Low-latency 5G architectures for mission-critical Internet of Things (IoT) services

Changsoon Choi, Jong-Han Park, Minsoo Na and Sungho Jo
SK Telecom

Abstract

This paper presents design methodologies for 5G architecture ensuring lower latency than 4G/LTE. Among various types of 5G use cases discussed in standardization bodies, we believe mobile broadband, massive IoT(Internet of Things) and mission-critical IoT will be the main 5G use cases. In particular, a mission-critical IoT service such as remote controlled machines and connected cars is regarded as one of the most distinguished use cases, and it is indispensable for underlying networks to support sufficiently low latency to support them. We identify three main strategic directions for end-to-end network latency reduction, namely new radio access technologies, distributed/flat network architecture, and intelligent end-to-end network orchestration.

I. Introduction

Recent years have seen rapidly increasing interests in 5th generation (5G) as a next-generation mobile network. A large number of mobile network operators (MNOs), equipment manufacturers, research organizations and governments have been actively engaged in the discussion on 5G and its key enabling technologies in various standardization bodies. For example, International Telecommunication Union (ITU) coined the term “IMT-2020” to describe 5G, announced its plan towards 5G and specified 5G key capabilities[1]. In the meantime, global MNOs have also started sharing their initial views and thoughts on 5G technologies and potential use cases in GSMA and NGMN[2][3][4].

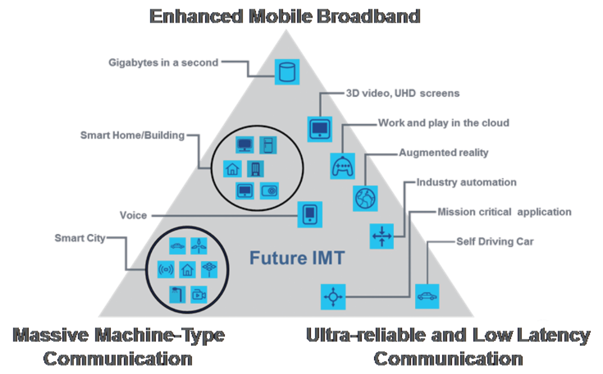
Enhanced Mobile Broadband
Gigabytes in a second
3D video, UHD screens
Work and play in the cloud
Smart Home/Building
Augmented reality
Industry automation
Voice
Mission critical application
Smart City
Self Driving Car
Future IMT
Massive Machine-Type Communication
Ultra-reliable and Low Latency Communication

Figure 1. 5G use cases and three main categories discussed in ITU-R WP5D

Although the degree of preference on different 5G use cases slightly varies amongst operators, there seems to be a consensus on the main categories into which most of suggested 5G use cases are organized. The three main 5G use cases are: 1) mobile broadband, 2) massive IoT and 3) mission-critical IoT. 〈Fig. 1〉 illustrates 5G use cases and three main categories discussed in ITU-R WP5D[1].

In 〈Fig. 1〉, enhanced mobile broadband is a use case category that extends the current broadband services towards more immersive services such as high quality multimedia, multi-angle/view video streaming, AR(Augmented Reality) and VR(Virtual Reality). This use case category calls for redesign of network architecture to efficiently support tremendous amount of real-time and immersive contents. Massive IoT (also referred to massive machine-type communication in ITU-R) use case category includes services built on the top of various sensors and things connected to the network at a massive scale. Even though data being sent and received by sensors and things are relatively small amount and may only require best-effort packet delivery, massive IoT requires mobile networks to be able

Figure 2. Remote–controlled robot demonstration with 5G radio in Mobile World Congress 2015, Barcelona

to handle massive number of simultaneous connections. Mission–critical IoT (also referred to ultra–reliable and low latency communication in ITU–R) includes mission–specific services such as remote controlled machine, autonomous driving with very strict requirement on latency, availability and reliability. For example, one of the highlighted 5G service demonstrations that SK telecom presented in the recent Mobile World Congress (MWC) 2015 event was a remote controlled robot with 5G radio, shown in 〈Fig. 2〉.

From operators' perspective, mission–critical IoT is seen as the most distinguished 5G use case compared to that of 4G, as it is expected to bring new possibility and to create opportunities for innovative 5G services which cannot be provided with 4G. To make mission–critical IoT practically available, it is critical for underlying network to be capable of providing low enough latency in the order of a few milliseconds from end–to–end perspective. Other benefits of lower latency include higher throughput, higher QoE, and low UE (User Equipment) buffer requirements.

In this paper, we present key strategic directions for achieving low latency from end–to–end network design perspective. Details of 5G key enabling technologies are also described with emphasis on how to reduce latency in 5G network architecture. It is expected that this guideline calls for a potential need to design and optimize end–to–end network for lower latency.

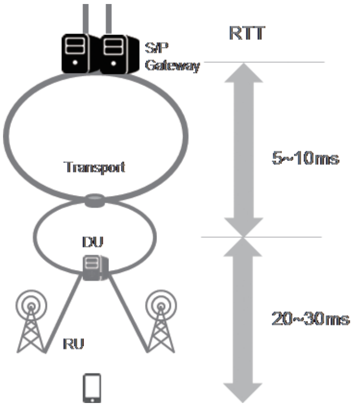

Figure 3. Network path and latency break–down in LTE–SAE architecture

Ⅱ. Low latency 5G architecture

〈Fig. 3〉 shows the legacy LTE–SAE network architecture between S/P (Serving/Packet Data Network) gateway and UE together with overall latency measurement results presented in [5]. Typically in current LTE networks, it takes approximately 25~40 milliseconds (one–way) for a given packet to travel from UE to P–gateway before reaching the Internet. Taking a closer look at each network node and link inducing additional latency with break–down of network segments, we identify the following three key directions for reducing end–to–end latency in 5G network architecture: 1) new radio access networks (incl. latency reduction techniques in LTE), 2) distributed/flat network architecture, 3) intelligent end–to–end network orchestration with unified and converged transport networks.

In the following subsections, we describe each of the above three key directions in more detail.

1. New Radio Access Technologies

As illustrated in Fig. 3, latency originated from radio access networks takes a dominant portion in end–to–end network latency. The latency requirement for 5G new radio access technology is still under discussion. However there seems to be a consensus on target latency of 1ms for 5G radio access network amongst operators and vendors. To achieve this, many research organizations and equipment manufacturers are currently looking into different aspects of the existing LTE technologies and new radio access technologies. Shorter TTI (transmission time interval) than 1ms of LTE system has been investigated while keeping the backward compatibility to the current LTE in 3GPP. Yet, LTE has a fundamental limitation that its TTI cannot be reduced below the OFDM symbol length. Therefore, 5G new RAT is expected to have shorter symbol length as well as shorter TTI than LTE to ensure lower latency in radio layer[6][7][8]. In addition, flexible duplex and new frame architecture have been extensively investigated for additional latency reduction. ⟨Fig. 4⟩ shows an example of 5G RAT(Radio Access Technology) frame architecture that allows shorter physical layer round trip time with fast control signaling and fast TDD (Time-Domain Duplex) data switching periodicity[8].

Radio protocol optimization is another approach for reducing radio latency. In current LTE networks, when a given UE has data in buffer to send in uplink, it first sends scheduling request (SR) to eNB and needs to wait for grant from eNB, as shown in ⟨Fig. 5⟩. If this process can be enhanced or even removed (by, e.g. uplink pre-grant), radio layer latency may be further reduced. To discuss L2 protocol enhancement in LTE and make a LTE standard amended, a study item for latency reduction techniques was recently approved in 3GPP RAN2 and now underway (it also consider backward-compatible TTI reduction in LTE)[9].

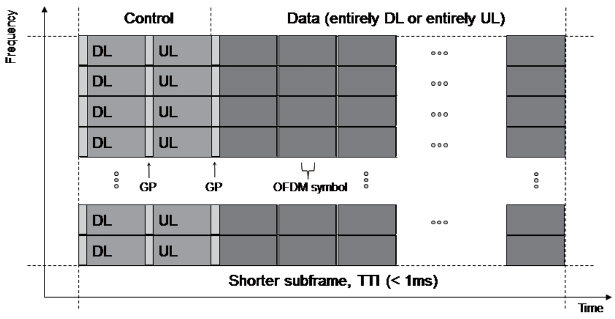
Frequency
Control
Data (entirely DL or entirely UL)
DL
UL
GP
GP
OFDM symbol
Shorter subframe, TTI (< 1ms)
Time

Figure 4. An example of 5G new RAT frame architecture providing lower latency than LTE

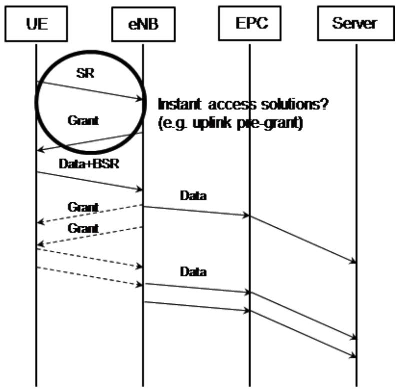

Figure 5. L2 protocol for LTE uplink access

Not all 5G radio access technologies improve latency. To boost up capacity in radio link, higher frequency such as millimeter-wave has been considered due to its high availability of broadband spectrum more than 100MHz. However, it suffers from higher atmospheric loss and higher constraint of NLOS(non-line-of-sight) while it allows more antenna to be integrated with the same form factor. Therefore, higher level of beamforming with more antennas is of high interest for millimeter-wave radio access links. Due to hardware complexities and power consumption issues, Analogue beamforming has typically been used for millimeter-wave, which cannot help sacrificing flexibility with less degree of freedom in intelligent beam searching as opposed to digital beamforming. This may induce increase in radio latency. For example, random access procedure needs more time to complete because PRACH needs to be transmitted multiple times to find best beams both in UE and in eNB as shown in ⟨Fig. 6⟩ [10]. In a higher mobility environment, beam searching procedure needs to be done more often, which would be another factor to increase radio latency in 5G. Therefore, careful design of beamforming architecture (e.g. hybrid beamforming architecture) and protocol (e.g. hierarchical beam searching) are indispensable for latency reduction in millimeter-wave 5G radio access networks.

2. Distributed architecture

As previously illustrated in ⟨Fig. 3⟩, the overall network path consists of mainly three parts: radio access,

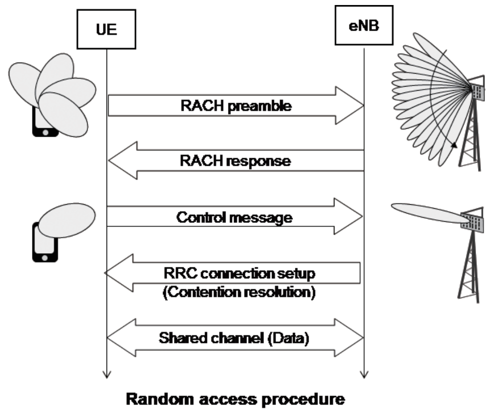

Figure 6. An example of random access procedure for beamforming–based radio access link

transport, and core network. In order to effectively reduce the overall end–to–end network latency, latencies associated with each of the three parts must be precisely identified and addressed. In the previous sub–section, we identified root causes of latency in radio access part. In this subsection, we investigate the core network, and discuss on how the core network architecture can be enhanced to reduce its latency.

In LTE–SAE architecture, all traffics must go through a single node named P–gateway. This centralized architecture is advantageous in terms of operations and managements. On the other hand, it may potentially pose a severe limitation for 5G network where lots of mobile backhaul traffics are expected. According to ITU–R discussion, maximum cell capacity for 5G base station is 20Gbps. Given that many base stations will be densely deployed and centralized for Cloud RAN, mobile backhaul traffic are expected to exceed even more than several hundred of Gbps.

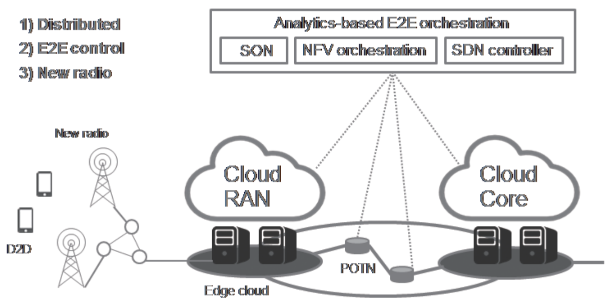

Figure 7. 5G end–to–end network architecture for low latency

One way to effectively avoid this potential traffic bottleneck is to transform the existing centralized architecture into a a more distributed architecture. For example, we can potentially place S/P gateways and have the traffic directly reach the Internet at the edge network closer to the user. ⟨Fig. 7⟩ shows an example of this approach, where future mobile network mainly consists of two types of clouds (i.e., Cloud RAN and Cloud Core). As opposed to today's LTE–SAE architecture where most of radio access functions are placed and run inside Cloud RAN, and most core functions (e.g., EPC) are placed and run inside Cloud Core, in 5G network architectures, the core functions may potentially be pushed out to the Cloud RAN. There are two main benefits of placing core functions at edge node where Cloud RAN is located. First, user packets can be directly sent out to the Internet, therefore reducing the overall latency. Second, as user packets are directly sent out to the Internet, the overall backhaul traffic being sent to the Cloud Core diminishes. This ultimately leads to the cost savings in mobile backhaul investments. Our preliminary experimental results showed an edge cloud would be able to reduce more than 30% of backhaul traffic with the help of mobile edge caching. There are various additional use cases and approaches to leverage this type of edge cloud, also referred as mobile edge computing discussed in ETSI MEC ISG and Fog computing[11][12]. Edge cloud can also eliminate significant amount of signaling traffic by placing relevant control and even analytics functions close to devices and things. Other benefits of edge computing include higher level of security, as all signals and data traffic can be processed locally, and this is particularly attractive for public or B2B applications. ⟨Figure 8⟩ shows comparison between LTE–SAE architecture and flat/latency–optimized networks we have been designed for 5G network architecture. Detailed information on the 5G architecture that SK telecom has been developing can be found in [13]

A lots of efforts have been done for the D2D (Device–to–Device) standardization in 3GPP Release 12 and we believe D2D will become increasingly widespread in 5G. This also implies 5G network becomes more

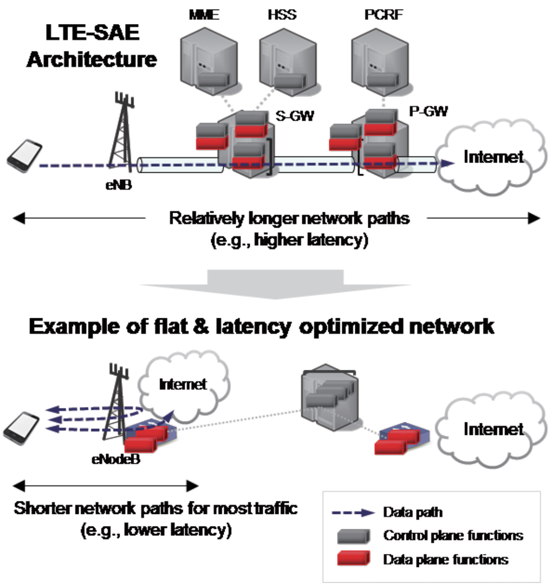
LTE-SAE Architecture
MME
HSS
PCRF
S-GW
P-GW
Internet
eNB
Relatively longer network paths (e.g., higher latency)
Example of flat & latency optimized network
Internet
Internet
eNodeB
Shorter network paths for most traffic (e.g., lower latency)
Data path
Control plane functions
Data plane functions

Figure 8. Evolution toward flatter network architecture for latency reduction

distributed with D2D although cellular-assisted operation is somehow required from the viewpoint of QoS management.

3. Intelligent E2E network orchestration

In order to provide service QoS, the network is typically overprovisioned and the resources used to implement the service are also sufficiently allocated such that QoS can be guaranteed even during the peak usage period. This has been the de-facto practice for guaranteeing QoS. However for 5G, this overprovisioning will no longer be a viable option as the peak usage is expected to be too high and ultimately lead to a cost prohibitive system. Therefore, the network and network functions must be designed to support policies for guaranteeing QoS even in the case of network failures and during the peak usage times.

〈Fig. 7〉 also illustrates an end-to-end policy-based network service orchestrator as a key enabler that intelligently manages QoS and minimize latency dynamically in real-time even during the network is congested. More specifically, a central controller would consist of SDN controller (or Network OS) and NFV orchestrator interworking based on policies to dynamically change the network in order to provide QoS (e.g., minimum latency) for a given service. Analytics-engine will play an increasingly important role and can also be integrated with PCRF (Policy and Charging Rule Function) to promise intelligent operation of E2E networks.

The central orchestrator can also be used to optimize transport network path, via transport SDN (T-SDN) that can dynamically modify configurations of routers in transport networks. Looking at an individual optical router, different network layers from L1 to L2~L3 are now being integrated with advanced optical technologies, representatively called POTN (packet optical transport networks). As shown in 〈Fig. 9〉, this allows network switching as much of lower (optical) layer processing as possible for low latency services. IP layer switching usually takes 40~50 microseconds, on the other hand, optical layer switching takes less than 50 nanoseconds. Technical challenges includes the limited flexibility for optical switching and wavelength allocations for beyond 100Gbps transmission and we expect elastic optical networking technologies play important roles to get through these problems[14]. Unified and converged dynamic transport networks realized by T-SDN and POTN can also reduce overprovisioning in backhaul/metro transports, attractive for CAPEX/OPEX reduction.

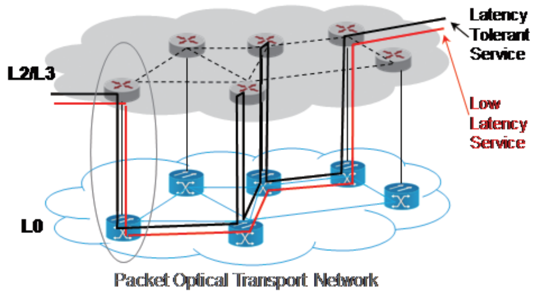

Figure 9. Packet optical transport network (POTN) and its optical routing for low latency services

III. Conclusion

Mission-critical IoT is one of the most attractive use cases in 5G as it creates myriad opportunities and challenges for the overall 5G stakeholders including operators, equipment vendors, and users. By carefully observing the existing LTE network architecture, we identified three main areas that need significant enhancements to support ultra-low latency requirements, which are 1) new radio access networks, 2) distributed/flat network architecture, 3) intelligent end-to-end network orchestration with unified and converged transport networks. This observation and study calls for careful design of the network to ensure 5G properly and efficiently supporting mission-critical IoT use cases.

Reference

[1] ITU towards “IMT for 2020 and beyond”, http://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2020/Pages/default.aspx

[2] NGMN – 5G initiative and whitepaper 1.0, https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf

[3] GSMA – Network2020, http://www.gsma.com/network2020/

[4] GSMA, “Understanding 5G: perspective on future technological advancements in mobile,” (available in https://gsmaintelligence.com/research/?file=141208-5g.pdf&download)

[5] NOKIA, “Technology vision 2020 reducing network latency to millisecond”, Jul. 2014

[6] E. Lahetkangas, K. Pajukoski, , E Tiirola, G. Berardinelli, I. Harjula and J. Vihriala, “On the TDD subframe structure for beyond 4G radio access network”, Future Network and Mobile Summit, Jul. 2013

[7] T. Levanen, J. Pirskanen, T. Koskela, J. Talvitie and M. Valkama, “Low latency radio interference for 5G flexibile TDD local area communication,” International Conference on Communication, 2014

[8] E. Lahetkangas, K. Pajukoski, J. Vihriala and E Tiirola, “On the flexible 5G dense deployment air interface for mobile broadband”, International Conference on 5G for Ubiquitous Connectivity, Nov. 2014

[9] RP-150310, “New SI proposal: Study on latency reduction techniques for LTE”, 3GPP RAN plenary meeting #67, Mar. 2015.

[10] C. Jeong, J. Park and H. Yu, “Random access in millimeter-wave beamforming cellular network: issues and approaches,” IEEE Communication Magazine, Jan. 2015

[11] ETSI mobile edge computing (MEC) ISG whitepaper, https://portal.etsi.org/Portals/0/TBpages/MEC/Docs/

[12] F. Bonomi, R. Milito, J. Zhu and S. Addepalli, “Fog computing and its role in the Internet of Things,” Sigcomm 12, 2012

[13] S1-150196, “SKT's 5G core network (SK Telecom's view), 3GPP TSG-SA WG1 Meeting #69, Feb. 2015

[14] M. Channegowda, R. Nejabati, and D. Simeonidou, “Software-defined optical networks technology and infrastructure: Enabling software-defined optical network operations,” IEEE Journal of Optical Communication and Networking, Oct. 2013

약력

최창순

1999년 연세대학교 공학사
2001년 연세대학교 공학석사
2005년 연세대학교 공학박사
2005년~2007년 일본 정보통신연구기구(NICT) Research Engineer
2008년~2010년 독일 라이프니츠 연구소(IHP) Research Scientist
2010년~2012년 독일 NTT DOCOMO 유럽 연구소 Senior Research Engineer
2012년~현재 SK텔레콤 종합기술원 5G TechLab
관심분야: 5G Architecture, 5G Radio Access, Millimeter–wave/Beamforming, NFV/SDN

박종한

2000년 University of California, San Diego 공학사
2006년 University of California, Irvine 공학석사
2011년 University of California, Los Angeles 공학박사
2011년~2014년 AT&T 연구소 Senior Member of Technical Staff
2014년~현재 SK텔레콤 종합기술원 5G TechLab 연구원
관심분야: 5G [R]evolution, NFV/SDN, Orchestration

나민수

2007년 고려대학교 공학사
2009년 서울대학교 공학석사
2009년~2014년 SK텔레콤 Network기술원 Access Network Lab
2014년~현재 SK텔레콤 종합기술원 5G Tech Lab
관심분야: LTE/LTE–A, Small Cell, SDN/NFV, 5G

조성호

1997년 경북대학교 공학사
2014년~현재 고려대학교 경영전문대학원 MBA
1997년 SKTelecom Network부문 입사
2000년~2014년 SK텔레콤 Network기술원 Access Network Lab
2015년~현재 SK텔레콤 종합기술원 5G Tech Lab장
관심분야: 5G, mmWave, Massive MIMO, Beamforming, SDN, NFV