Overview of 5G URLLC System and Security Aspects in 3GPP

Takahito Yoshizawa
KU Leuven – imec – COSIC
Kasteelpark Arenberg 10 Bus 2452,
B-3001 Leuven, BELGIUM
takahito.yoshizawa@kuleuven.be

Sheeba Backia Mary Baskaran
Huawei Technologies Sweden AB,
sheeba.backia.mary@huawei.com

Andreas Kunz
Lenovo, Oberursel, Germany,
akunz@lenovo.com

***Abstract*— With the emerging wide range of sophisticated 5G services and vertical business models, the mobile communication extends to include vehicles, high-speed trains, drones and industrial robots. Mission-critical applications in vertical industries have stringent service performance requirements in terms of latency, availability and reliability. Low latency seen as a critical deciding factor over service performance in some of the vertical industries, e.g., manufacturing or vehicular communications, thus the Ultra-Reliable Low Latency Communications (URLLC) is viewed as the enabling technology in the 5G system (5GS). This paper presents an overview of the 3GPP 5G system architectural and security enhancements to support URLLC services based on the recent standardization activities in 3GPP.**

Keywords—3GPP, 5G, URLLC, Security, Vertical industry, PSA.

I. INTRODUCTION

URLLC is the basis technology for an entirely new family of use cases in 5G system. Notable applications in vertical industries include autonomous driving for the automotive industry (Intelligent Transportation), remote surgery for eHealth (Remote Healthcare), and cloud robotics and deterministic communication for Industry 4.0 (Industrial Automation). In addition to these vertical industries, other URLLC nonexclusive applications are Tactile Interactions (TI), Augmented Reality (AR), Virtual Reality (VR), and Mixed Reality (MR). Collectively, the list of applications is listed in Table 1 and overall performance requirements [1] are summarized below:

- Improved latency - maximum of up to 1ms,
- Improved reliability - less than 10^{-5} packet drop rate,
- Higher availability - up to 99,999999999999 %,
- More stringent security – Guaranteed service availability in addition to confidentiality and integrity protection in the presence of denial-of-service (DoS) attacks.

These additional level of requirements pose new security challenges. In addition to the existing 5G robust security mechanisms, URLLC needs lower latency in access authentication, transmission protection, and security context handling.

Several studies and work items [2] culminated in extensive list of URLLC use cases and requirements from industry verticals, network operators and suppliers. Industry groups other than 3GPP, including 5G Americas [3], GSM Association (GSMA) [4], 5G Alliance for Connected Industries and Automation (5G-ACIA) [5] and 5G Automotive Association (5GAA) [6] are also developing requirements for specific markets and working with 3GPP to realize them.

Based on these service requirements, 3GPP SA2 (system architecture) and SA3 (security) working groups have conducted studies on system and security aspects of URLLC in the 5G system, respectively. Their studies have recently resulted in two Technical Report (TR) documents TR 23.725 [7] and TR 33.825 [9]. In this paper, we analyze these documents and highlight the key points.

The paper is organized as follows. We present the overall 5G URLLC key system and security aspects in Section II and Section III respectively. The Section IV outlines the open issues with potential way forward, followed by conclusion in Section V.

TABLE I. VERTICAL INDUSTRY AND ITS POTENTIAL URLLC USE CASES

	Use cases	Applications
Vertical Industry	Factory Automation	Motion Control
		Control-to-Control Communication
		Control-to-sensor/actuator communications
		Mobile Robots and Automated Guided Vehicles (AGVs)
		Closed-loop Control
		Process Monitoring
		Plant Asset Management
		Remote Access and Maintenance
	Health Industry	Remote Diagnosis
		Emergency Response
		Remote Surgery
	Electric Power Distribution	Primary Frequency Control
		Distributed Voltage Control
		Distributed automated switching for isolation and service restoration
		Smart grid millisecond-level precise load control
	Intelligent Transportation	Tele-Operated Driving (TOD)
		Collision Avoidance System
		Dynamic Traffic Light Sequence
	Entertainment	Immersive Entertainment
		Online Gaming

II. SYSTEM ASPECTS

3GPP SA2 (system architecture) working group has started the study on URLLC since early 2018. This work resulted in TR 23.725 [7]. It is near completion at the time of this writing, and the subsequent normative specification has

been included in the Release 16 (5G phase 2) version of the 5G system specification [8]. This section discusses the key points from [7].

A. *Supporting High Reliability by Redundant Transmission in User Plane*

The redundant transmission in User Plane (UP) is supported in the 5GS to increase the reliability. Depending on the condition of network deployment, e.g., which Network Functions (NFs) or segments cannot meet the reliability requirements, the Session Management Function (SMF) determines if the redundant transmission can be applied in the user plane path between the UE and the network. The three potential methods that can provide redundant user plane paths at different level includes:

1) Dual Connectivity (DC) based Redundant UP Path,
2) Redundant transmission on N3/N9 interfaces, and
3) Redundant transmission at transport layer.

For example, considering the article length, we describe here only the redundant user plane support using DC as listed in (1). In this case, the 5GS sets up the user plane paths of the two redundant Packet Data Unit (PDU) sessions to be disjoint as shown with Type 1 in the middle part of Fig. 1. During PDU session establishment procedure, the SMF determines whether the PDU Session is to be handled redundantly based on the combination of the Single Network Slice Selection Assistance Information (S-NSSAI), Data Network Name (DNN), user subscription and local policy configuration.

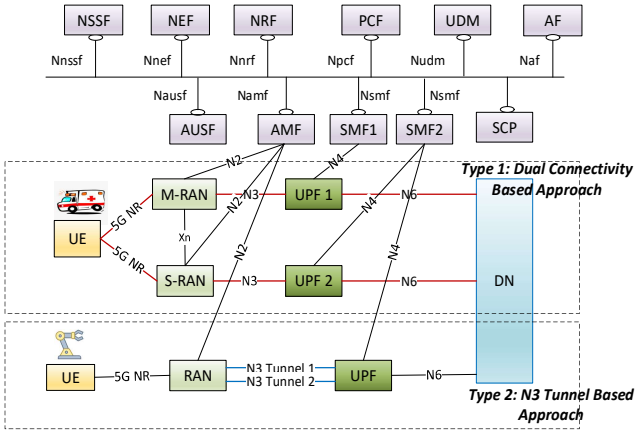

Fig. 1. Overview of End-to-End redundant User Plane paths using Dual Connectivity

The duplicated traffic originating from the same application are associated to two redundant PDU sessions based on the User Equipment (UE), i.e. 3GPP mobile device, Route Selection Policy or UE local configuration [11]. One PDU Session is established from the UE via Master Radio Access Network (M-RAN) to User Plane Function (UPF1) acting as the PDU Session Anchor (PSA), and the other PDU Session from the UE via Secondary RAN (S-RAN) to UPF2 acting as the PSA. Based on these two PDU Sessions, two independent paths to the UE are established. UPF1 and UPF2 connect to the same Data Network (DN), even though their traffic may be routed via different user plane nodes within the DN.

B. *Supporting Low Latency and Low Jitter During Handover Procedure*

Handover is a mechanism where the UE changes the radio connection from one cell to another while a call is in progress. Signaling associated with the handover procedure switches the bearer path from the source cell to the target cell. To maintain the connection of the ongoing call during handover, bearer payloads are buffered at the source cell and forwarded to the target cell while the UE establishes the radio connection with the target cell. This is to ensure all packets are preserved during this process (lossless handover). Due to the nature of this buffering and forwarding process, handover typically incurs some level of temporary degradation in the bearer performance, such as increased delay or jitter of the end-to-end bearer flow, packet loss, or retransmission. Depending on the application being used, this effect may be noticeable to the end user. To ensure low latency and low jitter in 5G during handover, a new solution is needed to either eliminate or minimize this negative effect.

3GPP has identified two possible solutions for this area in [7]:

1) duplicate the user plane tunneling during handover, and
2) handover timing coordination with multiple UEs in the device.

In the first solution, the core network (CN) establishes two tunnels with both the source and the target cells during the handover. Duplicate Downlink (DL) bearer is sent through both of these tunnels simultaneously while the UE establishes the connection with the target cell. This redundant transmission effectively eliminates the need of data forwarding from the source to the target cell during the handover procedure, thus minimizing the extra latency and jitter.

In the second solution, the end device consists of two independent UEs, each having its own PDU connection. In this context, the end device refers to a machine, an equipment, or a vehicle, which contains two UEs. This solution introduces the concept of Reliability Group (RG) where one or more UEs and gNBs (gNode B, the 5G base station) form a RG. In the RG concept, the UE preferably establishes radio connection with the gNB in the same RG group. By having 2 UEs in the device being assigned with two different RGs, the device establishes connection with 2 different gNBs of different RGs. The system is set up in such a way that the handover of different RGs does not occur simultaneously. This effectively results in the UE handover timing within a device to be staggered between two RGs. Consequently, when the UE's mobility condition requires handover to another cell, this mechanism ensures that only one of the RGs are engaged in the handover at any point in time, leaving the other RG in stable state with the gNB. This mechanism maximizes the stable latency and jitter even though one of the UEs in the device is in handover.

The conclusion of the study is not reached at the time of writing as these mechanisms require review by the 3GPP RAN group. Therefore, no conclusion is expected for this issue in Release 16 time frame.

C. Enhancement of Session Continuity during UE Mobility

This topic is regarding the application level session continuity when the Application Function (AF) relocates to another part of the data network. The trigger for the AF relocation is due to the UE's mobility in order to keep the AF anchor close to the UE to minimize the end-to-end (E2E) latency. The study identifies 3 sub-cases:

1) *Uplink Classifier (ULCL) relocation,*
2) *PSA relocation for Ethernet PDU session, and*
3) *Run-time coordination between the AF and the 5GC.*

For each of these sub-cases, solutions were discussed during the study. For the first case, the AF triggers the source Access Stratum (AS) to steer the UE to the target AS using, e.g., IP or HTTP level redirection. In the second case, the target UPF uses layer 2 Ethernet switch port learning mechanism, e.g., gratuitous Address Resolution Protocol (ARP), to update the forwarding table in the DN. In the third case, the SMF and AF coordinates to optimize the timing of UP path change to the target PSA by considering the AF relocation event.

D. QoS Monitoring to Assist URLLC Service

QoS monitoring is essential to ensure stringent E2E QoS requirements for URLLC services. There are several factors that can affect the E2E QoS performance such as radio coverage, network nodes (UPF/RAN/UE) resources, and the transport network. 5G system provides QoS monitoring on different levels of granularities subject to the operators' configuration, i.e. per QoS flow, per UE level or per node level. The QoS monitoring for the URLLC services can also be based on 3rd party application request and Policy Control Function (PCF) policy. Based on the request from the AF, the Policy Control and Charging (PCC) framework is used to activate or deactivate the QoS monitoring for the QoS flow. The SMF sends the QoS monitoring policy for the QoS flow to the PSA UPF and RAN via the PDU Session Establishment or Modification procedure. The PSA UPF and RAN node initiates the packet delay measurement based on the received QoS monitoring policy. The Uplink/Downlink (UL/DL) packet delay between UE and PSA UPF for per UE per QoS flow is a combination of the packet delay between UE and RAN node, and the packet delay between RAN node and PSA UPF. Both UL and DL packet delay need to be measured independently. The PSA UPF reports the QoS monitoring result to the SMF when specific thresholds are reached. The QoS monitoring functionality may evolve based on the RAN working group's decision.

E. Supporting Low Latency without Requiring UE to be Always in RRC Connected Mode

The idea of this topic is to reduce UE's power consumption and the usage of radio resources of the network, while supporting event-driven low latency scenarios. On the architecture level, it focuses on the mode transitions in RRC layer including scenarios which avoid the UE is going into Connected Mode between those transmissions. Only one solution was identified on this topic, also handling the scenario when the NG RAN node, where the UE wants to transmit data, is different from the one that stores the UE context. The solution describes how the UE changes state from *RRC_Inactive* to *RRC_Connected* state while keeping an anchor RAN node always in the loop for the UE context and data transmission.

In the end, SA2 group decided not to continue with this solution for normative work in the Release 16, leaving this key issue unsolved.

F. Division of E2E Packet Delay Budget

The E2E Packet Delay Budget (E2E PDB) consists of two parts:

1) *Access Network delay, and*
2) *Core Network delay.*

In 5G Phase 1 (Release 15), no consideration is given to different deployment scenarios in this respect. As a result, it is assumed that the RAN takes a fixed PDB for the CN for different 5G QoS Identifiers (5QIs). This topic included study whether there is a need for and how to distinguish different CN PDBs towards different UPFs acting as PDU Session Anchor (PSA).

Two solutions were identified for normative specification in Release 16, addressing different deployment options as follows:

- No Intermediate-UPF (I-UPF) used between RAN and PSA UPF – In this case, OAM configures the SMF with PSA-RAN PDB for different 5QIs. The RAN node retrieves, for each QoS flow, the PSA-RAN PDB from the SMF and derives the PDB for the RAN part.
- UPF/Uplink Classifier (ULCL) used between RAN and PSA UPF – in this case, the RAN cannot be aware of the delays for a PDU session between the PSA UPF and N3 UPF if there is no direct connection, i.e. an I-UPF is inserted in the path. The SMF can select the I-UPF and configures the RAN node with an average I-UPF to PSA UPF delay based on statistical analysis.

G. Automatic GBR Service Recovery after Handover

This topic is based on the scenario where a machine requires a Guaranteed BitRate (GBR) service but the current QoS level may not be kept due to reasons such as handover to a congested cell or a temporary overload condition. The goal of the solution is that the radio bearer and QoS flow is not dropped if the QoS cannot be preserved. This is because the machine may be able to adapt to the situation, e.g., a car or train reduces the speed according to the changing QoS condition. Once the network situation improves, the RAN tries to restore the original QoS level as soon as possible, preferably avoiding many signaling messages.

No specific solution is currently selected for this topic for the normative specification and the conclusions are still under discussions in SA2 working group. The solution may be left to implementation by defining specific timer and maximum number of retries in SMF and AF.

III. Security aspects

3GPP SA3 (security) working group has started the study on security aspect of URLLC since the end of 2018. This work resulted in TR 33.825 [9]. This study is currently in progress at the time of this writing, and the subsequent normative specification is expected to be included in the Release 16 version of the 5G security specification [10]. This section discusses the key points from TR 33.825 [9].

A. Security for Redundant Transmission

5G system considers two important criteria for the security of redundant transmissions:

1) Cryptographic separation for radio bearers serving redundant transmissions, and

2) Equal level of security provision comparable to single path transmission.

IPsec ESP and IKEv2 certificate-based authentication are implemented to protect the traffic on N3 reference point. If the redundant user planes are supported by two duplicated N3 tunnels, then the Network Domain Security (NDS)/IP framework to secure the network domain interfaces is reused.

For DC-based redundant user planes support, the solution for the security context is not finalized. There are two possible candidates that support redundant user plane security. The first option reuses the DC security mechanism, where the Master Node (MN) generates the Secondary Node (SN) seed key (K_{SN}) from the Access Stratum (AS) key (K_{gNB}) and a SN counter as a freshness parameter. In this solution, the MN sends the K_{SN} to the SN over the Xn-C interface and the SN uses it to derive further Radio Resource Control (RRC) and UP keys used between the UE and SN for signalling and redundant user plane protection. The second option introduces a new seed key (K_{UR}) specifically for the URLLC purpose. It is derived from K_{gNB} and is used for redundant user plane protection in URLLC services.

B. Support of Security for High Reliability by Redundant Data Transmission in User Plane

Redundant data transmission introduces additional threat surface for attackers to take advantage of the presence of multiple user plane paths. If any one of the two user plane path is compromised, then the whole proposition of URLLC security can collapse. To realize high level of reliability, the 5G system supports confidentiality and integrity protection for user plane data transmitted over multiple paths as the primary feature.

C. UP Security Policy Handling for Multiple PDU Sessions Established for Redundant Data Transmission

Use of redundant user plane data transmission introduces another aspect from the security perspective. The UP security policy for encryption and integrity protection provided by the 5G CN may be different between the user plane transmission paths. The UE may not know which policies are enforced in the network. An attacker may perform jamming on an integrity protected path to prevent forwarding of the user plane data from the gNB to the UPF and simultaneously modify the data on the non-protected path. To address this scenario, two solutions are proposed to address the potential security requirements:

- Encryption and/or integrity protection of user plane data between the UE and the gNB to be enabled for both redundant paths, if it is enabled for one.
- The Master gNB (MgNB) to ensure that the UP security policy is forwarded and used by the Secondary gNB (SgNB) for the two PDU sessions in the redundant data transmission.

D. Security Policy for URLLC Service

Two different scenarios of low latency and high reliability result in two security requirements of extremely fast and stringent security checks. These requirements may lead to different security policies for URLLC services in terms of, e.g., key length and key refresh time.

There is an inherent trade-off between adding security mechanisms and required computation power to achieve the target QoS requirements. For example, adding integrity protection may increase the computation complexity and thus adding an unacceptable delay to the URLLC service. At the same time, not providing integrity protection may open up the URLLC services to possible attacks.

The solution for this issue proposes to use only the integrity protection and encryption settings "Required" or "Not Needed". In other words, the value "Preferred" is not allowed for this solution.

E. Security Aspect of Low Latency Handover Procedure

Optimization in handover procedure is a likely area to ensure that the low latency service is maintained as the UE moves. At the same time, it needs to ensure there is no negative impact to the security aspects. From security perspective, handover procedure involves new key derivation and security algorithm selection as the serving gNB and/or AMF change as the result of handover. Both backward and forward security are required before/during/after the handover to prevent adversary to compromise the communication.

At the time of this writing, the study [9] does not have specific solution for this topic. This area may be addressed at a later time in Release 16 specification work.

F. Retaining AS Security Keys for Redundant Data Transmission in User Plane

In the 5G system, the gNB decides whether to retain or change the AS security keys between the UE and the gNB during intra-gNB handover. For URLLC services, re-generating AS keys at every intra-cell handover may introduce negative performance implications in the gNB, e.g., unwanted delay or latency for the URLLC service. Consequently, AS keys should be refreshed only if there is a security reason to maintain the strict performance constraints.

The concluded solution reuses the existing mechanism specified in TS 33.501 [10], allowing that the AS keys are retained based on the policy in the gNB for Intra-gNB handover.

G. QoS Monitoring Protection

Vertical applications with its E2E stringent QoS requirements expect awareness of the real time communication latency. The 5G E2E QoS monitoring is used to monitor the real time packet delay in 5GC and 5G-AN as discussed in section II-D. In this case, the QoS monitoring messages related to QoS activation and enforcement need to be protected. If there is no sufficient security mechanism in place to protect the E2E QoS monitoring procedure, an attacker can attempt to modify packets or message to report an incorrect latency. Currently, the security mechanism to address this potential open issue is not finalized.

H. Acceleration of Authentication and Key Agreement Procedure for Low Latency

Another security area for improvement for low latency service is the Authentication and Key Agreement (AKA) procedure itself in which the mutual authentication between the UE and the network are achieved and session keys are established between them. The AKA procedures used in 5G system always involve the home network to derive the Authentication Vector (AV) and authenticate the UE. There may be areas of improvement to achieve low latency service.

At the time of this writing, the study [9] does not have specific solution for this topic. This area may be addressed at a later time in Release 16 specification work.

I. Security Aspect of Low Latency Re-authentication Procedure

Release 15/16 5G system does not support fast re-authentication [10]. Fast re-authentication support can be considered to guarantee low latency for URLLC services. Otherwise, time consuming authentication procedure need to be invoked at every registration request from a UE irrespective of their previous authentication with the same network. This has a significant implication for the URLLC services. Currently Release 16 URLLC study has not reached conclusion on this topic.

J. UP Security Performance for Low Latency

At the time of this writing, SA3 group has not reached conclusion on the solution for this topic. The issue is that IPSec and Transport Layer Security (TLS) or Datagram TLS (DTLS) used on the interfaces between gNB and UPF for the UP path introduces transmit delays due to slow forwarding performance. This issue is not shared by all members in the SA3 group. As a result, no security requirements or threats were captured as well as no solution has been agreed at this time.

IV. Outlook and Conclusion

In this paper, we discussed the latest 3GPP work in URLLC from system and security perspectives based on the current status of the ongoing study items in SA2 and SA3 working groups. These study items are expected to be concluded soon followed by work items to start work on normative specifications. They are expected to be completed later in 2019 as a part of Release 16 content. There are topics and issues not concluded in the studies as discussed in this paper, e.g., topics in section II-B, E, G, and section III-A, C, E, G, H, I, J These topics may be left open as the future work beyond Release 16 time frame.

References

[1] 3GPP TS 22.104, v16.1.0, Service requirements for cyber-physical control applications in vertical domains, March 2019.

[2] 3GPP TS 22.261, v16.7.0, Service requirements for the 5G system; Stage 1, March 2019

[3] 5G Americas White Paper, New Services & Applications with 5G Ultra-Reliable Low Latency Communications, November 2018.

[4] GSMA White Paper, From Vertical Industry Requirements to Network Slice Characteristics, August 2018.

[5] 5G-ACIA White Paper, 5G for Connected Industries and Automation, November 2018.

[6] 5GAA White Paper, Timeline for deployment of C-V2X – Update, January 2019.

[7] 3GPP TR 23.725, v16.1.0, Study on enhancement of Ultra-Reliable Low-Latency Communication (URLLC) support in the 5G Core network (5GC), March 2019

[8] 3GPP TS 23.501, v16.0.2, System Architecture for the 5G System;Stage 2, April 2019

[9] 3GPP TR 33.825, v0.5.0, Study on the security of Ultra-Reliable Low-Latency Communication (URLLC) for 5GS, May 2019

[10] 3GPP TS 33.501, v15.4.0, Security architecture and procedures for 5G System, March 2019

[11] 3GPP TS 23.503, v16.1.0, Policy and charging control framework for the 5G System (5GS), June 2019.