Advanced Quality of Experience Monitoring Techniques for a New Generation of Traffic Types Carried by DOCSIS

ARRIS
Tom Cloonan, CTO Network Solutions; Ben Widrevitz, Sr. Architect System Architecture;
Jim Allen, Engineer System Architecture; Tony Cotter, Engineer System Architecture

ARRIS

Contents

Background Information

Introduction and Motivation

Quality of Experience (QoE) is a basic measure of the user's level of satisfaction, and Quality of Experience Monitoring is an extremely important tool that helps MSOs reduce subscriber churn and that helps MSOs quickly trouble-shoot subscriber problems. QoE Monitoring permits the MSO to determine when their network infrastructure is providing adequate bandwidth support for the offered services. Low QoE scores can be an indication that the network is no longer performing at adequate levels, and the existence of low QoE scores usually serves as an important trigger to the on-going evolution and modification of the HFC Plant as MSOs continually schedule node-splits and/or channel augmentations to address observed QoE issues.

Unfortunately, QoE Monitoring is becoming more difficult because the nature of traffic propagating over DOCSIS networks has been changing quite rapidly in recent years. These changes are taking place on four different fronts.

First, there is a much larger mix of traffic types that are now making up the aggregated bandwidth within typical DOCSIS networks. These traffic types now include:

Web-Surfing

Over-the-Top SD Streaming IP Video

Over-the-Top HD Streaming IP Video

MSO-Managed SD Streaming IP Video

MSO-Managed HD Streaming IP Video

Streaming Audio

Gaming

VoIP

Peer-to-Peer Services

The complex interactions between these different traffic types is making it more and more difficult to predict when different subscribers using different applications are receiving adequate service levels.

Second, each of the different traffic types listed above has its own unique and different traffic requirements and sensitivities. Some of the traffic types are sensitive to changes in packet stream bandwidth. Others are sensitive to changes in packet delay and jitter. Still others are sensitive to packet loss. The unique requirements for each traffic type make it difficult to easily ascertain QoE levels.

Third, Maximum Sustained Traffic Rates for cable modem subscribers now range over much larger values. The heterogeneous mix of subscribers with different bandwidth needs also complicates the problem of determining QoE levels.

Fourth, the arrival of Adaptive Bit-Rate IP Video traffic on the DOCSIS network introduces a new traffic type that can quickly expand to fill unused bandwidth capacity and can be compressed during periods of congestion. This also makes it more difficult to determine QoE levels.

While being able to quickly and correctly determine the DOCSIS QoE levels for individual subscribers is getting more difficult, it is also becoming more imperative. Why? There are several reasons. First, more and more service types carried by DOCSIS might be considered to be “mission critical” services. VoIP is an obvious example, but even IP Video services delivered to television sets might be considered critical in the future. In addition, MSOs are entering a decade of rapid change for many of their service bandwidths, which will cause rapid changes in their HFC spectrum mixes. Each MSO will experience these changes differently, as each MSO will be working with a different set of conditions and constraints on their HFC plants that lead them to different Spectral Maps over time. Figure 1 illustrates one example of how one MSO might migrate their spectral maps over the next decade. [Clo1]

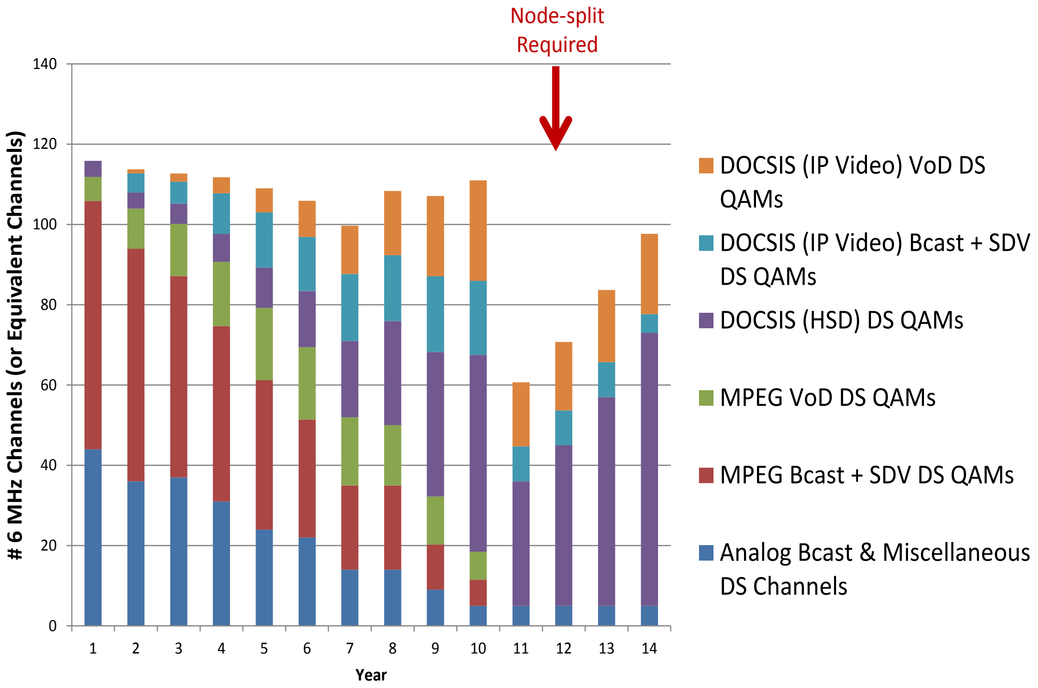


Figure 1: Example Spectral Map Changes

This figure illustrates how quickly certain service types are expected to grow. Required per-subscriber bandwidth growth for DOCSIS High-Speed Data service may continue to grow at a 50% CAGR in the future, and that would force the DOCSIS service tier to consume much larger portions of the HFC spectrum. The introduction of MSO-Managed IP Video over DOCSIS within the next decade is also expected to force swift and considerably larger changes on the HFC Spectral Map.

As MSOs increase the number of DOCSIS channels, they will from time to time be forced to make room for the new DOCSIS channels by eliminating channels from other service types within the HFC Spectral Map. This fact leads to a difficult dilemma- which services should donate their channels to DOCSIS and how quickly can they make these changes. There will undoubtedly be times when MSOs are forced to delay growth in their DOCSIS service tier due to constraints that temporarily preclude them from removing channels from the other tiers. Thus, MSOs may be forced to temporarily “squeeze” their DOCSIS services into over-subscribed DOCSIS spectra from time to time. This undesirable condition will make it imperative for MSOs to be able to determine the QoE of their DOCSIS subscribers while going through the upcoming decade of challenging transitions.

In addition, the *resolution* of QoE Monitoring performed by MSOs must be increased. Higher resolution QoE Monitoring implies the ability to monitor the QoE levels for different traffic types as well as for different Service Level Agreements (SLAs) at higher sampling rates. If DOCSIS subscribers are going to experience QoE issues, MSOs will undoubtedly be interested in knowing exactly which subscriber traffic types and SLAs are experiencing the issues and they will also be interested in knowing how severe the issues are for each of the traffic types and each of the SLAs. This could permit MSOs to make intelligent decisions about whether the issues warrant drastic changes like instantaneous node-splits or whether subscribers are likely to "live with" the issues for a short period of time.

As an example, if the only traffic type and SLA experiencing a performance issue is Peer-to-Peer traffic within the Bronze SLA level and the MSO has only a small percentage of subscribers within that Bronze SLA level, then it is possible that the MSO may want to risk living with the issue for a short while. If, on the other hand, the traffic types that are experiencing the issue are Web-Browsing and VoIP and IP Video services within the Bronze and Silver and Gold SLA levels, then rapid actions to mitigate the issues may be required.

As a result, it should be clear that high-resolution QoE Monitoring of the performance levels of different traffic types and different SLA levels could become an essential tool in the MSO toolkit as MSOs navigate the challenging, ever-changing waters of the upcoming decade.

Previous QoE Monitoring Tools

In the past, QoE Monitoring has taken on many different forms for the different services that were being monitored.

For VoIP services, monitoring was often implemented within the VoIP destination endpoints, monitoring performance parameters such as packet delay, packet jitter, and packet loss. In addition, VoIP monitoring oftentimes kept track of connection set-ups and the success of the associated signaling.

For Legacy Digital Video services, monitoring was often implemented with sampling points along the path of the Digital Video streams. These sampling points could be positioned in the head-end and/or in the subscriber endpoints (such as STBs). These sampling points could monitor packet loss, packet jitter, packet delay, and packet corruption. They could also perform deep-packet inspection types of functions to determine what type of video packet was corrupted or lost.

For both VoIP and Video monitoring, the monitoring tools would collect all of the aforementioned parameters. The parameters would then be combined together to create a reasonable prediction of the user's QoE level. Many of the monitoring tools had complex algorithms that determine the magnitude of any service disruption. For example, Video monitoring tools could determine whether a packet loss would corrupt an I-Frame or a P-Frame or a B-Frame to determine the duration of the resulting error. They also would attempt to determine if packet corruption would lead to macroblock tiling displays or frozen video displays or audio outages. The magnitude and duration of service interruptions were taken into account when creating these QoE level estimates.

The QoE level estimates from monitoring tools would oftentimes list all of the acquired parameters from above. However, a convenient technique for displaying the overall QoE level estimate with a single number uses the concept of a Mean Opinion Score (or MOS Score). A MOS Score is an average score (ranging from 1 to 5) that attempts to combine the many effects from the many different performance-affecting parameters defined above. A MOS Score of 1 corresponds to the lowest possible quality level as it might be perceived by the subscriber. A MOS Score of 5 corresponds to the highest possible quality levels as it might be perceived by the subscriber. Table 1 illustrates the typical definitions for the different MOS Score levels.

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly Annoying
2	Poor	Annoying
1	Bad	Very Annoying

Table 1: MOS Score Level Definitions

Since QoE levels (and their associated MOS scores) can vary over time, the continual monitoring of QoE MOS scores is important for MSOs, because the existence of lower MOS scores can be an indication that the network is no longer performing at adequate levels, and the existence of low MOS scores usually serves as an important trigger to the on-going evolution and modification of the HFC Plant as MSOs continually schedule equipment upgrades and/or node-splits and/or channel augmentations to address observed QoE issues.

MSOs have also performed QoE monitoring of their DOCSIS services for many years. These DOCSIS QoE Monitoring efforts were not as well-developed as the techniques used for VoIP and Digital Video services. For the most part, DOCSIS QoE Monitoring of packet streams was limited to a study of bandwidth utilizations on DOCSIS channels or DOCSIS service groups. Some MSOs would also monitor channel parameters such as SNR, uncorrectable FEC errors, and correctable FEC errors to determine if the quality of the DOCSIS channels was degrading with time. These simple techniques were often employed because they were good enough to ascertain a rough QoE metric for most of the DOCSIS services. They worked well in the previous era when less user applications existed, when users were more homogeneous, and when the distance between the highest and lowest Maximum Sustained Traffic Rate levels were much smaller than they are today. Some QoE monitoring tools would also provide more advanced Degraded Modem-Hour metrics (which are similar to MOS scores) for DOCSIS services. These types of scores were very good at predicting subscriber QoE levels when the number of different traffic types propagating over the DOCSIS network were limited.

However, with the rapid expansion of new traffic types being carried by DOCSIS and with larger variances in Maximum Sustained Traffic Rates for subscribers, it is becoming apparent that extensions may be required to the previously-available DOCSIS QoE Monitoring tools. Why is this the case?

Let us consider an example. Assume that an MSO has deployed a single ~40 Mbps DOCSIS channel with 100 DOCSIS subscribers sharing the channel, and assume that the channel is operating at ~100% utilization. Assume that there is 0.001% packet loss. Are the subscribers currently satisfied with their Quality of Experience levels?

The answer is not clear. For example, if there is currently only one active subscriber using the channel and that subscriber has a Maximum Sustained Traffic Rate of 40 Mbps, then that particular subscriber would be receiving the full ~40 Mbps of service and would probably be ecstatically satisfied with the level of the service. Since the other 99 subscribers are apparently not using the DOCSIS channel at this point in time, they are probably satisfied as well.

However, in the unlikely event that all 100 DOCSIS users were simultaneously sharing the resources of the channel at the same time, then (on average) each subscriber would be receiving (~40 Mbps)/100 = 400 kbps of bandwidth. Depending on the applications that the subscribers are using, the subscribers may or may not be satisfied. If everyone is making VoIP telephony calls requiring 120 kbps of bandwidth, then the 400 kbps level of bandwidth may be adequate and everyone might still be satisfied. If, however, a lot of the active subscribers are trying to access HD IP Video streaming services requiring (say) 6 Mbps of bandwidth, then 400 kbps level of offered bandwidth is inadequate and will undoubtedly result in many unsatisfied subscribers.

One of the other inadequacies of existing DOCSIS QoE Monitoring methods is the frequency with which monitoring is oftentimes implemented. Many DOCSIS QoE Monitoring systems of today are designed with typical sampling periods on the order of days or hours or minutes. While that gives some level of visibility into the QoE levels of subscribers, it can miss some of the high-frequency, short-duration events that might cause QoE degradation on the DOCSIS network (like micro-bursts of bandwidth causing frequent and temporary increases in latency).

Thus, from the above scenarios, it becomes clear that, in the future, MSOs may need to do more than just measure the channel utilization of their DOCSIS channels once per hour to ascertain whether their subscribers are currently satisfied or not. The remaining sections of this paper will outline some new ideas that might be useful as we consider the addition of extensions into the future DOCSIS QoE Monitoring tools.

Improvements in Future DOCSIS QoE Monitoring Tools

Categorizing DOCSIS QoE Metrics

As indicated in the previous sections, DOCSIS QoE Monitoring tools are likely to be receiving some upgrades over the next decade. In order to determine the likely forms of these upgrades, it is beneficial to first consider the various metrics of the DOCSIS packet streams that are already monitored and determine how they can be classified. These classifications will prove to be useful as we define new metrics.

The existing metrics used in QoE Monitoring tools and the methods for collecting those metrics can be classified using several different attributes.

First, metrics can be collected in a continuous fashion (ex: sampling every time a packet passes) or in a periodically sampled fashion (ex: sampling average values once every hour).

Second, metrics can be direct measurements of the Quality of Experience of a subscriber (ex: measuring file download times) or can be indirect measurements of the Quality of Experience of a subscriber (ex: measuring packet delay and trying to infer subscriber satisfaction for different applications when those delay conditions exist).

Direct methods can place "canary clients" inside of already-deployed modems or can place purpose-built "canary modems" out on the HFC Plant. These canaries periodically download HTTP files like a Web Browser or download video content like an IP Video client or perform many parallel TCP downloads of small files like a Peer-to-Peer client or transmit small UDP packets like a VoIP client or access DNS servers and measure the performance metrics of these activities. The direct performance metrics for canaries include metrics such as file download times, packet loss (for VoIP), packet delay and jitter (for VoIP), and delay (for DNS). These metrics are directly translatable into MOS scores that might be seen by other real subscribers on the same channel that might be using that same application.

Indirect methods do not require canaries and do not add any extra traffic to a channel that might already be congested. Instead, indirect methods sit passively inside of network elements like CMTS and CMs and measure the performance of "real-world" packets that are propagating through those devices. Indirect performance metrics that might be measured include packet throughput (for a user), packet loss, packet delay, packet jitter, and channel congestion. These metrics do not directly translate into MOS scores, but with intelligence, a QoE Monitoring tool can oftentimes infer the QoE levels of subscribers that might be on a channel where these indirect performance metrics were captured.

In general, direct metrics tend to be better measures of a subscriber's QoE level, because they measure "real-world" response times for their simulated traffic loads. However, direct metrics are more difficult to obtain (requiring canaries to be planted in the HFC plant and requiring servers to serve up the requested files). In addition, direct methods add more bandwidth to channels that may already be congested, so these measurement techniques can negatively affect the channel performance levels.

In general, the inference steps required for the use of indirect metrics tend to make them be less accurate measures of a subscriber's QoE level, because they can only infer what real performance levels might be from the measured variables. However, indirect metrics are much easier to obtain (requiring simple counts to be implemented in CMTS or CMs to monitor the real-world data that passes by). In addition, indirect methods do NOT add any extra bandwidth to channels that may already be congested, so these measurement techniques do not negatively affect the channel performance levels.

As a result of the above facts, QoE metrics can be classified into one of four categories:

(1) Continuous, Direct QoE metrics

(2) Sampled, Direct QoE metrics

(3) Continuous, Indirect QoE metrics

(4) Sampled, Indirect QoE metrics

In addition to this categorization of QoE metrics, the authors have also found it valuable to identify the scope of a particular metric. The scope of a metric helps answer the question "What layer or level of the data stream was monitored to collect the metric?" To define the scope levels, it is beneficial to view the aggregated traffic that makes up a DOCSIS network to be comprised of hierarchical layers of data streams, where each layer is built up from a set of elements from the underlying layers. An abstract illustration of this concept is shown in Figure 2, where DOCSIS Service Groups are comprised of DOCSIS Channels, DOCSIS Channels are comprised of Modem Traffic, Modem Traffic is comprised of Service Flow Traffic, Service Flow Traffic is comprised of TCP/UDP Sessions, and TCP/UDP Sessions are comprised of IP Packets.

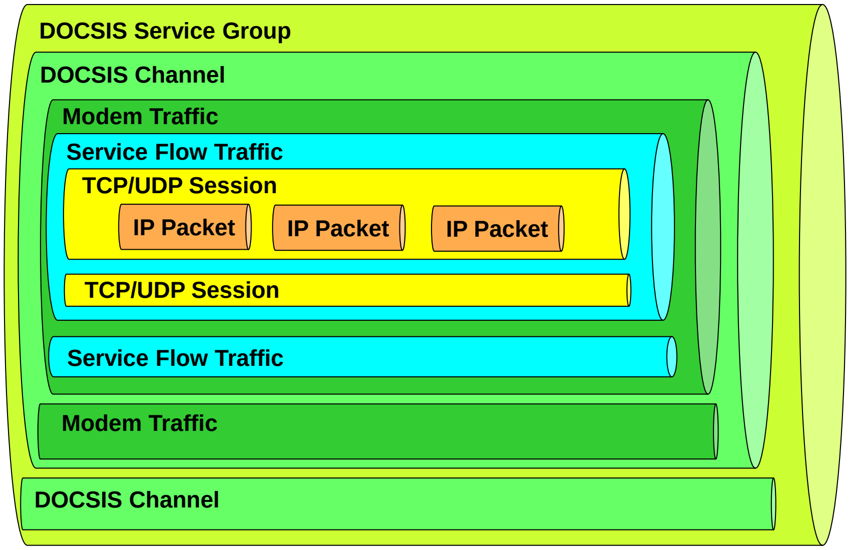


Figure 2: Traffic Hierarchy

Monitoring of QoE can take place at various levels in the hierarchy of Figure 2. For example, at the lowest level of the hierarchy (the IP Packet level), one can monitor packet-level characteristics such as packet loss, packet latency, and packet jitter. At the next level of the hierarchy (the TCP/UDP Session level)- which often corresponds to an HTTP Session for Web-Browsing or IP Video applications- one can monitor session-level characteristics such as average throughput, average download times, average packet loss, average packet latency, and average packet jitter within the session. All of these average session level characteristics can also be calculated for the higher levels in the hierarchy. In addition, at the Service Flow level, Modem level, DOCSIS Channel level, and DOCSIS Service Group level, one can also calculate aggregate characteristics such as Capacity Utilization.

Each of these metrics can, of course be categorized using the categories defined above. For example, average packet jitter that is sampled once per hour within a particular Service Flow would be categorized as a Sampled, Indirect QoE metric at the Service Flow scope. It is obviously sampled at a rate of one sample per hour. It is an indirect metric, because intelligence and calculations are required to determine whether the user experience level will be good or bad as a result of the measured jitter level, and it should be clear that the traffic type will be a key factor in determining that fact. It is also sampled from the many packets within the scope of a Service Flow.

All of the above metrics (packet loss, packet latency, packet jitter, throughput, capacity utilization, etc.) viewed at the different scope levels of the hierarchy are typical metrics that can be useful in developing a view of the subscriber QoE level at any instant in time. These metrics have been used by many traditional QoE Monitoring tools.

But the authors were interested in determining if there were other metrics that could be added to the list of existing metrics to help create a more insightful view of the subscriber QoE level in the future. The remainder of this paper will focus on this important topic.

Areas For Possible Improvement Within DOCSIS QoE Monitoring

The existing metrics available for QoE Monitoring tools cover many aspects of DOCSIS packet stream performance. But there is still much room for improvement in the future. Several areas are beckoning for improvement. These areas will receive special attention within this paper. They include:

(1) QoE MOS scores that are cognizant of the performance requirements of different traffic types

(2) QoE MOS scores that are cognizant of the performance expectations of different SLAs

(3) QoE MOS scores that are cognizant of both traffic types and SLAs

(4) QoE MOS scores that are cognizant of the number of active subscribers sharing a channel or service group

(5) QoE MOS scores that are cognizant of the traffic scheduling algorithms employed by the CMTS Downstream and Upstream traffic managers

(6) QoE MOS scores that are cognizant of Offered Loads

(7) QoE MOS scores that utilize metrics collected with high sampling rates

(8) QoE MOS scores that are cognizant of different scopes within the DOCSIS Network

We will explore each of these areas in the sections below.

MOS Scores Based On Traffic Types

At a minimum, MOS scores of the future will undoubtedly be required to recognize that different traffic types have different sensitivities to different traffic attributes. In addition, different traffic types can have extremely different performance requirements and can also have very different thresholds marking the important boundary line between acceptable QoE performance and unacceptable QoE performance.

As an example, if we consider the primary traffic types outlined in the first section of this paper, we will find that each traffic type is sensitive to a different set of traffic attributes. This is outlined in Table 2.

Traffic Type	Primary Sensitivities
Web-Surfing	Avg BW, Avg Delay
OTT SD IP Video	Avg BW
OTT HD IP Video	Avg BW
MSO SD IP Video	Avg BW
MSO HD IP Video	Avg BW
Streaming Audio	Avg Jitter
Gaming	Avg Delay, Avg Jitter
VoIP	Loss, Avg Delay, Avg Jiter

Table 2: Traffic Attributes to Which Each Traffic Type is Sensitive

One may argue that each of the traffic types in Table 2 might also be sensitive to other attributes, and that would be an accurate argument.

For example, one may argue that VoIP is also sensitive to Average Bandwidth levels- it requires the Average Bandwidth to support the minimum levels required for VoIP traffic (ex: 120 kbps) and it will not operate well at lower levels of Average Bandwidth. But these levels are so low that they are usually guaranteed by most DOCSIS Networks (unless exceptionally high congestion is occurring), so for most VoIP applications in today’s networks, the Average Bandwidth level is not a major issue for VoIP.

As another example, one may argue that IP Video is also sensitive to Average Delay and Average Jitter- that widely varying delays could lead to buffer underflows at the video display device. This is true if the IP Video Delivery Architecture does not buffer a large amount of video content before beginning the playout of the IP Video. However, most IP Video Delivery Architectures of today do, in fact, buffer a large amount of video content before beginning the playout, so this problem with Delay and Jitter is actually quite rare.

Nevertheless, these two examples illustrate that it is not always clear how to identify the specific traffic attributes to which a particular application is sensitive. Any assumptions will be valid for some cases and potentially invalid for other cases. So one must be cautious when interpreting the meaning of MOS scores that are defined for different traffic applications.

Other errors in MOS score predictions can occur whenever multiple traffic types are flowing to a single modem at the same moment in time. That particular condition can create unexpected interactions between the different traffic types. For example, TCP congestion windows and TCP bit-rates for a newly-started flow may take a longer period of time to grow if multiple traffic types are sharing the bandwidth capacity to a single modem. These types of effects make it more difficult to correctly predict the MOS scores for the interacting traffic types.

Regardless of the aforementioned issues, most MSOs will agree that some indicators of performance level are better than no indicators of performance level- even if those performance indicators have limitations.

For this reason, we will attempt to define appropriate formulae for MOS scores for the currently-dominant traffic types found on DOCSIS Networks. Those formulae will be developed in one of the sections below.

MOS Scores Based On SLAs

Service Level Agreements (SLAs) typically exist between subscribers and service providers. These SLAs define the maximum and minimum performance levels that the subscriber can expect to see for their purchased subscription to the service. DOSCIS permits MSOs to specify many different performance attributes for each of the service flows attached to each of their subscribers. These performance attributes can include parameters such as the Maximum Sustained Traffic Rate (aka Tmax) and the Minimum Reserved Traffic Rate (aka Tmin). MSOs usually offer different levels of service to their subscribers using different service tiers with different price tags and different performance attributes. As an example, an MSO may offer a Gold Service Tier for $70, with a Tmax of 50 Mbps, and may also offer a Bronze Service Tier for $40, with a Tmax of 10 Mbps. Once these parameters are defined and configured for a particular subscriber the MSO must try to satisfy the requirements of the SLA to ensure that the subscriber remains satisfied.

As a result, it seems apparent that QoE MOS scores could include information about the different SLA levels within the different Service Tiers. As an example, assume that an HD IP Video stream requires a minimum Average Bandwidth of 14 Mbps to yield acceptable results to the subscriber. If a subscriber has subscribed to the Gold Service Tier level (Tmax = 50 Mbps) and does not receive adequate bandwidth to support the required 14 Mbps HD IP Video stream, then that subscriber will obviously be less than satisfied with the service. The MOS score for the Gold subscriber will obviously be marked as being low. If, on the other hand, a subscriber has subscribed to the Bronze Service Tier level (Tmax = 10 Mbps) and does not receive adequate bandwidth to support the required 14 Mbps HD IP Video stream, then that subscriber has no right to be dissatisfied with the service, because their lower level of bandwidth is directly caused by their low service tier level. As a result, the MOS score for the Bronze subscriber may be marked as being low- but it should be marked as being low due to inadequate Tmax levels and not due to congestion. Alternatively, the MOS score can be marked high since the problem is not being caused by the MSO. Either approach can be deemed acceptable depending on how the MSO wants to interpret the information.

Thus, it should be apparent that MOS score calculations can make good use of any information provided on the subscribers' SLA levels. This means that every customer's QoE will now depend on their SLA and upon what service they are trying to obtain (e.g., Web-Browsing, IP Video viewing). Users with a higher tier SLA (and higher Tmax) will likely be happier than users with a lower tier SLA (and lower Tmax). Users requesting lower bandwidth services will tend to be happier than those needing more bandwidth. One might imagine every combination of SLA and traffic type receiving a different QoE. This type of scenario will be outlined below.

MOS Scores Based On Traffic Types And SLAs

The previous two sections described how MOS score calculations can benefit from information on the traffic types and SLAs being used by subscribers. In fact, MOS score calculations can make even better predictions about subscriber QoE levels if they combine these two attributes. One can imagine a good MOS score that displays all possible combinations of traffic types and SLAs.

Thus, if there are two SLA levels (Bronze and Gold) and if there are three primary traffic types (Web-Browsing, HD IP Video, SD IP Video), then one could envision a table with the three traffic types along the right-most column and with the two SLA levels along the top row. This results in six different combinations and six different MOS scores that can be calculated and displayed, as shown in Table 3.

	Gold Tier	**Bronze Tier**
Web-Browsing	MOS_1	MOS_4
HP IP Video	MOS_2	MOS_5
SD IP Video	MOS_3	MOS_6

Table 3: Traffic Type & SLA Combinations

MOS Scores Based On Active Subscriber Counts

In an earlier section of the paper, it was shown that knowledge of the channel utilization alone is inadequate to identify the QoE level for subscribers. In particular, it was shown that one cannot easily determine if a single ~40 Mbps DOCSIS channel with 100 DOCSIS subscribers sharing the channel and a channel utilization of ~100% yields high QoE levels or low QoE levels. If there is only one active subscriber, then the QoE level would likely be high (for that subscriber). If there are 100 active subscribers, then the QoE level would likely be low.

As a result, it should be clear that a measure of the number of active subscribers could also prove useful within MOS score calculations. For example, if the bandwidth is equally shared by all of the active subscribers, then the MOS score could use the following simple formula as an estimate of the bandwidth consumed by each subscriber:

Avg BW/Sub = C * U / N (1)

where:
C = Channel Capacity
U = Channel Utilization
N = No. Active Subscribers

MOS Scores Based On CMTS Scheduling Algorithms and SLAs Of Active Subscribers

The Average BW per Sub formula (formula 1) above assumes that the bandwidth was shared equally by all of the active subscribers. This assumption may not always be true, because the details behind any CMTS scheduler might be quite a bit more complicated. As an example, assume that a particular CMTS scheduling algorithm is designed to give each active subscriber an amount of bandwidth that is proportional to their Minimum Reserved Traffic Rate (Tmin) value divided by the sum of all of the Tmin values for all of the active subscribers. The bandwidth offered by the CMTS scheduling algorithm to a particular subscriber's Service Flow (ABW) would therefore be given by:

ABW = C * U * Tmin(i) / ▯[Tmin(j)] (2)

where:
ABW = Avg Service Flow BW for Sub #i
C = Channel Capacity
U = Channel Utilization
Tmin(i) = Tmin for i-th Subscriber Flow
▯[Tmin(j)] = Sum of Tmin(i) over Active Subs

If the QoE Monitoring tool was cognizant of the fact that the CMTS scheduling algorithm was using this formula, then the same formula could be used by the QoE Monitoring tool if/when it needs to determine the average bandwidth being offered to a particular subscriber within its MOS score calculations. In a certain sense, formula (2) is a more accurate description of the bandwidth offered to subscribers than the more simple version in formula (1). Using the more accurate formula (2), the SLA-based MOS score thresholds described in previous sections could be compared against the Average Bandwidths that would be offered by the CMTS scheduling algorithm, as predicted by formula (2).

It is important to note that these MOS score calculations would have to be altered if the CMTS scheduling algorithms were changed or if they were associated with different CMTSs from different vendors (since different CMTS vendors will likely use different scheduling algorithms within their proprietary solutions).

Thus, support for MOS scores of this nature requires tight coordination between the CMTS scheduling algorithms and the QoE Monitoring tool. In particular, the QoE Monitoring tool algorithms must be written with complete knowledge of the details behind the CMTS scheduling algorithms on each of the CMTSs that the tool would support.

MOS Scores Based On Requested Loads

While DOCSIS channel utilization is indeed a valuable tool that can (and should) be used in QoE Monitoring tools of the future for Network Planning and Fault Isolation activities, channel utilization has always suffered from an innate limitation that it is a measurement that "pegs" somewhat near 100%, as shown in Figure 3 below.

(Note: In order to compute the payload channel capacity to be used in our calculation of channel utilization we must discount the raw channel capacity by the bandwidth consumed by FEC overhead, protocol overhead and management messages. This discounting often involves some degree of estimation which introduces a small amount of error into the channel utilization measurement.)

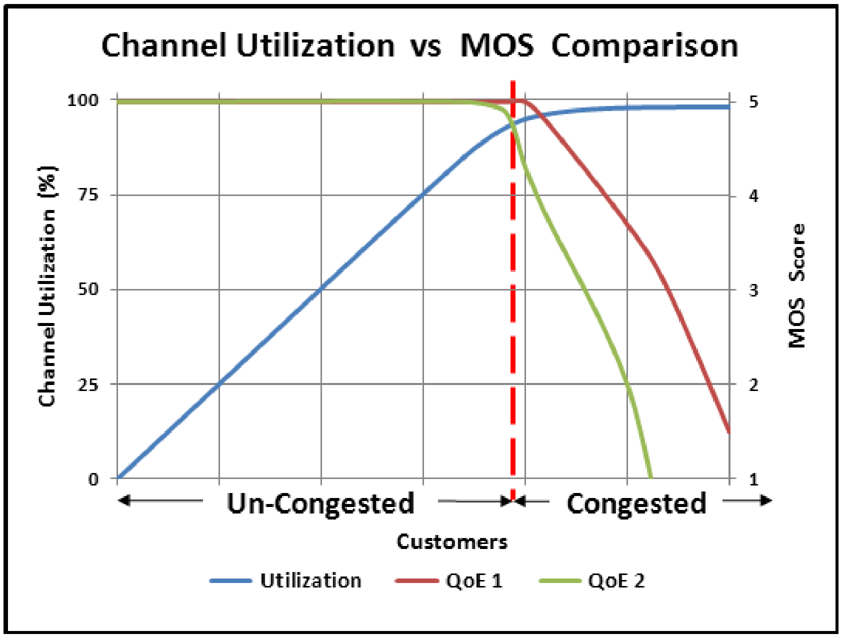


Figure 3: Channel Utilization & MOS Scores vs. # Customers

Figure 3 illustrates a hypothetical scenario where every new customer added to the channel adds a fixed amount of "requested load" to the channel. The requested load is the amount of bandwidth that the customer would like to receive to satisfy their desired high-speed data activities. The "composite requested load" on a channel is the sum of the requested loads from all active subscribers, and it should be obvious that (depending on the number of active subscribers at any given time) it can exceed 100% of the channel's capacity. But when the composite requested load exceeds 100% of the channel capacity, it is clear that the actual channel utilization will still be a value that is less than or equal to 100%. This difference between the composite requested load and the actual channel utilization is obviously a good indicator of subscriber QoE degradation, because it represents unsatisfied bandwidth demands.

It should be clear that if a particular customer's requested load is less than their Maximum Sustained Traffic Rate (Tmax) level within their SLA, then they will usually be offered 100% of their requested load if channel congestion (and channel utilization) is low enough. As channel utilization approaches the 95% level, though, congestion starts to develop during short transient periods of time when simultaneous requests for bandwidth from different customers occur. As channel utilization climbs higher than 95%, congestion issues worsen. It is at this point that CMTS congestion control algorithms (such as Weighted Random Early Discard) begin to turn on, delaying and dropping packets from selected Service Flows based on the priority, Tmax, and Tmin settings with each customer's SLA agreement. The dropping and delaying of packets obviously reduces the actual (utilization) load on the channel to be less than the requested load and typically results in degradations in the subscriber QoE levels.

In the past, these CMTS-induced packet drops and packet delays did not cause major QoE problems for subscribers of most network operators, because most network operators kept their channel utilizations well below the 100% level (by, for example, scheduling node splits if average utilization levels exceeded 70%). This had historically been a very viable network planning strategy since the QoE for traditional web traffic types (e.g. Web Browsing and Traditional Streaming Video) degraded very quickly under even small amounts of channel congestion, channel drops, and channel delays.

Today, however, new "compressible" traffic types, such as Adaptive Bit Rate (ABR) Video, have appeared, making it possible for networks to operate reasonably well at what used to be unacceptably high levels of channel congestion. Some of these ABR Video algorithms probe and sense the available bandwidth on the network channels and may even use more bandwidth than required (by switching to HD video resolutions) if extra bandwidth happens to become available. These ABR Video algorithms will also back down to lower resolution video streams if bandwidths are ever found to be lacking due to network congestion.

With continual economic pressures on MSOs, some may be considering ways to minimize their network equipment costs by moving to higher average levels of channel utilization within their DOCSIS networks, relying on the adaptive nature of ABR Video algorithms to benevolently throttle their traffic rates whenever excessive channel congestion occurs. These concepts are feasible, because ABR Video is making up larger and larger percentages of the total network capacity. But can/should MSOs trust the future QoE levels of their subscribers to ABR algorithms that are managed and maintained by third-party content providers on the Internet? Can/will these algorithms throttle their video resolutions (and bandwidths) fast enough to respond to network congestion and ensure that other traffic types (ex: Web-Browsing) are not negatively affected? A study of this topic was conducted by the authors, and it was concluded that Web-Browsing applications multiplexed with ABR Video streams can still suffer in the presence of network congestion. [Clo2] As a result, it seems clear that MSOs who plan to capitalize on the adaptive nature of ABR Video to increase their average channel utilizations must identify techniques to determine the QoE levels of their subscriber applications during the transient periods of time when the channel utilization hits 100%. In particular, they need their QoE Monitoring tools to find ways to guesstimate the composite requested load (in addition to the channel utilization) to determine how much bandwidth is being throttled. Obviously, as the composite requested load grows to be much higher than the channel utilization level (which may be pegged at 100%), more bandwidth is being throttled and MOS scores will drop. This is illustrated clearly by the dropping MOS scores on the right-hand side of Figure 3.

Determining the composite requested load on a particular CMTS channel is a challenging problem because the CMTS congestion control mechanisms that tend to throttle TCP traffic when it exceeds certain thresholds tend to reduce the traffic levels by throttling them at the TCP source. As a result, the actual requested bandwidth levels may not actually "show" themselves to the CMTS or the QoE Monitoring tool. But there are ways that the CMTS may be able to infer and approximate how high the requested load levels actually are. One way is to observe the behavior of queues as the CMTS congestion control mechanisms begin to drop and delay packets. The rate at which CMTS queues were growing can help to give some level of information regarding the current requested load levels. Other techniques could also be utilized. But in general, it should be clear that even an approximate measure of the composite requested load levels on a channel can help the QoE Monitoring tool determine how much QoE degradation is being caused by traffic throttling during periods of high congestion.

MOS Scores Based On High Sampling Rates

The measured bandwidth level or packet delay level or packet loss level for a particular subscriber (or for a particular type of subscriber with a particular SLA) is assumed to be accurate for only a relatively short period of time because bandwidth levels and delay levels and loss levels tend to fluctuate quite rapidly on DOCSIS channels. Ideally, these traffic metrics would be measured with very fine granularity (on the order of a few seconds or less), but for most QoE Monitoring tools, the ability to collect average channel bandwidth levels is limited by the rate at which statistics can be collected from the devices that are measuring the bandwidths. This precludes the use of exceptionally high sampling rates.

In the future, bandwidth sampling periods of, say, 1 hour may not be adequate. If a subscriber is not satisfied, they are most worried about the bandwidth levels that are occurring 'now' - not 1 hour ago when the last sample was taken.

Thus, it is possible that future QoE Monitoring tools will be more tightly integrated with the CMTSs and CMs and the other elements that are typically taking the bandwidth measurements.

As an example, future CMTSs and CMs may be required to collect measurements every few seconds and perform some post-processing of the data, sending the QoE Monitoring tools only anomalous results that they observed during their post-processing.

MOS Scores Based On Different Scopes

MOS scores within a QoE Monitoring tool can be created for many different scopes, and they all can have value. As an example, measurements can be taken to determine the traffic attributes (bandwidth, latency, jitter, and loss) at the TCP/UDP session level, at the Service Flow level, at the modem level, at the DOCSIS channel level, or at the DOCSIS Service Group level. Each of these measurements at each of the different scope levels yields different pieces of information.

For example, when monitored at the modem level, these measurements can be used to determine if there is congestion and QoE degradation that is being created by large amounts of traffic entering a single home. This can be useful information that an MSO can use to help determine when a subscriber is a candidate for a Service Tier upgrade recommendation.

When monitored at the DOCSIS channel level or DOCSIS Service Group level, these measurements can be used to determine if there is congestion and QoE degradation that is being created by large amounts of traffic generated within the subscriber's neighborhood. This can be useful information that an MSO can use to help determine when a node-split is required or when extra channels need to be added to augment the capacity of a DOCSIS Service Group.

These two situations (modem congestion vs DOCSIS channel congestion) are quite different from one another and may require different MSO responses, so it may be important for a QoE Monitoring tool to be able to monitor and differentiate between the two situations.

An Example of a Future QoE MOS Scores

Future QoE MOS Scores

In the previous sections, we described several ways to augment the typical measurements that are oftentimes utilized within QoE Monitoring tools. In this section, we will give some actual examples of future QoE MOS scores that attempt to make use of bandwidth levels, traffic types, and SLAs.

As suggested in the preceding material, measured per-subscriber metrics can be used as good predictors of QoE. One of the most obvious applications of this within a QoE Monitoring tool is to predict QoE for video viewing subscribers based on the bandwidth that each subscriber is receiving. Another application is to monitor the bandwidth and determine its impact on typical Web Browsing experiences.

From the channel capacity and the channel utilization levels (or from more advanced techniques that utilize active subscriber counts and SLA levels and CMTS scheduling algorithm knowledge), a QoE Monitoring tool can calculate the predicted amount of bandwidth that a particular subscriber (or type of subscriber) is likely receiving at a given instant in time. Once that predicted bandwidth level is calculated, the next obvious question that must be answered is whether that predicted bandwidth level is adequate to offer good QoE levels to the different applications that the subscriber might be utilizing. This question is answered by developing MOS scores for each of the applications, where the MOS scores are potentially a function of the application type, the predicted bandwidth level, and the SLA level for the particular subscriber.

In this section, we will focus on MOS scores for two popular application types (IP Video applications and Web-Browsing applications), but it should be clear that MOS scores can (and will) be developed for the many other types of applications that high-speed data subscribers utilize.

In developing these MOS scores, it was assumed that bandwidth bottlenecks for subscribers are often caused by the narrower bandwidths encountered on the DOCSIS network. This assumption may or may not be correct at all points in time, because it is, of course, quite possible that there are other bottlenecks in the data path between the source and destination of any Internet exchange. These other bottlenecks could occur in the servers or in the backbone network or in the subscriber's own home network. As a result, the development of MOS scores based solely on measurements taken on the DOCSIS network is, by nature, very subjective, and erroneous conclusions can sometimes be drawn about a user's QoE level. Nevertheless, the authors would argue that there is value in these DOCSIS-oriented MOS scores.

Another potential source of error in these MOS scores could result from low sampling rates (which was discussed above). The bandwidth levels offered to a particular subscriber are expected to vary over time and technological change. As a result, the QoE Monitoring tools must continually be adapted to keep up with the required sampling rates for the traffic types that appear on the networks.

Example Future MOS Scores For IP Video

The authors carried out first-order human factors experiments on several IP Video services. Various bandwidth limits were specified, and the QoE level of the resulting IP Video viewing experience was ascertained. These results are less than statistically perfect because many more subjects would have been required to carry out a more thorough study. Nevertheless, the results still provide some degree of information about the quality of IP Video viewing with different amounts of bandwidth being offered to the subscriber.

These results quickly revealed that many of the popular video aggregation services (such as Hulu, VUDU, Netflix, etc.) provide different qualities of service (Standard Definition, High Definition, etc.). Some (but not all) of these service can be *adaptive.* In other words, some of them sense the actual bandwidth available on their TCP session propagating over Internet connections, and they decrease or increase the quality of the video delivered based on this bandwidth.

Another important consideration is that some of these services sense the type of consumer device being used and adapt the BW needed accordingly. For example, a tablet with a small screen would be sent a lower resolution (and lower bandwidth) feed than a 'smart TV' with a 56" monitor.

Table 4 contains some experimentally determined, unconstrained bandwidths required for a number of aggregation services and target end-customer devices. These are the bandwidths which the services would (apparently) elect to utilize if there were no bandwidth limitation between the server and the client. For example, end customers who had chosen to use VUDU (a rental movie service), and who had opted and possibly paid extra to receive a super-high resolution video, would require approximately 9.5 Mbps to be fully satisfied with the service. However a viewer watching the same movie in Standard Definition Netflix on a tablet would not benefit from a bandwidth over ~1.5 Mbps.

(Note: This table is not exhaustive. As an example, Netflix advertises that they also offer higher-definition streams requiring rates of 3 Mbps, 5 Mbps, and 8 Mbps. However, the experiments performed for this paper were unable to access those higher tier services).

IP Video Application	**CPE**	**Avg Mbps**
Netflix	Tablet	1.5
	PC	2.9
	Roku	4.2
Hulu	Tablet	1.3
	PC	1.3
	Roku	2.8
VUDU	SD-PC	2.4
	SD-Tablet	2.4
	SD-Roku	2.3
	HD-Roku	4.7
	HDX-Roku	9.5
You Tube	PC	0.8
	Tablet	2.4

Table 4: Unconstrained Bandwidth Requirements for "Excellent" IP Video Service

Since nearly all of the services had mechanisms for providing *some* flavor of recognizable video down to bandwidths of around 0.8 Mbps (by way of adaptation), viewers would still be able to watch programs down to that bandwidth level, but they may not be very satisfied by the low resolutions. At bandwidth levels below about 0.8 Mbps, most of the IP Video services seemed to break down entirely, resulting in program halts and unsatisfactory QoE levels.

The authors devised a first-order MOS scoring algorithm to provide MSOs with bandwidth vs MOS score guidelines. (Note: It is important to note than any IP Video MOS scoring algorithm will become obsolete soon after it is released because the IP Video content aggregators are continually changing their ABR algorithms and changing the nature of their content. As a result, IP Video MOS score algorithms must be updated periodically to stay fresh and useful).

As a first-order approximation, the authors would suggest that MSOs utilize the required unconstrained bandwidth levels listed in Table 4 and associate it with the relatively high MOS score value of 4.8. (Note: Recall that a perfect MOS score of 5.0 corresponds to excellent service). As the predicted bandwidth level of a user falls below the unconstrained bandwidth level within Table 4, then the MOS score will obviously drop. For the purposes of this first-order model, we will assume that the MOS score drops roughly linearly as the predicted bandwidth level drops. We will assume that this linear drop approaches an MOS score of 2 (poor) as the predicted bandwidth level for a subscriber drops to 0.8 Mbps. If the predicted bandwidth level drops below 0.8 Mbps, then the MOS score is assumed to instantly fall to a value of 1 (bad), because the resulting IP Video is no longer deemed to be useable. These MOS scores tended to correlate fairly well with the IP Video viewing experiences of the authors.

From Table 4, it becomes clear that the MOS score for a particular subscriber would not only be dependent on the predicted bandwidth level available to the subscriber, but it would also be dependent on what particular IP Video service the subscriber is accessing and the device on which he or she is watching it. This is reflected in Table 4.

Example Future MOS Scores For Web-Browsing

The QoE for a Web-Browsing application is predominantly determined by the time between a subscriber’s 'click' on a web page’s hyperlink and the time when the end-user has the perception that the web page is (for the most part) downloaded and displayed on his or her screen. Since web pages vary widely in size, the QoE for different customers clicking on different URLs is a bit difficult to evaluate. We therefore used published average web page size estimates for our evaluations.

The size of the average webpage has grown over the last few years to roughly 1.3Mbytes, or about 10Mbits. [Inf1] (Note: As with IP Video, web page sizes also vary over time. As a result, QoE Monitoring tools must be periodically upgraded so that they can be trusted to perform their Web-Browsing calculations using appropriate web page sizes).

If we assume that an end subscriber is fully satisfied (i.e.- with a MOS score near 5.0) whenever they perceive the downloading of *an entire* web page of this size to occur within, say, 1 second, then he or she would require a minimum bandwidth of about 10Mbps to achieve this high QoE level.

Shorter web pages or longer web pages would, of course, require proportionately lower or higher bandwidths to achieve the desired one second download time goal.

However, complicating the relationship between bandwidth and Web-Surfing QoE is the fact that, very often, a customer will be quite happy with what he or she sees on the screen BEFORE the entire web page is downloaded. We name this time the 'splash time' and find that (most often) only about 1/2 of the total page's data content is required to create a readable screen, with the remaining bytes for the rest of the screen coming in after the splash time has passed.

We therefore assume that for the average-sized web page of 10 Mbits, the web page size associated with the splash time would be ~5 Mbits. Thus, if a single subscriber is receiving a predicted bandwidth of P, then the average splash time (in seconds) for that subscriber can be calculated by:

Average Splash Time = (5 Mbits)/P (3)

where the predicted per-subscriber bandwidth P is measured in units of Mbps.

Given these considerations, the authors would recommend that we associate a Splash Time of 1 second to an MOS score of 4.8, and we would also recommend that we associate a Splash Time over 10 seconds (unbearable to nearly all users) to an MOS score of 1.0. Note that these ratings are for average web pages of size 10 Mbits. A resultant table of Mbps vs. MOS would be represented by Table 5.

Predicted Subscriber Bandwidth (Mbps)	Average Splash Time (seconds)	MOS Score
>5	<1	4.8
4	1.3	4
3	1.7	4
2	2.6	3
1	5	2
.5	10	2
<.5	>10	1

Table 5: Subscriber Bandwidth vs. Splash Time vs. MOS Score For Web-Browsing

Combining Future MOS Scores For IP Video And Web-Browsing

For purposes of illustration, Figure 4 graphically summarizes the above information relating Predicted Subscriber Bandwidth to MOS scores for various applications. Note, for example, that the light green, rightmost line, represents the most demanding application shown- HDX (Very

High Definition) VUDU on a large screen Roku-Connected TV. This application would therefore require a subscriber bandwidth of ~10 Mbps to completely satisfy a customer who had paid for this service. In the middle of the graph, a user involved in Web-Browsing would very likely be exceedingly happy and have a high MOS score with a bandwidth of 6 Mbps. On the left side of the graph, a standard Netflix viewer on a PC with a bandwidth of 2.0 Mbps would be exceedingly happy and have a high MOS score.

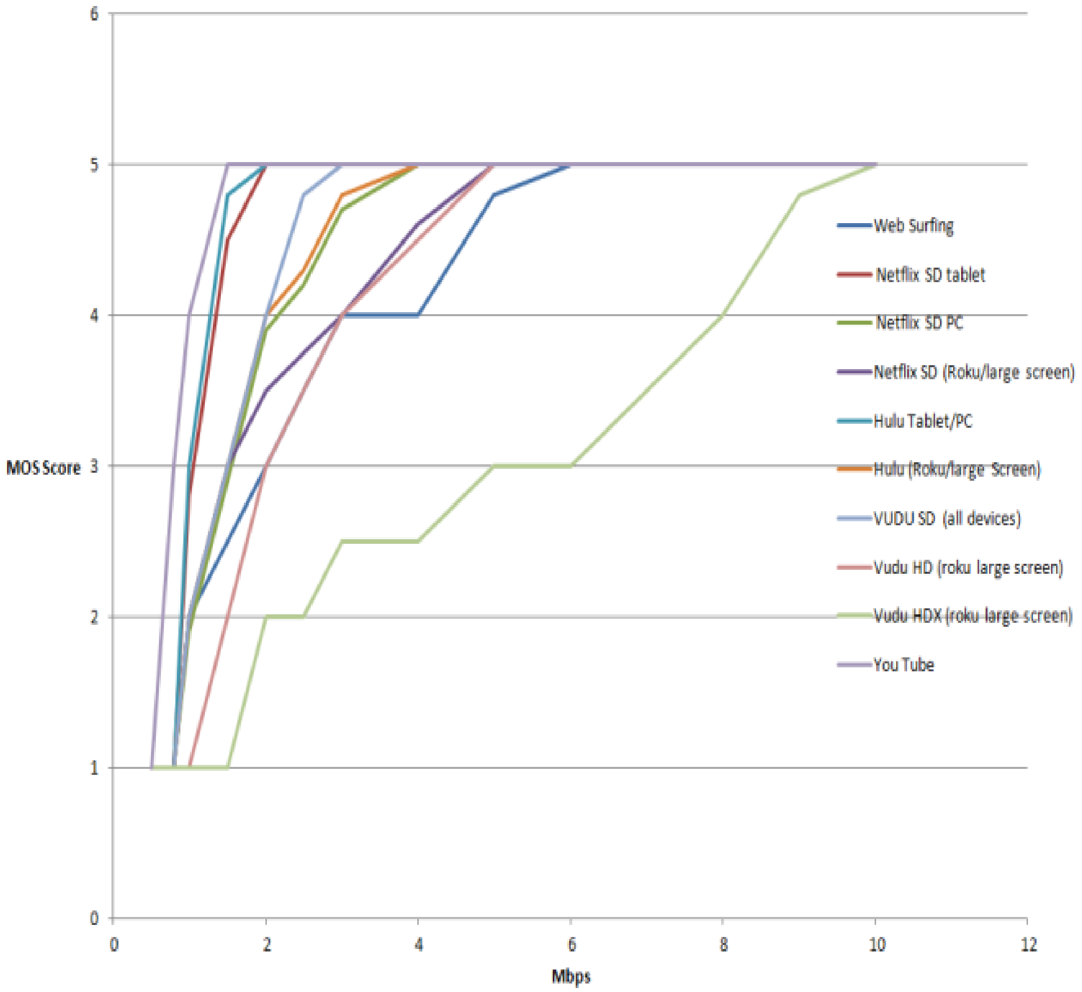


Figure 4: Bandwidth (Mbps) vs. MOS Score For Different Applications

A few final points should be made about the application of the plots in Figure 4. If we select a particular bandwidth value (say 4 Mbps), we find that each application can have a very different MOS score. As a result, MOS scores are very sensitive to the application being utilized.

The plots can be used once the bandwidth for a subscriber has been calculated. This calculation can be performed using simple algorithms (ex: total channel bandwidth divided by number of active users) or using complex algorithms (ex: bandwidth calculations based on CMTS scheduling algorithms and SLAs). But regardless of the type of algorithm used, the end result is the creation of a predicted per-subscriber bandwidth level on a particular channel or service group.

The plots can be used to predict the QoE levels for a particular subscriber, in which case the average bandwidth for that subscriber can be calculated by monitoring his or her actual usage levels.

The plots can also be used to predict the QoE levels for generic subscriber types, in which case the average bandwidth for typical users (that share a particular SLA level) would need to be calculated.

Once the average bandwidth (for a particular subscriber or for a generic subscriber type) is calculated, the QoE Monitoring tool can assume that all of the available bandwidth is used by one application, or the QoE Monitoring tool can assume that multiple applications within the subscriber's home are sharing the available subscriber bandwidth. If the latter is assumed, then more complicated formulae are required to determine how much of the available bandwidth should be allotted to each of the applications that are assumed to be operating within the home.

In any case, there will be a specific amount of bandwidth that is assumed to be available for a particular application. At this point, the MSO or the QoE Monitoring tool can access the plots of Figure 4.

Unless Deep Packet Inspection tools are being used in conjunction with the QoE Monitoring tool, it is unlikely that the QoE Monitoring tool will actually be cognizant of the traffic types that are actually passing to the subscriber at any point in time. As a result, the plots in Figure 4 can be utilized to describe the QoE levels for the subscriber IF they were currently accessing different traffic types (whether they are actually accessing those traffic types or not). Thus, the resulting MOS scores that are pulled out of the plots in Figure 4 are actually the MOS scores for hypothetical subscribers assumed to be using that service - they are NOT actual subscriber MOS scores. It is important to note this distinction.

If an MSO does not know exactly which application a subscriber is using at any given time, an obvious question comes to mind. How can the MSO predict the QoE levels of a subscriber without that knowledge? One way to do it is to create the concept of a "typical subscriber model." An MSO can construct this model using knowledge about the mixes of applications that their aggregate set of subscribers use on the Internet. For example, if an MSO knows that (on average) the mix of traffic types is given by 54% IP Video traffic, 18% Peer-to-Peer traffic, 11% Web-Browsing, 17% other traffic, the MSO can assume (for modeling purposes) that a "typical subscriber" might be temporally transitioning between the different traffic types such that his or her average bandwidth mix equals the percentages above. For the 54% IP Video Traffic mix listed above, the MSO could further try to determine the mixes of different types of IP Video traffic on the network (ex: Netflix, VUDU, etc.), and the MSO could then assume that the "typical subscriber" is temporally transitioning between video streams from those different IP Video traffic types whenever they are supposed to be viewing IP Video. As a result, a blended MOS score can be created by creating a weighted average of all of the MOS scores for all of the heavily used applications. This blended average can be useful in ascertaining whether a "typical subscriber" is currently experiencing high QoE levels or not. If more detailed information is desired, then the MSO can dive into the details of each individual MOS score that contributed to the weighted average.

Example of Future QoE Monitoring Tool Displays

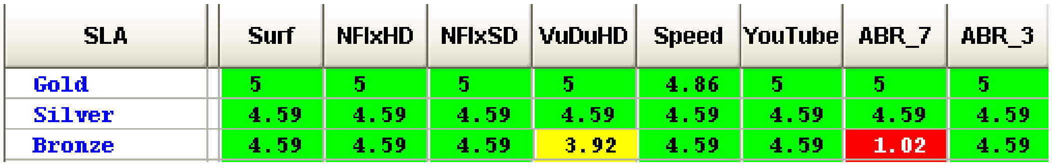
This section discusses an implementation of a QoE Monitoring tool that is possible to achieve with today's CMTS technology. We base this on the observation that every existing CMTS must currently incorporate some mechanism for resolving downstream channel congestion. We further observe that (regardless of the congestion resolution mechanism used) we should be able to model, to a reasonable degree, the current state of that mechanism as a mapping of the DOCSIS SLA parameters (Priority, Tmax, Tmin, etc.) from their SLA-specified values to a set of "Effective Values" that will limit the aggregate downstream data rate to fit into the available downstream bandwidth. In order for this process to be "fair," rather than random and arbitrary, we believe that most CMTS systems must currently maintain, in some fashion, a set of "state" information that would reveal this mapping for the current level of channel congestion.

While the volume and complexity of the information required to communicate this mapping may vary widely among CMTS vendors, we have shown, by the example described here, that in some cases this mapping can be summarized and communicated in a surprisingly small amount of data for each downstream channel. In the experiment described here we were able collect this congestion state information for all CMTS downstream channels at intervals as short as 15 seconds – which allows us to collect congestion-state information with a high degree of time resolution and without burdening the CMTS with a large volume of I/O.

Once we have this capability of mapping SLA-specified DOCSIS parameters (Priority, Tmax, Tmin, Tpeak, maxBurst, etc) into Average Bandwidth per Subscriber Service Flow values based on the current level of channel congestion, it is then a relatively simple matter to "predict" the length of time required to transfer a data block of any given size and, in turn, the resulting equivalent data rate and the resulting response time for various Web applications.

If we now select some data block size that is "typical" for a given type of internet service (e.g. a typical web page size for web browsing, or a 2-second block of video data) we can predict, with reasonable accuracy, both the response time and effective data rate that a customer would experience in requesting that data block (as a function of that customer's SLA parameters).

Finally, a table mapping the effective data rate or response time for that hypothetical internet service to an MOS Score (based on the subjective judgment of representative users) can yield a useful measure of a customer's expected QoE.



SLA	Surf	NFlxHD	NFlxSD	VuDuHD	Speed	YouTube	ABR_7	ABR_3
Gold	5	5	5	5	4.86	5	5	5
Silver	4.59	4.59	4.59	4.59	4.59	4.59	4.59	4.59
Bronze	4.59	4.59	4.59	3.92	4.59	4.59	1.02	4.59

Figure 5: MOS Score Matrix Display

The complete QoE status for a channel at any given point in time can be displayed in the form of a matrix as shown in Figure 5. This display is able to simultaneously display an MOS Score for every combination of SLA (Gold, Silver or Bronze) and Internet Service Type. (The service types corresponding to the eight column headings shown in this example are, respectively: Web Surfing, SD Netflix, HD Netflix, HD VuDu, Speed Test, YouTube, 7 Mbps ABR Video, and 3 Mbps ABR Video).

It is important to remember that a customer's quality of experience is affected not only by network congestion but also by their choice of SLA subscription. In the example in Figure 5, the customer's anticipated modest satisfaction (3.92) with VuDuHD and complete dissatisfaction (1.02) with any attempt to view 7 Mbps ABR Video is a consequence of the Tmax value of 5 Mbps in their SLA.

While the matrix format shown above can accommodate a large number of SLA and Service Type definitions, it lacks a time dimension that would permit the display of chronological trends for individual values. A chronological display, on the other hand, as shown in Figure 6, can display a wealth of time-dependent detail, but becomes confusing if we attempt to include too many SLA/Service Type combinations. (This example has only included five of the eight Service Types.)

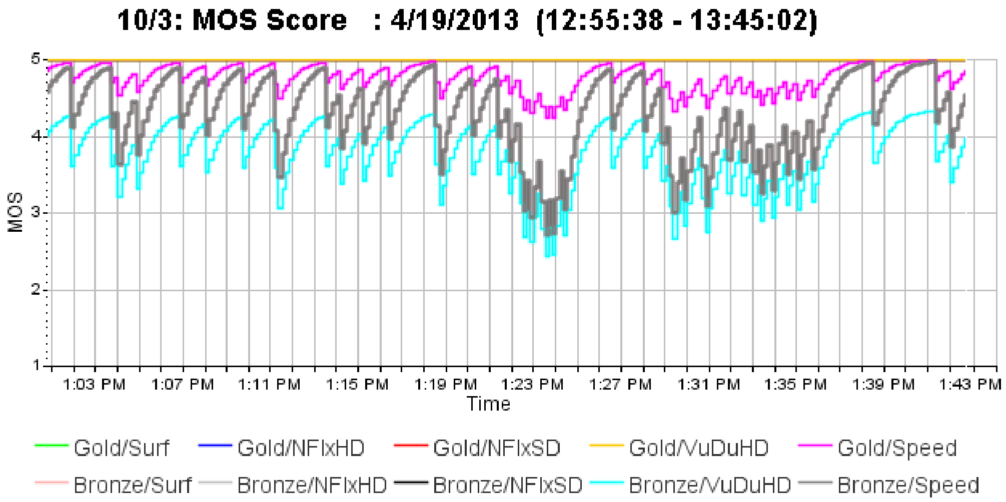


Figure 6: MOS Score Chronological Display

The very short sampling intervals (as low as 10 seconds) possible with this approach can provide a high degree of chronological resolution without massive data collection overhead – thus capturing congestion detail that would be completely lost with longer sampling intervals of 15 minutes or an hour.

Conclusions and Future Possibilities

This paper has attempted to answer several important questions about next-generation DOCSIS QoE Monitoring tools:

(1) Why should we suddenly be interested in a family of MOS scores and metrics that we have comfortably ignored for so long?

(2) What is it about our DOCSIS networks that has changed to require the use of more sophisticated MOS scores and metrics?

Both of these questions have the same answer. DOCSIS networks have historically used only a single primary metric (channel utilization) when attempting to determine QoE performance – and that metric is quickly losing its efficacy due to several facts:

(1) More bandwidth capacity is being offered within the SLAs of DOCSIS subscribers

(2) More applications with very different characteristics and a wider range of bandwidth needs are now sharing DOCSIS channels

(3) More bandwidth is being used by each subscriber application

(4) Each application is sensitive to different QoE metrics (bandwidth, delay, jitter, packet loss, etc.)

(5) Each application has different acceptability thresholds defining good QoE levels

(6) Benevolent applications (like ABR IP Video) that yield bandwidth to others are sharing channels with greedy applications (like Peer-to-Peer)

(7) ABR IP Video is a novel traffic type that is "sponge-like"- growing to use more bandwidth when the bandwidth is available and shrinking to use less bandwidth when the bandwidth is not available

(8) Single subscribers can now have a strong negative impact on all of the other subscribers sharing a channel

The authors believe that QoE Monitoring tools and QoE metrics will need to undergo a rapid change to help fill the instrumentation void caused by basic inadequacies of simple channel utilization measurements from the past. In the future, MSOs are likely to need multiple QoE metrics for their DOCSIS high-speed data service tier (instead of the single metric of channel utilization that is predominantly utilized today).

Because DOCSIS networks are carefully designed to always provide customers with the highest quality of service for which they have subscribed (unless it is physically impossible to do so), one might expect any measure of QoE to indicate full customer satisfaction unless global channel congestion levels or inadequate DOCSIS Tmax levels are limiting the subscriber's bandwidth to a level less than desired. If it is the former (channel congestion) that is causing the QoE problem, then the MSO may need to make changes to upgrade their network. If it is the latter (inadequate Tmax levels) that are causing the QoE problem, then the subscriber may need to make changes to upgrade their SLA level.

As a result, it is necessary for QoE Monitoring tools to make use of channel utilization as a metric to indicate whether the problem lies in the MSO's camp or the subscriber's camp. But monitoring of other metrics (either direct or indirect) can also be used to help the MSO know exactly which types of applications may be experiencing problems at any point in time. When coupled with more detailed information about subscriber SLA levels and CMTS scheduling algorithms, these tools can become quite intelligent as they predict the QoE levels for subscribers.

This paper has shown that many different improvements and augmentations can be added to future QoE Monitoring tools. In addition, the paper has shown how MOS scores of the future can take advantage of the many improvements that are listed. Finally, the paper showed an example of a QoE Monitoring tool and illustrated some useful display models.

Several key ideas were outlined within this paper. First and foremost, it was shown that knowledge of traffic types, subscriber SLA levels, and traffic metrics (such as bandwidth, delay, jitter, and packet loss) can help QoE Monitoring tools create much more specific and accurate predictions about subscriber QoE levels. In addition, it was shown that the ability of a QoE Monitoring tool to predict the QoE performance levels for different traffic types can be greatly improved if the QoE Monitoring tool is made cognizant of the particular scheduling algorithms used by the CMTS. With this knowledge, the QoE Monitoring tool can make more informed decisions about how the traffic will be handled by the CMTS during periods of congestion and non-congestion.

The fact that the QoE Monitoring tool is improved by receiving information from (and about) the CMTS leads the authors to wonder whether the opposite is also true. If information from the QoE Monitoring tool could be rapidly coupled back into the CMTS, could the CMTS make good use of that information? Is it possible that future CMTSs could dynamically modify their scheduling algorithms to try to assist a particular traffic type that might be experiencing low QoE levels at a particular instance in time? Or is it possible that load-balancing algorithms inside of CMTSs could make good use of the outputs from the QoE Monitoring tool to re-balance CMs attached to the channels on a CMTS? Even further, is it possible that future CCAP elements could initiate dynamic QAM sharing functions to modify the number of EQAM channels and DOCSIS channels in response to low DOCSIS MOS scores from the QoE Monitoring tool? Only time will tell if these interesting concepts might eventually find their way into real-world applications as QoE applications continue to evolve.

References

(1) [Clo1] T. Cloonan and S. Chernak, "HFC Spectrum Planning in the Era of the HSD Bandwidth Explosion," Proceedings, SCTE Cable Tec Expo '12, October, 2012.

(2) [Clo2] T. Cloonan and J. Allen, "Competitive Analysis of Adaptive Video Streaming Implementations," Proceedings, SCTE Cable Tec Expo '11, November, 2011.

(3) [Inf1] http://www.infodocket.com/2013/03/06/statistics-the-average-size-of-a-web-page-is-now-1-3-mb-up-33-in-last-year/