NVIDIA

# Advancing Memory and Storage Architectures for Next-Gen AI Workloads

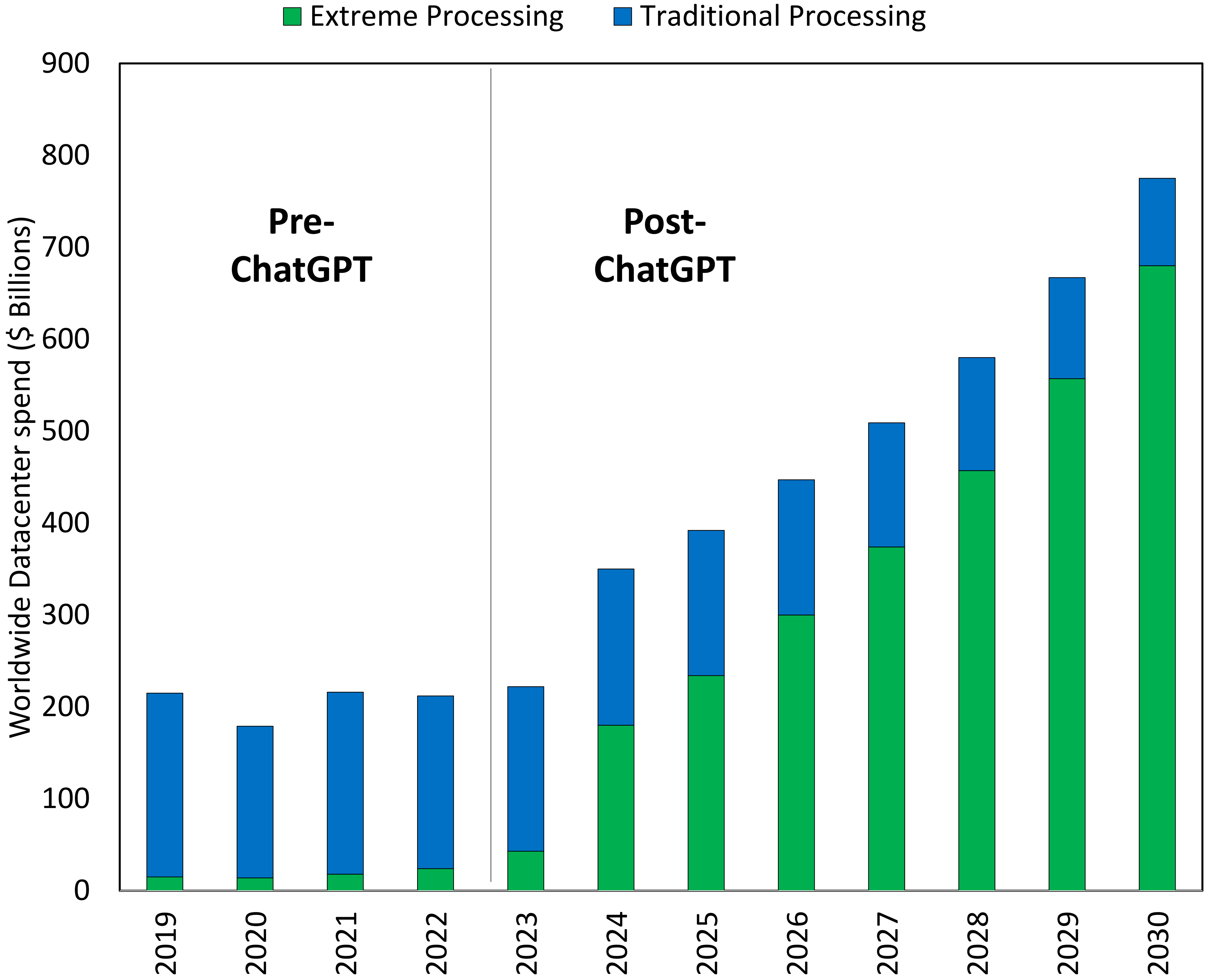
Vikram Sharma Mailthody

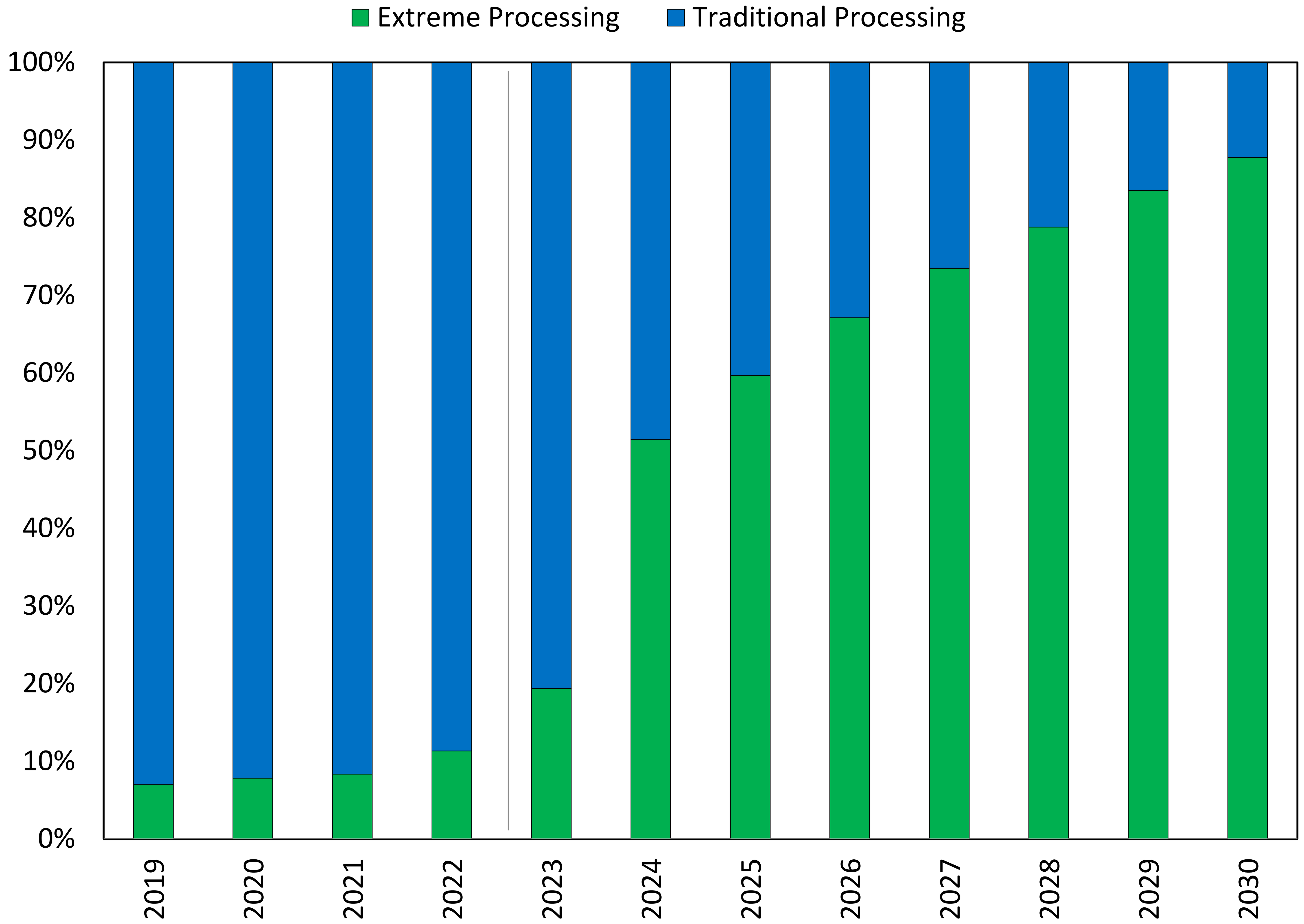
Sr. Research Scientist, NVIDIA Research

FMS

# Datacenter Compute Inversion

Extreme processing using GPUs and accelerators is now a norm in datacenter




Data source: theCUBE Research 2025, Research & Visualization by David Floyer

FMS

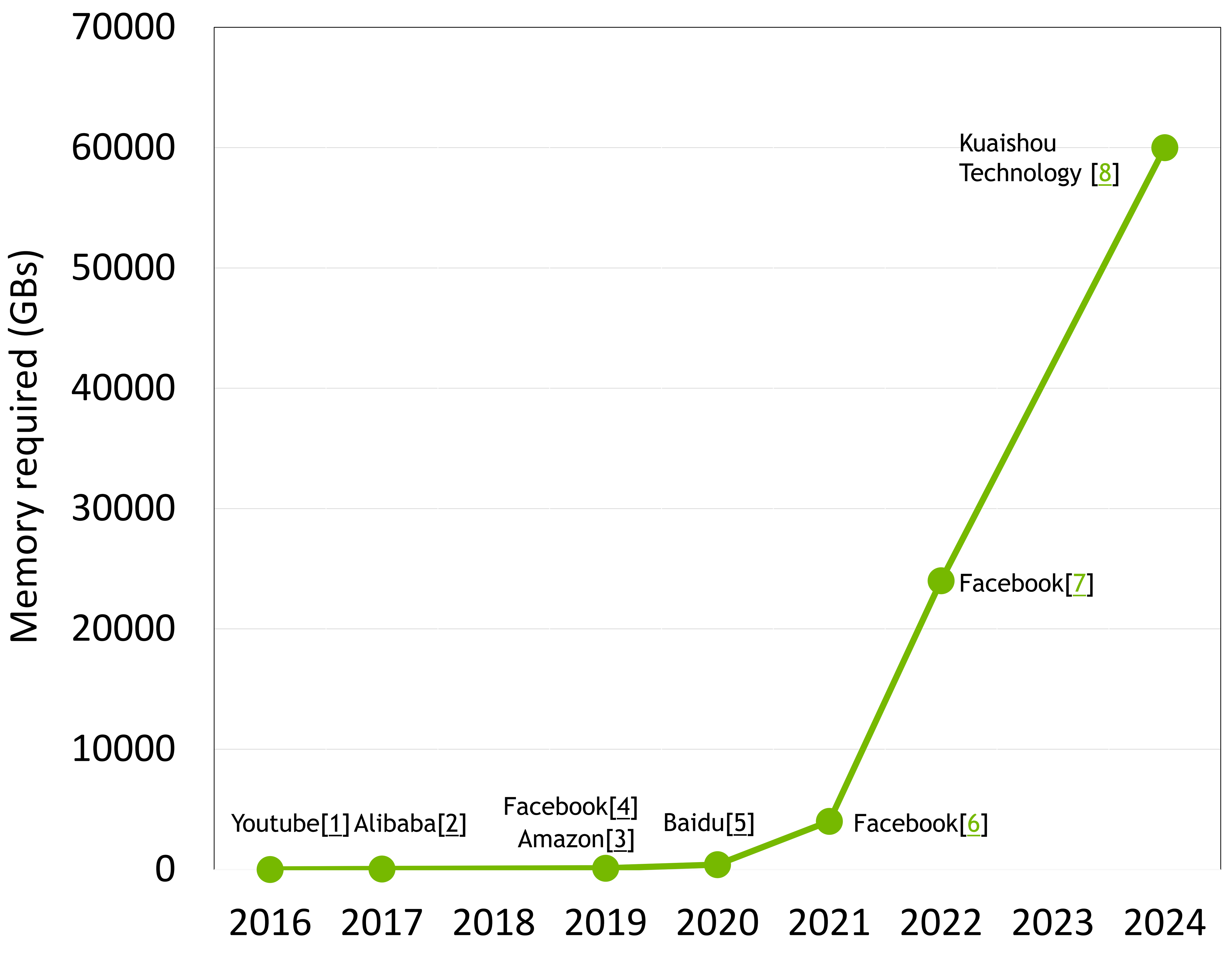
NVIDIA

# Emerging Generation And Prediction Based Applications

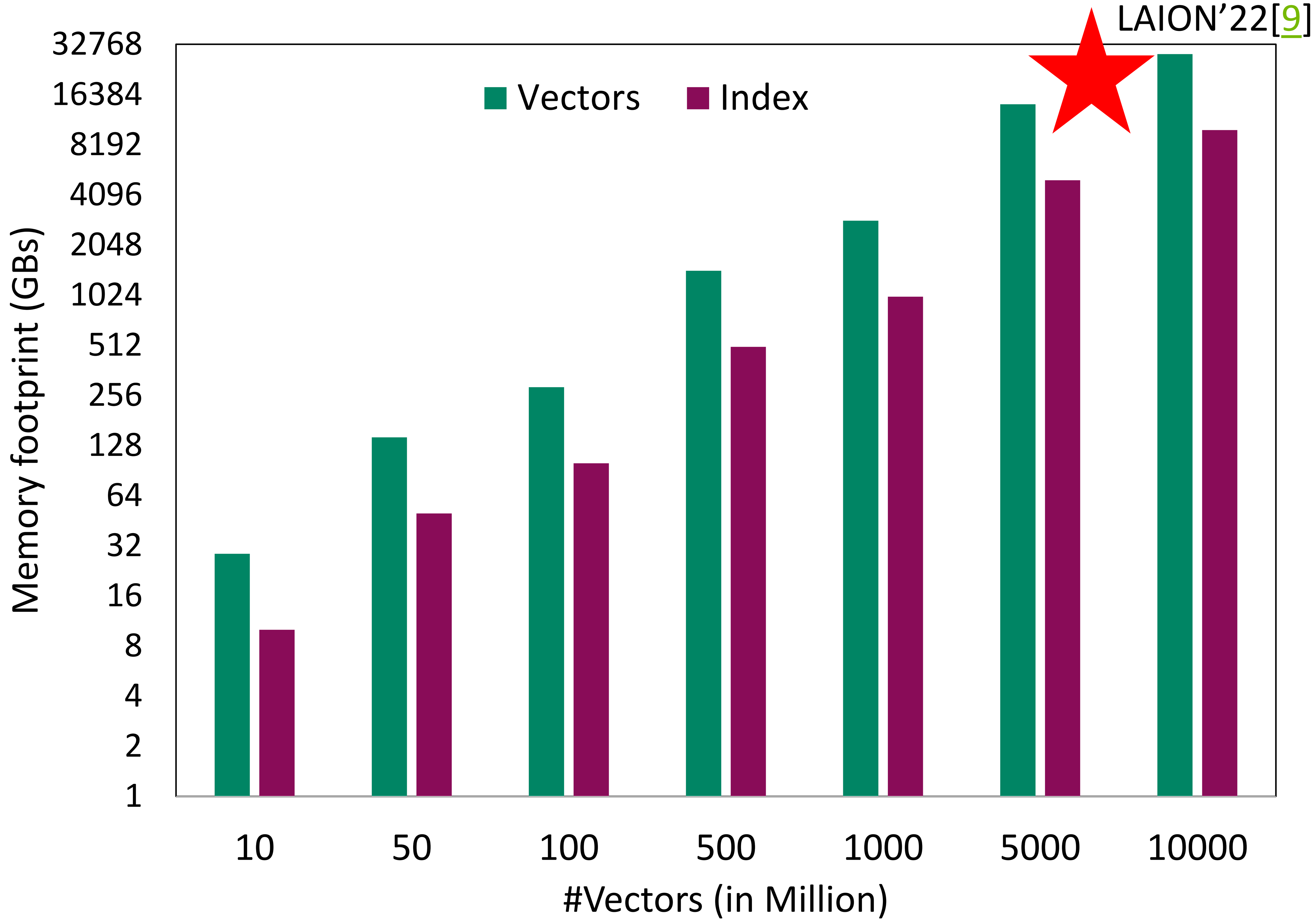
Prediction opens up new venues to revenue

| | Generation | Prediction |
|---|---|---|
| Output | Query response<br>Generated code<br>Art, e.g. designs, music | Recommendation<br>Filtered selection<br>TopK from vector search |
| Usages | Chatbots, reasoning | eCommerce, social media, enterprise market |
| Cost of error | Wrong, hallucinated answer<br>Debugging, escaped bugs<br>Artistic artifact | Mispredictions ok if usually right |
| Infrastructure | LLM | Two tower models, GNN, relational graph transformer,<br>Prefill heavy LLM inference |
| Computational complexity | High | Modest |
| Data needed in memory | Model weights | Knowledge/relational graph/embeddings |
| Data needed in storage | KV$ | Same, for bigger problems |
| Challenges | Running out of public data | Retrieval heavy, Many new horizons ("big data") |

# Emerging workloads demand memory




**Recommender Systems**
**1TB to 100TBs**




**Vector databases**
**1TB to 100TBs**

FMS

NVIDIA

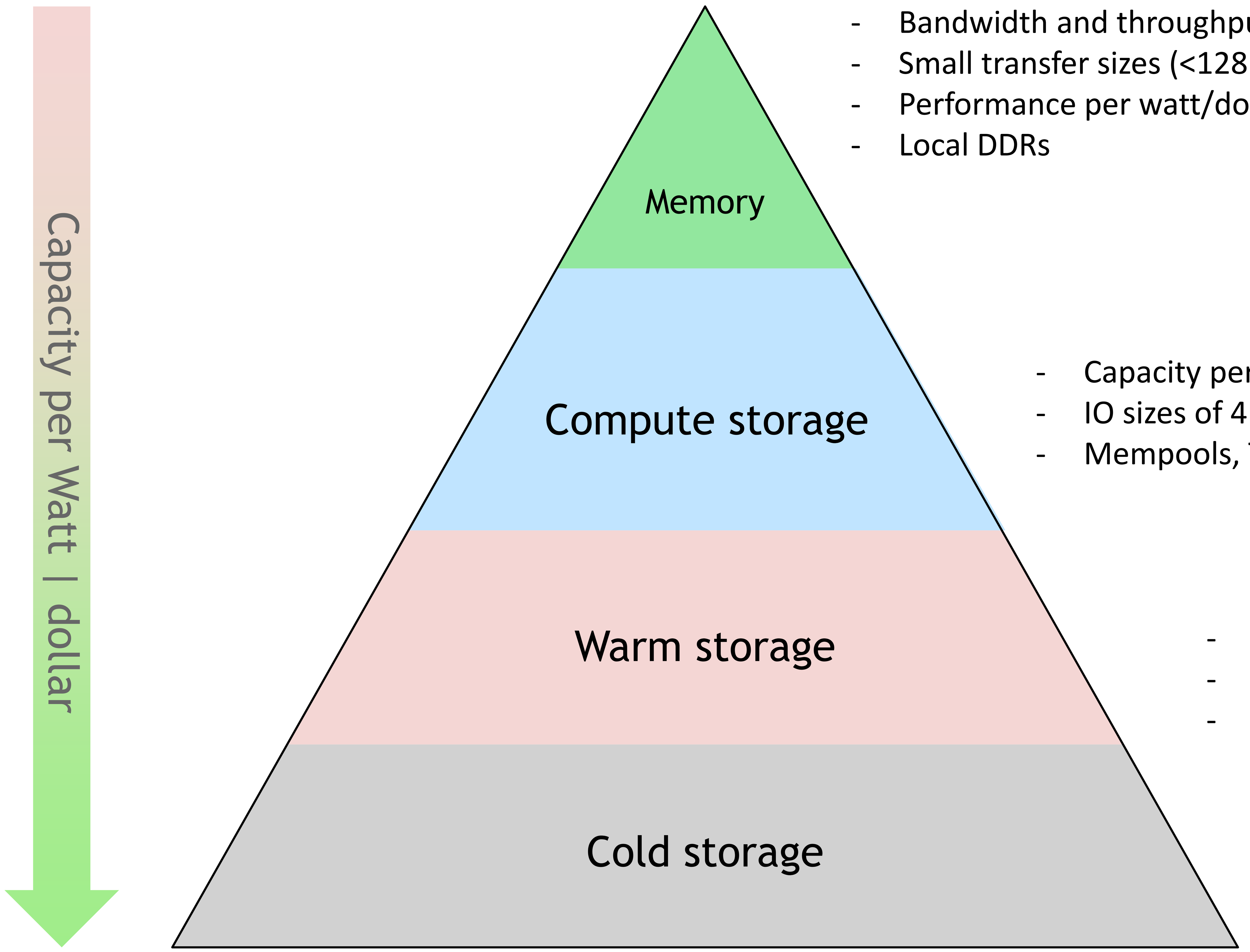
# AI application overview

Apps bifurcate by access pattern and IO intensity; TB/TCO persists, IOPS/TCO is emerging

| Area | Usage model | Applications | Access granularity | Total size /worker |
|---|---|---|---|---|
| Training | Checkpoint save/restore | LLM pretraining, fine tuning | 10MB – 1sGB | 1-10TB |
| Inference | KV context caching across queries, docs | LLM inference | 8KB – 4MB | >10sTB |
| | LLM+GNN, GNN+LLM | Contextual LLMs | 512B – 8KB | 5TB – 400TB |
| | Vector database | Dynamic Index build | 64B – 4KB | 6.4Gb – 20TB |
| | | LLM RAG doc retrieval | 512B – 8KB | 400GB – 1PB |
| Predictive AI | | Graph RAG | 64B – 8KB | 400GB – 1PB |
| | | Recommenders | 64B – 4KB | 5TB – 400TB |
| | GNN induced subgraphs | eCommerce, fraud, social networks | 512B – 8KB | >2TB |
| | Anomaly detection | eCommerce, fraud, social networks | 512B – 8KB | >10TB |
| | Relational graphs | Data Science Automation | 8B – 4KB | >100sTBs |

# Data hierarchy

## Pre-GenAI era




- Bandwidth and throughput
- Small transfer sizes (<128B)
- Performance per watt/dollar
- Local DDRs

- Capacity per watt/dollar
- IO sizes of 4K
- Mempools, TLC Flash locally attached

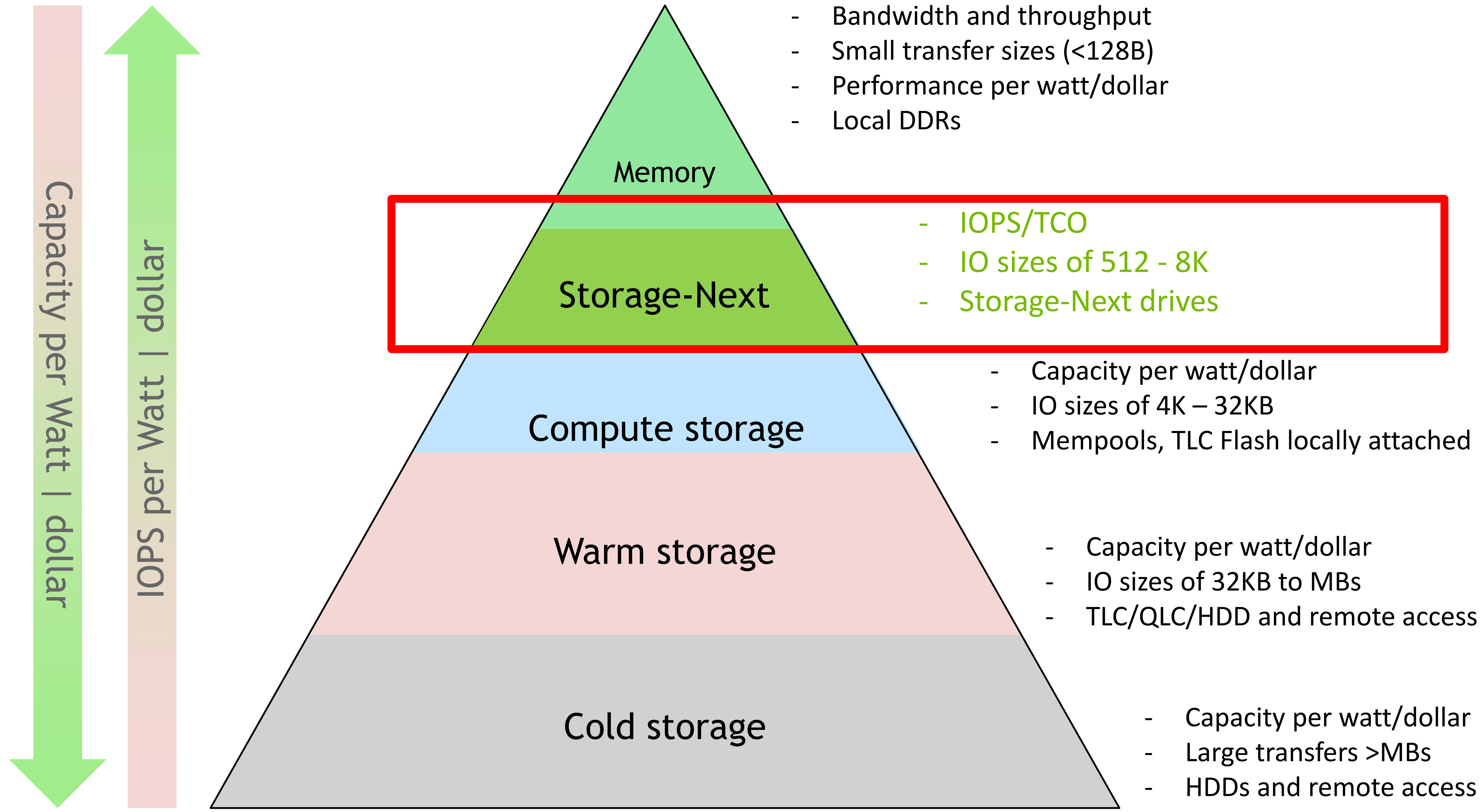
- Capacity per watt/dollar
- IO sizes of 32KB to MBs
- TLC/QLC/HDD and remote access

- Capacity per watt/dollar
- Large transfers >MBs
- HDDs and remote access

Figures modified from Ta-Yu OCP Storage June call

# Data hierarchy

Apparent Post-GenAI era




- Bandwidth and throughput
- Small transfer sizes (<128B)
- Performance per watt/dollar
- Local DDRs

- IOPS/TCO
- IO sizes of 512 - 8K
- Storage-Next drives

- Capacity per watt/dollar
- IO sizes of 4K – 32KB
- Mempools, TLC Flash locally attached

- Capacity per watt/dollar
- IO sizes of 32KB to MBs
- TLC/QLC/HDD and remote access

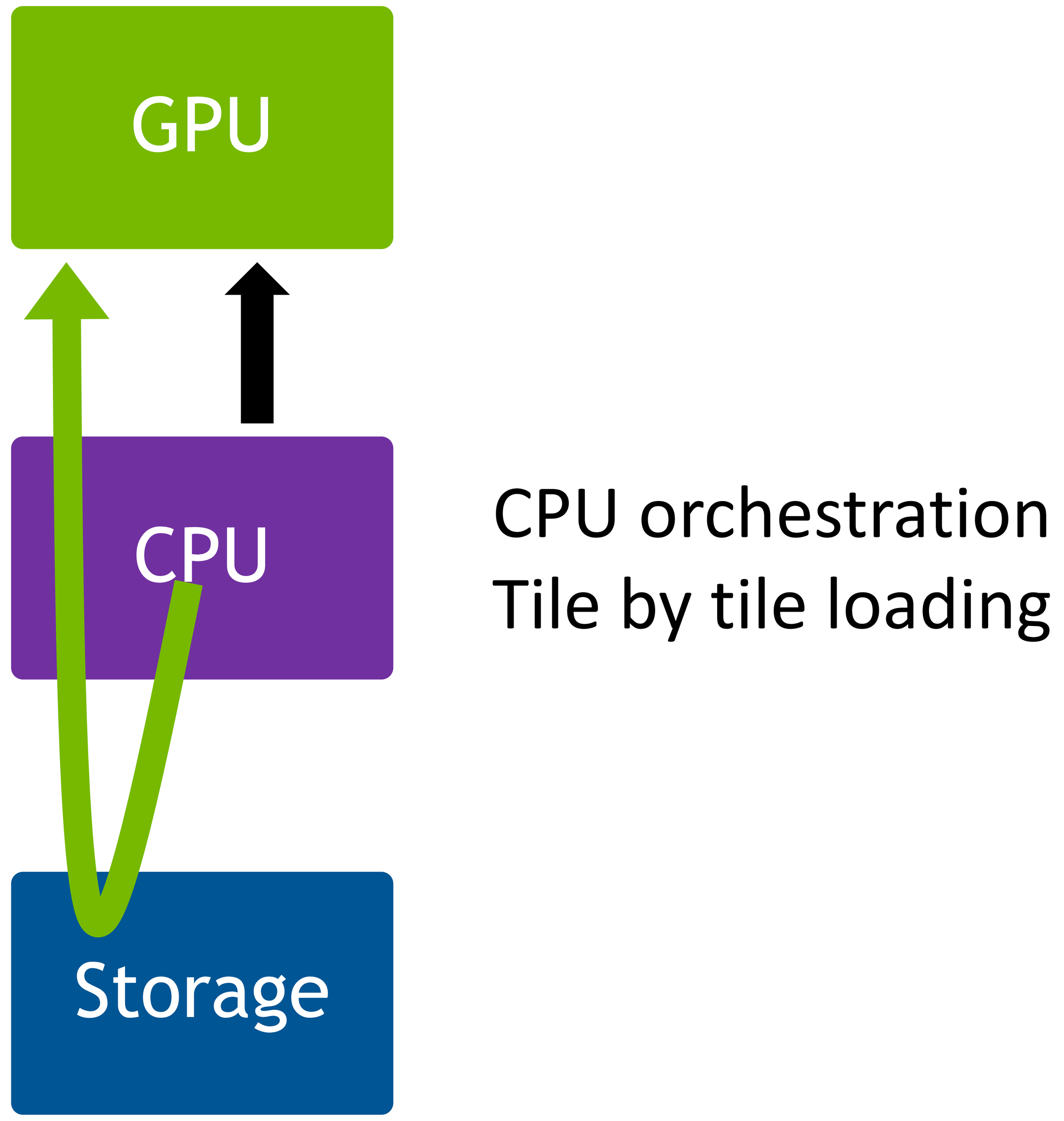
- Capacity per watt/dollar
- Large transfers >MBs
- HDDs and remote access

Trends existed in early 2016-2017

Figures modified from Ta-Yu OCP Storage June call

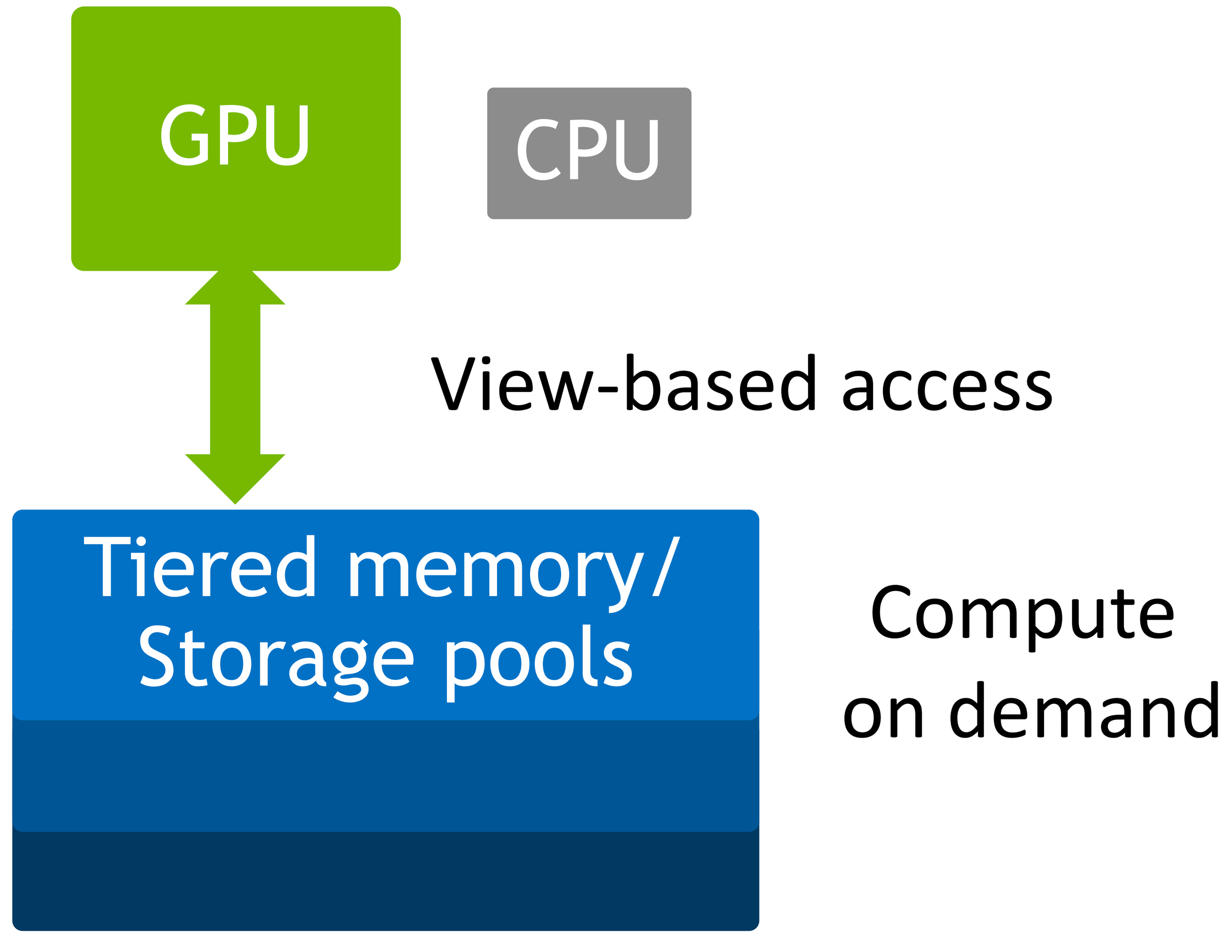
# From Offload Devices to Orchestrators

## Rethinking the Accelerator-Data Interface




**Current Approach**

- Current Approach
  - CPU has entire control path
  - GPU used as assistant
  - But majority of the work happens in GPU
  - Inefficient to load PB of data via tiling
- What applications need
  - A view-based access with a notion of infinite memory pool
  - Ability to fetch only needed data on-demand during computation
  - Ability to scale compute and data pipelines independently
  - Move majority of control to GPU while using CPU for house keeping




**What applications need**

Can GPU sustain enough parallelism to hide storage memory latency and provide a tiered memory/storage pool for applications?

# Software and Storage are the new bottleneck!

Little's Law

$$Q_d = T * L$$

Minimum steady state queue depth = Throughput * Latency

- PCIe x16 Gen6 = 104GBps
  - For 512B access :– T = 104GBps/512B = 208 M IOPS
  - For 4KB access :– T = 104GBps/4KB = 26M IOPS

- Assume SSD average access latency = 100us
  - Qd for 512B = 208 M * 100us = 20, 800
  - Qd for 4KB = 26 M * 100us = 2600

GPUs and emerging workloads have enough parallelism to issue these many requests in-flight[1].

But the software stack and SSDs can't keep up.

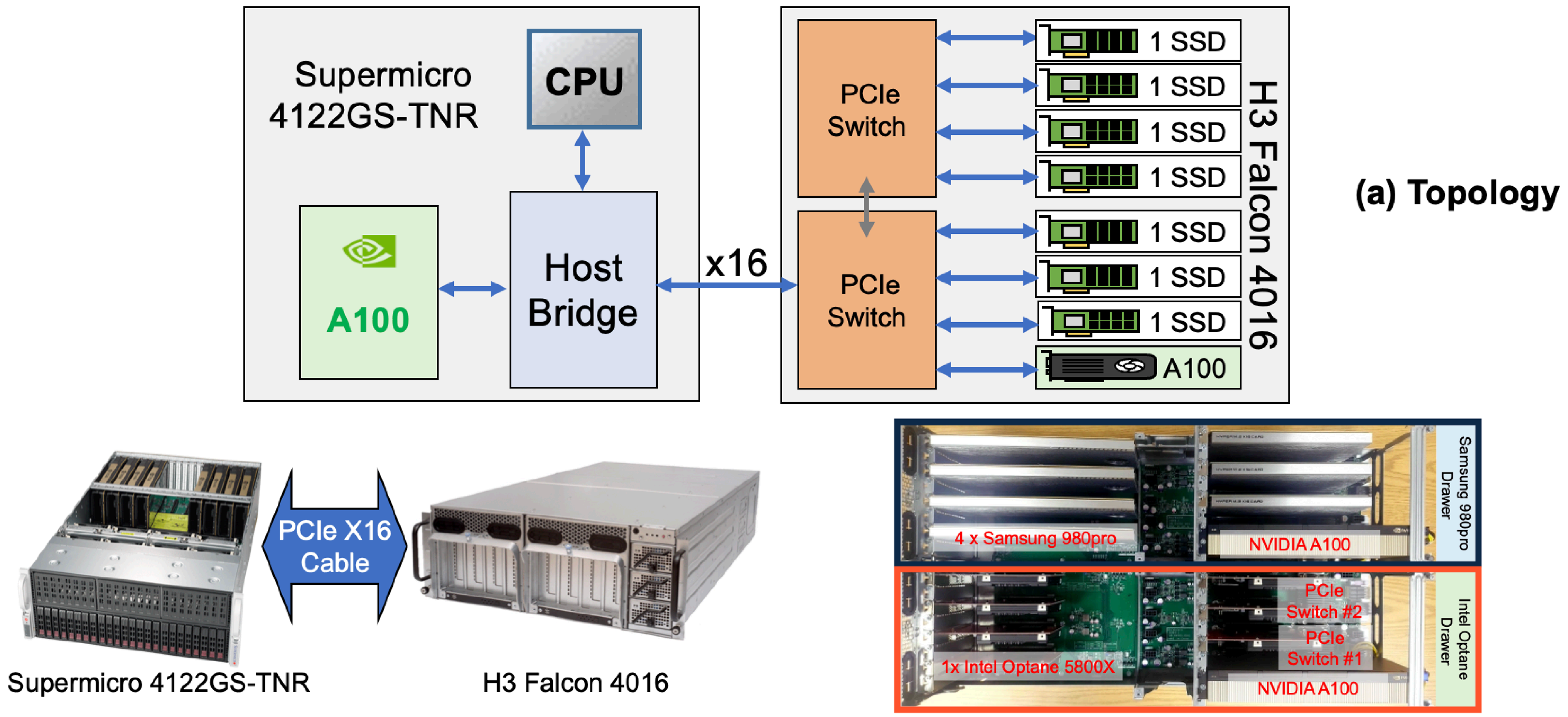
CPU-drive software serialize, batch, or block effectively limiting QD resulting in underutilized bandwidth and stalling accelerators

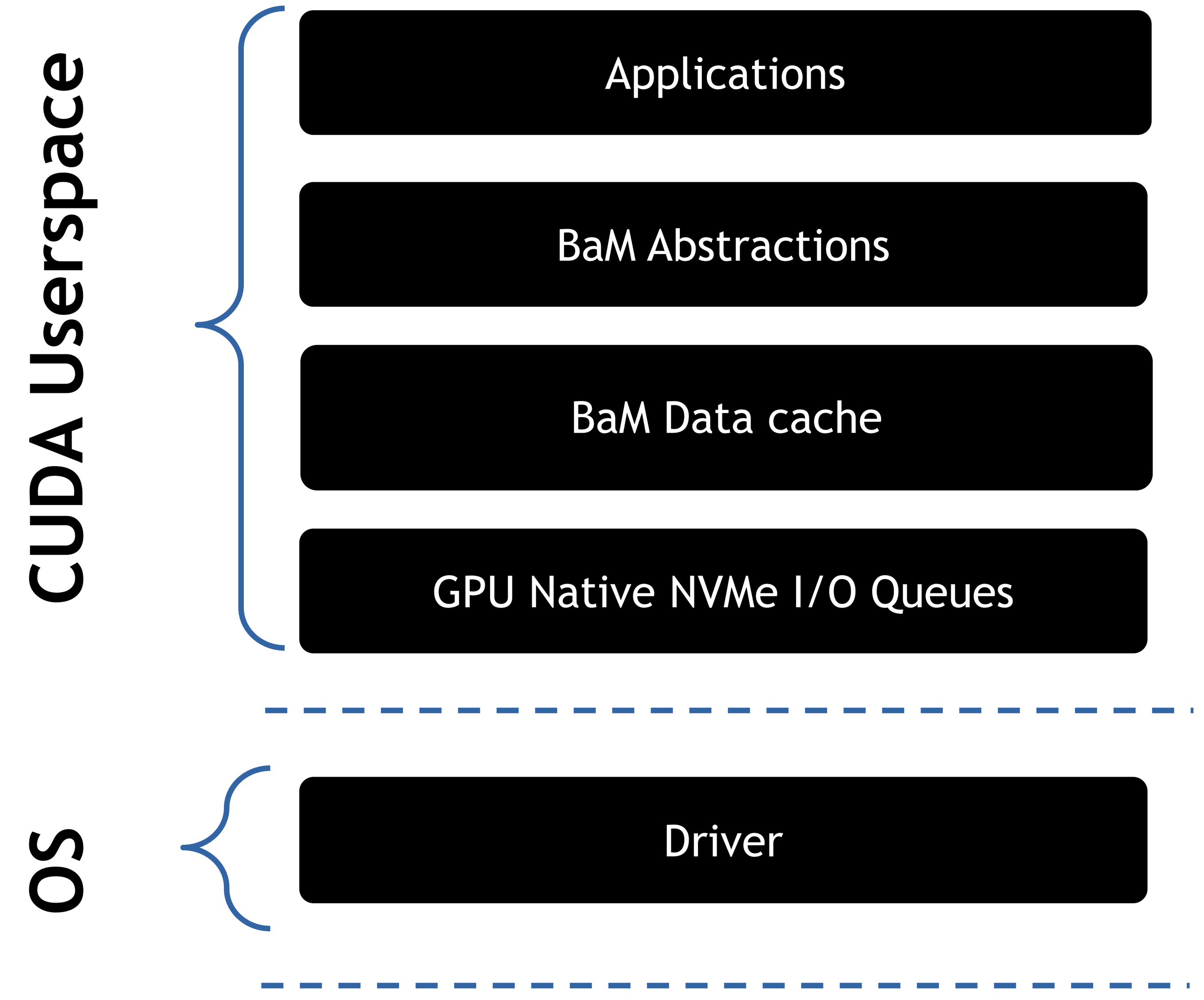
1. BaM: https://arxiv.org/abs/2203.04910

https://arxiv.org/abs/2203.04910
https://github.com/ZaidQureshi/bam

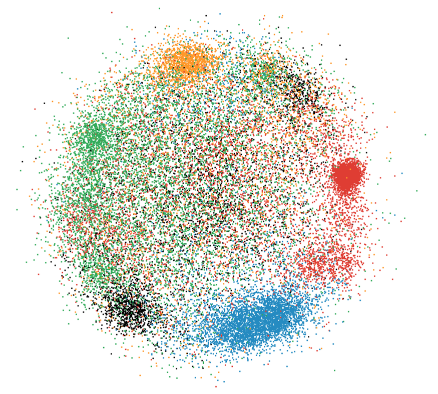
# BaM

## After many failures we got proof of concept prototype



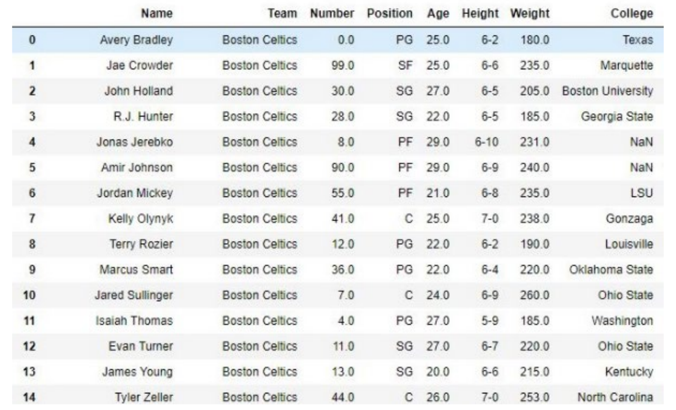




HW Prototype



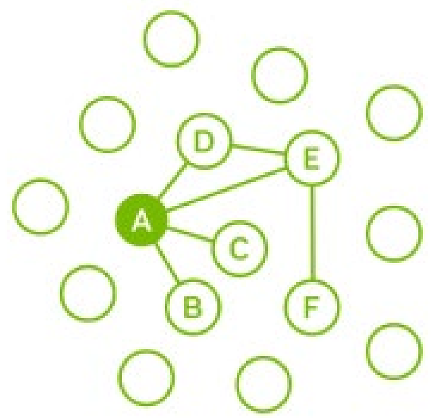
**Large Graphs**

- 1.0X for BFS
  1.5X for CC
- 4.6X Perf/$ for BFS
  6.6X Perf/$ for CC



**Data Analytics**

- Upto 5.3X perf
- Upto 23.6X Perf/$




**GNN**

- Upto 8.3X perf
- Upto 38.3X Perf/$

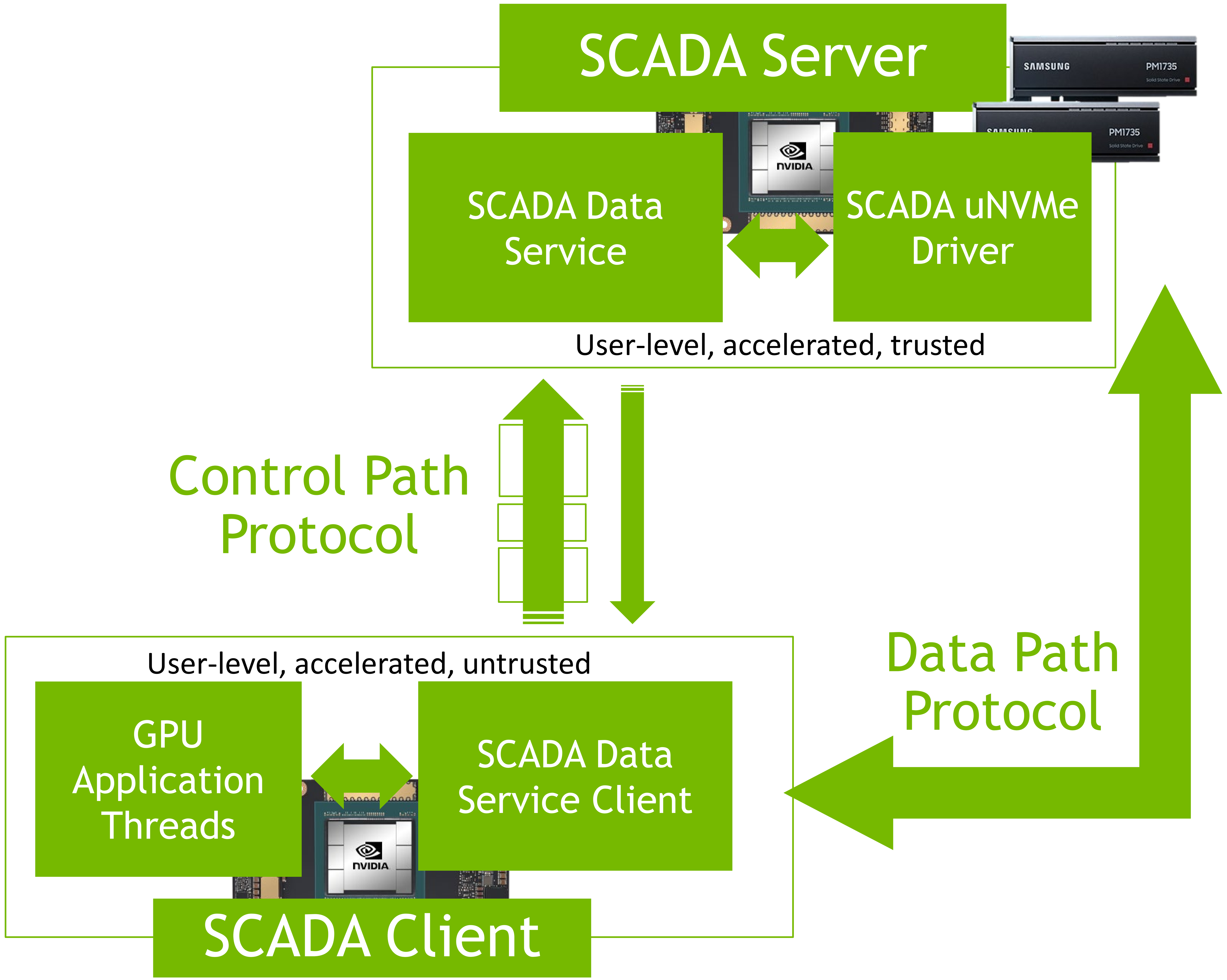
IBM, micron, KIOXIA, SK hynix, MICROCHIP, ByteDance, SAMSUNG, Graid Technology Inc., PHISON, PLiOPS EXTREME DATA PROCESSOR, SOLIDIGM

Many more…

# But …

# SCADA – SCaled Accelerated Data Access

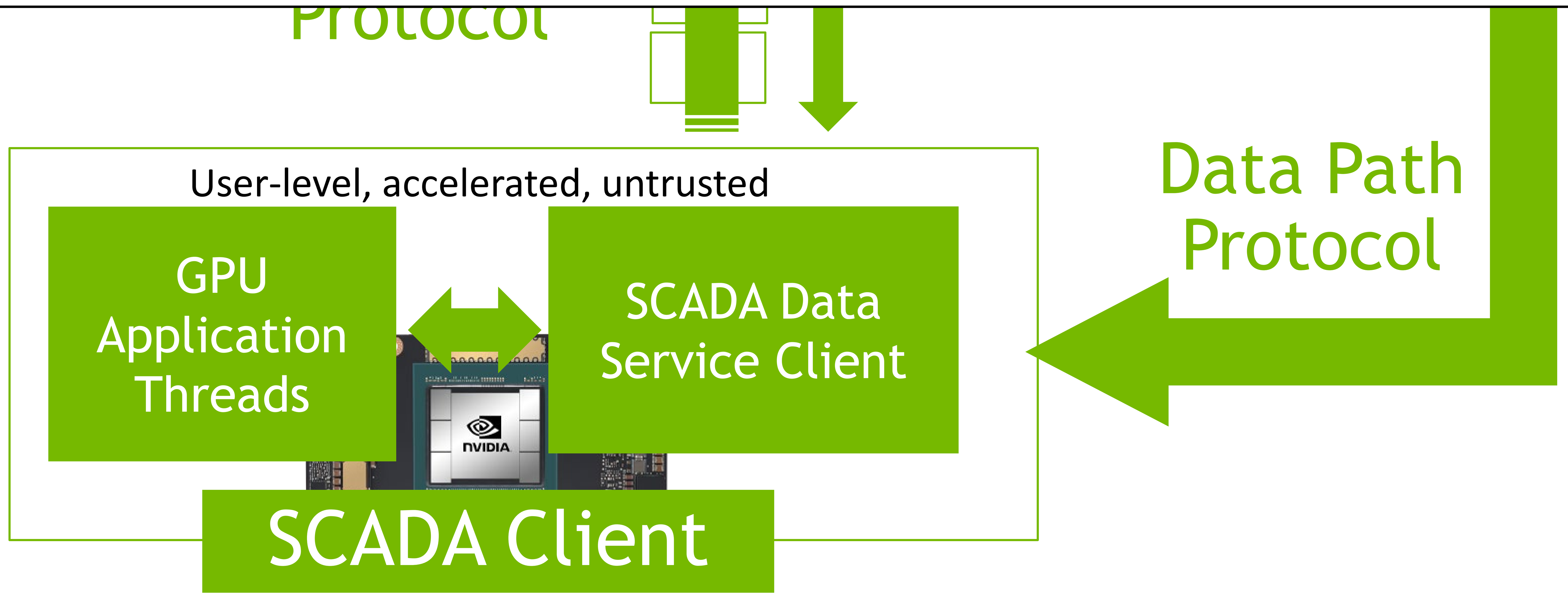
Software Architecture and Components – culmination of 8+ years of research and engineering



# SCADA – Scaled Accelerated Data Access

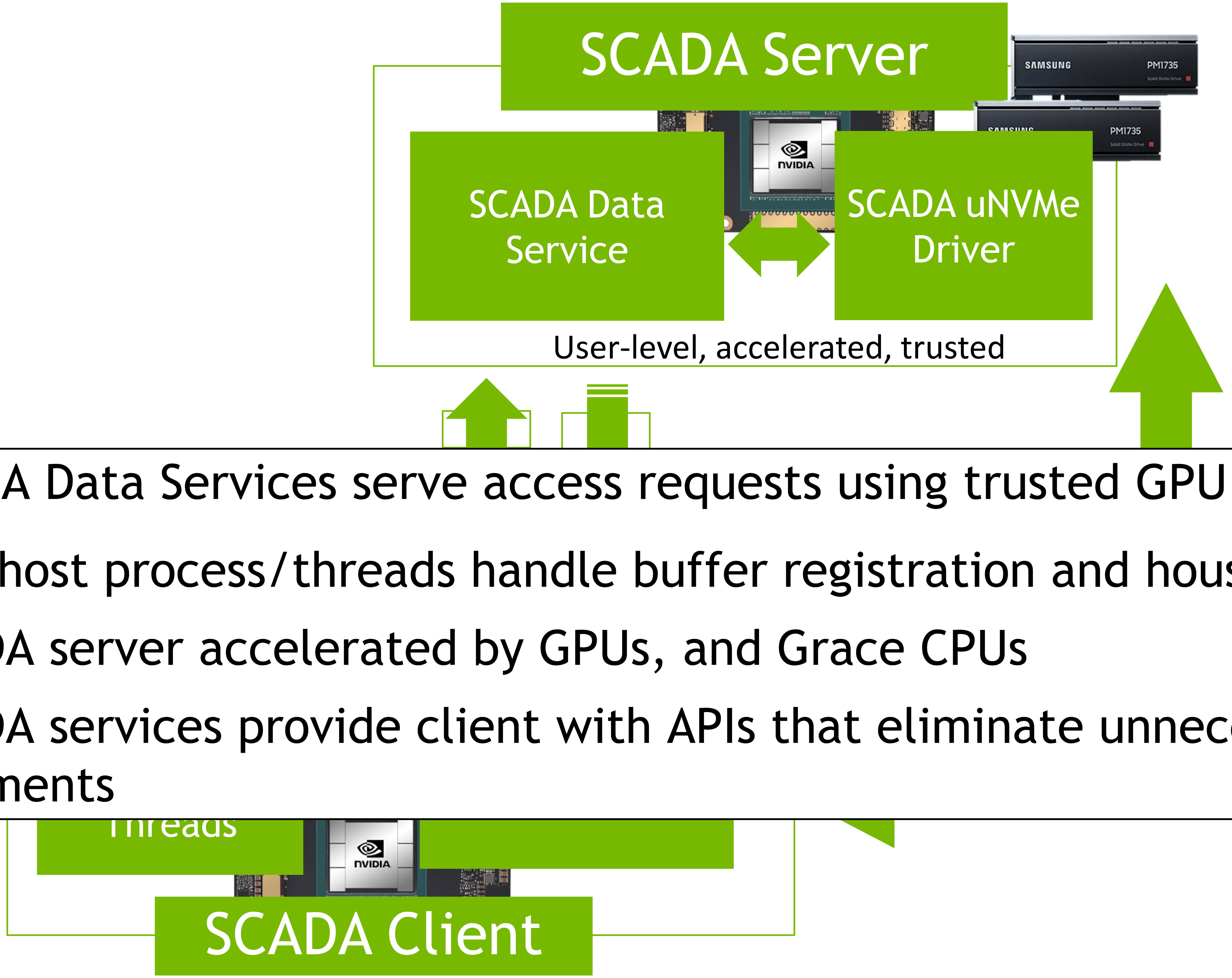
Software Architecture and Components

- ❑ Data Service Client is a header-only library compiled with applications
- ❑ Accelerate data buffer management through application defined (customized) software cache in HBM
- ❑ Provides simple abstractions that works with thrust library
  - ❑ mdspan array, keyvalue, graph, …



# SCADA – Scaled Accelerated Data Access

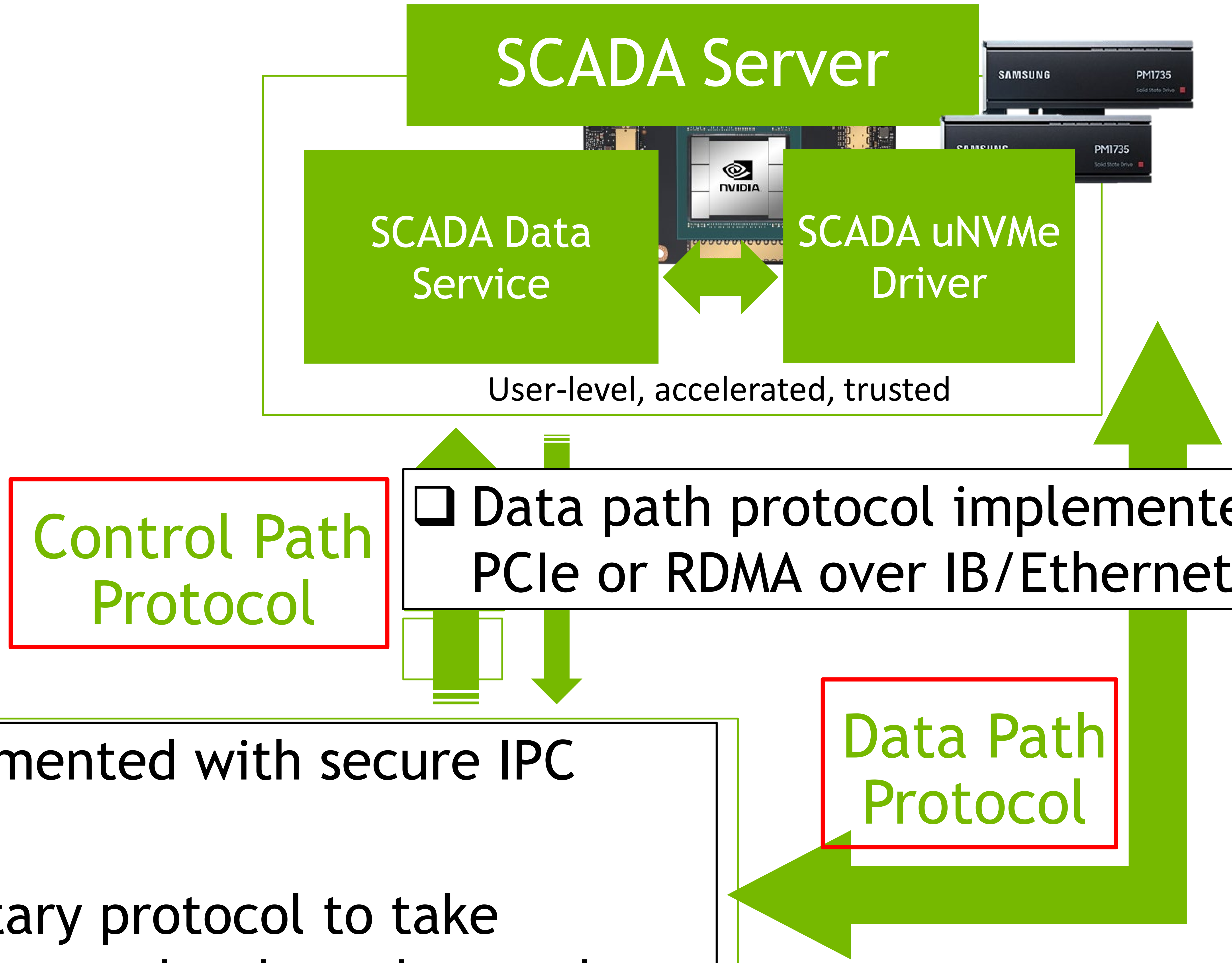
Software Architecture and Components




- ❑ SCADA Data Services serve access requests using trusted GPU kernels
- ❑ CPU host process/threads handle buffer registration and house keeping
- ❑ SCADA server accelerated by GPUs, and Grace CPUs
- ❑ SCADA services provide client with APIs that eliminate unnecessary data movements

# SCADA – Scaled Accelerated Data Access

Software Architecture and Components




- ❑ Data path protocol implemented with DMA over PCIe or RDMA over IB/Ethernet

- ❑ Control path protocol implemented with secure IPC and/or RDMA
  - ❑ New GPU oriented proprietary protocol to take advantage of GPU parallelism and reduce the number of doorbell rings

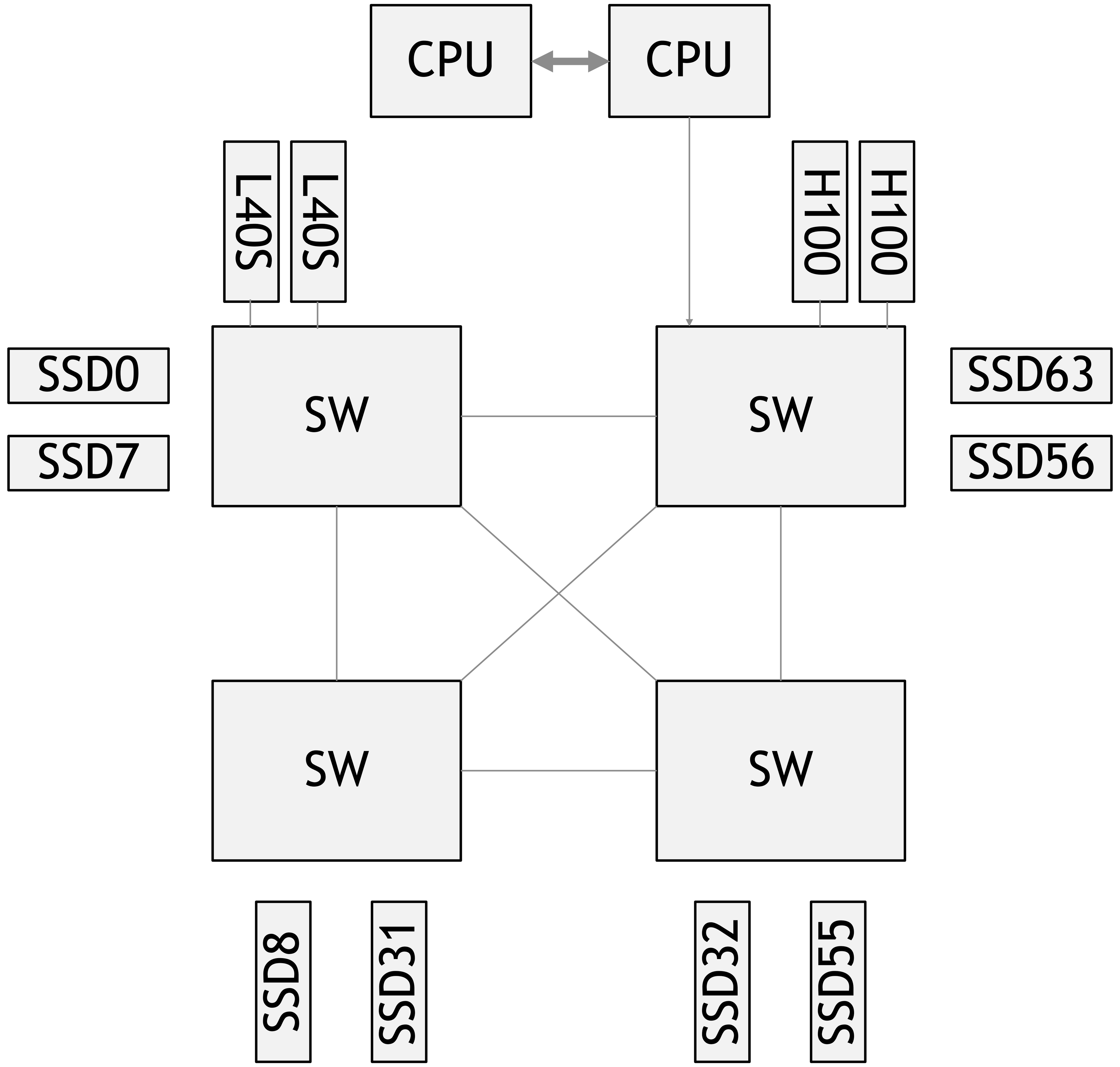
# H3 Platform Gen5 System

CPU BOX

SSD BOX

# H3 Platform Configuration

- INTEL(R) XEON(R) PLATINUM 8568Y
- 192 CPU cores
- 1 TB Memory
- 2 Micron SSD for OS.
- 2 Micron SSD for Data drives
- 64 Samsung E1.S PM9D3A SSDs for SCADA

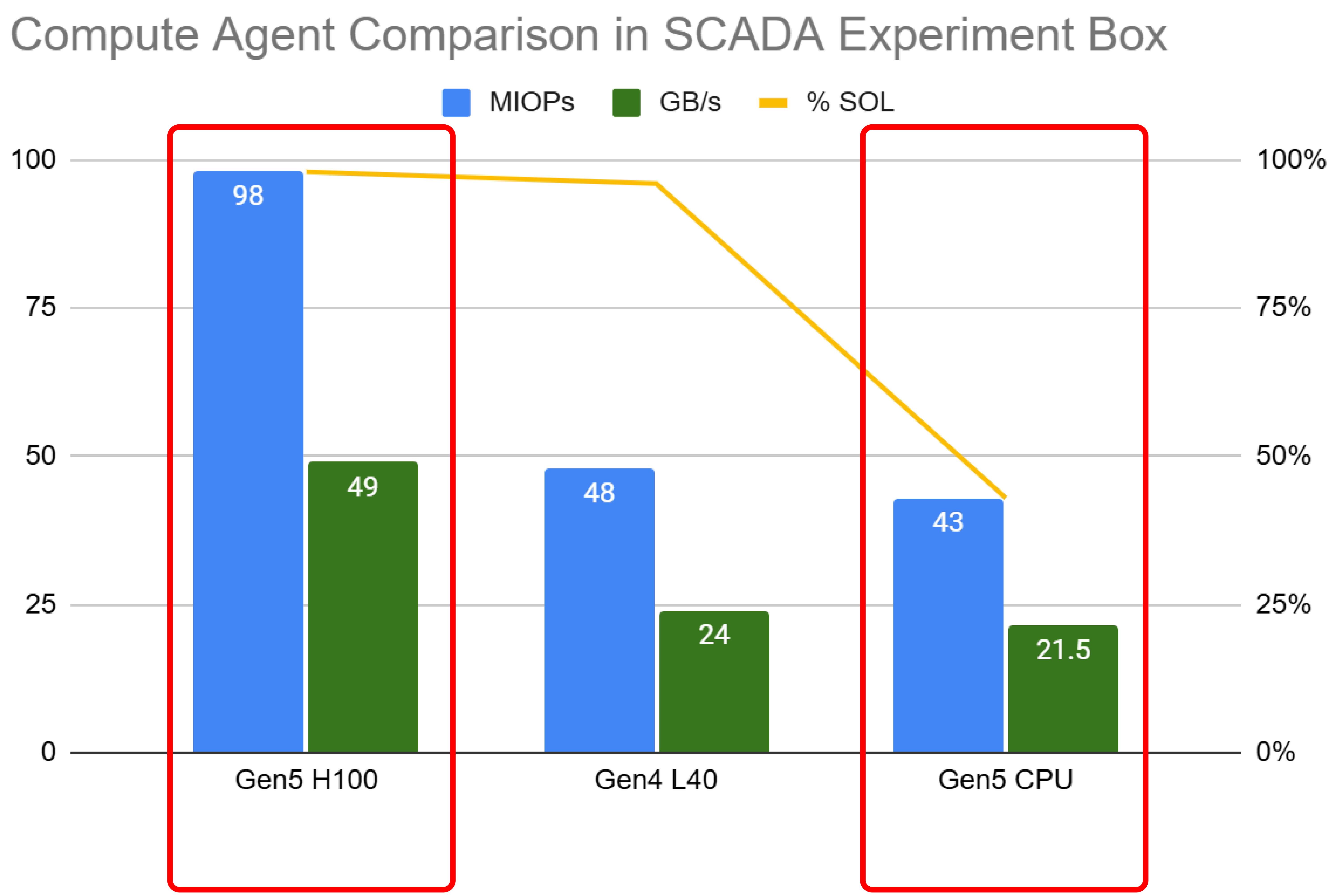



**Config as of today**

# Initial measurement summary

Implications for storage server: GPU vs. CPU (2x) to saturate IOPs, GB/s; PCIe switches in fabric node are efficient

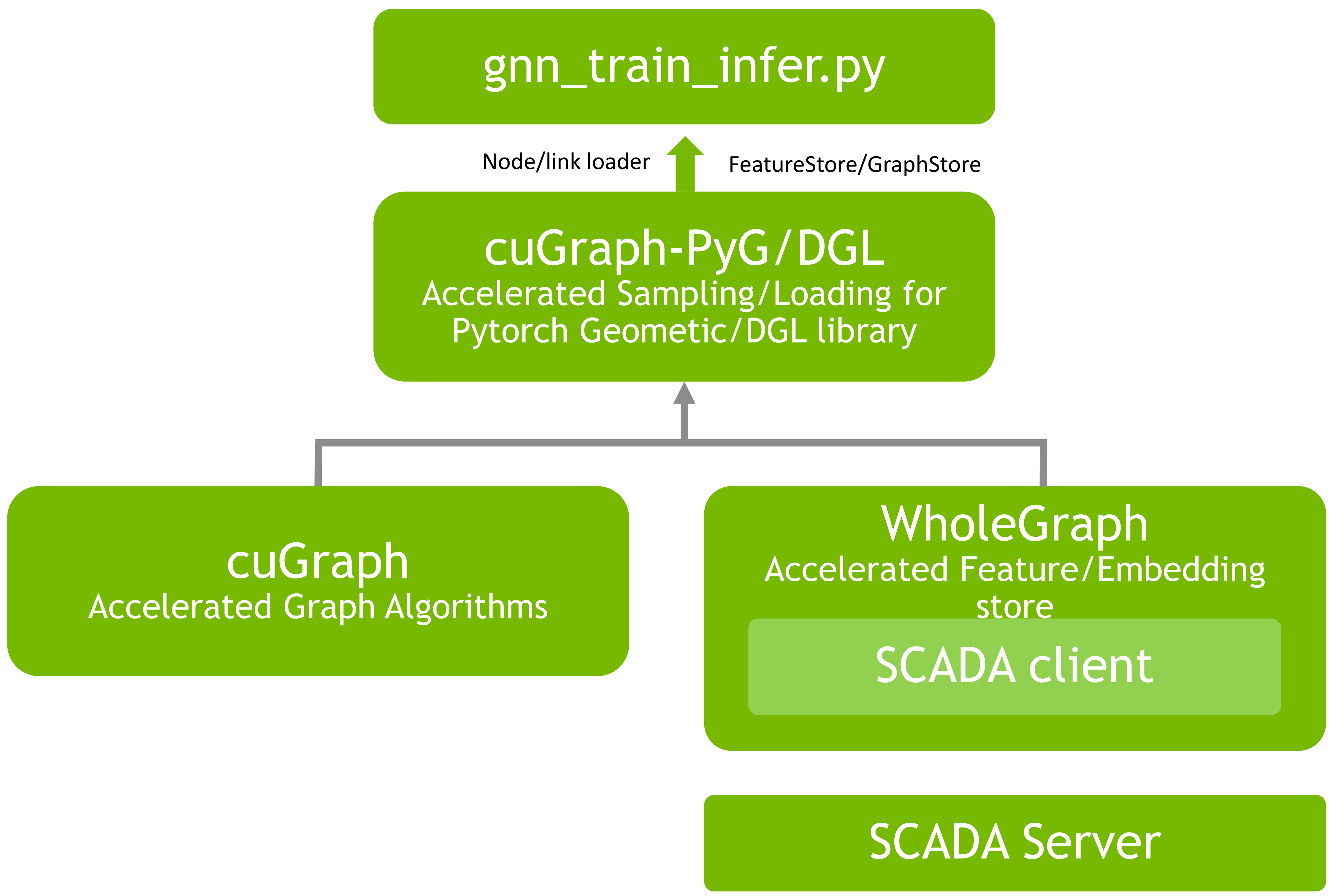
- Measurements from NVMe driver only
  - Full E2E results still being tuned to use more request buffers
  - Measurements made on H3 Storage-Next Experiment server
- H100 nearly saturates IOPs, GB/s
  - Broadcom PCIe switches in fabric mode very efficient
  - 96% of peak bandwidth and 512B IOPs
  - **Achieves that with 24% utilization, not yet tuned**
- Proof! CPU is insufficient to keep up
  - Gen5 GPU is >2x as effective as a Gen5 CPU
  - 1x Gen5 Intel Xeon Platinum, 48C 4GHz
- Modest GPU's perf isn't bad for its Gen
  - L40S is Gen4, there's no equivalent in Gen5
    - Based on Ada, not Hopper. Fewer SMs, lower HBM bandwidth
  - Cheaper GPU is 98% as efficient (48/98*2) for its generation



https://arxiv.org/abs/2302.13522

# SCADA GNN Training

Preliminary results using MLPerf GNN benchmark on IGBH dataset using Gen5 System





**SCADA DOES NOT USE CPU MEMORY IN THIS EXPERIMENT**

1 H100 GPU with 4 Samsung PM9D3A Gen5 drives

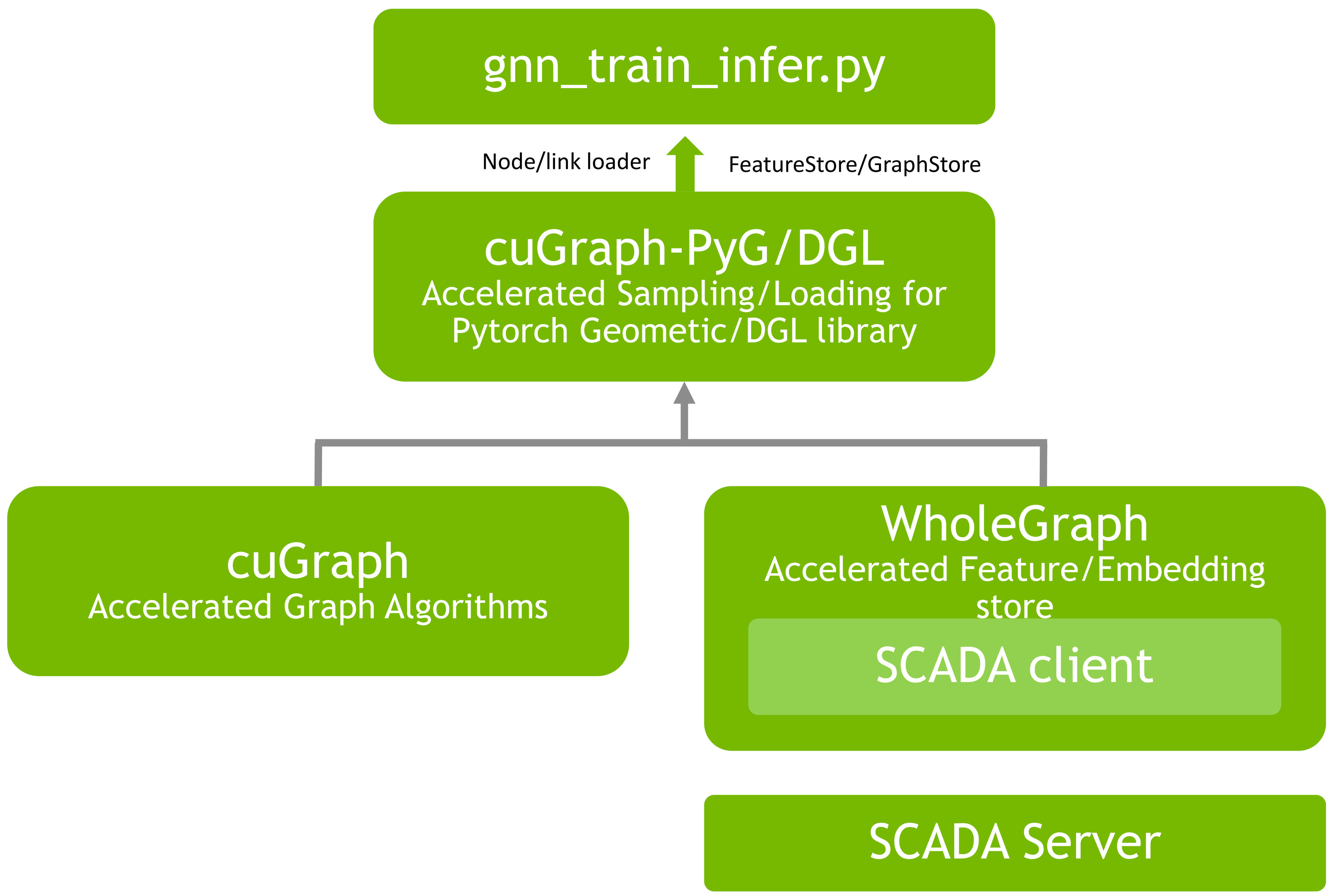
IGBH-medium run with bz=2K, IGBH-Full with bz=1K, 4KB access.

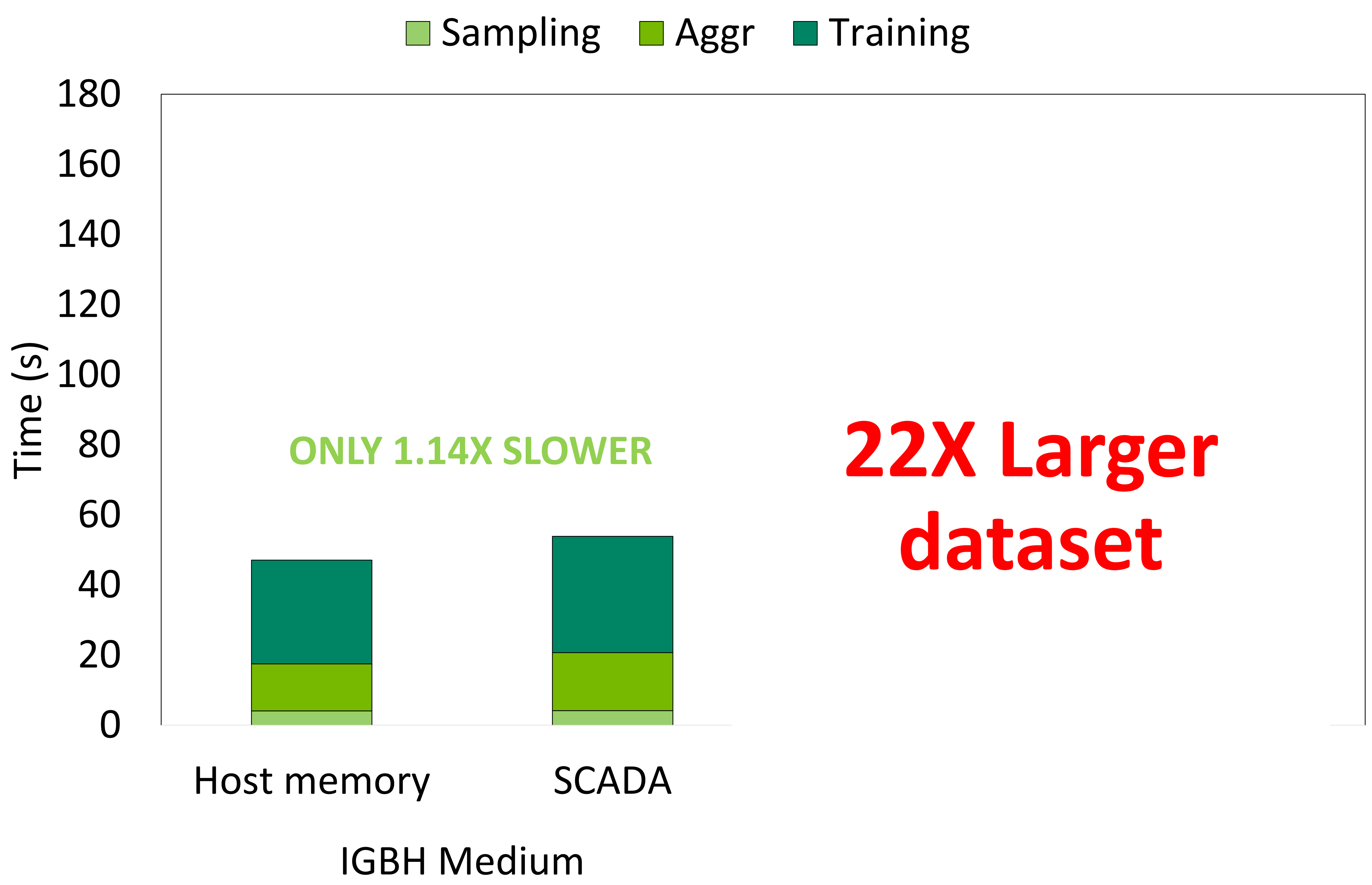
GNN RGAT model

https://arxiv.org/abs/2302.13522

# SCADA GNN Training

Preliminary results using MLPerf GNN benchmark on IGBH dataset using Gen5 System




**SCADA DOES NOT USE CPU MEMORY IN THIS EXPERIMENT**

1 H100 GPU with 4 Samsung PM9D3A Gen5 drives

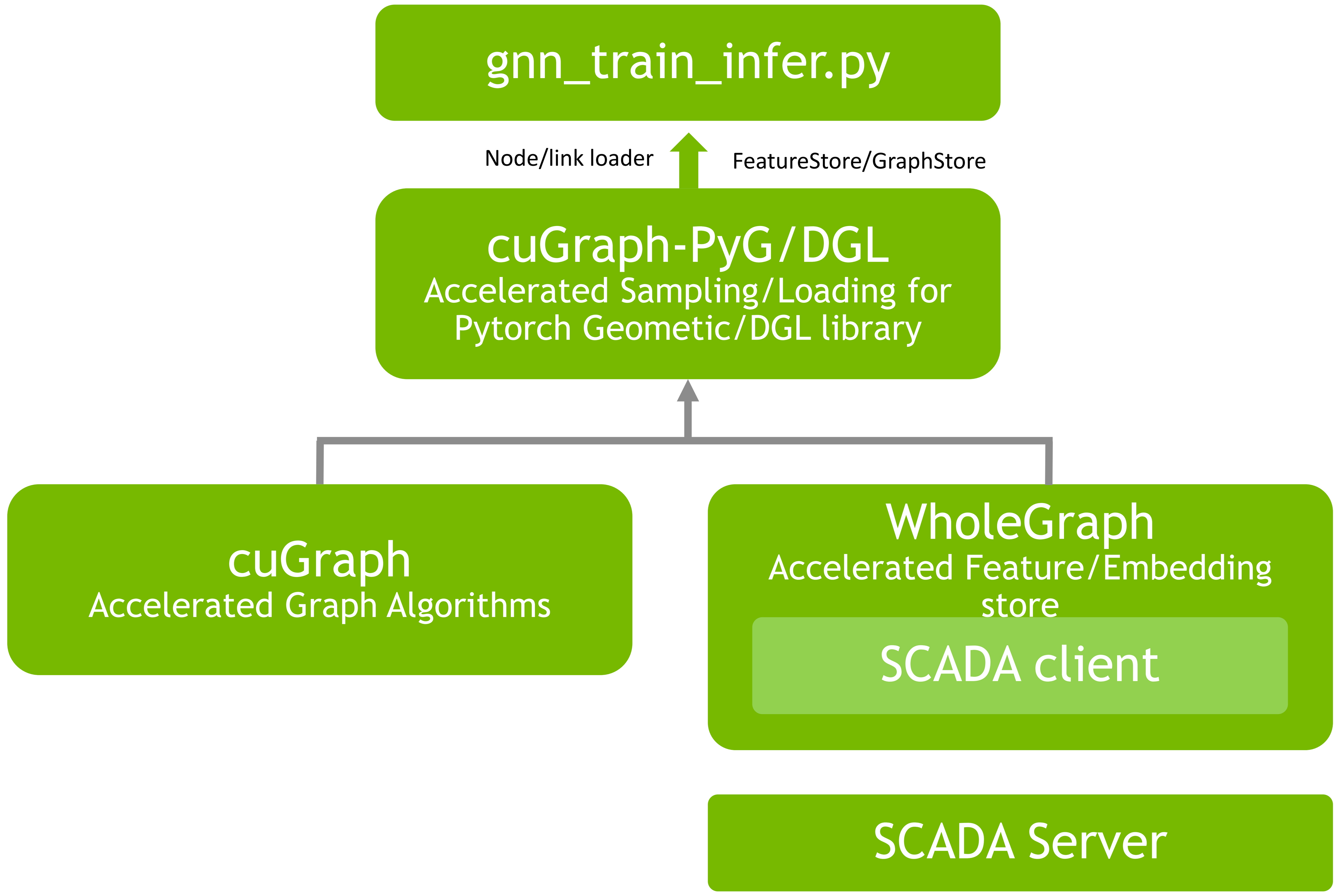
IGBH-medium run with bz=2K, IGBH-Full with bz=1K, 4KB access

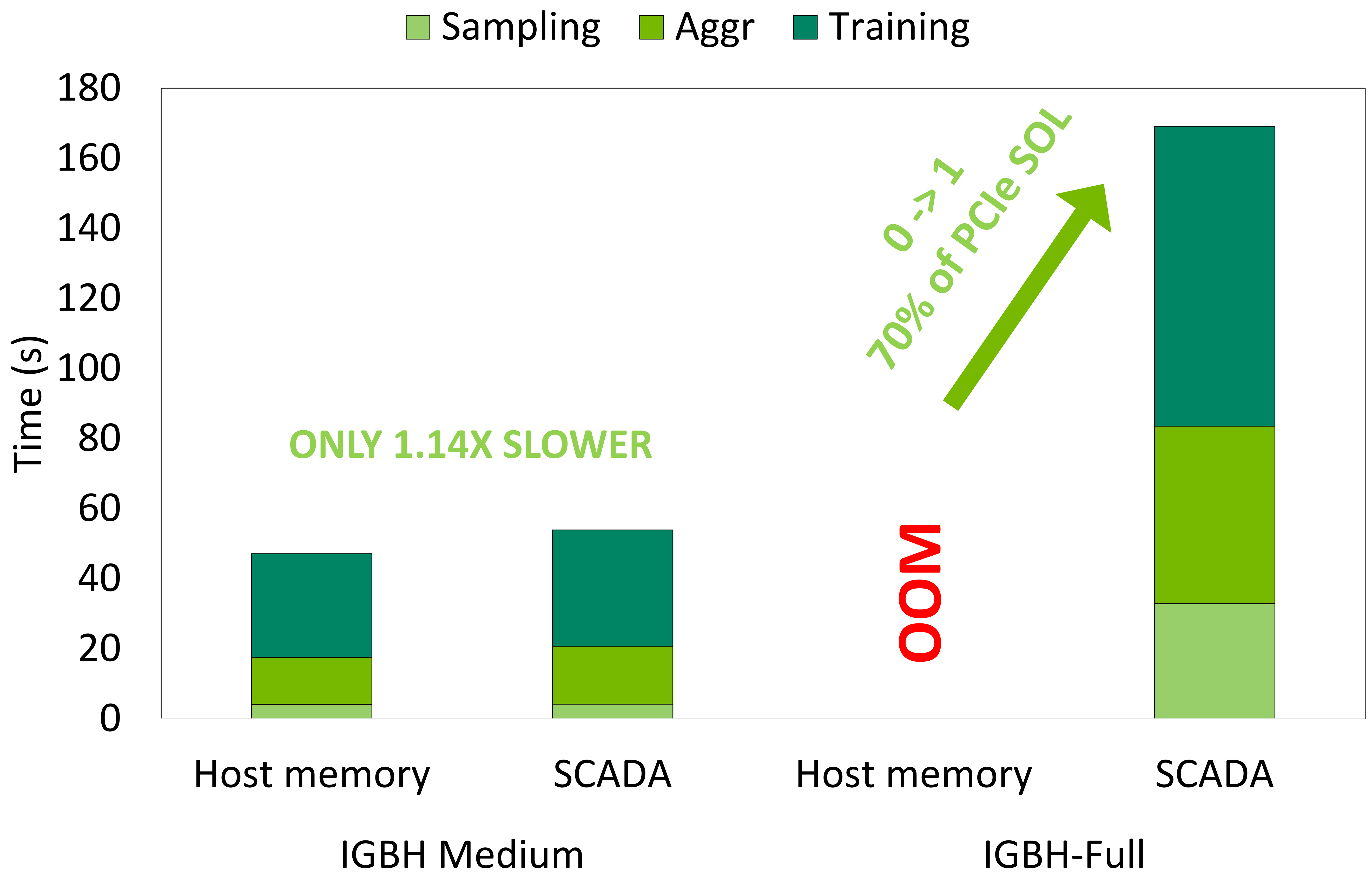
GNN RGAT model

https://arxiv.org/abs/2302.13522

# SCADA GNN Training

Preliminary results using MLPerf GNN benchmark on IGBH dataset using Gen5 System




**SCADA DOES NOT USE CPU MEMORY IN THIS EXPERIMENT**

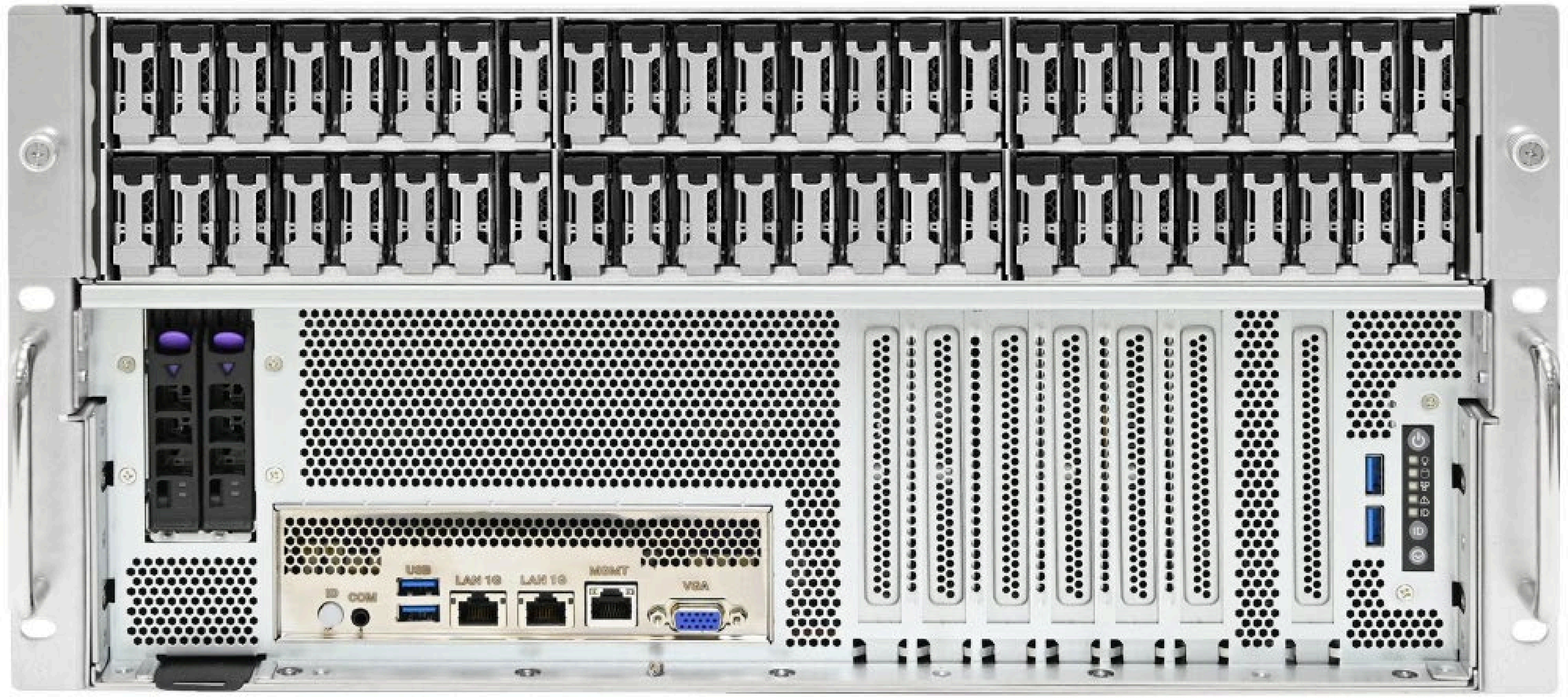
Single H100 GPU with 4 Samsung PM9D3A Gen5 drives

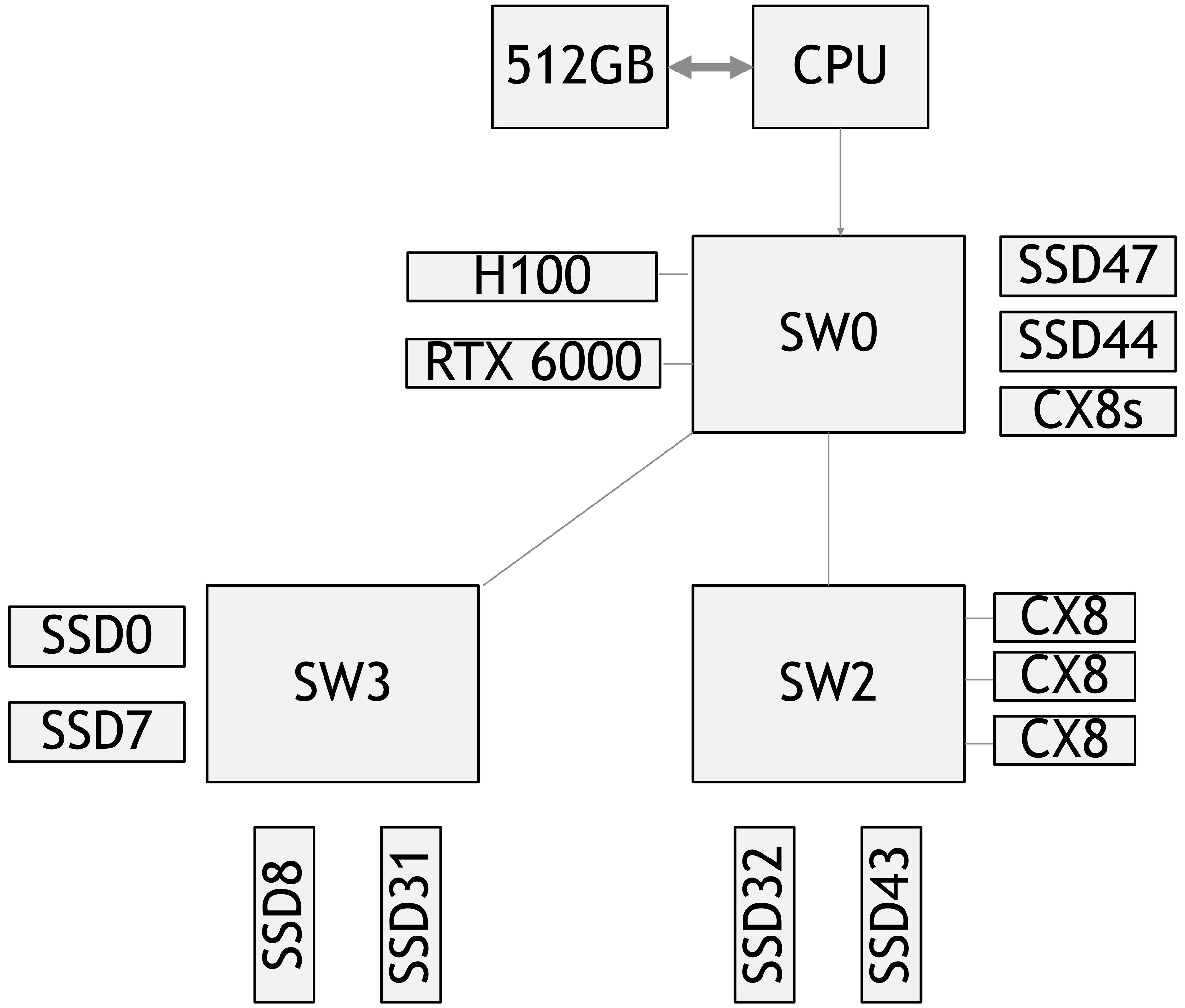
IGBH-medium run with bz=2K, IGBH-Full with bz=1K, 4KB access.

GNN RGAT model

# H3 Platform Gen6 System

48 E1.S SSD Platform with up to 4 GPUs and 6 CX8s





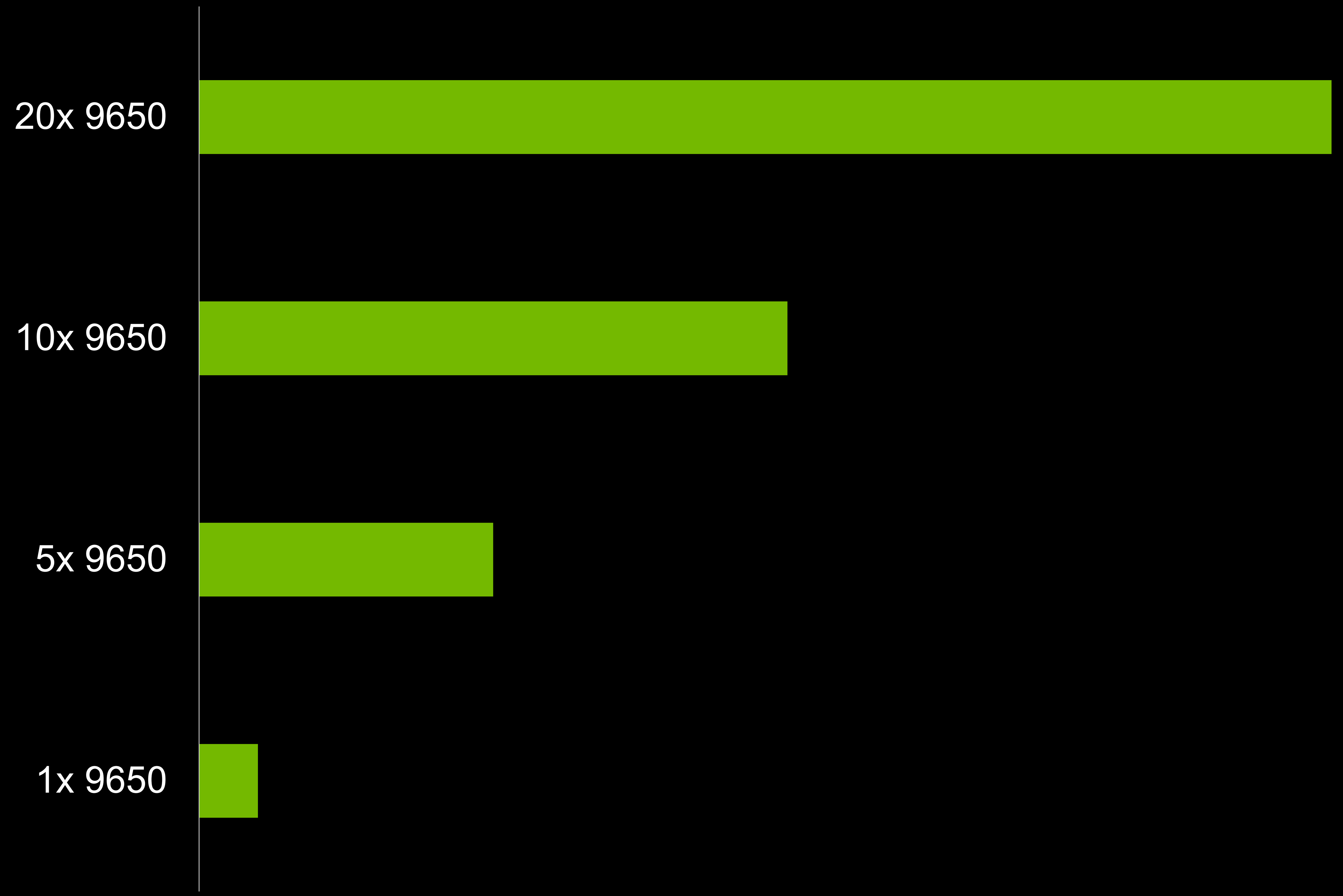

- Expected performance
  - 48 Micron Gen6 SSDs
  - 48 E1.S * 5.4 M IOPS per SSD = 260 M IOPS
  - Upto seq bandwidth – 600GBps

**Just arrived in our labs**

# Micron 9650 | NVIDIA SCADA

Early SCADA code shows strong performance and linear scaling on H3 Platform (Preliminary Results)

- **NVIDIA SCADA (Preview Software)**




**Early NVIDIA SCADA code drives impressive small block random read IOPs through 20 Micron 9650 Gen6 NVMe SSDs**

- **Linear performance scaling from 1 to 20 drives**
- H3 Platform System:
  - Intel 8568Y+, 512GB DDR5
  - 3x Broadcom 144 lane PCIe Gen6 A0 Switches
  - 20x Micron 9650 Gen6 NVMe SSD, E1.S 7.68TB
  - H100 NVL 96GB HBM3

* Preliminary Results:

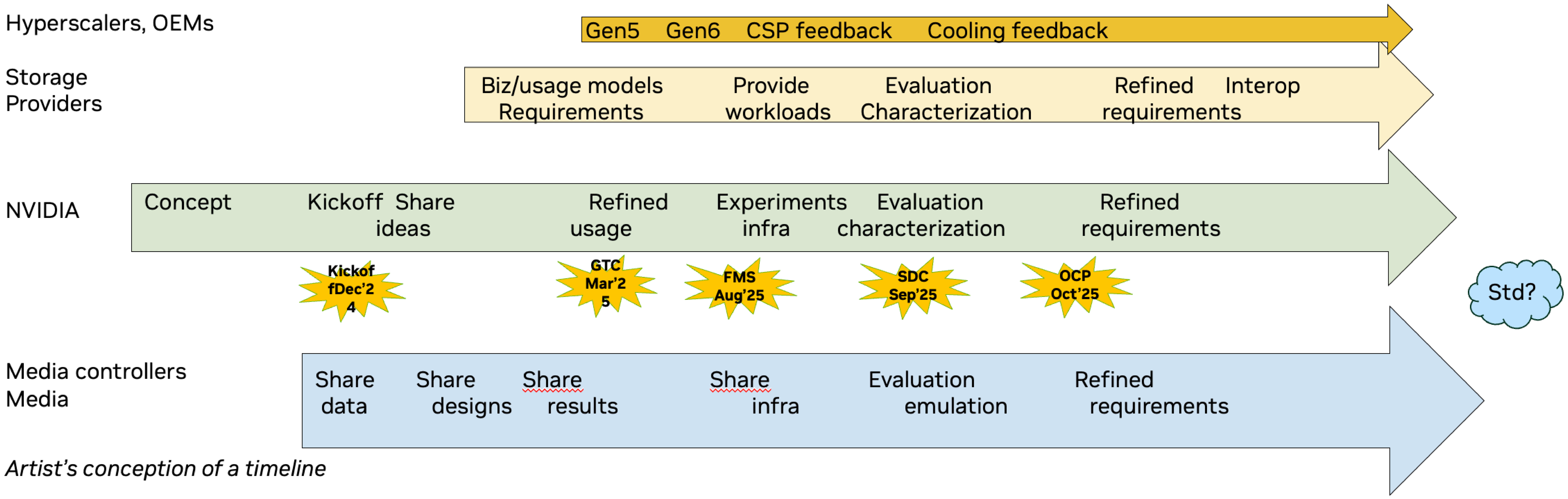
  * Hardware & software stack tuning for ongoing

SCADA test results collected on pre-production code.
System under test: H3 Platform PoC system, 1x Intel 8568Y+, 512GB DDR5, 3x Broadcom A0 PCIe Gen6 switches, 20x Micron 9650 E1.S 7.68TB, NVIDIA H100NVL-96GB PCIe Gen5x16,
Workload is 512B random read initiated from H100 GPU.
Performance testing completed by Micron's Data Center Workload Engineering team.

# Collaborators

This is Marathon and not a sprint!

- RAG, Fraud detection, ads, recsys, other apps…
- Framework providers
- CSPs, Hyperscaler and enterprise players
- OS, hypervisors, driver and library providers
- Storage Partners
- OEM/ODM partners
- SSD and DRAM media




*Artist's conception of a timeline*

FADU, Graid Technology Inc., KIOXIA, MARVELL, MICROCHIP, micron, PHISON
PLiOPS, SAMSUNG, SANDISK, ScaleFlux, SiliconMotion, SK hynix, SmartIOPS, SOLIDIGM, Western Digital
AIC, ddn, DELL Technologies, H3, Hewlett Packard Enterprise, Hitachi Vantara, IBM, NetApp, PURESTORAGE, VAST, WEKA

# We have started! Engage with us!

*Logos shown are owned by their respective companies.*

FMS NVIDIA

NVIDIA

# The path to Storage-Next

This is Marathon and not a sprint


*Artist's conception of a timeline*

FADU · Graid Technology Inc. · KIOXIA · MARVELL · MICROCHIP · micron · PHISON
PLiOPS EXTREME DATA PROCESSOR · SAMSUNG · SANDISK · ScaleFlux · SiliconMotion · SK hynix · SmartIOPS · SOLIDIGM · Western Digital
ddn · DELL Technologies · H3 · Hewlett Packard Enterprise · Hitachi Vantara · IBM · NetApp · PURESTORAGE · VAST · WEKA


*Artist's conception of a timeline*

FADU · Graid Technology Inc. · KIOXIA · MARVELL · MICROCHIP · micron · PHISON

PLiOPS · SAMSUNG · SANDISK · ScaleFlux · SiliconMotion · SK hynix · SmartIOPS · SOLIDIGM · Western Digital

AIC · ddn · DELL Technologies · H3 · Hewlett Packard Enterprise · Hitachi Vantara · IBM · NetApp · PURESTORAGE · VAST · WEKA

# Real-Time AI - Converged Data and Inference Applications

ms-level latency for fine-grained access to PB-scale data during inference