Optimizing Flash Controllers for the AI Data Center

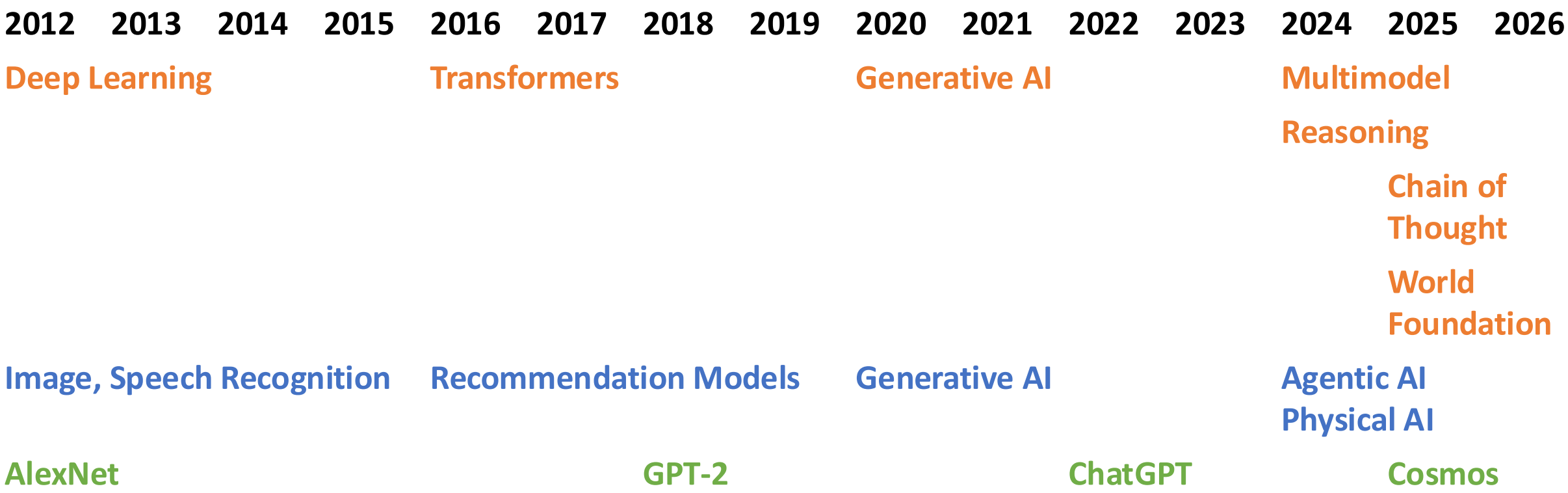
Erich F. Haratsch

Senior Director Architecture

Outline

- Evolution of AI
- AI Model Complexity
- AI Data Processing Pipeline
- Storage in the AI Data Center
- Performance Considerations for SSDs
- Conclusion

Recent Evolution of AI



- Compute, memory and storage requirements continue to increase

AI Model Complexity

Model	Release Year	Parameter Count	Model Size	Training Tokens	Raw training data
AlexNet	2012	60 million	240 MB	Not applicable	~1.2 TB
GPT-3	2020	175 billion	700 GB	300 billion	~45 TB
GPT-4	2023	1.76 trillion (*)	7 TB	13 trillion (*)	1 PB (*)
Llama2	2023	70 billion	280 GB	2 trillion	N/A
Llama3	2024	405 billion	1.6 TB	15 trillion	N/A

(*) estimated
Assumption: 4 bytes per parameter

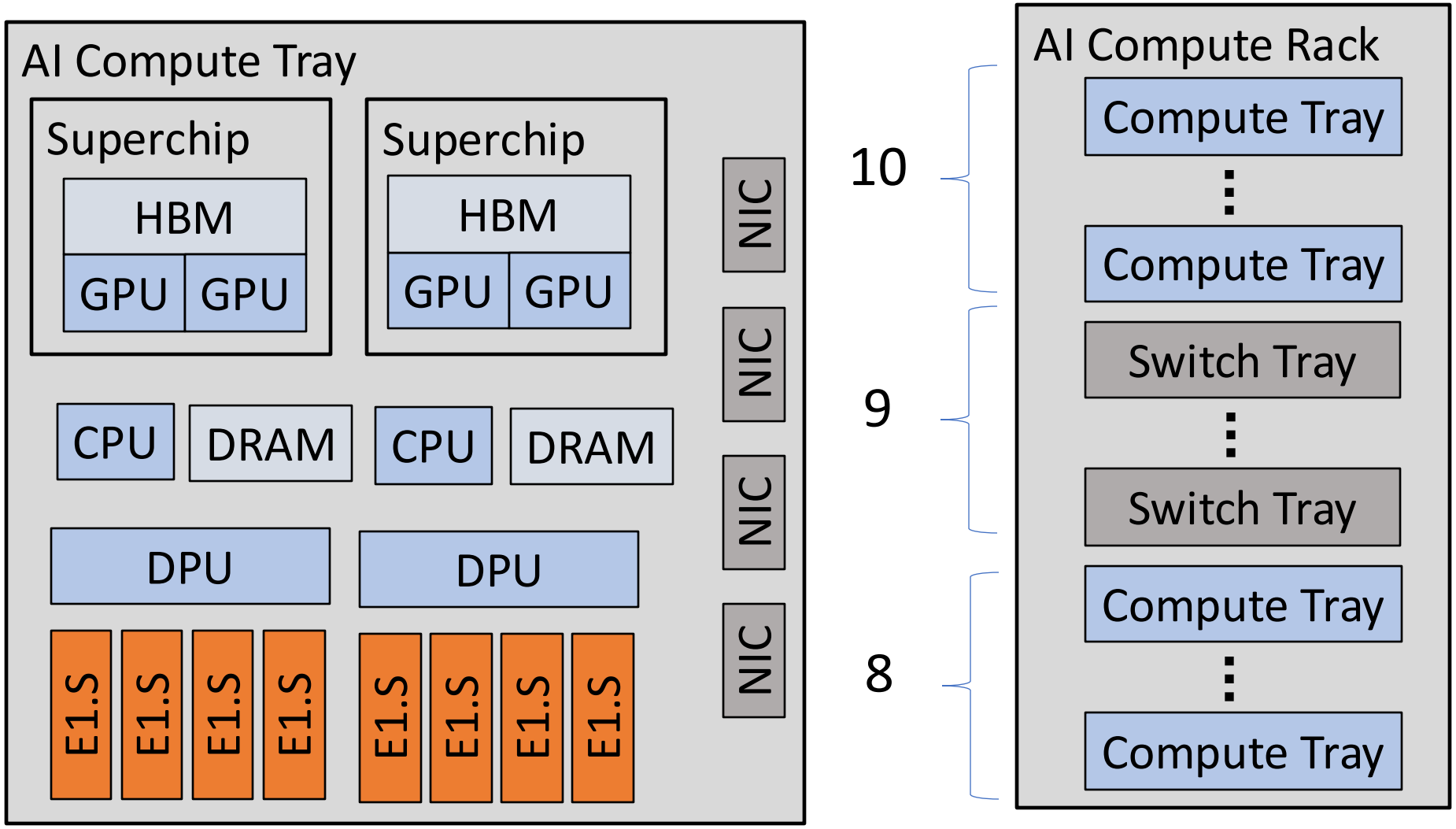
AI Processing Phases and Storage Workloads

Sequential	**Sequential Random**	**Sequential Random**	**Random**	**Sequential**
PBs	**TBs**	**TBs**	**TBs RAG: PB**	**PBs**

Training vs Inference

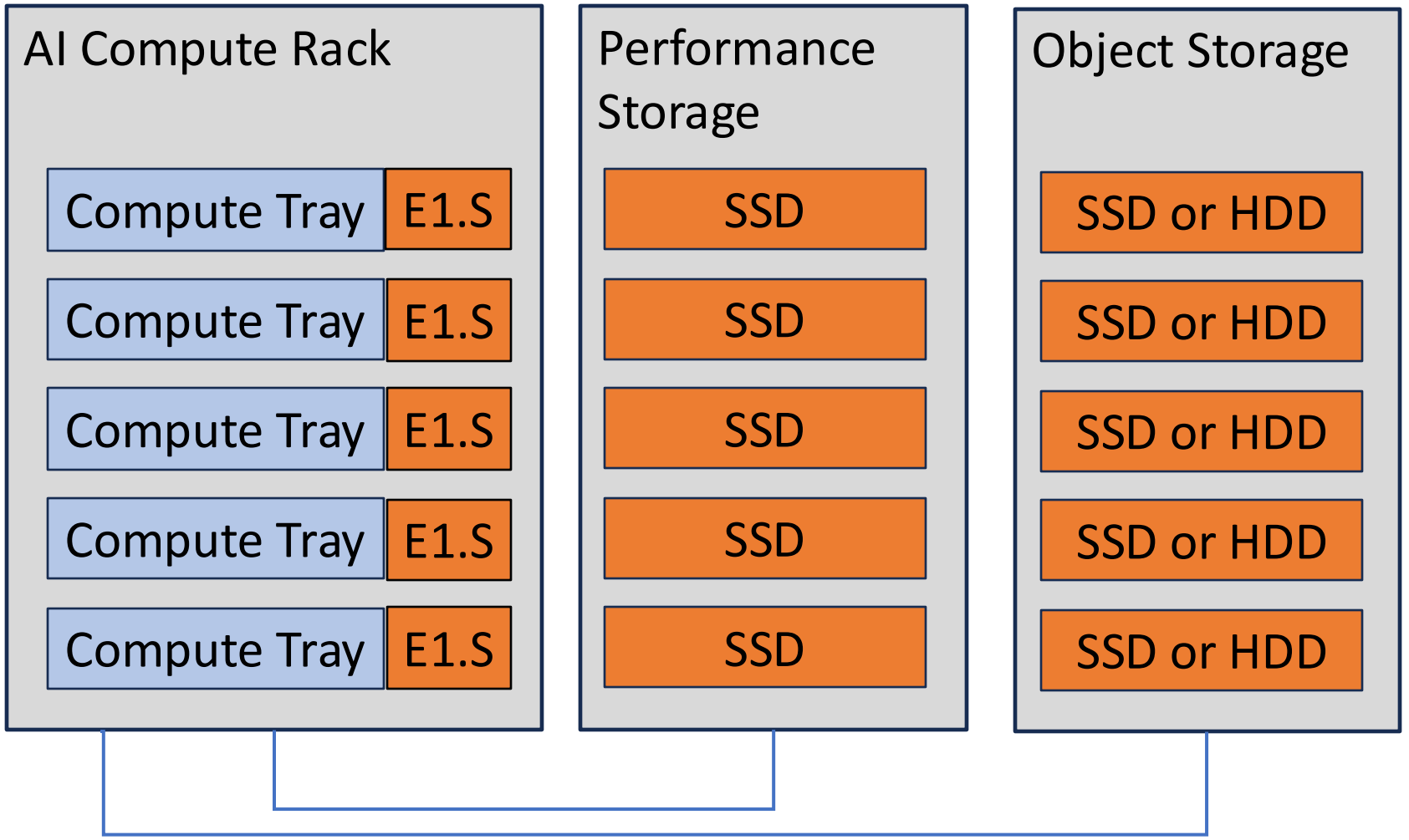
- Training
  - One large job on supercomputer with 10,000s or 100,000s of GPUs
  - Bandwidth is important
  - Frequent checkpointing to save model state in storage
- Inference
  - Many threads in parallel
  - Time to answer (latency) is important
  - RAG drives additional need for storage

Exemplary AI Compute Tray and Rack



- Per tray:
  - 4 GPUs
  - 2 CPUs
  - 8 E1.S SSDs
- Per rack:
  - 18 compute trays
  - 72 GPUs
  - 36 CPUs
  - 144 E1.s SSDs
- 32 PCIe lanes per tray for storage

AI Storage Tiers

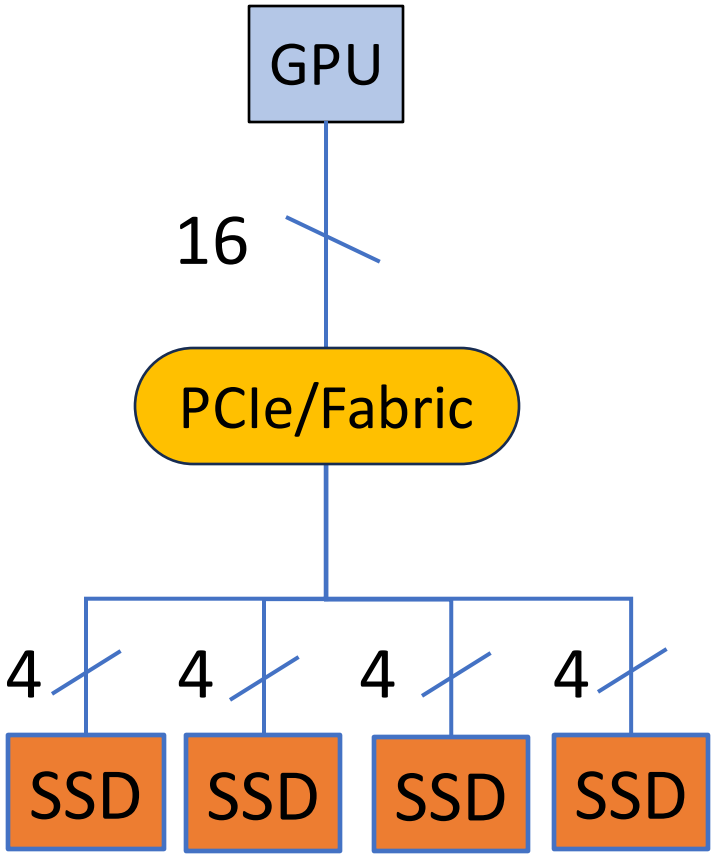


- TLC typical for compute rack and performance storage
- QLC or HDD typical for object storage
- Ethernet or Infiniband connectivity

SSD Performance Considerations

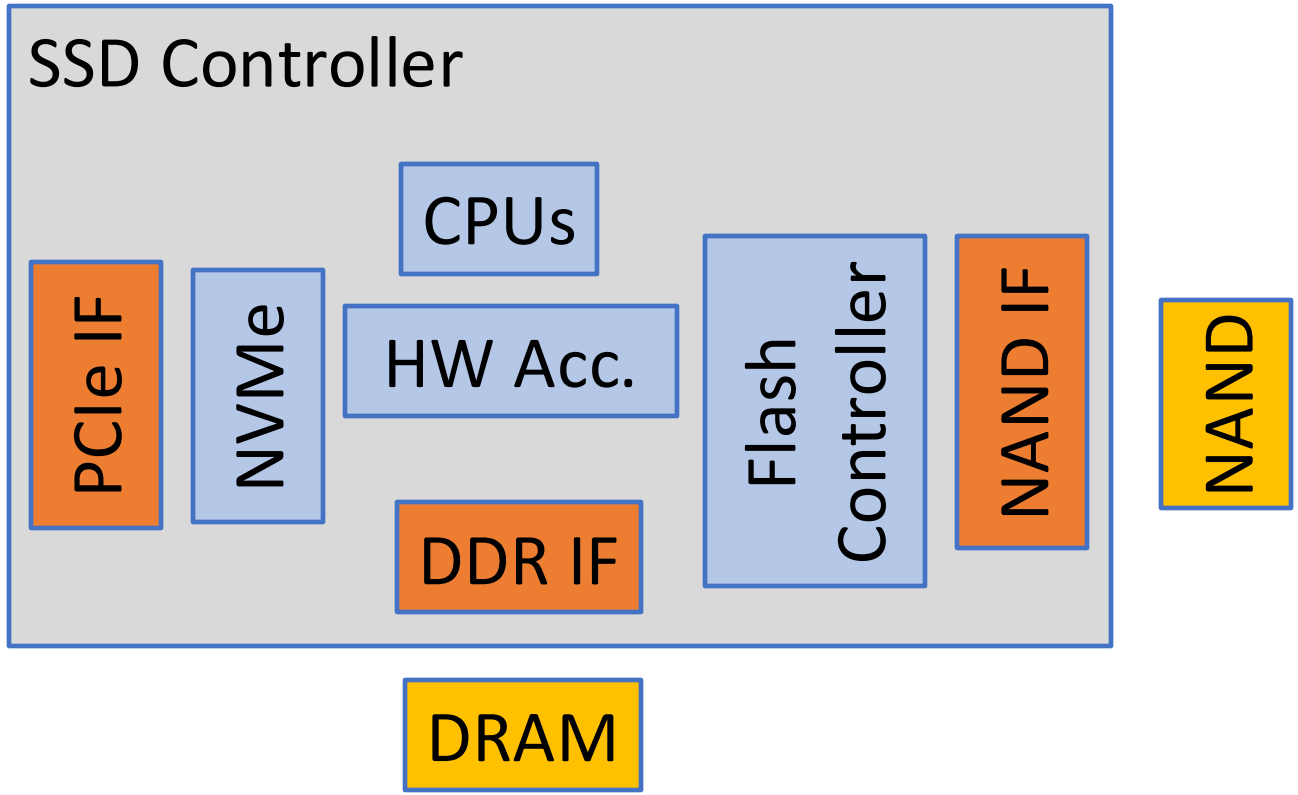
- In the past, SSDs adopted next generation PCIe interfaces later than CPUs
- AI is now driving adoption of next generation PCIe SSDs
- SSDs typically designed to saturate sequential and 4KB RR performance
- However, some AI workloads (especially for inference) have random accesses smaller than 4KB.

Saturating Read Performance For PCIe Gen6



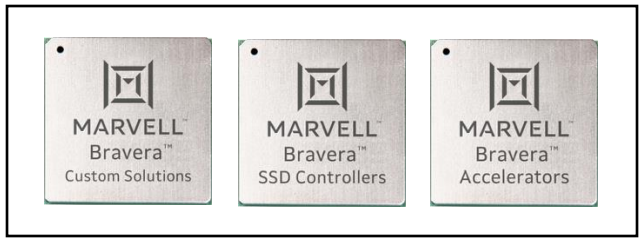
- Nvidia's Storage-Next initiative targets ~200 MIOPS (512B) for 16 PCIe lanes
  - See Nvidia presentations at OCP 2024, SC 2024, GTC 2025
- 16 PCIe lanes per GPU
  - 128 GB/s Raw
  - ~110 GB/s Effective
  - ~26.8 MIOPS (4KB)
  - ~215 MIOPS (512B)
- 4 PCIe lanes per SSD
  - 32 GB/s Raw
  - ~27.5 GB/s Effective
  - ~6.7 MIOPS (4KB)
  - ~53.7 MIOPS (512B)
- Current SSDs designed for 4KB IOPS performance
- Saturating 512B IOPS means 8x higher performance

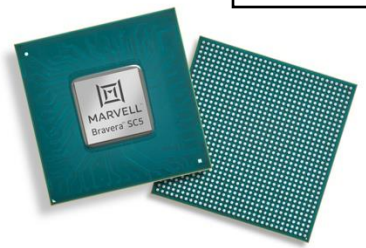
Optimizations for High IOPS SSDs



- NAND
  - Read Time
  - SLC, MLC vs TLC NAND
  - Page and Plane Architecture
- SSD Controller
  - CPUs
  - HW acceleration for FW Offload
- Host interfaces
- Form factors
  - 8x random read performance will increase power

March to Gen6 SSDs





3.2GB/s
16TB SSD
LDPC

PCIe 3.0

7.4GB/s
64TB SSD
LDPC+

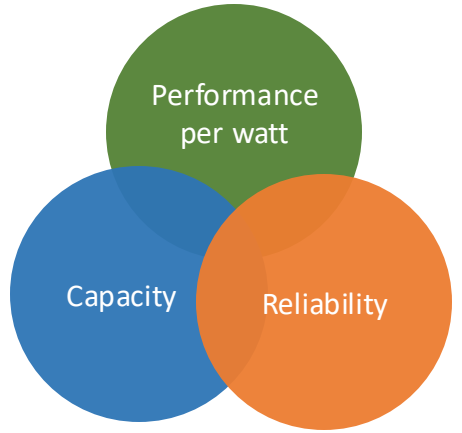
PCIe 4.0

14GB/s
128TB SSD
LDPC++

PCIe 5.0

28GB/s
256TB SSD
LDPC+++

PCIe 6.0



Critical Flash Controller AI workload needs

Gen6 closes the gap for AI workload needs

Conclusion

- New AI models continue to be released with increased capabilities and complexity
- Storage is an essential component in AI data centers
- AI drives adoption of next generation PCIe interfaces for storage
- Increasing Random Read IOPS by 8x requires optimizations in NAND media, SSD controller, host interfaces and potentially new form factors