SDC | StorageAI™

A SNIA Event

April 29, 2026 • Denver, Colorado

AI Server Clusters – Networking and Storage

David L. Black, Ph.D.
Sr. Distinguished Engineer, Dell

Topic Outline

- AI Workloads and Storage
- Networks for AI Server Clusters
- Storage Networking for AI Server Clusters
- Wrap-Up and Conclusions

SDC | StorageAI™

A SNIA Event

AI Workloads and Storage

AI Workloads: AI Data Pipeline Stages

AI Data Pipeline

Data Ingestion

Structured and unstructured raw data is stored.

Data Preparation

Wrangling: Processing of raw data to clean, merge, and transform it for use during model training and inference.

Model Training Fine Tuning, Validation (Model Development)

- **Dataloading (Feeding GPUs):** providing input data to the model training pipeline
- **Checkpointing**: saving model state to resume training after failures or pauses
- **Restoring**: reloading of model state from checkpoints to resume training

Model Inference (Model Deployment)

- **KV-Cache:** Caches attention states during inference to speed up generation
- **RAG:** Improves responses by retrieving relevant context from a **Vector DB** and injecting it into the prompt before generation.
- **Vector DB:** Stores embeddings for fast similarity search in RAG workflows.
- **GraphDB:** Provides structured, relationship-aware context for reasoning tasks

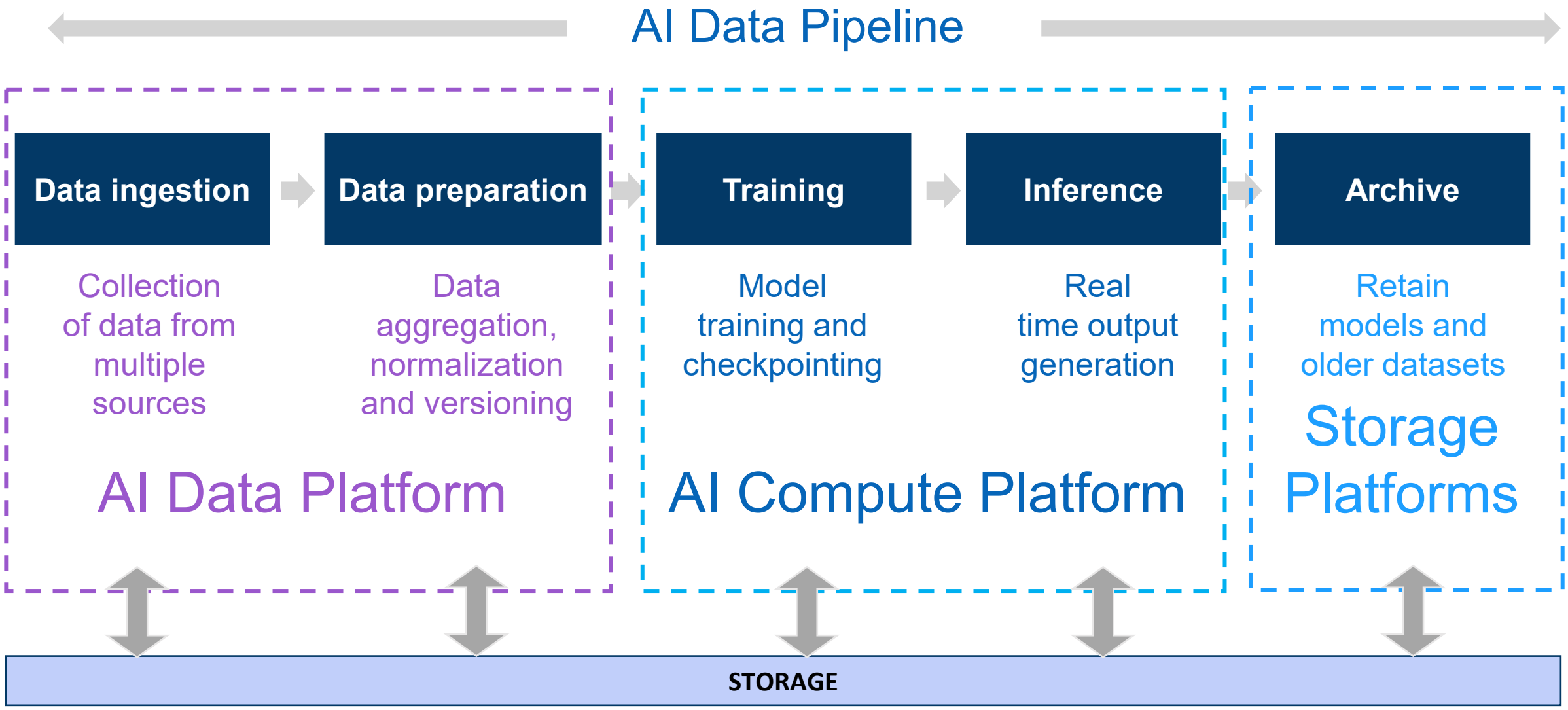
Archive

- **Store** datasets, model checkpoints, logs, and metrics for long-term retention.
- **Support** auditing, compliance, reproducibility, and future retraining needs.

AI Data Pipeline

AI Is Not One Workload: Storage at Every Stage

Each phase of Generative AI has different workloads and associated storage requirements



All of the Pipeline Stages use Storage … Differently!

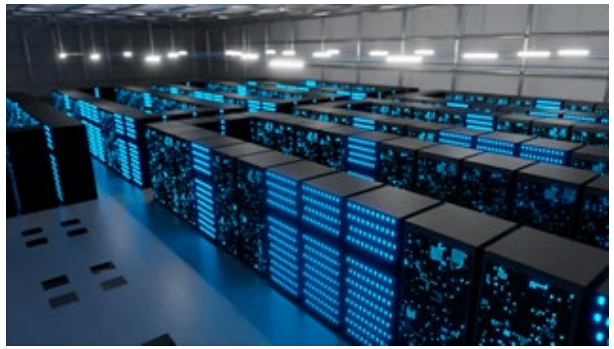
Storage Access Paradigms, Block, File, and Object

- Block (e.g., NVMe, NVMe over Fabrics)
- File (e.g., NFS, parallel NFS)
- Object (e.g., S3)

Workload characteristics

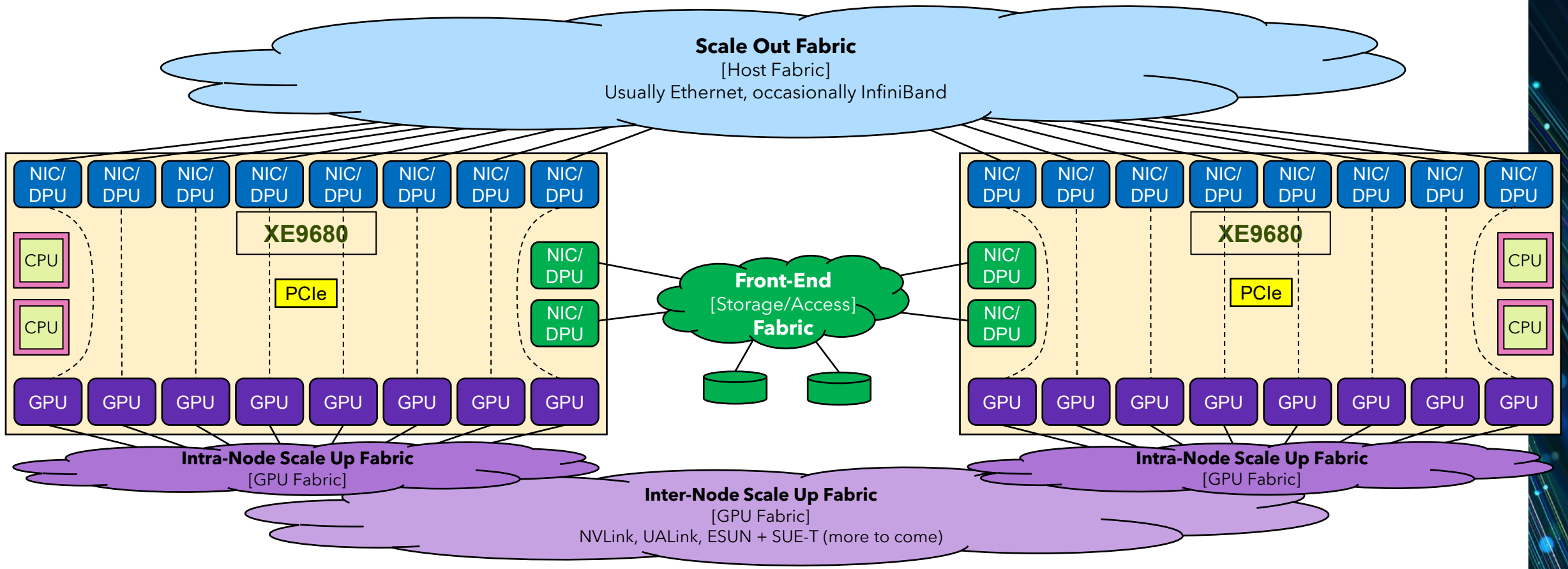
- Read/write mix, I/O sizes
- Sequential/random mix, locality

- Parallelism and concurrency
  - Data Preparation can be embarrassingly parallel
  - Archive access is often single stream




SDC | StorageAI™
A SNIA Event
Networks for AI
Server Clusters -

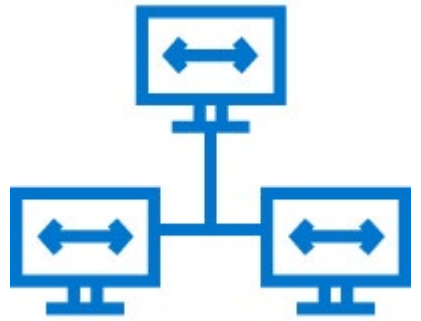
AI Server Cluster (using Dell XE9680s as example)



Three AI Network Fabrics:

1. Front-End [Storage/Access]: Connectivity to rest of datacenter and the outside world (Ethernet)
2. Scale-Out [GPU]: AI Cluster interconnect (usually Ethernet, occasionally InfiniBand),
3. Scale-Up [GPU subset]: AI Sub-Cluster very low-latency interconnect (specialized, e.g., NVLink, UALink)

1. Front-End [Storage/Access] Fabrics



Conventional data center networks

- Carry AI-related traffic … and everything else
- Scalable network topology/architecture (e.g., multi-spine) is important.

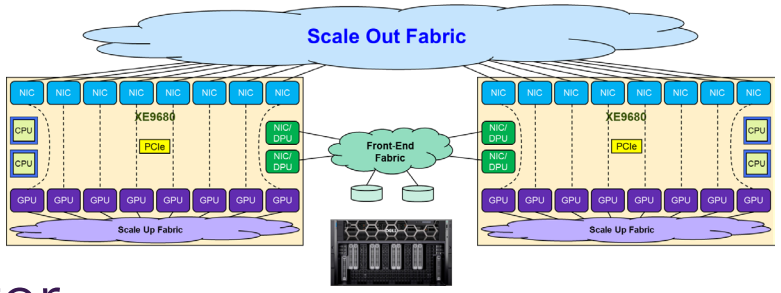
Designed for TCP and related traffic

- RDMA usage increasing, pay attention to TCP / RDMA coexistence
- Network may be converged, or separate network(s) for storage, etc.

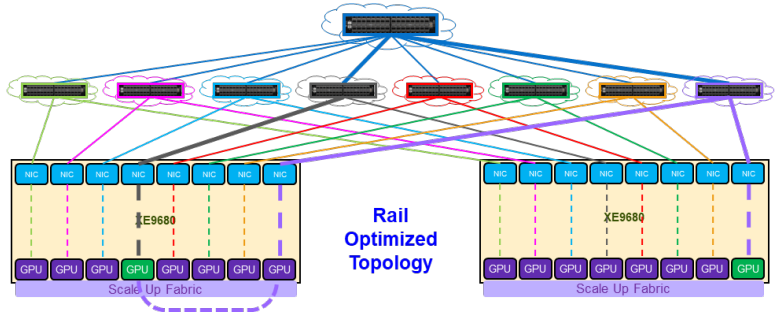
AI traffic impacts on Front-End networks

- Significant increases in network traffic
- Has not (yet) driven major changes in network topology/architecture

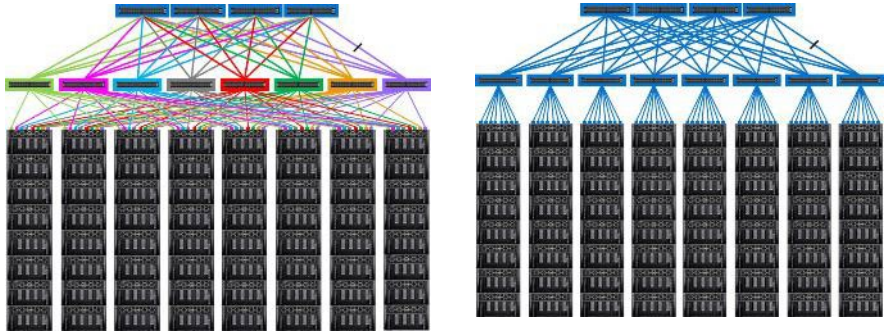
2. Scale-Out [GPU] Fabrics



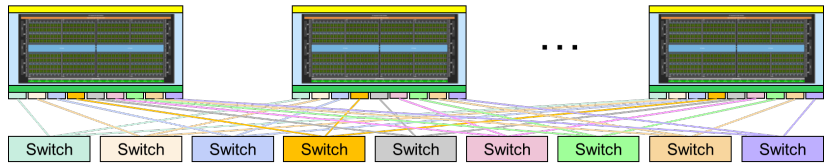
- GPU parameter exchange across entire Server Cluster
  - Scale: Span most or all of a data center
- Typically Leaf-Spine: Rail-Wired or TOR-Wired
  - Two layers of switches, cross-rack connectivity
  - Trending to stacked planes to avoid three-tier networks
- Scale Out Network Fabric characteristics:
  - Multipathing & congestion control with host involvement
  - Designed for RDMA Traffic, can carry TCP traffic
  - Usually IP Routable (IP addresses, IP headers)
  - Storage typically not attached
- Examples: Ultra Ethernet, Spectrum X



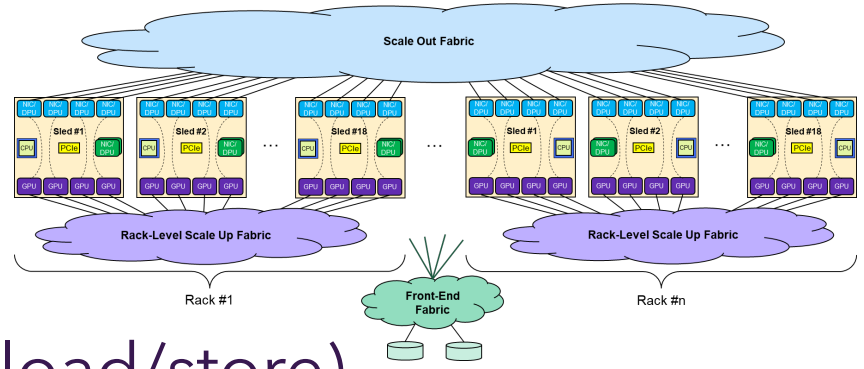
Rail-Wired vs TOR-Wired



3. Scale Up [GPU subset] Fabrics

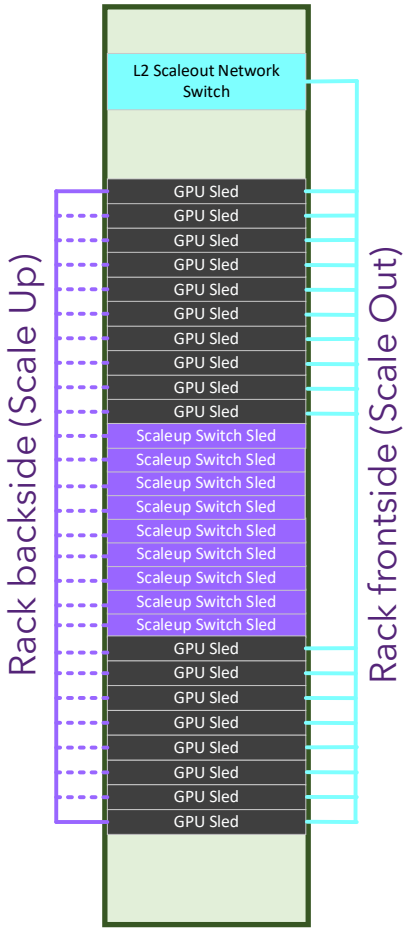


- High bandwidth, highly parallel GPU interconnect
  - Rack Scope (Scale-Out scope is much larger)
  - Used in conjunction with Scale-Out network for single AI application
- Scale Up Topology (current)
  - Single Layer of Switching
  - Stacked layers (network planes)
  - Server selects plane, no cross-plane connectivity



- Network traffic: Remote memory operations (load/store)
  - Packet typically contains multiple memory operations, coherent memory access
- Examples: NVLink, UALink, ESUN + SUE-T
  - Not IP routable (no IP addresses, no IP headers)
  - Storage not attached (currently)

Scale Up and Scale-Out: GPU Connectivity (1 rack)



Example: NVL72 rack architecture:

- 72 GPU chips (18 sleds x 4 chips/sled)
- Scale-Up: 18 network planes (9 sleds x 2 planes/sled)
  - Each GPU chip connects to each plane (72 x 18 = 1296 ports)
- Scale-Out: 2 top-of-rack switches (48 ports each)
  - Each GPU chip (indirectly) connects to one switch
  - 72 ports for GPUs, 24 uplink ports

Takeaways:

1. Scale-Up bandwidth much larger than Scale-Out
2. Scale-Up scope much smaller than Scale-Out
3. Scale-Up constrains # of tightly-coupled GPUs

Three AI Network Fabrics

1. **Front-End** [Storage/Access]: AI Compute Platform connectivity to rest of datacenter and the outside world (Ethernet)
2. **Scale-Out** [Host]: AI Cluster interconnect (usually Ethernet or variant, occasionally InfiniBand),
3. **Scale-Up** [GPU]: AI Sub-Cluster very low-latency interconnect (specialized, e.g., NVLink, UALink)

	Front-End	Scale-Out	Scale-Up
Connectivity Class	[Storage/Access]	[GPUs]	[GPU subset]
Typical Latency - less than:	20us (in data center)	5us	1us
Semantic model	Packets, RDMA optional	Packets with RDMA	Memory Load/Store
Scope	Datacenter + external	AI Cluster (rows to full DC)	AI Rack
Topology	Multi-Tier Data Center	Leaf-Spine	Stacked Planes
Traffic Classes	Multi-protocol	AI Cluster only	GPU-GPU only
Network Fabric Class	**General Purpose**	**Cluster Optimized**	**Highly Specialized**

Emerging network fabric classes:

a. Scale-Across: Extend/Connect data center Scale-Out networks across campus or metro distances

b. Scale-In: Coax cables and/or optics (e.g., co-packaged) to overcome PCB trace length limitations


SDC | StorageAI™
A SNIA Event
Storage Networking for AI Server Clusters

Storage Networking for AI

Predominantly File (e.g., NFS) and Object (e.g., S3)

- Strong trend towards Object, particularly at larger scale.
- Limited use of Block (e.g., NVMe-over-Fabrics)
- If SSDs used, each SSD usually has a physical file system (e.g., ext4)
  - File transfer to/from AI servers: Distributed or Parallel filesystem (e.g., NFS, parallel NFS)

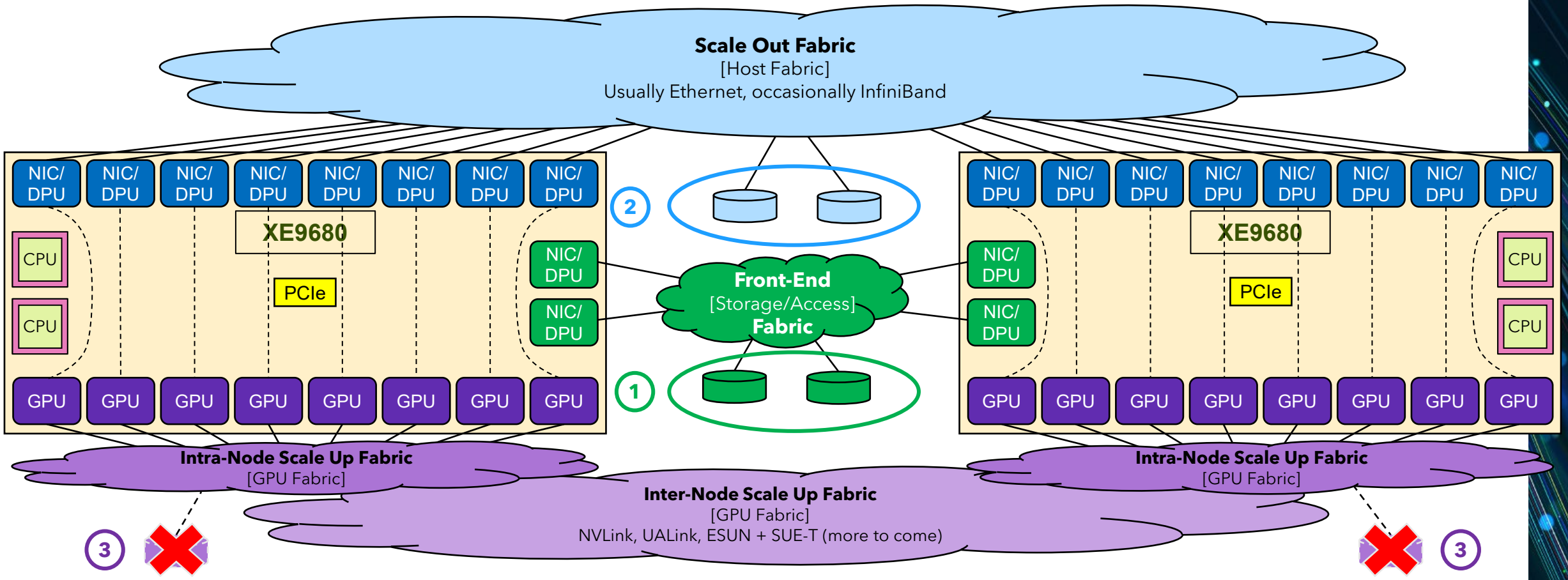
RDMA is essential for massive IOPS and bandwidth requirements

- RDMA data path is in hardware (e.g., NFS/RDMA, S3/RDMA, NVMe/RDMA)
- TCP partial offload optimizations in NICs do not perform as well

New Development: KV (Key-Value) pairs for KV Cache

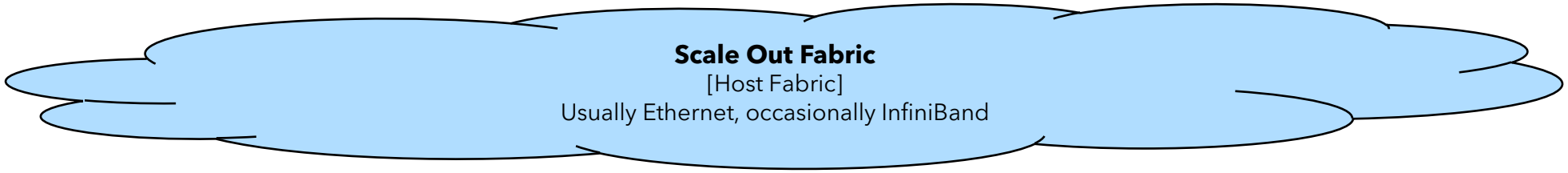
- Server KV cache extended to external KV Cache, shared among GPUs
- Stored entities: Simple key-value pairs, e.g., no S3 user metadata
- NVMe Key Value Command Set – good functional fit (surprise!)

Storage Networking Progression for AI



1. Initial situation: External Storage attached to Front-End (Data Center) Fabric e.g., NFS, NFS/RDMA, S3, S3/RDMA
2. Next Step: External Storage Attach to Scale-Out Fabric
3. Storage attach to Scale-Up [GPU subset] Fabric: Challenging due to Memory Access Semantics & Network Addressing

External Storage on Scale-Out Fabric



- Scale-Out Network Fabric: Usually Ethernet, Occasionally InfiniBand
  - RDMA traffic predominates
- Scale-Out Network Fabric: Highly Parallel
  - Opportunity: Scale-Out Server Cluster with Scale-Out access to Scale-Out Storage
- Storage on Scale-Out Fabric: Workloads Matter
  - Start with the most demanding workloads, e.g., Checkpoint, KV Cache
  - Motivation: If a workload is doing well on Front-End Fabric, why move it?
- AI workload demands evolve (typically increase) over time, e.g.:
  - Agents and Reasoning increase RAG coupling to LLM models
  - Results in more Vector DB access demand at lower latencies

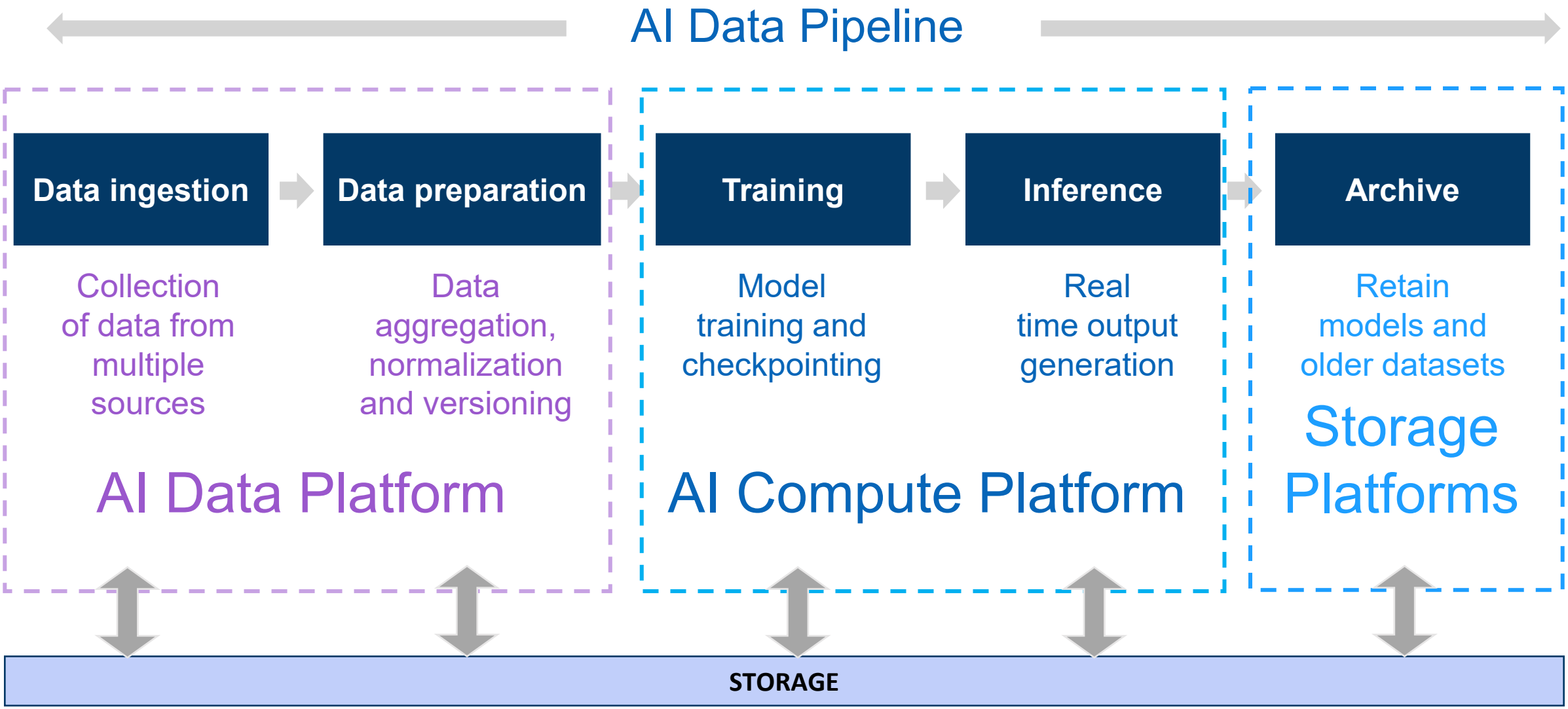
SDC | StorageAI™

A SNIA Event

Wrap-Up and Conclusions

AI Is Not One Workload: Storage at Every Stage

Each phase of Generative AI has different workloads and associated storage requirements



Conclusions

Image Credit: IconScout

- Storage Networking for AI: File & Object predominate
  - RDMA is Essential for high bandwidth
  - I/O will be increasingly GPU-driven, Integrate with GPU frameworks
- Network Fabrics for AI
  - Front-End: Connect AI Server Cluster to rest of data center
  - Scale-Out: Span the entire set of GPUs used by an AI Application
  - Scale-Up: Low latency GPU subset interconnect (e.g., all GPUs in a rack)
- Storage Attach to AI Servers: Focus on Scale-Out network
  - Storage Attach to Scale-Up Networks: Challenging

For More Information

Talks at this Storage.AI conference:

- “An Update on Accelerated Object Storage for AI/ML”, Jason Goldschmidt
- “Scaling Inference with KV Cache Storage Offload and RDMA Accelerated Architecture”, Ugur Kaynar
- “PNFS Past, Present, and Future: What’s all the excitement about?”, Gary Grider

S3/RDMA protocol development: SNIA Accelerated Object I/O TWG

SNIA presentations: Storage for AI 101, 102 and 103, Curtis Ballard


SDC | StorageAI™
A SNIA Event
Thank You
Slide Credits (it takes a talented village):
Joseph White, Claudio DeSanti, Ugur Kaynar