SDC | StorageAI™

A SNIA Event

April 29, 2026 • Denver, Colorado

Data Storage Innovations for Scalable AI Infrastructure

Erich F. Haratsch, Marvell

Outline

- AI Evolution and Scaling Trends
- Memory Pressures in AI Training and Inference
- Memory-Storage Hierarchies for Scalable AI Systems
- Flash Storage Innovations for the AI Data Center
- Key Takeaways

Boom in Data Storage Driven By AI

THE WALL STREET JOURNAL.

TECHNOLOGY | ARTIFICIAL INTELLIGENCE | HEARD ON THE STREET Follow

Hard Drives Are Making an AI Comeback. Yes, Hard Drives.

Data-storage needs are growing, lifting Seagate and Western Digital

By Asa Fitch Follow

Sept. 19, 2025 5:30 am ET

Forbes

Nvidia Dynamo And Storage Next Boost AI Storage, Performance And Lowers Costs

By Thomas Coughlin, Contributor. Covering Digital Storage Technology & ... Follow Author

Published May 07, 2025, 02:51pm EDT

tom's HARDWARE

Hard drives on backorder for two years as AI data centers trigger HDD shortage — delays forcing rapid transition to QLC SSDs

News By Hassam Nasir published November 9, 2025

The AI boom might help QLC overtake TLC in the next two years.

businesswire A BERKSHIRE HATHAWAY COMPANY

Aug 4, 2025 11:30 AM Eastern Daylight Time

SNIA Announces Storage.AI

Industry Leaders Combine Forces to Solve AI-Related Data Challenges

SDC | StorageAI™

The Evolution of AI: Key Milestones 2022-2026

One-shot inference	Multi-step inference	Multiple agents	World models	Local inference

- Reasoning and Agentic AI drive massive growth of token generation for inference
- Physical, Agentic and Local AI create new use cases in the edge and client

Evolution of AI Models

Era	Model Type	Model Size (Parameters)	Training Data Size
2012-2016	Vision (AlexNet, ResNet)	~25–60 M	~1 M images (~ 150 GB)
2018-2020	Early LLMs (GPT-2)	~1–2 B	~40 GB text
2020	GPT-3	~175 B	~300 B tokens
2023	GPT-4 class	~1–2 T	~10–13 T tokens
2024-2026	Frontier LLMs	~0.4–1.6 T	15–33 T+ tokens

- Parameters grew roughly 10–30× from GPT-3 → GPT-4/2026 frontier
- Training data grew 100×+
- This reflects scaling laws: larger models require disproportionately more data to realize their capacity

How much memory is needed for training?

- Training memory = model state + activations = 4*M*P + a*M*P
  - M: Number of model parameters
  - P: Precision, 2 Bytes for 16-bit floating point
  - Model state: model weights + optimizer states + gradients
  - Activations: a<1, scaling factor for active parameters

- Training Data = T*P
  - Number of training tokens: T
  - P: Precision, 2 Bytes for 16-bit floating point

Training: Size for Model and Training Tokens

Model	Year	Parameters	Model State	Tokens	Token Size
AlexNet	2012	60 million	500 MB	1 M images	150 GB
GPT-3	2020	175 billion	1.4 TB	300 billion	600 GB
GPT-4	2023	1.76 trillion (*)	14 TB	13 trillion (*)	26 TB
Llama2	2023	70 billion	560 GB	2 trillion	4 TB
Llama3	2024	405 billion	3.2 TB	15 trillion	30 TB
DeepSeek v4 Pro	2026	1.6 trillion	12.8 TB	33 trillion	66 TB

(*) estimated

Model and training data sizes are not huge but …

- Whole model state needs to fit in memory, ideally in one compute node
- In order to finish training within a reasonable time (~ 1 month), training is pipelined over thousands of compute nodes
- Checkpoints: 6-8B per model parameter
  - 1.6T parameter model, 30-day training, checkpoints every 30 mins
    - 50+ petabytes of transient data, ~200 TB of persistent data
  - Data flow: GPUs → NVMe → parallel filesystem→ object store

How much memory is needed for inference?

- Inference memory ≈ model weights + KV cache (context)
- Weights ≈ M × P
  - M: number of model parameters
  - P: Precision, 2 Bytes for BF16 or FP16, 1 Byte for int8
- KV cache ≈ U*2×L×H×T×P
  - U: number of concurrent users
  - 2: K + V
  - L: transformer layers (model depth)
  - H: hidden size (model width)
  - T: number of tokens in context
  - P: Precision, 2 Bytes for BF16 or FP16, 1 Byte for or int8

Inference (1 user) : Size for Model and KV Cache

Model	Parameters	Model Weights	Max context	Max KV Cache
GPT-3	175 billion	~350 GB	2–4k	~0.5–1 GB
GPT-4	1.76 trillion (*)	~3.4 TB	128k	~120–130 GB
Llama2	70 billion	~140 GB	4k	~1–1.5 GB
Llama3	405 billion	~810 GB	128k	~80–100 GB
DeepSeek v4 Pro	1.6 trillion	~3.2 TB	1M	~350 GB

(*) estimated

AI Processing Phases and Storage Workloads

Checkpoints

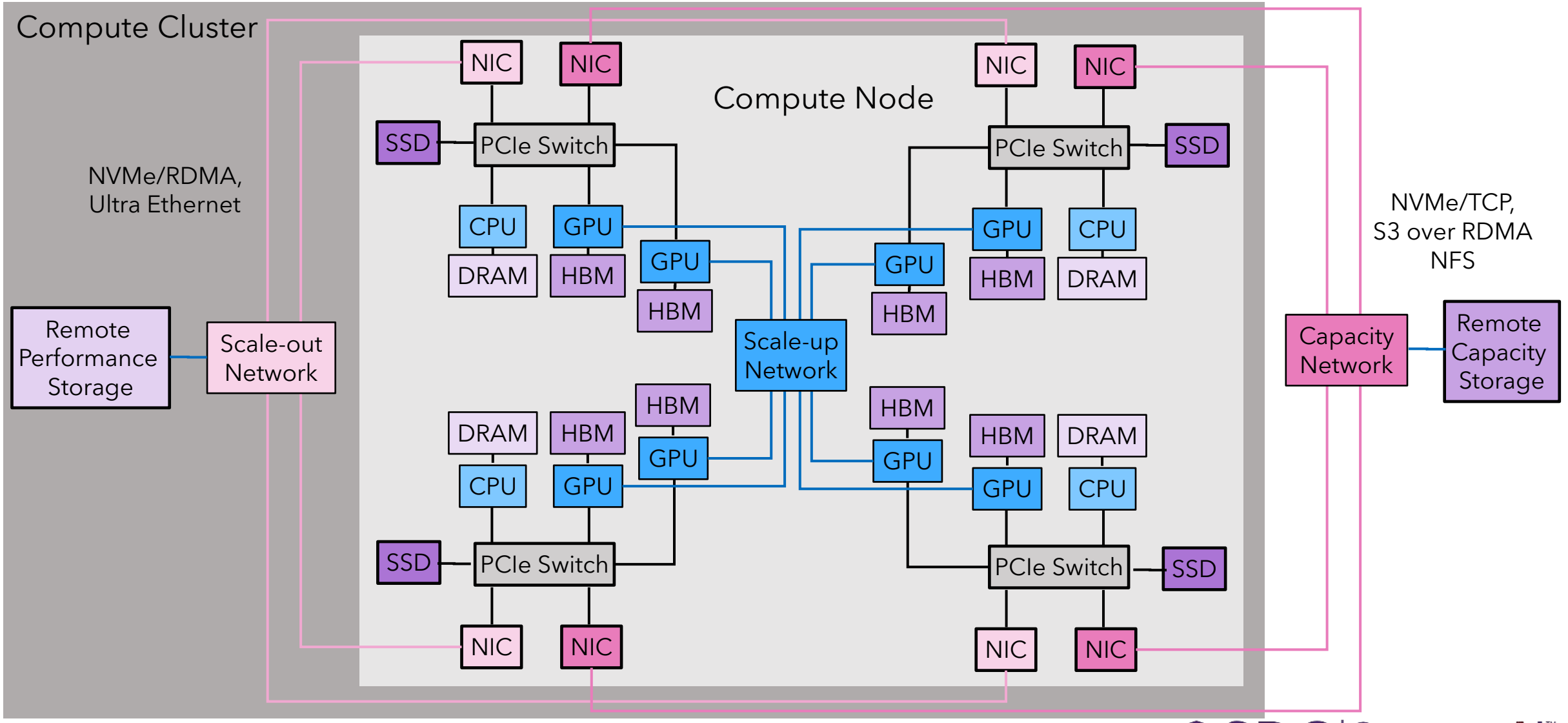
RAG
KV Cache
Small IOs

PB scale, Sequential, Object storage	PB scale, Throughput-oriented, Object and capacity flash	TB scale, High BW, Parallel access, PCIe, NVMe-oF	TB scale models and context, Latency-sensitive, Local and shared NVMe tiers	PB scale, Sequential, Object storage and HDD tiers

Compute Hierarchy in AI Data Center

Level	Physical Unit	Architectural Unit	GPU Scale	GPU Memory
0	Silicon Package	Superchip	2 GPUs, 1 CPU	~ 384 GB
1	Drawer	Compute tray	4 GPUs, 2 CPUs	~ 768 GB
2	Rack	Compute node scale-up	72 GPUs, 36 CPUs	~ 13.8 TB
3	Pod	Row of racks scale-out	576 GPUs, 288 CPUs	~ 110 TB
4	Cluster	Data hall, shared SW orchestrator	10k+ GPUs	~ 2+ PB
5	Campus	AI Factory	100k+ GPUs	~ 20+ PB

AI Cluster Components and Networks



SDC | StorageAI™

Storage Hierarchy in AI Data Center

Level	Media and Hardware	Typical Capacity	Training Workload	Inference Workload
1: Tray	4x E1.S SSDs	~30 TB	Ephemeral scratch	Hot KV Cache: spillover from HBM, LPDDR
2: Rack	72x E1.S SSDs	~540 TB	Aggregated scratch across compute node	Rack-local KV Cache
3: Pod	1x All-Flash NVMe-oF Racks	~10-40 PB	Active training tier: Feeds clean data to GPUs; absorbs parallel checkpoints	Warm KV Cache and fast vector DB for RAG
4: Cluster	Scale-out enterprise storage NFS/Flash/Hybrid	100 PB+	Data staging tier	Cold KV Cache & Telemetry
5: Campus	S3 Object Storage HDD/Tape	1 EB+	Data lake: Long-term retention of training and generated datasets	

Flash Storage Use Cases in the AI Data Center

Performance Tiers

- Optimized for 4KB and sequential performance
- In CPU or GPU compute nodes and remote performance tiers
- CPU centric workloads
- Support TLC and QLC
- Small to medium capacities – TBs per drive
- E1.S, E3.S

Capacity Tiers

- Optimized for capacity and read performance
- Remote capacity tiers
- CPU centric workloads
- Support QLC and Nearline Flash
- High capacities – approaching 1 PB per drive
- E1.L, E3.L, E2 (new)

Small IO Tiers

- Optimized for small IOs (512B+)
- Small IO performance tiers
- GPU centric workloads
- Support Low Latency or pseudo SLC
- Small to medium capacities – TBs per drive

Conclusion: Storage is a Critical AI Architecture Element

- **AI scaling is fundamentally a storage problem**
  - Orders-of-magnitude growth in parameters, tokens, checkpoints, and KV cache
  - Inference, not training, is becoming the dominant long-term storage driver
- **One-size-fits-all storage no longer works**
  - Distinct, workload-optimized tiers for training, inference, RAG, and archival
  - Performance, capacity, endurance, and latency must be explicitly balanced
- **Flash and NVMe form the backbone of AI data movement**
  - Local and remote NVMe tiers absorb bursty, parallel IO at PB scale
  - NVMe/RDMA enables GPU-efficient data paths at rack and pod level
  - Object storage (S3 over RDMA) anchors cost-efficient, exabyte-scale capacity
- **New architectures are emerging**
  - Novel form factors (E2, liquid cooling) enable density and power efficiency
  - Storage offload and data-path acceleration reduce CPU and GPU overhead
  - Memory–storage hierarchies blur as KV cache and embeddings spill beyond HBM
- **Bottom line**
  - Scalable AI requires co-designed compute, memory, storage, and networks
  - Storage innovation is now a critical enabler of AI scale, cost, and efficiency


SDC | StorageAI™
A SNIA Event
Thank You