Latency Analysis for Mobile Cellular Network uRLLC Services

Bin Liang
RMIT University

Mark A. Gregory
RMIT University

Shuo Li
RMIT University

Abstract: The fifth generation (5G) mobile network technologies include ultra-Reliable Low Latency Communications (uRLLC) capability. To fully exploit uRLLC, distributed Multi- access Edge Computing (MEC) is being developed and introduced at the network edge with an architecture that supports applications and services. Some of the MEC applications will benefit from uRLLC, including virtual reality, augmented reality, education, health, online gaming, automatic manufacturing and Vehicle-to-everything. However, unique challenges and opportunities exist for 5G cellular networks and MEC due to a range of factors, including end-user device mobility and the implementation of the network Control Plane (CP) and User Plane (UP). In this regard, there is a need to optimize protocols and network architecture. This paper investigates latency and related network elements in the next generation mobile cellular network. We also analyze the 5G network latency in the CP and UP. Finally, the paper identifies protocol optimization considerations for MEC integration with 5G to achieve low end-to-end latency.

Keywords: Ultra-reliable Low Latency Communications, Multi-access Edge Computing, End-to-End Latency, User Plane, Control Plane

Introduction

The evolution of mobile telecommunications and the development of information technology provides new horizons for the introduction of the Internet of Things (IoT), person-to-person and machine-to-machine communications. The innovative applications that are supported by the fifth generation (5G) mobile networks include remote surgery, smart cities, online gaming, virtual reality, augmented reality and autonomous vehicular networks. Specifications,

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447

architectures and standards are required for the new technologies, e.g., heterogeneous connections, high-throughput, ultra-reliability and ultra-low latency communications. The International Telecommunication Union (ITU), working with industry bodies, has introduced three areas of focus for 5G technology development, including enhanced Mobile Broadband (eMBB), massive Machine Type Communication (mMTC) and ultra-Reliable Low Latency Communications (uRLLC). With the introduction of 5G, significant improvements will be achieved for throughput, reduced latency, mobility, connectivity, density and energy efficiency.

For the 5G uRLLC scenario, the end-to-end (E2E) latency should be reduced to 1-10 ms (1 ms for user plane delay, 10 ms for control plane delay), which is 20% of that achieved for 4G. The 5G technology improvements are a big step forward; however, there is a need for compute and storage offloading at the network edge to ensure that the 5G improvements can be fully realized. Multi-access Edge Computing (MEC) is being developed and introduced at the network edge to fulfill this role. The European Telecommunications Standards Institute (ETSI) has made a significant contribution to the development of MEC and continues to provide guidance on specifications and architectures for MEC. MEC servers will be deployed at the edge of the 5G network to provide a facility for computation and data offloading and to reduce the current reliance on the Cloud with its associated high latency (Li, Wang & Zhang, 2020). MEC services, including Radio Network Information Service (RNIS), Location Service (LS) and Traffic Management (TM), can be used to make intelligent network-related decisions for effective service continuity and mobility.

To fulfil the potential of MEC with 5G, researchers have been rethinking network architectures and protocols. Changes to the core network architecture are a primary symbol of the evolution of the network. The mobile core network has evolved to become separated, virtualized and software-based. The Control and User Plane Separation (CUPS) architecture was introduced in 4G EPC nodes. In the 5G Core (5GC) architecture, the control and user plane separation has further evolved. Network slicing was introduced into the 5GC architecture. A network can be virtualized into multiple network slices. Each slice has its application scenario, network capability and quality assurance, and different slices are isolated from each other.

Network Function Virtualization (NFV) is being implemented in the core network to enhance communication network flexibility. NFV can be used to decouple hardware and software. Network functions can be rapidly and dynamically created and migrated utilizing modern software practices. Software Defined Networking (SDN) is a paradigm being introduced that separates the network control and forwarding planes, and the centralized control plane provides the 5G network with visibility, flexibility and dynamic deployment capability. However, the deployment of the SDN controllers in 5G networks can increase latency. The

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447

further separation between the Control Plane (CP) and User Plane (UP) components increases the number of network components involved, and the messages exchanged among them. Complex CP exchange increases the latency and overhead. Therefore, optimizing the network protocols and architecture is a current research focus that aims to significantly reduce latency in this scenario (Parvez *et al.*, 2018; Mohammadkhan, Ramakrishnan & Jain, 2020).

In this paper, the network latency and a selected latency-sensitive application (mobile online gaming) are analyzed in different domains. The latency is analyzed in the CP and UP. To emphasize the importance of optimized system architecture and protocols, selected CP protocols, e.g., User Equipment (UE) Registration and Handover procedures, are analyzed. The latency is analyzed in a real mobile network to identify the need for an optimized network architecture and protocols. Finally, the considerations for an MEC handover protocol are provided. The main contributions of this article are summarized as follows:

- We elaborated on the recent advances in lowering latency in the CP and UP planes across different domains.
- We emphasized the importance of an optimized network architecture and protocols to enhance service mobility and continuity.
- We analyzed the latency in the CP and UP in a real cellular network and identified the need for an optimized 5GC architecture.
- We identified CP latency as an ongoing challenge.
- We discuss the potential approaches for an optimized handover protocol based on the integration of 5G and MEC.

The rest of the paper is organized as follows. A review of related works is presented in the Current Research section. The Latency Definition and Use Cases section provides a review of the latency in different domains. The CP and UP Latency Analysis section comprehensively analyzes the latency in the CP and UP, the latency caused by mobility, and indicates the importance of the necessity of optimized CP protocols and network architecture. In the System Architecture and Protocol Consideration section, we identify considerations for an optimized handover protocol. The paper is concluded in a Conclusion section.

Current Research

MEC and 5G significantly benefit uRLLC applications. Recently, researchers have investigated how the integration of MEC and 5G affects latency and have explored the use cases. Shukla *et al.* (2021) discussed the recent latency reduction technologies, including Fog computing, machine learning and conventional techniques, in the cloud and IoT across computation. Parvez *et al.* (2018) presented a detailed latency survey by considering the different domains in 5G, including Radio Access Network (RAN), core network and caching. In the RAN domain,

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447

the emerging technologies were studied, e.g., short frame, new waveform design, modulation and coding scheme, to facilitate low latency. SDN, NFV and MEC deployments with 5GC were also investigated to identify how latency was affected. Distributed and centralized caching, e.g., cache placement and content delivery methods, were presented for latency minimization in the caching domain.

As a typical uRLLC application, mobile online gaming has been investigated to explore the challenges of the architecture. In Premsankar, Di Francesco & Taleb (2018), Premsankar, Di Francesco & Taleb evaluated the MEC performance for mobile gaming, and the results showed that MEC deployment effectively reduced the latency for mobile gaming. Kasenides & Paspallis (2020) compared the performance of the different game services in terms of latency, and the study showed that Infrastructure as a Service (IaaS) implementations outperformed Platform as a Service (PaaS) implementations. Huedo *et al.* (2021) proposed a distributed edge cloud platform, where an online gaming application was deployed across seventeen locations. The results showed that the latency achieved from the experiment was below 10 ms.

The recent changes to the 5G network architecture significantly improved network and service performance. One of the goals of the 5G technologies is to reduce network latency to meet the demands of the new specification for uRLLC services. Guimaraes *et al.* (2020) proposed an SDN-NFV based architecture in Wi-Fi networks by using the SDN and NFV orchestration functionalities. The proposed architecture significantly improved mobility management and reduced latency. Tomaszewski, Kukliński & Kołakowski (2020) proposed the three variants of the integration of MEC into 5GC architecture based on the consideration of service mobility and continuity. Khaturia *et al.* (2021) proposed a novel 5G architecture by separating UE control from network control to improve the flexibility of the 5G network architecture. Abdulghaffar *et al.* (2021) proposed an SDN-based 5GC network architecture to enhance the flexibility of network management. The results indicated that the proposed architecture reduced the latency by from 18% to 62% when compared with the previously proposed architecture.

Articles that respond to optimization of the CP protocols related to latency, including access and handover procedures, were reviewed. Mohammadkhan *et al.* (2016) proposed a simplified software-based architecture and a simplified control protocol for the mobile network by reducing the number of control messages to achieve lower latency. Gaur *et al.* (2018) proposed an efficient vertical handover based on containerization in IoT networks by considering criteria such as latency, reliability, and power consumption. The proposed method enhanced data precision for latency and reliability. Tan *et al.* (2018) identified that the signalling operation contributes a large portion of the latency and proposed a cross-layer design to reduce latency in the signalling operation. Zhou *et al.* (2021) proposed a handoff algorithm by

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447

considering the combined signalling and computational load. The result indicated that the proposed algorithm could significantly reduce the latency.

Latency Definition and Use Cases

Latency in the Next Generation Cellular Network

E2E latency across multiple networks is a focus for the next generation cellular networks. The latency of the E2E service delivery includes both processing delay and network delay. The processing latency is the computational latency (router, switch, VMs). The network latency is the sum of the transmission delay and the propagation delay over the link between two network entities. The transmission delay is the time for a forwarding device to receive or to send a packet (Martini *et al.*, 2015). In the 3rd Generation Partnership Project (3GPP), latency in a 5G network can be divided into CP latency and UP latency. In order to send or receive packets, the UE must be in the active mode following the idle mode. In the idle mode, the UE is in the Radio Resource Control (RRC) disconnected state The UE is in the RRC connected state in the active mode. The delay that happens in this process is defined as the typical CP latency. A signalling process is performed between the radio and the core network. Therefore, the CP latency in 5G networks is the time consumed by the signalling message exchange in the CP. In the next generation mobile network (NGMN) white paper (Hattachi & Erfanian, 2015), the UP latency is the time consumed to transfer and receive a data packet from the UE to the destination node at the 5G network interface between layer 2 and layer 3. In 3GPP TS 22.261 (G.Group.A, 2021), the E2E latency is the time consumed to transfer a given piece of information successfully from a source to a destination, which is measured at the communication interfaces. E2E latency is not only in the 5G network but also is in other networks outside of 5G networks. The latency definition is also considered in One-way delay (OWD) and Round -trip time (RTT). OWD is the time taken to transfer a packet from the source to the destination. RTT is the time taken by a packet that is transferred from the source to the destination and for a response from the source to be received at the destination. Overall, the network latency can be defined as a packet transferred over multiple networks or application nodes such as RAN, core, transport and data centre, along with the one-way or round-trip time (Kurian, 2018). Figure 1 depicts the typical route for the E2E delay of a packet transmission.

There are several approaches that might be explored for latency reduction in the RAN, including frame/packet structure, symbol detection, waveform/multiple access, mmWave, location-aware communication, and improved modulation. The backhaul latency occurs between the gNodeB and the core network over copper wires, optical fibres or microwave. The latency in the core network is the transmission and processing, which is a combination of

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447

delays associated with the network entities and control signalling. The transport latency is the transmission and signalling delay between the core network and end-points considering transport layer delays and signalling (Parvez *et al.*, 2018).

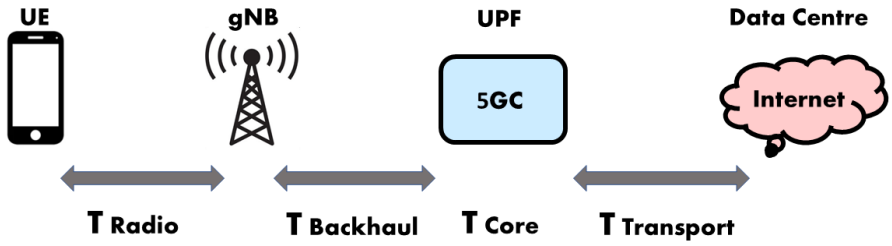

Figure 1. Latency contribution in packet transmission E2E delay

Latency-sensitive Application: Mobile Online Gaming

Mobile online gaming is likely to become one of the most popular 5G uRLLC applications. However, the user experience will degrade when the network latency is more than 30-50 ms. Ideally, 1 ms RTT is recommended for mobile online gaming, e.g., human interactive applications with high-quality visualization. An architecture for mobile online gaming is provided in Figure 2.

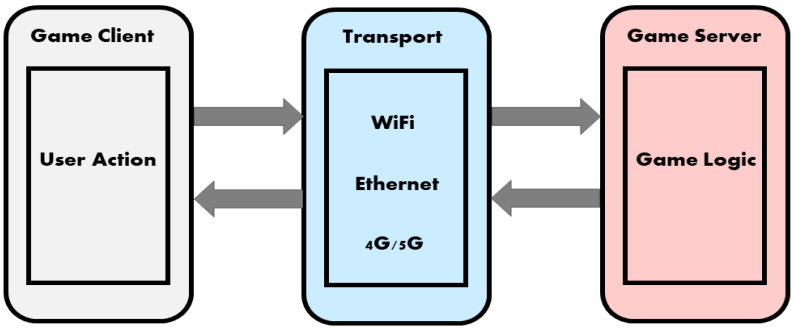

Figure 2. Mobile Online Gaming Architecture

Mobile online gaming has three key elements: the gaming client, the transport network and the gaming server. The players connect and use the gaming client, while the gaming server's role is to provide game logic management. The transport network provides the communications path between the gaming server and the gaming client and includes the mobile telecommunication network. The latency caused by the game operation can be divided into three parts: processing delay, playout delay and network delay. The time consumed for the gaming server to receive gaming client commands, to process and update the game state and respond is called the processing delay. Playout delay is the time consumed for a gaming client to receive, decode and render a frame on the screen. The network delay is the time required for a round of data to be exchanged between the gaming client and the gaming server (Peñaherrera-Pulla *et al.*, 2021; Sreenibha Reddy & Zaahid, 2018).

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447

Network latency is the essential network performance metric for mobile online gaming. The gaming server location and distance to the gaming clients will significantly affect the network latency. Network congestion will also affect latency. The decision on where to locate gaming servers is an essential consideration, affecting network architecture and energy utilization. The deployment of multiple gaming servers complicates the protocol and data exchange among the network components, thereby increasing latency. Therefore, the latency should be analyzed from different network domains and planes, combined with user behaviour, to achieve optimized protocols and network architecture for latency minimization (Peñaherrera-Pulla *et al.*, 2021; Sreenibha Reddy & Zaahid, 2018).

Control Plane and User Plane Latency Analysis

CP and UP Protocol Stacks

In 4G networks, the CP and UP are not completely separated, and some of the signalling messages and user data flow through the Serving Gateway (SGW) and Packet Data Network Gateway (PGW). Such a centralized system structure reduces network control and maintenance but limits backflows. In particular, the CP should focus on signal processing, while the UP should forward data. Thus, a fully decoupled CP and UP design should increase efficiency. Even though the CUPS was introduced in 4G, it is in 5G networks that CUPS has been fully implemented. CUPS in 5G allows network operators to separate the CP and UP, where the CP can be managed centrally. As we can see from Figure 3 and Figure 4, the 5G CP protocol stack is shown including UE, 5G-Access Network (AN), Mobility Management Function (AMF) and Session Management Function (SMF) in 5GC, and the UP protocol stack for UE, 5G-AN and User Plane Function (UPF) in 5GC. The data flows are transmitted through each sublayer of the UP. The signalling is exchanged between the CP sublayers. In 3GPP TS 23.501, the protocol stacks among 5G system elements are described, including the UE, the 5G-AN, the 5GC and 5G network functions. In Figure 3, the CP protocol stacks between the UE and the SMF via interfaces N2 and N11 over the multiple protocol layers are described, including NG Application Protocol (NG-AP), Stream Control Transmission Protocol (SCTP), 5G-AN Protocol layer, Non-Access-Stratum (NAS)—Mobility Management (MM) and NAS-Session Management (SM). AMF is the most important 5GC CP entity, which mainly supports registration, authentication and mobility management. The SMF is a session establishment, modification and release functional entity of the 5GC. In the CP, SMF is involved in SM parts of NAS message (G.Group.A, 2020). NGAP protocol is responsible for the application layer signalling protocol. SCTP protocol in the transport layer is over the Internet Protocol (IP) layer to guarantee the transport and application layer of the signalling messages. The IP layer provides the transmission of the signalling Protocol Data Unit (PDUs) (E.Group, 2019). The

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447

NAS protocol on N1 includes a NAS-MM and a NAS-SM. NAS-MM is responsible for registration, authentication, mobility management of NAS signalling and supports the UP activation and deactivation. NAS-SM is responsible for the PDU session management and is transparently transferred to the AMF. The 5G-AN protocol layer is a set of protocols related to the 5G access network.

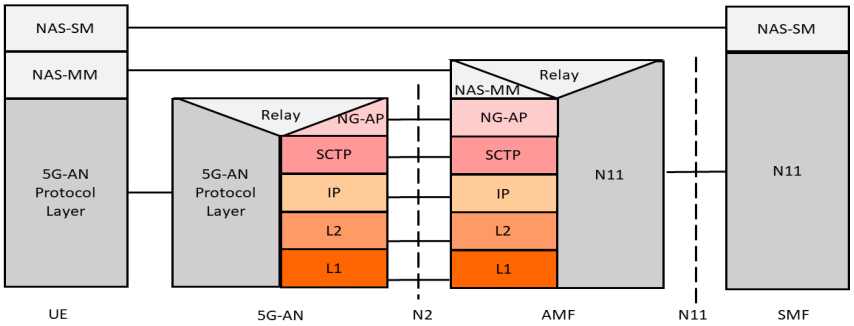

Figure 3. 5G Control Plane Protocol Stack [(G.Group.A, 2020)]

In the UP protocol stack, Figure 4 illustrates the UP transport for a PDU session from UE to UPF over UP protocol layers, including the PDU layer, GPRS Tunnelling Protocol for the User plane (GTP-U), 5G-AN and User Datagram Protocol (UDP)/IP via N3, N9 and N6 interfaces. UPF is a primary UP entity responsible for packet routing and forwarding. The PDU layer transmits PDU between the UE and the Data Network (DN) over the same PDU session. The GTP-U protocol layer provides user data tunnelling, the encapsulation of a PDU session, and a QoS flow definition. UDP is a transport layer protocol and supports unreliable datagram transmission between applications. IP is a network layer protocol and is responsible for a datagram service delivered between applications. TCP and UDP are the key network protocols used to transport traffic [(G.Group.A, 2020)].

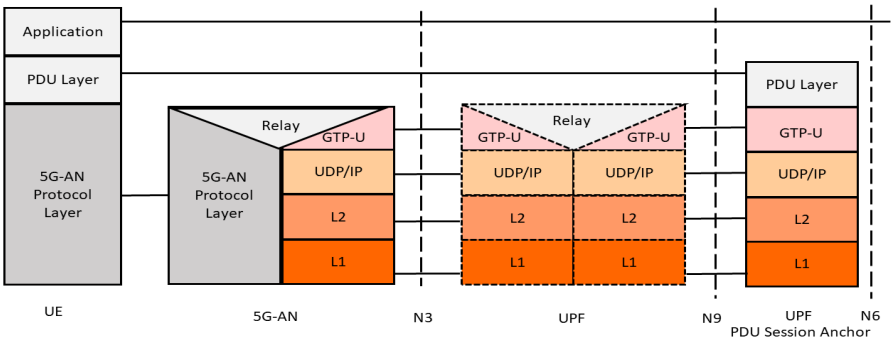

Figure 4. 5G User Plane Protocol Stack [(G.Group.A, 2020)]

Separation of the CP and UP enables flexible network operation and deployment and enhances the independent scaling of the network architecture, such as flexibly adding UP nodes without affecting the CP nodes, enabling SDN and NFV in the CP more efficiently, and providing a relatively independent evolution for the functions in the CP and UP.

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447

Control Plane Protocol Analysis

Due to the CP and UP separation and a set of distributed network components based on SDN and NFV, the cellular network has evolved into a new network architecture. However, this also increases the number of network components and the number of control messages exchanged, which significantly increases the protocol overhead and latency (especially in the CP). The CP latency is the key factor affecting E2E latency. Meanwhile, data packets in UP often wait for the CP functions to occur. Therefore, optimizing the CP latency will significantly reduce the E2E latency and improve user experience. We discuss and analyze selected 5G CP protocols (UE Registration and Handover) in the following sections.

Initial UE Registration Procedure

The initial UE registration procedure is a typical CP procedure. During an initial registration procedure, a user wants to make a connection with the cellular network to enable authentication, mobility tracking and service reachability. This event is quite similar in 4G and 5G protocols. It is called the initial attach procedure in 4G and the initial registration procedure in 5G.

The initial registration procedure in 5G is described in Figure 5 (E.Group, 2018). The UE first sends an RRC Connection Request to the gNB-Distributed Unit (gNB-DU). Once gNB-DU receives this message, the F1 Application Protocol (F1AP) Initial UL RRC Message Transfer message for UE will be transferred to gNB-Centralized Unit (gNB-CU) with Cell RNTI (C-RNTI) allocated by the gNB-DU. Then, gNB-CU will allocate a gNB-CU UE F1AP ID for the UE and creates an RRC Connection Setup message for the UE, which will be encapsulated in the F1AP DL RRC Message Transfer message. The gNB-DU sends the RRC connection Setup to the UE, then the UE sends back the RRC connection Setup Complete to the gNB-DU. Once gNB-CU has received the UL RRC message Transfer, it will trigger an Initial UE message to AMF. From steps 8 to 18, the UE Context, RRC security and reconfiguration will be set up and completed. After UE sends the RRC Connection Reconfiguration Complete message to the gNB-DU, gNB-DU completes encapsulation and sends it to the gNB-CU. The Initial UE Context Setup Response message will be sent to the AMF. Initial UE registration is then complete.

The function of the F1AP is to provide the signalling service between a gNB-CU and a gNB-DU. C-RNTI is responsible for identifying and scheduling RRC connection to a particular UE. In Jin, Zhong & Zhou (2016), an optimized access procedure was proposed for 4G networks by saving UE context information in the eNodeB in the idle state and sinking the control function of the radio bearer from the MME to the eNodeB and configuring the radio bearer to

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447

reduce signalling overhead and CP latency. The proposed access procedure also can meet the low latency requirements of the 5G network.

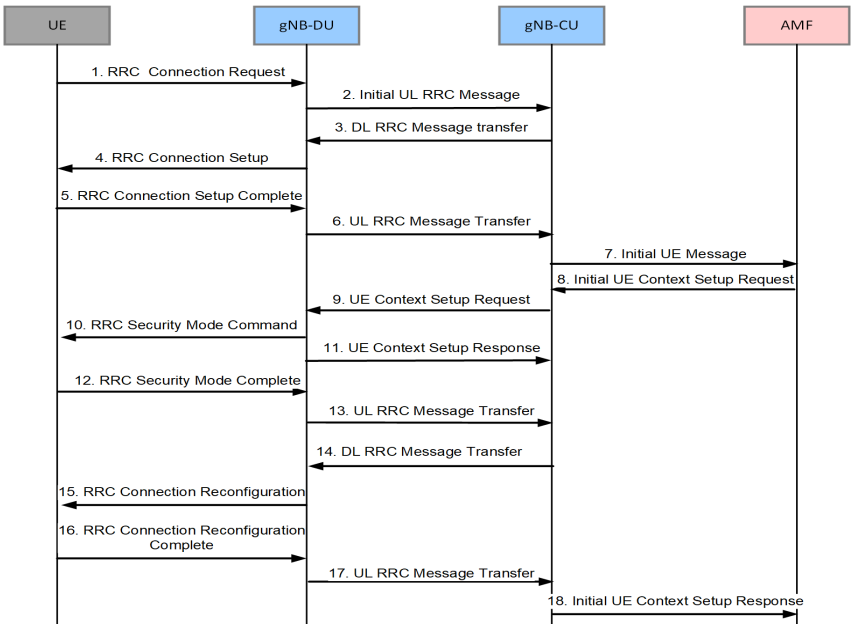

Figure 5. UE Initial Registration Procedure in 5G (E.Group, 2018).

Handover Procedure

When a UE moves across the network, handover happens from the serving base station to an adjacent base station. Handover is a significant cause of mobility latency. In order to reduce mobility latency, handover is optimized. There are two handover mechanisms, including hard handover and soft handover. A hard handover is a break-before-make approach, while a soft handover is a make-before-break approach. In 4G networks, 3GPP specifications include hard handover. A common strategy is X2 interface handover in 4G, in which the handover procedure occurs between eNodeBs and the signalling load towards the core network is reduced. On the other hand, in terms of S1 handovers, a reduced core network interaction was performed — this improved 4G network handover. The 5G uRLLC demands that handover latency be very close to zero. 5G handover includes Xn and N2 handover, and the differences between 4G handover (X2, S1) and 5G Handover (Xn, N2) are subtle. 5G Xn handover and 4G X2 handover are both hard handovers, where UE releases the connection to the source gNodeB before the connection to the target gNodeB is established. Here, before the UE starts to communicate with the target gNodeB, the uplink and downlink transmissions are finalized in the source gNodeB. This handover mechanism causes an interruption in communication between the network and the UE.

In order to optimize the mobility management procedure, reducing the handover interruption time between 5G cells is critical. To overcome this, a Dual Active Protocol Stack (DAPS) handover has been proposed by the 3GPP. DAPS handover maintains the source gNodeB

connection after reception of the RRC message for handover and releases the source gNodeB after successful random access to the target gNodeB (G.Group.B, 2021).

Research into the optimization of the handover procedure has occurred. Gimenez *et al.* (2017) proposed a data forwarding mechanism with the handover (make-before-break) procedure by analyzing handover time, UE and e-NodeB processing times and interface latency. The result indicated that the proposed data forwarding can reduce the interruption time by 18% and make-before-break handover is much more suitable for future latency-sensitive applications. Jain, Lopez-Aguilera & Demirkol (2018) proposed a new handover signalling method based on enhanced SDN and evaluated this method in 4G and 5G scenarios. The proposed mechanism effectively reduced latency and processing costs compared to the 3GPP handover specifications. Peltonen, Sasse & Basin (2021) presented a comprehensive analysis for 5G handover protocols and indicated that a formal and comprehensive analysis is essential for designing the reliability and security protocols and network architecture.

User Plane Protocol Analysis

In a cellular network, the UP latency is the time taken to transmit a packet/message on the application layer over the radio protocol layer 2 and layer 3 in between the uplink and downlink. For uRLLC applications, UP latency requirements are 0.5 ms for UL and 0.5 ms for DL. Improved reliability also needs to be associated with lowering latency. For the next generation cellular technologies, the UP latency should be much lower than the current target (G.Group.B, 2020). Evangelista, Kaddoum & Sattar (2021) formulated a mathematical model to analyse the latency and reliability performance of mmWave massive MIMO uRLLC applications by using the reactive and K-repetition Hybrid Automatic Repeat Request (HARQ) protocol. The result indicated that mmWave massive MIMO decreased the failure probability, but there is no significant improvement in reliability with a certain number of antennas. Thus, mmWave massive MIMO alone could not guarantee that the performance requirement is met for uRLLC applications.

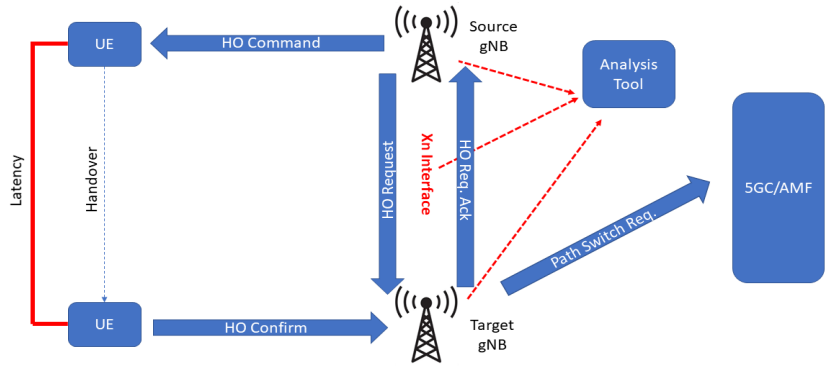

Figure 6. Xn Handover Signal Analysis and Monitoring Framework

Evaluation

In order to verify the latency among the network components, we conducted two empirical studies of the latency in the CP and UP over a national Chinese service provider 5G network. We collected and analyzed the latency from the two typical control signalling procedures (UE initial registration and Xn handover procedures). The signalling messages were captured from the gNodeB and filtered and analyzed utilizing a standard signalling message statistical and analytical tool. The data analyzed corresponds to 5G cells under good air interface conditions. The handover signal analysis and monitoring framework is specified in Figure 6.

Table 1. Latency in Control Plane for UE Access Procedure

Latency may belong to	Latency	Start	Signal Message	End
UE+RAN	0 ms	UE	RRC Connection Request	gNB-DU
RAN	0 ms	gNB-DU	Initial UL RRC message	gNB-CU
RAN	0 ms	gNB-CU	DL RRC message transfer	gNB-DU
UE+RAN	0 ms	gNB-DU	RRC Connection Setup	UE
RAN	31 ms	gNB-DU	UL RRC Message Transfer	gNB-CU
RAN+CORE	31 ms	gNB-CU	Initial UE message	AMF
RAN+CORE	36 ms	AMF	Initial UE Context Setup request	gNB-CU
RAN	36 ms	gNB-CU	UE Context Setup Request	gNB-DU
UE+RAN	1 ms	gNB-DU	RRC Security Mode Command	UE
RAN	1 ms	gNB-DU	UE Context Setup Response	gNB-CU
UE+RAN	5 ms	UE	RRC Security Mode Complete	gNB-DU
RAN	5 ms	gNB-DU	UL RRC Message Transfer	gNB-CU
RAN	4 ms	gNB-CU	DL RRC message transfer	gNB-DU
UE+RAN	4 ms	gNB-DU	RRC Connection Reconfiguration	UE
UE+RAN	11 ms	UE	RRC Connection Reconfiguration Complete	gNB-DU
RAN	11 ms	gNB-DU	UL RRC Message Transfer	gNB-CU
RAN+CORE	0 ms	gNB-CU	Initial UE Context Setup response	AMF

Table 2. Latency in Control Plane for Xn Handover

Latency may belong to	Latency	Start	Message	End
RAN	152 ms	S-gNB	Handover Request	T-gNB
RAN	9 ms	T-gNB	Handover Request Ack	S-gNB
UE+RAN	0 ms	T-gNB	RRC Connection Reconfiguration	UE
RAN	2 ms	S-gNB	SN Status Transfer	T-gNB
UE+RAN	34 ms	UE	RRC Connection Complete	T-gNB
RAN+CORE	0 ms	T-gNB	Path Switch Request	AMF
RAN+CORE	62 ms	AMF	Path Switch Request Ack	T-gNB
RAN+CORE	0 ms	T-gNB	UE Context Release Request	AMF
RAN+CORE	0 ms	AMF	UE Context Release Command	T-gNB
RAN+CORE	2 ms	T-gNB	UE Context Release Complete	AMF
UE+RAN	0 ms	T-gNB	RRC Release	UE

Tables 1 and 2 indicate the median processing time consumed for each signal message; more than 20 samples have been analyzed. The latency belonging to different network components

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447

in the CP is shown in Table 1 and Table 2. As we can see from Table 1 and Table 2, CP latency is more than the 5G CP target of 10 ms when the core network latency is around 31 ms-36 ms. In addition, the CP latency caused by the Xn handover procedure during the handover preparation phase is high, around 150 ms, which is much higher than the 10 ms we expected. Therefore, optimizing the 5GC architecture and CP latency is an ongoing challenge.

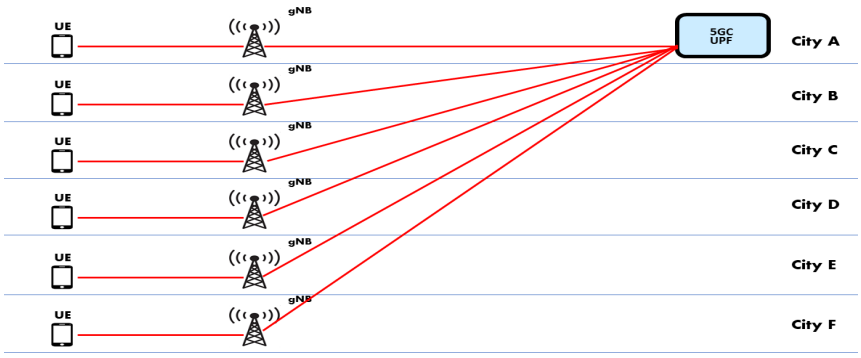

Figure 7. Ping Test Route

Table 3. Latency in User Plane

City	Latency between UE to gNB	Latency in gNB	Latency between gNB to UPF
City A	5 ms	3 ms	1-2 ms
City B	5 ms	3 ms	4-5 ms, 30 ms occurs
City C	5 ms	3 ms	4-5 ms, 30 ms occurs
City D	5 ms	3 ms	4-5 ms, 30 ms occurs
City E	5 ms	3 ms	3-4 ms, 30 ms occurs
City F	5 ms	3 ms	3-4 ms, 30 ms occurs

We also conducted a set of UP ping tests to analyze and verify the UP latency. We chose two cells per city to conduct multiple ping tests from UE to gNodeB, from gNodeB to core network UPF in the 5G network. The UPF of each city was deployed in city A. Figure 7 depicts the ping test route, and Table 3 shows the ping results. The UP median latency between UE and gNodeB in each city was almost the same, due to each pair of the UE and gNodeB being located in the same city. The UP median latency between gNodeB and UPF in city A is also stable for the same reason. For the rest of the cities, the latency is not stable, and 30ms latency occurs as their UPFs are located in city A, away from their gNodeB. Therefore, the backhaul distance and network deployment significantly affect performance and latency. Since the UP procedures need to be triggered and paused until the CP procedures are completed, complex CP protocols will directly affect the UP performance. Thus, cooperation and association with UP protocols also need to be considered when we optimize CP protocols.

System Architecture and Protocol Consideration

Examining the current protocols and latency in the CP and UP, there are many exchanged messages among the network elements, particularly for the evolved core network. A significant

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447

portion of this traffic is from the distributed network architecture in UP and complex CP protocols such as the handover procedure. Thus, it is essential to optimize the network architecture and lower the CP overheads. In the following section, we analyze the need for network optimization and propose a series of optimization considerations.

5GC and MEC Integration Architecture

The cellular core network is the brain of the telecommunication network and is mainly responsible for user and data control, management and distribution. The functions of each core element are different. To improve core network performance, changes to the core network architecture have occurred, from the separation of CP and UP, to the UP functions linking with the MEC nodes. The network elements on the UP can benefit from services offered by MEC nodes. The MEC technology enables applications by deploying capability to the edge of the network. A distributed architecture migrates the computing, storage and processing tasks close to the UE, thus significantly reducing transmission and processing latency. The new core network enabling technologies, e.g., SDN and NFV, have been proposed to provide hardware and software flexibility to ensure that service mobility and continuity occur.

Challenges related to the 5GC network and MEC include orchestration and management of the network and computation resources, which requires changes in the network architecture. Tomaszewski, Kukliński & Kołakowski (2020) proposed three variants of the integration of MEC into 5G slicing architectures, focusing on the integration of 5GC and MEC solutions. MEC APIs and 5GC CP capabilities from the Network Exposure Function (NEF) will extend the amount of communication information available for MEC and 5G tight integration to ensure smooth interaction and low latency. Fan & Huo (2021) provided an evolutionary perspective for low latency in cellular networks in two important aspects, including network architecture (RAN, core network, and bearer network) and physical layer air interface technologies (packet size, frame structure, modulation scheme, coding scheme, minimum transmission interval). A series of challenges in the network architecture are discussed, such as the trade-off between cost, maintenance and performance, standardized protocol and technologies, resource management and active cooperation. In addition, another challenge is CP scalability whilst balancing CP load to meet 5G latency requirements.

5G and MEC Handover

As one of the most critical procedures of the cellular network, handover can significantly impact the latency in CP and UP and the network performance. Mohammadkhan, Ramakrishnan & Jain (2020) proposed an optimized handover protocol based on a new CleanG system architecture to lower the latency, in which the control component duplicates

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447

the downstream packet for the source and target eNodeBs, while a sequence number will be added into the packets. Such packet duplication only takes a short period for handover actions. To optimize CP protocols, the authors eliminated synchronization and acknowledgement messages among the different core components. In addition, they used Generic Routing Encapsulation (GRE) tunnels instead of the GTP tunnels to eliminate the GTP tunnelling overheads. The authors use a series of measures to design the CP protocols, in which the protocols are designed for the specific scenarios to reduce the unnecessary messages in the standard scenarios by using a shared data structure and merging and reordering the messages to reduce exchanged messages among the core components, reassigning the tasks to the network components to reduce the unnecessary acknowledgement messages. Furthermore, besides optimizing the 5G system CP protocols, standardization of protocols is essential when considering the integration of MEC and 5G.

The integration of MEC and 5G is considered to be an effective solution to meet the requirement for uRLLC applications. Latency occurs between the MEC server and the cellular network UE; integrating with MEC services that are migrated to the edge network by deploying applications close to UE and enabling service migration between MEC hosts during the UE mobility procedures, which improve the computation and storage capacity and reduce the latency. Existing 5G mobility management procedures that do not consider the computational and processing load of MEC servers, are not suitable for real-time execution and integration with the standard protocol stack, which can affect the performance of the 5G MEC applications and increase the latency. Specifically, handover progress occurs between the base stations and MEC servers. The different MEC servers may have varying capabilities and loads, which will affect the user application performance. The current handover decisions only focus on signal strength using Signal-to-Interference-plus-Noise Ratio (SINR) and Reference Signal Received Quality (RSRQ), without concern for the state of neighbour MEC servers. However, the MEC server capacity and load can significantly affect the handover procedures and the latency.

Most importantly, in 5G networks, cell sizes in some parts of the network are small (<500 m), which causes more frequent handovers. Therefore, the challenges related to handovers are increasing (Zhou *et al.*, 2021). In Zhou *et al.* (2021), the authors proposed the Comp-Ho handoff algorithm, which jointly optimized signal strength in base stations and computational load in MEC servers. The base station will trigger a handover and reassign the communication and computation processes to another base station and MEC server when signal strength degrades to threshold values or the current MEC server does not have enough capacity and becomes overloaded. A balance is taken into consideration. Although this algorithm significantly reduced the latency caused by MEC congestion, developing an algorithm with the

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447

predictive capability to predict mobility where the resources can be pre-allocated before handover occurs is becoming very popular.

According to the analysis, we proposed optimization considerations for 5G and MEC integration. We identified an optimization consideration for handover procedures before and during the handover preparation pause, since the latency before handover request is quite high from our evaluation described earlier. We scale the MECs to deal with the UE directly by using the RNIS of the MEC, which will decrease the signal message overheads between the MECs and base stations over Xn links. The user context will start to prepare the migration from the source MEC when the source MEC receives a response from the UE, to see if it can reduce the latency before the handover request. Based on the optimized handover procedure and information sharing, the user context migration time between MECs may also reduce. Figure 8 depicts the proposed optimization considerations. Future work includes verifying the considerations by using a 5GC simulation, e.g., free5GC, to deploy 5GC into a Kubernetes environment, which orchestrates the MEC and 5GC network elements as containers.

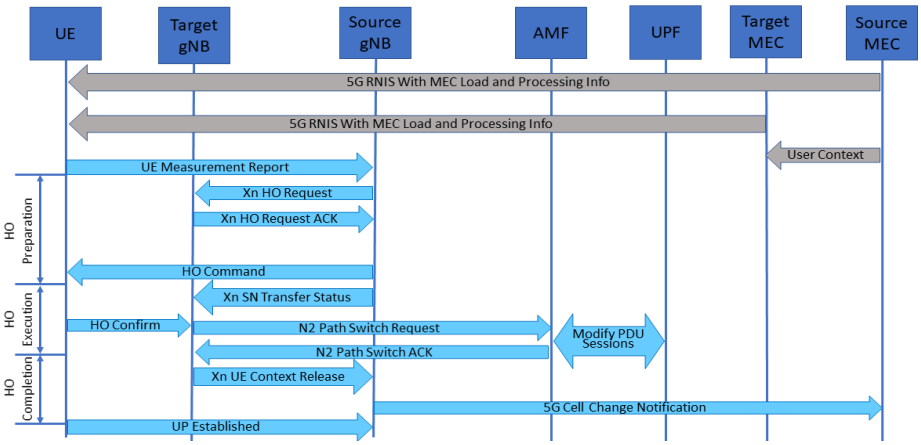

Figure 8. Optimization Considerations

Conclusion

This paper analyses CP and UP latency. A series of considerations for optimizing the CP protocol handover is proposed based on integrating 5G and MEC architectures. To achieve ultra-low latency, we can provide the following conclusions in the changes of network architecture: 1. core UP components keep moving towards the network edge; 2. network components continue to be virtualized on the Cloud platform; and 3. the integration of sliced 5GC and MEC is occurring based on Capital Expenditure (CAPEX) and Operating Expenditure (OPEX) considerations.

We can present the following conclusions for the optimization of CP protocols: 1. unnecessary exchanged messages in CP increased due to the evolution of the network, and security challenges persist; 2. the messages can be reordered depending on the network element

responsibilities; and 3. the CP protocol can be optimized, related to the different service requirements based on the association with the UP.

DAPS handover has been developed with the goal of reducing the UP latency to near 0 ms when a handover occurs, but complex CP protocols and network traffic during the mobility management procedure remain a challenge. The communication control protocols contribute to the significant latency. There is increasing use of SDN and NFV and the separation of CP and UP to provide scalability and flow distribution flexibility. However, adding an SDN controller to the network can be another latency source. Thus, there is a trade-off between the scalability of controllers and latency. Mobility management in a core network based on SDN can potentially introduce delays due to the complex CP exchange. Although uRLLC applications over 5G networks integrated with MEC can significantly improve application performance, the geographically distributed architecture and vast volumes of data further increase network complexity. Therefore, a challenge remains to optimize the CP exchange and the current network architecture to meet the demand for ultra-low latency.

References

Abdulghaffar, A., Mahmoud, A., Abu-Amara, M., & Sheltami, T. (2021). Modeling and evaluation of software defined networking based 5G core network architecture. *IEEE Access*, *9*, 10179–10198.

E.Group. (2018). 5G; NG-RAN; architecture description. ETSI, Technical Report.

E.Group. (2019). 5G; NR; overall description; stage-2. ETSI, Technical Report.

Evangelista, J. V., Kaddoum, G., & Sattar, Z. (2021). Reliability and User-Plane Latency Analysis of mmWave Massive MIMO for Grant-Free URLLC Applications. arXiv preprint arXiv:2107.08151.

Fan, X., & Huo, Y. (2021). An Overview of Low latency for Wireless Communications: an Evolutionary Perspective. arXiv preprint arXiv:2107.03484.

G.Group.A. (2021). Service requirements for the 5G system. ETSI, Technical Report.

G.Group.A. (2020). System architecture for the 5G system. ETSI, Technical Report.

G.Group.B. (2020). Study on scenarios and requirements for next generation access technologies. ETSI, Technical Report.

G.Group.B. (2021). Procedures for the 5G system. ETSI, Technical Report.

Gaur, A. S. (2018). Containerized IoT Solution for Efficient Mobile Vertical Handover, Doctoral dissertation, Carleton University.

Gimenez, L. C., Michaelsen, P. H., Pedersen, K. I., Kolding, T. E., & Nguyen, H. C. (2017, June). Towards zero data interruption time with enhanced synchronous handover. In 2017 IEEE 85th Vehicular Technology Conference (VTC Spring) (1–6).

Guimaraes, R. S., Martínez, V. M., Mello, R. C., Mafioletti, D. R., Martinello, M., & Ribeiro, M. R. (2020, June). An SDN-NFV Orchestration for Reliable and Low Latency Mobility in

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447

Off-the-Shelf WiFi. In ICC 2020-2020 IEEE International Conference on Communications (ICC) (1–6).

Hattachi, R. E., & Erfanian, J. (2015). 5G white paper. NGMN White Paper. Available at https://www.ngmn.org/wp-content/uploads/NGMN_5G_White_Paper_V1_0.pdf

Huedo, E., Montero, R. S., Moreno-Vozmediano, R., Vázquez, C., Holer, V., & Llorente, I. M. (2021). Opportunistic deployment of distributed edge clouds for latency-critical applications. *Journal of Grid Computing*, *19*(1), 1–16.

Jain, A., Lopez-Aguilera, E., & Demirkol, I. (2018, September). Improved handover signaling for 5G networks. In 2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) (164–170).

Jin, R., Zhong, X., & Zhou, S. (2016, December). The access procedure design for low latency in 5G cellular network. In 2016 IEEE Globecom Workshops (GC Wkshps) (1–6).

Kasenides, N., & Paspallis, N. (2020, September). Multiplayer game backends: A Comparison of commodity cloud-based approaches. In European Conference on Service-Oriented and Cloud Computing (41–55). Springer, Cham.

Khaturia, M., Manjeshwar, A. N., Jha, P., & Karandikar, A. (2021, March). 5G-Serv: Decoupling User Control and Network Control in the 3GPP 5G Network. In 2021 24th Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN) (75–79). IEEE.

Kurian, A. (2018). Latency analysis and reduction in a 4G network. Technical University of Delft, Master thesis. Available at http://resolver.tudelft.nl/uuid:e1badd8d-a384-49a1-b958-a0c1e499c539

Li, Z., Wang, X., & Zhang, T. (2020). *5G+: How 5G change the society*. Springer Nature.

Martini, B., Paganelli, F., Cappanera, P., Turchi, S., & Castoldi, P. (2015, April). Latency-aware composition of virtual functions in 5G. In Proceedings of the 2015 1st IEEE conference on network softwarization (NetSoft) (1–6).

Mohammadkhan, A., Ramakrishnan, K. K., & Jain, V. A. (2020). CleanG—Improving the Architecture and Protocols for Future Cellular Networks With NFV. *IEEE/ACM Transactions on Networking*, *28*(6), 2559–2572.

Mohammadkhan, A., Ramakrishnan, K. K., Rajan, A. S., & Maciocco, C. (2016, December). Cleang: A clean-slate EPC architecture and control plane protocol for next generation cellular networks. In Proceedings of the 2016 ACM Workshop on Cloud-Assisted Networking (31–36).

Parvez, I., Rahmati, A., Guvenc, I., Sarwat, A. I., & Dai, H. (2018). A survey on low latency towards 5G: RAN, core network and caching solutions. *IEEE Communications Surveys & Tutorials*, *20*(4), 3098–3130.

Peltonen, A., Sasse, R., & Basin, D. (2021, June). A comprehensive formal analysis of 5G handover. In Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks (1–12).

Peñaherrera-Pulla, O. S., Baena, C., Fortes, S., Baena, E., & Barco, R. (2021). Measuring key quality indicators in cloud gaming: Framework and assessment over wireless networks. *Sensors*, *21*(4), 1387.

Copyright © 2022

Premsankar, G., Di Francesco, M., & Taleb, T. (2018). Edge computing for the Internet of Things: A case study. *IEEE Internet of Things Journal*, *5*(2), 1275–1284.

Shukla, S., Hassan, M., Tran, D. C., Akbar, R., Paputungan, I. V., & Khan, M. K. (2021). Improving latency in Internet-of-Things and cloud computing for real-time data transmission: a systematic literature review (SLR). *Cluster Computing*, 1–24.

Sreenibha Reddy, B., & Zaahid, M. (2018). Performance Metrics Analysis of Gaming Anywhere with GPU accelerated NVIDIA CUDA. Blekinge Institute of Technology, Master's thesis. Available at https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1261108&dswid=-6847

Tan, Z., Li, Y., Li, Q., Zhang, Z., Li, Z., & Lu, S. (2018). Supporting mobile VR in LTE networks: How close are we? *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, *2*(1), 1–31.

Tomaszewski, L., Kukliński, S., & Kołakowski, R. (2020, June). A new approach to 5G and MEC integration. In IFIP International Conference on Artificial Intelligence Applications and Innovations (15–24). Springer, Cham.

Zhou, P., Finley, B., Li, X., Tarkoma, S., Kangasharju, J., Ammar, M., & Hui, P. (2021). 5G MEC computation handoff for mobile augmented reality. arXiv preprint arXiv:2101.00256.

Copyright © 2022
http://doi.org/10.18080/jtde.v10n3.447