Low-latency service

for enhanced broadband services

by CommScope

In this article, Ian Wheelock of the Home Networks CTO office describes the causes of network latency and its increasingly significant impact on end-user applications. He reviews the technology available to address it and how service providers can bring these techniques together to offer enhanced services.

What is latency?

One of the latest and most significant challenges in the broadband industry is network latency.

Put simply, it is a measure of time delay between endpoints on the internet—normally measured in milliseconds (ms). High latency or highly variable latency, also known as jitter, can impact application performance, especially interactive applications such as gaming, VR/AR, video conferencing, and wireless backhaul. Latency is introduced due to distances between users and websites as well as buffering of traffic between different types of networks with different capacities. Parallels can be drawn between the internet and the physical road network, where a mixture of roads with different capacities interconnect at various junctions.

Like the road network, traffic volumes and interconnections with different speeds can all lead to congestion, queuing, and a significant impact on network performance.

Broadband latency

Shared broadband networks (i.e., cable or fibre) act like a high-capacity highway with hundreds of on-ramps (subscribers) along its length. When there is very little traffic, all the on-ramps work well, and there is practically no latency in joining the highway. Broadband networks ensure that all these on-ramps can access the highway based on their service level agreement (SLA). When on-ramp traffic approaches its maximum access speed, traffic can queue at the on-ramp waiting to get onto the network. This queue build-up introduces latency. Like in the real world, certain high-priority traffic such as emergency vehicles need to get onto the highway with minimal delay.

Low latency can be achieved by organising the on-ramp into multiple queues, allowing emergency vehicles into a lightly loaded queue while the remaining traffic continues to join the busier general queue. Access to the highway is still controlled by the SLA (traffic lights), but now access is shared between the head of the multiple queues, resulting in the small number of emergency vehicles getting out ahead of the other big queues that have built up. Other options exist to address this issue in networking equipment, such as reducing the maximum size of the on-ramp queue by dropping traffic from the queue completely; however, this approach won't work in the real world. The longer the wait in the queue, the higher the latency.

Most internet applications can cope with some amount of packet loss, while some protocols rely on packet drops and network latency to help them identify a transmission rate equilibrium that works for the specific network.

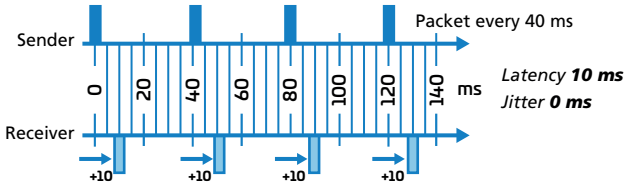

Figure 1: 10ms latency service with no jitter

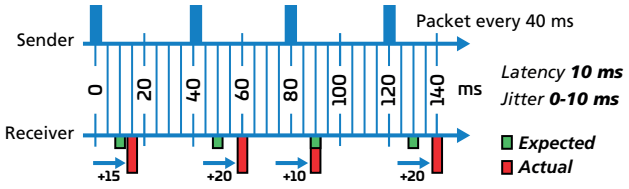

Figure 2: 10ms latency service with up to 10ms of jitter

Jitter

Jitter is a measure of latency variability. Comparing to vehicles on a road network, if 10 cars joined the network at 10-second intervals and arrived at their destination at the same 10-second interval, then there is no jitter. However, if they arrive at different intervals or even out of order, then they have experienced jitter.

Packets in the network operate similarly—if a traffic source sends 25 packets per second with 40ms spacing, and the receiver gets these packets in order with the same 40ms spacing, then there is no jitter. But if they arrive with different spacing or out of order, then the link has jitter.

To explain the key difference between latency and jitter: If every packet sent was received 10ms late, then this equates to 10ms of baseline latency (see Figure 1 and 2). However, if packets begin to arrive with variable delays such as +0ms, +5 ms, or +10ms, then the observed jitter is between 0 and 10ms as shown in Figure 2. Note that jitter can be much more variable than this example.

Round trip time

Latency on the internet is measured and reported using round trip time (RTT), or the travel time to a network location and back again. Like a road trip, the further you need to travel on the network, the greater the time to get there. Most speed test tools will report RTT to the closest server they use, but it is possible to test with more distant servers. RTT between London and New York is about 71ms, meaning each network crossing of the Atlantic takes 35.5ms, while the RTT between London and Dublin is about 11.5ms. In networking terms, this is known as the "propagation delay," and it is determined by the geographic distance and the speed of the networks used. Any delays that exceed the propagation time are considered increased latency, and any variability on that increased latency is, as explained earlier, considered jitter.

Wi-Fi latency

The above example talks about broadband networks, but wherever networks interconnect with different speeds or technologies, latency can be introduced. In the case of Wi-Fi, even though it offers very good connectivity, it can be difficult to support specific SLAs due to the nature of Wi-Fi channel access (or how devices connect to the Wi-Fi network). In a dense Wi-Fi deployment with a lot of devices transmitting, each device may need to queue traffic if it gets delayed accessing the Wi-Fi channel. Like the highway analogy, any traffic being queued will increase latency. And like the highway, each device supports different types of application traffic with unique latency demands.

The Wi-Fi WMM feature helps address this - effectively creating four different priority queues for all the traffic using a Wi-Fi device (see Figure 3). The Wi-Fi chipset transmits appropriate traffic from each queue in a prioritised order when it receives access to the Wi-Fi traffic channel, much like how the on-ramp queue analogy works. This approach reduces the amount of latency

Wi-Fi networks are based on each Wi-Fi device deciding when to access the Wi-Fi channel—in most cases by listening to hear if anyone else is transmitting, and sending if the channel is clear. The system is called Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA), and it works very well. Wi-Fi networks can support extremely high speeds based on this technology. However, it can suffer if there are large numbers of connecting devices, and throughput can drop for all devices if a portion of devices are far away from the Wi-Fi access point.

The Wi-Fi standards groups identified the latency issue many years ago and introduced a scheme to categorise traffic to different groups and prioritise the transmission of these groups. Access categories (ACs) exist for voice, video, best effort, and background traffic. This solution is called Wi-Fi Multi-Media (WMM) or 802.11e. Not only does the scheme queue traffic per category, but each category is configured to access the Wi-Fi channel with a different airtime priority and configuration (see Figure 3).

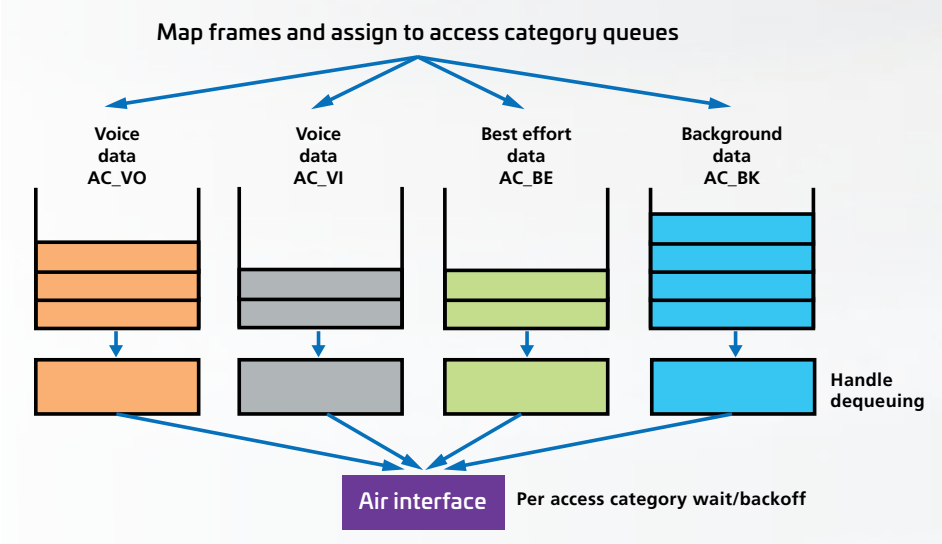

Figure 3: Wi-Fi Multi-Media queue scheme

encountered by some applications, while enabling latency-tolerant traffic to be supported as well. It is used extensively in enterprise settings, especially where voice calls are concerned.

Other improvements in Wi-Fi 6 include the ability for the Wi-Fi access point (AP) to schedule transmit times for stations to use. This uses a traffic scheduler within the AP (similar to schedulers in multi-access broadband networks such as DOCSIS or xPON) to instruct Wi-Fi devices as to what

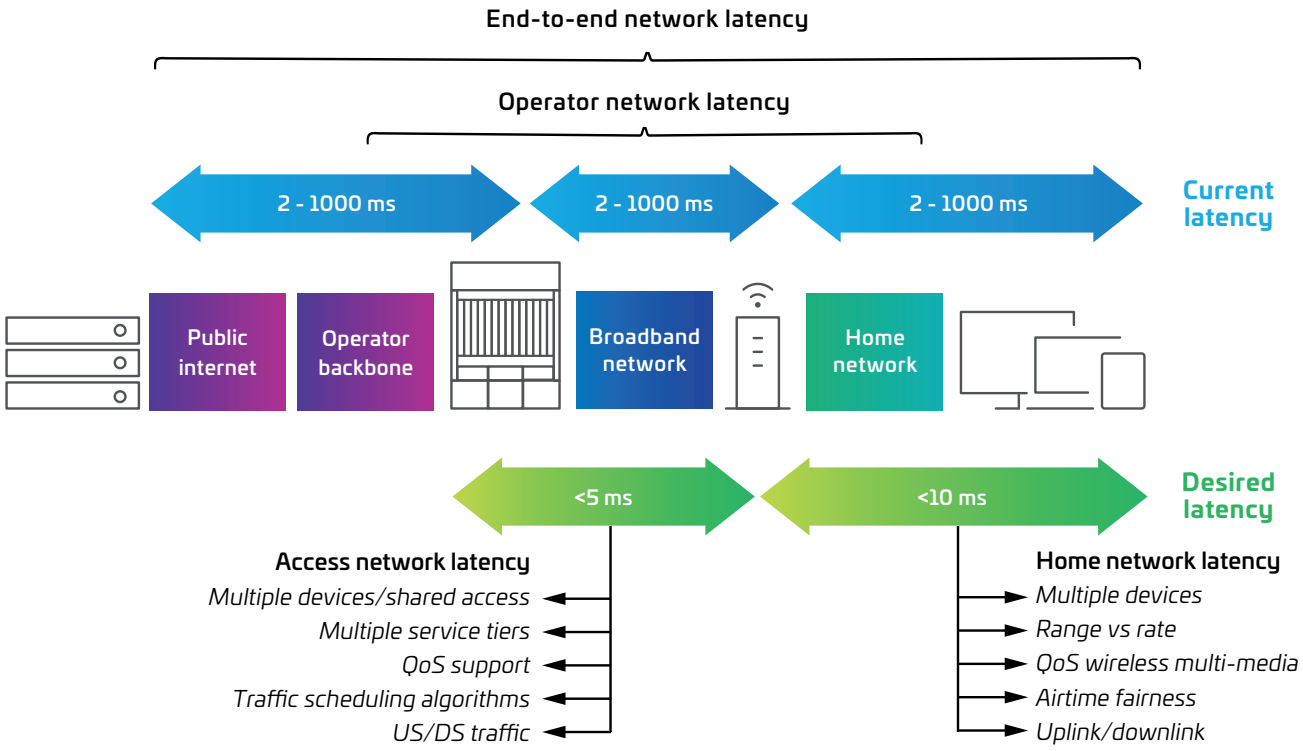

Figure 4: Sources of latency in broadband and home networks

time and for how long they can send traffic to the AP. Other complicated extensions in Wi-Fi 6 allow multiple devices to communicate concurrently with the AP on different parts of the available Wi-Fi channel. Combined, these two features enhance Wi-Fi performance to reduce both latency and channel access times for connected devices.

Another major enhancement, Wi-Fi 6E, improves available capacity by extending the amount of spectrum by 500 to 1,200MHz in the recently released unlicensed 6GHz frequency band. Having extra capacity for new Wi-Fi devices helps significantly—much like adding extra lanes in both directions to an existing highway. New devices are already available that are ready to use Wi-Fi 6E and its expanded spectrum. Specification work is already underway for the next version of Wi-Fi, called Wi-Fi 7, which aims to support up to a 30Gbps transmission rate along with more sophisticated bonding options to combine multiple bands simultaneously, while combining AP transmissions to individual devices as well. Wi-Fi 7 has significant potential in reducing latency between APs and connected devices.

Interconnected networks

Since the internet is a collection of interconnected networks, it must be pointed out that poorly performing networks within an end-to-end link (see Figure 4) will impact the overall link performance. The consequences of latency being added due to poor network hops can be remedied only by upgrading those hops. For example, if a poor Wi-Fi link is connected to an excellent broadband network, the latency of the Wi-Fi network will impact the overall performance. Similarly, if a Wi-Fi link operates very well, but connects to a poor broadband network, the performance will be impacted. Making sure that all networks across an end-to-end link operate at their best is the fundamental requirement to deliver low-latency service for applications that need it.

Why are latency and jitter important?

Often, as latency increases on a network, application performance degrades. This can be identified as a reduction in network throughput for bulk transfer applications (e.g., video streaming, file upload/downloads, and cloud sync) or interactive applications stalling, breaking up, or experiencing reduced quality (e.g., voice calling, video conferencing, and online gaming sessions). Activities such as web browsing can also be impacted by high latency, which is visible through long page load times and often occurs when many different page assets are downloaded over high-latency links. Other activities such as file sharing or the use of corporate applications can also be impacted by high latency—resulting in much longer times to access files or timeouts and the need to refresh screens frequently within business applications.

In the case of new extended reality (XR) applications (e.g., VR/AR), latency plays a major role. Motion to photon (MTP) describes the latency between a tracked object and how it is rendered in a headset. With a lot of XR applications operating in the cloud, reducing latency to the user's headset is a key feature requirement to enable XR. High MTP can cause a complete breakdown of the XR application; but, even if the application can work, the headset feedback to a server also needs to be delivered with minimal delay, and the corresponding XR data must be sent back to the headset quickly. Delays in getting information to the headset can negatively impact the experience, and worse, it can induce nausea for people wearing the headset. Depending on the XR application, a high bit rate may also be a requirement to support high-resolution rendered scenes—putting an additional constraint on the internet connection being used.

Any application interaction with a remote server will take at least 1 RTT, based on sending information to the server and waiting for a response. If an application needs to talk to a server 500 times, then, at 20ms RTT, it will take 10 seconds to complete. At 50ms RTT, it will take 25 seconds to complete. Most people have experienced application stalls at times due to latency across the network, such as a London-based user accessing SharePoint in San Diego where the RTT is 140ms (see Figure 5). Note that some advances in recently launched low-Earth orbiting (LEO) satellite arrays may reduce the latency for some of these long-distance RTT numbers due to intra-satellite laser links and optimised paths. Network protocol advances, such as QUIC from Google, have helped improve application performance by optimising these network interactions. These new advances are slowly being adopted, but not used by every application. Active queue management (AQM) techniques have also been developed and can reduce

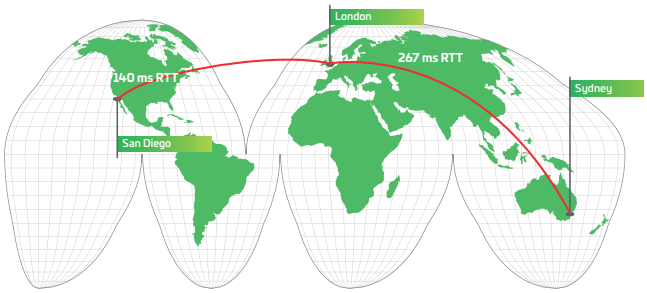

Figure 5: Example round trip times across the globe

latency build-up in network equipment. Solutions like fq_codel and CAKE have been contributed to the Linux kernel and used in some networking platforms, while another AQM technique, proportional integral controller enhanced (PIE), is implemented in DOCSIS 3.0 devices. AQMs are very effective, especially at the network edge, but they need more widespread adoption to help improve network latency.

Cable broadband latency improvement features

The earlier explanation of latency and the road network analogy included some information about queuing. Effective queuing and queue processing algorithms are key to low-latency systems. As mentioned, the internet serves multiple applications—some need low latency while others can cope with reasonable latency. Wi-Fi is not the only standard to have added specific features like WMM to address QoS/latency. The DOCSIS standard used by cable broadband networks has had updates for AQM and low latency features over the latest two major specification releases. The newest feature, DOCSIS Low Latency, uses dual queues (based around L4S)—one for low latency and the other for all the remaining traffic— aligning nicely with the on-ramp concept of the earlier analogy.

The feature makes sure that any latency-sensitive traffic that arrives for transmission can be delivered to the low-latency queue. The feature shares the bandwidth available at the DOCSIS modem among the dual queues—keeping within the customer SLA while ensuring latency-sensitive traffic from the home does not get stuck at the back of a long queue that creates high latency. Beyond the traffic queuing aspects, the key element to enable this solution is traffic classification—the ability to identify traffic that either needs low latency or is explicitly asking for it.

Identification of latency-sensitive traffic

The traffic being acted on by the DOCSIS modem originates in the home from devices connected by Ethernet, Wi-Fi, etc. (e.g., laptops, tablets, phones). In a perfect scenario, the applications on these devices (e.g., video streaming, voice calling, online gaming, email) will identify the latency needs of specific traffic by using certain network packet header fields. The local Wi-Fi subsystem can act on these packet header fields and categorise traffic to the appropriate WMM queue—ensuring it gets prioritised over the Wi-Fi link. The DOCSIS modem can then use the same network packet headers to identify traffic as low latency, and queue it accordingly. By enabling both the Wi-Fi and DOCSIS devices with low latency features and having applications correctly identify traffic, the network is well equipped to deliver low-latency traffic.

To classify traffic, CableLabs has suggested the use of a new quality of service (QoS) value to be applied to low latency or so-called nonqueue building (NQB) traffic. This is set with the Differentiated Services Code Point (DSCP) field in the network packet. Applications need to be updated to use this new QoS value. In the absence of the NQB value, DOCSIS devices can perform basic traffic analysis or deep packet inspection (DPI) to identify traffic that needs low latency treatment. In addition to the NQB value, the DOCSIS Low Latency feature also incorporates a safety measure that monitors traffic in the low latency queue for queue building. This Queue Protection (QP) feature analyses traffic microflows and moves any queue-building traffic from the low-latency queue to the other queue in the dual-queue system. Another feature, Explicit Congestion Notification (ECN) control, is also used to signal any congestion that may occur in the traffic queued in the cable modem.

Broadband service provider role

Latency is cumulative, in that—for every network hop encountered to the destination—the latency of that hop is added to the overall network latency. Like the road network, if you experience a 10-minute delay on part of a long trip, it is nearly impossible to recover that time without breaking the law. There is no single entity responsible for the end-to-end internet network, but broadband service providers have a lot of control over two significant components: the broadband access network and the home Wi-Fi network.

As most traffic is bidirectional, from a service provider view, there are four individual hops to be considered: the upstream (from devices) and downstream (to devices) of the access network as

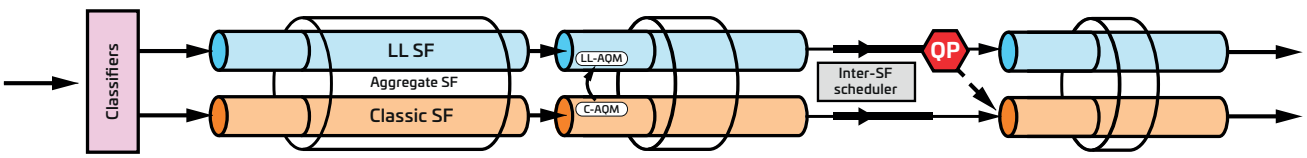

Figure 6: Low Latency DOCSIS service flow mode

Buffering can also occur in Wi-Fi to improve the efficiency of the network, where multiple packets are aggregated to form a long single transmission. Buffering 16 individual packets together allows a device to access the Wi-Fi channel once rather than 16 separate times.

well as the uplink (from devices) and downlink (to devices) of the Wi-Fi network. Most broadband networks are asymmetric, with a 10x difference between downstream and upstream speeds.

Some broadband access networks offer <10Mbps or <20Mbps upstream—the on-ramp speed mentioned earlier. Wi-Fi networks have continued to get faster, with sticker speeds of 1,000Mbps to 2,000Mbps quite commonplace (the actual speeds are lower, but still very fast). The difference between the Wi-Fi and broadband speeds illustrates the interconnection problem between networks connecting in the broadband gateway. In a perfect world, all uplink traffic from the Wi-Fi network would eventually match the broadband upstream speed, and there would be no need for buffering in the gateway. However, the network speeds available can change too quickly to enable a perfect speed match—resulting in some amount of buffering required to maintain optimal network performance. This buffering introduces latency, so employing WMM to ensure the Wi-Fi network is optimised for both the uplink and downlink as well as using DOCSIS Low Latency/optimised xPON DBA (dynamic bandwidth allocation) algorithms on the broadband upstream and downstream is essential.

CommScope low-latency system

The WMM and DOCSIS Low Latency features are available in the latest software loads running in many of today's broadband gateways. These features are normally direct implementations of what each specification calls for but, in some cases, these features are not enabled at all—or only the most basic aspects of the features are implemented to claim compatibility. The features are fundamental tools that need additional support and control to extract the most benefit from them, including the abilities to map specific traffic for low latency; to achieve continuity between the two different network domains; and to monitor, trust, and verify the operation of these features. Some features have been developed to provide QoS features and to

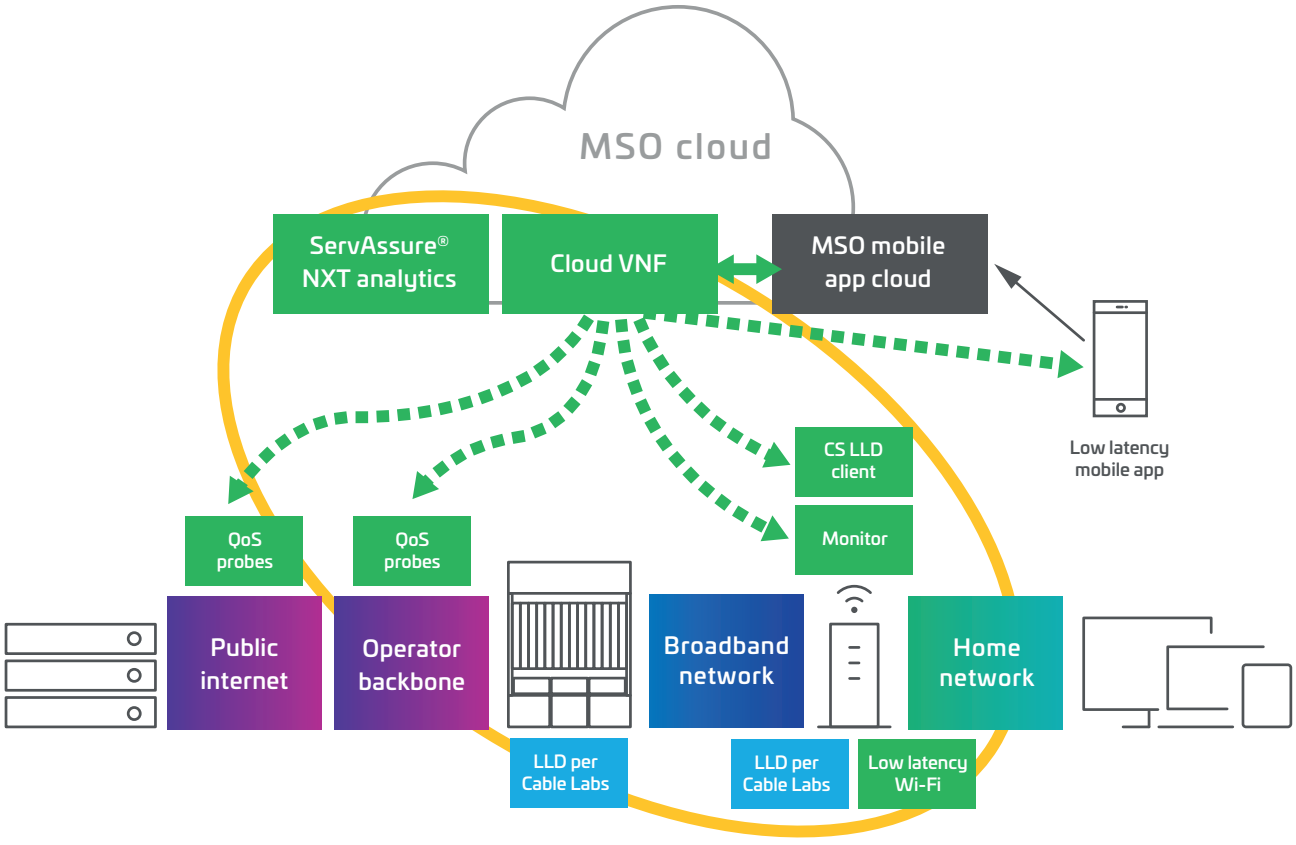

Figure 7: CommScope low latency system for broadband networks

get latency under control. The networks currently operating around the world tend to have QoS and latency issues. Access networks and Wi-Fi networks contribute a surprisingly high amount of latency considering they are generally within the control of service providers (compared to the rest of the internet). Some subscribers often take steps to add their own network equipment that includes features like fq_codel/CAKE to improve their home network and access network performance.

Rather than have subscribers take drastic measures like this, CommScope is of the view that these features can be enabled in the existing access and Wi-Fi networks (see Figure 7). CommScope has taken a system approach to address the latency problem—linking the standards-based features with additional support in the DOCSIS and Wi-Fi networks. The solution is based on a mixture of cloud and gateway applications that offer service control, monitoring, and controlled options for service provider and subscriber service selection for low-latency applications. Performance monitoring is essential to validate that these services receive the expected latency—especially if subscribers are paying a premium for low latency. The CommScope solution includes support for monitoring Wi-Fi, DOCSIS, and internet latency on a per-service and per-subscriber basis—giving visibility into overall performance with different reports available to both the service provider and subscriber.

The solution provides mechanisms to classify applications based on service provider or subscriber settings, utilising this to map applications to the upstream and downstream low-latency service, as well as across Wi-Fi networks. QoS marking based on DSCP is used for applications not already marked, or as a way of modifying existing marks, if required. Support for Wi-Fi WMM is automatic for downlink traffic, while additional extensions involving Wi-Fi Alliance QoS management are being factored in to enable control over uplink and downlink traffic.

Conclusion

By addressing latency and jitter, broadband service providers can improve consumer satisfaction when using latency-sensitive services. They can derive new revenues from low-latency service tiers targeting online gamers and other delay sensitive services.

For more information, see www.commscope.com