Latency Measurement: What is latency and how do we measure it?

A Technical Paper prepared for SCTE•ISBE by

Karthik Sundaresan
Distinguished Technologist
CableLabs
858 Coal Creek Circle, Louisville, CO, 80303
3036613895
k.sundaresan@cablelabs.com

Greg White
Distinguished Technologist
CableLabs
858 Coal Creek Circle, Louisville, CO, 80303
3036613822
g.white@cablelabs.com

Steve Glennon
Distinguished Technologist
CableLabs
858 Coal Creek Circle, Louisville, CO, 80303
3036613834
s.glennon@cablelabs.com

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

List of Tables

Title **Page Number**

1. Introduction

Low latency is gaining importance in the internet experience. Low Latency is being approached as an end to end solution by operators. This includes Wi-Fi links in the home, DOCSIS links in the access network and core network segments. Providing lower end to end latency is a top priority for operators in the coming years. Measuring the latency in the network then becomes a vital requirement.

Operators (and 3rd party speed-test websites) have metrics on latency which they have reported and discussed with the community. Yet there is confusion surrounding the latency numbers and the ability to compare them between networks. The language and meaning of latency metrics (latency vs jitter, one-way vs round-trip, average vs 99th percentile), the latency measurement methods, what is being measured and when (peak vs off-peak periods), are varied. This paper provides clarity around these topics and discusses latency measurement architectures as well as best in class measurement tools to streamline latency measurement for the cable industry.

Operators want the ability to measure the difference in latency that is actually being delivered, before and after they deploy a new technology in their network, like DOCSIS 3.1 AQM, Low Latency DOCSIS, Low Latency WiFi etc. The latency portion of measurement reports (e.g. FCC's Measuring Broadband America initiative) are not optimal, and without a consistent measurement approach to latency, this could become a customer perception problem for the internet service providers. For new technologies that differentiate traffic, there are also questions around how latency for unmarked traffic vs marked traffic can be measured and reported. Operators will be asked to help troubleshoot latency issues and it will be important for them to identify latency within their networks vs. outside of their networks. This paper discusses the latency measurement frameworks which an MSO can integrate into their network deployment.

1.1. Quality of Experience

Latency is the time that it takes for a packet to make it across the network from a sender to a receiver and for the response to come back. Network latency is commonly measured as round-trip-time and is sometimes referred to as ‘ping time’. As applications turn ever more interactive, network latency plays an increasingly important role for their performance. Applications that are real-time perform the best when latency is low, and adding more bandwidth without addressing latency doesn’t improve the user experience. Packet forwarding latency can have a large impact on the user experience for a variety of network applications. The applications most commonly considered as latency-sensitive are real-time interactive applications such as voice over Internet protocol (VoIP), video conferencing such as Zoom, and networked online gaming. However, other applications are sensitive as well; for example, web browsing is surprisingly sensitive to latencies on the order of hundreds of milliseconds.

Test results in [ITU-T G.114], show that highly interactive tasks (e.g. speech, video conferencing and interactive data applications) can be affected by delays beyond 100 ms and users report significantly reduced mean opinion scores (MOS) when the voice delays are beyond the 150ms mark. The current [ITU-T G.114] recommends a maximum of a 150 ms one-way latency for VOIP applications.

Online games have some models [QoE and Latency] that indicate the impact that network parameters have on user experience. Some data exists to indicate that end-to-end round-trip latency should be kept below 25 ms or 50 ms in order to provide a good user experience, depending on the type of game (first person shooters, massively multiplayer online games, e-sports, etc.). When the operational response delay is less than 50 ms, the MOS scores tend to be high; when the operation response delay is around 100 ms, the MOS decreases but is acceptable for some kinds of games, and when the operation response delay is beyond 200 ms, the interaction quality for the gamer is very poor.

If we assume that gaming servers centrally located in North America are serving gamers all over the continent, the round-trip time (RTT) on the fiber backhaul links for gamers in the west coast will be around 40 ms (assuming 4000 fiber kilometers between say San Francisco and Chicago, and speed of light in fiber as 0.67x speed of light in vacuum). These RTTs will be even higher for gamers across different continents, if they don't have separate gaming servers. So, for the games which require very low latency and latency variance, the 25 ms - 50 ms end-to end target implies that the access network latencies need to be consistently in the order of 5 ms – 10 ms target to meet the requirements for online games.

Web browsing performance is traditionally tracked using page load time. Web content can be sourced from different servers and web browsers typically fetch resources from each server by opening up multiple TCP connections to the server. As there are multiple handshakes/interactions in each of the underlying protocols (DNS, TCP, TLS, HTTP) and all of those handshakes are impacted by the RTT, higher RTTs increase the page load time. See the paper [Belshe M] "More Bandwidth Doesn't Matter (much)*"* for experiments on how RTT affects page load time.

1.2. Latency in the Internet

There are a few main contributors to the latency of a packet as it traverses the network. The switching/ forwarding delay, propagation delay, serialization/encoding delay are some of the factors which affect packets as they go across various network devices and links, from the source to the destination. Queuing delay is usually the biggest contributor to latency, and is mainly caused by the current TCP protocol and its variants. This delay is encountered at the bottleneck links like the home Wi-Fi network or the access network. The majority of TCP implementations use loss-based congestion control, where TCP ramps up the number of bytes "in-flight" (i.e. its congestion window) until it sees packet loss, cuts its congestion window in half, and then starts ramping back up again until it sees the next packet loss. (When the buffers in the device transmit queues are full, a new arriving packet has to be discarded). This way TCP automatically adjusts its transmission rate to fully utilize the available capacity of the bottleneck link.

The result of this congestion window ramp-up and cut-in half mechanism is a saw-tooth behavior for the buffer going between partially full and totally full. In every home there are multiple users and applications that will use the same connection to connect to the internet. Applications other than TCP will suffer as the packets from those applications will arrive to nearly full buffer that may take tens or hundreds of milliseconds to drain. This can make web browsing perform poorly, and make VoIP, video chat, or online games unusable when other TCP based applications (e.g. streaming video) are in use.

1.3. Common techniques to reduce latency

Setting the buffer sizes appropriately in each of the network devices is a first step to reduce the latency in the network. Active queue management (AQM) is the next step in mitigating queueing delay, where the basic idea is to detect the increasing queue created by TCP and, then drop a packet which will let TCP know to back off on its sending rate, much ahead of the time it takes to drop a packet when the buffer is completely full. There are variety of algorithms, such as random early detection (RED), Proportional Integral Controller Enhanced (PIE) etc., which an AQM system can implement.

The next stage in the evolution of latency reducing solutions is the dual-queue approach where the concept is to separate the traffic for queue-building applications from those applications/traffic flows which are non-queue building. See the paper [Greg W, SCTE 2019] Low Latency DOCSIS Overview and Performance Characteristics, for detail on these types of traffic flows and the dual queue approaches. Low

Latency DOCSIS and L4S technologies tackle the queueing delay by allowing non-queue-building applications to avoid waiting behind the full buffers caused by the current TCP or its variants.

2. View of latency measurements

Internet latency is crucial in providing reliable and efficient broadband services to customers who are connecting to servers across the country and the globe. The trend of real time gaming and other real time applications only accelerates the importance of accurately understanding the latency characteristics of the network. This bubbles up the task of latency measurement towards the top of an operator's priority list. Being able to accurately diagnose latency issues seen by residential or business customers is becoming more important. In order to support server selection in distributed /virtual computing environments measuring accurate latency becomes extremely important. Knowing the latency characteristics well allows an operator to make better decisions on which latency reducing technologies to deploy and where.

Accurately measuring network latency, however, is not an easy task due to lack of testing end points, lack of clock synchronization when needed, the sheer volume of collected data points, and aggregating and analyzing the data meaningfully. In addition, the time that latency is measured affects measurement results significantly due to network dynamics, volatile traffic conditions, and network failures.

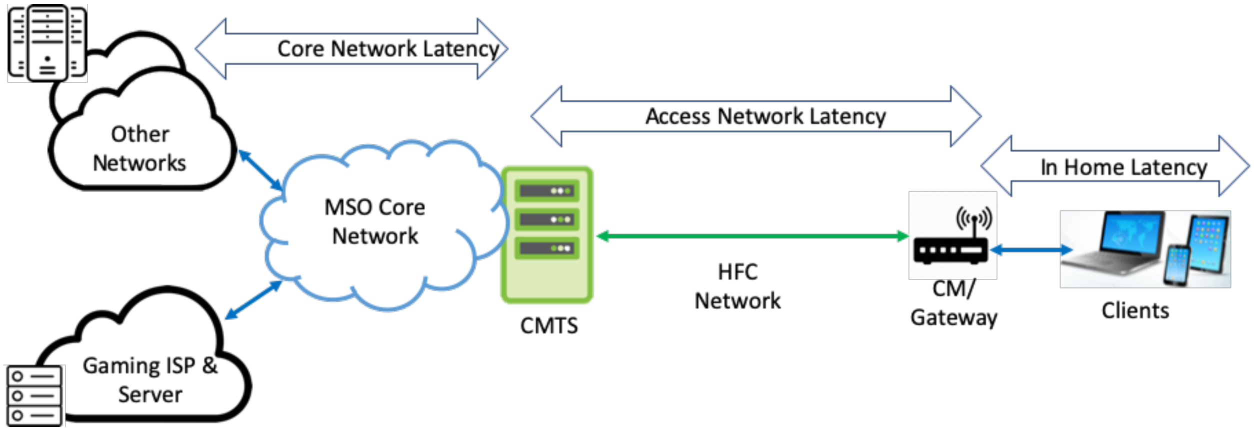


Figure 1 – MSO view of Latency Measurements

2.1. MSO Goals for Latency Measurement

Operators want to leverage existing available tools and standardized architectures to quickly set up a measurement infrastructure. Some of the common operator use cases and considerations are as follows.

- Identify Latency in 3rd party networks vs. MSO core network vs. Home network. In the Core network, there is a need to develop processes to identify routing issues, especially in the path to the egress point in the network which connects to a specific application server. For the Access & Home networks, it is extremely useful for an operator to be able to delineate latency from within the customer home (e.g. due to Wi-Fi) vs the access network latency vs the aggregation/core network.
- Operation Diagnostics Support: Operators would like to develop diagnostic tools, so that they can give meaningful information to their operations team. The use of latency measurements in NOC and field tester tools for live problem diagnosis is common at IP and Ethernet layers

- Operators would need to measure a variety of access and core architectures (R-PHY, FMA, Integrated) and need the measurement methods work across these range of deployments
- Network Architecture Analysis: The loss and delay performance metrics impact the scalability of the network and also on its behavior under load. For network architects, understanding both end-to-end network latency plus the contribution of the various links and nodes (network devices) that the network is comprised of is very useful.
- Understanding how to optimize the network deployments: e.g. with Distributed CCAP architectures an operator has to decide on a particular architecture, or where to place the physical or virtual components and decide on the location of certain functionality (e.g. MAC scheduler).
- There are many benchmarking purposes the latency measurement data can be used for e.g. different equipment (switches, routers) introduce different degrees of delay when processing packets. When moving from physical network elements to virtualized network elements, an operator needs to be able to quantify the latency difference.
- Lab latency measurements can compare the impact of introducing a new network element or configuration (e.g. a new technology like Low Latency DOCSIS) and verify the end user experience prior to deployment.
- Optimizing Network configuration: Appropriate latency measurement techniques can help diagnose intermittent issues (e.g. buffer overflows) and help fix them.
- Now with a goal of identifying per hop latency, operators need to identify the appropriate locations for the measurement end-points: end-device, gateway/CM, CMTS, router, interconnection point, etc.
- Any measurement architecture needs to support frictionless deployment of latency measurement infrastructure. This is dependent on how the specific measurement infrastructure is implemented and deployed (e.g. is it using hardware probes vs virtual probes). Scalability of the measurement platform across an entire operator becomes an important consideration.

2.2. Current National Latency Reports

Broadband infrastructure is gaining the attention of various national communications regulators, as countries focus on enabling their people with high speed internet connectivity. As a part of this many of these regulators measure the broadband deployments and report on various metrics such as houses covered, speed tiers available etc. and also conduct network measurements on actual upload and download speeds. Latency measurements are now also becoming an integral part of these reports.

2.2.1. Measuring Broadband America

In the United states, the Federal Communication Commission (FCC) runs the Measuring Broadband America (MBA) program. The MBA program is a nationwide study of consumer broadband performance and it collects network performance data from a representative sample of customers from each of the fixed Internet Service Providers (ISPs). See the paper [MBA FCC] 'Ninth MBA Fixed broadband report', for the latest speed and latency data reported. The MBA tests conducted are automated, direct measurements of the customers service during a single month and is done in collaboration with the measurement company SamKnows. Each volunteer customer connects a 'Whitebox' client device to their home network which performs the tests after finding the nearest test servers.

The MBA program measures latency by measuring the average round-trip time from the consumer's home to the two closest measurement servers, one server chosen from each of two "pools" of servers. The report shows the median latency for each participating ISP and includes aggregated information for each ISP and type of access network. It reports the measured latencies for various DSL, cable and fiber based ISPs on an individual basis as well as aggregated. The MBA program has a limited number of test server locations in each pool. Only six cities host test servers in both pools (an additional four cities host a server in only one pool). This means that client devices that are geographically distant from these six cities will report latency numbers that are more likely to be correlated to geography than to network capability. Difference in geographical distance to the server and also the distance in the of the number of hops internal and external to the ISP network, can make a difference in the number of network links the test packets have to travel across and ultimately the latency measured.

The MBA program latency and packet loss tests measure the round-trip times for approximately 2,000 packets per hour sent at randomly distributed intervals. Per the [MBA FCC] report, the latency and packet loss test records the number of packets sent each hour, the average round trip time and the total number of packets lost (a packet is considered lost if the packet's round-trip latency exceeds 3 seconds). The test computes the summarized minimum, maximum, standard deviation and mean from the lowest 99 percent of results. MBA determines the mean value over all the measurements for each individual's Whitebox and then computes a median value from the set of mean values for all the Whiteboxes.

2.2.2. Measuring Broadband Canada Project

The Canadian Radio-television and Telecommunications Commission (CRTC) has commissioned a study of the performance of broadband services sold to Canadian consumers. This project measures broadband Internet performance, including actual connection speeds, in Canadian homes. The CRTC collaborated with a number of Canadian Internet service providers (ISPs) and SamKnows, and produced a Measuring Broadband Canada Report, June 2020. See the paper at [MBC CRTC]. The report describes that, unlike in the US MBA program, the latency data was focused on Whiteboxes located within a 150km radius of the test server locations in order to minimize the effect of distance on measurements. See the paper [MBC CRTC] to understand the details on the average latency during peak hours for different Canadian service providers and access networks (Cable, DSL, Fiber) . Like the MBA report, the MBC report [MBC CRTC] also measures packet loss and average webpage loading times from a selection of websites.

2.2.3. EU Broadband Report

The European Commission has a vision around broadband connectivity and takes policy actions to turn Europe into a 'Gigabit Society' by 2025. In support of that the European Commission has commissioned a study to obtain reliable and accurate statistics of broadband performance across the different EU Member States and other countries.

2.2.4. Speedtest (Ookla)

Speedtest(Ookla) today publishes [SpeedTest] Market Reports as a guide to the state of fixed broadband and mobile networks around the world. Each report includes mainly speed (downstream and upstream) data and insights about country trends. Speed test data is based on the results of millions of tests run by Speedtest users. An individual user initiated Speedtest uses 'ping' to report the latency to the nearest Speedtest server. Speedtest is very relevant in the latency measurement landscape as that is how the majority of consumers understand what their service speed are and what latencies their connection achieves. Of course, consumers also tend to run Speedtests when they see an issue with the service or when they upgrade or get a new service, so this may also not be a representative sample across the consumers.

3. Latency Metrics

Each operator needs to track different metrics or KPIs when it comes to network latency. The network latency metrics important to operations teams will be different than what metrics are important to product or regulatory teams. Metrics can also be dependent on where the network is in the product life cycle. There are a variety of latency metrics to choose from and this section describes how to look at and understand latency.

As a packet travels across a network, the packet experiences different types of delays at intermediate hosts, routers, and network links. A host or a router needs time (processing/ switching/ forwarding delay) to process an incoming packet to determine its next hop. The packet also often waits in the transmit queue behind other packets (queuing delay). Transmission delay (serialization/encoding delay) is the time for a node to move out all the bits of the packet onto the link. Finally, it takes time for the packet to propagate over the link from one node to another. End-to-end latency is the sum of such delays at every step of the way.

3.1. One-way Latency (or Packet Delay)

One-way latency is the total time it takes for a packet of data to travel from the sender to the receiver, across one or multiple hops. The one-way delay will be dependent on congestion of the network at the time the packet was sent. It will also depend on the topology of the network and the distance and routing decisions between the two end points. Measuring one-way latency also implies that the sender and receiver have synchronized clocks, which sometimes is a challenge to set up and maintain when the end points are across multiple network domains.

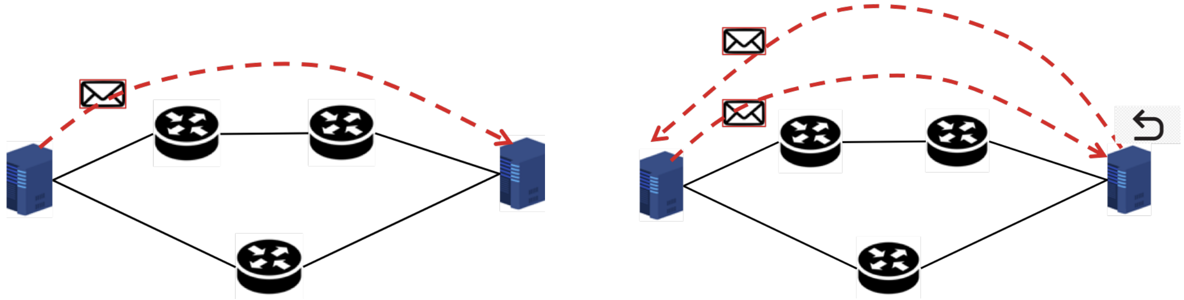


Figure 2 – One way Latency vs. Round trip Latency

3.2. Round Trip Latency

Round trip time (RTT) or round-trip latency, is the time taken for a packet of information to travel from the sender to the receiver and back again. RTT is the total time it takes for a packet of data to travel from the sender to the receiver, across one or multiple hops, plus the total length of time it takes for receiver to send a packet back to the sender, through one or multiple hops.

The Round-trip latency is more often quoted, as it can be measured from a single point. It requires a process running on the other end to mirror the packet back. The RTT can vary if the return path is different from the forward path. The most common example for round-trip measurements is the ICMP Echo Request/Reply, used by the ping tool.

3.3. Singleton Measurements vs Sets of Measurements

A singleton measurement test can send one packet and calculate the one way or round-trip latency of that packet. That is not the most interesting as that is just one sample on the network which is carrying millions of packets. Latency varies with different factors such as the location of the two measurement end-points, and with time (due to changes in route selection or due to congestion). So, most latency measurement tests use multiple packets in a test for one way or round-trip measurements. This gives an operator a sample distribution of latency measurements and paints a better picture of the latency behavior. A test would measure the latency of each of the test packets, and then an operator could understand what the behavior of the network latency is across that set of packets. Having more data samples allows the operator to observe the variations and better understand the network latency in a way that correlates to what the customer will perceive and experience.

Operators typically run each of these tests multiple times a day to get a feel for the network latency variation over time. These sets of measurements could be performed over time for one user, these could be tests done across multiple users or it could be both: tests done over time and for multiple users.

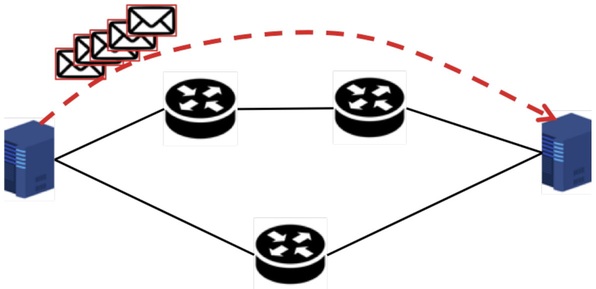


Figure 3 – Sets of Latency Measurements

3.4. Jitter or Delay Variation

The term 'jitter' is a commonly used term to refer to variation in the latency of arriving packets over time. Though prevalent in the networking parlance, the term is considered deprecated by technical bodies like the IETF. The IETF [IETF RFC 5481] now uses the term "delay variation" for metrics that quantify a path's ability to transfer packets with consistent delay. Note that jitter can also be used to convey undesired variation in signals in contexts beyond IP packet transfer. (e.g. frequency or phase variations in electronic circuits in reference to a clock, or sampling jitter in analog-to-digital conversion of signals etc.)

The term jitter can be defined within a specific context in order to provide a meaningful metric for a specific application, or it is sometimes defined simply in a manner that is convenient to calculate. Most real-time voice and video applications employ a (de-jitter) buffer to smooth out delay variation encountered on the path. Many of the commonly used jitter definitions are aimed at helping designers of such systems choose the size of the de-jitter buffer.

[IETF RFC 5481] notes that various standards for delay variation have allowed flexibility to formulate the metric and so the specific formulations of delay variation must be well understood. All definitions of delay variation are derived from the one way or round-trip delay metrics. The networking industry has predominantly implemented two specific formulations of delay variation: Inter-Packet Delay Variation and Packet Delay Variation.

3.4.1. Inter-Packet Delay Variation, IPDV

A latency test or application will send a sequence of packets to measure one way or round-trip latencies. Inter packet delay variation (IPDV) is derived from such a sequence of latency measurements. It is simply the difference in latency of each packet as compared to the previous packet.

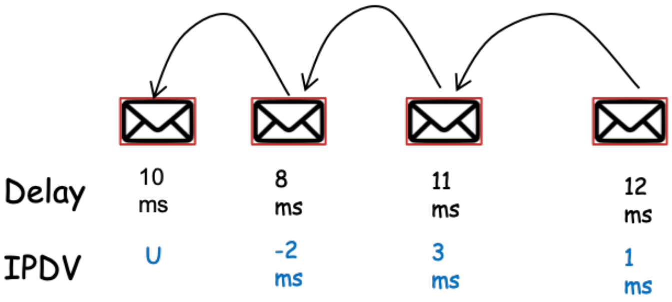


Figure 4 – IPDV Calculation

3.4.2. Packet delay variation, PDV

Packet Delay Variation, PDV, is also derived from a sequence of latency measurements where a single reference latency is chosen from the stream based on specific criteria. The most common criterion for the reference is the packet with the minimum delay in the sample. Other references such as average latency can be chosen as well. PDV is simply the difference in latency of each packet as compared to the one reference packet. In [ITU-T Y.1540] the ITU also chooses this definition of packet delay variation.

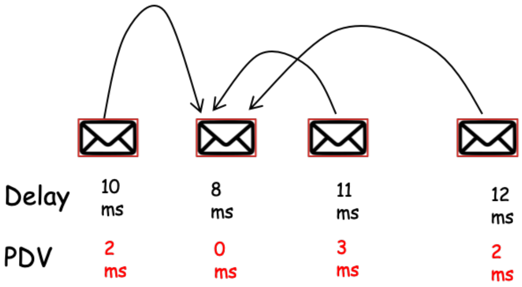


Figure 5 – PDV Calculation

3.4.3. Jitter Metrics in Use in industry

The formulations described in the previous sections result in a per-packet metric, which (across a set of packets) can then be summarized using descriptive statistics (e.g. mean, median, standard deviation, median absolute deviation, P99, P99.9, etc.) in order to come up with a summary of the delay variation across the set of packets.

There are different ways in which jitter definitions are used in different applications in the industry. [SamKnows], [Haste], [Excentis ByteBlower], [M-lab NDT] and [Network Next] use statistics derived from the PDV definition , while [WTFast], [3rdEchelon], [IETF RFC 3550] and [PingPlotter Pro] use statistics derived from the IPDV definition. The below table describes the way definition each of these entities use and how they aggregate it. As can be seen, there are significant differences in the meaning of the term from one entity to another.

Table 1 – Jitter definitions in the Industry

Entity	Parameter	Definition
	PDV based	
SamKnows *(Network performance measurement platform)*	Jitter	P99 PDV (PDV referenced against min latency)
Excentis *(Byteblower traffic generator)*	Jitter	Standard deviation of PDV
Haste *(Optimized routing for game traffic)*	Jitter	Standard deviation of PDV
Network Next *(Optimized routing for online games)*	Jitter	jitter = 3* RMS(PDV) (PDV referenced against min latency)
MLab NDT *(Network test)*	Jitter (round trip time variation)	max(PDV) (PDV referenced against min latency)
	IPDV based	
RTP protocol *(RFC3550)*	Interarrival jitter	Exponentially-weighted moving average of the absolute value of IPDV
PingPlotter Pro *(Ping statistics tool)*	Jitter	Average of the absolute value of IPDV
3rdEchelon *(Internet Services company)*	Jitter	Average of the absolute value of IPDV
WTFast *(Gaming VPN solutions)*	Jitter	Average of the absolute value of IPDV

3.5. Descriptive statistics

Once we have a set of measurements (each of which is an individual latency measurement), a network operator wants to easily aggregate and make sense of those sets of measurements across the whole network and over time. The question is how best an operator can analyze the data to guarantee that the latency meets service requirements.

3.5.1. Basic statistics

Many operators start with basic statistics like mean, median or min-max. Each of these numbers have their place, but for large populations of data they often hide the actual network behavior. Mean and median tend to hide outliers, especially the high latency events which may happen only during specific times. In contrast, the maximum is overly conservative and is easily distorted by a single outlier event.

- Average: The arithmetic mean or average, is the simply the sum of the set of the latency measurements divided by the number of measurements. The set of results of each experiment or an observational study can yield its own average number. For latency measurements, though the average maybe a starting point, it hides a lot of the variation in latency. Some of the much higher excursions are diluted by the mean, and thus averages hide high latency events which would ultimately impact the customer experience. Outliers also skew averages, so the average doesn't represent typical behavior either.
- Min/ Max: The maximum and minimum of a set of measurements are the largest and smallest value in the set of measurements. These are useful to understand the limits of the network. In the context of latency measurements one can separate lost packets as a separate measure, instead of considering it as infinite latency.

- Standard Deviation: The Standard Deviation is a measure of how spread out the latency measurement numbers are. The standard deviation is calculated as the square root of the variance (average of the squared differences from the Mean, for each sample). This gives a measure of the amount of variation or dispersion of a set of latency values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range.

3.5.2. Percentile Numbers

Many descriptive statistics like mean, standard deviation, and skew are most meaningful when the underlying data follows a roughly normal (Gaussian) distribution. In contrast, even simple latency distributions are often heavily skewed with a set of values around a certain range, and with many fluctuations and outliers. As a result, these traditional statistics offer very little value in capturing or describing latency, but percentiles can generally be much more effective.

Percentiles allow a better understanding of the latency distributions than averages. A percentile is a value below which are a certain percentage of observations. Percentiles show the point at which a certain percentage of observed values occur. For example, the 95th percentile is the value which is just greater than 95% of the observed values, i.e. 95 percent of packets got a lower latency than the P95 value. For example, to obtain the 99th percentile of a collection of latency measurements from a network, an operator can sort them and discard the highest 1% of values. The largest remaining value is the 99th percentile. This value is the largest latency that will be seen for 99% of the measurements. An operator can choose a measure like the 90th, 95th, 99.9th (or even more nines) percentiles, which are typically denoted as P90, P95, P99 etc.

Network latencies between machines can be low when the network path is idle, but when there is significant network activity packets can take anywhere from a few milliseconds to hundreds of milliseconds, or even seconds. Since many network segments (particularly broadband links) are idle, or nearly so, for a significant portion of the day, the median latency and the minimum latency are often pretty close to one another. Long tail latencies occur when the higher percentiles begin to have values that are many times greater than the median. In a long tail latency distribution, the 99th percentile can be fifty times greater than the median value, much beyond normal distributions.

Percentiles are often used to find outliers. When a range of percentiles are computed one can estimate the data distribution more accurately. Another use for latency percentiles is to implement a threshold beyond which issues are flagged to the operator. An operator could also track a combination of a few different percentiles, such as P50, P75, P95, P99 and flag issues when any of them change significantly with respect to previous measurements or thresholds.

Now the question is which latency response time metric is more representative of the user experience. Is it the 95th percentile or the 99.9th percentile? The below table tries to show how to think about the impact to an application like gaming. Gaming traffic flows are typically 60 packets per second at a rate of 100kbps-200kbps in the upstream direction and 60 packets per second at a rate of 500kbps-1Mbps on the downstream. Gaming clients or servers expect packets to arrive at that consistent rate of 60 times per second and any packets which arrive with a much higher latency cannot be used and are essentially thrown away. As an example, 99% of the gaming packets have a latency of 40ms or less, while 1% of packets are delayed for anywhere from 100ms to 500ms. For or a real-time game this 1% ‘latency event’ happens (on average) once every 1.6 seconds and such network behavior is unwelcome in gaming environment and may be a showstopper in other applications. Based on this view, perhaps the P99.9 value would be a good starting point to represent user experience for online gaming.

Table 2 – Understanding Latency Percentiles

Notation	Percentile Latency	Meaning	Implication	Impact for a gaming application
P50	50th percentile - median latency	50% of packets got this latency or better	50 of 100 of packets got worse than this latency	Every other packet!
P90	90th percentile	90% of packets got this latency or better	10 of 100 packets got worse than this latency	6 packets a second
P95	95th percentile	95% of packets got this latency or better	5 out of 100 packets got worse than this latency	3 packets a second
P99	99th percentile	99% of packets got this latency or better	1 of 100 packets got worse than this latency	1 packet every 1.6 seconds
P99.9	99.9th percentile	99.9% of packets got this latency or better	1 of 1000 packets got worse than this latency	1 packet every 16.6 seconds
P99.99	99.99th percentile	99.99% of packets got this latency or better	1 of 10,000 packets got worse than this latency	1 packet every 2 mins 46 seconds

3.6. Histograms

A histogram is a graphical method for displaying the shape of a distribution and is particularly useful when there are a large number of observations. To construct a histogram, the range of values observed in the measurement is divided into a series of intervals or bins. The measurements are then classified into bins counting how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins/intervals are contiguous and are often of equal size.

The goal is to collect enough data points for good latency characterization. This means an operator needs to collect data to obtain acceptable precision for different percentile levels. A simple process in latency measurement is to record all the latency data over multiple sets of tests and then later analyze and sort the data into traditional histograms to get the required percentile data. Some alternatives to the traditional histograms with linear bins are logarithmic bins, or arbitrary bins. Linear bins in histograms require lots of storage to cover the range with good resolution, while logarithmic covers wide range of values but does not have good precision. Arbitrary bins work only when the operator already has a good feel for the interesting parts of the data range.

3.7. Visualization of Latency

Data visualization can reveal patterns and trends in the data, allows quick absorption of large amounts of data by network operators, and ultimately lets the operator understand the information and make decisions. This section describes some of the ways an operator can visualize latency data.

3.7.1. Time series

A time series is a set of observations (x_t) ordered in time, observed at a discrete set of (approximately) evenly spaced time intervals: at times t =1,2,..., N , where N is the length of the time series. The figure below, created using [PingPlotter Pro], shows a time series of ping data, once every second for 10 minutes. While the average ping time is ~12ms, one can quickly see that it is not the normal case and there are many latencies of 15-20ms and occasionally even up to 25 to 30ms.

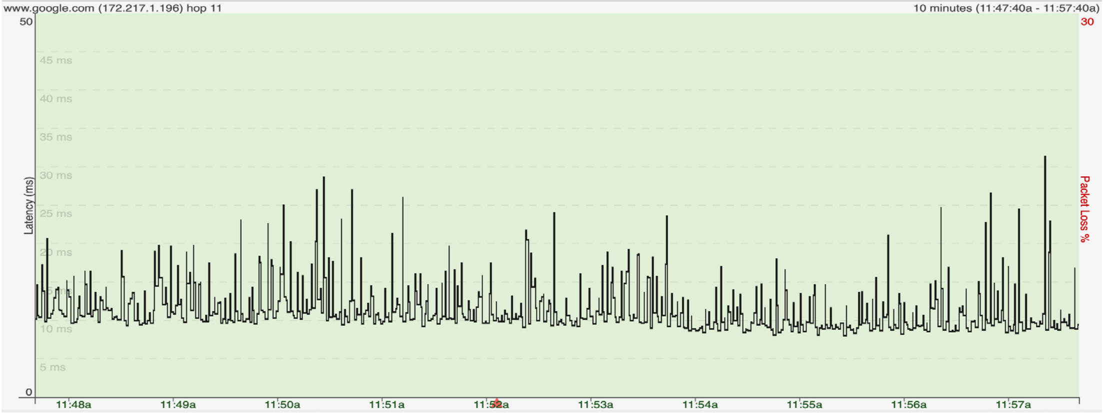


Figure 6 – Example time series of Latency Measurement

3.7.2. Probability density function (PDF)

A Probability Density Function (PDF) is a statistical expression used in probability theory as a way of representing the range of possible values of a continuous random variable. For a continuous function, the probability density function (pdf) is the probability that the variable has the value x. The area under the curve represents the probability that variable will fall within an interval; and is expressed in terms of an integral between two points. $\Pr[a \leq X \leq b] = \int_a^b f_X(x)dx$

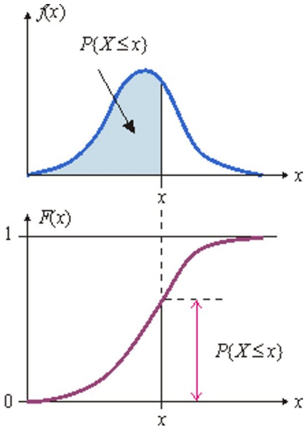


Figure 7 - PDF-CDF relationship

3.7.3. Cumulative Distribution Function, CDF

The cumulative distribution function (CDF) of a random variable is another method to describe the distribution of random variables. A cumulative distribution function describes probability that a random variable takes on a value less than or equal to x. That is $\Pr[X \leq x] = F_X(x)$

3.7.4. Complementary cumulative distribution function, CCDF

A complementary cumulative distribution function, answers the opposite question, i.e. how often is the random variable above a particular level x. To obtain the Cumulative Distribution Function (CDF), the integral of the PDF is computed. Then inverting the CDF results in the CCDF. (CCDF is the complement of the CDF or CCDF = 1 – CDF.) One can also plot the CCDF in a logarithmic scale so that the more interesting percentile values are easily discernible.

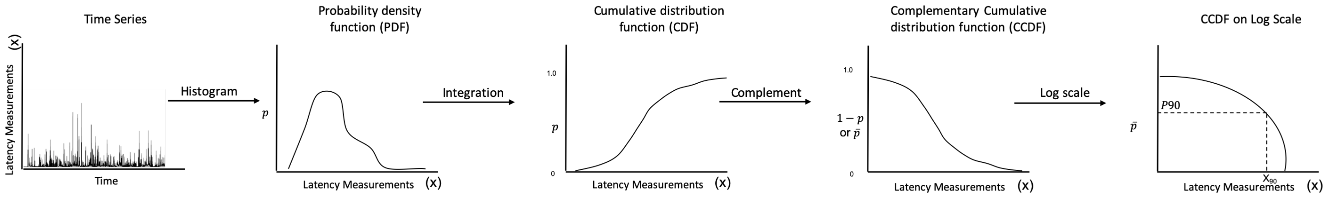
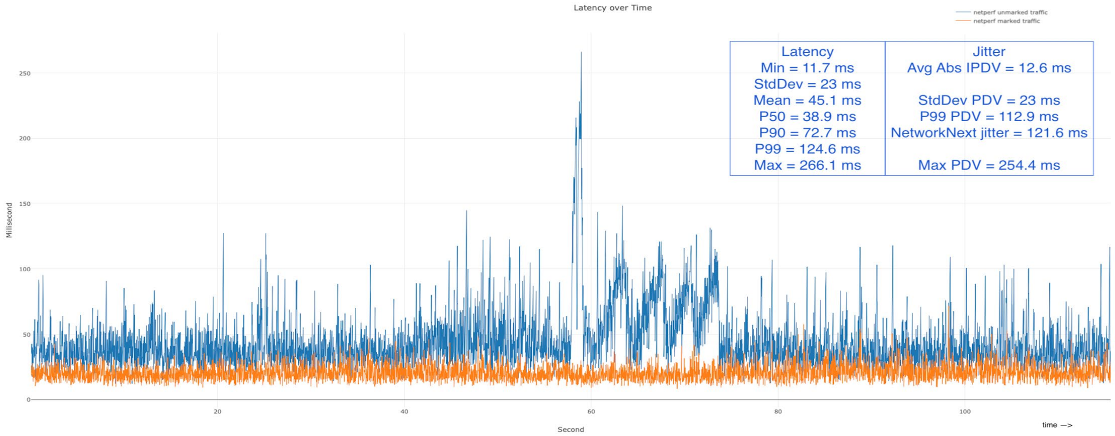


Figure 8 – Conversion from Time Series to PDF to CDF to CCDF to Logarithmic-CCDF

3.7.5. Example PDF/CDF/CCDF

Below is an example of some latency measurements performed in the lab, of round-trip times from a client to a server, which are separated by a Wi-Fi link and a DOCSIS link (CM and CMTS) with a pseudo Low Latency DOCSIS configuration. The time series figure below shows the latency measurements of unmarked traffic (in blue), while the latency of DSCP marked traffic is shown in orange. It also shows the various latency and jitter metrics of the unmarked traffic and how varied the numbers can be.



Latency	Jitter
Min = 11.7 ms	Avg Abs IPDV = 12.6 ms
StdDev = 23 ms	
Mean = 45.1 ms	StdDev PDV = 23 ms
P50 = 38.9 ms	P99 PDV = 112.9 ms
P90 = 72.7 ms	NetworkNext jitter = 121.6 ms
P99 = 124.6 ms	
Max = 266.1 ms	Max PDV = 254.4 ms

Figure 9 –Time series latency data of Marked vs unmarked traffic

The PDF figure below shows (using a histogram of 1ms bins) how different the two sets of latency measurements are, with the marked traffic flow(orange) having a lower and tighter latency numbers, while the unmarked traffic(blue) has latencies extending all the way from 100ms to 260ms.

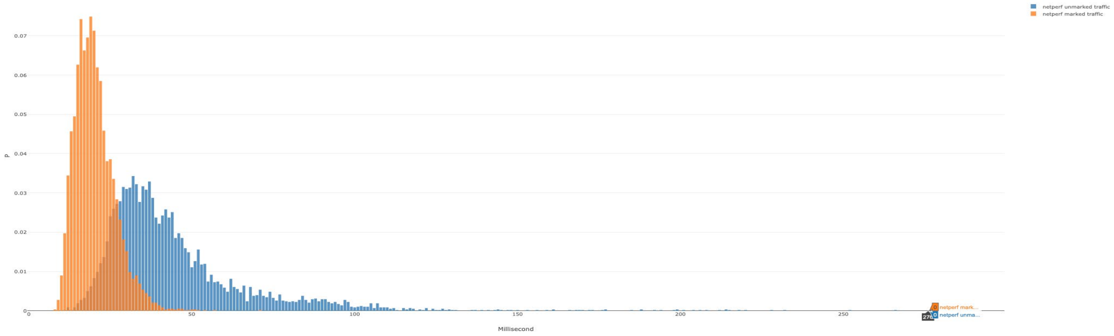


Figure 10 – Probability Distribution Function (PDF)

The Cumulative Distribution function (CDF) figure below for the same data set, show the marked traffic flow(orange) having a lower P99 (~38ms), while the unmarked traffic has a higher P99 (~125 ms).

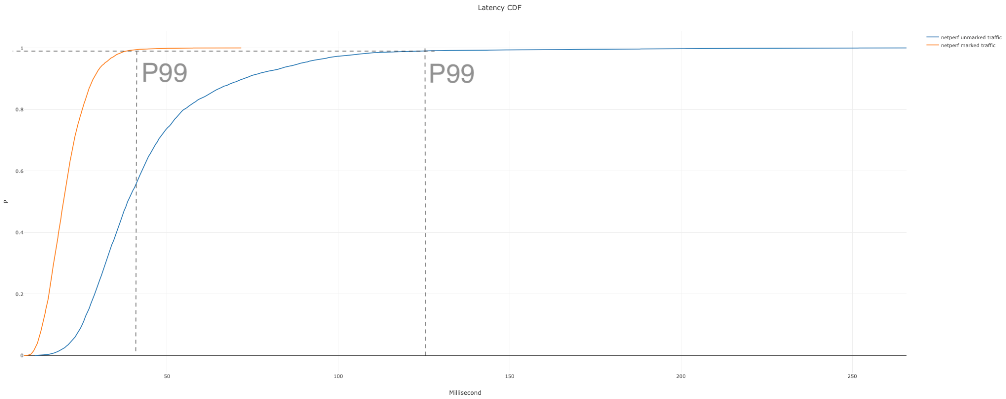


Figure 11 – Cumulative Distribution Function (CDF)

The CCDF figure below for the same data set, is essentially the same graph but inverted, with P99 readings closer to the bottom of the graph compared to the top.

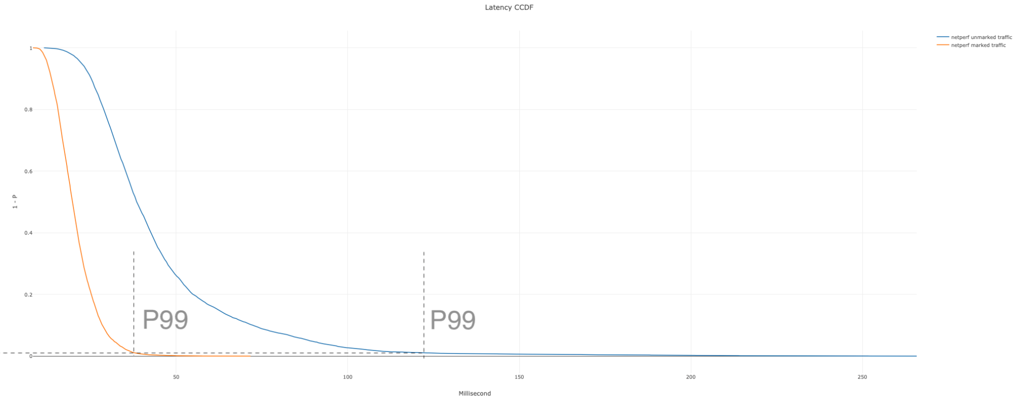


Figure 12 – Complementary Cumulative Distribution Function (CCDF)

The logarithmic CCDF figure, is the same CCDF graph but drawn on a logarithmic scale for both axes. Here we can compare the P90, P99 or P99.9 and see the differences in the percentiles we are interested in clearly at this scale.

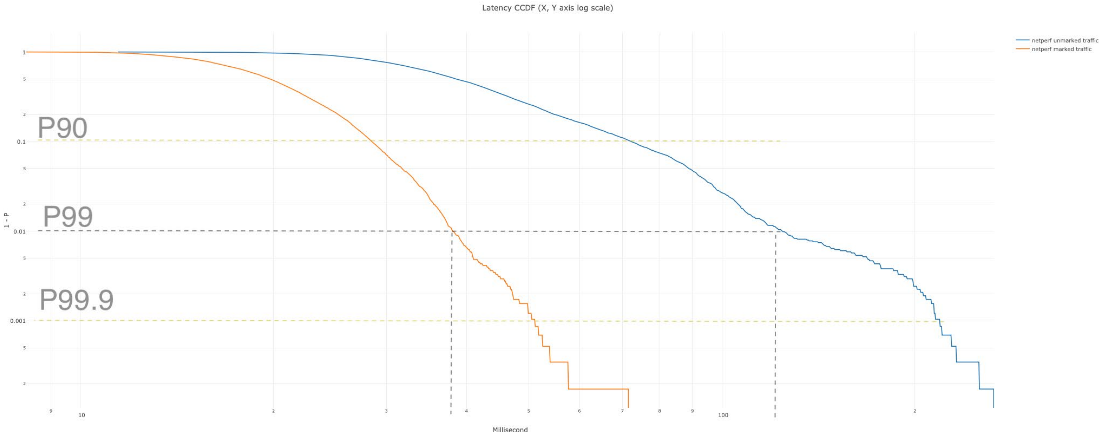


Figure 13 – CCDF on a Logarithmic scale

4. Latency Measurement architectures

4.1. Types of measurement

4.1.1. Active measurements

Active measurements are conducted by generating traffic between two end points for the sole purpose of measuring the latency. For example, with ICMP ping a sender sends an ICMP packet(s) to the receiver, who replies back; the sender calculates the time between sending and receiving the packet(s). The measurement is considered to be active, as the reason the traffic is created and sent is to measure the latency between the end points.

Active monitoring involves injecting test traffic into the network, typically with the same forwarding criteria as the user traffic being monitored, and then measuring its performance. These tests can either be one-way (from site 'A' to site 'D' or round trip (from site 'A' to site 'D' and back to site 'A'), depending on what the operator wants to measure. Since the test traffic mimics the user traffic, active testing gives a packet by packet view of the end-to-end performance of a network with regards to such things as latency, delay variation, or packet loss.

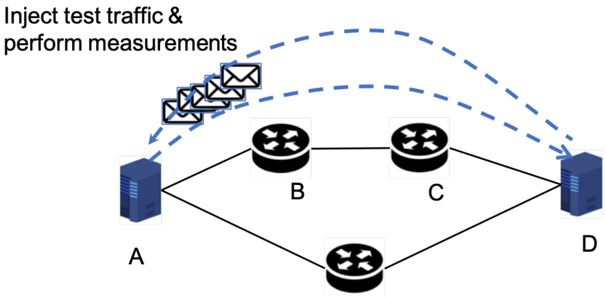


Figure 14 – Active Measurements

Active testing can be performed between successively longer paths along the network route, for example, from site 'A' to site 'B' or site 'A' to site 'C'. With this the operator can segment the overall end-to-end path so that performance indicators can be derived on a per segment basis, giving visibility into where issues might be located. Active monitoring is the primary method for policing service level agreements, since it provides a real-time view of performance. Active monitoring requires two end points to be able to create test traffic and respond back to complete the measurement

4.1.2. Passive measurements

Passive measurements are done simply by observing normal host-host interactions. Instead of measuring the latency of specially created test packets like in active measurements, passive measurements are based on the normal user packets that traverse the network. Passive measurements observe the traffic exchanged between two end-points and calculates the latency based on observed activity. For example, during normal interactions between host A and D, say during the initial handshake, a packet sent from A to D would be immediately responded by D as per the normal protocol interaction. If this transaction can be observed say at a location B, one can measure the time between sending the packet and receiving the response. Passive methods obtain similar measurements as an active measurement, without creating any new test traffic in the network, but are reliant on the presence of user traffic and can thus be skewed (for better or for worse) toward periods of time when more such traffic is present. Passive monitoring involves capturing and analyzing live network traffic, or traffic statistics, at a specific point in the network, for example at the network interface to an application server, or at an aggregation router.

Passive monitoring does not require another host in the network to be involved in the process. Passive monitoring involves capturing some, or all, of the traffic flowing through a port for detailed, offline analysis of things like signaling protocols, application usage or top bandwidth consumers. Passive monitoring is suited for in depth traffic and protocol analysis, and can give visibility into the customers actual quality of experience.

4.1.2.1. TCP Analysis

Analyzing the delay experienced by the TCP connection setup packets is an example of passive measurements. TCP uses a three-way handshake to establish a reliable connection. The TCP connection setup consists of a handshake with SYN, SYN+ACK, and ACK packets. The idea is to examine the data for outgoing connections, and compute the round- trip delay between the SYN & SYN+ACK packets as well as the SYN+ACK & ACK packet in the handshake. Since TCP connection endpoints normally respond immediately this is an easy way to compute the round-trip times.

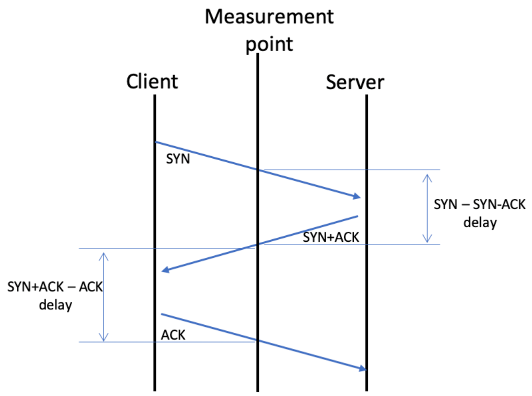


Figure 15 – Using the TCP handshake to measure latency

4.2. Industry measurement architectures

This section describes some of the commonly used measurement architectures.

4.2.1. SamKnows Whitebox (dedicated test device solution)

SamKnows has developed a “Whitebox”, a dedicated device with a test suite, for measuring internet performance. These Whiteboxes are used by service providers, government regulators etc. and the tests can also be incorporated into network devices like modems or routers. The [SamKnows] test methodology includes many aspects of measuring consumer broadband performance: providing consumer volunteers with the Whiteboxes to run tests on consumer internet connections, the mechanism for collecting and aggregating the data, and finally the format for presenting the data.

The following describes the overall latency measurement methodology followed by SamKnows Whiteboxes. As described in [SamKnows] literature, upon start up, the Whitebox runs a brief latency measurement to all measurement servers hosted by an operator, or hosted by Samknows on their behalf. The server with the lowest round-trip latency is selected as the target for all subsequent measurements.

Below are some of the latency specific tests that the SamKnows Whitebox, or routers with the test functionality can run, as described in the [SamKnows] documentation.

- Latency and packet loss (UDP): This test measures RTT of small UDP packets between the Whitebox and a target test server. Each packet consists of an 8-byte sequence number and an 8-byte timestamp. The test operates continuously in the background and randomly distributes the sending of the packets over a fixed interval, typically 2000 samples per hour. It then records the number of packets sent, the average round trip time of these and the total number of packets lost. The test uses the 99th percentile when calculating the summarized minimum, maximum and average results on each Whitebox.
- Contiguous packet loss / disconnections (UDP): This test records instances when two or more consecutive packets are lost to the same test server. Alongside each event it records the timestamp, the number of packets lost and the duration of the event. By executing the test against multiple diverse servers, an operator can begin to observe server outages or and disconnections of the user's home connection.
- Latency, jitter and packet loss (Fixed rate UDP test): This test uses a fixed-rate stream of UDP traffic, a bi-directional 64kbps stream (representative of the G.711 voice codec), running between the client and test nodes. The standard configuration uses 500 packets upstream and 500 packets downstream. The server and client record the loss rate and the jitter observed. Jitter is calculated using the PDV approach described in [IETF RFC5481]. The 99th percentile is recorded and used in all calculations when deriving the PDV.
- Latency and packet loss (ICMP): This test measures the mean round trip time (RTT) of ICMP echo requests in microseconds from the Whitebox to a target test node.

4.2.2. The M-Lab NDT (User initiated)

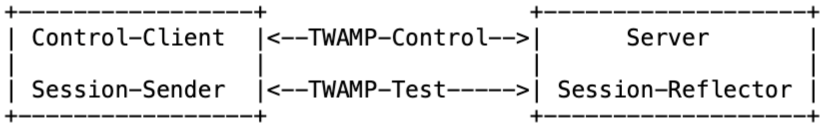
M-Lab is a consortium of research, industry, and public-interest partners, and provides an ecosystem for the open, verifiable measurement of global network performance. All of the data collected by M-Lab's global measurement platform are made openly available, and all of the measurement tools hosted by M-Lab are open source.

M-Lab defines a test suite known as Network Diagnostic Tool (NDT), which is a single stream performance measurement of a connection's capacity for bulk transport (as defined in IETF's RFC 3148). NDT reports upload and download speeds, and latency metrics. The M-Lab NDT is run by users to test their internet connections. As described in [M-lab NDT], when the test is run, the client attempts to pick the nearest server from the geographically distributed network of servers provided by the M-Lab platform. The test suite uses a 10-second bulk transfer from the server to the client. The server is instrumented with the TCP kernel instrumentation and captures several variables of the TCP state machine every 5 ms of the test. NDT uses the TCP RTT samples as the latency data points and reports the difference between the minimum and maximum RTT observed during a test run.

4.2.3. TWAMP

Two-way measurements are common in IP networks, primarily because synchronization between local and remote clocks is unnecessary for round-trip delay, and measurement support at the remote end may be limited to a simple echo function. [IETF RFC 5357] specifies the Two-Way Active Measurement Protocol (TWAMP), which provides a common protocol for measuring two-way or round-trip measurement between network devices.

The [IETF RFC 5357] TWAMP defines a standard for measuring round-trip network performance between any two devices that support the TWAMP protocols. TWAMP consists of two inter-related protocols: TWAMP-Control and TWAMP-Test. The TWAMP-Control protocol is used to set up performance measurement sessions, i.e. to initiate, start, and stop test sessions. The TWAMP-Test protocol is used to send and receive performance-measurement probes i.e. exchange test packets between two TWAMP entities. The TWAMP measurement architecture is usually comprised of two hosts with specific roles, shown below. The first host (controller) consists of the control-client which sets up, starts, and stops TWAMP-Test sessions, and the session-sender which instantiates TWAMP-Test packets that are sent to the session-reflector. At the second host (responder) the session-reflector reflects the measurement packet upon receiving the TWAMP-Test packet. The responder can also have the TWAMP server that manages one or more TWAMP sessions.

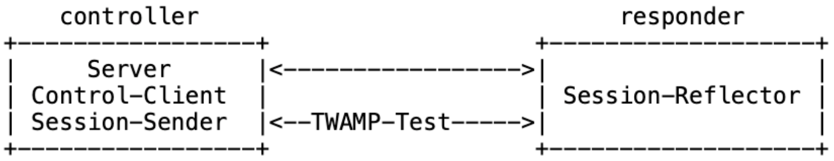


```
+-----------------+                 +--------------------+
| Control-Client  |<--TWAMP-Control-->|       Server       |
|                 |                 |                    |
| Session-Sender  |<--TWAMP-Test----->| Session-Reflector |
+-----------------+                 +--------------------+
```

[IETF RFC 5357]

Figure 16 – TWAMP reference Model

TWAMP Light is an alternative architecture which eliminates the need for the TWAMP-Control protocol and assumes that the Session-Reflector is configured and communicates its configuration with the Server through non-standard means. The Session-Reflector simply reflects the incoming packets back to the controller while copying the necessary information and generating sequence number and timestamp values. In TWAMP light, the roles of Control-Client, Server, and Session-Sender are implemented in one host (the controller), and the role of Session-Reflector is implemented in another host (the responder).



```
     controller                              responder
+-----------------+                 +--------------------+
|     Server      |<----------------->|                    |
| Control-Client  |                 | Session-Reflector |
| Session-Sender  |<--TWAMP-Test----->|                    |
+-----------------+                 +--------------------+
```

[IETF RFC 5357]

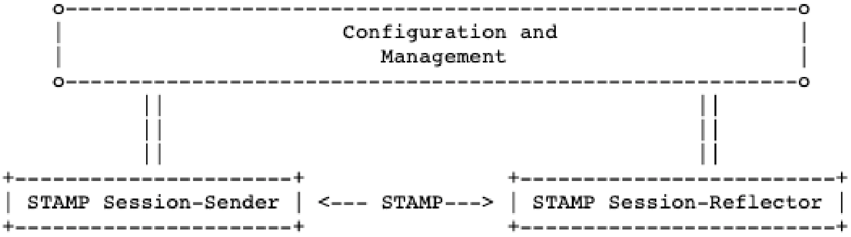
Figure 17 – TWAMP Light reference Model

TWAMP is more accurate than simple ping or traceroute measurements and is used by many operators in their transport, core and access networks. Several independent implementations of both TWAMP and TWAMP Light [IETF RFC5357] have been developed, deployed, and provide important operational performance measurements.

TWAMP is implemented in many of the core router products. TWAMP can provide accurate latency, jitter & packet drop KPIs, is supported by many probe vendors, and it can be integrated into network node equipment elements and CPE.

4.2.4. Simple Two-Way Active Measurement Protocol

Simple Two-way Active Measurement Protocol (STAMP), is a newer IETF standard [IETF RFC 8762] which provides a simpler mechanism for active performance monitoring. It separates the control functions (vendor-specific configuration or orchestration) and test functions. STAMP also enables the measurement of both one-way and round-trip metrics (delay, delay variation, and packet loss)



```
o-----------------------------------------------------------o
|                   Configuration and                       |
|                      Management                           |
o-----------------------------------------------------------o
         ||                                          ||
         ||                                          ||
         ||                                          ||
+----------------------+                +-------------------------+
| STAMP Session-Sender | <--- STAMP---> | STAMP Session-Reflector |
+----------------------+                +-------------------------+
```

Figure 18 – STAMP reference Model

4.3. Measurement considerations

4.3.1. Measurement under load vs quiet times

Latency measurement tests need to ensure that testing is done over a variety of times to understand the variation between when the network is relatively lightly loaded and peak load time. This can also be used to measure self-congestion vs. network-congestion. For example, measuring latencies at peak time in the evening when most of the subscribers on the plant are online is likely to catch incidents when the network is congested due to high overall network load. Another way latency numbers can be affected is the load within a single user’s home itself or even within the same client. If multiple devices in the home are using the network for various purposes like consuming video, voice calls, gaming, then a latency measurement will likely yield different numbers than running the test at quiet times. The path to various servers can change automatically to accommodate network/routing changes, so measured latency may vary over time and it may be appropriate to get a broad picture of latency including such situations.

4.3.2. Window over which the measurement is done.

Every latency measurement test can have a different purpose; one could be for a quick and immediate diagnostic purpose, while another could be to gather long term statistics. To diagnose issues in the network, an operator will need to consider the correct amount of time to run a test, how many latency samples will be collected in each run, and how often the test will be run. This could include sample rates of once per hour, once per minute, once per second, and as frequently as 50 times per second. The sampling rate and the number of measurements run will depend on the ultimate goal of the operator. If the goal is to reflect the worst gaming experience, then more measurements which mimic the game traffic flows will give us a better idea of the performance of the network.

To understand latency, one has to consider the entire distribution of latency measurements. While it is important for operators to look at latency numbers at the 99.9th percentile or higher, many monitoring systems stop at the 90th or 95th percentile. The reason is simply because it requires larger amounts of data to be collected, stored and analyzed. The data collected by most monitoring systems is usually summarized in small, five or ten second windows. Given we can’t meaningfully average percentiles or derive five nines from a collection of small samples of percentiles, there is no way to confidently know what the 99.99th percentile for the minute or hour was. A related question is how many total samples are needed to get valid statistics. If an operator wants to measure the 99.9th percentile latency, then at least 1000 latency measurements are required, and a lot more (at least 2000-8000) would be needed to have an accurate statistical estimate.

4.3.3. Off-net and On-net testing

Active measurement architectures (e.g. SamKnows) may use client devices which run bandwidth and latency tests to a specific measurement server. A majority of test servers used by SamKnows customers are off-net, i.e. hosted on the internet outside the operator network. Reporting results to target servers off an ISP's own network represents a 'real world' experience for end users. However, an ISP is not in control of the paths that get to the server and would also like to understand and debug issues within their own network. Hence many ISPs install test servers inside their network ("on-net") to allow them to segregate on-net and off-net performance.

With both on-net and off-net servers in use, operators can see the difference in performance internal to their network vs. external to it. The results can be used to troubleshoot peering links, routing issues, or simply rule out any capacity problems within the operator's own network. Consequently, any active measurement deployment should have a mix of on-net and off-net servers.

4.3.4. Latency Measurement Test definitions

When designing latency measurement tests, an operator needs to define the test and the associated parameters such as the test traffic flow (i.e. the packet size and rate used), whether to test under load or without load, and the periodicity of the measurements.

Many IP network switches and routers need the full packet to be clocked into the device before it can be forwarded to the next networking device in the path to the end destination. This delay is referred to as a serialization delay and these delays are often tested using 64-byte packets. For example, a 64-byte packet will have serialization delays of 5.12 µsec when clocked in using a 100 Mbps port. However, serialization delays are usually proportional to the size of the packet. If the size of the packet was 1280 bytes, the serialization delays would be twenty times bigger at 102.4 microseconds. Though this doesn't include the processing delays through a device (router, switch, CMTS, CM), it gives a sense of the interaction between packet size and link speed (interface bandwidth) that each node in the network could add as an absolute minimum

Small (say 64 byte) UDP packets sent every few seconds from a test node is a good place to start for RTT measurements. Latency Tests which mimic the gaming experience, (e.g. 150 Kbps upstream, 600 Kbps downstream, ~200-byte packets) would be a good data set to collect to understand the impact to gaming or other real-time audio-conferencing services. Latency tests with bigger size packets (1500 bytes) could also be used to expose any packet size limitations in the network.

When testing latency, it is also a good idea to understand the latency when the network is under load vs. when it is not. Latency testing with load is typically done by running both a downstream and upstream speed test or something equivalent at the same time as doing latency measurements. While the speed test is running, the latency under load test can send packets to a target server and measures the round-trip time and number of packets lost. The test packets should be sent equally spaced over the duration of the speed test.

4.3.5. Marked traffic vs Unmarked traffic.

Differentiated services or DiffServ [IETF RFC 2474] specifies a simple mechanism for classifying and managing network traffic and providing quality of service (QoS) on modern IP networks. DiffServ can, for example, be used to provide low-latency to critical network traffic such as voice or streaming media while providing simple best-effort service to non-critical services such as web traffic or file transfers.

The six most significant bits of the DiffServ field (previously 'type of service' (TOS) field) in the IP header are called as the DSCP (differentiated services Code point) and the last two bits are the Explicit Congestion Notification (ECN) field. Routers at the edge of the network classify packets and mark them with their DSCP value in a Diffserv network. Other network devices in the core that support Diffserv use the DSCP value in the IP header to select a per hop behavior for the packet and provide the appropriate QoS treatment. Various applications and services (typically UDP based) can also mark the traffic they generate with specific DSCP values. For example, the popular video conferencing application Zoom uses a default DSCP marking values of 56 for audio, 40 for video, and 40 for signaling.

In the Low Latency DOCSIS technology, by default, the traffic within an Aggregate Service Flow is segmented into the two constituent Service Flows by a set of packet classifiers that examine the DSCP field and the ECN field. Specifically, packets with a Non-Queue Building DiffServ value, 0x2A, per a current [IETF NQB PHB] draft, will get mapped to the Low Latency Service Flow, and the rest of the traffic will get mapped to the Classic Service Flow.

In the context of Low Latency DOCSIS and other technologies which support dual queue mechanisms, the question is how can we modify latency measurement tests to also report metrics on unmarked traffic as well as marked traffic. One solution is to run any test twice, once as currently designed without any packet marking, and once with marked DSCP packets, and report results on both. As more games and other applications start marking their packets, public internet measurement reports will also have to start reporting latencies on both types of traffic.

5. Conclusion

Interactive applications like gaming or real-time communication, where real-time responsiveness is required, are more sensitive to latency than bandwidth. These applications really stand to benefit from technology that can deliver consistent low latency. Operators need to understand the latency characteristics of their network and be able to delineate the latencies seen in the customer home, the access network, and the MSO-core network. Using a common set of metrics to describe latency is the first step in understanding the state of the networks. Round trip times are relatively easy to collect compared to one-way latencies. Multiple sets of measurements paint a better picture of the latency characteristics than single measurement. Using averages to measure latencies can be misleading, so an operator can choose better performance indicators such as the 99th or 99.9th percentile to track and understand latency behavior over time. Latency is being measured by national entities, raising the importance of operators to have their own latency measurement infrastructures. Active measurement techniques give an operator good control over the testing and a better understanding of the network over various times and conditions.

Abbreviations

bps	bits per second
ms	millisecond
RTT	Round trip time
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
IETF	Internet Engineering Task Force
RMS	Root Mean Square

Bibliography & References

[ITU-T G.114] *One-way transmission time https://www.itu.int/rec/T-REC-G.114 , Recommendation G.114 (05/03) & Annex B* ITU-T G.114 (05/2000)

[QoE and Latency] Saldana J., Suznjevic M. (2015) QoE and Latency Issues in Networked Games. Handbook of Digital Games and Entertainment Technologies. https://doi.org/10.1007/978-981-4560-52-8_23-1

[BelsheM] *"More Bandwidth Doesn't Matter (much)": Mike Belshe, Google https://www.belshe.com/2010/05/24/more-bandwidth-doesnt-matter-much/*

[Greg W, SCTE 2019] *Low Latency DOCSIS Overview And Performance Characteristics, SCTE 2019* , Greg White, Karthik Sundaresan, Bob Briscoe

[MBA FCC] Ninth Measuring Broadband America Fixed Broadband Report https://www.fcc.gov/reports-research/reports/measuring-broadband-america/measuring-broadband-america-program-fixed

[MBC CRTC] Measuring Broadband Canada Report https://crtc.gc.ca/eng/publications/reports/rp200601/rp200601.htm

[EU Broadband] European Commission Broadband Connectivity https://ec.europa.eu/digital-single-market/en/connectivity

[SpeedTest] Speed Test reports https://www.speedtest.net/global-index/united-states#fixed

[ITU-T Y.1540] *Recommendation ITU-T Y.1540, 2019, Internet protocol data communication service –IP packet transfer and availability performance parameters*

[IETF RFC 5481] *Packet Delay Variation Applicability Statement*

[IETF RFC 3550] *RTP: A Transport Protocol for Real-Time Applications*

[3rdEchelon] http://www.3rdechelon.net/jittercalc.asp

[Haste] https://haste.net/2017/08/23/what-is-jitter/

[Excentis ByteBlower] https://www.excentis.com/products/byteblower

[Network Next] https://www.networknext.com

[WTFast] https://www.wtfast.com/en/

[M-lab NDT] https://www.measurementlab.net/tests/ndt/

[SamKnows] *https://samknows.com* , *https://samknows.com/technology/tests/latency-loss-and-jitter#latency-jitter-and-packet-loss-udp*

[PingPlotter Pro] https://www.pingman.com , https://www.pingman.com/kb/article/what-is-jitter-57.html

[IETF RFC 5337] *A Two-Way Active Measurement Protocol (TWAMP)*

[IETF NQB PHB] *A Non-Queue-Building Per-Hop Behavior (NQB PHB) for Differentiated Services, draft-ietf-tsvwg-nqb-01, G.White*