Received: 9 May 2020 | Revised: 15 August 2020 | Accepted: 10 September 2020

DOI: 10.4218/etrij.2020-0200

SPECIAL ISSUE

ETRI Journal WILEY

Ultra-low-latency services in 5G systems: A perspective from 3GPP standards

Sunmi Jun | Yoohwa Kang | Jaeho Kim | Changki Kim

Network Research Division, Telecommunications & Media Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea

Correspondence
Yoohwa Kang, Network Research Division, Telecommunications & Media Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea.
Email: yhkang@etri.re.kr

Funding information
This research was supported by an Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2015-0-00042-005, Development of Wired-Wireless Converged 5G Core Technologies) and (No. 2020-0-00974-001, Development of ultra-reliable and low-latency 5G+ core network and TSN switch technologies).

Recently, there is an increasing demand for ultra-low-latency (ULL) services such as factory automation, autonomous driving, and telesurgery that must meet an end-to-end latency of less than 10 ms. Fifth-generation (5G) New Radio guarantees 0.5 ms one-way latency, so the feasibility of ULL services is higher than in previous mobile communications. However, this feasibility ensures performance at the radio access network level and requires an innovative 5G network architecture for end-to-end ULL across the entire 5G system. Hence, we survey in detailed two the 3rd Generation Partnership Party (3GPP) standardization activities to ensure low latency at network level. 3GPP standardizes mobile edge computing (MEC), a low-latency solution at the edge network, in Release 15/16 and is standardizing time-sensitive communication in Release 16/17 for interworking 5G systems and IEEE 802.1 time-sensitive networking (TSN), a next-generation industry technology for ensuring low/deterministic latency. We developed a 5G system based on 3GPP Release 15 to support MEC with a potential sub-10 ms end-to-end latency in the edge network. In the near future, to provide ULL services in the external network of a 5G system, we suggest a 5G-IEEE TSN interworking system based on 3GPP Release 16/17 that meets an end-to-end latency of 2 ms.

KEYWORDS

3GPP standardization, 5G system, mobile edge computing, time-sensitive communication, ultra-low-latency services

1 | INTRODUCTION

The demand for services such as three-dimensional interactive data exchange, augmented reality (AR)/virtual reality (VR), the tactile Internet, factory automation, and autonomous driving is growing rapidly in the vertical domain as well as the consumer domain. Such latency-sensitive services have not been supported in the 4G system, because it was only aimed at increasing the user service capacity. In addition to supporting broadband service, a fifth-generation system (5GS) considers how to support reliable and low-latency communication for latency-sensitive services.

To achieve this goal, the 3rd Generation Partnership Party (3GPP), a standardization organization for developing 5GS, classifies service scenarios and requirements using key performance indicators (KPIs) for 5GS, which are used to assess the performance of services. 3GPP technical specification (TS) 22.261 [1] presents various use cases and quantifies their service requirements with different combinations of KPIs (see Table 1) to support different services and different end-user communities.

TABLE 1 Different combinations of KPIs for 5G service use cases

KPIs	Use cases
High data rates High traffic densities	Urban/Rural macro—the general wide-area scenario in urban/rural areas Indoor hotspots Broadcast-like services High-speed trains/vehicles Airplane connectivity
High reliability Low latency	Cyber-physical control applications in vertical domains—industrial factory automation and energy automation Vehicle-to-everything (V2X) communication Rail communications Industrial automation
High availability	Medical monitoring
High data rates Low latency	AR/VR Interactive conversation Telesurgery

One of these KPIs, end-to-end latency, is the time it takes to transfer data in one direction, uplink or downlink, between the application server and the end user in a 5GS [1]. The fulfillment of low end-to-end latency is very critical for realizing emerging services in both consumer areas and vertical industries over 5GS. The time-critical operation of the factory automations presented in Industry 4.0 requires the lowest end-to-end latency, below 1 ms [2]. Automated driving systems are also required to have an end-to-end latency between 10 ms and 100 ms, and they should be able to exchange emergency messages within 10 ms for automated road safety. Robotics and telepresence services also require remote-controlled robots with real-time haptic feedback to satisfy an end user's sense of reality through an end-to-end latency of less than a few milliseconds [2,3]. Telesurgery systems require the end-to-end latency to be less than 10 ms to support surgeons who operate remotely and need haptic feedback urgently [3]. Therefore, 5GS has to provide an end-to-end latency from 1 to 10 ms to support the above services.

However, it has not been possible to achieve the performance of 1 ms latency before 5G. The average round-trip times in 3G and 4G are 63.6 ms and 53.1 ms, respectively [3]. In the case of a 4G radio access network (RAN), the overall radio access delay may take 1 ms optimally during downlink transmissions but may take up to 17 ms during uplink transmissions. Therefore, the 4G system cannot meet the requirement that the end-to-end latency should be between 1 ms and 10 ms, although the technologies for low latency at the RAN level were studied in Ref. [2–5].

In the next version, 5G New Radio (NR) can achieve this low-latency requirement because NR has a flexible orthogonal frequency-division multiplexing (OFDM) with short transmission time intervals (TTIs), short frame structure, and shorter OFDM symbols. As a result, 5G NR latency can be reduced to 0.5 ms one way [5]. In addition to 5G NR latency, end-to-end latency considers latencies in the backhauls and between the core network (CN) and data center (DC). Hence, a 5GS still has some challenging issues to achieve an end-to-end latency of sub-10 ms in the overall network. For example, if the distance between the CN and the DC is 3000 km, the latency will be more than 10 ms [6]. Therefore, a 5GS requires a revolutionary network architecture to reduce the end-to-end latency to below 10 ms.

In order to reduce the end-to-end latency, 3GPP introduces network-level solutions such as mobile edge computing (MEC) into 5GSs, which enables an operator and third-party services to be hosted close to the access point of attachment of the user equipment (UE) [7] to help the 5GS achieve an end-to-end latency of less than10 ms. In addition, 3GPP is making an effort to standardize time-sensitive communication (TSC), which is a solution for interworking the 5GS with IEEE time-sensitive networking (TSN) synchronized to a very accurate and precise clock source. 3GPP standardization is still ongoing, but TSC is expected to be a revolutionary network architecture to fulfill end-to-end ultra-low latency (ULL).

The rest of this paper is organized as follows. Section 2 presents the progress of 3GPP standardization to support end-to-end low latency focusing on MEC and TSC. Section 3 introduces a test bed that implements MEC based on 3GPP Release 15 and presents the actual amount of latency reduction. Section 4 describes some challenging issues to be considered when developing a TSC under standardization in 3GPP Release 16/17. Finally, concluding remarks are provided in Section 5.

2 | STANDARDIZATION OF ULL

As mentioned before, 5G radio network requires at least 1 ms of user plane latency for ultra-reliable low-latency communication (URLLC) services [8]. 5G NR met the latency requirements for URLLC service in 3GPP Release 15/16, but the 5G Core (5GC) network still lacks low-latency technology development. Currently, the 3GPP Service and System Aspects (SA) Working Group 2 (WG2) is developing standard technologies for low latency, and we will look at MEC and TSC technologies as standard technologies related to low latency in 3GPP SA WG2 in addition to other non-3GPP standard technologies.

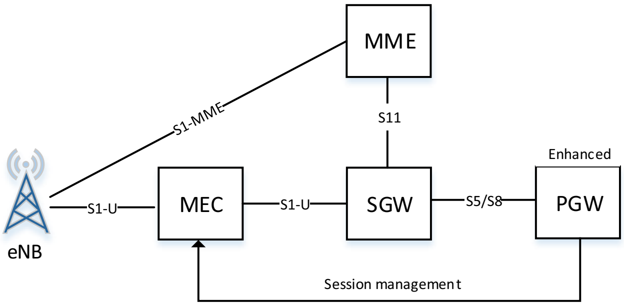
Mobile carriers have tried to introduce MEC from 4G for low-latency service, but the serving gateway (SGW) and packet gateway (PGW) are complex structures that process signals as well as data. Because of this structure, users who want to communicate locally also have a problem because they must transmit data to the PGW. Therefore, in a 4G system, rather than proceeding with MEC standardization, a third-party MEC solution such as Figure 1 was added to the edge located between the RAN and SGW to provide low-latency services [9].

On the other hand, unlike the 4G Evolved Packet Core (EPC), a 5GS has a structure in which the control plane and user plane are separated and a user plane function (UPF) for data transmission is distributed, so it is easy for 5GS to install a separate UPF for MEC in a local data network (DN). In addition, the session management function (SMF) can control the MEC packet and the packet routing to the DN together. We introduce in detail the enablers for 5G MEC standardized by 3GPP in Section 2.1.

In addition, 3GPP Release 16 defined cyber-physical control service requirements for IEEE TSN-based industrial networks where various Industrial Internet of Things (IIoTs) are connected [10]. To satisfy this, 3GPP 5GS has standardized TSC technology that can interwork with IEEE TSN. Release 17 defines the requirements of new low-latency services such as video, imaging, and audio for professional applications (VIAPA) beyond the factory domain, and such new services may require synchronization with the grand master (GM) clock of a 5GS. In the future, Release 18 is expected to provide time synchronization services by driving the 5GS itself as a 5G timing resiliency system. For this purpose, it is believed that standardization such as a mechanism that allows the 5GS to stably maintain the clock source and distribute it to users as well as a method to back up the external clock when a problem occurs with the GM clock of a 5GS will be necessary.

2.1 | 3GPP SA WG2 activity—MEC

Some efforts have been made to deploy low-latency services by deploying MEC servers in 4G networks as well [11–13]. Reference [10] introduces the concept of MEC and [12] deploys a fog gateway between the EPC and Long Term Evolution (LTE) RAN, which reduces transmission delay by filtering and processing low-latency packets at the fog gateway. Reference [13] showed that the end-to-end latency can be reduced to 5 ms by deploying a URLLC server in an edge network with a cellular environment (3.5 GHz, 800 MHz). Consequently, MEC is recognized as a major function for low-latency services, and 3GPP SA2 defines standards for 5G network architecture and functions in order to support MEC functions for low-latency services in a 5GS.



© European Telecommunications Standards Institute 2018. Further use, modification, copy and/or distribution are strictly prohibited.

FIGURE 1 MEC deployment in 4G [9]

2.1.1 | MEC in Release 15/16

Edge computing delivers low-latency service by placing the application server close to user access, thereby reducing the end-to-end delay. In a 5GC network, the UPF is located close to the UE to steer traffic to the local DN according to the UE's subscription data, UE location, information from application function (AF), or policy. In addition, a 5GS can support edge computing with a combination of the following functions:

(1) User plane (re)selection: 5GC (re)selects via the UPF to transmit traffic to the local DN.
(2) Local Routing and Traffic Steering: 5GC selects the protocol data unit (PDU) session anchor (PSA) with an uplink classifier (UL CL) and Internet Protocol version 6 (IPv6) multi-homing function and routes traffic to the local DN.
(3) Session and service continuity (SSC): SSC mode supports mobility between UE and application, and SSC mode 2/3 provides service continuity even when the PSA is changed.
(4) AF-influenced traffic routing: The AF performs (re) selection of the UPF through a policy control function (PCF) or network exposure function (NEF) to route traffic to the local DN.
(5) Network capability exposure: 5GC and AF can exchange information with each other through the NEF.
(6) Local Area DN (LADN): 5GC connects the LADN with applications.

The representative architectures supporting MEC in 5GS are the UL CL and branching point (BP). Figure 2 shows the architecture where a UL CL has been added; one PDU session has two PSAs in this architecture. In the figure, the UL CL can support edge computing by sending traffic to PSA 2 connected to the local DN in addition to PSA 1.

Figure 3 shows a BP, another 5GC network architecture that can support edge computing. A UPF with a BP function can transmit uplink traffic to different PSA 1's and PSA 2's using multiple IPv6 addresses of the PDU session. In this architecture, a multi-homed PDU session has the advantage of supporting make-before-break service continuity (SSC mode 3) during handover. Figure 3 shows a 5G network architecture that provides connectivity between BP and PSA 2 connected to the local DN to support edge computing.

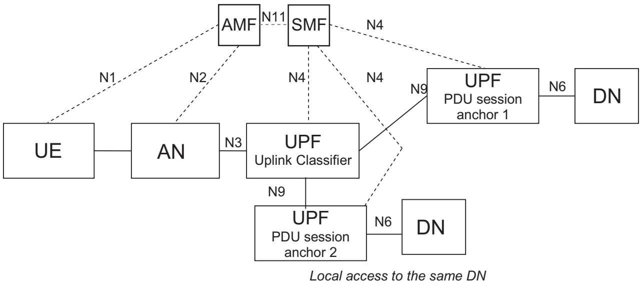


FIGURE 2 5GC network architecture for the UL CL [7]

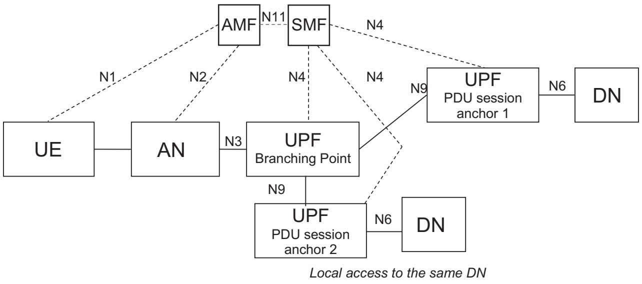


FIGURE 3 Multi-homed PDU session: access to a local DN [7]

2.1.2 | MEC in Release 17

Edge computing, which allows operators to place applications or content close to users, is considered to be one of the major technologies that can meet the requirements of the ultra-broadband and ULL services of 5G by deploying UPFs on the edge network in a distributed manner. Locally deployed UPFs supporting UL CL or BP was defined in Release 15 to support edge computing by performing LADN local routing, traffic steering, user plane (re)selection, and AF-influenced traffic routing [7]. However, issues still remain such as the IP discovery of local application servers and support for seamless application migration.

To solve these issues, a study on the enhancement of support for edge computing in 5GC (FS_enh_EC, 3GPP technical report (TR) 23.748) investigated 5GS enhancements for supporting edge computing [14]. First, the UE must be able to discover the IP address of the application server deployed in the edge computing environment in order to use applications/contents. In addition, it must be possible to seamlessly change the application server of the UE, and a method of providing a local application server capable of low-latency service is also needed. The UE can access the local site and the central site at the same time, and it may also have multiple PDU sessions as a local PSA and a central PSA. According to [14], a 5GS supports a "distributed anchor point" that exists in the local site for edge computing traffic, and one PDU session can have a PSA, which can be changed during service, at both the central site and the local site.

FS_enh_EC has proposed the following key issues (KIs) to be addressed, and it is expected to work on the Release 17 standard specifications after its completion in December 2020.

(1) KI: Discovery of an edge application server (EAS). Because one application service can be provided from multiple EASs existing at different sites, considering the service latency, traffic routing path, quality of experience (QoE), and other metrics, one optimal EAS is selected. It must be changeable to another optimal EAS because of user movement and server congestion.

(2) KI: Edge relocation.
In the edge computing environment, UE mobility and optimal application server relocation must be supported. Coordination with (local) PSA due to EAS relocation should also be improved to support seamless change.

(3) KI: Network information provisioning to local applications with low latency.
Interaction for network information exposure is required between a 5GS and edge computing functions, and the current Release 16 network exposure mechanism is designed so that network functions (eg, NEF and PCF) related to network exposure operate in the center. However, in practice, long exposure latency may occur because an EAS or AFs are locally deployed, so exposure information must be transmitted to edge computing functions with low latency.

(4) KI: Activating the traffic routing towards the local DN per AF request.
The AF requests a DN access identifier (DNAI), which is an identifier of a user plane access to one or more DNs where applications are deployed, to activate traffic routing. Then, the SMF activates traffic routing to the local DN by setting the requested DNAI.

2.2 | 3GPP SA WG2 activity—TSC

A 5GS supports a variety of vertical services such as vehicle-to-everything (V2X) communication, VR/AR, and factory automation, and these vertical service applications require time-sensitive (ie, deterministic) services as well as ultra-low delay services. Therefore, 3GPP 5GS is in the process of working on the standard specifications of the system extension function to support TSC [15] for ULL and time-deterministic services in Release 16. IEEE TSN is a standard technology that provides low latency and deterministic data transmission in Ethernet networks, and as shown in Figure 4, 5GS is intended to support integration with TSN networks. 3GPP Release 16, which should be completed in June 2020, is developing a 5GS extension function and time synchronization technology standard that allows 5GS to act as a TSN bridge, and Release 17 has been studying

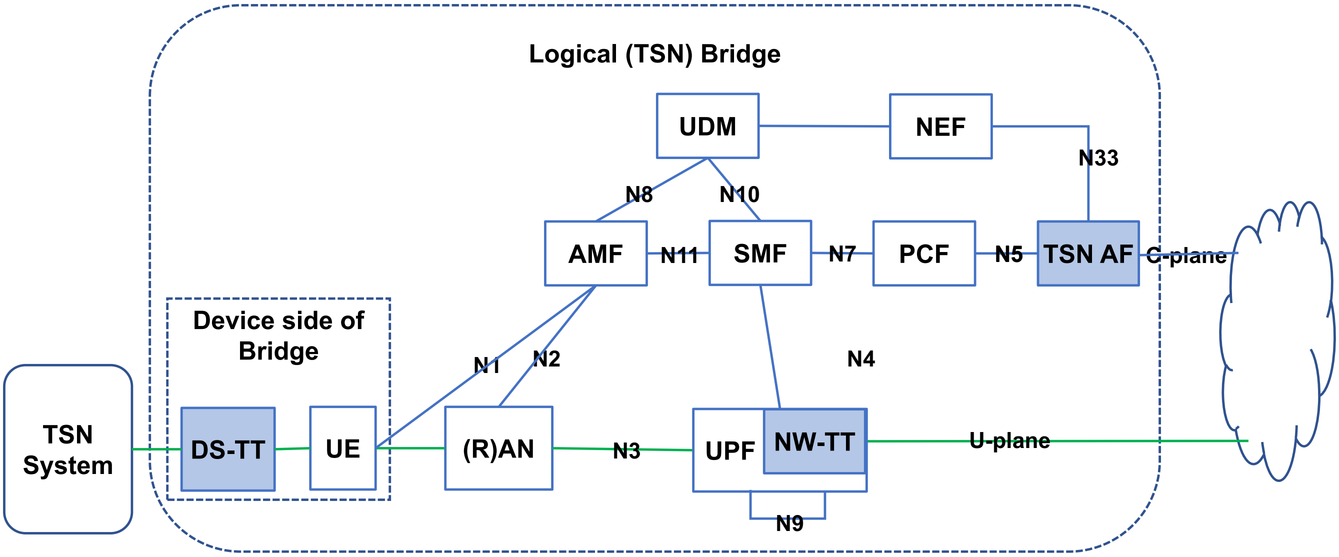


FIGURE 4 System architecture view with the 5GS appearing as TSN bridge [7]

a standard for the development of 5GS extensions for IIoT services.

2.2.1 | TSC in Release 16

3GPP Release 16 is working on standard specifications for the integration of 5GS and TSN networks. In Figure 4, 5GS has an additional TSN translator (TT) function to operate as one TSN bridge by connecting to the TSN network. For interworking between the TSN system and 5GS, the 5GS has a device-side TSN translator (DS-TT) and network-side TSN translator (NW-TT) in the UE and UPF, respectively, and it can provide TSN ingress and egress ports through DS-TT and NW-TT. A 5GS can operate as one logical TSN bridge per UPF, and the 5GS bridge is composed of NW-TT port, DS-TT port, and user plane tunnel between UE-UPF.

The TSN configuration model defined in IEEE 802.1Q provides three models: fully centralized, fully distributed, and hybrid, but Release 16 supports only the fully centralized model. A TSN AF is responsible for transferring information about the 5GS bridge and TSN network configuration to each other. The TSN AF also receives TSN traffic information and delivers QoS information of the TSN traffic to the UPF through the PCF and SMF.

In order to provide TSN time synchronization, 5GS operates as a TSN bridge compatible with IEEE 802.1AS, and TTs support the IEEE 802.1AS technology used in TSN domain synchronization. In the 5G time domain, UE, next-generation Node B (gNB), UPF, NW-TT, and DS-TT are synchronized to a 5G GM clock, and 5GS synchronization uses 5G RAN synchronization technology (Figure 5).

2.2.2 | TSC in Release 17

3GPP TS 22.104 [10] presents service requirements for cyber-physical control applications in the vertical domain, and various vertical services such as factory automation, smart grids, and robotic-assisted surgery presented here require URLLC. Therefore, a study on enhanced support of IIoT in the 5G System (5GS) (FS_IIoT, TR 23.700-20) [16] to improve the 5GS function has been conducted so that URLLC services can be provided by interworking with the TSN network in the industrial network. The main KIs to be discussed in Release 17 FS_IIoT are as follows.

(1) KI: Uplink time synchronization.
In Release 16, the TSN GM clock is located on the network side with UPF, and the packet for TSN synchronization is processed as downlink traffic within a 5GS. However, in Release 17, considering the case where the TSN GM is located on the device side, the synchronization packet must be processed as uplink traffic. As shown in Figure 6, the GM of TSN domain 1 is located on the UE side, and a generic Precision Time Protocol (gPTP) message for TSN synchronization must be transmitted to TSN domain 1 connected to another UE. The figure shows an example in which the gPTP message delivered to the UPF is transmitted to another UE in the same domain again as a downlink path.

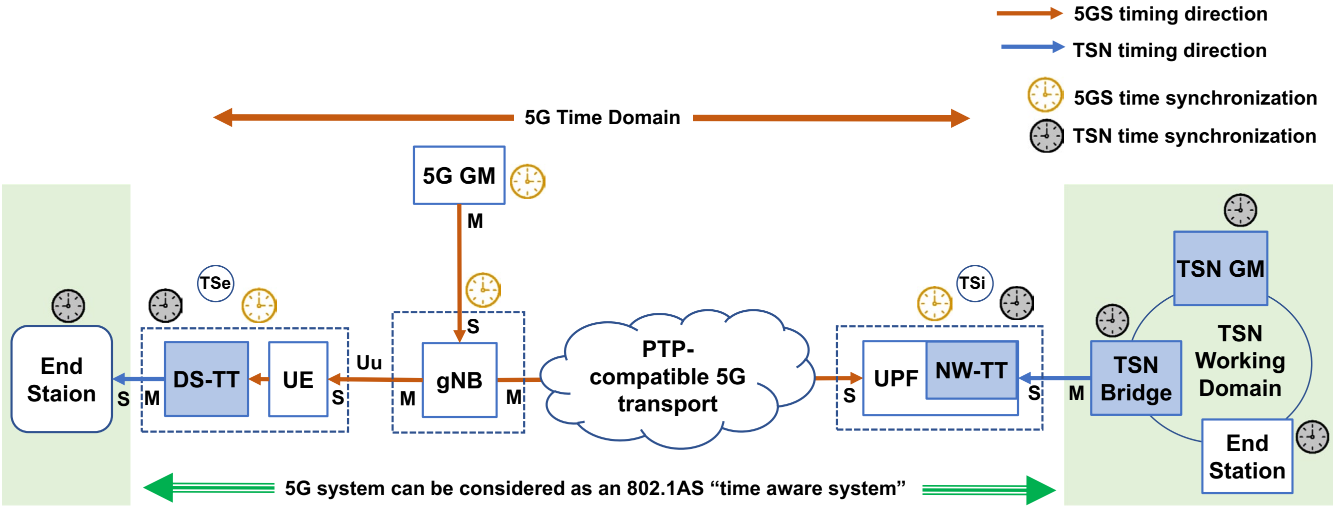


FIGURE 5 5GS for supporting TSN time synchronization [7]

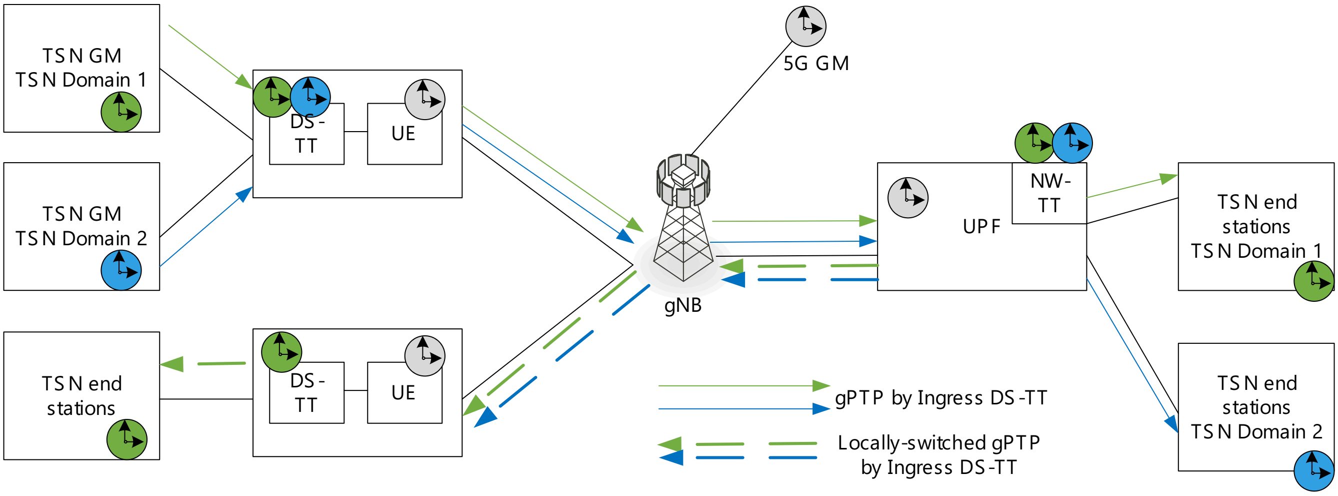


FIGURE 6 Distribution of UL time synchronization [10]

(2) KI: UE-UE TSC communication.

As shown in Figure 7, UE-UE communication is required when different UEs connected to one UPF belong to the same TSN domain. Existing mobile communication supports only communication between the UE and the DN, and this UE-UE communication method is a new type in mobile communication. In particular, in a delay-sensitive communication method such as TSC, a 5GS requires a new delay management method between UE and UPF and a new forwarding method for UE-UE traffic routing.

(3) KI: Exposure of TSC services.

Exposure of deterministic QoS and exposure of time synchronization. A 5GS has a network exposure framework based on the NEF, and the 5GS intends to provide network capabilities related to TSC and URLLC through the NEF to more flexibly provide TSC and URLLC services. In addition, AF requires a 5GS to apply delay, jitter, deterministic QoS, and other requirements of applications such as video and audio. For example, 5GS capability is exposed so that AF can turn ON/OFF the time synchronization function. To satisfy requirements from TS 22.263 [17] for VIAPA, 5GS provides a time synchronization service for an IP-type PDU session as well as Ethernet. At this time, the AF may request a method using a 5GS time source as well as the time synchronization method defined in Release 16.

(4) KI: Use of survival time for deterministic applications in a 5GS.

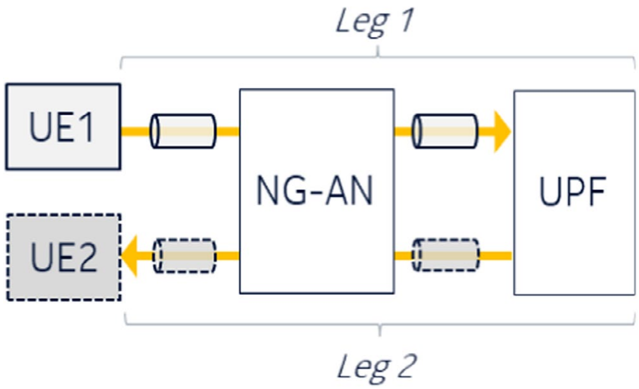


FIGURE 7 UE-UE TSC communication [16] [Colour figure can be viewed at wileyonlinelibrary.com]

Survival time, one of the periodic deterministic communication service performance requirements introduced in TS 22.104 [10], is used to support deterministic applications in 5GS. How a 5GS obtains survival time and transfers survival time to RAN should be studied.

2.3 | Non-3GPP standardization activity

This section introduces non-3GPP standardization activities for supporting low-latency services. These activities are mainly conducted by the European Telecommunications Standards Institute (ETSI), Institute of Electrical and Electronics Engineers (IEEE), and Internet Engineering Task Force (IETF).

The ETSI MEC Industry Specification Group has standardized multi-access edge computing. The purpose of this standard is to provide an environment that can easily run third-party applications, meet low-latency service requirements, and support a variety of access networks such as WiFi, FTTx, and 3GPP access. This group defines the MEC framework and reference architecture independently of the access network and mobile network evolution, which enables MEC servers deployed in a 4G network to be reused in 5G and other networks. This access-independent architecture allows MEC servers to be deployed at a wide range of edges, including cloud RAN, base stations, CNs, and DCs. Therefore, it is possible to support low-latency service by deploying the MEC server in the proximity of the user. In addition, it is possible to meet low-latency requirements by routing low-latency packets for MEC applications to the LADN if using traffic steering, one of the key functions of the MEC platform. This standard provides a solution for deploying and integrating MEC in 3GPP's 5G network using the AF as well as an environment in which MEC servers can allocate network resources more efficiently and improve QoE by utilizing the context information collected from radio, network, and devices [18,19].

In addition to studies by ETSI, there have been many studies to deploy MEC in 5GS. Referring to Ref. [20–23], Table 2 summarizes the technologies required to effectively implement and deploy MEC in a mobile communication network. While most of the proposed technologies have now been standardized, some of the technologies with high potential to

TABLE 2 MEC technologies related to 5GS

	Technologies	Related specifications
(1)	Framework and architecture	• 3GPP TS 23.558: Architecture for Enabling Edge Applications (EDGEAPP) • 3GPP TS 23.222: Common API Framework (CAPIF) for Enabling MEC Application • ETSI GS MEC 003: Common MEC Application Enablement Framework, Reference Architecture • ETSI GS MEC 009: General Principles of MEC Service APIs
(2)	Service APIs management APIs	• 3GPP TS 23.434: Service Enabler Architecture Layer for Verticals (SEAL) • 3GPP TS 23.286: Application Layer Support for V2X • 3GPP TR 23.745: Application Layer Support for Factories of the Future (FotF) • ETSI GR MEC 021: Service Specific Related APIs for Mobility • ETSI GR MEC 022: Service Specific Related APIs for V2X • ETSI GR MEC 033: Service Specific Related APIs for IoT • ETSI GS MEC 010: Management APIs
(3)	Enablers for 5G MEC	• 3GPP TS 23.501: Edge Computing for 5GS • 3GPP TR 23.748: Enhancement of Support for Edge Computing in 5G Core Network • ETSI is studying on MEC integration in 5G network [19]
(4)	Enablers for virtualization	• 3GPP TS 23.222: API invokers and service APIs on-boarding/off-boarding, register/deregister, discovery, and authorization by using CAPIF • ETSI GR MEC 017: the study on MEC in the network function virtualization (NFV) environment
(5)	Computation offloading	N/A
(6)	Resource allocation	N/A

be used have been mainly discussed in the research domain instead of being standardized.

3GPP or ETSI has been standardizing the following four technology groups. In Table 2, technology group (1) provides a common framework and architecture that can execute MEC services by deploying MEC systems independently in the mobile communication system environment. In particular, ETSI defines interfaces for harmonizing with the 3GPP architecture [24]. Technology group (2) is essential to enable developers to quickly develop MEC applications and conveniently manage MEC services using standardized APIs. Technology group (3) is required in order to efficiently deploy and operate MEC systems in 5GS, and 3GPP provides the basic functions of enablers for 5G MEC. Technology group (4) enables MEC systems to run in virtualization infrastructure and supports the portability of MEC service in addition to service continuity by providing migration of MEC service when user movement occurs.

Technology groups (5) and (6) have not yet been standardized, but they are essential technologies to realize MEC services. Technology group (5) guarantees a user’s QoE by offloading the intensive computational tasks to edge nodes or edge servers [21]. In a decentralized MEC environment, technology group (6) is required for distributed data processing among multiple edge nodes/servers that have limited resources and explosive mobile traffic growth [20]. In technology group (6), resources should be allocated to nodes based on radio and network conditions. Therefore, standards for these technologies are needed to define interface and exchange management information between 5GS AF and MEC systems.

IEEE and IETF are also working to standardize network systems for guaranteeing ULL services [25]. It is fundamentally difficult for Ethernet to support ultra-low and deterministic latency service because the medium access control (MAC) layer is based on a carrier-sense multiple access/collision detection (CSMA/CD) mechanism. Hence, the IEEE 802.1 TSN Task Group provides standardization of functions for deterministic services with low latency, low-latency deviation, and reliable transmission based on the Ethernet bridge and MAC layer. TSN guarantees the deterministic latency at bridges by adding a time-deterministic forwarding/queueing function into the existing IEEE 802.1 bridge, such as Per-Stream Filtering and Policing (802.1Qci) or Cyclic Queuing and Forwarding (802.1Qch). In addition, TSN provides the reliability to deliver TSN streams to a destination without loss, even if one or more failures occur in the network path through frame replication and elimination for reliability (FRER, 802.1CB). It also supports streaming service by improving the existing Stream Reservation protocol (802.1Qcc) [25,26]. Extending from the TSN, IETF deterministic networking (DetNet), whose aim is to standardize deterministic networking technology, is a network layer technology for an IP/multi-protocol label switching (MPLS)-based network. DetNet QoS mechanism for ULL follows the QoS mechanism of TSN to prevent the loss due to congestion and ensure deterministic latency. In the data plane, DetNet defines FRER, extended from TSN, to support lossless transmission. The detailed technology such as buffering, resource allocation, and data forwarding is being standardized for both IP and MPLS [25,27].

3 | MEC IMPLEMENTATION

In the previous section, MEC and TSC were closely reviewed as standard technologies for low latency in 5G networks. This section introduces the implementation of a 5GS supporting MEC. To build a 5GS test bed, we developed a 5G CN composed of 3GPP Release 15/16 CN functions (eg, the access and mobility management function (AMF), SMF, and UPF). We also built a RAN test bed including a UE by utilizing the Universal Software Radio Peripheral (USRP) platform (Figure 8). The RAN test bed was built using L1/L2/L3 provided by a 4G-based Software Defined Radio (SDR) platform and USRP [28]. The 5G Non-Access Stratum protocol was developed to operate on the RAN test bed.

In the test bed of Figure 9, the three types of UPFs deployed to support the MEC function are as follows: 1) UL CL UPF, 2) DN PSA (D-PSA) UPF, and 3) edge network PSA (E-PSA) UPF. UL CL transfers packets for LADN to E-PSA and packets for DN to D-PSA by filtering packets among the uplink user traffic. The operation administration maintenance (OAM) presents the monitoring of signal procedures and user traffic flows among 5G CN functions.

Figure 10 shows an OAM screenshot of the 5G CN test bed composed of 5G CN functions (eg, AMF, SMF, and UPF (UL CL or PSA)), 5G base stations, game streaming servers in the DN and the LADN, and UEs (Figure 11). The game streaming service requires a low round-trip latency time until the user's game manipulation input is transmitted to the server and the user terminal plays the game video [29]. It also needs a large bandwidth capacity to transmit the rendered

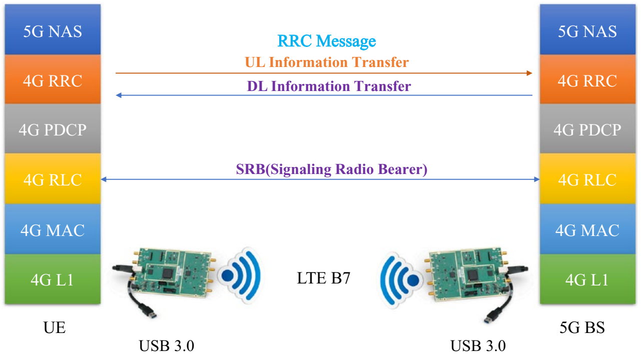


FIGURE 8 Implementation of RAN

game video [30]. Therefore, the game streaming service is most suitable for application to an MEC service that requires low latency and high bandwidth.

In this test, to compare the latency between DN and MEC services, the same game streaming servers were deployed in the DN and the LADN, respectively. UE 1 was connected to the game streaming server in the DN via D-PSA1, and UE 2 was connected to the server in the LADN via E-PSA (Figure 10). Instead of actually connecting the DN to the Internet, an Internet path emulator was installed in the DN to simulate the Internet. This emulator provides a configuration in which various latencies and packet loss can be set. The parameters were measured in the path between the actual test network and the external Internet and were set as follows [31].

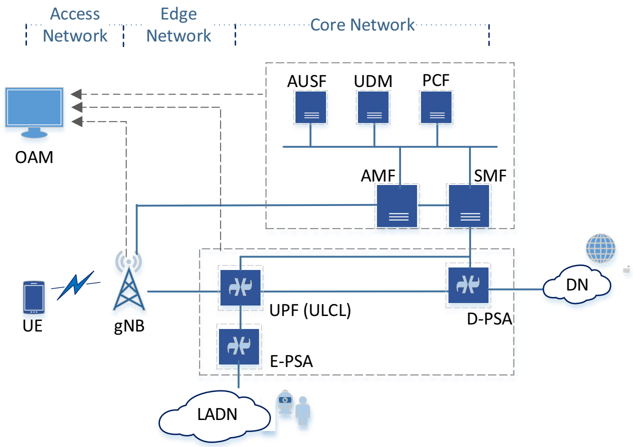


FIGURE 9 Test bed for supporting MEC

- Latency: 60 ± 5 ms with the next random element depending 25% on the last packet sent.
- Packet loss: 0.001% of the packets were randomly lost, and each successive probability depends by 25% on the last one.

First, the latency of our RAN test bed was measured, and then, the interactive latencies between the user input of the game streaming service and the game rendering video were measured in both the DN and LADN. The measurement results of game streaming service showed that the latency in the RAN was 4 ms on average, the round-trip latency between UE 1 and the DN was 128 ms, and the round-trip latency between UE 2 and the LADN was 6 ms on average.

In the previous section, we introduced several case studies showing the results of reducing service latency using edge computing technology in the 4G system [6,11–13]. In [6], a field test was performed to compare latencies between public and edge networks in a 4G-based test bed. The results showed the latency of the edge network to be 15 ms–17 ms. In addition, it showed that the latency at the edge network was reduced by 60% to 91% compared to that of the public network. In [32], as a result of comparing latencies between a fog gateway connected

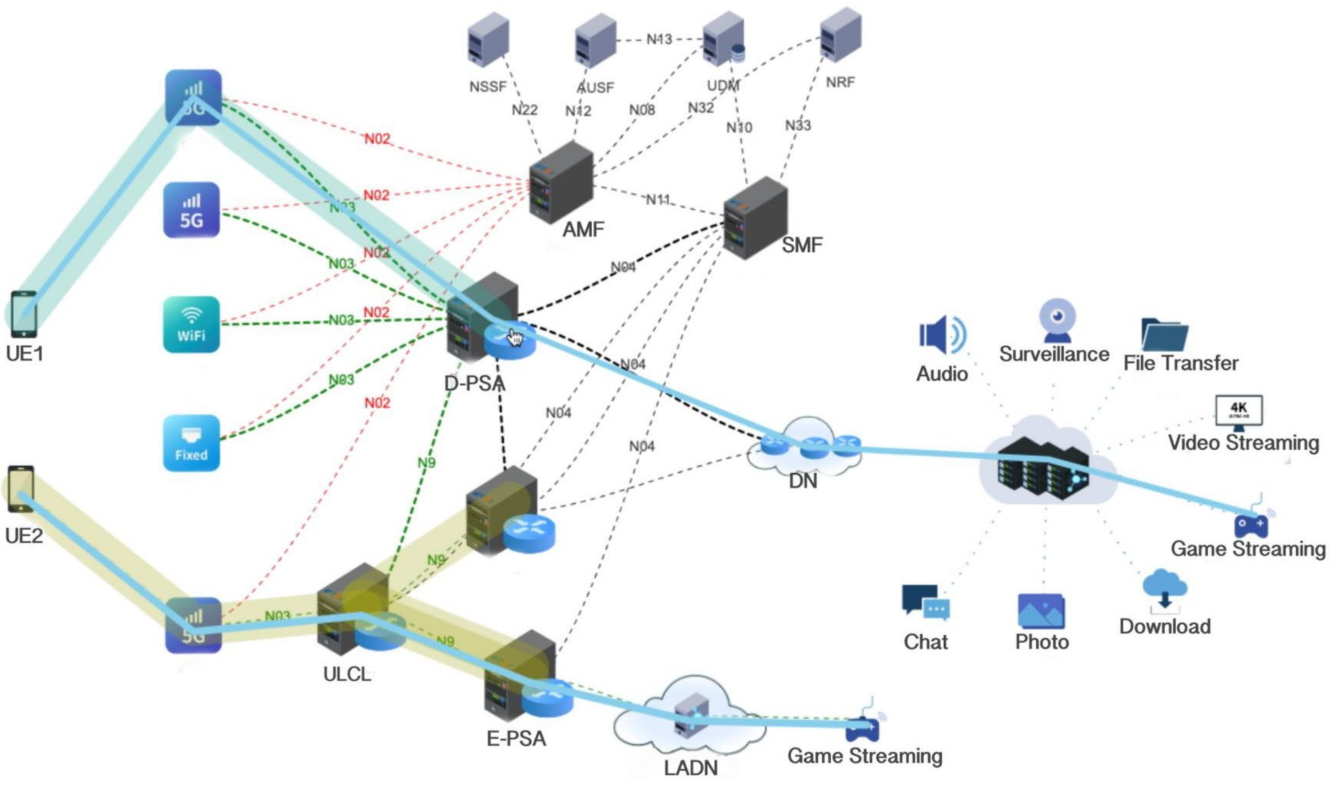


FIGURE 10 OAM screenshot of the 5G CN/MEC testbed

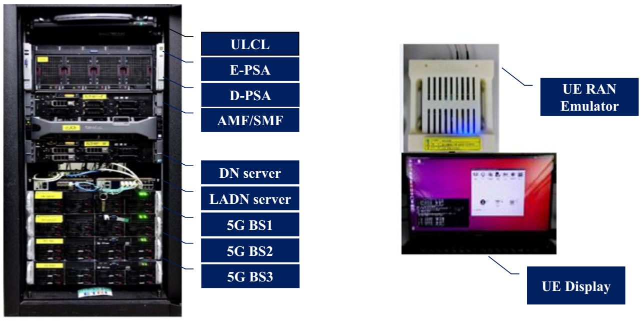


FIGURE 11 5G CN/MEC test bed: 5G CN functions, UEs, and game streaming servers in both DN and LADN

to a LADN and a serving gateway connected to public DN, the latency of the serving gateway was 58.179 ms on average and the latency of the fog gateway was 12.5 ms on average.

In the 5G CN/MEC test bed that we implemented, the measurement results show that the service had a latency of more than 100 ms in the DN simulating the Internet environment, but the MEC service had a low latency of sub-10 ms in the LADN. Although the latency in the 4G-based RAN test bed was measured at 4 ms, considering the current 5G NR URLLC latency is less than 1 ms, MEC service latency can be further reduced.

4 | CHALLENGING ISSUES IN TSC

Section 3 presented our implementation of 5GS test bed and the measurement results of MEC service latency in our test bed. It also showed that low-latency transmission is possible by supporting MEC function in the Release 15/16 5GS. However, the end-to-end latency with a server in the 5GS external network varies greatly from up to several tens of milliseconds to hundreds of milliseconds depending on the number of gateway hops and the routing configuration between the UE and server [33]. Therefore, a 5GS based on 3GPP Release 15 with MEC function cannot ensure the end-to-end latency of service in a 5GS external network. In particular, vertical services such as factory automation require ULL performance of sub-1 ms. To support these low-latency vertical services, further research is needed to satisfy the performance of end-to-end ULL in the industrial network outside the 5GS. For this, in Section 2, TSC was introduced as an effort to standardize interworking between 5GS and IEEE TSN in 3GPP Release 16/17.

In the interworking between 5GS and TSN, ensuring ultra-low and deterministic latency service is a challenging issue. For implementing this interworking, we would like to suggest that the following three KPIs be satisfied.

(1) KPI: 1-μs-level clock error.
The clock synchronization error between TSN end stations interconnected over a 5GS as a virtual bridge should be guaranteed to be 1 μs or less. This error is the highest level requirement of ULL applications among clock error levels defined in Ref. [25].

(2) KPI: 2-ms-level end-to-end latency.
3GPP defines the packet delay budget requirement of delay-critical guaranteed bit rate (GBR) service class to be less than 5 ms and that of the 5G CN to be 2 ms [7]. However, considering the latency of 0.5 ms at 5G NR, the end-to-end packet latency between TSN end stations in the industrial network should be considered less than 2 ms to provide ULL service.

(3) KPI: Sub-10-μs-level latency deviation.
The end-to-end latency deviation between TSN end stations should be within 10 μs. This latency deviation is based on the value defined in Ref. [25]. To satisfy low-level latency deviation, the packet arrival time should be deterministic.

In order to implement a TSN interworking 5GS that can satisfy the proposed KPIs, a new system architecture is proposed (Figure 12). The main considerations required in the design of the proposed system are as follows.

First, there is a need for a new system architecture that ensures a one-way end-to-end latency of 2 ms between TSN end stations. Since 5G NR provides sub-1 ms latency for URLLC service, it is necessary to design an enhanced 5G CN function so that the latency of the 5G CN section is within 1 ms.

Second, the new system architecture needs a high-precision time synchronization technology to be applied to 5GS and TSN, and this also needs a mechanism to deliver Precision Time Protocol (PTP) packets without latency. By applying this synchronization mechanism to 5GS and TSN, the individual TSN end stations should correct the time error to be synchronized with the clock, and the UE, RAN, UPF, and so on in 5GS must also be synchronized with the GM clock of 5G.

Third, additional functions such as NW-TT and DS-TT should be provided for interworking 5GS and TSN. To manage TSN end stations via a 5GS in TSN, centralized network configuration (CNC) must be able to recognize and control 5GS as a logical bridge in TSN. Therefore, a function of converting and transmitting TSN streams into the packets of a 5GS is required to deliver TSN streams of the data plane via a 5GS.

To implement the proposed system (Figure 12) according to these three design considerations, there is a need for an enhanced 5G CN including low-latency service control/data processing technologies in a 5GS. Table 3 lists the functions to implement this enhanced 5G CN, which includes 5GS-TSN interworking functions and PDU session/QoS management to control low-latency packets. In addition, it should be

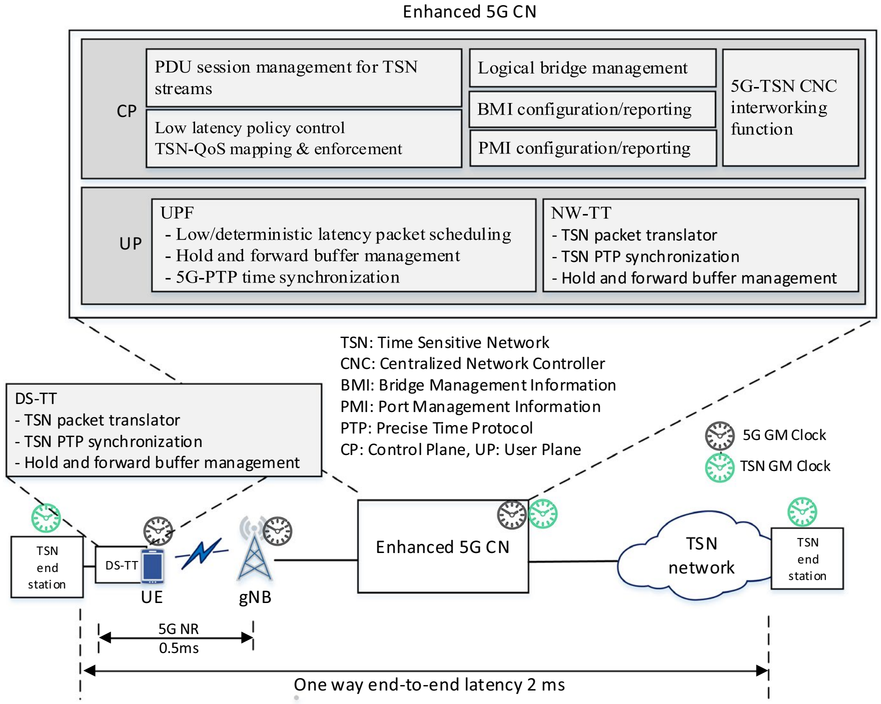

Enhanced 5G CN
CP
PDU session management for TSN streams
Low latency policy control
TSN-QoS mapping & enforcement
Logical bridge management
BMI configuration/reporting
PMI configuration/reporting
5G-TSN CNC interworking function
UP
UPF
- Low/deterministic latency packet scheduling
- Hold and forward buffer management
- 5G-PTP time synchronization
NW-TT
- TSN packet translator
- TSN PTP synchronization
- Hold and forward buffer management
TSN: Time Sensitive Network
CNC: Centralized Network Controller
BMI: Bridge Management Information
PMI: Port Management Information
PTP: Precise Time Protocol
CP: Control Plane, UP: User Plane
DS-TT
- TSN packet translator
- TSN PTP synchronization
- Hold and forward buffer management
5G GM Clock
TSN GM Clock
TSN end station
DS-TT
UE
gNB
Enhanced 5G CN
TSN network
TSN end station
5G NR
0.5ms
One way end-to-end latency 2 ms

FIGURE 12 Proposed enhanced 5GS for TSC

TABLE 3 Enhanced 5GS functions for ULL service

Functions	Descriptions
AF	5GS-TSN interworking control function, TSCAI management, Virtual bridge port management, Port path-pair management
PCF	Policy control and QoS enforcement for low-latency PDU session
SMF	Low-latency PDU session management for TSN streams
UPF/gNB/UE	Low-latency packet scheduling Packet transmission in a manner of delay-critical GBR
NW-TT DS-TT	Translating between TSN packets and 5G PDU Hold and forward buffer management for deterministic latency

possible to convert a TSN stream to 5GS traffic and perform time scheduled packet transmission so that the TSN stream can be delivered to the TSN end station mapped to the UE via 5GS with low and deterministic latency.

The TSN AF should determine TSN QoS information (ie, priority, delay, and maximum TSC burst size) based on the configuration information of the 5GS bridge received from the CNC, the bridge delay information, and the UE-DS-TT residence time [7]. Moreover, a TSN AF should generate time-sensitive communication assistance information (TSCAI), which is a set of {flow direction, periodicity, burst arrival time} appropriate to the TSN traffic pattern. The TSCAI is transmitted to the RAN via SMF so that it can be used for radio resource allocation, which guarantees low and deterministic latency in addition to the hold and forward buffering mechanism in the DS-TT and NW-TT. In addition, the SMF should provide new QoS management functions for ensuring low and deterministic latency packet transmission. To do this, the SMF should monitor the user plane path delay periodically, modify the QoS parameters of the PDU session, and provide control information to the packet scheduler of the user plane to apply the modified QoS parameters.

Meanwhile, it is important for the enhanced 5G CN to provide precise time synchronization between the TSN and 5GS to satisfy the KPI for the clock synchronization. Figure 12 shows two types of synchronization systems [7].

1. 5GS synchronization based on 5G GM.
2. TSN domain synchronization based on IEEE 802.1AS.

A 5GS should synchronize the gNB clock with the 5GS GM clock for 5GS synchronization as well as handle gPTP packets from the TSN DN to the TSN end stations for TSN domain synchronization. DS-TT and NW-TT are responsible for TSN domain synchronization. DS-TT measures bridge residence time, and NW-TT measures gPTP ingress link delay with adjacent ports to add them to the gPTP packet of the TSN domain. In addition, the time synchronization function of DS-TT/NW-TT can be extended to consider cases in which the TSN GM is located on the UE side as well as on the DN side.

5 | CONCLUSION

In order to satisfy the ULL performance requirement of 1 ms to 10 ms among various vertical services, the ongoing standardization progress of MEC and TSC of 3GPP to support ULL services at the network level of 5GS was introduced. This study showed the possibility of supporting ULL services in a LADN by implementing the 5GS test bed and introducing the results of measuring MEC service latency on the test bed. In addition, in order to support ULL in interworking with the external networks of a 5GS, this study presented major functions to be considered for ensuring low and deterministic latency performance through interworking with IEEE TSN. High-precision time synchronization of interworking systems is required. There is a need for a 5GS with enhancements such as interworking with TSN, ULL session management, TSN stream handling, and packet scheduling to ensure ultra-low/deterministic latency.

In further work, according to the design principles presented in this paper, we intend to develop a 5GS that can support ULL services such as factory automation, robot control, and telepresence over a wide area rather than a local area through interworking with a 5GS-IEEE TSN.

ORCID

Yoohwa Kang https://orcid.org/0000-0001-6311-7431

REFERENCES

1. 3GPP, *Service requirements for the 5G system; Stage 1,* V16.11.0, TS 22.261, Mar, 2020.
2. Parvez et al., *A survey on low latency towards 5G: RAN, core network and caching solutions*, IEEE Commun. Surveys Tutorials. **20** (2018), 3098–3130.
3. M. A. Lema et al., *Business case and technology analysis for 5G Low latency applications*, IEEE Access **5** (2017), 5917–5935.
4. 3GPP, *New WI proposal: L2 latency reduction techniques for LTE,* TSG RAN RP-160667, 2016.
5. 3GPP, *Study on Scenarios and Requirements for Next Generation Access Technologies,* TR 38.913, 2016.
6. J. Zhang et al., *Mobile edge computing and field trial results for 5G low latency scenario*, China Commun. **13** (2016), 174–182.
7. 3GPP, *System architecture for the 5G system; Stage 2,* V16.4.0, TS 23.501, Mar. 2020.
8. ITU, *Minimum requirements related to technical performance for IMT-2020 radio interface(s),* ITU-R M.2410-0, Nov. 2017, Retrieved Aug. 2019.
9. F. Giust et al., *MEC deployments in 4G and evolution towards 5G,* ETSI White Paper N.24, Feb. 2018.
10. 3GPP, *Service requirements for cyber-physical control applications in vertical domains; Stage 1,* V17.3.0, TS 22.104, July 2020.
11. M. Bennis, M. Debbah, and H. V. Poor, *Ultrareliable and low-latency wireless communication: Tail, risk, and scale*, Proc. IEEE **106** (Oct. 2018), 1834–1853.
12. E. Baştuğ et al., *Big data meets telcos: a proactive caching perspective*, J. Commun. Netw. **17** (2015), 549–557.
13. E. Bastug, M. Bennis, M. Debbah, *Living on the edge: The role of proactive caching in 5G wireless networks*, IEEE Comm. Mag. **52** (Aug. 2014), 82–89.
14. 3GPP, *Study on enhancement of support for edge computing in 5G core network (5GC),* V0.3.0, TR 23.748, Jan. 2020.
15. IEEE 802.1Q, Virtual LANs.
16. 3GPP, *Study on enhanced support of Industrial Internet of Things (IIoT) in the 5G System (5GS),* TR 23.700-20, Sept. 2019.
17. 3GPP, *Service requirements for Video, Imaging and Audio for Professional Applications (VIAPA); Stage 1,* V17.1.0, TS 22.263, July 2020.
18. C. Parada et al., *Multi-access edge computing: A 5G technology,* in Proc. Eur. Conf. Netw. Commun. (Ljubljana, Slovenia), June 2018, pp. 277–281.
19. S. Kekki et al., *MEC in 5G networks,* ETSI White Paper N.28, June 2018.
20. Q. Pham et al., *A survey of multi-access edge computing in 5g and beyond: fundamentals, technology integration, and state-of-the-art*, IEEE Access **8** (June 2020), 116974–117017.
21. C. Jiang et al., *Toward computation offloading in edge computing: a survey*, IEEE Access **7** (Aug. 2019), 131543–131558.
22. P. Mach et al., *Mobile edge computing: a survey on architecture and computation offloading*, IEEE Commun. Surveys Tutorials **19** (Mar. 2017), 1628–1656.
23. T. Choi et al., *Agile Management and interoperability testing of SDN/NFV-enriched 5G core networks*, ETRI J. **40** (Feb. 2018), 72–88.
24. N. Sprecher et al., *Harmonizing standards for edge computing-A synergized architecture leveraging ETSI ISG MEC and 3GPP specifications,* ETSI White Paper N.36, July 2020.
25. A. Nasrallah et al., *Ultra-low latency (ULL) networks: The IEEE TSN and IETF DetNet standards and related 5G ULL research*, IEEE Commun. Survey Tutorial **21** (2019), 88–145.
26. J. Farkas et al., *5G-TSN integration meets networking requirements for INDUSTRIAL AUTOMATION,* Ericsson Technology Review, Aug. 2019.
27. IETF, *Deterministic networking architecture,* Draft-IETF-Detnet-architecture-13, May 2019.
28. Ettus Research, *Universal software radio peripheral B210 SDR kit*, available at http://www.ettus.com/all-products/ub210-kit/
29. Wikipedia, *Cloud gaming, game streaming*, available at https://en.wikipedia.org/wiki/Cloud_gaming#Game_Streaming.
30. J. Engebretson, *Report: Google stadia cloud gamers poised to exceed internet data caps*, Oct. 2019, available at https://www.telecompetitor.com/report-google-stadia-cloud-gamers-poised-to-exceed-internet-data-caps/
31. Calomel.org, *Network latency and packet loss emulation*, available at https://calomel.org/network_loss_emulation.html.

32. C. A. Garcia-Perez, P. Merino, *Enabling low latency services in standard LTE networks,* in IEEE Int. Workshops Foundations Applicat. Self Syst. (Augsburg, Germany), Sept. 2016, pp. 248–255.
33. O. Al-Saadeh et al., *End-to-end latency and reliability performance of 5G in London,* in Proc. IEEE Global Commun. Conf. (Abu Dhabi, United Arab Emirates), Dec. 2018.

AUTHOR BIOGRAPHIES

Sunmi Jun received her BS and MS degrees in computer science from the Department of Computer Science, Pusan National University, Pusan, Republic of Korea, in 1998 and 2000, respectively. Since 2000, she has been a senior researcher at the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. Her main research interests are mobile communication systems, TSN, and URLLC in 3GPP 5G systems.

Yoohwa Kang received her MS degree from the Graduate School of Information Technology from Pohang University of Science and Technology, Pohang, Republic of Korea, in 2000. Since 2000, she has worked for the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. Her main research interest is the standardization of non-3GPP, ATSSS, URLLC, and TSN in 3GPP 5G systems.

Jaeho Kim received his BS and MS degrees in computer science from the Chung-Nam National University, Daejeon, Republic of Korea, in 1999 and 2001, respectively. Since 2000, he has been a principal researcher at the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. His main research interests are URLLC and TSN.

Changki Kim received his BS and MS degrees in electronics engineering from Pusan National University, Pusan, Republic of Korea, in 1995 and 1997, respectively. He also received his MS degree in software engineering from the Korea Advanced Institute of Science and Technology in 2006, and he received his PhD degree from the Department of Computer Engineering from Sun Moon University, Asan, Republic of Korea, in 2016. He worked as an engineer at Samsung Electronics from 1997 to 2000. Since 2001, he has been a principal researcher at the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. His main research interests are mobile communication systems including system design, protocol, and standardization.