Micron 3610 Enables High-Performance, Energy-Efficient AI Model Loading

Authors: Ranjith Kumar Nagisetty; Sandeep Kumar Goswami; Pradeep Kumar Jilagam; Ranjith Kumar Ravuri; Rui Zhou; Sharath Chandra Ambula

Introduction

On-device AI introduces a storage-gated phase in the user experience. Before inference can begin, multigigabyte model weights must be transferred from the SSD into system DRAM. This interval—referred to as the cold-start model load window—is predominantly read-intensive and affects perceived responsiveness.

Model Load Phase: Storage Requirements

As AI models grow in size and complexity, the storage subsystem faces increasing demand during the model load phase. The SSD must sustain high read bandwidth across mixed I/O patterns, including large sequential streams for model weights and smaller, more random reads for metadata and indices. Because this phase fully precedes inference, its efficiency directly determines the earliest possible user response.

Maintaining consistent throughput across varying block sizes is therefore critical. Larger model footprints and diverse file layouts further stress the SSD's ability to deliver data efficiently, making storage performance a key factor in optimizing on-device AI workloads.

Client and Edge Platform Considerations

Client and edge AI systems operate under strict power and thermal constraints. In configurations with limited DRAM capacity, paging and re-fetch behavior can extend reliance on SSD latency beyond a single cold start. Storage solutions that complete the load phase quickly—and with high performance per watt—help preserve energy and thermal headroom for the compute-bound inference phase.

Study Scope

This study evaluates the Micron® 3610 PCIe® Gen5 NVMe™ SSD against evaluated PCIe Gen4 SSDs using synthetic benchmarks and real LLM cold-start load tests. The analysis focuses on model load time and energy consumption during the storage-gated window.

Key Takeaways

55%

Faster read performance

The Micron 3610 SSD delivers *up to 55%* higher sequential read throughput than PCIe Gen4 SSDs, enabling significantly faster loading of large AI model files, high-resolution game assets, and bulk dataset transfers.

AI model load times improved by up to 25% allowing applications such as local LLMs, on-device copilots, and multimodal inference engines to become ready for use much sooner.

34%

Write performance improvement

The 2TB capacity variant provides *up to 34%* higher sequential write throughput than the 1TB model, accelerating real-world tasks such as saving checkpoints, exporting large files, caching datasets, and handling high-bandwidth content creation workloads.

81%

Random read performance improvement

With *up to 81%* stronger random read performance, the 2TB model improves responsiveness in workloads dependent on rapid small-block access—such as application startup, metadata lookups, KV-cache paging, and AI inference routines that stream parameter blocks on demand.

29%

Lower power consumption

Performance per watt improves by *up to 29%* versus Gen4 TLC SSDs, reducing storage power consumption during continuous AI and compute-heavy workloads. This translates into longer battery life, cooler device operation, and more sustained CPU/GPU/NPU performance without throttling.

Ideal for AI PCs, gaming, and client edge systems

Key Findings

- In the evaluated tests, the Micron 3610 Gen5 NVMe SSD reduced cold-start model-load time compared with the evaluated Gen4 SSDs, shortening the storage-bound portion of the workload.
- Across CrystalDiskMark (CDM), IOmeter, and LLM cold-start scenarios, the 3610 demonstrated higher read performance and improved performance per watt versus the evaluated Gen4 drives.
- These gains translated into shorter model-load windows and lower or comparable energy consumption during the load phase, depending on the comparison drive.

Technical Description

The Micron 3610 is a PCIe Gen5 NVMe client SSD based on Micron G9 QLC NAND, with controller and firmware optimizations targeted at read-dominant, mixed-I/O workloads common in on-device AI. At capacities suitable for storing multiple local models and datasets, the 3610 delivered improved read performance and efficiency in the evaluated scenarios, making it well-suited for AI PCs and edge platforms.

NAND Technology Background: SLC vs TLC vs QLC

NAND flash stores information by programming multiple threshold-voltage levels within each cell. The bits per cell (SLC/TLC/QLC) drive trade-offs among performance, endurance, density, and cost per bit.

- **Single-Level Cell (SLC)** stores 1 bit per cell and uses 2 voltage levels. It delivers the lowest latency, highest endurance, and best reliability. However, its premium cost per bit limits its use in modern SSDs, where it typically appears as a small, fast write buffer rather than the primary storage medium.
- **Triple-Level Cell (TLC)** stores 3 bits per cell across 8 voltage levels and offers a balanced combination of performance, endurance, and cost efficiency. TLC remains the choice in client and data center SSDs, providing higher P/E cycle counts and lower read latency than older QLC generations.
- **Quad-Level Cell (QLC)** stores 4 bits per cell using 16 voltage levels, enabling high density and lower cost per bit. This can be beneficial for storing large AI models, embeddings, and datasets locally. Compared with SLC/TLC, QLC typically has lower endurance and can have higher read/program complexity; SSD controller/firmware and NAND-generation improvements can mitigate these effects depending on the workload and implementation.

These characteristics make QLC-based SSDs a cost- and capacity-optimized option for read-dominant AI workloads where device-local storage footprints continue to expand. In the Micron 3610, QLC's capacity and cost advantages are paired with PCIe Gen5 bandwidth and controller/firmware features intended to improve read performance and efficiency. In the evaluated tests, this combination supported shorter model-load time and lower energy during the load window versus the evaluated Gen4 SSDs, supporting QLC as a practical foundation for scalable, on-device AI storage.

Micron G9 QLC NAND — Architectural Advantages

- The Micron® 3610 NVMe™ SSD is the industry's first PCIe® Gen5 QLC client SSD, powered by Micron G9 NAND. This latest generation brings multiple architectural enhancements that strengthen both performance and value:
- **High-bandwidth interface (ONFI 5.0 @ 3.6 GT/s)**
  The faster interface improves internal data movement, enabling higher effective read throughput—critical for streaming large AI model weights.
- **Improved program/erase behavior**
  Enhancements to cell design and P/E algorithms help reduce write amplification, allowing more consistent write performance even under mixed workloads.
- **Optimized read-latency characteristics for AI-dominant block sizes**
  G9 QLC is tuned for the 128KB–2MB block ranges common in AI tensor loading, contributing directly to faster cold model initialization.

PCIe Gen5 Interface Benefits

The PCIe Gen5 ×4 interface doubles the theoretical bandwidth of Gen4, and when combined with Micron's controller and G9 NAND, it delivers meaningful real-world benefits for AI workloads:

- **Significantly higher sequential read/write speeds**
  Faster transfer of multigigabyte model weights reduces storage-bound delays during cold starts.
- **Better queue depth scaling for multithreaded applications**
  Modern AI frameworks spawn multiple I/O threads; Gen5 improves utilization across these parallel flows.
- **Enhanced parallelism for hybrid AI tasks**
  Workloads that interleave sequential weight streaming with smaller random reads (KV-cache activity, attention block loads) benefit from higher available bandwidth.

Together, these improvements help the 3610 sustain high performance across the diverse access patterns common in AI inference pipelines.

Firmware Innovations for AI Workloads

Client-side AI I/O is typically **read dominant** (streaming weight files), punctuated by **bursty writes** (caches, checkpoints), and it uses a **mix of block sizes** (large sequential transfers plus smaller random metadata reads). The 3610 firmware is tuned to keep this load window short and consistent by reducing internal work that does not contribute to host visible progress.

- **Adaptive write handling for bursty AI writes**
  During short write bursts (logs, cache updates, checkpoint artifacts), adaptive write handling adapts buffering and placement to limit write amplification. Outcome: less background flash movement competing with reads during subsequent model loads.
- **Host memory buffer for read heavy, metadata sensitive access**
  AI initialization mixes large reads with many smaller lookups (file system metadata, tensor index structures). Caching mapping structures in host DRAM reduces flash reads for metadata, lowering tail latency and improving consistency for small random reads that interleave with weight streaming.
- **Queue handling for parallel model-load streams**
  Frameworks often issue concurrent reads across threads (multiple files, layers, or prefetch streams). Queue scheduling prioritizes read completion and maintains throughput under concurrency, reducing stalls that would otherwise extend the storage-bound window.

Micron's G9 QLC NAND builds on multiple generations of QLC experience, combining high density 2Tb dies with a 3.6 GT/s ONFI 5.0 interface and the PCIe Gen5 architecture of the 3610. Together, these platform level and firmware level innovations enable the driver to deliver the cost and capacity benefits of QLC while achieving the performance, efficiency, and responsiveness required for mainstream AI PC workloads.

Table 1: Key Architectural Attributes of Evaluated SSDs

Feature	3610 QLC	3500 TLC	2600 QLC
Interface	PCIe Gen 5 x4	PCIe Gen4 x 4	PCIe Gen4 x 4
Form Factor	Mx2 22 x 80	Mx2 22 x 80	Mx2 22 x 80
NAND	Micron G9 QLC	Micron G9 QLC	Micron G9 QLC
Capacity	1TB/2TB	1TB/2TB	1TB

Note. Table 1 summarizes the key architectural attributes of the SSDs evaluated in this study, providing a baseline for the performance, efficiency, and workload comparisons presented in the following sections.

Methodology

To understand how storage influences AI workload performance in client systems, we adopt a structured methodology that evaluates both device-level characteristics and real AI workload behavior. Synthetic benchmarks provide controlled measurements of throughput, latency, and efficiency, while real LLM workloads reveal how these characteristics translate into user-visible responsiveness, such as time-to-first token and model initialization time.

This dual-layer approach ensures that our results reflect not only what the SSD can do under idealized conditions, but also how it behaves when executing the types of AI tasks users run on an AI PC.

Benchmark-Driven Device Characterization

Synthetic benchmarks are the first step because they allow us to isolate and measure core SSD behaviors independent of other system components. These tests quantify:

- **Sequential throughput**
  Critical for streaming multi-gigabyte model weights into memory during cold starts.
- **Random I/O characteristics**
  Relevant for metadata fetches, KV-cache paging.
- **Block-size sensitivity**
  Essential for understanding performance across AI-dominant read sizes (128 KB – 2 MB).
- **Performance per watt**
  A key differentiator for edge and mobile AI devices with limited power budgets.

While these metrics define the capabilities of the storage device, they do not fully predict user experience in AI workloads. Therefore, we complement them with real AI model tests.

Real AI Workload Evaluation

To capture realistic access patterns, we evaluate the SSD using representative on-device AI models, including:

- Llama 3.1 8B
- Mistral 7B
- Bakllava 7B
- MiniCPM 8B

These models differ in tensor dimensions, graph complexity, and metadata structure, generating diverse read patterns during initialization. Although LLM inference is compute-bound once the model is loaded, cold starts are entirely storage-bound, making model load time the most direct indicator of how SSD performance affects AI responsiveness.

For clarity and focus, this paper includes results from:

- CDM for peak sequential/random behavior
- IOmeter for performance per watt
- LLM cold-start model load tests for real AI behavior

This combination allows us to correlate synthetic behavior with real user-facing outcomes.

Test Setup

All SSDs were tested on the same controlled platform:

- Intel Core Ultra 7 processor
- ASUS ProArt Z890 motherboard
- 64GB DDR5
- Windows 11

This configuration represents a typical AI PC, where CPUs, GPUs/iGPUs, and NPUs operate alongside storage to support high-volume model movement. Keeping the compute platform constant ensures that all differences in behavior originate from the SSDs themselves.

Benchmark Tools

Table 2: Benchmark and Analysis Tools

Benchmark	Definition
CDM	It is a synthetic storage benchmark tool used to measure peak read and write performance.
IOmeter	IOmeter is a professional I/O subsystem measurement and workload characterization tool used to evaluate IOPS, bandwidth, and latency of storage.

Note. Table 2 summarizes the benchmark and analysis tools used in this study, along with the specific aspects of SSD behavior each tool was selected to measure.

Table 3. Performance Analysis and Tracing Tools Used for AI Workload Characterization

Tools	Usage
Diskmon	• Captures kernel-level disk I/O events using Windows event tracing • Displays read/write operations, offsets, and execution timing
Procmon	Built-in Windows monitoring tool that provides counter-based measurement of system performance in real time or over extended periods. It collects and logs metrics related to CPU, memory, disk, network, and process behavior.
Windows Performance Analyzer (WPA)	WPA is an advanced performance analysis tool included in the windows performance toolkit. It analyzes the event tracing for windows trace files recorded by the windows performance recorder.

Note. This table summarizes the system-level tracing and analysis tools used to observe I/O behavior, access patterns, and execution dynamics during AI model initialization and inference. These tools enable correlation between storage activity and CPU/GPU/NPU utilization under real device AI workloads.

Case Studies and Analysis

SSD Performance Comparison Across Generations

CDM 9 results indicate that, in the evaluated tests, the Micron 3610 Gen5 QLC SSD demonstrates higher sequential and random read/write performance versus the evaluated Gen4 TLC and Gen4 QLC SSDs. The 3610 shows up to ~55% higher sequential read bandwidth versus the evaluated Gen4 SSDs in CDM 9 testing, and up to ~34–42% higher sequential write bandwidth (depending on the compared drive and workload). These differences can reduce the time spent on the storage-bound critical path during AI model loading and initialization.

Figure 1: Gen5 vs. Gen4 SSD Performance — CDM 9

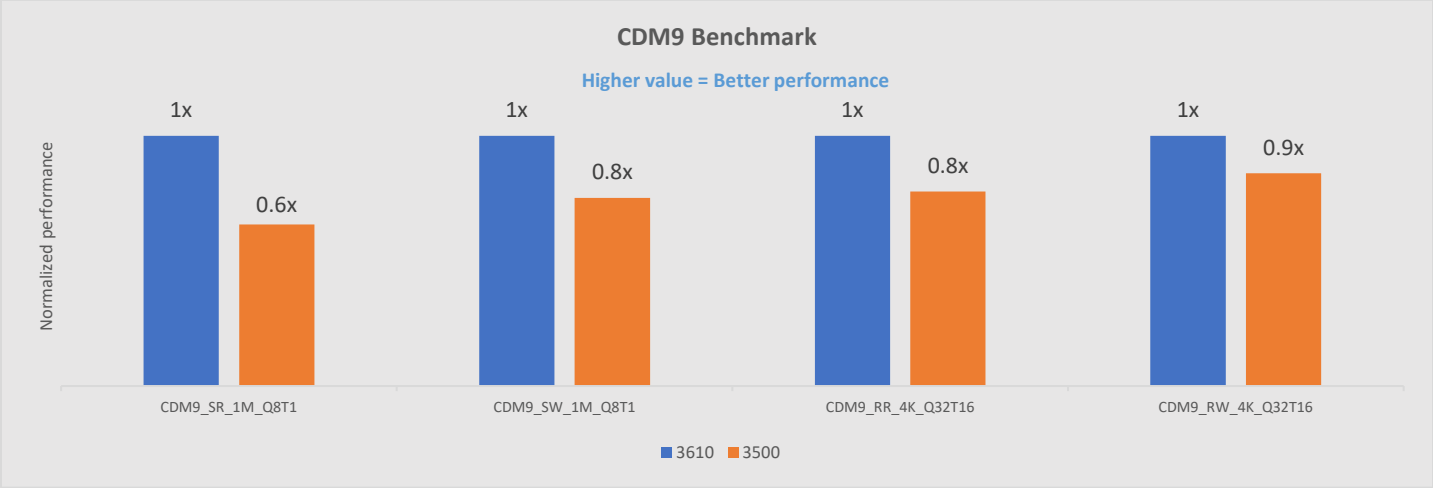
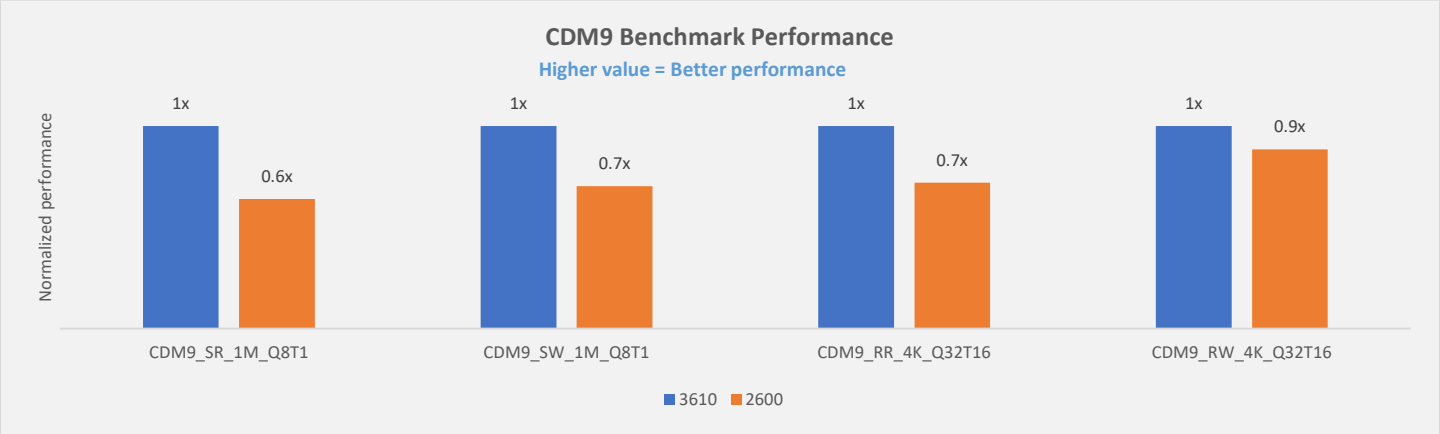


Figure 2: Gen5 vs. Gen4 SSD Performance — CDM 9



AI Model Load Time and Energy Comparison

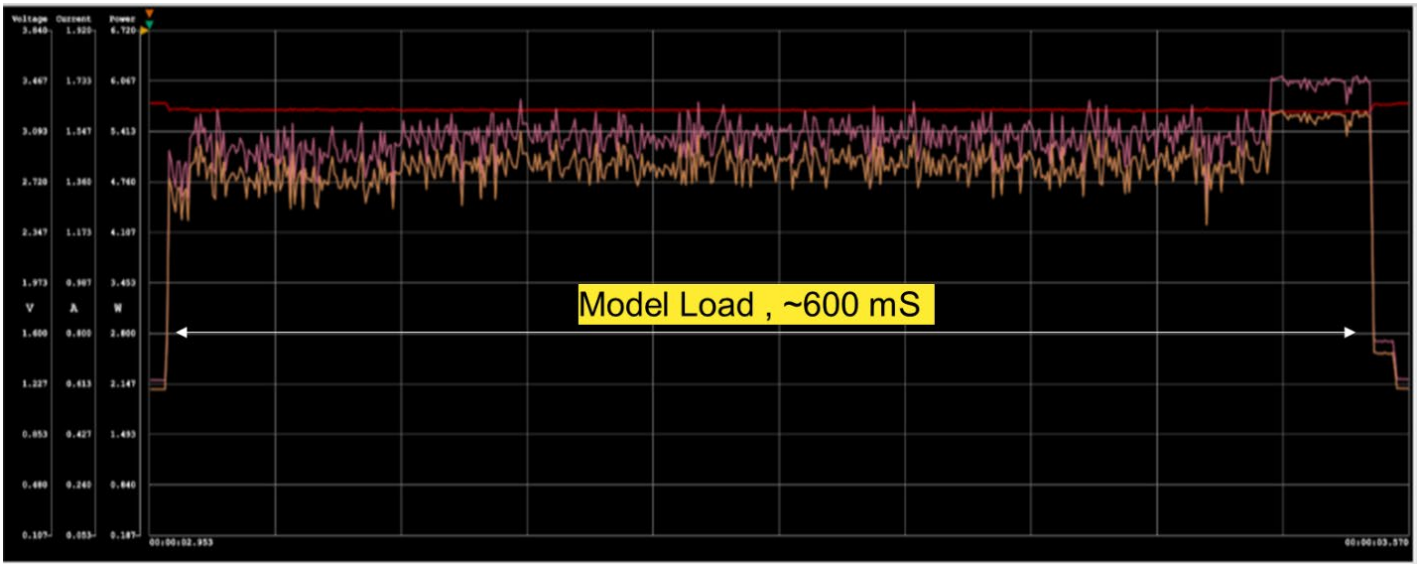
In the evaluated cold-start tests, the Gen4 SSDs required up to ~43% more time to reach a model-ready state at a similar measured energy level, resulting in faster user-perceived AI responsiveness in storage-gated scenarios for Micron 3610.

To understand the storage contribution to AI cold-start behavior, we first examine SSD power activity during model initialization. Figure 3 shows a time-aligned snapshot of input power measured across three SSD supply rails while loading an AI model from storage into DRAM. Monitoring multiple rails ensures that both controller and NAND-side activity are captured, providing a complete view of SSD energy behavior during the load phase.

The highlighted window (~600 ms) represents model loading. Power remains elevated and relatively steady, consistent with sustained high-bandwidth reads while weights are transferred from the SSD into DRAM. During this window, inference cannot begin.

Two points follow from this trace: (1) time to first token (TTFT) in cold-start scenarios is bounded by completion of the load window, and (2) energy during loading depends on both average power and load duration.

Figure 3: SSD Input Power Rail Activity During AI Model Loading



Snapshot of SSD input power during AI model loading, captured by monitoring three independent SSD input power supply rails. The highlighted interval (~600 ms) corresponds to the model load phase, during which model weights are

streamed from SSD into system memory. The sustained power draw across all monitored rails indicates a storage-dominated execution window, where inference cannot begin until model loading completes.

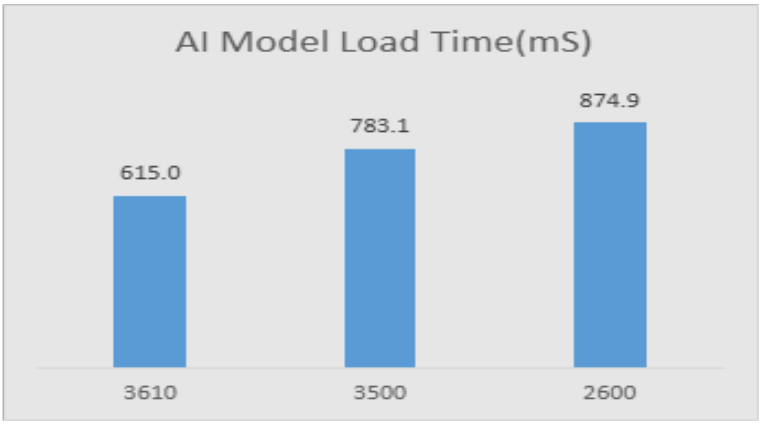
Takeaway: Because inference cannot begin until model loading completes, reducing the duration of this power-elevated window can improve TTFT and reduce total energy during initialization.

Building on the power-trace observations, Figure 4 quantifies model-load time across SSD generations. The Micron 3610 Gen5 QLC SSD completes model loading in approximately 616 ms, while the Gen4 TLC-based 3500 requires ~784 ms, and the Gen4 QLC-based 2600 extends to ~878 ms.

These results highlight the impact of storage throughput on AI cold-start latency. In the evaluated setup, the 3610's higher effective read bandwidth—enabled by PCIe Gen5, NAND interface bandwidth, and controller/firmware behavior—reduced the storage-bound phase by up to ~21% versus the evaluated Gen4 TLC SSD and up to ~30% versus the evaluated Gen4 QLC SSD.

From a system perspective, reducing model load time also shortens the period during which the SSD operates at elevated power, setting the stage for improved energy efficiency, as explored in the next section.

Figure 4: AI Model Load Time Comparison (ms)



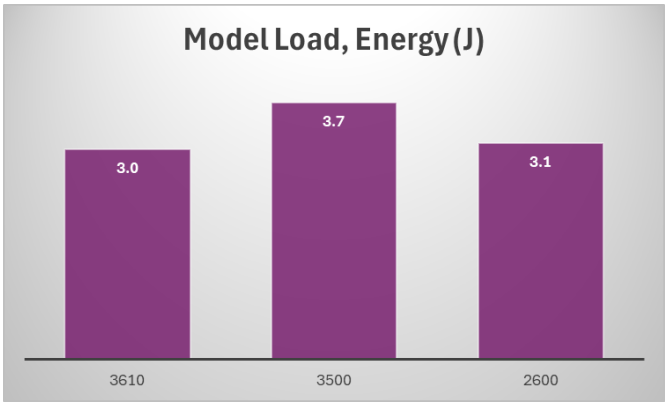
Measured AI model load time for Micron 3610 Gen5 QLC, Micron 3500 Gen4 TLC, and Micron 2600 Gen4 QLC SSDs. The Micron 3610 completes model loading in approximately 616 ms, compared to 784 ms for the Gen4 TLC SSD and 878 ms for the Gen4 QLC SSD.

Takeaway: Faster storage reduces the storage-bound portion of TTFT in cold-start scenarios by shortening the model-load interval.

While peak power is often a concern in mobile and edge systems, total energy consumption is the more meaningful metric for storage-bound AI workloads. Figure 5 shows the total energy consumed by each SSD during the model loading interval. The Micron 3610 Gen5 QLC SSD consumes approximately 3.0 J during model loading, compared to ~3.7 J for the Gen4 TLC-based 3500 and ~3.1 J for the Gen4 QLC-based 2600. Despite higher throughput, the 3610 reduces total energy by shortening the time spent at elevated power.

This result underscores a key insight for AI PCs and client-edge devices: performance and energy efficiency are not opposing goals. By completing storage-bound work more quickly, the 3610 minimizes the duration of elevated power draw, freeing energy and thermal headroom for subsequent compute-bound inference phases. This behavior is especially important in battery-powered systems, where sustained AI workloads must balance responsiveness, thermals, and power consumption.

Figure 5: Energy Consumed During AI Model Loading (J)



Total energy consumed by the SSD during AI model loading, derived from integrated power measurements across monitored SSD input rails. Despite higher instantaneous performance, the Micron 3610 Gen5 QLC SSD consumes less total energy than the Gen4 TLC SSD due to its shorter load duration.

Takeaway: Total energy during model loading depends on both power and time; higher throughput can lower energy if it shortens the load window.

Performance per Watt

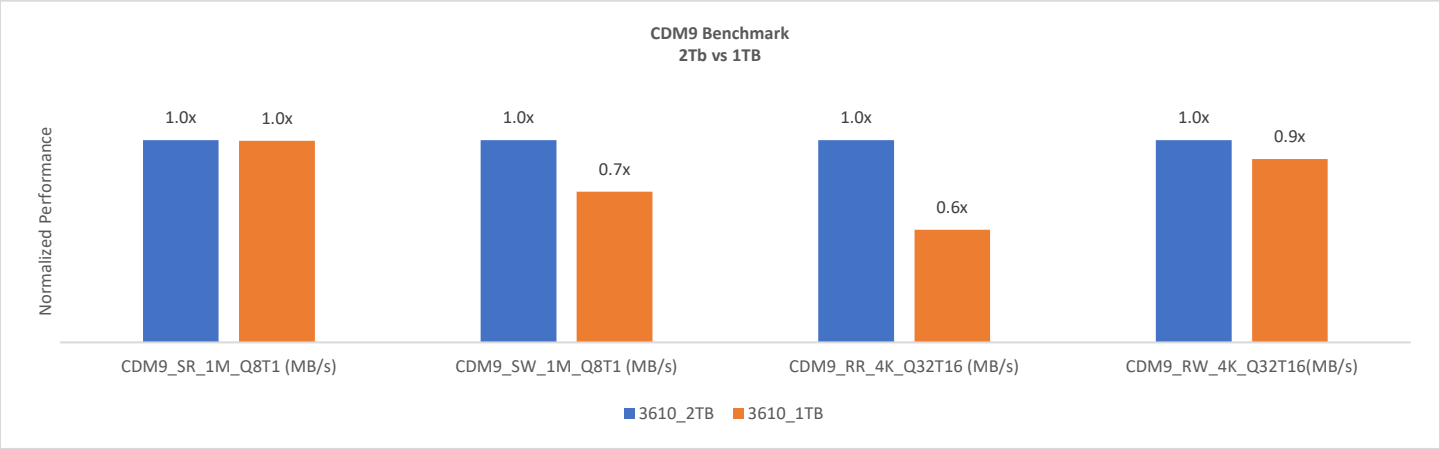
In the evaluated IOmeter tests, the Micron 3610 Gen5 QLC SSD improved performance per watt by approximately ~21–29% versus the evaluated Gen4 TLC and Gen4 QLC SSDs across the sequential and random workloads measured. Higher efficiency can enable AI model data to be moved faster while reducing energy used for storage-bound transfer, which can benefit power- and thermally constrained client systems.

Figure 6: SSD Performance per Watt with CDM 8 Benchmark

Capacity Impact: 1TB vs 2TB

CDM 9 results comparing 2TB and 1TB Micron 3610 SKUs show that while sequential read performance is similar, the 2TB drive demonstrates higher write and random read performance in the evaluated tests. The 2TB SKU provides up to ~34% higher sequential write throughput and up to ~81% higher random read performance (in CDM 9 testing), which can improve responsiveness for workloads that rely on frequent small-block access.

Figure 7: 3610 Capacity Scaling — 2TB vs 1TB



Conclusion

The data in this study show that AI cold-start latency is strongly influenced by the storage-bound model load window: the sooner the SSD completes weight streaming into DRAM, the sooner the system can begin generating tokens. For client and edge designs, this makes sustained read throughput and energy during the load interval key determinants of user-perceived responsiveness.

Across CDM, IOmeter, and real LLM cold-start tests, the Micron 3610 (Gen5, G9 QLC) demonstrated higher read performance and improved performance per watt versus the evaluated Gen4 SSDs, translating into shorter model-load time and lower energy consumed during loading in the evaluated tests. The implication for AI PCs and client-edge systems is practical: faster, more efficient storage can reduce the time and energy spent before inference can start, preserving thermal and battery headroom for sustained compute.

Author Bios

Ranjith Kumar Nagisetty is a Client Edge System Architecture Manager at Micron Technology, supporting the Mobile and Client Business Unit. Based in Hyderabad, he specializes in client Memory and Storage system architecture, workload engineering, and performance alignment, working closely with engineering, product, and field teams to enable customer engagements and platform strategy. He brings over 13 years of semiconductor industry experience, with a decade of focused work on system power and performance optimization in relation to thermal management.

Sandeep Kumar Goswami works as a Senior Engineer at Micron Technology, based in Hyderabad, India. He focuses on system validation for client and mobile storage solutions, with expertise in performance and power validation across real-world workloads. His role involves end-to-end system characterization, workload engineering, and competitive benchmarking to evaluate and align storage behavior against platform and usage requirements.

Pradeep Kumar Jilagam is the director of Systems & Workload Engineering for Micron's Mobile & Client Business Unit. He drives activities responsible for developing Micron's next gen memory and storage solutions, especially for edge AI, with impact across product definition, ecosystem enablement, product launch and thought leadership. He has strong expertise in multimedia and AI solutions across compute (CPU/GPU/NPU), memory and storage with a special focus on systems architecture. Prior to joining Micron, he worked on Android, IoT, XR and AIML at Qualcomm.

Sharath Chandra Ambula is a Staff Systems Engineer and Mobile Architect at Micron Technology, where he focuses on system architecture and performance analysis for mobile and client platforms, with an emphasis on memory-subsystem behavior. He brings over 10 years of industry experience spanning SoC system performance engineering and memory-centric system design across compute, interconnect, and memory subsystems. Sharath's work bridges system architecture and memory technology, with a focus on how standards, platform design choices, and real-world workloads influence performance, power, and scalability in modern client systems.

Ranjith Kumar Ravuri is a Lead Systems Architect in MCBU, driving workload engineering across Memory and Storage by translating system-level insights into product and business decisions. He leads AI-enabled workload engineering, cross-functional insight synthesis, and executive reporting to accelerate product strategy and workflow transformation. He brings ~20 years of industry experience, shaped by startup environments requiring agility, ownership, and end-to-end delivery, with deep expertise in NVMe SSD validation, NVMe-oF, and data center solutions.

Rui Zhou is a Principal Product Marketing Manager for the Mobile Client Business Unit at Micron Technology. With over 15 years combined experiences of marketing and engineering in high-tech industry, Rui is an expert in product positioning, value proposition, and growth strategy. Rui has an MBA from Portland State University, and both a M.Sc. in IC design and a BEng degree in Electrical and Electronics Engineering from Nanyang Technology University in Singapore.

References

Micron.com/3610

micron.com

©2026 Micron Technology, Inc. All rights reserved. All information herein is provided on as "AS IS" basis without warranties of any kind, including any implied warranties, warranties of merchantability or warranties of fitness for a particular purpose. Micron, the Micron logo, and all other Micron trademarks are the property of Micron Technology, Inc. All other trademarks are the property of their respective owners. No hardware, software or system can provide absolute security and protection of data under all conditions. Micron assumes no liability for lost, stolen or corrupted data arising from the use of any Micron product, including those products that incorporate any of the mentioned security features. Products are warranted only to meet Micron's production data sheet specifications. Products, programs and specifications are subject to change without notice. Rev. A 05/2026 CCM004-1681249710-11882.