Contributing workpackages: | WP2 |
| Dissemination level: | PUBLIC |
| Due date: | 31-12-2023 |
| Actual submission date: | 21-12-2023 |
| Responsible organization: | KTH |
| Editor(s): | James Gross and Gourav Prateek Sharma |
| Version number: | v2.0 |
| Status: | Final |
| Short abstract: | This deliverable reports on the 6G-centric enablers essential for dependable time-critical communications in 6G. First, an analysis is presented on the transmission latency of 5G systems. Several approaches for Packet Delay Correction are presented to address delay variations in 6G. Additionally, data-driven methods are introduced to enhance predictability in 6G systems. The report also delves into efficient resource allocation methods for supporting XR applications. |
| Keywords: | 5G, 6G, RAN, latency, delay, synchronization, TSN, DetNet, ML, PDC, PDV, federated learning, URLLC, XR |


| | |
|---|---|
| Contributor(s): | Joachim Sachs (EDD),<br>Arash Sahbafard (SAL),<br>James Gross (KTH),<br>Gourav Prateek Sharma (KTH),<br>Balázs Varga (ETH),<br>Marilet De Andrade Jardim (EAB), |

## Revision History

| | |
|---|---|
| 29/03/2023 | Table of Contents |
| 06/11/2023 | First internal draft |
| 05/12/2023 | Second internal draft |
| 16/12/2023 | Final version |

## Disclaimer

This work has been performed in the framework of the Horizon Europe project DETERMINISTIC6G co-funded by the EU. This information reflects the consortium's view, but the consortium is not liable for any use that may be made of any of the information contained therein. This deliverable has been submitted to the EU commission, but it has not been reviewed and it has not been accepted by the EU commission yet.

## Executive summary

Dependable time-critical communications are foreseen to play a vital role in future 6G networks. This is driven by the need to support a wide range of applications emerging from domains such as Extended Reality (XR), Occupational exoskeletons (OEs), adaptive manufacturing and mobile automation. For these applications, communication requirements go beyond optimizing average throughput and latency but strict guarantees in Packet Delay (PD) and Packet Delay Variation (PDV). Unlike previous generations of communication systems, certain enablers are needed in 6G to realize dependable time-critical communication. This document reports on a set of enablers focused on dependable time-critical operations in 6G.

A comprehensive analysis of transmission latency in 5G systems is provided along with a systematic description of various sources of latency. We also summarize various enhancements proposed under Ultra-Reliable Low Latency Communications (URLLC) focusing on improving both latency and reliability in 5G networks. Packet Delay Correction (PDC) is presented as a solution to ensure bounded and predictable latency behavior within 6G networks. To this end, multiple PDC methods are proposed, including a generalized timestamp-based solution, an advanced approach using the 3GPP protocol stack and an approach based on the number of radio retransmissions. In the context of dependable time-critical communications, data-driven latency characterization will serve as a bridge for the interworking between wired and wireless systems. To this end, this report presents Mixture Density Networks (MDNs) based models that leverage extreme value theory for accurate tail latency characterization. Furthermore, we present an analysis to compare various overheads between the centralized and federated latency prediction architectures. Finally, efficient Radio Access Network (RAN) resource allocation methods are required for dependable time-critical communication that optimize resource efficiency while ensuring latency and reliability requirements. To this end, a problem optimizing the number of Hybrid Automatic Repeat Request (HARQ) retransmissions is presented to support the requirements of AR/XR applications.

# Contents

# 1 Introduction

Digital transformation of industries and society is resulting in the emergence of a larger family of time-critical services with needs for high availability and which present unique requirements distinct from traditional Internet applications like video streaming or web browsing. Time-critical services are already known in industrial automation; for example, an industrial control application that might require an end-to-end "over the loop" (i.e., from the sensor to the controller back to the actuator) latency of 2 ms and with a communication service requirement of 99.9999% [3GPP16-22261]. But with the increasing digitalization similar requirements are appearing in a growing number of new application domains, such as extended reality, autonomous vehicles and adaptive manufacturing [DET23-D11]. The general long-term trend of digitalization leads towards a Cyber-Physical Continuum where the monitoring, control and maintenance functionality is moved from physical objects (like a robot, a machine or a tablet device) to a compute platform at some other location, where a digital representation – or digital twin – of the object is operated. Such Cyber-Physical System (CPS) applications need a frequent and consistent information exchange between the digital and physical twins. Several technology developments in the ICT-sector drive this transition. The proliferation of (edge-) cloud compute paradigms provide new cost-efficient and scalable computing capabilities, that are often more efficient to maintain and evolve compared to embedded compute solutions integrated into the physical objects. It also enables the creation of digital twins as a tool for advanced monitoring, prediction and automation of system components and improved coordination of systems of systems. New techniques based on Machine Learning (ML) can be applied in application design, that can operate over large data sets and profit from scalable compute infrastructure. Offloading compute functionality can also reduce spatial footprint, weight, cost and energy consumption of physical objects, which is in particular important for mobile components, like vehicles, mobile robots, or wearable devices. This approach leads to an increasing need for communication between physical and digital objects, and this communication can span over multiple communication and computational domains. Communication in this cyber-physical world often includes closed-loop control interactions which can have stringent end-to-end KPI (e.g., minimum and maximum packet delay) requirements over the entire loop. In addition, many operations may have high criticality, such as business-critical tasks or even safety relevant operations. Therefore, it is required to provide *dependable time-critical communication* which provides communication service-assurance to achieve the agreed service requirements.

## 1.1 DETERMINISTIC6G Approach

Time-critical communication has in the past been mainly prevalent in industrial automation scenarios with special compute hardware like Programmable Logic Controller (PLC), and is based on a wired communication system, such as EtherCat and Powerlink, which is limited to local and isolated network domains which is configured to the specific purpose of the local applications. With the standardization of Time-Sensitive Networking (TSN), and Deterministic Networking (DetNet), similar capabilities are being introduced into the Ethernet and IP networking technologies, which thereby provide a converged multi-service network allowing time critical applications in a managed network infrastructure allowing for consistent performance with zero packet loss and guaranteed low and bounded latency. The underlying principles are that the network elements (i.e. bridges or routers) and the PLCs can provide a consistent and known performance with negligible stochastic variation, which allows to manage the network configuration to the needs of time-critical applications with known traffic characteristics and requirements.

It turns out that several elements in the digitalization journey introduce characteristics that deviate from the assumptions that are considered as baseline in the planning of deterministic networks. There is often an assumption for compute and communication elements, and also applications, that any stochastic behavior can be minimized such that the time characteristics of the element can be clearly associated with tight minimum/maximum bounds. Cloud computing provides efficient scalable compute, but introduces uncertainty in execution times; wireless communications provides flexibility and simplicity, but with inherently stochastic components that lead to packet delay variations exceeding significantly those found in wired counterparts; and applications embrace novel technologies (e.g. ML-based or machine-vision-based control) where the traffic characteristics deviate from the strictly deterministic behavior of old-school control. In addition, there will be an increase in dynamic behavior where characteristics of applications, and network or compute elements may change over time in contrast to a static behavior that does not change during runtime. It turns out that 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. Instead, a revolutionary approach to the design, planning and operation of time-critical networks is needed that fully embraces the variability but also dynamic changes that come at the side of introducing wireless connectivity, cloud compute and application innovation. DETERMINISTIC6G has as objective to address these challenges, including the planning of resource allocation for diverse time-critical services end-to-end over multiple domains, providing efficient resource usage and a scalable solution [SPS+23].

DETERMINISTIC6G takes a novel approach towards converged future infrastructures for scalable cyber-physical systems deployment. With respect to networked infrastructures, DETERMINISTIC6G advocates (I) the acceptance and integration of stochastic elements (like wireless links and computational elements) with respect to their stochastic behavior captured through either short-term or longer-term envelopes. Monitoring and prediction of KPIs, for instance latency or reliability, can be leveraged to make individual elements plannable despite a remaining stochastic variance. Nevertheless, system enhancements to mitigate stochastic variances in communication and compute elements are also developed. (II) Next, DETERMINISTIC6G attempts the management of the entire end-to-end interaction loop (e.g. the control loop) with the underlying stochastic characteristics, especially embracing the integration of compute elements. (III) Finally, due to unavoidable stochastic degradations of individual elements, DETERMINISTIC6G advocates allowing for adaptation between applications running on top such converged and managed network infrastructures. The idea is to introduce flexibility in the application operation such that its requirements can be adjusted at runtime based on prevailing system conditions. This encompasses a larger set of application requirements that (a) can also accept stochastic end-to-end KPIs, and (b) that possibly can adapt end-to-end KPI requirements at run-time in harmonization with the networked infrastructure. DETERMINISTIC6G builds on a notion of time-awareness, by ensuring accurate and reliable time synchronicity while also ensuring security-by-design for such dependable time-critical communications. Generally, we extend a notion of deterministic communication (where all behavior of network and compute nodes and applications is pre-determined) towards dependable time-critical communication, where the focus is on ensuring that the communication (and compute) characteristics are managed in order to provide the KPIs and reliability levels that are required by the application. DETERMINISTIC6G facilitates architectures and algorithms for scalable and converged future network infrastructures that enable dependable time-critical communication end-to-end, across domains and including 6G.

## 1.2 Background on 5G and TSN integration

In the past, industrial networking comprised of multiple fieldbus technologies and their subsequent real-time industrial Ethernet versions (e.g., PROFINET, EtherCAT), delivering time-critical communication services. However, these technologies are applied in certain isolated network segments and are limited in their capabilities to serve a broader communication platform. Today, three major standards suites specify dependable time-critical communications based on their scope: IEEE's TSN, IETF's DetNet, and 3GPP's 5G Time-Sensitive Communication (TSC) and Ultra-Reliable Low Latency Communications (URLLC). On the wireline side, TSN provides the required functionality as a toolbox of standard capabilities for the bridged Ethernet standard specified in IEEE 802.3 and IEEE 802.1 [BSB+19]. TSN toolbox includes functional building blocks, e.g., time synchronization, guaranteed low latency transmissions and high reliability, etc, to enable time-critical communications in Ethernet. For wired IP-based routed packet networks, DetNet is emerging as an extension to IP networking technology specified in IETF. On the wireless side, the 5G mobile communication system is specified in the Third Generation Partnership Project (3GPP). The 5G network as an alternative to a wired connectivity solution supports communication with unprecedented reliability and very low latency through the URLLC enhancements introduced in Release 16. Currently, TSN and 5G are considered as complementary technologies in providing dependable time-critical communication services, thereby paving the way towards future advanced manufacturing systems and other vertical areas. The integration of the TSN system to the 6G system provides converged communication on the same network infrastructure for a wide range of services including time sensitive applications that require deterministic, reliable, and low latency communication.

The 3GPP standardization work to integrate with TSN started in Release 16 with 5G. The 5G system is represented as a set of IEEE-compliant virtual TSN bridges (also referred to as virtual TSN bridges, see Figure 1). The 5G virtual TSN bridge can be connected to TSN nodes (also be referred to as wired TSN nodes/bridges) in a seamless way. 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 TSN network. The RAN can span over the whole production plant to provide wireless connectivity to one or more User Equipments (UEs).

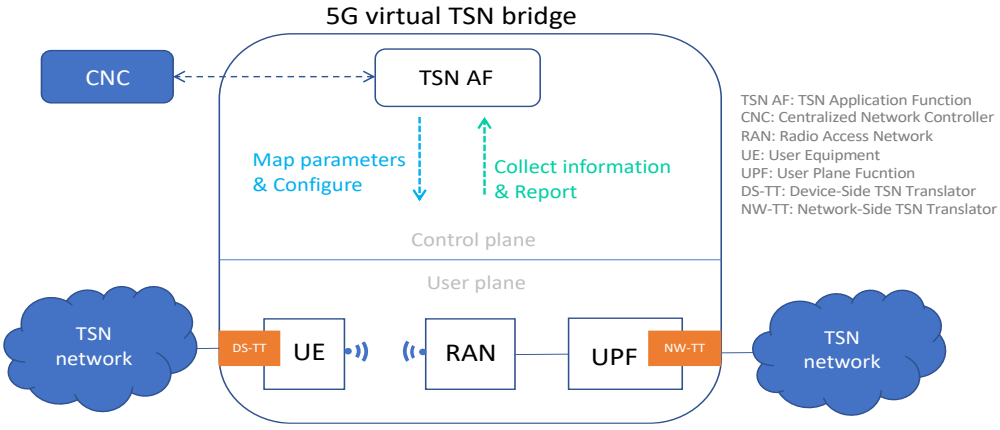


Figure 1. 5G integration with TSN

The 5G network/virtual TSN bridge defines several gateways between the TSN system and the 5G network, as shown in Figure 1. The gateways include a TSN Application Function (AF), device-side TSN translators (DS-TTs) on the UEs and network-side TSN translators (NW-TT) on the UPF. The TSN AF

connects a Centralized Network Controller (CNC) and a 5G control plane. The TSN AF will be responsible to collect information from the 5G system and report it to the CNC via managed objects. The CNC can configure the 5G virtual TSN bridge through the TSN AF, which maps parameters and sets the configuration in the 5G control plane.

End-to-end time-critical communication provided by the integrated 5G-TSN system requires bounded latency between an ingress port and an egress port of the 5G system. The deterministic transmission latency may be described as an upper bound/maximum allowed packet delay together with a maximum tolerated PDV. An Ethernet-based TSN system can provide a small PDV due to wired connectivity characteristics. A minimum and maximum delay between port pairs of the TSN node are key characteristics for the computation of time-triggered schedules to achieve the deterministic transmission latency. However, there are some substantial differences between the 5G virtual TSN bridge and (wired) TSN bridges. One of the key differences is that the PDV of the 5G system remains considerable higher, for example, 1-2 orders of magnitude compared to the wired TSN nodes where latencies can be controlled to the order of tens of microseconds. Thus, a key challenge in achieving the deterministic transmission latency in the integrated TSN-5G network is the higher PDV of the 5G network. In addition, the higher PDV of the 5G network makes it difficult to practically apply time-scheduled transmission for time schedule configurations, even though support for 802.1Qbv has been targeted in the 5G standard via a hold and forward mechanism. In addition to the remarkably higher PDV in 5G system, there are additional challenges with related to the scalability of the control plane and the complexity of managing end-to-end traffic flows in the 5G-TSN system [DET23-D31].

## 1.3 Contributions of the Report

The objective of this report is to describe *6G-centric enablers* to realize dependable time-critical communications in future. In fact, this ambition extends beyond the pure communication capabilities of 6G, but even includes compute capabilities of 6G for enabling applications to be realized in an edge compute solution. In practice, dependable time-critical communications need to be provided end-to-end and may stretch beyond what will be covered by the 6G system and comprise further domains like wired TSN or DetNet network segments. This end-to-end integration challenge of these technologies is highlighted in [DET23-D11]; in this integration of 6G with other domains we describe the 6G system also as a (virtual) 6G Deterministic communication (DetCom) node. As 6G is still an area of active research, no design choices for a 6G system have yet been made and many options and solutions are still being explored. A systemization of 6G with a detailed realization of capabilities and functionality will only happen when 6G enters standardization and design choices are selected from various options according to their technical merits. 6G standardization will largely be based on what has been defined for 5G, but design choices are not limited by a need to provide backward compatibility to the existing 5G standard. Beneficial capabilities of 5G will be a baseline for 6G, but also new capabilities will be integrated based on learnings of 5G shortcomings and novel technology components.

Our approach to defining and analyzing 6G capabilities is based on an assessment of 5G. 5G and 5G-Advanced (5G-Adv) have already introduced capabilities to support time-sensitive and ultra-reliable and low latency communication. We explore how well these solutions provide a foundation for dependable time-critical communications for future services according to the requirements described in [DET23-D11]. We also identify and propose enhancements to the functionality and capabilities provided by 5G-Adv as standardized until now. These enhancements are here referred to as 6G enablers for dependable time-critical communications. It may be possible to introduce some of these

6G enablers already into 5G-Adv systems. Other 6G enablers may need further maturity or may not match well with some design decision defined in the 5G-Adv standards – those enablers are candidates for consideration in upcoming 6G standardization.

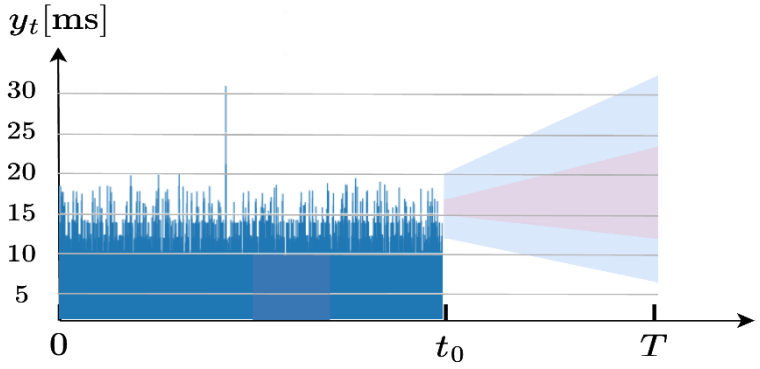


Figure 2. An illustration for the time of predictability limit for latency prediction.

Predictability is one such enabler that will be crucial for incorporating dependability in 6G. The ability to estimate the evolution of a system metric or state in the future is the predictability of the system. In general, the confidence in our estimation of the future decreases with the forecast lead time, i.e., the further we are estimating in time the lower confidence in the estimated system state should be. The level of confidence is determined using a probability density function (PDF) of the predicted metric. Authors in [DT07] have also described a "time of predictability limit", which is the moment when we can no longer distinguish between the forecast and statistical distributions, indicating that predictability diminishes. Here, statistical distribution refers to, e.g., the marginal latency distribution corresponding to inherent stochasticity of the system, which is impacted by the design and implementation of time-critical communications functionality. The forecast distribution is a conditional distribution obtained using, e.g., data-driven latency prediction based on prevailing conditions. This above concept of predictability is illustrated for the problem of latency prediction in Figure 2, which shows latency forecast (e.g., at 40-60 and 10-90 percentiles) at $t_0$. The time limit for latency prediction is the time in the future ($t = T$) when the estimates about the latency are no longer differentiable from the statistical distribution of latency. The loss of confidence in latency prediction can arise from different sources: (i) the error in initial state observation and (ii) inherent limitation of prediction approaches. It is worth pointing out that system design choices (e.g., features in URLLC and PDC) are intricately intertwined with the confidence of the prediction approaches.

To enhance predictability in future communication systems, enhancements are required in both (a) the communication system as well as (b) the characterization methods used to analyze these systems. The first requirement is that the system functions regulate and confine uncertainty in the system performance (i.e., latency and reliability) within required thresholds. Including URLLC optimizations, these functions attempt to minimize variability to maintain relatively consistent performance. This report proposes *Packet Delay Correction* (PDC) as a system-level enabler aimed at controlling the stochastic uncertainties associated with the latency. In addition to system-level enablers, capabilities are required to accurately characterize stochastic uncertainties in the system thus ensuring consistent and dependable time-critical communications service for applications. To this end, this report describes data-driven approaches aimed at accurately estimating latency in a 5G or 6G system. These prediction tools provide insights into the evolving characteristics of the communication system enabling proactive measures to adjust operations according to the required performance levels.

Together, these two elements constitute the foundation of predictability, crucial for delivering a dependable time-critical communications service. Finally, with an objective to support dependable time-critical communications for time-critical applications in 6G, the resource allocation problem in the context of RAN is presented.

Given that time-critical communication is important to this document, it is important to delve into the terminology related to latency before discussing the transmission latency analysis of 5G and the proposed enablers for time-critical communication. This deliverable addresses two crucial metrics related to latency: Packet Delay (PD) and Packet Delay Variation (PDV). Those terms are defined in ITU-T Recommendation Y.1540, section 6.2 for IP packets but can be generalized for all packet-based transport technologies [ITU19-1540]. On one hand, there is a wide consensus on the definition of PD – i.e., PD is the time (t2 – t1) between the occurrence of two corresponding packet reference events associated with packet transmission (t1) and reception (t2), respectively. On the other hand, there are multiple valid definitions for PDV. In one definition, PDV is defined based on the observations of corresponding packet arrivals at ingress and egress measurement points. These observations characterize the variability in the pattern of packet arrival events with respect to a reference delay. Such a reference delay can be the minimum delay of the population of interest. As an alternative definition example, Appendix II of ITU-T Y.1540 provides a definition – based on RFC3393 – of Inter-Packet Delay Variation (IPDV) being a measure of the network's ability to preserve the spacing between packets. In this document we use the ITU-T Y.1540 definitions for PD and PDV (unless indicated otherwise).

## 1.4 Relation to other Work Packages

The work presented in this document is related to various other work packages of the project as follows. The proposed PDC approaches take their input in terms of PDV requirements from the use cases investigated in WP1 and from WP3 with respect to the requirements of end-to-end dependable communications, as described in [DET23-D11] and [DET23-D31], respectively. Latency measurement data collected in WP4 will serve as an input for the training of the latency prediction models presented in this report. The developed models will also be used as an input for the modeling of the 6G DetCom node in the simulator framework developed in WP4 [DET23-D41]. For both the PDC approaches, as well as data collection for latency predictors, to work, an accurate time reference is required at various locations in the network. The time synchronization approaches and challenges with respect to dependable time-critical communications are described in [DET23-D22]. Finally, the concepts developed within this report will be evaluated using the simulation framework of WP4 and also provide input for the E2E architecture developed in WP1.

## 1.5 Structure and Scope of the Document

The first report on 6G centric enablers for dependable time-critical communications consists of six sections. The technical contributions that map to 6G centric enablers are shown in yellow boxes in Figure 3. After the introduction, section 2 delves into the comprehensive analysis of transmission latency in current 5G/5G-Adv systems. Section 3 introduces the idea of PDC and discusses three different solutions. Data-driven approaches for latency characterization in 5G-Adv/6G are described in section 4. This is followed by a mathematical formulation of the RAN resource management problem for XR applications in section 5. Finally, the conclusions and future works of this report are presented in section 6.

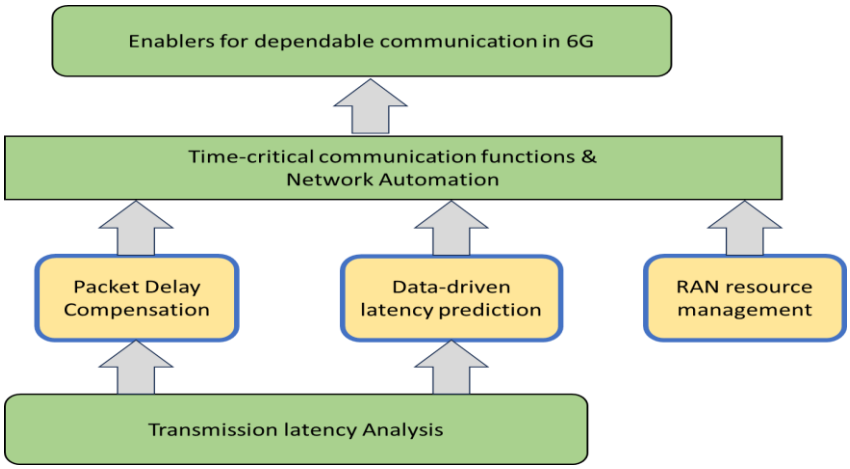


Figure 3 Structure of the deliverable and relationships with dependable time-critical communications.

# 2 5G Transmission Latency Analysis

A comprehensive analysis of 5G transmission latency is important for several reasons. First, it provides an assessment of opportunities to exploit in 5G-Adv/6G and current limitations that should be addressed for dependable time-critical communications. Furthermore, this analysis will provide valuable insights for the development of PDC methods as well as data-driven latency characterization approaches.

## 2.1 Background

Wireless communication is inherently exposed to stochastic performance variations, which may be caused by radio channel fluctuations, mobility, or interference variations. Sophisticated radio transmitters and receivers and radio resource management are adopted to handle stochastic variations in radio communication and to achieve high performance. 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 to provide URLLC. The ambition of 5G URLLC has been to be able to provide low-latency communication while being able to provide 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, as discussed in more detail later in the section. However, HARQ leads to an increase in latency due to multiple transmissions causing a significant PDV. For URLLC, the 5G standard introduced a toolbox to reduce the latency bound by two means: (i) reducing the radio transmission structure for lower latencies (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.

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.

While URLLC provides reduced latencies, this comes at the inherent cost that larger amounts of resources are allocated to the transmission in order to reduce the tail probabilities of transmission

failures. This leads to a loss in spectral efficiency, higher transmission energy and battery consumption for devices. Still, URLLC is the main tool in 5G to reduce latency and thereby also reduce packet delay variation and a trade-off needs to be made between required latency bounds that shall be achieved and the corresponding price in spectral efficiency. To give an example, an uplink transmission in a millimeter wave carrier can be made in two different configurations (see [5GS21-D15] and Figure 14):

- Normal 5G New Radio (NR) configuration with up to 3 retransmissions for reliability with packet delay from ~500 µs to 2.8 ms, with low resource usage,
- 5G URLLC NR configuration with single-transmission reliability with packet delay from ~500 µs to 900 µs, involving high resource usage.

Apart from the high resource costs, 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.

Even though URLLC capabilities allow to significantly reduce the latency for the 5G transmission and lower the upper bound for the transmission, the PDV of the 5G system remains considerable. 5G system PDV is 1-2 orders of magnitude larger compared to fixed TSN bridges where latencies can be controlled to the order of tens of microseconds or even lower [DET23-D31]. Some TSN traffic shaping mechanisms, like time-scheduled transmission (i.e., 802.1Qbv) on the end-to-end path, expect very deterministic latency behavior in every node on the transmissions path. The high PDV of a 5G system is so large to make it impracticable for time-scheduled transmission for some time-schedule configurations, even though support for 802.1Qbv has been targeted in the 5G standard (see [DET23-D31]). Therefore, to ensure integration and interworking with wired deterministic technology such as TSN and DetNet, it is desirable to limit the packet delay variation to a similar level as found in TSN bridges, i.e., in the order of 10’s of microseconds.

The combination of URLLC concepts as known from 5G, and 6G packet delay correction, which is described in section 3, provides means of resource-efficient latency control. Not all time-critical communication services require ultra-low latency, but still the latency needs to be predictable. As URLLC allows to control the upper bound and PDC the lower bound of the packet delay, the latency of 6G transmission can be configured for different target delay values and still have deterministic latency with bounded variation. The required ambition for PDC in 6G is to be able to provide packet delay variations that can be bounded down to a level of microseconds or 10’s of microseconds with high probabilities (e.g., 99.999%).

## 2.2 Transmission Latency Breakdown

In order to enable dependable time-critical communication end-to-end with 5G or 6G, the mobile network typically needs to interwork with external frameworks for time-critical communication such as TSN and DetNet. The latency and PDV of the mobile network need to be well understood. In this section we provide a breakdown of the 5G network architecture, in order to understand where in the system, latencies and PDV are being introduced; we analyze the network mechanisms that impact packet delay, their objectives, and their location. These insights help in the understanding of the work described later: methods to improve the PDV of the mobile network (section 3), ways to characterize and predict the latency of the mobile network (section 4) and radio resource management algorithms that effect the transmission latency of the mobile network (section 5).

The packet delay that is observed by applications is end-to-end between the transmitting and receiving application entities. In the use cases considered here, some part of the connectivity is provided by a mobile 5G or 6G network. One component in packet delay is the transport time of data over the end-to-end distance. If the application endpoints are far apart (e.g., in some tele-operation use cases over large distance) the transport of data over a wide-area transport network (e.g., a fibre network) can make a large contribution to end-to-end latency. As an increasing number of applications are executed in cloud environments, there is the possibility to select the location of the application within a distributed cloud infrastructure so that it is close to the connected devices or machines and thereby reducing transport latency. In combination with 5G mobile communication, edge computing allows to select the location of the 5G network edge gateway at a data center site where the application can be hosted, and thereby minimizing transport network related latencies [ABS+20], see Figure 4. In particular, when use cases are confined to local environments the transport network latencies become negligible.

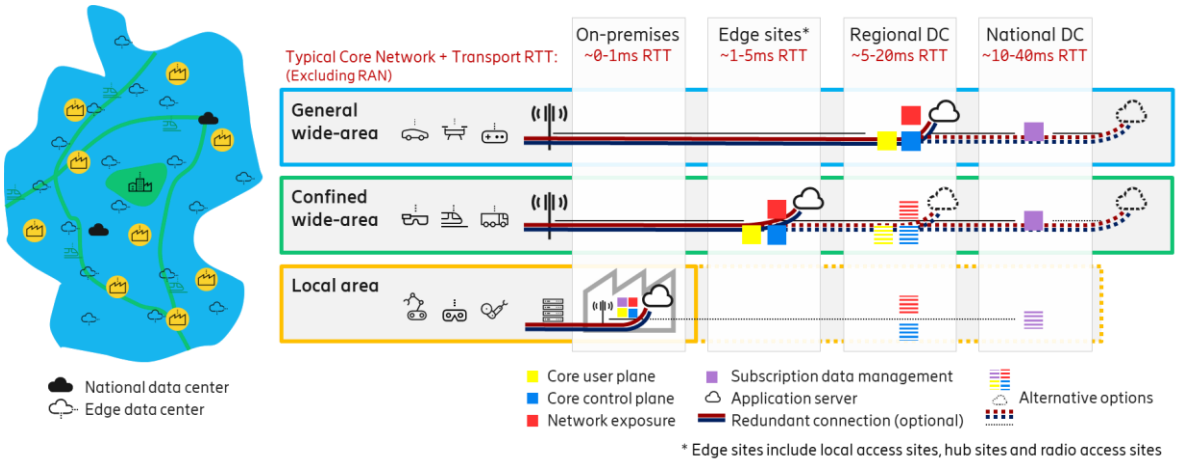


Figure 4. 5G network deployment options with edge computing [ABS+20].

The latencies contributed by the 5G network for end-to-end data transmission can be separated into different components, as shown in Figure 5. The core network comprises the UPF, which is the gateway of the 5G network towards (upstream) data networks or an edge data center. It maps the 5G network internal handling of traffic handling towards the traffic handling of the external data network (e.g., IP or Ethernet). The UPF routes traffic via a transport network to the radio base station (gNB) of the RAN. The RAN is responsible for the wireless communication to and from User Equipments (UEs). Most functions of the UE are related to the wireless communication of the RAN, but it also has components to interface to the application or a (downstream) network (e.g., IP or Ethernet).

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

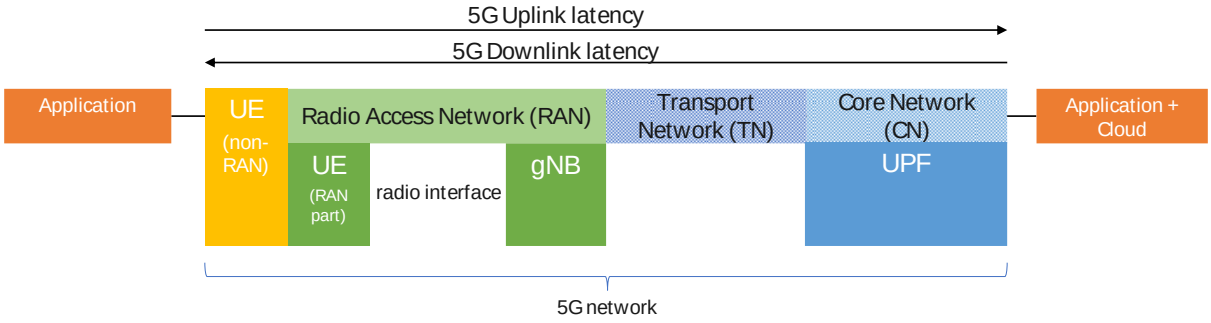


Figure 5. Contributors to latency in the 5G network.

In the 5G RAN the main latency contributors are:

1) Time-domain reliability based on HARQ

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 (e.g. 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 (and at most 0.001% exceed the latency found). Instead HARQ allows for a better utilization of the resources while being robust for a defined upper bound. Retransmissions inherently contribute to the latency of the packet with defined probability given the number of retransmissions. 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).

2) Mobility with handover interruptions

During handover, a UE switches connection from one base station to another, which can lead to handover interruption times. There are some defined optimization tool options to minimize this impact, e.g., L3 make-before-break handover where the resources are allocated and ready before performing the handover, L1/L2 mobility with multiple transmission-reception point (multi-TRP), multi-connectivity. These options are dependent on deployment and spectrum.

3) Time-division duplex structure

The 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.

4) Contention due to resource sharing and queuing

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.

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 [3GPP17-23501][5GA-QOS21][5GS21-D54], as summarized in Figure 6. 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.

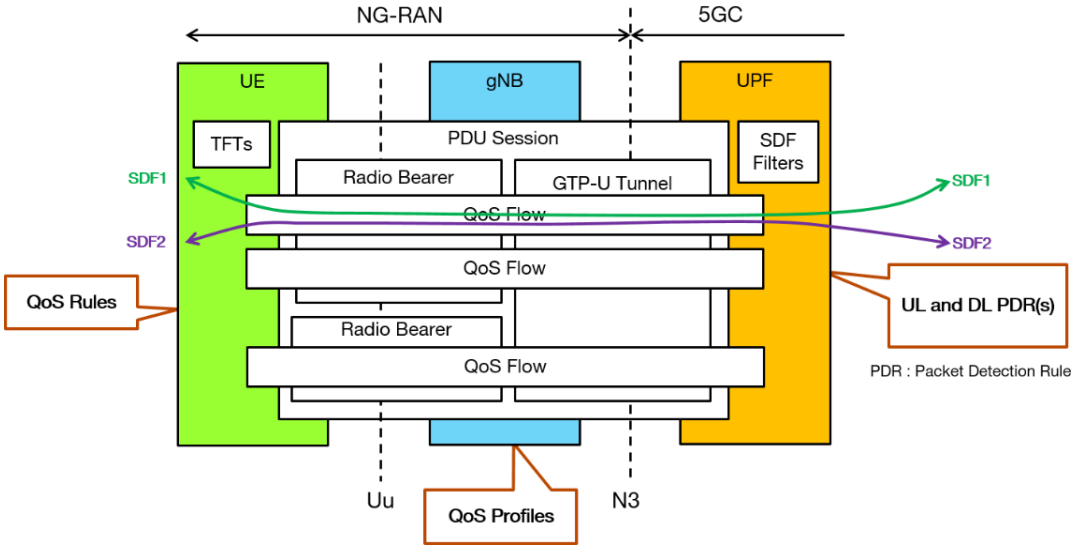


Figure 6. 5G QoS architecture [5G-SMART D5.4].

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 [ABS+20]. In local deployments 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.

In the context of dependable time-critical communications, it is important to analyze the latency end-to-end through the 5G system. On the core network side, a typically well dimensioned fixed transport network is used to connect UPF to gNB, and the latency is typically relatively small and consistent compared to the RAN. RAN is influenced by the stochastic and dynamic nature of wireless communication and applies a number of mechanisms to efficiently transmit data reliably over the radio link. Typically, the packet delay and PDV contributions of the RAN dominate the total packet delays and their variations of the 5G system. Therefore, a comprehensive investigation of the latency sources in the 5G NR RAN and its radio protocol stack are needed. In the following we investigate latency contributions in the different layers of the radio protocols, as shown in Figure 7 and Figure 8.

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 8); 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.

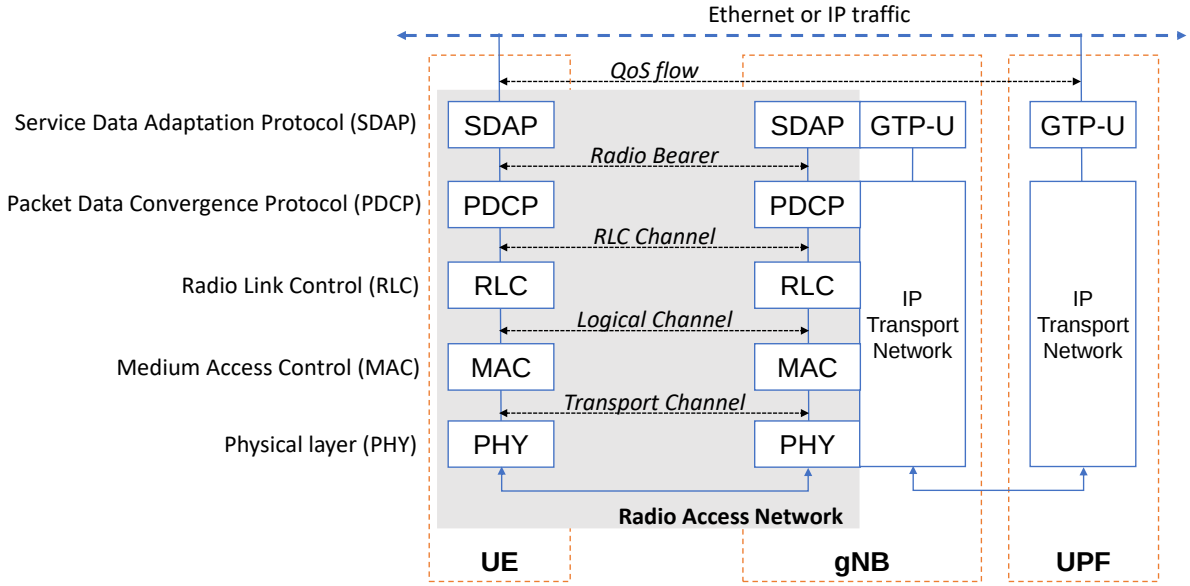


Figure 7. 5G protocol stack for user plane with focus on RAN.

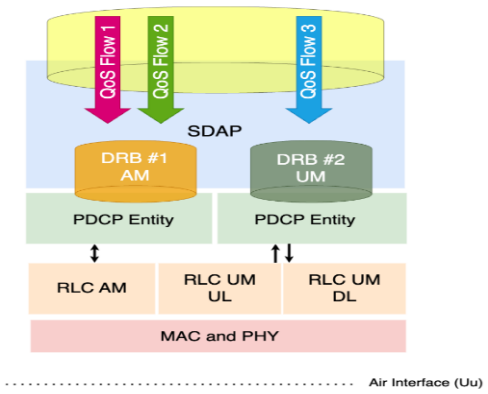


Figure 8. 3GPP 5G protocol stack and data flow [DPS21].

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 & 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, see Figure 9. 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 [DPS21]. 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 deliverable. 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.

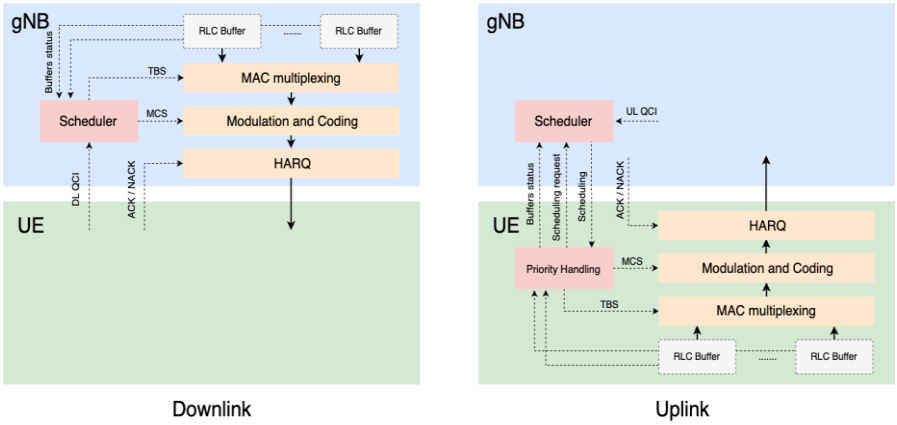


Figure 9. Transport format selection in (a) downlink and (b) uplink [DPS21].

Summarizing, several areas as contributors to the latency:

- Data transmission over the radio interface
- Processing delays at gNB and UE
- Traffic handling / queuing
- Reliability mechanisms (like HARQ)

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

**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 [SWD+18] [LSW+19]. The common slot lengths in deployed 5G networks have a length of 0.5 ms long (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. Most spectrum bands for 5G are licensed by public mobile network operators, and the available spectrum depends on the holdings of an operator and the bands available for 5G in a country. In addition, a spectrum for local deployments is available in several countries [NHB+21], with one intended use case being industrial networks. The majority of 5G network deployments are based on TDD spectrum allocations, particularly the midband (approx. 3-5 GHz) and highband (approx. 26-28 GHz) spectrum allocations. As a consequence, TDD is the most common deployment for 5G networks.

In TDD, a TDD pattern is used, which is a sequence of slots allocated for downlink transmission and a sequency of slots allocated for uplink transmission. Additional special slots are partly allocated to downlink and partly to uplink transmission, with some short guard period in-between, as shown in Figure 10. 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 if the networks are in close vicinity and use the same spectrum (i.e., co-channel coexistence) or are overlapping in area using adjacent spectrum carriers (adjacent channel coexistence). TDD coexistence is a serious concern in wide-area public mobile networks within countries (adjacent channel) and at country borders (co-channel), and consequently different wide-area networks are typically synchronized to the same TDD pattern, which may even be mandated by national regulators [ECC19-296]. 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 [5GS21-D14] [5GS21-D15] [CAS+22] [CAS+23]. 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. One such common TDD pattern called DDDSU is depicted in Figure 10. It contains three downlink slots (D), followed by one special slot (S) and one uplink slot (U). The special slot contains mostly downlink OFDM symbols, a short guard period and some uplink symbols. With each slot lasting 0.5ms, the TDD pattern repeats itself after 2.5ms. 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 [SWD+18] [LSW+19].

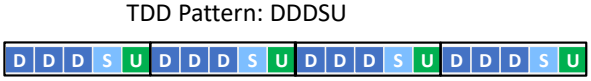


Figure 10. A TDD pattern of a 5G NR radio interface with Downlink (D), Uplink (U) and Special (S) slots.

**Processing delays in gNB and UE**

For the 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.

**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 of utter importance. 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 (see e.g., Figure 6). 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.

**Wireless transmission reliability**

The latency of the RAN is also impacted by the way reliable transmission is provided over the radio link. In wireless communication, the wireless propagation environment is inherently comprising dynamic variations of signal strength due to e.g., multi-path radio propagation with fast fading, shadowing and blocking by obstacles, distant-dependent path loss and possible variations of interference. Dips in signal strength due to fading can be up to several 10’s of decibels and can lead to fading-induced outage with losses of the transmitted data [JWE+15]. To counter such losses, 5G implements a large set of mechanisms to alleviate channel variations, such as adaptive link adaptation with adaptive modulation and coding, fast power control, channel-dependent and frequency selective scheduling, adaptive MIMO transmission. A new paradigm has been introduced with the 5G standard to address time-critical communications, for which features for ultra-reliable and low latency communication 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 [LLM+09] [MWR+06]. 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/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 11, 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.

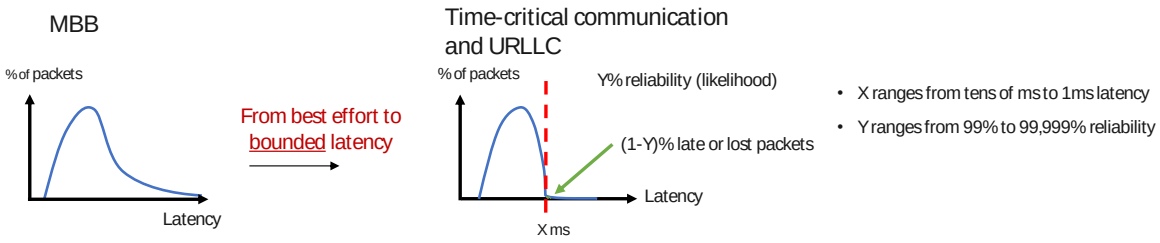


Figure 11. Time-critical communication with URLLC: from best effort to bounded latency performance.

The support of time-critical communications requires a re-thinking on how wireless communication reliability is provided. This can be best understood by looking at a good RAN configuration for mobile broadband services, where spectral efficiency and data rates are the key performance metrics and low latency is provided on a best-effort basis. A good configuration could look as follows:

- The link adaptation is configured for a target block error rate (BLER) of 10%.
  With this configuration not too many radio resources are allocated to a data transmission, so that a high spectral efficiency and system capacity can be achieved. The margin of “extra” radio resources to cover for variations of the wireless transmission is limited, which means that in approximately 10% of transmissions a packet cannot be successfully decoded at the receiver.
- The MAC HARQ is configured for several (e.g., four) retransmissions.
  Only for 10% of unsuccessful transmissions are additional radio resources invested to provide additional reliability by means of a HARQ retransmission. For the retransmission only incremental redundancy to the previous transmission is sent. The retransmission is soft-combined at the receiver with the already received but not yet decodable earlier transmission, which provides additional coding gain and energy combining gains compared to normal ARQ retransmissions. With four HARQ retransmissions at 10% BLER the percentage of remaining unsuccessful transmissions is less than 0.0001% with a maximum latency of four HARQ RTTs plus the time for the initial transmission. In typical TDD configurations a HARQ RTT is a few milliseconds.

- RLC ARQ is configured for several retransmissions.
  RLC ARQ is configured to recover from remaining packet losses due to HARQ failures. Several RLC ARQ retransmissions are permitted, with typical ARQ RTTs of some 10's of ms.

A sketch of the latency profile for such a configuration can be found in Figure 12. It shows several peaks with decreasing heights corresponding to the first transmission and up to four retransmissions. Due to the frame alignment delays caused by the frame structure and TDD pattern each peak has a certain width.

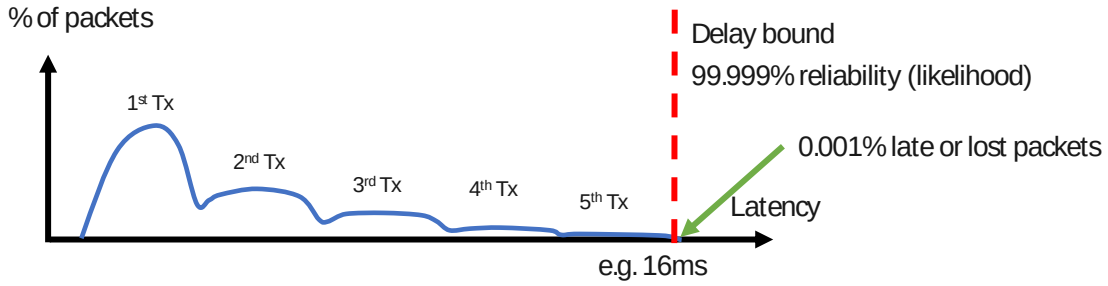


Figure 12. Sketch of latency histogram with HARQ

For time-critical communication services, 5G has been standardized with a tool of features that allow to provide high reliability even for very low latency bounds. A consequence of a low latency bound is that fewer HARQ retransmissions are possible, and in the extreme only a single transmission attempt is permitted that needs to be provided with very high reliability. Which latency bound is achievable depends on the configuration of the frame structure, the TDD pattern, the URLLC features and the availability processing capability (e.g., accelerated processing of "UE capability 2"). For targeted latencies in the low single-digit millisecond range, a specific 5G RAN configuration may be required. While many URLLC features can be configured per device, or per radio bearer, a configuration of frame structure and TDD pattern affects the entire RAN. This means that all UEs in the network need to comply with these configurations. An assessment of the latency of a 5G network has been made by 3GPP in [3GPP22-37910] following an agreed latency evaluation methodology according to Figure 13. A more extensive analysis of 5G latency performance can be found in [3GPP22-37910] [5GS21-D15] [5GS21-D14] [LSW+19] [SWD+18]. The achievable latencies for different configurations of the 5G RAN are depicted in Figure 14, a more extensive discussion of these options is provided in [5GS21-D15].

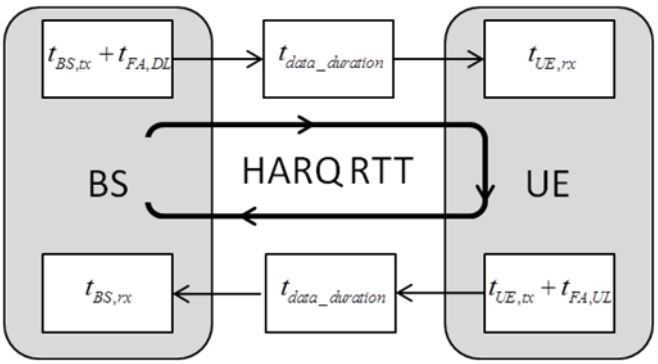


Figure 13. 5G user plane procedure for latency evaluation ([3GPP22-37910]) between a base station (BS, in 5G called gNB) and a UE.

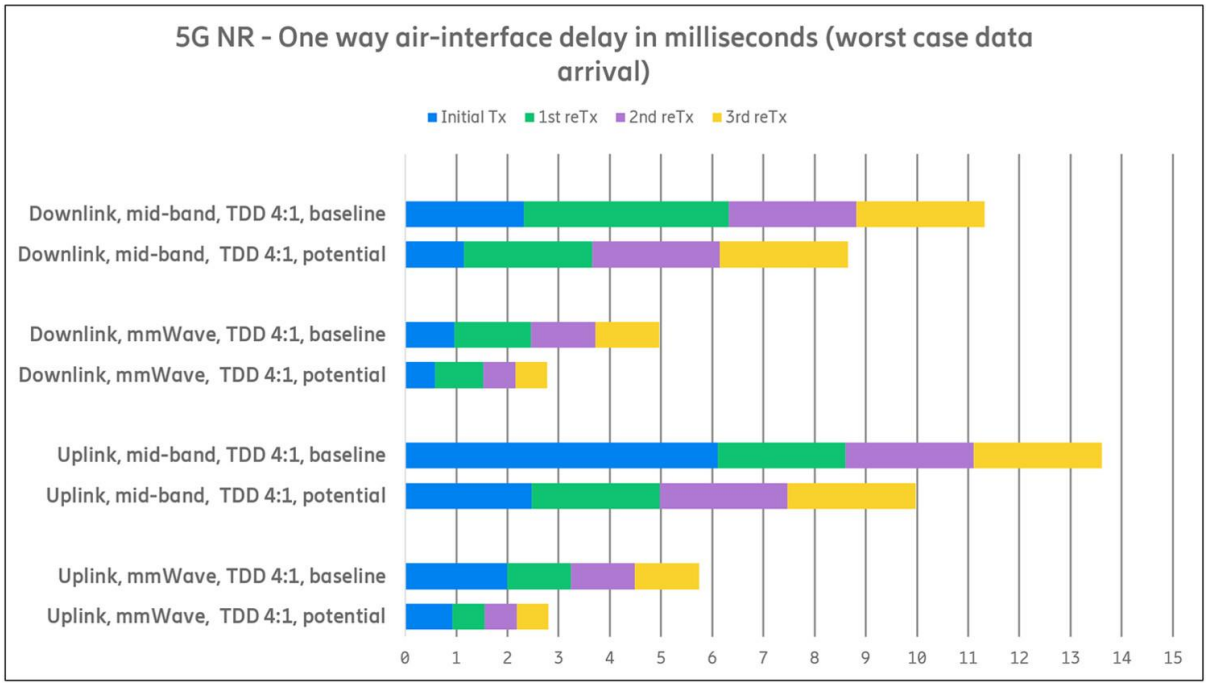


Figure 14. 5G NR one-way latencies for different NR configurations (see [5GS21-D15]).

The latency characteristics for a very low latency bound without HARQ retransmission is depicted in Figure 15. The latency distribution shows a single peak for the first transmission attempt. The configured functionality for ultra-reliable radio transmissions ensures that the single transmission attempts succeed with high reliability. A suitable 5G RAN configuration could look as follows:

- The radio bearer is configured for bounded transmission with 99.999% reliability on the first transmission attempt. Very robust transmission modes are configured for both the data channel and the control signaling. Radio resources can be pre-configured to provide many transmission opportunities for the data flow to shorten the channel allocation delay. Specific RAN configurations (e.g., frame structure, TDD pattern) may be needed for very low latency bounds.
- The latency introduced by MAC HARQ retransmissions may be too high to allow for HARQ retransmissions. In this case the HARQ may be disabled. However, since in this case the first transmission is configured with very high reliability, HARQ retransmissions have a negligible impact even when configured, since they would occur only in rare occasions. It shall be noted that if the latency bound permits for HARQ retransmissions, it is always advisable to utilize HARQ with an appropriate configuration as discussed below.
- RLC ARQ retransmissions are typically too slow for the targeted latency. With a very reliable configuration of the PHY and MAC, RLC retransmission will not be triggered in practice. RLC can also be operated in an unacknowledged mode.

Even if the radio bearer is configured for a single reliable transmission as in Figure 15, there is still a range of latencies that is perceived by the individual transmissions, which is mainly due to frame alignment: some packets arrive in the 5G RAN exactly at the time for the next transmission opportunity

and are immediately scheduled. Other packets may just miss the transmission opportunity and may need to wait, e.g., for an entire cycle of the TDD pattern.

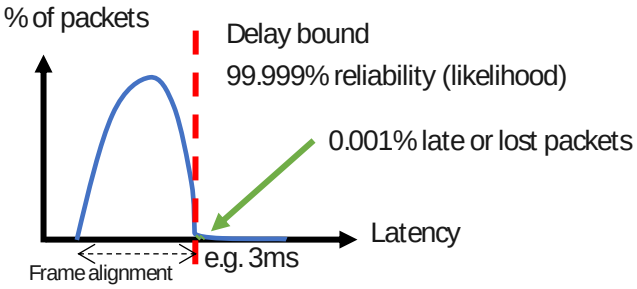


Figure 15. Sketch of latency histogram for time-critical communications with low latency bound using standardized URLLC features without HARQ retransmissions.

It is interesting to understand the conceptional description above with regard to real 5G networks. The packet delay in 5G trial networks has been measured in various 5G trial networks, see e.g. [AAB+22] [KAJ+22] [AVK+22] [DET23-D41].

Figure 16 shows the downlink packet delay distribution for a non-latency-optimized 5G network in which largely HARQ is used for reliability, see (a), and for a 5G network optimized for latency and reliability with URLLC features, see (b). It can be seen that by applying URLLC both the absolute latency and also the PDV can be reduced.

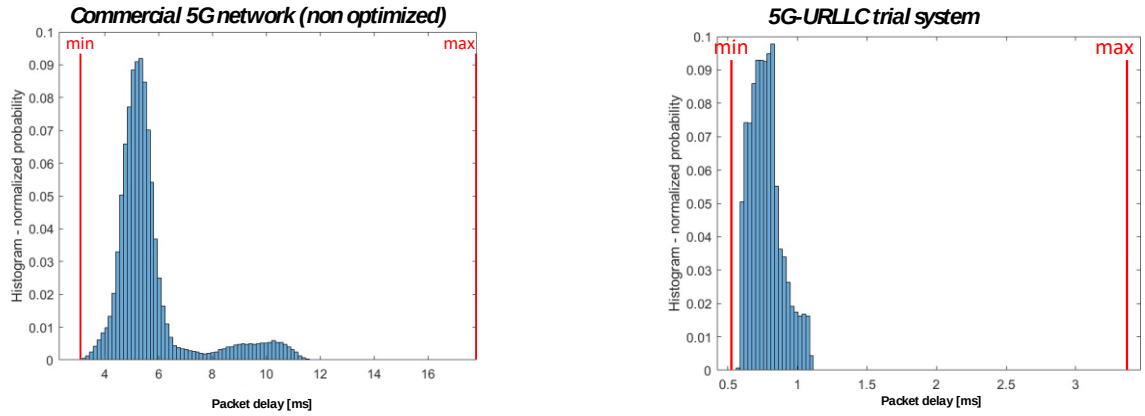


Figure 16 Histogram of downlink packet delay from two 5G trial networks for periodic data transfers (see also [DET23-D41]): (a) a 5G trial network operated at 3.7 GHz [AAB+22] (b) a 5G testbed with advanced URLLC features operated at 28 GHz [KAJ+22] [AVK+22].

The toolbox of 5G time-critical communications features allows to either push down the latency bound that can be provided with a certain reliability, or to increase the reliability from e.g., 99.9% to 99.9999%. However, applying such functionality comes at a cost. This can be understood by comparing Figure 12 with Figure 15 as explained in Figure 17. To achieve lower bounded latencies, a fewer number of HARQ retransmissions are permitted. This means that the packet transmissions which perceive the worst transmission conditions (e.g., due to channel variations) must be improved by boosting the reliability of their transmission. As it is not known beforehand which transmissions are most affected by e.g., drops in channel conditions, all transmissions must be generally protected to a

higher level. For a latency bound targeted with a reliability level of 99.999% this means that to ensure that the worst performing packet transmission reaches its reliability target also the other one million packets need to be boosted with an increased reliability margin. This implies that the majority of packet transmissions are transmitted with more allocated transmission resources than would actually be needed. For this reason, the spectral efficiency of 5G decreases for lower latency bounds that are provided. In other words, for the same type of traffic, a 5G network can support more simultaneous connections for providing a latency of 16 ms at 99.999% reliability than what can be supported for a latency of 3 ms at 99.999% reliability. With regard to spectral efficiency the application of HARQ is a very efficient way to boost reliability as additional transmission resources are only invested in cases where they are needed (i.e., a packet could not be received correctly). Therefore, when the delay bound required by the application allows to cater for one or more HARQ retransmissions it is advisable to utilize HARQ in the reliability planning.

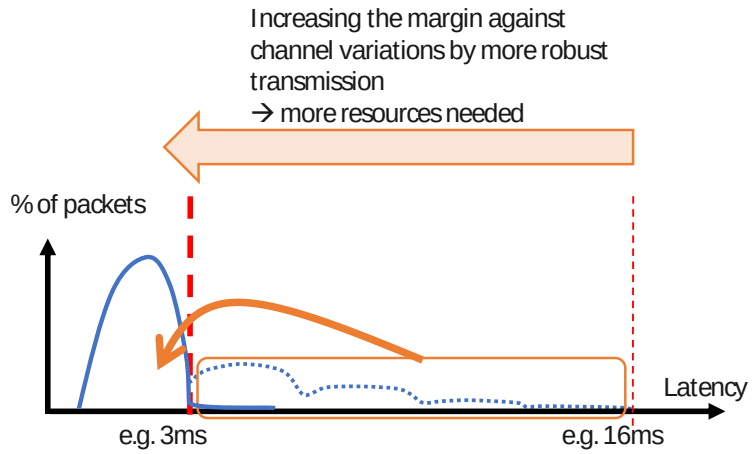


Figure 17. Concept of applying 5G time-critical communication features for reducing latencies.

The discussion so far has focused on the latency bound and the reliability for time-critical services. This has also been the focus of 5G standardization. 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 switches. Constructive measures to compensate for the packet delay variation can be found in the next section.

# 3 Packet Delay Correction

As seen in section 2, it is possible to achieve low latency and high transmission reliability by means of intrinsic 5G features. However, the PDV is considerably large compared to wired TSN bridges. It is desirable to limit the PDV of the virtual TSN node/6G network to a similar level as determined in the wired TSN nodes to ensure integration and interworking of the TSN system with the wireless communication network/6G network. To enable a dependable time-critical 6G service will require an additional mechanism to guarantee a limited PDV (e.g., down to tens of microseconds). The idea is to force packets to be transmitted to the next TSN node right on time, i.e., not earlier and not later. The upper delay bound can be set with a defined reliability level by the 5G/6G system, hence ensuring packets are not delivered later than they should. A correction mechanism is used to ensure the lower

delay bound is as close as possible to the upper bound, such that PDV is extremely small, and this is what we define as Packet Delay Correction (PDC).

This section offers an account of previous work and discussions in 3GPP related to de-jittering the TSN streams. Then, a set of concepts related to PDC are proposed and described. To conclude, the gains and challenges of PDC mechanisms are discussed. Even if the 5G network architecture is used to explain the concepts in this chapter, it is important to keep in mind that PDC is not currently available for 5G and should be expected for the next generation, 6G.

## 3.1 Background

The task to compensate delay variation in 5G initiated in 3GPP at the System Architecture WG 2 (SA2) where impact to the 5G was the target. Hence 3GPP defined a "hold and forward buffering" mechanism to be applied at the TSN Translators (TT), i.e., DS-TT and NW-TT. This mechanism was intended originally to pace out the packets such that it looked like there was no delay variation in the packet transmission, or such that the PDV was bounded through the 5G system.

Even when a precise way to guarantee a very low or near-zero PDV in the 5G was necessary for a viable CNC schedule, 3GPP decided to leave it out for implementation. Hence, the details on how the hold and forward buffering mechanism is provided by the DS-TT and NW-TT were not defined, and merely indicated that this mechanism will mimic the behavior of the timed gates of TSN scheduled traffic (a.k.a. Qbv).

As the envisaged de-jittering would require receiver actions that are aligned with the transmitter actions – and given the heterogeneity of 5G network and device vendors, a working solution of the hold and forward mechanism needs to be standardized. At the minimum both DS-TT and NW-TT should apply the same mechanism. For a future 5G release or 6G, it is expected to solve this issue by specifying a common mechanism to achieve a compensation of the PDV such that time-critical communication may be enabled.

Among the proposals discussed in 3GPP Release 16, the general idea was to keep holding every packet until they reach a predefined maximum delay which is based on the Packet Delay Budget (PDB), a QoS parameter to be delivered by the 5G system for a traffic flow.

A solution was proposed and based on timestamping a packet at ingress and egress of the 5G system [3GPP16-2002055] [3GPP16-2002056], i.e., at DS-TT (ingress) and NW-TT (egress) for uplink, and at NW-TT (ingress) and DS-TT (egress) for downlink. When the time that a packet has spent within the 5G system is known (by subtraction of egress and ingress timestamps), then the time to hold the packet in a buffer to reach its predefined maximum delay value is also known. Wire speed timestamping on legacy hardware is usually a very challenging task unless solutions using programmable packet processors [KSB+20] can be implemented in DS-TT and NW-TT. The proposed solution defined the use of virtual time slots instead of hardware timestamping, which is based on the fact that the 5G clock is common for both NW-TT and DS-TT. Only the 5G system is aware of the virtual time slots, and the packet will be marked with the virtual time slot that corresponds to the arrival of the packet and will later be marked as well at the egress of the 5G system. At that point, it is possible to know the number of virtual time slots that the packet has spent in the 5G system (a.k.a. residence time). The difference of the maximum delay and the 5G residence time (predefined for every flow belonging to the same traffic class and port pair) determines the number of time slots that the packet shall be held before it is transmitted. In this way, all packets in flows belonging to the same traffic class

and port pair will experience the same predefined maximum delay, and therefore resulting in small delay variation.

The original motivation for compensating packet delay variations came with the standardization of supporting Ethernet-TSN with 5G. The proposed solution [3GPP16-2002056] was to use the Ethernet R-Tag (as defined in IEEE 802.1CB) at the Ethernet frame header for transferring packet timing information between ingress and egress TSN translators. In this case, the "Sequence Number" field, as shown in Figure 18, will carry the value of the virtual time slot. Nothing prohibits the use of multiple R-Tags in an Ethernet frame. In this case, an R-Tag already defined for other purposes can be reused to insert a virtual timestamp.

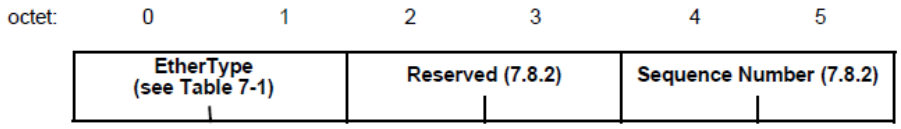


Figure 18. R-TAG format, as defined in IEEE 802.1CB [x]

Virtual slots are the same on all the 5G system components (including DS-TT and NW-TT) as they are all synchronized with the 5G grandmaster clock. Slot ID refers to the arrival slot that is encoded in the R-Tag. The tasks at the ingress are: (i) identify the time slot with specific Slot ID when the packet arrives, (ii) encode Slot ID in the added R-Tag (as Sequence Number). At the egress the tasks are: (i) identify the egress Slot ID, (ii) based on R-Tag and stream specific targeted residence time within the 5GS, calculate the number of time slots that the packet should be held before transmitting (i.e., targeted delay), (iii) buffer the packet for the calculated targeted delay, (iv) remove R-Tag and transmit.

In Figure 19, the use of the virtual timeslots is shown using an example. Assuming that there are three packets arriving at Slot ID 2, 6, and 10, the ingress TT will encode the slot ID in each respective R-tag field of the packets. When the packets arrive at the egress TT, the residence time (3 slots) is added to the encoded slot ID of every corresponding packet in order to obtain the egress slots: 5, 9, and 13, respectively. Note that communications can go from and ingress TT, such as DS-TT at UE, to an egress TT, such as NW-TT at UPF, or vice versa, from NW-TT to DS-TT. UE-UE communication (between DS-TT and another DS-TT) is considered as well.

This additional R-Tag is used only within the 5G system. It is added at ingress and removed at egress of the 5GS. Therefore, the R-Tag used inside the 5G system is not visible to the outside world. Ethernet frames may use other R-Tag(s) for which the 5G system is transparent, and hence these R-Tag(s) remain untouched by the 5G system.

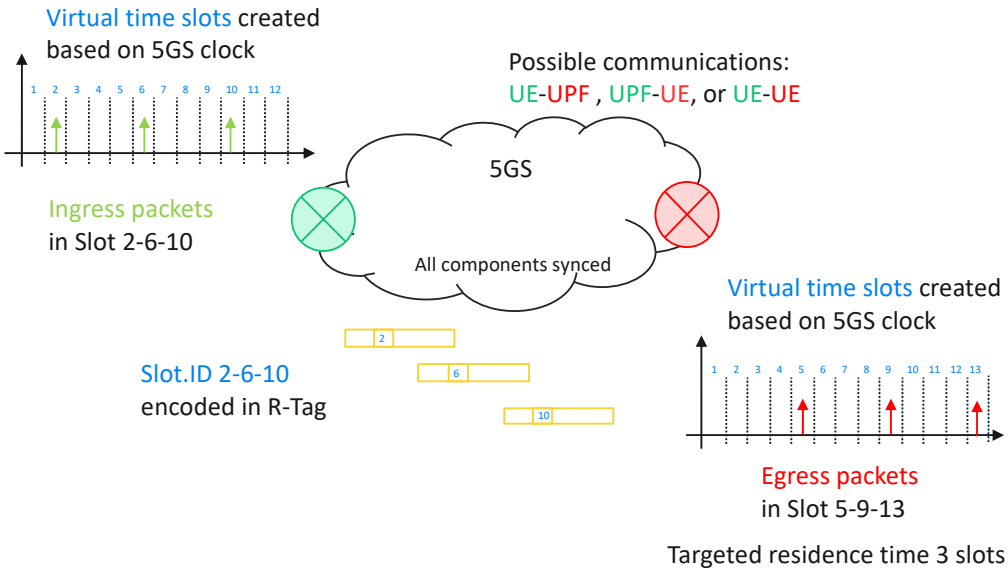


Figure 19. Timestamp-like solution: example on how it works.

## 3.2 Proposed PDC Methods

In this section a number of solutions are proposed to PDC-related problems. One problem is to generalize the proposal of virtual timeslots to any type of communication, and not only be applicable to Ethernet. A second problem is how to transfer the virtual timeslot or pseudo timestamp within the 6G system. Also, different alternatives are discussed based on to which segment within the 6G system the PDC is applied as well as where the calculations of forwarding time can be made. Finally, a different PDC approach than using timestamps is proposed, and it is based on the number of radio retransmissions.

### 3.2.1 Generalized Timestamp-based Packet Delay Correction

The proposed approach described in section 3.1 can be leveraged to achieve a more general solution that can apply to any communication technology. Indeed, using the R-tag of Ethernet frames is limited to Ethernet-based communication. The main idea is to propose a mechanism that can be applicable to any type of communication, IP, Ethernet, etc., and regardless of using a specific real-time technology such a TSN or DetNet.

In this section we propose a general mechanism based on the same principle of the virtual timeslot that implements a sort of pseudo-timestamp. This has the benefit of avoiding the implementation of wire-speed egress timestamping at physical layer (hardware timestamps). Instead, the focus is to implement pseudo-timestamping at higher layers (software timestamps). Even when these timestamps have lower resolution or precision, this is still useful to minimize the PDV, without the burden of hardware timestamping for every single packet.

The general solution is illustrated in Figure 20, and described in the following steps:

1. First, an external management controller entity (e.g., CNC, CUC, DetNet controller, via Open Platform Communications Unified Architecture (OPC UA) semantics, generic AF for TSC, etc.) sends a request to the 5G/6G control plane regarding the required maximum delay (Max delay) and maximum jitter (Max jitter) that a specific application can tolerate. These parameters characterize the traffic stream that will be transferred through the 5G system.

Note that currently CNC does not provide such parameters, CUC does not communicate directly with the 5G bridge, and the DetNet controller does not include such parameters in the YANG model. However, this type of exchange is envisioned for future 6G time-critical communication.

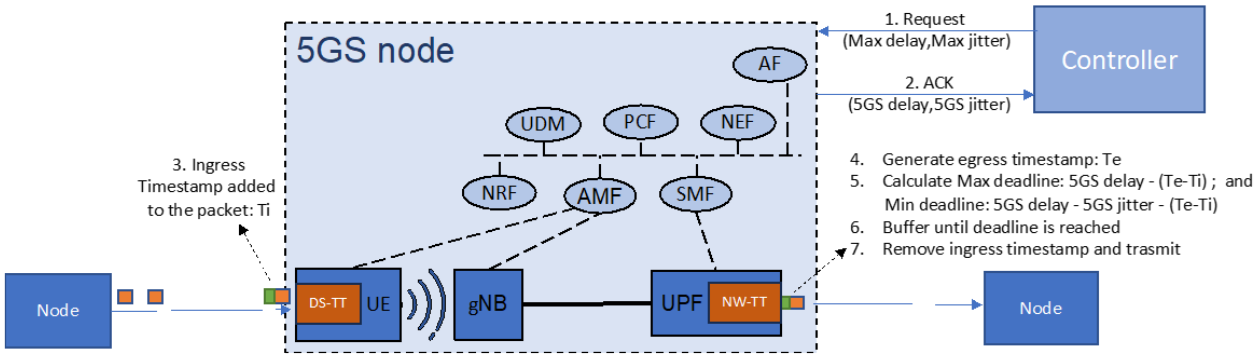


Figure 20. Generalizing timestamping for packet delay correction

2. The request is handled by the 5G core network, and an acknowledgement is provided back from the 5G control plane towards the controller indicating the 5G system (5GS) delay and 5GS jitter. The acknowledged 5GS parameters may be lower than the required parameters, or the same. Otherwise, if 5GS cannot provide at least the maximum required values for delay and jitter, then the 5GS may notify that the support delay and jitter bounds cannot be met, or the communication request is rejected. The negotiation of parameters between CNC and 5GS can be extended further such that it includes more parameters e.g., the feasible limits for the 5GS jitter.
3. Once a packet of the traffic stream in question arrives either at DS-TT or NW-TT port, an ingress pseudo-timestamp or virtual timeslot Ti is generated and added to the packet's header. For simplicity we assume an example where the flow is uplink and therefore the packet is received at the DS-TT. Note that in the next section, the matter of which header and communication protocol to be used withing the 3GPP protocol stack is discussed.
4. After the packet was transferred transparently through the 5GS, the egress port will generate an egress pseudo-timestamp or virtual timeslot Te. In this example the egress port is the NW-TT, however it could be another DS-TT for the case of UE-to-UE communication. This pseudo-timestamp is used in the next step but not added to the packet's header.
5. The egress TT, in this case the NW-TT, calculates the packet's transmission deadlines, i.e., Max deadline and Min deadline, as follows:

   Max deadline = 5GS delay - (Te-Ti);

   Min deadline = 5GS delay - 5GS jitter – (Te-Ti).

   The Max and Min deadlines correspond to the interval in which the time to transmit the packet is allowed.

   The value of 5G delay and 5G jitter can be expressed in terms of virtual timeslot size. 5G jitter is the amount of jitter that can be allowed in the system, it could be set to an appropriate value, e.g. in the order of 10's of microseconds or larger values according to the tolerated jitter by the end-to-end time-critical communication. Note that the virtual time slot inherently includes a jitter depending on the resolution or granularity of the implemented virtual

timestamps Te and Ti values are a specific virtual timeslot number/id, which is known in both DS-TT and NW-TT due to the fact that these TTs are synchronized with the 5G clock.

6. If the pseudo timestamp of the packet is not currently within the deadline interval (Max deadline – Min deadline), then the packet shall be buffered and held until the time ticks till the virtual timeslot that is within the deadline interval.
7. When the packet has reached a virtual timeslot that is within the packet's deadline interval, then the timestamp is removed and the packet enqueued for transmission over the port.

### 3.2.2 Advanced PDC Using 3GPP Protocol Stack

In this section, it is proposed an approach to insert the pseudo-timestamp in the header of the packet. Timestamp based PDC requires the transfer of (pseudo) timestamps along with user-plane packets in some protocol header. In section 3.1 a PDC was proposed which is useful only for Ethernet-based communication. Instead in this section we focus on generalizing the way pseudo-timestamps can be delivered through the 5G system. An important remark is that the PDC method is only to be applied within the 5G system (at the DS-TT and NW-TT usually), therefore there is no need to use network protocols (and their packet headers) that can span outside the 5G system. Instead, the approach would be using the internal 3GPP protocol stack.

The idea is to use a protocol layer that spans between UE and RAN node (base station or gNB), and a protocol layer that spans between RAN node and UPF. The protocol sub-layer that serves the segment UE to gNB is Service Data Adaptation Protocol (SDAP), and the protocol layer that serves the segment gNB to UPF is the GPRS Tunneling Protocol for User plane data (GTP-U), as shown in Figure 21.

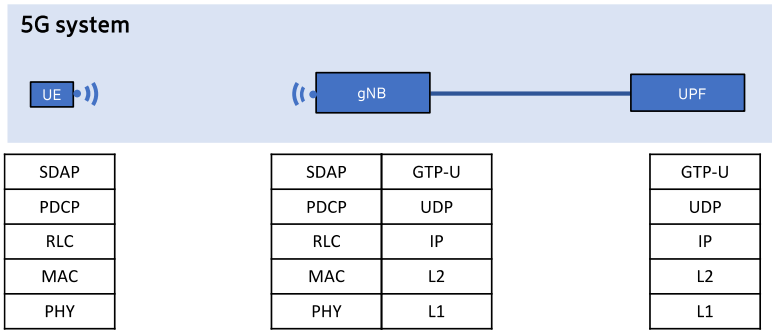


Figure 21. 3GPP protocol stack.

The GTP-U protocol header at the packet already contains a "Next Extension Header Type" field that can be reused (see highlighted octet 12 in Figure 22). The "Next Extension Header Type" of a GTP-U PDU can be enhanced to indicate when timestamp information is included in a GTP-U PDU that relays a downlink packet (e.g., the UPF receives the downlink packet and corresponding timestamp from a NW-TT) or an uplink packet (e.g., a gNB receives the uplink packet and corresponding timestamp from a UE). In Figure 23, the new type of header extension is added to support an ingress timestamp.

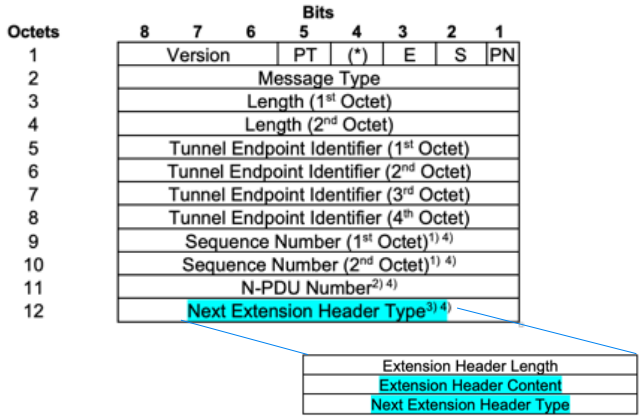


Figure 22. Reuse of the GTP-U header for timestamping

| Next Extension Header Field Value | Type of Extension Header |
|---|---|
| 0000 0000 | No more extension headers |
| 0000 0001 | Reserved - Control Plane only. |
| 0000 0010 | Reserved - Control Plane only. |
| 0000 0011 | Long PDCP PDU Number. See NOTE 2. |
| 0010 0000 | Service Class Indicator |
| 0100 0000 | UDP Port. Provides the UDP Source Port of the triggering message. |
| 1000 0001 | RAN Container |
| 1000 1001 | Ingress Timestamp |
| 1000 0010 | Long PDCP PDU Number. See NOTE 3. |
| 1000 0011 | Xw RAN Container |
| 1000 0100 | NR RAN Container |
| 1000 0101 | PDU Session Container. See NOTE 4. |
| 1100 0000 | PDCP PDU Number [4]-[5]. See NOTE 1. |
| 1100 0001 | Reserved - Control Plane only. |
| 1100 0010 | Reserved - Control Plane only. |

Figure 23. Adding a new type of extension header for the ingress timestamp

The base station or gNB supports both SDAP and GTP-U protocol stacks, therefore it can transfer the ingress timestamp value from GTP-U to SDAP in downlink or from SDAP to GTP-U in uplink direction.

SDAP (for UE-gNB segment) has a simpler header compared to GTP-U header. The proposed solution is to add the same header extension header in GTP-U in the SDAP header for downlink and for uplink. Also, a new type of extension header needs to be defined which also includes the ingress timestamp.

*Alternative*

Other equivalent header changes are possible, the essential idea is to allow transporting the pseudo timestamp using the 3GPP protocol headers. Alternatives might be considered with regards to which segment of the 5GS will be targeted. It is well understood that the major component of PDV is the RAN, while the transport network is fixed and suffers little or minimal variation in the per-packet delay. Therefore, an alternative is to apply PDC between gNB and UE only, e.g., based on the SDAP protocol extension described above. This may lead to some uncertainty since the PDV at the transport network (gNB to UPF) is not being compensated, however it might be acceptable for the majority of applications as the variations are considered to be quite stable in this segment. It is also useful to apply

the QoS monitoring capability standardized for GTP-U in order to estimate the delay variation of the transport network (gNB to UPF) segment. This may give a sufficient estimation of the PDV in the 5G transport network.

### 3.2.3 PDC Based on Number of Radio Retransmissions

This section describes a method, which corrects the packet delay variations based on the RAN internal states/processes. For the correction it considers the packet specific RAN events (i.e., radio re-transmissions).

#### *State of the Art*

There is a long history of various queuing techniques, as they are under discussion since packet transport was developed. The history started with "static queues", where Streams/Flows and Queues were paired statically (this is a 1:1 mapping). The target of queuing was to provide appropriate bandwidth for the services. Queue selection is done by mapping traffic based on frame/packet header fields (typically implemented with ACL: Access Control List). Service of the queues is based on pre-defined rules (e.g., PQ: Priority Queueing, WFQ: Weighted-Fair Queuing, CB-WFQ: Class-Based Weighted-Fair Queuing).

With the advent of TSN/DetNet a new epoch started: the era of timed queues. Streams/Flows are more dynamically mapped to queues (this is a 1:n mapping), based on header fields and the actual time of the system. Similarly, the serving of the queues is driven not only by bandwidth but by time parameters as well. IEEE defined multiple standards to describe these timed queues for TSN: 802.1Qci (aka, PSFP: Per-Stream Filtering and Policing) allows time-based queue selection and 802.1Qbv (aka, Scheduled Traffic) describes time-based queue service.

Recent research and standardization work focuses on a third era dealing with "dynamically controlled meta-queues", which can add further flexibility to queueing systems. In these solutions the selection of queue is based on metadata (it can be local to the node or travel with the packet e.g., in special header fields). Furthermore, the serving of the queues can be also based on metadata e.g., packet-timestamp, urgency indicator.

#### *Proposed Approach*

Here a new packet delay correction (PDC) mechanism is proposed for the 6G system which ensures an upper bound for both the wireless transmission latency and the packet-delay variation of the transmission. These characteristics are achieved without a fundamental change of RAN components, but via reusing their internal states during the transport of a given packet.

Figure 24 shows a possible strategy to set bounds on the packet delay distribution, via additional delay during the forwarding.

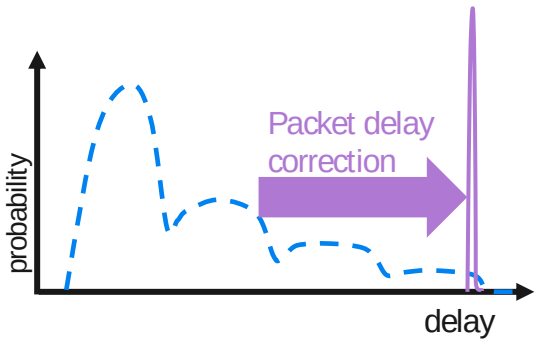


Figure 24. Packet delay correction solution via additional delay.

The new PDC function is placed after the air transmission and controls the lower bound of the packet delay distribution to become tighter toward the upper bound and thereby controlling the maximum variation of packet delay. Practically, it decreases PDV of RAN segment caused by HARQ process. PDC provides significant improvements for time-critical communication compared to legacy 5G systems. While 5G only provides control of the upper bound of the latency, such a PDC function enables the next step in determinism with bounded and predictable latency behavior for 6G.

The proposed method uses two components to decrease the PDV of forwarded packets: (1) the no. of retransmission information for each packet and (2) a specific egress interface service queuing system (CQF: Cyclic Queuing and Forwarding). The basic idea of this approach is to assign a packet that has waited a long time for transmission due to multiple unsuccessful re-transmission attempts to a queue that is served in near future. The selection is adaptive and depends on the number of re-transmission attempts since each attempt adds delay to forwarded packets.

Below is the procedure in detail:

1. Packet received over the air after a number of re-transmissions (#ReT={0, 1, 2, …})
2. A queue (“j”) of the PDC function is selected based on the number of retransmission(s): “j=i+(N-(#ReT+1))”,
   where “N” denotes the number of queues used by the CQF system, and “i” denotes the actually served queue.

Note: queue numbering is {1, 2, … , N}, and calculation for selecting which queue has to store the received packet (”j”) is a modulo N operation.

Figure 25 shows the CQF system of the PDC.

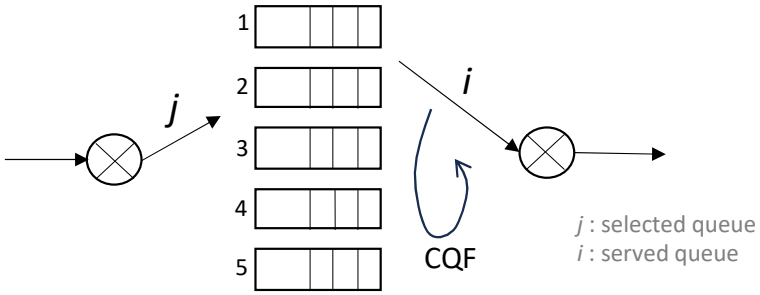


Figure 25. Packet delay correction by proper queue selection in CQF (example).

Note: there are multiple valid strategies to select the required number of queues (N).

The queue(s) of PDC are at the RAN e.g., in PDCP/SDAP (SDU level) at the receiver side. The serving time of queues in the CQF system equals to the time needed for a HARQ retransmission cycle over the air interface:

$$T_{\text{CQF-service-time-per-queue}} = T_{\text{HARQ-RTT}}.$$

Therefore, the resulted packet delay (PD) is bounded and it is in the following range:

$$\{ (N-1) \times T_{\text{HARQ-RTT}} ; N \times T_{\text{HARQ-RTT}} \},$$

so, packet delay variation (PDV) is also bounded $PDV=T_{\text{HARQ-RTT}}$.

Note that depending on the traffic situation and CQF service rate the PDV can be significantly lower than $T_{\text{HARQ-RTT}}$, (e.g., in a lightly loaded system).

For proper operation the receiver component has to know the number of transmissions done for the packet. This can be achieved by various methods and is subject to further studies:

- Solution1: Knows it from the radio scheduler
- Solution2: The number of tx is signaled to the rx component
- Solution3: By other means

The above described PDC method is 6G system internal functionality and its details are not visible to the remaining part of the TSN/DetNet network. Per-stream guarantees given by this approach and its interworking with other queueing methods used in the wired domain are subject to further studies. However, exposing the internal states of 6G component(s) could be useful to create new e2e architectures (to be discussed/studied in WP3).

### 3.2.4 Gains and Challenges of PDC Methods

The combination of URLLC concepts as known from 5G and 6G PDC provides means of resource efficient latency control. Not all time-critical communication services require ultra-low latency, but still the latency needs to be predictable. As URLLC allows to control the upper bound and PDC the lower bound of the packet delay, the latency of 6G transmission can be configured for different target delay values and still have deterministic latency with bounded variation, as illustrated in Figure 26. The combination of URLLC and PDC we denote as dependable time-critical 6G transmission, with bounded packet delay variation around a configured target latency.

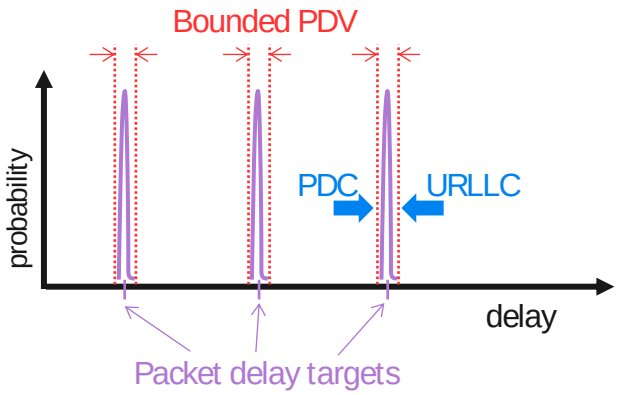


Figure 26. Dependable 6G transmission with configurable packet delay targets in terms of upper delay bound and maximum PDV.

The required ambition for PDC in 6G is to be able to provide packet delay variations that can be bounded down to a level of 10's or 100's of microseconds with high probabilities (e.g., 99.999%).

An ideal PDC does not require redesign of RAN and has no cost regarding spectrum efficiency.

The main challenge for PDC is standardization. Agreement to agree on a single PDC mechanism is challenging in the 3GPP arena. As described, in order to achieve PDC it is required that both ingress and egress TTs implement the mechanism, or the UE and the UPF, or in the case of RAN-centric PDC have the same mechanism in both UE and gNB. Different vendors of the mentioned devices will have different and even proprietary PDC mechanisms that will not be compatible for interoperation. Hence, in such case, PDC would be only functional with TTs from the same vendor.

Other aspects to consider are the precision of the PDC considering differences between hardware timestamps and using virtual timeslots. However, a low variation of the per-packet delay can be acceptable using virtual timeslots if PDV is limited to 10's of microseconds. Defining the size of virtual time slots will directly affect the level of PDV, and further study is required.

The PDC based on the number of radio retransmissions may require further investigations with regards to defining the buffer size in relation to the head of line blocking.

# 4 Data-driven Latency Characterization

The system enhancements discussed until now are aimed at improving PD and/or PDV. In addition to these enhancements, methods are required for seamless integration of wireless (5G/6G systems) with wired (TSN/DetNet) communication domains. In this section, we will describe a data-driven approach for conditional density estimation of latency and also touch upon architectural aspects of latency prediction relevant to 5G-Adv/6G networks.

## 4.1 State-of-the-art in Latency Characterization

In literature, several approaches have been proposed for the latency characterization of communication networks. These approaches can be roughly divided into two major categories: (i) Model-driven approaches and (ii) Data-driven approaches as shown in Figure 27. Next, we discuss these two categories of approaches in detail.

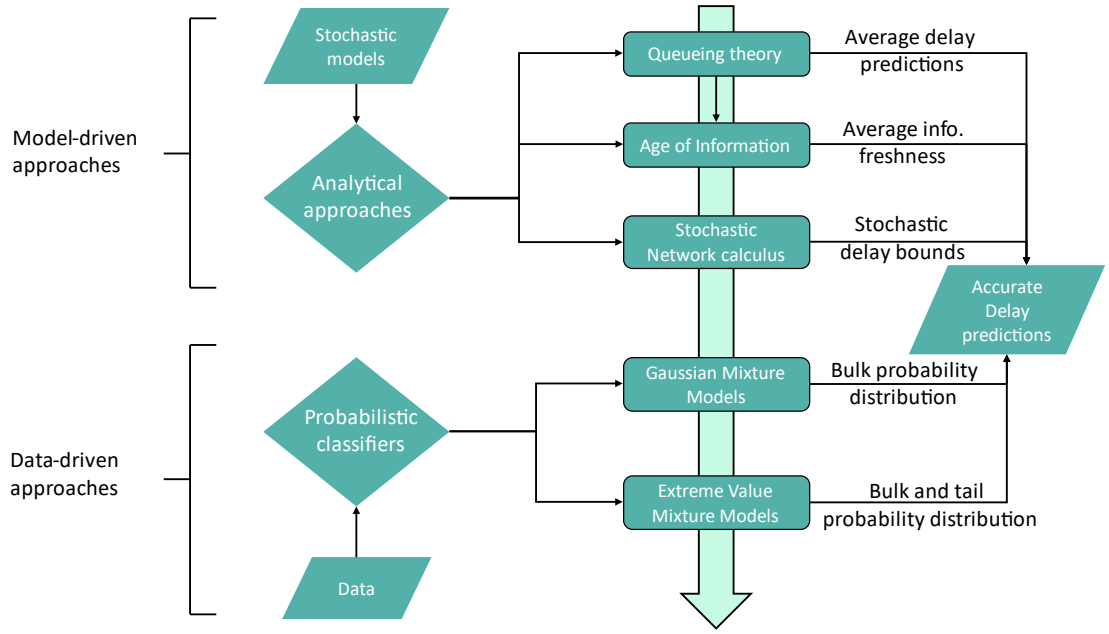


Figure 27. Classification of the state-of-the-art approaches for latency characterization.

### 4.1.1 Model-driven Approaches

Model-driven approaches for latency prediction involve using mathematical models or simulations to estimate, e.g., latency or response time of a communication system. In model-driven approaches, mathematical models are built using the knowledge and rules of the system. These mathematical models can then be used to derive bounds on the system performance, e.g., average BER can be derived using a given Signal-to-noise-plus-interference ratio (SINR). Models of wireless networks can be developed using queueing theory, Age-of-Information (AoI) and stochastic network calculus.

Mathematical models from queuing theory have been used to estimate the packet latency in wireless networks [BIA00][TS08]. The idea is to represent various elements (e.g., UE or gNB) of the network as a queue and derive KPIs such as latency and throughput using the framework of queueing theory. In order to simplify the derivation of average KPIs, these models take certain assumptions on the arrival and service processes, e.g., modeling the arrival and service processes using Poisson distributions (with interarrival times and service times), respectively. However, these assumptions are not representative of the complex dynamics present in wireless networks. Furthermore, the average KPIs (e.g., average latency) might not be useful in scenarios where instead of average values high quantile guarantees (e.g., 99.999%) on latency are required. In contrast to queuing theory, AoI framework KPIs describe the timeliness or freshness of information at the destination instead of the latency through the system. However, the output of the AoI analysis provides average timeliness and still does not provide bounds on the higher quantiles of latency. Finally, Stochastic Network Calculus (SNC) is another mathematical framework that incorporates probability theory in queueing theory to analyze system performance. The idea is to characterize traffic and service using mathematical curves and apply corresponding operations (e.g., max-plus algebra) to derive stochastic bounds on packet latency. For instance, authors in [CAG20] analyzed multi-hop wireless network using SNC to obtain end-to-end latency delay and backlog. Like other model-based approaches, SNC also takes certain assumptions on traffic arrivals as well as service.

### 4.1.2 Data-driven approaches

Unlike model-driven approaches, data-driven approaches utilize machine learning techniques to identify and learn relationships between system KPIs and other variables from the measurement data collected in the real-world systems or simulations. Authors in [KFR19] introduced a neural network-based approach that was trained on packet dispersion values, effectively identifying bottleneck links and approximating residual bandwidth in the network. However, such methods often yield single point estimates of performance metrics, like average throughput and delay, which again prove insufficient when applications demand high-quantile performance assurances [BDP18]. Data-driven techniques have also been employed to produce probability distributions for diverse performance metrics. A histogram-based approach to estimate conditional Round Trip Time (RTT) probability distributions in IoT systems is presented in [FYJ20]. Mixture Density Networks (MDNs) with Gaussian and Log-normal distributions, along with histograms, were employed to construct conditional distributions, each resulting in distinct trade-offs. Another work in [SSI22] proposed an MDN based on Laplace distributions to predict delay jitter in high-frequency and mobile communication. It is worth noting that these estimators are good at predicting the bulk/central portion of latency distributions but may fall short when handling tail probabilities—an aspect that cannot be ignored in applications requiring stringent latency guarantees.

## 4.2 Conditional Density Estimation

Next, we formulate the problem of latency prediction problem as a Conditional Density Estimation (CDE) problem.

### 4.2.1 Problem Overview and Approaches

Latency CDE is the problem of inferring the PDF of latency, considering various conditions present in the system. In other words, the objective is to estimate the likelihood that packets undergo specific delays (latency), given the circumstances under which the system operates. For instance, one might be interested in estimating the PDF for link latency, given channel (e.g., RSRP, RSRP, etc.) and traffic (e.g., packet arrival rate, 5QI) conditions. The conditional density estimation problem for such systems can be represented mathematically as:

$$\hat{p}(\mathrm{Y}\,|\mathrm{X}=\mathrm{x}) \approx \mathrm{P}[\mathrm{Y}|\mathrm{X}=\mathrm{x}].$$

Here, the latency through the system is denoted by the random variable $\mathrm{Y}$, whereas the system conditions random variable is represented by $X$. The objective is to estimate $\hat{p}$ conditioned on $X = x$ while the true probability is denoted by $\mathbb{P}$.

Solving such problems involves using a parametric model, usually a Gaussian Mixture Model (GMM), to capture the distribution of the latency samples. In a non-conditional density estimation scenario, a GMM with finite-dimensional parameters is fitted to a given dataset containing latency samples. The parameters of the GMM are represented by $\theta$, including weights, centers and variances of the Gaussians. In the case of CDE, the challenge is to is to map the system condition values that $X$ takes into the parameter space $\theta$ of the density function. This mapping is achieved through a function $h_\omega$ that connects the conditions to the density parameters such that $\theta = h_\omega(x)$. This mapping could be obtained by using a fully connected neural network, referred to as Mixture Density Network (MDN) whose parameters are represented by $\omega$. The values for $\omega$ are determined through maximum likelihood estimation (MLE), ensuring that the likelihood of observed samples is maximized, which corresponds to minimizing the Kullback-Leibler divergence (KL-divergence) between the actual data distribution and the parametric density $\hat{p}$ [BIS94]. To enhance the accuracy, particularly at the tail of

the PDF, the Extreme Value Theory (EVT) models have been proposed to be used alongside GMMs. The combined approach leverages a parametric mixture function that combines a GMM with the generalized Pareto distribution (GPD) tail model, which we refer to as Extreme Mixture Model (EMM). This combination of the two distributions helps capturing both the bulk distribution and the tail behavior accurately instead of just using GMMs. The combination of the two distributions can be expressed mathematically as follows:

$$\hat{p}(y|\theta = h_\omega(x)) = \begin{cases} f(y|\varphi) & y \le u \\ [1 - F(u|\varphi)]g(y|\beta, \xi, u) & y > u \end{cases}$$

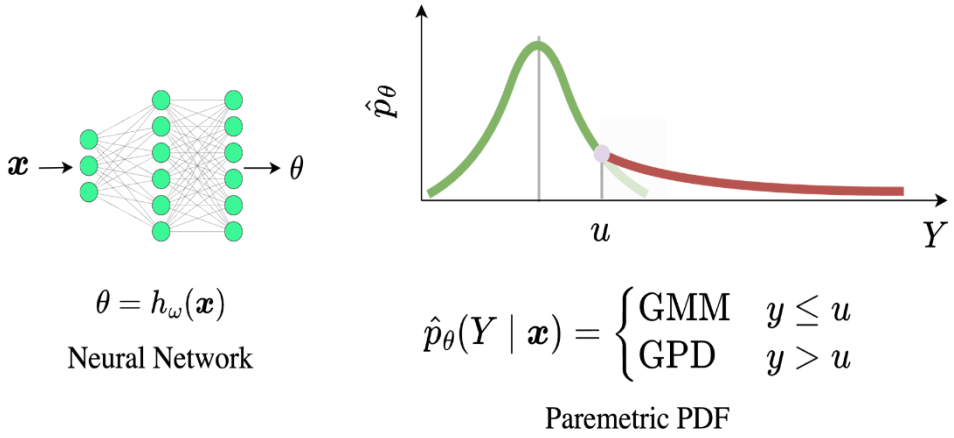


Figure 28 Block diagram showing mapping between the input (conditions) and the output (PDF) through a neural network.

Where $f(y|\varphi)$ and $F(u|\varphi)$ denote the probability density function (PDF) and cumulative density function (CDF) of the Gaussian distribution, respectively; $g(y|\beta, \xi, u)$ denotes the PDF of the GPD distribution where $u$ is the tail threshold, $\xi$ is the tail index, and $\beta$ denotes the tail scale. For simplicity, we define θ as the collection of all parameters $\varphi$, $\beta$, $\xi$, and $u$. It is worth pointing out that the parameters of the two distributions are obtained by the neural network shown in Figure 28.

### 4.2.2 Application in Queueing Systems

The above approach to solve the Conditional Density Estimation (CDE) problem could be applied to various systems to obtain an accurate estimate of the latency PDF. Next, we describe how this approach can be utilized to characterize latency in a multi-hop queuing-theoretic system. Figure 29 illustrates a three-hop tandem queuing system, each queue having a single server and an infinite buffer. This queuing model can be used to represent a closed-loop control application [MDG21]. Here, the tasks generated at the end node are queued for transmission to the compute node. After arriving at the compute node, they are queued again until the compute service becomes available to process them. The computed responses are similarly queued before being sent back to the end node. The CDE problem in this system concerns estimating the likelihood that a task's end-to-end delay will surpass a certain threshold, considering both the task's progress through the network and the overall state of the queues (e.g., number of backlogs) at a given time.

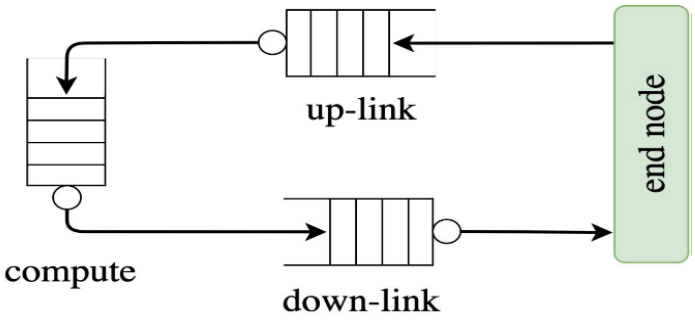


Figure 29. An illustration of the queuing network model for closed-loop control applications.

The tail probability distributions of the three-hop queueing system estimated using (i) the GMM predictor and (ii) the EMM predictor along with empirical (test and training) data are shown in Figure 30. The empirical data is generated by simulating a three-hop queuing system in MATLAB. In the simulation, tasks periodically arrive at the first hop at a rate of $\lambda = 0.9$. The service time distribution for all three queues was identical, characterized by service rates $\mu = 1$. It can also be observed that the error in the GMM predictor begins to increase when there is an absence of training data (where the tail probability is below $10^{-2}$). Consequently, GMM predictions exhibit an exponential decline, as anticipated. In contrast, EMM predictions closely track the actual data until reaching $10^{-4}$ thus highlighting EMM's capability to accurately estimate tail estimation with limited training data.

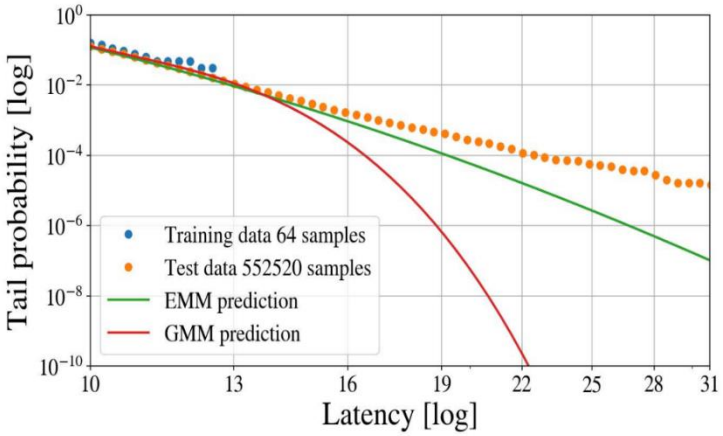


Figure 30. Sojourn time tail probability estimation of the tail probability through a three-hop.

### 4.2.3 Active Queue Management

Active Queue Management (AQM) is a class of approaches that are used to address network congestion by dropping packets from network queues to improve the overall end-to-end delays. Data-driven approaches for latency prediction offer the potential to be harnessed for improving AQM. The idea is to incorporate PDF predictions in AQM algorithms that are otherwise configured with a predefined threshold independent of the Delay Violation Probability (DVP) of packets. The objective of such an approach is to obtain a policy that maximizes the expectation of successful packets (i.e., packets processed before deadlines) by deciding about dropping some of the packets of the stream. *Delta* is a DVP-aware AQM scheme that utilizes data-driven latency predictions to decide when to drop packets as illustrated in the example shown in Figure 31. Using the DVPs for the queued packets, which

are obtained through the estimated PDFs, a packet drops decision vector is obtained that maximizes the successful packets.

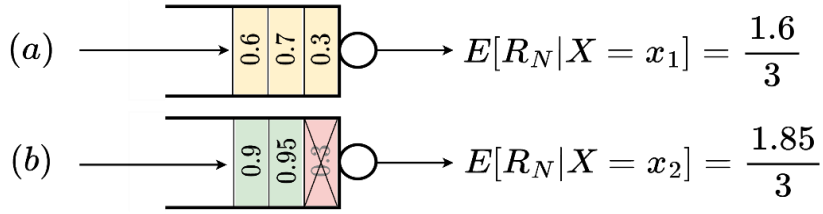


Figure 31. An example on calculating the aggregate packet success for the dropping vectors x1 and x2 which consider if packets in the queue should be dropped.

The simulation reveals that the performance of Delta is superior to no-AQM as well as state-of-the-art AQM schemes such as CoDel and DeepQ, as shown in Figure 32.

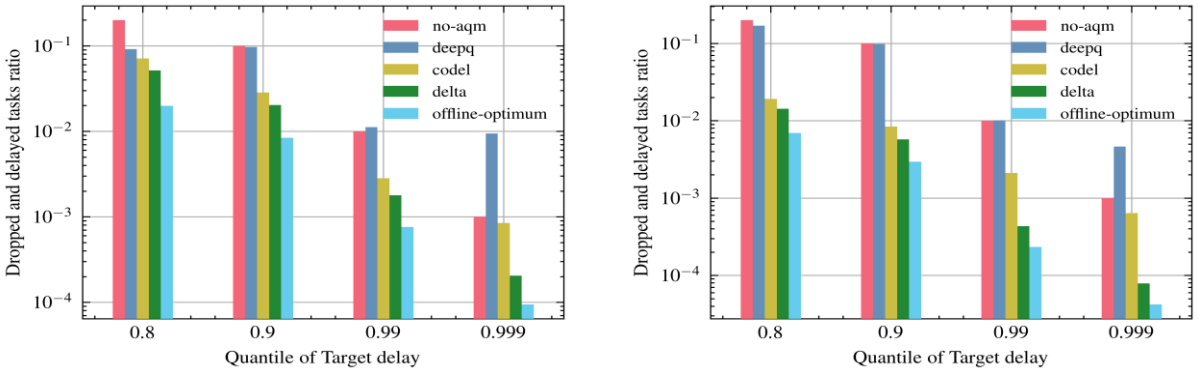


Figure 32. Comparison of the state-of-art and the proposed AQM schemes at utilization of (a) 91.6% and (b) 96.7%.

### 4.2.4 Wireless Networks

In the previous sections, we discussed how theoretical systems (e.g., multihop queueing systems) can be characterized and managed using data-driven approaches. In particular, accurate latency PDFs can be obtained by exploiting EMM-based predictors. Furthermore, this understanding allows us to optimize network operations to improve overall network performance.

However, the effectiveness and accuracy of these approaches needs to be validated on real-world scenarios, particularly in 5G/5G-Adv networks is essential.

Based on this characterization in the bulk as well as the tail portion, more informed decisions about wireless transmissions can be made to fulfill extreme reliability targets (>99.999%). Wireless networks, in contrast to queueing theoretic systems where latency was conditioned on only the number of backlogs in the queue and the time of in service of the current task, consist of complex dependencies between the latency PDF and network / traffic conditions. Therefore, characterizing latency in 5G/5G-Adv networks could be challenging. Especially the tail portions could be subjected to extremely rare outliers caused by specific combinations of network / traffic conditions.

In the CDE problem for wireless networks, the objective is to estimate the PDF of the latency of packets traversing a wireless network. The comprehensive analysis of 5G latency presented in section 2 revealed that 5G latency is subject to the inherent stochastic influences. For instance, the retransmissions arising from HARQ have a significant impact on the overall latency experienced by packets. The number of re-transmissions itself are dependent on various conditions such SINR and the Modulation Coding Scheme (MCS) index. From now onwards, the set of all conditions in the system (both network and traffic) are represented as the variable $X$. To capture this phenomenon, we introduce a random variable $Y$ defined within the domain $Y \subset \mathbb{R}_+$ to represent the latency of packets. Its probability density can be expressed based on the observed transmission conditions, denoted as $x$, leading to the conditional probability $\mathrm{P}[\mathrm{Y}|\mathrm{X} = \mathrm{x}]$.

#### 4.2.4.1 Methodology

The proposed data-driven approaches are validated by predicting latencies for real 5G network latency ($Y$) across varying network conditions ($X$). To this end, latency samples are collected from the wireless network, resulting in a dataset for training and evaluating the MDN-based proposed predictors, i.e., Gaussian Mixture Model (GMM) and Gaussian Mixture Extreme Value Model (GMEVM). Subsequently, the accuracy of these predictors is evaluated, particularly in predicting the tail latency distribution.

Measurement Setup: The experimental setup involves two nodes connected through a 5G network as illustrated in Figure 33. Latency samples are gathered from a 10 ms, 172B periodic transmission of timestamped packets between nodes, both uplink and downlink. To ensure synchronization between the nodes on different machines, their clocks are synchronized using the PTP protocol on a dedicated interface. For the measurements on the 5G network, a commercial off-the-shelf (COTS) 5G setup and an OpenAirInterface (OAI) based software-defined radio (SDR) 5G setup was established.

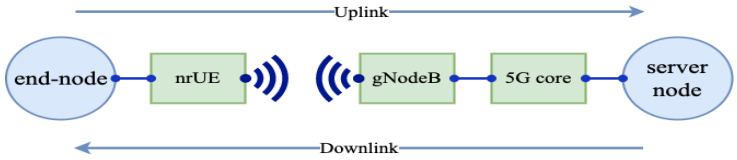


Figure 33. Illustration for the setup used to gather measurements data from both COTS5G and OAI 5G.

Network Measurements: Both 5G networks operate in band 78, using 106 physical resource blocks (PRBs) over 40 MHz bandwidth at 3.5 GHz. In the COTS 5G experiment, we collected 4 million uplink latency measurements from 20 different locations in an indoor environment. The OAI 5G experiment gathered 5 million uplink latency measurements under controlled conditions.

Data Preprocessing: Data normalization and standardization are essential for accurate training. The measured latency values are first scaled to milliseconds and then centered around zero by subtracting the average latency. The values of the conditions are normalized between 0 and 1. For improving the fit accuracy of the predictors, noise regularization is also applied to MDNs during the training phase. The different noise levels of 1 ms and 3 ms are compared against the raw data.

Training Phase: The MDN model is constructed with four hidden layers that are fully connected, featuring neuron counts of [10, 100, 100, 80], respectively. It governs parameters for 15 Gaussian centers and potentially incorporates GPD for tail modeling. The output size of the neural network

corresponds to the parameters ($\theta$) of the parametric density function, totaling 48 with GPD and 45 without. When the GPD component is integrated, the model is referred to as the GMEVM. The training process employs the Adam optimizer, spanning four rounds, each comprising 200 epochs and distinct learning rates of $[10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$, respectively. To maintain a suitable balance, the batch size is set at 1/8 of the training dataset size. Notably, including training rounds with minimal learning rates, such as $10^{-4}$ or $10^{-5}$, is found to prevent underfitting in the tail region. Table 1 lists the time required for training over different dataset sizes. The MDN models are implemented and evaluated using CPU-only TensorFlow, running on an Intel Core i9-10980XE CPU clocked at 3.00GHz.

Table 1. Training parameters for 1000 epochs.

| Training dataset size | 1M | 256k | 64k |
|---|---|---|---|
| Batch size | 128k | 32k | 8k |
| Epoch duration | 4 s | 1 s | 300 ms |
| Step duration | 500 ms | 100 ms | 27 ms |
| Total training time | 66.6 m | 16.6 m | 5 m |

#### 4.2.4.2 Evaluation Results

As illustrated in Figure 34, the predicted tail probability for COTS 5G latency is compared among different predictor models and varying training sample sizes. It can be observed that the tail latency distribution from the GMM predictor exponentially deviates from the empirical distribution with increasing delay, whereas the GMEVM predictor closely follows the empirical distribution. Next, we present the results for predictors performance in the context of conditional scenarios and a more challenging latency profile.

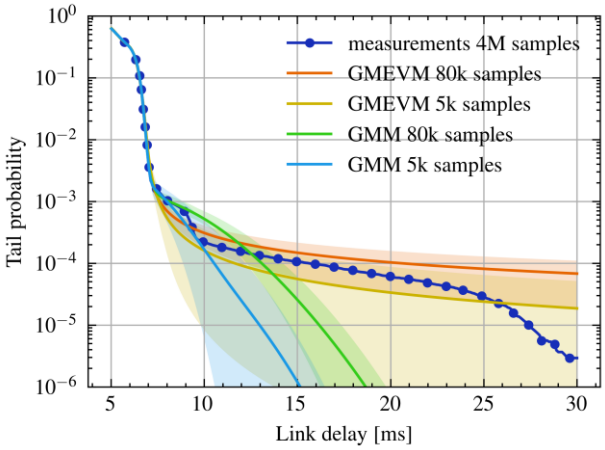


Figure 34. COTS 5G uplink latency measurements vs parametric density fits with different number of samples and models.

Figure 35 shows the tail probability distribution measured from OAI 5G for different values of MCS indices using all 5M samples (1.6 million per MCS index). The two predictors were trained with 1M total training samples covering all three MCS conditions. As expected, increasing the MCS index resulting in decreasing latency arising from the increased link capacity. It can be observed from the figure that the GMEVM scheme exhibits high variance and poor prediction accuracy across all three cases, while GMM performs better. This can be attributed to the measurements' bumpy and non-smooth tail profile.

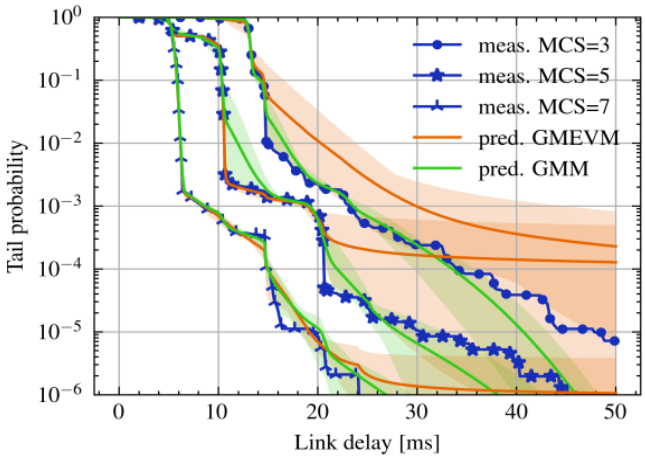


Figure 35. Performance of MDN models trained without noise regularization. Number of samples: 1M (20%).

This issue can be addressed by incorporating noise regularization in the training phase of the predictors by introducing random Gaussian noise with a variance of 1 ms in latency samples. Figure 36 demonstrates that this approach considerably enhances GMEVM's performance, even with reduced training samples, by smoothing out the tail. However, it is important to note that noise regularization also leads to reduced accuracy in the bulk of the latency distribution. Therefore, careful selection of the noise's variance is vital to make predictions in the bulk as well as the tail portion of the PDFs. Furthermore, higher uncertainty and error from the models in the case of a smaller number of training samples can be observed.

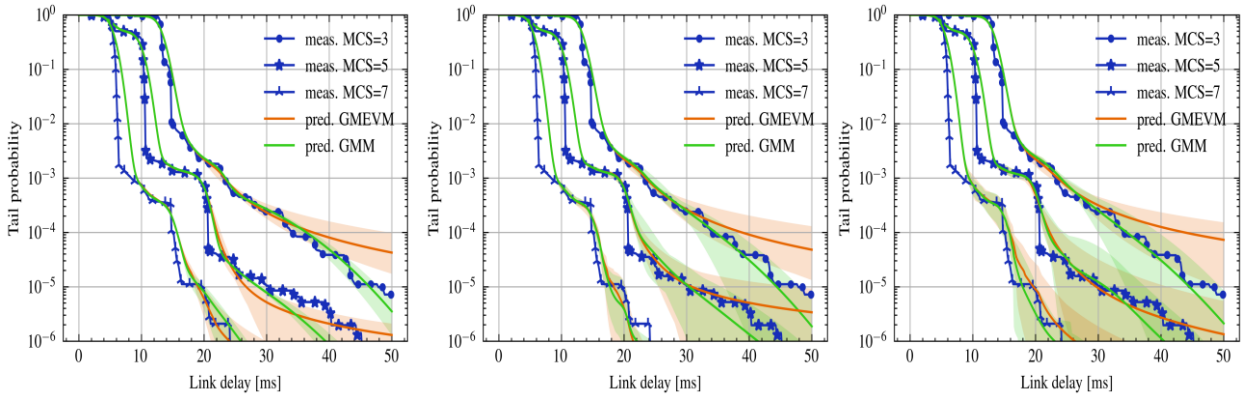


Figure 36. OAI 5G uplink latency tail probability measurements with added Gaussian noise conditioned for different MCS indices with 5M samples vs predictions with (a) 1M, (b) 250k and (c) 64k samples.

In Figure 37, we present the generalization capability of the proposed predictors. To this end, models were trained on a dataset excluding latency samples with MCS=5, making MCS=5 an unseen condition for the predictors. Both GMM and GMEVM follow the empirical tail but with noticeable variance. However, GMM outperforms GMEVM with lower variance. This demonstrates a prospect for improving predictors' ability to handle previously unseen conditions in future research endeavors.

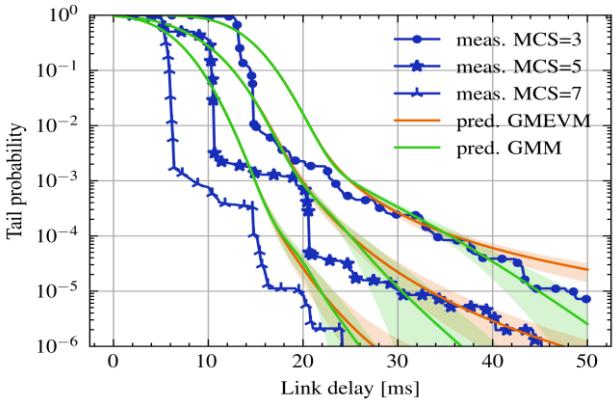


Figure 37. Generalization capability of MDN models trained without MCS=5 samples. Number of samples: 0.6M (20%).

## 4.3 Latency Prediction Architectures

The presented framework lays the foundation for integrating data-driven (AI/ML) approaches for latency prediction within the 6G architecture. The framework shown in Figure 38 is adapted from the one described in TR 37.817 for AI-enabled RAN Intelligence [3GPP22-37817]. First, the Data Collection function serves as the initial step, responsible for gathering relevant data corresponding to the traffic flowing in both uplink (UL) and downlink (DL) directions from different network entities. The data collection function should configure these entities to timestamp packets at different layers of the 5G protocol stack as well as capture the corresponding network (e.g., CSI) and traffic (background traffic, packet rate). A crucial assumption in data collection requires relevant entities within the 5G system to be sufficiently synchronized (i.e., sub-microsecond synchronization accuracy) [DET23-D22]. Time synchronization in these entities ensures accurate latency calculations from the timestamps as well as correlation of events and conditions in the network to the resulting latencies.

The training of latency prediction models can be carried out at distinct locations within the 5G system, such as the 5G core or the gNB. The task here is to train, validate and rigorously test latency prediction models using the training data. Training datasets typically consist of a substantial number of samples, ranging from 10k to 100k samples collected over time spans of tens of minutes to several hours. Each sample in the dataset consists of the system state, i.e., network conditions (Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ), 5G QoS Identifier (5QI), etc.) and traffic conditions (packet length, interval, etc.) along with corresponding latency. These datasets facilitate a rich representation of network and traffic conditions and corresponding latencies. Two fundamental approaches to training latency prediction models exist: online and offline. Offline training would involve deploying latency prediction models only after completing the training, validation, and testing procedures. Conversely, online training allows for immediate deployment and continuous training of models to converge towards sufficient performance. The option of utilizing pre-trained models or employing continual and/or transfer learning schemes further enhances adaptability in the training process, enriching the predictive capabilities of the models with faster training time.

Once the latency prediction models are trained, the model Inference function, primarily located within the gNodeB, can take control. This function leverages the trained models, current network conditions, and past latency values (especially in the case of recurrent predictors) to estimate the latency distribution in the future. A crucial component of model inference is the monitoring of model

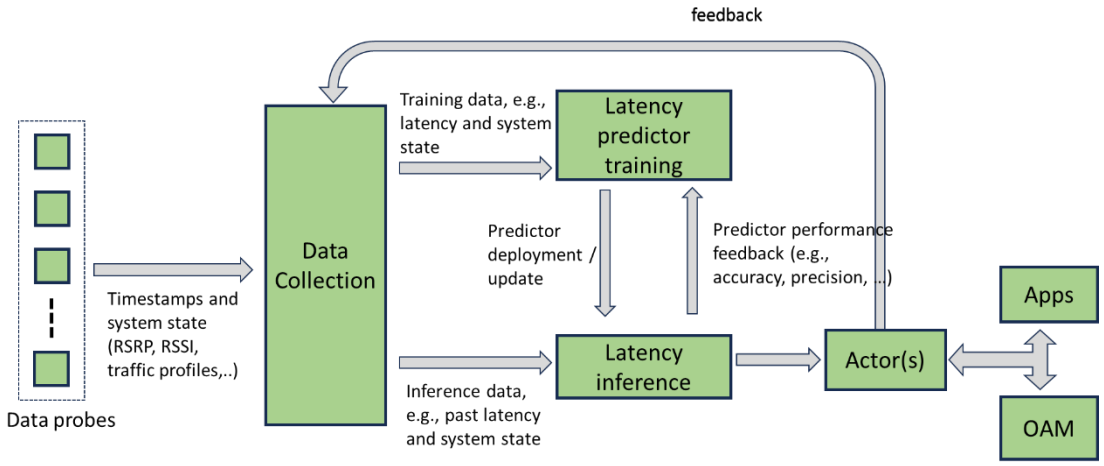


Figure 38 Architectural aspects of data-driven latency prediction.

performance. If discrepancies between latency predictions and ground truth are identified, retraining should be triggered to maintain the reliability of latency predictions.

Lastly, the actor function utilizes the latency predictions from model inference to achieve specific objectives, such as optimizing resource allocation within the network at the same time fulfilling the application requirements. The actor function also provides valuable feedback to the data collection function, ensuring that the data gathering strategies are refined and optimized, ultimately enhancing the quality and relevance of the training dataset.

# 5 RAN Resource Management

5G transmission latency analysis presented in section 2 revealed the importance of resource allocation in RAN in managing latency and reliability, in particular concerning HARQ retransmissions. In this section, we will delve into the resource allocation schemes, exploring various techniques that have been at the forefront of this field. Subsequently, we will review the resource allocation considering the repetition schemes, addressing the issues and challenges that arise in this evolving landscape in the context of 5G to 6G. The objective is to provide an overview of the current state of the art, shedding light on strategies and methodologies employed in resource allocation. Following this exploration, we will transition into a problem formulation, dissecting the challenges posed by resource allocation in 6G. Through this review, we aim to gain an understanding of the dynamics and efficient distribution of resources in the next generation of 6G.

## 5.1 State-of-the-art in Resource Allocation

The development of time-critical communications in 6G has spurred significant interest in developing resource allocation techniques that can meet the stringent requirements of URLLC applications. In this section, we present a review of the state-of-the-art resource allocation techniques for dependable time-critical 6G networks. We highlight advancements and cutting-edge approaches that aim to optimize resource utilization and ensure reliable and predictable communication services in 6G networks. We categorize the resource allocation techniques in dependable time-critical 6G into the following domains. In the time domain, several novel resource allocation approaches such as introducing a dynamic time slot assignment algorithm based on traffic patterns and ensuring real-time allocation for URLLC services are proposed. The algorithm uses a queuing model to predict the traffic demand for each URLLC flow, enabling timely resource allocation. Another approach considers the

time-varying channel conditions and user mobility to dynamically allocate transmission resources, effectively reducing packet latency and improving overall system performance [MHL+23]. Frequency domain techniques have received significant attention for improving spectrum utilization in dependable time-critical networks. [MGR+23] presents sub-band full duplex (SBFD) as a duplexing scheme to improve the uplink (UL) throughput in 5G–Advanced networks, as an alternative to traditional TDD. SBFD provides opportunities to transmit and receive simultaneously on non-overlapping frequency resources. To accomplish this, SBFD time slots include both UL and downlink (DL) transmission. Concurrent UL and DL transmission create different types of interference, which makes cancellation approaches essential for appropriate performance. One of the other aspects is proposing an intelligent subcarrier allocation scheme using machine learning algorithms to optimize spectral efficiency while maintaining stringent guarantees. The other approach leverages deep learning models to predict channel conditions and allocate subcarriers dynamically, resulting in enhanced system capacity. Moreover, exploring the spectrum-sharing strategies to enable efficient coexistence between URLLC and enhanced mobile broadband (eMBB) services. Spatial resource allocation techniques play a vital role in achieving reliable and low-latency communication. One approach is developing a novel beamforming approach that leverages massive-MIMO technology to enhance spatial resource utilization and reduce interference. The proposed technique employs the concept of dynamic beam steering to track URLLC users and adapt beamforming vectors, accordingly, resulting in improved reliability and reduced signal fluctuations. Moreover, one can propose an antenna selection algorithm that optimizes the spatial diversity gain for URLLC applications [YGA+20]. By selecting the most appropriate subset of antennas, the algorithm ensures reliable and low-latency communication for URLLC users, even in highly dynamic and interference-prone environments.

The performance of resource allocation techniques is evaluated using various metrics, including throughput, latency, reliability, spectral efficiency, fairness, and energy efficiency. Time domain approaches are particularly effective for meeting strict latency requirements, while frequency domain techniques can improve spectrum efficiency. Spatial domain methods offer spatial diversity gain. Despite the progress made in resource allocation, challenges remain, such as scalability for massive IoT deployments, effective interference management, and real-time optimization. By optimizing resource utilization and ensuring reliable communication, the identified techniques pave the way for achieving the ambitious goals of 6G technology, enabling transformative applications in diverse sectors.

## 5.2 Problem Formulation

In the dynamic aspect of cellular networks, technological progress is a constant force driving evolution. 3GPP releases represent the principles of current delay reduction advancements in wireless networks, introducing transformative enhancements across various domains. However, the deployment of these innovations poses several challenges and warrants a thorough problem formulation that needs to be addressed. First of all, how can the reduced latency advancements introduced in 3GPP releases be integrated into the vision of dependable time-critical 6G networks, where ultra-reliable, low-latency communication is a fundamental requirement? What role do these advancements play in ensuring predictability? While the emphasis on retransmissions methods is critical for enhancing reliability and reducing latency, there is a need to formulate solutions for streamlining the repetition and retransmission processes. This includes reducing retransmission delays, minimizing resource utilization and optimizing the retransmission strategy for different types of applications. Furthermore, managing the trade-off between reliability and latency for various use cases poses a significant

problem. While placing a strong emphasis on enhancing spectral efficiency through flexible scheduling, the problem at hand revolves around devising scheduling algorithms and network management strategies that can dynamically allocate resources to accommodate varying traffic demands remains critical. This includes the challenge of balancing the trade-off between spectral efficiency and fairness among users while ensuring efficient utilization of the available frequency spectrum. An overarching challenge lies in optimizing network performance while balancing the quality of service provided to users with efficient resource utilization. This involves trade-offs between reliability, latency, data rates and network capacity, necessitating the development of adaptive and intelligent network management strategies.

In addressing above mentioned challenges, we aim to unlock the full potential of optimizing network performance and fostering an enhanced user experience across various sectors and applications. One of these challenging domains is HARQ retransmissions in combination with the mentioned resource allocation method which stands out as a key challenging component. Figure 39 presents a simple illustration of the k-repetitions HARQ process that takes place in the time interval $T_{rep}$. When a packet enters MAC layer, the first delay starts. The packet will be assigned to a transmission channel (like HARQ channel). The waiting time in the queue to allocate the available resource blocks and the preparation time for encoding and modulation are modeled by $T_1$. It is assumed that the resources are available in every TTI, where one TTI duration corresponds to a number of consecutive symbols for one transmission. The next latency block is the initial transmission block, which, based on packet size, channel quality, and scheduling strategy, the transmission time can vary from one to multiple TTIs which is modeled by $T_2$, which includes the packet transmission time in the UE, the processing time in the gNB and the repetition time duration in every transmission. If the packet is successfully decoded, gNB sends ACK feedback, otherwise, it sends a NACK, where the ACK/NACK feedback time and the processing time in the UE are modeled by $T_3$.

Table 2 Latency components.

| Parameter | Value |
|---|---|
| $T_{symbol}$ | configurable |
| $T_1$ | Zero to one TTI |
| $T_2$ | Zero to multiple TTI |
| $T_3$ | Zero to one TTI |
| Scheduling priority | Every TTI |

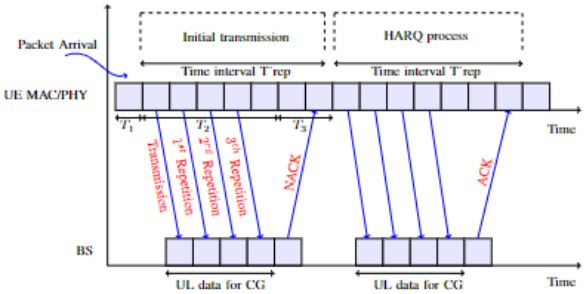


Figure 39 k-repetition scheme with k=4.

We will assume that every TTI duration is equal to 4 symbols and the duration of the symbol is based on the TDD configuration in the system. The relevant latency components of this model are listed in Table 2. A fixed number of UEs are deployed in cells and are assumed to be synchronized and connected. As an example, for the XR traffic models, the packet arrival rate is determined by the frame generation rate, e.g., $60fps$ [3GPP17-38838]. Accordingly, the average packet arrival periodicity is given by the inverse of the frame rate,

e.g., $16.6667ms = 160fps$. The periodic arrival with jitter gives the arrival time at gNB for a packet with index $k'$ $(1,\ 2,\ 3,\ \ldots)$

$$offset + \frac{k'}{F} \cdot 1000\ + J\ [ms]$$

where F is the given frame generation rate (per second) and J is a random variable capturing jitter. Note that the actual traffic arrival timing for each UE could be shifted by the UE specific arbitrary offset. Packets are generated by each UE according to the traffic model in the above expression. We formulate the SINR at the BS as:

$$SINR_k = \frac{g_k h_0}{I_{intra} + \sigma^2}$$

, where $g_k$ denotes $k_{th}$ repetition, $h_0$ is the power gain of the channel, which we assume to be exponentially distributed as is the case for Rayleigh fading channels, $\sigma^2$ is the noise power and $I_{intra}$ is the interference caused by the UEs in the same cell [NYA+18]. We consider intra-cell interference because the UEs in the same cell associated with the same BS may cause interference with each other. We derive the probability $P_{st,k}(n)$ of successful transmission at the $k_{th}$ repetition in the case of $n$ interfering UEs in the same cell considering the channel conditions. We assume successful transmission if the SINR is above a certain threshold. Successful reception of packets depends on the MCS used and the post-processing SINR achieved. If reception fails, the UE issues a HARQ retransmission after processing the feedback from the BS. We derive the success probability under the k-repetition scheme conditioned on the $n$ number of intra-cell interfering following the steps in [YYM+21]. Therefore, we need to propose a system model to evaluate the network performance such as packet arrival rate and number of required repetitions, considering the state of the communication channel and the available time budget. By combining a channel model, we establish a framework to control the performance of the network under varying conditions, such as the number of active UEs, and numerology of the network, the packet size and more. Given that the traffic models prevalent in the majority of 6G applications lack periodicity, formulating a framework capable of accommodating

diverse applications with distinct traffic patterns and varied QoS requirements becomes a formidable challenge. Addressing this intricate challenge is the focal point of our endeavor in this study.

# 6 Conclusions & Future Work

As we look ahead to 6G and its role in supporting emerging application domains, it is expected that dependable time-critical communications are going to play a significant role. This document delved into the key 6G-centric enablers that will be vital for realizing dependable time-critical communications in future communication and computer systems.

We provided a detailed analysis of transmission latency in the context of current 5G / 5G-Adv / URLLC systems. This analysis lays the foundation for understanding the challenges with respect to fulfilling the requirements that arise in ensuring dependable time-critical communications.

One of the key issues in realizing dependable time-critical communications is controlling the PDVs of wireless communication links to enable smooth and efficient interworking between wired (e.g., TSN) and wireless communication technologies (e.g., 5G-Adv/6G). To this end, three PDC-based solutions were presented that target controlling PDV to tens of microseconds with no cost regarding spectral efficiency. In the future, the proposed PDC schemes will be evaluated via simulations in the context of end-to-end dependable time-critical communications.

Despite time-critical communications functionalities, wireless systems are today subject to spatial and temporal stochastic variations that need to be accurately characterized dynamically. To this end, data-driven latency characterization approaches have been proposed and evaluated. The proposed latency prediction approaches did not account for any temporal dependency. As a future work, advanced latency prediction methods utilizing recurrent neural networks will be investigated to predict such time-dependencies. Additionally, we will delve further into the 6G architectural aspects of data-driven latency prediction schemes and investigate temporal aspects in the latency forecasting problem.

Lastly, the problem of RAN resource allocation modeling is presented. Modeling of traffic shapes, considering repetition and re-transmission schemes, has led to important takeaways such as adaptive repetition schemes based on the available resources and a traffic shape-based scheduling methodology. As a future work, the presented model will focus on 6G applications such as XR traffic shape, which are challenging due to the aperiodic behavior of the traffic shapes.

By addressing these areas, we aim to further advance the understanding and implementation of k-repetitions for data transmission reliability. This will contribute to more robust and dependable time-critical communications systems in a wide range of XR applications.

Collectively, technological enablers outlined in this document provided solutions for realizing dependable time-critical communications in 6G.

## References


| | |
|---|---|
| [3GPP16-2002055] | 3GPP S2-2002055, "Clarifying hold and forward buffering mechanism," change request for TS 23.501, 24-27 February, 2020, Electronic Meeting.v16.3.0 |
| [3GPP16-2002056] | 3GPP S2-2002056, "Clarifying hold and forward buffering mechanism," discussion paper, 24-27 February 2020, Electronic Meeting. |
| [3GPP16-22261] | 3GPP TS 22.261, "Service requirements for the 5G system," v19.4.0 |
| [3GPP17-23501] | 3GPP TS 23.501, "System architecture for the 5G System (5GS)," v15.0.0 |
| [3GPP17-38838] | 3GPP TR 38.838, "Study on XR (Extended Reality) Evaluations for NR," v17.0.0 |
| [3GPP22-37817] | 3GPP TR 37.817 "Study on enhancement for data collection for NR and ENDC," v17.0.0 |
| [3GPP22-37910] | 3GPP TR 37.910 "Study on self evaluation towards IMT-2020 submission", V17.0.0 |
| [3GPP22-37910] | 3GPP TR 37.910 "Study on self evaluation towards IMT-2020 submission", V17.0.0 |
| [5GA-QOS21] | 5GACIA, "5G QoS for Industrial Automation," 5GACIA Whitepaper, Nov. 2021, https://5g-acia.org/whitepapers/5g-quality-of-service-for-industrial-automation/ |
| [5GS21-D14] | 5G-SMART deliverable D1.4, "Radio network deployment options for smart manufacturing", Nov. 2020, https://5gsmart.eu/deliverables/ |
| [5GS21-D14] | 5G-SMART deliverable D1.4, "Radio network deployment options for smart manufacturing", Nov. 2020, https://5gsmart.eu/deliverables/ |
| [5GS21-D15] | 5G-SMART deliverable D1.5, "Evaluation of radio network deployment options", Dec. 2021, https://5gsmart.eu/deliverables/ |
| [5GS21-D15] | 5G-SMART deliverable D1.5, "Evaluation of radio network deployment options", Dec. 2021, https://5gsmart.eu/deliverables/ |
| [5GS21-D54] | 5G-SMART deliverable D5.4, "Second report on 5G network architecture options and assessments", Nov. 2021, https://5gsmart.eu/deliverables/ |
| [AAB+22] | J. Ansari, C. Andersson, P. de Bruin, J. Farkas, L- Grosjean, J. Sachs, J. Torsner, B. Varga, D. Harutyunyan, N. König, R. H. Schmitt, "Performance of 5G Trials for Industrial Automation. Electronics", 2022; 11(3):412. https://doi.org/10.3390/electronics11030412 |
| [AAB+22] | J. Ansari, C. Andersson, P. de Bruin, J. Farkas, L- Grosjean, J. Sachs, J. Torsner, B. Varga, D. Harutyunyan, N. König, R. H. Schmitt, "Performance of 5G Trials for Industrial Automation. Electronics", 2022; 11(3):412. https://doi.org/10.3390/electronics11030412 |

| | |
|---|---|
| [ABS+20] | F. Alriksson, L. Boström, J. Sachs, Y. . -P. E. Wang and A. Zaidi, "Critical IoT connectivity Ideal for Time-Critical Communications," in *Ericsson Technology Review*, vol. 2020, no. 6, pp. 2-13, June 2020, doi: 10.23919/ETR.2020.9905508 |
| [ABS+20] | F. Alriksson, L. Boström, J. Sachs, Y. . -P. E. Wang and A. Zaidi, "Critical IoT connectivity Ideal for Time-Critical Communications," in *Ericsson Technology Review*, vol. 2020, no. 6, pp. 2-13, June 2020, doi: 10.23919/ETR.2020.9905508 |
| [ALR+04] | A. Avizienis, J. Laprie, B. Randell and C. Landwehr, "Basic Concepts and Taxonomy of Dependable and Secure Computing," in *IEEE Transactions on Dependable and Secure Compututing*, January 2004, doi: https://doi.org/10.1109/TDSC.2004.2 |
| [ATP19] | M. Angjelichinoski, K. F. Trillingsgaard and P. Popovski, "A Statistical Learning Approach to Ultra-Reliable Low Latency Communication," in IEEE Transactions on Communications, vol. 67, no. 7, pp. 5153-5166, July 2019, doi: 10.1109/TCOMM.2019.2907241 |
| [AVK+22] | J. Ansari, B. Varga, P. Kehl, N. König, R. H. Schmitt "5G and TSN integrated prototype for flexible production", presentation at TSN/A conference, Stuttgart, September 28-29, 2022 |
| [AVK+22] | J. Ansari, B. Varga, P. Kehl, N. König, R. H. Schmitt "5G and TSN integrated prototype for flexible production", presentation at TSN/A conference, Stuttgart, September 28-29, 2022 |
| [BDP18] | M. Bennis, M. Debbah and H. V. Poor, "Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," in Proceedings of the IEEE, vol. 106, no. 10, pp. 1834-1853, Oct. 2018, doi: 10.1109/JPROC.2018.2867029 |
| [BIA00] | G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," in IEEE Journal on Selected Areas in Communications, vol. 18, no. 3, pp. 535-547, March 2000, doi: 10.1109/49.840210 |
| [BIS94] | C. Bishop, "Mixture density networks," Aston University, 1994. |
| [BSB+19] | D. Bruckner, M.-P. Stănică, R. Blair, S. Schriegel, S. Kehrer, M. Seewald, T. Sauter, "An Introduction to OPC UA TSN for Industrial Communication Systems," in *Proceedings of the IEEE*, vol. 107, no. 6, pp. 1121-1131, June 2019, doi: 10.1109/JPROC.2018.2888703 |
| [CAG20] | J. P. Champati, H. Al-Zubaidy and J. Gross, "Transient Analysis for Multihop Wireless Networks Under Static Routing," in IEEE/ACM Transactions on Networking, vol. 28, no. 2, pp. 722-735, April 2020, doi: 10.1109/TNET.2020.2975616 |
| [CAS+22] | J.B. Caro, J. Ansari, J. Sachs, P. de Bruin, S. Sivri, L. Grosjean, N. König, R.H. Schmitt, "Empirical Study on 5G NR Cochannel Coexistence". Electronics 2022, 11, 1676. https://doi.org/10.3390/electronics11111676 |
| [CAS+22] | J.B. Caro, J. Ansari, J. Sachs, P. de Bruin, S. Sivri, L. Grosjean, N. König, R.H. Schmitt, "Empirical Study on 5G NR Cochannel |

| | |
|---|---|
| | Coexistence". Electronics 2022, 11, 1676. https://doi.org/10.3390/electronics11111676 |
| [CAS+23] | J. B. Caro, J. Ansari, A.R. Sayyed, P. de Bruin, J. Sachs, N. König, R.H. Schmitt, "Empirical study on 5G NR Adjacent Channel Coexistence," 2023 IEEE Wireless Communications and Networking Conference (WCNC), Glasgow, United Kingdom, 2023, pp. 1-6, doi: 10.1109/WCNC55385.2023.10119074 |
| [CAS+23] | J. B. Caro, J. Ansari, A.R. Sayyed, P. de Bruin, J. Sachs, N. König, R.H. Schmitt, "Empirical study on 5G NR Adjacent Channel Coexistence," 2023 IEEE Wireless Communications and Networking Conference (WCNC), Glasgow, United Kingdom, 2023, pp. 1-6, doi: 10.1109/WCNC55385.2023.10119074 |
| [DET23-D11] | DETERMINISTIC6G, Deliverable 1.1, "DETERMINISTIC6G use cases and architecture principles," Jun. 2023, https://deterministic6g.eu/index.php/library-m/deliverables |
| [DET23-D22] | DETERMINISTIC6G, Deliverable 2.2, "First Report on the time synchronization for E2E time awareness," Dec. 2023, https://deterministic6g.eu/index.php/library-m/deliverables |
| [DET23-D31] | DETERMINISTIC6G, Deliverable 3.1, "Report on 6G convergence enablers towards deterministic communication standards," Dec. 2023, https://deterministic6g.eu/index.php/library-m/deliverables |
| [DET23-D41] | DETERMINISTIC6G, Deliverable 4.1, "Digest on First DetCom Simulator Framework Release," Dec. 2023, https://deterministic6g.eu/index.php/library-m/deliverables |
| [DPS21] | E. Dahlman, S. Parkvall and J. Skold, "Chapter 6 - Radio-Interface Architecture," 5G NR (Second Edition), 2021 |
| [DT07] | T. DelSole and M.K. Tippett, "Predictability: Recent insights from information theory,"in Reviews of Geophysics, 45(4) |
| [ECC19-296] | CEPT ECC, ECC Report 296, "National synchronization regulatory framework options in 3400-3800 MHz: a toolbox for coexistence of MFCNs in synchronised, unsynchronised and semi-synchronised operation in 3400-3800 MHz," March 2019 |
| [ECC19-296] | CEPT ECC, ECC Report 296, "National synchronization regulatory framework options in 3400-3800 MHz: a toolbox for coexistence of MFCNs in synchronised, unsynchronised and semi-synchronised operation in 3400-3800 MHz," March 2019 |
| [FYJ20] | C. Flinta, W. Yan and A. Johnsson, "Predicting Round-Trip Time Distributions in IoT Systems using Histogram Estimators," NOMS 2020 - 2020 IEEE/IFIP Network Operations and Management Symposium, Budapest, Hungary, 2020, pp. 1-9, doi: 10.1109/NOMS47738.2020.9110315 |
| [ITU19-1540] | ITU Recommendation Y.1540, "Internet protocol data communication service - IP packet transfer and availability performance parameters" |
| [JWE+15] | N. A. Johansson, Y. . -P. E. Wang, E. Eriksson and M. Hessler, "Radio access for ultra-reliable and low-latency 5G communications," 2015 IEEE International Conference on |

| | |
|---|---|
| | Communication Workshop (ICCW), London, UK, 2015, pp. 1184-1189, doi: 10.1109/ICCW.2015.7247338 |
| [JWE+15] | N. A. Johansson, Y. . -P. E. Wang, E. Eriksson and M. Hessler, "Radio access for ultra-reliable and low-latency 5G communications," 2015 IEEE International Conference on Communication Workshop (ICCW), London, UK, 2015, pp. 1184-1189, doi: 10.1109/ICCW.2015.7247338 |
| [KAJ+22] | P. Kehl, J. Ansari, M. H. Jafari, P. Becker, J. Sachs, N. König, A. Göppert, R. H. Schmitt, "A Prototype of 5G Integrated with TSN for Edge-Controlled Mobile Robotics" Electronics 11, no. 11: 1666, 2022. https://doi.org/10.3390/electronics11111666 |
| [KAJ+22] | P. Kehl, J. Ansari, M. H. Jafari, P. Becker, J. Sachs, N. König, A. Göppert, R. H. Schmitt, "A Prototype of 5G Integrated with TSN for Edge-Controlled Mobile Robotics" Electronics 11, no. 11: 1666, 2022. https://doi.org/10.3390/electronics11111666 |
| [KFR19] | S. K. Khangura, M. Fidler, and B. Rosenhahn. "Machine learning for measurement-based bandwidth estimation." *Computer communications* 144 (2019): 18-30 |
| [KSB+20] | R. Kundel, F. Siegmund, J. Blendin, A. Rizk and B. Koldehofe, "P4STA: High Performance Packet Timestamping with Programmable Packet Processors," NOMS 2020 - 2020 IEEE/IFIP Network Operations and management Symposium, Budapest, Hungary, 2020, pp. 1-9, doi: 10.1109/NOMS47738.2020.9110290 |
| [LLM+09] | A. Larmo, M. Lindström, M. Meyer, G. Pelletier, J. Torsner and H. Wiemann, "The LTE link-layer design," in IEEE Communications Magazine, vol. 47, no. 4, pp. 52-59, April 2009, doi: 10.1109/MCOM.2009.4907407 |
| [LLM+09] | A. Larmo, M. Lindström, M. Meyer, G. Pelletier, J. Torsner and H. Wiemann, "The LTE link-layer design," in IEEE Communications Magazine, vol. 47, no. 4, pp. 52-59, April 2009, doi: 10.1109/MCOM.2009.490740 |
| [LSW+19] | O. Liberg, M. Sundberg, Y.-P. E. Wang, J. Bergman, J. Sachs, G. Wikström, Cellular Internet of Things - From Massive Deployments to Critical 5G Applications, Academic Press, second edition, ISBN: 9780081029022, October 2019 |
| [LSW+19] | O. Liberg, M. Sundberg, Y.-P. E. Wang, J. Bergman, J. Sachs, G. Wikström, Cellular Internet of Things - From Massive Deployments to Critical 5G Applications, Academic Press, second edition, ISBN: 9780081029022, October 2019 |
| [MDG21] | S. S. Mostafavi, G. Dán and J. Gross, "Data-Driven End-to-End Delay Violation Probability Prediction with Extreme Value Mixture Models," 2021 IEEE/ACM Symposium on Edge Computing (SEC), San Jose, CA, USA, 2021, pp. 416-422, doi: 10.1145/3453142.3493506 |
| [MGR+23] | M. Mokhtari, G. Pocovi, R. Maldonado and K. I. Pedersen, "Modeling and System-Level Performance Evaluation of Sub-Band Full Duplexing for 5G-Advanced," in IEEE Access, vol. 11, pp. 71503-71516, 2023, doi: 10.1109/ACCESS.2023.3295121 |

| | |
|---|---|
| [MGR+23] | M. Mokhtari, G. Pocovi, R. Maldonado and K. I. Pedersen, "Modeling and System-Level Performance Evaluation of Sub-Band Full Duplexing for 5G-Advanced," in IEEE Access, vol. 11, pp. 71503-71516, 2023, doi: 10.1109/ACCESS.2023.3295121 |
| [MHL+23] | M. U. A. Siddiqui, H. Abumarshoud, L. Bariah, S. Muhaidat, M. A. Imran and L. Mohjazi, "URLLC in Beyond 5G and 6G Networks: An Interference Management Perspective," in *IEEE Access*, vol. 11, pp. 54639-54663, 2023, doi: 10.1109/ACCESS.2023.3282363 |
| [MMD+17] | H. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. AISTATS 2017*, April 2017. |
| [MMR+23] | S. Mostafavi, V.N. Moothedath, S. Ronngren, N. Roy, G.P. Sharma, S. Seo, M.O. Muñoz and J. Gross, "ExPECA: An Experimental Platform for Trustworthy Edge Computing Applications," Nov. 2023, doi: 10.48550/arXiv.2311.01279 |
| [MWR+06] | M. Meyer, H. Wiemann, M. Renfors, J. Torsner and J. -f. Cheng, "ARQ Concept for the UMTS Long-Term Evolution," IEEE Vehicular Technology Conference, Montreal, QC, Canada, 2006, pp. 1-5, doi: 10.1109/VTCF.2006.442 |
| [MWR+06] | M. Meyer, H. Wiemann, M. Renfors, J. Torsner and J. -f. Cheng, "ARQ Concept for the UMTS Long-Term Evolution," IEEE Vehicular Technology Conference, Montreal, QC, Canada, 2006, pp. 1-5, doi: 10.1109/VTCF.2006.442 |
| [NHB+21] | M. Norin, R. Högman, M. Buchmayer, G. Lemne, F. Pedersen, A. Zaidi, "5G spectrum for local industrial networks," Ericsson white paper, April 2021 |
| [NHB+21] | M. Norin, R. Högman, M. Buchmayer, G. Lemne, F. Pedersen, A. Zaidi, "5G spectrum for local industrial networks," Ericsson white paper, April 2021 |
| [NYA+18] | N. Jiang, Y. Deng, A. Nallanathan, X. Kang, and T. Q. S. Quek, "Analyzing random access collisions in massive iot networks," IEEE Transactions on Wireless Communications, vol. 17, no. 10, pp. 6853–6870, 2018 |
| [NYA+18] | N. Jiang, Y. Deng, A. Nallanathan, X. Kang, and T. Q. S. Quek, "Analyzing random access collisions in massive iot networks," IEEE Transactions on Wireless Communications, vol. 17, no. 10, pp. 6853–6870, 2018 |
| [PSS+22] | J. Park, S. Samarakoon, H. Shiri, M. K. Abdel-Aziz, T. Nishio, A. Elgabli, M. Bennis, "Extreme Ultra-Reliable and Low-Latency Communication," *Nature Electronics*, vol. 8, no. 3, pp. 133-141, March 2022, doi: 10.1038/s41928-022-00728-8 |
| [SPS+23] | G. P. Sharma, D. Patel, J. Sachs, M. De Andrade, J. Farkas, J. Harmatos, B. Varga, H. -P., Bernhard, R. Muzaffar, M. Ahmed, F. Duerr, D. Bruckner, E.M. De Oca, D. Houatra, H. Zhang and J. Gross, "Toward Deterministic Communications in 6G Networks: State of the Art, Open Challenges and the Way Forward," in IEEE Access, vol. 11, pp. 106898-106923, 2023, doi: 10.1109/ACCESS.2023.3316605 |

| | |
|---|---|
| [SSI22] | A. Sawabe, Y. Shinohara and T. Iwai, "Delay Jitter Modeling for Low-Latency Wireless Communications in Mobility Scenarios," GLOBECOM 2022 - 2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 2022, pp. 2638-2643, doi: 10.1109/GLOBECOM48099.2022.10001637 |
| [SWD+18] | J. Sachs, G. Wikström, T. Dudda, R. Baldemair, K. Kittichokechai, "5G radio network design for ultra-reliable low-latency communication", IEEE Network, vol. 32, pp. 24-31, Mar./Apr. 2018 |
| [SWD+18] | J. Sachs, G. Wikström, T. Dudda, R. Baldemair, K. Kittichokechai, "5G radio network design for ultra-reliable low-latency communication", IEEE Network, vol. 32, pp. 24-31, Mar./Apr. 2018 |
| [SWD+18] | J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair and K. Kittichokechai, "5G Radio Network Design for Ultra-Reliable Low-Latency Communication," in IEEE Network, vol. 32, no. 2, pp. 24-31, March-April 2018, doi: 10.1109/MNET.2018.1700232 |
| [TS08] | O. Tickoo and B. Sikdar, "Modeling Queueing and Channel Access Delay in Unsaturated IEEE 802.11 Random Access MAC Based Wireless Networks," in IEEE/ACM Transactions on Networking, vol. 16, no. 4, pp. 878-891, Aug. 2008, doi: 10.1109/TNET.2007.904010 |
| [YGA+20] | Y. Liu, X. Wang, G. Boudreau, A. B. Sediq and H. Abou-Zeid, "A Multi-Dimensional Intelligent Multiple Access Technique for 5G Beyond and 6G Wireless Networks," in IEEE Transactions on Wireless Communications, vol. 20, no. 2, pp. 1308-1320, Feb. 2021, doi: 10.1109/TWC.2020.3032631 |
| [YYM+21] | Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis,<br>“Analyzing Grant-Free Access for URLLC Service,” IEEE Journal on Selected Areas in Communications, vol. 39, no. 3, pp. 741–755, 2021 |
| [YYM+21] | Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis,<br>“Analyzing Grant-Free Access for URLLC Service,” IEEE Journal on Selected Areas in Communications, vol. 39, no. 3, pp. 741–755, 2021 |

# 7 List of abbreviations

| | |
|---|---|
| eMMB | Enhanced Mobile Broadband (eMBB) |
| HoL | Head-of-line |
| 3GPP | 3rd Generation Partnership Project |
| 5G-Adv | 5G-Advanced |
| 5GS | 5G System |
| 5QI | 5G QoS Identifier |
| ACL | Access Control List |
| AF | Application Function |
| AGV | Automatic Guided Vehicle |
| AM | Acknowledged Mode |
| AoI | Age-of-Information |
| AQM | Active Queue Management |
| AR | Augmented Reality |
| ARQ | Automatic Repeat Request |
| BER | Bit-error Rate |
| BLER | Block Error Rate |
| CB-WFQ | Class-based Weighted-Fair Queueing |
| CDE | Conditional Density Estimation |
| CDF | Cumulative Distribution Function |
| CG | Computer Graphics |
| CNC | Centralized Network Controller |
| COTS | Commercial-off-the-shelf |
| CP | Cyclic Prefix |
| CPS | Cyber-Physical System |
| CQF | Cyclic Queueing and Forwarding |
| CSI | Channel State Indicator |
| CUC | Centralized User Configuration |
| DetCom | Deterministic Communication |
| DetNet | Deterministic Networking |
| DL | Downlink |
| DRB | Data Radio Bearer |
| DS-TT | Device-side TSN Translator |
| DVP | Delay Violation Proabality |
| EMM | Extreme-value Mixture Model |
| FDD | Frequency Division Duplexing |
| FEC | Forward Error Correction |
| GMEVM | Gaussian Mixture Extreme Value Model |
| GMM | Gaussian Mixture Model |
| gNB | Next Generation Node B |
| GPD | Generalized Pareto Distribution |
| GPRS | Generalized Packet Radio Service |

| | |
|---|---|
| GTP | GPRS Tunnelling Protocol |
| GTP-U | GTP for User plane data |
| HARQ | Hybrid Automatic Repeat Request |
| IIoT | Industrial Internet of Things |
| IoT | Internet of Things |
| IPDV | Inter-Packet Delay Variation |
| KPI | Key Performance Indicator |
| MAC | Medium Access Control |
| MBB | Mobile Broadband |
| MCS | Modulation and Coding Scheme |
| MDN | Mixture Density Network |
| MIMO | Multiple-Input Multiple-Output |
| ML | Machine Learning |
| NMSE | Normalized Mean Square Error |
| NR | New Radio |
| NW-TT | Network-side TSN Translator |
| OE | Occupational Exoskeletons |
| OFDM | Orthogonal Frequency Division Multiplexing |
| OPC UA | Open Platform Communications Unified Architecture |
| PD | Packet Delay |
| PDC | Packet Delay Correction |
| PDCP | Packet Data Convergence Protocol |
| PDF | Probability Density Function |
| PDU | Protocol Data Unit |
| PDV | Packet Delay Variation |
| PDV | Packet Delay Variation |
| PHY | Physical |
| PQ | Priority Queueing |
| PRB | Physical Resource Block |
| QoS | Quality of Service |
| RAN | Radio Access Network |
| RLC | Radio Link Control |
| RSRP | Reference Signal Received Power |
| RSRQ | Reference Signal Received Quality |
| RTT | Round Trip Time |
| SBFD | Subband Full Duplex |
| SCS | Sub-carrier Spacing |
| SDAP | Service Data Adaption Protocol |
| SDR | Software-defined Radio |
| SDU | Service Data Unit |
| SINR | Signal-to-interference-plus-noise Ratio |
| SNC | Stochastic Network Calculus |

| | |
|---|---|
| SNR | Signal-to-noise Ratio |
| TB | Transport Block |
| TDD | Time-division Duplexing |
| TRP | Transmission and Reception Point |
| TSN | Time-sensitive Networking |
| TT | TSN Translator |
| TTI | Transmission Time Interval |
| UE | User Equipment |
| UL | Uplink |
| UM | Unacknowledged Mode |
| UPF | User Plane Function |
| URRLC | Ultra Reliable and Low Latency Communication |
| WFQ | Weighted-Fair Queueing |
| XR | Extended Reality |

Table 3: List of abbreviations