Internet Draft
Intended status: Informational
Expires: April 2018

J. Xia
Huawei

October 30, 2017

Architectural Considerations for Delivering Latency Critical Communication over the Internet
draft-xia-latency-critical-communication-00.txt

Abstract

There is clear demand for Internet applications requiring critical low-latency and reliable communication - Latency Critical Communication (LCC).

This document is intended to stimulate LCC discussion and is not expected to be published as an RFC.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on April 30, 2018.

Copyright Notice

Table of Contents

1. Introduction

Latency Critical Communication (LCC) applications are increasingly important, requiring guaranteed low-latency communication high-reliability and ensuring quality of user experience.

Several on-going mechanisms exist for delivering LCC services within multiple Standards Development Organizations, including: Time-Sensitive Networking Task Group [TSN8021] in IEEE 802.1, 5G requirements for next-generation access technology [TS38913] in 3GPP and Broadband Assured IP Service [BAS-Architecture] in the BBF.

This draft identifies common service requirements in heterogeneous networks for delivering LCC services, and outlines specific uses across a spectrum of applications, specifically: cloud-based virtual reality, remote surgery, and live-TV distribution in virtualized CDN environments.

We may scope LCC application requirements by explicitly focusing on end-to-end (E2E) service characteristics and capability requirements for delivering each specific use case. Furthermore, as the E2E service usually traverses multiple domains and involves multiple layers. Yet, existing standards and current discussion typically focuses on a specific layer, protocol, or link layer technology. This focused view lacks an integrated approach or system view on solving the LCC problem space.

This document is intended to stimulate discussion and outlines specific LCC application requirements, and proposes an architecture and enabling functional components to address the common requirements discussed in each use case.

2. The Need for Low Latency Communications

Fundamentally, latency is a time interval between the stimulation and response, or, from a more general point of view, a time delay between the cause and the effect of change in the system being observed.

Network latency in packet networks is measured either one-way (the time from the source sending a packet to the destination receiving it), or round-trip delay time (the one-way latency from source to destination plus the one-way latency from the destination back to the source). Some packets will be dropped, i.e., never delivered, due to buffer overflows, synchronization failures, etc. Moreover, we assume that packets that are decoded in error are also dropped either by the protocol itself or by higher layers. Using the convention that dropped packets have infinite latency, we can define

the reliability as the probability that the latency does not exceed a pre-described value.

Our community has recognized low latency networking as an important research problem, and devoted much attention to tackle the issue from various perspectives, these include:

o Processing delays

o Buffer delays

o Transmission delays

o Packet loss

o Propagation delays

There are a number of common requirements across low latency use cases (including 3GPP on Cloud RAN, front haul, back haul and by various application layers use cases. Additional useful documents exist that provide background and motivation for low latency networks, including [I-D.arkko-arch-low-latency] and [I-D.dunbar-e2e-latency-arch-view-and-gaps].

3. Quantifying Latency

LCC Applications exist for a variety of deployments, use cases are assigned into the following categories:

o Extreme Mobile Broadband (xMBB): high speed and low latency mobile broadband;

o Ultra-reliable Machine-type Communication (uMTC): reliability is the key service requirement of these services;

o Massive Machine-Type Communication (mMTC) and Massive IoT (mIoT): massive M2M and IoT connectivity;

o Critical Connections/ Ultra Reliable Low Latency Connections (CriC/URLLC): low latency and ultra-reliable communications.

The focus of this document is to outline requirements for CriC/URLLC use cases, specifically:

o Cloud-based virtual reality;
o Remote surgery;
o Live-TV distribution in virtualized CDN environments.

3.1. Determinism

- Difference between predictable and reliable bounds.

- Behavior of packet flow, and loss, and/or packets allowed outside of the bounds.

3.2. Network KPIs

For each category of use case, specific KPIs are identified for clustering requirements:

Device density:
- High: 10000 devices per km2
- Medium: 1000 - 10000 devices per km2
- Low: < 1000 devices per km2

Mobility:
- No: static users
- Low: pedestrians (0-3 km/h)
- Medium: slow moving vehicles (3 - 50 km/h)
- High: fast moving vehicles, e.g. cars and trains (> 50 km/h)

Infrastructure:
- Limited: no infrastructure available or only macro cell coverage
- Medium density: Small number of small cells
- Highly available infrastructure: Big number of small cells available

Traffic type:
- Continuous
- Bursty
- Event driven
- Periodic
- All types

User data rate:
- Very high data rate: 1 Gbps
- High: 100 Mbps - 1 Gbps
- Medium: 50 - 100 Mbps
- Low: < 50 Mbps

Latency
- High: > 50 ms
- Medium: 10 - 50 ms
- Low: 1 - 10 ms

Reliability
- Low: < 95%

o Medium: 95 - 99%
o High: > 99%

Availability (related to coverage)
o Low: < 95%
o Medium: 95 - 99%
o High: > 99%

3.3. Service KQIs

Application requirements, can be modelled by user experience (QoE), and qualified by service KQI. From users' point of view, QoE is the overall performance of a system. It is a measure of E2E performance at the services level from the user perspective and an indication of how well the system meets the user's needs. There are many factors affecting QoE, such as user expectations, end-to-end system effects, etc. It is essential to establish a relation between user expectations and QoS, considered as the ability of the network to provide a service at a guaranteed performance level.

Network's performance can be evaluated with network KPIs such as delay, jitter, packet loss, etc. For URLLC services, existing KPIs are insufficient to forecast the service quality and reflect end-users' QoE. Hence, it is important to identify useful KPIs to quantify end-users' experiences and build the connections between network KPI and service KQI, as shown in Figure 1. The KQI for a given service can be expressed as a function/combination of the KPIs, and can be expressed as follow: KQI=f(KPI1, KPI2,...,KPIn).

3.4. Correlating KQI and KPI

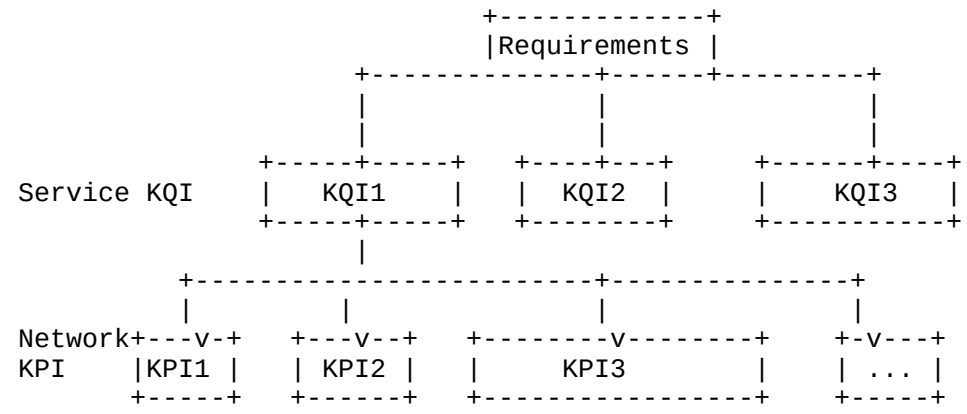


Figure 1: KQI-KPI Correlation

The emerging LCC application services have led to composite KQIs use, providing network measurement of specific application service aspects (i.e., the performance of the application or service). As there is limited experience to guide how to deliver the new LCC services, the mapping between the KPI and KQI will require specific

3.5. Application Use Cases

3.5.1. Cloud-based Virtual Reality

Virtual Reality (VR), also known as immersive multimedia or computer-simulated reality, is a computer technology that replicates an environment, real or imagined, and simulates a user's physical presence and environment to allow for user interaction.

Although some aspects of VR are becoming promising, there is still bottleneck that prevents it from being popular. High cost, especially for higher-end systems that try to reproduce a high-quality experience, is one barrier to success for VR. One way to reduce the cost of local VR computing, to make it more affordable and increase its popularity and general usage, is to offload the computations to a cloud-based server. This especially fits the cloud-based VR gaming environment when connecting with multiple parties.

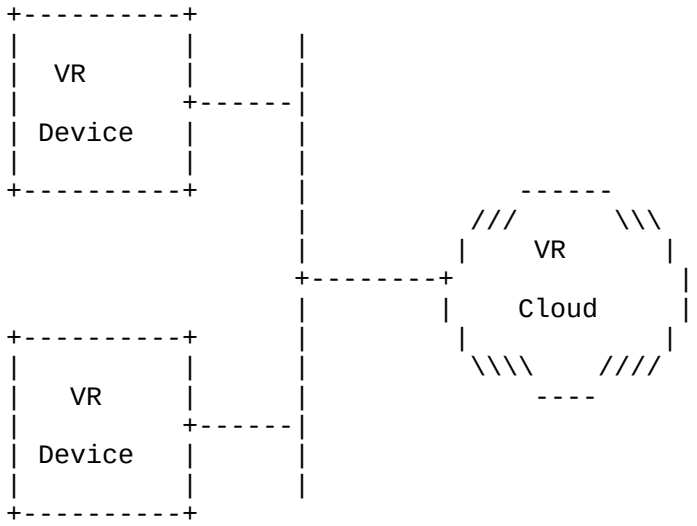


Figure 2: Cloud-based VR Scenario

But then, additional stringent requirements for the underlying network are being introduced into the system, including high bandwidth, low latency and low packet loss ratio.

3.5.1.1. Quality of Experience Requirements

To make the VR world realistic, the VR motion-to-photon latency is recommended to be less than 20ms. However, the network transmission latency is limited within 5ms because other VR processing (i.e., tracking, rendering, and displaying) latency almost consumes 15ms. To achieve this, the VR cloud is proposed to be deployed at the edge of operator network.

Regarding bandwidth requirements, high-end VR systems typically use a display frame rate of 75-90 frames per second on dual HD or 4K displays, which can result in traffic rates four to eight times that for regular HD or 4K video respectively. Of course, low packet loss is required to prevent video freezes or dropouts.

Name	Elements
Service type	CriC/URLLC
Bandwidth [Mb/s]	4K 25Mb/s
	8K 100 Mb/s
	12K 418 Mb/s
Bitrate(Mbps)	4K 16 Mbps
	8K 64 Mbps
	12K 279 Mbps
Latency	5 ms
Reliability	High (five 9)

Figure 3: Cloud VR Service Type

3.5.2. Remote Surgery

Remote surgery (also known as telesurgery) is the ability for a doctor to perform surgery on a patient even though they are not physically in the same location. It further includes the high-speed communication networks, connecting the surgical robot in one location to the surgeon console at another location manipulating the robot through an operation.

Remote telesurgery allows the specialized surgeons to be available to the patients worldwide, without the need for patients to travel beyond their local hospital. Imagine a doctor in an urban city, performing an urgent procedure on a patient in an inaccessible rural area.

3.5.2.1. Quality of Experience Requirements

In order to ensure a telesurgery procedure to be highly safe, a particularly unique demand is required on the network, at least including very reliable connection (99.999% availability), sufficient bandwidth to ensure adequate video resolution as required by the remote surgeon controlling the robot, as little as possible latency allowing the feel of instantaneous reaction to the actions of the surgeons and of course as little as possible latency variation (i.e., jitter) allowing system or human correction of the latency.

Name	Elements
Service type	CriC/URLLC
Bandwidth [Mb/s]	Up to 1Mb/s for control commands
Bitrate(Mbps)	8K 64 Mbps
Latency	30 ms
Reliability	High (five 9)

Figure 4: Remote Surgery Service Type

3.5.3. Live-TV Distribution in Virtualized CDN environments

Live-TV signal distribution is a growing service that a network operator needs to support. The bandwidth needed to convey a video stream is determined by its quality. Evolution from standard definition (SD) and high definition (HD) quality formats towards Ultra-High Definition (UHD) formats, including 2K and 4K UHD will have to be carried across an IP network, thus requiring the migration from traditional Serial Digital Interfaces (SDI) transmission to all-IP environments.

Various paradigms exist to provide cost-effective scalable live-TV distribution. Specifically, in live-TV distribution, uncompressed video stream formats are used before the video is produced. Once the video has been produced, distribution to end-users is based on compressed video streams, which quality is adapted to the one that fits better the user’s device (i.e., compressed SD, HD or UHD formats).

Content Delivery Networks (CDN) can be considered as a suitable option for live-TV content delivery by means of the standardized MPEG Dynamic Adaptive Streaming over HTTP (MPEG-DASH) technique.

3.5.3.1. Quality of Experience Requirements

Transport quality (packet loss, jitter) highly impacts on users' quality of experience (QoE). Undesired effects such as pixelization, tiling, frame freezing, or blue screen can appear if transport quality is slightly degraded.

Monitoring at different levels (network, computing, service) and applying local/global KDD procedures enable dynamic adaptive CDN reconfiguration, i.e. scaling up/down HTTP servers, reassigning users, increasing CDN links capacity, etc.

Name	Elements
Service type	CriC/URLLC
Bandwidth [Mb/s]	Low 1-4 Mb/s SD Med 10 Mb/s HD High 25 Mb/s UHD
Latency	High 50-100s ms
Jitter	Stringent <1 ms
Reliability	High (five 9)
Availability	Moderate (99%)
Mobility - UE Speed	Up to 50km/h
Area Traffic	Normal 1s Gb/s Hotspot 10s Gb/s
Sensor Network	No
Massive Type	No
Device Direct	No
Coverage Required	Standard
Energy Consumption Critical	No
Type of Use Equip.	Conventional

4. Measurement of Latency

Various Internet measurement methods have been proposed to identify latency between end hosts. Active network measurements, which involve sending a stream of measurement data traversed along arbitrary paths over a network, including the Internet, are amongst the more popular methods to provision end-to-end quality-of-service.

Accurate network measurement would require mesh measurement of all network links to collect sufficient network latency information for network path construction based on active measurement methods. It takes a longer time; thus, it may be possible for a small group of nodes but not for larger number of nodes. Inaccurate measurement over lossy network with long inter-packet delays would become an issue, and not support real-time applications that require time sensitive information for network path decisions.

In the [I-D.dunbar-e2e-latency-arch-view-and-gaps], several key latency factors are listed as below:

o Generation: delay between physical event and availability of data

o Transmission: signal propagation, initial signal encoding

o Processing: Forwarding, encoding/decoding, NAT, encryption, authentication, compress, error coding, signal translation

o Multiplexing: Delays needed to support sharing; Shared channel acquisition, output queuing, connection establishment

o Grouping: Reduces frequency of control information and processing; Packetization, message aggregation

From the network point of view, only the last four latency factors are highly relevant to the network characteristic and need to be measured.

The E2E performance has been focused on connection-less technologies, the requirements of measuring and maintaining "flow" state for end-user have gaps.

Measurement of network delay, performance guarantees, dynamic path adaption, and throughput optimization, all exist but are generally technology specific.

4.1. End-to-end Latency

A One-way Delay Metric for IPPM is defined for packets across Internet paths based on notions introduced in the IPPM framework with detailed introduction on the measurement methodology, error analysis and relevant statistics. The result can be used to indicate the performance of specific application and the congestion state on the path traversed.

IP Flow Information Export (IPFIX) Protocol serves as a means for transmitting Traffic Flow information over the network from an IPFIX Exporting Process to an IPFIX Collecting Process.

IPPM or IPFIX should be sufficient for the controller of distributed control plane to make the necessary optimization or bandwidth control, assuming IPFIX and IPPM can measure segment, interface, and chassis/fabric time. But if not, the extension of existing IPPM (metrics) may be needed.

In addition, other existing technologies, such as One-Way Active Measurement Protocol (OWAMP) and Two-Way Active Measurement Protocol (TWAMP), are focused on providing one way and two way IP performance metrics. Latency is one of metrics that can be used for End-to-End deterministic latency provisioning.

Using OWAMP/TWAMP protocols or extension on that to support measurement of flow latency performance is also feasible.

4.2. Per Link Latency

Latency related to link can be computed as the ratio between the link length and the propagation speed over the specific medium of the link.

The link capacities along the path as well as the way in which the available capacity is used can have impact on the latency. Whenever the link capacity is low, the time of getting data out of network card to onto the medium will be high. Furthermore, capacity is often shared and only a small proportion of the capacity may be available to a specific flow, this is the case when links are congested.

4.3. Per Node Latency

The links along a network path are connected by network nodes, such as core switches and routers. Transit through each node adds to the path latency, this type of latency is referred to as switching/forwarding latency.

To achieve optimized end-to-end low latency services, each network node along the path needs to measure the latency metric on it. Using OWAMP /TWAMP and/or extension on that, each network node needs to record accurate measurements and thus requires accurate time synchronization, which also contributes latency on the network node.

4.4. Reporting Per Link and Per Node Latency

Basically, the latency information needs to be reported from the network node to the controller or OAM handler [RFC7491] to keep the end-to-end latency under bound. A template that defines the LCC connection attributes, latency, loss and etc, must be used.

In addition, an interface or mechanism to report such latency performance information is necessary. A simple approach can be an interface from network device to controller, which collects all the latency performance information of each network node, and then make a decision how to serve the flow at each network node.

4.5. Isolating Latency Disruption

When congestion occurs, it is often not being detected until it has already induced latency. Early detection of the onset of congestion allows the controllers to reduce their transmission rate quickly. This could use delay based inference of congestion or early explicit notification of congestion by the network.

However, the congestion occurred link should be separated with other links and thus will not disrupt the other links. One feasible way is to reserve dedicated network resources to the specific link (for a specific application) and thus isolate the usage of the dedicated network resources from other links.

5. Mechanisms to achieve low latency flows

The network infrastructure will need advanced interaction with LLC applications. The network will need insight into which types of applications are being transported, and traffic classification and path control to ensure SLAs expected by the applications are met. Several techniques exist to achieve this, and are discussed in the following sub-sections.

5.1. Path Computation

The Path Computation Element (PCE) was developed to provide path computation services for path controlled networks. The may be used to provide path computation and policy enforcement for LCC applications and services.

The PCE operates on a view of the network topology stored in the Traffic Engineering Database (TED). The TED is a data store of topology information about network nodes and links, and capacity and metrics such as link performance (latency, latency-variation, and packet loss). The TED may be further augmented with status information about existing services as well.

The PCE would facilitate the setting up of LCC application paths by computing a path based on the end-to-end network performance criteria.

5.2. Traffic Engineering

Traffic engineering techniques, including Multiprotocol Label Switching (MPLS) Traffic Engineering (MPLS-TE), are commonly deployed by operators.

MPLS-TE allows for a TE scheme where the ingress node of a label switched path (LSP) can calculate the most efficient route (including latency minimization) through the network toward the egress router of the LSP.

The operator typically has a pre-planning task to monitor the physical layout of the network for capacity planning and network visualization followed by estimation of possible TE settings of the links, knowing how much an IGP setting affects the traffic flow and path. Modification of TE settings to reduce latency based on network conditions is possible, but introduces potential network instability if changes are frequent.

Overall, TE technologies come with limitations such as scalability, operational complexity, protocol overhead, and supervised network optimization. Although, recent enhancements to MPLS-TE exist, the interaction between applications and network infrastructure is still not sufficient for the LLC challenges.

5.3. Coloring

It is possible to build colored paths through the network with the colors representing low bandwidth, low delay, high cost, affinities. Application traffic can then be assigned to those paths based on traffic placement profile.

Link coloring could be used to classify specific low latency links for LLC applications, and assigned to a logical topology for the delay-sensitive application traffic.

MPLS-TE also supports this function, often described as administrative groups "colors" or "link colors". Furthermore, link coloring is supported in IP networks with the use of MT-aware IGPs.

5.4. Queue Management

Deploying queue management techniques, such as Active Queue Management (AQM), in the network may facilitate latency reduction, reduce network latency. It may be useful to distinguish between two related classes of algorithms: "queue management" versus "scheduling" algorithms.

o Queue management algorithms manage the length of packet queues by marking or dropping packets when necessary or appropriate

o Scheduling algorithms determine which packet to send next and are used primarily to manage the allocation of bandwidth among flows.

The two mechanisms are loosely related, they address different performance issues and operate on different timescales.

As interactive applications (e.g. voice over IP, real time video streaming and financial transactions) run in the Internet, high latency and latency variation degrade application performance.

Deploying intelligent queue management and scheduling schemes to control latency and latency variation, would provide desirable and predictable behavior to end-to-end connections for applications

5.5. Latency Management

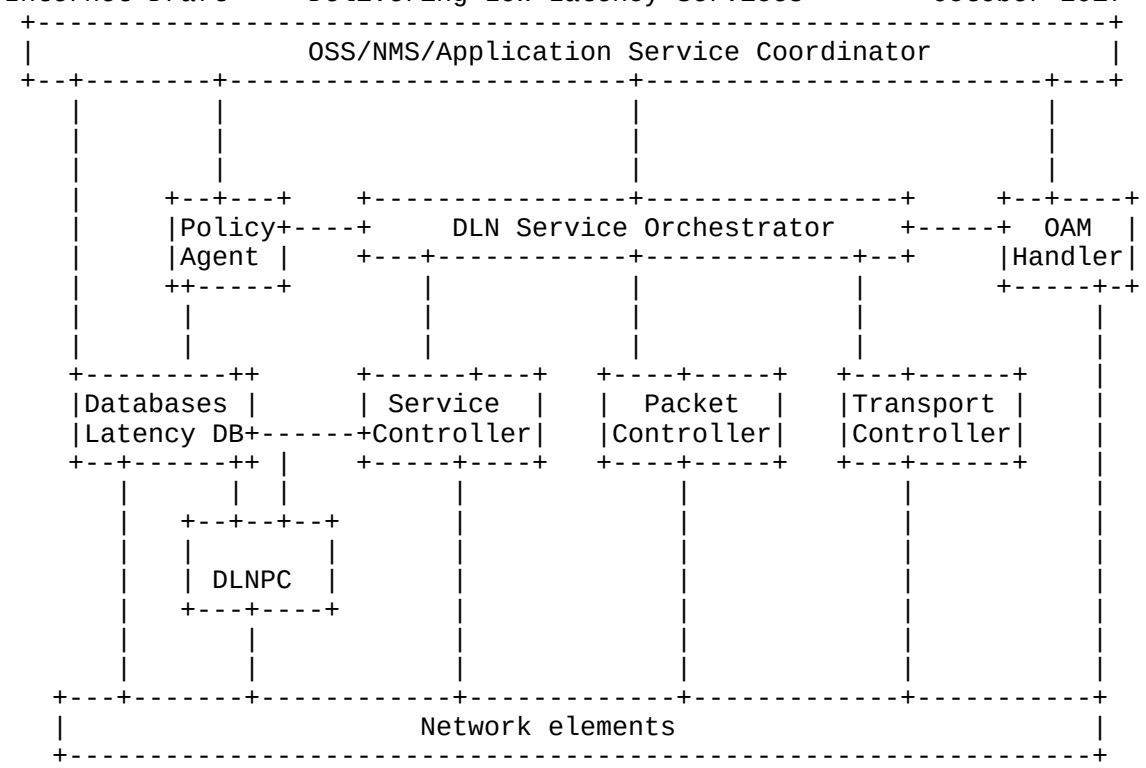
Latency management techniques include:

o XoDel (Controlled Delay) and FQ-CoDel (FlowQueue-CoDel) Controlled Delay (CoDel) are queue management technologies to set limits per packet for delay

o FlowQueue-CoDel (FQ-CoDel) is a hybrid packet scheduler/AQM algorithm for fighting application latency across the Internet. It is based on a two-tier Deficit Round Robin (DRR) queue scheduler, with the CoDel AQM algorithm operating on each sub-queue.

6. Functional Architecture for LCC

A basic architecture for LCC operation will be required. These will include the necessary components to manage the latency service, underlay packet and transport communications infrastructure.



6.1. LCC Functional Components

7. Alternatives to Low Latency Networking

7.1. Mitigation

Several strategies and techniques exist for reducing or negating network latency for some time sensitive applications.

7.2. Resource Placement

There is a trend of placing resources in locations that would reduce or negate service and application latency.

One approach to support more dynamic placement of functions, enabling the LLC application, close to the user is to introduce Virtualized Network Functions (NFV) at the edge of the network, near the LCC application users to reduce end-to-end latency, time-to-response, and unnecessary utilization of the core network infrastructure.

Specific technology threads developing edge resource placement, via virtual functions, include Mobile Edge Computing (MEC) and Fog Computing.

7.3. Application and Resource Pipelining

To be discussed.

7.4. Prediction

To be discussed.

7.5. Buffering

Reducing switch queue length, or buffer occupancy, is the most direct way to tackle the latency problem. Packet forwarders could use deep buffers to handle bursty traffic. However, they must ensure that this does not become detrimental to latency performance. As TCP relies on packet drops for congestion control, it introduces overhead for the application.

8. Security Considerations

The following list provides some security challenges and considerations in designing and building network infrastructure for LLC applications:

o Identification and authentication of the entities involved in the LLC service

o Access control to the different entities that need to be connected and capable of creating LLC services

o Processes and mechanisms to guarantee availability of LLC network resources and protect them against attack

8.1 Privacy and Regulatory Issues

o Identification of endpoints

o Data protection to guarantee the security (confidentiality, integrity, availability, authenticity) and privacy of the information carried by the network for the LCC service

9. IANA Considerations

10. References

10.1. Normative References

10.2. Informative References

[TSN8021] "Time-Sensitive Networking Task Group", IEEE (http://www.ieee802.org/1/pages/tsn.html).

[BAS-Architecture] Y.L. Jiang, "Broadband Assured IP Services Architecture", draft WT-387-00, broadband forum (BBF), July, 2016.

[TS38913] "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on Scenarios and Requirements for Next Generation Access Technologies; (Release 14)"

[I-D.arkko-arch-low-latency] J. Arkko, "Low Latency Applications and the Internet Architecture", draft-arkko-arch-low-latency (work in progress), 2017.

[I-D.dunbar-e2e-latency-arch-view-and-gaps] Dunbar, L., "Architectural View of E2E Latency and Gaps", draft-dunbar-e2e-latency-arch-view-and-gaps (work in progress), 2017.

[MEC_White_Paper] ETSI, "Mobile-Edge Computing - Introductory Technical White Paper", 2014.

[RFC7491] D. King and A. Farrel, "A PCE-Based Architecture for Application-Based Network Operations ", RFC 7491, March 2015, <http://www.rfc-editor.org/info/rfc7491>.

11. Acknowledgments

Authors' Addresses

Jinwei Xia
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: xiajinwei@huawei.com

Contributors

Ning Zong
Huawei Technologies

Email: zongning@huawei.com

Daniel King
Lancaster University

Email: d.king@lancaster.ac.uk