REGIONAL SDC 25
BY Developers FOR Developers

Regional SDC Denver
April 30, 2025

Storage Devices for the AI Data Center

Erich F. Haratsch, Marvell

SNIA

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

2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026
Deep Learning				Transformers				Generative AI				Multimodel Reasoning	Chain of Thought World Foundation	
Image, Speech Recognition				Recommendation Models				Generative AI				Agentic AI Physical AI		
AlexNet						GPT-2				ChatGPT			Cosmos	

- Compute, memory and storage requirements continue to increase

REGIONAL SDC

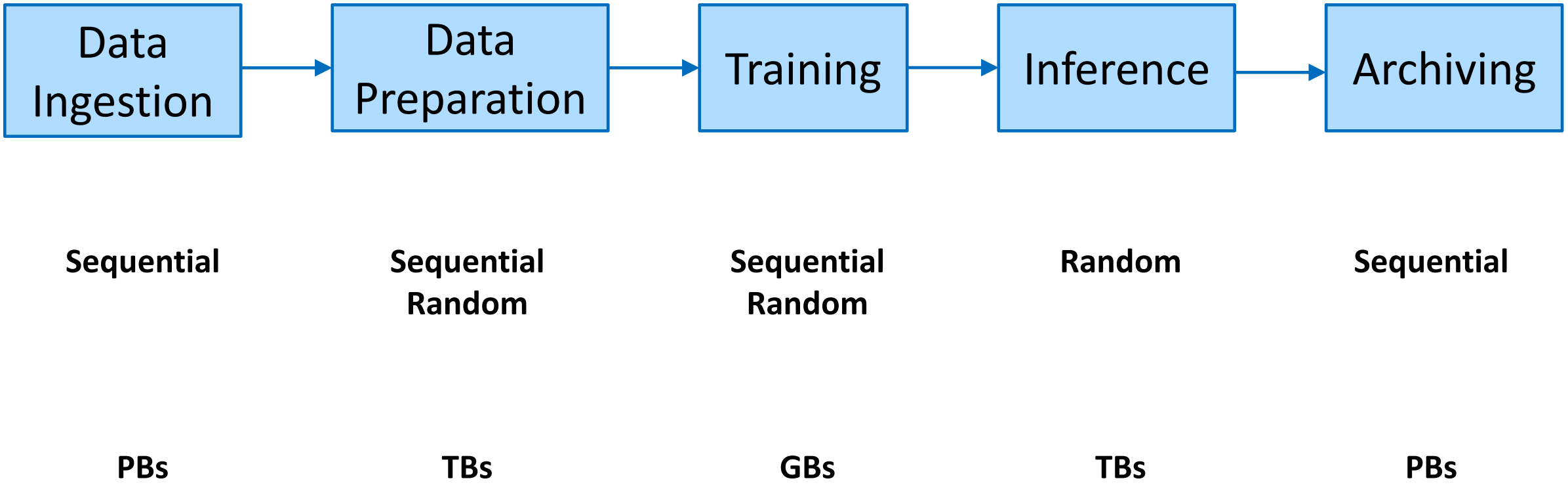
AI Model Complexity

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

REGIONAL SDC

AI Processing Phases and Storage Workloads

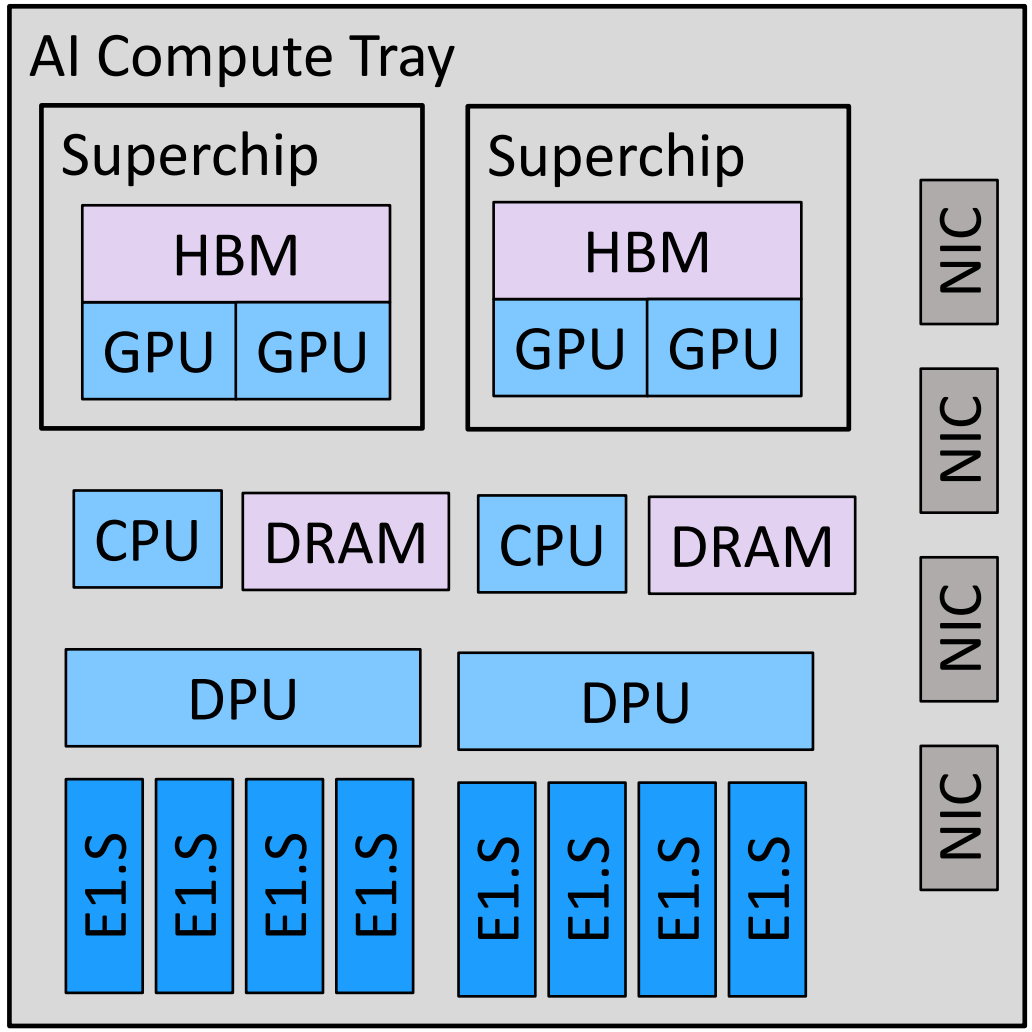


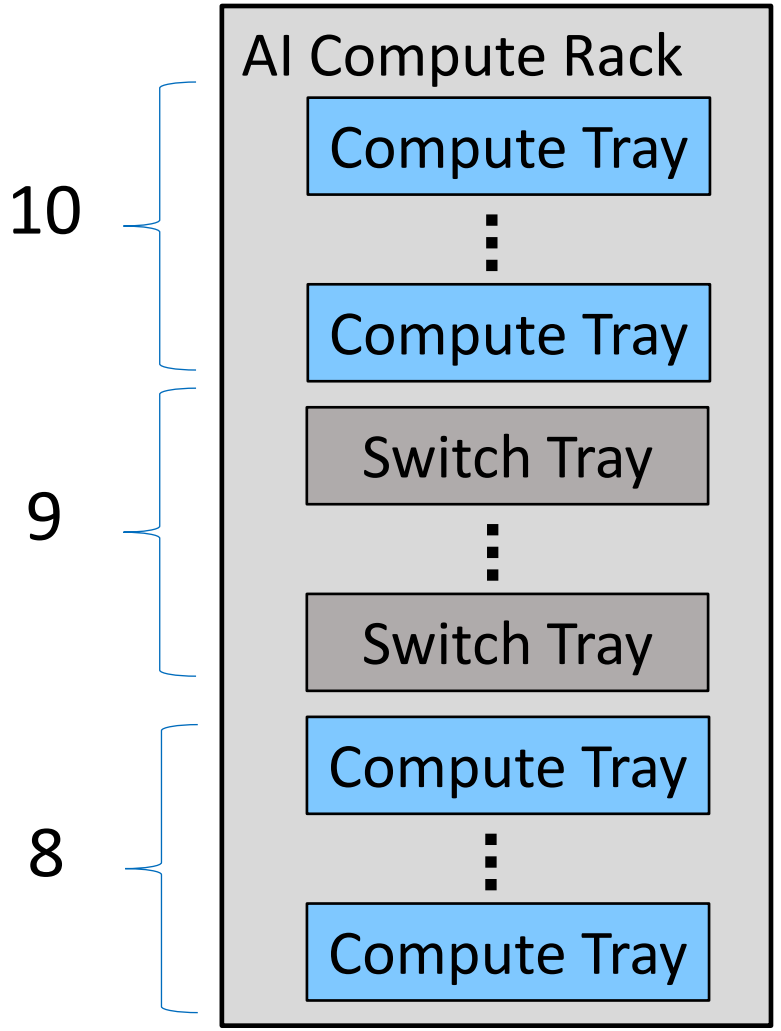
REGIONAL SDC

Training vs Inference

- Training
  - One large job on supercomputer with 10000s or 100000s 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 Configuration: GB200 NVL72

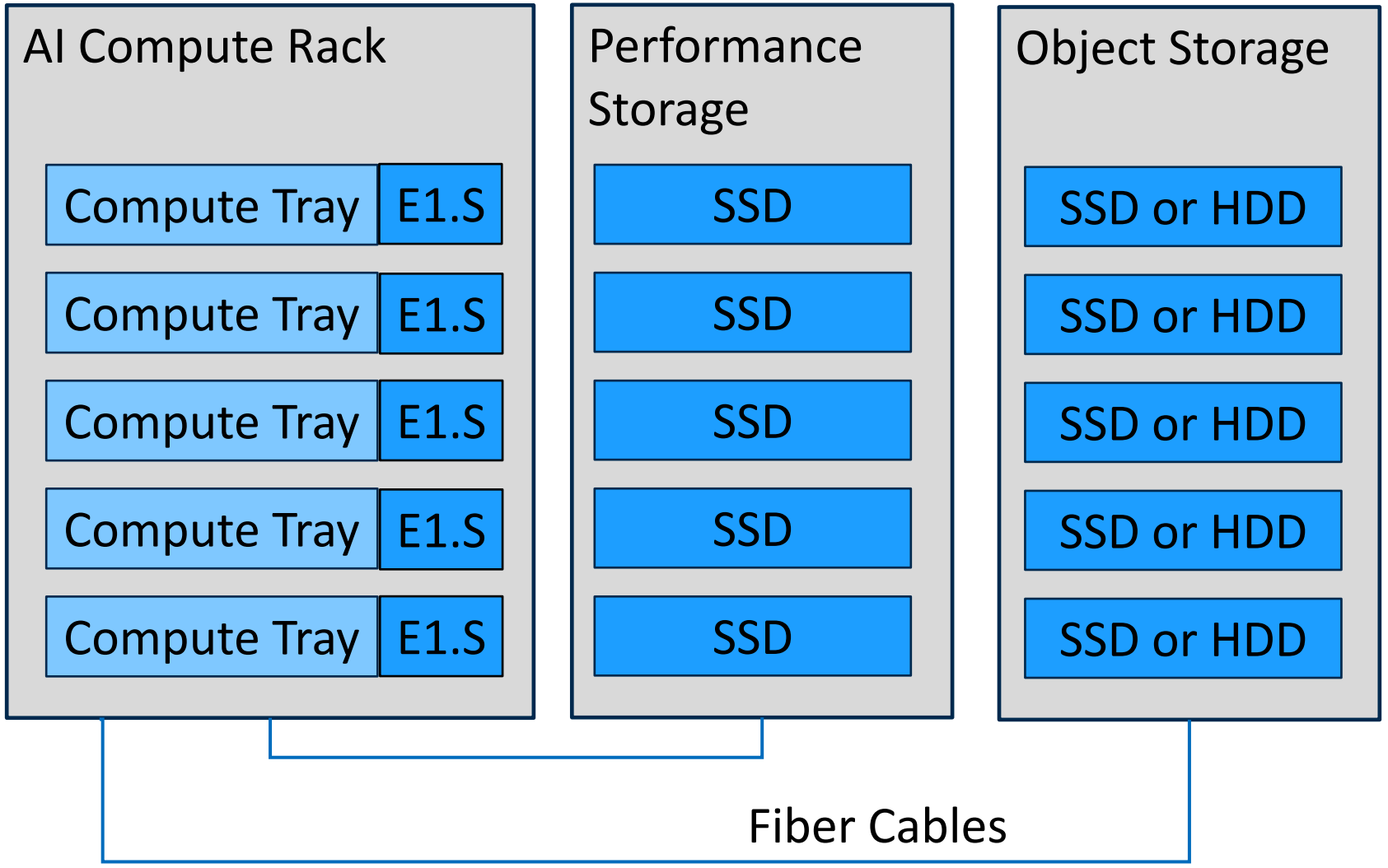




- 16 PCIe lanes per Superchip for storage
- 4 E1.S per Superchip
- 8 E1.S per compute tray
- 18 compute trays per rack
- 144 E1.s per AI compute rack

REGIONAL SDC

AI Storage Tiers



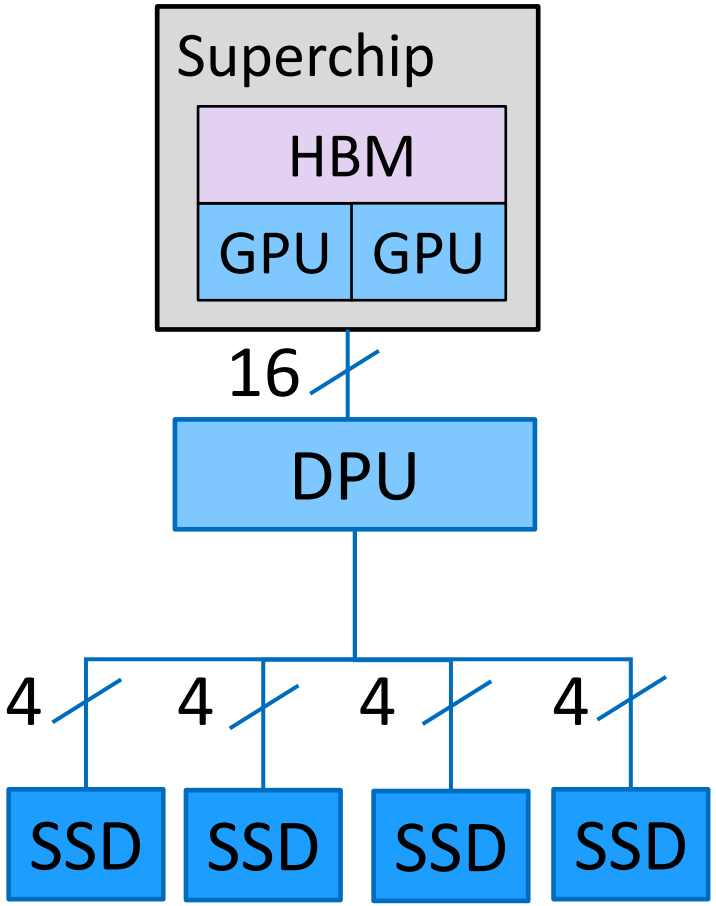
- TLC typical for compute rack and performance storage
- QLC or HDD typical for object storage
- Ethernet or Infiniband connectivity

REGIONAL SDC

SSD Performance Considerations

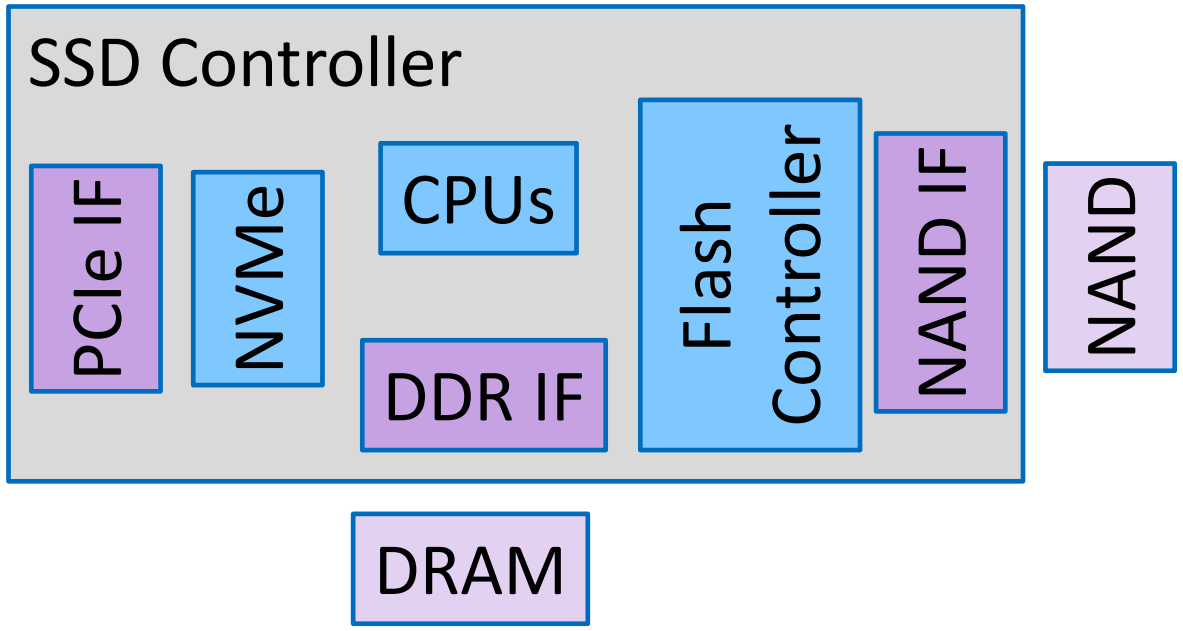
- In the past, SSDs adopted next generation PCIe interfaces later than CPUs
- AI is now driving adoption of next generation PCIe SSDs
- SSDs typically designed to saturate sequential and 4KB RR performance
- However, AI workloads (especially Inference) have random accesses smaller than 4KB.

Saturating Read Performance For PCIe Gen6



- Nvidia's Storage-Next initiative targets ~200 MIOPS (512B) for 16 lanes
  - See Nvidia presentations at SC 2024, OCP 2024 and GTC 2025
- 16 PCIe lanes seen by GPU
  - 128 GB/s Raw
  - ~110 GB/s Effective
  - ~26.8 MIOPS (4KB)
  - ~215 MIOPS (512B)
- 4 PCIe lanes seen by SSD
  - 32 GB/s Raw
  - ~27.5 GB/s Effective
  - ~6.7 MIOPS (4KB)
  - ~53.7 MIOPS (512B)

Optimizations for High IOPS SSDs



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

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

