Pipeline
Technology for Service Providers.
www.pipelinepub.com Volume 14, Issue 2

Resolving Latency Issues That Affect Carrier Networks and Quality of Experience

By: Scott St. John

The speed with which data travels from one side of the network to a subscriber's end point can mean the difference between a positive or negative Quality of Experience (QoE) and subsequent loyalty or churn.

Delivering the required transport performance — and making the customer happy with a solid QoE — has become more challenging with the insatiable demand for data and the expanding number of data-hungry digital services.

Transport performance can suffer for many reasons: excess packet loss, latency (delay) and jitter (delay variation). Of those variables, latency appears to have the biggest impact on customers.

Latency can be simply thought of as the time it takes to send a unit of data between two points in a network. It is not fixed and varies over time, and is a natural consequence of utilization of the entire network — not only subscribers' services.

By its nature, latency is highly asymmetrical, just as traffic on a busy urban highway is congested in one direction in the morning commute, but slowed in the other direction in the afternoon commute.

Increasingly, the QoE of Internet-based applications is sensitive to latency. Packet loss is relatively rare and easy to measure and manage: either the packet shows up, or it does not. But, latency is trickier, and you can't tell a network, "No delays, please." Without proper attention paid to the problem, QoE can suffer greatly due to latency.

Enterprise customers become frustrated with their cloud apps, background tasks time out or freeze, calls and sessions are dropped, and potential customers with ever-shorter attention spans click away when a website or service appears to be unresponsive.

According to Gartner, "High latency tends to have greater impact than bandwidth on the end-user experience in interactive applications, such as web browsing." Amazon once reported that a 100-millisecond delay in serving web pages decreased online sales by one percent. Similarly, Google has said that slow response time reduced the number of searches and, therefore, reduced the ability to serve ads. This is a big deal, especially when some studies say that as much as 80 percent of network traffic is affected by latency issues. Carriers can no longer rely on the most common method of understanding delays: global positioning systems (GPS) synchronization (see "GPS-Based Clock Synchronization," below).

Network Latency and Asymmetry

There are four main causes of latency across end-to-end carrier networks:

- **Transport —** The longer the links, the more delay experienced by packets as they traverse the links. Network topology can affect the transport, as well as physical distances between nodes. It also takes time for TCP to establish connections. Transport-related delay cannot be managed under most circumstances.
- **Congestion—**As more traffic traverses links or routers, there can be bandwidth contention in the transport layer, or resource congestion in forwarding devices. Congestion can be caused when an unexpected amount of traffic inundates a node, when high CPU or memory

utilization exists, or when packet congestion ensues because of problems elsewhere on the network — all of which require substantial routing changes.

- **Processing —** Although many network forwarding devices are billed as "wire-speed," in reality, many are plagued by processing delays when determining routes and how to best forward packets. This is particularly true in highly virtualized networks, such as those managed by SDN/NFV, because of the intense processing required on busy devices.
- **Routing changes —** In an IP-based network, not all packets take the same path, so if some traffic traverses longer routes, it results in latency issues, excess jitter, and out-of-order packet transmission. In extreme cases, packets may be considered lost, and the service may be instructed to resend, even if those packets show up later, which can further exacerbate jitter and latency issues.

Bear in mind that in any network, the network congestion and even the routes will be asymmetrical, so that, for example, streaming a game's video quality might be satisfactory when the user is passive, but at the point the server becomes less responsive to the player's clicks or mouse movements, the gaming experience begins to feel less responsive.

"Not only that, but the carrier services may be intentionally asymmetric in terms of bandwidth and directional performance," explains Thierno Diallo, a product manager and expert in packet technologies for EXFO.

"Your Internet provider may have told you that you have 10 megabits downstream, but only four megabits upstream," Diallo explained, adding that this is okay because "we consume more data traveling from the network toward us than that which we push up to the network. So when we look at delay loss, and performance in general, it is essential to look at it from a directional perspective. It's not sufficient to look at round-trip performance."

Most services are asymmetrical: games, broadcasting, movie streaming, software updates, and business applications, with far more traffic heading downstream toward the customer, and relatively little heading upstream toward the server or cloud provider.

It is critical to take into account this asymmetry. It is not enough to focus on downstream latency. Both upstream and downstream latency must be measured, and measured accurately, in order to understand the overall session end-to-end latency, and to use that data to improve QoE.

GPS-Based Clock Synchronization

Time is relative. When measuring latency in sub-milliseconds, it is essential to ensure that accurate timestamps from synchronized clocks are used when measuring unidirectional packet delays.

The most common approach has been to use accurate time signatures from GPS, which requires extremely accurate timings, down to the microsecond, based on satellite transmissions. Because GPS does not require that network nodes sync with each other, it is the *de facto* standard; however, the technology is costly to acquire, deploy and manage.

First, not all endpoints possess GPS capability or the capability to be directly connected to a GPS device. Second, distance is a factor: the farther away the GPS, or the more complex the network topology, the less accurate the clock synchronization. Third, even GPS-equipped network devices are not always active or usable because some utilize GPS for networking functionality as opposed to functionality testing.

And last but not least, a major downside of GPS-driven approaches is the fact some carriers measure round-trip packet latency and then divide that measurement by two when calculating an approximation of one-way delay. While this avoids the need for direct clock synchronization, it can be very inaccurate when measuring upstream or downstream performance independently. It can provide a false representation of QoE, and make it difficult to determine where a problem may lie, or how to rectify it.

Active Testing of Latency in Both Directions

“As we get into more transactional services and become a more mobile society,” said EXFO’s Diallo, “the importance of looking at everything from a unidirectional perspective is significantly more important for carriers and enterprise customers. Carriers must gather those metrics in as efficient a manner as possible.” Based on the data produced from those metrics, carriers can gain a real understanding of where delay occurs in the network and subsequently improve traffic engineering to reduce latency.

This can be done without a true end-to-end clock synchronization, even through mobile backhaul, indoor deployments, or mobile devices may lack reliable clocks or access to GPS. Through active testing, it becomes possible to conduct GPS sub-microsecond measurements, even when devices lack access to GPS receivers or other similar clock sync resources.

EXFO is leading the development of this technology, which it terms “Universal Virtual Sync.” According to Diallo, the technology leverages EXFO’s existing Active Verifier physical or virtual network probe devices. “It has the ability to learn about the endpoint, and about the responder on the far end,” which means the software can locally provide a correction factor (a clock offset) that provides a very accurate view of one-way latency.

Using this technology, the responder has a hardware-based timing system that leverages the RFC-5357 TWAMP (Two-Way Active Measurement Protocol) or Y.1731 SOAM-PM standards found in most switches, routers, cell-site routers, and other similar devices.

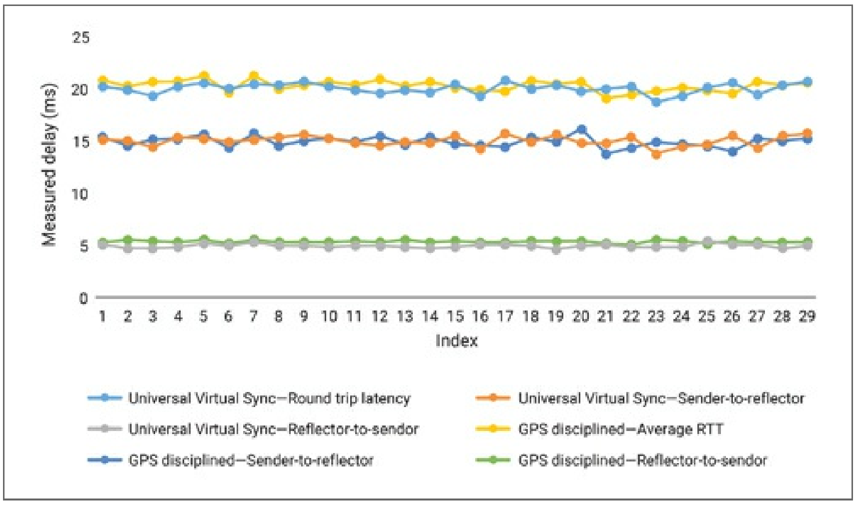

Fig. 1 – GPS vs. One-way Latency Testing

“We’ve set up a mechanism that allows for the very accurate measurement of one-way delay to any standards-based responder,” he explained. “You can leverage the same monitoring used to assure your backhaul or Layer 2/3 circuits.” Based on that data, network managers can still calculate two-way traffic metrics when needed, but have the one-way traffic data required for spotting problems and remediating issues.

EXFO’s new approach has the potential to reduce CAPEX by reducing the need for GPS-based time synchronization devices, which can be expensive to install and maintain. It also has the potential to reduce OPEX by helping to identify problems more accurately and to reduce troubleshooting time. And of course, there are bottomline benefits of more accurately measuring and maintaining a positive QoE to improve customers' loyalty and reduce churn.

Evolving From Round-Trip Metrics To Bidirectional Latency Measurement

Cisco’s Visual Networking Index predicts IP-video traffic to be 82 percent of all consumer traffic by

2021. That means more than 80 percent of network traffic can be significantly affected by latency, and if the delays raise the potential to affect QoE, traffic engineering or other steps can be taken to address the underlying problems and reduce that latency. While the latency caused by transport distances generally cannot be resolved, other causes of latency can be, but only if latency metrics are based on accurate unidirectional metrics that not only show the proper upstream or downstream delay, but also their origin.

For those still measuring network latency by dividing in half round-trip packet time, the time has come to abandon that approach. It has become too expensive and inaccurate for today’s digital service providers. The better path is to embrace more accurate bidirectional latency measurement, which can not only improve customer satisfaction, but also potentially reduce both OPEX and CAPEX.