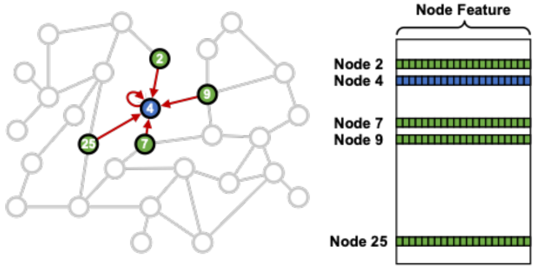
Near-Compute Storage and GPU Software Stack for Predictive AI Applications

Wen-mei Hwu

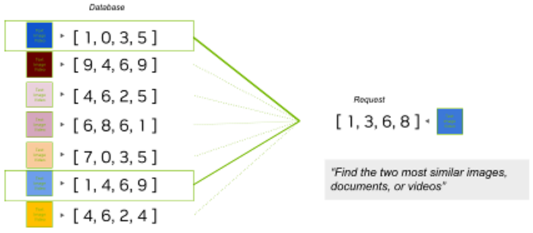
New Applications Demand Fast, Sparse Access to Massive Data

Compute-Directed Fine-grain Data Access



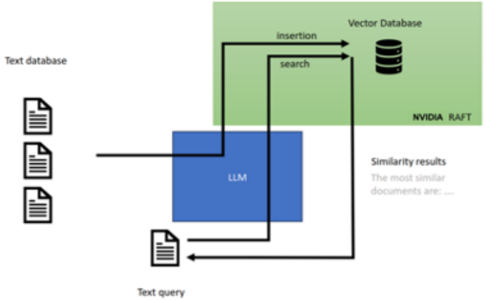
Graph Analytics and Graph Transformers (100GB-100TB) nodes/edges/embeddings

Need e.g.: AWS, Amex, PayPal, VISA, MasterCard, Block, ...



Semantic Search (up to 40PB) specialized algos on embeddings and files

Need e.g.: Google, Baidu, OpenSearch

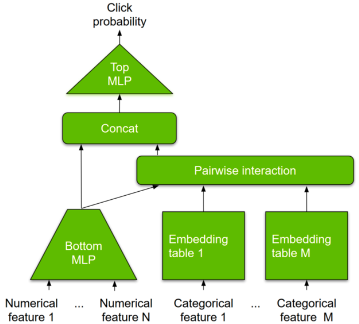


RAG/VectorDB (>600GB) ANN indexing algos on embeddings

Need e.g.: cuVS, Milvius, Pinecone

Data Analytics (100GB-1PB) select row/column based on compute

Need e.g.: RAPIDS



Recommender Systems (5-10TB) MLP and hash-table lookup on embeddings

Need e.g.: Merlin/HugeCTR HKV, Baidu

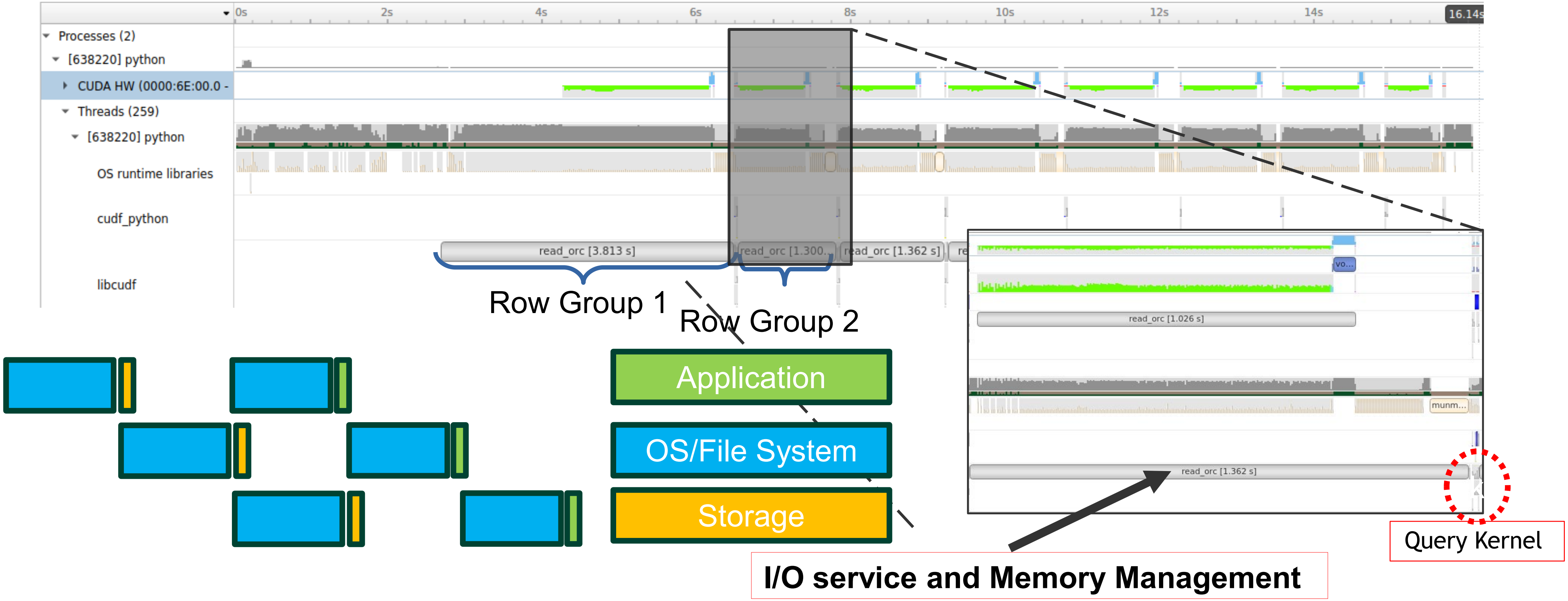
Computing on such data is currently orders of magnitude off in Cost/Throughput/Power

NVIDIA

Data Intensive Applications - Software Stack Overheads Dominate

GPU Accelerated Data Frame Analytics on New York Taxi Dataset using RAPIDS

Query: Get average cost per mile for trips that are at least 30 miles



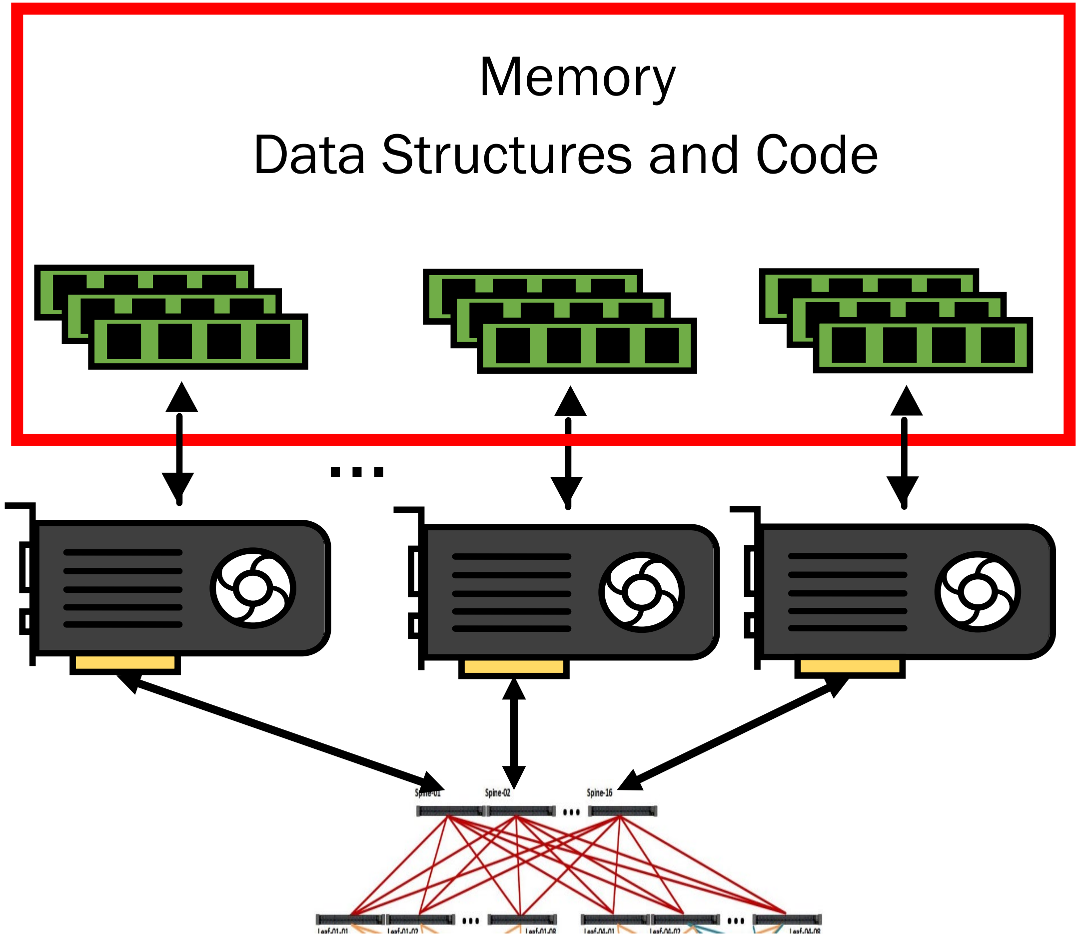
Further Acceleration of Storage Devices

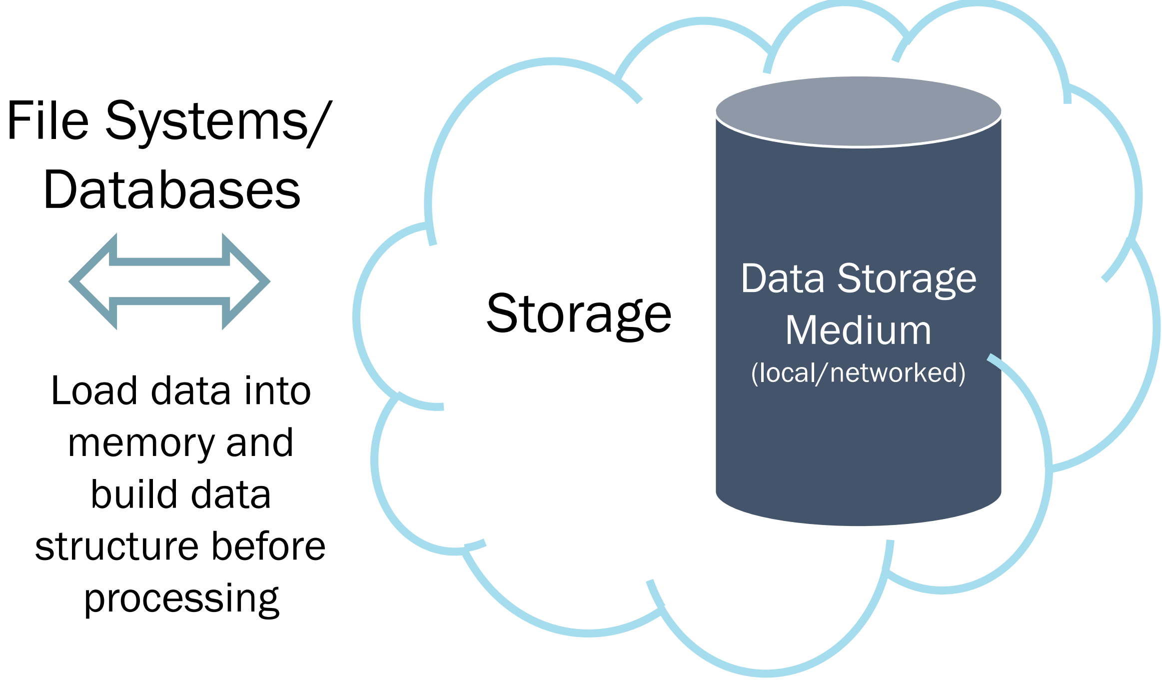
The medium access time and data transfer time will continue to decrease.

The end-to-end application time will be virtually all due to the software stack for data-intensive applications!!

NVIDIA.

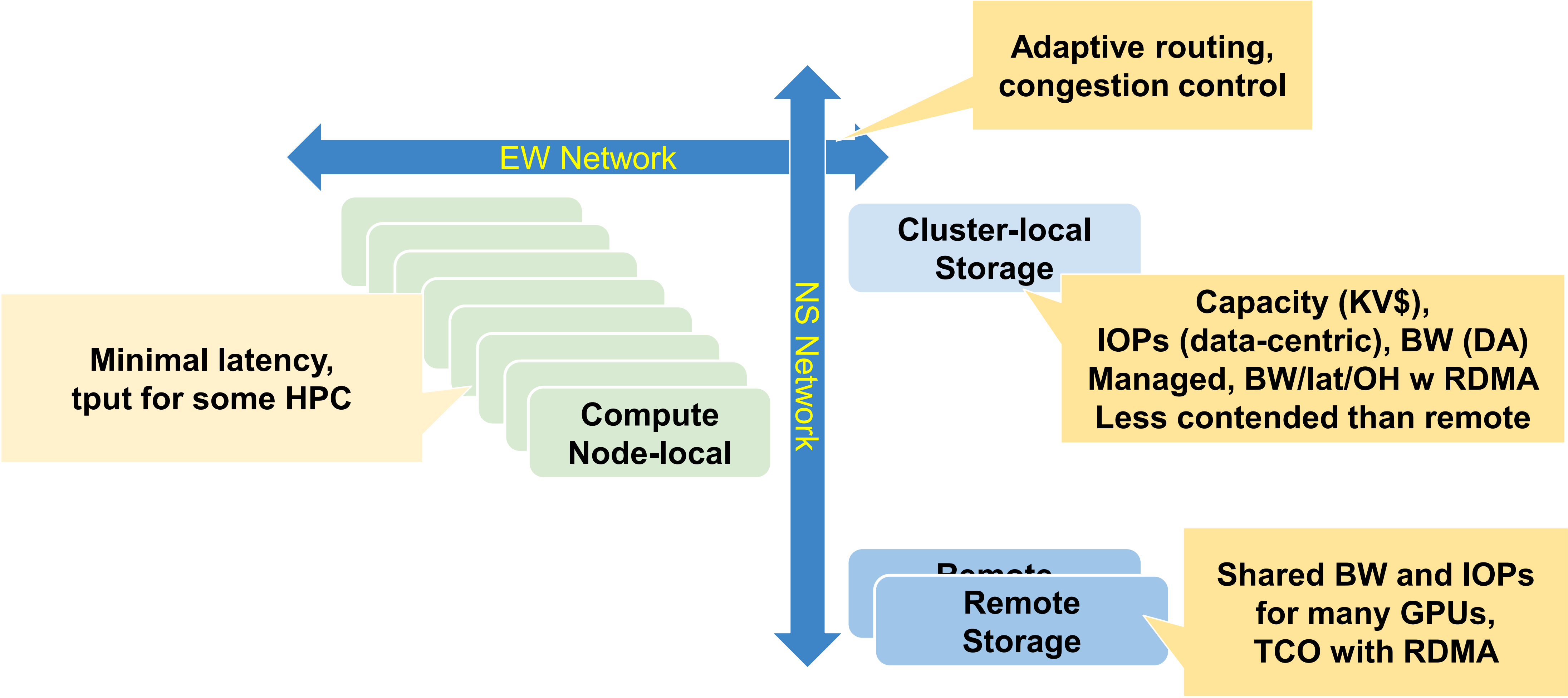
The Memory-Storage Divide



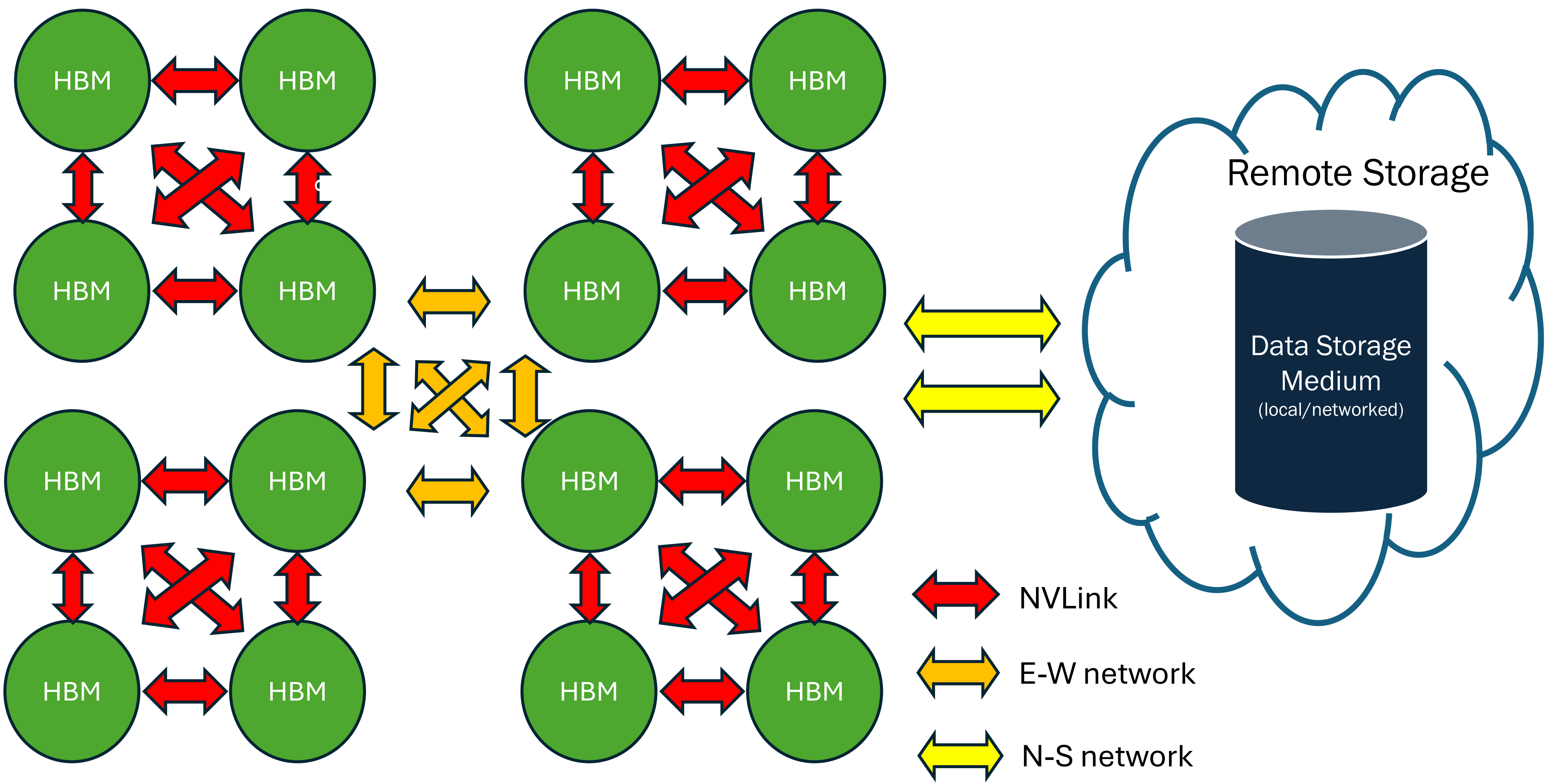


Opaque to applications except through memory-mapped files.

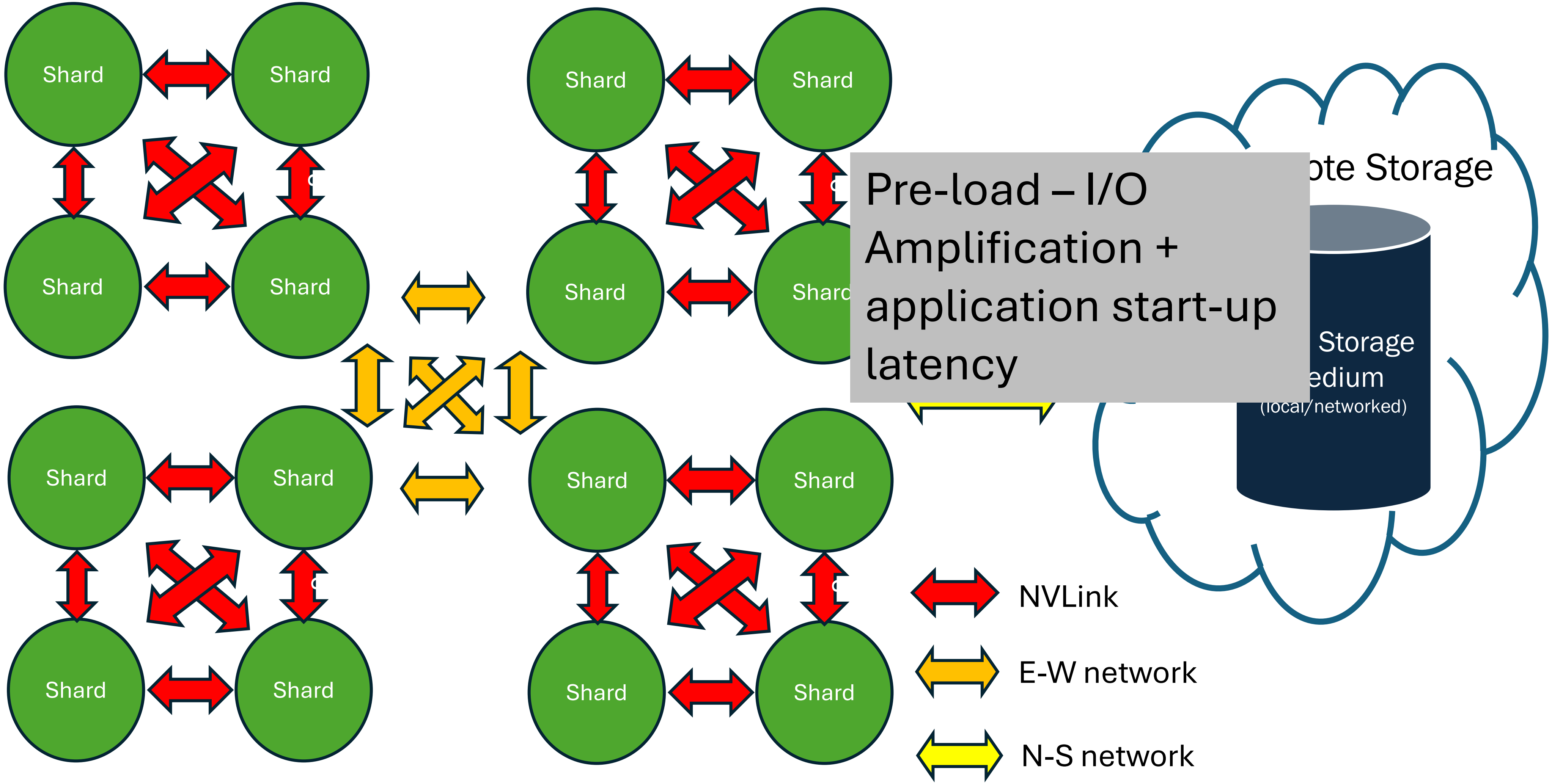
Storage in data center architecture



Old Way - Pooled HBM with Data Preload



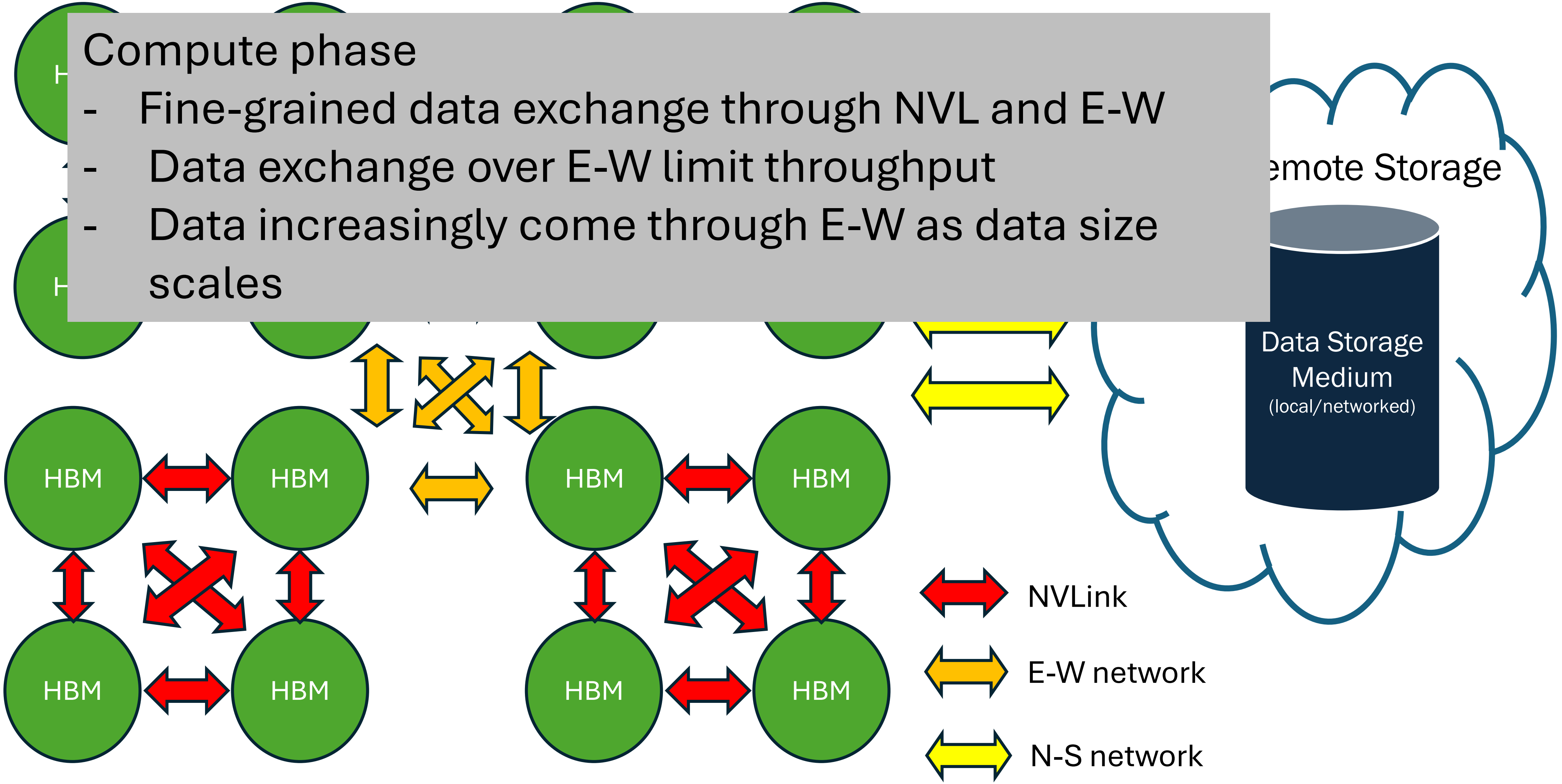
Pooled HBM with Data Preload



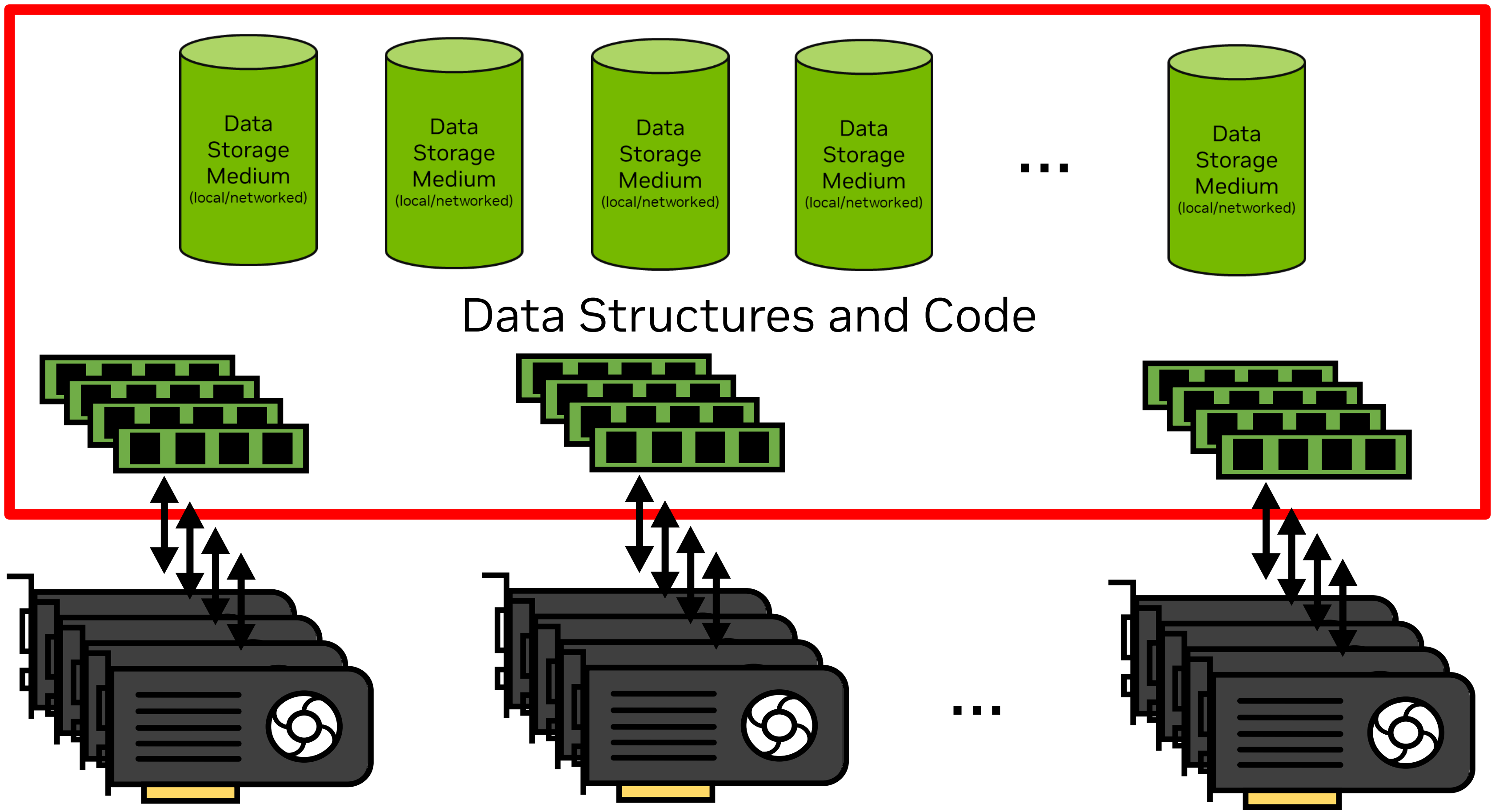
Pooled HBM with Data Preload

Compute phase

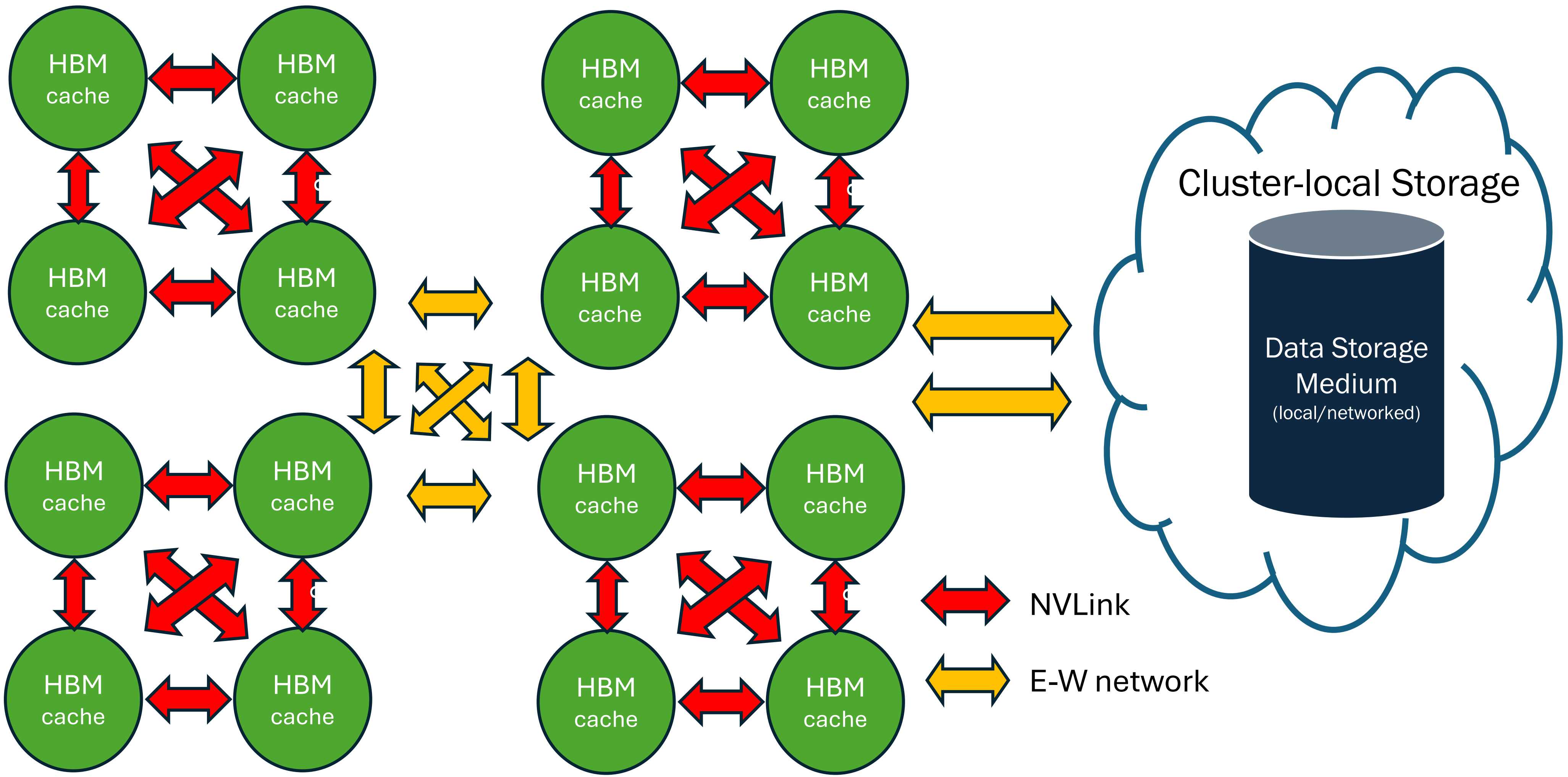
- Fine-grained data exchange through NVL and E-W
- Data exchange over E-W limit throughput
- Data increasingly come through E-W as data size scales



SCADA - SCaled Accelerated Data Access

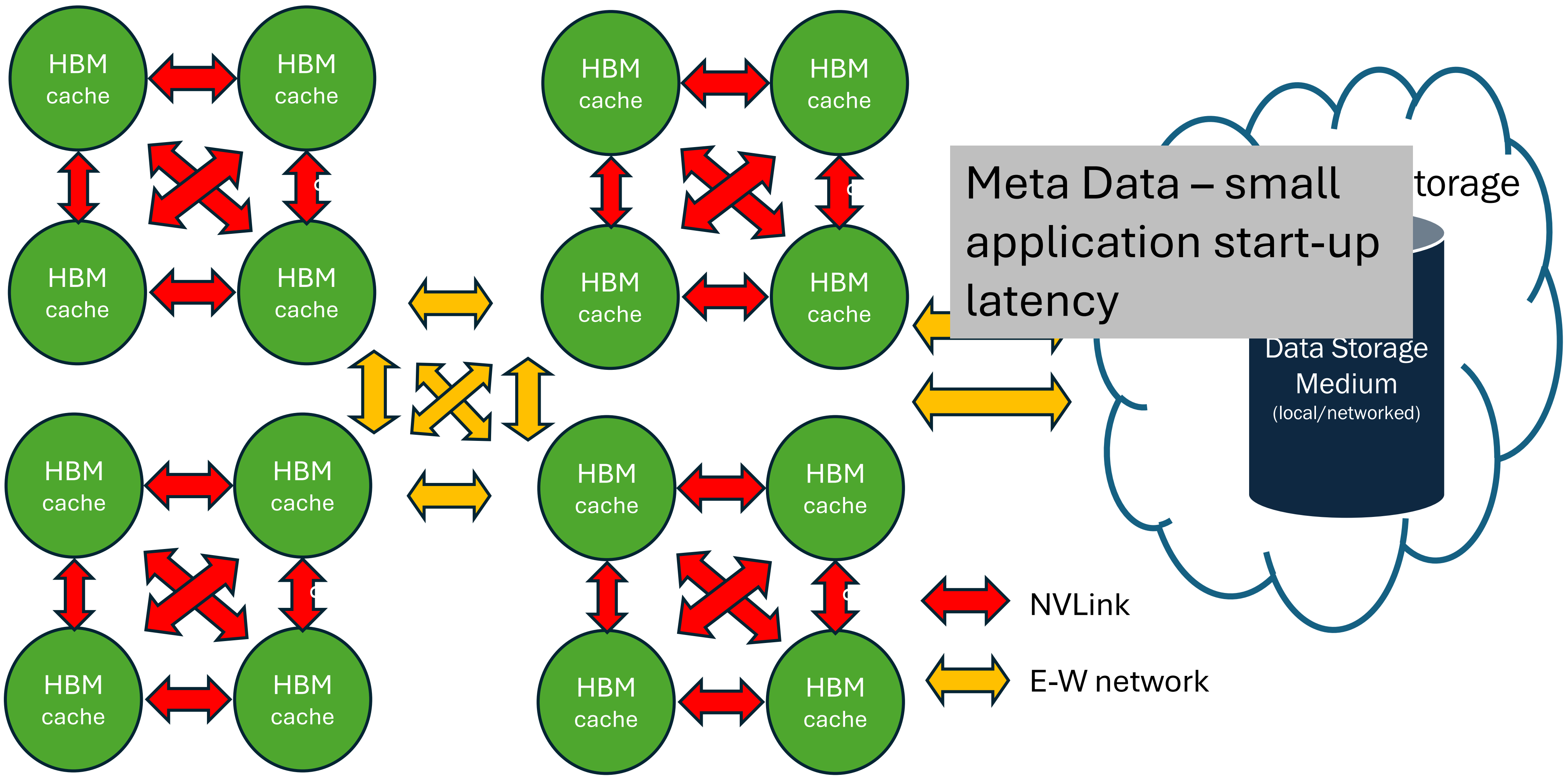


SCADA

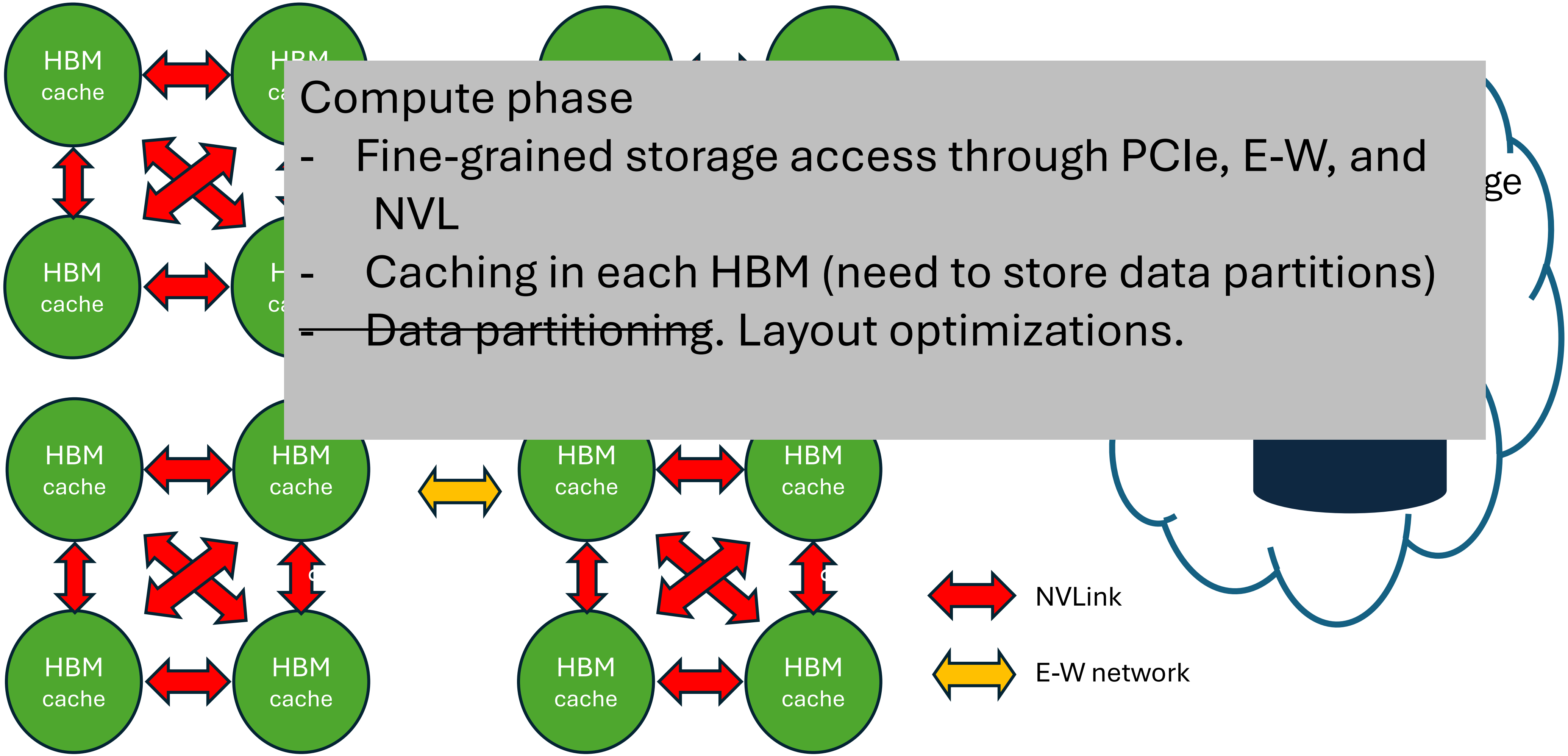


HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
Cluster-local Storage
Data Storage Medium
(local/networked)
NVLink
E-W network

SCADA


HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
HBM cache
Meta Data – small application start-up latency
torage
Data Storage Medium
(local/networked)
NVLink
E-W network

SCADA



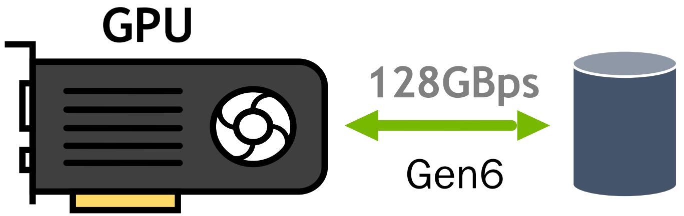
Compute phase

- Fine-grained storage access through PCIe, E-W, and NVL
- Caching in each HBM (need to store data partitions)
- ~~Data partitioning~~. Layout optimizations.

Tolerating Storage Latency with (lots of) Parallelism

Little's Law: $L = \lambda W$

Needed Queue Depth (parallelism) = Storage Throughput $\times$ Storage Latency



Raw PCIe Bandwidth (Gen 6): 128 GB/s each direction

Usable Bandwidth ~100 GB/s

Max Throughput for 512-Byte Deliveries: $\frac{100\ GB/s}{512 Byte/delivery}$ = 200M deliveries/sec (IOPS)

NAND SSD Example:

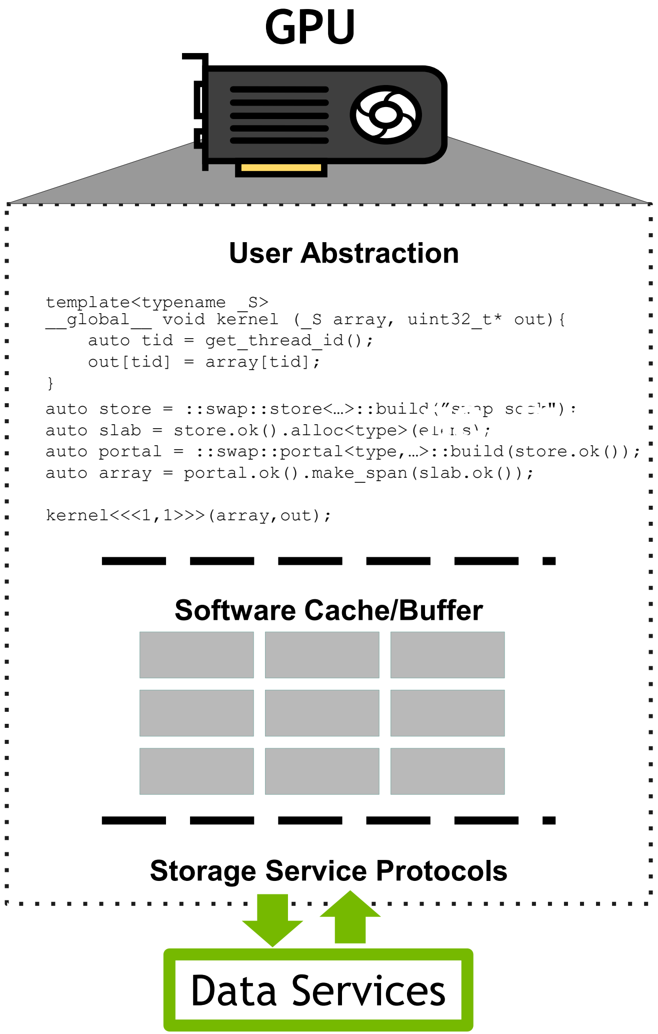
Throughput at 512-Byte: 10M deliveries/sec per SSD (requires 20 SSDs to achieve 200M deliveries/sec)

Access Latency: 300 us = 250 us (media) + 30 us (interconnect) + 20 (software)

Little's Law: 300 us × 200M deliveries/sec = 60,000 ← Mininal Parallelism to sustain over time

SCADA – Breaking the Storage-Memory Divide

CUDA threads access storage and remote data as data structure objects.



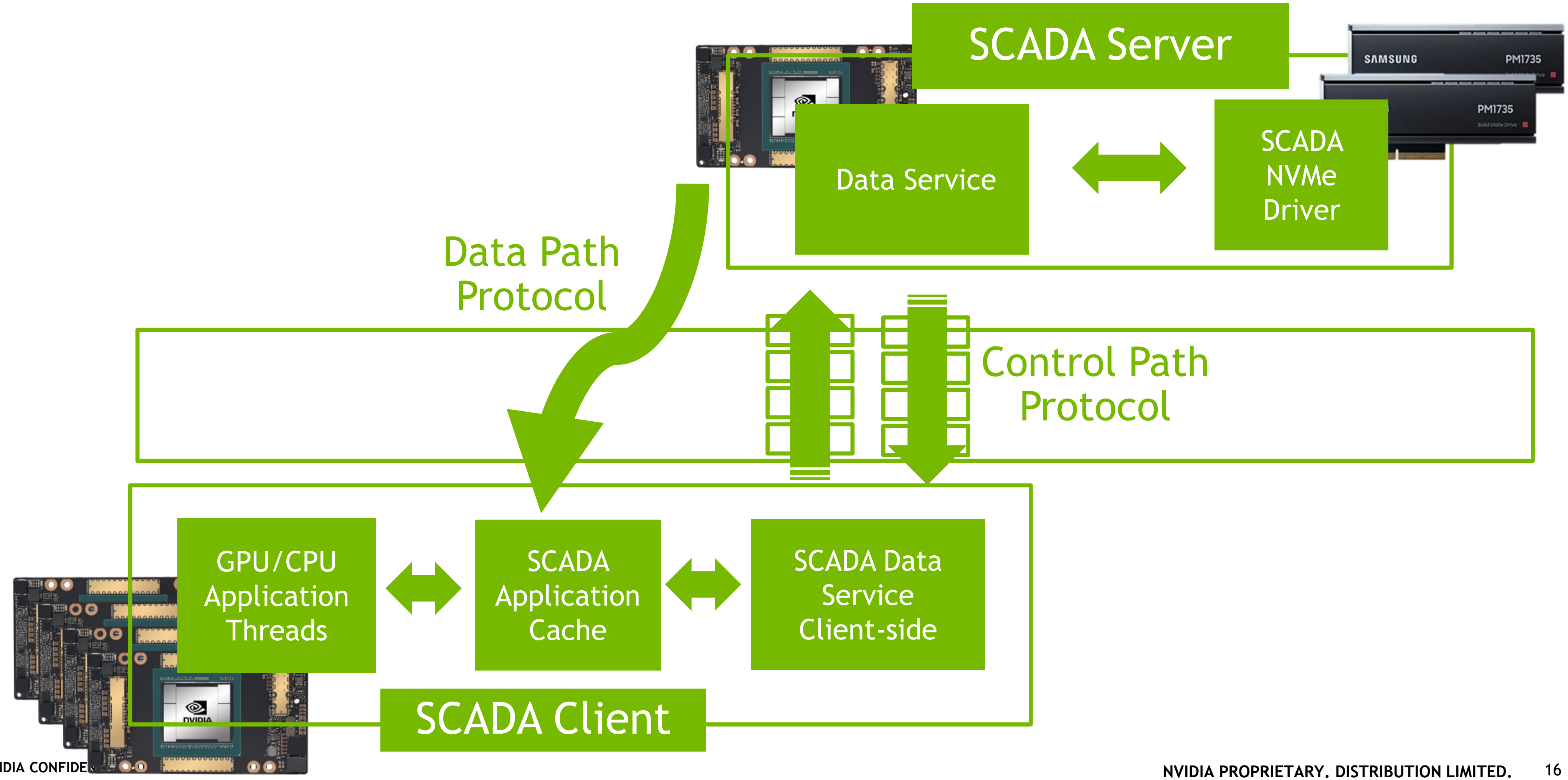
With SCADA, GPU threads can directly access data where it is, be it memory or storage!

C++ std:mdspan and KV abstractions

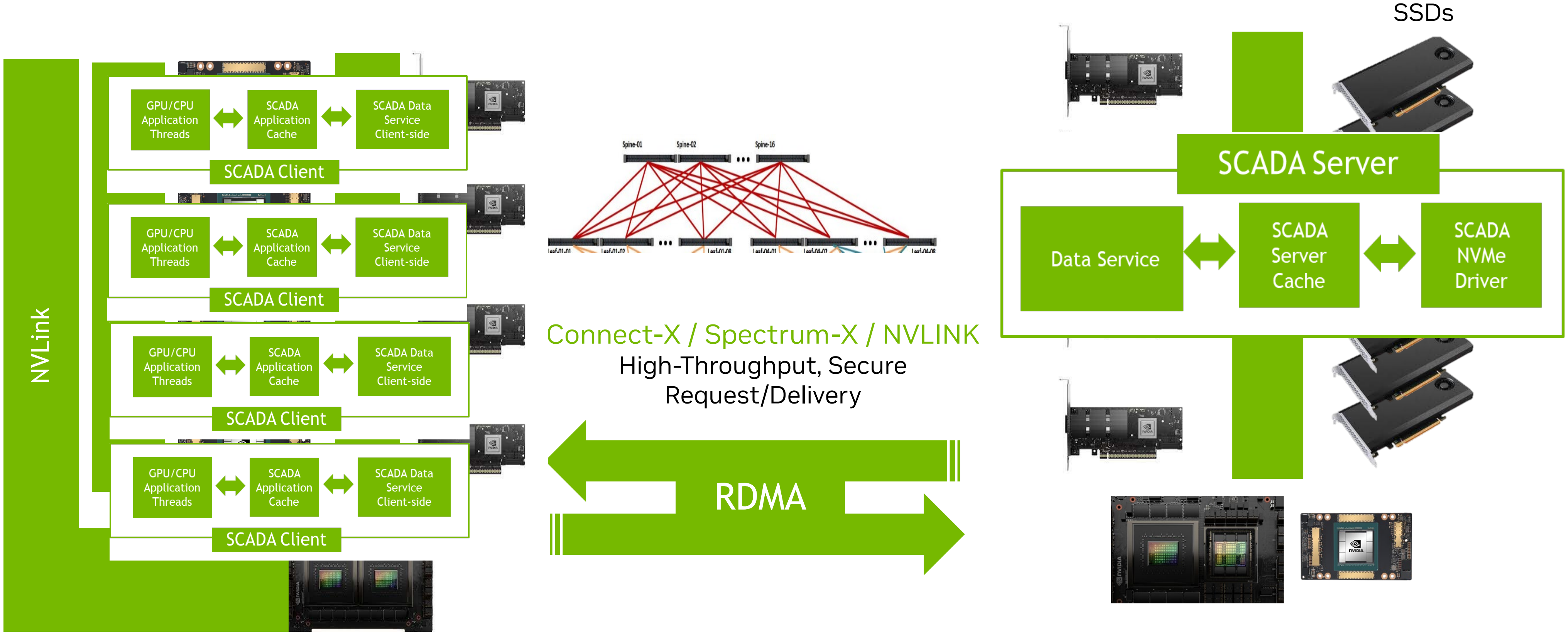
Leverage GPU memory & optimize storage bandwidth utilization

Enable GPU threads to directly access data in storage

SCADA Software Architecture and Components



GPU Accelerated Data System Architecture



Scalable Number of Accelerated Compute Nodes

Scalable Number of Accelerated Data Nodes

AI Application Overview

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

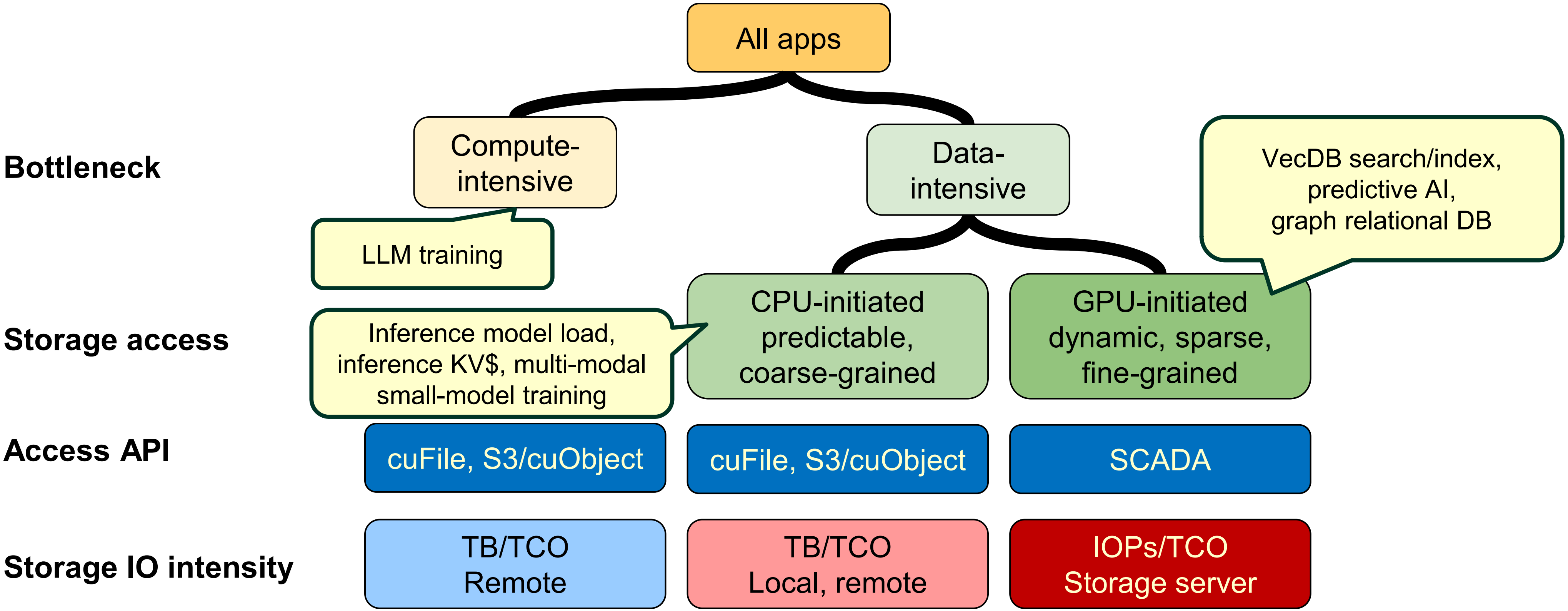
SCADA *

cuFile/S3oRDMA

Area	Usage model	Applications	Criticality @ node type		
			Compute	Storage	SKU objective
Training	Ingest	LLM pretraining, fine tuning	Low	High	TB/TCO
	Checkpoint save/restore		Low	High	TB/TCO
Inference	KV context caching across queries, docs	LLM inference	Usually low	High	TB/TCO
	LLM+GNN, GNN+LLM	Contextual LLMs	High	High	IOPS/TCO
	Vector database	Dynamic Index build	High	High	IOPS/TCO
		LLM RAG doc retrieval	Low	High	TB/TCO
Predictive AI		Graph RAG	Low	High	IOPS/TCO
		Recommenders	High	High	IOPS/TCO
	GNN sampling induced subgraphs	eCommerce, fraud, social networks	High	High	IOPS/TCO
	Anomaly detection	eCommerce, fraud, social networks	High	High	IOPS/TCO
	Small World Graphs	Vector Search Index	High	High	IOPS/TCO
	Relational graphs	Data Science Automation	High	High	IOPS/TCO

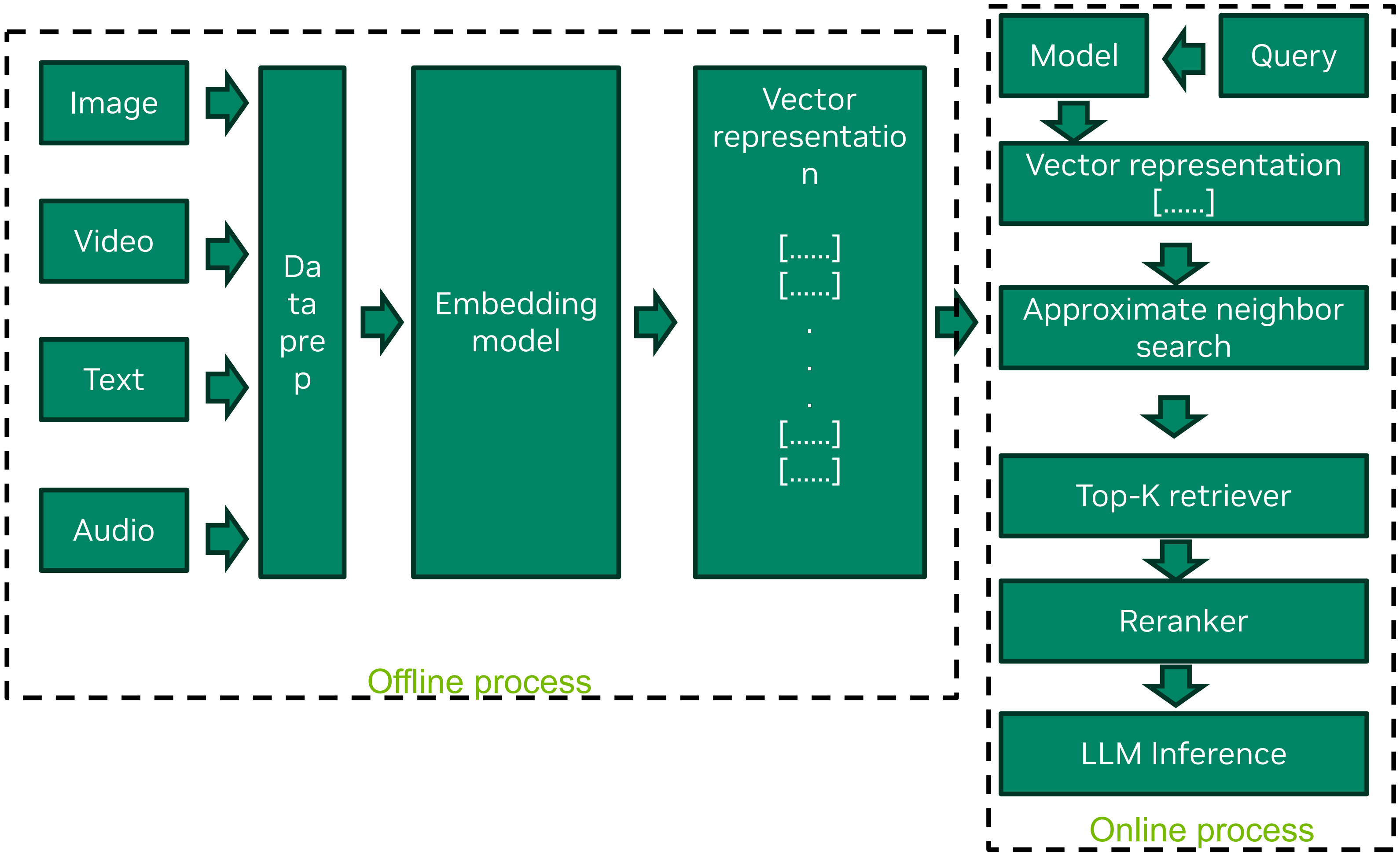
App Taxonomy: intensity, bottlenecks and APIs

Designing storage solutions hinges on understanding app requirements



- Traditional focus for GPUs has been compute-intensive apps like Gen AI
- The explosion of innovation for VecDB, predictive AI drives new technologies to fill the gaps

Vector (semantic) search pipeline for RAG

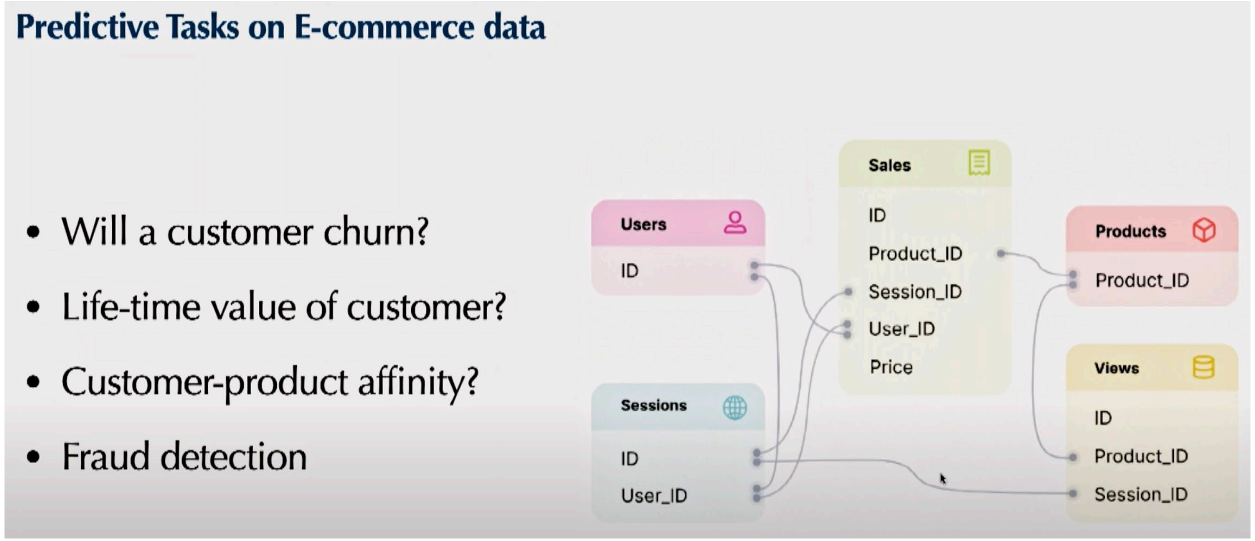


Offline process

Graph Relational DB

Predictive AI using relationship information is gaining popularity

- Raw Data: SQL tables w/ keys linking them
- Node Data: Row in a table w/ time stamp
  - Each table: users, sessions, products, views, sales, ...
  - Each column is a feature
  - Each feature:
    - integers, floating points, or text/image embeddings
- Edge Data: Unique ID's linking nodes

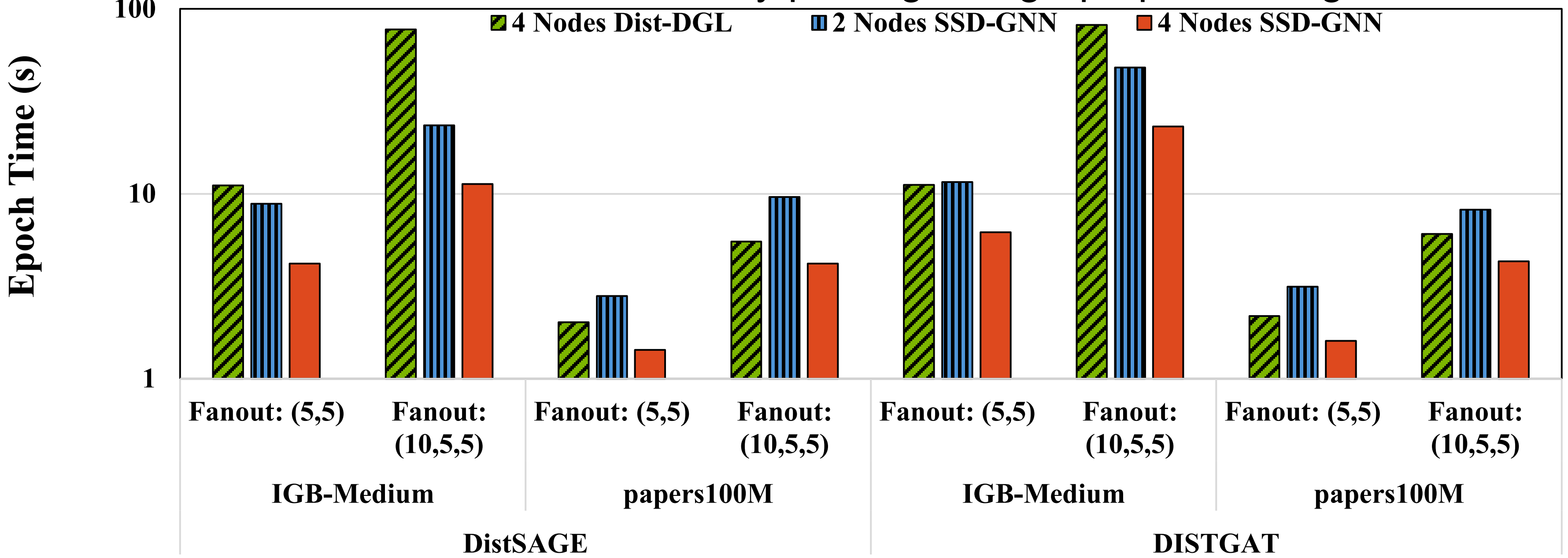


Heterogeneous graph – massive in size for both structure and embeddings

https://www.youtube.com/watch?v=s_xGhXX_F3c

Application Performance Example – GNN Training

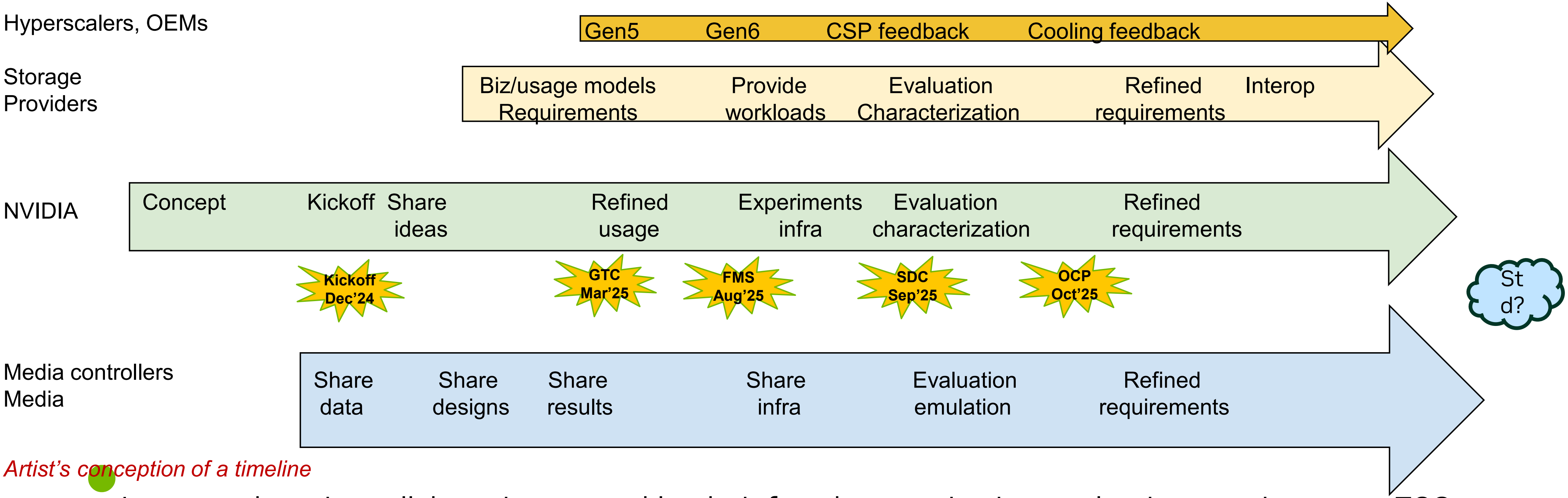
Dist-DGL uses memory pooling and graph partitioning.



SSD-GNN is faster with fewer GPUs for large graphs

The path to Storage-Next

IOPs and TB with improved TCO from BOM, volume, power improvements



Artist's conception of a timeline

- Interested parties collaborating on workloads, infra, characterization, evaluation, requirements, TCO
- Open to all direct contributors: media vendors, storage controller vendors, OEMs, hyperscalers

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