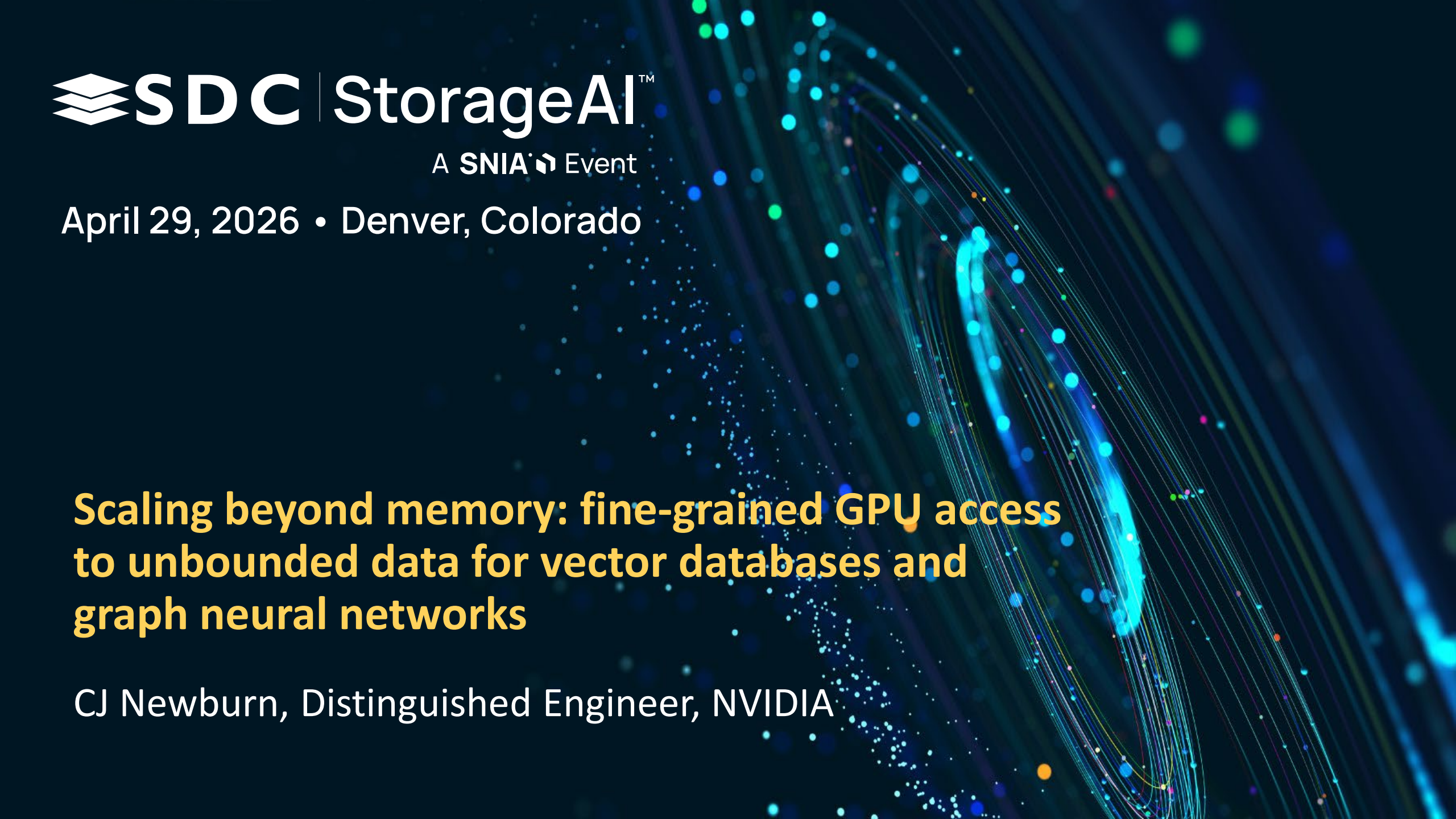
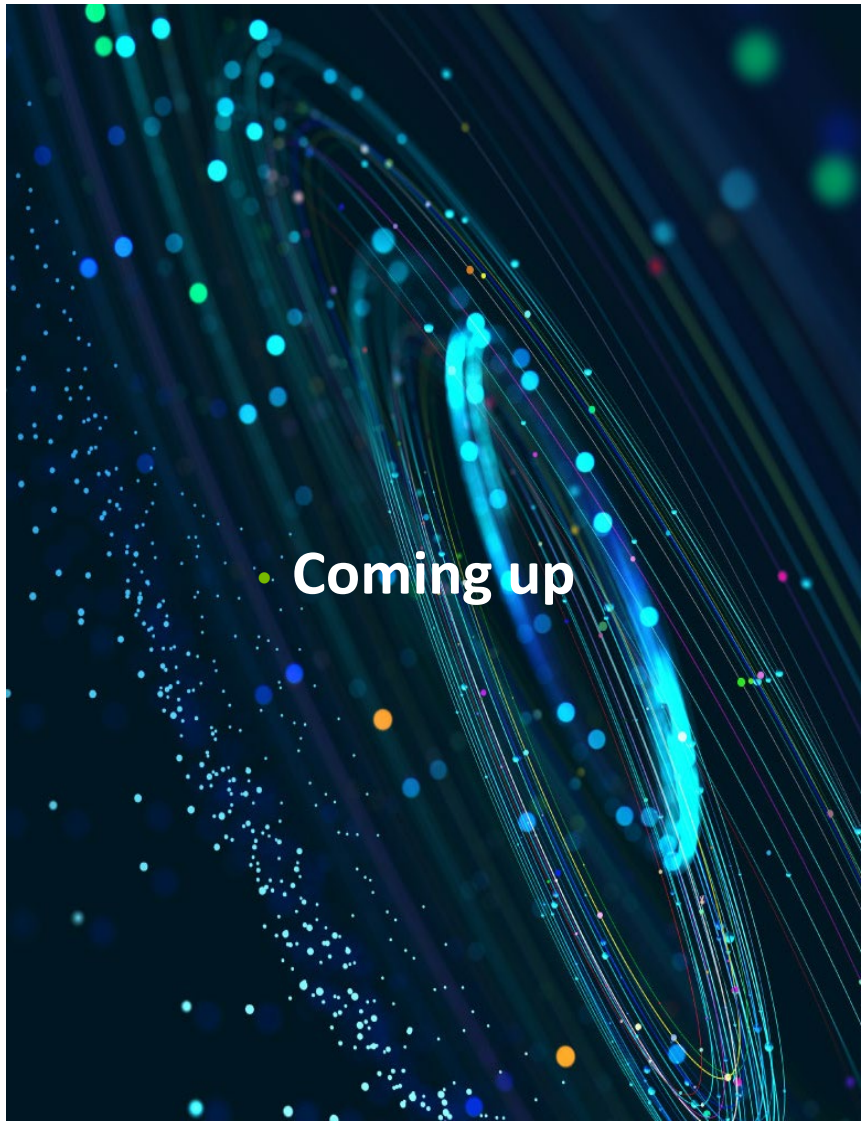


Scaling beyond memory: fine-grained GPU access to unbounded data for vector databases and graph neural networks

CJ Newburn, Distinguished Engineer, NVIDIA

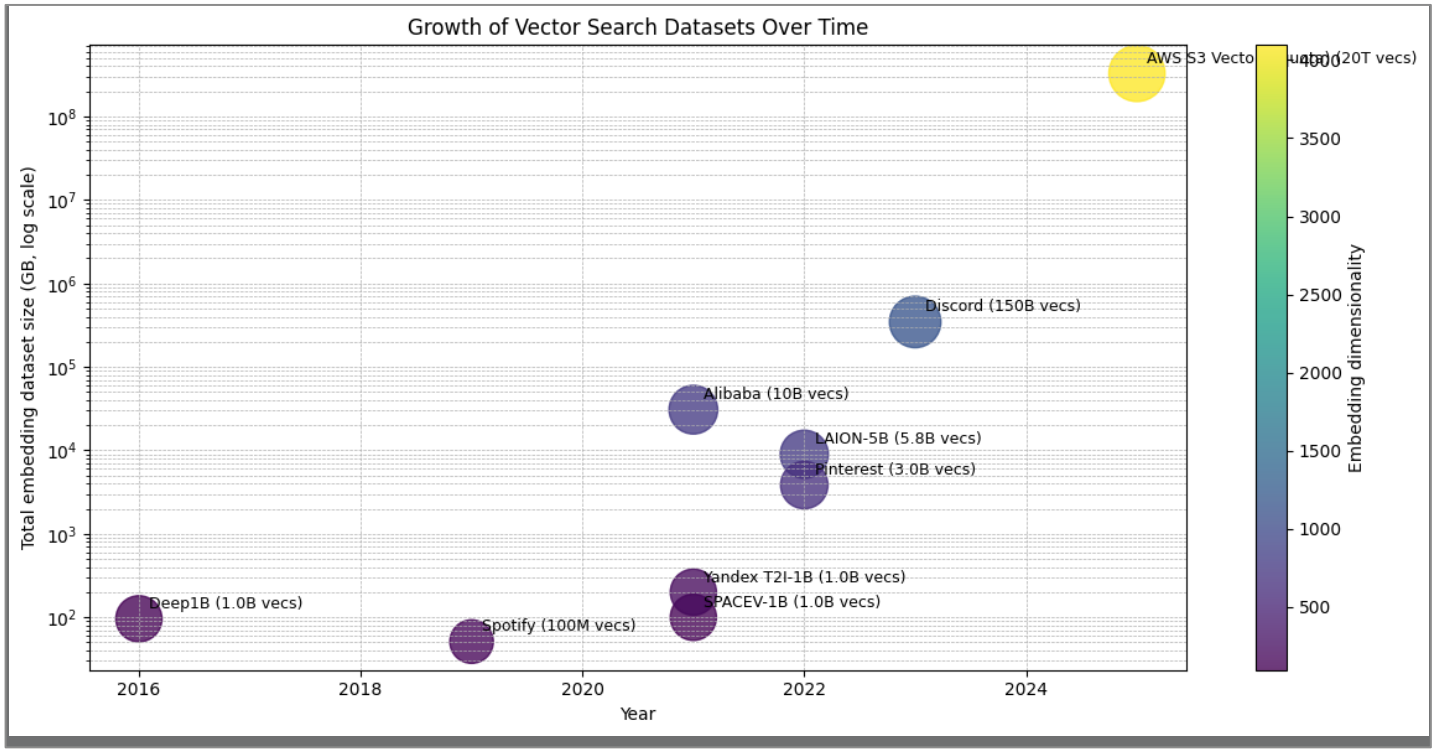


- Big, fine, sparse: vector DBs and GNNs
- Trade-offs
- Architecture
- Data orchestration
- Ecosystem

 ©2026 SNIA. All Rights Reserved.

Vector Search: Evolution of embedding volume over time

1000x growth in overall dataset size in 10 years → scale-out memory is not enough



1. Embedding vector dimensionality has grown 10x, number of vectors has grown 100x+
2. Datatypes vary: FP32, FP16, INT8, even binary!
3. Brute force search = GEMV (single query) and GEMM (batch) --> not feasible for larger datasets (size + latency constraints)
4. Vector search algorithms: from brute force to ANNs (approximate nearest neighbors)

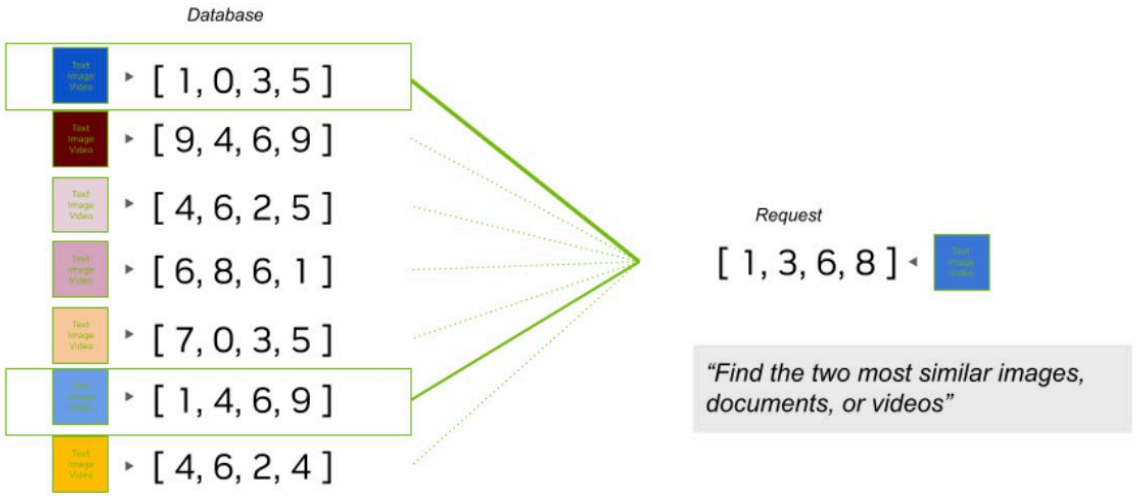
Vector Database Size = num_vecs * dtype * dimensionality + index_size (not modelling above)

Original slide from Srinivasan Ramesh, NVIDIA and modified

 ©2026 SNIA. All Rights Reserved.

Vector search

Find nearest neighbors given an input query (embedding)



Vector Search (Credits: NVIDIA Blog)

What is Vector Search?

Finding similar items by comparing dense numerical representations (embeddings) instead of exact keyword matching. Powers the "understanding" layer of modern AI.

Application	Description	Scale Examples
Retrieval-Augmented Generation (RAG)	Chatbot, enterprise QA, support bots, Ground LLMs with external knowledge	1M-1B doc chunks
Semantic Search in Web/eCommerce apps	Natural language product/content search; legal, medical, eCommerce	MSFT Bing Search, Trillion(s!) of vectors, Alibaba product catalog (billion-scale)
Recommendation Systems	User-item matching; collaborative filtering at scale	Netflix, Pinterest

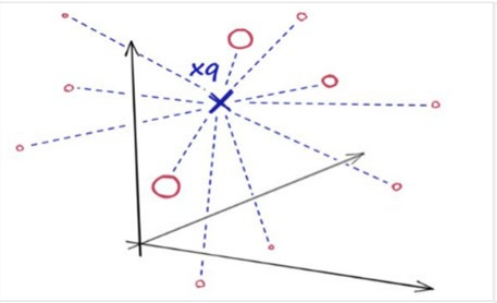
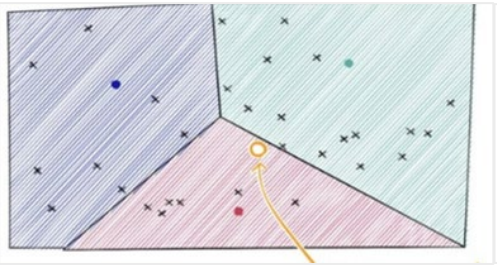
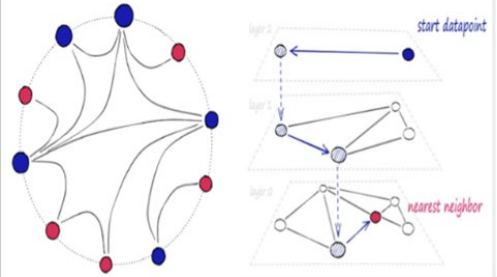
 ©2026 SNIA. All Rights Reserved.

Graph neural networks and relational graphs

- Complex vs. pairwise relationships
  - Fraud, social networks
  - Predictive AI for retail and eCommerce
- Scale to 1T edges and beyond
  - Too big to fit in the memory of a whole rack
- Involve sparse, unpredictable traversal through a graph
- Fine-grained: 128B-4KB; 512B is common
  - Richer data may drive up dimensions
  - But potential for higher IOPs drives granularity down

 ©2026 SNIA. All Rights Reserved.

Approximate methods scale to large size

		Advantages	Characteristics	Examples
Flat		Quality	Exhaustive search. Exact, but slow – especially at scale	Brute-force
Tree-based		Scale	Semantically partitions data to reduce distances during search	IVF SCaNN HNSW-IF
Graph-based		Low latency	Constructs a neighborhood graph to improve searh latency	HNSW CAGRA Vamana

Slides credit: https://www.nvidia.com/en-us/on-demand/session/gtc25-dlit71710/

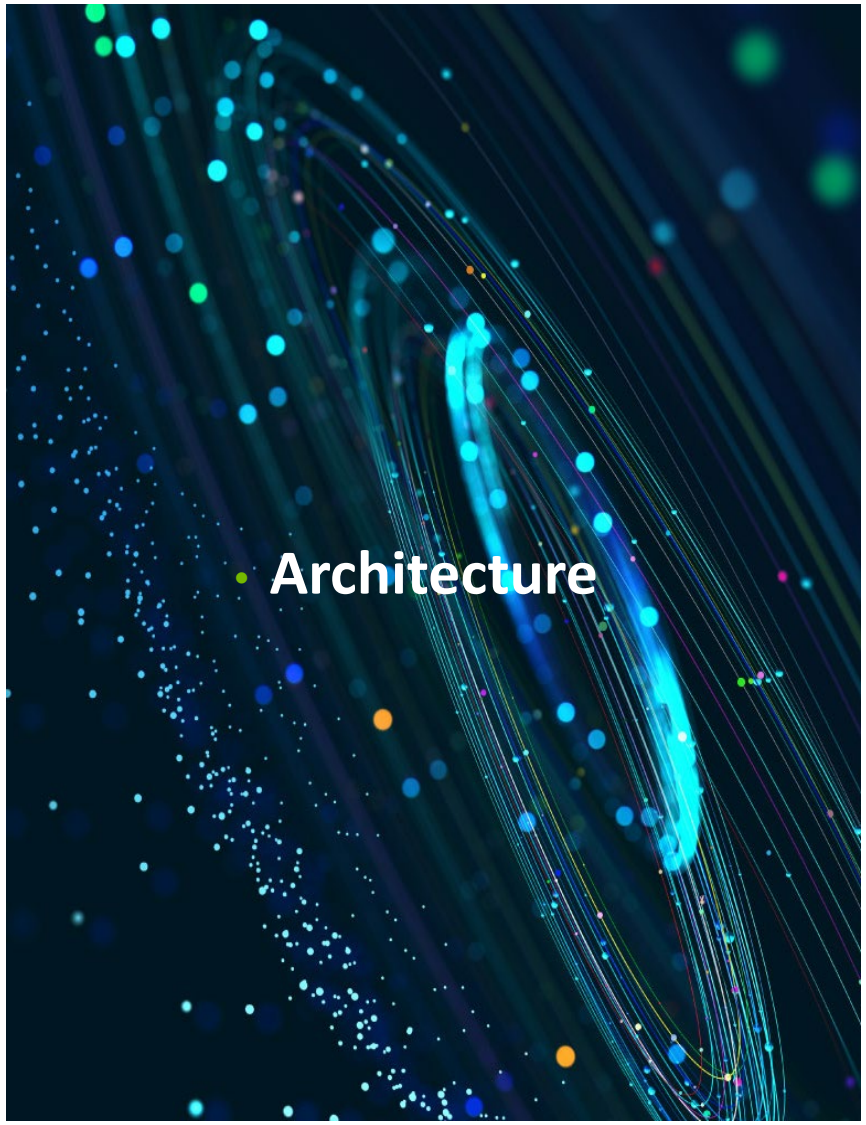
 ©2026 SNIA. All Rights Reserved.

Fine, sparse

- When you no longer fit in memory, bottleneck is storage IO
- Access fewer data items with pruned vs. brute force search
- Access smaller data items directly from GPU to avoid RAF
- Graph and VecDB accesses are sparse and unpredictable
- Looks like pruned graphs help with both of these

- Ultimate measure besides perf is bytes moved to get accuracy
- Better indexing/training reduces search/inference time

 ©2026 SNIA. All Rights Reserved.



- SCADA client/server
- GPUs and performance
- Storage servers and back ends
- Cmp: Context memory storage
- Reference designs

 ©2026 SNIA. All Rights Reserved.

Problems → solutions for concurrency

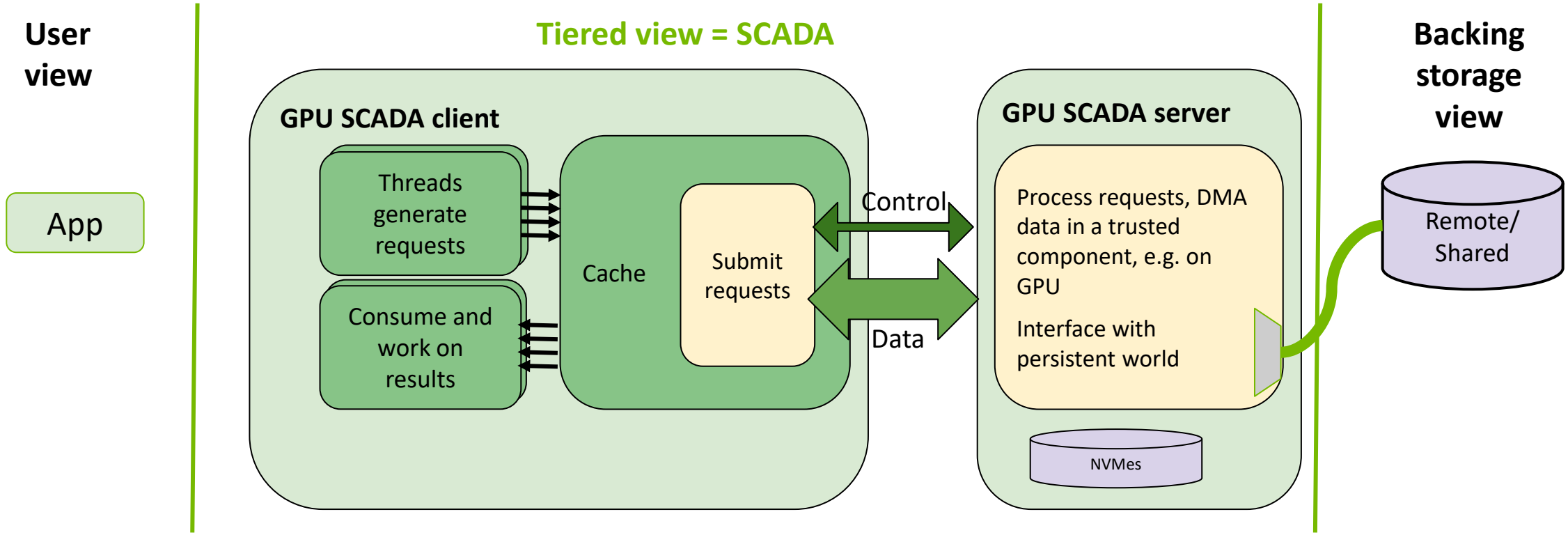
Ushering in a new era of data-centric computing on large problems with GPUs

SC	Scaled	• O(100) thread CPUs limit concurrency → O(100K) GPU threads • Memory fabric breaks at scale → introduce an API, support errors
A	Accelerated	• GPUs can't run a file system, can't ask CPU for help → GPU initiation • Can read into registers from IO → buffer in a cache on GPU • CPUs can't tolerate latency → GPUs do
D	Data	• Limited capacity in memory → unbounded storage
A	Access	• RAF problem from sparse access to big chunks read in → direct access

 ©2026 SNIA. All Rights Reserved.

SCADA™ abstraction

GPU becomes an autonomous highly parallel data access engine



- User view: app-centric API, e.g. contiguous linear/mdspan, key-value (coming)
- Tiered view: implement caching, block access, many disks, map to outside world
- Backing storage view: interop with file/object systems

 ©2026 SNIA. All Rights Reserved.

GPUs and vector search

When do GPUs make the most sense?

Batch parallel processes run much faster on GPUs

- ANN index builds
- Throughput-intensive search
- Preprocessing (e.g. quantization)

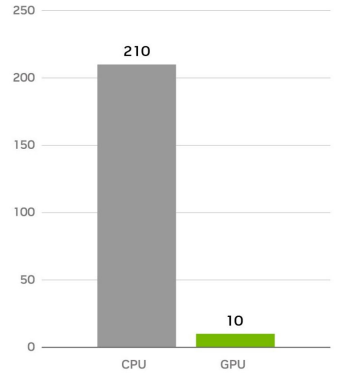
Streaming processes can also run faster on GPUs

- Latency sensitive search, especially at high volume

Higher performance can lead to lower costs

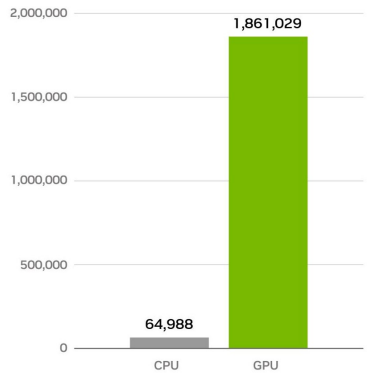
- Scaling up index builds that can be deployed to CPU for search

21x Faster Index Build

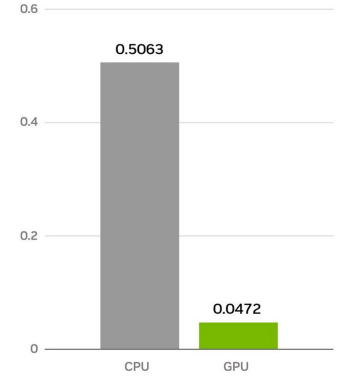


12.5x Lower Index Build Cost

29x Higher Throughput



11x Lower Latency



Slides credit: https://www.nvidia.com/en-us/on-demand/session/gtc25-dlit71710/

 ©2026 SNIA. All Rights Reserved.

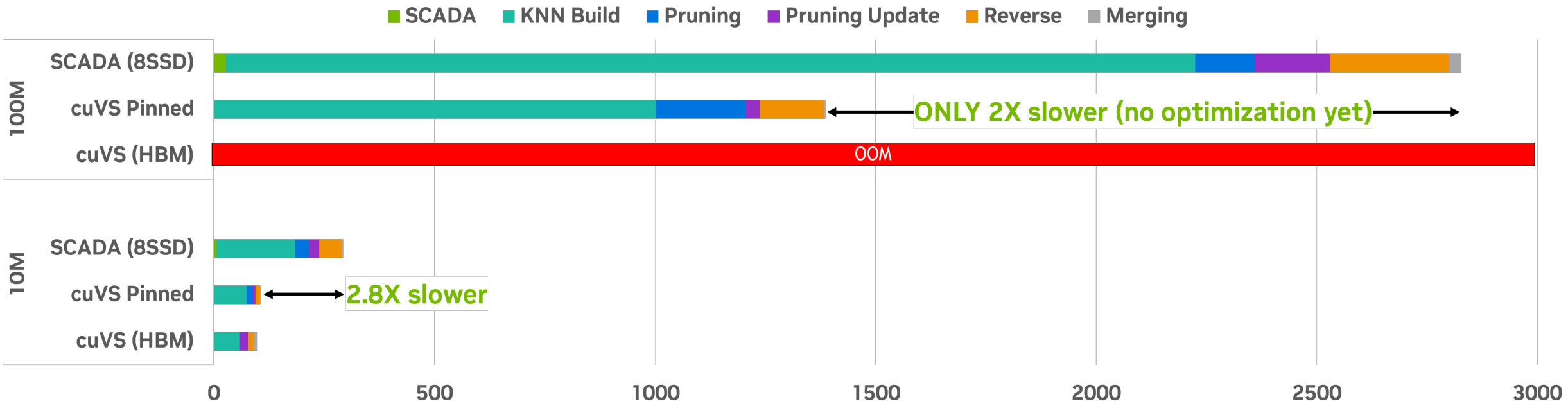
cuVS + SCADA Status

Early, unoptimized research results show promise for B and T scale

- cuVS (CAGRA) +SCADA is showing promising trends especially for scale up story
  - cuVS SCADA is about 3X slower than cuVS CAGRA index build (10M dataset) from host-memory
  - 10X larger dataset → cuVS execution time scales by 13X.
  - 10X larger dataset → SCADA + cuVS execution time scales by 9.6X.

→ cuVS+SCADA 2X slower than cuVS+HostMem at 10X scale.

H3P with H100, Donated Micron 9650s. Thank you!

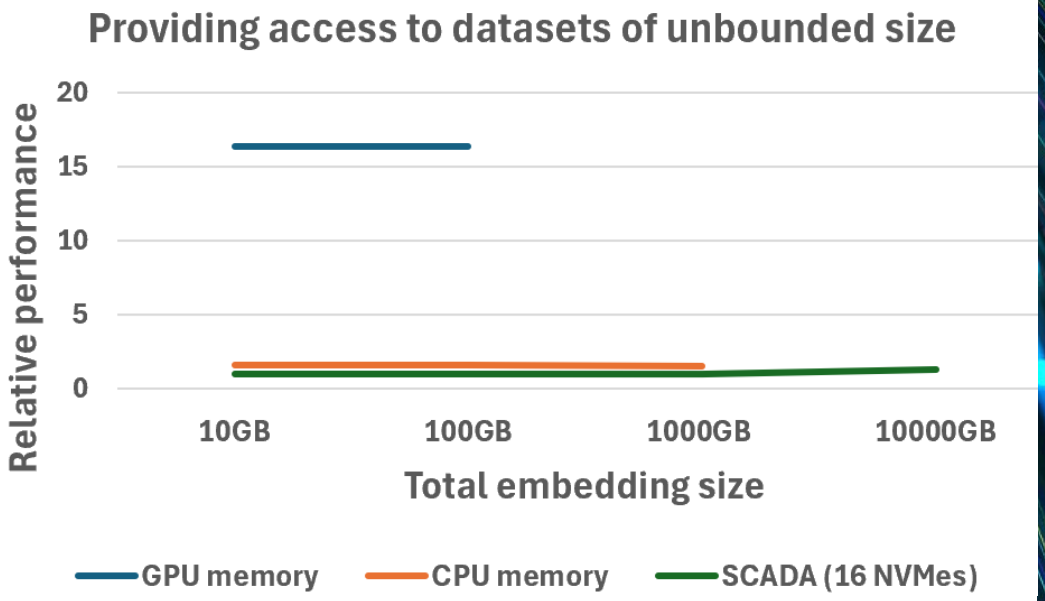


 ©2026 SNIA. All Rights Reserved.

GNNs: End to end results with SCADA-enabled WholeGraph

Using one H100, Gen5, 16x Micron® 9550s

- E2E results on 1 GPU for GAT node prediction
  - GPU memory with ld/st (blue)
  - CPU memory across PCIe to x86 (orange) – **10x larger than GPU**
  - SCADA to NVMes (green) – **100-1000x larger than GPU**
- GPU vs. CPU memory from GPU is 10x faster
  - Limited GPU memory per node, fits 12GB IGBH-small
  - 100GB IGBH-medium won't fit in some GPUs
- **Trade off 20% perf to get unlimited size**
  - 2100 GB IGBH-full won't fit in CPU memory per node
  - IGBH-full vs. medium is a little faster, perhaps from more exposed parallelism
  - Early result 4with client/server on same GPU, IGBH 4KB, still in tuning
- New: multi-GPU
  - Size: Scale # GPUs to hold larger graph structure
  - Perf: Scale # GPU to increase storage bandwidth and IOPs
  - Deployment: Can map multiple GPU clients to a shared storage server
  - Perf: Scale # GPUs to reduce processing time
  - Data is forthcoming

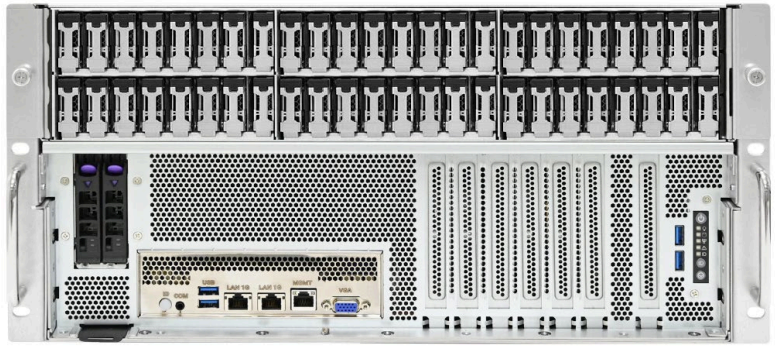


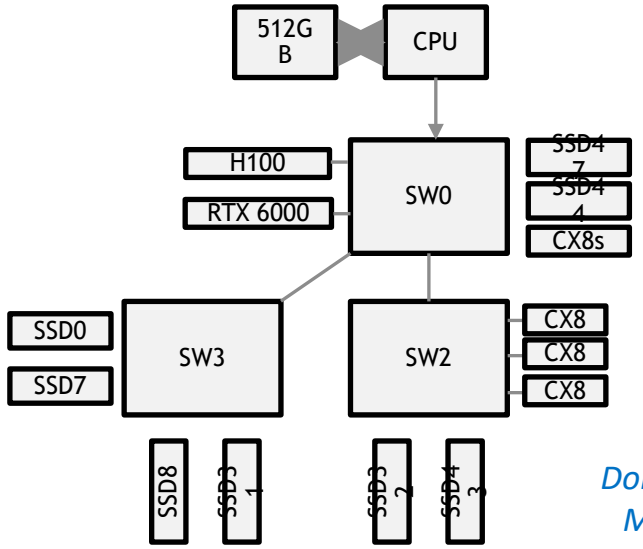
H3P with H100, Donated Micron 9550s. Thank you!

 ©2026 SNIA. All Rights Reserved.

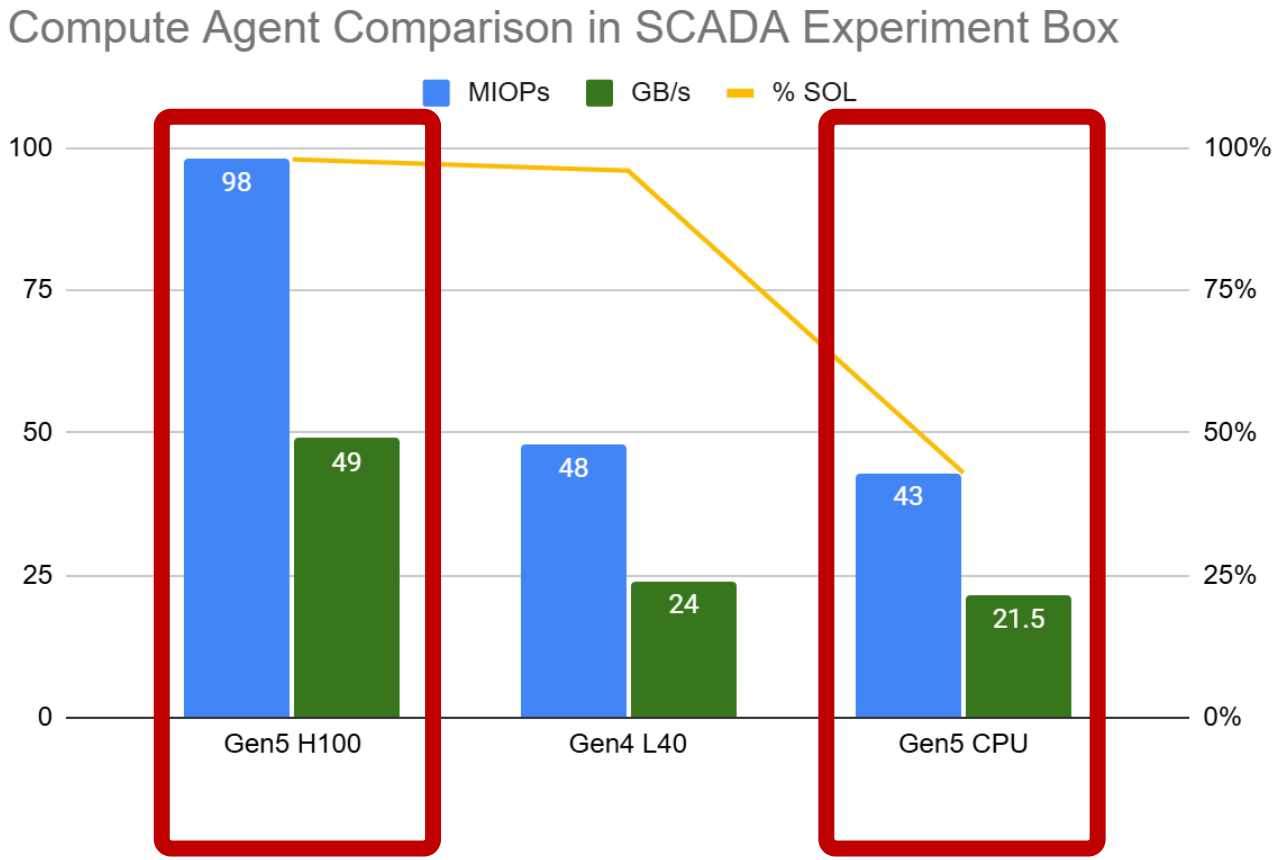
Initial microbenchmarks summary

H100 saturates PCIe without breaking a sweat @10% utilization; high-end CPU can't keep up





Donated drives:
Micron 9650
Thank you!



Consuming ~10% of H100 resources

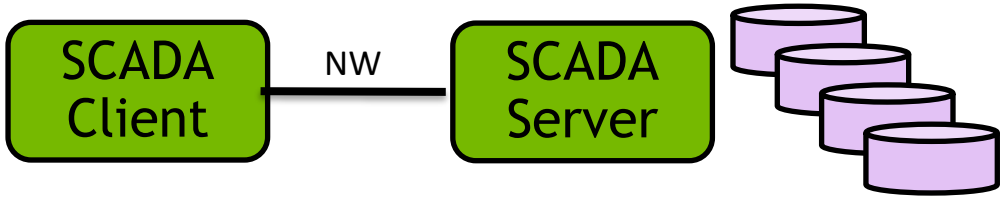
Remeasuring with Samsung Gen6 drives and Gen6
GPU: linear scaling with drives;
>90% Gen6 (200 MIOPs) saturation at 32 drives

 ©2026 SNIA. All Rights Reserved.

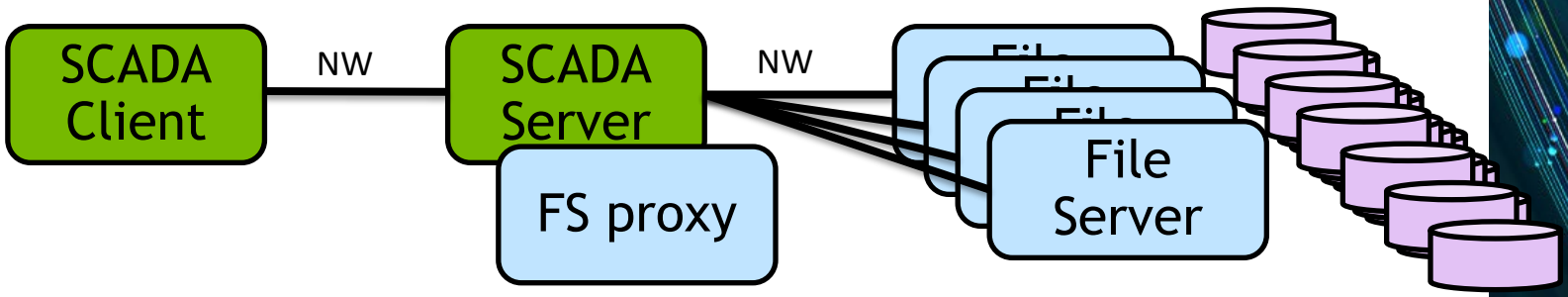
Storage servers and back ends

- Incremental
  - Client knows FS
  - Too much burden on client

- SCADA today
  - Steer to local NVMe
  - FS logic in SCADA server



- FS interop alternative
  - Steer to back end FS server
  - Learn to speak our interface



 ©2026 SNIA. All Rights Reserved.

Toward reference storage server designs

- Two storage server flavors
  - Fewer, high-IOPs drives with modest capacity
  - More, potentially lower-IOPs drives for larger total capacity
- NVIDIA is aligning on recommended components, design
  - CPU/GPU
  - Air → liquid cooling
  - E1.S → E3.S/L

 ©2026 SNIA. All Rights Reserved.

NVIDIA Inference Context Memory Storage Platform

NVIDIA Inference Context Memory Storage Platform

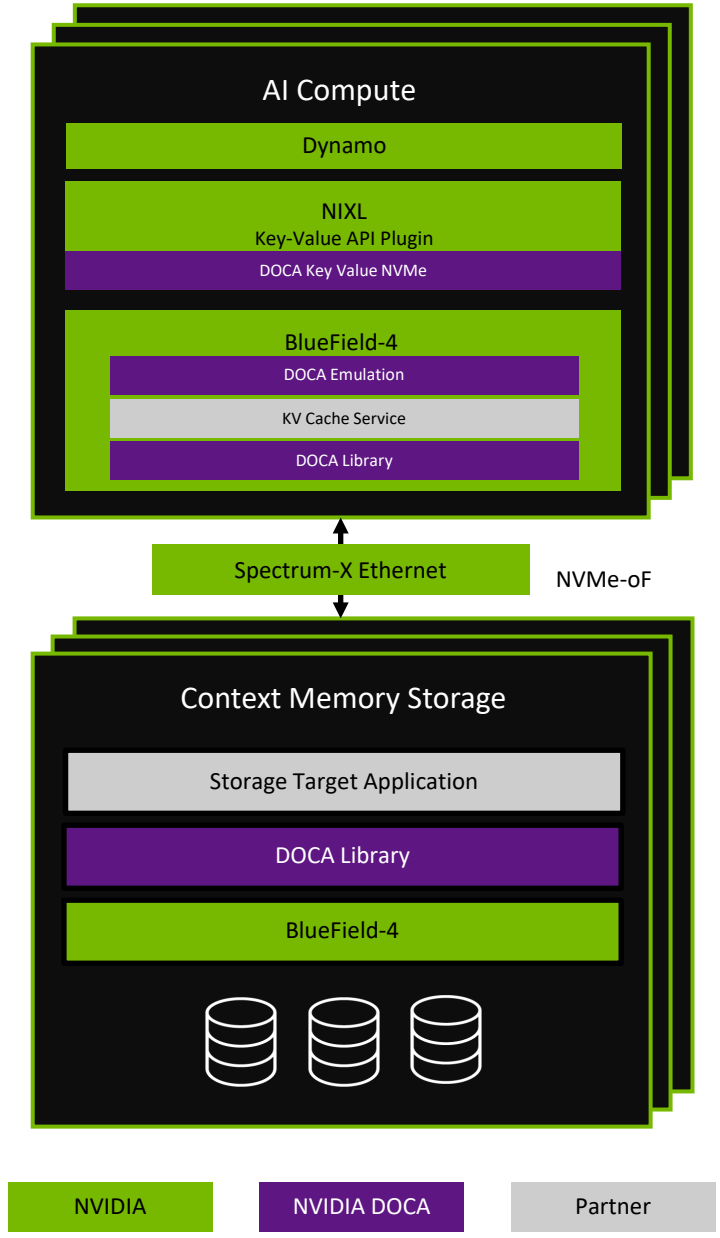
- AI-native, pod-level context storage tier
- Shared across GPUs for efficient context reuse
- Complements GPU memory, local, and networked storage
- Improves performance and scale, reducing TCO:
- 5x higher throughput (tokens-per-second) than baseline without additional capacity
- 5x better power efficiency than traditional storage
- Accelerated by BlueField-4, Spectrum-X Ethernet, DOCA, NIXL, and Dynamo
- Reference Design by NVIDIA, Storage Partner Solution

AIC CLOUDIAN ddn DELL Technologies Everpure HPE
Hitachi Vantara IBM NUTANIX QCT SUPERMICRO VAST WEKA

 ©2026 SNIA. All Rights Reserved.

NVIDIA DOCA Accelerates Inference Context Memory

Secure, low-latency KV cache access at pod scale



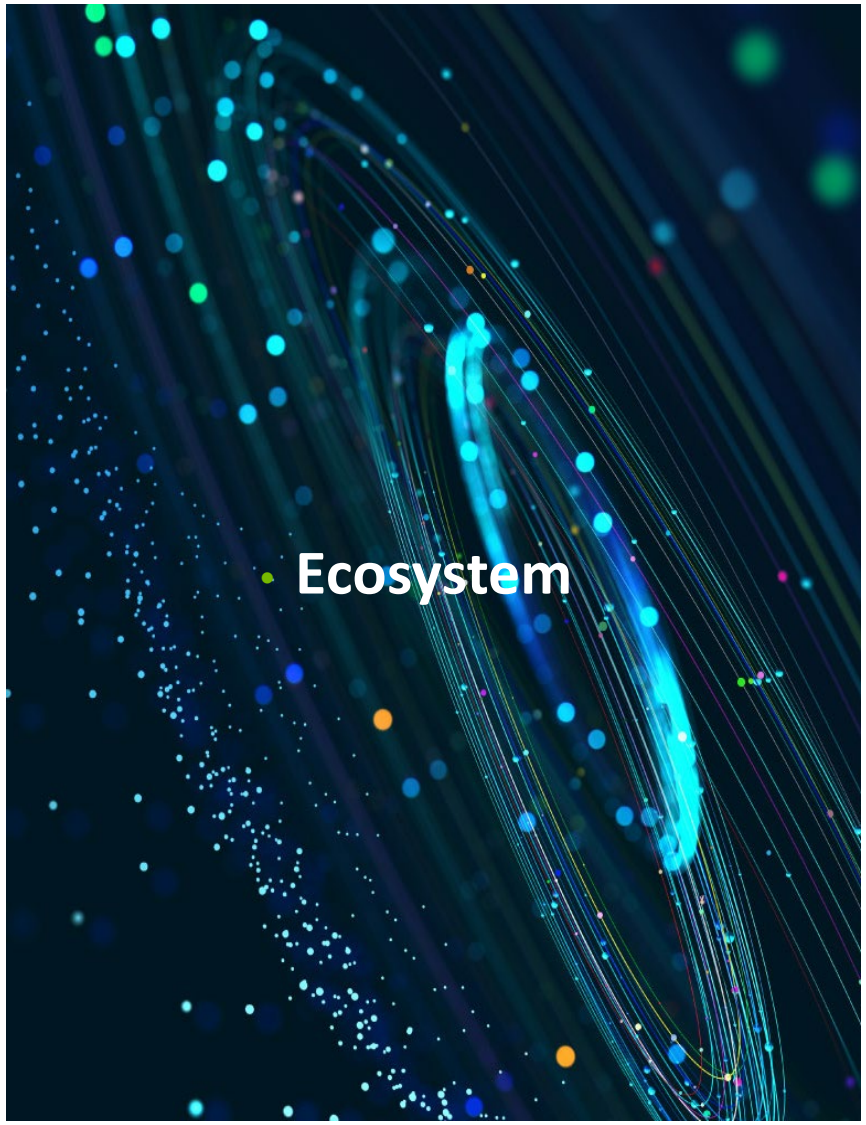
- Efficient KV cache management and sharing
- Runs on BlueField-4 at compute and context memory nodes
- POD level scalable – application remain stateless
- Developer-ready, extensible framework
- Key-value NVMe APIs expose context directly to AI compute
- Secured and isolated KV Cache access
- Hardware-accelerated integrity and encryption

NVIDIA CONFIDENTIAL. DO NOT DISTRIBUTE.

Inference context memory storage's relation to SCADA, Storage-Next

- ICMS improves benefits from caching
  - Recover prompt-specific KV context previous created by prefill
  - LLM$ manager identifies previously-seen prompt prefix, fetches context
- Usage models for KV$/ICMS and SCADA are very distinct
  - Induce different priorities among storage server configs (see below)
- Storage-Next™ contributes to overall storage architecture
  - Drive toward convergence
  - Sharpen definitions of minimal # storage servers and variations

 ©2026 SNIA. All Rights Reserved.



- Storage-Next
- File interoperability
- Emulation
- Profiling and tracing

 ©2026 SNIA. All Rights Reserved.

Storage-Next™ Ecosystem

NVIDIA-driven NDA effort to define next-generation storage architecture and components for AI
Develop and vet the tech, then publicly drive standards

Hyperscalers

Storage & RAID providers

OEMs
Server Component providers

SSD controllers

NAND providers

 ©2026 SNIA. All Rights Reserved.

Usage models and data orchestration

WIP; subject to refinement and change

Usage model	GNNs via WholeGraph	Vector database index build/ search via cuVS
• Starting point	• Raw data	• Vector database

• Ingress
  • Ingress target
	• ETL'd edge +node embedding data
	  • CPU
		• Database of encoded vectors
		  • Used via SCADA on GPU

• Usage
  • Write
  • Read
	• Swap store
	  • Spill embedding data onto disk
	  • Traversal, node aggregation
		• Swap store + persistent store
		  • Generate vector database index
		  • Search vector database

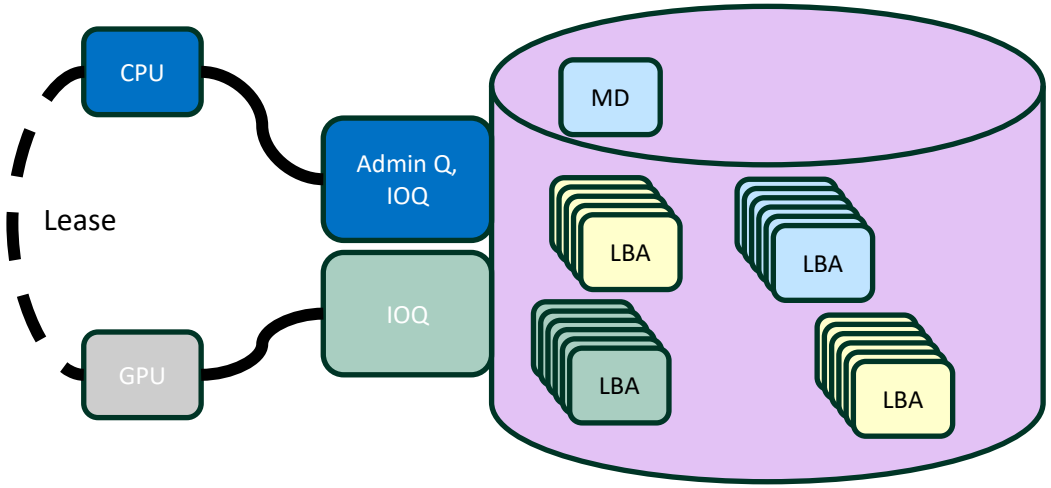
• Actions	• Sample, node aggregation, train	• Index build

• Egress
  • Egress source
	• Model weights (trivial)
	  • GPU
		• Vector database index
		  • Generated via SCADA on GPU

• File reads	• None	• Spread across 1/Nth of many disks
• File writes	• None	• Written to pre-allocated locations • No change to file size (confirm), no new files

 ©2026 SNIA. All Rights Reserved.

File system interoperability



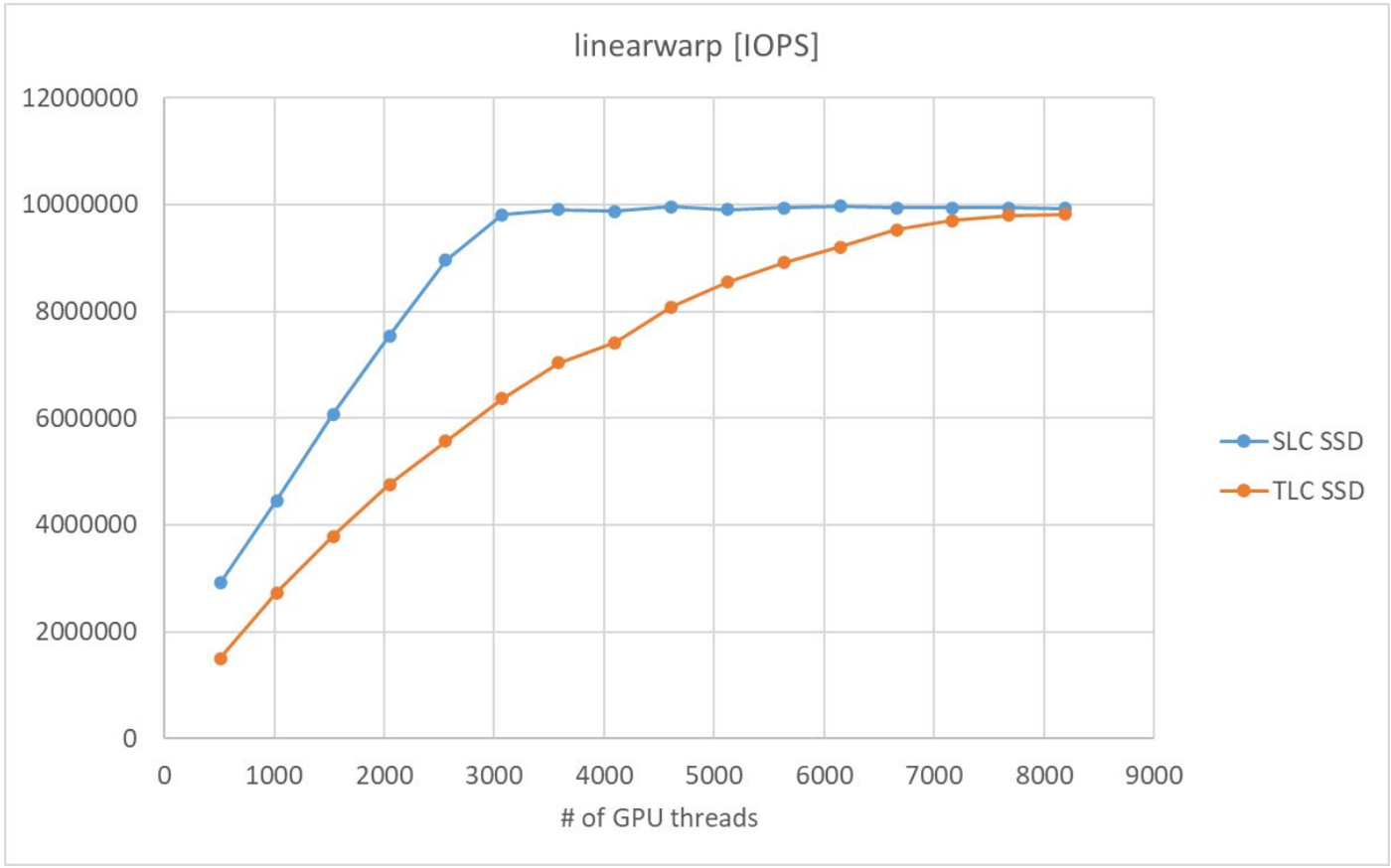
- Orchestration
  - Prepare/defragment drives as necessary
  - Map from object to file(s) on drives: shard, use file system
- CPU/file system
  - File system owns the files and the drives
- GPU/SCADA
  - Harness O(100K) threads, HBM BW, lat tolerance to directly access storage as blocks
  - GPU receives bundles of O(16K) requests, maps to drive and LBA within drive
  - Leverages LBA map from the FS

 ©2026 SNIA. All Rights Reserved.

Kioxia emulator results

Lower drive latency increases robustness wrt varying available concurrency

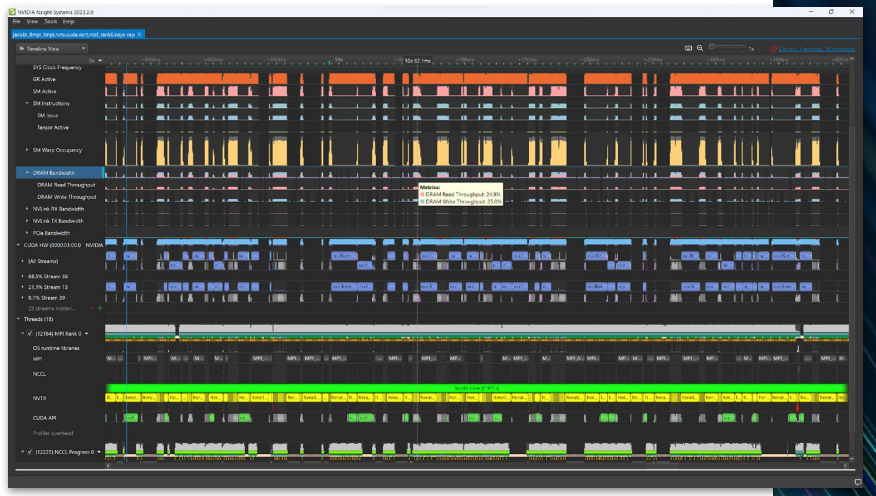
- Concurrency: requests between sync points
  - Approximated with # GPU threads
- Perf is linear with concurrency up to asymptote of 10 MIOPs design point
- Higher latency flattens out; more concurrency to hit peak
- Tail latency flattens more
- Lower latency and fewer structural hazards are closer to roofline



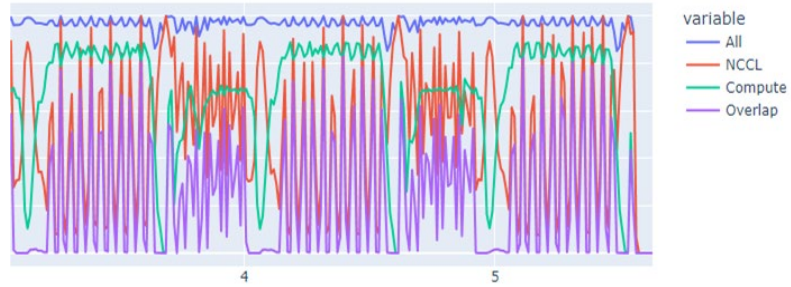
 ©2026 SNIA. All Rights Reserved.

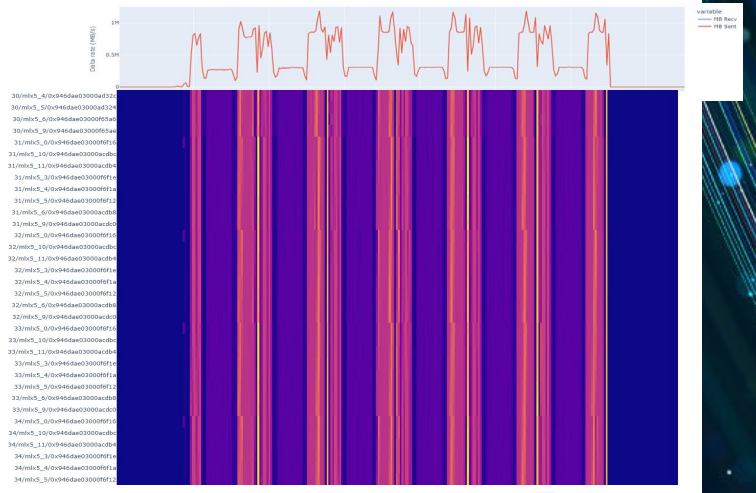
Nsight Systems

System and Cluster Profiler



- System-wide, multi-node, multi-process workload tuning
- Visualize millions of events, counters and metrics
- Balance your workload across multiple CPUs, GPUs, DPUs, NICs and storage volumes
- Locate optimization opportunities
  - GPU starvation
  - Idle HW utilization gaps
  - Suboptimal synchronizations
- Linux, Windows, x86-64, Arm, Tegra
- Supports client storage volumes counters
  - Lustre, NFS, GPFS, S3, NVMe, MVMe-oF





Credits: Yaki Tebeka, Roee Milo, Doron Ofek; Javier Vera

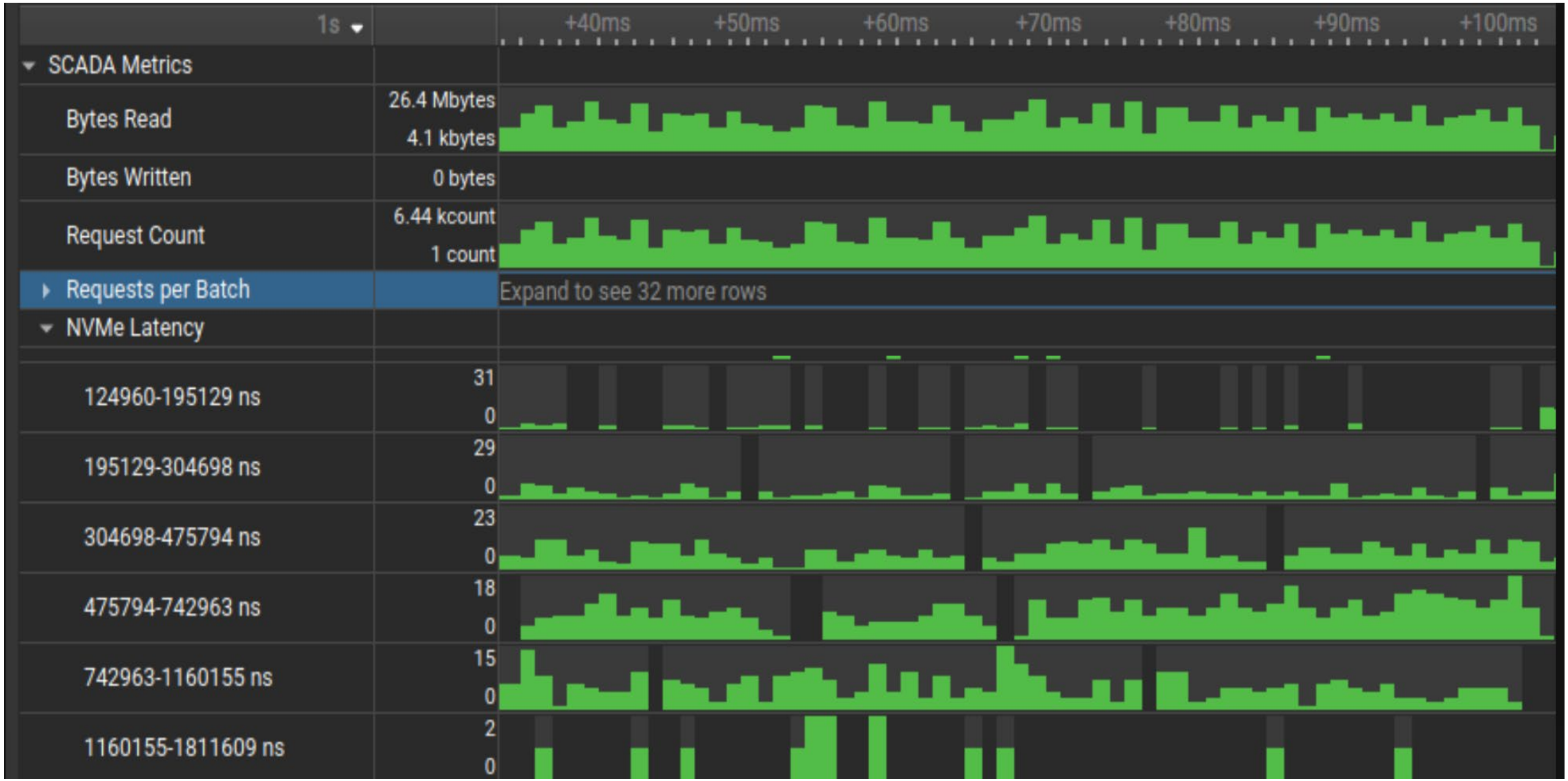
Status: POC, considering for product

 ©2026 SNIA. All Rights Reserved.

SDC Storage NVIDIA

And Now - SCADA Support!

Thank you H3P team



 ©2026 SNIA. All Rights Reserved.

SCADA request tracking

Profiling

- Capture and analyze batches of requests passed from client to server
- Count reads, writes, commands
- Assess locality
- Assess granularity and sequentiality

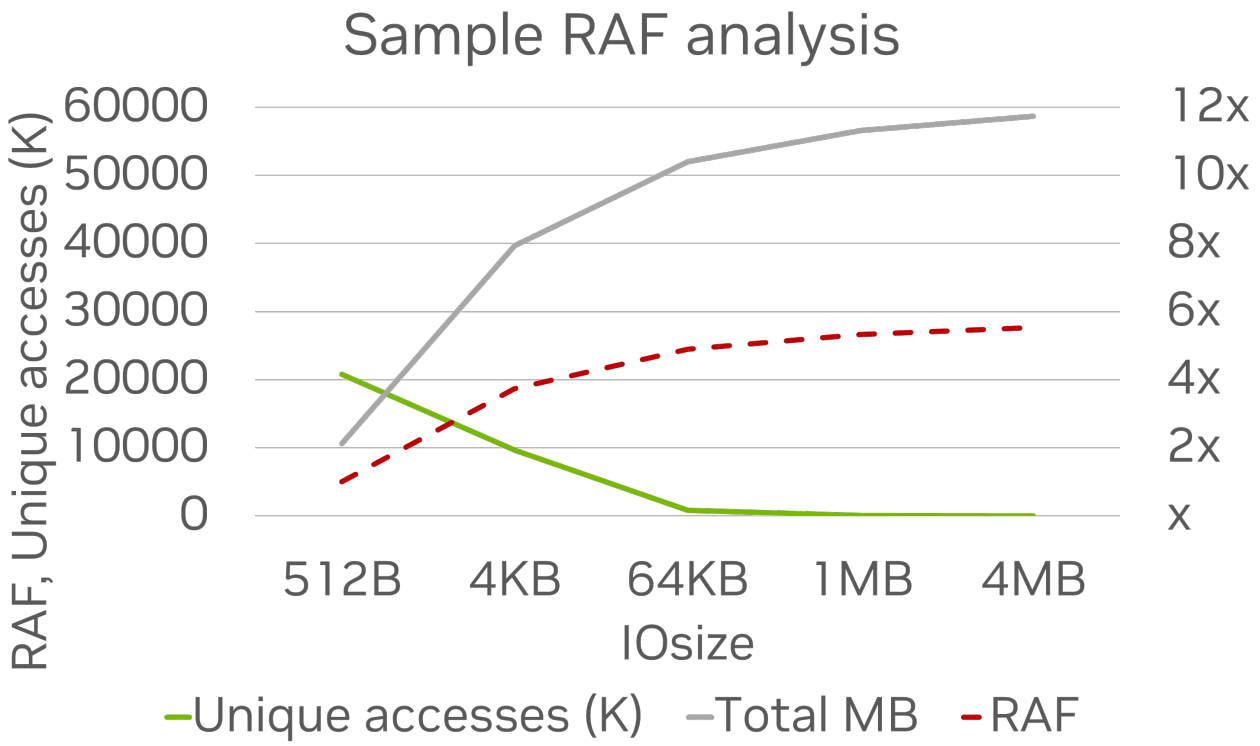
Tracing

- Store trace of requests to disk, postprocess
- Assess potential RAF

- Credits: Javier Vera
- Status: prototype

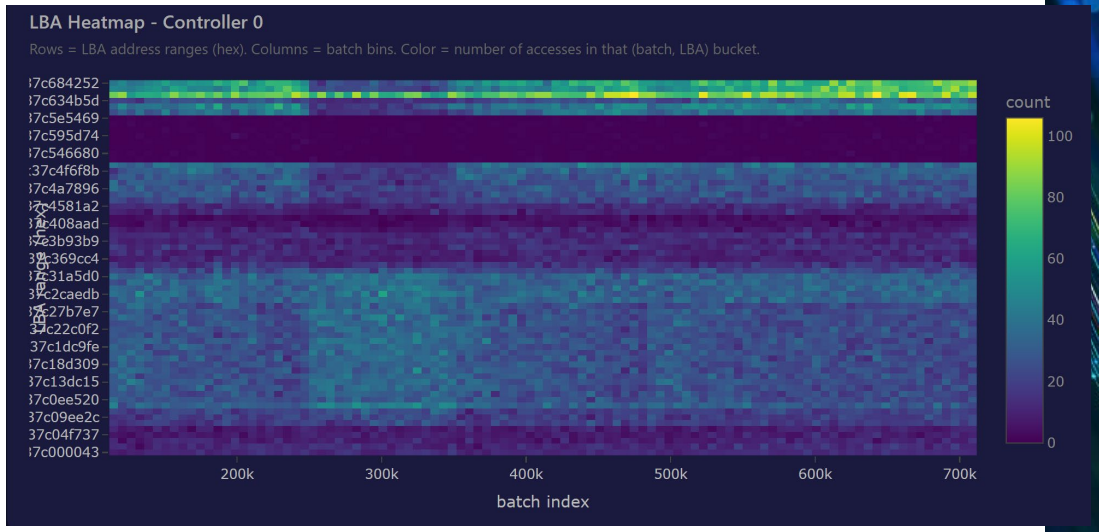
 ©2026 SNIA. All Rights Reserved.

What if accesses were coarser grained? RAF analysis



- 100M papers data set
- WholeGraph GNN
- 1.2M starting set
- 2-layer sampling: x5,x5
- Cache bypassed, 512B IOs
- Status: prototype

- This is data from last night; revising



Credits

Phil Carns – RAF tool

Javier Vera – Tracing infra

Alex Sloboda – data collection, analysis

 ©2026 SNIA. All Rights Reserved.

Nsight Integration for GPU, NVMe, NIC Profiling

Show the performance of all devices attached to the PCIe switches

NVIDIA

Nsight Systems Plugins

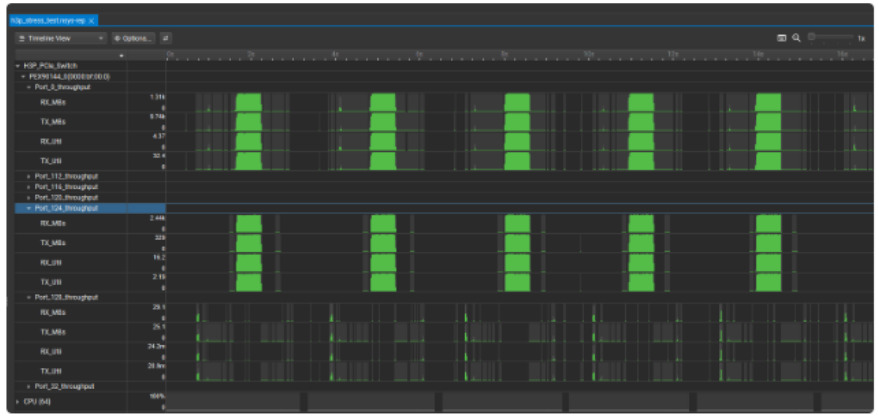
This site lists Nsight Systems third-party plugins.

- For information about Nsight Systems, visit the Nsight Systems website.
- To add a third-party plugin to this list, refer to the instructions provided in the ADD_PLUGIN.md file. file.

Table of Contents

h3_sw_counters

A plugin to monitor and record H3P PCIe switch utilization, throughput, and error counters onto the Nsight Systems timeline.



Architectures
x64

Operating systems
Linux

Minimal Nsight Systems version
2026.2.1

Setup Notes
Requires libh3ppci.so. Set NSYS_PLUGIN_SEARCH_DIRS to point to the plugins directory.

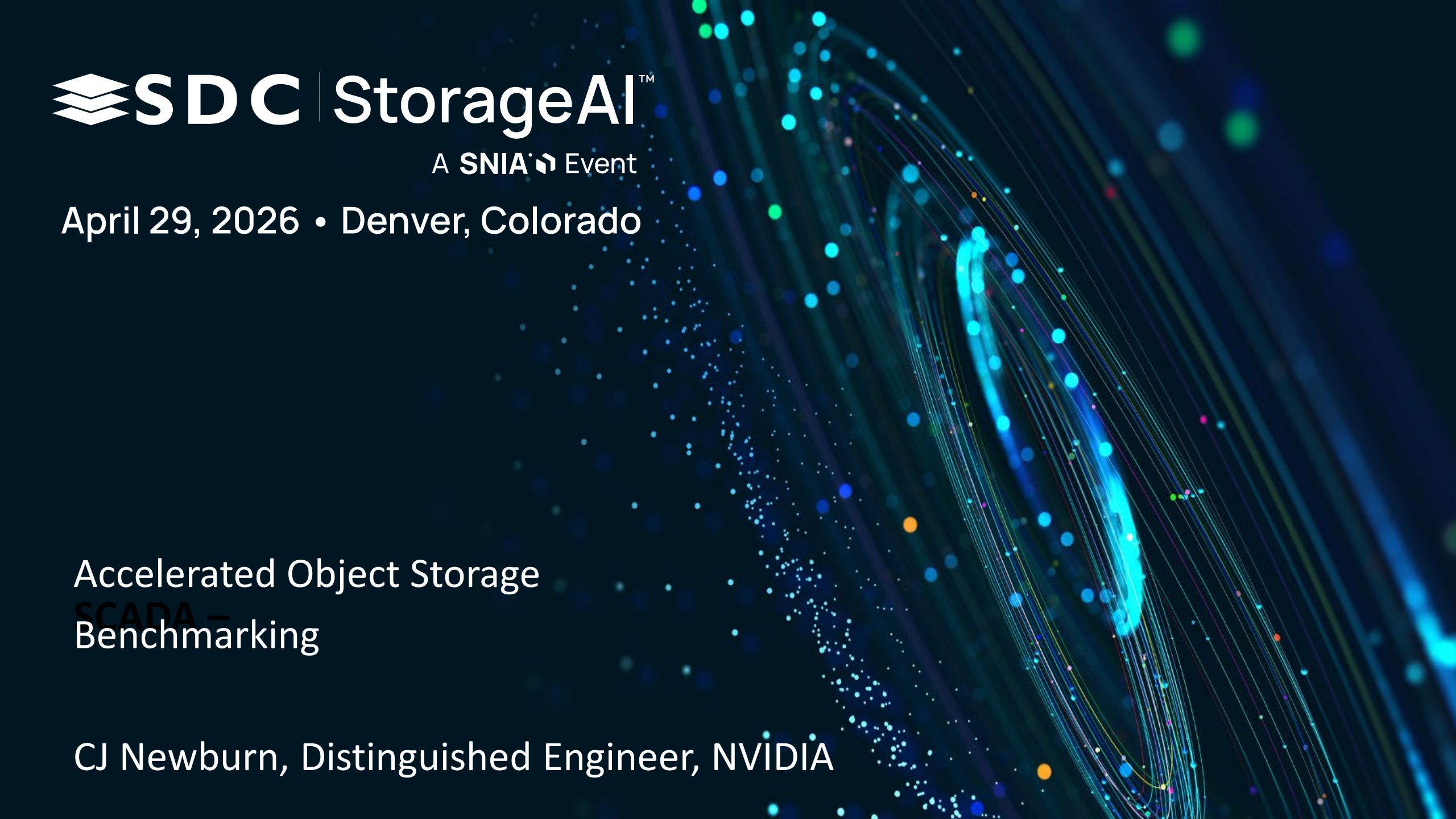
Site URL
https://github.com/H3Platform/SW_NSYS_Plugin

Company
H3 Platform Inc.

Call to action

- Track usage models and workloads
  - Storage criticality, throughput, latency, data orchestration
- Evaluate benefits of fine-grained, GPU-initiated storage
  - Ex: Refined vector indexing and search
- Explore effective storage architectures
  - Ex: where to place and manage interoperable storage
- Provide feedback on desired tools
- Capable, power-efficient components
  - 10x higher IOPs and power efficiency for some usages

 ©2026 SNIA. All Rights Reserved.



Accelerated Object Storage SCADA – Benchmarking

CJ Newburn, Distinguished Engineer, NVIDIA

Accelerating Object Storage

Contacts: CJ Newburn, Harish Arora (PM), Kiran Modukuri (Architect)

- RDMA for object is worth doing
  - Accelerates back-end storage and makes it more efficient
  - Relieves performance bottlenecks on compute side when they exist
- What NVIDIA has been doing
  - cuObject libraries for client and server to enable RDMA for S3 on MLNX
  - Published wire protocol
  - Proposed plugin architecture and reference implementation for AWS
  - Working with storage partners to assure resilient, robust perf at scale
- SNIA opportunity
  - Focus on amplifying available solutions to meet immediate needs

 ©2026 SNIA. All Rights Reserved.

Benchmarking Methodology

- High-performance storage as essential to supporting AI workloads
  - Core requirements: Performance, resilience, and consistency
- Disciplined benchmarking methodology and assessment framework
  - Provide greater customers confidence, enable more predictable outcomes in production
- Evaluations should reflect the demands of modern AI Factory workloads
  - Training, fine-tuning, inference, and specialized inference scenarios
- As AI workloads continue to evolve, benchmarking should also expand
  - New use cases, GPU-accelerated AI frameworks, synthetic workloads, and targeted stress testing aligned to representative AI data patterns
- NVIDIA intends to share its perspective and help inform broader industry discussions on storage benchmarking and assessment

 ©2026 SNIA. All Rights Reserved.