

From Connectivity to Intelligence: AI-Driven Wi-Fi Systems

MediaTek Filogic White Paper

Release Date: 29 May 2026

Use of this document and any information contained therein is subject to the terms and conditions. This document is subject to change without notice.

© 2026 MediaTek Inc. All rights reserved.

Unauthorized reproduction or disclosure of this document, in whole or in part, is strictly prohibited.

Contents

1. Key Insights	3
2. Introduction	4
3. The Evolution of AI: From Rules to Reasoning	4
3.1 Traditional AI and Machine Learning (ML)	4
3.2 Generative AI and Language Models	5
3.3 Agentic AI	5
4. AI and Wi-Fi Convergence	6
4.1 Demand Inflection: Toward AI Token-Centric	6
4.2 From Throughput to Reliability	6
4.3 Ecosystem Momentum	6
5. Hybrid AI Architecture: Edge, Cloud and Telemetry	7
5.1 Edge AI (On-AP)	7
5.2 Cloud AI	8
5.3 The Telemetry Pipeline	8
6. Closed-Loop Intelligence: The Agentic AI Paradigm	8
7. Use Cases and Customer Impact	9
8. MediaTek Filogic AI in Practice	10
8.1 Intelligent Quality of Experience	11
8.2 AI-Assisted Adaptive Power Saving with Intelligent Power Management	12
9. Enabling AI at the Silicon Level	13
10. Beyond TOPS: Memory, Bandwidth, and Silicon Enablement	14
10.1 The TOPS Metric: Useful but Incomplete	14
10.2 Memory Capacity: What Models Can You Run?	14
10.3 Memory Bandwidth: How Fast Can Models Run?	15
10.4 A Framework for Evaluation	15
11. Challenges and Considerations	16
12. Conclusion	17
13. MediaTek in the Wi-Fi Industry	17
14. Acknowledgements	17

1. Key Insights

- Wi-Fi networks are scaling in size and intelligence, creating strong momentum toward autonomous, AI-driven network management.
- AI capability in access points is defined by system balance, where compute, memory, bandwidth, and telemetry together determine practical AI workloads, not TOPS alone.
- Wi-Fi 8 marks a shift toward experience-centric performance, prioritizing reliability, intelligence, and real-world operation over peak throughput.
- Closed-loop AI enables proactive network optimization, continuously translating intent into action and refining outcomes through telemetry feedback.
- Access points are evolving into intelligent edge platforms, extending beyond connectivity to support AI services, natural language interaction, and autonomous decision-making.

2. Introduction

Wi-Fi has become the primary method of network access for the connected world, supporting more than 20 billion active devices. At the same time, AI-driven applications such as cloud inference, edge computing, AR/VR, and autonomous IoT systems are placing unprecedented demands on the wireless connectivity layer. These applications require consistent low latency, high reliability, and deterministic quality of service to function correctly and deliver a seamless user experience.

This creates a bidirectional value proposition that defines the thesis of this whitepaper: advanced Wi-Fi technologies are essential to unlocking the full potential of AI, while the rapid growth and evolution of AI workloads are, in turn, accelerating the need for more intelligent, capable, and efficient Wi-Fi networks. Table 1 summarizes this relationship.

AI FOR Wi-Fi	Bidirectional Value	Wi-Fi FOR AI
<ul style="list-style-type: none"> • Self-optimizing, self-healing networks • Autonomous troubleshooting & management • Intelligence integrated across the full stack 		<ul style="list-style-type: none"> • Low-latency connectivity for edge AI • Reliable transport for cloud inference • Deterministic QoS for AI workloads

Table 1. The bidirectional relationship: AI for Wi-Fi and Wi-Fi for AI

This whitepaper examines how artificial intelligence is evolving the Wi-Fi access point from a connectivity endpoint into an intelligent network platform. It outlines the architectural approaches, hardware requirements, practical use cases, and industry challenges associated with this transformation, providing a framework for evaluating AI readiness in next-generation Wi-Fi infrastructure.

3. The Evolution of AI: From Rules to Reasoning

Understanding AI's role in Wi-Fi starts with recognizing that "Artificial Intelligence" encompasses many techniques, each with different strengths, limits, and system requirements. The industry is moving through three main phases, each building on the previous one.

3.1 Traditional AI and Machine Learning (ML)

The first generation of AI in networking used classical ML: supervised models for anomaly detection, and rule-based systems for automated RF optimization. These have been used for years in enterprise WLAN management, usually running in the cloud on data collected from large fleets of access points.

Traditional ML is effective within well-defined scopes but is inherently specialized. Individual models are typically designed to address specific tasks—such as channel selection, client steering, or interference classification—and rely on manual feature engineering and labeled training data. As a result, these models tend to have limited ability to generalize beyond their original problem domains, offer only constrained human interpretable explanations, and often require retraining or retuning to adapt to new environments or operating conditions.

3.2 Generative AI and Language Models

Large language models (LLMs) and small language models (SLMs) introduce new capabilities for Wi-Fi management by operating on unstructured data, reasoning across complex system states, and interacting through natural language. For Wi-Fi management, generative AI enables:

- Natural language querying, for example: “Why are users on Floor 3 experiencing slow video calls?” with a clear, contextual answer instead of raw graphs.
- Log and event analysis, correlating system logs and client events to find patterns that are hard and time-consuming for humans to detect.
- Automated report generation, including network health summaries, incident reviews, and capacity recommendations.

3.3 Agentic AI

The next phase is agentic AI: systems that can continuously observe the network, determine appropriate actions, change configurations through APIs or CLIs, and validate outcomes through telemetry. This shifts AI from an advisor to an autonomous operator.

In these systems, a language model is the reasoning core. It interprets high-level intent, such as “optimize video conferencing quality during business hours,” and breaks it into specific actions like tuning QoS policies, adjusting channel plans, or updating client steering rules. A built-in feedback loop monitors the impact of changes and can roll back if performance worsens.

This progression from traditional ML to general language reasoning, to autonomous action mirrors the evolution of Wi-Fi operations itself: moving from reactive troubleshooting toward proactive, selfoptimizing networks.

Table 2 provides a concise comparison of Traditional AI and ML, Generative AI, and Agentic AI, highlighting differences in operational scope, deployment location, adaptivity, and level of human dependency

AI Type	Scope	Runtime Location	Adaptivity	Human Dependency
Traditional AI & ML	Task-specific	Cloud	Limited	High
Generative AI	Cross-domain	Edge + Cloud	Medium	Medium
Agentic AI	System-level	Edge-first	High	Low

Table 2. High-level comparison of Traditional AI & ML, Generative AI, and Agentic AI system characteristics

4. AI and Wi-Fi Convergence

The integration of AI into Wi-Fi is already underway; driven by standards development, vendor innovation, and growing ecosystem collaboration. Emerging efforts—such as standardized telemetry for AI consumption, distributed intelligence across access points, and AI-assisted PHY and MAC optimization—illustrate how these technologies are beginning to converge in practical deployments.

4.1 Demand Inflection: Toward AI Token-Centric

The network demand profile is changing in ways that make this convergence urgent. Agentic AI, where models proactively plan and execute multi-step tasks on behalf of users, is driving a shift from traditional mixed traffic (including video, voice, and web) to AI token-centric traffic. Daily AI token consumption in high-adoption markets grew 300 times in 18 months, exceeding 30 trillion tokens by mid-2025¹. Hybrid AI architectures, where inference is split between on-device and cloud, are particularly demanding: they require not just throughput but deterministic low latency to maintain coherent behavior across the device-cloud boundary. Next-generation Wi-Fi is essential infrastructure for this model, delivering the bandwidth headroom and latency consistency that hybrid AI requires at scale.

4.2 From Throughput to Reliability

Wi-Fi 8 is being defined as the “Ultra High Reliability” (UHR) generation, reflecting a fundamental industry reorientation. As MediaTek has articulated in its Wi-Fi 8 whitepaper series, the focus is shifting from maximizing theoretical PHY rates to delivering reliable, consistent performance in real-world environments.

AI is a natural complement to this reliability focus. While standards-based features create the PHY and MAC layer mechanisms for improved reliability, AI provides the intelligence layer that dynamically optimizes how these mechanisms are deployed in response to real-time network conditions.

4.3 Ecosystem Momentum

Across the Wi-Fi ecosystem, AI integration is becoming a competitive differentiator. Cloud management platforms are incorporating AI-assisted troubleshooting and predictive analytics. Access point vendors are exploring on-device inference capabilities. Enterprise customers are increasingly requesting AI-driven features in their RFP requirements.

This momentum creates a window of opportunity. The access point’s chipset determines what AI capabilities are fundamentally possible: the available compute for inference, the telemetry data that can be extracted from the radio, and the memory subsystem that constrains which models can run. Silicon level AI enablement is the foundation upon which the entire AI-driven Wi-Fi ecosystem is built.

¹ <https://counterpointresearch.com/en/insights/Agentic-AI-Driving-Paradigm-Shift-in-Mobile-AI-Transforming-5G-Network-Evolution>

5. Hybrid AI Architecture: Edge, Cloud and Telemetry

Deploying AI in Wi-Fi access points is not one single, monolithic task. The most effective designs distribute intelligence across three main components, each optimized for different workloads and latency requirements, as illustrated in Figure 1. As a general architectural principle, decisions that directly affect the packet path and real-time traffic behavior must be executed locally. Latency sensitivity and privacy considerations therefore make on-device execution essential for these functions.

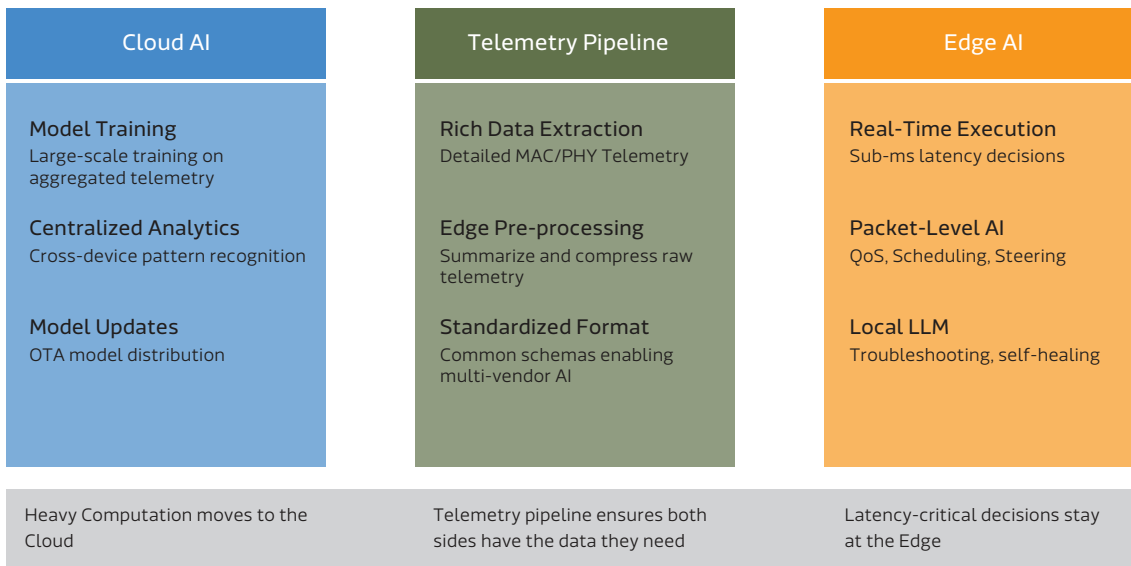


Figure 1: The general architectural principle

5.1 Edge AI (On-AP)

Edge AI runs directly on the access point, without needing the cloud. It is used for real-time decisions that cannot tolerate the round-trip latency of a cloud-based service.

Typical edge AI workloads include:

- Real-time anomaly detection: Small ML models that watch radio telemetry for interference, rogue APs, or unusual client behavior will instantly trigger reactions.
- Traffic classification and QoS: The AP intelligently identifies application types (video, gaming, bulk download) and applies QoS policies by itself, instead of needing user input.
- Local SLM inference: Small language models that can answer basic natural language questions about AP status, analyze local logs, and create short diagnostic summaries.

Edge AI needs dedicated compute (such as an NPU), enough memory for models and inference, and fast access to detailed radio telemetry. Because APs are limited by power, thermals, and cost, efficient silicon design is crucial.

5.2 Cloud AI

Cloud AI supports workloads that need large compute resources, global visibility, or models too big to run on the AP.

Cloud AI workloads include:

- Model training and refinement: Training and improving models using anonymized, aggregated telemetry from thousands of APs to analyse deployment-specific patterns. Private-cloud deployment should be available for customers with regulatory requirements.
- Enterprise-wide analytics: Cross-site comparisons, long-term trends, capacity planning, and predictive maintenance.
- Complex reasoning: Using large language models for advanced diagnostics, multi-site correlation, and natural language reports.
- Model distribution: Sending updated edge models over the air to the deployed APs so their ondevice AI keeps improving.

5.3 The Telemetry Pipeline

The telemetry pipeline connects edge and cloud and is often the most critical design element. It must support:

- Rich data extraction: Detailed telemetry from MAC and PHY layers, including channel conditions, client signal quality, frame statistics, interference, and environment measurements. Better telemetry enables better AI, but strong privacy and security mechanisms are also fundamental to building trust in buyers/operators. Telemetry data should be anonymized and aggregated on the edge before leaving the device. Cloud data isolation per customer using private-cloud and on premises options.
- Edge pre-processing: Summarizing and compressing raw telemetry on the AP before sending it to the cloud, to reduce bandwidth while keeping key information.
- Standardized formats: Common telemetry schemas to support multi-vendor AI systems. This is an active topic in the standards group.

With this architecture, latency-critical decisions stay at the edge, heavy computation moves to the cloud, and the telemetry pipeline ensures both sides have the data they need.

6. Closed-Loop Intelligence: The Agentic AI Paradigm

The most transformative application of AI in Wi-Fi is the transition from open-loop advisory systems, where AI provides recommendations that humans must interpret and act upon, to closed-loop autonomous systems that operate with minimal human intervention.

The Intent → Execution → Feedback Loop

A closed-loop AI system operates through three phases:

- Intent: A high-level objective is defined, either by human input (natural language command, policy configuration) or by the AI system itself based on observed network conditions. Example: "Prioritize video conferencing quality on the 3rd floor during business hours."
- Execution: The AI reasoning engine decompiles the intent into specific, concrete actions: adjusting QoS parameters, modifying channel assignments, implementing EDCA access categories, or applying client steering policies. These actions are executed through the AP's configuration APIs.
- Feedback: After execution, the system monitors telemetry metrics to verify that the intended outcome was achieved. Did video conferencing latency decrease? Did jitter improve? If the observed metrics do not improve or degrade, the system can autonomously roll back the change and attempt an alternative approach.

Agentic Frameworks for Network Intelligence

The Agentic AI paradigm provides a practical framework for implementing closed-loop intelligence. In this architecture, a language model serves as the reasoning engine, equipped with a set of “tools”, structured API calls that it can invoke to observe network state, execute configuration changes, and retrieve telemetry data.

Protocol frameworks such as Model Context Protocol (MCP) illustrate how AI systems and network devices can communicate and work together in a structured and interoperable manner. MCP provides a standard way for the AI to use different tools, collect telemetry data, make configuration changes, and handle responses. An MCP-enabled Wi-Fi Access Point can stream client activity, RF environment data, and usage statistics as telemetry to an MCP-aware cloud analytics platform. By implementing MCP endpoints or agents on an AP, the AP can act as a rich telemetry data source for network management and AI analytics systems. MCP can enable seamless integration of both cloud-based management APIs for business-wide operations, and local AP interfaces for edge-autonomous decisions, making it possible to connect the reasoning engine directly to access points as rich telemetry sources.

The closed-loop paradigm fundamentally changes the value proposition of AI in Wi-Fi networks. Rather than simply surfacing insights that require human action, the system becomes a self-optimizing network element that continuously improves its own performance.

7. Use Cases and Customer Impact

AI-driven intelligence enables distinct value propositions across consumer and enterprise Wi-Fi segments. The following use cases are representative examples (not an exhaustive list) illustrating how AI capabilities described earlier translate into practical value

Consumer

- **Intelligent QoS:** After classifying the traffic, the AP can identify which stations or flows require higher priority. Unlike static QoS rules, AI-driven prioritization automatically adapts to changing usage patterns without manual tuning. By leveraging on-device AI, the system can derive optimal QoS configurations based on environmental conditions and telemetry data, ensuring that latency-sensitive traffic, such as video calls and online gaming, consistently receives the resources it needs for a better user experience.
- **Self-healing home networks:** When a consumer reports “my Wi-Fi is slow,” the AP’s on-device AI can autonomously diagnose the issue (interference from a neighboring network, suboptimal channel selection), take corrective action, and confirm resolution without requiring the user to contact technical support.
- **Energy-aware Wi-Fi operation:** Reduces radio activity at night or when nobody is home.
- **Wi-Fi sensing and presence-aware features:** Uses changes in Wi-Fi signals to detect motion or presence, enabling occupancy-based automations.
- **Natural language Wi-Fi assistant:** Lets users ask questions and give commands in plain language such as “Why is my Wi-Fi slow?” or “Make my work laptop the priority for the next hour.”

Enterprise

- **AI-assisted troubleshooting:** When a help desk ticket reports connectivity issues in a conference room, AI can correlate the complaint with radio telemetry, client event logs, and environmental data to identify root causes in minutes rather than hours. Natural language reporting enables IT staff to perform diagnostics that previously required RF experts
- **Anomaly detection:** AI models baseline normal behavior, then detect and report hardware, system, and Wi-Fi anomalies, such as rogue devices, interference, or abnormal traffic patterns, by monitoring key metrics and events to improve reliability and security.

- Autonomous RF optimization: Continuous AI-driven optimization of channel assignments, transmit power levels, and spatial reuse parameters across a multi-AP deployment, adapting to changing occupancy patterns, new interference sources, and evolving traffic demands.
- AP as a Platform: Perhaps most significantly, AI-capable access points create the foundation for a platform model. Third-party applications can leverage the AP's edge compute, radio telemetry, and AI inference capabilities to deliver services that go beyond traditional connectivity.

AI's most immediate impact is in transforming how different user personas interact with Wi-Fi networks. Natural language interfaces, powered by on-device or cloud SLMs, significantly reduce operational complexity for network management.

End Users

For consumers and enterprise end users, AI eliminates the black box of Wi-Fi troubleshooting. Instead of cryptic signal strength numbers and channel utilization percentages, users can ask natural language questions and receive actionable responses:

"Why is my Zoom call dropping?"

An AI-enabled AP can correlate the timing of video quality degradation with radio events: a neighboring AP switched channels causing co-channel interference, or the client's signal strength dropped below a threshold as the user moved. The system can explain the cause in plain language and take corrective action (steer the client to a less congested band or channel).

IT Administrators

For enterprise IT teams, AI transforms the daily workflow from reactive ticket management to proactive network oversight:

"Show me the top 10 clients with the worst experience scores across Building A this week."

Rather than navigating through multiple management console views and manually correlating data, the administrator receives a prioritized list with root-cause annotations, recommended actions, and the option to execute fixes directly through natural language commands.

Operators and MSPs

For managed service providers overseeing thousands of APs across hundreds of customer sites, AI provides the scalability that human-only operations cannot:

"Generate a weekly health report for all managed sites, highlighting any SLA, and capacity concerns."

AI aggregates fleet-wide telemetry identifies patterns across sites and produces formatted reports that would previously require hours of manual analysis. This operational efficiency directly translates to margin improvement for the service provider.

The common thread across all personas is that AI lowers the barrier between intent and outcome. Users no longer need to understand the technical mechanics of Wi-Fi to get value from their network. They simply express what they want, and the AI system handles the translation to network operations.

8. MediaTek Filogic AI in Practice

The use cases described above are not abstract aspirations. MediaTek has implemented and characterized two AI-driven Wi-Fi 8 platform features that directly translate the architectural principles in this white paper into measurable outcomes: Intelligent Quality of Experience and adaptive AI-assisted power saving. Both features run on-device using edge compute and rich MAC/PHY telemetry; both of which are representative of the closed-loop intelligence model.

8.1 Intelligent Quality of Experience

MediaTek’s Intelligent Quality of Experience capability addresses three common sources of Wi-Fi degradation - congestion, interference, and weak signal, using a combination of adaptive algorithms and ML running on edge AI compute using the MediaTek Wi-Fi SoC. Figure 2 illustrates the overall architecture.

The system includes four capabilities:

1) an intelligent scheduler that dynamically controls airtime allocation and traffic priority, 2) active queue management (AQM) that protects latency-sensitive applications under heavy load, 3) an airtime efficiency engine that uses a 5th receive antenna for continuous background channel scanning and automatic channel switching, and 4) coverage extension for stations with weak or obstructed signals. These capabilities work together to continuously adapt the AP’s behavior to changing network conditions, replacing static default configurations with real-time, environment-aware optimization.

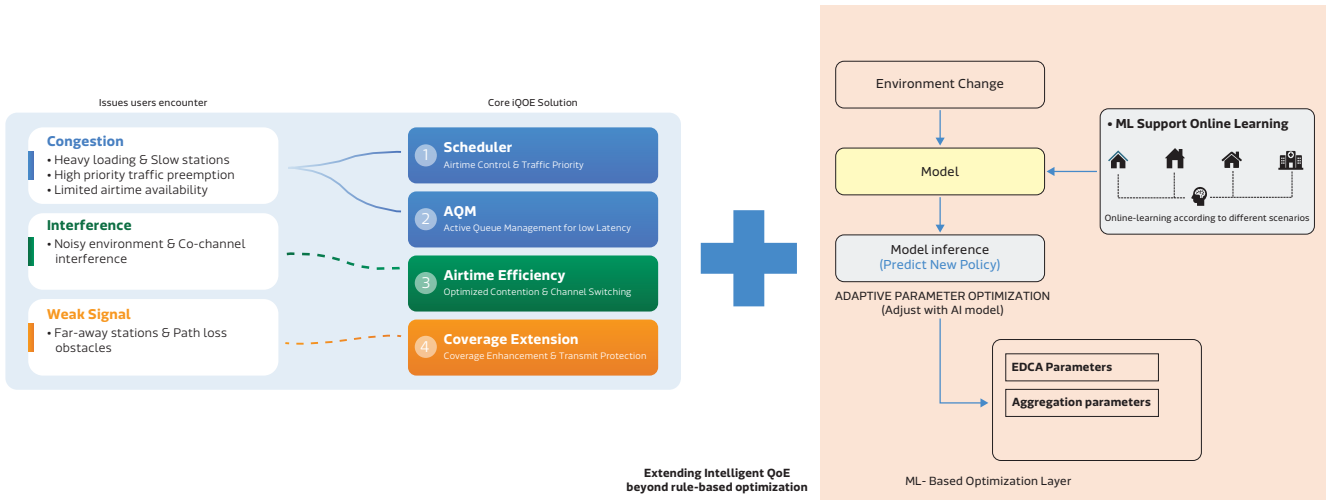


Figure 2: Intelligent QoE Solution

Complementing these capabilities, MediaTek is developing an adaptive parameter optimization layer driven by ML. This ML-based capability focuses on optimizing EDCA and aggregation parameters across varying environmental conditions. Rather than relying on fixed parameter tables, it automatically learns the best contention window parameters, TXOP, and aggregation settings for each deployment by evaluating different configurations against live network telemetry. Whether the environment is dominated by dense client congestion, co-channel interference, or mixed traffic patterns, the model converges on parameter combinations tuned to the specific scenario, achieving gains that fixed or rule based approaches cannot reach.

iQoE Test Result(50% congestion)

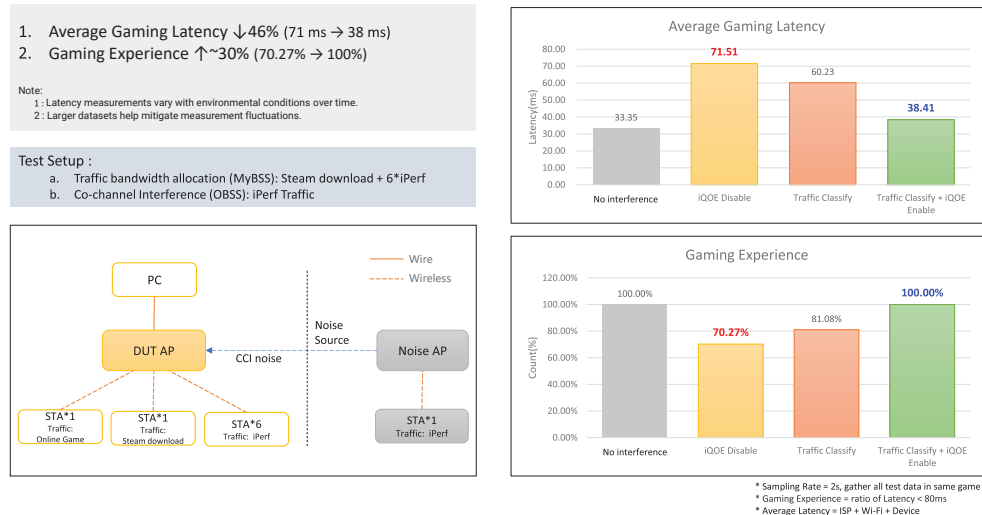


Figure 3: Gaming latency and experience improvement with Intelligent QoE enabled.

As shown in Figure 3, during congestion testing, which comprised of a Steam download, six iPerf flows, and a co-channel interferer consuming 50 percent of available airtime—the Intelligent QoE system with traffic classification reduced average gaming latency by 46 percent (from 71.51 ms to 38.41 ms) and improved the gaming experience metric from 70 percent to 100 percent. Game Experience is defined as the percentage of time during which end-to-end latency remains below 80 ms, a threshold commonly regarded as comfortable and acceptable for real-time gaming.

In separate testing, the ML-based adaptive parameter optimization reduced gaming latency by 30 to 40 percent. Together, the rule-based capabilities deliver measurable quality of experience improvements today, with ML-based optimization extending these gains further as it reaches production maturity.

8.2 AI-Assisted Adaptive Power Saving with Intelligent Power Management

Energy Saving - Intelligent Power Management for ultra power saving 24/7

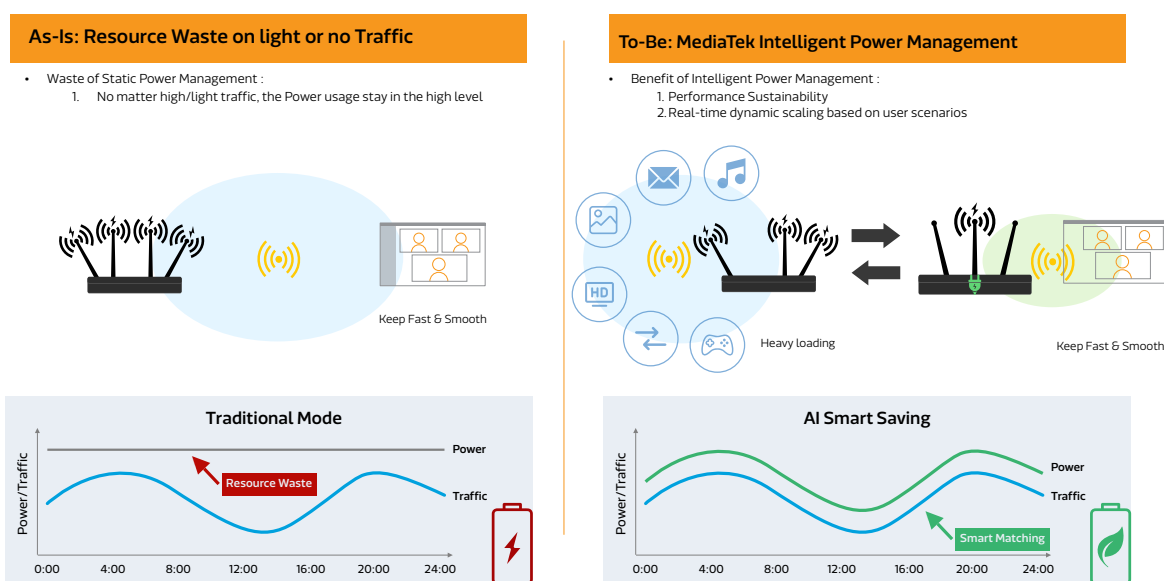


Figure 4: Dynamic power scaling matches real-traffic demand

Conventional access points often keep radio subsystems operating in a high-power state regardless of actual traffic demand, leading to continuous energy consumption during extended periods of light network activity. This approach overlooks the predictable daily structure of real-world traffic, where streaming, gaming, and large data transfers cluster into defined peak windows, while overnight and mid-day periods are dominated by low-intensity background usage. Static power management therefore provisions for worst-case load on a 24/7 basis. MediaTek’s AI-assisted adaptive power saving addresses this inefficiency by applying on-device inference to learn the access point’s traffic profile and dynamically scale power consumption in response to real-time demand, sustaining performance during peak periods while significantly reducing energy draw during idle windows, as illustrated in Figure 4.

Energy Saving - Incredible energy saving gain in light traffic environment

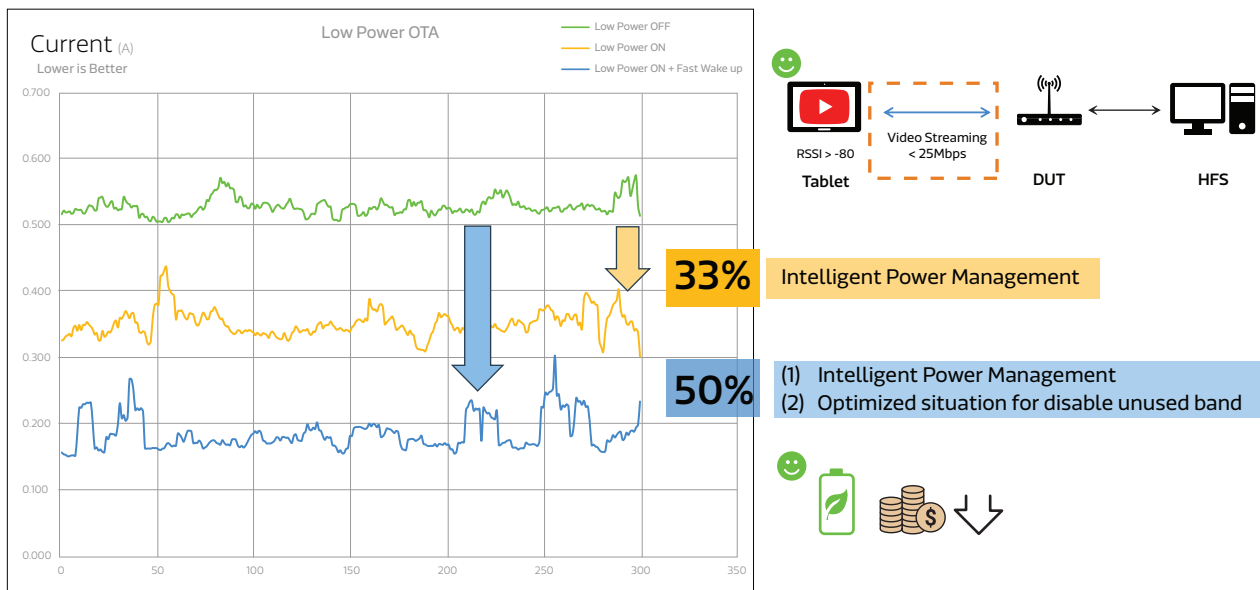


Figure 5: Current reduction with Intelligent Power Management

In an OTA video streaming test (RSSI > -80 dBm, sustained throughput under 25 Mbps), enabling Intelligent Power Management on the Filogic platform reduced device current draw by approximately 33 percent compared to the baseline static power profile. When combined with adaptive disabling of unused radio bands, total current reduction reached approximately 50 percent, as shown in Figure 5. Importantly, a fast wake-up latency was preserved throughout, ensuring that user-perceived responsiveness was unaffected. These results demonstrate that AI-driven power management is not a performance tradeoff, but rather a structural efficiency improvement enabled by closed-loop telemetry and on-device inference operating continuously across the full 24-hour traffic cycle.

What These Results Show

Across both features, the common pattern is that on-device AI extracts measurable value from telemetry that already exists at the MAC and PHY layers, which static configurations cannot exploit. Latency, interference resilience, user experience, and power efficiency improvements are not achieved by adding throughput; they are achieved by adding intelligence.

9. Enabling AI at the Silicon Level

The AI capabilities described in this whitepaper are ultimately enabled or constrained by the access point’s chipset. The silicon layer determines the ceiling for edge AI performance, telemetry richness, and power efficiency.

Neural Processing Unit (NPU)

A dedicated NPU integrated into the Wi-Fi SoC provides hardware-accelerated inference for neural network workloads. Unlike running AI models on general-purpose CPU cores which consumes application processing headroom and increases power draw, a purpose-built NPU delivers higher inference throughput per watt, with predictable latency characteristics that do not compete with the AP’s primary networking functions.

Telemetry Hooks

AI is only as good as the data it can access. Chipset-level telemetry hooks at the MAC and PHY layers provide raw data for AI intelligence: per-client signal quality metrics, per-frame transmission statistics, channel occupancy measurements, interference event records, and radio environment characterizations. The depth, granularity, and accessibility of these

telemetry hooks vary significantly across chipset platforms. A platform that exposes rich, structured telemetry at high sampling rates provides a fundamentally better substrate for AI than one that offers only aggregated, coarse-grained statistics.

Memory Subsystem Design

The memory subsystem is a critical enabler for AI workloads. The choice of memory technology - total capacity, bus width, and memory controller design collectively determines what AI models can be hosted and how fast they can run. The significance of this will be elaborated in subsequent sections.

10. Beyond TOPS: Memory, Bandwidth, and Silicon Enablement

As AI becomes a key differentiator in Wi-Fi access points, evaluating a platform's AI capability requires looking beyond a single metric such as TOPS (Tera Operations Per Second), which alone can be misleading. Real-world AI performance is determined by how silicon integrates compute, memory hierarchy, bus fabric, and software. Memory capacity, sustained memory bandwidth, and NPU throughput must be considered together, with platform-level integration as the deciding factor.

10.1 The TOPS Metric: Useful but Incomplete

TOPS measures the peak compute throughput of an NPU: how many operations it can do per second.

On an access point, a high-TOPS NPU can sit idle if the surrounding silicon - memory controllers, on-chip SRAM, bus fabric, and software stack - cannot feed it with sufficient data. A well-integrated platform with balanced SoC including appropriately sized NPU can deliver better real-world inference performance and latency, than one pursuing headline TOPS alone.

10.2 Memory Capacity: What Models Can You Run?

For edge processing, memory capacity decides what models can fit on the AP. Key components include:

- **Model weights:** A 1-billion parameter model quantized to INT8 (8-bit integer) precision requires approximately 1 GB of storage for weights alone. A 500M-parameter model at INT8 requires roughly 500 MB.
- **KV cache:** During inference, language models maintain a key-value cache that grows with sequence length. For a typical SLM, this can add 100–300 MB of memory demand depending on context window size.
- **Activation memory:** Intermediate computation results during inference require additional memory, typically 50–200 MB depending on model architecture.
- **System overhead:** The Wi-Fi stack, operating system, and application software on the AP have their own memory requirements, which must be satisfied concurrently with AI workloads.

Combined, running even a modest SLM plus the normal AP software can require 1.5–2 GB or more of DRAM. Memory generation and total capacity directly limit how large and how many models the AP can support.

10.3 Memory Bandwidth: How Fast Can Models Run?

For LLM/SLM inference, sustained bandwidth between the NPU and memory is often the main bottleneck. Generating each token requires streaming model weights from DRAM, so performance is usually limited by bandwidth, not compute.

A practical rule of thumb for memory-bound inference workloads is:

$$\text{Tokens/second} \approx \text{Effective Bandwidth (GB/s)} \div \text{Model Size (GB)}$$

Effective bandwidth is the operative term. It depends on memory controller design, bus width, prefetch and compression logic, and software scheduling - all silicon and platform choices.

Example: A 500M-parameter INT8 model (0.5 GB) on a memory subsystem delivering 12.8 GB/s effective bandwidth would yield approximately 25 tokens/second which is sufficient for interactive natural language responses.

The same model on a system delivering only 6.4 GB/s would generate approximately 12 tokens/second which is noticeably slower and potentially inadequate for real-time interaction.

For LLM/SLM inference workloads, sustained bandwidth is the performance ceiling, and the silicon design determines how close a platform gets to that ceiling.

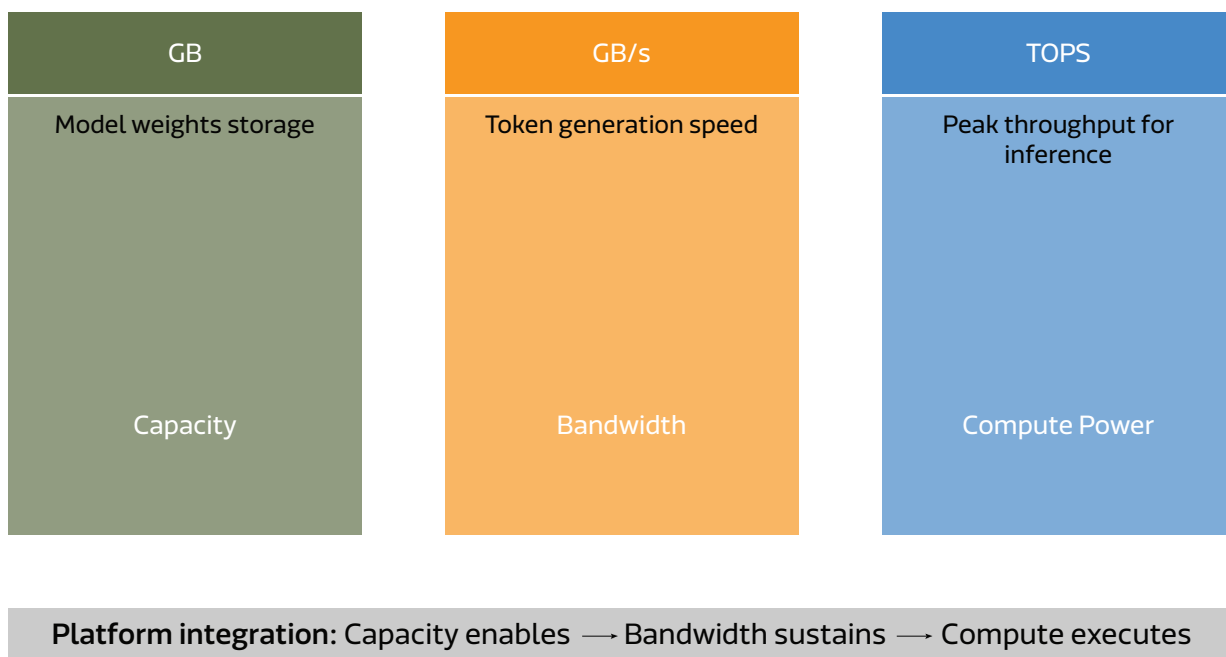


Figure 6: Three Pillars of Edge AI compute

The three pillars must be designed together. Capacity sets what models can run, bandwidth sets how fast they run, and compute sets the ceiling - but it is the SoC's integration of all three that determines realworld performance.

10.4 A Framework for Evaluation

When evaluating the AI readiness of an access point platform, network architects and product managers should consider a holistic set of metrics rather than TOPS alone:

Metric	What It Determines	What to Ask Vendors
NPU TOPS	Peak compute throughput for AI workloads	What inference frameworks are supported? What is the sustained (not peak) throughput?
DRAM Capacity	Maximum model size that can be hosted alongside the AP software stack	Total DRAM? How much is available after OS and Wi-Fi stack reservation?
Memory Bandwidth	Token generation speed for LLM/SLM inference	Memory generation and capacity? Bus width? Effective sustained bandwidth under load?
Telemetry	Richness and granularity of radio data available for AI consumption	What MAC/PHY metrics are exposed?

Table 3: AI Readiness Evaluation Framework

11. Challenges and Considerations

It is important to acknowledge the major challenges that still exist in putting AI into Wi-Fi access points. These are engineering problems that can be solved over time, and they represent active engineering frontiers that shape today's platform design and architectural choices.

- **Power and Thermal:** Every extra watt used by the NPU competes with radios, CPUs, and other subsystems and must be safely dissipated as heat. This makes highly efficient NPU designs, fine-grained power management, and thermally aware system architectures essential to running meaningful AI workloads on the AP.
- **Model Lifecycle:** Running AI at scale means managing models across thousands of APs in the field. This requires strong infrastructure for version control, over-the-air (OTA) distribution, safe rollback if a new model misbehaves, and continuous monitoring for performance degradation. Without this lifecycle management, AI features become challenging to maintain.
- **Privacy and Interoperability:** Wi-Fi telemetry can reveal sensitive information about devices, users, and usage patterns, so strict data governance is required: clear rules for collection, anonymization, retention, and consent. At the same time, multi-vendor networks need open, standardized AI and telemetry interfaces, so that AI solutions can work across different vendors' hardware and software.
- **Edge vs. Cloud Trade-offs:** The best split of AI workloads between edge and cloud depends on the specific deployment: link reliability, latency needs, privacy rules, operational complexity, and cost. There is no single ideal architecture. Platforms that allow flexible placement of workloads will fit more use cases.
- **Standardization:** The IEEE AI/ML Topic Interest Group and related work in bodies such as the WiFi Alliance (WFA) and Wireless Broadband Alliance (WBA) are still in early stages toward aligning the industry on AI and Wi-Fi integration. Much work remains to standardize telemetry formats, AI-driven coordination protocols, and interoperability test methods. Until standards mature, early AI solutions will include proprietary pieces that can limit interoperability.

12. Conclusion

Wi-Fi is entering a new phase of evolution, from a connectivity utility to an intelligent infrastructure platform. This transformation is driven by a clear need (network complexity), enabled by maturing technology (efficient AI silicon, practical small language models, rich radio telemetry), and validated by industry direction (the standards group focus, vendor ecosystem momentum, customer demand).

Several key insights emerge from this analysis:

- AI integration in Wi-Fi is inherently a systems-level design effort, requiring co-optimizing NPU compute, memory subsystem, telemetry infrastructure, and power management.
- The three-tier architecture (edge, cloud, telemetry pipeline) provides an effective balance of latency, capability, and cost for practical AI deployment.

Closed-loop intelligence, where AI autonomously observes, decides, acts, and validates, represents the highest-value application of AI in Wi-Fi, transforming reactive troubleshooting into proactive, self-driving network optimization. As the world's largest supplier of Wi-Fi solutions, MediaTek is committed to leading this transformation through continued investment in AI-enabled silicon, active participation in standards development, and deep collaboration with ecosystem partners across the consumer, enterprise, and broadband segments.

The access point of tomorrow will extend beyond connectivity. It will understand network conditions, anticipate demands, and act autonomously to deliver the reliable, intelligent connectivity that the modern connected world demands.

13. MediaTek in the Wi-Fi Industry

MediaTek is the world's largest supplier of Wi-Fi solutions, including standalone networking products such as routers, repeaters, and mesh access points, and devices with embedded Wi-Fi connectivity such as smartphones, tablets, TVs, IoT, smart home devices, PCs and laptops, games consoles, and many others.

Besides delivering high performance and low power integrated solutions to these platforms, MediaTek is actively participating in IEEE and Wi-Fi Alliance certification development to ensure top performance and industry interoperability. Some recent examples include selection of MediaTek's Filogic platforms as WiFi 6E and Wi-Fi 6 R2 test bed devices. With Wi-Fi 7, and soon Wi-Fi 8, MediaTek continues to contribute technical expertise and knowledge of diverse market segment standards for improved Wi-Fi performance in daily applications.

14. Acknowledgements

Author:

- **Shahnawaz Siraj** (Director of Technology, intelligent Connectivity Business)

Contributor:

- **Yaling Hu** (Director of intelligent Connectivity Business)
- **Terry Chen** (Technical Marketing Manager of intelligent Connectivity Business)
- **Bruce Chuang** (Technical Marketing Manager of intelligent Connectivity Business)
- **Stone Zhang** (Sr. Engineer of intelligent Connectivity Business)

Editor:

- **James Chiang** (Sr. Technical Marketing Manager of intelligent Connectivity Business)