Breaking the Wi-Fi Latency Barrier by Combining L4S and EDCA Reconfiguration

Chia-Yu Chang*, Werner Coomans†, and Koen De Schepper*
*Nokia Bell Labs, Antwerp, Belgium, email: {chia-yu.chang, koen.de_schepper}@nokia-bell-labs.com
†Nokia, Antwerp, Belgium, email: werner.coomans@nokia-bell-labs.com

***Abstract*—Interactive applications such as video conferencing, remote control, and cloud gaming require not only low, but consistently low latency. In today's increasingly fiber-based networks, Wi-Fi can often act as the bottleneck. Wi-Fi latency is dominated by two factors: queue buildup before over-the-air transmission, and medium access contention among multiple devices. Supporting the IETF's latest Internet protocol, L4S (Low-Latency, Low-Loss, Scalable throughput), makes it possible to enable minimal queuing delay. This requires a design of the L4S dual-queue coupled Active Queue Management (AQM) that properly takes into account the channel access dynamics of the Wi-Fi link. Medium Access Control (MAC) delay, which then becomes the dominant contributor to Wi-Fi latency, can be reduced by designing new Enhanced Distributed Channel Access (EDCA) parameter sets. These new parameter sets reduce long-tail MAC delays and network outage events, while maintaining a fair throughput share relative to best-effort traffic. Testbed experiments and large-scale simulations show more than an order-of-magnitude improvement in worst-case latency and outage frequency, together with improved scalability as the number of contending Wi-Fi devices increases. Crucially, these improvements do not necessarily require new generation Wi-Fi devices; they can be supported on existing Wi-Fi 6/7 devices through firmware updates and reconfigurations.**

***Index Terms*—Wi-Fi, L4S, DualPI2, EDCA, QoS**

I. Introduction

Real-time applications, such as video conferencing, remote control, and cloud gaming, require *consistently* low latency to ensure a good user experience. To achieve such consistency, latency should remain low even under dynamic network conditions, including congestion and contention. This property is referred to as maintaining a low "working latency" (as opposed to merely low "idle latency") [1].

In practical Wi-Fi deployments, even a few active STAtions (STAs) and Access Points (APs) operating on the same channel can trigger performance degradation. As traffic intensity or device load increases, several user-visible artifacts may appear, e.g., conversational glitches and video freezes. The root causes are twofold: (i) queue buildup before over-the-air transmission, which increases queuing delay, and (ii) medium access contention among multiple devices, which leads to large Medium Access Control (MAC) delay and occasional network outages. As a result, the high raw capacity of a Wi-Fi link often fails to deliver low latency in practice.

Next-generation Wi-Fi promises several features for high throughput and low latency. For example, Wi-Fi 8 introduces multi-AP coordination, Transmission Opportunity (TXOP) preemption, and high-priority Enhanced Distributed Channel Access (EDCA) [2], but they are effective mainly in selected scenarios rather than as a general Wi-Fi Quality of Service (QoS) solution. Thus, latency spikes can persist even with these Wi-Fi features.

Recently, IETF introduced Low-Latency, Low-Loss, Scalable Throughput (L4S) as a new Internet architecture to address excessive queuing delay [3]. It marks an evolution from a non-AQM baseline (i.e., without Active Queue Management (AQM), using tail-drop), through dropping-based AQM (loss as a congestion signal), to a marking-based approach via fine-grained Explicit Congestion Notification (ECN). L4S also requires application support [4] to adjust the sending rate in response to the ECN marking probability. This synergy reduces queue backlogs even at a shared bottleneck (e.g., per-destination transmission queues at an AP).

While L4S can mitigate queuing delay before transmission, medium access contention (e.g., exponential backoff and air-interface collisions) remains unavoidable for securing TXOPs. Therefore, after deploying L4S, MAC delay dominates Wi-Fi latency in multi-STA scenarios. This delay is largely driven by the Wi-Fi Multimedia (WMM) framework, where all traffic flows are classified into different Access Categories (ACs), and each AC receives differentiated channel access through specific EDCA parameter sets. For example, high-priority ACs (e.g., video, voice) are assigned more favorable parameters to increase their chance of winning contention with low-priority AC (e.g., best-effort).

This article therefore puts forward two complementary mechanisms. First, based on the dual-queue coupled AQM model [5], we introduce L4S into Wi-Fi to reduce excess queuing delay. Note that due to dynamic radio conditions and varying user (dis)associations of Wi-Fi networks, the conventional queuing-delay-based scheme should be modified to adapt to the TXOP-driven characteristic of IEEE 802.11 links. Second, we address MAC delay by revisiting EDCA parameter design. We propose new EDCA parameter sets and the parameter adaptation scheme to not only reduce long-tail MAC delay and outage events in multi-STA scenarios but also preserve near-fair throughput with best-effort traffic. Together, these two mechanisms target the respective sources of Wi-Fi latency and enable consistently low latency under real-world loaded scenarios.

In summary, this article makes the following contributions:

- A Wi-Fi solution architecture integrating L4S by adapting the dual-queue model [5] to 802.11 characteristics, enabling minimal queuing delay.
- A redesigned EDCA parameter set and adaptation scheme that reduce long-tail MAC delay even with a larger

number of active STAs.
- Introduction of "*outage duration*" and "*outage occurrence interval*" as new Wi-Fi QoS metrics.

We validated these improvements through both experiments and simulations. Crucially, they do not require next-generation Wi-Fi devices and can be supported on existing Wi-Fi 6/7 devices through firmware updates and reconfigurations. For example, all experiments in this article were conducted using Wi-Fi 6 equipment.

II. Related work

A. Wi-Fi Latency Components and the Role of Wi-Fi 8

Wi-Fi latency comprises several components: (i) *propagation delay* limited by the speed of light; (ii) *PHY delay* incurred during the transmission and reception of preamble, header, payload; (iii) *MAC delay* introduced by contention, backoff, and collisions under distributed coordination; and (iv) *queuing delay* caused by Wi-Fi frames accumulating in buffers (e.g., device memory) before transmission. Without an AQM mechanism, queuing delay can easily reach or exceed hundreds of milliseconds. In addition, under medium access contention among a few Wi-Fi devices, the tail of MAC delay can inflate to more than hundreds of milliseconds, causing network outages. Accordingly, both queuing and MAC delays contribute to significant latency variability.

Wi-Fi 8 introduces several features to improve latency [6], such as multi-AP coordination, TXOP preemption for prioritized traffic, and high-priority EDCA for protected contention. While effective in reducing MAC delay in specific scenarios, these features may deteriorate latency for de-prioritized traffic or legacy devices. Therefore, fundamental mechanisms are still required to address buffer management and long-tail MAC delay.

B. L4S Architecture

L4S [3] provides a scheme to eliminate excessive queuing delay through scalable congestion control [4] at the transport layer with ECN marking in the IP header. It leverages two ECN bits therein: the sender sets one bit to indicate its capability to adjust the sending rate based on the network congestion level, and the network infrastructure sets the second bit for an appropriate proportion of packets, enabling fine-grained rate control by the sender without causing delays or dropping packets. Notably, this network-application collaboration requires the receiver to provide frequent and accurate status feedback to complete the control loop.

Importantly, L4S traffic can fairly coexist with non-L4S (i.e., classic) traffic by redesigning the queue model that reflects bottleneck characteristics. Hence, L4S is a good fit for existing Wi-Fi devices to reduce queuing delay and complements lower-layer improvements that reduce MAC delay.

C. EDCA Parameters and Limitations

WMM provides differentiated channel access through four ACs (*AC_VO*, *AC_VI*, *AC_BE*, *AC_BK*), corresponding to voice, video, best effort, and background traffic, respectively. Each AC is assigned a different priority queue and a set of contention parameters: the Arbitration Inter-Frame Space Number (AIFSN) determines the waiting time before initiating and resuming backoff; the minimum and maximum Contention Window (CW) values (*CW_min*, *CW_max*) define how the backoff window grows after collisions[1]; and the TXOP limit specifies the maximum duration a device can retain medium access after winning contention. High-priority ACs (voice and video) use smaller AIFSN, CW_min, and CW_max than best-effort traffic to reduce MAC delay, while TXOP limits are longer for video to handle high-bandwidth bursts and shorter for voice due to its small packet nature.

In addition, several studies have identified shortcomings of existing EDCA parameter sets. Results in [7] show that static EDCA parameters can lead to excessively high MAC delay for best-effort traffic, regardless of whether it is competing with the same or different ACs. Some extensions have been proposed for improvements, such as adaptive EDCA parameter tuning [8] and model-driven admission control [9]. However, these solutions rely on centralized coordination for reconfiguration [8] or additional signaling for runtime adaptation [7], making them difficult to deploy universally.

More recent extensions based on Orthogonal Frequency Division Multiple Access (OFDMA) in Wi-Fi 6E/7 improve spectral efficiency but still rely on initial contention outcomes before allocating resources. Therefore, the long-tail MAC delay can persist and even deteriorate due to legacy devices or signaling overhead [10].

D. Summary

Existing work on low-latency Wi-Fi typically targets a single source of delay, either queuing delay through AQM [11] or MAC delay through contention management [7], [8]. Addressing only one dimension, however, is insufficient for consistently low latency, motivating a joint consideration of buffer management and medium access control.

III. L4S for Wi-Fi Networks

The goal of applying L4S to Wi-Fi is to reduce queuing delay before Wi-Fi frames are transmitted over the air interface. However, the IEEE 802.11 MAC layer and most existing Wi-Fi chipsets do not support L4S-aware queue management or ECN marking. Therefore, its functionalities need to be implemented at a higher layer, typically in the host network stack (e.g., Linux traffic control/qdisc used in OpenWRT) or device driver, without requiring firmware or hardware modifications on existing Wi-Fi devices.

In our design, the host or driver maintains separate transmit queues for each destination to prevent the MAC layer from backlogging large numbers of undelivered packets in hardware (e.g., aggregation buffer shown in Fig. 1). In addition, the amount of data released from each transmit queue is controlled to ensure efficient utilization of Wi-Fi link capacity while avoiding excessive buffering at the MAC layer. In practice,

[1] Each device picks a random value in the range $[0, \mathrm{CW}]$ as its backoff timer before transmitting. CW starts at CW_min and doubles after each collision (up to CW_max). Upon a successful transmission, CW is reset to CW_min.

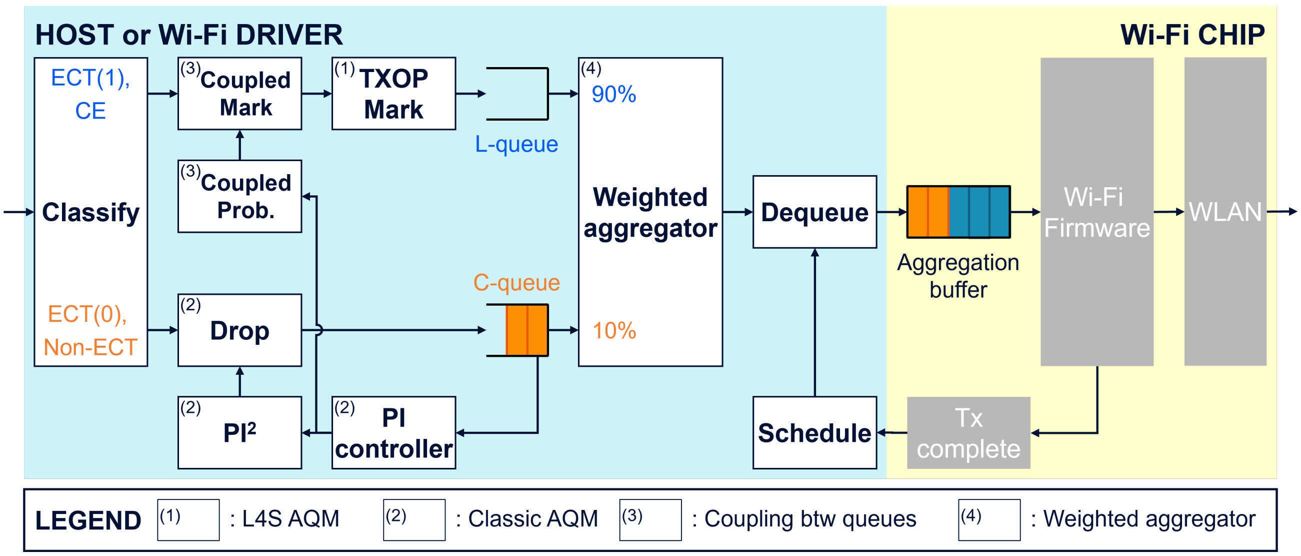


Fig. 1: Redesigned dual-queue with coupled AQM for Wi-Fi.

when the Wi-Fi chipset reports a successful delivery (e.g., upon reception of a Block Ack), the host or driver retrieves the latest PHY rate and TXOP limit to refill the aggregation buffer with a bounded number of bytes and keep the chipset buffer shallow. In summary, this host/driver-based design enables L4S on existing Wi-Fi devices, whether they are APs or STAs.

A. *Dual-queue Coupled AQM for Wi-Fi*

In our dual-queue design for Wi-Fi, traffic is classified into two queues: the *C-queue* for classic traffic and the *L-queue* for L4S traffic. As illustrated in Fig. 1, this design consists of: (1) an L4S AQM for the L-queue, (2) a classic AQM for the C-queue, (3) a coupling mechanism between the two queues, and (4) a weighted aggregator.

L4S AQM marks L4S traffic, typically based on the queuing-delay metric. However, in Wi-Fi, volatile link capacity and unpredictable channel access can cause a delay-based metric to underestimate congestion (e.g., sudden PHY rate increases due to spatial multiplexing) or overestimate it (e.g., long intervals between TXOPs). Therefore, the L4S AQM uses a TXOP-capacity metric by comparing the L-queue occupancy to the number of bytes that fit into the TXOP size for L4S traffic (i.e., no marking occurs if the queued bytes fit into the allocated TXOP). An optional scaling factor (e.g., 1–2) allows L-queue occupancy to slightly exceed the TXOP size for L4S traffic, balancing the trade-off between throughput and latency as a representative operating point.

Classic AQM computes the “dropping probability” for the C-queue using a Proportional Integral (PI) controller. To tailor it to the Wi-Fi characteristics, we apply one of two approaches depending on the TXOP size: (1) when the TXOP size for classic traffic is small, we control queuing delay toward a delay target; (2) otherwise, we control the C-queue length towards the TXOP size for classic traffic. The former queuing-delay-based method is effective when low PHY rate dominates the delay, while the TXOP-capacity-based approach is suitable when delay is driven by sparse channel access. In addition,

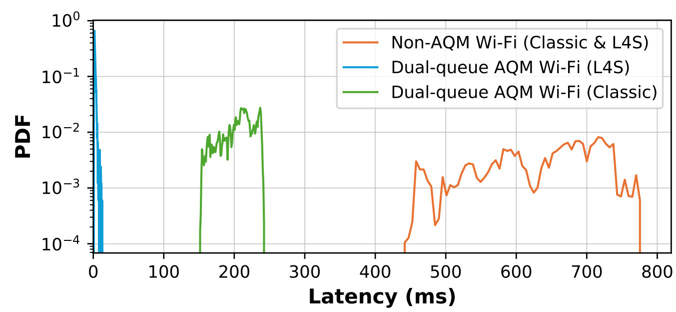


Fig. 2: Measured latency PDFs for non-AQM Wi-Fi and dual-queue Wi-Fi when the link is fully congested.

we apply the squared rule (PI^2) [5] to calculate the dropping probability for the C-queue and derive a coupled marking probability for the L-queue. This coupling mechanism ensures rate fairness between L4S and classic traffic.

Finally, a weighted aggregator assigns the L-queue a higher weight to work in conjunction with the coupling mechanism. When both queues are active, rate fairness is achieved by coupled marking, while weights mainly prevent transient blocking. A weight of 90% for the L-queue can avoid unnecessary L4S traffic limitation (if too low) and prevent C-queue from building up to trigger coupled marking (if too high). This weight can be programmable [5].

B. *Performance Evaluation*

We measure the latency of Classic and L4S traffic transmitted to a Wi-Fi 6 STA via a Nokia Wi-Fi 6 AP over a 20 MHz channel, with a Modulation and Coding Scheme (MCS) index around 8. The latency Probability Density Function (PDF) is shown in Fig. 2. Without AQM, both traffic types experience large queuing delay, peaking at 760 ms and fluctuating between 440 ms and 760 ms. This peak delay can increase or decrease due to different service rates for the same queue length, but large delay fluctuations always persist.

Next, the dual-queue is introduced with one-TXOP target for L4S AQM and a 200 ms PI controller target for Classic AQM. A target value between 50 ms and 200 ms is a representative operating point for Wi-Fi devices: smaller targets can reduce airtime utilization, while larger targets mainly increase classic queuing delay with limited impact on L4S. Latency is significantly improved for both classic and L4S traffic. Compared to the non-AQM baseline, classic traffic exhibits lower peak latency and smaller latency fluctuations, remaining close to the 200 ms target. For L4S traffic, the peak latency is drastically reduced to around 10–20 ms. Extra results for deploying L4S over Wi-Fi are shown in [12].

IV. Redesigned EDCA Parameters & Parameter Adaptation Scheme

With queuing delay reduced by L4S, MAC delay becomes the main source of Wi-Fi latency. This delay primarily arises from medium access contention among multiple STAs using the same or different ACs.

A. Quantifying MAC Delay: Outage Duration

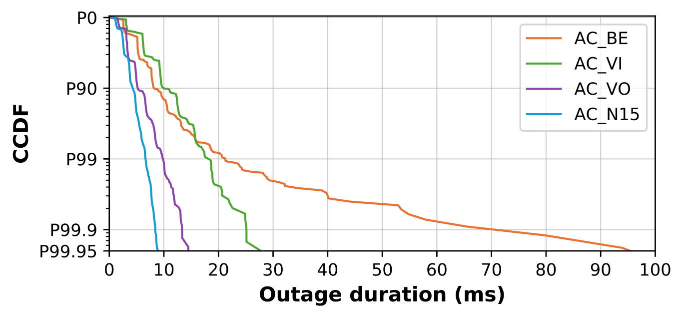
To quantify MAC delay, we measure the time interval between the starts of two consecutive TXOPs won by the same STA, defined as the *outage duration* — the time during which the medium is inaccessible to that STA.

We conduct experiments with two STAs using the same AC. Two Nokia Wi-Fi 6 APs operate on the same 20 MHz channel, each serving one Wi-Fi 6 STA transmitting full-buffer uplink traffic. As shown by the Complementary Cumulative Distribution Function (CCDF) in Fig. 3, when both STAs use AC_BE, the outage duration exceeds 95 ms at the 99.95th percentile. In practical terms, this means a STA occasionally waits nearly 100 ms before regaining medium access. For high-priority ACs, large MAC delays persist: P99.95 outage duration is 27 ms for AC_VI and 14 ms for AC_VO.

B. New EDCA Parameter Set

The above results reveal inherent limitations of legacy EDCA configurations. Our objective is therefore to design a new EDCA parameter set with the following properties. First, it should maintain low latency for itself as well as for other traffic sharing the same medium (e.g., other STAs). Second, because we target a general low-latency solution, it should be airtime-fair and scalable to multiple STAs (compared to AC_VI and AC_VO). Furthermore, considering coexistence with legacy EDCA configurations, the new parameter set must be more resilient to AC_VI and AC_VO in terms of latency and throughput (compared to AC_BE), while also remaining sufficiently friendly to AC_BE (compared to AC_VO). Finally, the new parameter set is not intended to obtain throughput advantages, but to reduce MAC delay.

To meet these goals, we introduce a new EDCA parameter set, AC_N15 (Fig. 3). First, it uses a static contention window range (CW_min = CW_max), eliminating exponential backoff growth and reducing prolonged MAC delays after collisions. Second, the CW value is chosen to lie between the CW_max of AC_VI and the CW_min of AC_BE (i.e., CW = 15), yielding a configuration that is neither more aggressive than AC_VI nor more conservative than AC_BE. Third, the TXOP limit is shortened to prevent any single STA from occupying the medium for a long period. Finally, a smaller AIFSN minimizes the waiting time before entering backoff, partially compensating for the small TXOP limit.



	AC_BE	AC_VI	AC_VO	AC_N15
AIFSN	3	2(STA)/1(AP)	2(STA)/1(AP)	1
CW_min	15	7	3	15
CW_max	1023	15	7	15
TXOP limit	2.5ms	3ms	1.5ms	1ms

Fig. 3: EDCA parameter sets and outage duration CCDFs for AC_BE, AC_VI, AC_VO, and AC_N15.

As shown in Fig. 3, AC_N15 reduces P99.95 outage duration to 8.8 ms, substantially mitigating long-tail MAC delay compared to all legacy EDCA configurations.

C. Comparison across AC Combinations

Using the same experiment setup, we evaluate all pairwise combinations of ACs. Fig. 4 shows the average throughput and P99.95 outage duration for the two STAs. When AC_VI or AC_VO coexists with AC_BE, the high-priority AC achieves lower latency and higher throughput, but at the cost of increased latency and reduced throughput for AC_BE. For example, when AC_BE coexists with AC_VO, its throughput drops to roughly one-tenth of that of AC_VO, and its P99.95 outage duration reaches 300 ms. These advantages diminish when high-priority ACs compete with each other; when AC_VI and AC_VO coexist, their P99.95 outage durations double compared to when each competes with AC_BE.

In contrast, when both STAs use AC_N15, their P99.95 outage durations are minimized simultaneously. Although this results in a modest throughput reduction (83.5% of that achieved by two AC_BE STAs), such a reduction can be offset by increased bandwidth or extra spatial streams.

Importantly, AC_N15 maintains its low-latency characteristic when coexisting with legacy EDCA parameter sets. When competing with AC_BE, throughput remains nearly evenly distributed while AC_N15 achieves lower latency. When competing with AC_VI or AC_VO, AC_N15 accepts reduced throughput but retains 27–30 ms P99.95 outage durations, while high-priority ACs preserve their throughput advantages. This demonstrates that low MAC delay does not require throughput privilege.

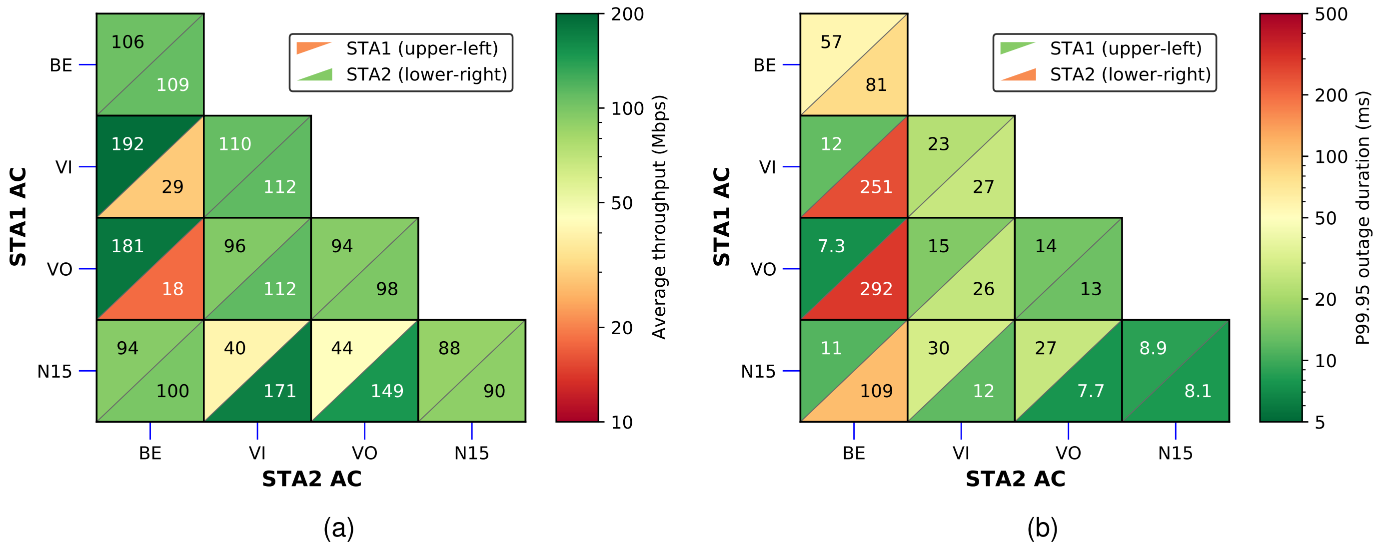


Fig. 4: (a) Average throughput and (b) P99.95 outage duration for pairwise AC combinations of two STAs.

D. Quantifying outage frequency: Outage Occurrence Interval

While outage duration characterizes the severity of each individual MAC delay, it does not capture how often such events occur during a Wi-Fi session. For interactive applications, the frequency of long outages is often critical to user experience. For example, repeated 300 ms delays can make applications sluggish and unresponsive, while frequent 1 second stalls can render a service unusable.

To quantify it, we define the *outage occurrence interval* as the average time interval between two consecutive outages exceeding a given duration. This metric directly reflects how frequently users encounter performance issues. For instance, if the occurrence interval for 300 ms outages is 20 seconds, the application will likely fail to respond approximately three times during a one-minute Wi-Fi session.

E. Scalability Evaluation

To evaluate the scalability of different EDCA parameter sets under dense contention (up to 128 active STAs), we conduct long-duration simulations to model Wi-Fi sessions exceeding one day. Specifically, we focus on the occurrence intervals for both 300 ms and 1 second outages. The results for different numbers of active STAs are shown in Fig. 5 to compare whether different EDCA parameters can consistently provide low latency at different contention levels.

For AC_BE, even under light contention with only four STAs, outages exceeding 300 ms and 1 second occur frequently, with average occurrence intervals of about 20 seconds and 2 minutes, respectively. As the number of STAs increases, outages rapidly become more frequent. With 16 competing STAs, 1 second outages occur every 7 seconds on average, resulting in continuous service disruptions that are unacceptable to users.

High-priority ACs (AC_VI and AC_VO) provide low latency when the number of contending STAs is small. With four STAs, no 300 ms or 1 second outages are observed over day-long sessions. However, this benefit does not scale. With eight STAs, 300 ms outages already occur every 30–60 seconds, and with 16 STAs such outages appear every 1–2 seconds. These results indicate that high-priority ACs fail to maintain low latency as density increases.

F. Adaptive EDCA based on AC_N15

The proposed AC_N15 configuration reveals a challenge in dense deployments. As the number of active STAs increases, the probability that more than one STA chooses the same backoff value from a small range rises sharply. Therefore, we introduce an adaptation mechanism that dynamically adjusts the contention window parameters (CW_min and CW_max) for AC_N15. Specifically, several relaxed-CW variants of AC_N15 are used (e.g., AC_N255, where CW_min = CW_max = 255), thereby increasing contention window range and reducing collision probability. In practice, the AP can estimate the number of active STAs by using existing MAC layer statistics (e.g., association table and recent uplink activity over a sliding time window)[2], without requiring hardware update or extra signaling.

As shown in Fig. 5, the adaptation scheme progressively increases the contention window (e.g., 31, 63, 127, 255, 511, 1023) as the estimated number of active STAs grows. With 32 STAs, AC_N255 results in only one 300 ms outage during a one-hour session. As for 128 STAs, AC_1023 relaxes the occurrence of 1 second outage to a 30-minute interval, while legacy EDCA parameter sets experience same outages every few seconds, severely degrading user experience.

G. Summary

These results show that AC_N15 and its relaxed-CW variants can provide consistently low MAC latency, and that the adaptation scheme is deployable on existing Wi-Fi devices.

[2]Advanced schemes based on collision probability, idle-slot estimation, or machine learning depend on new runtime MAC statistics and AP resources.

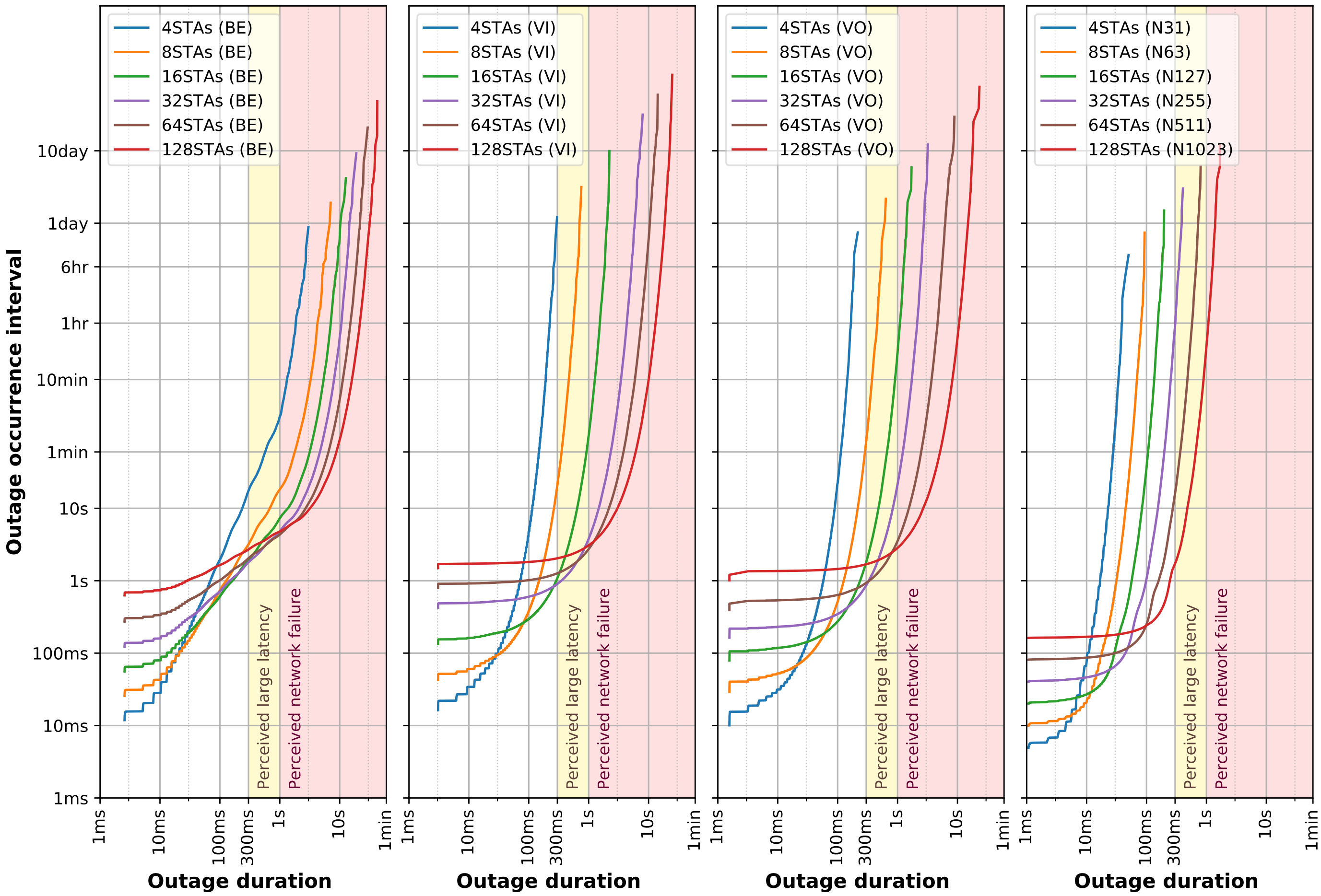


Fig. 5: Outage occurrence intervals versus number of active STAs for multiple EDCA settings.

V. Joint L4S and EDCA Evaluations

A. Performance Evaluation

To evaluate the effectiveness of jointly applying the dual-queue coupled AQM and the redesigned EDCA parameters, we conduct experiments on existing Wi-Fi devices. In our setup, a Nokia Wi-Fi AP serves different numbers of STAs: one STA receives downlink L4S traffic, while the remaining STAs generate full-buffer uplink traffic. We measure the latency of the L4S traffic to assess how queuing delay and MAC delay contribute to Wi-Fi latency.

Fig. 6 shows the latency CCDFs for scenarios with 1, 2, and 4 active STAs, comparing *L4S-only* and *joint L4S and new EDCA* configurations. Results for the non-AQM baseline are omitted, as Wi-Fi latency is dominated by excessive queuing delay and fluctuates between 450 ms and 800 ms (similar to Fig. 2). In addition, packet loss increases rapidly with contention, reaching 0.03%, 3.86%, and 7.11% for 1, 2, and 4 STAs, respectively. After applying L4S, the P99.99 latency is about 20 ms in the 1-STA case, but increases to nearly 60 ms and 105 ms in the 2-STA and 4-STA cases, respectively. This shows that once queuing delay is suppressed, MAC delay grows rapidly with the number of active STAs and becomes dominant. When both L4S and redesigned EDCA parameters are applied, the P99.99 latency increases only mildly with contention and remains about 35 ms in the 4-STA case. Overall, the combined approach effectively reduces both queuing delay and MAC delay.

B. Implementation and Deployment Discussion

As discussed earlier, the dual-queue coupled AQM can be implemented in the host network stacks or device drivers on a per-destination basis. Since existing Wi-Fi devices already maintain multiple priority queues under the WMM framework, this mechanism can, in principle, be applied to all WMM access-category queues, provided that computing and memory resources permit. However, we recommend it at least on the best-effort transmission queue, since Differentiated Services Code Point (DSCP) bleaching may cause latency-sensitive traffic to fall into AC_BE.

The redesigned EDCA parameters and adaptation scheme can be deployed directly on existing Wi-Fi platforms, including open-source implementations (e.g., WMM/EDCA profiles in `hostapd`) and most enterprise-grade APs. In the longer term, we envision AC_N15 (or its variants) acting as the default WMM profile and operating alongside multiple AQM-enabled transmit queues to support Wi-Fi QoS (e.g., based on DSCP values in the IP header). This approach can naturally integrate low latency into future Wi-Fi features for higher throughput.

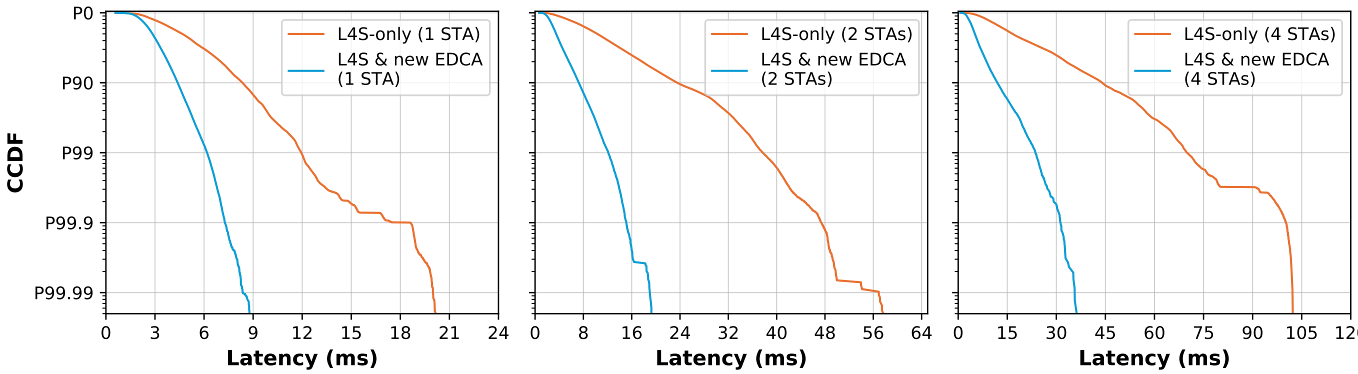


Fig. 6: Latency CCDFs comparing L4S-only and joint L4S and redesigned EDCA for different numbers of active STAs.

VI. Conclusions

This article demonstrates that combining an L4S-enabled dual-queue coupled AQM with a redesigned EDCA scheme can significantly reduce Wi-Fi latency. The dual-queue AQM effectively limits queue buildup at the Wi-Fi air interface by maintaining only enough bytes in the L-queue and C-queue to fill the next TXOP. We observe an order-of-magnitude reduction in queuing delay, making MAC delay the dominant latency component. MAC delay is mainly caused by contention among multiple STAs or ACs. Legacy high-priority categories (AC_VI, AC_VO) reduce latency at the expense of AC_BE, while struggling to scale beyond a few STAs. Our proposed low-latency category, AC_N15 significantly shortens long-tail outage durations while maintaining fair throughput with legacy best effort traffic. Additionally, the dynamic extension of AC_N15 adapts contention windows based on the number of active STAs, substantially reducing the frequency of perceived outages and achieving an additional order-of-magnitude improvement. Together, these mechanisms provide a practical path toward low-latency Wi-Fi at scale.

To support reproducibility, simulation scripts are available upon reasonable request, or via restricted repository access when limited by licensing or IP constraints.

References

[1] Broadband Internet Technical Advisory Group (BITAG), “Latency explained,” Technical Working Group Report, Tech. Rep., Jan. 2022, accessed on: 2026-05-12. [Online]. Available: https://bitag.org/documents/BITAG_latency_explained.pdf

[2] W. Ning, F. Tang, M. Zhao, and N. Kato, “A Survey on IEEE 802.11bn Wi-Fi 8: Advantages of Ultra High Reliability for Next-Generation Wireless LANs,” *IEEE Communications Surveys & Tutorials*, vol. 28, pp. 4215–4247, 2026.

[3] B. Briscoe, K. D. Schepper, M. Bagnulo, and G. White, “Low Latency, Low Loss, and Scalable Throughput (L4S) Internet Service: Architecture,” RFC 9330, Jan. 2023, accessed on: 2026-05-12. [Online]. Available: https://www.rfc-editor.org/info/rfc9330

[4] L4S Team. (2025) Udp prague reference implementation. GitHub repository. Accessed on: 2026-05-12. [Online]. Available: https://github.com/l4steam/udp_prague

[5] K. D. Schepper, B. Briscoe, and G. White, “Dual-Queue Coupled Active Queue Management (AQM) for Low Latency, Low Loss, and Scalable Throughput (L4S),” RFC 9332, Jan. 2023, accessed on: 2026-05-12. [Online]. Available: https://www.rfc-editor.org/info/rfc9332

[6] A. Karamyshev, I. Levitsky, D. Bankov, and E. Khorov, “A Tutorial on Wi-Fi 8: The Journey to Ultra High Reliability,” *Problems of Information Transmission*, vol. 61, no. 2, pp. 164–210, 2025.

[7] X. Chen, I. Akinyemi, and S.-H. Yang, “A control theoretic approach to achieve proportional fairness in 802.11 e EDCA WLANs,” *Computer Communications*, vol. 75, pp. 39–49, 2016.

[8] P. Serrano, A. Banchs, P. Patras, and A. Azcorra, “Optimal configuration of 802.11 e EDCA for real-time and data traffic,” *IEEE Transactions on Vehicular Technology*, vol. 59, no. 5, pp. 2511–2528, 2010.

[9] N. C. Taher, Y. Ghamri-Doudane, B. El Hassan, and N. Agoulmine, “An efficient model-based admission control algorithm to support voice and video services in 802.11 e wlans,” in *2009 Global Information Infrastructure Symposium*. IEEE, 2009, pp. 1–8.

[10] K. Montgomery, M. K. Hany, and R. Candell, “Latency-sensitive networked control using 802.11 ax ofdma triggering,” in *2024 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE, 2024, pp. 1088–1095.

[11] T. Høiland-Jørgensen, M. Kazior, D. Täht, P. Hurtig, and A. Brunstrom, “Ending the Anomaly: Achieving Low Latency and Airtime Fairness in WiFi,” in *2017 USENIX Annual Technical Conference (USENIX ATC 17)*, 2017, pp. 139–151.

[12] Wireless Broadband Alliance, “Implementation Guidelines for Low Latency, Low Loss, and Scalable Throughput in Wi-Fi Networks (L4S),” [Online], accessed on: 2026-05-12. [Online]. Available: https://wballiance.com/implementation-guidelines-for-l4s/