Breaking the Barriers: Abstracting Traffic Management for Superior Quality of Experience in Multi-Technology Networks

A Technical Paper prepared for SCTE by

Sebnem Ozer, Ph.D.
Principal Architect III
Charter Communications
sebnem.ozer@charter.com

Ramneek Bali
Principal Wireless Engineer III
Charter Communications
ramneek.bali@charter.com

Kamran Yousuf, Principal Engineer P.ENG | PMP, Charter Communications

Moutaz Elkaissi, Director, Product Management, Charter Communications

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

1. Introduction

Today, cable service providers deploy a variety of network technologies to offer a wider range of services to their customers and to tailor their services to specific geographic areas or customer needs. These technologies include Hybrid Fiber-Coaxial (HFC), Passive Optical Networks (PON) and Mobile Networks, with WiFi as the final leg of the connection to end devices or as a hotspot network. Deploying different network technologies can be challenging due to the increased complexity and cost of managing multiple networks. Operators may also face interoperability issues, which can result in poor service quality or higher maintenance costs. Distinct standards govern each of these technologies, which are further characterized by disparate network and service components. Consequently, the development of Quality of Experience (QoE) platforms becomes inherently tailored to specific technologies, creating barriers to achieving a cohesive and unified service framework.

Within this paper, we present a compelling use case for Internet Engineering Task Force - Low Latency, Low Loss, and Scalable Throughput (IETF L4S) and 5G midhaul support [1] over DOCSIS networks in order to optimize the customer's QoE for emerging and future use cases. Our focus delves into the exploration and evaluation of the mapping of 5G Stream Control Transmission Protocol (SCTP) signaling traffic onto the low latency Data Over Cable Service Interface Specification (LL DOCSIS/LLD Service Flow, substantiating the capacity of MSOs to accommodate this emerging service with the LLD support. Furthermore, we address the support for the IETF L4S traffic over multiple network segments. The study encompasses crucial facets like traffic classification, resource mapping and monitoring metrics, each pertaining to the corresponding network segments.

The LLD test system described in this paper uses manual processes for traffic mapping and key performance indicator (KPI) monitoring in a controlled lab environment. Building upon this foundation, we propose an automated solution to support end-to-end QoE by abstracting traffic management with a high-level orchestration platform and a digitized and hierarchical component structure. Ultimately, our paper provides guidelines to ensure optimal QoE delivery while simultaneously curbing the complexities and costs associated with traffic management.

2. The Diverse Network and Service Technologies Within ISP Architecture

Multiple System Operators (MSOs) employ an array of network technologies to broaden the spectrum of services accessible to subscribers, adapting offerings to specific geographical locations or consumer needs. The main technologies [15] are wireline networks such as Hybrid Fiber-Coaxial (HFC), Passive Optical Networks (PON) and Mobile Networks, with WiFi constituting the final link between the connectivity and end-user devices (depicted in Figure 1). Historical architectural design is predominantly centered on purpose-built infrastructure featuring centralized and integrated data, control and management systems. However, propelled by advancements in Software Defined Networking (SDN), Network Function Virtualization (NFV), cloud integration, computational capabilities, switching and data analysis technologies, MSOs have transitioned towards deploying distributed and virtualized configurations. Notable among these are Remote PHY/Virtual Cable Modem Termination System (vCMTS) for HFC (Figure 2), Remote Optical Line Terminal (rOLT)/Virtual OLT for Passive Optical Networking (PON) (Figure 3), Remote Radio Unit (RRU) in tandem with Virtual Radio Access Network (vRAN) (Figure 4) and Open RAN (ORAN) for Mobile Networks. These innovations are also being harnessed within the core and WiFi network segments, where NFV and SDN are being strategically applied.

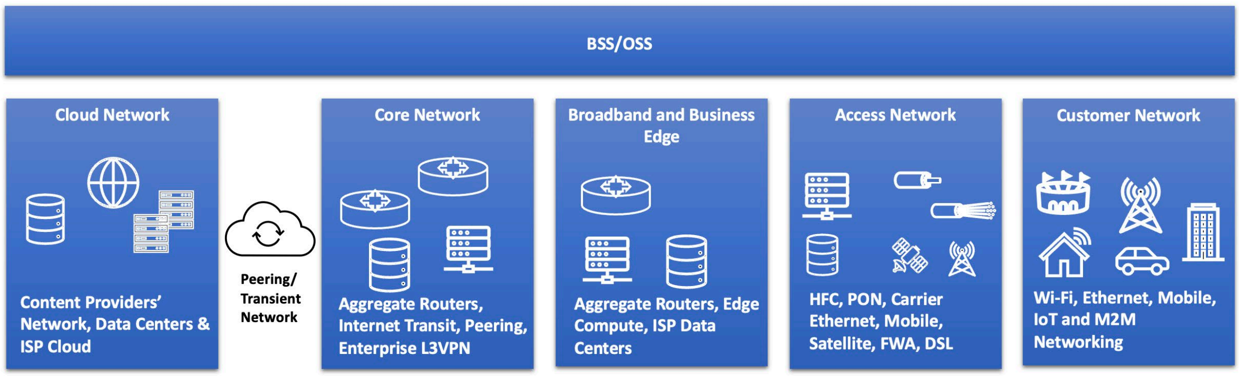

Figure 1- MSO Architecture With Diverse Networking Technologies

In these network paradigms, a prevalent strategy involves leveraging commercial-off-the-shelf commercial (COTS) hardware to deploy containerized applications and microservices, all seamlessly orchestrated by a management platform. Figure 2 displays an example of DOCSIS DAA where Remote PHY (RPD) is connected to vCMTS POD over fiber via a switch.

Here, containerized software materializes vCMTS and other access VNFs. While initial implementations encompassed a monolithic vCMTS data and control paradigm, recent design strategies emphasize microservices. Data forwarding functionality includes classifiers, traffic shaping, scheduling, queueing, filtering, relaying and routing, while control features include subscriber management, service provisioning, device configuration and activation.

Virtual routing interconnects numerous vCMTS PODs with the core network. Similarly, Figure 2 illustrates the PON DAA scenario, where Remote OLT (ROLT) interfaces with vOLT PODs. ROLT integrates PHY and lower MAC functionalities, while PON network provisioning and management functionality is virtualized. The modular separation of monolithic functions offers MSOs the avenue to disentangle data and control VNFs, harmonizing common attributes across DOCSIS, PON and ethernet technologies, while remote PHY/MAC components nestle within a remote switch node [3]. The orchestration of services and resources encompasses lifecycle management, automation and data analysis, all empowered by cloud-based telemetry.

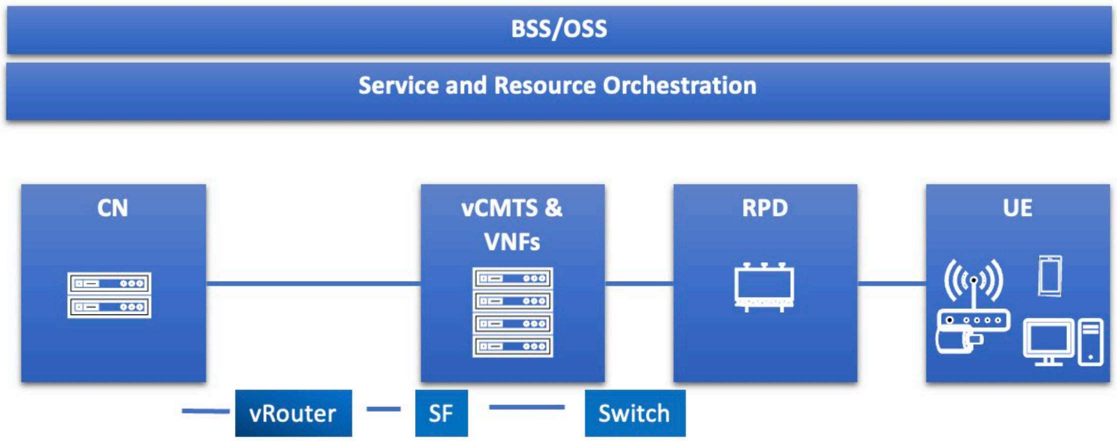

Figure 2- DOCSIS DAA Example

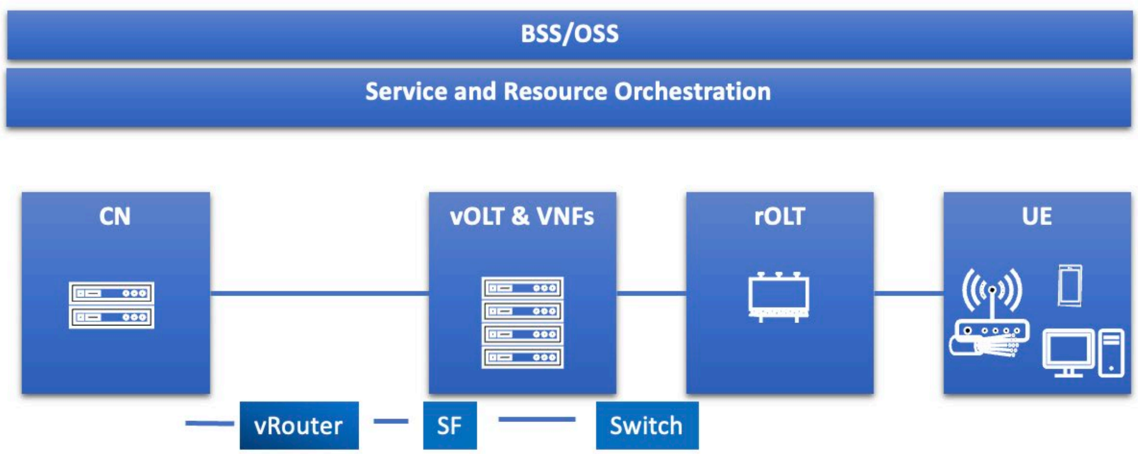

Figure 3- PON DAA Example

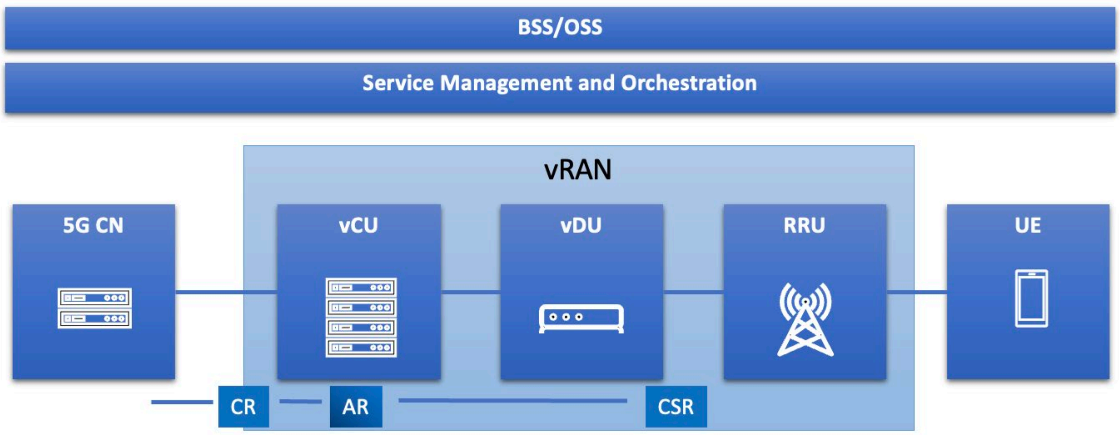

Figure 4- vRAN Architecture Example

Figure 4 shows a VRAN architecture. Virtual Centralized Unit (vCU) provides non-real-time processing and access control. It manages higher layer protocols including Radio Resource Control (RRC) from the control plane, and Service Data Adaptation Protocol (SDAP) and Packed Data Convergence Protocol (PDCP) from the user plane. The vCU is connected between the 5G core network and the Virtual Distributed Unit (vDU). One vCU can be connected to multiple vDUs. The vDU provides real-time processing and coordinates lower layer protocols including Physical Layer, Radio Link Control (RLC) and Media Access Control (MAC). Virtualization shifts the vCU and vDU from dedicated hardware to software components, allowing for flexible scaling, as well as rapid and continuous evolution. With vDU, all the baseband functions of real-time RLC/MAC/PHY layers are executed over the commercial-off-the-shelf (COTS) server Remote Radio Unit (RRU), which does the physical layer transmission and reception, supporting technologies such as Multiple Input Multiple Output (MIMO). Session management, usage metering, common policy, authentication and network slicing are implemented in the virtualized core network. Service management and orchestration (SMO) includes life cycle management, automation, analytics and non-real time intelligent controller. ORAN architecture defines open interfaces between RRU and DU and between DU and CU.

Amidst these breakthroughs in ISP network and service architecture innovation, concurrent efforts are still underway to enhance subscribers' quality of experience regardless of location (e.g. in the home/office or on the go), time (e.g. peak and off-peak utilization times) or device (e.g. wired and wireless devices), while simultaneously streamlining expenditures by enhancing operational and business support systems. These endeavors collectively harmonize diverse access and service technologies through unification and convergence. In this paper, we address this framework, with a concentrated focus on the facilitation of low-latency services. Notably, while our discourse employs the terminology of "low latency services," our underlying perspective interprets this as addressing the Quality of Service (QoS) requirements across distinct services via a comprehensive approach aimed at bolstering Quality of Experience (QoE). Our investigation thus addresses how MSOs can channel investments into future-proof network and service architectures.

3. Everchanging Landscape of Quality of Experience Framework

In the ever-evolving landscape of the Internet, extensive research revolves around the critical factors of Quality of Service (QoS) that affect subscribers' Quality of Experience (QoE), which is a multidimensional and multi-disciplinary phenomenon. One QoS metric that has been monitored and managed extensively is bandwidth, although it's widely used interchangeably with speed, capacity, throughput or bit rate [2]. The QoS criteria, including parameters such as throughput, latency, jitter and packet loss, are contingent upon the specific characteristics of the service, the application in question, and the unique requirements of the customer use cases.

Embedded within this narrative is a fundamental premise: the intricate interplay of bandwidth, latency and reliability necessitates a holistic approach. This complexity arises from the cyclic interdependency encompassing network and service technologies, application evolution and evolving customer expectations, all propelled by induced demand. Technological enhancements usher in higher speeds, reduced latency and augmented reliability. This, in turn, spurs novel applications that mold consumer behavior and aspirations, thereby fostering increased demand for advanced services and network technologies, forming a self-perpetuating cycle. Reflecting upon the last decade's Internet development and its assimilation into diverse demographics, we glean insights into the path that the ensuing 10 years may traverse.

While projections regarding bandwidth requirements may differ, consensus emerges around the ascendant significance of attributes beyond mere speed. Reliability, consistent low latency and responsiveness occupy the forefront. To complement advancements in self-healing and automated solutions, customer service that adeptly resolves issues will wield equal significance.

Within the scope of Quality of Experience (QoE), the foundational elements remain steadfast, yet the dynamic consumer behavior and expectations continue to shift in tandem with the rapid evolution of technology. A compelling illustration of this phenomenon emerges when examining the statistics related to the abandonment rate of subpar streaming content on connected TVs and virtual Multichannel Video Programming Distributors (vMVPDs) in the United States during the years 2017 and 2018 [5]. Evidently, there was a significant surge in this rate over one year, underscoring a discernible decline in consumer tolerance for videos that exhibit glitches or compromised quality.

Extending our focus to a distinct study conducted in early 2021, a revelation unfolds – half of U.S. consumers opt out of online purchases when confronted with a slow and complex checkout process [5]. It is noteworthy that research has brought to light the impact of network delays on the loading time of web pages and associated activities. A revealing example is a study [6] which indicates that a mere 10ms increase in network latency engenders an additional 0.8 seconds for a page to load. A 60ms delay prolongs loading time to 1.7 seconds, while a 510ms lag culminates in a 17-second delay.

As expectations grow, so do the number of users. Anticipating a burgeoning future, it's anticipated that approximately 187.5 million U.S. users will have conducted at least one online purchase via web browsers or mobile apps on their mobile devices by 2024 [5]. This projection, a notable climb from the 167.8 million mobile U.S. buyers recorded in 2020, augments the prospect of a digitally immersive future.

A glimpse into the rapidly expanding realm of Internet of Things (IoT) devices, as depicted in [5], emphasizes the transformative role of digital innovation in every facet of modern existence. Notably, consumer internet and media devices, followed closely by smart grids and connected vehicles, are at the forefront of IoT applications, as portrayed in Figure 5. This extensive survey epitomizes the broad spectrum of services that necessitate dynamic quality of experience due to the ever-shifting landscapes of residential, enterprise and industrial services.

In the wake of cutting-edge developments, novel services like Augmented Reality/Virtual Reality (AR/VR) usher in fresh paradigms for sensory evaluation, while the realm of Machine-to-Machine

(M2M) services introduces new dimensions of criteria. Another thought-provoking survey [5] encapsulates consumers' unchanging expectation for consistent service quality, regardless of the underlying technology. Thirty-nine percent of surveyed gamers in 2022 anticipated that mobile adaptations of games would rival their PC or console counterparts in terms of storyline and narrative quality.

The widespread demand for seamless experiences across devices is palpable, as evidenced by a survey from Q3 2022, revealing that a staggering 81.9 percent of global internet users actively engage in video gaming across diverse platforms. A staggering 66.2 percent of respondents endorsed smartphone gaming, crowning it as the most popular form of gaming worldwide. In a close second, PC gaming via laptops or desktops claimed a substantial 37.9 percent share.

The far-reaching ramifications of latency, jitter and packet loss on various applications, including traditional and cloud-based gaming, as well as videoconferencing, is explained in [2]. The impending advent of immersive applications like AR/VR across domains such as entertainment, education, healthcare and more underscores their pronounced sensitivity to latency, jitter, packet loss and throughput performance. In tandem with these Quality of Service (QoS) metrics, reliability, security, privacy and accessibility collectively shape subscribers' satisfaction with services. A pressing imperative for cable operators lies in adapting to the flexible QoE frameworks as discerning subscribers increasingly demand services that transcend temporal, spatial and device constraints.

As technology advances and network and service technologies elevate user expectations, the tolerance for disruptions diminishes. This intricate interplay underscores the central importance of staying attuned to evolving performance standards. The subjective nature of QoE requires personalized solutions, but MSOs need to support it without requiring custom solutions. This is why network and service programmability with automated platforms that enable self-optimization continues to drive design improvements.

Pioneering strategies in network monitoring and operations come to the forefront, aimed at precisely measuring QoS metrics encompassing latency and jitter behaviors for a diverse range of traffic characteristics. The crucial significance of the IETF L4S and Non-Queue-Building Per-Hop Behavior (NQB-PHB) specifications and drafts lies in their mandate for measuring scalable TCP and NQB-PHB traffic, both at idle and working latency levels, in conjunction with classic traffic measurements [7]. Hop-by-hop measurements are used to find bottlenecks and other impairments at each network segment, while end-to-end measurements map all the findings to QoE.

Strategically opting for both active and passive latency measurement platforms, catering to both hop-by-hop and end-to-end analyses, emerges as a pivotal pursuit. It's important to capture network and service states during specific timeframes that precipitate latency fluctuations, as this enables the identification of key correlations and causations. For instance, latency sources in DOCSIS networks primarily stem from queuing and media access, while factors such as propagation, serialization, encoding and switching also contribute to latency. PON networks exhibit notable distinctions, especially from a media access and queuing standpoint, but the major latency and jitter factors are similar. Although propagation delay differs between coaxial and fiber mediums, the introduction of DAA has brought fiber closer to the subscriber site in HFC networks, resulting in an increased portion of the distance being covered by fiber.

While 5G stands apart due to its distinct medium and protocols, the same underlying factors impact both latency and jitter. In the case of WiFi, inter-AP and external interference, as well as downstream and upstream contention, the dynamics of latency and jitter factors are magnified. An important consideration is that detecting latency variations and their origins is challenging until in-packet measurements become feasible, given the potentially sub-millisecond or lower durations of these events.

Embracing automation and data analytics emerges as a promising avenue for effectively fine-tuning measurement and analysis protocols, ushering in a more efficient era of QoS evaluation. Delving into the heart of network segments, and keenly understanding their inherent characteristics, proves indispensable in tailoring QoS measurement tools. A notable instance is the nuanced role of core networks, which wield a more pronounced influence on latency than jitter, while WiFi, on the other hand, significantly impacts jitter. This comprehension paves the way for precision-tailored tuning strategies, enhanced by judicious application of data analytics and automation.

As depicted in Figure 6 where an aggregate view of internet usage is collected for network analysis, a DOCSIS service group's utilization at various resolutions showcases microbursts occurring at different aggregate levels. However, most traditional and modern distributed MSO architectures typically gather data in 15- to 300-second measurement windows, potentially missing microbursts and their impact on jitter. Similarly, unless Received Signal Strength Indicator (RSSI) or other link issues, along with microbursts arising from home network applications, are measured at smaller resolutions, their influence on WiFi latency and jitter might remain unnoticed. MSOs can leverage data analytics to detect such events.

The dawn of new Simple Network Management Protocol Management Information Bases (SNMP MIBs), encompassing latency histograms and congestion metrics (i.e. docsQosSfLatencyStats and DocsQosSfCongestionStats) within CableLabs DOCSIS specifications, enhance the QoS monitoring for MSOs. Employing similar measurements in other network segments can also improve hop-by-hop and end-to-end analyses if executed in a coordinated manner. This coordination entails synchronized measurements, standardized metrics and integrated dashboards alongside comprehensive data analysis.

In conclusion, the unceasing evolution of technology beckons us to explore the ever-shifting terrain of Quality of Experience (QoE). As technology reshapes expectations, it's imperative for stakeholders to navigate this landscape with agility, embracing the tools of analytics, automation and adaptability to ensure that the march towards optimal quality remains uninterrupted.

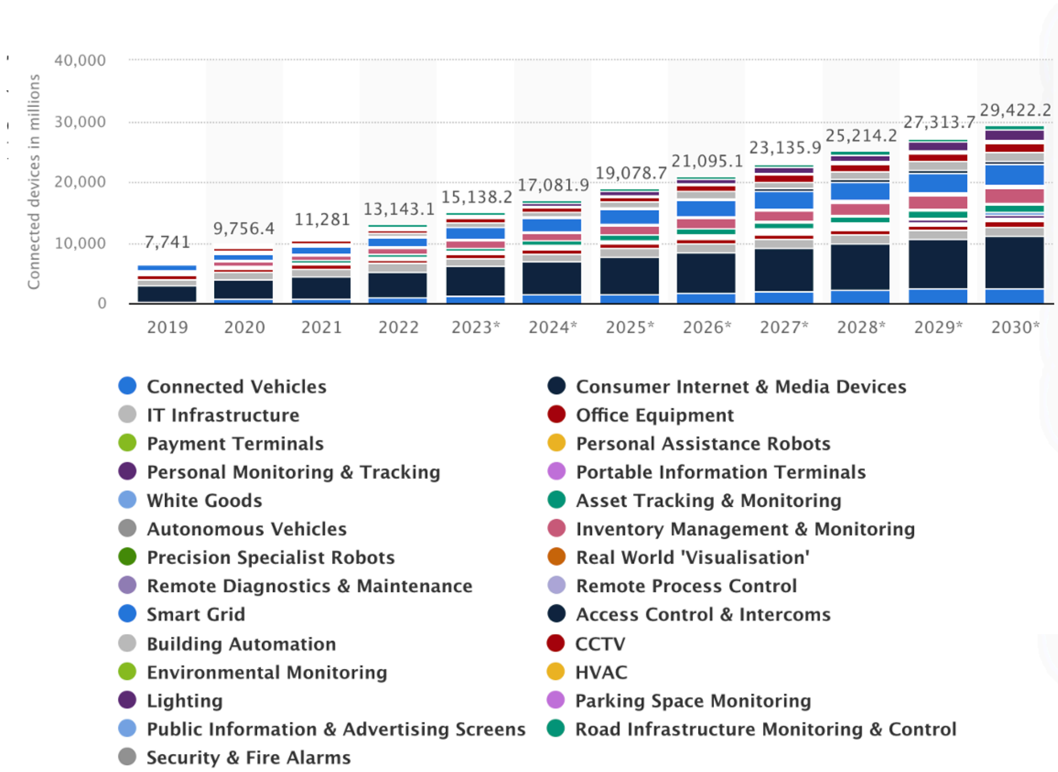

Figure 5- IoT Usage

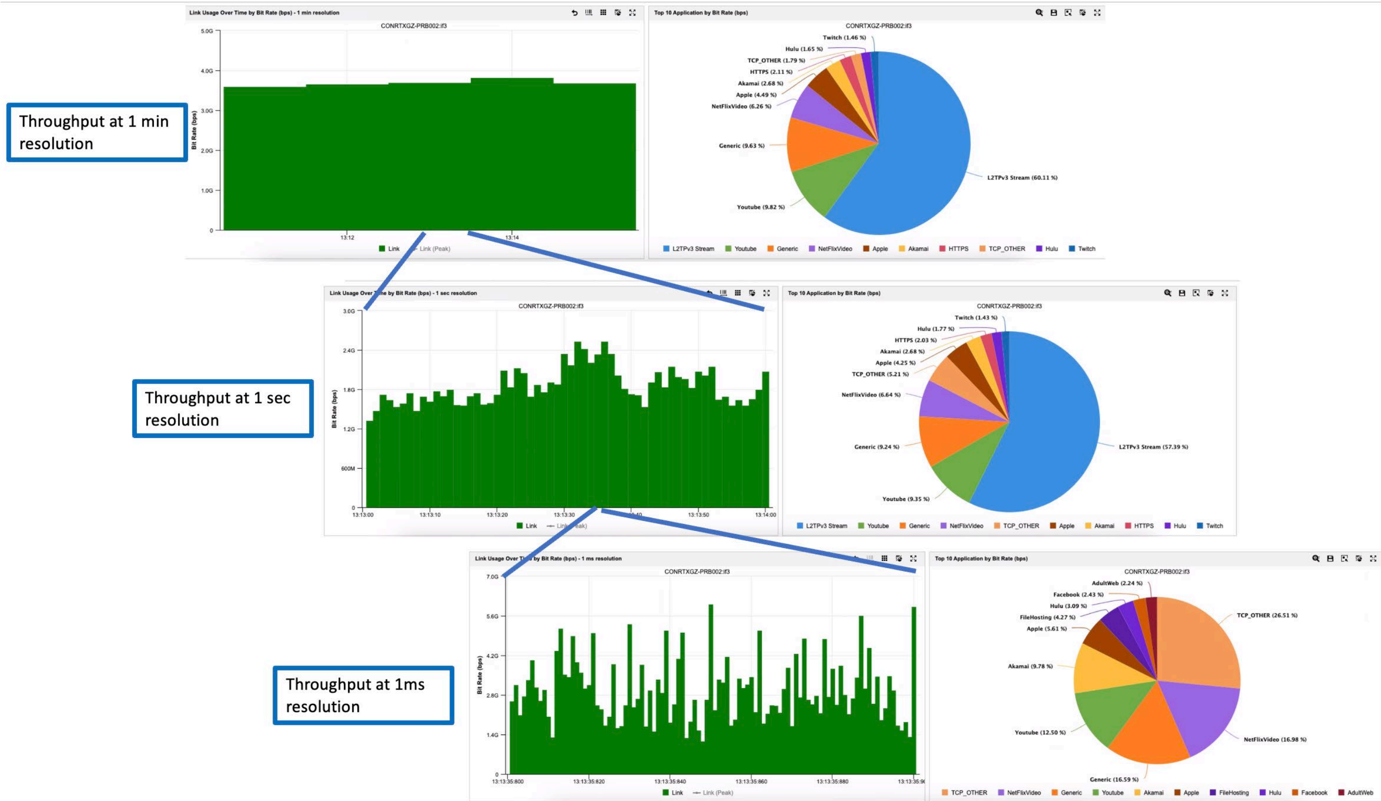

Figure 6- DOCSIS Serving Group Utilization Characteristics per Observation Time Windows (source: NetScout Packet Inspection Tool)

4. Case Study: Supporting IETF L4S and 5G Midhaul Over DOCSIS Networks

The LLD improvements are first tested for Low Latency traffic (i.e. IETF L4S and PHB-NQB) [10,11] to assess the wireline quality improvements over DOCSIS. Figure 7 shows a lab result where LLD ensures downstream L4S traffic to keep consistent low latency and stable throughput changes during bursts created by other household utilization that saturates the home network capacity. Even when full home bandwidth is used, IETF L4S traffic RTT stays low while total capacity is efficiently used. Figure 8 illustrates the one-way latency percentiles of upstream and downstream low rate UDP traffic when home network is saturated for different bandwidth rates. The latency and jitter improvements for LL marked traffic can be observed when compared to unmarked traffic results. The findings indicate that increased bandwidth does not necessarily ensure consistent reductions in latency and jitter, particularly when a home network's bandwidth becomes saturated. The phenomenon of induced demand introduces scenarios in which microbursts may occur, even within non-latency-sensitive applications that can derive advantages from transmitting substantial traffic volumes. One important aspect is that this may happen in any network segment candidate to be the bottleneck over the service path.

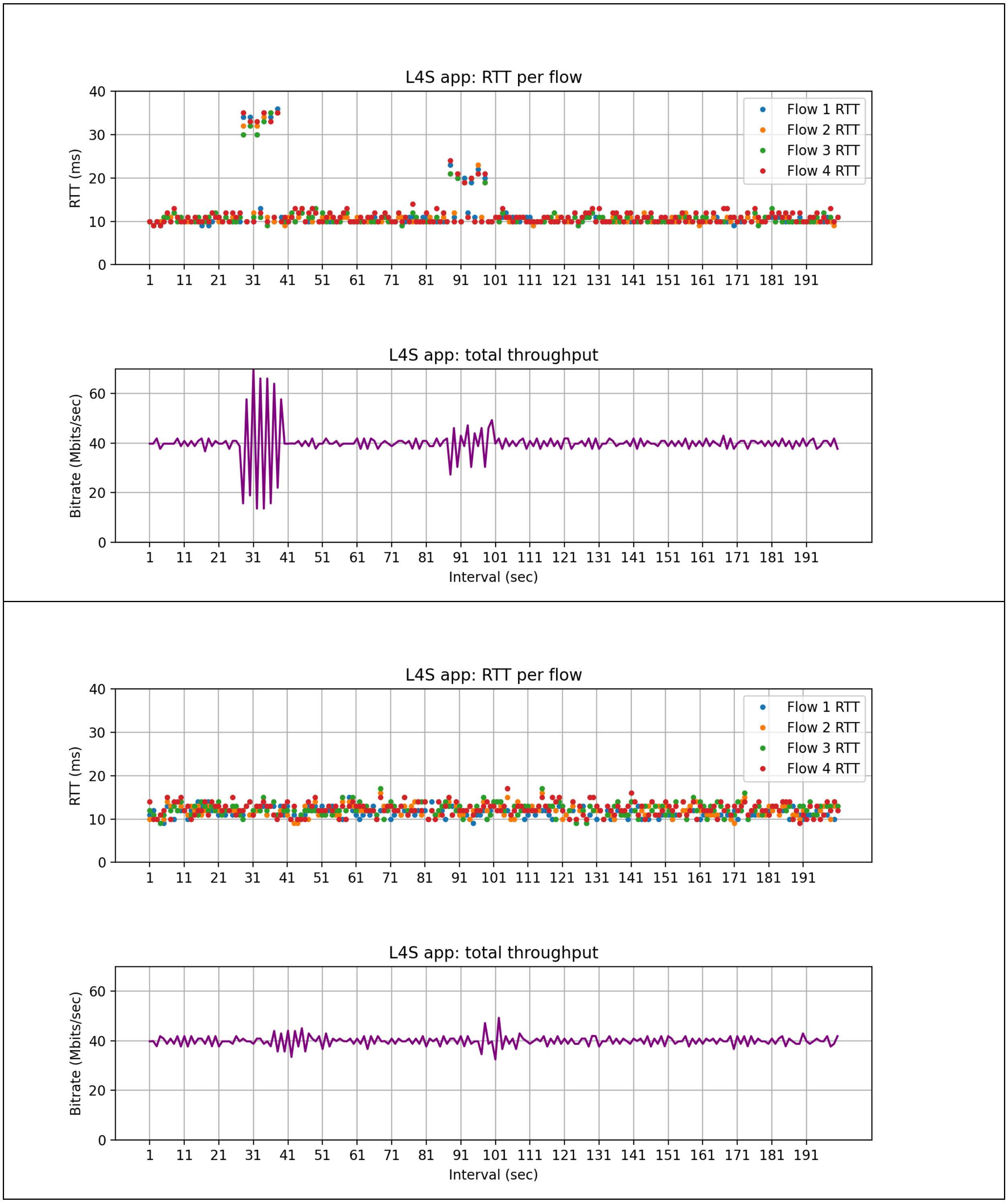

Figure 7- Downstream IETF L4S application throughput and RTT when home network is saturated. Top: Without LLD; Bottom: With LLD

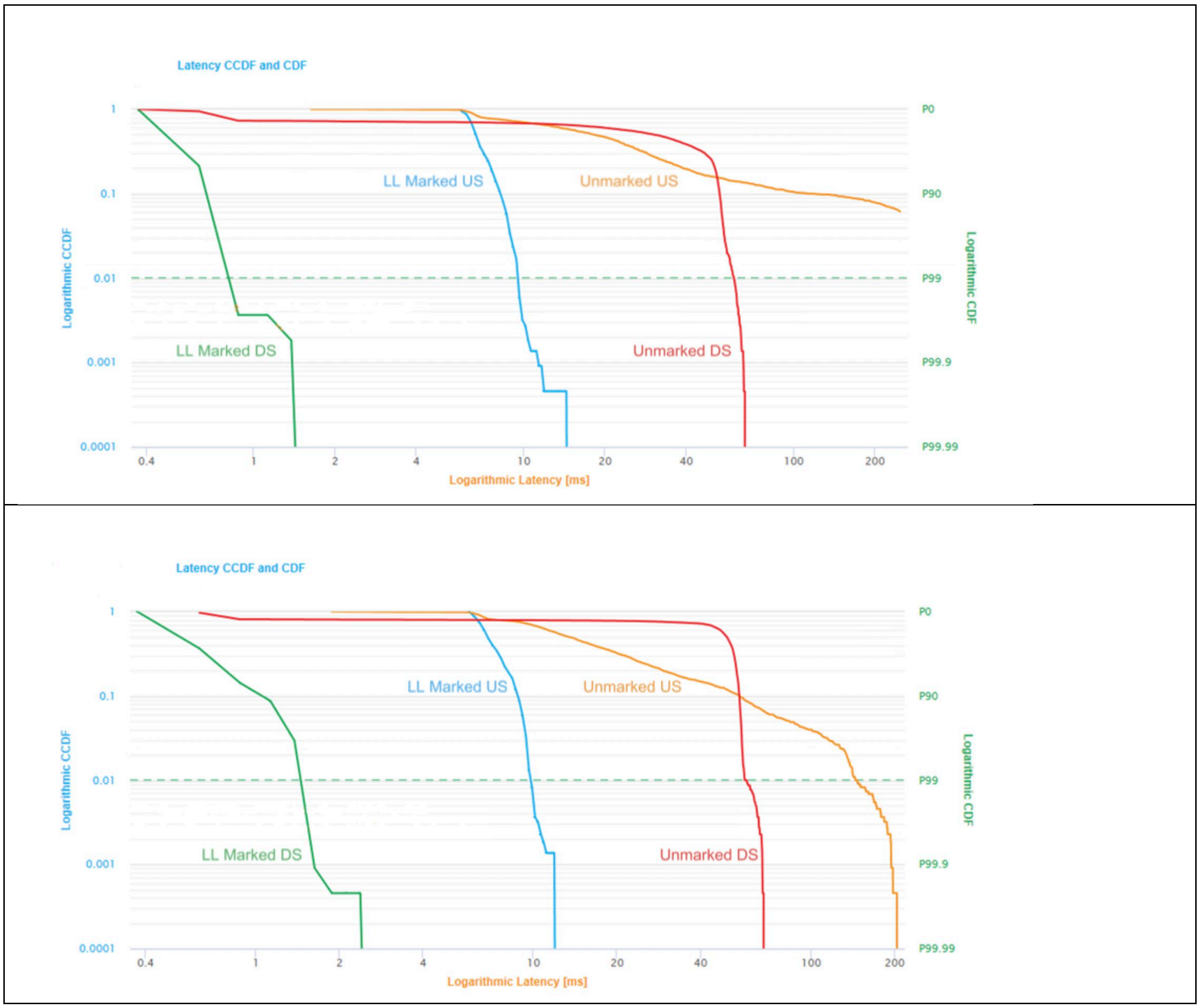

Figure 8- IETF PHB-NQB Traffic One Way Latency Top: 300/10Mbps ; Bottom: 1G/200Mbps

Virtualization of RAN offers the flexibility of varied deployment options based on the latency requirements. The deployment possibilities for MSOs are discussed in [1]. MSOs can leverage their HFC network to build a network of 5G small cells to offload MVNO traffic, thereby reducing service cost. Today most of the HFC networks can support backhaul use case (Figure 9) without additional upgrades or architecture changes. With the LLD support, the feasibility of midhaul (Figure 10) use cases becomes also achievable by only firmware changes [4], which is studied in this section. Fronthaul options would require specific architectural design and protocols, like LLX.

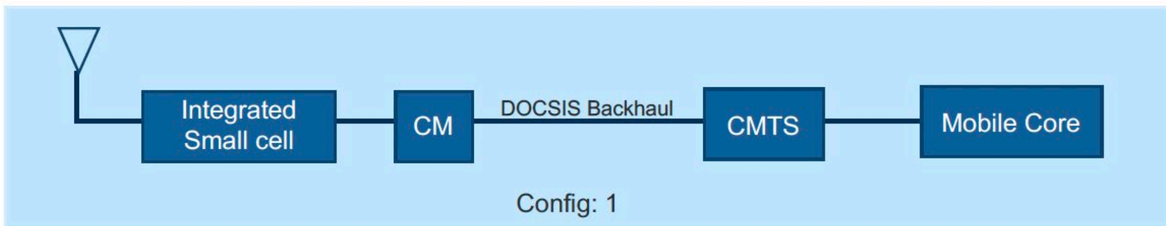

Figure 9- RAN Architecture - Backhaul

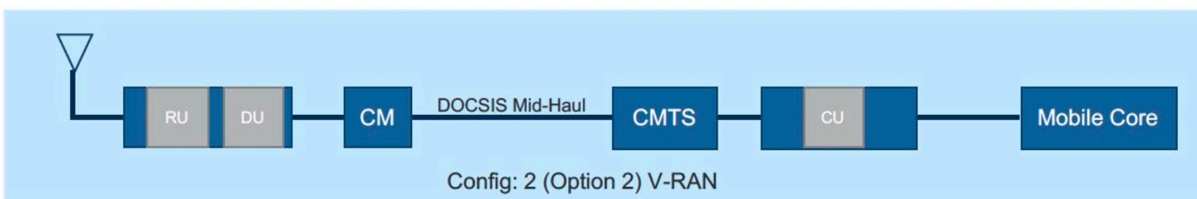

Figure 10- vRAN Architecture - Midhaul

The F1 interface (midhaul) connects a gNB vCU to a gNB vDU. This interface is applicable to the vCU-vDU Split gNB architecture. The control plane of the F1 (F1-C) allows signaling between the vCU and vDU, while the user plane of the F1 (F1-U) allows the transfer of application data. The latency performance of DOCSIS with LLD is best suited for this type of deployment with vDU at the edge and vCU centralized or at the edge.

A lab setup with 5G virtual core, V-RAN on open platforms and HFC network was done to evaluate the performance of the midhaul use case. The main challenges of the lab setup were interoperability issues with the integration of OS, vCU and vDU SW, along with routing and messaging on standard interfaces, while underlaying DOCSIS network setup and configuration was relatively easy. Unloaded DOCSIS throughput performance is shown in Figure 11 for benchmarking purposes of achievable peak VRAN (100Mhz) rates with selected configuration parameters. Total channel capacity is ~180Mbps in upstream and ~2.25Gbps in downstream. Figure 12 shows latency values when utilization is close to channel capacity.

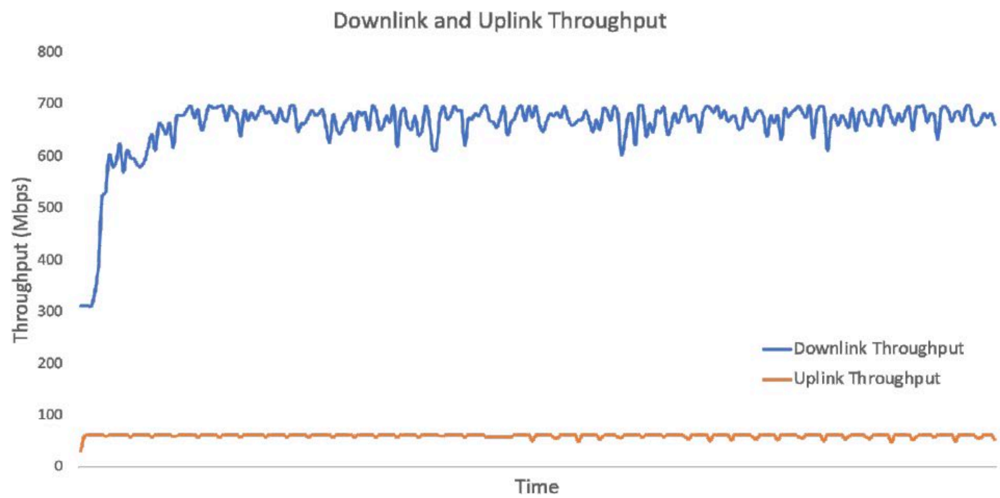

Figure 11- Unloaded DOCSIS Throughput in the Lab Setup

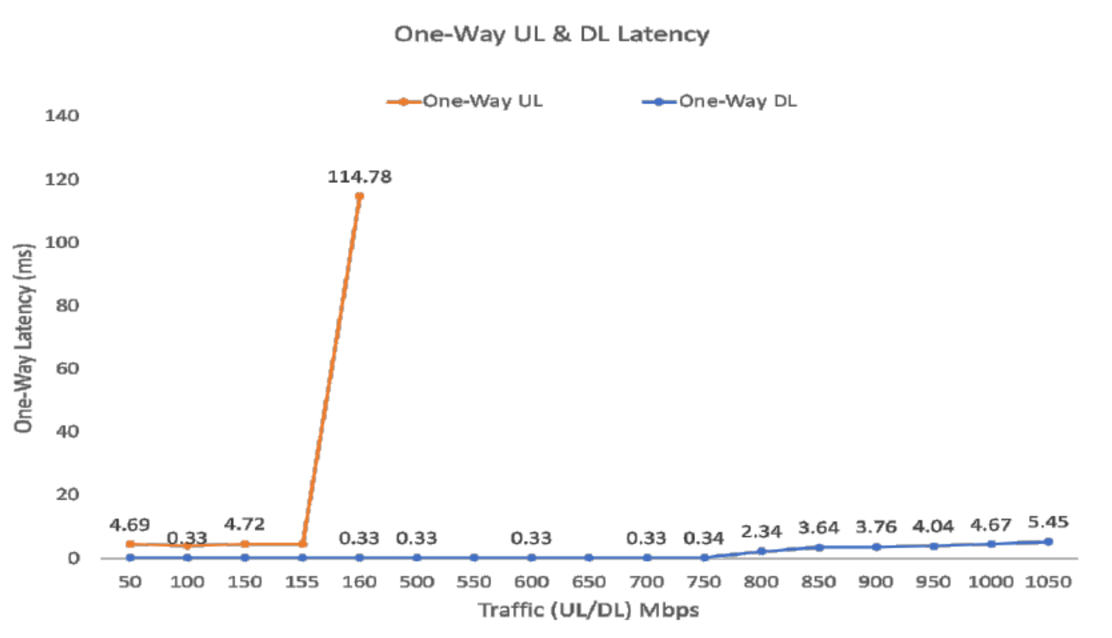

Figure 12- DOCSIS Latency Benchmark in the Lab Setup

Two use cases are used for each test suite. “Without LLD” refers to the case where signaling traffic is not classified to DOCSIS LL SF. In this case, both signaling and user data are sent through the same DOCSIS service flow as illustrated in Figure 13 [10, 11]. Both flows of traffic use the same queue with classic Active Queue Management (AQM) and they are subject to the same scheduling decisions. “With LLD” refers to the case where signaling is classified as LL SF and user data as classic SF in the same Aggregate Service Flow (Figure 14). In this case, user data is sent through Classic AQM while signaling traffic is forwarded through IAQM defined in [10, 11]. A weighted intra-SF scheduling is applied in the CMTS for upstream and downstream traffic. Queue protection detects and sanctions misbehaving traffic while coupling aims throughput fairness between classic and LL traffic.

The emphasis was given on the upstream performance due to the limiting factor in terms of latency and jitter, which varies with the utilization and other factors. Figure 15 displays the results for the multiples of the number of simultaneously active UEs per small cell, which increases with the LLD deployment. When SCTP is classified as LL traffic, more users can be supported in the small cell in a reliable way even when DOCSIS load percentage increases without requiring architecture or hardware changes as long as DOCSIS 3.1 is deployed with LLD.

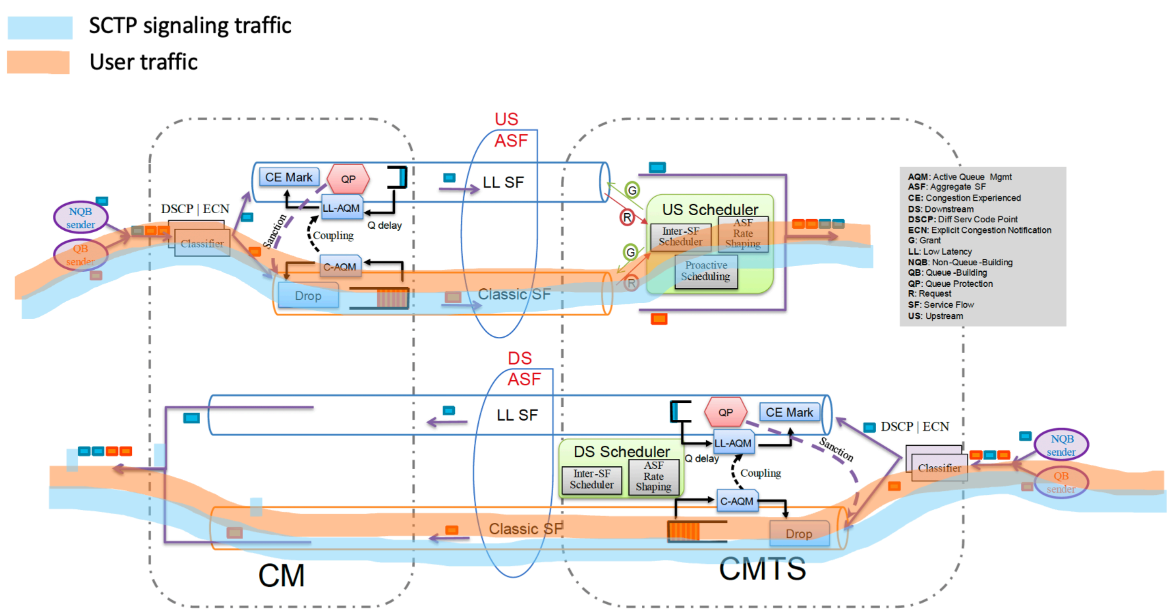

Figure 13- “Without LLD” Use Case

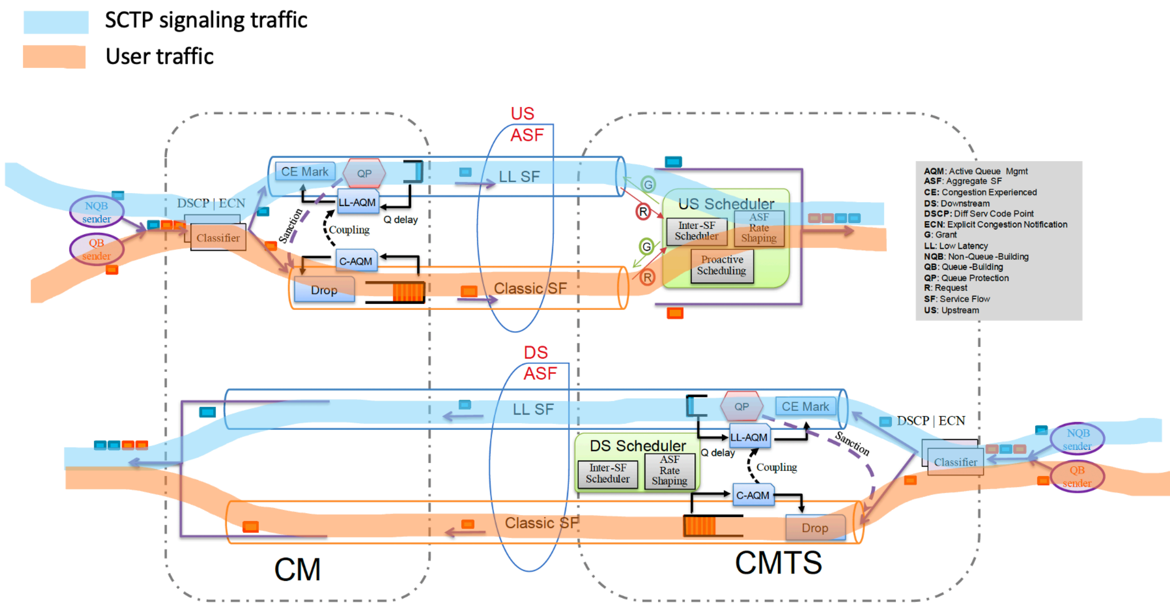

Figure 14- "With LLD" Use Case

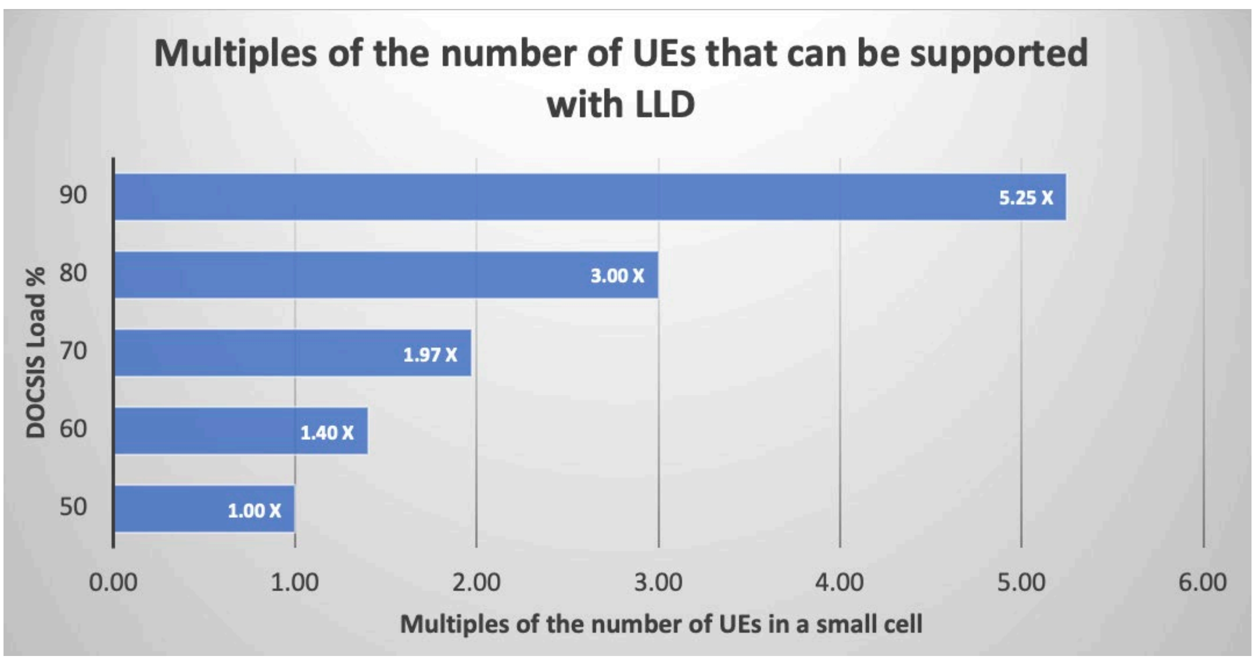

Figure 15- Multiples of the Number of UEs Supported for Different DOCSIS Load Percentages with Additional LLD Firmware in a Midhaul Use Case

Creating this proof of concept required several critical considerations:

- Mapping SCTP signaling to DOCSIS LL Service Flow
    - SCTP signaling is differentiated in the 5G domain, logically from user data, but normally it is aggregated in the same DOCSIS SF that carries all xhaul traffic
- Measurement of QoS metrics and mapping them to subscriber experience
    - Employing throughput, latency, jitter and other QoS measurement methodologies capable of distinguishing between various traffic categories
    - Analyzing end-to-end KPIs (e.g., service availability/connectivity for UEs)
- Understanding QoS functionalities at different network segments
    - Forwarding of signaling traffic in 5G versus forwarding of DOCSIS Service Flows (e.g., queueing and scheduling)
- Analysis of factors contributing to QoS impairments
    - Identifying impairments specific to 5G versus DOCSIS networks, including considerations like utilization ratios

Our laboratory analysis holds the potential for broad applicability across other traffic categories and their corresponding demands for an enhanced quality of experience. Although our lab procedures included manual steps for testing purposes (e.g., classifying SCTP signaling traffic as LL SF and using different monitoring metrics from vRAN and DOCSIS), MSOs can extend support to generalized use cases by integrating a traffic abstraction layer within the framework of service and network orchestration, a topic covered in a subsequent section.

5. Future-Proofing the Network and Service Platforms

In this section, we discuss several design items that MSOs should consider for future-proof QoE support in heterogeneous networks and service architectures. Although convergence of multiple access technologies and BSS/OSS systems are being investigated [8], functionalities like traffic abstraction can be supported before full convergence is feasible and incorporated to the convergence in the future phases.

One critical design choice is the digitization of network and service components defined with a unified set of metrics so that they can be defined through models that are agnostic of access technology and abstracted from low layer attributes (Figure 16). Network components can be various Virtualized Network Functions (VNF) and Physical Network Functions (PNF) [9]. The service request is then conveyed by using an inventory system of mapped digitized components. Resource attributes are maintained in a common knowledge center for each digitized network resource component. Based on service and resource data, an orchestration layer creates configuration models that can be adaptive to network and resource states. Such a system can provide automated rules to support 5G backhaul service by configuring CM and vCMTS components with corresponding classifiers that are mapped to 5G traffic classification by using programmable and automated systems. Data analytics can be part of the orchestration system that evaluates the service and network states, e.g., by observing UEs that have reliable service with required metrics versus UEs that are denied service or lost connectivity.

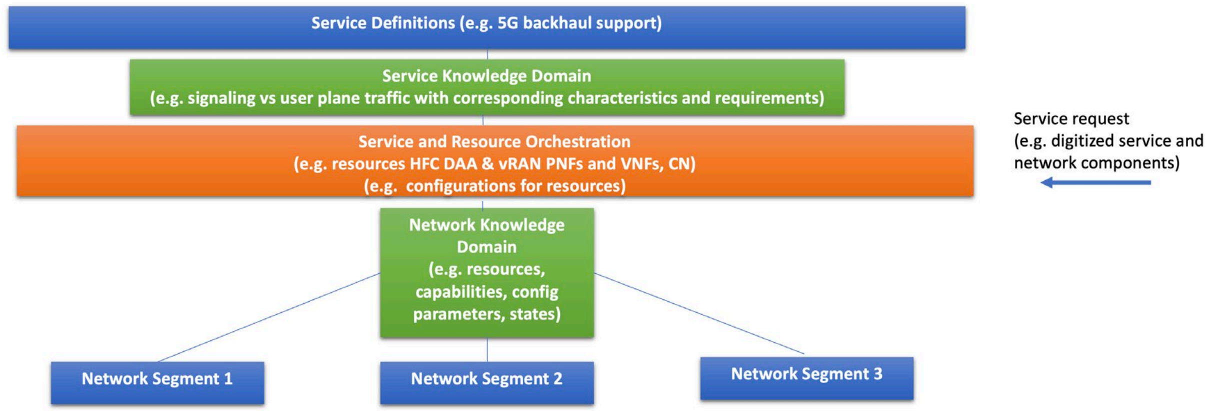

Figure 16- Traffic Abstraction for Automated and Programmable Systems

An essential element in supporting a system, as depicted in Figure 15, is the alignment of standards to ensure convergence. Our focus delves into three key facets of standardization endeavors that guide the mapping from QoS to QoE:

- End-to-end QoS path (Figure 17): Automating traffic classification mapping and translation at network boundaries
    - This translation encompasses various standards such as DOCSIS and PON Service Flows (SFs) and Aggregate SFs, 5G QoS identifiers (5QI), IEEE 802.11's user priority (UP) and similar traffic classifications. Multiple standardization efforts have tackled the concept of end-to-end traffic classification. IETF L4S and NQB-PHB specifications and drafts propose the use of ECN and DSCP bits for low latency service flows, endorsing end-to-end marking utilization. For instance, Low Latency (LL) traffic can be a subset within a DOCSIS ASF or WiFi MultiMedia (WMM) Access Category, correlated to a UP while mapped to a single 5QI. Such, mapping at network segment boundaries, while preserving LL markings, streamlines traffic abstraction. The WiFi Alliance [12] suggests QoS translation/mapping to bolster low latency traffic across wired, WiFi and 5G networks. DSCP mapping is proposed to convert WiFi QoS (UP and WMM ACs) to 5QI. However, the proposal missed an important feature for LL services, which is effective AQM with granular congestion notification.
- End-to-end QoS layering (Figure 17): Employing application marking and MAC-to-transport layer congestion notification
    - Each layer, spanning from PHY and MAC to routing, transport and application, integrates a suite of features aimed at accommodating QoS metrics tailored to diverse traffic characteristics and requirements. The absence of interlayer visibility within this vertical dimension constitutes a barrier to automated traffic management and robust QoE support. The deployment of Low Latency Active Queue Management (LL AQM), paired with precise congestion notification and scalable transport protocols as outlined in IETF L4S specifications, seeks to remove this obstacle. Cablelabs DOCSIS 3.1 specifications encompass L4S support, while both Cable Modem (CM) and Cable Modem Termination System (CMTS) vendors have realized effective implementations. 3GPP Release 18 specifications include L4S ECN marking and API-driven exposure to radio congestion information, with ongoing initiatives to extend this support. Emerging proposals within IEEE 802.11 [13, 14] include L4S integration within the WiFi domain. Achieving end-to-

end L4S support holds the potential to exert tangible influence on QoE, given that bottlenecks may manifest across distinct network segments and temporal intervals. To illustrate, if WiFi acts as the bottleneck and only WMM is applied for L4S traffic, accommodating congestion dynamics proves challenging, potentially amplifying fluctuations and diminishing QoE for latency-sensitive, high-rate NQB applications.

- End-to-end QoS monitoring (Figure 18): Harmonizing Key Performance Indicators (KPIs) and telemetry platforms across the entire path and layering
    - Orchestrating traffic abstraction hinges on the efficient and timely collection and analysis of hop-by-hop and end-to-end metrics such as throughput, latency, jitter, packet loss and reliability metrics. Coordinated monitoring of latency and jitter at each network segment serves to discern and forecast the end-to-end state, complemented by scrutinizing additional network and service KPIs for detecting impairments and optimizing performance. This necessitates harmonized telemetry systems that span both dimensions of the network architecture.

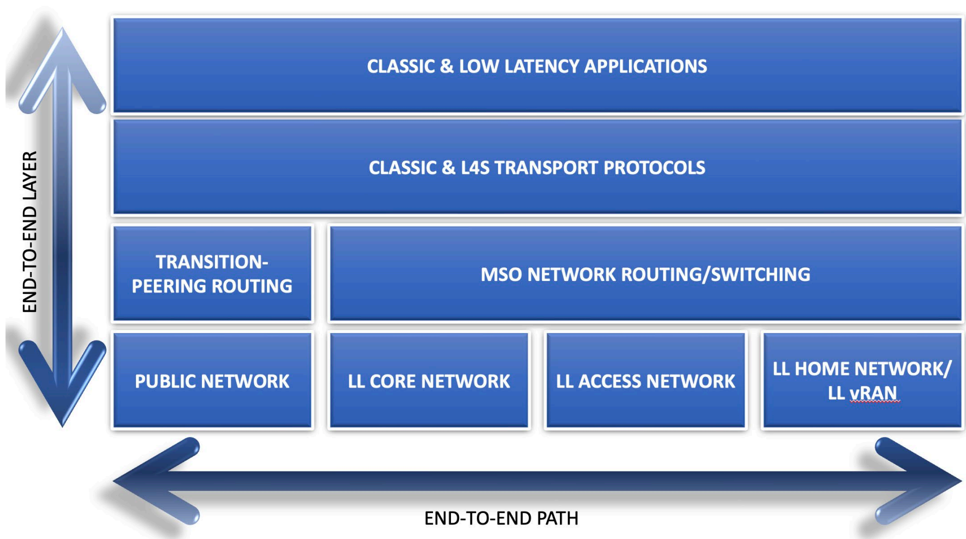

Figure 17- End-to-End Traffic Management

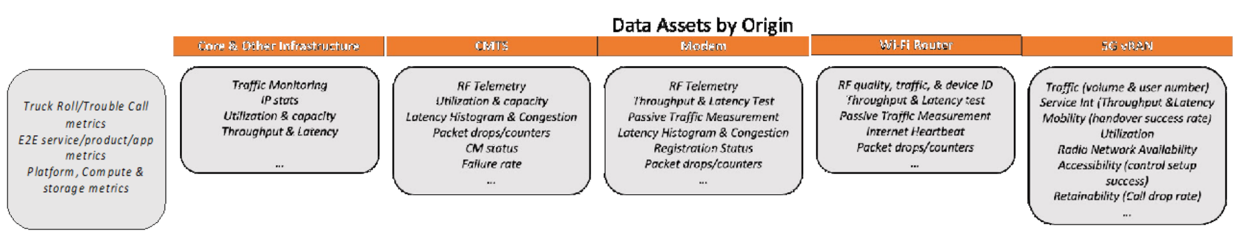

Figure 18- End-to-End and per hop QoS Monitoring: KPI examples

6. Conclusion

Distributed architectures incorporating SDN, NFV and cloudification empower MSOs to implement programmable and automated network and service platforms. However, while MSOs utilize a range of network technologies to expand the array of services available to subscribers, these platforms are currently not access technology agnostic (i.e. tailored for specific network such as HFC, PON or Mobile). Conversely, the diverse service landscape mandates a dynamic quality of experience, owing to the ever-evolving nature of residential, enterprise and industrial services.

Consequently, MSOs must embrace access-agnostic, converged platforms that can ensure optimal QoE for a diverse array of services. To address this concern, we delved into the issue by examining low-latency support within MSO networks. We specifically analyzed the viability of IETF L4S and 5G midhaul services over DOCSIS networks, showcasing that MSOs can effectively deploy such services using LLD specifications. This deployment is facilitated by the design of a traffic abstraction architecture, enhancing automation and programmability for efficient operations and a superior QoE.

Abbreviations

AQM	Active Queue Management
ASF	Aggregate Service Flow
CMTS	Cable Modem Termination System
CU	Centralized Unit
DOCSIS	Data Over Cable Service Interface Specification
DU	Distributed Unit
gNB	Next-Generation Node B (5G Base Station)
HFC	Hybrid Fiber-Coaxial
IAQM	Integrated Active Queue Management
IETF	Internet Engineering Task Force
ISP	Internet Service Provider
L4S	Low Latency, Low Loss, and Scalable Throughput
LL	Low Latency
LLD	Low Latency DOCSIS
MIB	Management Information Base
MSO	Multiple System Operators
NFV	Network Function Virtualization
NQB	Non-Queue-Building
ORAN	Open Radio Access Network
OS	Operating System
PHB	Per-Hop Behavior
PNF	Physical Network Function
PON	Passive Optical Networks
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RSSI	Received Signal Strength Indicator
rOLT	Remote Optical Line Terminal

RRU	Remote Radio Unit
SCTP	Stream Control Transmission Protocol
SDN	Software Defined Networking
SF	Service Flow
UP	User Priority
vCMTS	Virtual Cable Modem Termination System
vCU	Virtual Centralized Unit
vDU	Virtual Distributed Unit
VNF	Virtualized Network Function
vOLT	Virtual Optical Line Terminal
vRAN	Virtual Radio Access Network

Bibliography & References

[1] *Virtual RAN (V-RAN) Over DOCSIS: DOCSIS As A Midhaul*, Ramneek Bali, SCTE 2023

[2] https://www.bitag.org/documents/BITAG_latency_explained.pdf

[3] *A Virtual Broadband Network Gateway (vBNG) Approach for Cable Operators in a Distributed Access Environment*, Jason Combs, SCTE 2020

[4] *Configuring and Deploying Low Latency DOCSIS Networks*, Greg White and Karthik Sundaresan, SCTE 2021

[5] https://www.statista.com

[6] https://www.excentis.com/blog/latency-in-a-nutshell-and-what-makes-it-so-important/

[7] *Fastest Path to Low Latency Services How Can Cable Operators Deliver Consistent Latency by Following an Efficient and Future-Proof Design Path?*, Sebnem Ozer et. al., SCTE 2021

[8] *How VCMTS Paves The Way For 5G Over DOCSIS Exploring Software-centric Solutions for 5G Xhaul and FMC,* by Brendan Ryan et. al. , SCTE 2021

[9] The Future of Operations: Building a Data-Driven Strategy by Sebnem Ozer, Sinan Onder and Nagesh Nandiraju, SCTE Journal of Network Operations, 2019

[10] IETF RFC 9330 Low Latency, Low Loss, and Scalable Throughput (L4S) Internet Service: Architecture

[11] CableLabs DOCSIS Specification CM-SP-MULPIv3.1-I25-230419

[12] Analyzing operator QoS deployment scenarios between WiFi and 5G, WiFi Alliance, 2022

[13] https://mentor.ieee.org/802.11/dcn/23/11-23-0650-01-0uhr-qos-re-visited.pptx

[14] https://mentor.ieee.org/802.11/dcn/23/11-23-0679-00-0uhr-low-latency-qos-based-on-l4s.pptx

[15] https://www.fcc.gov/consumers/guides/getting-broadband-qa

Acknowledgments

The authors would like to thank and acknowledge all those who helped to make this paper possible. This paper includes several tests done by the lab supported by Philip Anderson, Lei Zhou, Jiaxin Zhao, Timothy Welch, Deependra Rawat, and Daniel Lynch.

