Future architectures for AI workloads

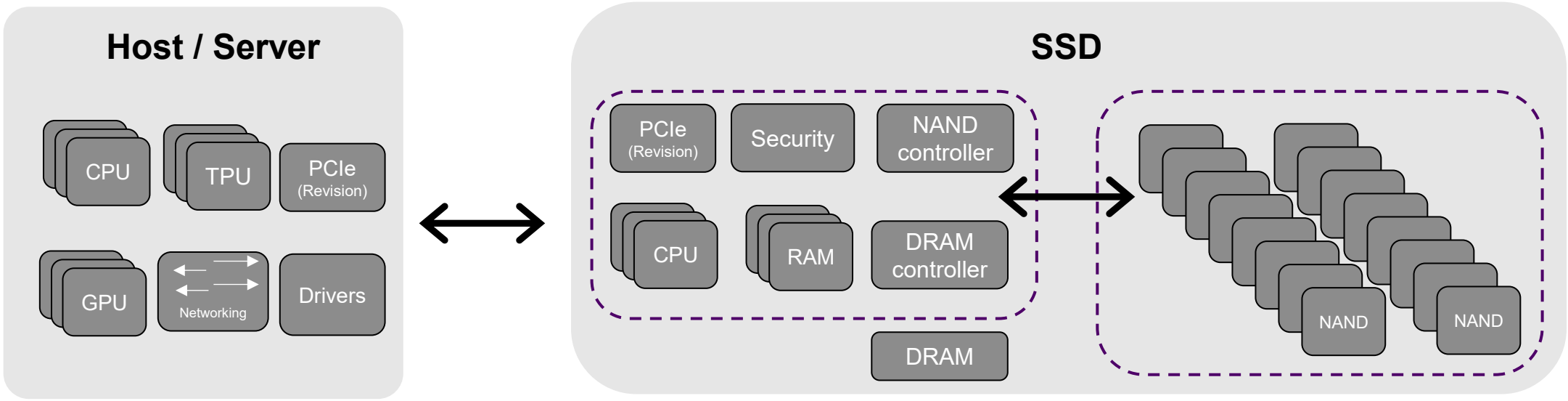
Rob Sykes
Director ASIC TPM, Micron

micron Intelligence Accelerated™

FMS

Micron Confidential

Understanding the bottlenecks



Host system bottlenecks

- PCIe
- Queue depth
- # Queues
- Packet overheads
- Network / switch performance

SSD bottlenecks

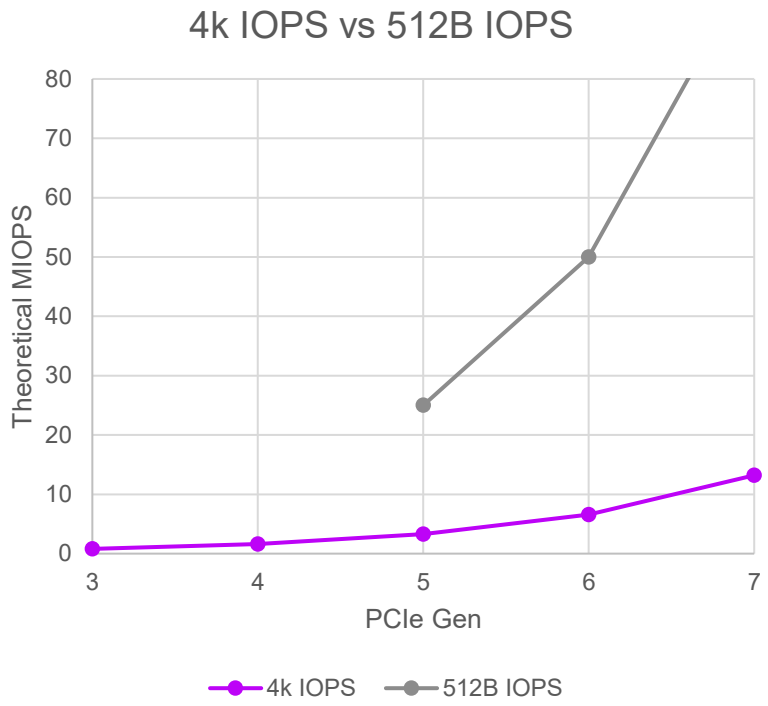
- PCIe
- Power
- NAND channels / planes
- NAND ONFI/Toggle
- Controller

Micron Confidential

Host interface performance and overheads

Host PCIe performance / overhead

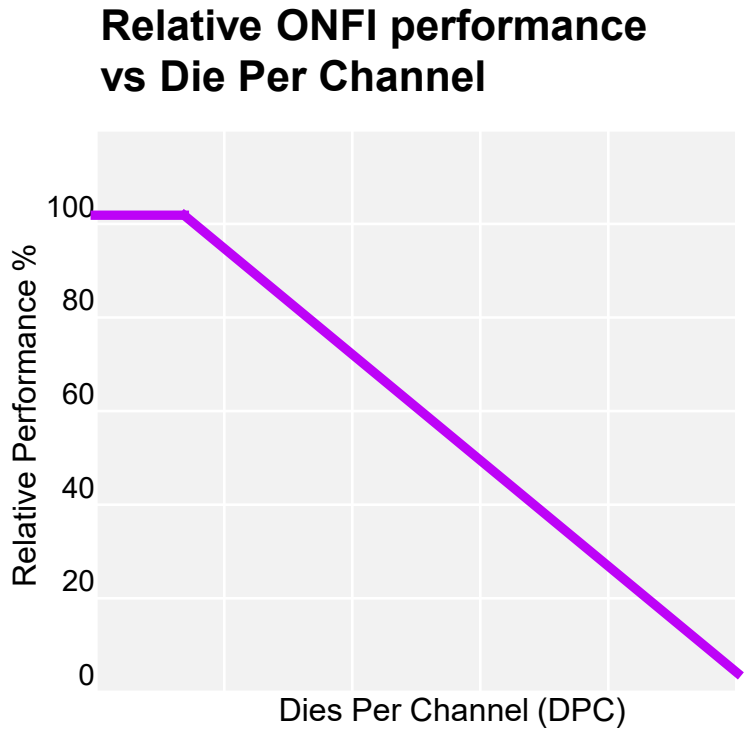
Specification	X4 Lanes (bidirectional)	Burst Performance (Encoding)	98% Packet Efficiency	95% Host Efficiency	Random Reads (MIOPS) 4KB	Random Reads (MIOPS) 512B
PCIe 4.x	8 GB/s	7.88 GB/s	7.72 GB/s	7.3 GB/s	1.89 MIOPS	15.3 MIOPS
PCIe 5.x	16 GB/s	15.75 GB/s	15.45 GB/s	14.7 GB/s	3.7 MIOPS	29.6 MIOPS
PCIe 6.x	32 GB/s	30.11 GB/s	29.54GB/s	28 GB/s	7.2 MIOPS	62.4 MIOPS
PCIe 7.x	64GB/s	60.24 GB/s	59.08 GB/s	56. 13 GB/s	14.4 MIOPS	115.2 MIOPS



- Controller overhead will reduce the numbers further – architecture specific
- 512B numbers will be lower due to increased operations / overhead

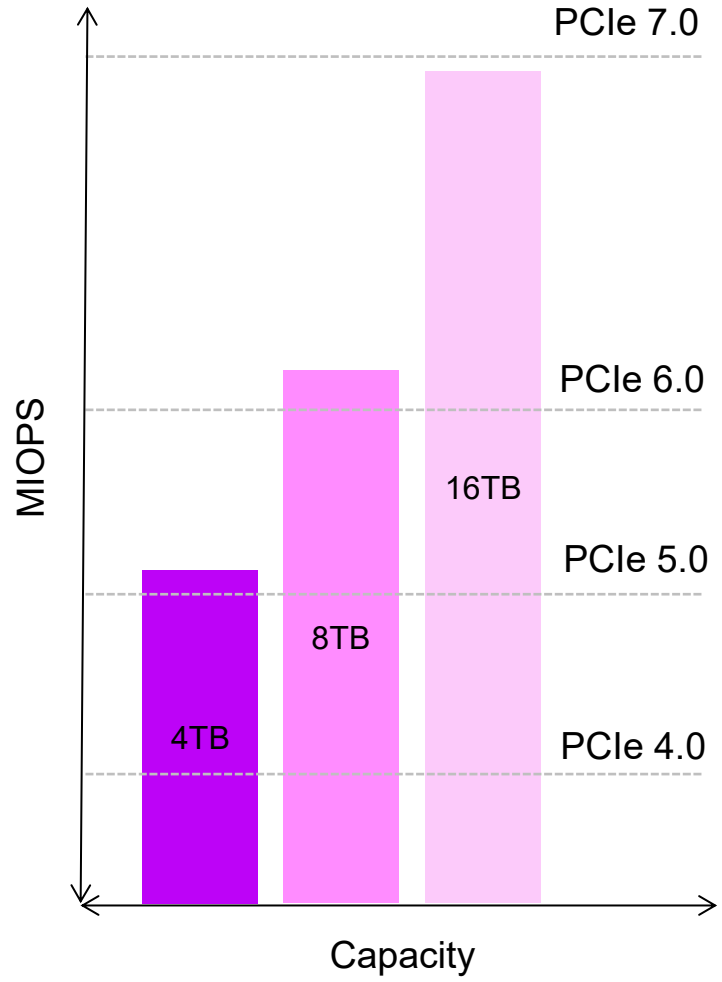
Relative ONFI performance

- NAND parallelism provides high performance for small capacity, but as more die per channel are added (channel loading) the relative performance decreases
- Adding more die doesn’t provide a linear improvement in performance
- Channel loading will decrease the performance of the drive
- Power is also a critical factor on larger capacity drives, performance within a power budget



PCIe generation vs 4KB Random Read (TLC)

- As PCIe bandwidth improves over time, SSD vendors have had to move to larger capacities to keep up with the PCIe Generation
- For some AI workloads requiring maximum Random Read Performance, keeping the $/IOPS under control is a battle against capacity and performance
- PCIe 6 Random Read saturation can be achieved on an 8TB drive (may be either air cooled or liquid cooled)
- PCIe 7 Random Reads cannot be easily met on a 8TB drive and in some cases a 16TB drive may not have enough die to achieve the performance
- Consideration to number of NAND Die, NAND Type (TLC, MLC, SLC), Power and Capacity will have to be considered to achieve maximum PCIe Gen7 throughput



GPU based hosts and performance

The GPU working dataset is growing

- AI applications (RAG, Agentic workloads, Vector dB, GNNs, LLM Inference) all need constant access to a large pool of data

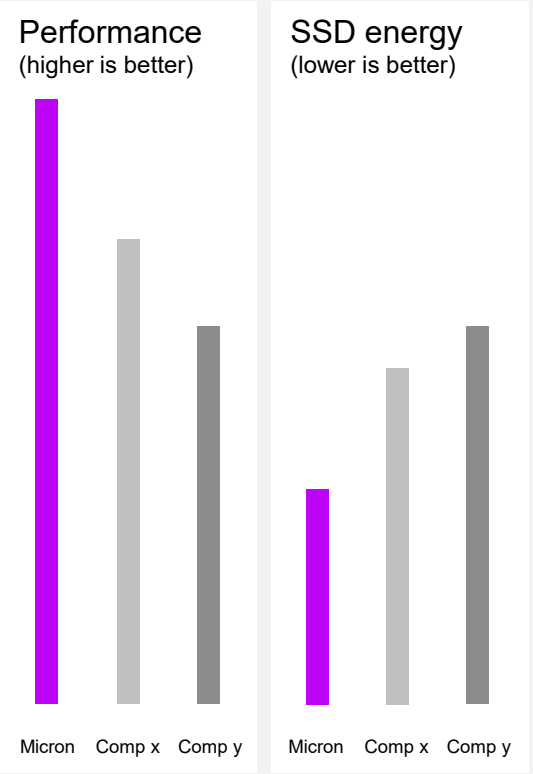
High IO requests that are concurrent and often random

- Driving high IOPs requires a significant amount of compute
  - 3M IOPs on 1 drive can consume 4-8 CPU cores
- Traditionally, CPUs initiated IO requests
  - Modern CPUs have up to 192 cores and typically 1 to 2 sockets
  - Maximum 384 physical cores, 768 logical cores
- GPUs can drive much larger IO workloads
  - H100 has 16,000 cores
  - Can easily saturate Gen5x16 in IO and still execute complex training and inference workloads
  - Can drive storage queue depth over 10,000

Micron Confidential

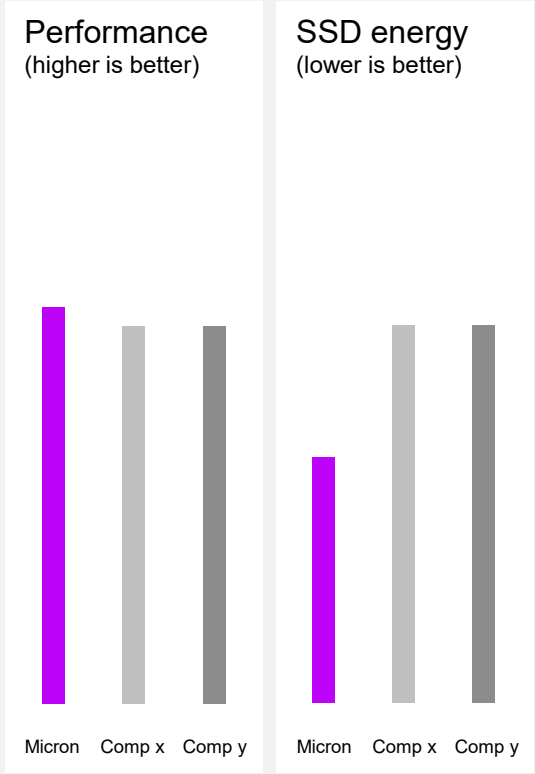
The importance of Pwr/Perf efficiency

Graph neural network training
(Big accelerator Memory)



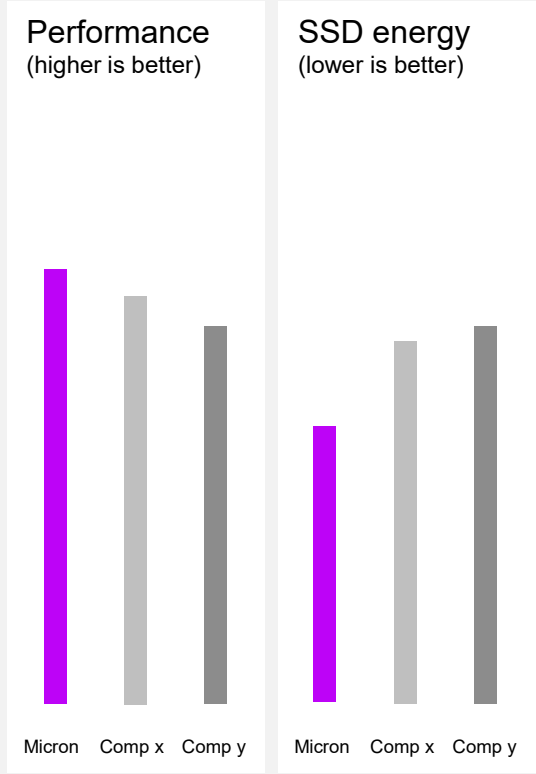
Up to
60% higher performance
43% less energy

Unet3D medical image training
(Deep learning IO)



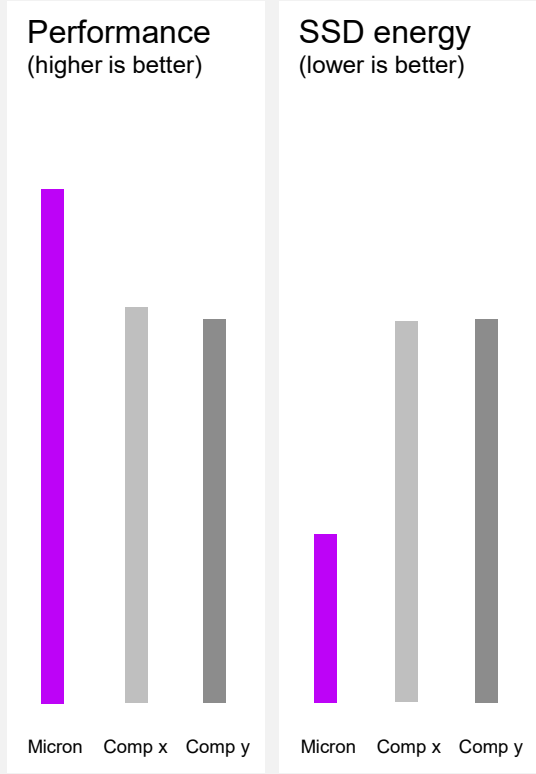
Up to
5% higher performance
35% less energy

Large language model inference
(DeepSpeed ZeRO-Inference LLM)



Up to
15% higher performance
27% less energy

NVIDIA GPUDirect® Storage



Up to
34% higher performance
56% less energy

Micron Confidential

Architectural considerations

Power Efficiency

Random Workloads consume more power than sequential reads. To achieve the best performance, all aspects of an SSDs power budget needs to be scrutinized to maintain low TCO / Performance.

Performance

Implications to all aspects on the controller architecture. Consideration to #NAND Channels, die loading, and power.

Complexity

One Size fits all vs targeted solutions. The landscape is forever changing, and all require specific ASIC resources, with impact to power, performance, and potentially die size / cost. Difficult architectural decisions need to be made to meet sometimes difficult schedules in aligning to PCIe Generations and releasing competitive products with demanding schedules.

Micron Confidential