

August 30, 2018

Architectural View of E2E Latency and Gaps

[draft-dunbar-e2e-latency-arch-view-and-gaps-02.txt](#)

Abstract

Ultra-Low Latency is a highly desired property for many types of services, such as 5G MTC (Machine Type Communication) requiring E2E connection for V2V to be less than 2ms, AR/VR requiring delay less than 5ms, V2X less than 20ms, etc.

This draft examines the E2E latency from architectural perspective, from studying how different OSI layers contribute to E2E latency, how different domains, which can be different operators' domains or administrative domains, contribute to E2E latency, to analyzing the gaps of recent technology advancement in reducing latency.

By studying the contributing factors to E2E latency from various angles, the draft identifies some gaps of recent technology advancement for E2E services traversing multiple domains and involving multiple layers. The discussion might touch upon multiple IETF areas.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on February 23, 2019.

Copyright Notice

Copyright (c) 2018 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the [Trust Legal Provisions](#) and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1.	Introduction.....	4
2.	Terminology.....	4
3.	AR/VR Use Case.....	5
4.	Contributing Factors to E2E Latency.....	5
5.	Application Layer Initiative in reducing E2E latency.....	6
	5.1. Content Placement mechanisms need visibility to Network.	6
6.	Transport Layer Initiatives in reducing Latency and gaps.....	7
	6.1. TCP Layer Latency Improvement Alone is not enough.....	7
	6.2. LTE Latency Impact on TCP Performance.....	8
	6.3. Low Latency via Multipath TCP Extension.....	8
7.	Network and Link Layer Initiatives in reducing E2E Latency...	9
8.	Radio Channel Quality Impact to flows with High QoS.....	10
9.	E2E Latency Contributed by multiple domains.....	10
10.	Conclusion.....	11
11.	Security Considerations.....	11
12.	IANA Considerations.....	11
13.	Acknowledgements.....	11
14.	References.....	11

14.1. Normative References.....	11
-------------------------------------------------	--------------------

14.2. Informative References.....	11
15. Appendix:.....	12
15.1. Example: multi-Segments Latency for services via Cellular Access.....	12
15.2. Latency contributed by multiple nodes.....	13
15.3. Latency through the Data Center that hosts S-GW & P-GW	14
Authors' Addresses.....	15

Internet-Draft E2E Over Internet Latency Taxonomy

1. Introduction

Ultra-Low Latency is a highly desired property for many types of services, such as 5G MTC (Machine Type Communication) requiring E2E connection for V2V to be less than 2ms, AR/VR requiring delay less than 5ms, V2X less than 20ms, etc.

This draft is to examine E2E latency from architectural perspective, from studying how different OSI layers contribute to E2E latency, how different domains, which can be different operators' domains or administrative domains, contribute to E2E latency, to analyzing the gaps of recent technology advancement in reducing latency.

The primary purpose of studying E2E Latency from architectural perspective is to help the IETF community identify potential work areas for reducing E2E latency of services over the Internet.

In recent years, the internet industry has been exploring technologies and innovations at all layers of the OSI stack to reduce latency. At the upper (application) layer, more contents are distributed to the edges closer to end points and more progress in Mobile Edge Computing (MEC) has been made. At the Transport layer, there are QUIC/L4S initiatives. At the network layer, there are IP/MPLS Hardened pipe ([RFC 7625](#)), latency optimized router design, and BBF's Broadband Assured Services (BAS). At the link layer, there are IETF DETNET, IEEE 802.1 TSN (Time Sensitive Networking), and Flex Ethernet (OIF).

By studying the contributing factors to E2E latency from various angles, the draft identifies some gaps of recent technology advancement for E2E services traversing multiple domains and involving multiple layers. The discussion might touch upon

multiple IETF areas.

2. Terminology

DA: Destination Address

DC: Data Center

E2E: End To End

GTP: GPRS Tunneling Protocol (GTP) is a group of IP-based communications protocols used to carry general packet

radio service (GPRS) within GSM, UMTS and LTE networks. In 3GPP architectures, GTP can be decomposed into separate protocols, GTP-C, GTP-U and GTP'. GTP-C is used for signaling. GTP-U is used for carrying user data.

LTE: Long Term Evolution

TS: Tenant System

VM: Virtual Machines

VN: Virtual Network

3. AR/VR Use Case

The E-2-E delays of AR/VR system come from delay of multiple systems:

- Tracking delay
- Application delay
- Rendering delay
- Display delay

For human beings not to feel dizzy viewing AR/VR images, the oculus delay should be less than 19.3ms, which includes display delay, computing delay, transport delay, and sensing delay. That means the "Network Delay" budget is only 5ms at the most.

4. Contributing Factors to E2E Latency

Internet data is packaged and transported in small pieces of data. The flow of these small pieces of data directly affects a user's internet experience. When data packets arrive in a smooth and timely manner, the user sees a continuous flow of data; if data packets arrive with large and variable delays between packets, the user's experience is degraded.

Key contributing factors to E2E latency:

- Generation: delay between physical event and availability of data

- Transmission: signal propagation, initial signal encoding
- Processing: Forwarding, encap/decap, NAT, encryption, authentication, compress, error coding, signal translation
- Multiplexing: Delays needed to support sharing; Shared channel acquisition, output queuing, connection establishment
- Grouping: Reduces frequency of control information and processing; Packetization, message aggregation

The 2013 ISOC Workshop [[Latency-ISOC](#)] on Internet Latency concluded that:

- o Bandwidth alone is not enough in reducing latency
- o Bufferbloat is one of the main causes for high latency in the Internet.

Figure 1 of the 2013 ISOC workshop report showed that the timing of download of an apparently uncluttered example Web page (ieeexplore.ieee.org), actually comprised of over one hundred

objects, transferred over 23 connections needing 10 different DNS look-ups. This phenomenon just further proves that reducing E2E latency will need multiple layers coordination and interaction.

5. Application Layer Initiative in reducing E2E latency

More and more End to End services over internet are from end users/devices to applications hosted in data centers.

As most content today is distributed, E2E services usually do not traverse the globe but rather more often than not, the network segments that the E2E service traverses are from end users to regional data centers. The practice of content distribution to the edge has transformed reaching low latency goals from fighting against the speed of light to optimizing communication between end users and their desired content.

However, without awareness of latency characteristics of network segments, the content distribution mechanisms & algorithms might not achieve their intended optimal result.

[5.1](#). Content Placement mechanisms need visibility to Network

To be added.

6. Transport Layer Initiatives in reducing Latency and gaps

IETF QUIC, L4S are some of the initiatives in reducing E2E latency at the Transport Layer.

IETF QUIC focus on the improvement from end points. It doesn't take into consideration of the network latency that the data packets traverse.

The IETF L4S uses AQM for network nodes to purposely drop packets or send indication to end points when their queues are above certain thresholds. The goal is for the end nodes to reduce transmission rate when intermediate nodes buffers are almost

full. It has following issues:

As network aggregates many flows from many different end points and most flows have variable data rate, an intermediate network node+port's buffer being almost full at one specific time doesn't mean that the same amount of traffic will traverse the same port a few microseconds later. If all end (source) points reduce transmission rate upon receiving the AQM indication (or experiencing packets drop), traffic through the network can be greatly reduced (i.e. leaving no queue in the buffer). Then all end points can increase their rate, causing traffic pattern oscillation and buffer congestion again.

6.1. TCP Layer Latency Improvement Alone is not enough

The following example shows why simply optimizing transport layer alone is not enough. More details can be found at <https://www.w3.org/Protocols/HTTP/Performance/Pipeline.html>.

Typical web pages today contain a HyperText Markup Language (HTML) document and many embedded images. Twenty or more embedded images are quite common. Each of these images is an independent object in the Web, retrieved (or validated for change) separately. The common behavior for a web client, therefore, is to fetch the base HTML document, and then immediately fetch the embedded objects, which are typically located on the same server.

The large number of embedded objects represents a change from the environment in which the Web transfer protocol, the Hypertext Transfer Protocol (HTTP), was designed. As a result, HTTP/1.0 handles multiple requests from the same server

inefficiently, creating a separate TCP connection for each object.

6.2. LTE Latency Impact on TCP Performance

HTTP/TCP is the dominating application and transport layer protocol suite used on the internet today. According to HTTP Archive (<http://httparchive.org/trends.php>), the typical size of

HTTP based transactions over the internet are in the range of a few 10's of Kbytes up to 1 Mbyte. In this size range, the TCP slow start period is a significant part of the total transport period of the packet stream.

During TCP slow start, TCP exponentially increases its congestion window, i.e. the number of segments it brings into flight, until it fully utilizes the throughput that LTE (Radio + EPC) can offer. The incremental increases are based on TCP ACKs which are received after one round trip delay in the LTE system. Thus, as it turns out, during TCP slow start the performance is latency limited in Radio Network (LTE). Hence, improved latency in LTE can improve the perceived data rate for TCP based data transactions, which in its turn reduces the time it takes to complete a data down-load or upload.

Despite rather small (in terms of milliseconds) improvements that can be achieved over the radio round trip time, the total increase in the perceived throughput and delay savings of downloading an item below 1MB is significant due to the additive effect of LTE latency improvements in the TCP slow start[LTE-Research].

6.3. Low Latency via Multipath TCP Extension

There are some research work on how to use multi-path TCP to reduce E2E latency, such as <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7510787>. The paper proposes an MPTCP extension that sends data redundantly over multiple paths in the network, which basically exchanges bandwidth for latency. The integration into the MPTCP protocol provides benefits such as transparent end-to-end connection establishment, multipath-enabled congestion control, and the prevention of head of line blocking. The research paper claims that their proposed Multipath TCP extension can halve the average round-trip time and reduce its standard deviation by a factor of 19 for a real world mobile scenario in a stressed environment.

Those kind of researchers should be invited to the "Reducing latency over Internet Deep-Dive" workshop or cross-area BOF (to be organized by IAB).

7. Network and Link Layer Initiatives in reducing E2E Latency

Several industry initiatives already exist for improving latency at the Link and Network layers:

- Link Layer: IEEE 802.1 TSN (Time Sensitive Networking), and Flex Ethernet (OIF).
- The network layer: IETF DETNET, IP/MPLS Hardened pipe ([RFC 7625](#)).

Gaps:

IEEE 802.1 TSN (Time Sensitive Networking) requires stringent synchronous timing among all the nodes, which is suitable for small scoped network, but not suitable for the internet because most routers/switches in the network don't support synchronous timing.

IP/MPLS hardened pipe can guarantee no congestion and no buffering on all nodes along the path, therefore, ensure the lowest latency along the path. The hardened pipe is ideal for flows with steady bandwidth requirement.

But for applications that don't have steady flow size, the hardened pipe requires reserving the peak rate dedicated channels, which, like TDM, will incur bandwidth waste when application traffic goes below peak rate.

Traffic Engineering is one of the most commonly used methods to reduce congestion at the network layer. However, it doesn't completely prevent transient congestion. Depending on the tunnel sizing, there could be momentary traffic bursts that exceed the tunnel size, thus causing congestion if there isn't adequate headroom on the trunk carrying the tunnel to absorb the burst. Or a link or node outage, that reroutes the tunnel onto a secondary path that becomes overloaded, could cause congestion.

8. Radio Channel Quality Impact to flows with High QoS.

QoS is one of the key methods employed by fixed IP network to reduce latency for some flows. However, in Radio network, if a UE's channel condition is poor, the eNB may schedule more frames to other UEs whose flow are marked with much lower QoS.

There are many studies showing how Radio quality negatively impact to the TCP performance.

It is beneficial to the whole industry if there is a workshop to get people or SDOs working on different layers of Internet service together to showcase their work or their pain points.

IESG can make much more informed decision on creating useful initiatives when the community is aware of other work and obstacles.

9. E2E Latency Contributed by multiple domains

All of the latency improvement initiatives in the link layer have been within a single domain, such as IETF DETNET, IEEE 802.1 TSN (Time Sensitive Networking), and Flex Ethernet (OIF). The network layer latency improvement, such as IP/MPLS Hardened pipe ([RFC 7625](#)) is also within a single domain.

But E2E services usually traverse more than one domain, which can be administrative domains or multiple operators' networks.

Yet today, there is no interface between domains to:

- Inquire about the latency characteristics or capabilities from another domain
- Negotiate or reserve latency capabilities from another domain.
- Have a standardized method to characterize latency

IETF/IAB is an ideal organization to tackle those issues because IETF has the expertise.

10. Conclusion

As end to end services traverse multiple types of network segments and domains, and involve multiple layers, more informed decision in each layer technological improvement is important.

- Need across domain coordination
- Need across layer coordination

11. Security Considerations

As the trend is going more encryption, it is getting more difficult for various network segments to detect applications sessions. Therefore, it is more important to create ways for better coordination among different layers, for improved latency, trouble shooting, restoration, etc.

12. IANA Considerations

This section gives IANA allocation and registry considerations.

13. Acknowledgements

Special thanks to Jari Arkko for encouraging writing this draft. And many thanks to Andy Malis, Jim Guichard, Spenser Dawkins, and Donald Eastlake for suggestions and comments to this draft.

14. References

14.1. Normative References

14.2. Informative References

[LTE-latency] <https://www.ericsson.com/research-blog/lte/lte-latency-improvement-gains/>

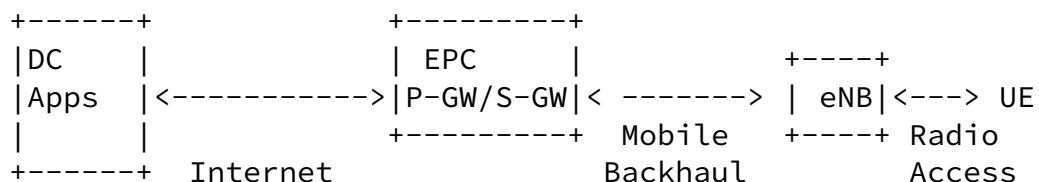
[Latency-ISOC] 2013 ISOC organized Latency over Internet workshop report

[15.](#) Appendix:[15.1.](#) Example: multi-Segments Latency for services via Cellular Access

Via Cellular network, there are User Plane Latency and Control Plane Latency. Control plane deals with signaling and control functions, while user plane deals with actual user data transmission.

The User Plane latency can be measured by the time it takes for a small IP packet to travel from the terminal through the network to the internet server, and back. The Control Plane latency is measured as the time required for the UE (User Equipment) to transit from idle state to active state.

User Plane latency is relevant for the performance of many applications. This document mainly focuses on the User Plane Latency. The following diagram depicts a logical path from an end user (smart phone) application to the application controller hosted in a data center via 4G Mobile network, which utilize the Evolved Packet Core (EPC).



Mobility Management Entity (MME) is responsible for authentication of the mobile device. MME retains location information for each user and then selects the Serving Gateway (S-GW) for a UE at the initial attach and at time of intra-LTE handover involving Core Network (CN) node relocation.

The Serving Gateway (S-GW) resides in the user plane where it forwards and routes packets to and from the eNodeB (eNB) and packet data network gateway (P-GW). The S-GW also serves as the local mobility anchor for inter-eNodeB handover and mobility between 3GPP networks.

P-GW (Packet Data Network Gateway) provides connectivity from the UE to external packet data networks by being the point of exit and entry of traffic for the UE. A UE may have simultaneous connectivity with more than one P-GW for accessing multiple Packet Data Networks. The P-GW performs policy enforcement,

packet filtering for each user, charging support, lawful interception and packet screening. Another key role of the P-GW is to act as the anchor for mobility between 3GPP and non-3GPP technologies such as WiMAX and 3GPP2 (CDMA 1X and EvDO).

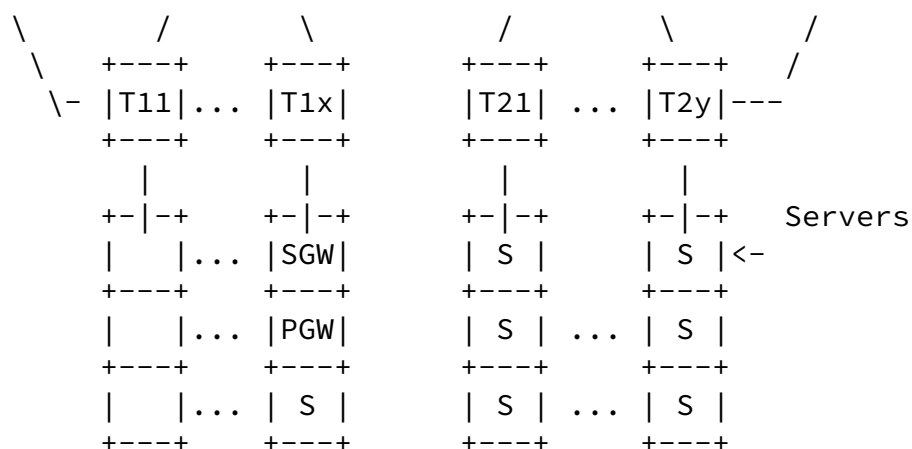
Very often P-GW and S-GW are co-located. The data traffic between eNB and S-GW is encapsulated by GTP-U.

The figure above shows that the end to end services from/to UE consists of the following network segments:

- Radio Access network - RAN
- Mobile Backhaul network that connect eNB to S-GW.
- Network within the DC that hosts S-GW & P-GW
- Packet Data Network, which can dedicated VPN, internet, or other data network.
- Network within the DC that hosts the App.

The RAN (Radio Access Network) is between UE (e.g. smart phone) and eNB. 3GPP has a group TSG RAN working on improving performance (including latency) of the Radio Access network. There are many factors impacting the latency through RAN.

The Mobile Backhaul Network connects eNBs to S-GW/P-GW, with data traffic being encapsulated in GTP protocol. The number of UEs that one eNB can handle are in 100s. The number of UEs that one S-GW/P-GW can handle are in millions. Therefore, the mobile backhaul network connects 10s of thousands of eNBs to S-GW/P-GW.



As the distance within data center can be small, the transmission delay within data center can be negligent. The majority of latency within data center is caused by the switching within the gateway routers, traffic traversing through middleware boxes such as FW, DPI, IPS, value added services, the top of the rack switches, and aggregation switches.

If the S-GW and P-GW are hosted in large data center, there could be latency contributed by the encapsulation/decapsulation such as work specified by NV03.

Authors' Addresses

Linda Dunbar
 Huawei Technologies
 5430 Legacy Drive, Suite #175
 Plano, TX 75024, USA
 Phone: (469) 277 5840
 Email: linda.dunbar@huawei.com

