NEW SERVICES & APPLICATIONS WITH 5G ULTRA-RELIABLE LOW LATENCY COMMUNICATIONS

5G Americas Whitepaper

NOVEMBER 2018

Table of Contents

1. INTRODUCTION

New services and applications requiring lower latency, better reliability, massive connection density and improved energy efficiency are emerging in an unprecedented fashion. A variety of advanced features make 5G uniquely well positioned to meet all of these requirements and capitalize on these market opportunities. A prime example is Ultra-Reliable Low-Latency Communication (URLLC), a set of features designed to support mission-critical applications such as industrial internet, smart grids, remote surgery and intelligent transportation systems.

With 4G LTE, latency is currently in the 4-millisecond range under 3GPP Release 14. URLLC is part of Release 15 and has a target of 1-millisecond. URLLC also is ideal for applications that require end-to-end security and 99.999 percent reliability, and it's almost deterministic in time bounds on packet delivery. This combination of capabilities requires almost a fundamentally different approach to system design and operations compared to previous mobile wireless technologies.

The physical layer is unquestionably the most challenging because URLLC must satisfy two conflicting requirements: low latency and ultra-high reliability. This combination is a vastly different type of quality of service (QoS) compared to traditional mobile broadband applications.

This white paper describes the principles of achieving URLLC while comparing them to the traditional methodology used in information and communication theory and explains why a new approach is required. It also highlights key requirements of URLLC services and provides an overview of URLLC communications with an emphasis on technical challenges and solutions. Further, this paper discusses the following:

- Design principles to enable URLLC services in 5G, many of which have been considered as work items in the 3GPP Release 15 standards and that will be part of the first release for 5G New Radio (NR)
- Physical-layer issues, enabling technologies, packet and frame structure, multiplexing schemes, coding and reliability improvement techniques
- Theoretical queueing analysis and performance data that support systems design
- The building blocks in a wireless communication system for supporting URLLC connections in the context of key applications and services

1.1 SOPHISTICATED NEW ARCHITECTURES TO MEET DEMANDING NEW REQUIREMENTS

A growing number of mission-critical applications have stringent communication performance and reliability requirements. Communications with vehicles, high-speed trains, drones and industrial robots are just a few examples of applications where wireless must meet either high reliability (for example, $<10^{-5}$ packet drop rate) or low latency (for example, ~1 ms) requirements, or both at the same time. These applications frequently have strong security requirements, too.

To meet all of these requirements, 5G combines URLLC with enhanced Mobile Broadband (eMBB) services under a unified 5G air interface framework. To achieve the 1-millisecond goal, the basic problem that needs to be addressed is the end-to-end network latency. This is the time period from when, for example, an Internet of Things (IoT) sensor transmits data to the point that processing is complete at the back end of the network, and the subsequent communications are generated by the network in response and received at the sensor. Figure 1.1 illustrates this process, which URLLC shortens by reducing the user plane latency.

This involves the communications from application processing at the device modem to the application processing in the base station modem.

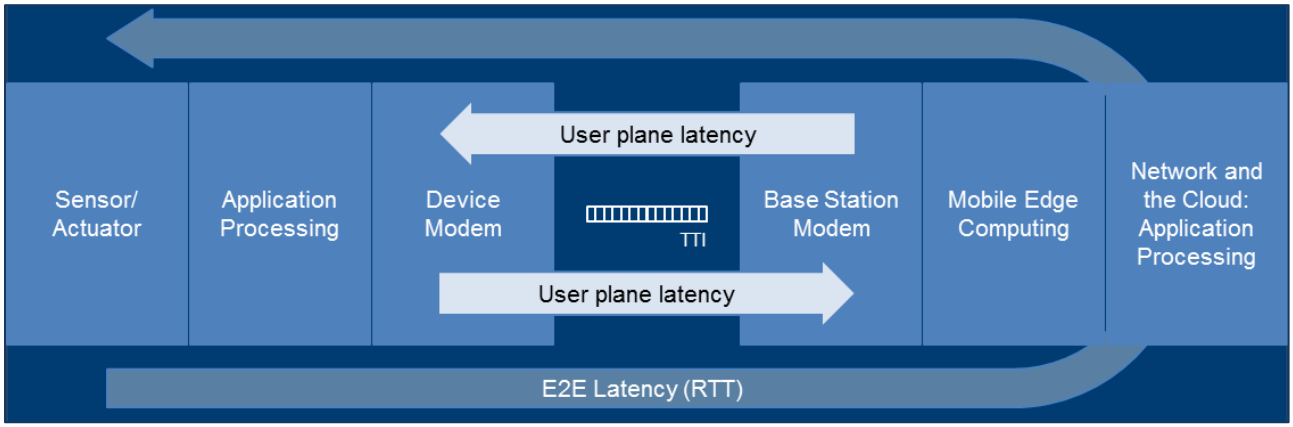


Figure 1.1. Low Latency Problem – Latency involved in a packet from origination at the source to destination in a 5G network.

User plane latency is the time it takes to successfully deliver an application layer packet or message at the radio protocol layer from the service data unit (SDU) ingress point to the corresponding egress point (TR 38.913). Round trip time (RTT) includes user plane latency contributions, application processing times and transport network delays.

Modem processing times, radio Transmission Time Interval (TTI) and an averaged contribution from Hybrid Automatic Repeat Request (HARQ) retransmissions all contribute to the user plane latency. 5G is expected to significantly reduce user plane latency to less than1 ms. For example, typical 5G parameters of 60 kHz subcarrier spacing and a 2 OFDM symbol TTI allow for a user plane latency of significantly less than 1 ms.

It is expected that Multi-Access Edge Computing (MEC), a solution deployed today in many private 4G/LTE networks, can eliminate network delays of approximately 100 ms from end-to-end latency.

There is a general consensus that the future of many industrial control, traffic safety, medical and internet services depends on wireless connectivity with guaranteed, consistent latencies of 1 ms or less and exceedingly stringent reliability of Block Error Rates (BLERs) as low as 10^{-9}. The increased capacity would be achieved by using higher spectrum bands and through network densifications. At the same time, wireless network and device technologies also must advance to do their part to minimize latency and maximize reliability. This paper explains several URLLC techniques already adopted and currently being considered by 3GPP 5G NR.

Low-latency and high-reliability design: The paper provides a summary of key elements of the 5G air interface design including the following:

- Integrated frame structure (for example, self-contained slot structure and low-latency, mini-slot structures)
- Fast turnaround and flexible HARQ design
- Efficient control and data resource sharing and multi-link/multi-carrier diversity control/data transmission
- Low-latency, multiple access scheme and grant-free based uplink transmission with autonomous retransmission and
- Advanced channel coding schemes to support URLLC traffic

Unified air interface framework for URLLC: It is important to be able to quickly schedule mission-critical traffic in order to meet stringent latency requirements without consuming excessive radio interface resources. This paper describes a flexible and scalable unified air interface design to satisfy both URLLC and eMBB requirements based on the following:

- Scalable and unified eMBB and URLLC design framework to cover different scenarios, including indoor, small cell, urban macro/micro and high-speed trains
- Dynamic multiplexing of URLLC and eMBB for efficient spectrum utilization
- Preemption indication to ensure robust performance
- Code-block-group (CBG)-based retransmission for enhanced HARQ design

Strong Security: In addition to protecting the confidentiality and integrity of the information transmitted over the cellular network, it is also important to guarantee service availability in the presence of denial-of-service (DoS) attacks preventing network access. This paper will provide an overview of security techniques for supporting mission-critical services.

3GPP has specified URLLC as a key feature for Release 15 5G NR, in addition to eMBB. This paper provides an update on 3GPP standardization progress and a summary of URLLC-related designs. The paper is organized as follows:

- Section 2 describes the current market landscape for URLLC, including the industry status and the planned use cases
- Section 3 explains the key requirements and KPIs necessary for adoption of 5G communications for URLLC use cases
- Section 4 provides details of the physical and mac layer innovations in designing the 5G air interface
- Section 5 reviews the 5G network architecture and the upper-layer protocols planned to implement URLLC features
- Section 6 provides examples and references of performance results as currently seen and evaluated by the industry proponents
- Section 7 covers the key security implementation aspects necessary for URLLC
- Section 8 gives a window into the current status of the 3GPP standards

2. MARKET DRIVERS AND USE CASES FOR 5G URLLC

Over the past several years, a variety of marketplace trends have resulted in a large and growing number of applications that require URLLC. The NGMN (Next Generation Mobile Networks) 5G white paper in 2015[1] highlighted low-latency and high-reliability applications including video, high-speed trains, extreme real-time communications, tactile internet, collaborative robots, automation and augmented reality. 5G Americas (www.5GAmericas.org) has also published several white papers on use cases requiring 5G capabilities.

3GPP picked up the ball in 2015 with the SMARTER project to define technology for new services and markets in the 5G time frame. Since then, several studies and work items have produced an extensive list of URLLC use cases and requirements from industry verticals, network operators and suppliers. Industry groups outside of 3GPP—including 5G Americas (www.5gamericas.org), GSMA (Groupe Speciale Mobile Association), the new 5G Alliance for Connected Industries and Automation (www.5g-acia.org) and the 5G

[1] *NGMN 5G White Paper,* NGMN Alliance. February 2015. www.ngmn.org.

Automotive Association (www.5Gaa.org)—are also developing requirements for specific markets in the 5G time frame and working with 3GPP to realize them.

New business opportunities are unquestionably in the offing in the wake of the integration of health care, industrial processes, transport services and entertainment applications that require low latency and ultra-reliability features. These applications, however, pose difficult requirements for current network deployments and major challenges to the design and the implementation of the future 5G network.

2.1 APPLICATIONS

Applications that require ultra-low latency networks across the different industries are summarized in Table 2.1.

Table 2.1. URLLC Applications.

Industry Vertical	Application
Smart Factory/Industrial Automation	Industrial Control Robot Control Machine to Machine Process Control
Healthcare Industry	Remote Diagnosis Emergency Response Remote Surgery
Entertainment Industry	Immersive Entertainment Online Gaming
Transport Industry	Driver Assistance Applications Enhanced Safety Autonomous Driving Traffic Management
Manufacturing Industry	Motion Control Remote Control AR and VR Applications
Energy Sector	Smart Energy Smart Grid

The market sizing of some of the key verticals that are rooted in 5G low latency features is well captured in the 2017 publication *Business Case and Technology Analysis for 5G Low Latency Applications.*[2] The digital health care vertical is estimated to rise to a global market size of GBP $43 billion, almost doubling in worth compared to 2014. The global automotive vertical or the connected car market offers new applications in driver assistance, basic safety technologies and autonomous driving, and is forecasted to reach GBP $104.2 billion by 2019 and to continue to grow at a CAGR of 35 percent.[3] The global market for Augmented Reality (AR) and Virtual Reality (VR) is expected to reach GBP 118.5 billion; low latency features are critical for these applications.[4] The AR and VR areas[5] are among the largest growth segments and are widely expected to fuel the wearable industry. Importantly, the ICT sector in the manufacturing vertical, based on

[2] *Business Case and Technology Analysis for 5G Low Latency Applications,* by Maria A. Lema, Andres Laya, Toktam Mahmoodi, Maria Cuevas, Joachim Sachs, Jan Markendahl and Mischa Dohler, IEEE Access. April 2017.
[3] *Intelligent Transportation Systems Report for Mobile,* White Paper, GSMA. April 2015 and *Connected Car Forecast: Global Connected Car Market to Grow Threefold Within Five Years*, White Paper, GSMA. February 2013.

[5] "*The wearable future*," by M. Pegler, J. Lamano, and K. Shahabi' PwC, White Paper, 2014.

robotic controls and industrial automation technologies, is estimated to provide a market of GBP 755.7 billion in the European Union (EU) alone.

Clearly, a compelling case has emerged for 5G to support low latencies and ultra-reliability features. Achieving those requirements is regarded as a major challenge, one that requires major financial investment. The following sub-sections are devoted to reviewing the emerging mission-critical applications where latency and reliability requirements will be identified.

2.2 INDUSTRIAL AUTOMATION

Industrial automation is a key application for URLLC features. Some industrial processes have extremely tight Key Performance Indicators (KPIs) for 5G communications links between sensors, actuators and controllers. Example use cases in this category include the following:

- Motion control
- Industrial Ethernet
- Control-to-control communication
- Process automation
- Electric power generation and distribution

URLLC is one of the enabling technologies in the fourth industrial revolution. In this new industrial vision, industry control is automated by deploying networks in factories. Typical industrial automation use cases requiring URLLC include factory, process and power system automation. Use cases involve communication transfers enabling time-critical factory automation that are required in many industries across a wide spectrum that includes metals, semiconductors, pharmaceuticals, electrical assembly, food and beverage.

To enable these applications, an end-to-end latency lower than 0.5 ms and an exceedingly high reliability with BLER of 10^{-9} is required. Traditionally, industrial control systems are mostly based on wired networks because the existing wireless technologies cannot meet the industrial latency and reliability requirements. Nevertheless, replacing the currently used wires with radio links can bring substantial benefits:

- Reduced cost of manufacturing, installation and maintenance
- Higher long-term reliability as wired connections suffer from wear and tear in motion applications
- Inherent deployment flexibility

2.3 GROUND VEHICLES, DRONES, ROBOTS

5G will support communications with and among ground vehicles, drones and robots. Automated guided vehicles are common in factory applications, and they have requirements for coverage and mobility in addition to URLLC. Robots include fixed and mobile applications, with varying KPIs for communications links depending on the specific application. Example use cases in this category include the following:

- Mobile and industrial robots
- Connectivity for the factory floor
- Factories of the future, including automated guided vehicles

2.4 TACTILE INTERACTION

Tactile interaction refers to a level of responsiveness that works at a human scale. For example, remote health care or gaming applications may require very low round-trip times to convince human senses that the perceived touch, sight and sound are lifelike.

These use cases involve interaction between humans and systems, where humans wirelessly control real and virtual objects, and the interaction requires a tactile control signal with audio or visual feedback. Robotic controls and interaction include several scenarios with many applications in manufacturing, remote medical care and autonomous cars. The tactile interaction requires real-time reactions on the order of a few milliseconds. Example use cases in this category include these:

- Tactile internet
- Extreme real-time communications

2.5 AUGMENTED REALITY (AR) AND VIRTUAL REALITY (VR)

Augmented Reality (AR), Virtual Reality (VR) and Mixed Reality (MR) bring higher bandwidth requirements in addition to URLLC constraints. The main difference between AR and VR is in the uplink requirements. VR needs low-data-rate pose estimates from the headset, while AR requires images of the view experienced by the user.

These use cases are the critical IoT applications that will have very high demands on reliability, availability and low latency, with lower demands on the volume of data, but significantly higher business value. These use cases also fall into the category of mission-critical Machine-Type Communication (MTC).

The mission-critical MTC is envisioned to enable real-time control and automation of dynamic processes in various fields, such as industrial process automation and manufacturing, energy distribution and intelligent transport systems. These use cases and applications feature interactions across all categories, human-to-human, human-to-machine and machine-to-machine.

2.6 EMERGENCY, DISASTERS AND PUBLIC SAFETY

The use cases in this category require robust and reliable communications in case of natural disasters such as earthquakes, tsunamis, floods and hurricanes. The use cases may require accurate position location and quick communication exchanges between users and systems. Energy efficiency in user battery consumption and network communications are critical in these use cases. The public safety organizations require enhanced and secured communications with real-time video and the ability to send high-quality pictures.

2.7 URGENT HEALTH CARE

These use cases are envisioned around applications involving remote diagnosis and treatment. There is a need for remote patient monitoring and communications with devices measuring vital signs such as ECG, pulse, blood glucose, blood pressure and temperature. The remote treatment and response based on monitored data can be life critical for a patient, requiring immediate, automatic or semi-automatic response. The URLLC features are used for two aspects here: remote surgical consultations and remote surgery.

Remote surgery is about applications in a mobile scenario in ambulances, disaster situations and remote areas requiring providing precise control and feedback communication mechanisms for surgeons in terms of low latency, high reliability and tight security. In a remote surgery scenario, the entire treatment procedure of patients is executed by a surgeon at a remote site, where hands are replaced by robotic arms. In these two cases, the communication networks should be able to support the timely and reliable delivery of audio and video streaming.

Moreover, the haptic feedback enabled by various sensors located on the surgical equipment is also needed in remote surgery such that the surgeons can feel what the robotic arms are touching for precise decision making. Among these three types of traffic, it is haptic feedback that requires the tightest delay requirement with the end-to-end RTTs lower than 1 ms. In terms of reliability, rare failures can be tolerated in remote surgical consultations, while the remote surgery demands an extremely reliable system (BLER down to 10^{-9}) because any noticeable error can lead to catastrophic outcomes.

2.8 INTELLIGENT TRANSPORTATION

The realization of URLLC can empower several technological transformations in the transportation industry, including automated driving, road safety and traffic efficiency services. These transformations will get cars fully connected such that they can react to increasingly complex road situations by cooperating with others rather than relying on their local information. These trends will require information to be disseminated among vehicles reliably within extremely short time duration.

For example, in fully automated driving with no human intervention, vehicles can benefit by the information received from roadside infrastructure or other vehicles. The typical use cases of this application are automated overtake, cooperative collision avoidance and high-density platooning, which require stricter end-to-end latencies and high reliabilities.

3. REQUIREMENTS AND KPIS FOR 5G URLLC

Based on use cases and applications described above, 3GPP has specified normative requirements for the 5G system.[6] This includes service requirements and KPIs for private networks, industrial automation, AR, VR and the tactile internet. These requirements will be fully specified with URLLC capabilities defined in 3GPP Radio Access Network (RAN) and Services and Systems Aspects (SA) groups. 3GPP has recently benefited from an influx of non-traditional participants who would like to use 5G to serve new markets such as industrial automation, entertainment and transportation systems. Their research and requirements have been incorporated in recent studies and work items.

For 3GPP Release 16, SA1 augmented this work with studies on 5G integration with Local Area Networks (LAN),[7] communications for automation in verticals such as factory automation, transportation and program making[8], and business models for network slicing[9]. For the latest normative 3GPP KPIs for URLLC, we can point directly to 3GPP specifications[10] [11]. See the following summary table[12] (including approved change requests (CRs) up to May 2018). These numbers will be further refined for 3GPP Release 16.

[6] 3GPP TS 22.261: *Service requirements for next generation new services and markets.*
[7] 3GPP TR 22.821: *Feasibility Study on LAN Support in 5G.*
[8] 3GPP TR 22.804: *Study on Communication for Automation in Vertical domains (CAV).*
[9] 3GPP TR 22.830: *Study on Business Role Models for Network Slicing.*
[10] 3GPP TS 22.261: *Service requirements for next generation new services and markets.*
[11] 3GPP TS 22.104: *Service requirements for cyber-physical control applications in vertical domains.*
[12] 3GPP TS 22.261: *Service requirements for next generation new services and markets.*

Table 3.1. Performance Requirements for Low-Latency and High-Reliability Scenarios.

Scenario	Max. allowed end-to-end latency (note 2)	Survival time	Communication service availability (note 3)	Reliability (note 3)	User experienced data rate	Payload size (note 4)	Traffic density (note 5)	Connection density (note 6)	Service area dimension (note 7)
Discrete automation	10 ms	0 ms	99,99percent	99,99percent	10 Mbps	Small to big	1 Tbps/km^2	100 000/km^2	1000 x 1000 x 30 m
Process automation – remote control	60 ms	100 ms	99,9999 percent	99,999percent	1 Mbps up to 100 Mbps	Small to big	100 Gbps/km^2	1 000/km^2	300 x 300 x 50 m
Process automation – monitoring	60 ms	100 ms	99,9percent	99,9percent	1 Mbps	Small	10 Gbps/km^2	10 000/km^2	300 x 300 x 50
Electricity distribution – medium voltage	40 ms	25 ms	99,9percent	99,9percent	10 Mbps	Small to big	10 Gbps/km^2	1 000/km^2	100 km along power line
Electricity distribution – high voltage (note 2)	5 ms	10 ms	99,9999 percent	99,999percent	10 Mbps	Small	100 Gbps/km^2	1 000/km^2 (note 8)	200 km along power line
Intelligent transport systems – infrastructure backhaul	30 ms	100 ms	99,9999 percent	99,999percent	10 Mbps	Small to big	10 Gbps/km^2	1 000/km^2	2 km along a road

NOTE1: Currently realized via wired communication lines

NOTE 2: This is the maximum end-to-end latency allowed for the 5G system to deliver the service in the case the end-to-end latency is completely allocated to the 5G system from the UE to the Interface to Data Network

NOTE 3: Communication service availability relates to the service interfaces, reliability relates to a given node. One or more retransmission over the radio interface may take place in order to satisfy the reliability requirement

NOTE 4: Small: payload typically ≤ 256 bytes

NOTE 5: Based on the assumption that all connected applications within the service volume require the user experienced data rate

NOTE 6: Under the assumption of 100percent 5G penetration

NOTE 7: Estimates of maximum dimensions; the last figure is the vertical dimension

NOTE 8: In dense urban areas

NOTE 9: All the values in this table are targeted values and not strict requirements. Deployment configurations should be considered when considering service offerings that meet the targets

4. PHY/MAC LAYER DESIGN FOR 5G URLLC

As explained in Section 3, URLLC use cases have very stringent requirements in terms of latency and reliability. This creates significant challenges for cellular networks due to variables such as interference levels, channel fading and user equipment (UE) movements. This section discusses the enabling technologies in 5G physical and medium access control (PHY/MAC) design that support low latency, high reliability and the efficient multiplexing between URLLC and other traffic in the system.

4.1 LOW LATENCY

URLLC use cases can have very stringent latency requirements, as low as 5 ms for end-to-end latency, as shown in Section 3. Considering the different components of the end-to-end network, the budget for the air interface delay can be very limited (for example, 1 ms or even less). To achieve such low latency, every step of the data delivery needs to be optimized. Figure 4.1 illustrates the latency components in each step of a downlink (DL) data transmission and the corresponding mechanisms to reduce the latency.

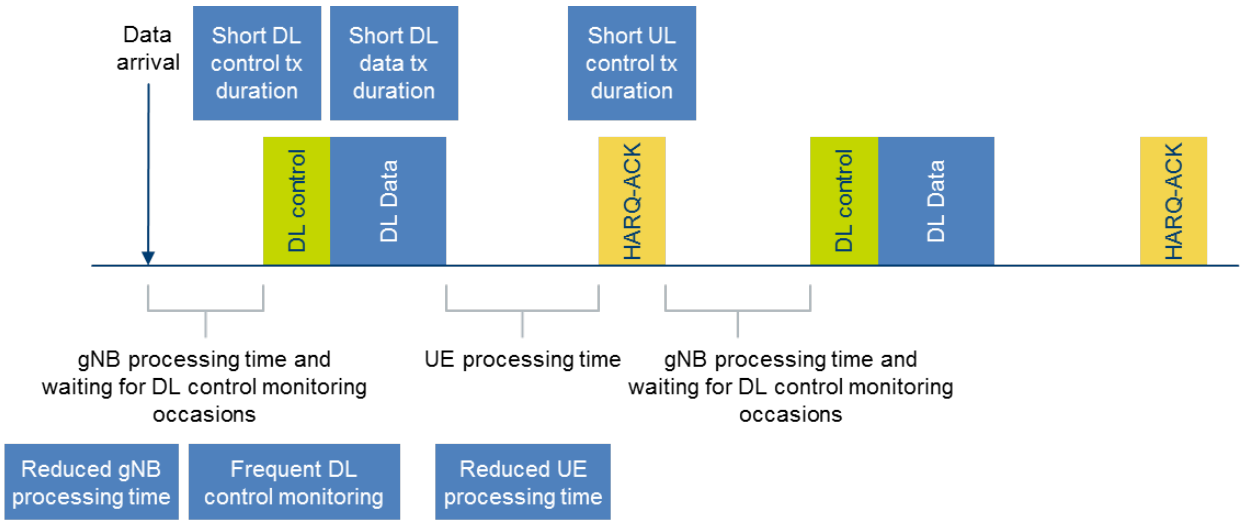


Figure 4.1. Latency Components of a DL Transmission and the Mechanisms to Reduce the Latency.

Some key techniques in the PHY/MAC layer to reduce the latency include the following:

- Frequent transmission opportunities that minimize waiting time
- Flexible transmission duration (short duration for both data and control channel)
- Short UE processing time
- Short next-generation NodeB (gNB) processing time
- Grant-free (or configured grant) UL transmission
- Flexible frame structure for Time Division Duplexing (TDD)

Frequent Transmission Opportunities:

For a scheduling-based system, when a data packet comes, it needs to wait for its transmission opportunities. The key aspects are these:

- Frequent monitoring occasions for DL control channel
- Frequent opportunities for UE to transmit Scheduling Request (SR) on the uplink (UL)
- Flexible scheduling timing between the Physical Downlink Control Channel (PDCCH) and the Physical Downlink Shared Channel/Physical Uplink Shared Channel (PDSCH/PUSCH)

On the DL, the DL control channel is used to carry scheduling information for DL and UL data transmission. Typically, a UE does not continuously monitor DL control for power consumption considerations. However, for URLLC service, to reduce the waiting time for delivering the control information, the UE needs to monitor the DL control channel frequently. The monitoring periodicity can be as low as one or a few Orthogonal Frequency-Division Multiplexing (OFDM) symbols, if necessary.

On the UL, for scheduling-based transmission, when a data packet arrives at the UE, the UE needs to send a SR to the gNB to request UL resource allocation. To minimize the waiting time for sending the scheduling request, the periodicity of the SR resource configuration should be sufficiently low.

In addition, flexible scheduling timing between DL control channel and DL/UL data channel can allow the gNB to schedule a data packet as soon as the resource becomes available. This is especially important for Time Division Duplex (TDD) systems where the DL and the UL are time-division multiplexed.

Flexible Transmission Duration:

The actual transmission time is also an important component of the over-the-air latency, which can be reduced by supporting short duration for the data and control channels.

In 5G NR, flexible transmission duration is supported. For the DL and UL data channels, the short transmission durations can be achieved by using larger subcarrier spacing (therefore, shorter symbol/slot duration) and/or small scheduling units such as mini-slots, which can be as short as one OFDM symbol, as illustrated in Figure 4.2.

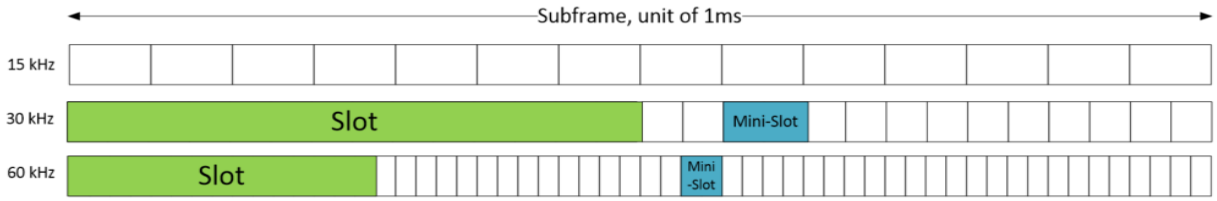


Figure 4.2. Illustration of Slot and Mini-slot Structure for Different Subcarrier Spacings.

Some examples of mini-slot flexible data transmission duration are illustrated in Figure 4.3, with mini-slot length of 2, 4 and 7 OFDM symbol for PDSCH transmission. Different data transmission durations can be used to achieve different latency versus spectrum efficiency tradeoffs. For instance, when subcarrier spacing (SCS) = 30kHz is used, transmission duration of 70us, 140us, 250us is achieved by 2-symbol, 4-symbol, 7-symbol mini-slot as opposed to 0.5 ms based on slot-based transmission, which demonstrates substantial latency reduction by adopting a short transmission duration.

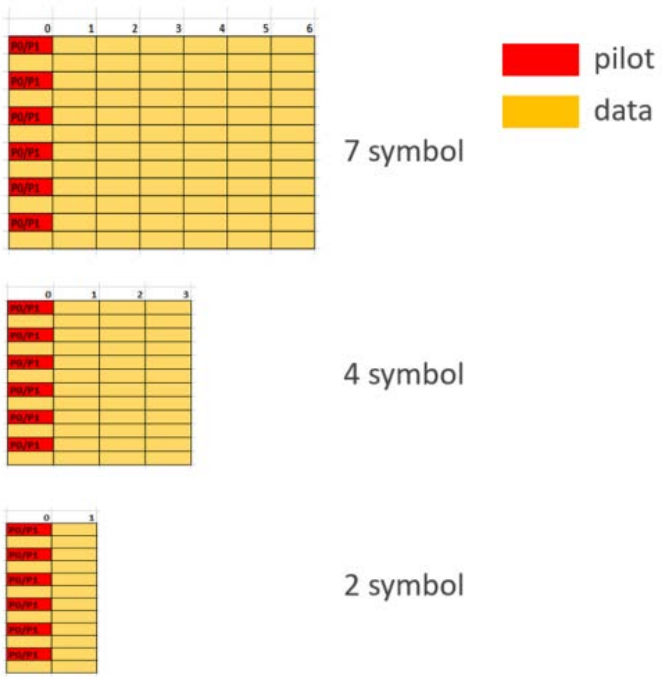


Figure 4.3. Illustration of Different Data Transmission Durations.

Figure 4.4 provides an example of using different scheduling intervals for URLLC and eMBB traffic, where mini-slot-based scheduling is used for URLLC for fast packet delivery, while slot-based scheduling is used for eMBB for better spectrum efficiency.

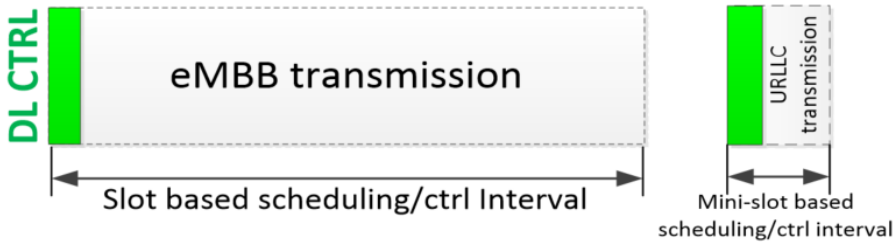


Figure 4.4. Illustration of Different Scheduling Intervals for eMBB and URLLC Traffic.

Similarly, short transmission durations for the DL control channel (which carries the scheduling information for DL and UL data channels) and UL control channel (which provides UL control information such as Hybrid Automatic Repeat Request – Acknowledgment (HARQ-ACK) feedback) can be achieved by larger subcarrier spacing and/or a short resource unit in time. The DL/UL control channel can be transmitted using as short as one OFDM symbol.

Reduced Processing Time at the UE/gNB:

HARQ is a key technique to ensure high reliability with good link-level efficiency with feedback. HARQ turnaround processing time is another important aspect of enhancements to achieve both low latency and high reliability at the same time: fast processing turnaround reduces latency, while multiple transmission opportunities with Acknowledgment/Negative Acknowledgment (ACK/NAK) feedback guarantees high reliability of data packet delivery.

Many aspects need to be considered to allow efficient pipeline processing at the UE/gNB, which is critical for reducing the processing time. These include but are not limited to the following:

Structure of the Data Channel:

- Front-loaded DeModulation Reference Signal (DMRS): As shown in Figure 4.3, the location of pilot (therefore, DMRS) in the data channel is important for the demodulation and decoding processing latency at the receiver. Putting all DMRS at the beginning of the data transmission (instead of being distributed over the slot as in LTE) allows the receiver to start the channel estimation right away, without waiting for the later symbols.
- No time domain interleaving: Interleaving across different OFDM symbols provides good time diversity when the time duration is long enough. However, it prevents the pipelining processing at the receiver, as the receiver needs to wait to receive all the symbols to be received before they can be de-interleaved and then proceed to decoding. No interleaving across the time would allow the receiver to process each symbol once it is received.
- Frequency-first mapping: Similarly, frequency-first mapping also allows symbol-by-symbol processing at the receiver.

Channel Coding

- The channel code design for URLLC should consider the decoding complexity and processing time. It should allow efficient parallelization of decoding process to reduce the latency. Low-Density Parity Check (LDPC) code is a good channel coding candidate for 5G URLLC due to its high parallelizable decoder, thanks to the quasi-cyclic structure of 5G NR LDPC code.

By significantly reducing the processing time at the UE and the gNB, the HARQ Round Trip Transmission (RTT) can be greatly reduced. Figure 4.5 provides examples of RTT for 15 kHz and 30 kHz subcarrier spacings. For 15 kHz, the processing time at both the UE and gNB is assumed to be 3 symbols, which results in an RTT of 10 symbols (< 1 ms). For 30 kHz, the RTT is 14 symbols (0.5 ms), assuming the processing time of 4.5 symbols at the UE and the gNB. This is significantly faster turnaround than the RTT of 8 ms in LTE. Note that the gNB processing time of ACK/NAK feedback detection and retransmission scheduling is not specified, and it is up to gNB implementation. In theory, the N3 value can be as low as the gNB is able to support.

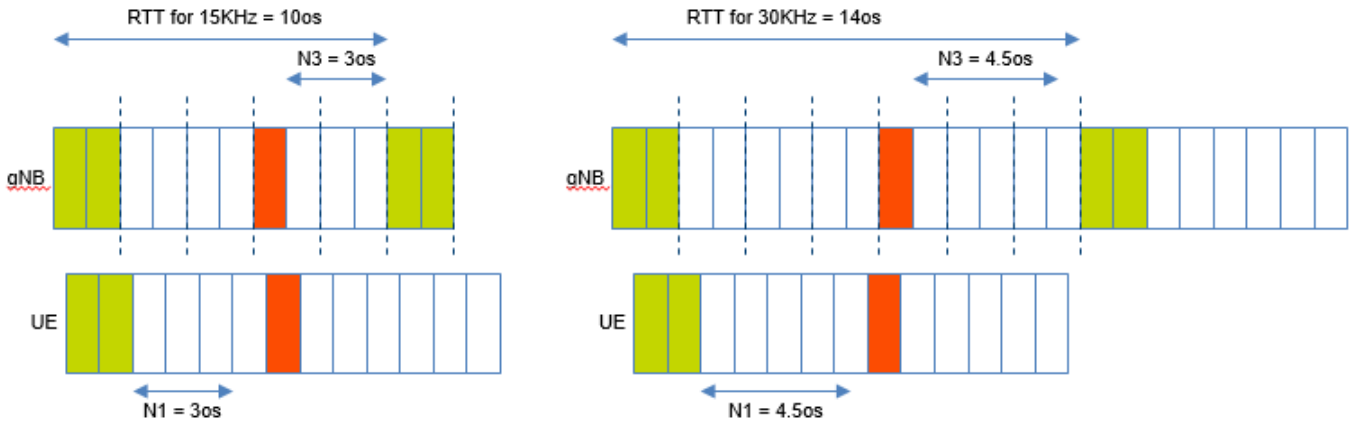


Figure 4.5. Examples of HARQ Round-trip Time with Aggressive Processing Timeline at UE and gNB.

Some examples of different HARQ timelines are shown in Figure 4.6. It can be seen that tightened HARQ timeline can substantially improve URLLC system capacity.

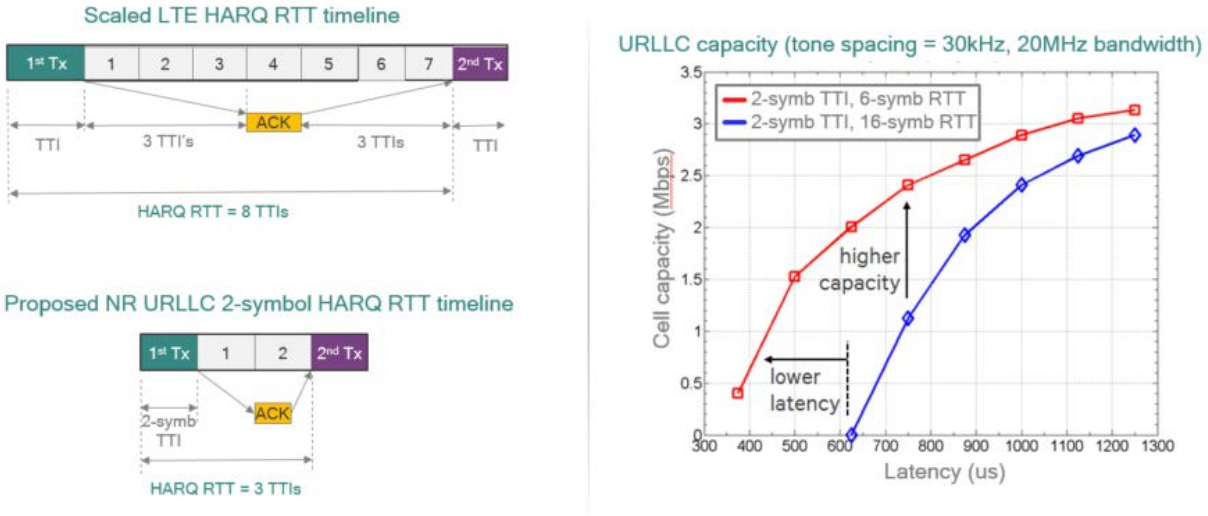


Figure 4.6. Examples of Performance Gain from Faster HARQ Timeline.

Grant-free UL Transmissions:

The normal scheduling-based (or grant-based) UL transmission requires the UE to transmit a scheduling request first and then wait for the UL grant from the gNB. Although the timeline in 5G has been reduced significantly compared to LTE, this additional handshake between UE and gNB can still make it challenging to meet the URLLC latency requirement of 1 ms or less in some scenarios. For example, in a TDD single carrier scenario, UE and gNB handshake turnaround can involve an excessive initial time overhead. Grant-free (aka configured grant in 3GPP) UL transmissions allow the gNB to configure periodic UL resources for a UE; when the UE has data, it can transmit on the configured resources without the need for dynamic UL

grant. This improves the UL latency, and the additional savings sometimes may be critical for meeting the most stringent latency requirement.

Flexible Frame Structure for TDD:

URLLC, especially its low-latency aspect, is possibly more challenging for TDD than for Frequency Division Duplexing (FDD). That's because the TDD DL and UL need to share the resources in a TDM fashion, which introduces additional delay. Most traffic is unpredictable. So, in order to be able to deliver a DL or UL packet and the corresponding control signaling with minimum delay, it is important to have a flexible frame structure that allows fast switching between DL and UL and also allows the gNB to dynamically determine the DL or UL direction based on the traffic.

4.2 HIGH RELIABILITY

HARQ transmissions have been a traditional and effective way in wireless systems to achieve efficient transmission and low residual BLER after multiple HARQ transmissions. However, when the high reliability requirement comes together with low latency, there may be very limited opportunities for HARQ retransmissions within the latency budget. To achieve high reliability on the order of 10^{-5} or even lower for data delivery with a limited number of HARQ transmissions, each air interface channel needs to be designed with a high reliability target. At the 5G NR PHY layer, the following techniques aspects have been used to improve reliability:

- Data channels
    - Channel coding
        - Channel code that facilitates efficient HARQ support
        - Channel code that is designed with error floor optimization
    - Channel State Information (CSI) report enhancements
        - For eMBB services, the BLER target for CSI report is 10 percent. This would be a mismatch if a lower BLER target is required for scheduling. Therefore, it would be desirable to support lower BLER target (for example, 10^{-5}) for CSI reporting and the corresponding CQI table in addition to the regular 10 percent CSI report
    - MCS table enhancements
        - Lower BLER can be achieved at the expense of lower spectral efficiency. In addition, low latency results in fewer HARQ transmissions. Therefore, URLLC requires lower spectral efficiency entries in the Modulation Coding Scheme (MCS) table
    - Time/frequency/spatial diversity
        - Time diversity can be difficult for low-latency applications, as the packet cannot span over a long time due to the latency requirement. However, frequency and spatial diversity can be used to improve the reliability. This includes frequency hopping and spatial diversity transmission schemes (for example, precoding cycling). One packet can also be transmitted from multiple, non-collocated Transmission Points (multi-TRP) to achieve a different level of spatial diversity

- Control channel
    - For the DL control channel, the payload size matters because it affects the required Signal-to-Interference-plus-Noise Ratio (SINR) needed to achieve a certain BLER target. Therefore, a compact Downlink Control Information (DCI) with small payload size is useful for improving the reliability

- Similarly, for the UL control channel, smaller payload size is also helpful. It can be potentially achieved by, for example, not multiplexing HARQ-ACK feedback with other UL control information to guarantee the reliable delivery of HARQ-ACK feedback
- In addition, higher aggregation levels can be supported for the DL control channel to reduce the effective code rate

- Repetitions for data and control channels
  - Repetition is a common approach to improve coverage and reliability. It is especially useful when there is not sufficient time for the UE to process and provide HARQ ACK/NAK feedback. Repetitions can be applied on all the DL/UL data and control channels

Reliability can additionally be improved by enabling higher-layer packet duplication, such as Packet Data Convergence Protocol (PDCP) packet duplication, as explained in Section 5.

4.3 MULTIPLEXING OF URLLC AND OTHER TRAFFIC

Queueing Effect of URLLC Traffic

In a RAN, the L2-to-L2 end-to-end latency of a successful data transmission is comprised of: scheduling delay (the time between packet arrival and the next scheduling instant); queueing delay; transmission delay; receiver-side processing and decoding delay; and multiple HARQ RTTs. The queueing delay results from the statistical multiplexing of data flows destined for multiple URLLC users. The data flows may also be sporadic and bursty because of the traffic models of various URLLC use cases.

The queueing effect needs to be considered in the systems design of URLLC because of the hard latency requirement[13] [14]. In general, a sufficient number of HARQ retransmissions are needed to achieve high reliability while leaving enough delay margin from the hard latency bound to mitigate the queueing effect, which is worsened as the admitted traffic load increases. This approach maximizes the spectral efficiency for the URLLC services.

To obtain qualitative insights about the relationship between the URLLC capacity and the hard latency requirement, an M/M/m/k queueing model is explained:

- The first M means Poisson packet arrivals
- The second M means exponential service times, which reflect the fact that the time to decode a data packet after multiple HARQ retransmissions follows approximately a geometric distribution (This is only an approximation in the sense that a packet is continuously serviced in the queueing model, but the resources are made available in between HARQ retransmissions of the packet in the wireless network)
- The notation m indicates the number of allowed concurrent transmissions, which is proportional to the system bandwidth available to the URLLC services
- The notation k means that if an arriving packet observes k outstanding packets in the system for some k>m, including the packets both being queued and undergoing HARQ retransmissions, it will be dropped from the network. The value of k increases with the hard latency requirement (for example, d milliseconds) in the sense that if an arriving packet sees d milliseconds' worth of packets

[13] *5G ultra-reliable and low-latency systems design*", C. Li, J. Jiang, W. Chen, T. Ji, J. Smee, 2017/6/12, Networks and Communications (EuCNC), 2017 European Conference on, p 1-5

[14] "*5G-Based Systems Design for Tactile Internet*", C. Li, C. Li, K. Hosseini, S. Lee, J. Jiang, W. Chen, G. Horn, T. Ji, J. Smee, J. Li, 2018/8/31, Proceedings of the IEEE

awaiting in the system, it will surely miss its deadline and shall be discarded from the network. In this queueing model, the probability of dropping packets, therefore, the loss of system reliability, is

$$p_{\text{block}} = \left(G p_0 \frac{m^m}{m!} \right) \rho^k = \Theta(\rho^k), \quad \rho = \lambda/(m\mu),$$

where G is a constant, p_0 is the probability that the system is empty, μ is the mean service time and λ is the Poisson arrival rate.

The relationship between the loss of system reliability (packet error rate), resource utilization and the packet arrival rate in an M/D/m/m queueing model with m = 10 is shown in Figure 4.7. The URLLC capacity for a given reliability/latency target is the admitted arrival rate in the M/M/m/k queue:

$$\lambda_{\text{URLLC}} = (1 - p_{\text{block}})\lambda \approx \lambda = \Theta\left(\sqrt[k]{p_{\text{block}}}\right) = \Theta\left(p_{\text{block}}^{(1/\text{latency})}\right)$$

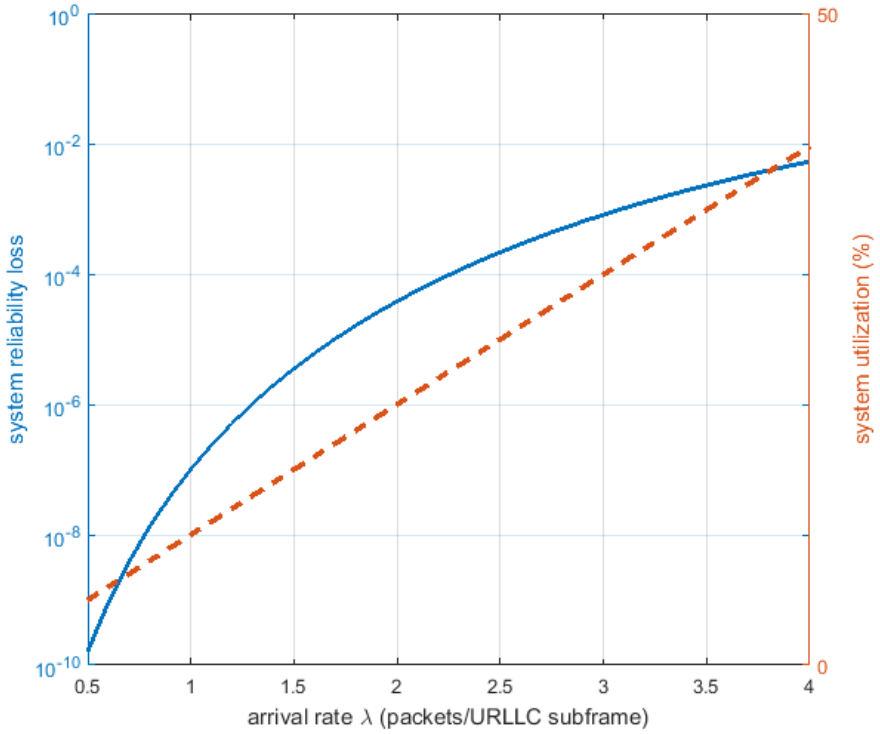


Figure 4.7. Relationship between Loss of System Reliability, Resource Utilization & Packet Arrival Rate for M/D/m/m Queueing Model with m = 10.

The tradeoff of available resource vs. supportable URLLC capacity is illustrated in Figure 4.8. It shows that as the number of allowed concurrent transmissions (therefore, the available bandwidth for URLLC) decreases, the system capacity decreases exponentially with low resource utilization. At the same time, when the amount of resource increases, system utilization improves due to better trunking efficiency among different users.

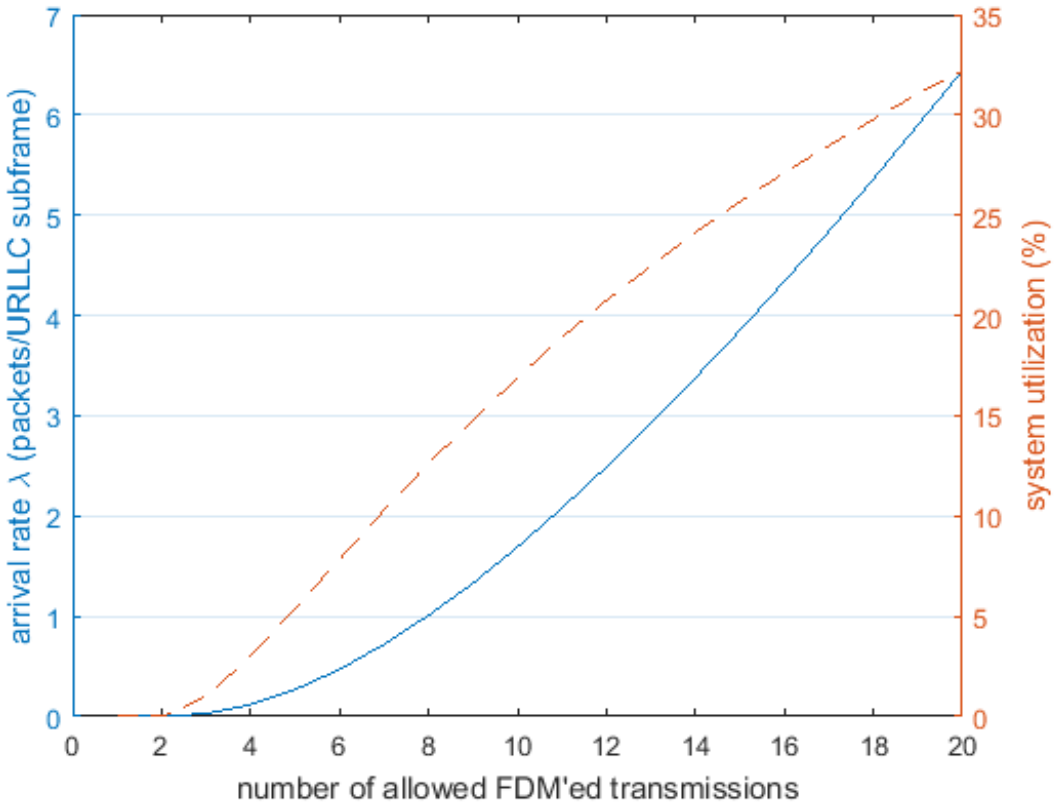


Figure 4.8. Maximally Supportable Poisson Arrival Rate and the Resource Utilization under the Reliability of p_loss = 1e-5 in an M/D/m/m Queueing Model.

The first "m" is the number of allowed simultaneous data transmissions and scales with the reserved bandwidth for URLLC.

Based on the queueing analysis, to achieve a target BLER with a low-latency bound, a large amount of instant system resource (bandwidth) is needed. Trunking efficiency is key for 5G NR URLLC and eMBB design. Therefore, it is clear that dynamic multiplexing plays a pivotal role in 5G URLLC design.

Dynamic Multiplexing Design:

The 5G system is designed to be able to support different types of services in the same network efficiently. When considering multiplexing of URLLC and other types of traffic such as eMBB, the performance of URLLC needs to be guaranteed, and efficiency should be considered at the same time.

From the gNB scheduling point of view, to avoid URLLC performance being affected by other traffic, one simple approach is to semi-statically partition the resources for URLLC and other traffic. However, this is inefficient from a resource utilization point of view. That's because the resources cannot be dynamically shared between URLLC and other traffic, and the trunking efficiency is lost, as shown in the queueing analysis previously. Therefore, some enhancements would be desirable to allow the dynamic resource sharing between the traffic while still being able to guarantee the URLLC performance. DL and UL pre-emption is one such type of enhancement.

With dynamic sharing of resources among different traffic, it may occur that when URLLC traffic arrives, there are no resources immediately available because they are already allocated to other traffic. Pre-emption allows the gNB to signal the UE(s) with ongoing transmission/reception to stop or temporarily pause the transmission/reception and use the freed-up resources to schedule URLLC traffic quickly.

On DL, preemption itself can be done by gNB scheduler implementation. But some signaling to the pre-empted UEs would be useful to mitigate the effect of preemption. This is illustrated in Figure 4.9, where there is an ongoing eMBB transmission with a long duration, the incoming URLLC packet cannot wait and preempts part of the eMBB transmissions. For an eMBB transmission with a large transport block size, using code block group-based HARQ retransmission can improve the efficiency by enabling the retransmission of only the code block groups in error (not the entire transport block).

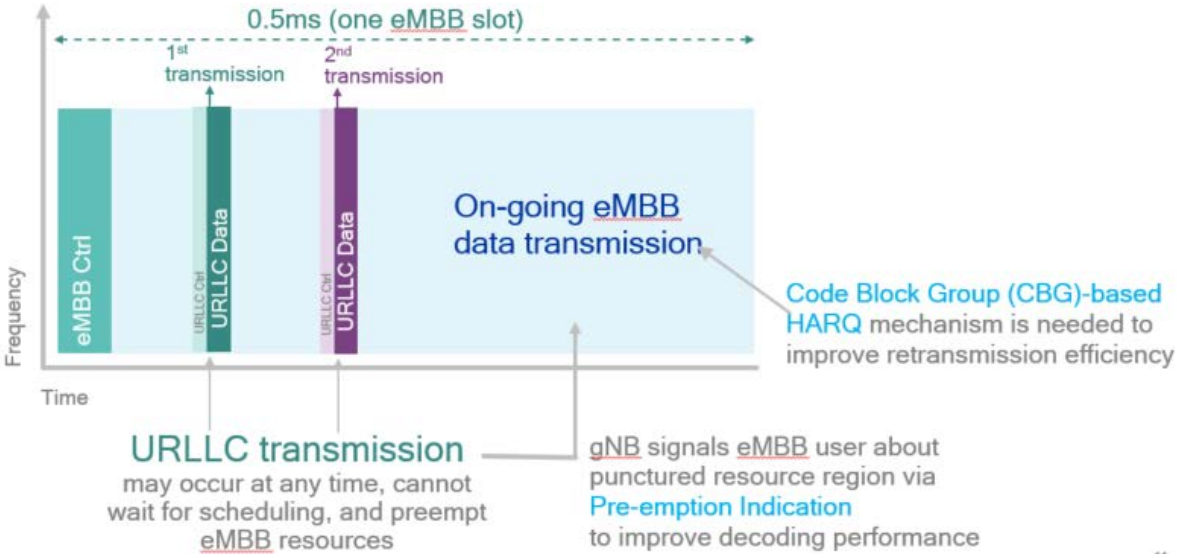


Figure 4.9. Illustration of DL Pre-emption.

Alternatively, or complementarily, the gNB can transmit a preemption indication to the eMBB UEs to inform about the punctured resources, as shown in Figure 4.10, so that the UE can take it into account in the decoding procedure to improve the decoding performance. The preemption indication in Figure 4.10 is transmitted at the beginning of the next slot, instead of being transmitted at the time of preemption, in order not to create additional burden on the eMBB UEs to monitor the control channel too frequently.

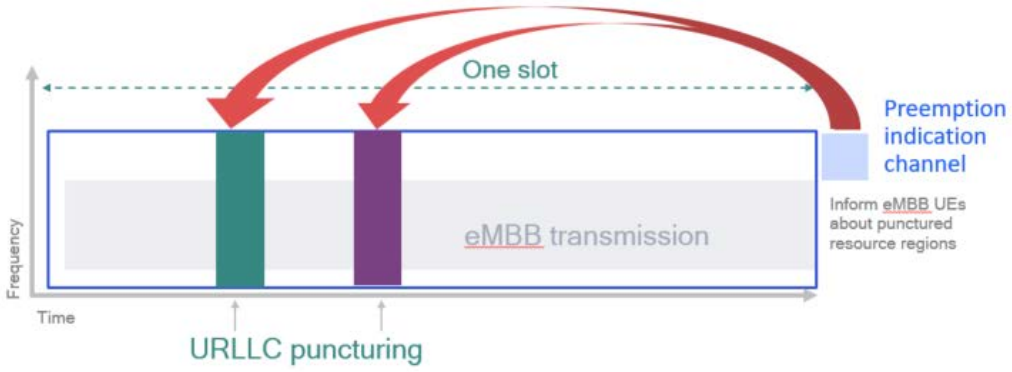


Figure 4.10. Illustration of Pre-emption Indication.

The concept of UL preemption is similar. The gNB can schedule URLLC traffic on the UL by using the resources that have already been allocated to other UEs. In this case, it is important for the gNB to signal to the other UE(s) to stop the ongoing transmission(s) at the time when the preemption decision is made. That way, the ongoing transmission(s) can be stopped in time and not interfere with the newly scheduled URLLC traffic.

This means that these eMBB UEs need to monitor the preemption indication frequently, such as in every mini-slot or every few mini-slots, while transmitting on UL. The concept of UL preemption is illustrated in Figure 4.11.

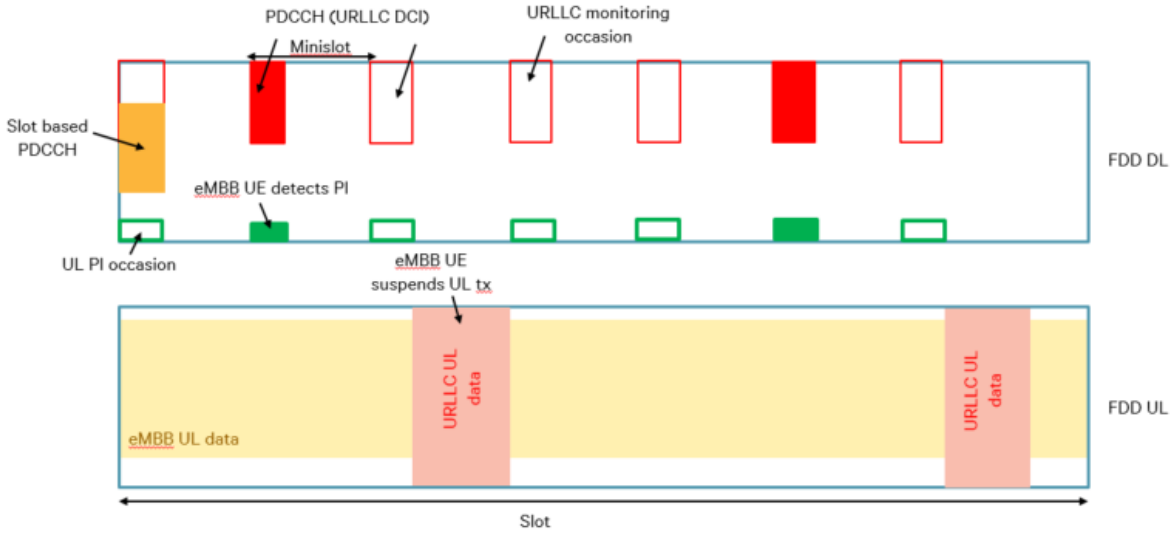


Figure 4.11. Illustration of UL Pre-emption.

For a UE supporting both URLLC and other types of traffic, intra-UE multiplexing and prioritization mechanisms would be necessary. The Medium Access Control (MAC) layer should provide the mechanism to prioritize and multiplex different types of the traffic.

In addition, traffic differentiation in the physical layer is also beneficial for channel prioritization and multiplexing, even though traditionally the physical layer is not aware of the service type. For example, HARQ-ACK for URLLC may be allowed to take higher priority than lower-priority data transmission, and the

grant-free UL transmission for URLLC may take higher priority than the grant-based UL transmission for lower-priority traffic. Shorter duration low-latency URLLC data transmission may take priority over long duration already scheduled data transmission.

5. UPPER LAYER DESIGN AND NETWORK ARCHITECTURE FOR 5G URLLC

3GPP is enabling several system and core network enhancements in Releases 15 and 16 to support URLLC requirements for latency, jitter and reliability aspects in the 5G system. Some of these enhancements include the following:

- Flow-based QoS mode with support for reflective QoS and standardized 5QIs
- Support for MEC, which enables efficient service delivery through reduced end-to-end latency and load on the transport network. There are several MEC enablers in the 5G system:
    - User Plane Function (UPF) selection to route the user traffic to the local data network. The 5G core network selects the traffic to be routed and steered to the applications in the local data network
    - Multiple data paths with redundant transmission in the user plane to ensure reliable delivery of application data. This helps in transmission of data with reliability higher than with single user plane tunnel of N3 and N9 and Network Functions (NFs) in the user plane path
    - Session and service continuity to enable UE and application mobility. Multi-homed IPv6 Protocol Data Unit (PDU) sessions to support make-before-break service continuity to support Session and Service Continuity (SSC) mode 3 and concurrent access to local services and internet with different IPv6 prefixes
    - Application function influence on UPF (re)selection and traffic routing via Point Coordination Function (PCF) or Network Element Function (NEF)
    - Network capability exposure with 5G core network and application function providing information to each other via NEF
- Support for UE and network-controlled, always-on PDU sessions to enable low-latency transmissions.
- Enabling a new Radio Resource Control (RRC) state, RRC_INACTIVE, which allows a UE in connected state when not transmitting or receiving data to achieve power efficiency comparable to that of an idle UE.

Some of these aspects are described in more detail in the following sections.

5.1 QOS

One of the key requirements for URLLC services is the stringent end-to-end QoS goals that include low latency and high reliability. Meeting these is challenging for 5G systems because various factors affect the end-to-end QoS performance, such as coverage, judicious use of system and network resources (UPF/RAN/UE) and transport network design.

The QoS differentiation within a PDU session is defined by QoS Flow, which is identified by a QoS Flow ID (QFI). Traffic associated with the same QFI receives the same QoS forwarding treatment. QFI is used as U-plane marking on N3/N9 interfaces and is unique within a PDU session. Each QoS Flow is associated with a set of QoS characteristics (packet delay budget, packet error rate, priority level) and Address Resolution Protocol (ARP) value. Guaranteed Bit Rate (GBR) QoS Flows, in addition, are associated with Guaranteed Flow Bitrate (GFBR), Max Flow Bit rate (MFBR) and Notification control. A standardized set of 5G QoS Indicators (5QIs) are defined and point to a set of QoS characteristics. The 5QI is similar to the

QoS Class Identifier (QCI) in Evolved Packet System (EPS). A new “resource type,” Delay Critical GBR, is also defined. There are two ways to control QoS flows:

1. For non-GBR QoS flows with standardized 5QIs, the standardized 5QI value can be used as QFI, and a default ARP value is assumed. In this case, no additional N2 signaling is required at the time traffic for the corresponding QoS flows start

2. For GBR and non-GBR QoS flows, all the necessary QoS parameters corresponding to a QFI are sent as a QoS profile to the RAN either at the PDU session establishment or during the QoS flow establishment or modification

In 5G, there is no end-to-end bearer as in the Evolved Packet System (EPS). The per-packet QoS marking is carried out in the N3 encapsulation header in both the UL and DL to differentiate QoS Flows. The RAN continues to use Digital Radio Broadcasting (DRBs). All packets in a DRB receive the same QoS treatment. A new aspect in 5G is the flexibility in RAN to bind QoS flows onto DRBs. The per-packet QoS marking is carried out in the radio header Service Data Adaption Protocol (SDAP). The UE determines UL data QoS binding either via Reflective QoS (RQoS) based on DL data QoS marking, or via explicit QoS signaling. 5GS enables several standardized 5QIs to support URLLC-based applications.

RQoS is used to minimize the need for control-plane signalling (N1). RQoS is achieved by creating a *derived* QoS rule in the UE based on the received downlink traffic. The UE inspects the IP 5-tuple in the downlink packet, creates a “mirror” packet filter and associates the QoS of the downlink packet to uplink packet. The mirror packet filter and the associated QoS constitute the derived QoS rule.

The UE uses the derived QoS rule to bind corresponding uplink packets on the same QoS flow. The UE indicates support for RQoS upon PDU Session establishment and also indicates the number of supported packet filters for derived QoS rules. Reflective QoS Attribute (RQA) is an optional parameter associated with a QoS Flow indicating that certain traffic (not necessarily all) carried on this QoS Flow is subject to Reflective QoS. Only when the RQA is signalled for a QoS Flow, the RAN enables the transfer of the Reflective QoS Indicator (RQI) (therefore, adds the SDAP header).

When the 5G core network determines that RQoS is to be used for a specific Service Data Fow (SDF), it sets the RQI in each packet of this SDF. RQI is carried in the packet encapsulation header on N3 and over-the-air interface in the SDAP[15] header.[16]

5.2 MOBILE EDGE COMPUTING (MEC)

5G systems are driven by the use of software and IT virtualization technology toward the development of telecommunication infrastructure, functions and applications. Edge computing enables operator and third-party services to be hosted close to the UE's access point of attachment. MEC enables application developers and content providers to use cloud computing capabilities and IT services at the network edge.

This results in users having greater proximity to contextual information with low latency and real-time access to their data, as well as high bandwidth and location awareness. This also results in efficient utilization of radio and network resources by reducing the load on the transport network.

[15] 3GPP TS 23.501: *System Architecture for the 5G System; Stage 2.*
[16] 3GPP TS 24.501: *Access-Stratum (NAS) protocol for 5G System (5GS); Stage 3.*

MEC uses a virtualization platform for running applications at the mobile network edge. This Network Functions Virtualization (NFV) infrastructure may be dedicated to MEC or shared with other network functions or applications. Where possible, MEC uses the NFV infrastructure management entity.

The current 3GPP systems allow operators to expose some network capabilities such as QoS policy to third-party Internet Service Providers/Internet Content Providers (ISPs/ICPs). With 5G, new network capabilities are being exposed to the third party in new ways where a dedicated network slice is used for a specific use case. This allows a third party to manage a trusted third-party application in a Service Hosting Environment to improve user experience and efficiently utilize backhaul and application resources.

Several mechanisms are incorporated for minimizing user plane resources utilization, including in-network caching and putting an application in a Service Hosting Environment closer to the end user. These optimization efforts help lower latency and increase reliability.

5.3 HIGH RELIABILITY BY REDUNDANT TRANSMISSION IN USER PLANE

3GPP provided the basic support for URLLC by introducing TTI structures and L2/L3 features. The following L2 functionality is defined:

Logical Channel Prioritization (LCP) Restrictions: This helps reserve the services required for URLLC aspects. By using LCP restrictions in MAC, the mapping of a logical channel can be restricted and reserved to a subset of resources such as cells, numerologies (subcarrier spacing), PUSCH transmission durations and control logical channel to utilize the resources allocated for the uplink direction.

Packet Duplication at the RAN Layer: Using packet duplication at the RAN layer allows the packet to be transmitted with two independent radio paths in the air interface. This is achieved by introducing a radio bearer mapping to two Radio Link Control (RLC) entities and two logical channels. The packet duplication mechanism further increases reliability and reduces the latency that would serve URLLC services. The packet duplication mechanism is applicable only for PDCP data PDUs but not for PDCP control PDUs (therefore, they are always sent on primary path). The original packets and duplicated packets are transmitted on different carriers, respectively. Therefore, one of the paths is expected to succeed in terms of reaching the receiver side.

5.4 EDGE COMPUTING ENABLERS

The 5G systems enable edge computing through a variety of enablers,[17] some of which are the following:

Flexible placement of UPF: In the 5G system, the N3 tunnel is provided with PDU session granularity. In EPS, all Packet Data Network (PDN) connections to the same PDN are terminated on the same PDN Gateway (PGW). But in the 5G system, multiple PDU sessions to the same data network need not be terminated at the same User Plan Function (UPF) (N6). The number of UPFs for a PDU session is not imposed by the specification. For a UE with multiple PDU sessions, there is no need for a mandatory "convergence point" similar to the Serving Gateway (SGW). Thus, beyond the Access Network (AN), the user plane paths of different PDU Sessions (to the same or to different Data Network Name (DNN) belonging to the same UE may be completely disjointed. This also implies that for idle mode UEs, there can be a distinct buffering node per PDU Session. The 5G core network selects a UPF close to the UE and executes the traffic steering from the UPF to the local data network via a N6 interface. This may be based

[17] 3GPP TS 23.501: *System Architecture for the 5G System; Stage 2.*

on the UE's subscription data, UE location and information from application function, policy or other related traffic rules. For the low-latency requirements of URLLC, the UPF can be collocated with the RAN.

User plane (re)selection: The 5G core network (re)selects UPF to route the user traffic to the local data network. The UPF selection and re-selection is performed by the SMF by considering UPF deployment scenarios such as a centrally located UPF and distributed UPF located close to or at the access network site. The UPF selection also enables deploying the UPF with different capabilities.

The SMF may consider different parameters and information during UPF selection and re-selection. These include the UPF's dynamic load, UPF location, UE location information, UPF capability and the functionality required for the particular UE session, Data Network Name (DNN), PDU session type, Session and Service Continuity (SSC) mode selected for the PDU Session, UE subscription profile in Unified Data Management (UDM) and local operator policies.

Local Routing and Traffic Steering: The 5G core network selects the traffic to be routed to the applications in the local data network. This includes the use of a single PDU session with multiple PDU session anchor(s) (UL CL / IP v6 multi-homing). In order to support selective traffic routing to the Data Node (DN), or to support SSC mode 3 as described below, the SMF may control the data path of a PDU session so that the PDU session may simultaneously correspond to multiple N6 interfaces. The UPF that terminates each of these interfaces is said to support PDU session Anchor functionality. Each PDU session anchor supporting a PDU session provides a different access route to the same DN.

Multi-homed PDU Session with Uplink Classifier: The UP function contains Uplink Classifier (UL CL) functionality that allows steering of local traffic to local services (for example, local CDN server) and the rest of the traffic towards central services. The UL CL applies filtering rules (for example, to examine the destination IP address of IP packets sent by the UE) and determines how the packet should be routed. The UL CL also supports connectivity to a local data network (for example, tunneling), as well as charging, lawful intercept and bitrate enforcement. The UL CL is controlled by SMF over N4.

An UL CL for a given PDU session may be inserted or removed by the SMF on the fly and is transparent to the UE. The UE uses the same IP address for traffic bound to either a local DN (where the edge computing server is located) or a remote DN. As Figure 5.1 illustrates, the N9 forwarding tunnel created between the source and target UL CLs would help achieve low latency in terms of forwarding data packets locally without impacting UE signaling in mobility.

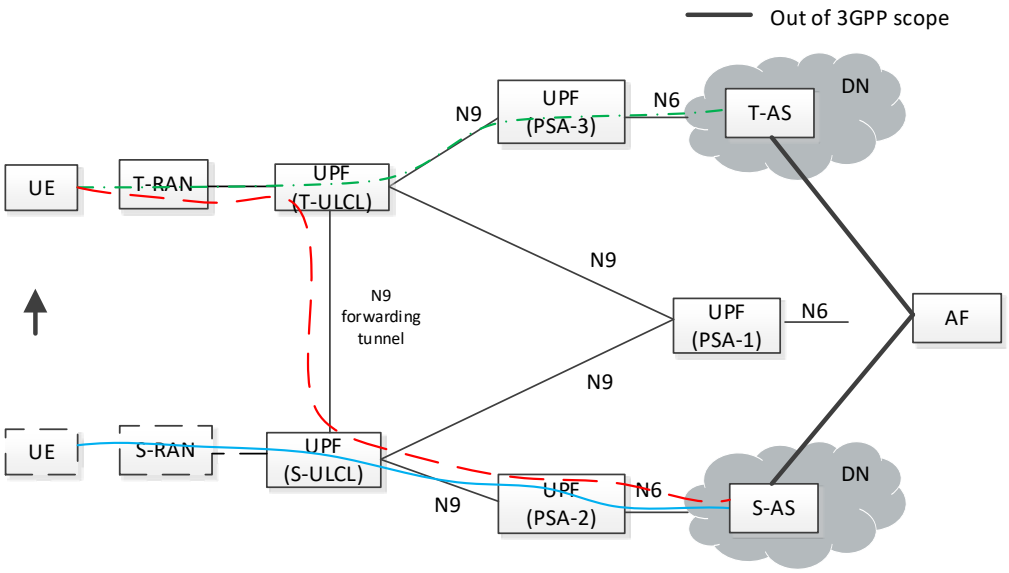


Figure 5.1. Sample Network Architecture with N9 IF Forwarding Tunnel between Source and Destination ULCL.

Session and service continuity to enable UE and application mobility: In a 5G system, the PDU session can be associated with multiple IPv6 prefixes. The UE uses different IPv6 addresses to access different DNs. A "common" UPF, referred to as a branching point, steers the UL traffic toward one or the other IP anchor based on the packet's source prefix. The Internet Engineering Task Force (IETF) RFC 4191 is used to configure rules into the UE to influence the source address selection.

This corresponds to Scenario 1 defined in IETF RFC 7157 "IPv6 Multi-homing without Network Address Translation." This can be used to support make-before-break service continuity to support SSC mode 3. It also can be used to support concurrent access to a local service (for example, local CDN server) and the internet with a different IPv6 prefix.

The 5G system supports the following Session and Service Continuity (SSC) modes, which Figure 5.2 illustrates:

- **SSC mode 1**: The same UPF is maintained regardless of the network access technology a UE is using
- **SSC mode 2**: The same UPF is only maintained across a subset (therefore, one or more, but not all) of the access network attachment points (for example, cells and RATs), referred to as the serving area of the UPF. When the UE leaves a UPF's serving area, it will be served by a different UPF suitable for the UE's new point of attachment to the network
- **SSC mode 3**: In this mode, the network allows the establishment of UE connectivity via a new UPF to the same data network before connectivity between the UE and the previous UPF is terminated. When trigger conditions apply, the network selects a target UPF suitable for the UE's new point of attachment to the network. While both UPFs are active, the UE either actively rebinds applications from the previous to the new address/prefix, or alternatively, the UE waits for flows bound to the previous address/prefix to end

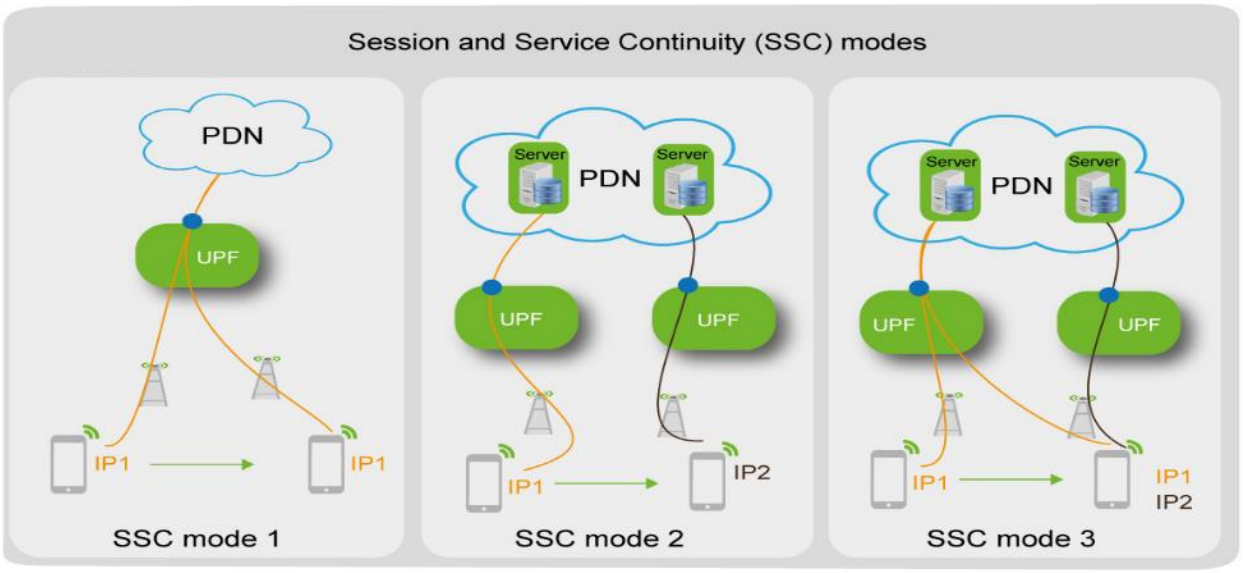


Figure 5.2. User Plane Architecture for the Uplink Classifier.

Application Function (AF) influence on UPF (re)selection and traffic routing: The logical AF in the 3GPP architecture may correspond to an MEC application server. An AF may send requests to influence SMF routing decisions for traffic of the PDU session, for event subscription or both. The AF may influence UPF (re)selection and allow routing user traffic to a local access to a data network (identified by a Dynamic Network Access Identifier (DNAI)). The AF may issue requests on behalf of applications not owned by the Public Land Mobile Network (PLMN) serving the UE. If the operator does not allow an AF to access the network directly, the AF shall use the NEF to interact with the 5GC.

The AF influence doesn't apply to the home-routed roaming case. The PCF does not apply AF requests under such cases. The AF may make a request for a group of UEs by providing Group Identifiers. In such case, the AF request will influence multiple PDU sessions possibly served by multiple SMFs and PCFs. The group identifiers provided by the AF are mapped to International Mobile Subscriber Identity (IMSI)-Group identifiers by PCF at PDU session setup.

The AF may subscribe to notifications about UP path management events. Two examples are when the request becomes active or inactive, or when a change of DNAI occurs for the PDU session. When notifications about UP path management events are sent to the AF via the NEF, if required, the NEF maps the UE identify information (for example, SUPI) to the General-Purpose Serial Interface (GPSI) and the AF transaction internal identifier to the AF transaction identifier before sending the notifications to the AF.

Network Capability Exposure: The 5G core network and Application Function provide information to each other via the NEF, which supports external exposure of capabilities of network functions. External exposure can be categorized as monitoring capability, provisioning capability and policy/charging capability. The monitoring capability is for monitoring of specific events for UEs and making such monitoring events information available for external exposure via the NEF. The provisioning capability is for allowing an external party to provision information that the UE can use. The policy/charging capability is for handling QoS and charging policy for the UE based on the external party's request.

Monitoring capability is comprised of means that allow the identification of the 5G network function suitable for configuring the specific monitoring events, detecting the monitoring event and reporting the monitoring

event to the authorized external party. Monitoring capability can be used for exposing the UE's mobility management context such as UE location, reachability, roaming status and loss of connectivity. Provisioning capability allows an external party to provision the foreseen UE behavioral information to 5G NF via the NEF.

Policy/charging capability is comprised of means that allow the request for session and charging policy, enforces QoS policy and applies accounting functionality. It can be used for specific QoS/priority handling for the session of the UE and for setting the applicable charging party or charging rate.

Enhancing topology of Session Management Function (SMF) and User Plane Function (UPF): In 3GPP Release 16, there is an ongoing study on enhancing the topology of SMF and UPF[18]. One goal of the study is to support deployments where a single SMF is not able to control the UPFs throughout the PLMN. In Release 15, when the UE has moved outside of the service area of the old UPF, the Inter-User Plane Function (I-UPF) may need to be inserted, relocated or removed.

This causes delay to the UL data and especially to the DL data because the DL data is buffered in the old UPF, and the data needs to be retrieved from there to the new target UPF. This delay to set up the user plane when the UE returns from idle may be critical (for example, for Low-Latency Communications (LLC) and IP Multimedia Subsystem (IMS) voice services). The outcome of the Release 16 study may be that, in this case, the SMF controlling this I-UPF also may need to be inserted, relocated or removed.

To solve this issue, 3GPP is considering using a deployment-based solution. The UPF that supports the latency-sensitive services can be planed properly in order to avoid the I-SMF/I-UPF change during service request procedure (therefore, to align the UPF service area same as the Access Mobility Function (AMF) service area.) Thus, when the UE does not change the serving AMF, no matter how it allocates the Tracking Area (TA) list to the UE, the SMF/UPF will not be changed when the mobility procedure is triggered.

Another alternative being considered at 3GPP is to update the Radio Access (RA) based on UPF service area of the PDU sessions that are latency sensitive. When a PDU session is to be established, the SMF determines whether the PDU session is latency sensitive (for example, based on indication received from UE, PCF or AF, or based on local policy). If the SMF determines a PDU session is latency sensitive, the SMF subscribes the UE mobility and provides the AMF with the UPF service area information. The AMF recalculates the RA based on the UPF service area to ensure that the UE will not change the UPF when the service request procedure is triggered. If the recalculated RA is different than the RA that has been provided to UE, the AMF triggers UE configuration update procedure to provide the new RA to UE.

5.5 ALWAYS-ON PDU SESSIONS

To support low-latency communication, the 5G system supports always-on PDU sessions, which may be UE controlled or network controlled. The network-controlled always-on PDU session mechanism can co-exist with the UE-controlled always-on PDU session mechanism.

An always-on PDU session is used when user plane resources are activated during every transition from idle mode to connected mode. A UE requests a PDU session to be established as an always-on PDU session based on indication from upper layers. Then the network decides whether a PDU session is established as an always-on PDU session.

The UE requests a new PDU session as an always-on PDU session by including the always-on PDU session requested IE in the PDU session establishment request message. Based on local policies or

[18] 3GPP TR 23.726: *Study on Enhancing Topology of SMF and UPF in 5G Networks*.

configurations in the SMF, the SMF determines whether the requested PDU session needs to be established as an always-on PDU session. The SMF indicates this to the UE in the PDU session establishment accept message. The activation of resources during transition to connected mode enables the UE to conduct user data transmission with low latency and jitter.[19]

5.6 SUPPORTING LOW LATENCY WITH A NEW STATE RRC_INACTIVE

The battery-efficient RRC_INACTIVE state introduced in Release 15 addresses URLLC's low-latency requirement to some extent. The RAN-level UE tacking is based on the RAN Notification Areas (RNAs) which are similar to the Universal Terrestrial Radio Access Network (UTRAN) Registration Areas (URAs) in UTRAN. RNA is UE specific and consists of one or more cells and is smaller than the tracking area. Figure 5.3 illustrates the transitions between different states INACTIVE, IDLE and CONNECTED in Core Network (CN) and RAN for EPS/E-UTRAN and 5GS/NR.

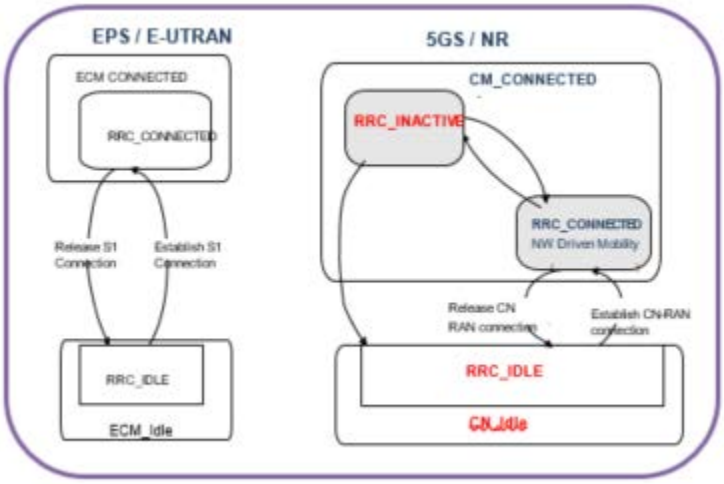


Figure 5.3. CN and RAN States for EPS/E-UTRAN and 5GS/NR.

One potential use case of low-latency mobility is wherein the UE moves from RRC_INACTIVE state to RRC_CONNECTED state and connects to another Next Generation Radio Access Network (NG-RAN) node in which the UE context is NOT available but mobile originated/mobile terminated Mobile Originating/ Mobile Terminating (MO/MT) data need to be delivered with ultra-low latency. In this situation, the potential solution is to transition to CM_CONNECTED from RRC_INACTIVE with the following enhancements so as to be able to transmit user data with low latency:

- RAN uses the QoS information to detect if URLLC is requested
- When the UE sends the RRC resume request to a NG-RAN node that does not have the UE context, then the RAN may decide that the anchor NG-RAN node will not do a path switch of N3 to the new NG-RAN node immediately
- Mobile Switching Center (MSC) depicts how ultra-low latency user data can be delivered for this use case

[19] 3GPP TS 24.501: *Access-Stratum (NAS) protocol for 5G System (5GS); Stage 3.*

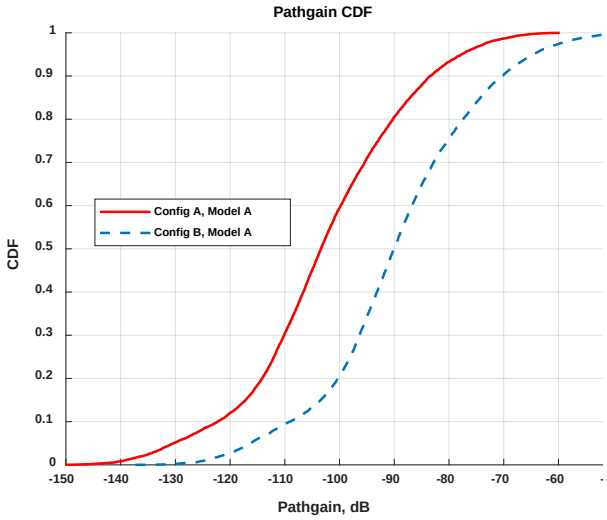
6. PERFORMANCE, POSSIBLE IMPROVEMENTS AND CHALLENGES

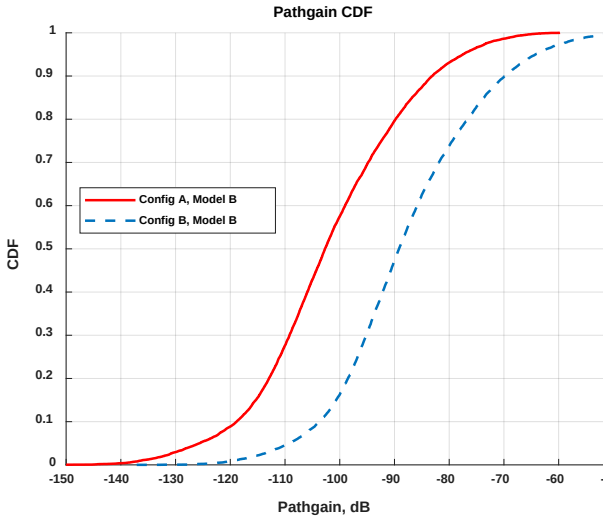
This section presents some preliminary performance results on reliability evaluation using IMT-2020 methodology.[20] The IMT-2020 methodology for reliability is defined by ITU and specified by 3GPP.

Section 6.1 outlines the assumptions. Section 6.2 discusses DL reliability and how it's evaluated at the system level and then by link-level modeling of PDCCH and PDSCH.

The system-level assumptions used the for the performance evaluation are summarized in the appendix in Table B.1. Essentially, two carrier frequencies are considered for performance analysis: 4 GHz (Config A) and 700 MHz (Config B).

The following results depicted in Figures 6.1 and 6.2 are the coupling loss/path gain statistics for Channel Model A and Model B for DL and UL setups.





[20] *Reliability Evaluation for URLLC*, Intel Corp. 3GPP TSG RAN WG1 Meeting #94.

Figure 6.1. UE Useful Path Gain Statistics for DL Setup, IMT-2020 URLLC UMa Config A/B in Channel Model A/B.

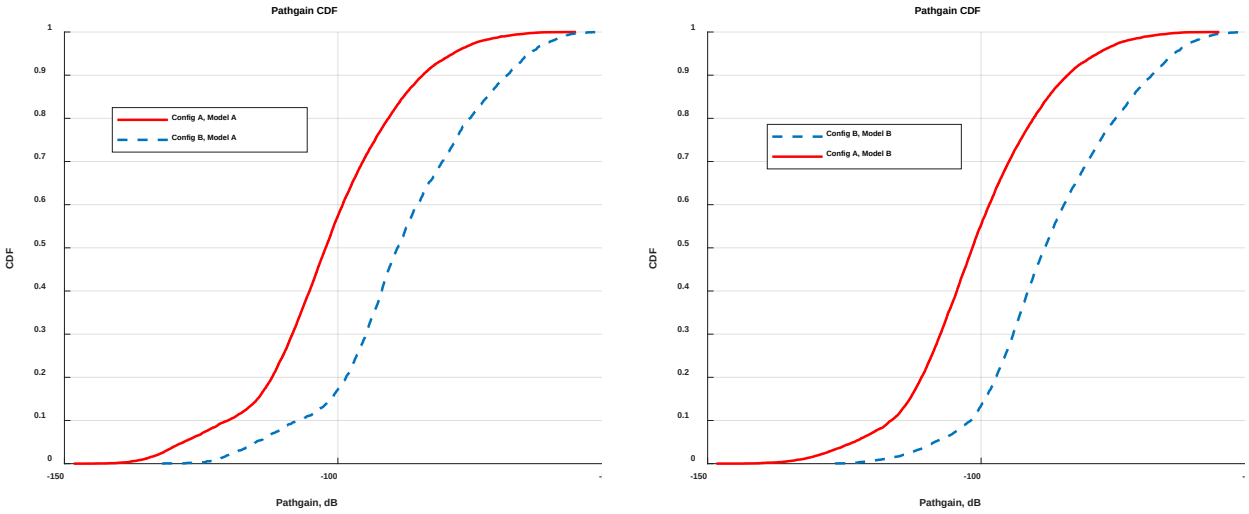


Figure 6.2. UE Useful Path Gain Statistics for UL Setup, IMT-2020 URLLC UMa Config A/B in Channel Model A/B.

Link-level evaluation assumptions are summarized in Table 6.1.

Table 6.1. Common Link-Level Evaluation Parameters.

Parameters	Values
Channel model	TDL-C, 300 ns delay spread
UE speed	3 km/h
System BW	40 MHz (106 PRB)
Numerology	30 kHz
UE antenna	Config A: 4 RX, 1 TX, low correlation Config B: 2 RX, 1 TX, low correlation
BS antenna	8 RX, 2 TX, low correlation
TX diversity	PDCCH and PDSCH - based on precoder cycling
Channel estimation	Practical
PDCCH	1 symbol CORESET, AL 16
DCI payload	40 bit + 24 bit CRC

PDSCH	6 symbols after PDCCH Mapping type B 65 PRB
PDSCH DMRS	Type 1 No additional DMRS
PDSCH MCS	MCS#0 of low SE table (QPSK, CR = 30/1024) TBS = 256 bit
PUSCH	7 symbols Mapping type B 24 PRB, distributed over two parts
PUSCH DMRS	Type 1 No additional DMRS 3 dB power boosting
PUSCH MCS	MCS#4 of low SE table (QPSK, CR = 78/1024) TBS = 256 bit

6.1 DL RELIABILITY ANALYSIS

Based on the evaluation assumptions in Table 6.1, the derived DL geometry for both cases of carrier frequency are shown in Figure 6.3.

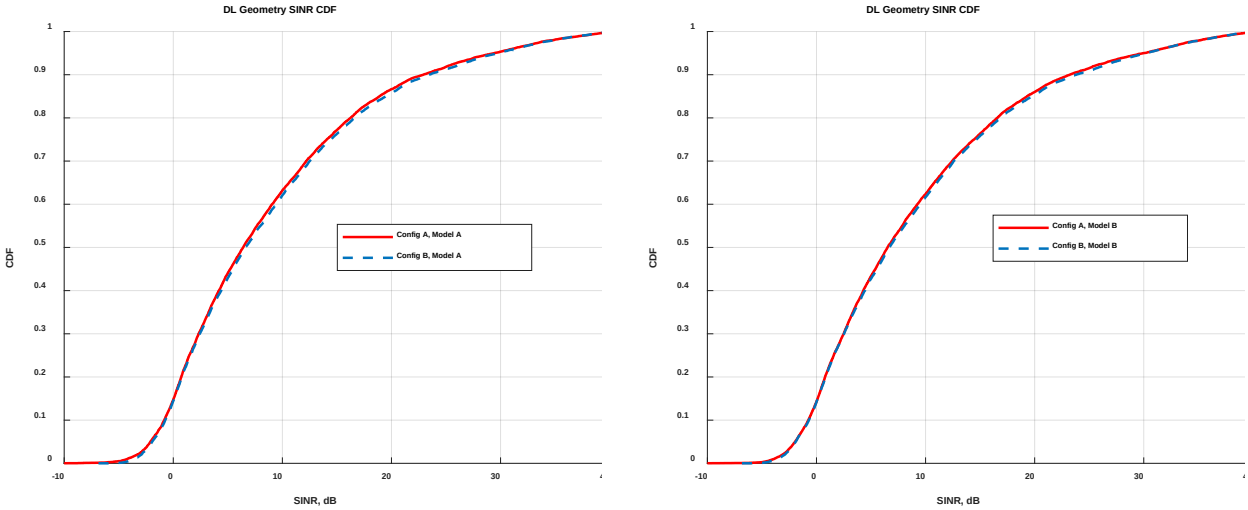


Figure 6.3. DL Geometry SINR, IMT-2020 URLLC UMa Config. A/B in Channel Model A/B.

From the geometry presented above, the 5 percent CDF point provides target SINR as summarized in Table 6. below:

Table 6.2. DL SINR CDF.

	Configuration A		Configuration B	
	Model A	Model B	Model A	Model B
5percent SINR, dB	-2.091	-1.965	-1.936	-1.902

The results show there is no noticeable difference in results between the A and B configurations as the overall geometry is interference limited.

It should be noted that this is a single-port SINR, wherein multi-antenna gains are assumed to be accounted in link level evaluations.

For the link-level part, the worst of the two values (therefore, -2.091 dB and -1.936 for Config A and B respectively) is considered further for analysis.

6.2 LINK-LEVEL RESULTS

The performance analysis is conducted for a slot configuration with a 30 kHz subcarrier spacing in 40 MHz system bandwidth.

The required 1 ms latency budget includes frame alignment delay, Transmit (TX) delay and UE processing delay for single-shot transmission. Figure 6.4 depicts a sketch of DL transmission structure considered for the reliability performance evaluation.

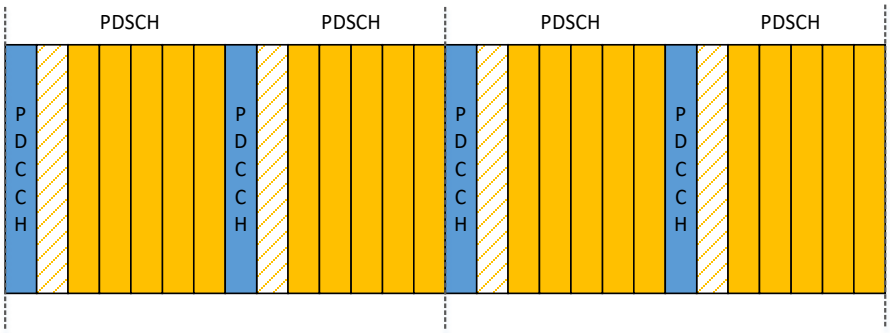


Figure 6.4. Sketch of DL Transmission Structure for Reliability Evaluation.

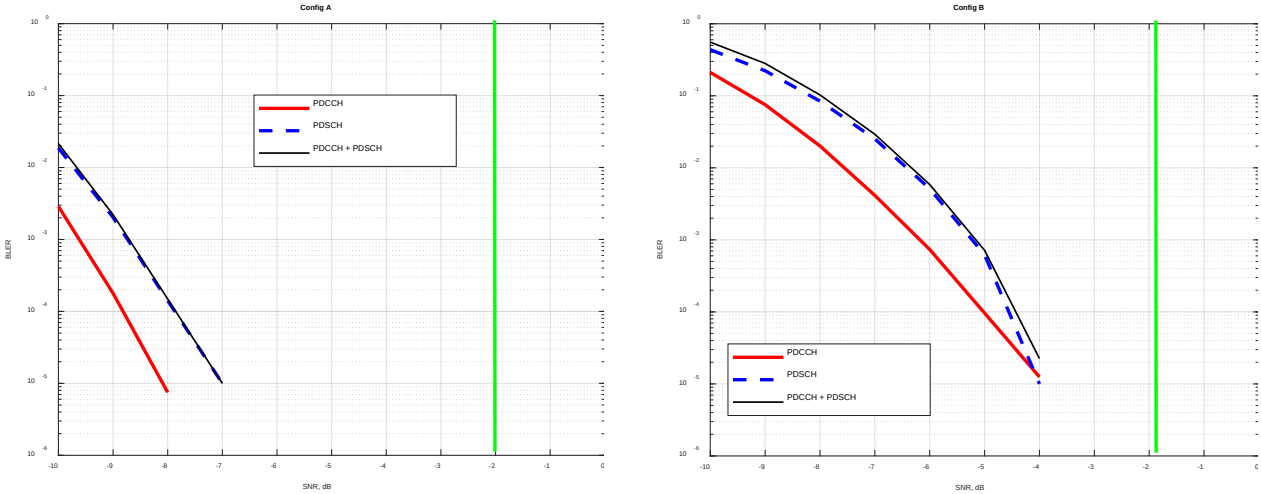


Figure 6.5. BLER vs SNR for PDCCH and PDSCH for Config A (4 RX) and Config B (2 RX).

The results shown in Figure 6.5 indicate that single-shot performance successfully achieves the requirements in cases of both A and B configurations. Configuration A obviously has superior performance due to the high carrier frequency where 4 Receive (RX) antennas are available at a UE.

Further, the total reliability presented as joint BLER is dominated by PDSCH performance and therefore not very different from single-shot PDSCH curve.

The results summarized in Table 6.3 show that the derived 5-percentile point of DL SINR CDF indicates the total DL reliability achieved within 1 ms is better than 99.999 percent for considered test cases.

Table 6.3. Summary of DL Results for Configurations A & B.

	Configuration A		Configuration B	
	Model A	Model B	Model A	Model B
Target SINR, dB	-2.091	-1.965	-1.936	-1.902
SINR at 99.999percent reliability, dB	-7	-7	-3.8	-3.8

The performance results presented on reliability evaluation of Rel-15 NR URLLC based on IMT-2020 methodology for DL show that the requirements are fulfilled for single-shot transmission of PDCCH+PDSCH for both Configuration A and B in both channel models A and B.

6.3 UL RELIABILITY ANALYSIS

As for the UL reliability analysis, UL geometry for both cases of carrier frequency is shown in Figure.6.6. The geometry was derived using round-robin scheduling so that there is no optimization of inter-cell interference. Each UE has 48 Physical Resource Block (PRB) in 15 kHz Sub Carrier Spacing (SCS) (corresponding to 24 PRB in LLS part) and is power controlled according to the assumptions in Table 6.1.

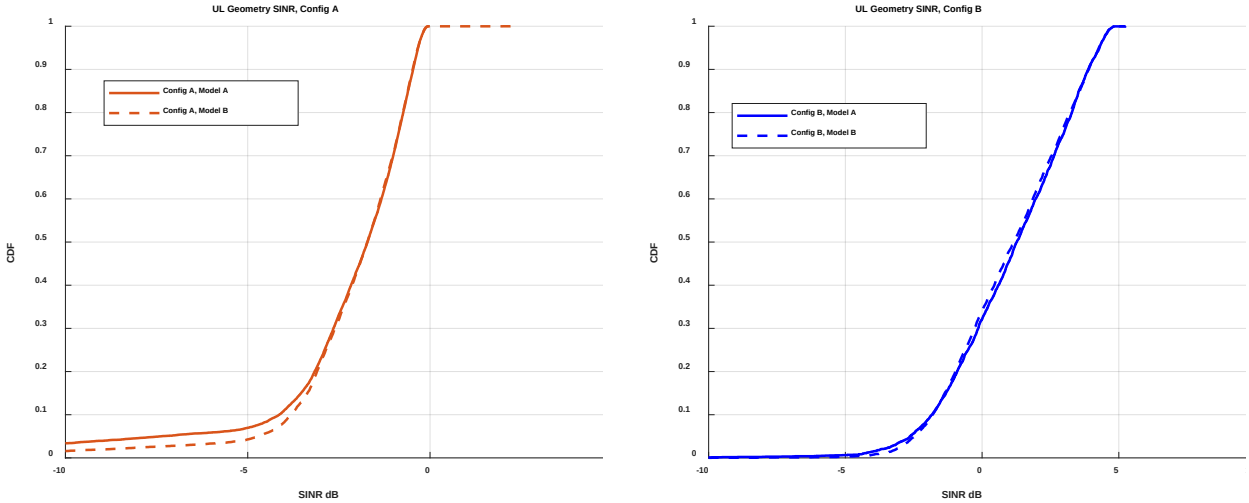


Figure 6.6. UL Geometry SINR, IMT-2020 URLLC UMa Config. A/B in Channel Model A/B.

From the geometry presented above, the 5 percent cumulative distribution function (CDF) point provides target SINR summarized in Table 6.4:

Table 6.4. UL SINR CDF.

	Configuration A		Configuration B	
	Model A	Model B	Model A	Model B
5percent SINR, dB	-7.3769	-4.71	-2.6378	-2.4747

As for the UL Link Level Results, it is assumed that 30 kHz subcarrier spacing is taken in 40 MHz system bandwidth. The required 1 ms latency budget needs to at least account for frame alignment delay, TX delay and gNB processing delay for single-shot transmission.

Thus, dividing two slots of 30 kHz onto four parts of 7 symbols from a UE perspective may result in maximum scheduling at most 7-symbol one-shot PUSCH in configured grant manner.

Assuming the 7-symbol transmission unit, wherein 1 symbol is allocated to DeModulation Reference Signal (DMRS), the resource allocation to carry TBS = 256 bit using MCS#4 from the low SE 64 Quadrature Amplitude Modulation (QAM) table requires around 24 Physical Resource Block (PRBs), which are taken for further Low Latency Socket (LLS) evaluation.

The bandwidth and MCS were selected to have reasonable percentage of power limited UEs, which may achieve good Signal-to-Noise Ratio (SNR) for the given bandwidth.

The results shown in Figure 6.7 indicate that single-shot performance for UL successfully achieves the requirements in the cases of both Configuration A and B where 1e-5 BLER is achieved at ~ - 8.6 dB.

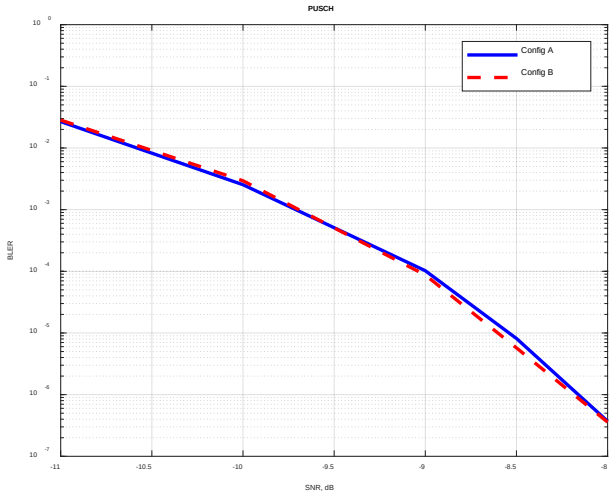


Figure 6.7. BLER vs SNR for PUSCH for Config A and Config B.

Results are summarized in Table 6.5.

Table 6.5. Summary of UL Results for Configurations A & B.

	Configuration A		Configuration B	
	Model A	Model B	Model A	Model B
Target SINR, dB	-7.3769	-4.71	-2.6378	-2.4747
SINR at 99.999percent reliability, dB	-8.54	-8.54	-8.6	-8.6

6.4 SYSTEM-LEVEL RESULTS

To illustrate the queueing effect and the URLLC QoS requirements, Figure 6.8 provides a simplified queueing model that describes the PHY/MAC layer behavior when a gNB schedules a user on the downlink. In particular, user data packets arrived at the gNB are buffered at the first-transmission queue of the user and await to be scheduled for their first HARQ transmissions. If the first HARQ transmission of a data packet fails, the packet is available at the second-transmission queue for future retransmission opportunities after an RTT.

Whenever a packet that is buffered at the gNB misses its deadline, the packet is of no use and dropped by the system, resulting in the loss of reliability for this user. Additionally, a HARQ failure for a data packet may be declared by the gNB if the packet cannot be decoded after n HARQ transmissions, resulting in the loss of reliability as well. At every scheduling instant, the gNB allocates time and frequency resources to new transmissions and retransmissions of the buffered packets, in order to satisfy the QoS requirements of all users.

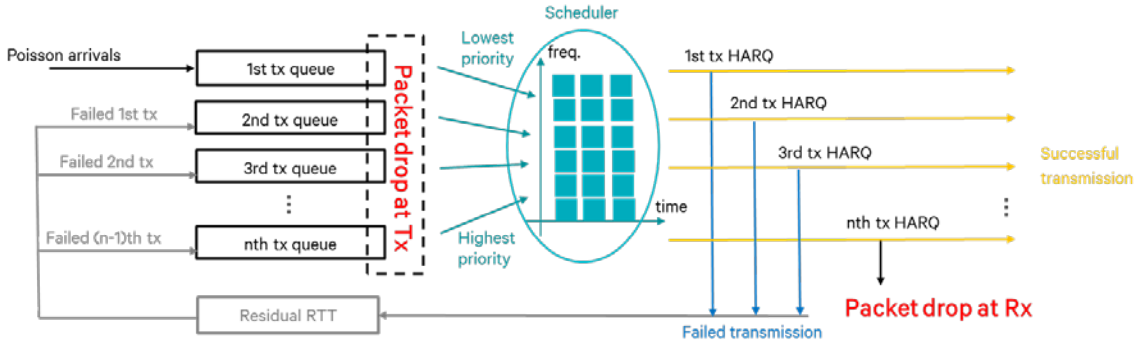


Figure 6.8. Queueing Model that Describes the PHY/MAC Layer Behavior at the gNB for URLLC.

The following discussion uses queueing models to get qualitative insight into the performance scaling between reliability, delay and outage capacity, and to justify the observations from the system-level simulation results. The finite buffer aspect of the queueing model can capture the fact that packets are dropped from the transmitter side if queueing delay is longer than the latency requirement.

Some qualitative URLLC capacity results are shown in Figure 6.9, which shows the following:

- As latency requirements tighten, URLLC channel capacity quickly drops to zero
- High-reliability requirement also reduces URLLC capacity
- URLLC capacity grows super linearly as a function of available frequency bandwidth resources

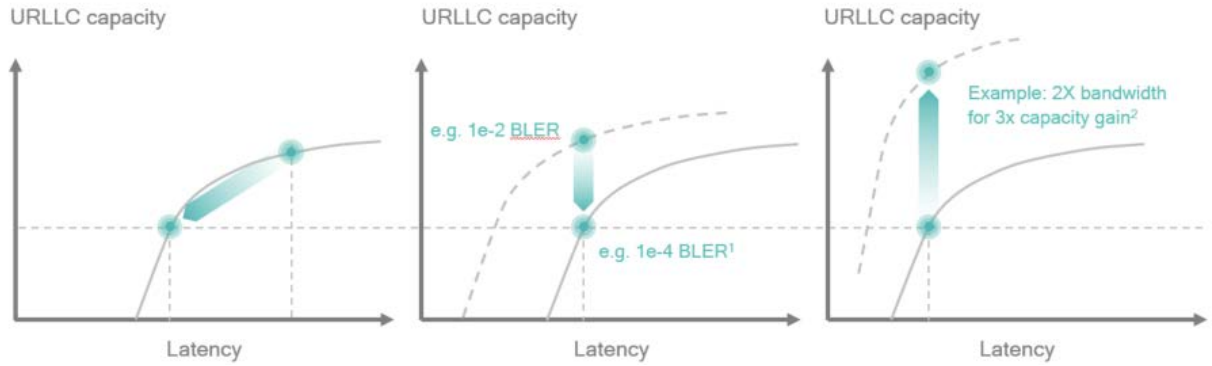


Figure 6.9. Qualitative Queueing Analysis of URLLC Capacity.

More quantitative results of URLLC capacity for both UL and DL are shown in Figures 6.10 and 6.11, respectively, where similar results are seen. These include super-linear growth of capacity as a function of bandwidth. Another result is a drastic drop in system resource utilization as deadline tightens. It can be seen that, with a given amount of resource, it is important to dynamically share resources between eMBB and URLLC to achieve the best efficiency.

Finally, some results of TDD vs. TDD+FDD Carrier Aggregation (CA) are shown in Figure 6.12. It has been shown that high-bandwidth TDD can be very useful to support URLLC, especially with the assistance of FDD channel (for ACK channel/Physical Uplink Control Channel (PUCCH) anchoring) and to schedule retransmission across different component carriers. For practical purposes, aggregating FDD and TDD spectrum resources to support high-capacity URLLC is important to achieve high efficiency.

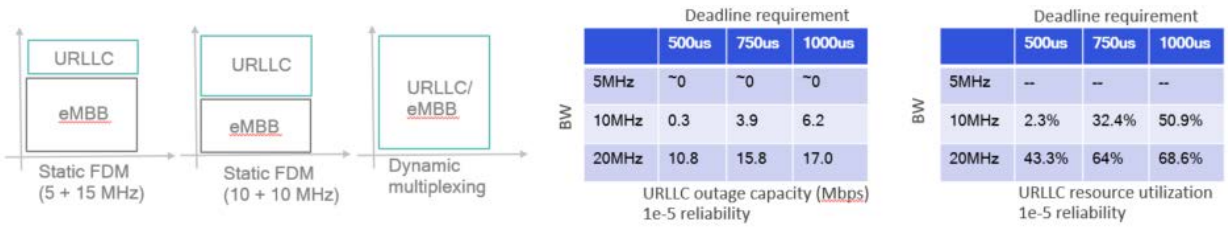


Figure 6.10. DL Evaluation Results.

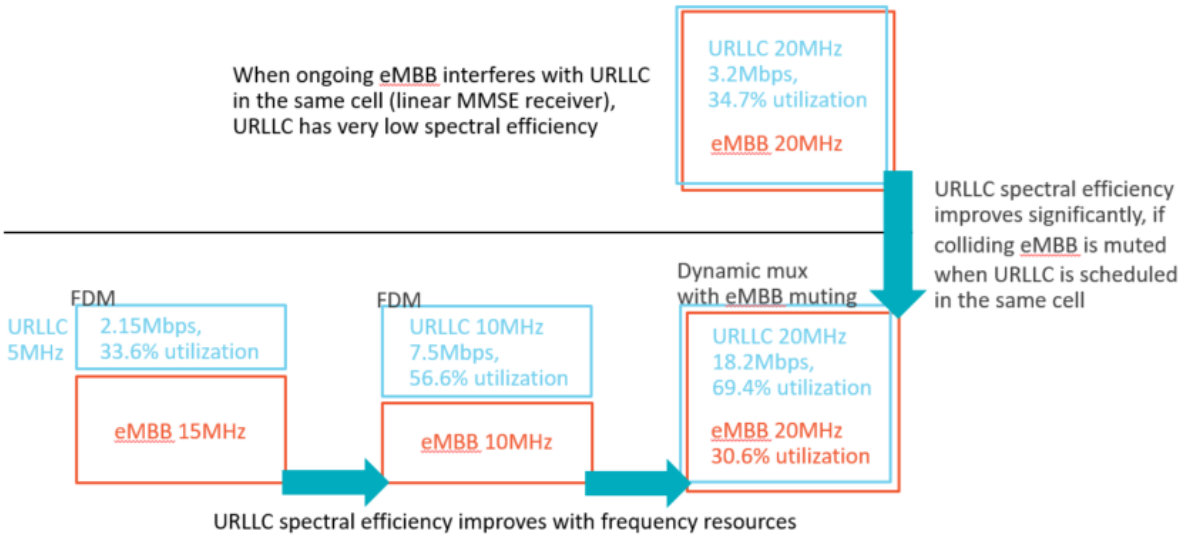


Figure 6.11. UL Evaluation Results.

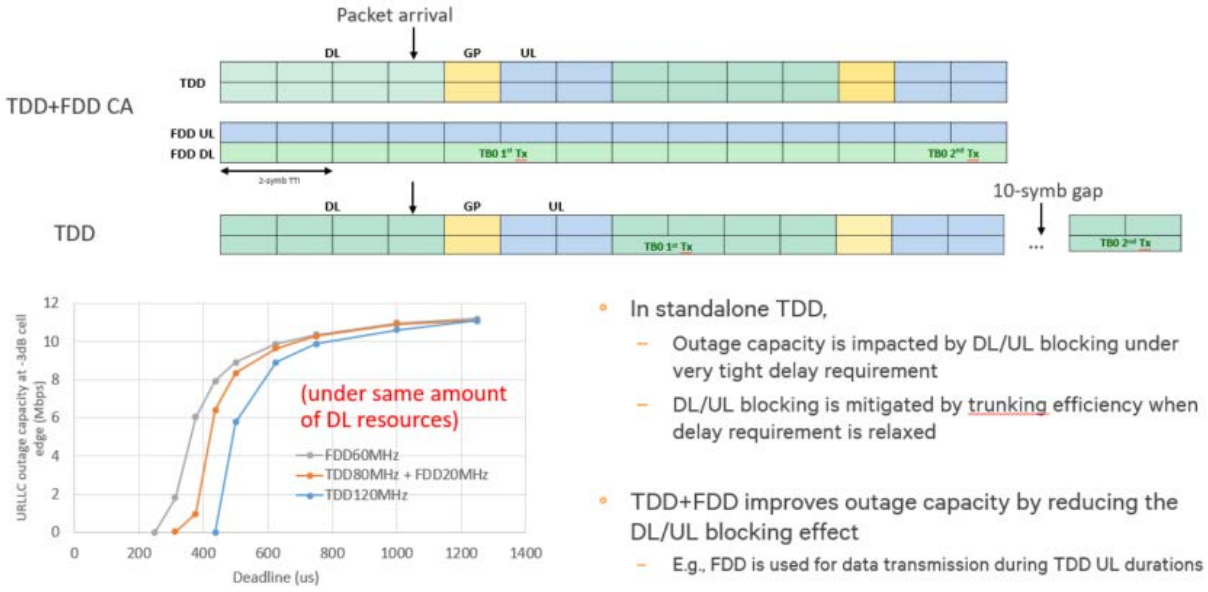


Figure 6.12. TDD+FDD Carrier Aggregation Evaluation Results.

7. SECURITY CONSIDERATIONS

In this section, we discuss security aspect of URLLC in the context of 5G.

7.1 5G SYSTEM SECURITY

In 5G security, the supported crypto algorithms for Access Stratum (AS) and Non-Access Stratum (NAS) security are identical to LTE as of Release 15 but can evolve (therefore, be added) independently.

7.1.1 5GS SECURITY ARCHITECTURE

The 5GS security architecture is shown in Figure 7.1.

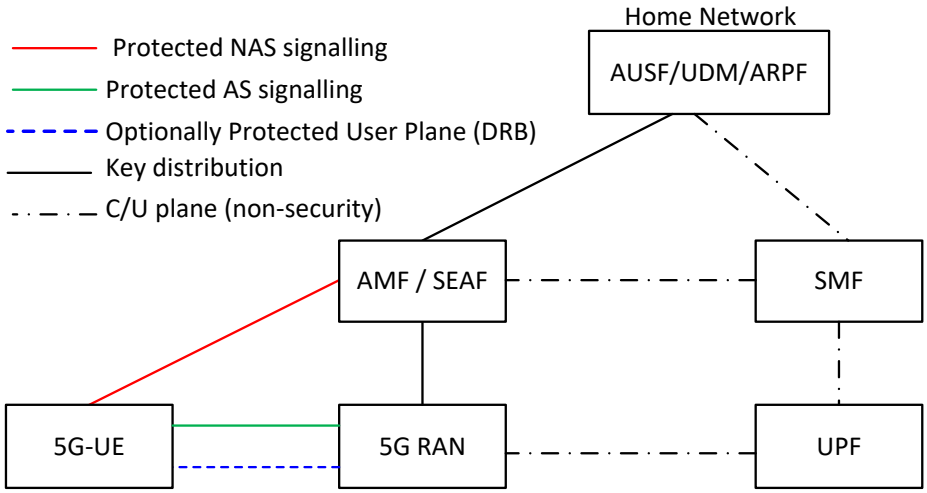


Figure 7.1. 5GS Security Architecture across C/U Plane.

The Authentication Server Function (AUSF), Unified Data Management (UDM) and Authentication credential Repository and Processing Function (ARPF) are all located in the home network in the case of roaming. The Authentication Server Function (AUSF) is responsible for UE authentication. The Unified Data Management (UDM), including the Authentication credential Repository and Processing Function (ARPF), is responsible for subscription information management and storage, including generation of the 3GPP Authentication and Key Agreement (AKA) Authentication Credentials. The UDM also is responsible for user identification handling (for example, storage and management of Subscriber Permanent Identifier) (SUPI) for each subscriber in the 5G system) and access authorization based on subscription data (for example, roaming restrictions). The subscription credentials include the set of values in the Universal Subscriber Identity Module (USIM and the ARPF, consisting of at least the long-term key(s) and the subscription identifier SUPI, used to uniquely identify a subscription and to mutually authenticate the UE and 5G core network.

The Access Mobility Function (AMF) is located in the serving network controls and terminates NAS security with the UE. The AMF uses the NAS Security Mode Command to signal the chosen algorithms and protect them from bid-down attacks. The AMF receives the Key Security & Anchor Function (KAMF) from the Security Anchor Function (SEAF) as part of a primary authentication run and provides keys to the gNB. The AMF also holds fresh keys that are fetched after/during handovers. The AMF controls when the key hierarchy in serving network is re-keyed (for example, to trigger a primary authentication).

The SEAF is also located in the serving network and is co-located with the AMF (for Release 15) but may be separated later. Separation allows the proper isolation of AMFs without requiring a fresh primary authentication, as well as control of security features that are outside of the AMF's control (for example, UE-to-UPF security, if introduced). The SEAF receives key generated for SEAF KSEAF from AUSF as part of a primary authentication run.

7.1.2 MUTUAL AUTHENTICATION & ESTABLISHMENT OF KEYING MATERIALS BETWEEN UE & NETWORK

The 5G authentication framework includes 5G AKA and Extensible Authentication Protocol (EAP-AKA). It is up to the home network to decide whether to use 5G AKA or EAP-AKA.

5G AKA is similar to EPS AKA but is enhanced with the addition of Authentication Confirmation (AC) from the Security Anchor Function (SEAF) to the Authentication Server Function (AUSF). The addition of AC is for increased home network control (therefore, confirmation that the UE is present in the serving network) to avoid certain types of billing fraud during roaming scenarios. EAP-AKA and other EAP methods natively provide this control as the EAP methods terminate at the AUSF in the home network.

In the 5G EAP framework, the UE takes the role of EAP peer. The SEAF takes the role of EAP pass through authenticator, and the AUSF takes the role of back-end authentication server. In addition, there is now optional EAP-based secondary authentication for access to specific data networks/service.

Another distinct feature of 5G authentication is Subscription Permanent Identifier (SUPI) privacy. The SUPI in 5GS can be either an IMSI or Network Access Identifier (NAI). The protected version of the SUPI that is sent over the air is termed the Subscription Concealed Identifier (SUCI). The UE calculates a fresh SUCI every time from SUPI and sends it, for example, in the response to the NAS identity request. In the SUCI, the home network identifier part of the SUPI (for example, Mobile Country Code/Mobile Network Code (MCC/MNC) in IMSI) is included in the clear, and only the Mobile Subscription Identification Number MSIN/username part of SUPI is encrypted.

7.1.3 AS SECURITY

The same Access Stratum (AS) security algorithms are used for Radio Resource Control (RRC) and the user plane. The AS security algorithms are Radio Access Technology (RAT) dependent.

The AS security algorithm selection at initial AS security context establishment consists of the UE providing the AMF with the UE security capabilities over NAS and the AMF then providing these capabilities to the serving gNB. The gNB then indicates to the UE the chosen algorithm in the AS Security Mode Command (SMC).

The gNB activates ciphering and integrity protection of user data based on the security policy sent by the SMF. The security policy consists of an indication of User Plane (UP) integrity and UP ciphering activation per Protocol Data Unit (PDU) session (therefore, for all DRBs belonging to the PDU session). For UP integrity protection, the maximum data rate is based on the UE capability. The gNB shall not overrule the UP-security policy provided by the SMF.

In addition, there is AS security algorithm selection during Xn- or N2-based handover, where any algorithm change is signaled to the UE using the RRC Connection Reconfiguration procedure. The UE's UP security policy is in the handover request message to the target gNB. The target gNB activates UP confidentiality and/or UP integrity protection per Digital Radio Broadcasting (DRB), according to the received UE's UP security policy, and rejects all PDU sessions for which it cannot comply with the received-UP security policy and indicates the reject cause to the core network. For Xn-based handover, the same key derivation and algorithm selection procedure occur as for LTE X2 handover.

In addition, AS security accounts for the following key differences from LTE due to the support of Central Unit -- Distributed Unit (CU-DU) split in the RAN:

- Key retainment, which allows the K_{gNB} to be retained in cell change if the PDCP anchor does not change
- CU-CP and CU-UP separation, which is transparent to the UE. CU-UPs are assumed to reside in the same security domain

7.1.4 OTHER KEY DIFFERENCES FROM EPS (4G SYSTEM)

5G security includes several additional enhancements compared to EPS security that may be applicable to the URLLC in some deployments.

Initial NAS protection has been introduced to protect confidentiality of Information Elements (IEs) that are not required for establishing (or identifying) NAS security contexts between UE and the serving network. Once the UE is authenticated, the AMF shall allocate the 5G- Globally Unique Temporary UE Identity (5G-GUTI), which is used in NAS messages by the UE instead of the SUCI and by the AMF to identify the UE's 5G security context.

5G security also includes the following enhancements:

- Bid-down protection for network security features that may be introduced in future releases
- Per-Protocol Data Unit (Per-PDU) session-level negotiation of DRB confidentiality and/or integrity protection of user plane traffic
- Mobility Anchor (AMF) key change without requiring fresh authentication
- Enhanced Key Hierarchy to support future deployment models and services
- Introduction SEcurity Anchor Function (SEAF) in serving network (KSEAF)
- Key left at the home network after primary authentication (KAUSF)

7.2 RADIO RESOURCE ISOLATION FOR URLLC

The 5G system primarily uses network slicing to enable resource isolation, especially at the core network level. At the RAN level, slice resource isolation is primarily an implementation and network operation/management issue.

Slicing is initiated by the UE and optionally provides the Network Slice Selection Assistance Information (NSSAI) as part of the registration procedure. The network (AMF) responds with the accepted NSSAI. The NSSAI consists of one or more Single NSSAI (S-NSSAI), where each network slice is uniquely identified by a S-NSSAI, as defined in TR 23.799. An S-NSSAI may have a standardized value or a value specific to the PLMN.

The RAN supports resource management, QoS and differentiated handling of traffic within and across slices, as well as policy enforcement for different network slices based on implementation.

Resource isolation between slices is also based on implementation and can include Radio Resource Management (RRM) differentiation across slices, as well as enforcement of service level agreements across slices for shared resources. In addition, it is possible (based on implementation) to dedicate specific RAN resources to a slice.

An additional mechanism to enable resource isolation (within a single device) is Logical Channel Prioritization (LCP) Restriction at the MAC layer. While there is no explicit separation due to slicing provided in the RAN, this feature introduced for NR allows traffic from different flows to be restricted (therefore, isolated) to the following:

- A set of usable numerologies (for example, URLLC LCH can use only 60 KHz Sub Carrier Spacing (SCS) or higher)
- Maximum Transmission Time Intervals (TTIs) (for example, URLLC LCH can use only TTIs of 250us or shorter)
- Transmission mode: dedicated vs. shared grant (for example, eMBB can't use shared grants)

Based on configuration, each logical channel is mapped to one or multiple TTIs and/or numerologies. For a UL grant of particular TTI and/or numerology, Logical Channel Prioritization (LCP) restriction specifies which set of logical channels (LCHs) (therefore, which traffic flows) are eligible for the grant. LCP restriction is performed before the prioritization procedure. Whether to use LCP Restriction to isolate the URLLC traffic is a decision left to configuration.

7.3 SECURE RADIO RESOURCE SCHEDULING

The 5G system aims at high reliability and low latency over the radio link even in the presence of attacks against availability, especially radio jamming attacks. In general, it is extremely difficult to counter the radio jamming attacks due to the wireless medium's intrinsic vulnerability. At the same time, it is equally difficult for an attacker to effectively jam the wide spectrum over the wide area that a cellular base station covers. However, this does not necessarily mean the cellular system is relatively safe against such jamming. A sophisticated attacker may exploit the radio resource scheduling mechanisms employed in the 3GPP system to maximize attack effects.

Intelligent attacks can be classified into two broad categories: control channel attacks and data channel attacks.

7.3.1 CONTROL CHANNEL JAMMING ATTACK

Jamming uplink and/or downlink control signals can effectively deny packet transmissions over the data channels.

For the uplink, the Physical Uplink Control Channel (PUCCH) is used to transmit a Scheduling Request (SR), ACK/NACK or Channel State Information (CSI) by a UE. As a result, jamming the PUCCH will result in preventing/disrupting data transmission or triggering Remote Line Failure (RLF). Because the PUCCH resources are located at the edge of the system bandwidth and utilize narrow bandwidth, it can be selectively jammed by an attacker.

For the downlink, the Physical Downlink Control Channel (PDCCH) is used to carry the Downlink Control Information (DCI), such as DL scheduling assignments or UL scheduling grants for a specific UE. Jamming PDCCH would prevent UEs from being scheduled. However, the effect of PDCCH jamming is relatively localized if the attacker's jamming signal power is limited. Nonetheless, PDCCH offers very useful information for the attacker who is targeting a set of UEs because PDCCH contains information about the UE specific uplink and downlink Resource Block (RB) allocation.

7.3.2 DATA CHANNEL JAMMING ATTACK

Selective jamming of the shared data channels (therefore, PUSCH and PDSCH) targeting a set of UEs is possible by decoding PDCCH. Because the PDCCH contains information about the RBs allocated to the target UEs, an attacker can concentrate jamming signal on those RBs in the data channels, thereby optimizing the attacker's power utilization.

Control channel jamming, especially against PUCCH, impacts all UEs connected to a base station. The presence of the attacks can be identified by the base station and can trigger an alarm or a reactive measure (for example, by locating the jamming source(s)). On the other hand, data channel jamming can be more selectively launched against targeted UEs, hence is stealthier than the control channel jamming. Note that the entire RBs allocated to the target UEs do not need to be jammed. Damaging only a portion of RBs (or certain RBs) may be sufficient to disrupt services or at least degrade the service qualities.

7.3.3 POTENTIAL COUNTERMEASURES

The effects of PUCCH jamming can be mitigated as follows:

- Spreading PUCCH to a wider bandwidth would make it more difficult for an attacker to jam the PUCCH
- Scheduling PUCCH with PUSCH would substantially randomize the PUCCH region, thereby disallowing narrow band PUCCH jamming

3GPP Release 9 (LTE) introduced simultaneous PUCCH and PUSCH transmission by multiplexing PUCCH with data on PUSCH.

7.4 DEVICE PLATFORM SECURITY

Besides the AS and NAS/UP security, 5G systems need to be equipped with more intrinsic security functionalities tailored for URLLC services. URLLC devices such as sensors and actuators may be deployed and operated in unattended environments and managed remotely. Due to this nature, those devices are susceptible to malicious attacks either launched locally (for example, based on physical attacks) or remotely (for example, by exploiting software vulnerabilities).

To overcome such attacks, devices need to be security hardened. They also should support the capability for the network to verify the integrity of the applications running in the device remotely. In other cases, the remote servers that collect data from the devices and process the data need to verify the data provenance, which guarantees the data origin and its authenticity and the data governance that manages the access and processing privilege of the data. These security features, in effect, rely on the strong security of the device platform that is preferably established based on the hardware root of trust in the device and is verifiable based on the hardware root of trust.

8. 3GPP URLLC DESIGN AND SPECIFICATION STATUS

URLLC is one of the key aspects of the 5G system being specified at 3GPP. The requirements for enhanced URLLC work are captured in 3GPP TS 22.804, as well as in others referenced in Section 3. 3GPP has two ongoing study items related to this feature. One is “Study on enhancement of URLLC supporting in 5GC” in SA2 for Release 16. The other is the “SID on Physical Layer Enhancements for NR URLLC” in RAN for Release 16. In terms of URLLC support in LTE, there is a completed work item “EPC support for E-UTRAN URLLC” in 3GPP SA2 for Release 16. In addition, the eV2X study in SA2 (Study on architecture enhancements for 3GPP support of advanced V2X services, TR 23-786) also addresses some of the QoS aspects for V2X services that demand URLLC services.

8.1 GENERAL TECHNICAL APPROACH FOR URLLC IN 5G NETWORK

A challenge for creating a URLLC-capable network is in the nature of diverse and sometimes competing use cases that need to be covered in 5G networks. Some of these core network challenges are the increasing complexity, increasing capacity demands for the network nodes (and interactions between the nodes) and the dynamic nature of topology and the heterogeneity of E2E elements in the network (from UE, RAN, core and AS). To meet these challenges, several technologies are being studied and developed for the 5G core network and defined in the 3GPP standard. The following are some of techniques that are applicable:

- Modular network design that supports dynamic service chaining and enables the use of NFV and Software-Defined Networking (SDN) technologies
- Separation of control plane and user plane to enable different topological design to reduce latencies that can be tailored to specific use cases or applications
- Use of NFV and SDN technologies that enable flexible placement of network elements and optimize their interactions
- Increasing NF’s internal reliability and availability, as well as optimizing interactions between NFs
- Providing multiple data paths to ensure the delivery of application data (redundant transmission in user plane)
- Optimizing handover procedures to ensure low latency and low jitter for URLLC services
- Enhance coordination between application function (AF) and 5GC to support application relocation (therefore, to an edge data network) without impacting control plane functions and service continuity. This is also related to the second bullet above
- URLLC has stringent requirements on latency, jitter and other QoS requirements, which demand the monitoring of QoS in order for the network to take actions to meet these requirements, when possible
- Possible ways to reduce RAN resource utilization to support event-driven, low-latency use cases. This is mainly a RAN-related enhancement

8.2 URLLC FOR 5G CORE

The goal of the ongoing SA2 study “Study on enhancement of Ultra-Reliable Low-Latency Communication (URLLC) support in the 5G Core network” (TR 23. 725) is to investigate further potential and new enhancements to the 5GC architectural specifications that can support URLLC services. This study item was approved as one of the priority study items in Release 16 for SA2 at the 3GPP Plenary #80 in La Jolla, and the work is ongoing.

8.3 URLLC FOR NR

In 5G NR Release 15, many features enabling URLLC traffic were already present in overall design. These include the following:

- NR supports different numerologies with subcarrier spacings of 15, 30, 60 or 120 kHz. Larger subcarrier spacing implies shorter slot duration and lower latency
- NR supports mini-slot scheduling, where the transmission occupies only some of the Orthogonal Frequency-Division Multiplexing (OFDM) symbols. This allows shorter alignment time between data arrival and transmission, and earlier decoding
- To achieve fast uplink transmission compared to the scheduling request-based procedure, the UL transmission is facilitated by configured grants, where either Type 1 Configured Uplink Grants and Type 2 Configured Uplink Grants can be used. When either Type 1 or Type 2 is configured, the UE can autonomously start the uplink data transmission according to the configured periodicity and radio resources. Type 1 UL transmission relies only on the RRC configured parameters, while Type 2 UL transmission relies on both RRC-configured parameters and DCI (de)activation. In addition, K-slot repetition (therefore, slot aggregation) can be configured to achieve the high reliability requirement
- Aggregation level 16 is supported for PDCCH transmission, supporting reliable PDCCH reception
- Aggressive UE processing capabilities are defined, which URLLC can leverage
    - One example of higher UE capability is multiple PDCCH monitoring occasions within one slot are supported to achieve finer time domain scheduling granularity supporting the low latency requirement of URLLC. Specifically, a PDCCH monitoring pattern within a slot can be configured, indicating first symbol(s) of the control resource set within a slot for PDCCH monitoring, by higher layer parameter monitoring symbols within a slot
    - Another example of higher UE capability is that low-latency CSI reports can be triggered. This improves the efficiency for URLLC resource utilization by obtaining more accurate and in-time CSI
    - Yet another example of higher UE capability is that the enhanced UE capability #2 is defined in addition to UE capability #1, supporting faster data transmission and reception
- Data duplication on higher layers to ensure high reliability

In addition, some features were introduced specifically to support URLLC that can also be used for other kinds of traffic:

- A CQI table targeting BLER 1e-5 was introduced to support reliable link adaption for DL data
- MCS tables supporting spectral efficiencies down to 0.0586 bits per modulation symbol (compared to 0.2344 bits per modulation symbol for eMBB tables) were introduced to support reliable data transmission using very low code rates
- Dynamic switching between the low spectral efficiency table and an eMBB table is supported for UEs supporting both eMBB and URLLC traffic, or UEs with quickly changing radio conditions

3GPP Release 16 includes a RAN1-led study item studying enhancements for further improved latency and reliability. Expected to be finished in February 2019, the study item contains three parts:

- Layer 1 improvements
    - PDCCH enhancements, focusing on compact DCI, PDCCH repetition, increased monitoring capability

    - Universal Communications Identifier (UCI) enhancements, focusing on enhanced HARQ feedback methods and CSI feedback enhancements
    - PUSCH enhancements focusing on mini-slot level frequency hopping and enhancements to retransmission and repetition schemes
    - Scheduling/HARQ/CSI processing timeline enhancements
- Enhanced multiplexing considering traffic with different latency and reliability requirements
- Enhanced configured grant transmissions studying, for example, explicit HARQ-ACK, mini-slot repetitions within a slot and ensuring the same number of repetitions when repeating

There is also a RAN2-led study item studying industrial IoT, considering enhancements to FR1, FR2, TDD, and FDD. Expected to be finished in February 2019, the study focuses on:

- L2/L3 enhancements
    - Data duplication enhancements, including higher layer multi-connectivity
- UL/DL intra-UE prioritization and multiplexing
- Time sensitive networking
    - Delivery processes for providing an accurate reference timing
    - Scheduling enhancements to support different traffic patterns, QoS for wireless Ethernet, cyclic traffic
    - Ethernet header compression

8.4 HRLLC (HIGHER-RELIABILITY AND LOW-LATENCY COMMUNICATION) FOR LTE

HRLLC LTE was specified during 3GPP Release 15. The design is now stable and, with the Release 15 cycle complete, has entered maintenance.

8.4.1 DESIGN OVERVIEW

The LTE URLLC design is built on the 3GPP feature for low latency known as Short TTI and Short Processing Time. Thus, short TTI, sub-ms physical layer latency can be achieved with LTE. The design for URLLC LTE is based on requirements from ITU 2020 and thus allows to achieve transmission of a 32 bytes payload within 1 ms at a success rate of 99.999 percent (error rate of 1E-5).

The main additional features specific to URLLC are:

- PDSCH repetitions: Enabled by RRC. A new DCI with a dynamic parameter allows the scheduler to use 1, 2, 3, 4 or 6 repetitions
- PUSCH repetitions with uplink semi-persistent scheduling. Multiple configurations of UL Signaling Protocols & Switching (SPS) can be enabled to allow flexible activation/UL transmission time. This mitigates the lack of UL SPS transmission opportunity caused by the repeated transmission
- New DCI formats carried by PDCCH / SPDCCH to schedule repetitions
- Physical Control Format Indicator Channel (PCFICH)-free transmission
    - Support an optional, CFI configuration per TTI length through UE and serving-cell-specific semi-static configuration
    - If CFI is semi-statically configured for TTIs of different lengths, the UE does not expect the configured CFI values to be different

    - The semi-static CFI value could be configured separately for Multimedia Broadcast multicast service Single Frequency Network (MBSFN) and non-MBSFN subframes over each cell
    - When a UE is configured with a semi-static CFI for a given TTI length, the UE is not expected to decode PCFICH for that TTI length
- PDCCH enhanced reliability by virtual CRC: 1 bit of the DCI in PDCCH can be fixed (semi-static configuration) to enhance the reliability of PDCCH, if needed

In summary, 3GPP standards work to support URLLC services in 5G is ongoing with study items in both the 5G RAN and the 5G core network. The related specifications are built on Release 15 architecture, with additional techniques to enhance 5G networks' capability to support URLLC services. These techniques leverage both fundamental technologies such as 5G phase I architecture and MEC and, in combination with specific core network and RAN techniques, provide flexible deployment options that can be tailored to different use cases. These works are targeted for completion in the Release 16-time frame.

9. CONCLUSIONS AND RECOMMENDATIONS

This is a time when mobile networks and vertical industries are going through a major technological transformation aided by critical communication capabilities in terms of low-latency and high-reliability features, particularly in the health care, automotive, industrial automation, energy and entertainment sectors. As the first commercial 5G deployments are focused on enhanced mobile broadband use cases, the future is gearing up for use cases characterized by ultra-high reliability and/or low-latency features.

Low latency is seen as a crucial ingredient to ensure applications are usable and interactive whether communication is human-to-human, human-to-machine or machine-to-machine. Significant design, standardization and engineering challenges are being overcome to deliver networks that are both reliable and provide low latencies.

This white paper presents a brief overview of emerging applications, design challenges, and potential approaches in the design of URLLC. This paper describes upcoming use cases of URLLC in tele-surgery, smart transportation and industry automation, and presents the latency and reliability requirements for these applications. The paper also specifies key latency bottlenecks in current cellular networks, provides a breakdown of the various delay sources in 5G networks and lays out the necessary implementation blocks for achieving end-to-end latency reduction required to support mission-critical applications.

This white paper also summarizes the recent performance evaluation results of the basic designs and implementation of the 5G physical layer, multiple access layers and air interface blocks essential to reducing latency and achieving the desired reliability.

Furthermore, the paper also discusses other potential latency reduction measures, including MEC.

10 LIST OF FIGURES

11 LIST OF TABLES

APPENDIX A: ACRONYM LIST

3GPP	Third-Generation Partnership Project
4G	Fourth Generation mobile communications
5G	Next generation mobile communications
5Qis	5G Quality of Service Indicators
AC	Authentication Confirmation
ACK	Acknowledgment
AF	Application Function
AKA	Authentication and Key Agreement
AMF	Access Mobility Function
AN	Access Network
AR	Augmented Reality
ARP	Address Resolution Protocol
AS	Access Stratum
ARPF	Authentication credential Repository and Processing Function
AUSF	Authentication Server Function
BLER	Block Error Rate
CA	Carrier Aggregation
CBG	Cod Block Group
CDG	Cumulative Distribution Function
CN	Core Network
CQI	Channel Quality Indication
CSI	Channel State Information
CU	Control Unit
DCI	Downlink Control Information
DL	Downlink
DMRS	DeModulation Reference Signal
DN	Data Node or Data Network

DNAI	Dynamic Network Access Identifier
DNN	Data Network Name
DRB	Digital Radio Broadcasting
DU	Distributed Units
EAP	Extensible Authentication Protocol
eMBB	Enhanced Mobile Broadband
EPS	Evolved Packet System
FDD	Frequency-Division Duplexing
GBR	Guaranteed Bit Rate
GFBR	Guaranteed Flow Bit Rate
gNB	Next-Generation NodeB
GPSI	General Purpose Serial Interface
GUTI	Global Universal Transport Interface
HARQ	Hybrid Automatic Repeat Request
I-UPF	Inter User Plane Function
ICP	Internet Content Provider
IE	Initial UE identity
IEs	Internet Elements
IETF	Internet Engineering Task Force
IMS	IP Multimedia Subsystem
IMSI	International Mobile Subscriber Identity
IoT	Internet of Things
ISP	Internet Service Provider
KAMF	Key Access & Mobility Function
KSEAF	Key Security & Anchor Function
KPI	Key Performance Indicator
LAN	Local Area Network
LCP	Logical Channel Prioritization

LCH	Logical Channel
LDCP	Low-Density Parity Check
LLC	Low Latency Communications
LLS	Low Latency Socket
LTE	Long Term Evolution
MAC	Medium Access Control
MBSFN	Multimedia Broadcast multicast service Single Frequency Network
MCC	Mobile Country Code
MCS	Modulation Coding Scheme
MEC	Multi-access Edge Computing
MFBR	Max Flow Bit Rate
MNC	Mobile Network Codes
MO/MT	Mobile Originating/Mobile Terminating
MR	Mixed Reality
MSC	Mobile Switching Center
MSIN	Mobile Subscription Identification Number
MTC	Machine-Type Communications
NAI	Network Access Identifier
NAK	Negative Acknowledgment
NAS	Non-Access Stratum
NEF	Network Element Function
NF	Network Functions
NFV	Network Functions Virtualization
NG-RAN	Next Generation Radio Access Network
NSSAI	Network Slice Selection Assistance Information
OFDM	Orthogonal Frequency-Division Multiplexing
PCFICH	Physical Control Format Indicator CHannel
PDCP	Packet Data Convergence Protocol

PDN	Packet Data Network
PDSCH	Packet Data Shared CHannel
PDU	Protocol Data Unit
PGW	PDN Gateway
PHY	Physical Layer
PCF	Point Coordination Function
PDCCH	Physical Downlink Control CHannel
PLMN	Public Land Mobile Network
PRB	Physical Resource Block
PUCCH	Physical Uplink Control CHannel
PUSCH	Physical Uplink Shared Channel
QAM	Quadrature Amplitude Modulation
QCI	QoS class identifier
QFI	Quality of service Flow ID
QoS	Quality of Service
RA	Radio Access
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Resource Block
RFC	Remote Function Cell
RLC	Radio Link Control
RLF	Remote Line Failure
RNA	RAN Notification Area
RQA	Reflective QoS Attribute
RQoS	Reflective QoS
RRC	Radio Resource Control
RRM	Radio Resource Management
RTT	Round-Trip Transmission

TA	Tracking Area
Tx	Transmitter
RAT	Radio Access Technology
Rx	Receiver
S/NR	Signal to Noise Ratio
SCS	Subcarrier Spacing
SDAP	Service Data Adaption Protocol
SDF	Service Data Flow
SDN	Software-Defined Networking
SDU	Service Rata Unit
SEAF	SEcurity Anchor Function
SGW	Serving GateWay
SINR	Signal-To-Interference-Plus-Noise Ratio
SMC	Security Mode Command
SMF	Session Management Function
SNR	Signal-To-Noise Ratio
SPS	Signaling Protocols & Switching
SR	Scheduling Request
SSC	Session and Service Continuity
SUCI	SUbscription Concealed Identifier
SUPI	SUbscription Permanent Identifier
TA	Timing Advance (Parameter)
TBS	Transmission Block Selection
TDD	Time-Division Duplexing
TTI	Transmission Time Interval
TX	Transmit
UCI	Universal Communications Identifier
UDM	Unified Data Management

UE	User Equipment
UL	Uplink
UL CL	Uplink Classifier
UP	User Plane
UPF	User Plane Function
URA	UTRAN Registration Areas
URLLC	Ultra-Reliable, Low-Latency Communications
USIM	Universal Subscriber Identity Module
UTRAN	Universal Terrestrial Radio Access Network
VR	Virtual Reality

APPENDIX B: GLOSSARY OF TECHNICAL CONCEPTS

The system-level assumptions used the for the performance evaluation are summarized in Table B.1.

Table B.1. System-Level Evaluation Parameters for SINR Derivation.

Parameters	Urban Macro
Carrier frequency for evaluation	Config A: 4 GHz Config B: 700 MHz
BS antenna height	25 m
Total transmit power per TRxP	46 dBm per 10 MHz bandwidth
UE power class	23 dBm
min UE power	-40 dBm
Percentage of high loss and low loss building type	100percent low loss (applies to Channel model B)
Inter-site distance	500 m
Number of antenna elements per TRxP	64 Tx/Rx, (M,N,P,Mg,Ng) = (4,8,2,1,1), (dH,dV) = (N/A, 0.8)λ +45°, -45° polarization
Number of UE antenna elements	2 Tx/Rx, (M,N,P,Mg,Ng) = (1,1,2,1,1) 0°, 90° polarization
Device deployment	80percent outdoor, 20percent indoor Randomly and uniformly distributed over the area
UE mobility model	Fixed and identical speed |v| of all UEs of the same mobility class, randomly and uniformly distributed direction
UE speeds of interest	3 km/h for indoor and 30 km/h for outdoor
Inter-site interference modeling	Explicitly modelled
BS noise figure	5 dB
UE noise figure	7 dB
BS antenna element gain	8 dBi
BS antenna element pattern	According to TR 36.873
UE antenna element gain	0 dBi
UE antenna element pattern	Omni-directional
Thermal noise level	-174 dBm/Hz
Simulation bandwidth	40 MHz
UE density	10 UEs per TRxP
UE antenna height	1.5 m
Channel model variant	Channel model B, IMT-2020 Urban Macro
TRxP number per site	3
Mechanic tilt	90° in GCS (pointing to horizontal direction)
Electronic tilt	99° in LCS
Handover margin (dB)	0 (therefore, the strongest cell is selected)
TRxP boresight	30 / 150 / 270 degrees
UE attachment	Based on RSRP (formula (8.1-1) in TR36.873) from port 0
Wrapping around method	Geographical distance-based wrapping
Minimum distance of TRxP and UE	d2D_min= 10m
Traffic model	Full buffer (Note: it is for SINR CDF distribution derivation)

ACKNOWLEDGMENTS

The mission of 5G Americas is to advocate for and foster the advancement of 5G and the transformation of LTE networks throughout the Americas region. 5G Americas is invested in developing a connected wireless community for the many economic and social benefits this will bring to all those living in the region. 5G Americas' Board of Governors members include AT&T, Cable & Wireless, Cisco, CommScope, Ericsson, Intel, Kathrein, Mavenir, Nokia, Qualcomm Incorporated, Samsung, Shaw Communications Inc., Sprint, T-Mobile USA, Inc., Telefónica and WOM.

5G Americas would like to recognize the significant project leadership and important contributions of project co-leaders Rao Yallapragada from Intel and Jing Jiang of Qualcomm, as well as representatives from member companies on 5G Americas' Board of Governors who participated in the development of this white paper.