Blueprint for Mobile Xhaul over DOCSIS®

How Low Latency Xhaul (LLX) and Other Technologies Make DOCSIS an Ideal Solution for Mobile Xhaul

A Technical Paper prepared for SCTE•ISBE by

Jennifer Andreoli-Fang, PhD
Distinguished Technologist
CableLabs
Boulder, Colorado
j.fang@cablelabs.com

John T. Chapman
CTO Cable Access & Cisco Fellow
Cisco
San Jose, California
jchapman@cisco.com

Tong Liu, PhD
Principal Engineer
Cisco
San Jose, California
tongliu@cisco.com

Damian Poltz
Vice President, Technology Strategy & Networks
Shaw Communications
Calgary, Alberta
Damian.Poltz@sjrb.ca

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

List of Tables

Title **Page Number**

Introduction

One of the next big frontiers for the cable operators is to transition into becoming mobile operators and doing so in a cost-effective way that reutilizes their main assets – the hybrid-fiber coax (HFC) plant. Native DOCSIS technology provides a good starting point for cable operators to get into (backhaul, midhaul, and fronthaul) xhaul for mobile, but may not be enough to support the ultimate latency and timing requirements needed for future mobile traffic. The cable industry needs readily deployable technologies to take the DOCSIS network from a merely cost-effective xhaul network to a great high-performance solution, all without expensive upgrades and without moving to a higher cost fiber-only solution.

Low Latency Xhaul (LLX) is a part of a group of technologies that facilitate xhaul over DOCSIS. LLX is key to significantly reducing the upstream latency that mobile traffic may experience while traversing the DOCSIS link. Other companion technologies that are part of this technology suite are precision timing and synchronization, predictive granting, and centralized or distributed hierarchical QoS.

This paper starts with a baseline case of transporting mobile traffic over native DOCSIS technologies. It will then introduce four key technology areas that improve the performance and the cost of xhaul over DOCSIS networks. These technologies are:

- LLX, which utilizes the bandwidth report (BWR) message,
- DOCSIS distributed hierarchical QoS,
- DOCSIS predictive scheduling (DPS), and
- DOCSIS Time Protocol (DTP) for timing and synchronization.

Tradeoffs between performance and implementation complexity will be discussed.

The paper includes key results from a recent LLX trial conducted jointly with Shaw, Cisco, and Sercomm, and CableLabs. The results demonstrated that *LLX provides the latency performance for DOCSIS that is virtually indistinguishable from fiber, thereby enabling a host of low latency 5G applications such as ultra-reliable low-latency communication (URLLC), mobile gaming, and video conferencing, at a fraction of the cost of fiber installations.*

It is our hope that operators can use this paper as a high-level guideline to determine the right technologies for their xhaul needs.

Why Should Cable Operators Get into Mobile?

The answer is quite simple – it's all in the economics.

Cable operators have been successful at growing their fixed broadband offerings. But with an already high market penetration, it is unlikely they will see the same levels of unparalleled subscriber growth that they enjoyed. Entering the mobile market provides cable operators with new opportunities to grow subscribers and revenue.

Traditional cellular deployments have been in the form of high powered macrocells. However, new generations of mobile standards like LTE-A and 5G are able to offer even higher speeds and lower latency. Adding spectrum is certainly one way for new entrants into the wireless space to achieve the higher speed their customers desire, but comes at a cost. The alternative approach is to borrow a page from the cable industry's old trick of node splitting. In node spitting, when you run out of capacity, you split the area into a small area with less customers. This allows physical spectrum re-use. The most aggressive version of increasing capacity and decreasing node size has been the transition from N+5 to N+0 where one node is replaced by 15 to 20 nodes.

The equivalent technique to this in the mobile is to replace or supplement a macrocell with a substantial number of small cells to increase data capacity. Small cells are interesting because they enable higher frequency reuse due to their small coverage area. This means that instead of hundreds of users sharing a single block of spectrum on a macrocell, a magnitude fewer number of users get to share the same block of spectrum on a small cell, which results in higher data throughput per user.

Densification presents a challenge to mobile operators. All that infrastructure build require investments. For outdoor deployments, about 80% of total spending is attributed to backhaul, siting, and powering, while only 20% is spent on network equipment [9]. How can mobile operators scale their network an order of magnitude or more without doing the same to their economics?

The answer lies in the HFC plant. Simply put, cable operators own the infrastructure that can cover the 80% of the small cell deployment cost. By virtue of its large footprint as well as reduced siting, installation, and utility charges, this "gold in the ground" allows the cable operators to offer their customers wireless services at a significantly reduced cost [10]. Cable operators are increasingly realizing the increased revenue opportunity that comes with offering wireless services, which is why about half of CableLabs' members are already mobile operators and we anticipate that number to increase.

Having favorable economics creates the opportunity. Is the DOCSIS technology up for the challenge?

An Operator Experience with Mobile Backhaul over DOCSIS

HFC infrastructure provides many strategic advantages for wireless backhaul when compared to pure fiber backhaul or microwave/wireless backhaul approaches. These include:

1. **Ubiquity** – HFC networks run down every street and to every building in the city. This gives significant flexibility to wireless teams to design optimal small cell deployments with virtually no civil build requirements. For cable operators, fiber assets typically stop at the fiber node which can be hundreds of meters from the ideal small cell location. Coax, however, is typically within just a few meters of ideal outdoor small cell locations, making it the perfect choice for backhaul builds.
2. **Capacity** – Thanks to the DOCSIS technology, HFC networks can handle impressive broadband capacities. Leveraging DOCSIS 3.1 and mid-split, a typical HFC serving area has gigabits of downstream capacity and up to 500 Mbps of upstream. This makes it possible to serve small cells today with tiers such as 500 Mbps down / 100 Mbps up which maps very well to the throughput capabilities of existing mid-band based LTE small cells. In the future, as 5G and mm-wave technologies come to market, cable operators will be able to leverage DOCSIS 4.0 to provide multi-gigabit symmetrical backhaul over existing HFC infrastructure.
3. **Power** – One of the most notable advantages of HFC over fiber and wireless backhaul is its ability to transport power to small cells. Large passive optical network (PON) and fiber deployments typically eliminate the need for active field equipment. In order to deploy an

outdoor small cell in a PON or wireless topology, operators would typically need to obtain permitting, add temperature/weather hardened transformers, batteries, wiring, and complete significant electrical work. HFC, however, is an active architecture, meaning that power is typically transmitted down the coax cable to power amplifiers. This power is perfect for feeding small cell deployments, and often eliminates the need for onerous and complicated power upgrades.

4. **Speed & Simplicity** – HFC aerial architecture provides an ideal medium for fast small cell backhaul deployments. Attachment of a small cell to HFC "aerial strand" typically does not require any permitting, greatly reducing the time to build. A simple cut in coupler is needed to feed the small cell and backhaul modem/powering unit with data and power [14]. The small cell equipment is then attached the strand and the build is complete. All of this work typically takes just a few hours to complete.
5. **Automation** – Operators have invested heavily to enable automated provisioning and enablement over their DOCSIS infrastructure. These same systems can be leveraged to quickly and easily provision, turn-up and document new small cell deployments, often without any human intervention. This can further accelerate the speed of these deployments while also reducing manpower and waste.

The reasons outlined above should make choosing HFC for backhaul an easy decision for both multiple service operators (MSOs) and mobile network operators (MNOs). However, because of its capacity and dedicated wavelength properties, fiber is often touted as the ideal wireless transport technology. Shaw Communications wanted to further quantify the real-world benefits of using coax versus extending their fiber to feed their small cell deployments.

To quantify the difference in technologies, Shaw selected a high-density urban location in one of its Tier 1 markets. The location was a likely candidate for upcoming small cell deployments, and Shaw's wireless team had already provided ideal small cell deployment locations for the neighborhood. These locations were designed for their RF characteristics only and did not take into consideration proximity to HFC plant, power or fiber (see Figure 1 below).

Figure 1 – Ideal Small Cell Locations - Shaw Model

As designed, these 15 small cell locations fell across 13 node boundaries. This was important, as powering small cells from HFC infrastructure adds load to existing power systems. It was estimated that 1-2 small cells per node would not require power upgrades, which significantly reduces the cost and time of build. In this analysis, it was assumed that HFC power would be used to power small cells in both the fiber backhaul and coax backhaul scenarios.

The fiber wireline design team then designed a build to connect each of the ideal small cell locations to the nearest fiber location, which is typically the nearest fiber node (see Figure 2 below). It was assumed that 4x8 channel dense wavelength division multiplexing (DWDM) muxes with two wavelengths per small cell would be needed. It was also assumed that sufficient wavelengths or backbone fibers were available at the node location, eliminating the need for an overbuild. If overbuilds were required, the cost of the build would increase dramatically.

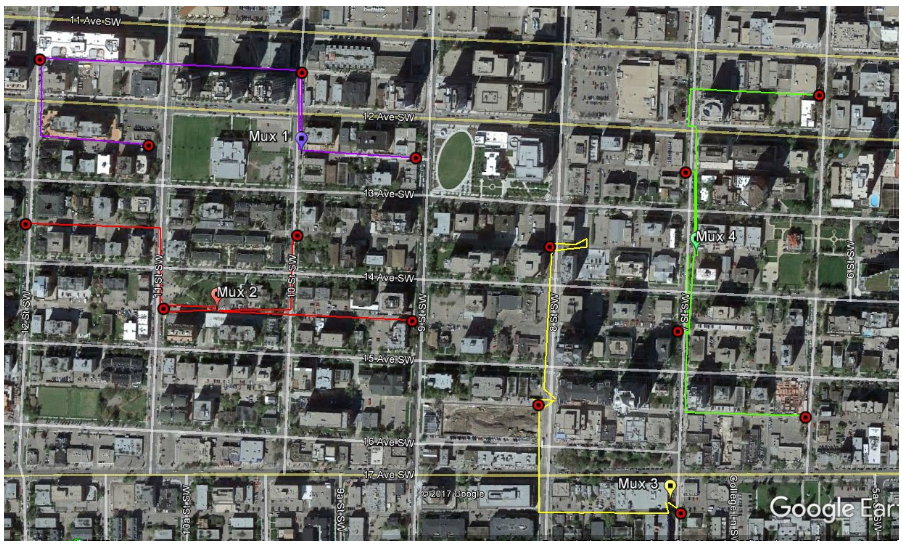


Figure 2 – Fiber Build Required for Backhaul - Shaw Model

Based on these assumptions, Shaw's plant design teams estimated a civil build cost of $182,500 and a build time of 4-6 months, with all required permitting and field work. And remember, this was already leveraging the fiber portion of the HFC plant.

The design team then completed a design that connected the small cells to existing coaxial infrastructure (see Figure 3 below). The small cells would be connected to existing coax through a simple cut-in coupler. They found that all the ideal small cell locations were within 10 meters of coax. As this was deemed "close enough" for small cell coverage, the need for any civil build was eliminated.

Figure 3 – Coax Build Required for Backhaul - Shaw Model

The design team estimated to total build cost in this scenario to be only $1,500 and time required to be one week. It's important to note this cost covers the civil build and components for the backhaul build only, and none of the small cell/backhaul electronic or labor to attach the small cells, which would have been the same in either scenario. Proximity of locations to existing coax dramatically reduced the cost of the build.

This simple design example demonstrates the strategic value of coax backhaul for MSOs and MNOs. (See Table 1 below).

Table 1 – Backhaul Build Comparison – Shaw Model

Small Cell Count	Backhaul Option	Backbone Fibers	Estimated Civil Build Cost	Estimated Build Time
15	DWDM	1	$183k	4-6 months
	Coax w/ couplers	0	$1.5k	1 week

In this simple model, when looking at backhaul civil build costs only, leveraging coax proved to be 100 times cheaper and 20 times faster than attempting to extend fiber to small cell locations (See Figure 4 below). It's important though to highlight again that these costs were estimates for civil build only and did not include any small cell or backhaul electronics, or multi-year total cost of ownership calculations. Adding those components would likely reduce this gap, however, is beyond the scope of this basic analysis.

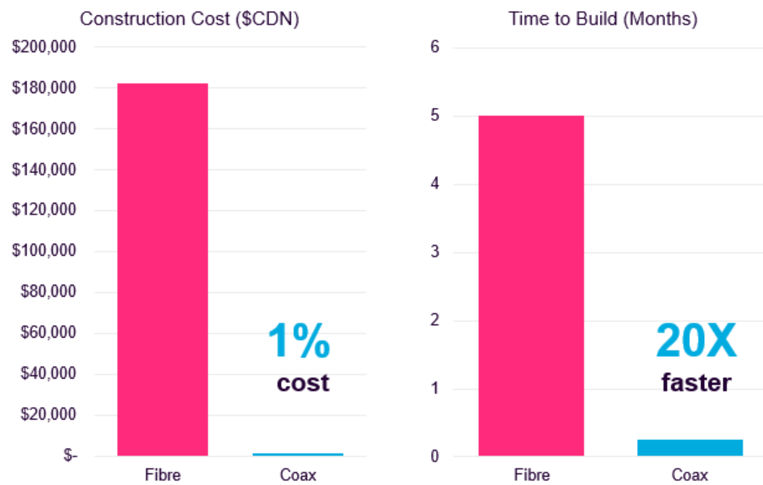


Figure 4 – Backhaul Build Comparison – Shaw Model

While DOCSIS infrastructure provides many strategic advantages for small cell backhaul, it does have a few key challenges. The first challenge is latency and jitter. As a shared medium, DOCSIS must schedule traffic across many endpoints. This can result in latency spikes during times of high utilization or contention. Furthermore, DOCSIS is challenged in serving the extreme bandwidth and latency requirements of protocols like Common Public Radio Interface (CPRI). In the sections below, we will address how these remaining challenges can be solved.

DOCSIS and Mobile Architectures

1. CMTS Architectures

The classic DOCSIS system consists of a cable modem termination system (CMTS) and a cable modem (CM). A CMTS system can be deployed with different options are shown in Figure 5. The CMTS scenarios are:

1. A traditional integrated CMTS,
2. a DAA system with a CMTS core (or flexible MAC architecture [FMA]) and a Remote PHY device (RPD), and
3. a DAA Remote MACPHY device (RMD).

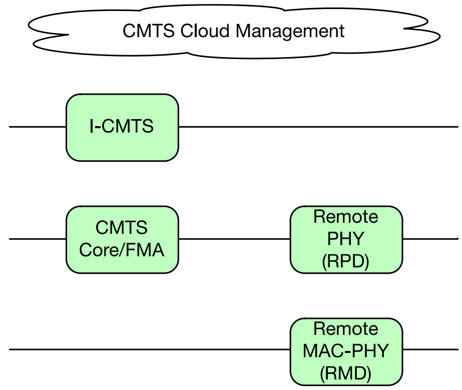


Figure 5 – CMTS Architectures

2. RAN Architectures

The small cell product that we have been referencing is known in the mobile specifications as a Evolved Node B (eNB) in Long Term Evolution (LTE) mobile network and a gNodeB (gNB) in fifth generation (5G) mobile network. The eNB or gNB can be further functionally decomposed into three components: a central unit (CU), a distributed unit (DU), and a radio unit (RU). Specific functions residing in each component are under discussion and standards organizations such as the O-RAN Alliance have defined a certain flavor for the lower layer split. The RU consists of the RF and all or a portion of the PHY, the CU consists of the PDCP layer and above, and the middle portion is the DU. A backhaul, midhaul, or fronthaul network, collectively known as xhaul, interconnects the different functional components together.

One may assume that xhaul is always done over a dedicated fiber. This is not always true. The connection of the two sides is often done over an IP network. In this paper, the transport is provided by a DOCSIS network and an IP backbone network. This is shown in Figure 6.

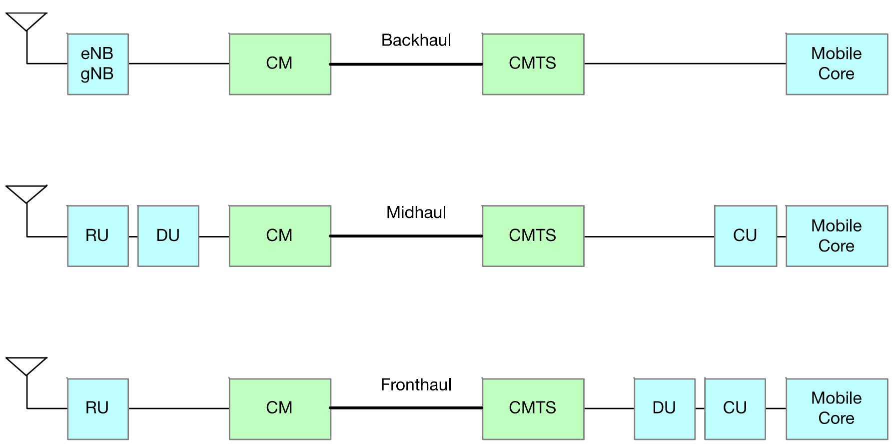


Figure 6 – Mobile Xhaul over DOCSIS Architectures

3. Mobile Backhaul Over DOCSIS with Native DOCSIS Techniques

To xhaul mobile traffic, the transport network must be capable of supporting the appropriate capacity, QoS, latency, synchronization and timing requirements for each xhaul scenario. In this section, we describe how QoS can be supported by native DOCSIS technology available today using the concepts of classifiers and service flows (SF). In the next section, we describe how Low Latency Xhaul (LLX) and other technologies can significantly improve the latency performance over native DOCSIS technology.

3.1. Traffic Types

Table 2 provides a comprehensive list of mobile traffic by type for LTE. 5G has similar traffic types but has renamed some of the interfaces. While this paper focuses on LTE, the same concept applies to 5G.

Mobile traffic to be backhauled by a DOCSIS network can be divided into user equipment (UE) traffic and radio access network (RAN) signaling. The RAN signaling includes S1-AP and X2 traffic that originates at the eNB. RAN signaling and latency-sensitive UE traffic require special handling on the DOCSIS network.

The UE traffic is carried on the S1-U interface. This is the user plane interface between the eNB and the serving gateway (S-GW) which is part of the evolved packet core (EPC). The UE traffic is further divided into UE signaling, IP Multimedia Subsystem (IMS) voice signaling and data, and UE data. One or more of a UE's application data flows are carried in a single bearer. A UE can have multiple bearers. In LTE, each bearer corresponding to a GTP-U tunnel. In 5G, all bearers of a UE are aggregated into a single GTP-U tunnel. In LTE, each bearer is assigned a QoS class indicator (QCI) during bearer establishment that is based on the type of traffic and qualification by the MNO. While this section focuses on LTE, the same concept is extended in 5G to 5G QoS indicator (5QI) [14].

The 3GPP specifies QoS characteristics, including latency and jitter, for a standardized set of QCI values along with a set of example services [13]. For example, IMS signaling is assigned QCI 1. To ensure QoS on the xhaul, the MNO extends the QCI table to include a DSCP value for each QCI. These QCI-to-DSCP mappings are configured on the eNB by the MNO to mark the IP header outside of the GTP-U header. The QCI-DSCP mappings are specific to each MNO and are not standardized.

The control plane interface (S1-MME) connects the eNB and the mobility management entity (MME) in the EPC. The S1 Application Protocol (S1-AP) runs over S1-MME. The S1-AP traffic is carried separately with the stream control transmission protocol (SCTP) and is not encapsulated in GTP-U tunnels. The interface is mainly used for bearer setup and release, paging, and handover related operations. A combination of SCTP protocol ID and IP destination address (IP DA) can be used to separate the S1-AP traffic from the rest.

The X2 interface connects eNBs. X2-AP handles the operations related to inter-eNB handover, dual connectivity, load management, etc. [9] The X2-U interface handles the transfer of user data related to handover, among other functionalities [12]. During inter-eNB handover procedure, both X2-AP and X2-U interface run simultaneously. For the procedure to run smoothly, both types of traffic need to be afforded a certain level of QoS [15] on the xhaul. A combination of SCTP protocol ID and IP DA can be used to separate the X2-AP traffic from the rest, while the MNO needs to configure the eNB to mark the X2-U traffic with a DSCP.

As a final step of the QoS provisioning, the mapping between the DSCP and the QoS characteristics, as well as the mapping between the IP DA / protocol ID and the QoS characteristics must be communicated to the CMTS. This is essentially the LLX Common QoS framework, where the two discreet systems align their common notion of QoS. Details are discussed in Section 5. The CMTS can then use these mappings to provision DOCSIS classifiers on the CM.

Table 2 – Mobile Traffic Carried Over DOCSIS Networks and Their QoS Support

<table>
<tr><th>Traffic Type</th><th>Traffic SubType</th><th>What Is It?</th><th>Traffic Identification</th><th>QoS Support by Native DOCSIS?</th><th>Low Latency Support by LLX?</th></tr>
<tr><td rowspan="2">Voice</td><td>IMS Voice</td><td>Voice data</td><td rowspan="5">Radio must mark traffic with DSCP.

Otherwise no.</td><td rowspan="5">DSCP to QoS mapping must be configured on CMTS</td><td>✓</td></tr>
<tr><td>IMS Signaling</td><td>Voice signaling</td><td>✓</td></tr>
<tr><td rowspan="2">End User Traffic (UE)</td><td>UE Signaling (S1-U)</td><td>UE configuration & reconfiguration (bearer setup, etc.)</td><td>✓</td></tr>
<tr><td>UE Data (S1-U)</td><td>Application data (mobile gaming, video conferencing, etc.)</td><td>✓</td></tr>
<tr><td rowspan="3">RAN Signaling</td><td>X2-U</td><td>User data transfer during HO, etc.</td><td>Potentially</td></tr>
<tr><td>X2-AP</td><td>Handover (HO), CoMP, etc.</td><td rowspan="2">SCTP, IP DA</td><td rowspan="2">IP DA to QoS mapping configured on CMTS</td><td>Potentially</td></tr>
<tr><td>S1-AP</td><td>RAN to mobile core signaling</td><td>Potentially</td></tr>
<tr><td>RAN Timing</td><td>PTPv2 / IEEE-1588</td><td>Timing for RAN</td><td>Downstream only.

May be tagged with DSCP</td><td>Should be given highest priority in converged interconnect network (CIN)</td><td>Supported by DTP (D3.1 or later)</td></tr>
</table>

3.2. QoS Model

Figure 7 shows the CM QoS configuration. The S1-AP traffic may be segregated and classified into a separate DOCSIS SF by classifying based on IP DA and SCTP. The X2-AP traffic may be segregated based on SCTP and IP DA, while the X2-U traffic must be marked by the radio with special DSCP. For the UE traffic that requires low latency, it is up to the radio to mark this traffic with special DSCP(s). For the remaining UE traffic, the radio can either mark them with separate DSCPs or leave them unmarked.

Due to the latency-sensitive nature of the signaling traffic, S1-AP, X2-AP, X2-U, and low latency UE traffic should be served with a DOCSIS scheduling service that is not best effort (BE). However, since these traffic can be bursty and unpredictable in nature, real-time polling service (RTPS) or proactive grant service (PGS) could be a good compromise in terms of providing latency performance and bandwidth efficiency.

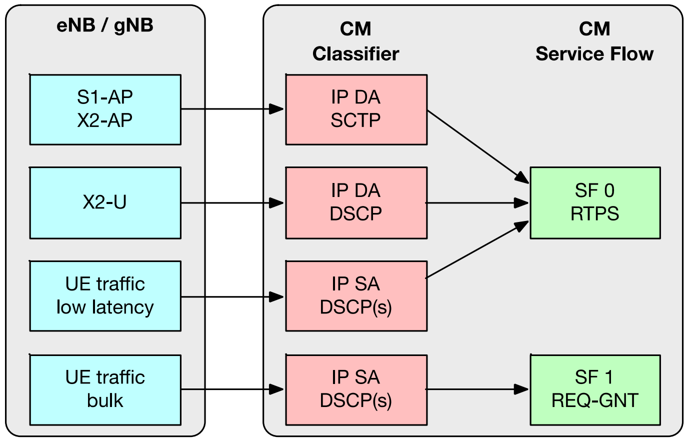


Figure 7 – CM Configuration Using Native DOCSIS Upsteam QoS

Native DOCSIS can provide QoS differentiations for mobile traffic. Now the question is, can latency on the DOCSIS network be improved to support mobile traffic, with minimal disruption to the mobile and DOCSIS technologies? The answer is yes, using the Low Latency Xhaul (LLX) technology.

Low Latency Xhaul (LLX)

LLX enables a range of use cases for cable operators. For a cable operator that provides wholesale xhaul transport for another mobile operator, or a converged cable operator who also owns both cable and mobile operations and xhauls its own mobile traffic, LLX provides a means to significantly lower the latency of all traffic coming from the UE to the level that is comparable to fiber. These traffic include signaling, IMS voice (data and signaling), low latency applications such as mobile gaming, video conferencing applications such as Zoom, WebEx and FaceTime, and URLLC for 5G.

LLX has the following fundamental components:

1. scheduler pipelining using the BWR message, and
2. a common QoS framework that matches the DOCSIS QoS to the mobile network QoS.

In addition, there needs to be system level configuration and operation to align the DOCSIS and mobile systems.

4. LLX Scheduler Pipelining

Scheduler pipelining is a very unique and inventive aspect of LLX and the heart of what creates a low latency transport. LLX uses the decisions made by the mobile scheduler to inform the CMTS what is about to happen next.

So how does that work and what are the results? Let’s start with a simple example of a 4G/LTE eNB that is backhauled over a DOCSIS network. This is shown in Figure 8.

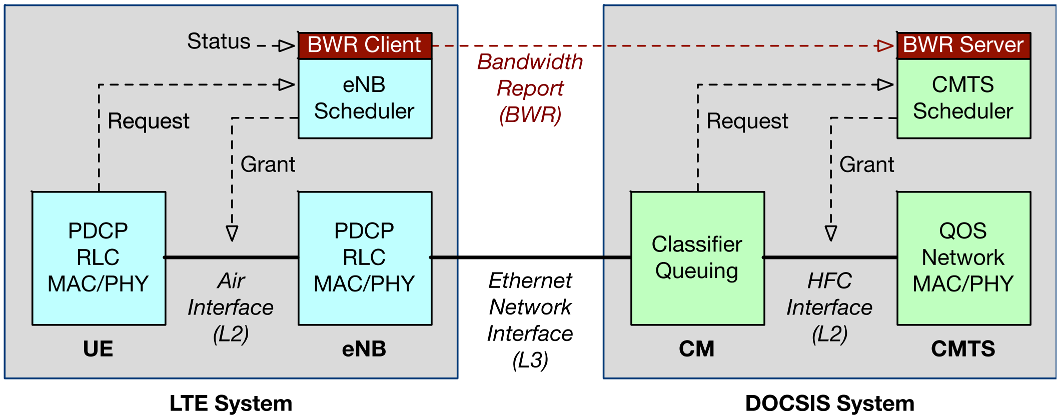


Figure 8 – Linking DOCSIS and LTE Schedulers with BWR

The first thing to notice is how similar the LTE and DOCSIS systems are. Specifically, both systems are multi-point to point in the upstream and both systems have upstream paths that are centrally scheduled. This means that both systems have an inherent latency due to the request-grant delay in the upstream. In LLX, that latency is incurred once in the mobile system. The results of the request-grant process, in the form of a BWR message, are then passed to the DOCSIS system so the CMTS can grant the CM directly without waiting for a native layer 2 DOCSIS request.

Let's look at an example.

1. The UE has an application that wants to send 1000 bytes. It sends a request to the eNB scheduler.
2. The eNB schedular responds and says that the UE may sent the 1000 bytes 8 ms from a reference time.
   - The eNB scheduler, now that it knows what is about to transpire on its air interface, makes a determination of what will happen across the network interface that is shares with the DOCSIS system. In our example, the eNB adds 12 ms of engineering margin to cover any buffering and internal path delays.
3. The eNB sends a BWR message to the CMTS system that says that 1000 bytes will be arriving on the shared network port 9 ms from the reference time.
   - The CMTS scheduler, now that it knows when the bytes will arrive in the CM, determines when it wants to send a grant to that CM. In this example, it adds 1 ms of engineering margin to cover any buffering or scheduling jitter.
4. The CMTS sends a DOCSIS MAP to the CM at the correct time telling the CM to transmit the 1000 bytes 10 ms from the reference time.

The BWR message is essentially a layer 3 request message that replaces the native layer 2 request message in the CM. The concept is shown in Figure 9. Without scheduler pipelining, there are two independent request-grant cycles, one in mobile and one in DOCSIS. With scheduler pipelining, the CM layer 2 request is replaced with the eNB layer 3 request (BWR). The two requests happen in rapid succession and then the two grants occur in rapid succession.

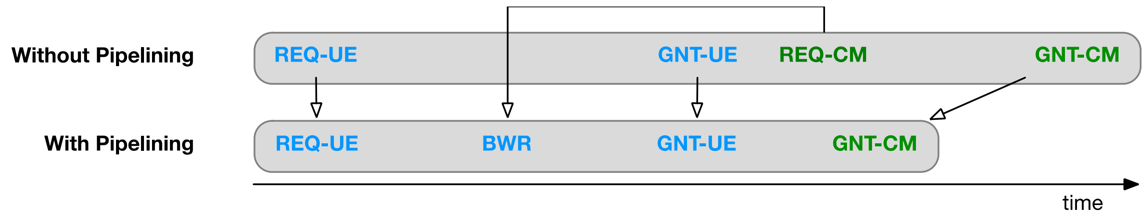


Figure 9 – Pipelining Requests and Grants

The net result is that the latency of the DOCSIS system is effectively reduced by hiding it under the LTE system. This is illustrated in Figure 10.

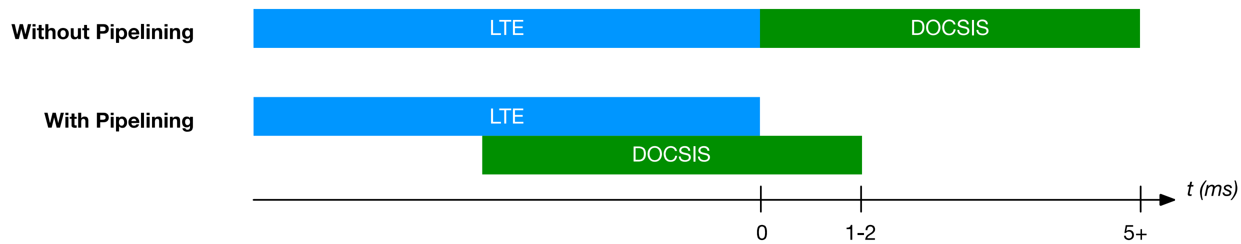


Figure 10 – Reducing DOCSIS Effective Latency Under LTE

This concept can be applied to midhaul and fronthaul system as well where the eNB has been replaced with separate components, namely:

- a radio unit (RU) that contains a portion of the PHY and generates the data stream over the network interface,
- a distributed unit (DU) that contains the scheduler and encapsulation, and
- a control unit (CU) that contains software that manages the RU and CU.

The transmission system can also be either a DOCSIS system of a PON system. In O-RAN Alliance terminology, the elements would be:

- a transport unit (TU) that receives the packets from the RU. In DOCSIS, this would be the CM. In a PON system, this would be the optical network unit (ONU). A TU can be an ONU from a PON system or a CM from a DOCSIS system; and
- a transport node (TN) that receives the packets from the TU and contains the scheduler for the transmission system. A TN can be an OLT from a PON system or a CMTS from a DOCSIS system.

These network elements are shown in Figure 11.

- For backhaul, the access entity (AE) could be an LTE eNB or a 5G NR gNB. The scheduler would be contained locally in the AE and the BWR signaling traffic would travel over the transport network.
- For midhaul, the CU has been centralized. This case would be similar to backhaul from a LLX viewpoint.
- For fronthaul, the CU and DU are moved centrally. The scheduler is now centralized and the BWR packets do not traverse the same network that the data packets do.

Note that in the fronthaul case, the transmission time for the BWR message form the BWR client to the BWR server is typically much less compared to backhaul. This is good as the fronthaul case is usually associated with more stringent latency requirements.

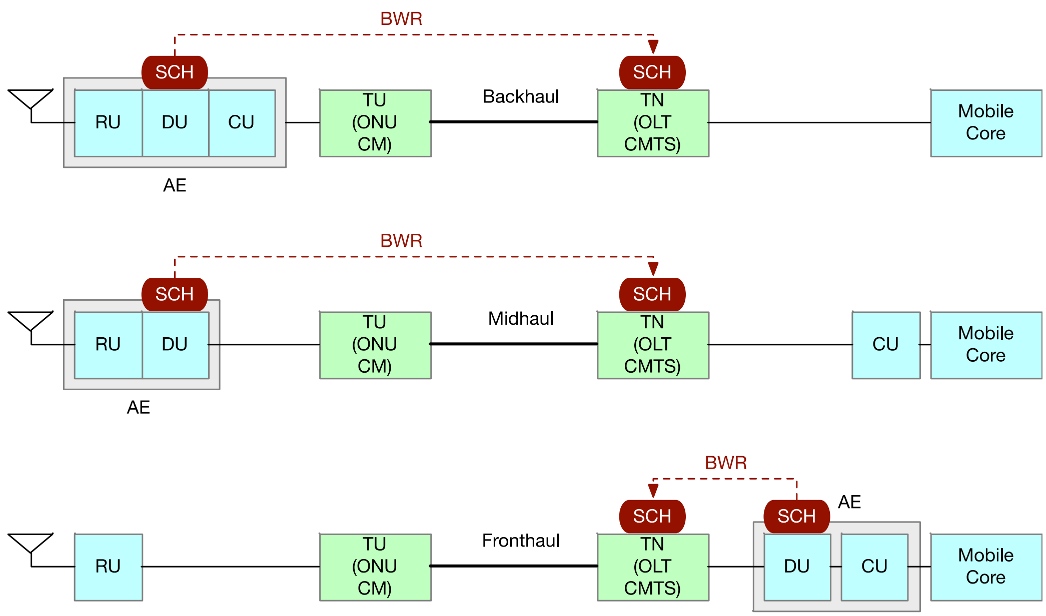


Figure 11 – Mobile Xhaul over DOCSIS Using BWR

What we have established at this point is that we can take a flow of bytes packets that are crossing the mobile air interface and move those across the transport network with lower latency. Next, we are going to go into the detail at another level and see what happens when that stream of bytes and packets may be composed of multiple flows with multiple queues and multiple scheduling mechanisms.

5. LLX Common QoS Framework

If all traffic move across an interface together, then all high priority traffic, such as signaling, and low priority traffic, such as a file transfer, would experience the same latency. If the latency of that interface was infinitely low and the bandwidth was infinitely high, then there would not be a problem.

But that is rarely the case. The mobile bandwidth is limited as is the bandwidth of the transport network. In such systems, the file transfers can saturate a system, causing buffers to fill up and system latency to increase. For this fundamental reason, it is common to separate traffic into multiple flows that are sorted or classified into multiple queues. This is a fundamental property of QoS. With a QoS mechanism, services such as signaling or 5G URLLC can be provided to have the absolute minimum latency.

The air interface for LTE has four layer 2 request queues known as logical channel groups (LCGs). The UE requests per LCG, the eNB grants based on the requested bytes per LCG, and each grant result is tracked separately within the BWR message as a **BWR flow**. In 5G, the number of LCGs per UE expands from 4 to 8 for even more granular QoS treatment. Ironically, when in the 5G fronthaul cause, these eight LCGs are combined together into one eCPRI flow, although there is work underway to separate some of the traffic out of the main eCPRI flows.

Consider the following scenario: all the flows from the mobile data plan are aggregated together onto a common Ethernet and passed to the transport network. How does the transport network disaggregate the flows and put the right flows on the right queues? How are these queues associated with the BWR flows? How many grants should each queue get?

To address these concerns, LLX proposes a common QoS framework between the mobile system and the transport system. There are many variations of how this could be done, but the most fundamental and pure system is to do the following:

1. Use the same number of queues in the transport system as there is in the mobile system
2. Use the same classifier mechanism in the transport system as there is in the mobile system
3. Use the same policy/queue-weighting mechanism in the transport system as there is in the mobile system

Let's look at this with our original LTE and DOCSIS backhaul example as shown in Figure 12. Let's follow a packet through this system.

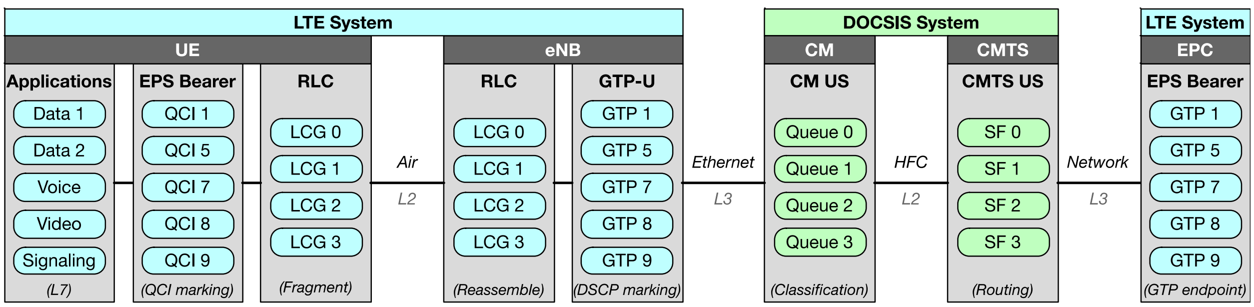


Figure 12 – QoS Model of LTE System with DOCSIS Xhaul

An application in the UE creates a packet. That application has an associated QCI. The packet and the QCI marking are placed into a radio bearer and into a logical channel. The logic channels are placed into one of four LCGs. Each LCG has its own request-grant exchange with the eNB and its own flow entry in the BWR message. When the packets are received at the eNB, they are placed into GTP. The GTP packet is market with a DSCP that is chosen based on the QCI.

When the packets arrive at the DOCSIS system, they need to be classified into different service flows. If the CM is provided with the correct DSCP to service flow mapping, the contents of the original four LCGs can be recreated on the four DOCSIS service flows. This is shown in Figure 12.

Other options are available. Even though the four LCGs are reported as four separate flows in the BWR message, at a system level, both the request information in the BWR message and the classifiers in the CM could be combined in different ways. An example of this is shown in Figure 13.

In both of these examples, the signaling traffic is separated from the data traffic. The signaling traffic is comprised of both native eNB signaling traffic and UE signaling traffic. In the non-BWR example, the signaling traffic is placed onto a DOCSIS RTPS flow; the data traffic is placed onto a regular request-grant flow. The advantage of the RTPS slow is that request slots are dedicated and there are no lost request which eliminates long tail latency. In the BWR example, the native eNB signaling is left on an RTPS flow assuming that it truly real-time and may not always be accurately predictable. All UE traffic, including signaling and low latency application flow, however, is predictable since it is granted by the eNB and described by a BWR flow. So it is placed into its own DOCSIS service flow. The bulk traffic is then placed in a second service flow. Both service flows are managed using BWR. Of course, there could be multiple low latency application flows and multiple bulk flows to further differentiate the QoS among the flows.

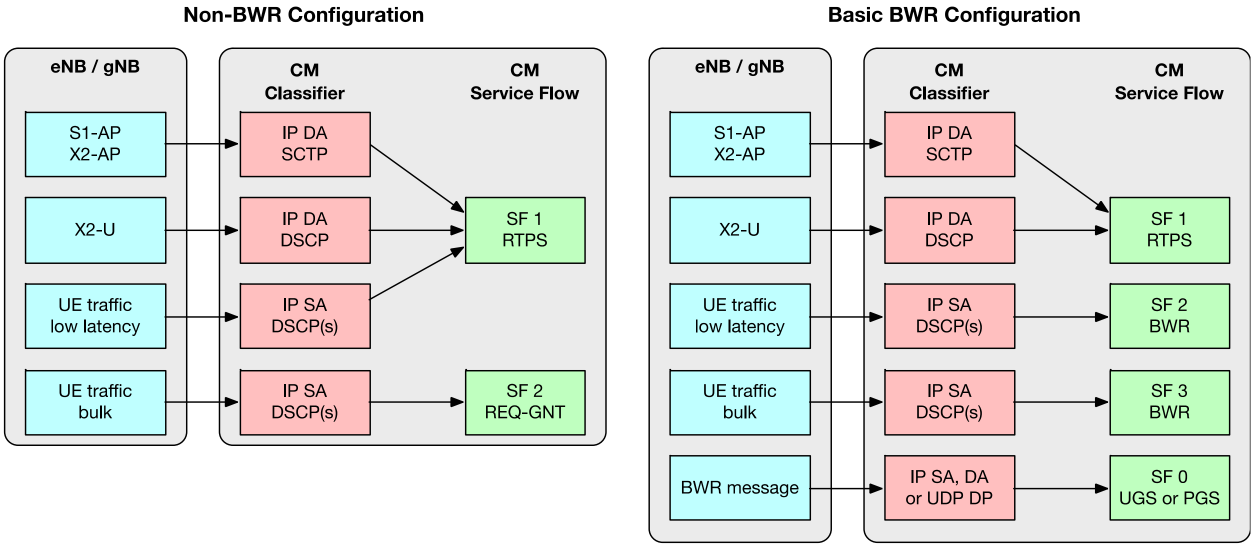


Figure 13 – CM Configuration Using Native DOCSIS QoS (Left) and Using BWR (Right)

6. LLX Performance

The LLX project started as a joint development project between CableLabs and Cisco. As part of the project, CableLabs and Cisco built a proof of concept testbed using a Cisco cBR-8 CMTS and the Open Air Interface (OAI) LTE RAN platform. We have previously published numerous test results with this test setup and reported that BWR achieves 1-2 ms of DOCSIS US latency with a low to medium channel loading on the DOCSIS network ([1][3][4][5]).

6.1. Shaw BWR Trial

In July 2019, Shaw, Cisco, Sercomm, and CableLabs jointly conducted a BWR trial. The LTE portion of the network used a Sercomm CBRS F208 small cell with BWR software that gathers the LTE scheduler outputs and generates the BWR messages. The DOCSIS portion of the network used a Cisco cloud native broadband router (cnBR) with a BWR API that receives and interprets the BWR messages and translates them into native DOCSIS REQs as input to the existing DOCSIS scheduler.

Figure 14 shows the test setup. Table 3 and Table 4 show the configuration parameters.

Although the BWR message can describe 4 or 8 BWR flows corresponding to LTE or 5G systems, to simplify testing, only 1 BWR flow is instantiated and tested.

For the BWR flow, the UE was configured to send 500-byte packets every 50 ms. The tests were conducted with four different DOCSIS channel utilizations (see Table 3) to assess the performance of BWR under various deployment conditions. The channel utilization was achieved by sending traffic via two background CMs configured with one or multiple background DOCSIS SFs (BG-SF). For medium and high utilization, the US channel loading were achieved with two separate setups: one BG-SF and 20 BG-SFs. In the latter case, background loading was distributed over the 20 BG-SFs, and is designed to test the scenarios where channel loading comes from a number of users rather than just one. For the high utilization case, an additional pair of tests were run to give higher scheduling priority to the BWR flow.

For each of the eight sets of test cases, DOCSIS US latency experienced by the BWR flow was measured between CM 1 and cnBR, with BWR disabled and enabled.

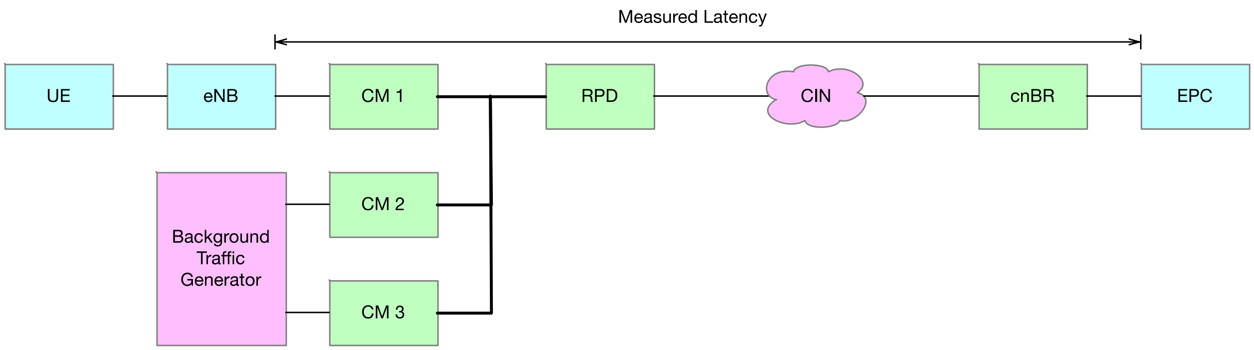


Figure 14 – Shaw-Cisco-Sercomm-CableLabs Trial Setup

Table 3 – DOCSIS Parameters

DOCSIS parameters	
Number of CMs	1 CM performing LTE backhaul (CM 1) 2 CMs generating background traffic (CM 2, CM 3)
Channel configurations	1 DOCSIS 3.0 US ATDMA channel 1 DOCSIS 3.0 DS SC-QAM channel MAP interval = 1 ms MAP advance time = 3 ms CIN delay = 256 µs (25 km plant)
Number of background service flows (BG-SF)	0 for no load 1 for low load 1 for medium load 20 for medium load, split between CM 2 and CM 3 1 for high load 20 for high load, split between CM 2 and CM 3
Channel utilization due to background load	0 for no load 20% for low load 40% for medium load 70% for high load
Scheduling service for BWR messages	UGS Grant interval = 1 ms
Scheduling service for BWR flow & background traffic	Best effort, for all flows
Scheduling priority	Default is all SFs have same priority. 2 additional tests were run with CM 1's backhauled traffic with higher priority compared to BG-SF.

Table 4 – LTE Parameters

LTE parameters	
LTE channel bandwidth	20 MHz
Duplexing method	TDD
Spatial multiplexing	SISO
HARQ	ON (10% BLER)
Number of UE	1
BWR periodicity	2 ms
Traffic rate	500 bytes, every 50 ms

6.2. Trial Results and Key Conclusions

Table 5 shows the latency experienced by the BWR flow for each test case (TC). Figure 15 shows the cumulative distribution function (CDF) plots of the results.

Summary of findings:

- BWR significantly reduces the latency experienced by the BWR flow in all scenarios, regardless of channel utilization on the DOCSIS network.
- With the exception of TC403 (high load, 20 BG-SFs, same priority of BWR flow and BG-SFs), BWR consistently achieves 95th percentile latency of less than 5-6 ms.
- When the channel utilization is high, with the same DOCSIS traffic priority, BWR effectively reduces the 95th percentile latency to 1/3 when compared to native DOCSIS, as shown in TC405.
- When the channel utilization is high and a large number of users trying to access the channel, BWR ensures a 1-2 ms latency with a higher DOCSIS traffic priority applied to the BWR flow, as shown in TC407. At the 95th percentile, BWR reduces DOCSIS upstream latency almost an order of magnitude, from 22 ms to 2.5 ms in TC407.

Absolute latency achievable by BWR: The trial was conducted on the DOCSIS 3.0 ATDMA channel. With DOCSIS 3.1 OFDMA US, granting resolution can be reduced to sub-millisecond range. Thus, the absolute latency achievable by BWR can be reduced significantly further.

Higher scheduling priority: In TC403, equal priority was given to the background traffic and the BWR flows when they were granted. The BWR message itself, however, does not have to go through a contention channel whereas regular layer 2 requests do. This results in BWR flows have lower latency.

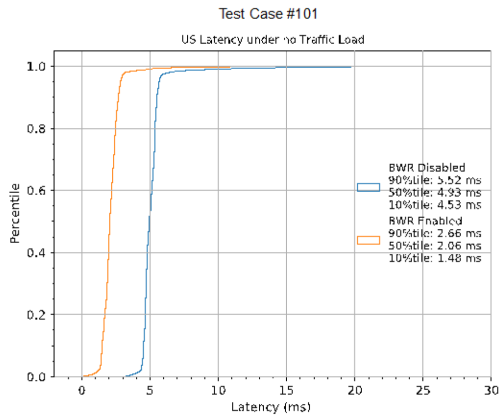
In TC407, BWR flows where given higher priority than background traffic at the time of granting. This, when combined with the more reliable BWR requesting flow, results in a consistent 2 ms observed upstream latency up to the 95th percentile.

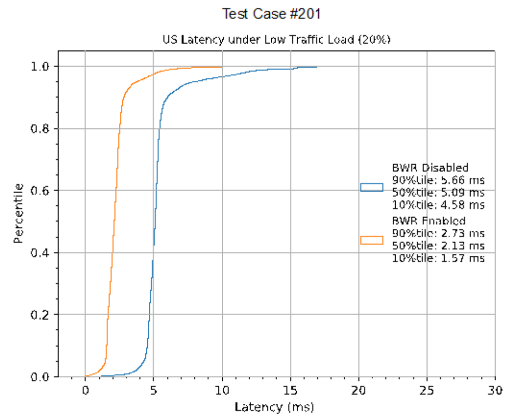
Impact of bandwidth efficiency: Throwing extra bandwidth at a flow with a scheduling service such as PGS helps reduce latency. But at what cost? The efficiency of the BWR approach is what makes it an attractive solution. BWR provides a way for the CMTS scheduler to predict almost exactly the amount of traffic to arrive at the CM at a precise time in the future. BWR messages themselves incur very little grant overhead.

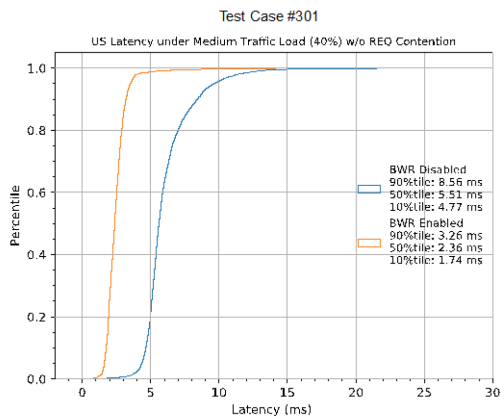
Impact of CIN delay: The trial used a 25 km CIN which is in-line with the Shaw network. BWR works with larger CIN delay. If the CIN is expected to be large, such as several hundred miles, BWR can reduce the latency using a combination of CMTS grant tracking and prediction, at a slightly reduced efficiency.

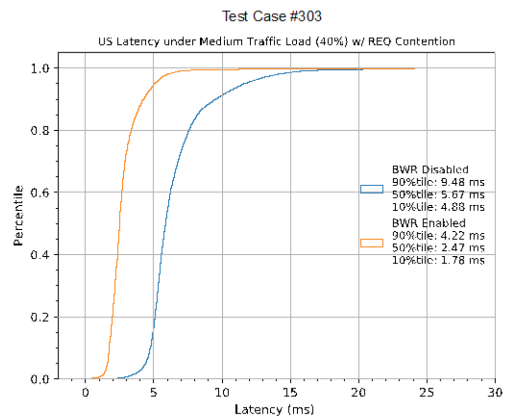
Table 5 – Test Cases and Results

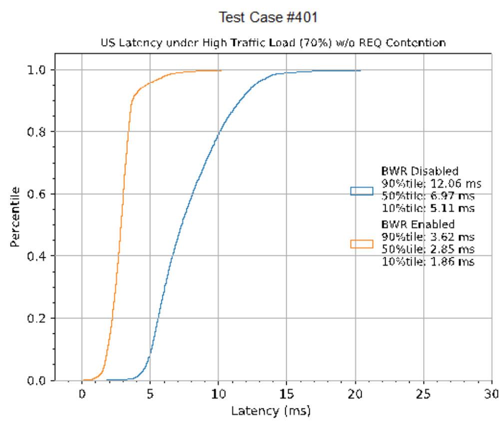
Test Case #	Test Case	BWR	10^{th} Percentile	50^{th} Percentile	95^{th} Percentile
101	No load	Off	4.53	4.93	5.61
		On	1.48	2.06	2.82
201	Low load (20%)	Off	4.58	5.09	7.34
		On	1.57	2.13	3.5
301	Medium load, 1 BG-SF	Off	4.77	5.51	10.09
		On	1.74	2.36	3.51
303	Medium load, 20 BG-SFs	Off	4.88	5.67	12.57
		On	1.78	2.47	5.37
401	High load, 1 BG-SF	Off	5.11	6.97	13.06
		On	1.86	2.85	4.65
403	High load, 20 BG-SFs	Off	5.86	8.03	21.61
		On	1.88	3.34	7.5
405	High load, 1 BG-SF, higher priority	Off	4.64	6.66	12.91
		On	1.59	2.02	2.64
407	High load, 20 BG-SFs, higher priority	Off	4.8	6.92	21.93
		On	1.6	2.05	2.58

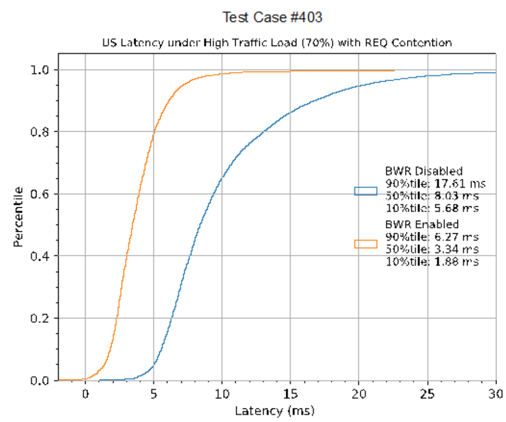

Test Case #101
US Latency under no Traffic Load
BWR Disabled
90%tile: 5.52 ms
50%tile: 4.93 ms
10%tile: 4.53 ms
BWR Enabled
90%tile: 2.66 ms
50%tile: 2.06 ms
10%tile: 1.48 ms
Percentile
Latency (ms)

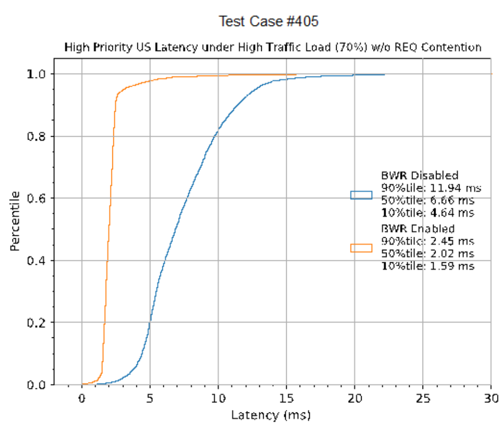

Test Case #201
US Latency under Low Traffic Load (20%)
BWR Disabled
90%tile: 5.66 ms
50%tile: 5.09 ms
10%tile: 4.58 ms
BWR Enabled
90%tile: 2.73 ms
50%tile: 2.13 ms
10%tile: 1.57 ms
Percentile
Latency (ms)


Test Case #301
US Latency under Medium Traffic Load (40%) w/o REQ Contention
BWR Disabled
90%tile: 8.56 ms
50%tile: 5.51 ms
10%tile: 4.77 ms
BWR Enabled
90%tile: 3.26 ms
50%tile: 2.36 ms
10%tile: 1.74 ms
Percentile
Latency (ms)


Test Case #303
US Latency under Medium Traffic Load (40%) w/ REQ Contention
BWR Disabled
90%tile: 9.48 ms
50%tile: 5.67 ms
10%tile: 4.88 ms
BWR Enabled
90%tile: 4.22 ms
50%tile: 2.47 ms
10%tile: 1.78 ms
Percentile
Latency (ms)


Test Case #401
US Latency under High Traffic Load (70%) w/o REQ Contention
BWR Disabled
90%tile: 12.06 ms
50%tile: 6.97 ms
10%tile: 5.11 ms
BWR Enabled
90%tile: 3.62 ms
50%tile: 2.85 ms
10%tile: 1.86 ms
Percentile
Latency (ms)


Test Case #403
US Latency under High Traffic Load (70%) with REQ Contention
BWR Disabled
90%tile: 17.61 ms
50%tile: 8.03 ms
10%tile: 5.68 ms
BWR Enabled
90%tile: 6.27 ms
50%tile: 3.34 ms
10%tile: 1.88 ms
Percentile
Latency (ms)


Test Case #405
High Priority US Latency under High Traffic Load (70%) w/o REQ Contention
BWR Disabled
90%tile: 11.94 ms
50%tile: 6.66 ms
10%tile: 4.64 ms
BWR Enabled
90%tile: 2.45 ms
50%tile: 2.02 ms
10%tile: 1.59 ms
Percentile
Latency (ms)

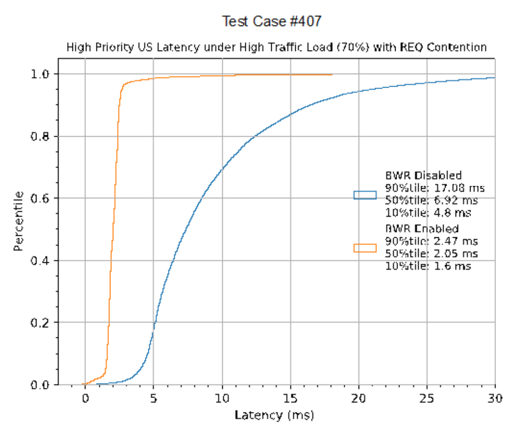

Test Case #407
High Priority US Latency under High Traffic Load (70%) with REQ Contention
BWR Disabled
90%tile: 17.08 ms
50%tile: 6.92 ms
10%tile: 4.8 ms
BWR Enabled
90%tile: 2.47 ms
50%tile: 2.05 ms
10%tile: 1.6 ms
Percentile
Latency (ms)

Figure 15 – Trial Results in CDF

DOCSIS Scheduling Optimizations

7. Distributed Hierarchical QoS

This paper proposes a new distributed Hierarchical Quality of Service (HQoS) mechanism that builds upon the existing centralized HQoS system that exists today.

HQoS provides an intermediate level in the scheduling hierarchy between DOCSIS SFs and channels / bonding groups and introduces aggregate QoS treatment for an aggregation of SFs. Distributed HQoS is a technique designed to improve the DOCSIS upstream latency and bandwidth efficiency by enabling the CM to schedule bandwidth among individual service flows within the aggregate QoS envelope enforced by the CMTS.

HQoS is defined to extend the traditional flat QoS model into a two-layered QoS hierarchy so that the operator could define QoS policies on an aggregation of service flows using the aggregated service flow (ASF) construct. Specifically, the root bandwidth resource, provided by a bonding group or a RF channel, is first scheduled among ASFs and non-ASF individual service flows at the top layer; then the bandwidth resource allocated to each ASF is further scheduled among its member SFs for a more granular QoS control.

DOCSIS HQoS today is centrally enforced by the CMTS, both at the inter-ASF and intra-ASF level. From the latency and bandwidth sharing efficiency point of view, this centralized QoS scheme does not offer any improvement compared to the traditional flat QoS model. Since the packet queues are physically located at the CM, the intra-ASF scheduling decision that is made at the CMTS has to rely on the queue depth information sent by the CM via the DOCSIS request messages. This implies a built-in request-grant delay for any bandwidth sharing among the CM packet queues. Additionally, since the grant is associated with an individual SF, any unused grant has to be wasted, as the CM cannot apply it to other SFs even when there are packets waiting to be sent in another SF.

Distributed HQoS on the other hand, turns the HQoS into a low latency scheduling technique by distributing the lower-layer, intra-ASF scheduling down to the CM. As shown in Figure 16, the CMTS enforces the top layer QoS among ASFs and any non-ASF individual service flows. The CM then enforces the intra-ASF scheduling policy to provide more granular QoS among the member SFs within the ASF. In this construct, the request and grant interface between the CMTS and the CM is at the ASF level, and the CM's intra-ASF scheduler decides which packet queue to serve based on the scheduling policy. This technique eliminates the request-grant delay and grant wastage as experienced in the traditional flat QoS model or with the centralized HQoS scheme.

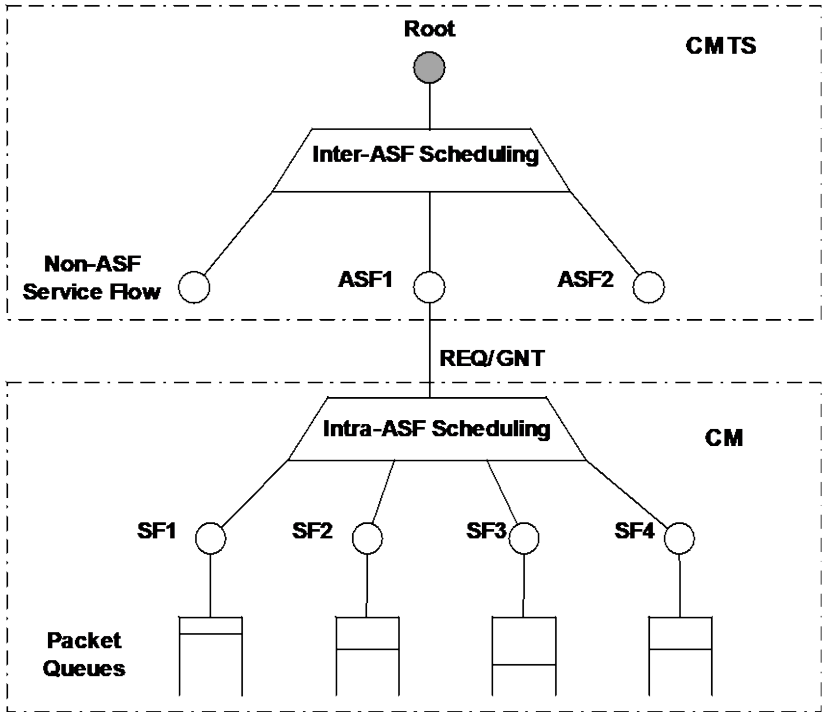


Figure 16 – Distributed Hierarchical QoS for Low Latency and High Bandwidth Efficiency

Figure 17 and Figure 18 provide the lab test results showing the latency and efficiency improvement with the distributed HQoS scheme compared to the centralized scheduling scheme. In this test, a single CM is attached to the CMTS via a DOCSIS 3.0 upstream channel. Two upstream traffic streams were sent through the CM. The first stream was a high priority traffic stream at 1 Mbps that was then classified to a low latency SF. The second stream was a low priority traffic stream (emulating a speed test) at 10 Mbps that was classified to a best effort SF.

On the CMTS side, proactive granting scheduling (PGS) was used to generate grants at one 1500 byte maximum transmit unit (MTU) size every 1 ms, equivalent to a grant rate of 12 Mbps. In the centralized scheme, PGS grants were applied to the low latency SF only. In the distributed scheme, PGS grants were shared between the low latency SF and the best effort SF based on priority.

Figure 17 shows the latency and the best effort traffic latency test results in microseconds. The distributed scheme shows significant latency improvement especially for the best effort traffic as the best effort traffic was able to share the remaining PGS grants after the 1 Mbps low latency traffic was served.

In contrast, the centralized scheme where the PGS grants were not sharable, the best effort traffic had to go through the request-grant delay to access the upstream. This resulting in longer queueing delays and lower channel efficiency. As the channel utilization reached 90% of the 25 Mbps channel capacity, with half of it (12 Mbps) was reserved for PGS for low latency traffic but most of those grants went unused.

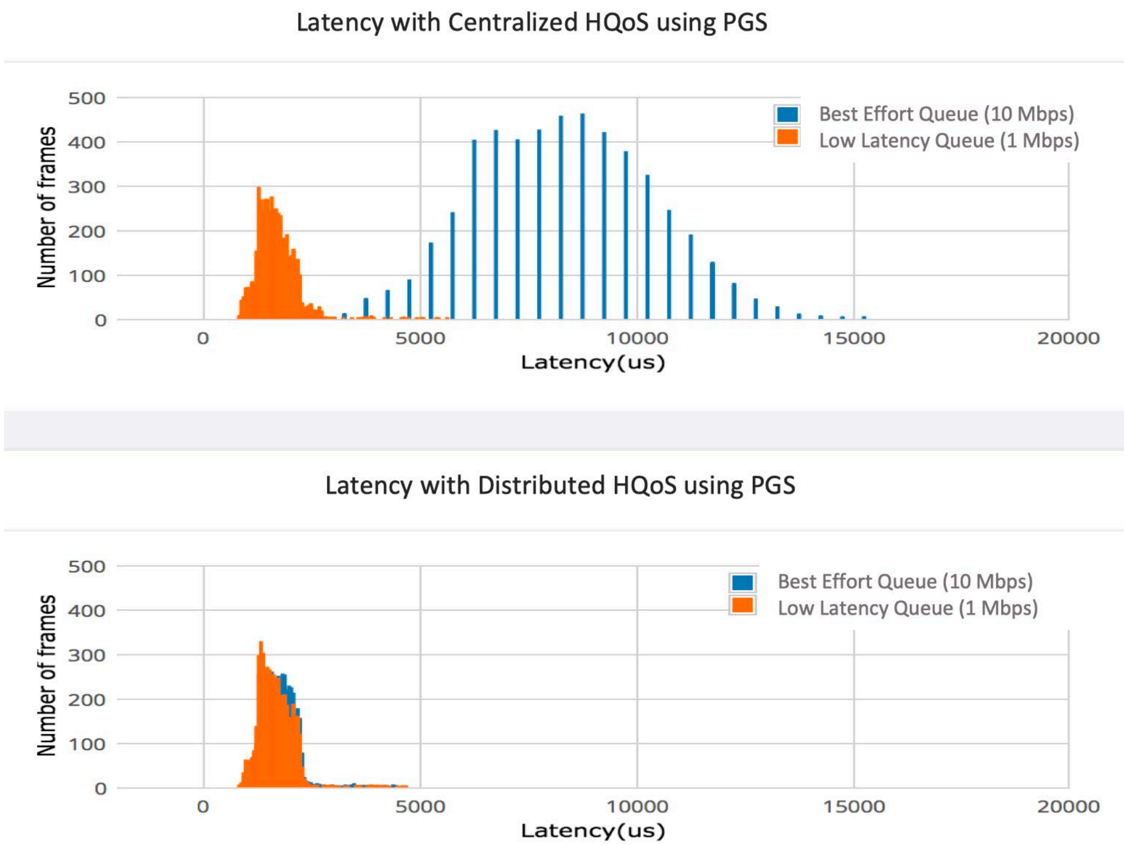


Figure 17 – Latency Improvement with Distributed HQoS

Figure 18 compares the bandwidth efficiency in terms of the channel utilization of the centralized and the distributed HQoS schemes. The channel utilization of the centralized scheme was around 90% on the 25 Mbps channel while the channel utilization of the distributed scheme was significantly lower at around 50%.

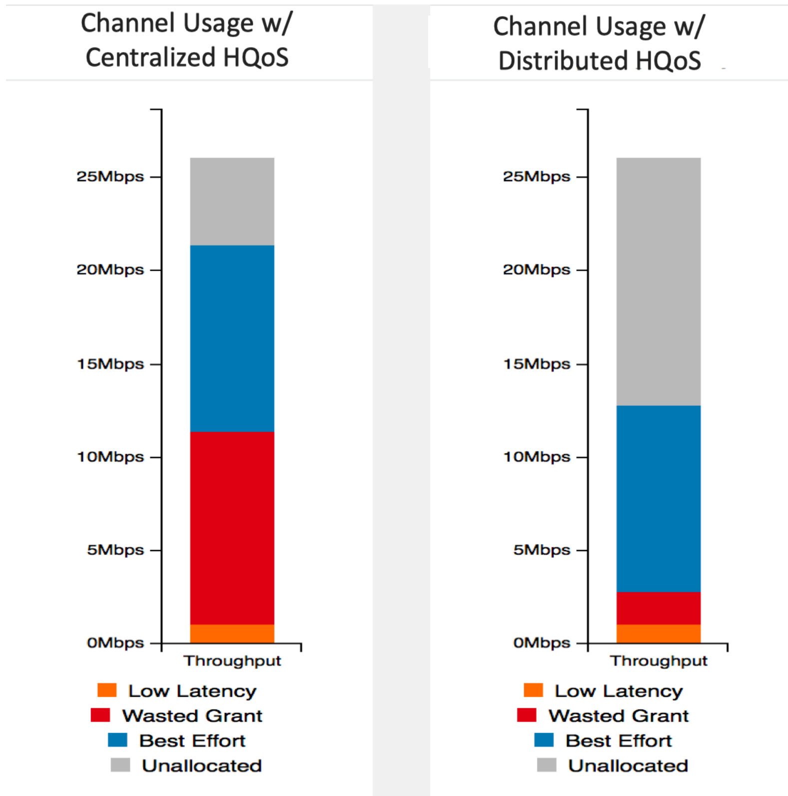


Figure 18 – Bandwidth Efficiency Improvement with Distributed HQoS

The lab results indicate that for latency based upstream scheduling such as PGS where the CMTS provides proactive grants to anticipate packet arrivals, *the distributed HQoS scheme can significantly and effectively improve bandwidth efficiency and reduce latency for all traffic types*.

8. DOCSIS Predictive Scheduling

DOCSIS Predictive Scheduling (DPS) is a technique that CMTS can use to reduce the US latency by giving grants based on the projected packet arrivals. The logic behind this scheme is that those predicted grants will allow CMs to bypass the conventional steps of the DOCSIS request-grant process and send packets right away so that US latency can be reduced.

DPS projects future bandwidth requests based on the following three types of information:

- **Configuration**: specifies the service flow QoS parameters and upstream channel capacity.
- **Explicit signaling**: includes the DOCSIS request message and LLX BWR message. The DOCSIS request message reflects the bandwidth need in the past when the request message was sent. The LLR BWR message on the other hand reflects the bandwidth need in the future when the data is expected to arrive at the CM.
- **Measurement:** includes service related metrics, such as the requested traffic rate, granted traffic rate and actual traffic rate and latency incurred per service flow.

DPS then uses the projected bandwidth requests to generate grants for the CM to send the upstream traffic. From the CM's point of view, there is no difference between the predicted grants and the requested grants.

Figure 19 illustrates the DPS framework which includes two control loops to adapt to both traffic and service policy changes. The inner traffic control loop is between the CMTS and the CM. The CMTS predicts requests and generates grants. The CM uses the grants to send traffic including data frames and BWR signaling to indicate future packet arrivals. If under granted, the CM sends DOCSIS request messages. The CMTS handles the requests and measures per service flow data throughput, request rate and data rate and use these data to predict future bandwidth requests.

The outer service policy control loop is between the service policy server and the CMTS. The operator defines service policy and pushes it to the CMTS as service flow, channel and data profile configurations. The CMTS uses the configurations to direct bandwidth request predictions. The prediction results as reflected in the measured service metrics are feedback to the policy server for data analysis and performance tuning.

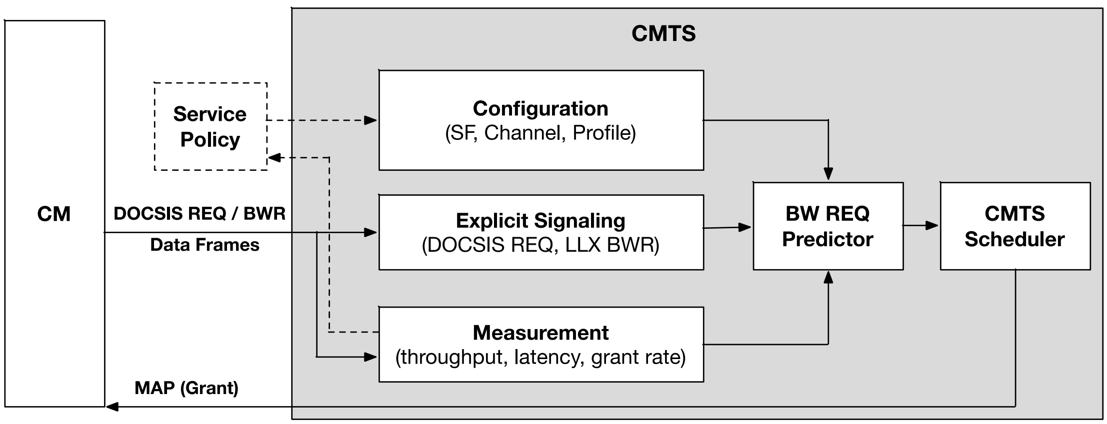


Figure 19 – DOCSIS Predictive Scheduling Framework

Essentially, DPS achieves traffic prediction by expanding the DOCSIS upstream scheduling knowledge base to include cross-domain traffic information such as BWR that is not native to DOCSIS and using modern data analytics for pattern based traffic predictions.

DOCSIS Synchronization

9. IEEE1588

The mobile network is by nature a synchronous network. The radios have overlapping spectrums and they use a central clock to align their clock frequencies and to align framing so that they can coordinate shared airspace. To achieve the goal of sharing a common clock, radios and their controllers often connect to the Global Navigation Satellite System (GNSS).

An equivalent global clock signal may also be transported over the IP network using the precision time protocol (PTP) which is described in IEEE 1588v2. A PTP timestamp is sent over the IP network. In a fully participant network, the clock is regenerated at each network node (usually an Ethernet switch or router). The delay through that network hop is deterministic and constant since it is over a wire. The PTP protocol is run between the network nodes to measure the delay. Once all delays are constant and measured, the timestamp that has been sent to the far end through the same constant delay network can be corrected and then match the original timestamp. A phase locked loop is required to filter out any packet delay variation (PDV).

The new emerging DAA architecture is also a synchronous system where common timing information is required between the CMTS core and the Remote PHY node. While this timing is not as critical in performance for DAA operation, it is derived from PTP and can be upgraded to the tolerances required by the mobile system.

The advantage of providing 1588 timing over the network is that it can lower the cost of the eNB by eliminating the GPS connection. This may lower the in-house installation cost by eliminating a GPS antenna run. The timing also provide a time reference for the eNB and the CMTS system to communication for LLX.

10. Over-the-Top (OTT) Mobile Xhaul over DOCSIS with GPS or 1588 Sync

The DOCSIS network is asymmetrical. If the PTP protocol is sent OTT of the DOCSIS network, it may incur variable buffer delay which in turn cause PDV.

For some mobile configurations such as LTE FDD, sync provided OTT might still yield a workable solution with some mitigating work. First, the PTP packets are given the higher priority DSCP and the DOCSIS system is configured to provide those packets with the lowest latency. Next, the grandmaster clock is inject at the CMTS or within a few network hops of the CMTS. This reduces the number of network elements that contribute timing error, so there is more timing error budget available for the DOCSIS system. The DOCSIS system is then configured with close to the same interleaver depth in the downstream and upstream as the interleaver depth is the dominant factor that leads to asymmetry.

While this is a good starting point, the better way to propagate the network timing is not OTT of a DOCSIS system, but with the DOCSIS Time Protocol (DTP) through the DOCSIS network.

11. DOCSIS Time Protocol (DTP)

The full end-to-end timing system that includes a DOCSIS network is shown in Figure 20. The GNSS system is received by the Primary Reference Time Clocks (PRTC). The PRTC becomes a grandmaster clock and generates PTP. The PTP messages are sent through a number of Ethernet switches which operate as telecom boundary clocks (T-BC). When PTP arrives at the CMTS, the DOCSIS system takes over. A DOCSIS network is already a synchronous system with its own timestamp. DTP finds the round trip delay of the DOCSIS system. Together, they provide the basic functionality of PTP. Both the CMTS and the CM each act as the equivalent of a T-BC. The CM then generates PTP that is sent to the end point, e.g., an eNB.

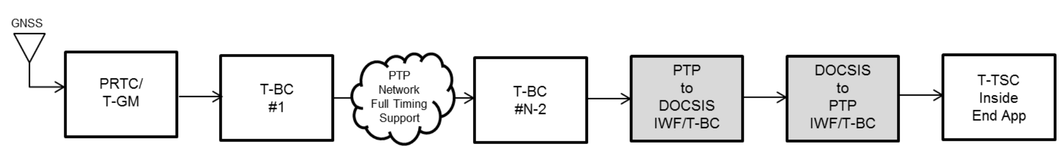


Figure 20 – PTP Deployment with DOCSIS Equipment

DTP determines the round trip delay of a DOCSIS network by reverse engineering the ranging information. It then compares a zero length HFC network to the current HFC network to determine the HFC forward path one-way delay. This algorithm is explained in [7].

Conclusion

LLX uses a pipelined scheduling mechanism to provide lower latency. It does this by connecting the mobile schedular to the CMTS scheduler using the BWR message. LLX also defines a common quality of service framework for both mobile and DOCSIS so that the relative importance of different traffic streams is maintained across the two systems. The result is that the DOCSIS system has fiber-like performance and latency.

LLX can be further enhanced by using a new distributed hierarchical QoS structure that allows for grant sharing in the CM. There are two primary use cases. The first is when a scheduling service like PGS is used that is wasteful of grants. The second use case is then the traffic being presented to the CM by the eNB is place on different flows that what was indicated in the BWR message.

DOCSIS predictive scheduling (DPS) can be useful for flows that are not completely covered by BWR, such as flows that originate at the eNB. It can also help out if the BWR message does not arrive in time at the CMTS.

The DOCSIS time protocol (DTP) can characterize the round-trip and one-way delay for a DOCSIS network. When combined with the native DOCSIS 3.1 timestamp, it can work with a IEEE 1588/PTP network to provide highly accurate timing to a CPE.

LLX is implemented with minimal impact on existing RAN and DOCSIS equipment. No hardware changes are required. The software can be engineered in ways to have minimal impact to existing implementation. This should facilitate rapid industry adoption.

Implementations are available today on commercially deployed equipment.

Abbreviations

5QI	5G QoS indicator
AE	access entity
ASF	aggregated service flow
BE	best effort
BWR	bandwidth report
CDF	cumulative distribution function
CIN	converged interconnect network
CPRI	common public radio interface
CU	central unit
DAA	distributed access architecture
DPS	DOCSIS predictive scheduling
DS	downstream
DSCP	DiffServ codepoint
DTP	DOCSIS time protocol
DU	distributed unit
DWDM	dense wavelength division multiplexing
EPC	evolved packet core
GTP	GPRS tunneling protocol
GTP-U	GTP user plane
HQoS	hierarchical QoS
IMS	IP multimedia subsystem
LCG	logical channel group
LLD	low latency DOCSIS
LLX	low latency xhaul
MNO	mobile network operator
MSO	multiple service operator
MTU	maximum transmit unit
MVNO	mobile virtual network operator
OAI	OpenAir interface
OLT	optical line terminal
ONU	optical network unit
OSS	operation support system
OTT	over-the-top
PON	passive optical network
PGS	proactive grant service
PRTC	primary reference time clock
PTP	precision time protocol
QCI	QoS class indicator
QoS	quality of service
RAN	radio access network
RMD	remote MACPHY device
RPD	remote PHY device
RTPS	real-time polling service
RU	radio unit
SCTP	stream control transmission protocol
S-GW	serving gateway

T-BC	telecom boundary clock
TN	transport node
TU	transport unit
UE	user equipment
US	upstream
URLLC	ultra-reliable low-latency communication

Bibliography & References

[1] "Low Latency Mobile Xhaul over DOCSIS Technology," CM-SP-LLX-I01-190628, CableLabs.

[2] John T. Chapman, Jennifer Andreoli-Fang, "Low Latency Techniques for Mobile Backhaul over DOCSIS," SCTE Fall Technical Forum, October, 2017.

[3] Jennifer Andreoli-Fang, John T. Chapman, "Mobile-aware scheduling for low latency backhaul over DOCSIS," *Proc. of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Montreal, Oct 2017.

[4] Jennifer Andreoli-Fang, John T. Chapman, "Latency reduction for mobile backhaul over DOCSIS through pipelining," *Proc. of IEEE Globecom*, Singapore, Dec 2017.

[5] John T. Chapman, Jennifer Andreoli-Fang, Michel Chavin, Elias Chavarria Reyes, Zheng Lu, Dantong Liu, Joey Padden, Alon Bernstein, "Low latency techniques for mobile backhaul over DOCSIS," *Proc. of IEEE Wireless Communication and Networking Conference (WCNC)*, Barcelona, April 2018.

[6] Jennifer Andreoli-Fang, John T. Chapman, "Synchronization for Mobile Backhaul over DOCSIS," SCTE Fall Technical Forum, October, 2017.

[7] John T. Chapman, et. al., "The DOCSIS Timing Protocol (DTP), Generating precision timing services from a DOCSIS system," INTX/SCTE Spring Technical Forum, 2011.

[8] Jennifer Andreoli-Fang, "DOCSIS-Mobile Convergence," Broadband Library, summer edition 2019. [link]

[9] Kyung Mun, "Spectrum – Will its intrinsic value decrease in a small cell world?" [link]

[10] Shahed Mazumder, Joey Padden, Jennifer Andreoli-Fang, "DOCSIS Technology for Mobile Xhaul – Enabling Superior TCO." CableLabs Strategy Brief, August 2019. [login required]

[11] 3GPP TS 36.423, "E-UTRAN X2 Application Protocol (X2AP)."

[12] 3GPP TS 36.425, "E-UTRAN X2 Interface User Plane Protocol."

[13] 3GPP TS 23.203, "Policy and charging control architecture."

[14] 3GPP TS 23.501, "System Architecture for the 5G System."

[15] K. Alexandris, et. al., "Analyzing X2 Handover in LTE/LTE-A," IEEE International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), Tempe, May 2016. [link]

[16] Dave Morley, "5G Small Cells and Cable – Realizing the Opportunity", SCTE Fall Technical Forum, October, 2018.