2022 International Symposium on Networks, Computers and Communications (ISNCC) | 978-1-6654-8544-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISNCC55209.2022.9851705

New IP Enabled End-to-End Latency Guarantee for Downlink Traffic in 5G

Lijun Dong, Lin Han, Richard Li
Futurewei Technologies Inc.
2330 Central Expressway
Santa Clara, CA, U.S.A
{lijun.dong, lin.han, richard.li}@futurewei.com

Abstract— Precise end-to-end latency guarantee is predicted to be required by many emerging applications. On the other hand, the network traffic will continue to be dominated by mobile devices. Therefore, the end-to-end latency is composed of the latency incurred in the Internet as well as in the mobile networks. In this paper, we target to address the end-to-end latency guarantee requirement for downlink traffic by leveraging a new type of 5G slice namely, LGS (Latency Guarantee Service) slice. The mechanisms and procedures are proposed by taking the compatibility of 5G architecture into consideration. The simulation results show that the downlink flows which are admitted by the LGS slices could satisfy the end-to-end latency constraint consistently.

Keywords— latency guarantee service, in-time, QoS, 5G, downlink traffic, slice, New IP

I. INTRODUCTION

Existing network applications might suffer from a gradual deterioration of the Quality of Experience (QoE) when underlying network condition is degraded. For example, audio may be broken, or it takes longer time to load web pages, but the application as a whole may still fundamentally be workable. In contrast, many emerging network applications [1] are mission-critical and extremely sensitive to violation of required end-to-end latency guarantee, any misses would not result in graceful degradation, instead would cause a disastrous breakdown, and even endanger human lives. Today's Internet technologies built on the best effort (BE) principle cannot readily support these applications with characteristics of high precision, low latency, and low tolerance to degradation. Traditional techniques that provide certain degree of compensation at upper layers for any deficiencies, such as retransmission (to recover from packet loss) or adaptive rate control (reducing sending rate to fight ongoing network congestion) may no longer be sufficient.

Taking remote driving as an example, which is one emerging variant of driverless vehicle technology, remote operations require every piece of information from vehicle-driving commands to continuous feeds of video and telemetry data to reach the remote operator in time. Unexpected obstacles or events can happen randomly, e.g., a deer running out of the forest to the highway. At a speed of 100 km/h, a vehicle travels roughly 27 meters per second. The end-to-end latency includes multiple folds, e.g., processing time required for rendering, latency incurred by the network, the latency incurred by human or automated reaction time, which adds to the distance required for the vehicle to come to a stop. High reliability and avoidance of packet loss are critical as retransmissions would result in unacceptable increases in delay.

In order to support those emerging applications like remote driving, in-time service is needed to ensure delivery of packets within a bounded latency, the end-to-end latency is precisely guaranteed (namely Latency Guarantee Service, i.e., LGS). The end-to-end latency is defined as the time period which elapses from when a data packet is sent by a sender (i.e., from when the transmission of the packet's first bit) until when the data packet (i.e., the packet's last bit) is received by a receiver across the network.

It is envisioned that mobile device access would continue to increasingly dominate the Internet traffic [2]. Thus, with one end in the cellular network (e.g., 5G for broadband cellular networks in recent years and near future) and the other end likely locating in the Internet, the end-to-end latency guarantee would require the latency guarantee from the Internet as well as the cellular network to work collaboratively. Many previous works have mainly focused on improving Quality of Service (QoS) or achieving the latency guarantee of traffic that is within the Internet scope. To the best of our knowledge, our previous work [3] was the first in the literature to consider this problem. This paper continues the path of pursuing the concept of LGS slice dedicated for flows that require in-time guarantee and develops the mechanisms and procedures for downlink traffic.

The rest of the paper is arranged as follows: Section II presents the related works, including the existing mechanisms of improving QoS and achieving latency guarantee in the Internet, a specially designed type of 5G slice called LGS slice [3] to address the gap of achieving the latency guarantee in 5G domain; Section II describes the procedures and mechanisms for latency guarantee of downlink traffic with LGS slice. Section III presents the simulation results. Section IV concludes the paper.

A. Overview of Internet QoS Mechanisms

IETF has defined two complementary foundational QoS architectures, DiffServ [4] and IntServ [5]. When congestion already happens, it is impossible for DiffServ to mitigate such situations, in which packet dropping and packet retransmission are inevitable. Diffserv does not adapt to changing behaviors of

978-1-6654-8544-9/22/$31.00 ©2022 IEEE

applications and its QoS guarantee granularity is very raw at the traffic class level. The IntServ QoS model allows a flow to request resources from the network, correspondingly the network would perform deterministic admission control of the flow based on its available resources. However, its scale is limited in small networks, so very hard to implement or scale to large networks like in the whole Internet scope.

More recently, the IETF DetNet Working Group has proposed the DetNet architecture (Deterministic Networking Architecture) [6]. The DetNet architecture intends to provide guarantee on bounded delay as well as packet loss ratio at per flow basis. Similar to IntServ, DetNet also targets reservations according to the applications' requested CBR (Committed Bit Rate), while many emerging and future applications involve highly variable bitrates.

Resource reservation and flow admission control with in-band signaling in the Internet are proposed in [7][8] to make sure an admitted flow would have an end-to-end bounded latency. However, such guarantee focused on the traffic within the Internet scope.

LBF (Latency-Based Forwarding) [9] is a recent proposal that provides support for in-time and on-time services with concise end-to-end latency objectives at packet level. Each intermediate router in the Internet assesses whether the packet is on track to meet its latency objective and determines a local time budget (i.e., dwell time of the packet in the router) before the packet leaves the current router and is forwarded to the next hop. [10] builds on LBF and proposes an optimal scheduling algorithm that minimizes the average dwell time for all packets in the queue, but assuming that all packets can meet their deadlines under the scheduling algorithm. Although packet-level latency guarantee is more granular compared to flow-level latency guarantee service, the above proposals were still under the assumption that the packets only flow within the Internet scope.

B. Overview of LGS Slice in 5G

Network slicing [11][12] addresses the issues of conventional "one-size-fits-all" design, i.e., lack of flexibility, scalability, and isolation. A 5G network slice is self-contained and logically isolated networks, such that over the same network infrastructure, diversified applications can be accommodated simultaneously with their service requirement being met respectively and without any affect from other slices. Network slice orchestration can be undertaken in different domains namely: RAN (Radio Access Networks), core, DN (Data Networks), or end-to-end as shown in Fig. 1.

In our previous work [3], we briefly introduced a new slice type, namely LGS, which is specifically configured to provide stringent latency guarantee service. An LGS type slice is associated with the maximum latency that could be incurred between the UE (User Equipment) and the UPF (User Plane Function) that terminates the N6 interface, which corresponds to the PDB (Packet Delay Budget) as one of the 5G QoS characteristics. The LGS type slice in 5G also involves the 5G-RAN slice, Transport slice, Core slice. The latency generated in the LGS slice is the combination of the latency incurred in 5G-RAN slice, Transport slice, and Core slice. As shown in Fig. 2, the appropriate 5G-RAN slice, Transport slice and Core slice are chosen to construct an LGS slice with the sum of their corresponding PDB values equal to the PDB requirement of the LGS slice. An application in an end device could select one of the available LGS slice instances, whose end-to-end PDB is small enough to satisfy the application's stringent end-to-end latency requirement.

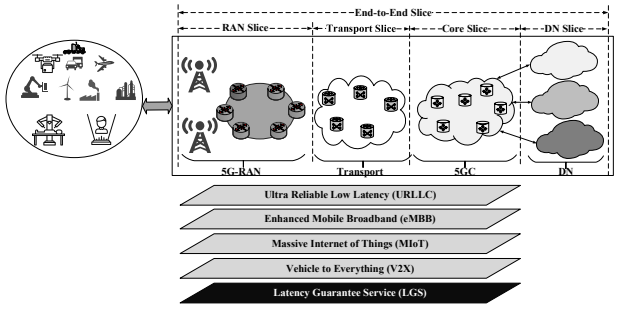

Fig. 1. End-to-end slice for Latency Guarantee Service (Some icons made by Freepik from www.flaticon.com)

We may reuse the design of the 5G-RAN slice in URLLC (Ultra Reliable Low Latency Communication) or V2X (Vehicle to Everything), which is optimized to provide bounded latency within the 5G-RAN. The following constraint on PDB in the Transport slice and Core slice must be satisfied:

$$PDB_{TP+5GC} = E2EPDB_{requsted} - PDB_{DN} - PDB_{5GRAN}$$

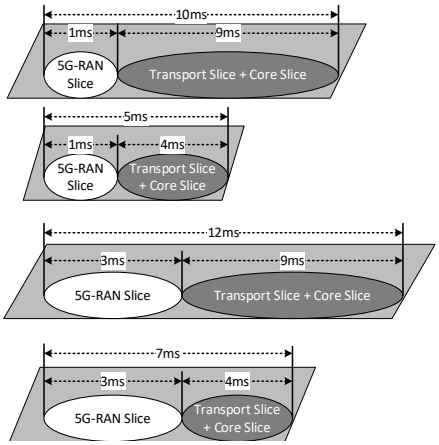

Fig. 2. LGS Slices with various PDB

II. Latency Guarantee for Downlink Traffic

In this paper, we mainly focus on downlink flows that require in-time guarantee. The similar mechanisms could apply to other type of traffic, e.g., uplink flows, UE-to-UE flows. The source node in the DN (we use Internet as the example) needs to initiate a request-to-send message, which contains the following information:

- Latency guarantee service indicator: it indicates that the source node in the DN requests for the end-to-end latency guarantee service for the data transmission towards the UE.
- End-to-end PDB requirement ($E2EPDB_{requsted}$): it denotes the end-to-end latency budget between the time

when the source node sends the packet and when it reaches the UE.

- Maximum flow rate (MBR): it denotes the maximum flow rate that the source node will generate towards the UE.
- Destination UE default IP address: it is associated with the default PDU session that UE establishes towards the DN.

If New IP [13][14] is used as the unified framework for the data plane as shown in Fig. 3 and the "Request-to-Send" is regarded as data plane message that is initiated by the node in DN for the downlink data, the contract clause will include the Action as "***LatencyGurantee***" and the Metadata as the above information. The New IP header including the action and metadata could be accessed by each intermediate node, i.e., every on-path router in the Internet, network function nodes in 5G.

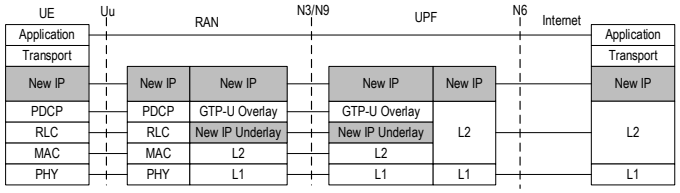

Fig. 3. 3GPP Protocol Stack with New IP at UE/Backhaul

When the Request-to-Send message is routed in the Internet, the $E2EPDB_{requsted}$ is updated by subtracting PDB_{DN} from it. Therefore, when the Request-to-Send message reaches the 5G domain (i.e., UPF), the $E2EPDB_{requsted}$ denotes the total PDB of 5GC, transport network and 5G-RAN ($PDB_{TP+5GC} + PDB_{5GRAN}$), which is referred to be $5GPDB_{requsted}$ in the following text.

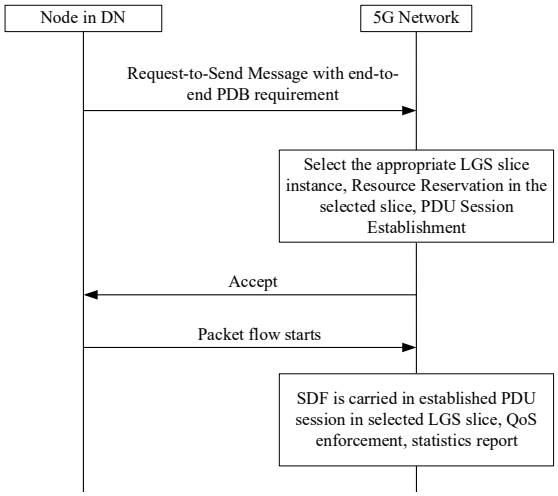

Fig. 4. High level procedure for downlink latency guarantee traffic (accepted secenario)

In the 5G domain, based on $5GPDB_{requsted}$, the decision on whether the Request-to-Send is accepted or rejected is going to be made. If there exists a LGS slice instance that the UE subscribes to, which can guarantee the latency to be smaller than $5GPDB_{requsted}$, the Request-to-Send is accepted. Otherwise, the Request-to-Send is rejected. Fig. 4 shows the high-level procedure for downlink latency guarantee traffic when Request-to-Send is determined to be accepted. Accordingly, an appropriate LGS slice instance is chosen, resource is reserved in this slice. Either a new PDU session is established between the UE and the Internet in the selected LGS slice instance, or an existing PDU session within this selected LGS slice instance is activated. The upcoming SDF (Service Data Flow) is associated with the QoS of latency guarantee, with the PDB characteristic set to be equal to $5GPDB_{requsted}$.

Fig. 5 shows the high-level procedure for downlink latency guarantee traffic when Request-to-Send is determined to be rejected, because none of the UE-subscribed LGS slice instances can satisfy $5GPDB_{requsted}$. Under this circumstance, the smallest latency that can be provided by the LGS slice instances that the UE subscribes to is calculated and attached within the rejection response to the source node in the Internet. The source node may take the suggestion to adjust the latency guarantee requirement and start a new round of Request-to-Send procedure.

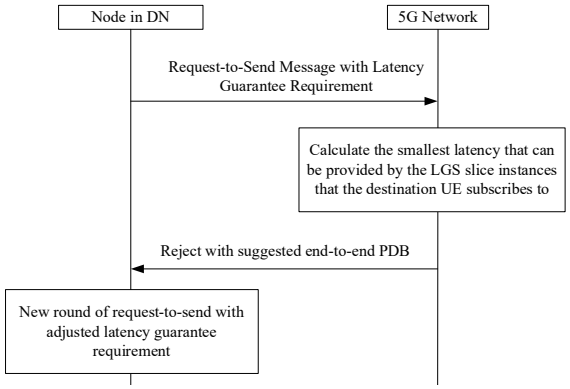

Fig. 5. High level procedure for downlink latency guarantee traffic (rejected secenario)

1) LGS Slice Instance Selection and Rescource Reservation

Fig. 6 shows the major Network Functions and interfaces in 5G architecture [15][16]. The proposed mechanisms and procedure mainly involve the following Network Functions:

- AMF stands for "Access and Mobility Management Function". It interacts with the radio network and the devices through signaling over the N2 and N1 interface respectively.
- SMF is the "Session Management Function", which oversees the establishment, modification and release of individual sessions, and allocation of IP addresses per session.
- UPF stands for "User Plane Function", which has the major tasks of processing and forwarding user data. The functionality of the UPF is controlled from the SMF. It connects with external IP networks and acts as a stable IP anchor point for the devices towards external networks. This means that IP packets with a destination address belonging to a specific device is always routed

from the Internet to the specific UPF that is serving this device even though the device is moving around in the network.

- UDM is the "Unified Data Management Function". It acts as a front-end for the user subscription data stored in the UDR (Unified Data Repository) and executes several functions on request from the AMF.
- NSSF is the "Network Slice Selection Function", which can be leveraged by the AMF to assist with the selection of the network slice instances that will serve a particular device. As such, the NSSF will determine the allowed NSSAI (Network Slice Selection Assistance Information) that is supplied to the device.

Before a UE can access an LGS Slice, the UE needs to register it with the network and that is done using the registration procedure. Multiple or all LGS Slices may be registered by a UE for latency guarantee service, here we assume that the UE has subscribed to such latency guarantee service. The UE's subscription information can be retrieved from UDM.

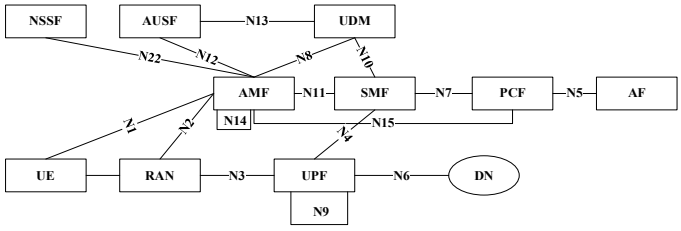

Fig. 6. 5G Nodes, Interfaces

The AMF or NSSF may be configured for LGS Slice selection. We illustrate the procedure as shown in Fig. 7, assuming the NSSF is in charge of LGS Slice selection for downlink traffic. When the UPF receives "Request-to-Send" message with latency guarantee requirement from a node in DN, the UPF could obtain the remaining latency budget for 5G network. The remaining end-to-end PDB is reduced while the "Request-to-Send" message traverses the DN, the corresponding New IP Metadata (end-to-end PDB requirement) is updated for each hop. Once the UPF receives the "Request-to-Send" request, the remaining PDB for the 5G network can be extracted from this New IP Metadata. Based on the standard 5G interfaces as shown in Fig. 6, the remaining PDB for the 5G network and the destination UE identifier is forwarded through the SMF, AMF to NSSF for LGS slice selection. Each LGS slice which the destination UE subscribes to and establishes PDU session in is associated with a maximum latency value that may be incurred in 5G network. By comparing the maximum latency value with the remaining PDB included in the LGS slice selection request, the NSSF can return the NSSAI of the selected LGS slice, which generates a latency in 5G network smaller than the remaining PDB and can accommodate the MBR of the upcoming downlink flow, i.e.:

- In the selected LGS slice, for an intermediate router with ingress rate as $R_{ingress}$, the MBR needs to satisfy: $MBR \leq R_{ingress} - \sum_{i=1}^{n} MBR_i^{LGS}$. $\sum_{i=1}^{n} MBR_i^{LGS}$ is the total MBR of already admitted n number of LGS flows in the selected LGS slice.
- In the selected LGS slice, for an intermediate router with egress rate as R_{egress}, the MBR needs to satisfy: $MBR \leq R_{egress} - \sum_{i=1}^{n} MBR_i^{LGS}$.

If there are multiple LGS slices that satisfy the above criteria, the LGS slice that could yield the smallest end-to-end latency in 3GPP 5G network is chosen.

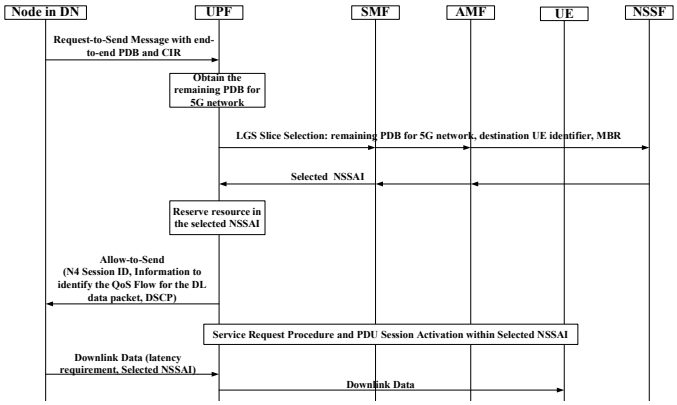

Fig. 7. Request-to-Send, LGS Slice instance selection, Allow-to-Send

The UPF receives the selected NSSAI returned from the NSSF via the AMF, SMF. The resource in the 5G core network (intermediate routers), and in the 5G RAN network is reserved to accommodate the upcoming downlink traffic. The QoS of the flow is set to be "Latency Guarantee", which may be assigned with a new 5QI to identify the QoS flow and a new DSCP (Differentiated Services Code Point) identifier could also be allocated for "Latency Guarantee" in 5G core network and in the DN. A PDU session needs to be established or activated for the upcoming traffic. Assuming a PDU session for the destination UE in the selected LGS slice has established before but is under deactivated status, instead of waiting for the first downlink packet arrival to trigger the service request procedure, we propose that the service request procedure is carried out right after the UPF is acknowledged with the selected LGS slice information from the NSSF. In this way, there is no delay for service request procedure after the first downlink packet arrival, the transmission towards the destination UE could happen right away, which further ensures the latency guarantee. If none of the LGS slices that the UE subscribes to could satisfy the criteria listed above, the "Request-to-Send" will be rejected, some suggested end-to-end PDB in 3GPP 5G network will be proposed by the NSSF based on the currently available LGS slices and their maximum end-to-end latency characteristics.

2) Service Request and PDU Session Establishment/Activation in LGS Slice Instance

If the UE is in CM-IDLE state or CM-CONNECTED state in 3GPP access, the network initiates a Network Triggered Service Request procedure right after UPF gets confirmation from NSSF that the Request-To-Send is allowed with a selected LGS Slice instance (i.e., Selected NSSAI is returned to the UPF from NSSF).

If the UE is in CM-IDLE state, and asynchronous type communication is not activated, the network sends a Paging Request to (R)AN/UE. The Paging Request triggers the UE Triggered Service Request procedure in the UE.

The UE sends a 5GSM (5G Session Management) NAS (Non-Access-Stratum) PDU Session Establishment message to the AMF, including PDU Session Id, selected NSSAI (The node in DN will request this NSSAI for the downlink data), PDU Session type, etc. The Request Type indicates "Initial request" if the PDU Session Establishment is a request to establish a new PDU Session.

The AMF then forwards the 5GSM container (containing the PDU Session Establishment message) to the SMF. The SMF retrieves the Session Management related UE subscription data from UDM. It is assumed that the UE subscribes to the latency guarantee service as a recipient, thus is eligible to use the selected LGS slice instance in the 5G infrastructure.

3) Data Transmission and Statistics Report

When the downlink traffic that requires latency guarantee arrives at the UPF, the PDU session lookup identifies the PDU session by matching the PDR (Packet Detection Rule) with highest precedence. If the DN is Internet, then the PDR could be IP header filter by matching the source/destination address pair, port number and protocol type. The selected LGS slice and the activated PDU session would be used to transport the downlink traffic.

4) PDU Session Deactivation in LGS Slice

The UPF detects that the PDU Session has no data transfer for an inactivity period specified by the Inactivity Timer, it reports PDU Session Inactivity to the SMF. The SMF may decide to release the UPF of N3 terminating point. The resources reserved (including bandwidth and cache storage) in the LGS slice for the flow are released as well.

III. Performance Evaluation

OMNeT++ [17] is used for simulation and performance evaluation. In the simulation scenario as shown in Fig. 8, we have servers located on the one end of the Internet sending downlink traffic to UEs in 5G network.

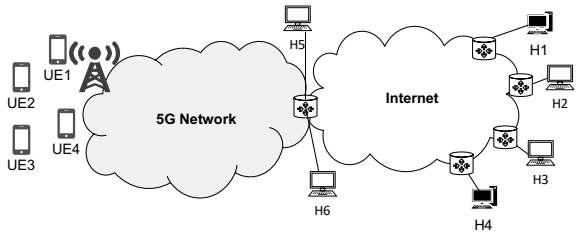

Fig. 8. Simulation scenario

We assume the routers in the Internet follow the design as proposed in [8] to make sure that the latency incurred in the Internet would be within a bounded value. In the following, we would like to show the end-to-end latency and throughput performance of the downlink flows and prove that the proposed LGS slice and procedures in 5G domain do achieve the performance goals. Multiple UDP (User Datagram Protocol) downlink traffics are configured between pairs:

- One downlink LGS flow originated from H1 and terminated at UE1 (corresponds to App0 at H1), with H5 residing on the other side of Internet and UE1 as a mobile user device. The MBR of this flow is 12 Mbps.
- One downlink LGS flow originated from H4 and terminated at UE4 (corresponds to App0 at UE4), with H4 residing on the other side of Internet and UE4 as a mobile user device. The MBR of this flow is 8 Mbps.
- Some other downlink flows configured other than LGS (Best Effort), from H2 to UE2, from H5 to UE2, from H3 to UE3 and from H6 to UE3. They do not require end-to-end latency guarantee.

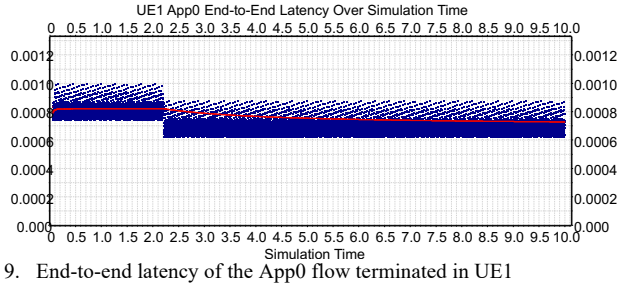

Fig. 9. End-to-end latency of the App0 flow terminated in UE1

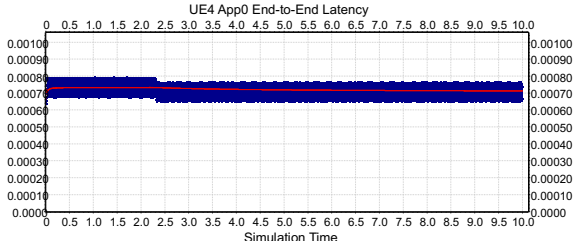

Fig. 10. End-to-end latency of the App0 flow terminated in UE4

Firstly, we take a look at the end-to-end latency performance of the UDP flows. Fig. 9 shows the end-to-end latency of UE1 App0. Fig. 10 shows the end-to-end latency of UE4 App0. Both downlink traffic flows require the precise end-to-end latency guarantee. We can observe that the end-to-end latency of packet delivery for both flows is within the specified deadline with a steady mean value. The proposed LGS slice mechanisms and procedures ensure that the two flows are only admitted when the selected LGS slice has enough resources to accommodate the requested MBR and the resulted latency satisfies the constraint.

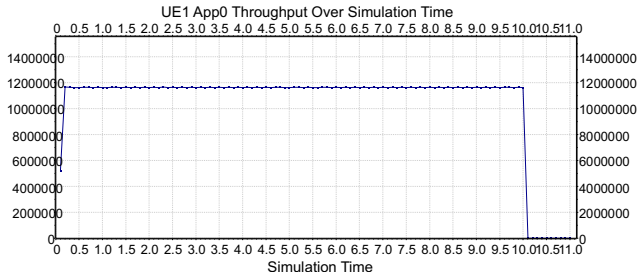

Fig. 11. Throughput of the App0 flow terminated in UE1

In the meantime, Fig. 11 and Fig. 12 confirm that the throughput of the App0 flow terminated in UE1 is maintained at 12 Mbps and the throughput of the App0 flow terminated in UE4 is maintained at 8 Mbps as requested. The routers in the selected

LGS slices reserve the ingress and egress bandwidth to be at least 12 Megabytes and 8 Megabytes, respectively.

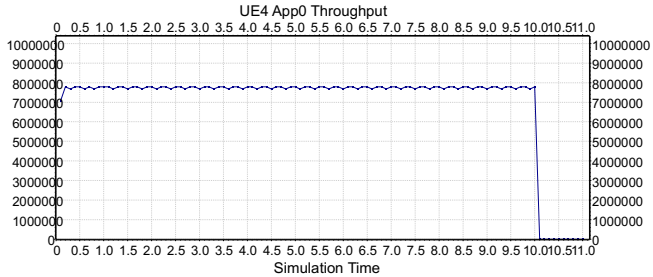

Fig. 12. Throughput of the App0 flow terminated in UE4

We also check the end-to-end latencies of other non-LGS downlink traffic flows. Fig. 13 shows the end-to-end latency of App0 flow terminated in UE2 as the example. The BE class traffic is pumped into network with higher bit rate than the remaining bandwidth after the routers have reserved resource (bandwidth) for LGS traffic. This will severely congest all routers from host to host. We can see the end-to-end latency fluctuates over the time, with most of them much higher than expected. BE traffic is scheduled with the residual resources. Thus, its throughput cannot be guaranteed at all as shown in Fig. 14. Although each individual BE flow is pumped into the network with pre-configured 5 Mbps bit rate, the maximum throughput it can achieve is as low as 0.4 Mbps for App0 terminated in UE2.

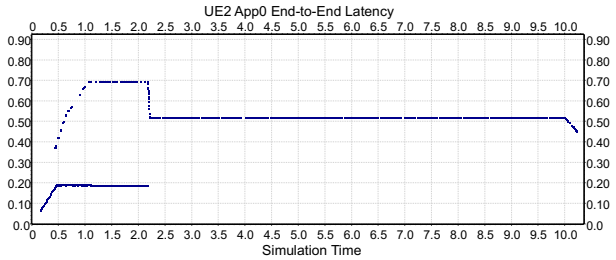

Fig. 13. End-to-end latency of the App1flow terminated in UE3

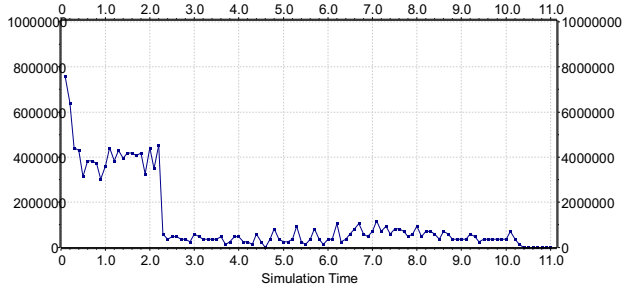

Fig. 14. Throughput of the App1 flow terminated in UE3

IV. Conclusions

With mobile devices dominating the Internet access, the end-to-end latency guarantee which is required by many emerging and future applications needs the support and enabling technologies for both the Internet and the cellular network. This paper focuses on the downlink traffic which terminates at the mobile devices (e.g., UEs) in 5G network. By utilizing a special 5G slice, namely LGS slice, all downlink traffic flows that require end-to-end latency guarantee need to be admitted by one of the LGS slices registered by the UE. The paper introduces the relevant procedures that involve various network functions in 5G architecture. The performance evaluation shows the end-to-end latency for downlink flows admitted by LGS slices always achieves the required upper bound. In the meantime, the throughput of those flows can reach the MBR consistently over the simulation time. Comparably, for other classes of flows (e.g., BE flows), neither the end-to-end latency nor the throughput could be guaranteed. The similar mechanisms could also be applicable to uplink flows, UE-to-UE flows that require LGS guarantee.

References

[1] 2030: Description, Technical Gap and Performance Target Analysis," 2019.

[2] E. Enge, Mobile vs. Desktop Usage in 2020, https://www.perficient.com/insights/research-hub/mobile-vs-desktop-usage.

[3] L. Dong, R. Li, "Latency Guarantee Service Slice in 5G and Beyond," IEEE CCNC 2022.

[4] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Services," RFC 2475, IETF, December 1998.

[5] R. Braden, D. Clark, S. Shenker, "RFC 1663: Integrated Services in the Internet Architecture: An Overview," IETF, Jun. 1994.

[6] N. Finn, P. Thubert, B. Varga and J. Farkas: "Deterministic Networking Architecture." RFC 8655, IETF, October 2019.

[7] L. Han, Y. Qu, L. Dong, R. Li, "A Framework to Realize the Guaranteed Service for Bandwidth and Latency for Future IP network," 2020 Infocom workshop on New IP: The Next Step.

[8] L. Dong and L. Han, "New IP Enabled In-Band Signaling for Accurate Latency Guarantee Service, " 2021 IEEE Wireless Communications and Networking Conference (WCNC).

[9] A. Clemm, T. Eckert, "High-Precision Latency Forwarding over Packet Programmable Networks," IEEE/IFIP Network Operations and Management Symposium, 2020.

[10] L. Dong, R. Li, "Packet Level In-Time Guarantee: Algorithm and Theorems", IEEE Globecom 2020.

[11] 5G Specifications in 3GPP: North American Needs for the 5G Future, ATIS-I-0000078, July 2020.

[12] How is 5G Slicing different from QoS, 5GWorldPro.com, July 24, 2019.

[13] R. Li, K. Makhijani and L. Dong, "New IP: A Data Packet Framework to Evolve the Internet," IEEE HPSR 2020.

[14] R. Li, A. Clemm, U. Chunduri, L. Dong, and K. Makhijani, "A New Framework And Protocol For Future Networking Applications," ACM Sigcomm Workshop on Networking for Emerging Applications and Technologies (NEAT 2018), pp. 637–648, May 2018.

[15] 3GPP TS 23.501, Technical Specification Group Services and System Aspects, System architecture for the 5G System (5GS).

[16] 3GPP TS 23.502, Technical Specification Group Services and System Aspects, Procedures for the 5G System (5GS).

[17] OMNET++, https://omnetpp.org/.