SNIA DEVELOPER CONFERENCE

SDC 25

By Developers FOR Developers

Hyatt Regency Santa Clara, CA
September 15-17, 2025

Storage Devices for the AI Data Center

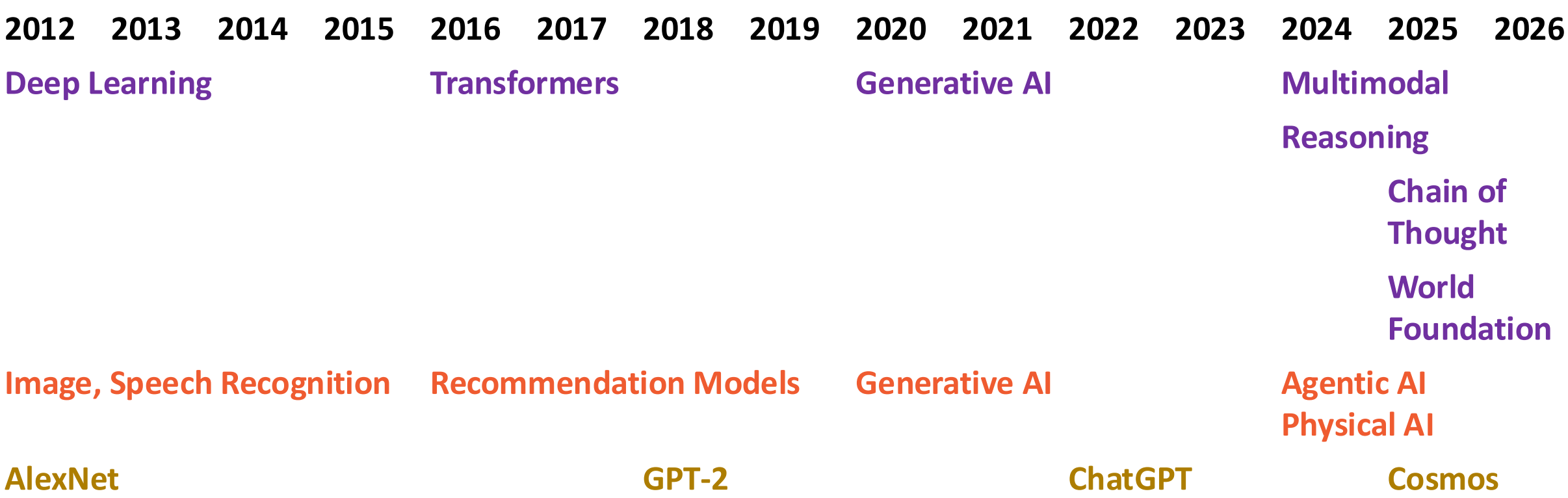
Erich F. Haratsch
Marvell

www.sniadeveloper.org

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

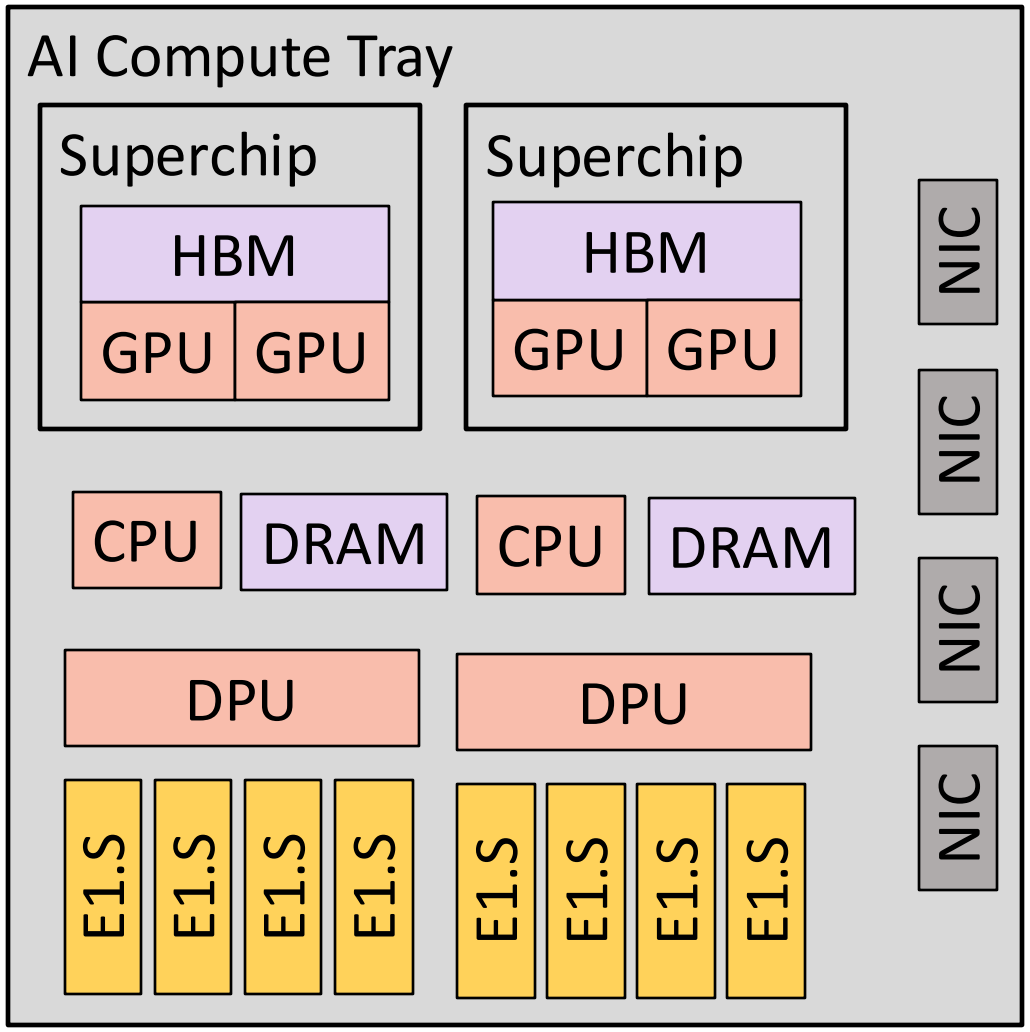
AI Processing Phases and Storage Workloads

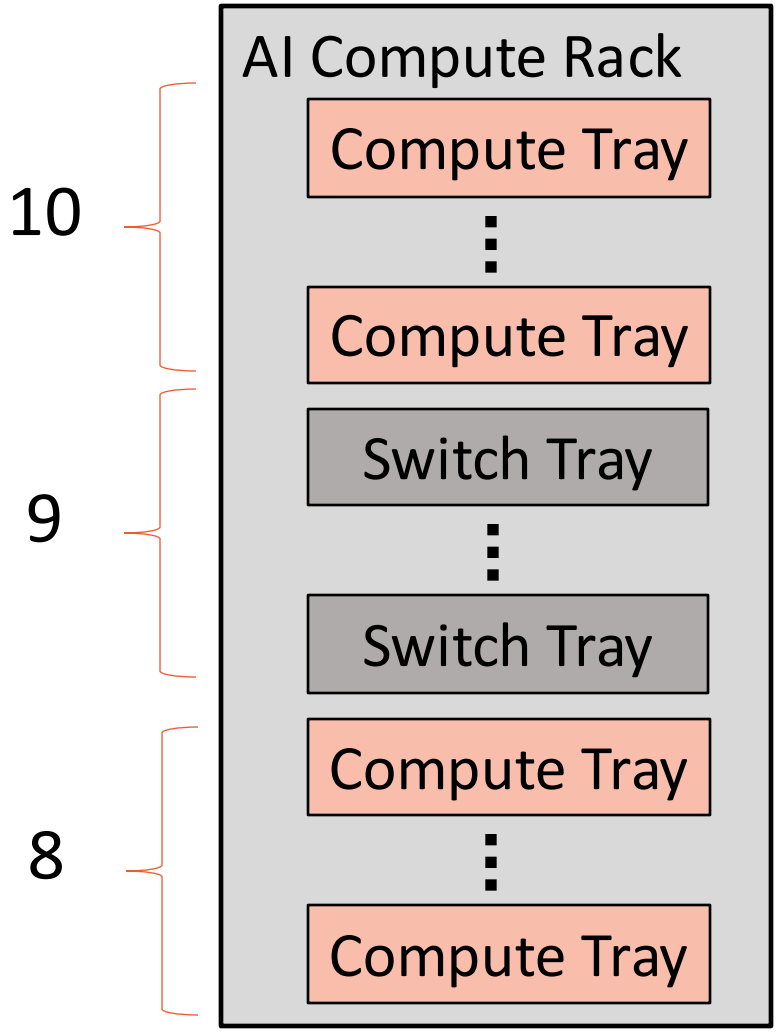
Data Ingestion	Data Preparation	Training	Inference	Archiving
Sequential	Sequential Random	Sequential Random	Random	Sequential
PBs	TBs	TBs	TBs RAG: PBs	PBs

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

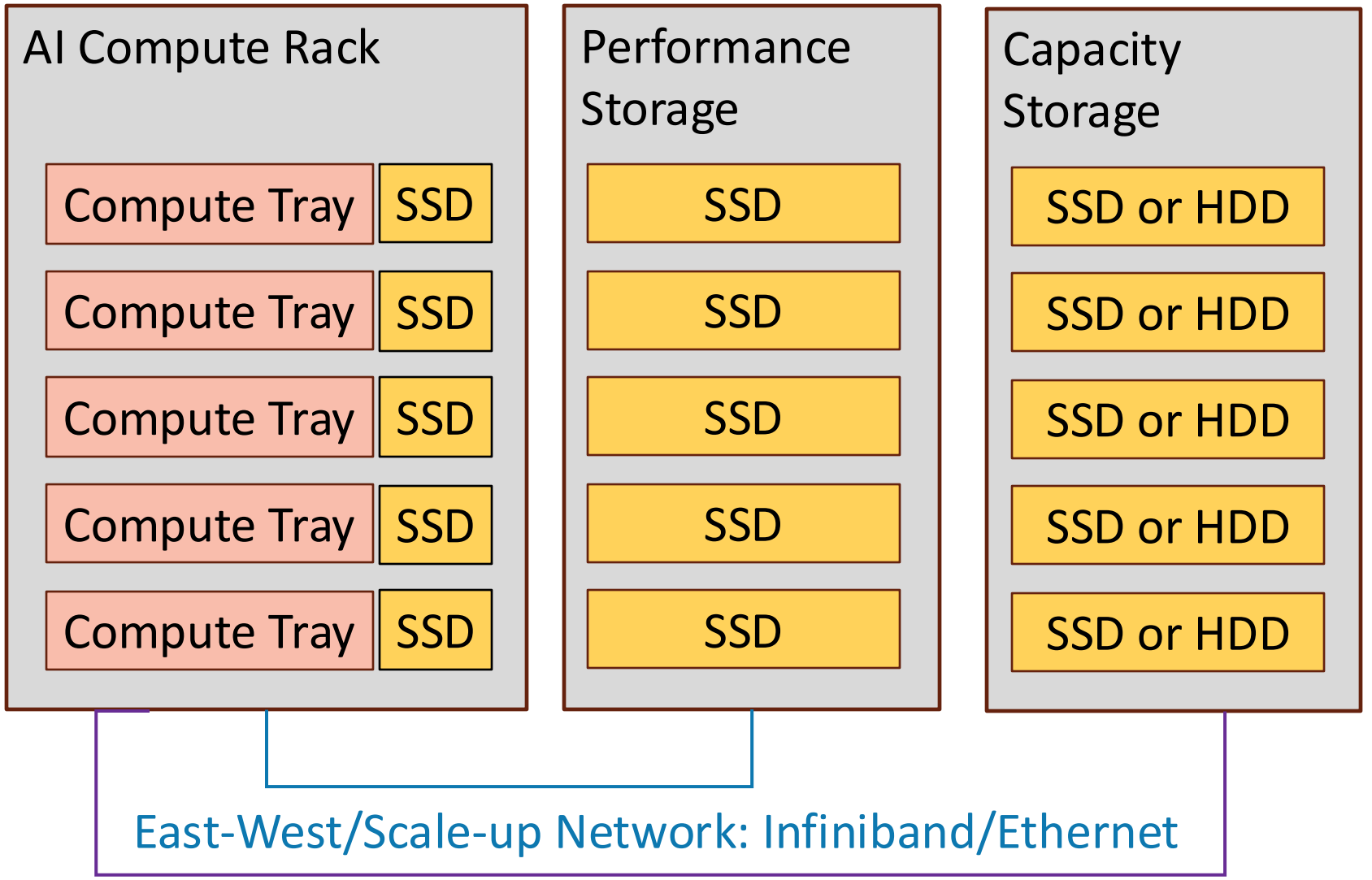
Exemplary AI Compute Tray and Rack





- Per tray:
  - 4 GPUs
  - 2 CPUs
  - 8 E1.S SSDs
- Per rack:
  - 18 compute trays
  - 72 GPUs
  - 36 CPUs
  - 144 E1.s SSDs
- SSD to GPU ratio: 2:1

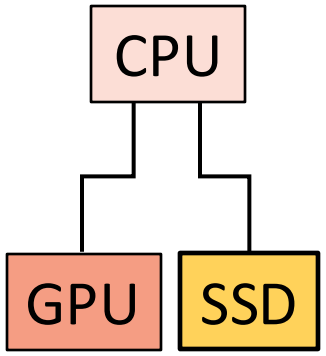
AI Storage Tiers



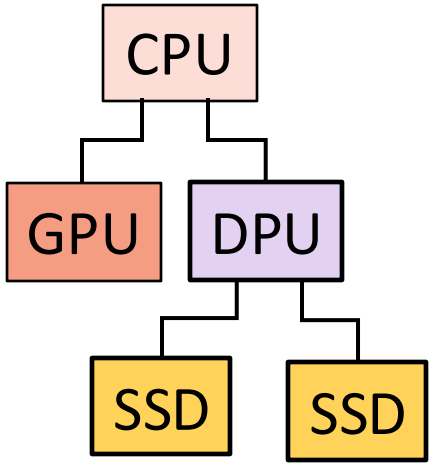
- TLC typical for compute rack and performance storage
- QLC or HDD typical for object storage
- Ethernet or Infiniband connectivity

Topology Options for Local NVMe Storage In AI Compute Nodes

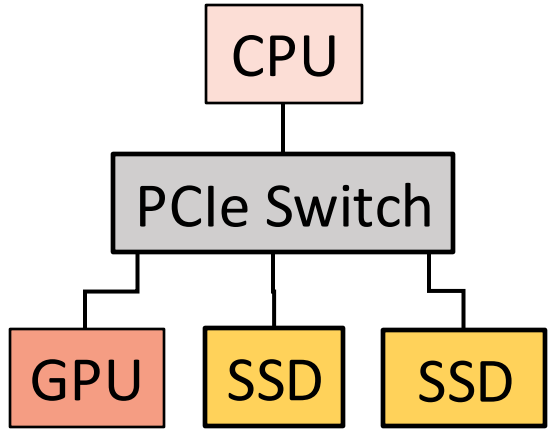
- PCI Bus
  - Number of SSDs limited by PCIe lanes of CPU



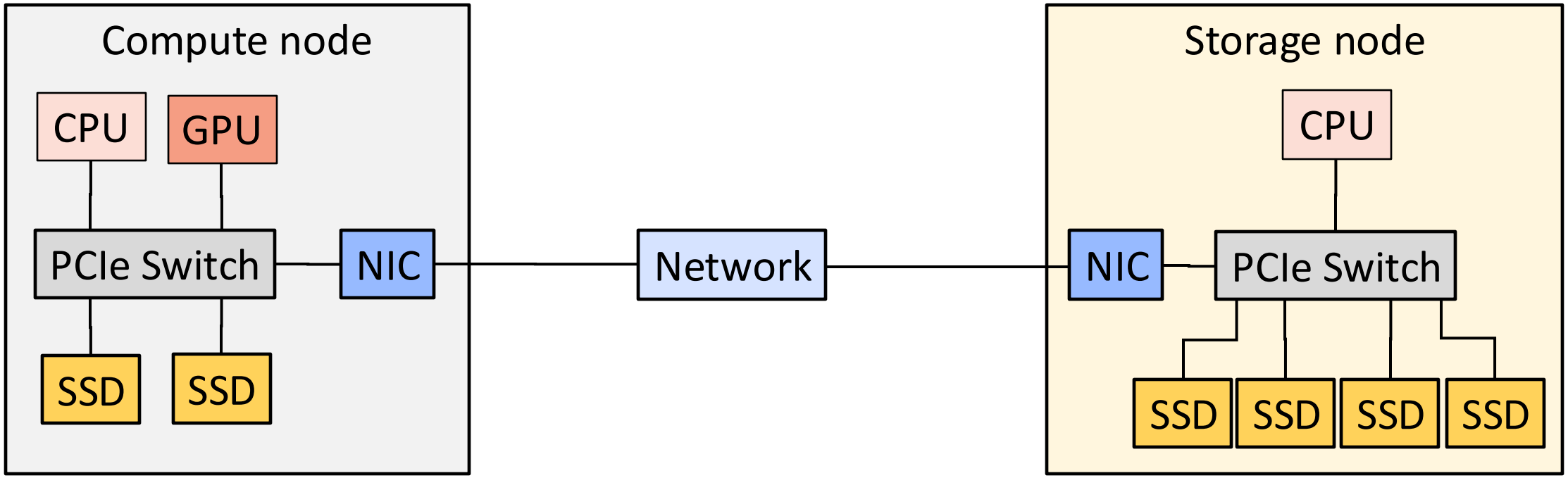
- DPU
  - Aggregates and virtualizes SSDs
  - Accelerates storage functions



- PCIe Switch
  - Scales up number of SSDs

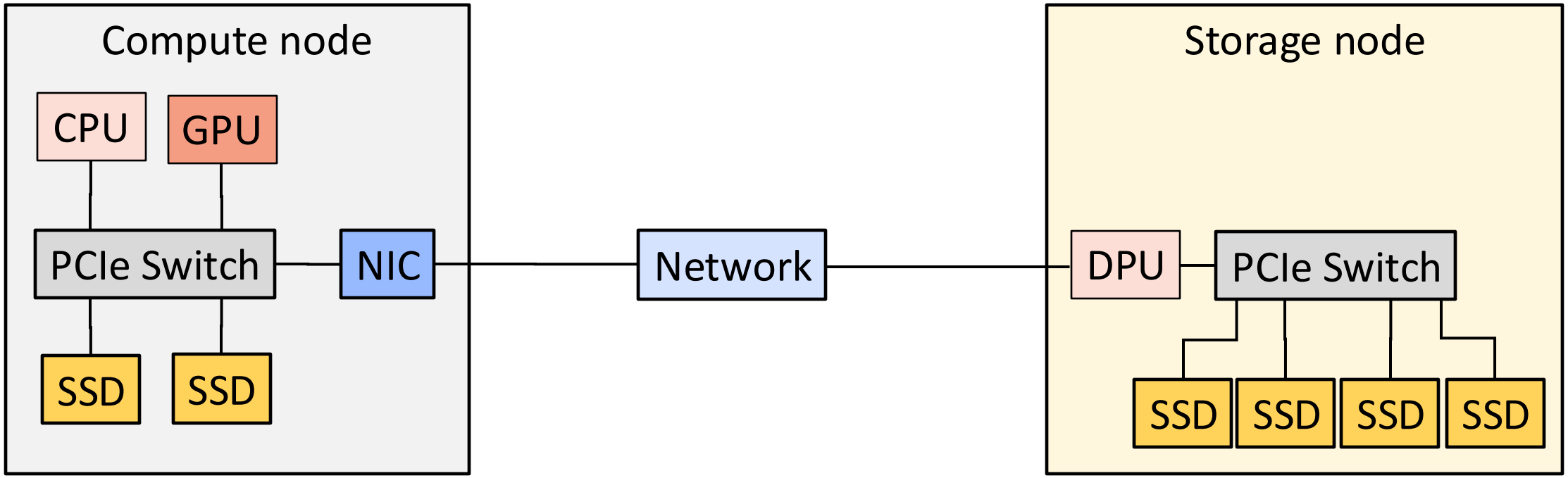


Topology for Remote Storage



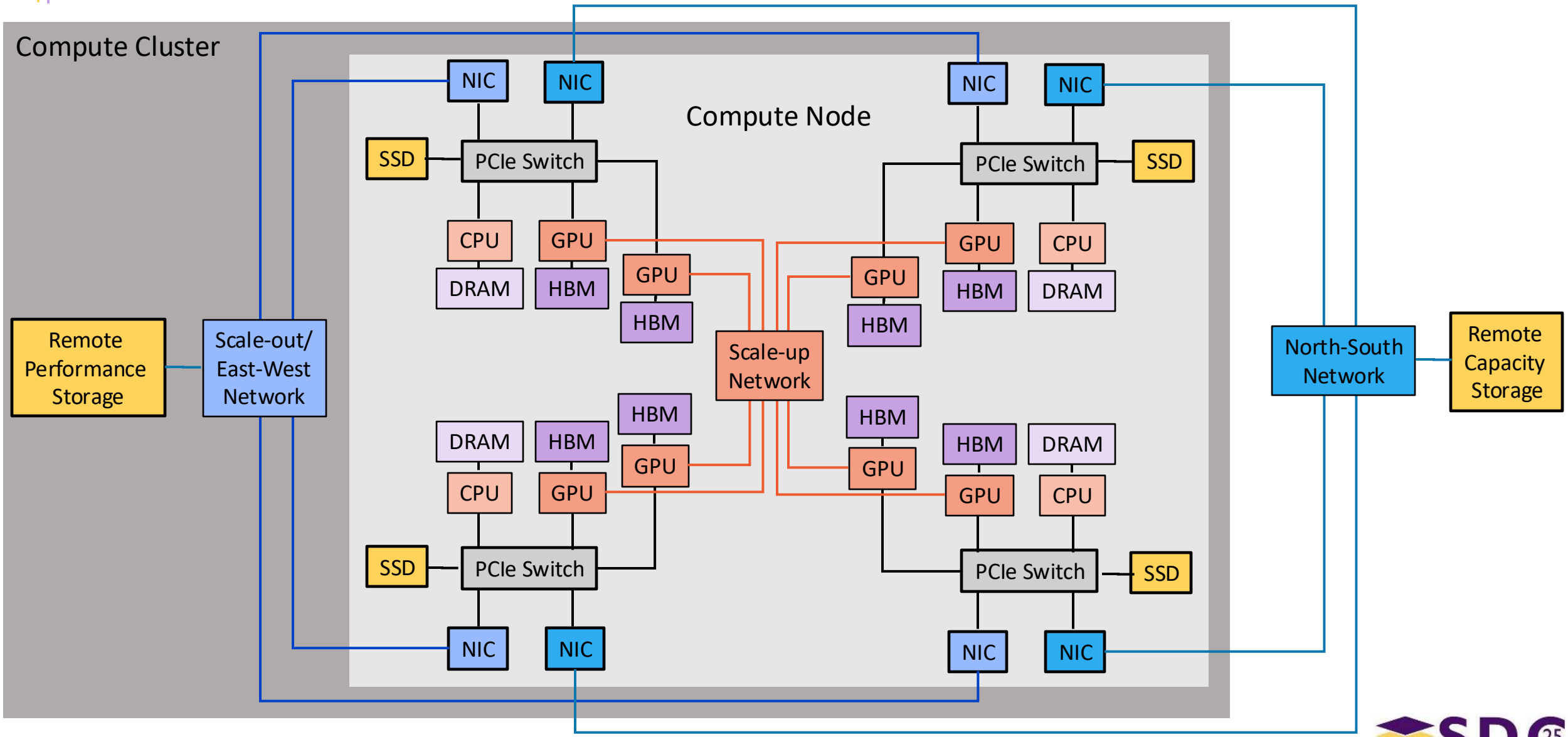
- (Ultra-)Ethernet or Infiniband for Performance Storage in East-West Network within AI Cluster
- Ethernet for Capacity Storage in North-South Network

Topology for Remote Storage

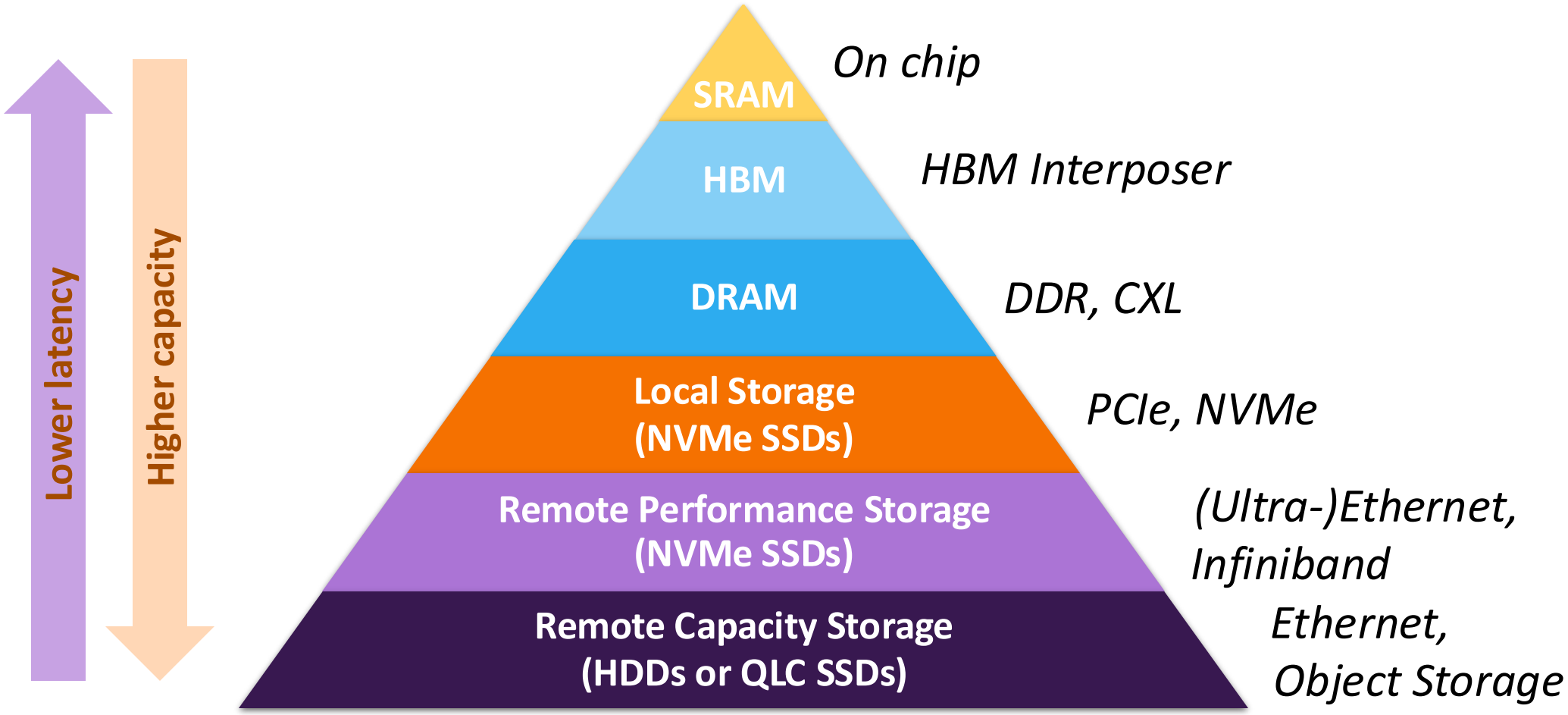


- DPU can combine function of NIC and CPU in storage node

AI Data Center Components and Networks



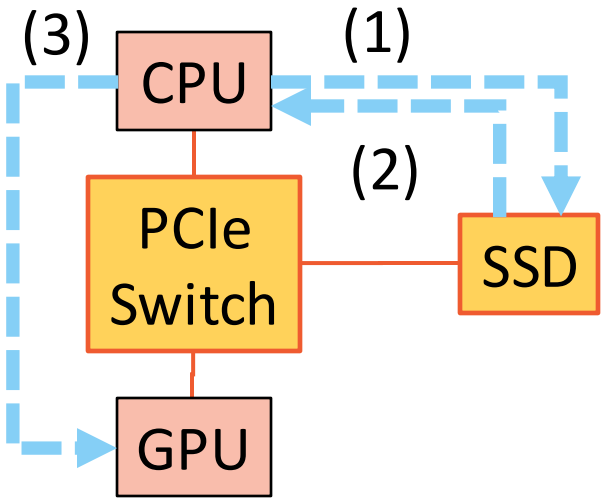
Memory and Storage Hierarchy in AI Data Center

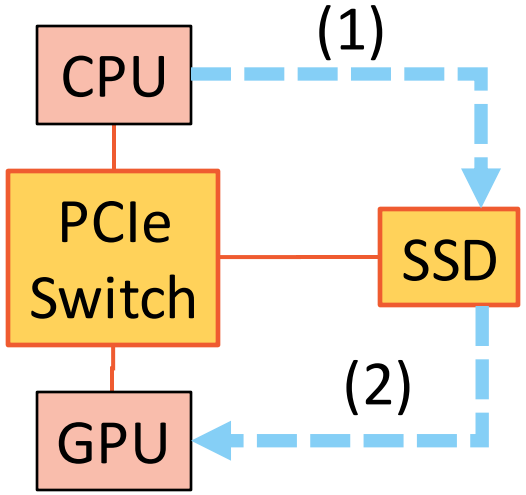


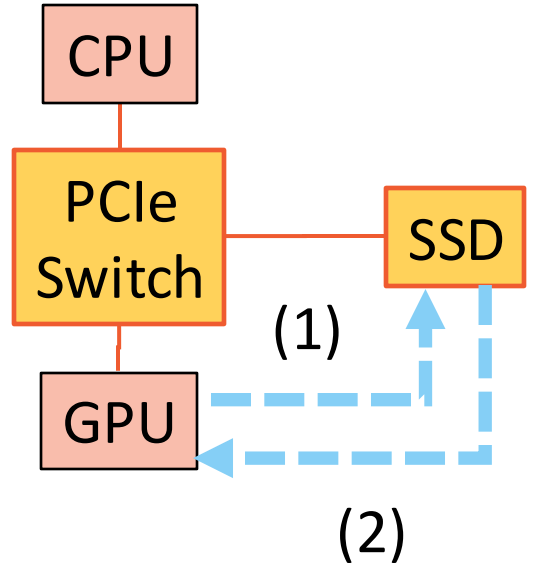
Multiple Memory and Storage Tiers

GPU Access to Storage

	CPU Storage IO	CPU Initiated GPU Storage IO	GPU Initiated Storage IO
Initiation	CPU (1)	CPU (1)	GPU (1)
Data Flow	SSD -> CPU -> GPU (2), (3)	SSD -> GPU (2)	SSD -> GPU (2)
CPU Overhead	High	Low	None
Implementation	Legacy	GPU Direct Storage	BaM (*)





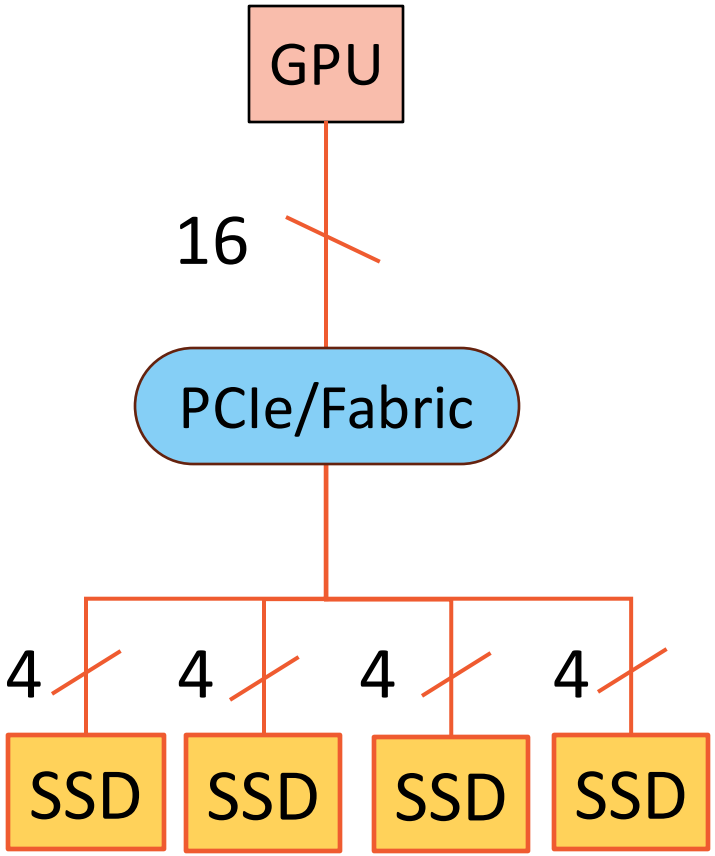


(*) Z. Qureshi et al., "GPU-Initiated On-Demand High-Throughput Storage Access in the BaM System Architecture", ASPLOS '23

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



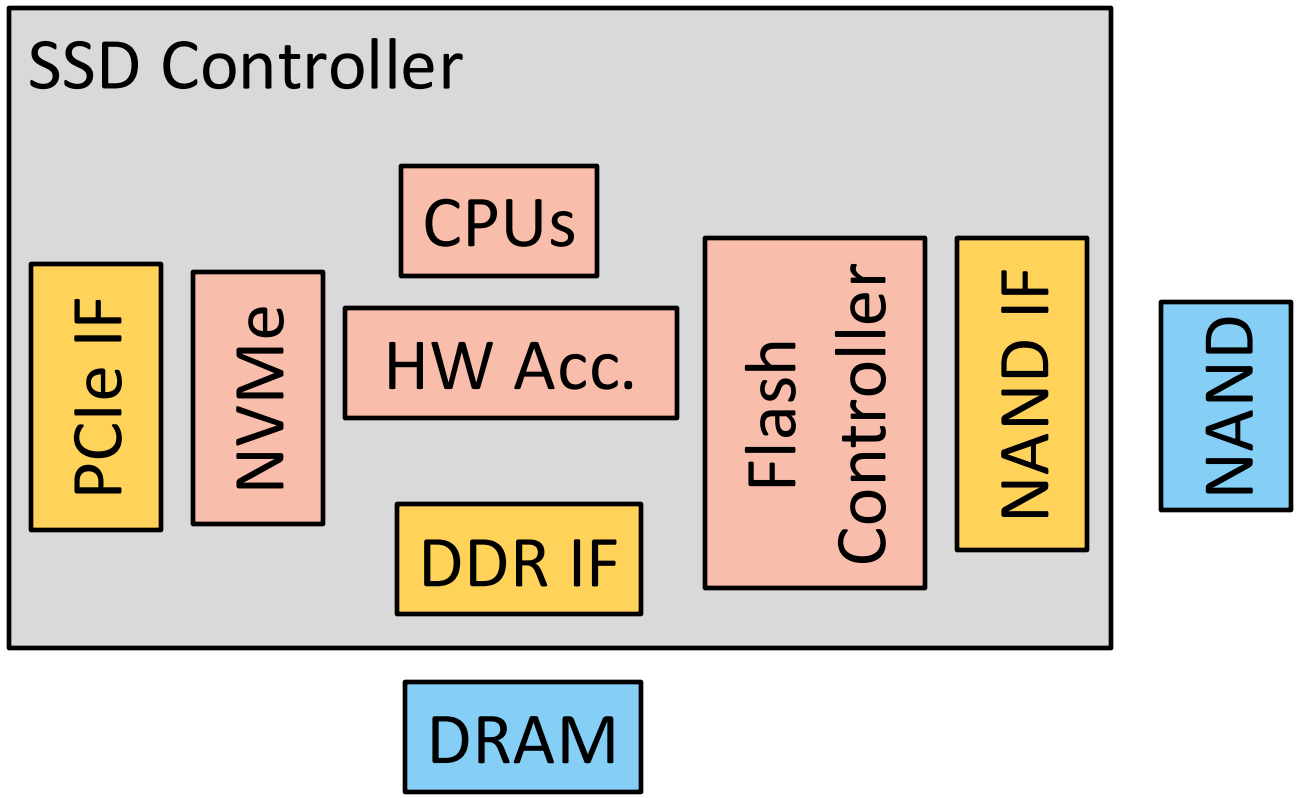
- Nvidia's Storage-Next initiative targets ~200 MIOPS (512B) for 16 PCIe lanes
  - See Nvidia presentations at OCP 2024, SC 2024, GTC 2025
- 16 PCIe lanes per GPU
  - 128 GB/s Raw
  - ~110 GB/s Effective
  - ~26.8 MIOPS (4KB)
  - ~215 MIOPS (512B)
- 4 PCIe lanes per SSD
  - 32 GB/s Raw
  - ~27.5 GB/s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

Saturating GPU Performance with Multiple SSDs

PCIe	MIOPS Target per GPU	4KB Optimized SSDs		512B Optimized SSDs	
		MIOPS (eff)	Number of SSDs	MIOPS (eff)	Number of SSDs
Gen5	100	3.125	32	25	4
Gen6	200	6.25	32	50	4
Gen7	400	12.5	32	100	4
Gen8	800	25	32	200	4

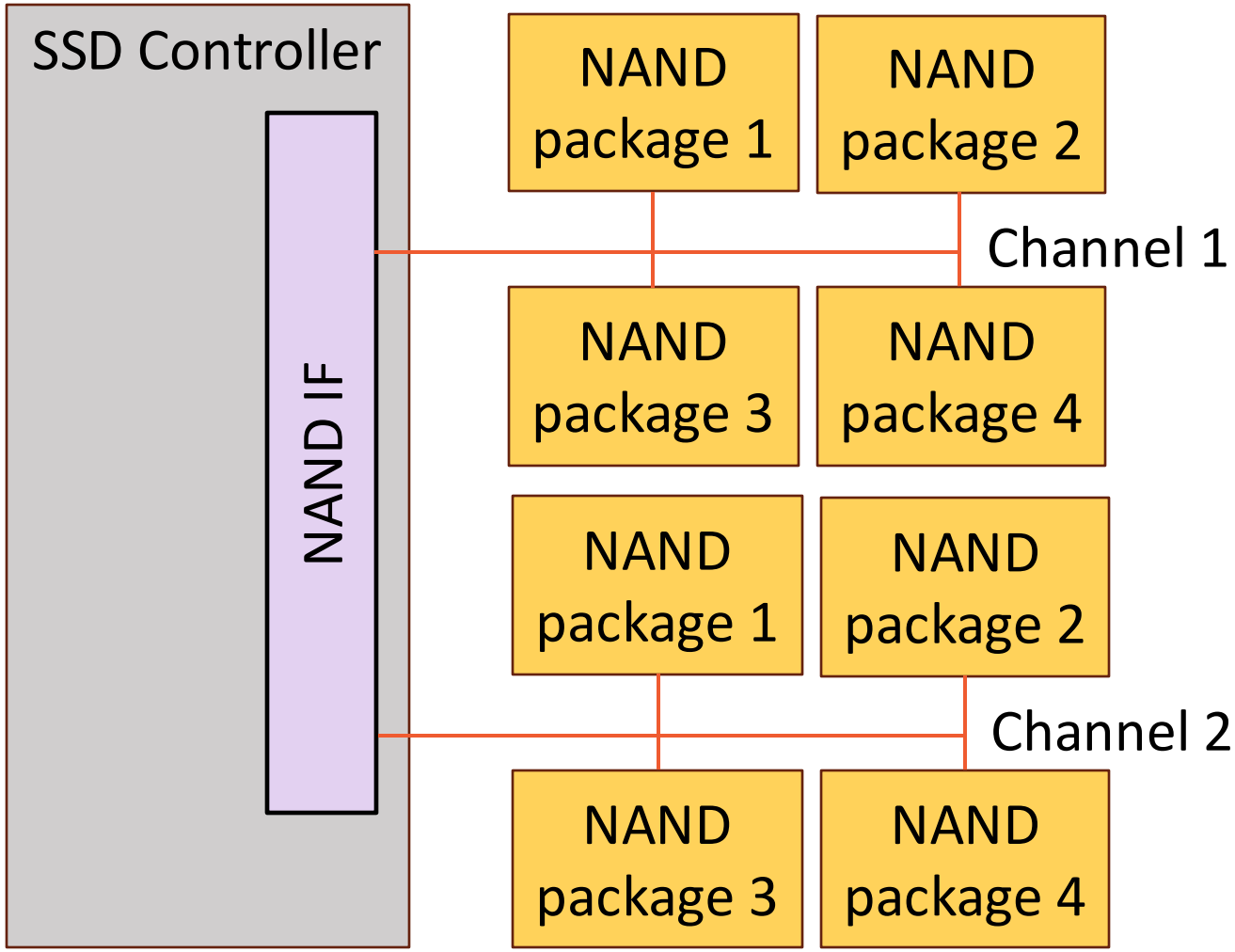
- Improving IOPS over current design points reduces number of SSDs needed to saturate GPUs
  - For example,16 SSDs with 12.5 MIOPS can saturate Gen6 512B performance

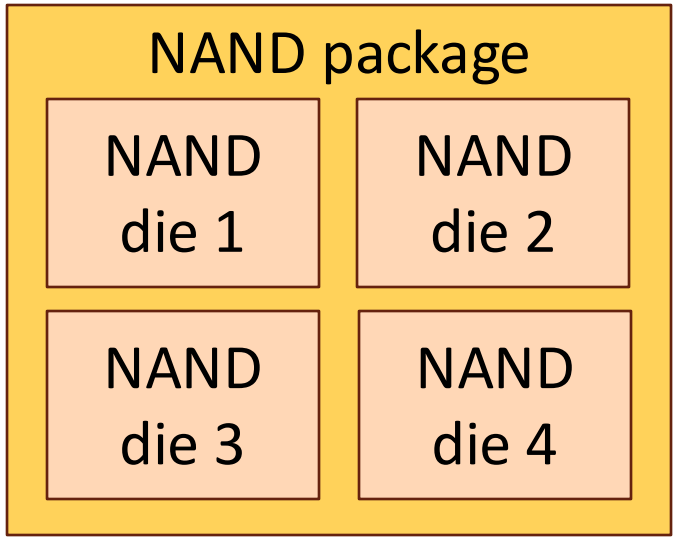
Optimizations for High IOPS SSDs

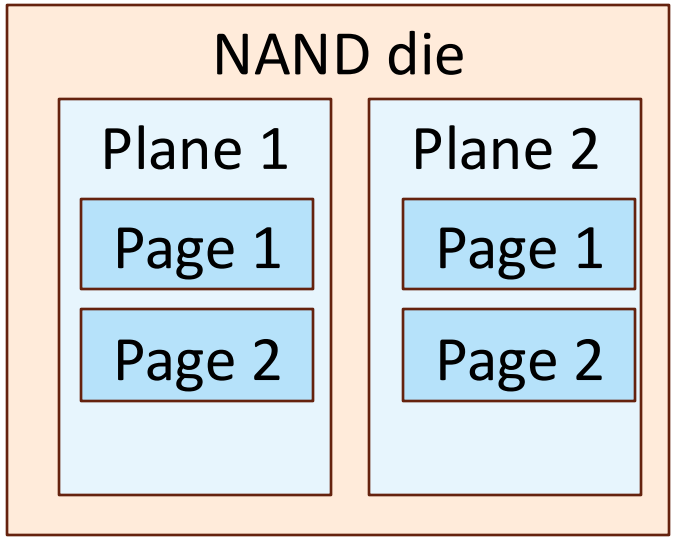


- NAND
  - Read Time
  - SLC, MLC vs TLC NAND
  - Page and Plane Architecture
- SSD Controller
  - CPUs
  - HW acceleration for FW Offload
- Host interfaces
- Form factors
  - 8x random read performance will increase power

IOPS and Read Parallelism







Random Read Performance

- Read time for one page
  - Page size and read unit size
  - ECC size
  - Low latency (SLC) vs TLC NAND flash
- Read parallelism
  - Number of concurrent page reads within die
  - Number of dies per channel
- NAND channel BW
  - NAND channel speed
  - Number of NAND channels
  - NAND channel efficiency
- SSD Controller
  - IOPS capability

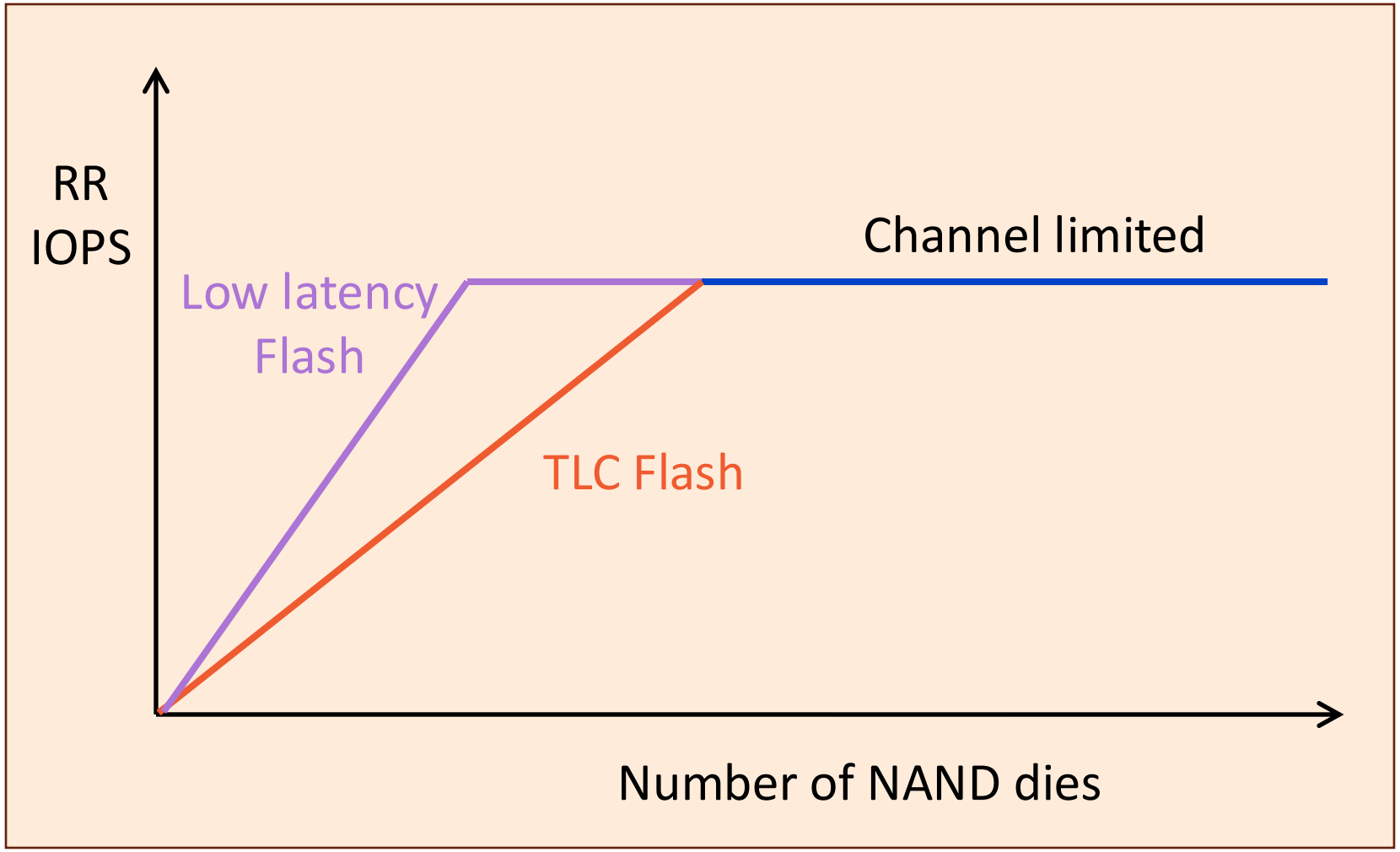
SSD Performance and NAND Speed

PCIe	Performance per SSD (4 PCIe lanes)		Typical NAND Speed
	GB/s (*)	4KB IOPS (**)	MT/s
Gen5	16	3.125	1600
Gen6	32	6.25	3200
Gen7	64	12.5	6400 (***)
Gen8	128	25	12800 (***)

(*) raw
(**) effective
(***) extrapolated

- Small IOs over NAND channel decrease NAND channel efficiency due to command overheads
- Need to either
  - Reduce NAND command overhead
  - Increase NAND channel speed
  - Use more NAND channels

Random Read Performance and SSD Capacity



- IOPS at high capacities limited by effective NAND channel BW
- Low latency NAND reduces required SSD capacity to saturate performance
- However, low latency flash more expensive than TLC

RR: Random Read

Form Factor Considerations

	E1.S	E1.L	E3.S	E3.L	E2 (New)
Use case	High density	High density and capacity	High performance	High performance and capacity	Near line storage
Max Capacity (2024)	~8TB	~64TB	~16TB	~128TB	>256TB (planned)
Recommended power dependent on FF height	12/16/20/25W	25/40W	25/40W	40/70W	<80W

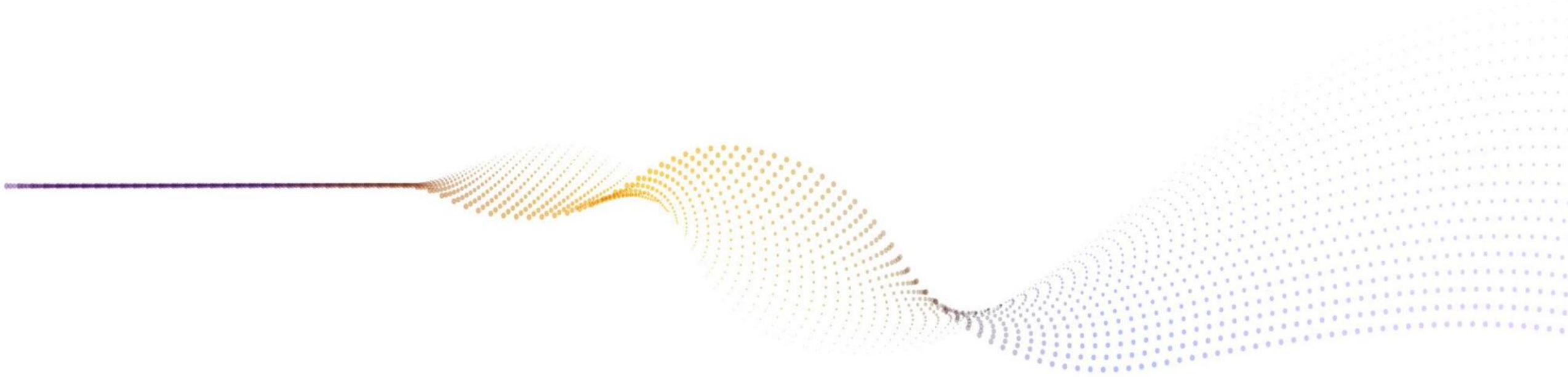
- Current SSDs optimized for 4KB IOPS
  - Increasing IOPS capability by 8x will increase power
- Need to determine best form factors for high IOPS SSDs
  - Dimensions
  - Power budgets
  - Cooling
  - Targeted capacities

Flash Storage: Standards and Industry Working Groups

- Interfaces and protocols
  - PCIe, NVMe, JEDEC NAND and DDR
- Form factors and connectors
  - SNIA SFF
- NVMe datacenter SSDs
  - OCP Storage
- Storage for AI
  - Storage-Next (Nvidia led)
  - SNIA Storage.AI
  - Ultra Ethernet

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



Thank you for attending!

Please remember to rate this session. You get access the presentations at
http://sniadeveloper.org/conference

