Latency

The Final Frontier to a Great Broadband User Experience

by Bjørn Ivar Teigen, Head of Research at Domos and Broadband Forum member

As the world daringly flirts with real-time everything, latency can no longer be a mere afterthought; it's emerging as the next frontier for an innovative broadband industry. There is a misconception that latency – the amount of time it takes for a data packet to travel from one point to another – can be solved with even more bandwidth, and there is an onus on the broadband industry to heighten awareness of latency for consumers.

Bjørn Ivar Teigen, Head of Research at Domos and Broadband Forum member

Latency is one of the network performance metrics when determining customer Quality of Experience (QoE). Despite the common misconception that it only relates to gaming, the end-user has become progressively more aware of its significance as they navigate their digital lives. Several well-known capacity tests (so-called speed tests), such as speedtest.net by Ookla, and Cloudflare's network test, are increasingly emphasising the latency in the measurement results. It is important to grasp that latency is not a set number — it changes and fluctuates based on how the network is being used, the applications in use and the kind of connection being utilised. When one person downloads something, it can affect another person on the same network having a voice call. Similarly, when many active users are connected to the same cell tower, latency tends to grow larger and can degrade the user experience.

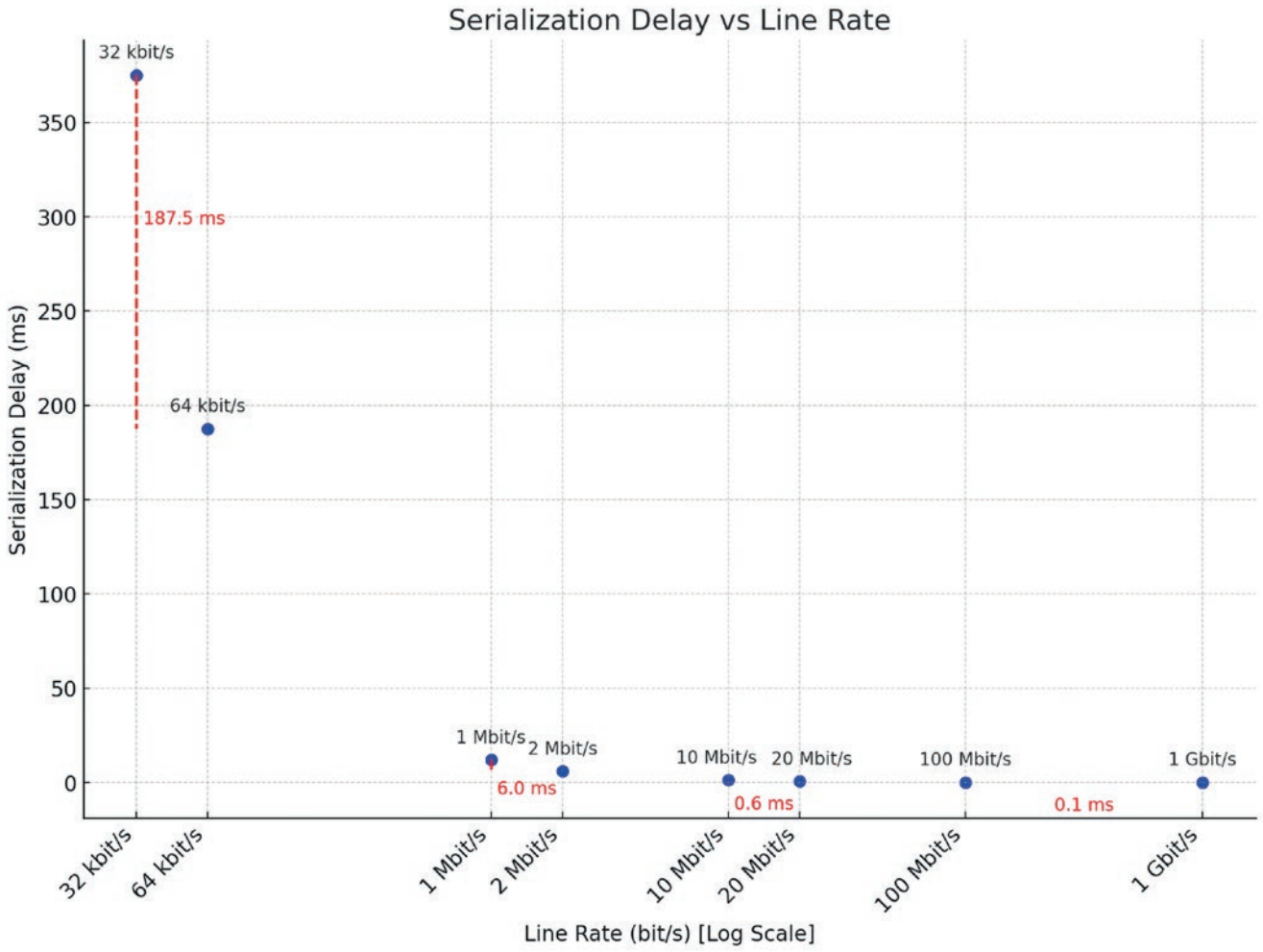
So why is latency so important for broadband user experiences? It boils down to reliability. People do not necessarily remember to thank their Internet Service Provider when the network is working flawlessly. What stays with them is the call that kept breaking up, or the game they lost because of lag.

Don't disregard distance

The simplest cause of latency is distance. The speed of light limits how quickly you can communicate and, on the Internet, that matters. The time it takes to go around the earth at the speed of light is roughly 133 milliseconds, which is about the time it takes to blink. While that may not seem very long, it is significant because most interactions on the Internet require more than one round-trip to complete. When we have lots of round trips, the delay due to distance adds up.

Fool's errand: let's add more bandwidth!

The amount of bandwidth in broadband networks has been steadily increasing for decades, as has the amount of Internet traffic. Of course, this higher bandwidth results in each piece of information being sent more quickly. So, more bandwidth should mean that data can be sent in less time? Well, yes and no.



When sending a large file, the time to transmit the data roughly correlates with the bandwidth of the broadband connection. But when it comes to latency, things are not so simple. When our connections were 32kbit/s, it took 375 milliseconds to send a 1500 byte packet. So doubling the bandwidth to 64kbit/s would shave 375 milliseconds off the round-trip delay (assuming a symmetrical connection). At today's connection rates, this is no longer true. Upgrading your bandwidth from 100Mbit/s to 1Gbit/s only shaves off around 0.1 milliseconds off the round-trip of a 1500-byte packet. Therefore, in a modern network more bandwidth does not necessarily impact best-case end-to-end latency.

MIT published a report last year outlining many examples of real-world measurements that highlight how more bandwidth can have an adverse effect on latency in some instances. The report showcased a few examples of more bandwidth increasing latency. That can happen because even though the time needed to send each packet over the physical wire gets a little bit shorter with higher rates, other parts of a high-bandwidth network sometimes make other sources of delay worse. We will dig deeper into two major sources of such delays: Bufferbloat and the latency spikes caused by capacity variation.

Understanding latency – what is Bufferbloat?

Many types of Internet traffic will use all the bandwidth at their disposal regardless of how much is available at the time. This type of traffic is called capacity-seeking. Capacity-seeking traffic can create delays in two ways: they temporarily overload the network when probing for the maximum capacity, and once the maximum capacity is found they may create a "standing queue" somewhere in the network. This standing queue is what's known as Bufferbloat. As you would expect, a standing queue means that the traffic being transmitted gets backed up. Bufferbloat can cause many seconds of delay (there are even reports of minutes!).

> **Regardless of whether capacity increases or reduces, there is a period of time, albeit a short one, when the network becomes overloaded, a queue builds up, and the user experiences a latency spike.**

Bufferbloat is much more common than you might think, and it is not a stretch to say that the vast majority of broadband consumers worldwide are affected by Bufferbloat. This performance issue was highlighted at the 2021 17th International Conference on Network and Service Management (CNSM) and was described in an article by Hackaday as a 'dreaded disease' plaguing service providers.

Increasing the peak network bandwidth can cause more Bufferbloat. This is one reason why getting a higher-capacity link can, counterintuitively, result in worse user experiences.

Capacity variation – a formidable culprit

Similarly, capacity-seeking traffic is the culprit when we consider another source of high and unstable latency: capacity variation. Capacity-seeking traffic will, of course, attempt to use all of the available bandwidth. But what happens when the amount of available bandwidth changes over time?

Whenever the available capacity changes, capacity-seeking traffic inevitably takes some time to adjust. If the capacity increases then the traffic will probe the boundaries to gauge what the new limit is. Capacity-seeking traffic does this by (briefly) sending more traffic than the network can actually support, and this causes queuing for a short period. On the other hand, if the capacity-seeking traffic is sending at the maximum possible rate and the network suddenly loses some capacity, the network will be overloaded until the traffic sender adapts to the new capacity. This period of overload caused by sudden drops in bandwidth can cause latency spikes in the hundreds of milliseconds and even seconds.

Regardless of whether capacity increases or reduces, there is a period of time, albeit a short one, when the network becomes overloaded, a queue builds up and the user experiences a latency spike.

Latency spikes – tell me more, tell me more

Latency spikes are prevalent in networks whenever capacity changes, but how common is this? Wireless links, such as Wi-Fi or 4G/5G have become the most common way of connecting to the Internet. Wireless links are inherently highly variable. Something as simple as rotating your phone, someone walking between the Wi-Fi router and your computer, or even just opening or closing a door can impact the link conditions enough to cause a significant change in capacity.

More bandwidth will not fix the latency spike problem because the spikes are caused by changes in rate. These spikes are typically measured in hundreds of milliseconds, which is long enough to cause issues with real-time applications and services. This includes video conferencing, gaming, and Artificial Reality / Virtual Reality / Spatial Computing.
The combination of both Bufferbloat and capacity variation can cause headaches for network operators and their subscribers. The traditional approach of adding more bandwidth is likely to make latency issues worse. A higher maximum bandwidth increases the potential for capacity variation, meaning bigger latency spikes and more end-user frustration.

Calling for new approaches, beyond conventional measurements

Building out endless bandwidth will never reliably deliver high-quality network connectivity for real-time interactions, so we must consider other approaches. Latency is a crucial and underappreciated part of creating a good user experience on the Internet.

The importance of latency is underappreciated by end-users and even by many in the broadband industry. Some passionate individuals, however, continue to bang the latency drum and this has resulted in a movement from difficult-to-understand network latency statistics to easy-to-understand measurements like Real-time performance monitoring and Quality of Outcome.

Broadband Forum's Broadband Quality Experience Delivered (Broadband QED) project, published as part of the TR-452 specifications, probes the network to diagnose performance issues. This includes congestion, latency and delay within the subscriber premises and at any location in the network.

The project uses Quality Attenuation (written ΔQ) methods and the initiative looks beyond conventional measurements to improve the overall broadband experience. Using these principles, performance problems can be characterised in a way that makes it possible to manage network latency, consistency, predictability and reliability.

This mathematical framework defines metrics to capture variability and analyses network quality to help network operators navigate complexity, optimise networks and deliver greater end-user QoE. Quality Attenuation measurements can break down latency into distinct components, identifying the source of performance degradation and evaluating the nature and cause of network performance issues.

For example, the packet delay is broken down into a constant component (due to distance travelled), a variable component (caused by queuing or buffering), and a serialisation delay (tied to link capacity).

The latency mission - should you choose to accept it

To build on this and deliver the enhanced broadband experience end-users crave, the broadband industry has to look beyond conventional measurements of throughput and latency to improve the management of network performance. The measurements and metrics leveraging the QED standard will deliver great network insights and diagnostics for operators.

With access to a greater understanding of individual latency components, network operators can unlock an enhanced user experience in the quest to seamlessly support demanding applications and services.

For more information, visit https://domos.no/