Taking the Guesswork Out of Capacity Management

A technical paper prepared for presentation at SCTE TechExpo24

Sung-eun Kim
Principal Architect
Cox Communications, Inc.
Sung-eun.Kim@cox.com

Carol Ansley
Senior Lead Architect
Cox Communications, Inc.
Carol.ansley@cox.com

Ted Boone
AVP, Technology
Cox Communications, Inc.
Ted.boone@cox.com

Table of Contents

List of Figures

1. Introduction

This paper provides results from an in-depth analysis of premises gateway speed test performance testing aimed at facilitating robust capacity management strategies within broadband networks. The study focuses on downstream and upstream speed distributions vs advertised maximum speeds, latency characteristics, and their variations across different parameters such as service tiers, time of day, access technologies (Hybrid Fiber-Coaxial (HFC) vs. Passive Optical Network (PON)), customer-initiated vs. random events, site locations, and customer premises equipment (CPE).

The analysis revealed insights into the speed distribution patterns, exhibiting variations across service tiers and hourly fluctuations. Latency distributions scrutinized with respect to access technologies, discerning disparities between HFC and PON deployments, as well as differentiating between customer-initiated and random samples. Moreover, latency distributions across various sites and CPE configurations are analyzed to understand localized performance dynamics.

Furthermore, the paper discusses potential applications for effective network capacity management. Insights gleaned from speed distributions aid in optimizing bandwidth allocation, especially during peak hours, ensuring equitable service delivery across different service tiers. Analysis of speed performance and latency enables the identification of bottlenecks and metrics for the optimization of resource allocation to enhance network responsiveness and reliability. Additionally, the data can inform proactive maintenance strategies by identifying areas prone to performance issues or CPE-related performance constraints.

Overall, this study underscores the significance of comprehensive network performance analysis in informing capacity management strategies. Leveraging insights derived from speed distributions, latency characteristics, and associated parameters enables operators to proactively optimize network resources, enhance service quality, and ensure optimal customer experience in broadband networks.

2. Automated Performance Testing Infrastructure and Results

Cox Communications chose to invest in an automated testing infrastructure utilizing an integrated application in residential gateways, a standardized network of test servers, and a test management infrastructure. The tests include both downstream and upstream speed tests as well as latency characterization.

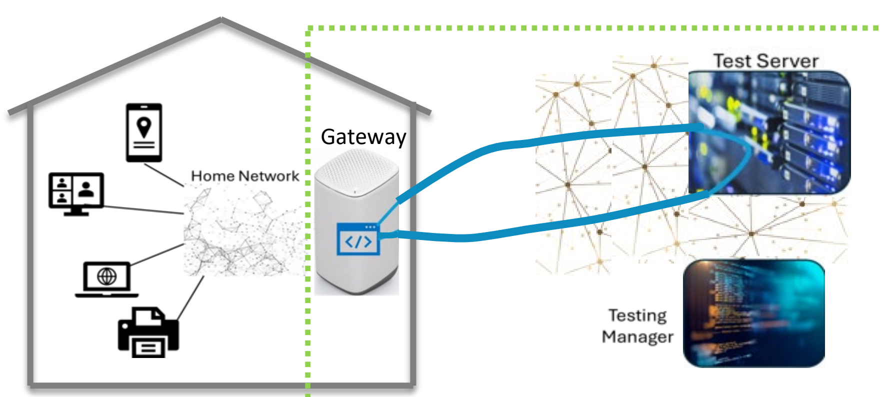

Figure 1 – Testing architecture

The integration of a test application into widely deployed residential gateways, connected directly to the access network, allows testing to isolate the performance of the network without adding the variability introduced by home networking equipment and the customer's connected devices. Results from some of the external testing services can have a large amount of variability depending upon the connected device used for the testing. For example, a test run with an older phone using Wi-Fi to connect to the home network might have worse results than a test run with a computer that is connected to the home network using wired Ethernet.

The testing infrastructure provides the ability to control the scheduling of tests across the network. For example, specific tiers can be targeted to evaluate the performance of a tier across regions as well as across time-of-day variations. The testing infrastructure can also be accessed by technicians and customers themselves allowing both groups to benefit from the testing infrastructure. No matter what source initiated a test, the results are all stored in a single database for future analysis and study.

Test servers are configured in a uniform fashion and utilize similar hardware so that the test results can be constructively compared with each other.

3. Network Performance Metrics

In this section, we summarize the performance metrics studied including actual downstream speed over advertised download speed, actual upstream speed over advertised upload speed, and latency.

Roughly 2.5 million test results were collected across 1.1 million gateway devices. Each test event measured download speed, upload speed and latency. Measuring period was May 1st to May 31st 2024. Some additional test results are included that were run in June. The tests were run from 4 AM to midnight, with an aim of an even time distribution throughout this time window. Midnight to 4 AM local time is excluded to avoid any conflict with operational activities during the maintenance time window. We scheduled a high number of additional tests for 2 Gbps service, our highest HFC service tier. As advertised speeds increases, achieving good performance becomes more difficult. This allows us to test the most challenging cases and gather a sufficient sample size for the speed tier in spite of low penetration.

3.1. Downstream Speed Distributions

Figure 2 represents the complementary cumulative distribution of the ratio of actual download speed to advertised speed by each product tier. It displays not only the median values but also the entire range of the test results including tails of the distribution.

In the Figure, the X-axis represents the ratio of actual to advertised download speed, and the Y-axis represents the cumulative percentage of total tests.

The median of actual/advertised download speed ratio is 117%. 80% of the tests had better than 113% of the Advertised speed as the test's actual speed. The curve shows a steep drop-off around 117%, which means the majority of the tests achieved 117% or better performance. 91% of test results are within 110%-120% range of actual/advertised download speed. 96% of the tests achieved better than 100% download speed over advertised speed.

The median result for the 2Gbps tier ratio is 114%, and 93% of the tests achieved better than 100% performance.

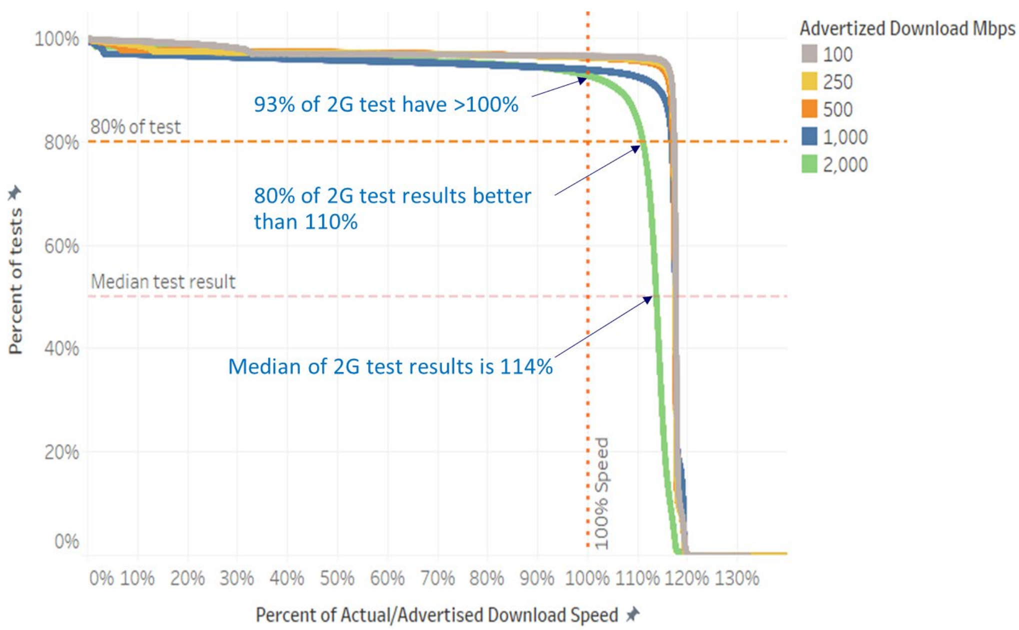

Figure 2 - Complementary cumulative distribution of the ratio of download actual speed to advertised speed

Figure 3 presents the ratio of download actual speed to advertised speed versus the time of day of the test. The left chart shows the median values over the time. The right chart shows the bottom 20% of the values over the time of day, which means 80% of the tests had equal or better than these values.

The test results show consistent speed performance across the day including peak hours. The 2Gbps tier had consistent performance in median values and had 109% during peak hours, which is a slight decrease for the bottom 20% performance.

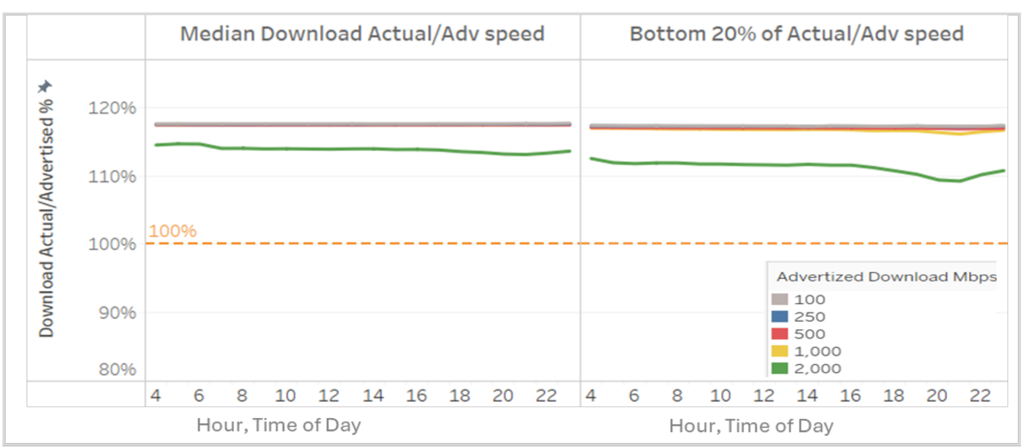

Figure 3 - The ratio of download actual to advertised speed over the time of day, median and bottom 20%

3.2. Upstream Speed Distributions

Figure 4 represents the complementary cumulative distribution of the ratio of actual upload speed to advertised speed by product tier upstream speed. The X-axis represents the ratio of actual to advertised upload speed, and the Y-axis represents the cumulative percentage of total tests.

The median of the actual/advertised upload speed ratio is 113%. 80% of the tests had better than 110% of the advertised upload speed as their actual test result. 92% of the tests achieved better than 100% upload speed over advertised speed.

As the advertised speed increases, it is difficult to achieve 100% performance consistently. Based on the test results, 100Mbps upload speed tier performance still shows consistent performance. The median of 100Mbps tier test results was 112% of advertised speed, and 80% of the tests had better than 107% of the advertised speed. 88% of the tests achieved better than or equal to 100% of the actual/advertised upload speed.

Upload speed offerings differ by node type in terms of DOCSIS® spectrum configurations. 100 Mbps is the maximum upload speed in mid-split nodes and 35 Mbps is the maximum in sub-split nodes. We manage the node status and tier offerings to ensure strong speed performance consistently.

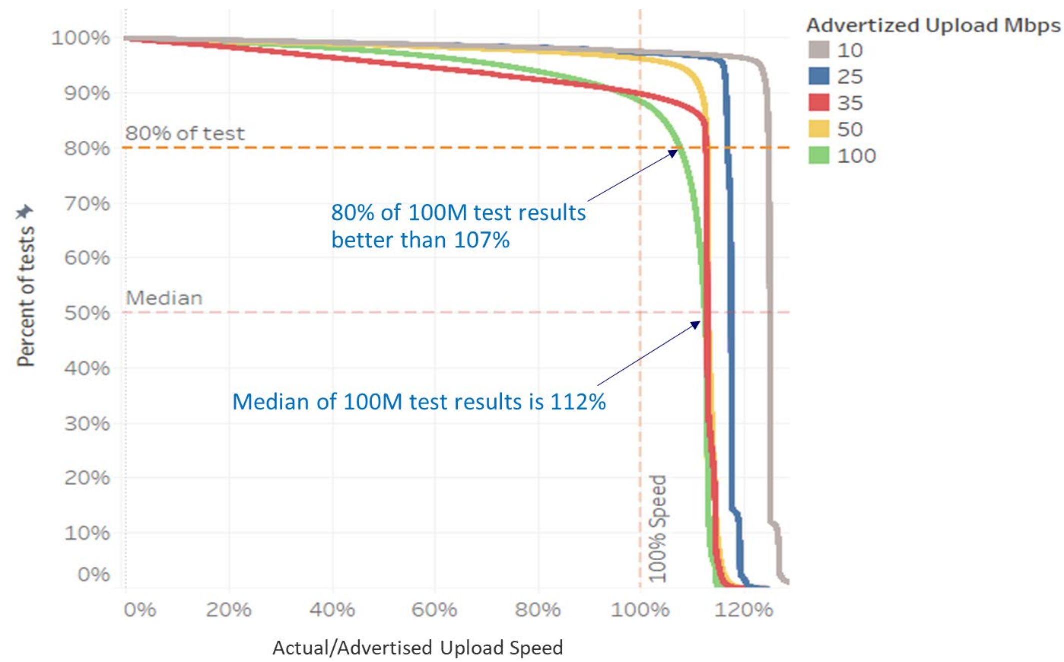

Figure 4 - Complementary cumulative distribution of the ratio of Upload actual to advertised speed

Figure 5 displays the median of the ratio of actual to advertised upload speed across the time of day and bottom 20% performance for various advertised upload speeds.

The left figure shows the median of the ratio of actual over advertised upload speed across the measuring time of day. Median speeds are consistent across all tiers throughout the day including peak hours.

The right figure shows the percentage of the bottom 20% of the ratio of actual upload speeds to the advertised speed over the time of day, which means 80% of the test results are equal or better than these values.

It indicates that the bottom 20% upload speeds also remain relatively consistent throughout the day. However, the highest speed tiers, 100 Mbps in mid-split nodes and 35 Mbps in sub-split nodes, had slight variations in performance, with slight decreases during the peak hours. All of the tiers' bottom 20% performance far exceeds 100% of the advertised speeds.

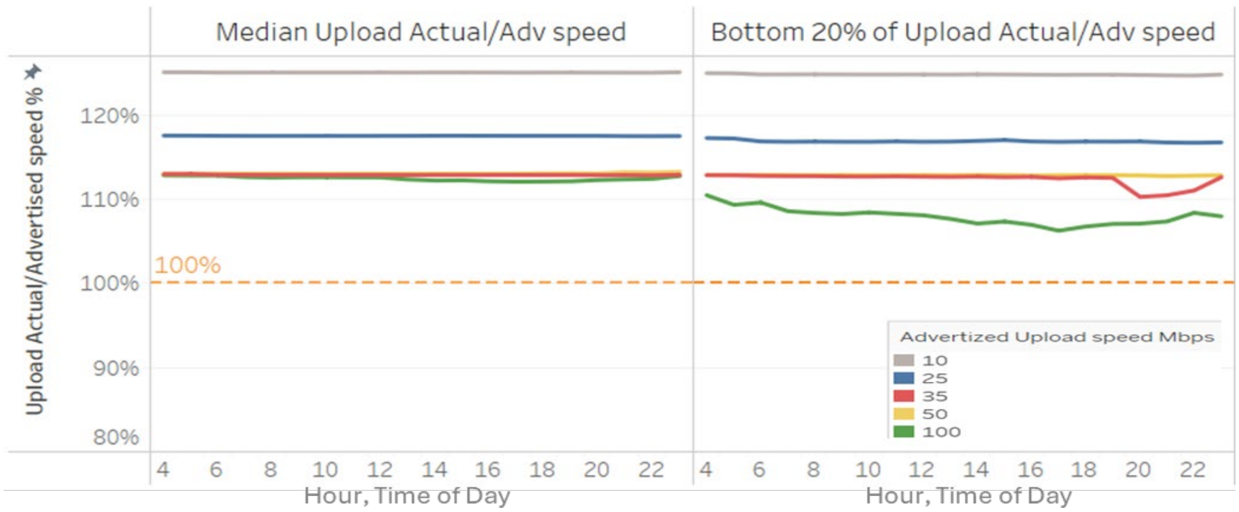

Figure 5 – The ratio of upload actual to advertised speed over the time of day

3.3. Latency Distribution

Latency is one of the most important performance metrics since it directly affects customer experience, especially when they use real-time applications such as online gaming, video conferencing, and live streaming.

As generally known among gamers and game developers, latency around 40-60ms or lower is considered acceptable for most online games, while latency over 100ms can introduce noticeable lag in gaming. The ideal latency is 20-40ms or lower for optimal gaming experiences. Lower latency generally leads to better user experiences in real-time applications.

Figure 6 presents the distribution of latency results. The X-axis shows the latency results in milliseconds. The Y-axis indicates the percentage of tests that had latency results longer than or equal to the corresponding value on the X-axis. The median latency across all service tiers was 13.3ms. 93% of the test results fell within the range of 10ms and 20ms.

93% of the test results had latency under 20ms, while 99% of the test results had latency under 30ms. Conversely, 1% of the test results had latency over 30ms, and 0.5% of the results had over 40ms. 0.25% of the results were greater than 100ms during the measuring period.

In general, higher speed service tiers have lower latency. The measured median latencies ranged from 13ms to 16ms across different tiers. The 2 Gbps service tier had a median latency of 12.5ms, while 100 Mbps service tier had a median of 16ms. Figure 7 presents the latency results across different service tiers. The Y-axis shows the latency on a logarithmic scale to display the bottom 1% of the tail distribution more clearly.

Since even the occasional high-latency packet can have negative impacts on customer experience for some applications, network latency for the 99th percentile should be 40 milliseconds or lower; and the 99.9th percentile should be less than 100 milliseconds.

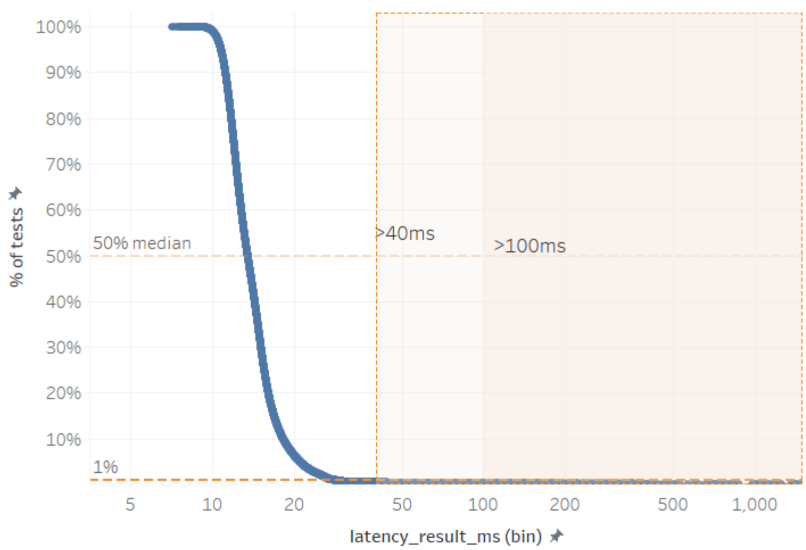

Figure 6 – Latency distribution

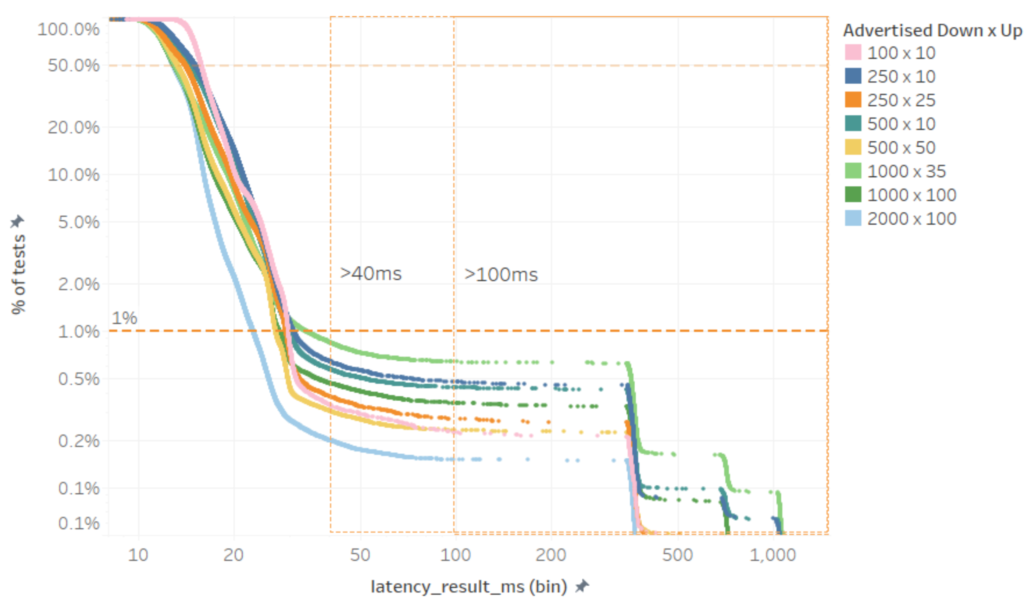

Figure 7 – Latency distribution by Service Tier

4. Potential Applications for Effective Network Capacity Management

The test results from the automated systems apply in many areas to improve network capacity management. The following sections discuss areas in which Cox has found advantages from analysis of the automated system testing results.

4.1. Insights Gleaned From Speed Distributions

Classifying the test results by the service tier of the subscriber provides information about the actual performance of each tier. We found generally that median speed test results for both downstream and upstream speed tiers consistently exceeded advertised speeds. Only the highest speed tiers (in our case 2Gbps downstream and 100Mbps upstream for HFC) showed any portion of test results less than advertised tier, and even that was minor. Satisfying broad customer expectations for these highest tiers presents both a challenge and an opportunity for network capacity management.

For example, if customers broadly expect a 2 Gbps service tier to deliver between 1.5 Gbps and 2.2 Gbps with a median above 2 Gbps, there may be an opportunity to optimize the cost and pace of network expansion.

4.2. Optimizing Bandwidth Allocation, Especially During Peak Hours

Sorting the test results by time of day provided another view of network performance. We found when test results were combined across the entire network or even across a site, the level of performance by time of day had little variation. If the results are sorted even further down to the node level, more variation can be seen.

4.3. Ensuring Equitable Service Delivery Across Different Service Tiers

One interesting insight from the test data is that lower speed tiers' performance is more robust than the highest tiers. This effect can be seen through comparing the performance of different tiers in the same node.

By comparing the median performance from different speed tiers, we found that downstream median performance was generally very consistent versus the different upstream speeds, as shown in Figure 8. Interestingly, we found that the upstream median performance varied by tier, as shown in Figure 9. This result is not unreasonable when one considers that the highest speed tiers make more intensive use of the bandwidth in the upstream and the downstream.

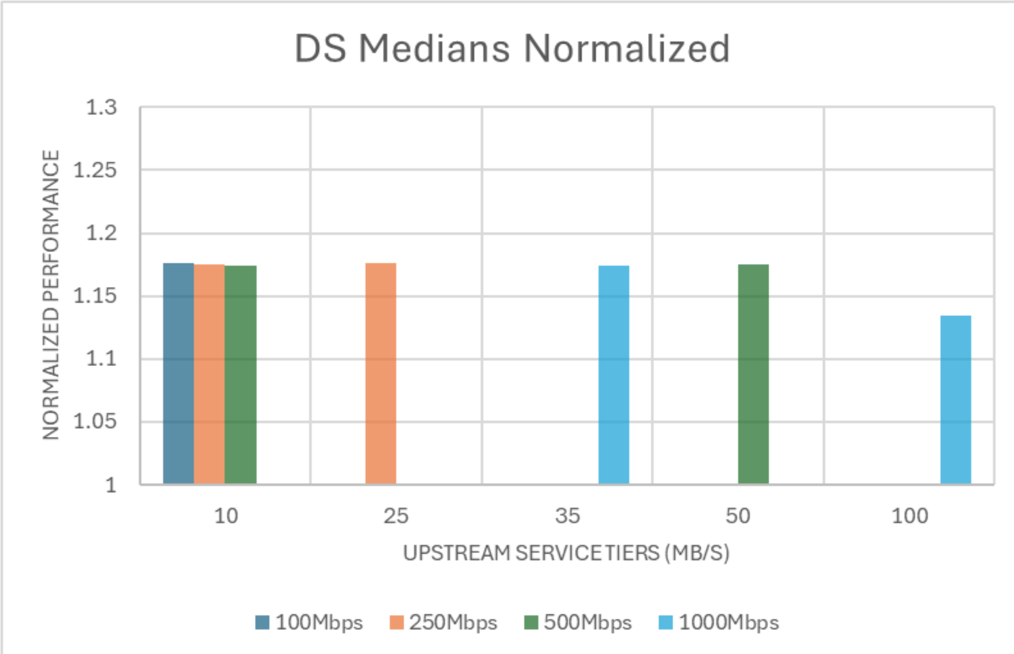

Figure 8 –Downstream median distribution by Service Tier

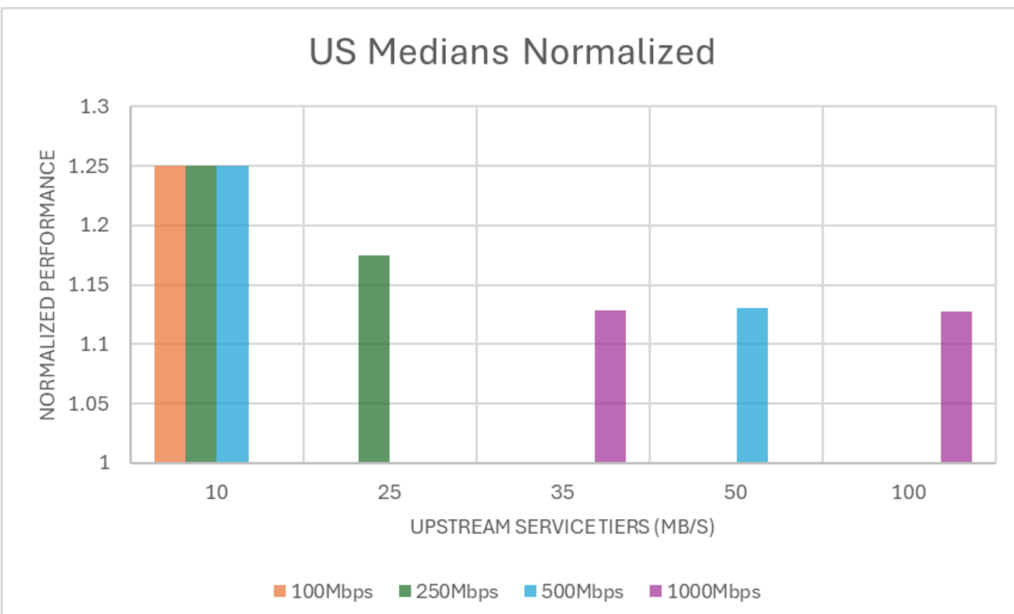

Figure 9 – Upstream median distribution by Service Tier

4.4. Identification of Bottlenecks and Optimization of Resource Allocation from Latency Analysis

Our analysis of latency results indicates issues with ingress and congestion can result in latency increases, but because latency is also directly affected by the distance between the gateway and the test server, latency results need to be examined carefully.

Test server placement should track with the actual traffic flow for a gateway to the nearest Internet connection. For example, as we analyzed test results, we found an area with much higher latency than would otherwise be expected. Upon further examination, we found that devices in that area had a much longer distance to reach a test server due to metro and backbone network configuration, but the actual network traffic was routed more efficiently, so the latency results were not representative of actual performance.

4.4.1. PON vs HFC Latency

Test results show the median PON latency typically ~5 milliseconds lower than HFC. This is likely due to higher interleaver depth for HFC error correction relative to PON.

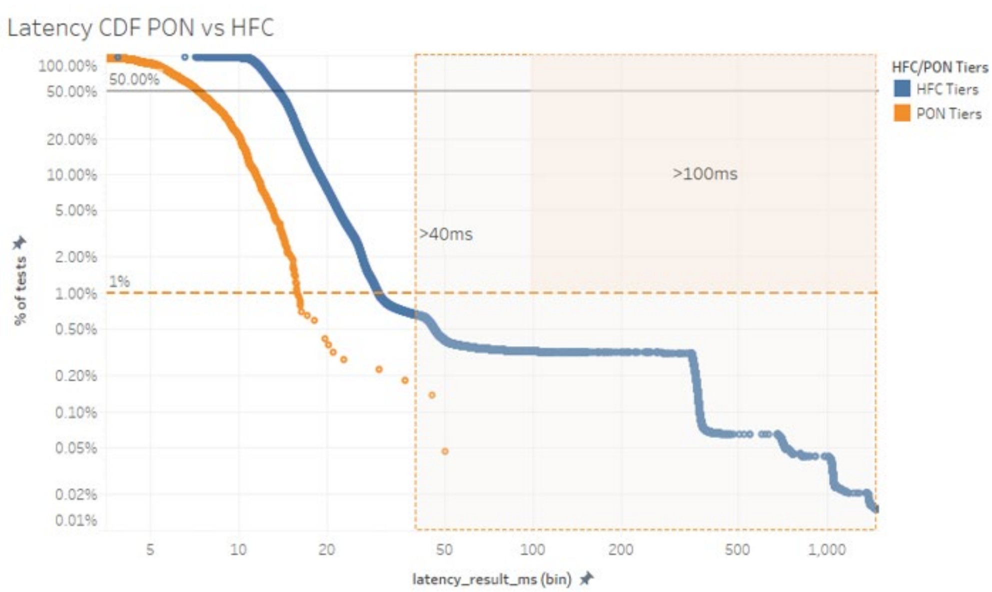

Figure 10 Latency distribution by HFC and PON

4.4.2. Long Tail Latency for HFC

Long-tail latency on the HFC network occasionally stretches much longer than 40 milliseconds. The increase in latency beyond about 40ms is surprisingly steep. This suggests some retry timeout or back-off function of the DOCSIS protocol might be adjusted to remediate. Judging by the measured latency curves, long-tail latency might be improved as much as 250 ms!

4.4.3. Customer-Initiated vs Cox-Initiated Latency

Latency for customer-initiated tests show a long-tail latency roughly 8 times more often than the much broader provider-initiated tests. This confirms that customers are far more likely to initiate speed and latency tests when their network is performing poorly. This suggests an opportunity to enhance support procedures.

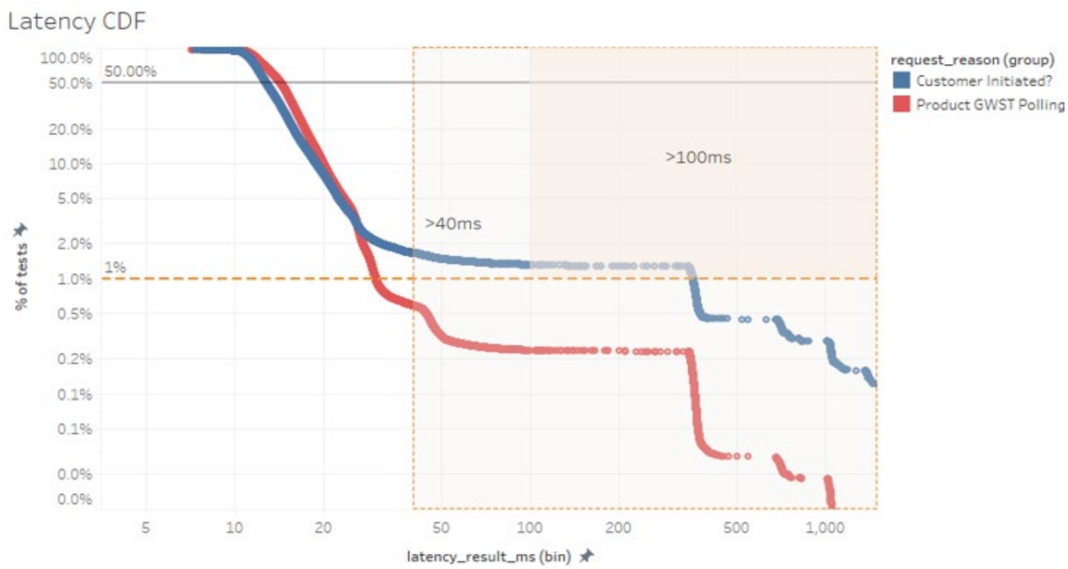

Figure 11 Latency distribution for Customer Initiated Tests

4.5. Proactive Maintenance Strategies

A collection of gateways have been selected for more intensive testing. These gateways are either deployed in headends or employee homes. They are tested once an hour continuously. From these test results, additional observations can be made for network performance. Because not all of the gateways are deployed in the outside plant network, the ones deployed in the headends show very few issues. The ones deployed in the network show more variation that can be traced to network activity. In the graphs below, the test samples have been sorted by device, so each block of roughly 800 samples corresponds to a different device.

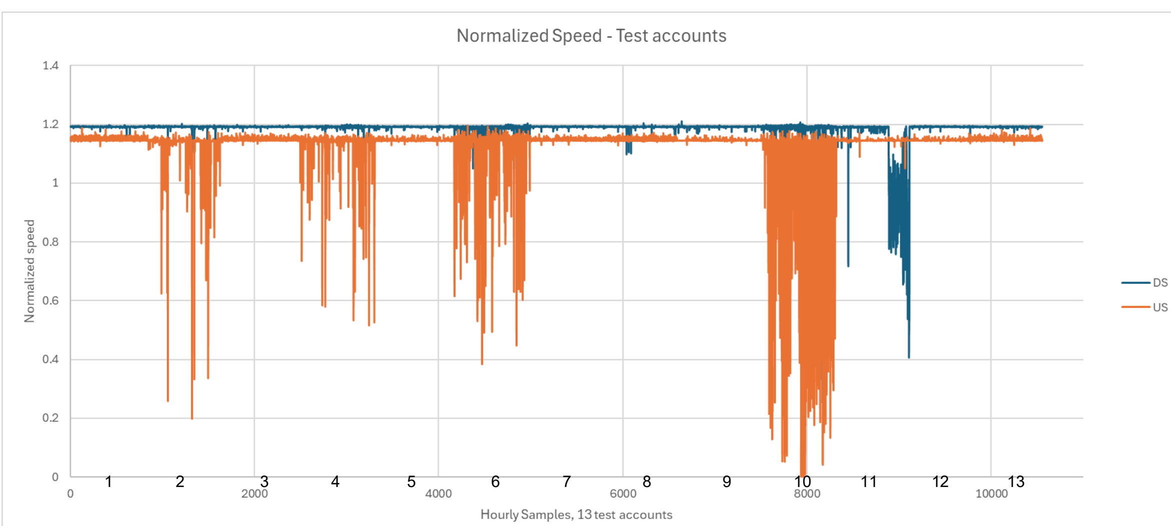

Figure 12– High Frequency Test Devices – Speed Results

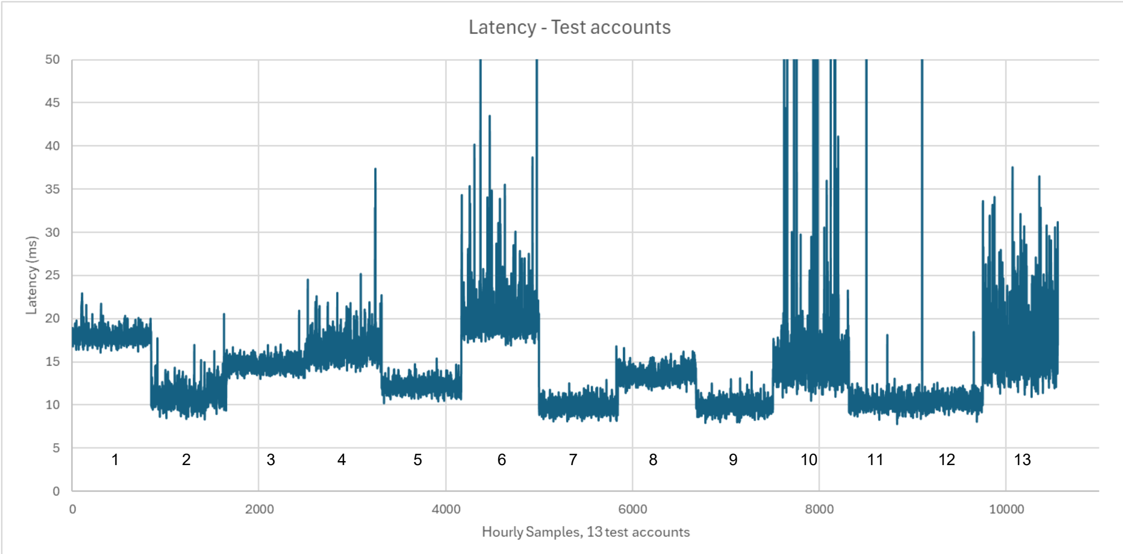

Figure 13– High Frequency Test Devices - Latency Results

One can see variances between different test devices that may be due to the nodes they are attached to or because of their specific network connections. The graphs below are zoomed in to highlight 2 specific devices that may warrant further attention. They are both provisioned the same but show different behaviors across the month of testing.

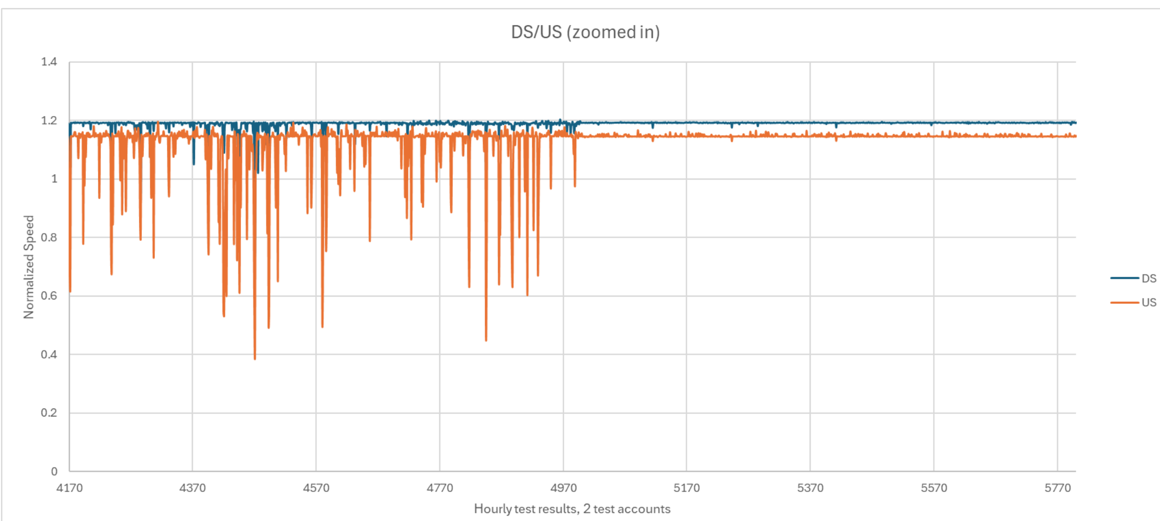

Figure 14– Zoomed in – 2 High Frequency Devices, speed results

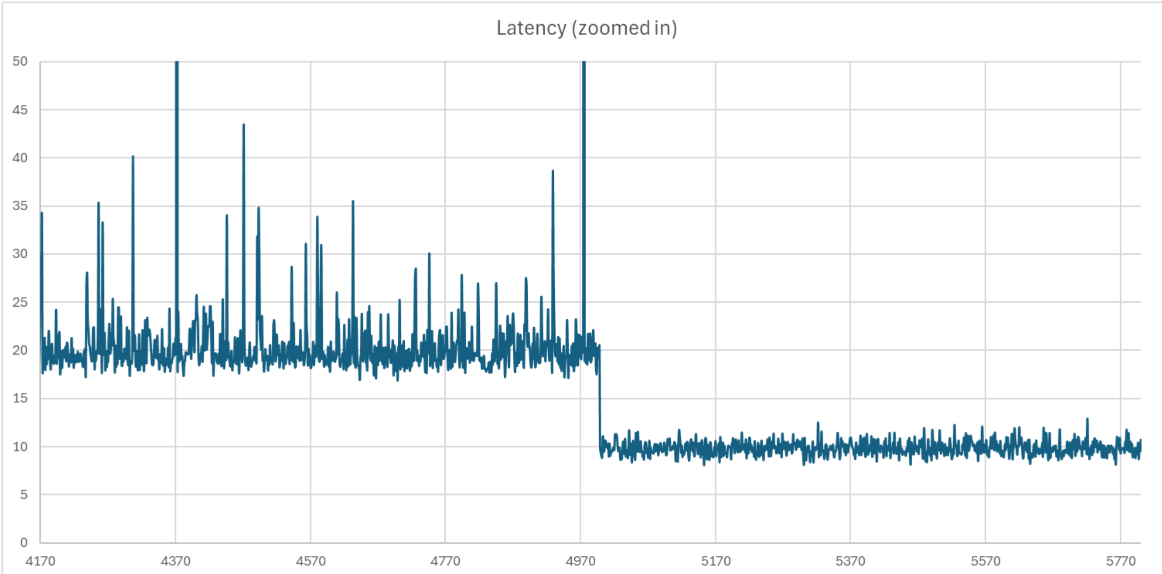

Figure 15– Zoomed in – 2 High Frequency Devices, latency results

Another interesting observation from the test accounts was that results from test accounts on mid-split nodes were more uniform and generally better behaved than those from sub-split nodes. This finding was likely not too surprising since a mid-split node was typically reworked as part of ongoing network capacity management. Compare Figure 16 to Figure 12, where Figure 16 contains results from mid-split nodes and Figure 12 contains results from sub-split nodes. Similarly, Figure 17 can be compared to Figure 13.

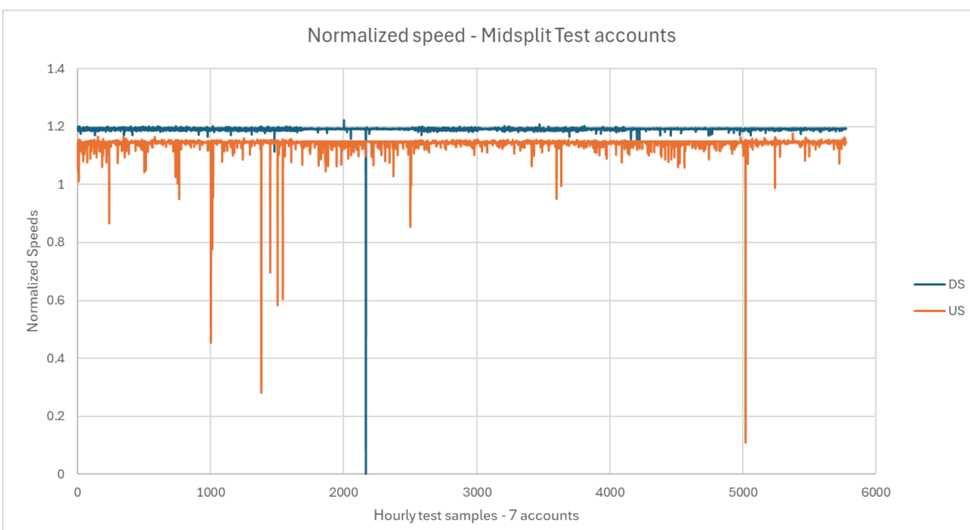

Figure 16– High Frequency Test Devices, Mid-split Nodes - speed

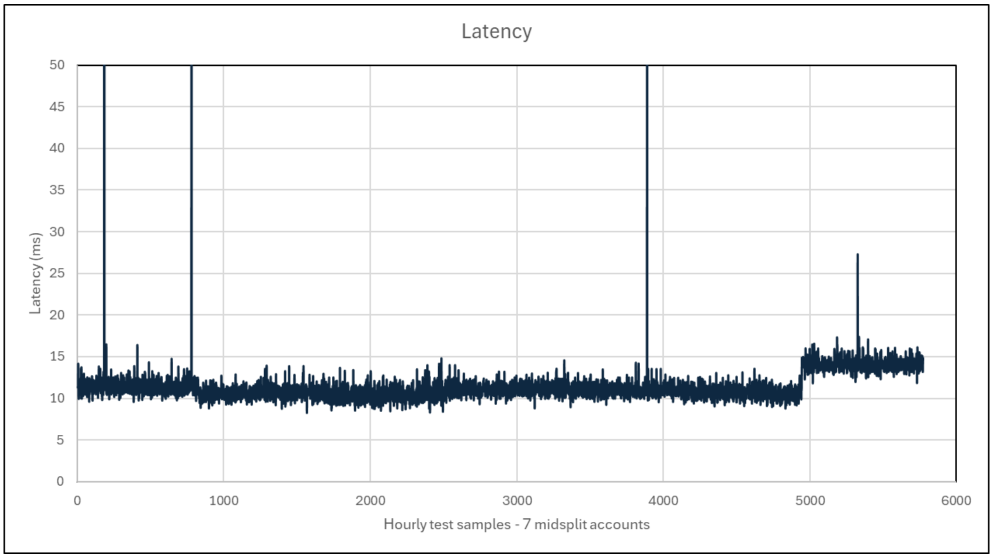

Figure 17– High Frequency Test Devices, Mid-split Nodes - latency

5. Observations and Takeaways or next steps

5.1. Correlations Between Gateway Speed Test Results and Network Performance Metrics

In theory, speed performance decreases as node utilization increases, especially beyond a certain threshold percentage. The gateway speed test results can provide valuable reference data for determining a node utilization threshold as a decision criterion for a node action plan.

Figure 18 presents the relationship between node utilization and download speed performance in terms of the ratio of actual to advertised speed, showing how node utilization and congestion impact speed performance. The Y-axis in the upper part of Figure 18 displays actual download speeds as a percentage of the advertised speed, while the X-axis presents node utilization divided into 10% bins across different service tiers.

As shown in the upper part of Figure 18, the median of actual to advertised download speed begins to degrade roughly starting at 60-70% node utilization, depending on the service tier speed. The 80 percent of speed performance, which corresponds to the bottom 20% of performance, experiences higher degradation as node utilization increases. This degradation is more significant in higher speed tiers. The bottom part of Figure 18 displays a histogram of the number of test results in each 10% node utilization decile. As shown in Figure 18, over 95% of nodes have a downstream bandwidth utilization of less than 40%. Therefore, the sample size of test results from highly utilized nodes is significantly smaller. Although we can observe a degradation pattern, more data is required to draw a definitive conclusion and accurately define an ideal threshold value for planning. Additionally, degradation may occur not only due to bandwidth utilization, but also due to factors such as signal to noise ratio or forward error detection ratio. We will continue to study this data and continue to analyze ongoing data in future work to develop meaningful use cases and thresholds.

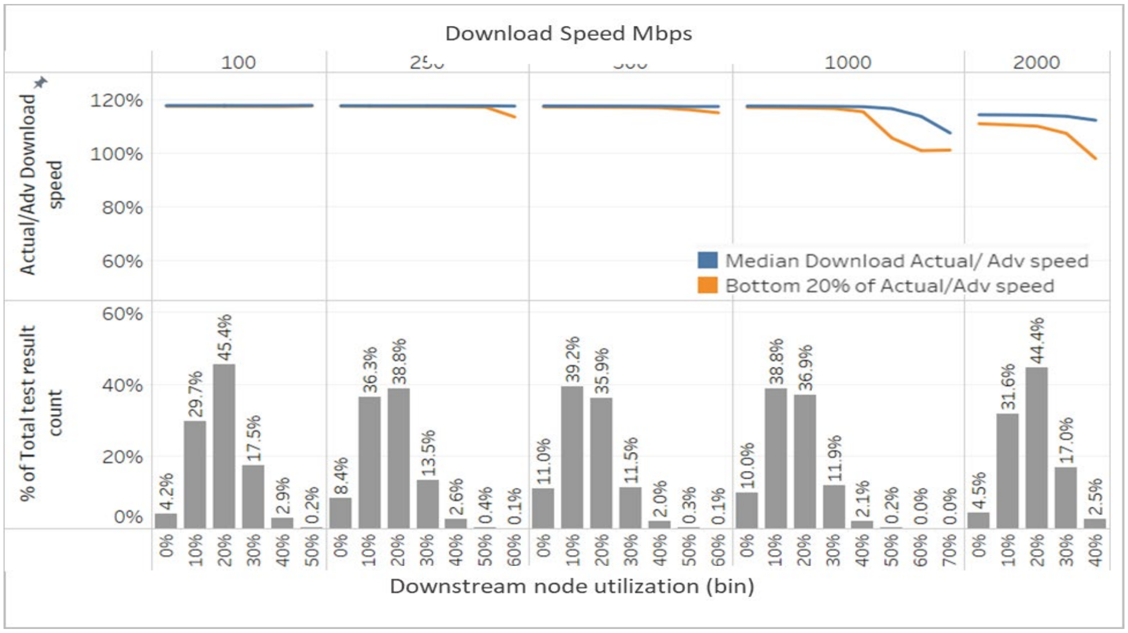

Figure 118 - Actual download/advertised download speed by node utilization

Upstream actual to advertised speed performance shows a trend similar to that of downstream speed. As shown in Figure 19, median speed performance remains consistently good, but the bottom 20 percent value shows degradation beyond a certain node utilization level. It happens mostly in the maximum tier speed, which includes 35Mbps upstream speed for sub-split nodes and 100Mbps upstream speed for mid-split nodes. The test results from highly utilized nodes were insufficient during the measurement period for statistical significance. We will be able to expand the data set as more data becomes available over time.

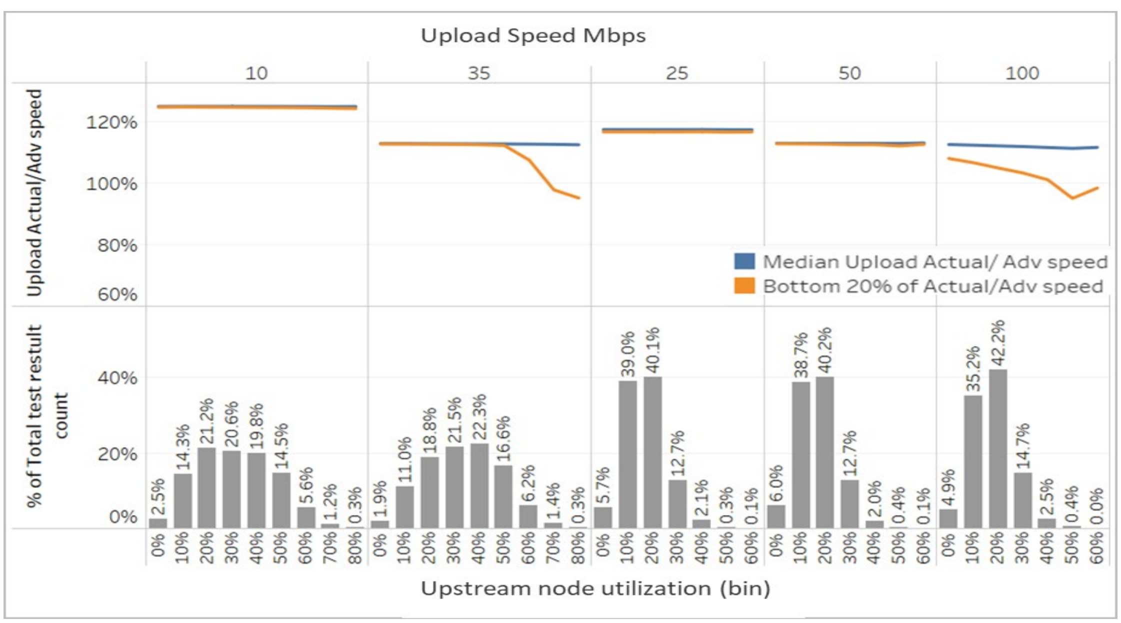

Figure 19 - Actual upload/advertised upload speed by node utilization

Figure 20 shows latency trend over node utilization. Latency increases as the offered speed decreases across different service tiers. Latency increases as node utilization increases.

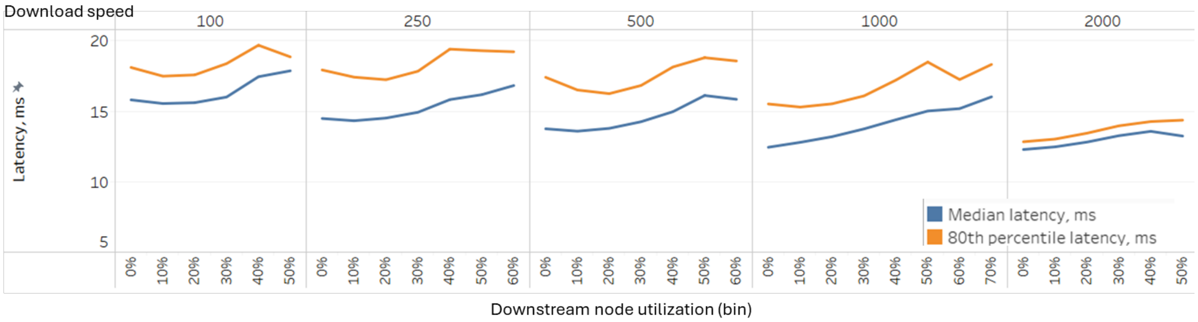

Figure 20- Latency by node utilization

5.2. Correlations Between GWST Results and Network Configuration

Another result of interest is that test results for Node+0 nodes were not noticeably different from other more traditional nodes. It is possible that the number of Node+0 nodes was too small to show differentiation from the much larger number of traditional nodes.

6. Topics for Future Investigation

The sheer volume of test results presents many opportunities for future studies to hopefully enable more proactive network capacity management and operational preventative maintenance.

As an example, consider two nodes with similar high-level statistics, both had a little over 100 test results in June, with a similar rate of questionable test results. However, upon close examination, those results were caused by different issues. One node's test results were dominated by a gateway that had an issue. The user or a customer service technician did a large number of tests with many results failing in the upstream. The other node had the same number of failing tests, but the failing test results were generated by many different gateways. In the second case, the node is more likely to have a systemic issue versus the problematic gateway on the first node.

In Figure 21, the test results of the two nodes are sorted by the date of the tests.

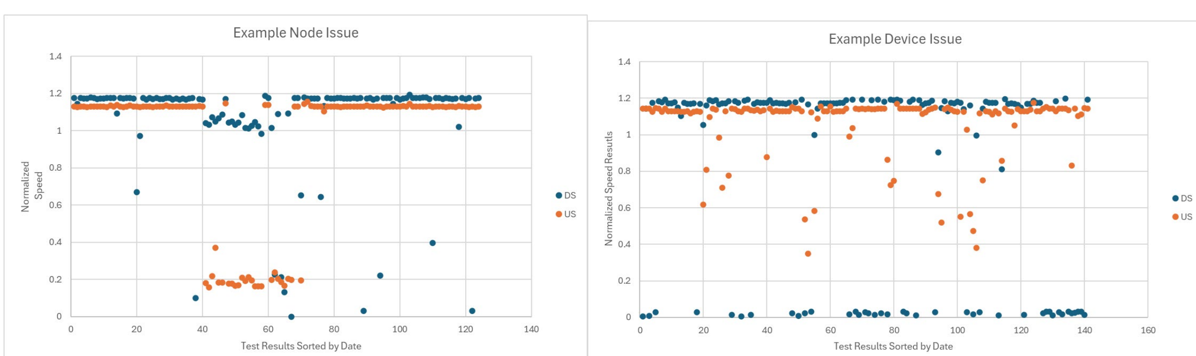

Figure 21– Comparison of Test Results Sorted by Date

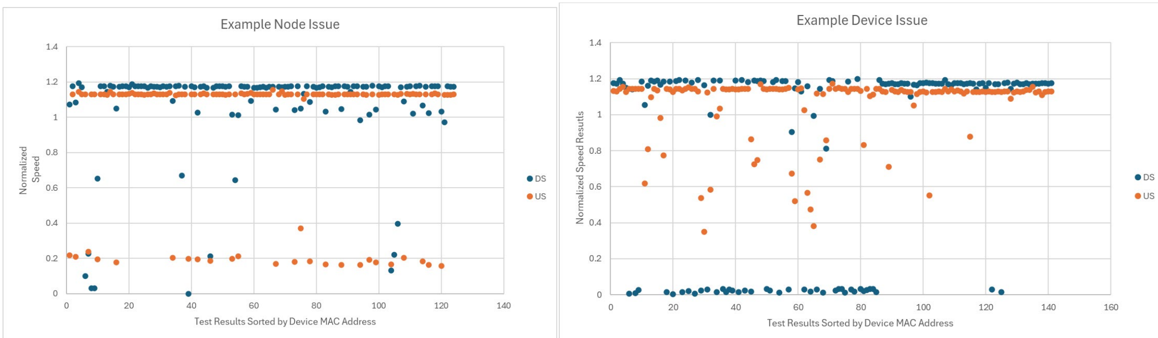

Figure 22– Comparison of Test Results Sorted by Device ID

When you consider the graphs of the node on the left, the time sorted results show an outage for a period of time that affected all devices more or less equally. The graphs of the node on the right show little time correlation to the poor results, but when sorted by device ID, the errors are clustered to a particular device, which happened to have run about half of all the tests run on that node across that month.

Opportunities are available to create automated algorithms to detect poor performing devices before the customer notices, or to enact proactive node maintenance activities before the subscribers on that node notice that their service has been impacted. But these topics will be further explored in a future paper.

7. Conclusion

In this paper, we have described a new tool in our toolbox for network capacity management and on-going network performance management: automated and user-initiated gateway performance testing. We have shared insights that we have gained from careful analysis of these on-going test results and pointed out areas where we anticipate additional studies will provide further insights.

Overall, we have found that utilizing actual test results has shown that our holistic approach to network capacity management is providing excellent service for our customers. We hope to use this continuing data source to further improve our network monitoring to include more proactive network and subscriber actions.

Abbreviations

CPE	Customer Premises Equipment
DS	Downstream
Gbps	Gigabits per second
GWST	Gateway Speed Test
HFC	Hybrid Fiber Coaxial
Mbps	Megabits per second
PON	Passive Optical Network
SCTE	Society of Cable Telecommunications Engineers
US	Upstream