Opening pathways to storage-driven insight

CJ Newburn, Distinguished Engineer, Data Center and IO Architect | IBM Scale User's Group, ISC June '25

Overview

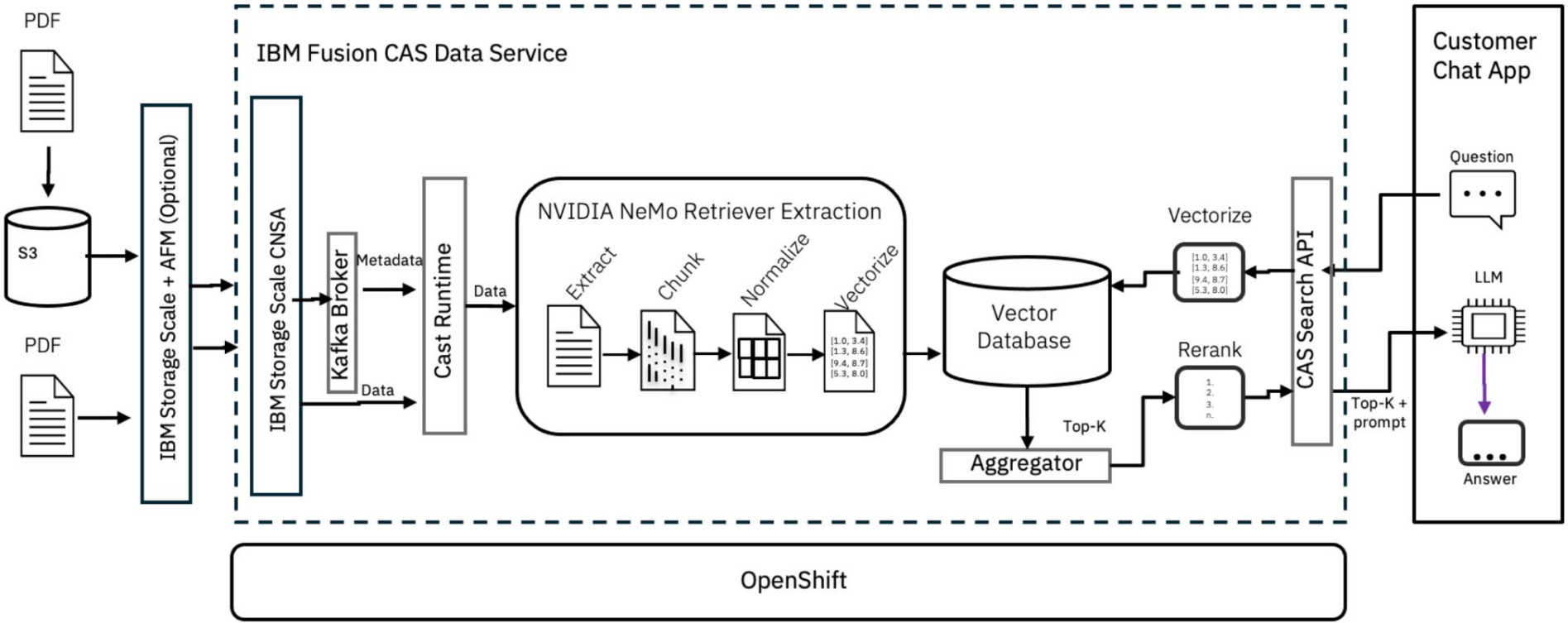
- Partnerships
- Usage models
- Architecture
- Technologies

Initiatives and partnerships

- NVIDIA AI Data Platform
  - NVIDIA is introducing a reference design for building intelligent AI Data Platforms to usher in a new generation of enterprise storage
  - Integrated storage solution to accelerate demanding inference agentic and reasoning workloads
  - SW components: NVIDIA AI Enterprise: NVIDIA blueprints with AI-Q, NIMs, NeMo Retriever; coming soon: NVIDIA Dynamo, NIXL
  - HW components: GPUs, Spectrum-X for adaptive routing and congestion control, and BlueField DPUs for lighter storage solutions
- Storage-Next, led by NVIDIA
  - NDA partnership, closing the gap for fine-grained GPU-initiated storage and NVMes with better IOPs/$, TCO (IOPs, space, power)
- Storage certifications

Content-aware storage platform

Opening a pathway to storage-driven insight



NVIDIA storage certification

Ensuring that AI factories are built on a foundation of high-performance, reliable storage solutions

NVIDIA DGX Systems

Storage certification for NVIDIA DGX for enterprise and cloud AI deployments

Certifications: DGX BasePOD and DGX SuperPOD

Scale: 10's-1000 GPUs

NVIDIA-Certified Storage

Storage certification for NVIDIA-Certified servers for enterprise AI and HPC deployments

Certifications: Foundation and Enterprise

Scale: 10's-1000 GPUs

NVIDIA Cloud Partner Storage

Designed for cloud service providers that want to offer AI and HPC services to their customers

Pre-requisite: NVIDIA-Certified Enterprise or NVIDIA DGX SuperPOD certification

Scale: 1000-10,000 GPUs

NVIDIA-Certified and NCP Storage Certification Programs for AI factories

- Perf, security, RAS, multitenancy, QoS, data services, and scalability
- File now, S3 future; consistent IO APIs across NCPs
- Enterprise use cases from HPC, big data analytics, classical ML/DL, gen AI fine tuning and inference, and agentic reasoning

The NVIDIA Cloud Partner (NCP)

- Reference architecture: blueprint for high-performance, scalable and secure data centers for generative AI
- Data architecture is based on data lake and HPS tiers
- Today: CSP data lake with S3 or NFS, HPS with certified file

Credits: John Fragala, NCP Storage Lead and
Satheesh Iyer/Matthew Hausmann for Enterprise Storage Certification

Usage models

- GenAI
  - LLM inference and training
  - Model loading
  - Vector search, vector database indexing
  - GNNs, GNN+LLM
  - Relational Graphs
- HPC
- Data analytics
  - ETL
  - Visualization

Usage models

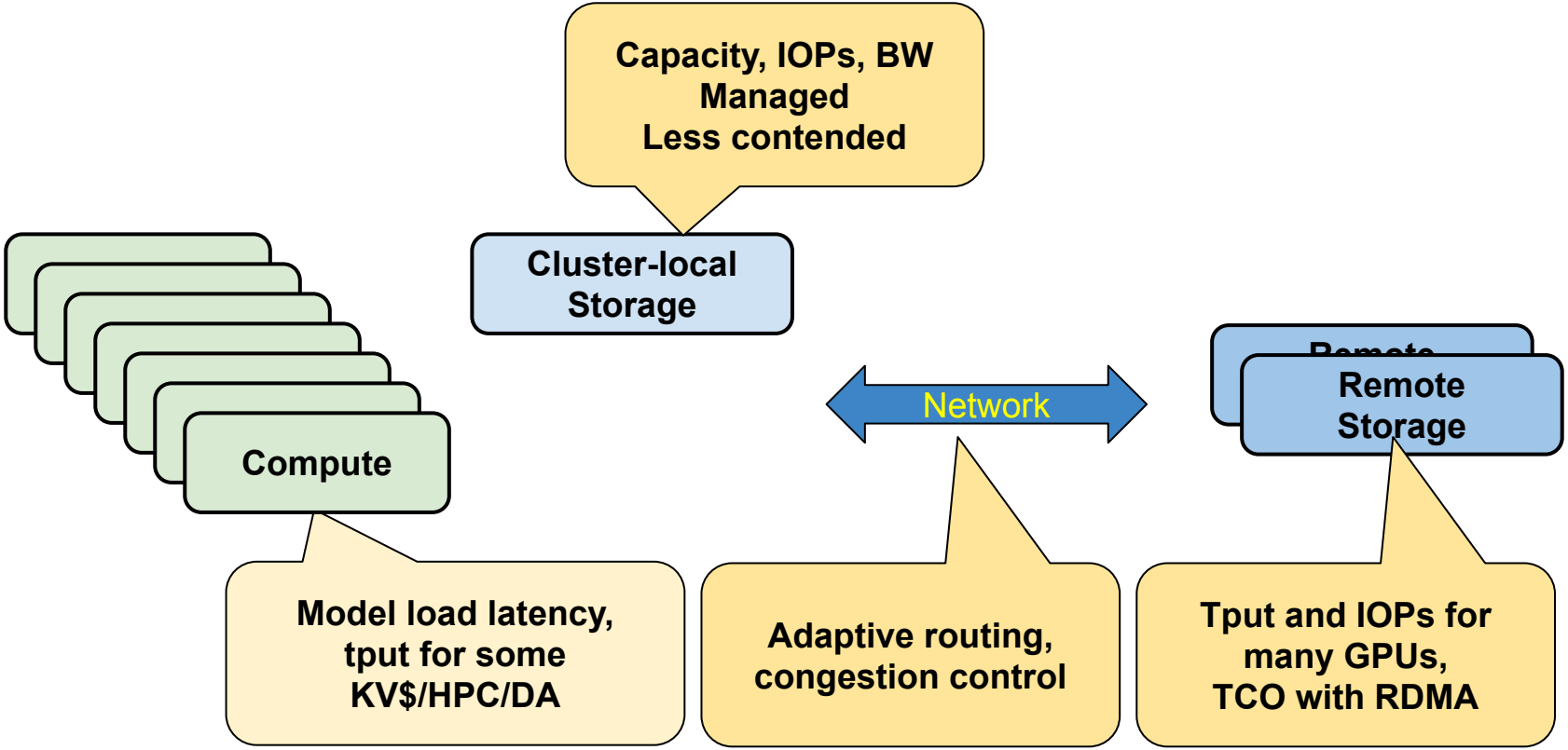
Area	Usage model	Figures of merit	Conditions for storage criticality
Inference	Summarization	Ease, TTFT, E2E, tput	Perf: Large ISL
	Multi-turn queries	Ease, TTFT, E2E, tput	Perf: Large ISL, small OSL
	Multi-step agentic	Ease, E2E, tput	Perf: Large ISL, small OSL
	Model loading	TTFT	More model space than can fit in memory
	(Ingest)		Storage not relevant
Training	Checkpoints	Resource cost, tput	Perf: legacy SW
	Ingest	Resource cost	Choice: object data not available as file
GNNs, relational graphs	Fraud, social networks, DB	Sparse fine-grained IOPs	Problem sizes don’t fit in memory
VecDB index, srch		Sparse fine-grained IOPs	Problem sizes don’t fit in memory
HPC	Healthcare Seismic, RTM	Resource cost, tput	Perf: doesn’t fit in memory
Data analytics	ETL Visualization	Throughput	Larger problems, not compute intensive

Figures of merit/metrics

- Non-performance
  - Ease of use
    - KV state referenced by key'd lookup
    - KV$ manager has lower complexity for object vs. file
  - Resource costs - TCO for provider
  - Customer choice - more options that align with customer or partner preference
- Performance
  - TTFT - perception of responsiveness for user; impact threshold is 100ms
  - E2E - latency of response for user, SLA for provider
  - Tput - efficiency for provider; impact threshold is 2x if enabling required

Storage in data center architecture

Different concerns and opportunities for each part of the data center storage architecture

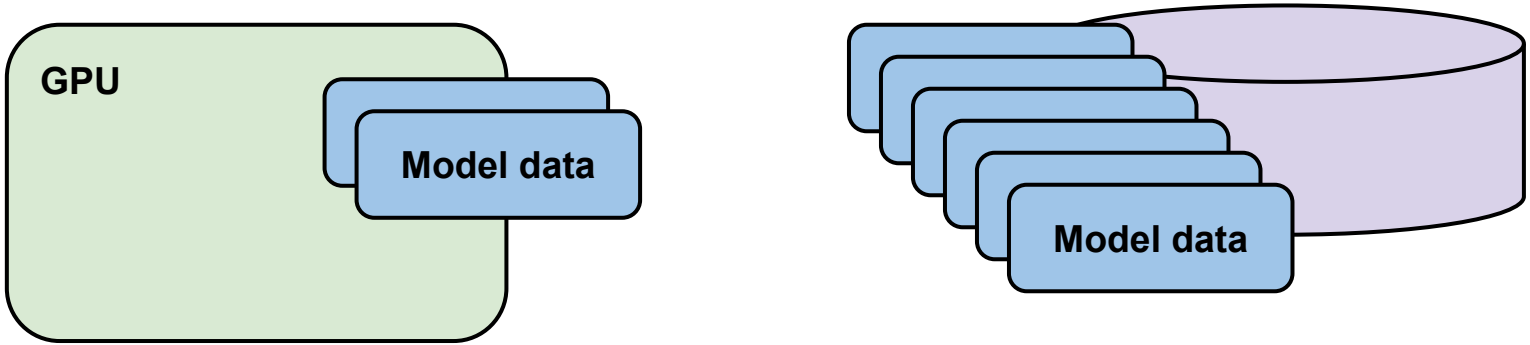


Main performance takeaways

- Performance at compute server/network/storage server differ
- Understand and track shifting usage models' requirements
- Performance impact of acceleration may vary greatly
- There are cases where performance can really matter
- Bring the best tech to the table so it's available when it does matter

Inference model loading

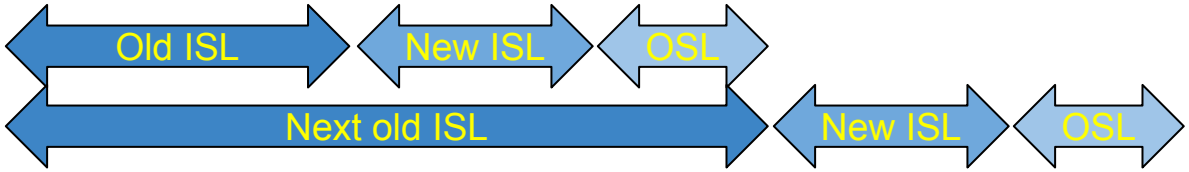
Key driving usage model for storage IO performance in GenAI



- Model load time directly affects TTFT
- Models' weights occupy ~2GB per billion parameters
  - Examples: 7*2=14GB, 405*2=810GB → not many fit in GPU
- Mixture of experts may involve a couple of defaults but many others
- May not be able to fit all models in memory → spill to nearby disk

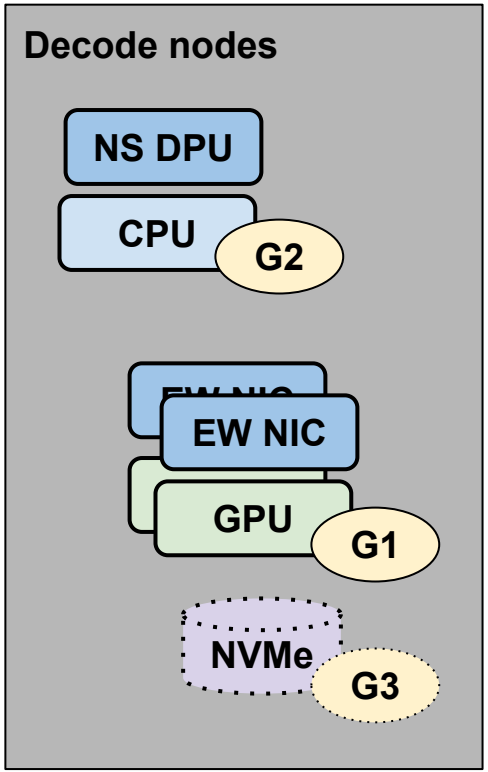
KV caching synopsis

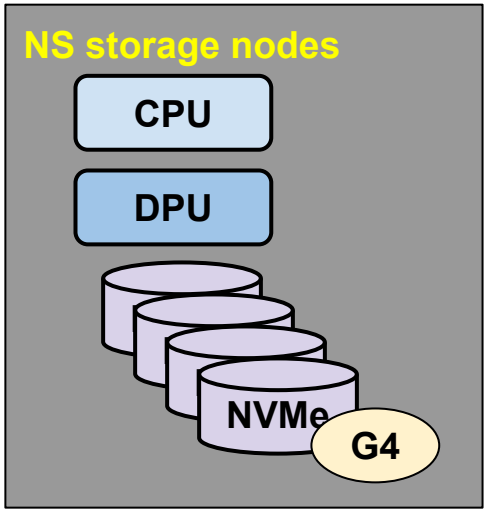
Key play for available capacity, and in some cases, bandwidth and latency

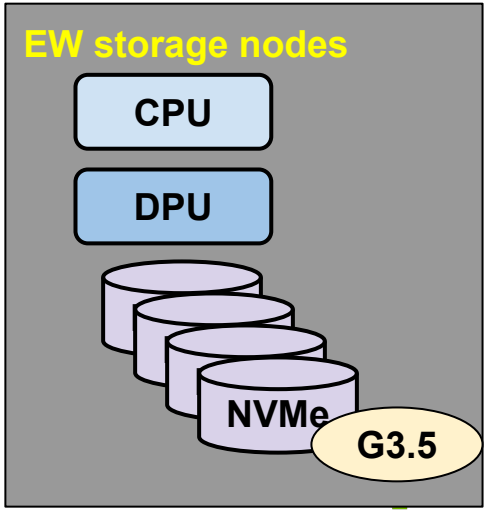


- Each query includes both 1) all previous queries, 2) new turn's query
- This makes up the ISL = input sequence length
- ISLs often have a common prefix that can be cached
- For larger ISLs, it's faster to reload from storage than recompute
- Many turns or steps that may be spaced out → huge caching context

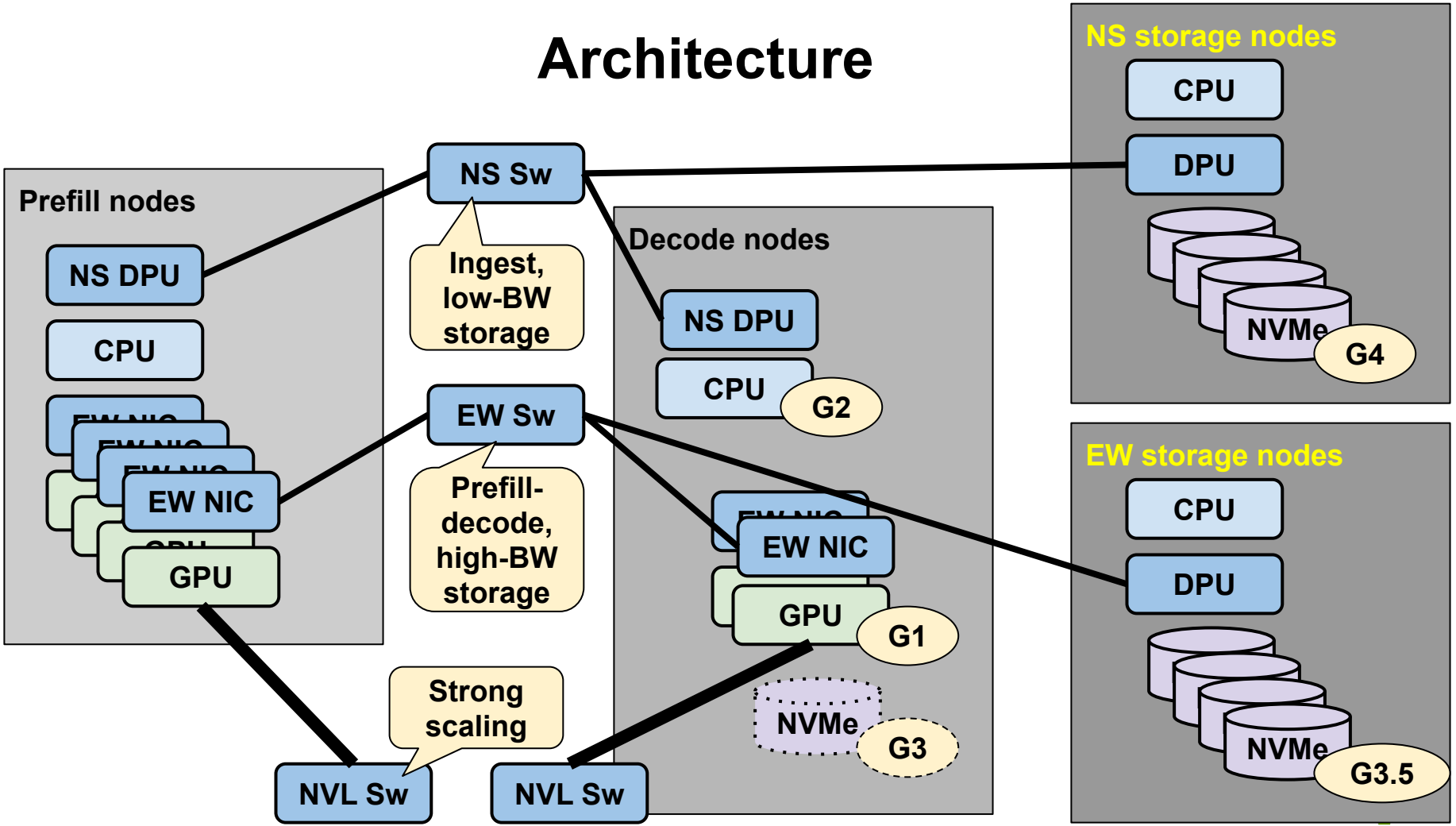
Architecture



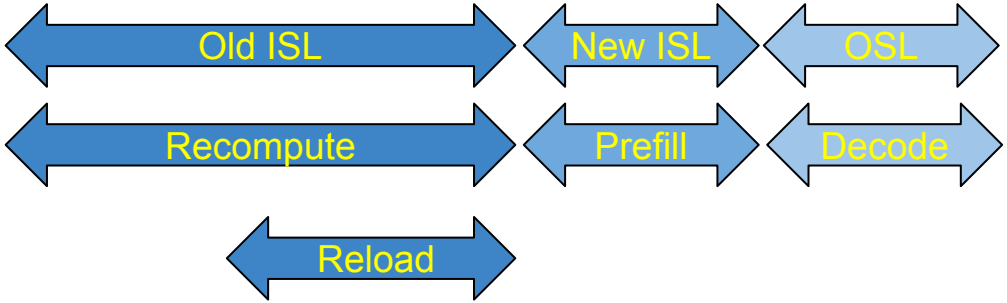




Architecture



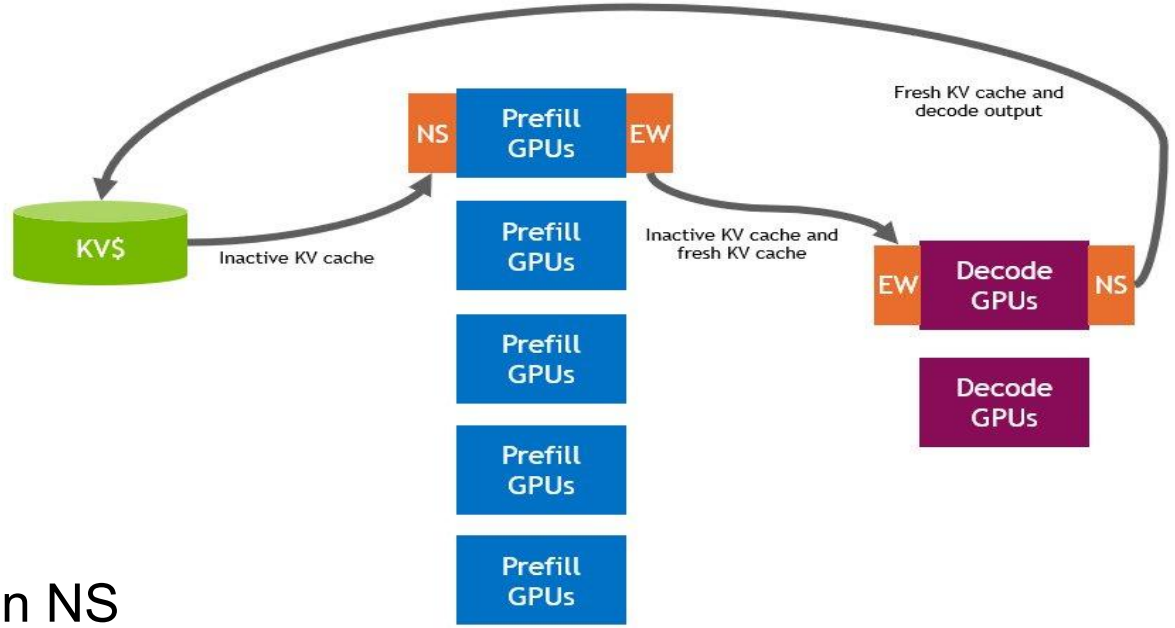
When does storage performance matter for KV?



- Significance
  - For TTFT, (load or recompute time) > (human response time of 200ms)
  - For E2E, fraction F is a sizable % of recompute+load+decode time
    - F = (load time)/((recompute time)+(load time)+(decode time))
- KV$ used when loading becomes faster than recomputing
  - (ISL)(compute/token) > (KV context size)/(load bandwidth)
  - True for relatively long ISLs

KV cache management

NS storage IO bandwidth scaled by # prefill GPUs



- Spill from decode on NS
- If disaggregated prefill and decode and not using G3.5 on EW
- Fill into more prefill GPUs on NS, EW is not the bottleneck
  - Prefill, 2x GPUs, NS NICs: GPU:NIC=2, 100-200 Gbps
  - Decode, EW NICs: GPUs:NICs=1, 400-800Gbps

What affects storage performance significance for KV?

- For E2E, fraction F is a sizable % of recompute+load+decode time
  - F = (load time)/((recompute time)+(load time)+(decode time))
- Storage IO performance matters less for
  - Long mandatory recompute: fraction of new ISL is high
  - Small state: short old ISL, low state/token
  - Fast state reload: fast network, many GPU load targets
  - Insignificant wrt decode: large OSL, slow decode, no speculative decode
- Trends
  - FP8 moving to FP4, reducing state
  - OSL: 1 page is 250 words, 300 tokens, limits dilution from decode time
  - ISLs: growing, inceases state
  - Network: 6.25 GB/s for HGX B200 - 25 GB/s for GB200/300, scaled by GPUs
  - Decode: faster; but there are efficiency trade-offs

What-if analysis shows IO acceleration not 1st order for KV

Storage IO acceleration is applied to the fraction of load time to TTFT or E2E time, which is **never very big**

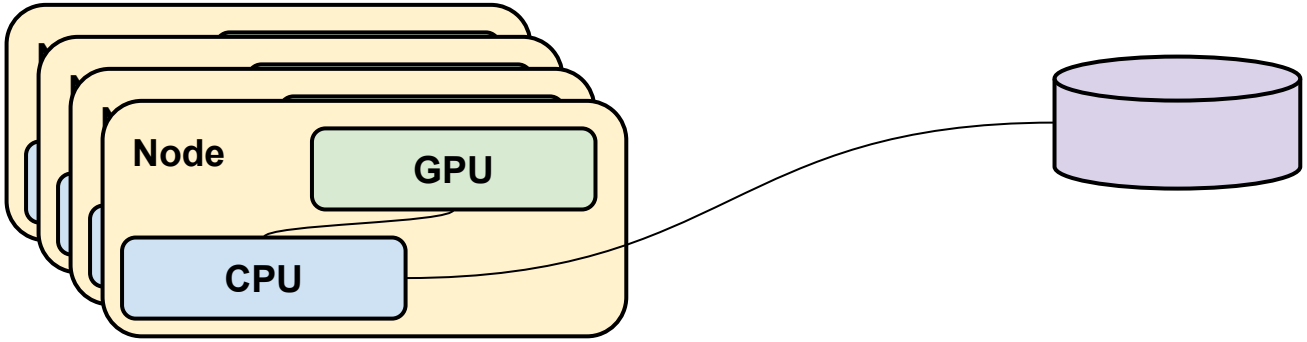
IO % of decode	IO % total	ISL	Model	KB/ token	State size	Network	# PF GPUs	Decode/ token	OSL
0%	0%	64K	DSv3 R1 FP4	17	1GB	12.5 GB/s	32	30ms	1000
0.7%	0.1%	64K	DSv3 R1 FP4	17	1GB	12.5 GB/s	**4**	30ms	**300**
0.1%	0.1%	64K	**Llama 405B FP4**	123	**8GB**	12.5 GB/s	**64**	30ms	300
0.7%	0.6%	64K	Llama 405B FP4	123	8GB	12.5 GB/s	**32**	**10ms**	300
9.5%	8.8%	128K	Llama 405B FP4	123	8GB	**6.25** GB/s	**8**	10ms	300

- Careful analysis and debate still underway
- Scenarios are hypothetical/what-if to illustrate spanning a range of parameters
- % impact: (load time)/((optional prefill time)+(load time)+(decode time))
- Impact largest with fewer GPUs, larger ISL, shorter OSL, slower NW
- 300 tokens ~250 words for 1 output page

Assume prefill time of 250ms for new ISL of 2K tokens on 32 prefill GPUs

Checkpoint save/restore

Storage IO performance matters on storage server side
On compute side, may matter only for recovery at small scale, some under-tuned systems/infra



- Optimal arrangements take save time off of the critical path
  - Fast transfer from GPU to CPU; async from CPU to storage
- Save/restore time negligible at scale
  - 8TB state for Llama 405B, spead across GPUs in system
  - <100MB/GPU → 8 ms
- Save % negligible at small scale
  - Failure/GPU/s 5.61E-09, from MLPerf
  - MMTF for system: (1-5.61E-09)^(# GPUs)s
  - GPUs: 206 days @10, 2.1 days @ 1000, 179s @ 100K, 18s @ 1M

GNNs

Enable larger problems, scale linearly with IOPs

- Problems solved and figures of merit
  - Fraud ($44B in ‘23, $172B in ‘31, $250B by 2032)
    - accuracy as f(capacity, recency for dynamic data): length of history, # of customers
  - Social networks
    - accuracy as f(capacity): whole graph accuracy vs. segmented for relationships; perf from click-through throughput
- Sampling for extraction of induced subgraphs for training
  - Start with a subset of 100000s of sampled node, visit 5-10 neighbors in 3 iterations based on node embeddings
  - High concurrency → O(100K) GPU threads, each initiating its own access to data of unbounded size (1B → 1T edges)
  - Fine granularity: with compression, embeddings are often 512B

  → In Gen6, 50 GB/s / 512B = can be 100 MIOPs per GPU

GPU-Accelerated Vector Databases

NVIDIA enables massive scale with accelerated extraction, retrieval, and vector search

Flexible and Scalable

Integrations
Available in major vector databases and libraries, including FAISS, Milvus, Solr, Elastic

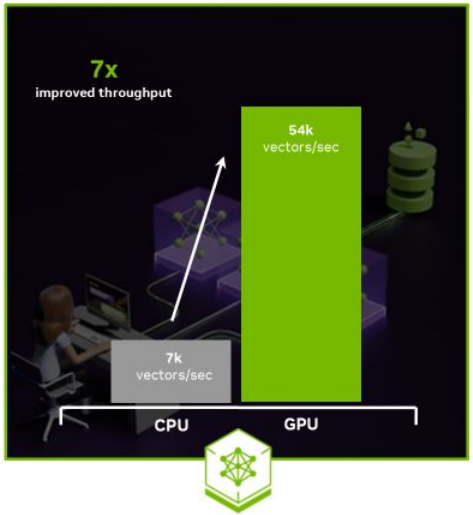
Advanced Algorithms
GPU Indexing optimized for high accuracy and low latency at all batch sizes

Interoperable
interoperable between CPU and GPU enabling index building on a GPU and searching on a CPU

Scalable
Enables high-volume vector indexing and search

Flexible Integration
Supports multiple languages including C, C++, Python, and Rust, for easy integration into vectorized data applications

Higher Index Build Throughput



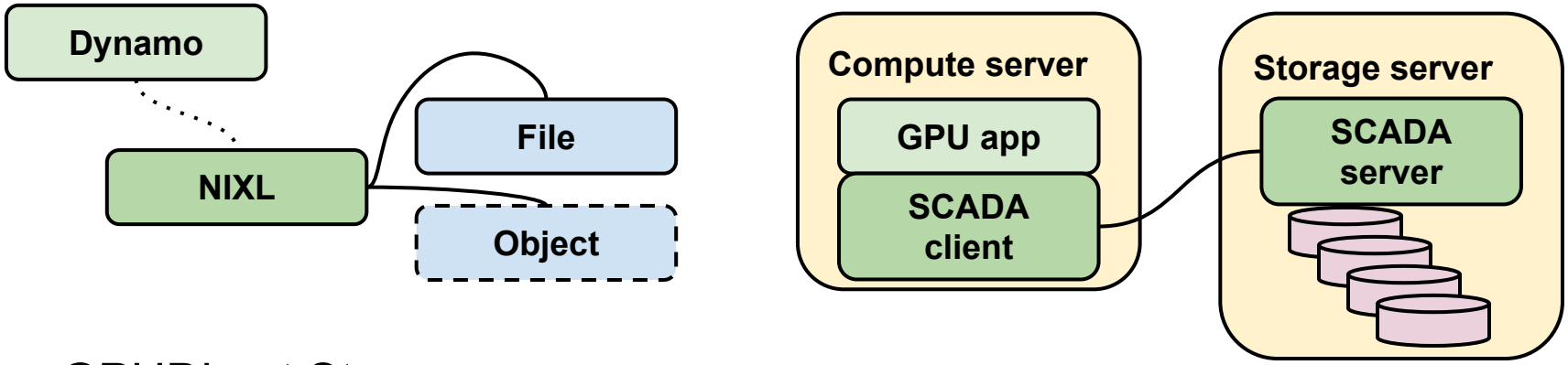
NV-EmbedQA-E5-V5

English text embedding + indexing for question-answering retrieval

CPU indexing HW - 5th gen Intel Xeon (192vCPU); GPU indexing HW - 8xL4
Embedding - nv-embedqa-e5-v5; segment size - 240K vectors (1024 Dim, fp32);
Indexing - CAGRA (GPU), HNSW (CPU); Target Recall - 98%

NVIDIA technologies

Get the best programming models in your system to be ready



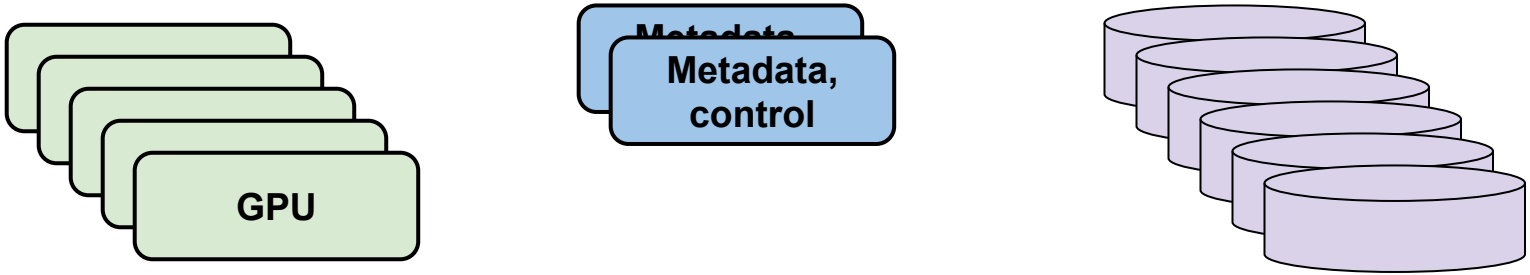
- GPUDirect Storage
  - cuFile for file - enables async for NIXL
  - cuObject for object storage - seamlessly integrated under S3 client for RDMA
  - Either of these can be to CPU or direct to GPU
- SCADA, for scaled accelerated data access
  - New programming model enabling fine-grained GPU-initiated access storage
  - Able to saturate PCIe: 98 MIOPs from NVMe driver in Gen 5

NVIDIA technologies

Get the best system and monitoring components in your system to be ready

- Spectrum X
  - Improved throughput and jitter at cluster level
  - Delivered through adaptive routing and congestion control
  - IBM hitting full bandwidth across two switches @ 79.8 GB/s IOR tput
- SuperNIC
  - Scale connections without reverting to host
  - Scaling with SuperNIC in IBM 6000 in SC25 timeframe
- BlueField
  - Plans for a JBOF with AIC
- NSight
  - Perf analysis tools integration with IBM Storage Scale

Disaggregation

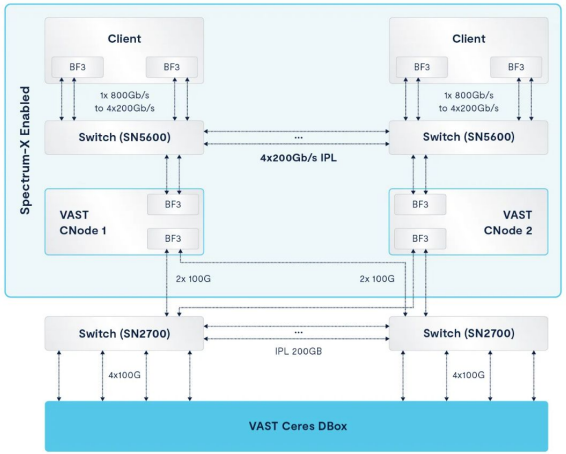


- In compute server
  - Pressure to squeeze out local storage to enable higher density
  - Trend toward managed storage for ephemeral data
  - For KV$, can’t fit enough capacity in compute server
  - For some GenAI apps, can’t get high enough IOPs either
- In storage server
  - Decouple compute/metadata from storage capacity
  - Adds flexibility, e.g. scaling for use of S3oRDMA
  - Contrast with fixed scaling, and implications of excess capacity, more cores with SW licenses than needed

Role of the network

Spectrum-X avoids congestion to remote storage, GDS enables a direct path

- NVLink: GPU-GPU within server/rack
- East-West: node-node, maybe specialized storage
- North-South: Storage, external users, external content
- Enable efficient multi-job/multi-tenancy with Spectrum-X-based congestion control
  - up to 1.48x for read/1.41x for write on NV IL-1
  - See tech blogs from NVIDIA, DDN, VAST, and Weka



Courtesy of VAST, from blog
CNodes hold metadata, direct accesses to DNodes, which hold data and connect directly

25

Call to action

- Anticipate rapid shifts in usage models
- Follow the data, dig into usage models to understand storage requirements
  - Capacity, bandwidth, IOPs
- Distinguish between requirements and acceleration
  - at each of compute, storage servers
- Tune storage architecture near compute servers and storage servers
- Leverage the best-available technology in partnering with NVIDIA
  - GPUDirect Storage, Spectrum-X, SCADA
- Watch for where new usage models may drive new SSD SKUs