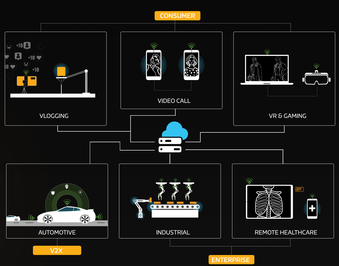
Improving FWA Performance with Advanced Congestion Avoidance Techniques

Nature of Latency in Telecom Networks and Ways for Improvement

In today's hyper-connected world, the importance of low latency in telecom networks cannot be overstated. As businesses and individuals increasingly rely on real-time data transmission for everything from video conferencing, to online gaming, to vehicle connectivity, to enterprise and industrial applications, the demand for seamless communications from anywhere has never been higher. In an era where milliseconds can make a significant difference, optimizing network latency is essential for maintaining an operator's competitive advantage andensuring the reliability and efficiency of digital services.



Total latency, the time it takes for a data packet to travel from the source to the destination, is influenced by several factors. Four primary delays contributing to the total latency are:

- Application processing delay
- Propagation delay
- Interface delay
- Queuing delay

Application Processing Delay

This refers to the delay introduced by the software applications themselves. This latency occurs when an application processes data before sending it over the network or after receiving it. Factors contributing to application latency include the efficiency of the code, the processing power of the hardware, and the complexity of the tasks being performed.

Propagation Delay

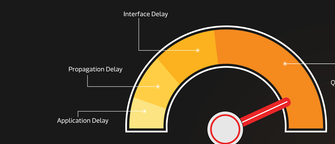
The time it takes for a signal to travel from the sender to the receiver. This latency is primarily influenced by the physical distance between the two points and the speed at which the signal travels through the medium. Even at the speed of light, significant distances can introduce noticeable delays. Minimizing signal travel latency often involves optimizing the physical layout of the network, placing data centers closer to end-users.

Interface Delays

These are introduced by network interfaces, and depend on Physical and MAC layers implementations. New communication technology standards, such as PON, and 5G, deal with optimizations in these areas.

Queuing Delays

These represent the largest source of latency and occur when data packets are temporarily stored in buffers within network devices while waiting to be processed or transmitted. This type of latency is influenced by the size of the buffers, the traffic load on the network, and the efficiency of the queuing algorithms. During periods of high network traffic, buffers can become congested, leading to increased latency as packets wait in line to be processed.



AQM and L4S

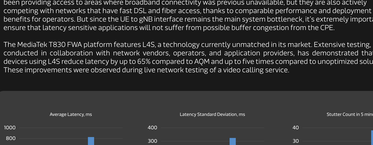
AQM (Active Queue Management) is a technique designed to manage congestion, and has significant advantages over traditional queue management schemes. Unlike those typically relying on simple drop-tail mechanisms, where packets are dropped only when the queue is full, AQM proactively monitors the queue length or queuing delay to make more informed decisions about when to drop packets. This proactive approach helps to prevent congestion before it becomes problematic, thereby reducing latency and packet loss. AQM algorithms, such as Controlled Delay (CoDel), aim to maintain a balance between high throughput and low delay, ensuring a smoother and more efficient flow of data across the network. By affecting the traffic originator to dynamically adjust the sending rate, AQM helps to mitigate the negative effects of congestion, leading to a more stable and responsive network environment.

However, AQM is mainly based on implicitly informing the traffic originator about buffer congestion happening along the way by dropping certain packets. This causes some extra delay due to the retransmission of those dropped packets, which can be sensible in real-time applications such as AR/VR. Additionally, the performance of AQM is highly dependent on the latency and bandwidth targets set, as these parameters influence the algorithm's ability to maintain balance between them. An aggressive latency target leads to better delay management but poorer resource utilization, as more bandwidth is sacrificed. A more relaxed target saves bandwidth utilization but ends up with insufficient queue improvements, still allowing buffer bloating.

To address these AQM drawbacks, L4S (Low Latency, Low Loss, Scalable Throughput) congestion avoidance technique was introduced by IETF as a set of RFCs , including RFC9330, RFC9331 and RFC9332, that are defining the "Prague L4S Requirements". L4S is an advanced networking framework that builds upon the principles of AQM to achieve better performance. It re-uses the AQM congestion detection but utilizes this information for packets marking with L4S Explicit Congestion Notification (ECN), a part of TOS byte in the IPv4 IP header, or the Traffic Class byte in that IPv6, rather than applying actions on packets in queue.

This helps to eliminate unnecessary packet drops and replaces them with explicit informing of the traffic originator, which can proactively modify the packet sending rate to manage the queue on the reporting node. Such an updated approach provides better flexibility and faster reaction to the early signs of possible buffer congestion, eventually preventing it by sacrificing as little bandwidth as possible.

Obviously, not all applications would be able to support L4S from the start, which triggers demand for additional mechanisms to provide smooth co-existence of L4S and traditional traffic flows. The Dual Queue (DualQ) specification was developed to deal with this challenge. DualQ enables separate queues for L4S and Classic traffic, treating them via AQM accordingly – marking packets with ECN in the L4S queue and drop them in the Classic queue. L4S queue has initial scheduling priority, but subsequent faster growth of Classic queue is due to a less effective technique. Classic queue growth impacts the marking probability of L4S AQM thanks to the DualQ coupling mechanism. This results in eventual fair distribution, not giving priority to any of the queues.



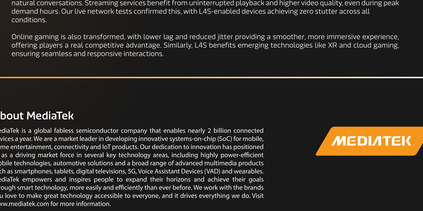
L4S Performance in Live Network Test

An end-to-end implementation is preferred for L4S, with all nodes in the way being able to detect congestion and apply the packets marking with ECN. However, the most important nodes are those placed in possible bottlenecks. For a 3GPP wireless network, one such bottleneck is the air interface. Two network elements placed at the edges of this interface are the UE (User Equipment) and the base station: for 5G this is a gNodeB.

The UE controls uplink congestion, while gNodeB can do the same for the downlink. Due to reasons such as the subscribers' traffic profile and more transmit power at the base station, downlink performance has always prevailed in wireless networks, and 5G is no exception. This leaves UE to gNodeB interface as the most probable bottleneck in a 5G network, where buffer congestion may appear, impacting delays on latency-sensitive applications.

Significant growth of Fixed Wireless Access (FWA) have impacted on the use of 5G deployments. 5G CPEs have not only been providing access to areas where broadband connectivity was previous unavailable, but they are also actively competing with networks that have fast DSL and fiber access, thanks to comparable performance and deployment benefits for operators. But since the UE to gNB interface remains the main system bottleneck, it's extremely important to ensure that latency sensitive applications will not suffer from possible buffer congestion from the CPE.

The MediaTek T830 FWA platform features L4S, a technology currently unmatched in its market. Extensive testing, conducted in collaboration with network vendors, operators, and application providers, has demonstrated that CPE devices using L4S reduce latency by up to 65% compared to AQM and up to five times compared to unoptimized solutions. These improvements were observed during live network testing of a video calling service.



The tests covered three network conditions: no congestion, a 10 Mbps margin to congestion, and a fully congested network. The result was L4S delivered consistent and stable performance regardless of bandwidth, with significantly lower latency variance than AQM and unoptimized traffic. Service data rates and video call quality remained largely unaffected. Lab results showed even greater performance gains, indicating further optimization potential.

By significantly reducing latency, L4S enhances real-time communication (video conferencing) and entertainment experiences (online gaming, XR). For video calls, this means clearer audio, smoother video, and minimal lag, ensuring natural conversations. Streaming services benefit from uninterrupted playback and higher video quality, even during peak demand hours. Our live network tests confirmed this, with L4S-enabled devices achieving zero stutter across all conditions.

Online gaming is also transformed, with lower lag and reduced jitter providing a smoother, more immersive experience, offering players a real competitive advantage. Similarly, L4S benefits emerging technologies like XR and cloud gaming, ensuring seamless and responsive interactions.

About MediaTek

MediaTek is a global fabless semiconductor company that enables nearly 2 billion connected devices a year. We are a market leader in developing innovative systems-on-chip (SoC) for mobile, home entertainment, connectivity and IoT products. Our dedication to innovation has positioned us as a driving market force in several key technology areas, including highly power-efficient mobile technologies, automotive solutions and a broad range of advanced multimedia products such as smartphones, tablets, digital televisions, 5G, Voice Assistant Devices (VAD) and wearables. MediaTek empowers and inspires people to expand their horizons and achieve their goals through smart technology, more easily and efficiently than ever before. We work with the brands you love to make great technology accessible to everyone, and it drives everything we do. Visit www.mediatek.com for more information.

mediatek.com