SCTE CABLE-TEC EXPO FAST FORWARD 2021
VIRTUAL EXPERIENCE
OCTOBER 11-14

2021 Fall Technical Forum
SCTE • NCTA • CABLELABS

Fastest Path to Low Latency Services

How Can Cable Operators Deliver Consistent Latency by Following an Efficient and Future-Proof Design Path?

A Technical Paper prepared for SCTE by

Sebnem Ozer, Ph.D.
Senior Principal Architect
Comcast
1800 Arch St., Philadelphia, PA 19103
2152868890
Sebnem_Ozer@comcast.com

Aaron Tunstall, Engineer 3, Enterprise Data & Analytics, TPX NGAN Access Eng/Comcast

Carl Klatsky, Principal II Engineer, Prodt Dev Engineer, TPX CPT NCE/Comcast

Dan Rice, VP, HFC Architecture, TPX NGAN Access Eng/Comcast

Jason Livingood, VP - Technology Policy & Standards, TPX CPT NCE/Comcast

John Chrostowski, Executive Director, NGAN Access Eng, TPX NGAN Access Eng/Comcast

John Raezer, VP, XCS Strategy, Planning, Connectivity & Consumer Experience Eng/Comcast

Joshua Gerson Sr. Mgr., XCS Strategy & Planning, XCS Strategy & Planning/Comcast

Mulbah Dolley, Eng 2, Technl Research & Dev, TPX NGAN Access Eng/Comcast

Priyan Sarathy Sr Mgr, Enterprise Data & Analytics, TPX NGAN Access Eng/Comcast

Sarulatha Subbaraj Engineer 4, Enterprise Data & Analytics, TPX NGAN Access Eng/Comcast

Soomin Cho, Data Engineer, TPX CPT NCE/Comcast

Trevor Graffa, Engineer 4, Software Dev & Engineering, TPX RDK/Comcast

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

List of Tables

Title **Page Number**

1. Introduction

The requirements of emerging interactive real-time services and changes in online usage patterns impose entirely different network challenges that Internet Service Providers need to overcome. Cable networks are going through a big transition to the next-generation 10G technologies with substantial speed increases, that can meet the online traffic volumes created by these services accumulatively. However, the new quality of experience judge will not praise or condemn the network operators only by their speed but also by their consistent support of low latency. Therefore, 10G technologies need to address a fundamental redesign of traffic classification and latency monitoring, prediction and optimization. Unprepared Internet Service Providers (ISPs) that design their architecture for mean and median values instead of peaks cannot support the interactivity of real-time services and cannot avoid the impact of these huge data volumes on other services.

In this paper, we will discuss the low latency services that are still evolving today, such as cloud gaming, video/voice conferencing and live video streaming, as well as emerging applications with progressively more interwoven human and machine interactions. We will first cover current network features and tools that can be used to measure, monitor and manage the latency of today's networks. We will then describe the next steps to support new Low Latency (LL) services by applying D3.1 features and a LL service differentiation framework. To support the LLD (Low Latency DOCSIS) features, traffic classification and monitoring must be redesigned and inherent rules may need to be replaced. Lastly, we will provide guidelines to deliver low latency services with the most efficient and future-proof investments.

2. Low Latency Services and Requirements

Not only do we experience continuous technological advancements and breakthroughs but we also observe faster democratization of technologies. Improved products and user experiences on interactive real-time services, IoT and sensor-based systems with big data learning, immersive applications, and autonomous systems altered the landscape of network traffic as consumers have easier access to these products and services. Mass production, digitization, software-defined, virtualized and cloud-based systems with open source software, platform models with partners and co-innovators have been key in the democratization of these technologies and building blocks of digital native companies. Legacy companies cannot survive if legacy chains are not broken for a digital transformation to meet the consumers' demands. Consumers are so immersed in the new technologies that they expect good quality, and nothing infuriates them more than if they don't work [1]. A technology that doesn't work for a consumer means bad Quality of Experience (QoE) [2].

Cable operators continue increasing connectivity speeds to improve the QoE of their subscribers through a series of new technologies and deployments, such as widening the upstream path via mid-split and high-split spectrum, and applying Full Duplex (FDX) architectures and distributed access networks. Speed as a performance metric has been regularly measured by MSOs, but speed is only one of the performance indicators. Different services and applications require different levels of Quality of Service (QoS) metrics, such as speed, latency, jitter, packet loss, reliability and security. Recently, many specifications and standardization documents from various networking technology organizations have been updated to include new traffic categories and QoS levels. Latency and latency variation (jitter) definitions and measurements are not as unified and standardized as speed (throughput) and packet loss. In the following sections, we will describe latency and jitter metrics that ISPs should monitor and improve for low latency services and thresholds, as defined in various standards and specifications.

2.1. Low Latency Services

For an efficient and future-proof design, cable operators must have a solid understanding of current and emerging services with reliable traffic forecasting. However, as we have seen during the pandemic, traffic forecasting and emerging services may not be always foreseeable. Therefore, it is crucial to design an agile system that can adapt to the changes faster and establish an accurate assessment of consumers' new QoE factors. Below we define the key points for current and emerging low latency services.

Real-time Gaming: Gaming lag means a delay between pressing a command button and the game responding on-screen. Network latency is one of the sources of gaming lag. Gamers are particularly sensitive to latency performance (a.k.a. jitter) when competing in multi-player online games, such as *League of Legends*, *Rocket League,* and *Fortnite*. A common complaint from gamers about latency is that the connection "lags out" during gameplay; even millisecond-level differences can make an impact for players competing in multiplayer online games. Latency variation (jitter) that lag-compensation algorithms cannot mitigate may reduce the gamer's QoE significantly. Additionally, the time difference of responses received by different multi-players causes unfairness [4]. Many game applications monitor and report measured latency and jitter, as shown in Figure 1, where players can compare the gaming performance to monitored latency.

Figure 1 – Ping Reports on the Gaming App

Cloud Gaming: Gamers send commands from a mobile device to cloud gaming platforms that execute those commands and then stream the results back to the gamer [6]. Network latency and jitter in the downstream affect the streaming and chat quality and while latency and jitter in the upstream affect the user input reception time and chat quality. Overall, high latency and jitter cause lag during play, choppy audio, poor video and distorted chat.

VideoConferencing: As more people have been working or studying from home, the quality of video and audio conferencing became essential in everyday life. High latency and jitter cause time lags, loss of lip synchronization and choppy or frozen video and audio [4]. Media and file sharing time can be also affected if the network quality is low.

Real-time interactive video streaming: While buffered streaming can be affected by high latency and jitter, the requirements for real-time interactive video streaming are stricter [3]. Services such as sports

betting, watch parties, shopping and synchronized second screen depend on an end-to-end platform with ultra-low latency and jitter. High latency and jitter can cause a subscriber to miss the hard cut-off times to place a bet before a game or a shopping window time. It can also cause spoiler issues during watch parties [8].

New services with web browsing are also sensitive to latencies on the order of hundreds of milliseconds [5]. Other emerging low latency services such as Holographic Type Communications, Multi-Sense Networks, Time Engineered Applications and Critical Infrastructure Services exposed several deficiencies in current network technologies that need to be addressed for future-proof deployments. [7]

2.1. Latency and Jitter Definition

Network latency (delay) is defined as the total time it takes for a data packet to travel between two networking points. One-way latency is the time required for a packet of data to travel from the sender to the receiver while Round trip time (RTT) is the time required for a packet of data to travel from the sender to the receiver and back again [5].

Network jitter or latency/delay variation refers to variation in the latency of arriving packets over time. Inter packet delay variation is the difference in latency of each received packet as compared to the previously received packet, while packet delay variation is the difference in latency of each received packet as compared to one reference value such as minimum or average latency [5].

The QoE depends on the end-to-end network latency and jitter while each network hop can be monitored and managed as discussed in [2].

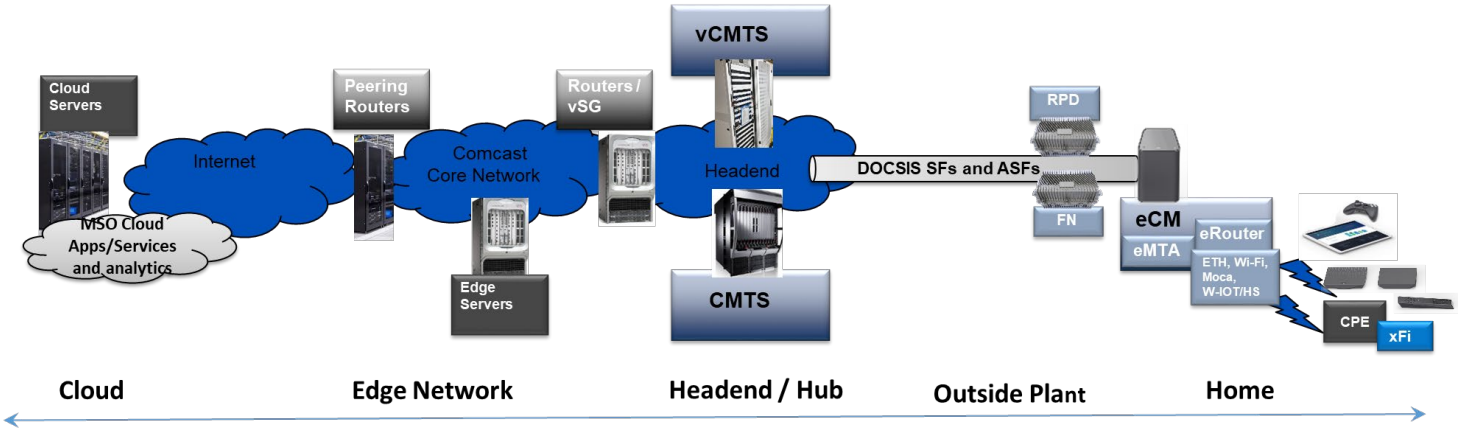


Figure 2 – Cable Network Segments

The low latency services in Section 2.1 require consistent latency, hence well-bounded jitter levels. It is important to assess the worst-case latency in the network while other latency and jitter measurements can help to analyze the latency sources and components. **Idle latency** measures responsiveness when a network connection is unused. It is mostly correlated to access network and distance of the path (round

trip time) e.g. fiber vs. Hybrid fiber-coaxial (HFC) vs. satellite. Many wireline network differences are insignificant. **Working latency (a.k.a. latency under load)** is the real-world measure of responsiveness when a network connection is actively used. Responsiveness of real-time applications during moderate usage of a network connection, whether upstream or downstream. When it gets really bad, it is often called "Buffer Bloat.", e.g. when gaming or video conference is interrupted by large file download or many devices in homes. The worst-case latency can be measured by the maximum allowed load (e.g. speed tier rate).

A common property of low latency services is that packets are useless when they are received with latency higher than an acceptable level. Therefore, they benefit from fast transmissions over shallow queues. This traffic type is called **non-queue building traffic (NQB)**. They do not benefit from increasingly consuming resources beyond need. They perform well in idle network conditions and good link quality, but with larger queues or dynamic movement between Idle Latency and Working Latency QoE can be variable without the right technology [9].

On the other hand, large down/uploads, buffered video streaming, speed tests, email, etc. rely on protocols that fairly use as much of the network capacity as possible to transfer the data at a high rate. This traffic type is called **queue building traffic (QB)**. The applications often open many TCP sessions in parallel and they are not latency sensitive. When these applications traffic is present on the network, latency is Working Latency. Small network queues lower latency but can make high speed QB traffic not hit peak rates without the right technology.

An overview of QB and NQB traffic is displayed in Figure 3 [9]. Real-time gaming control data and some audio/videoconferencing data are low-data-rate NQB traffic while cloud gaming streaming, real-time streaming and some videoconferencing applications are being implemented by developing new scalable congestion control algorithms (e.g. defined in [9]) to conform to high-date-rate NQB traffic type.

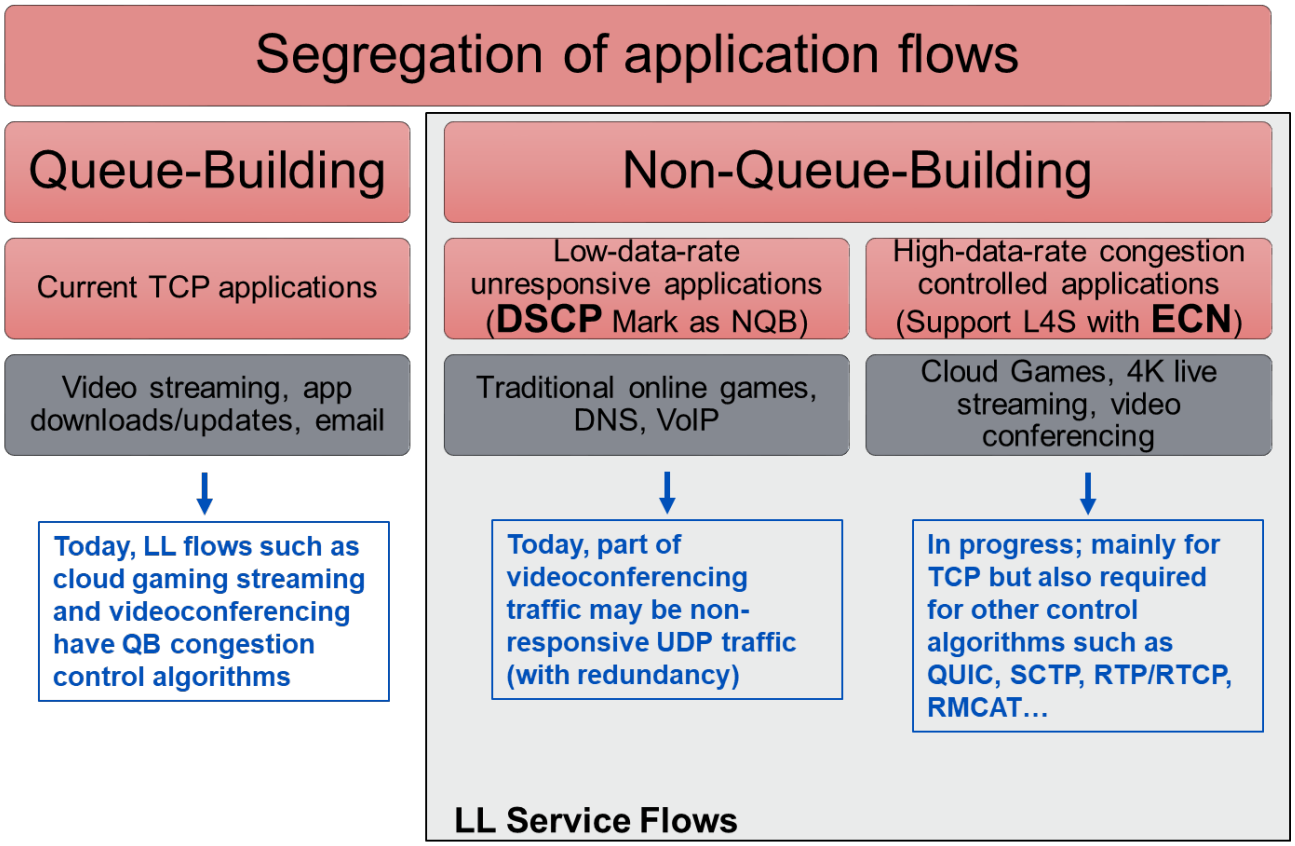


Figure 3 – QB vs NQB traffic

2.2. QoS requirements for Low Latency Services

DOCSIS 3.1 Specifications and Standards

DOCSIS 3.1 specifications include Low Latency Services support based on the coupled dual queue and proactive grant scheduling algorithms. The current target sets for applications such as real-time and cloud gaming and videoconferencing are <10ms RTT between the Cable Modem Termination System (CMTS) and Cable Modem (CM) for 99th percentile of packets. 1ms RTT can be achieved with proactive grant scheduling with the tradeoff of efficiency based on the currently available solutions. More details on the D3.1 low latency services support are provided in Section 4.1.

IEEE 802.11 Specifications and Standards

Time Sensitive Network support in 802.11 includes bounded 802.1Q traffic classification and stream reservation, low latency capabilities with 802.1Qbv over 802.11, scheduled operation with 802.11ax and 802.11be low latency channel access enhancements. Lower worst-case latency and jitter is a key feature for Wi-Fi 7 with multi-link and multi-AP operation and wider channels [10].

In [4], latency, jitter and packet loss requirements are proposed for several low latency services for the Wi-Fi networks (Table 1). Latency is defined as the RTT between the station (STA) and Access Point (AP), and jitter is defined as the standard deviation of latency. Since worst case latency is a key issue for these services, the definitions are based on the latency spikes that can also cause packet loss when certain

thresholds are exceeded, hence causing lagging and other issues. The document suggests new areas for further enhancement. Potential enhancements and new capabilities to address requirements of emerging real-time applications that can be grouped in the following categories:

Table 1 – Requirements Metrics of RTA Use Cases

Use cases		Intra BSS latency/ms	Jitter variance/ ms	Packet loss	Data rate/ Mbps
Real-time gaming		< 5	< 2	< 0.1 %	< 1
Cloud gaming		< 10	< 2	Near-lossless	< 0.1 (Reverse link) > 5 (Forward link)
Real-time video		< 3 ~ 10	< 1~ 2.5	Near-lossless	100 ~ 28,000
Robotics and industrial automation	Equipment control	< 1 ~ 10	< 0.2~2	Near-lossless	< 1
	Human safety	< 1~ 10	< 0.2 ~ 2	Near-lossless	< 1
	Haptic technology	<1~5	<0.2~2	Lossless	<1
	Drone control	<100	<10	Lossless	<1 >100 with video

3GPP Specifications

Technical specifications produced by the 3rd Generation Partnership Project (3GPP) and adopted by regional standards organizations use QoS class identifier (QCI) as a reference to a specific packet forwarding behavior (e.g. packet loss rate, packet delay budget) to be provided to a service data flow [3]. A subset of QCIs with a one-to-one mapping of standardized QCI values to standardized characteristics is shown in Table 2 for guaranteed and non-guaranteed bitrate resources (G/Non-GBR). A standardized QCI and corresponding characteristics are independent of the user's current access (3GPP or Non-3GPP). The characteristics describe the packet forwarding treatment that a service data flow aggregate receives edge-to-edge between the UE and the Policy and Charging Enforcement Function / Packet Data Network (PCEF/PDN) Gateway that is the interconnect point to the external network backbone.

Table 2 - A subset of 3GPP QoS Class Identifiers

QCI	Resource Type	Priority Level	Packet Delay Budget	Packet Error Loss Rate	Example Services
1	GBR	2	100 ms	10^{-2}	Conversational Voice
2		4	150 ms	10^{-3}	Conversational Video (Live Streaming)
3		3	50 ms	10^{-3}	Real Time Gaming, V2X messages Electricity distribution - medium voltage Process automation - monitoring
4		5	300 ms	10^{-6}	Non-Conversational Video (Buffered Streaming)
67		1.5	100 ms	10^{-3}	Mission Critical Video user plane
5	Non-GBR	1	100 ms	10^{-6}	IMS Signalling
6		6	300 ms	10^{-6}	Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)
7		7	100 ms	10^{-3}	Voice, Video (Live Streaming) Interactive Gaming
8		8	300 ms	10^{-6}	Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)

For instance, real-time gaming can be supported optimally if end-to-end latency is <60ms as working latency while <100ms latency can provide a good QoE [5, 6]. If Wi-Fi and DOCSIS networks can provide 15-20ms working latency as described above, 45-85ms latency can be allocated for the network segments outside of the ISP domain, encoding/decoding, rendering and/or cloud computing.

3. Current Latency Measurement, Monitoring and Management in the DOCSIS Networks

As discussed in a technical paper published at the 2020 SCTE Expo, and written by many of the same authors as this year's contribution [2], current DOCSIS features such as buffer control, Active Queue Management (AQM) along with better scheduling and efficiency implementations can help to reduce working latency and jitter for all High Speed Data services. Additionally, increasing speed tier rates through upcoming Mid-Split, High-Split and FDX technologies will help to improve the QoS support. However, these advancements alone are not adequate to provide bounded low jitter aimed for low latency services. QB applications can increase their traffic rates as speed tier rates increase and still cause high latency and jitter for NQBQ traffic. On the other hand, most NQB services do not require high speed, and network cost and operations can benefit from latency and jitter improvements to provide only the required

resources to each traffic type. To achieve better QoE for all services while creating an efficient and cost-effective network resource management, Cable Operators must have accurate and manageable latency measurement and monitoring tools.

We have been deploying with optimal configurations both D3.1 US AQM and DS AQM in our networks. The initial results have been captured in our 2020 latency paper [2]. Our internal measurement techniques for working latency (LUL) are similar to Samknows LUL data that is available as part of the FCC's Measuring Broadband (MBA) database [13]. We analyzed DS LUL results using FCC MBA data before and after our optimization. As displayed in Figure 4, DS working latency/LUL has been improved after optimal AQM deployments, which we were able to monitor and manage using our own platform.

As we extended our latency management platform, we optimized some of our methods and tools as we learned more with each deployment. In the following sections, we describe several parts of our platform by emphasizing key points that may be useful for other operators considering their latency strategies.

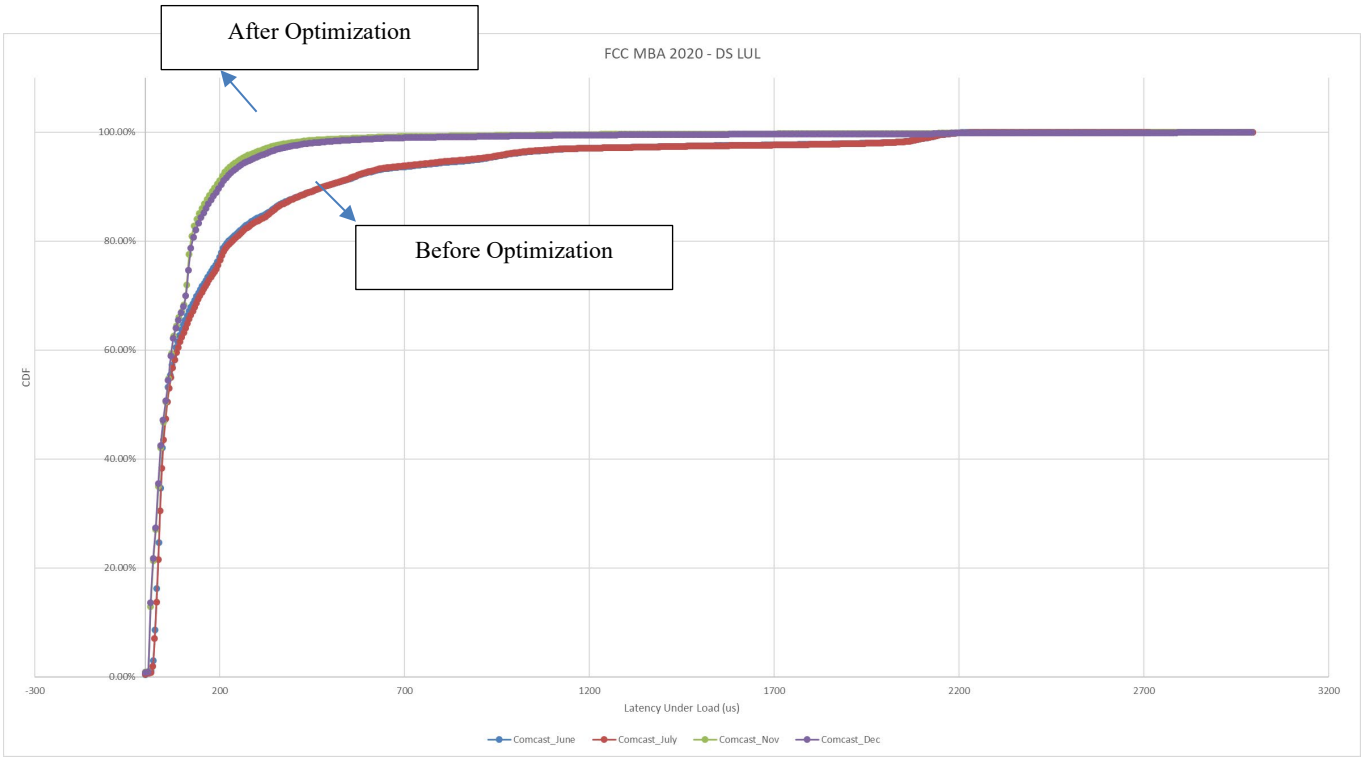


Figure 4 – FCC MBA 2020- Comcast LUL Results

3.1. Latency Measurement

As network operators focus more on the latency portion of the Quality of Experience, the challenge becomes how to measure the latency mitigation techniques being deployed on the network. Historically, the Internet Control Message Protocol (ICMP, RFC 792) has been used to perform a network layer latency check between two network endpoints. The limitation of ICMP is that it is a network layer check. Ideally, the latency check would be included as part of the application layer. Since that layer is predominantly outside the scope of the network operator, an alternate approach is to conduct the latency measurement at the transport layer. Further, most ICMP latency measurements are conducted in a

standalone instance, whereas the typical latency that impacts the customer QoE occurs while other users are accessing the network concurrently. Idle latency through ICMP pings can provide limited information.

Comcast has developed the Internet Measurement Platform (IMP) which provides a platform for measuring throughput and latency concurrently. Adapting the model used in other open source network measurement tools like Flent (https://flent.org), Comcast's IMP conducts both an "idle" latency measurement, meaning no other concurrent traffic from the test user, and a "latency under load (LUL)" measurement, meaning a latency measurement at the same time a throughput measurement is conducted, where the throughput measurement is trying to maximize its data consumption. Comparing the idle latency vs working latency (latency under load) measurements enables a clearer picture of the effectiveness of a deployed latency mitigation technique.

This new platform is implemented with an embedded agent on cable modems based on the Reference Design Kit/Broadband. The IMP agent interacts with the IMP control servers to process test requests using the specific IMP data plane test servers. The results are reported back from the client & server. By launching the tests from within the modem itself, the network operator is able to measure as closely as possible to the in-home devices that utilize the Internet service.

The idle latency portion of the measurement uses an HTTP CURL request / response, which uses TCP as its transport protocol. The latency under load portion of the measurement uses Netperf's request / response test, which uses UDP as its transport protocol. The throughput portion of the measurement uses the Iperf3 open source measurement tool, which uses TCP as its transport protocol. Both measurements are run concurrently. The TCP based data transfer will attempt to maximize its throughput up to the available provisioned capacity, potentially filling up node buffers along the network path while the UDP based request / response will attempt to complete its transaction competing against the load from the throughput measurement. In this fashion, the test is simulating the user's in-home experience where one user may be conducting a bulk data transfer (e.g. large file photo downloads) while another user is trying to complete quick request / responses (e.g. real-time gaming).

The latency reports can include min, mean, max, 50%, 75%, 95% and 99% and standard deviation values. Packet loss can be estimated by monitoring the successful transactions. Methods such as iRTT can be extended for more accurate packet loss and latency values, depending on an efficient implementation that can be supported by limited resources in the gateways.

The IMP platform has been audited by NetForecast, a 3rd party who independently reviewed the test results generated by the platform https://www.netforecast.com/netforecast-design-audit-report-of-comcasts-network-performance-measurement-system/

The IMP platform is currently in use on Comcast's production network, providing a comprehensive data set of latency measurements. Future enhancements to the platform include:

- Measuring the impact of different TCP congestion control algorithms
- As QUIC protocols are widely used, implementing UDP based data loading for speed test and working latency measurements
- Marking test data to measure latency for different services, such as low latency HSD flows.
- Exploring various control protocols to standardize test requests & results reporting

In addition to ping and working latency/LUL, there are other latency options that can be explored by the Cable Operators [5] such as actual customer traffic latency by monitoring TCP connections [2], monitoring queuing latency at DOCSIS with new D3.1 latency histogram recordings and Wi-Fi queuing metrics and two-way active measurement techniques.

The Two-Way Active Measurement Protocol (TWAMP), specified by IETF RFC 5357, provides a common protocol for measuring two-way or round-trip measurement between network devices. Today many routers used in Cable Operators' core network and CIN have TWAMP measurements capabilities. These capabilities may be extended to cover DOCSIS and other access networks and home networks as well [5]. The extensions can enable end-to-end latency profiling for QoE assessment of low latency services. A simplified version called Simple Two-way Active Measurement Protocol (STAMP) by IETF RFC 8762 can be used for one-way and round-trip latency, jitter and packet loss metrics.

As we measure the speed and latency for higher speeds, we see the shortcomings of certain measurement techniques. For example, when we measure symmetric 1Gbps DOCSIS network speed and latency for High-Split architectures, current TCP algorithms are not adequate. One key issue is that optimal test parameters that measure the speed tier vs working latency are not the same, depending on the protocol used for data load. Concurrent TCP flows may fill up the pipe but not the available buffer depending on the test parameters. When the goal is to measure the speed and working latency accurately and during a short time interval, other measurement options such as using UDP may be more efficient [11]. Cable operators can try these techniques by using Odroids connected to HS CMs as open source implementations (e.g. https://github.com/BroadbandForum/obudpst) are available before integrating them to their gateways for automated and stand-alone measurement capabilities.

3.1. Latency Visualization and Dashboarding

Visualization has always been a challenge for data. Of course, there are stated rules to follow also known as the graphic continuum. The challenge comes in telling a specific story that the data outlines in a readable form for the audience. Many can often make a scatter plot, bar chart, line chart etc., but readability becomes the challenge. Does your audience understand the story you portray?

For example, Figure 5 and Figure 6 can be used effectively to explain the difference between idle and working latency, to assess the best CM model and configuration for certain services. It is clear from the latency visualizations that, although idle latency performance does not vary significantly among the CM models, working latency is very different depending on the model and configuration. These visualizations can be used to compare the measured latency levels with the requirements given in Section 2.2.

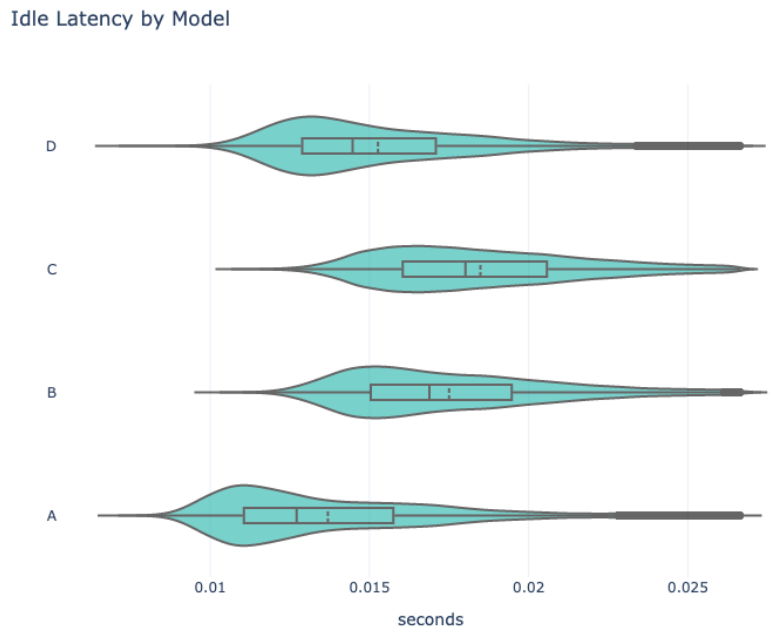


Figure 5 – US Idle Latency

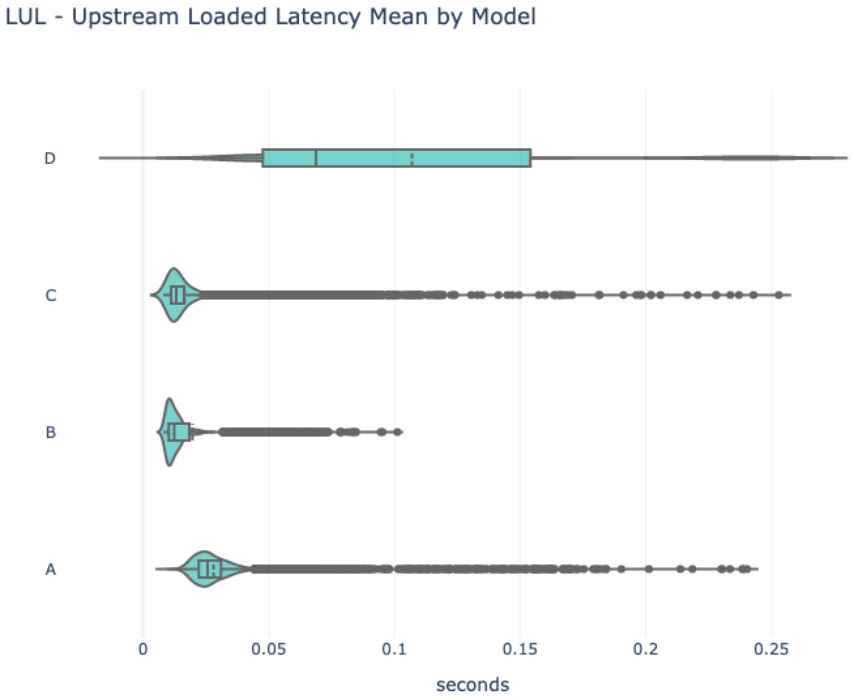


Figure 6 – US Working Latency (LUL)

Latency measurements over time as shown in Figure 7 can be used to monitor the latency improvements of a certain model that may have new FW or configurations, or of a new model that is deployed recently. Figure 8 is another example for DS working latency monitoring over time where new configurations are deployed to achieve a common improved latency range for all models.

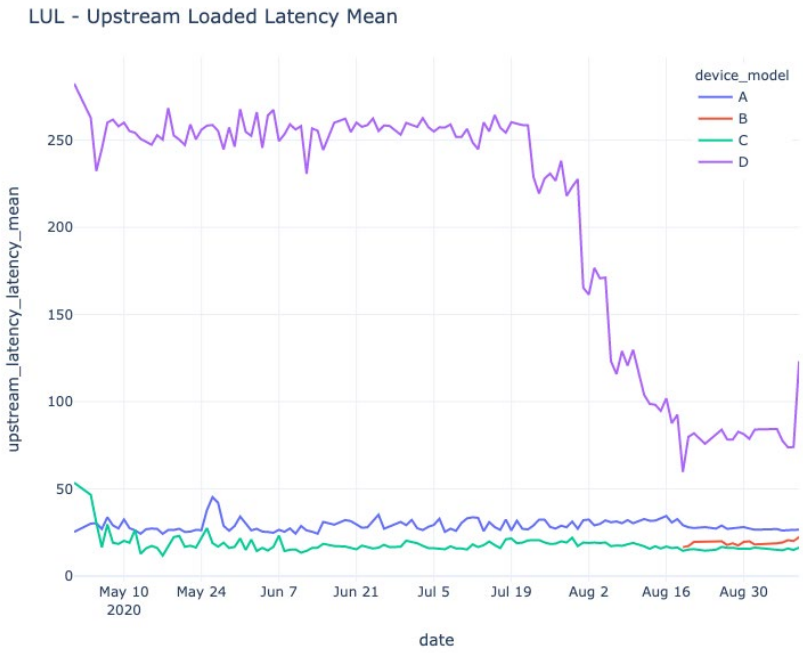


Figure 7 – US Working Latency (LUL) over time

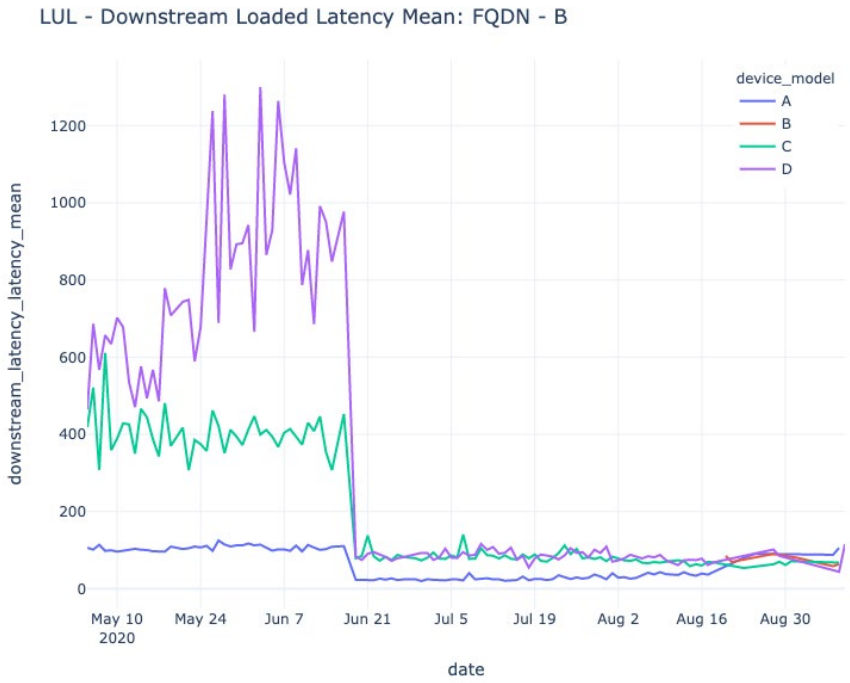


Figure 8 – DS Working Latency (LUL) over time

These initial visualizations are then used to create dashboards for continuous latency monitoring and management.

3.1.1. Current dashboards and data analysis

Figure 9 is an "executive" view of the latency data. Understanding your audience is what helped to shape the data into this view. The dashboard went through many iterations before becoming the high-level view shown above. The data shown is an aggregation of millions of rows of data that are processed daily. Breaking down the data with simplified thresholds allows an easy look into the performance. This also allows the end user to see spikes, anomalies, and trends from the data. This does not allow for a deep dive into the data and isolation of specific problems.

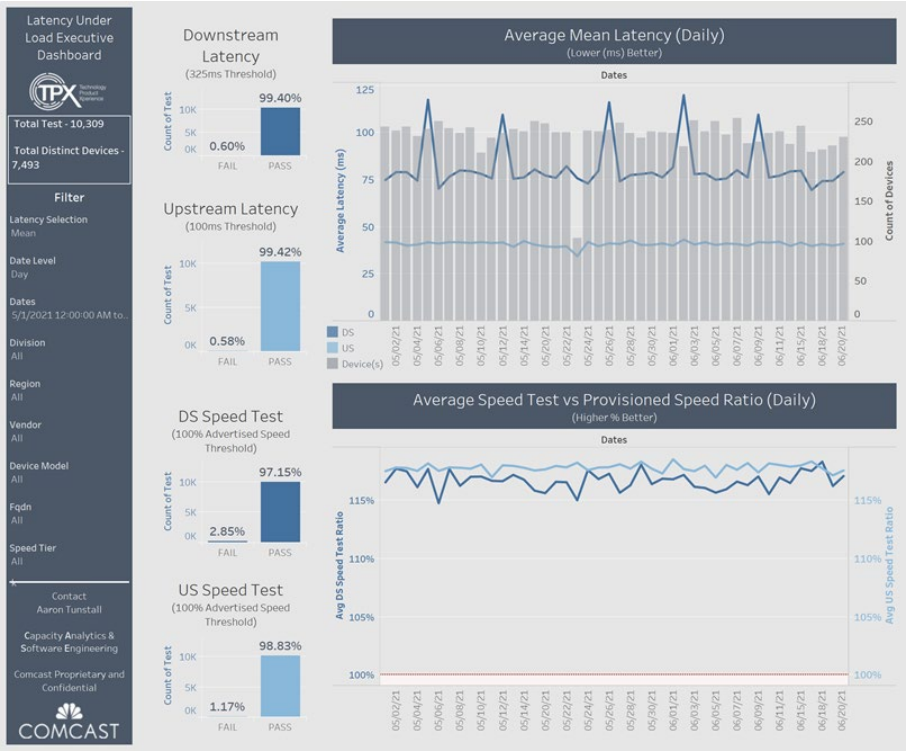


Figure 9 – Executive Dashboard for Latency & Speed test

To enable a more granular view of the performance we created a detailed version of the dashboard (shown in Figure 10). This version was the main source of truth for identifying latency performance and outliers. Each portion of the dashboard is interactive and allows the user a plethora of filters and customizations. Most users will never use all the functionalities included in this view, hence the "executive" view is the main view given to customers when they first load the dashboard. The main issue with data, especially latency, is that you have to balance what is readily available so all customers' needs are met and a proper analysis can be obtained.

Figure 10 – Latency Detailed Dashboard

3.2. Latency Management

3.2.1. Monitoring Latency

It is important to have a smart Business Intelligence (BI) tool when monitoring data. It must be dynamic enough to allow easy understanding and manipulation of the data. For latency we decided to use Tableau dashboards as they allow easy filtering and customization of views. The customer can easily click to change, aggregate or highlight specific data to find outliers and information.

3.2.2. Joint Analysis of latency and speed tests

Once latency data is processed and network configurations are set, the next analysis needed is to see if there are direct correlations between latency and network congestion (utilization). Correlation can be difficult when the measurements of latency can't be directly aligned with metrics that measure network congestion. This makes finding correlations difficult.

The easier method comes with comparing high latency to speed test results. When we compared "failed" speed tests (tests that don't achieve 100% of advertised speed), we found little correlation to high latency. The main reason is that latency is affected by small bursts of delay or packet loss that can be very challenging to detect in a granular way. When the utilization measurements are averaged between long measurement intervals compared to required granularity, the correlation cannot be seen. Therefore, we started a new trial where more granular utilization measurements are implemented to correlate their impact on the speed and latency tests.

3.2.3. Challenges and Guidance

When given the task to analyze data, the main challenge is balancing what is needed and what is not. It is an even harder challenge when you have data and no clear understanding on thresholds. What makes

latency good or bad? What is the expected customer experience? How can we isolate issues and outliers? Initially, when data was received, it was a small sample size and the dashboard was created with that in mind (Figure 11).

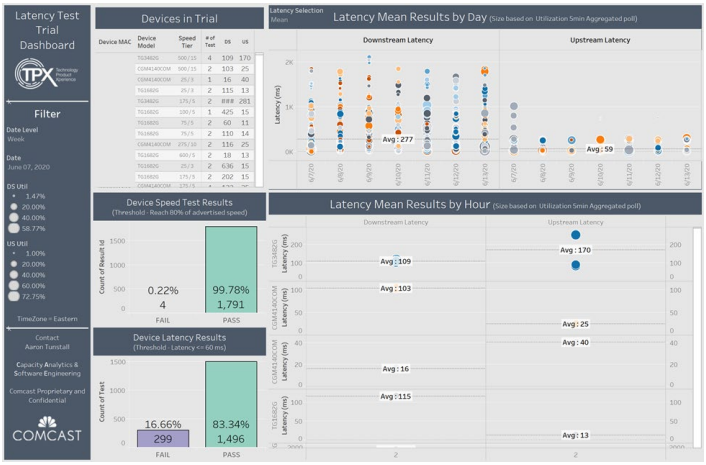


Figure 11 – Original dashboard for latency trial

At this point, there was no real threshold, just a small sample of test device data. We had an understanding that high latency is bad and lower latency is good. As the trial began and the scale of devices increased, the first version of the dashboard had to evolve. We created a detailed dashboard (Figure 10). This allowed a deep dive to analyze latency in multiple aspects. The dashboard showed how CMTS, device model, device firmware, vendor, and speed tier performed. This allowed us to isolate differences between all of the above, and how latency varied for each. The functionality made outlier identification a lot easier. This dashboard served its purpose, but when reporting findings it was not dynamic enough for presentation.

Newer temporary views were created, each to serve a different purpose for our Operations and Testing teams. These views were high-level comparisons of vendors, device types, and more. One of the lessons learned was to understand what your customer needs. This can be difficult when those needs are truly not yet understood. As time progressed, data became more readable, and thresholds started to be defined the needs of the customer become easier to define. What we learned allowed us to create the Executive dashboard (Figure 9) as well as easier views that tell the story the data shows. This allows a high level understanding of the overall customer experience.

In conclusion, data can be difficult to decipher when there is no clear understanding of how to read it. Understanding simple things, like high latency is bad and low latency is good, does not assist with telling the data's story. Sometimes it takes multiple views and analysis to find the needle in the haystack. As a data engineer, you have to understand that your initial idea or understanding may not be the final. Lastly, it is important to transform your data. As the analysis progressed the data sample grew, causing performance issues for customers viewing it. Summarizing data through aggregation is very important to giving the best customer experience when reviewing the data.

Another challenge is the storage, maintenance and query of the available data. We started getting LUL measurements from Elastic search production. The data was downloaded in Json format using Python script and stored it into the SQL server. The real-time data is visualized using Kibana - the data visualization dashboard software (open search dashboard) for Elasticsearch. The data download frequency was daily basis for the previous day. Due to ELK cluster workload maintenance, the LLD data is moved

to our internal streaming data platform, and Kinesis streaming for download [12]. The raw data download frequency is close to real-time. Scalability issues can arise later, as more data is collected and more queries are done by different teams. Therefore, the design must be flexible and extensible.

These large-scale measurement comparisons should provide additional data to justify the future deployment of AQM by ISPs and customer premise equipment manufacturers. This may be of interest to people working on the Low Latency, Low Loss Scalable Throughput (L4S) protocol or other TCP/UDP congestion controls. We believe, there is an opportunity to better standardize/define how working latency is measured. There is a need for open global internet measurement platforms to focus on working latency (or create new platforms & beyond access network segments) and/or sharing of such measurement data.

4. New Low Latency DOCSIS Features and Latency Management

As we discussed briefly in Section 1, broadband has historically differentiated based on download speed, however, customers are increasingly interested in additional characteristics when shopping for broadband. These emerging differentiators include faster upstream speeds, whole home WiFi coverage, and low latency.

Gamers are a large market and for serious gamers, latency influences ISP switching behavior and product selection. Gamers are also heavy streamers and in general, they tend to buy the best product that can support low latency services and high speed streaming. Beyond gamers, customers are generally interested in the concept of "no-lag broadband," suggesting that low latency DOCSIS could resonate with a larger audience. Specifically, broadband users that are employed and/or working from home have higher interest compared to other non-gamers, implying that the employed/work-from-home segment is another group to potentially target.

Enabling low latency DOCSIS would give cable providers an advantage over other ISPs that cannot compete with such consistently low levels of latency under load. Given the interest level in low latency among certain segments, go-to-market approaches could include incorporating low latency features across all internet households (where available), incorporating low latency features in premium speed tiers, or upselling low latency features within a separate premium data product.

To support these business cases, a new architecture that can differentiate low latency services and provide required QoS requirements must be defined and implemented. Although latency and jitter improvements described in Section 3 improve the overall QoE for several services, they are not adequate for current and emerging real-time interactive services.

Low latency will enable the creation of major new classes of applications where delay to local storage is equivalent to delay to a network-based resource. The cable industry is poised to take some leaps ahead of the competition as in the post-gigabit future, latency will be equally important to speed in marketing.

If we apply the same network resources with the same functionalities to each traffic with different QoS requirements (Section 2.2), then we provide equal resources but the outcome, which is the quality of experience, will be very different for each traffic type and unfair. AQM provides equitable performance because some traffic is managed (Figure 12 explains the difference between equality and equity nicely). Both large QB flows and smaller NQB flows perform better. LLD with L4s and dual queue promise even better. Early results show that we can hit the LLD target for working latency. As well, equity increases for all traffic because they are not having to share the same queue. This also helps to make sure that no app/service is harmed while, for example, providing different AQMs to low latency traffic groups because the equal quality of experience score is targeted.

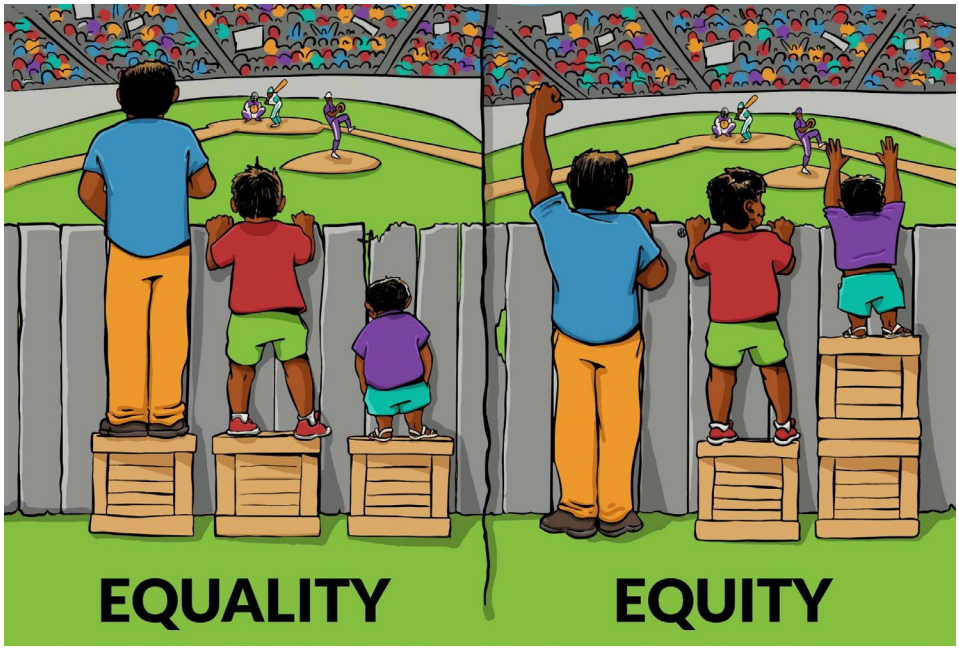


Figure 12 – Equality vs Equity

4.1. D3.1 LLD Features

New D3.1 LLD Dual Queue Features promise <10ms RTT between the CM and CMTS for 99th percentile of LL service packets. The main functionalities are described in Figure 13and Figure 14 [14].

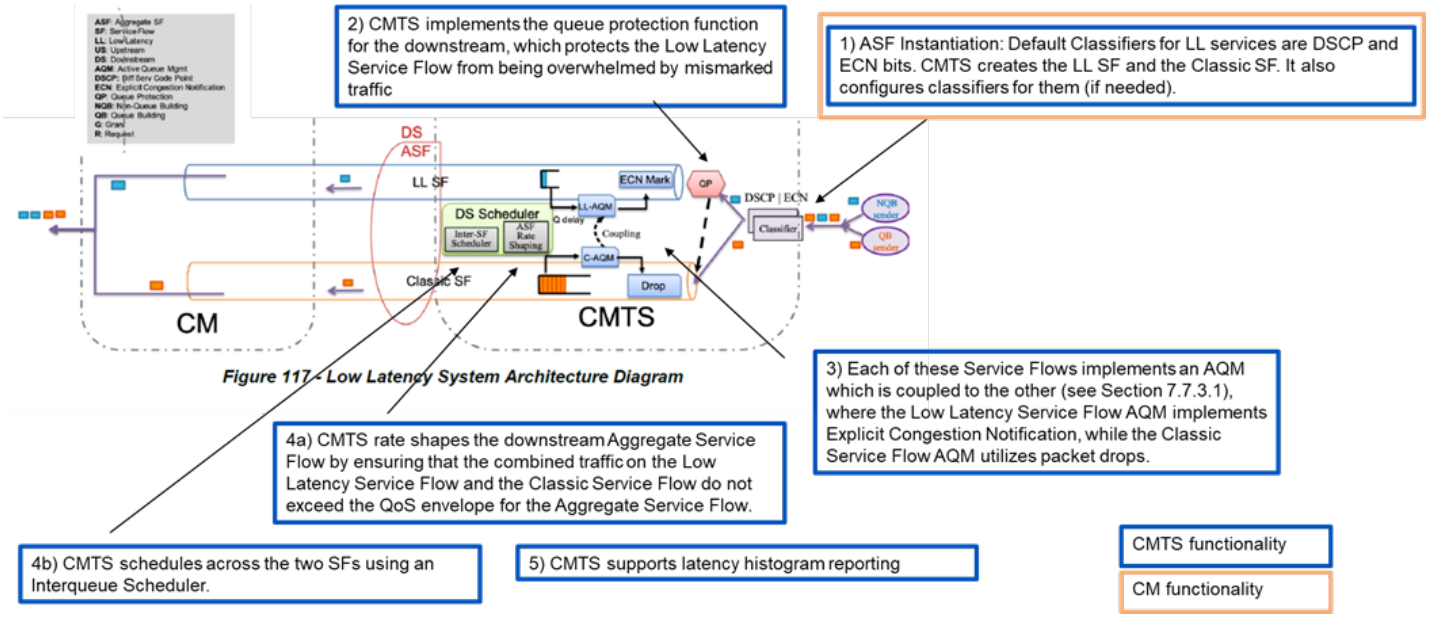


Figure 13 – D3.1 DS LLD Features

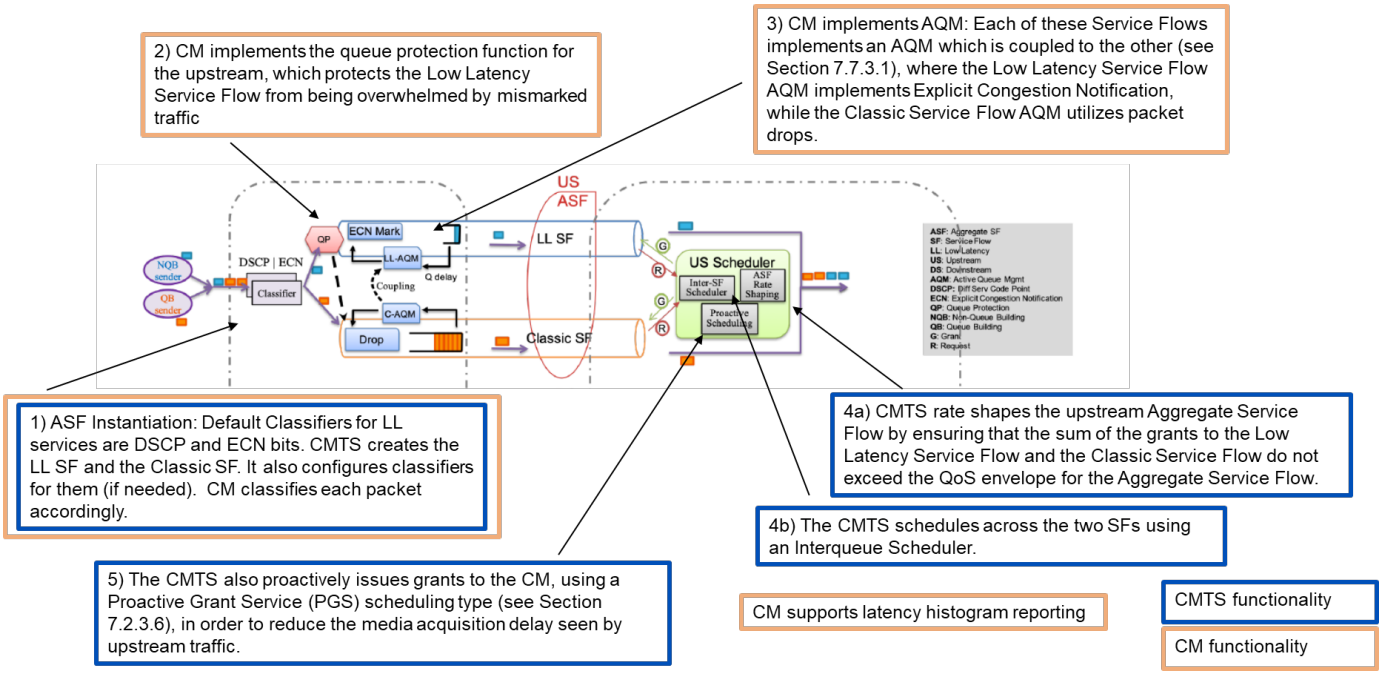


Figure 14 – D3.1 US LLD Features

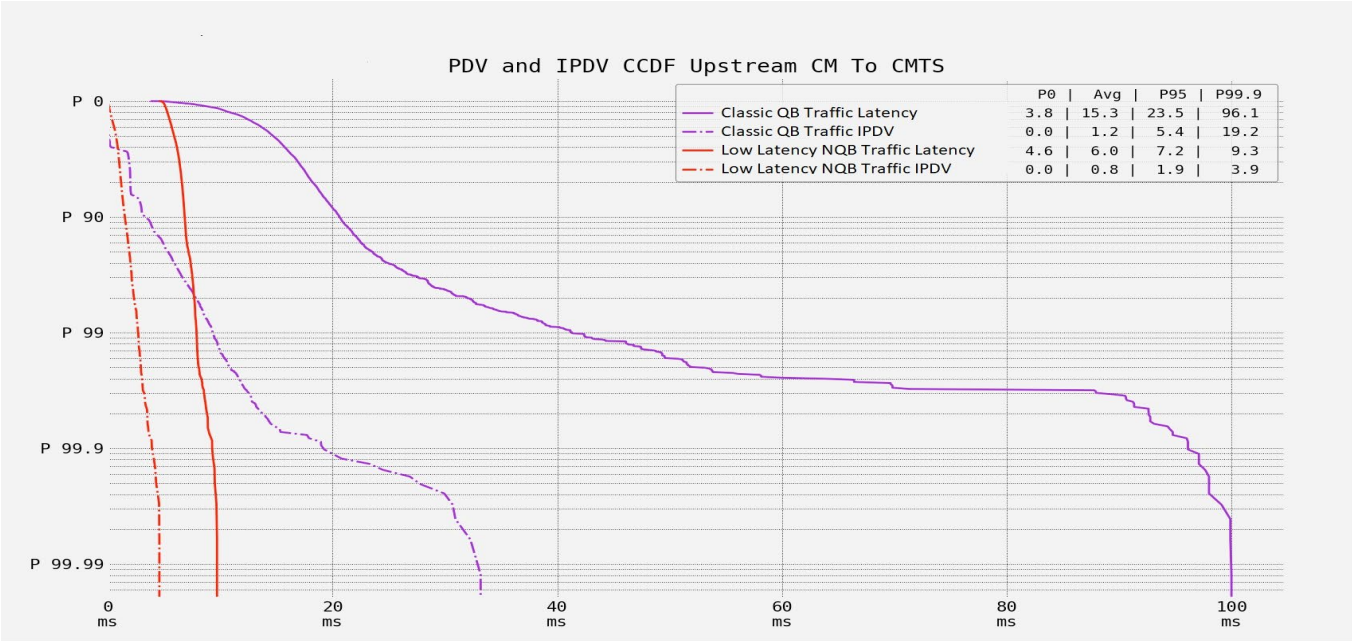
We collaborate with Cablelabs closely for the new LLD features. The dual queue approach will enable us to support the latency and jitter requirements described in Section 2.2 for low latency services. In this section, we present an example scenario tested by Cablelabs[1] with dual queue approach and PGS.

Figure 15 shows the results for upstream LLD without PGS, with a mix of TCP upload file transfers as cross traffic. QB and NQB traffic packets are queued to classic and low latency queues respectively. In the latency distribution charts, both the raw latency distribution and the distribution of Inter-Packet Delay Variation are plotted. Latency distribution is plotted with solid lines, IPDV plotted with dot-dash lines. The 99.9th percentile of LL traffic latency is less than 10 ms.

Figure 16 shows upstream LLD with 2Mbps PGS, with the same cross traffic. The 99.9th percentile of LL traffic latency is less than 6 ms. The NQB traffic has different packet sizes while PGS grants for the selected settings are for 250 bytes. Therefore different latency values for different packet sizes are observed in this test.

We can conclude that the current DOCSIS network latency that may be in the order of 100 ms can be decreased to less than 10 ms for 99th percentile of LL service flow packets with the new LLD features.

[1] We would like to thank and acknowledge Greg White from Cablelabs for his LLD tests presented in this paper.



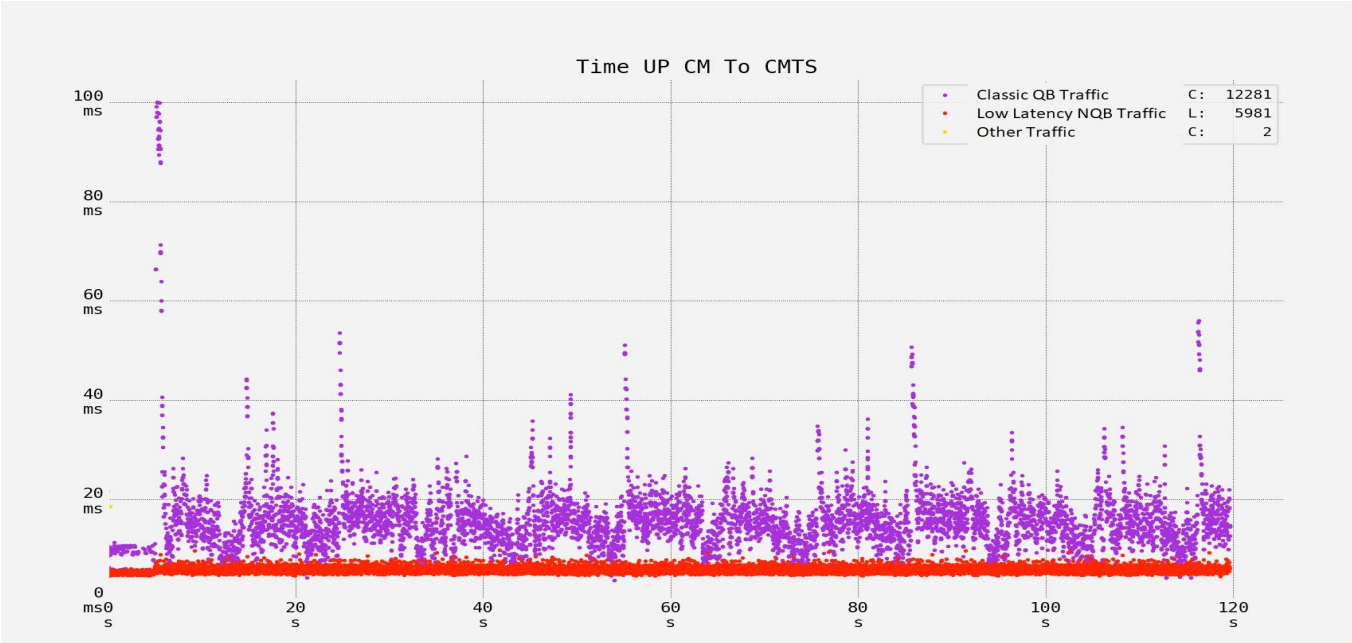
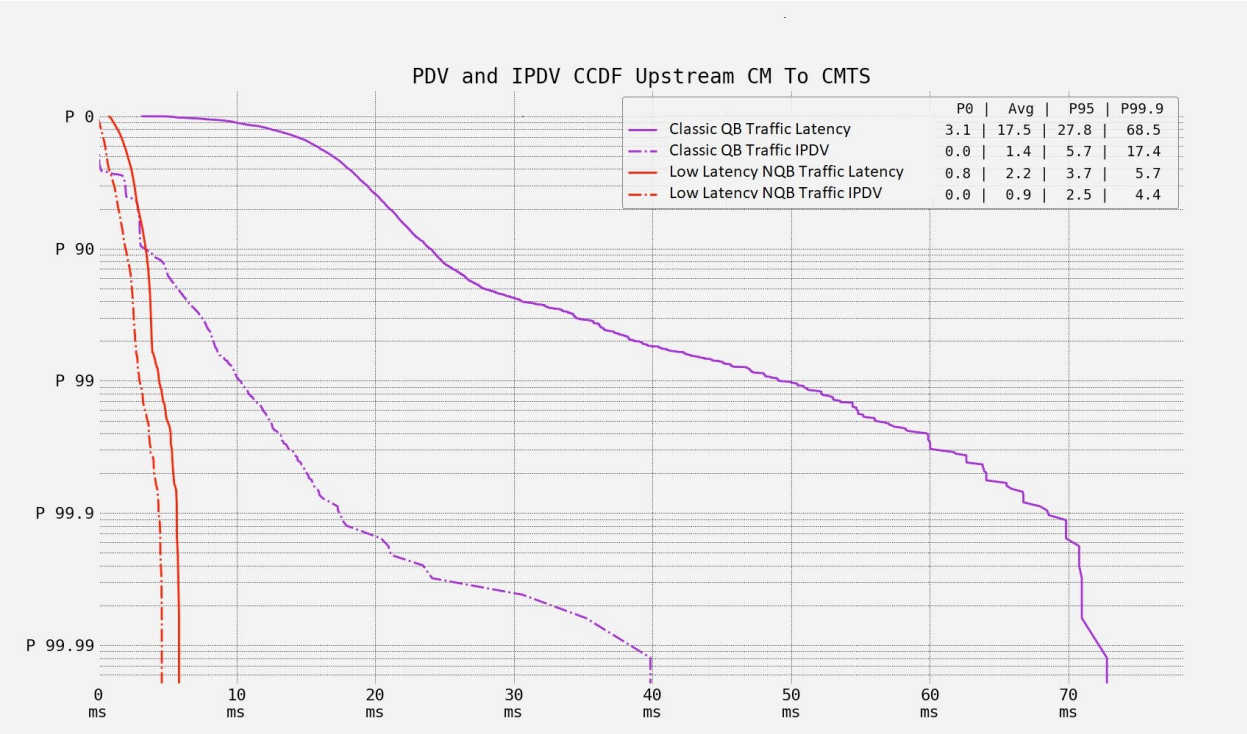


Figure 15 – US LLD With Dual Queue and no PGS



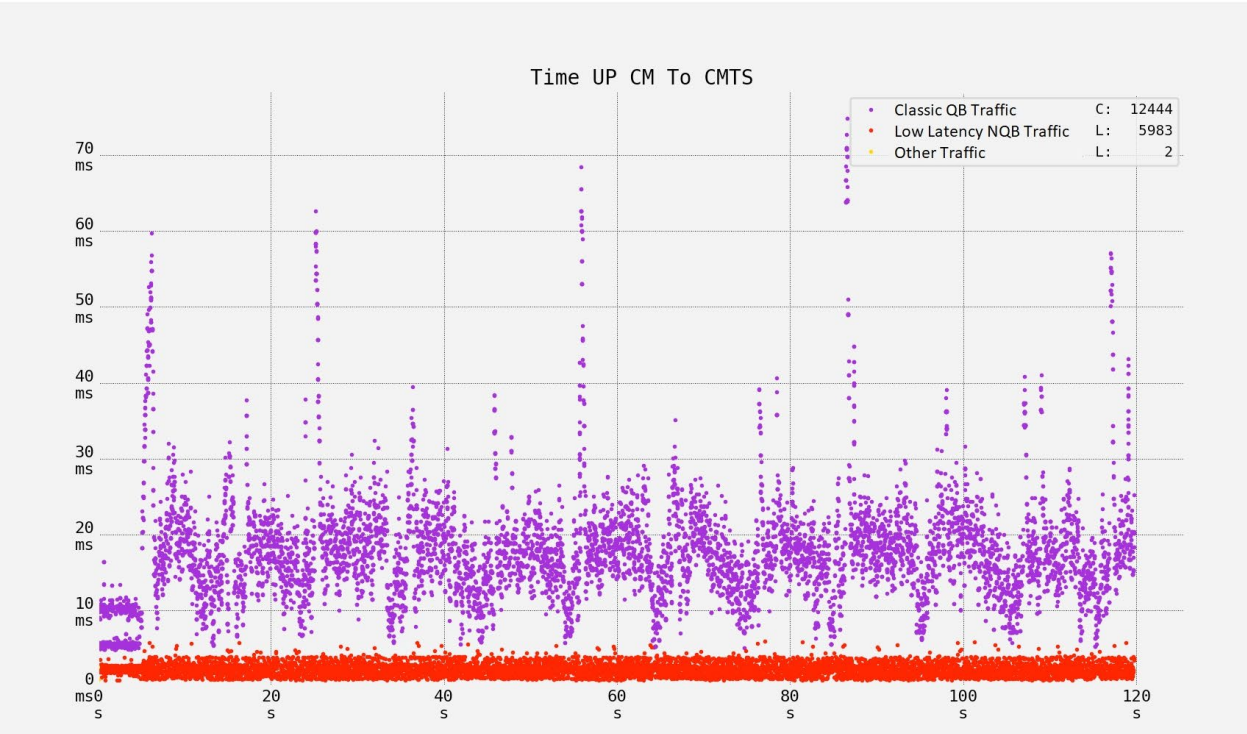


Figure 16 – US LLD With Dual Queue and PGS

These features can be enabled only if Low Latency services are differentiated. Traditionally, Cable Operators do not trust any marking from the user space due to security and operations challenges. The new framework suggests Differentiated Services Code Point (DSCP) marking for low-date-rate NQB traffic and ECN Capable Transport ECT(1) bit marking for high-data-rate NQB L4S traffic (Figure 3 and Figure 17). NQB-DSCP and L4S IETF documents address the challenges of this new framework.

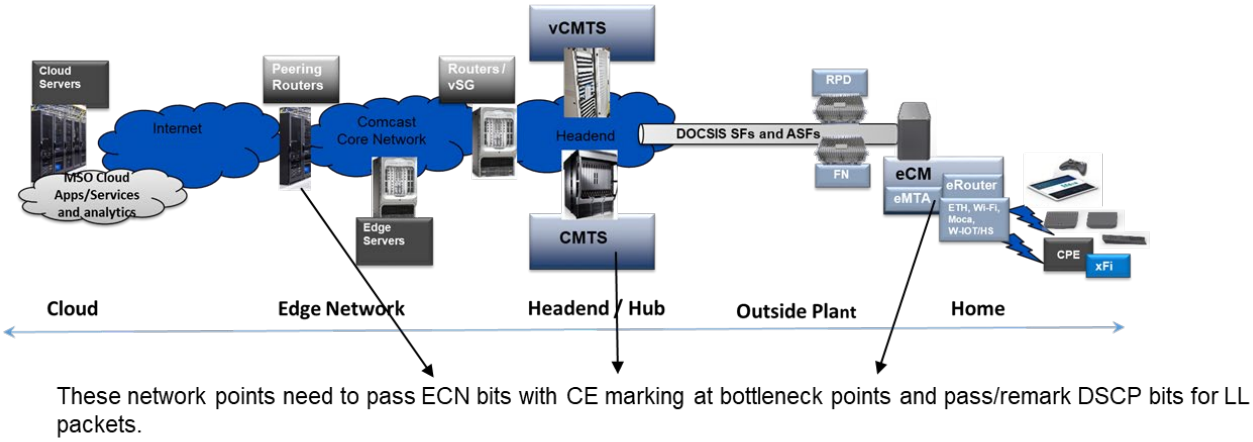


Figure 17 – Marking for LL Services

5. Conclusion

Current latency measurement, monitoring, and management frameworks are fundamental to support the next generation of Low Latency services. These platforms and network architectures must be extended to differentiate LL services to apply D3.1 LLD features, Wi-Fi and Core Network enhancements and to manage their performance. The faster democratization of technologies pushes ISPs to support new services at a faster pace. This can be enabled by achieving three main points within the industry:

1) Better standardize/define how latency, jitter, packet loss and other QoS metrics are measured and create open global internet measurement platforms to focus on end-to-end QoE assessment.
2) Start breaking legacy chains through digitization, software defined, virtualized and cloud based systems with open source software, platform models with partners and co-innovators to meet the consumers' demands in an agile way.
3) Apply an end-to-end approach for traffic differentiation and QoE management with new upcoming 10G technologies.

Abbreviations

AP	Access Point
AQM	Active Queue Management
CE	Congestion Encountered
CM	Cable Modem

CMTS	Cable Modem Termination System
DSCP	Differentiated Services Code Point
ECN	Explicit Congestion Notification
ECT	ECN Capable Transport
FDX	Full Duplex
GBR	guaranteed bitrate resources
HFC	Hybrid fiber-coaxial
IPDV	Inter-Packet Delay Variation
ISBE	International Society of Broadband Experts
ISP	Internet Service Provider
L4S	Low Latency, Low Loss Scalable
LL	Low Latency
LLD	Low Latency DOCSIS
NQB	Non-queue-building
PCEF	Policy and Charging Enforcement Function
PDN	Packet Data Network
QB	Queue-building
QCI	QoS class identifier
QoE	Quality of Experience
QoS	Quality of Service
RTT	Round Trip Time
SCTE	Society of Cable Telecommunications Engineers
STA	Station

Bibliography & References

1. *The Democratization of Technology*; Mihir Shukla, Forbes Technology Council Post, https://www.forbes.com/sites/forbestechcouncil/2019/11/07/the-democratization-of-technology/?sh=765aa8643796
2. *Approaches to Latency Management: Combining Hop-by-Hop and End-to-End Networking,* Sebnem Ozer, Carl, Klatsky, Daniel Rice, John Chrostowski, SCTE-ISBE Workshop 2020
3. *Policy and charging control architecture,* 3GPP TS 23.203 V17.1.0 , 3GPP, 2021
4. *IEEE 802.11 Real Time Applications TIG Report,* 2019.
5. *Latency Measurement: What is Latency and How Do We Measure It?,* Karthik Sundaresan, Greg White & Steve Glennon, SCTE-ISBE Workshop 2020
6. *Mobile cloud gaming: the real-world cloud gaming experience in Los Angeles*, RootMetrics, 2020
7. *A Blueprint of Technology, Applications and Market Drivers Towards the Year 2030 and Beyond*, ITU-T FG-NET-2030
8. *Four reasons why low latency streaming matters*, https://nscreenmedia.com/4-reasons-low-latency-streaming-matters/, 2021
9. *Low Latency DOCSIS: Overview And Performance Characteristics*, White, G., Sundaresan, K. and B. Briscoe, SCTE-ISBE Workshop 2019.
10. *Wi-Fi TSN Capabilities and Evolution Towards Deterministic Low Latency,* Dave Cavalvanti and Ganesh Venkatesan, 2020
11. *Maximum IP-Layer Capacity Metric, Related Metrics, and Measurements*, TR-471, BBF, 2020

12. *High Performance Data Streaming with Amazon Kinesis: Best Practices* (ANT322-R1) - AWS re:Invent 2018
13. *FCC MBA Raw Data Releases*, https://www.fcc.gov/oet/mba/raw-data-releases
14. Cablelabs DOCSIS 3.1 MULPI Specifications

