Storage Optimizations for the AI PC

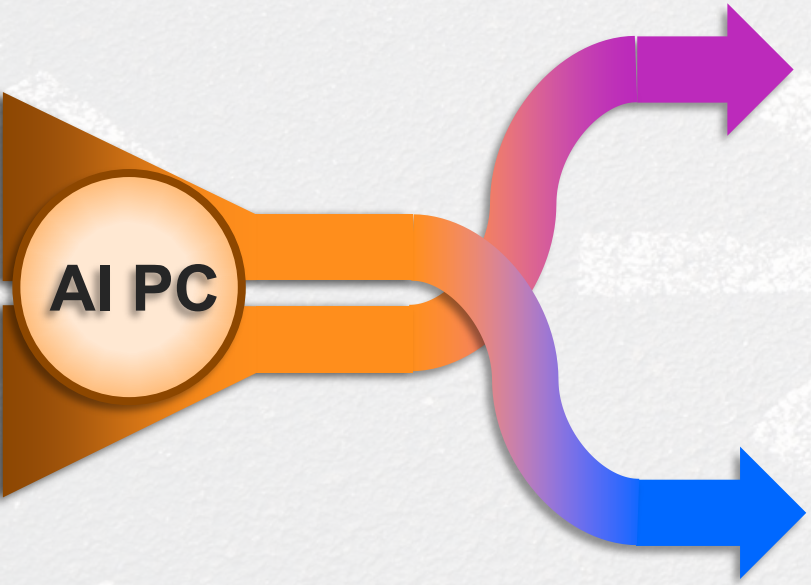
Rahul Jairaj (Micron), Cory Steinmetz (Micron) and Scott Lee (Microsoft)

micron Intelligence Accelerated™

FMS

AI PC's

Emerging Trends



Cloud based inferencing and the edge devices serve as conduits for the information. Works well for a vast array of cost sensitive PCs and casual PC use cases.

Most inferencing to happen at the edge at low latencies, necessitating a massive need for low power HW, large capacity memory and high speed NVMe **Storage**.

AI PC's

Emerging Trends

Edge Inferencing

Energy Efficiency to conserve Battery life

Generative AI applications take a significant toll on battery life for mobile systems and expected to extend to the PC as well – Link
Tirias Research for Enovix

Doing More with the same Hardware - Cost Optimization

AI PC & Storage

LLMs & VLMs on PCs

Large Language Models are anticipated to exceed

OR

Have multiple models running in parallel competing for Memory resources

This would *ideally* need over

7-13B+
Parameters

16-32GB+
DRAM

Model Parameters partitioned **between Storage** & **Memory**, rotating the parameters on demand

Storage

1TB+
Large Capacity storage to host advanced LLMs, OS, apps and generated data

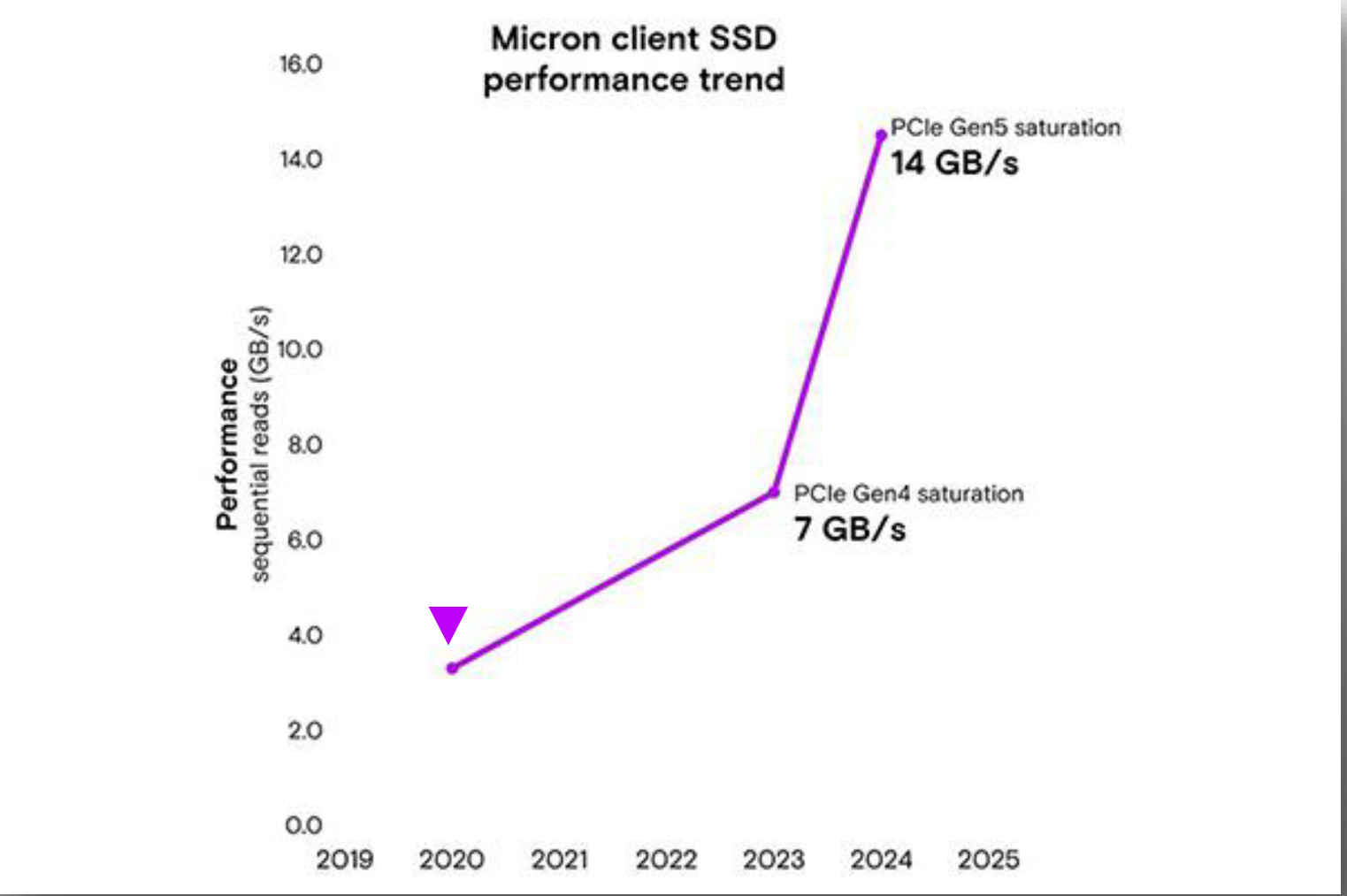
14GB/s+
High performance & Low Latency SSDs for seamless data transfer between Storage and Memory

Memory

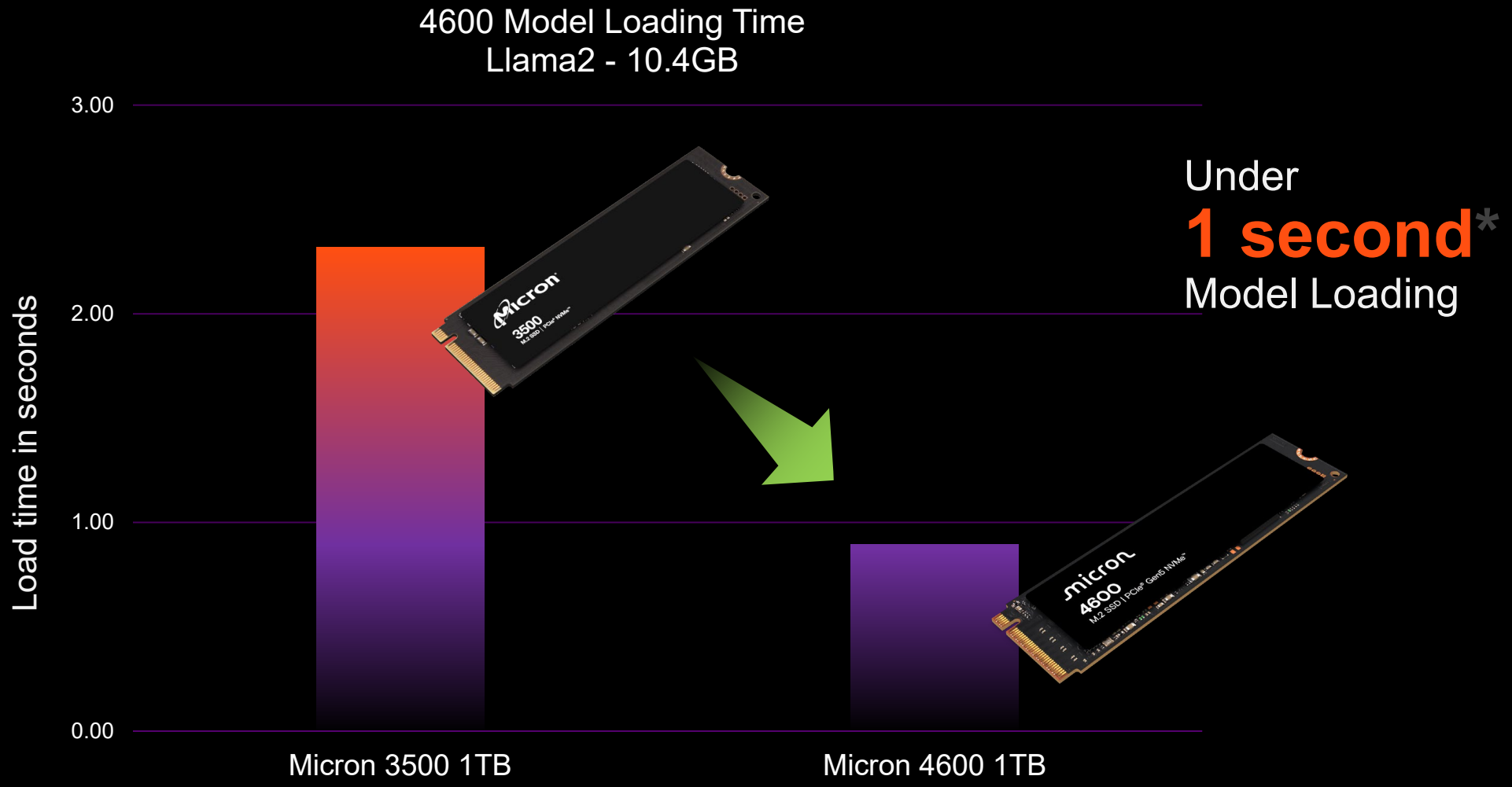
AI in PC: Why Not? – Link
Blog By Prasad Alluri

Reference Article : LLM in a flash: Efficient Large Language Model Inference with Limited Memory (arxiv.org)

Do Storage Performance Trends Matter?



Ollama Model Loading 4600 vs 3500

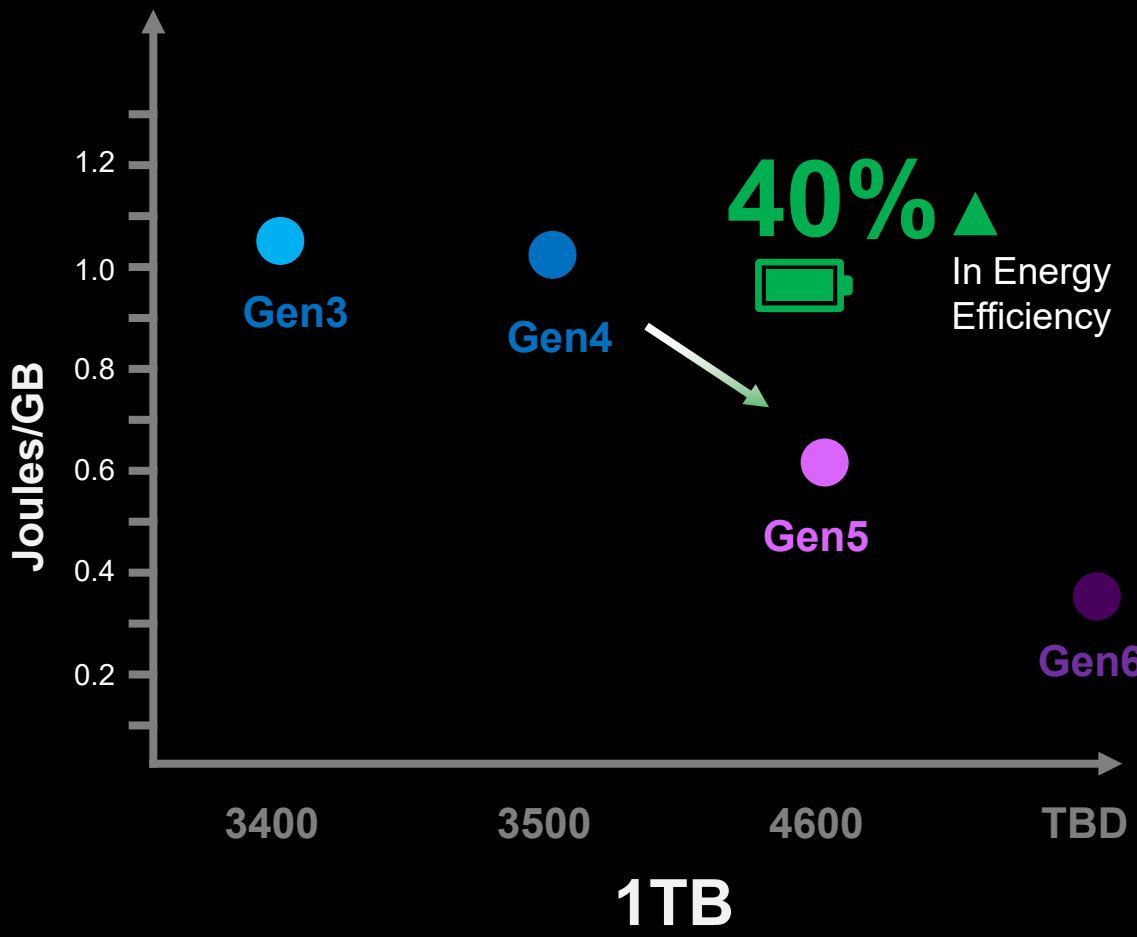


*Experimental model fragmentation setup to drive full Gen 5 BW saturation

Energy Savings from High Speed NVMe Storage

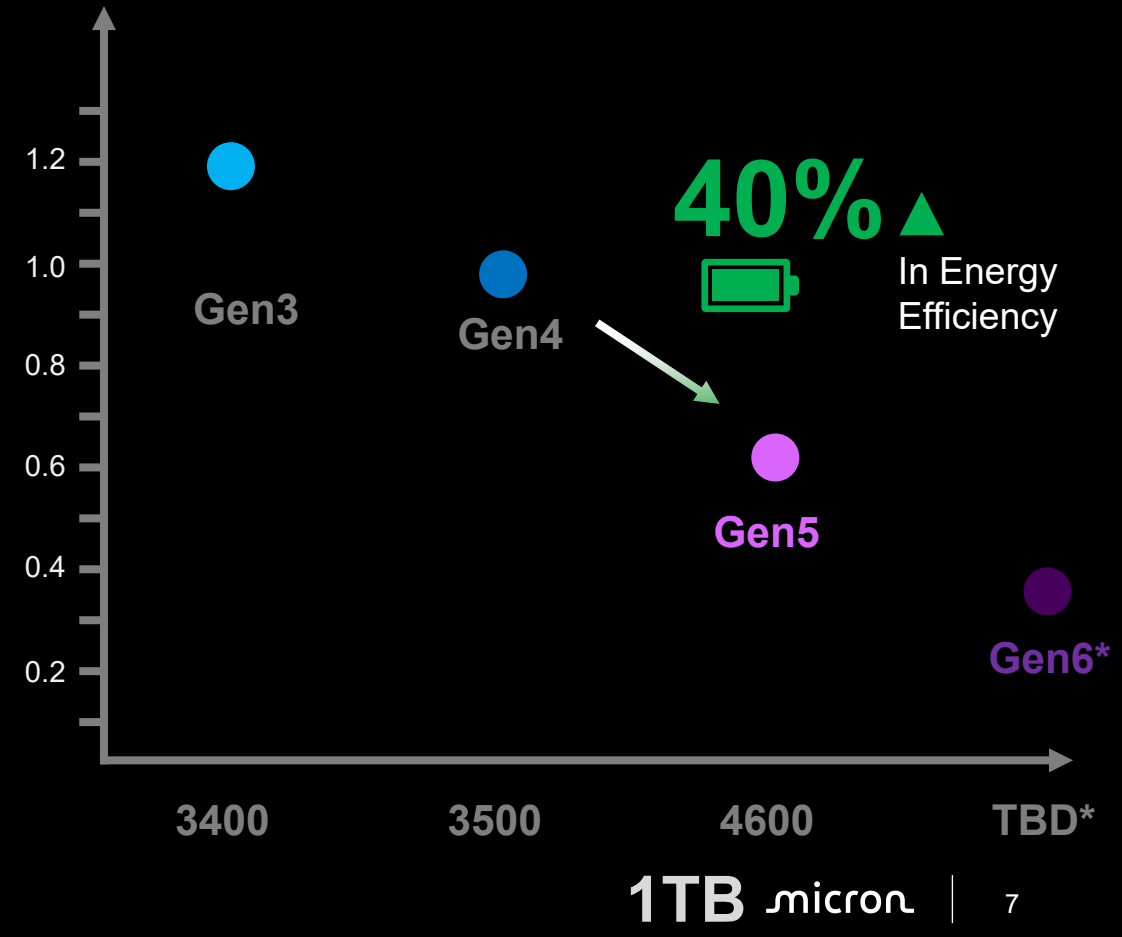
Read

Energy to Read
1GB of Data



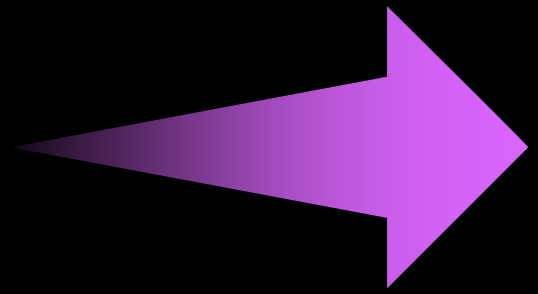
Write

Energy to Write
1GB of Data



Testing Lower Latency Storage Options

QLC



SLC

Via Firmware

AI PC & Storage

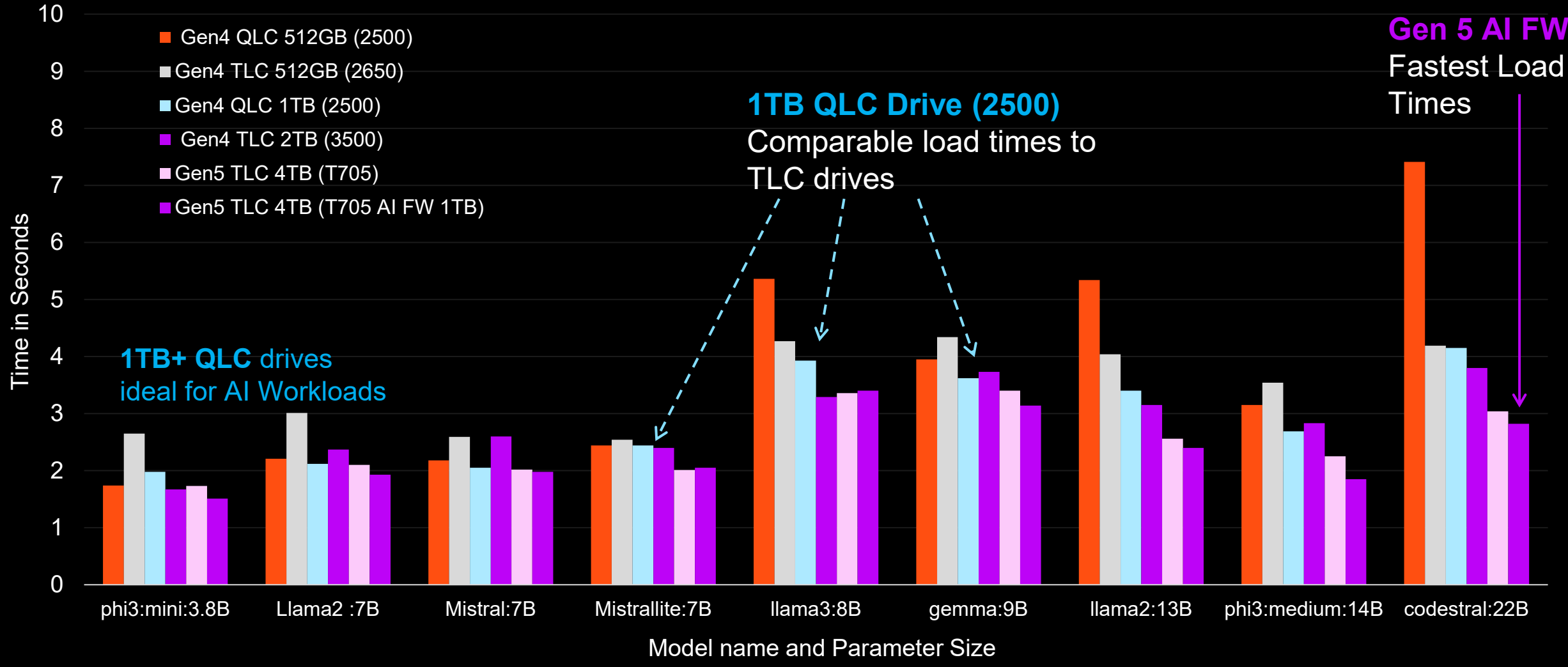
Measured Model Load Times: SLC vs QLC

Impact of having models reside only in SLC

Model Name	Model Size	Micron 512GB 2500 with **LLM in QLC**	Micron 512GB 2500 with **LLM in SLC**	Time Delta Percentage
Phi3:mini	2.2 GB	1.66 Seconds	1.58 Seconds	4%
Llama2 (7B)	3.8 GB	2.83 Seconds	2.15 Seconds	24%
Mistral	4.1 GB	2.42 Seconds	2.48 Seconds	2%
Mistrallite	4.1 GB	2.97 Seconds	2.76 Seconds	7%
Llama3	4.7 GB	3.78 Seconds	2.56 Seconds	32%
Gemma	5.0 GB	3.57 Seconds	2.29 Seconds	36%
Llama2 (13B)	7.4GB	5.07 Seconds	2.90 Seconds	43%
Phi3	7.9GB	4.71 Seconds	2.84 Seconds	40%
Codestral	12GB	7.21 Seconds	4.01 Seconds	44%

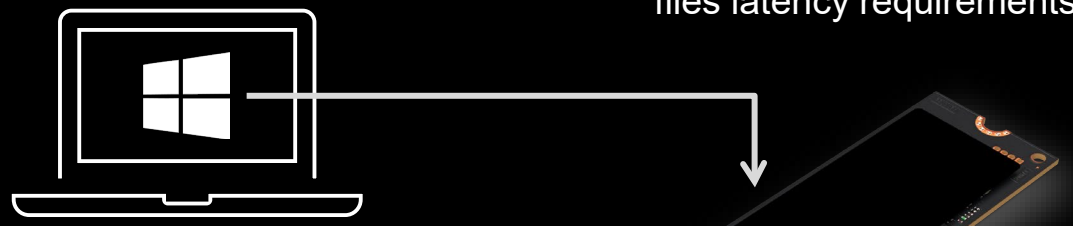
Having the models reside only in **SLC regions** on average improves model load times by **~26%** ▲

Ollama Model Loading Time Comparison

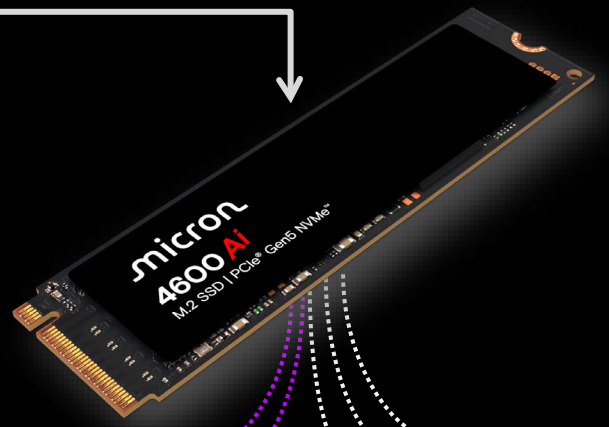


SSD Firmware Optimizations

OS provides "**Hints**" about files latency requirements



1TB+



Data marked as needing low latency access is pinned to the Dynamic SLC region and not folded over – guaranteeing short access times always

FMS 2024

Future of Memory and Storage

Microsoft Presentation

1. DSM Hints from the OS

DSM Hints

Scott Lee announced the proposal for Dataset Management Hints

Micron and Microsoft Codeveloped this concept to help optimize storage for the AI PC and beyond.

NVMe Dataset Management (DSM) Hints

AI PC & Storage

This feature driven by Micron in collaboration with Microsoft

Why?

1. Allows SSD to manage data in different ways
   - Improve latency (lower)
   - Group data (sequential)
   - Improve wear leveling (smarter folding)
2. Gives priority to most critical OS functions
   - Smoother user experience
   - Improved performance
3. Improves LLM Load times by optimizing for low latency execution of model files.

How?

Micron innovating with existing NVMe features

The OS gives extra information to SSDs about importance of data

Bits	Description
07:00	**Dataset Management (DSM):** This field indicates attributes for the LBA(s) being read.

Bits	Attribute	Definition
07	Incompressible	If set to '1', then data is not compressible for the logical blocks indicated. If cleared to '0', then no information on compression is provided.
06	Sequential Request	If set to '1', then this command is part of a sequential read that includes multiple Read commands. If cleared to '0', then no information on sequential access is provided.
05:04	Access Latency	

Value	Definition
00b	None. No latency information provided.
01b	Idle. Longer latency acceptable.
10b	Normal. Typical latency.
11b	Low. Smallest possible latency.

Bits	Attribute	Value	Definition
03:00	Access Frequency	0h	No frequency information provided.
		1h	Typical number of reads and writes expected for this LBA range.
		2h	Infrequent writes and infrequent reads to the LBA range indicated.
		3h	Infrequent writes and frequent reads to the LBA range indicated.
		4h	Frequent writes and infrequent reads to the LBA range indicated.
		5h	Frequent writes and frequent reads to the LBA range indicated.
		6h	One time read. E.g., command is due to virus scan, backup, file copy, or archive.
		7h	Speculative read. The command is part of a prefetch operation.
		8h	The LBA range is going to be overwritten in the near future.
		9h to Fh	Reserved

Dataset Management Hints in Windows

Windows NVMe driver has been updated to pass DSM hints for some reads and writes commands

Available DSM Hints

- DSM.SequentialRequest
- DSM.Incompressible
- DSM.AccessLatency = NVME_ACCESS_LATENCY_LOW
- DSM.AccessLatency = NVME_ACCESS_LATENCY_LOW & DSM.AccessFrequency = NVME_ACCESS_FREQUENCY_FR_WRITE_INFR_READ
- DSM.AccessLatency = NVME_ACCESS_LATENCY_IDLE

Support available in Windows client preview builds through Windows Insiders program or Collaborate if you are part of a Windows ecosystem partner program

Seeking feedback on the usefulness of these DSM hints. Will add more if it is useful

Prototyping Effort

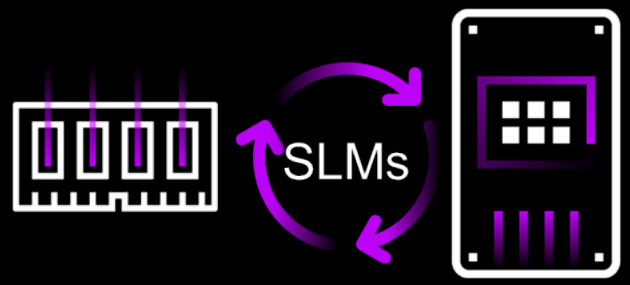
Micron is currently prototyping DSM Hints to explore a multitude of "What if" scenarios

1. When the drive is 100% full and SLC space is limited what happens to LBAs marked with smallest possible latency?
2. How are the context attributes stored?
3. What is the impact to Garbage Collection, endurance, and WAF?
4. What is the impact to other benchmarks such as CDM and PCMark10?

Micron advances AI storage innovations

SSD feature developments to improve user experiences

PERFORMANCE



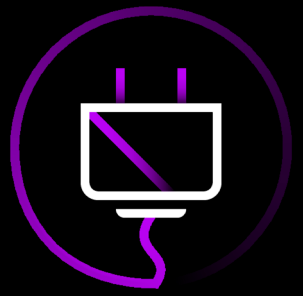
AI PC needs

- Reduce Time To First Token (TTFT) and seamlessly transition from one AI application to the next

Micron development

- NAND timings & ONFI
- Advanced caching & heuristics for faster model access

POWER



AI PC needs

- Energy efficient SLM transfer from storage to memory

Micron development

- Micron PCIe Gen5 for 45% improvement over Gen4 in picojoules per bit
- NAND and controller optimizations

PROTECTION

AI PC needs

- Data path protection

Micron development

- Controller features
- Ecosystem engagement

micron Intelligence Accelerated™