Closing the Gap

Strategies For Ultra Low Latency in Wi-Fi

A technical paper prepared for presentation at SCTE TechExpo24

Pratyusha Malladi
Principal Engineer II
Charter Communications
Pratyusha.Malladi@charter.com

Dileep Kumar Soma
Principal Engineer I
Charter Communications
DileepKumar.Soma@charter.com

Table of Contents

List of Figures

1. Introduction

Traditionally, network performance has been evaluated by metrics like speed and throughput. With broadband networks delivering multi-gigabit speeds, latency is growing as a key metric of network performance.

Latency typically refers to round-tripdelay. This round-trip encompasses the time it takes for the data to traverse the network. Today, a large majority of internet traffic's first hop is via either a mobile wireless network or a provider's managed Wi-Fi network. This first hop is often the largest contributor to latency. Wireless technologies, such as cellular and Wi-Fi, have implemented various mechanisms to minimize the latency in the first hop. This paper investigates the factors that contribute to latency in Wi-Fi networks and examines various technologies that can be used to further reduce latency. The aim is to enable cable operators to extend Low Latency DOCSIS® (LLD) networks to their customers who are connected via Wi-Fi. In this paper, we will use the terms Wi-Fi 7 representing the latest IEEE 802.11be standard interchangeably.

2. Understanding Latency

Latency typically refers to round-trip delay. This round-trip encompasses the time it takes for the data to traverse the network from a source to a destination and for the response to get back to the source. It is a metric in network performance because it impacts the responsiveness and efficiency of data transfer. In today's networks, attention is given to latency for several reasons. Firstly, with the increasing reliance on real-time applications such as video conferencing, online gaming, and cloud computing, low latency provides smooth and seamless user experiences. Secondly, the rise of Internet of Things (IoT) clients and autonomous systems requires near-instantaneous communication for tasks like remote monitoring, control, and data analysis.

Latency in wireless networks refers to the delay experienced when transmitting data from a source client to the nearest access point or router. Several factors can contribute to higher latency, including signal propagation issues, contention for wireless medium access, channel congestion, queueing delays, interference from other clients or environmental factors, and handover delays when switching between access points. These factors can cause delays, signal loss, or packet degradation, resulting in increased latency in the wireless networks.

3. Collision Avoidance In Wi-Fi

This section will delve into the technology behind Wi-Fi and the progressive enhancements made in the 802.11 standards to minimize latency. Wi-Fi networks were initially designed as "best effort" technologies, prioritizing the efficient delivery of data packets without guaranteeing specific service levels or latency. The use of unlicensed frequency bands in Wi-Fi networks, which can be shared by multiple clients, can lead to varying latency due to factors such as network congestion, signal strength, and the number of connected clients. However, advancements in Wi-Fi technology, including the introduction of newer standards like IEEE 802.11ax (Wi-Fi 6), have aimed to improve performance and reliability and reduce latency. Techniques like Orthogonal Frequency Division Multiple Access (OFDMA) and Multi-User MIMO (MU-MIMO) have been implemented to mitigate interference and enhance network efficiency.

Collision avoidance is an aspect of Wi-Fi networks. It plays a role in ensuring efficient and reliable data transmission. Without collision avoidance mechanisms, multiple clients within a Wi-Fi network may attempt to transmit data simultaneously, potentially leading to collisions and disruptions in communication. Collision avoidance techniques, such as CSMA/CA (Carrier Sense Multiple Access with

Collision Avoidance), RTS/CTS (Request to Send/Clear To Send) and MBCA (Mesh Beacon Collision Avoidance) play a role in enhancing Wi-Fi networks.

3.1. The CSMA/CA Mechanism

Collision avoidance in Wi-Fi networks can have an impact on latency, which refers to the delay in data transmission. While collision avoidance mechanisms, like CSMA/CA, help in preventing collisions and ensuring reliable data transmission, they can introduce latency to the network.

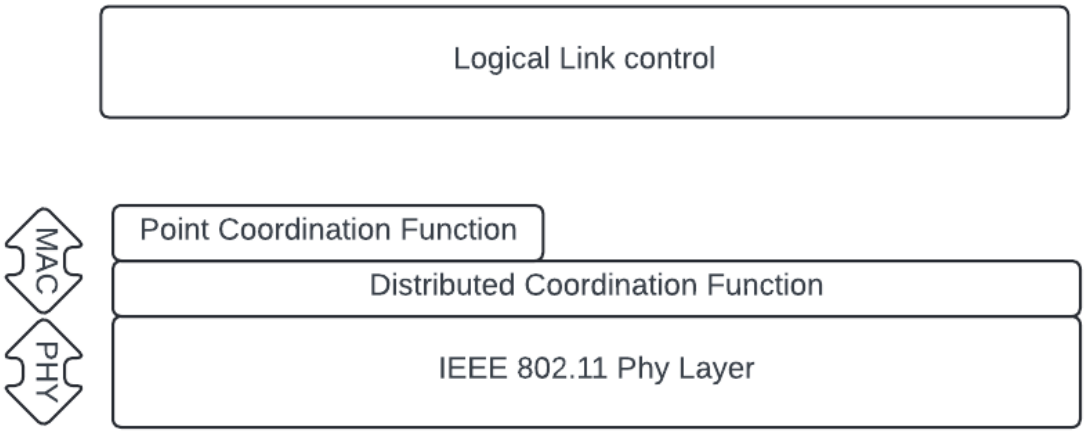

Figure 1: 802.11 PHY and MAC layers

The Distributed Coordination Function (DCF) was introduced in Wi-Fi as part of the original IEEE 802.11 standard in 1997. It is the mechanism by which CSMA/CA is applied to Wi-Fi networks. DCF is the basic access method used in Wi-Fi networks to manage the transmission of data between multiple clients.

The DCF mechanism begins by the station performing:

- Physical carrier sense: The physical carrier sense mechanism in wireless networks involves listening to the channel to detect RF transmissions. It uses two thresholds: Energy Detect (ED) for non-802.11 transmissions and Signal Detect (SD) for 802.11 transmissions. The ED threshold detects any energy in the channel, while the SD threshold specifically looks for 802.11 signals. If the energy or signal strength exceeds these thresholds, the channel is considered busy, and transmission is deferred to avoid collisions. Typically, SD is set to 4dB more than the noise floor and the ED is about 20dB higher than the SD. The physical carrier sense mechanism is an essential part of the CCA (Clear Channel Assessment) process.
- Virtual carrier sense: Virtual carrier sense is another component of the CCA process in wireless networks. It operates by examining the Network Allocation Vector (NAV) field in the control frames. The NAV field contains the duration for which the channel is expected to be busy due to ongoing transmissions by other clients. By checking the NAV field, a client can determine if the channel is currently in use and defer its own transmission to avoid collisions.

The combination of physical carrier sense, which involves listening to the channel, and virtual carrier sense, which involves examining the NAV field, allows clients in a wireless network to check if the channel is available before sending data. If the channel is found to be busy, the client will postpone transmission and the countdown timer will be inactive. The backoff timer is influenced by the contention window values, CWmin and CWmax. Initially, the contention window starts at CWmin, and if collisions occur, it doubles until it reaches CWmax. The device selects a random backoff value from the range of 0 to CW - 1, where CW is the current contention window. After the backoff timer expires, the client checks the channel again using Clear Channel Assessment (CCA) to confirm if it is still busy. This process repeats until the channel is confirmed to be idle. After detecting an idle channel, the station employs a period called DIFS (Distributed Inter-Frame Space) or SIFS (Short Interframe Space), followed by a backoff timer, before initiating transmission. DIFS is used for generic 802.11 frames and SIFS for high-priority frames like ACKs (Acknowledgements). SIFS is typically a shorter interval than the DIFS. Another interval known as AIFS (Arbitration Inter-Frame Space) is used in Wi-Fi. AIFS is typically longer than the DIFS and SIFS intervals. It is used to provide priority access to different traffic classes or categories in a wireless network. Each traffic class is assigned a specific AIFS value, which determines the wait time before transmission. AIFS allows stations with higher priority traffic to have shorter access delays compared to stations with lower priority traffic. This helps in achieving quality of service (QoS) requirements for diverse types of traffic, such as voice, video, or data. The inter-frame spacings are measured in microseconds(μs).

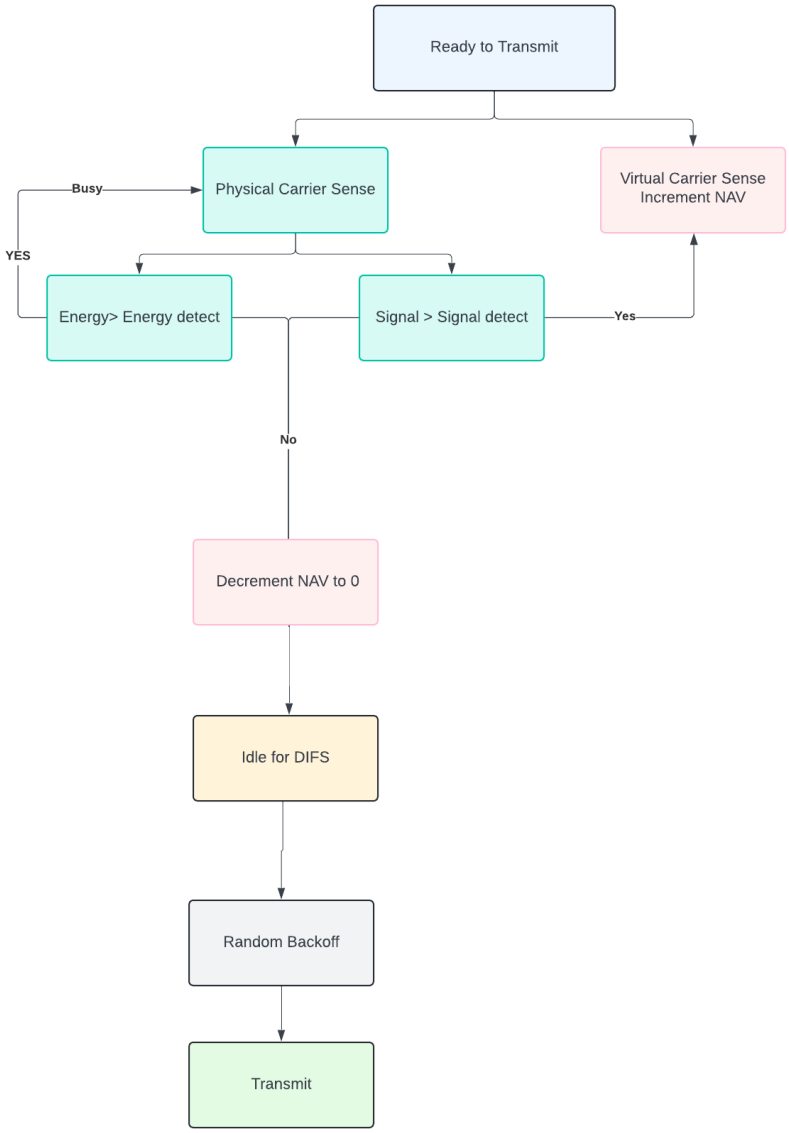

Figure 2: DCF Mechanism

3.2. RTS/CTS

The hidden node problem occurs when two stations that are far apart transmit to the same access point. In this situation, they struggle to perform carrier sense effectively and cannot accurately determine the channel's status (idle or busy). As a result, collisions can occur when both stations transmit simultaneously to the access point. To address this problem, the RTS/CTS mechanism can be enabled. When a station wants to transmit, it sends an RTS frame to the access point, requesting permission. The access point replies with a CTS frame, granting permission and specifying a reserved transmission duration.

By using RTS/CTS, stations coordinate their transmissions to avoid collisions. Other stations within range receive the frames and defer their transmissions. This ensures only one station transmits at a time, reducing collisions and improving network performance. The RTS/CTS mechanism does add overhead and latency but improves data transmission reliability and efficiency in scenarios with hidden node problems.

3.3. MBCA

MBCA stands for Mesh Beacon Collision Avoidance, which is a collision avoidance technique used in wireless mesh networks. Wireless mesh networks consist of multiple devices, called nodes, that communicate with each other to extend network coverage and improve connectivity. In a wireless mesh network, each node periodically transmits a beacon frame. The beacon frame contains information about the node's identity, network status, and the links it has with other nodes in the network. This information helps other nodes in the network to discover and establish connections with each other. MBCA is a mechanism that attempts to ensure efficient beacon transmission in wireless mesh networks by avoiding collisions. Collisions occur when multiple nodes attempt to transmit their beacons simultaneously, which can lead to data corruption and reduced network performance. MBCA aims to prevent such collisions and maintain smooth communication within the network.

To achieve collision avoidance, MBCA utilizes a distributed algorithm that coordinates the beacon transmissions among the nodes. The algorithm assigns specific time slots to each node for transmitting its beacon. Each node follows the assigned schedule and only transmits its beacon during its designated time slot. By carefully coordinating the beacon transmissions, MBCA minimizes the chances of collisions and attempts to ensure that the beacon information is reliably shared among the nodes in the network.

4. Latency: Cellular and Wi-Fi

Cellular networks operate in exclusive licensed spectrum, which provides exclusionary rights, meaning no other networks may operate in an exclusive licensed band and thus there is no risk of contention from other networks which results in lower latency compared to Wi-Fi networks. Exclusive licensed spectrum provides a controlled environment where cellular operators have sole control through which to prioritize traffic, allocate resources, and manage interference. Wi-Fi networks, which operate using shared unlicensed spectrum, use a “polite” operating protocol that requires clients to contend for access to the medium by “listening before talking” and then backing off at exponentially increasing time increments when a channel is already occupied. That may lead to increased latency compared to an exclusive mobile service. Collisions become less likely when more contiguous spectrum is available. For that reason, advances in Wi-Fi technology benefit from new unlicensed spectrum (i.e. 6 GHz) and from the use of larger bandwidth to reduce the amount of time needed for individual Wi-Fi transmissions. The 7 GHz spectrum band is a critical opportunity for continued Wi-Fi innovation and growth, as it could enable additional next-generation wide-bandwidth channels to reduce latency, support the growing number of devices and deliver higher speeds and capacity for data-intensive applications. Federal regulators are currently evaluating the 7 GHz band to determine the feasibility of allowing unlicensed sharing, licensed sharing or exclusive commercial mobile use.

An attribute of operating in exclusive licensed spectrum is the elimination of collision avoidance mechanisms like CSMA/CA used in Wi-Fi networks. By eliminating the need for collision avoidance mechanisms, and managing resource contention via a centralized scheduler, the cellular network can provide more predictable latency.

However, it is important to note that the federal government has acknowledged that there is no additional greenfield spectrum available for commercial or federal use, meaning there is no spectrum available for

exclusive licensing without the significant expense and delay of removing incumbent users – which also assumes removing them is even possible given competing federal priorities. With spectrum sharing increasingly necessary for access to additional spectrum bandwidth, collision avoidance mechanisms and contention management are likely to be increasingly relevant.

5. Ultra-Reliable Low Latency Communication

3GPP Release 15 5G-NR introduces a new feature called Ultra-Reliable Low Latency Communication (URLLC), which provides reliable and low latency communication in the licensed spectrum. URLLC is designed to support applications that require reliable and near-instantaneous transmission of data, such as mission-critical applications, industrial automation, remote surgery, and autonomous vehicles. These applications often have stringent requirements for reliability, latency, and availability.

To meet the latency requirements of URLLC, significant enhancements have been implemented in the physical (PHY) layer and medium access control (MAC) layer. These include the following techniques:

- Minimization of waiting time with frequent transmission opportunities: The downlink (DL) control channel is used to carry scheduling information both for the uplink (UL) and DL data transmission. This channel is therefore frequently monitored by the UE to reduce the wait time to receive control information. When data is received by the UE, the UE needs to send a scheduling request (SR) to the gNB for UL resource allocation. To further reduce the wait time for resource allocation, the SR must be sent in more frequent intervals.
- Reduce transmission duration: Another factor that adds to the latency experienced over the air is the duration of the transmission. To decrease this duration, there are two approaches that can be utilized. The first approach involves increasing the subcarrier spacing, which in turn reduces the symbol duration. The second approach involves using mini slots, which enables an increase in the transmission frequency. By implementing these methods, the overall transmission duration can be reduced, leading to a decrease in over the air latency.
- Hybrid automatic repeat request (HARQ) enhancements: HARQ improves the reliability of data transmission by enabling the receiver to request retransmissions of erroneous or lost packets. It allows for error detection and correction, minimizing the impact of channel impairments and improving data integrity. Faster feedback to the transmitter provided by reducing HARQ processing time results in reduced latency.
- Grant-free or configured grant for uplink (UL) transmission: Grant-based handshakes require additional signaling between the UE and the gNB, which can introduce latency on the air interface. On the other hand, grant-free transmissions allow for preconfigured UL resources for the UE, which eliminates the need for explicit grants and helps reduce latency.

5.1. 5G NR-U

When considering latency in wireless networks, it is more appropriate to compare Wi-Fi with 5G NR-U as both technologies operate in unlicensed spectrum. Therefore, contention from other systems and the additional cost of the Listen Before Talk (LBT) are additional challenges for URLLC in unlicensed spectrum.

LBT is a collision avoidance mechanism employed in 5G NR-U, which functions similarly to CSMA/CA in Wi-Fi. In frequency bands where 5G and Wi-Fi coexist without licenses, 5G utilizes a channel access scheme known as LBT. This scheme ensures that 5G clients actively listen for ongoing Wi-Fi transmissions before initiating their own transmissions. By implementing this process, 5G clients can effectively prevent interference with Wi-Fi signals and facilitate a harmonious coexistence between the two technologies.

To ensure effective operation of 5G NR-U in unlicensed frequency bands, four distinct categories of LBT protocols have been established:

- CAT1-LBT (Type 2C): A gNB can access the channel immediately without performing LBT.
- CAT2-LBT (Type 2A and 2B): When operating in NR-U mode, a client is required to monitor the channel for a specified duration. If the channel remains unused during this time, the client is then permitted to utilize the channel for communication.
- CAT3-LBT: An NR-U client must back off for a random period before accessing the channel. This random period is sampled from a fixed-size contention window.
- CAT4-LBT (Type 1): An NR-U client must back off according to the CSMA/CA procedure with exponential backoff. This mechanism is utilized by LTE- Licensed Assisted Access (LTE-LAA) and is also considered as the baseline NR-U operation for shared spectrum access.

Since CAT4-LBT uses the same mechanism as CSMA/CA on Wi-Fi the latency between both technologies is comparable.

6. Latency Improvements in Wi-Fi Networks

The challenge of latency on Wi-Fi is linked to situations where multiple clients are trying to access the air interface. As discussed earlier, each client must compete for access. This section examines the technological improvements made in different generations of Wi-Fi to ensure more efficient use of the medium.

To gain a deeper understanding of the strategies used to improve Wi-Fi latency, consider a scenario where several Wi-Fi clients are connected to a Wi-Fi router operating on the same channel. For each client to send data at the same time, they must first detect if the channel is used by other clients or the AP. If it is, they must patiently wait for an idle period before transmitting their own data. This waiting time inevitably contributes to the overall latency experienced in the network.

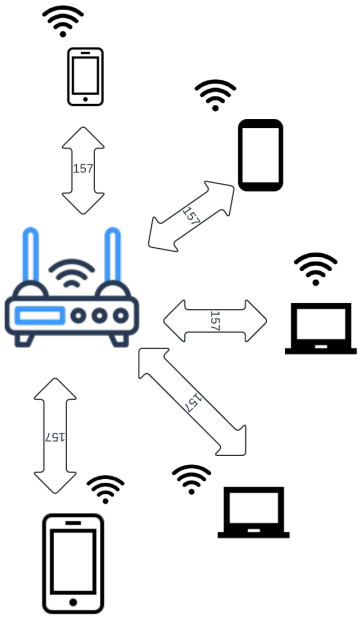

Figure 3: Scenario

6.1. OFDMA

In the scenario above, latency can be reduced by implementing orthogonal frequency division multiple access (OFDMA). This allows multiple clients to transmit and receive data simultaneously by dividing the available channel into smaller sub-channels, each of which can be assigned to a different client. This enables more efficient use of the channel and reduces the need for clients to wait for idle periods. As a result, the overall latency in the network is reduced, allowing for faster and more simultaneous data transmissions.

Orthogonal Frequency Division Multiple Access (OFDMA) was introduced as one of the key features in the 802.11ax standard, also known as Wi-Fi 6.

OFDMA operates as follows:

- Each 20 MHz channel is divided into 78.125 kHz wide subcarriers. Total number of subcarriers is given by:

$$N_{subcarrier} = 20\ MHz / 78.125\ kHz = 256$$

- Out of the 256 subcarriers, 14 are used for guard and pilot tones.
- The remaining 242 subcarriers are divided into resource units (RU) comprising of 26 subcarriers each.
- A minimum of one RU can be allocated to each client. This ensures that every client has access to at least one RU for communication.
- Therefore, the maximum number of users that can simultaneously transmit on a 20 MHz channel is given by:

$$N_{maxusers} = 256/26 \sim = 9$$

Therefore, it can be deduced that with OFDMA, nine users can transmit simultaneously using a single contention window, each utilizing a single resource unit, thereby reducing latency.

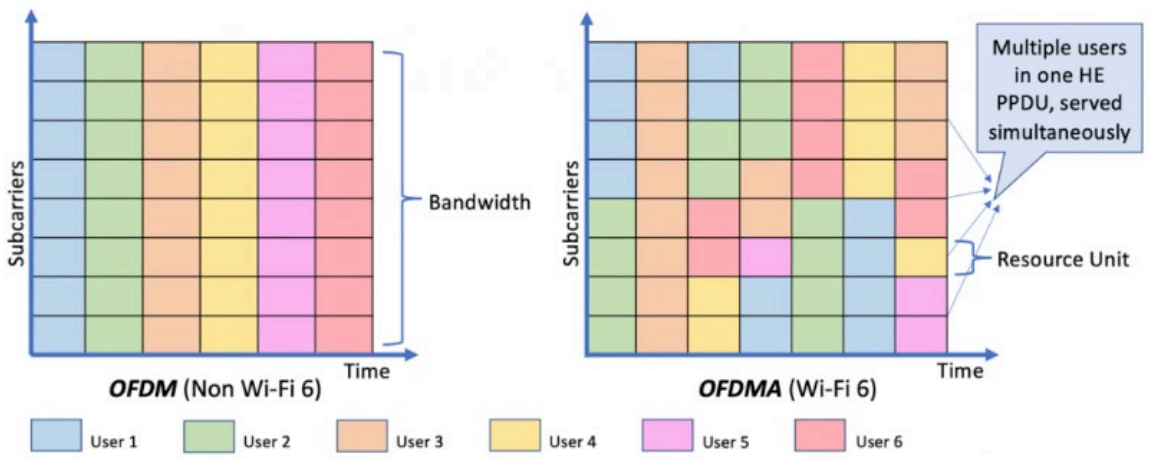

Figure 4: OFDMA Mechanism

(Courtesy: https://blogs.cisco.com/networking/WiFi-6-ofdma-resource-unit-ru-allocations-and-mappings)

6.2. QoS Prioritization

To enhance latency reduction in the described scenario, network administrators can allocate Quality of Service (QoS) priority to clients utilizing real-time applications like voice or video. QoS empowers administrators to categorize and prioritize traffic based on its significance and specific needs. By assigning higher priority to real-time applications over other forms of traffic, they can enforce the allocation of essential network resources and shield them from congestion or delays.

QoS prioritization in Wi-Fi networks is performed using WMM (Wi-Fi Multimedia). WMM is a Wi-Fi Alliance ® certification program that provides enhanced QoS features to prioritize different types of traffic and ensure better performance for specific applications. WMM defines four access categories, each with its own priority level:

- Voice (AC_VO): This category is used for real-time voice traffic, such as VoIP (Voice over IP) calls. It has the highest priority to ensure low latency and minimal packet loss.
- Video (AC_VI): This category is used for real-time video traffic, such as video streaming or video conferencing. It has the second highest priority.
- Best Effort (AC_BE): This category is used for typical data traffic, such as web browsing or file downloads. It has a medium priority and shares the remaining bandwidth after voice and video traffic.
- Background (AC_BK): This category is used for low-priority or background traffic, such as software updates or file backups. It has the lowest priority and only utilizes the remaining bandwidth after all other categories.

WMM uses a contention-based mechanism known as Enhanced Distributed Channel Access (EDCA) to prioritize different kinds of traffic across the medium. EDCA utilizes different Arbitration Interframe Spaces (AIFSs) and Contention Window (CWmin and CWmax) sizes for each access category. This assures that the access category with the highest priority, such as voice, has the lowest wait times before transmission, as shown below.

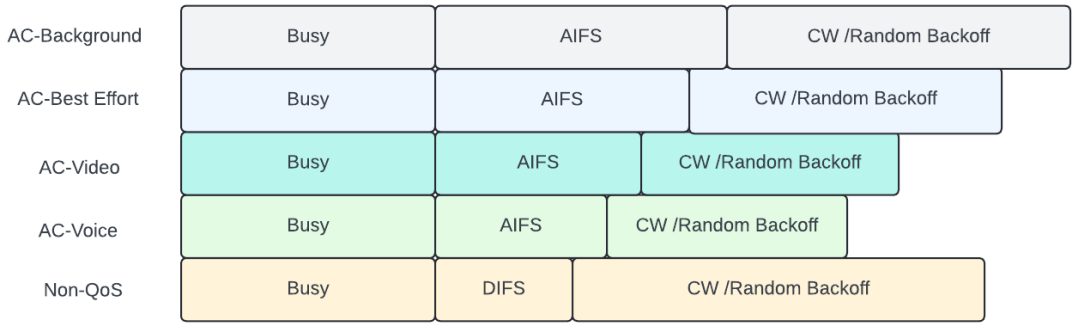

Figure 5: QoS management using WMM

The introduction of the QoS Management certification program by the Wi-Fi Alliance builds upon the existing WMM technology, further enhancing the quality of service provided by Wi-Fi networks. Wi-Fi QoS management introduced two new features:

Mirrored Stream Classification Service (MSCS): MSCS allows clients to negotiate downlink quality of service (QoS) based on QoS mirroring. Both the Wi-Fi QoS Management Access Point (AP) and Station (STA) need to support MSCS as defined in the specified standards. MSCS is activated at the Media Link Descriptor (MLD) level when Multi-Link Operation (MLO) is enabled. In simple terms MSCS works by mirroring the QoS settings from the sender to the receiver.

For example, consider a Wi-Fi router and a smartphone connected to it. The router supports MSCS, which means it can negotiate the quality of service with the smartphone. This negotiation is based on the settings of the smartphone. When the smartphone sends a request to the router, it includes information about the quality of service it wants. The router checks if it supports MSCS and if it can provide the requested quality of service. If everything matches, the AP mirrors the QoS of uplink flows from the smartphone to the downlink flows. However, there are some conditions required for MSCS to work properly. The AP and STA should support MSCS negotiation with the Classifier Type field set to 4 for IP and higher layer parameters and should classify both IPv4 and IPv6 packets. If the router doesn't have enough resources or if it doesn't support MSCS, it may reject the requests from the smartphone. In that case, the smartphone may not get the desired quality of service. The router can use specific codes to explain the reason for rejection. It’s important to note that the MSCS feature is designed to ensure fair and efficient use of the wireless network. If the router detects that certain priority levels are being used excessively and causing problems for other clients, it can automatically adjust the QoS settings or even disconnect the clients that are using too much bandwidth.

6.3. Multi Link Operation (MLO)

In the IEEE 802.11be standard, also known as Wi-Fi 7, Multi-Link Operation (MLO) is offered. Consider a scenario where we focus on the potential of using one of the MLO links exclusively for low latency traffic, while utilizing different links for handling all other types of data. This approach aims to prioritize the seamless and immediate transmission of time-sensitive applications, such as real-time financial transactions or critical command and control signals. At the same time, it allows for effective management of other network activities on separate links.

To implement this scenario successfully, it would be necessary to carefully configure and allocate resources in order to achieve optimal performance for both low latency and non-low latency data transfer. The IEEE 802.11be standard defines multiple modes of MLO, including (Enhanced) Multi Link Single

Radio (MLSR/eMLSR), Multi Link Multi Radio – Simultaneous Transmit and Receive (MLMR-STR), and Multi Link Multi Radio – Non-Simultaneous Transmit and Receive (MLMR-NSTR). Among these modes, Enhanced Multi-Link Single Radio (eMLSR) and Multi-Link Multi-Radio Simultaneous Transmit and Receive (MLMR-STR) have emerged as the most favored by industry implementations. These two variants offer advantages in terms of performance, efficiency, and practicality for a wide range of devices and use cases. Consequently, our discussion will focus on the impact of these two features in reducing latency in Wi-Fi networks

6.3.1.1. Enhanced Multi-Link Single Radio (eMLSR)

eMLSR offers latency improvements over Single Link Operation (SLO) by enabling rapid switching between multiple links using a single radio.

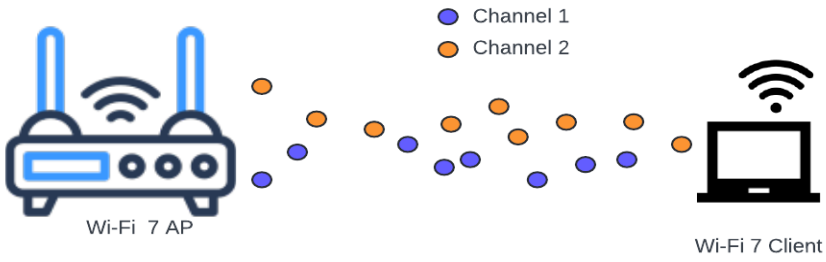

Figure 6: eMLSR

Key latency reduction mechanisms:

- Fast link switching: eMLSR allows devices to switch between links in microseconds, dramatically reducing the time spent waiting for a clear channel.
- Increased spectrum access: By monitoring multiple channels, eMLSR increases the probability of finding an available transmission opportunity, reducing media access delays.
- Dynamic interference avoidance: Devices can quickly switch away from congested or interfered channels, minimizing retransmissions and associated latency.
- Load balancing: Traffic can be distributed across different bands based on current conditions, preventing any single link from becoming a bottleneck.

Latency performance under different scenarios:

- Low congestion: Modest latency improvements over SLO, as channel access is generally available.
- Moderate congestion: Significant latency reductions as eMLSR leverages its ability to find clearer channels quickly.
- High congestion: Substantial latency benefits, with eMLSR maintaining lower and more consistent latency compared to SLO.

6.3.1.2. Multi-Link Multi-Radio Simultaneous Transmit and Receive (MLMR-STR)

MLMR-STR builds upon the benefits of eMLSR and provides even greater latency reductions through simultaneous multi-link operation.

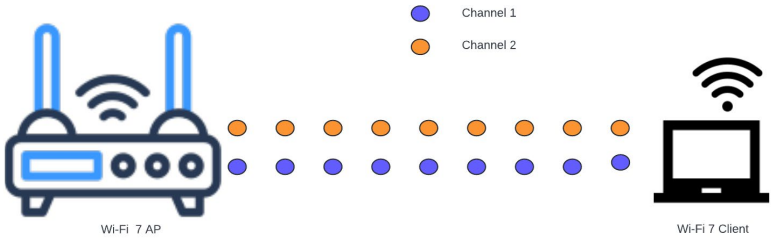

Figure 7: MLMR

Additional latency reduction mechanisms:

- Parallel transmissions: MLMR-STR can send and receive data simultaneously on multiple links, effectively reducing eliminating queueing media access delays for multi-link capable flows.
- Link aggregation: By combining multiple links, MLMR-STR can reduce the transmission time for large packets, lowering overall latency.
- Redundant transmissions: Critical or latency-sensitive packets can be sent over multiple links simultaneously, ensuring the fastest possible delivery.
- Optimized link selection: MLMR-STR can choose the best link(s) for each packet based on current conditions and QoS requirements, minimizing latency for all traffic types.

Latency performance comparison:

- Low congestion: MLMR-STR shows more noticeable latency improvements over eMLSR, particularly for large data transfers or multiple simultaneous flows.
- Moderate congestion: Significantly lower latency than eMLSR, as MLMR-STR can utilize multiple clear channels concurrently.
- High congestion: MLMR-STR maintains the lowest and most consistent latency, with the ability to leverage any available spectrum across multiple links simultaneously.

6.3.1.3. Comparison between eMLSR and MLMR-STR

In conclusion, while both eMLSR and MLMR-STR offer latency improvements over SLO, MLMR-STR provides slightly more latency reductions across various network conditions. The choice between these technologies will depend on factors such as device capabilities, power consumption requirements, and specific use cases. For devices where power consumption and complexity are less of a concern, MLMR-STR is the likely option for minimizing latency in Wi-Fi 7 networks. The devices with power consumption constraints and lower computational complexity will probably prefer eMLSR for latency minimization.

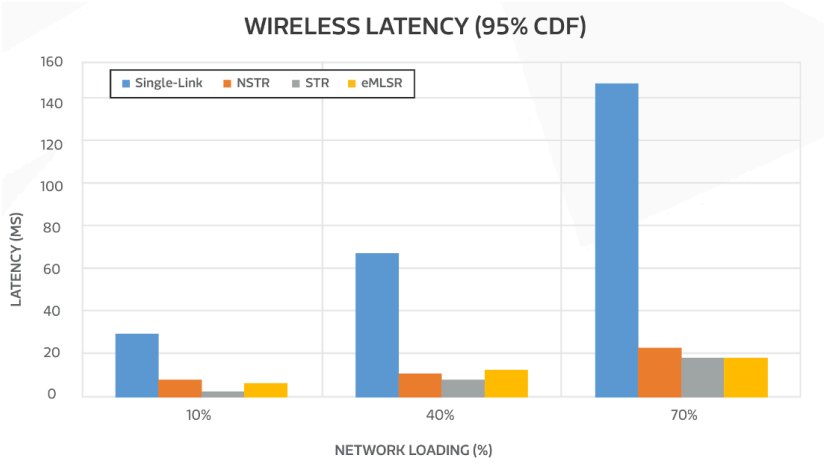

Figure 8: Latency comparison against load

(Courtsey: https://www.mediatek.com/technology/mlo-infographic)

Internet Service Providers (ISPs) can enhance the Wi-Fi experience for their customers by incorporating Wi-Fi service as part of their internet offerings. By doing so, they can provide low latency Wi-Fi access to the internet. This might be made possible in part through the utilization of a combination of technologies that have been previously described. These technologies could potentially work together synergistically to help ensure that users can enjoy better connectivity, likely allowing them to browse the web, stream content, and engage in online activities with minimal delays or interruptions. By leveraging these technologies, ISPs might be able deliver an enhanced Wi-Fi experience that meets the growing demands and expectations of their customers.

7. Future of L4S in Wi-Fi

In a Wi-Fi network, congestion is typically concentrated at specific locations, such as the access network. The two primary factors that contribute to latency in Wi-Fi connections are; the increased delay caused by queuing and buffering under load, and the delays associated with the 802.11 media access control protocol. To tackle the queing delay it may be beneficial to implement Low Latency and Low Loss Scalable (L4S) throughput support in these congested regions. L4S effectively mitigates queuing delays caused by traditional congestion control protocols and enhances the previously mentioned Quality of Service (QoS) features.This involves deploying L4S Active Queue Management (AQM) systems and isolation mechanisms to enable coexistence with traditional congestion controllers. L4S operation requires isolating L4S flows from classic flows to protect queuing delay and using Explicit Congestion Notification (ECN) marking to signal congestion. Successful L4S deployment depends on correctly handling the ECN bits in IP packet headers. In the access network and in-home network, L4S support is preferred to mitigate queuing delays. In the aggregation networks and metro/core IP networks, sufficient link capacity can potentially minimize queuing delay, but isolation and prioritization of L4S traffic may be required. In fixed access networks, L4S support can be offered through L4S-capable devices or remote configuration of end-user devices. In mobile access networks, L4S support is usually needed in the Radio Access Network (RAN) and can be implemented through ECN marking in the CU. L4S can enable large-scale service offerings of real-time applications, while Ultra-Reliable Low Latency Communications (URLLC) is suited for strict end-to-end SLAs in controlled areas.

8. Conclusion

The growing demand for latency-sensitive applications like video conferencing, cloud gaming, and the Internet of Things has led to a focus on reducing latency in wireless networks. The Wi-Fi ecosystem has made significant changes with the goal of improving latency performance. The introduction of OFDMA in 802.11ax (Wi-Fi 6) has arguably enabled more efficient utilization of the wireless medium by allowing multiple clients to transmit simultaneously. Additionally, the implementation of QoS prioritization through WMM ensures that time-sensitive traffic, such as voice and video, receives preferential access to the network. Further enhancements, such as MLO, can likely be combined with techniques used in DOCSIS 4.0, like LLD and L4S, to extend low latency access for ISPs' customers end-to-end across their network, conceivably providing a seamless low latency experience.

Through a combination of technology and deployment strategies, Wi-Fi could bridge the latency gap with cellular networks, potentially allowing cable operators to deliver a seamless, low-latency experience to their customers across both wired and wireless access networks.

Abbreviations

AP	access point
AIFS	arbitration interframe space
AQM	active queue management
CSMA/CA	carrier sense multiple access with collision avoidance
CTS	clear to send
DIFS	distributed interframe space
DL	downlink
EDCA	enhanced distributed channel access
ECN	explicit congestion notification
EMLSR	extremely low latency single radio
HARQ	hybrid automatic repeat request
IoT	internet of things
L4S	low latency, low loss, scalable throughput
LBT	listen before talk
LLD	low latency docsis
MLD	media link descriptor
MLO	multilink operation
MSCS	mirrored stream classification service
MU-MIMO	multiuser multiple input multiple output
OFDMA	orthogonal frequency division multiple access
QoS	quality of service
RAN	radio access network
RTS	request to send
SIFS	short interframe space
STA	station
URLLC	ultrareliable low latency communication
UL	uplink

Bibliography & References

1. 5G Americas white paper on "5G Evolution: 3GPP Releases 16-17" (2020) - Provides an overview of the URLLC (Ultra-Reliable Low Latency Communication) features introduced in 3GPP Release 15 for 5G NR, including techniques to minimize latency such as frequent transmission opportunities, reduced transmission duration, and HARQ enhancements.

2. 5G Americas white paper on "URLLC in Unlicensed Spectrum" (2019) - Discusses the challenges and feasibility of implementing URLLC in the unlicensed spectrum using 5G NR-U, including the different LBT (Listen Before Talk) categories defined to enable coexistence with Wi-Fi.

3. Wi-Fi Alliance "Wi-Fi QoS Management Specification v3.0" (2021) - Details the Mirrored Stream Classification Service (MSCS) feature introduced by the Wi-Fi Alliance, which allows negotiation of

downlink QoS between Wi-Fi access points and client devices to ensure prioritization of latency-sensitive traffic.

4. IETF draft on "Low Latency, Low Loss, Scalable Throughput (L4S) Internet Service" (2022) - Discusses the deployment of L4S Active Queue Management (AQM) and isolation mechanisms in congested areas of the network, such as the access network and in-home network, to enable low latency and low loss for real-time applications.

5. Intel: "Wi-Fi 7 and Beyond" [Online]. Available: [Link](https://www.intel.com/content/dam/www/public/us/en/documents/pdf/wi-fi-7-and-beyond.pdf)

6. MediaTek: "MLO Infographic" [Online]. Available: [Link](https://www.mediatek.com/technology/mlo-infographic)

7. Netgear Community Forum: "MLO Multi-Link Operation WiFi7 of RS700" [Online]. Available: [Link](https://community.netgear.com/t5/Nighthawk-with-WiFi-7-BE/MLO-Multi-Link-Operation-WiFi7-of-RS700/m-p/2371909)

8. Netgear Community Forum: "MLO Multi-Link Operation WiFi7 of RS700" [Online]. Available: [Link](https://community.netgear.com/t5/Nighthawk-with-WiFi-7-BE/MLO-Multi-Link-Operation-WiFi7-of-RS700/m-p/2353727)

9. SNBForums: "Wi-Fi 7 Multi-Link Operation (MLO) Discussion" [Online]. Available: [Link](https://www.snbforums.com/threads/wi-fi-7-multi-link-operation-mlo-discussion.87598/)

10.Slalmi, Ahmed & Chaibi, Hasna & Chehri, Abdellah & Rachid, Saadane & Jeon, Gwanggil & Hakem, Nadir. (2020). On the Ultra-Reliable and Low-Latency Communications for Tactile Internet in 5G Era. Procedia Computer Science. 176. 3853-3862. 10.1016/j.procs.2020.09.003.