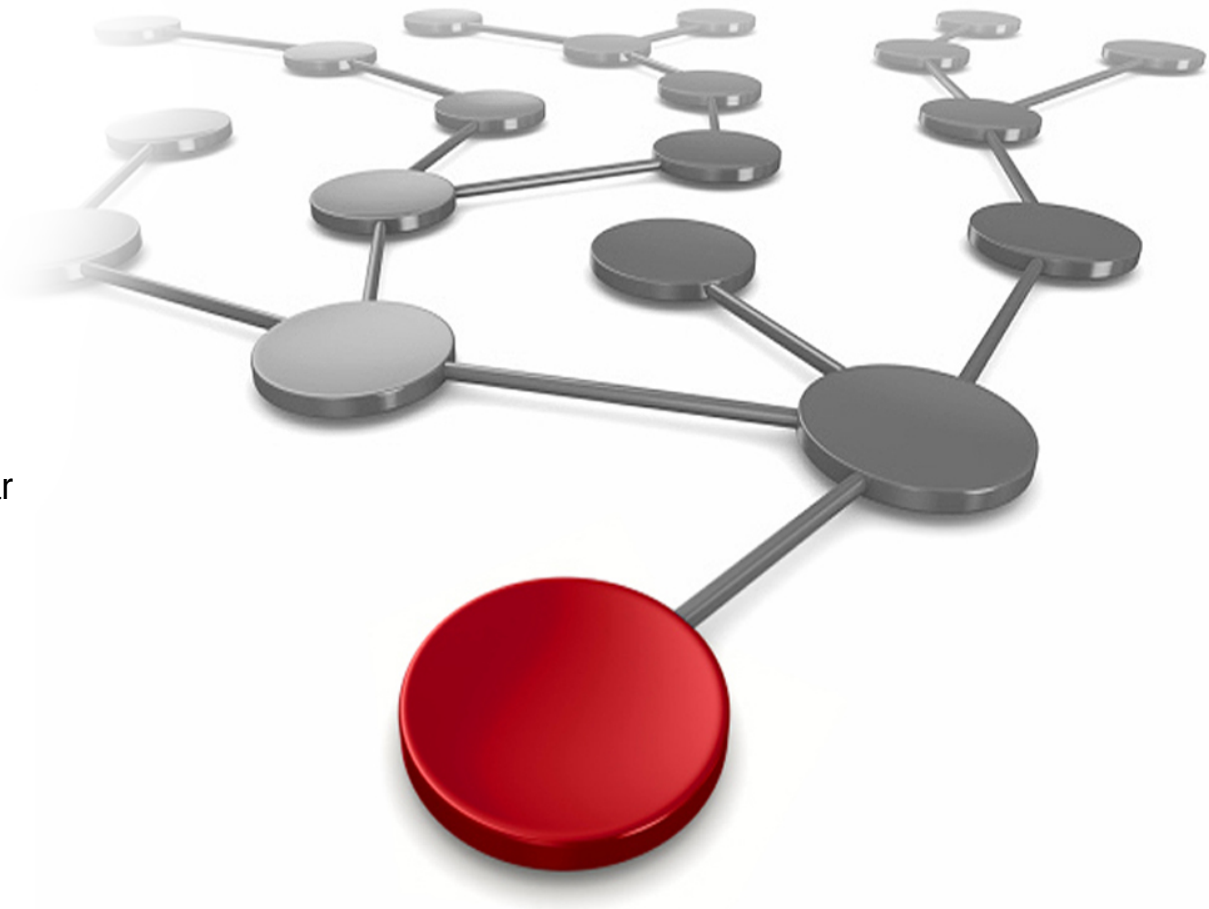
Redbooks
ibm.com/redbooks

Context Without Limits: A High-Performance KV Cache Platform for Large-Scale AI Inference



Ka Wai Leung
Guy Margalit
Anthony Hsu
Nikhil Khandelwal
Khanh Ngo
Shaunak Karmarkar
Allen Lui
Sherry Lin
Kumar Mahesh
Ronan How
Patrick Riel
James Tau
Jeff Shao
Hrishikesh Gaikwad
Jaya Mondal

IBM SUPERMICRO NVIDIA

Context Without Limits: A High-Performance KV Cache Platform for Large-Scale AI Inference

Executive Summary

As enterprises scale generative AI and agentic AI deployments, a critical infrastructure challenge is emerging: the rapid growth of inference state is outpacing GPU memory capacity, creating a bottleneck that directly impacts service quality and cost. Long-context workloads, including multi-turn assistants, retrieval-augmented generation (RAG) applications, and autonomous agent pipelines, generate large volumes of key-value (KV) cache data that must be retained across requests. When GPU memory is exhausted, inference platforms are forced to discard this cached context and recompute it from scratch, consuming expensive GPU cycles, increasing response latency, and driving up the total cost of inference at scale.

The answer is not simply more GPUs. What's needed is infrastructure that allows KV cache to be persistently stored, shared, and reused across requests, sessions, and GPU nodes — eliminating redundant computation and enabling organizations to serve far more concurrent users without a proportional increase in GPU spend. This paper presents a validated reference architecture that delivers exactly that: NVIDIA Dynamo for intelligent distributed KV cache management, IBM® Storage Scale Erasure Coding Edition (ECE) as the high-performance shared storage tier, Supermicro Petascale servers as the storage and networking foundation, and NVIDIA Spectrum-X Ethernet to tie it all together with the low-latency, high-bandwidth fabric that production AI inference demands.

Authors and Contributors

IBM

- **Guy Margalit**, Storage CTO Office
- **Anthony Hsu**, Storage Engineering
- **Nikhil Khandelwal**, Performance Engineering
- **Khanh Ngo**, Storage CTO Office
- **Ka Wai Leung**, Storage AI Solutions
- **Shaunak Karmarkar**, Storage Partner Engineering
- **Hrishikesh Gaikwad**, Storage Engineering
- **Jaya Mondal**, Storage Engineering

Supermicro

- **Allen Lui**, Architecture Product Management
- **Sherry Lin**, Architecture Product Management
- **Kumar Mahesh**, Storage Product Management
- **Ronan How**, Architecture Product Management

NVIDIA

- **Patrick Riel**, Technical Marketing
- **James Tau**, Networking Software Applications Engineering

- **Jeff Shao**, Spectrum-X Marketing

Solution objectives

- **Demonstrate scalable inference performance** by validating sustained low and predictable time-to-first-token (TTFT) as prompt lengths and context windows increase
- **Quantify the impact of KV cache persistence and sharing** by comparing re-computation-based inference against cached-context inference across multiple prompt sizes and request patterns
- **Validate high concurrency inference behavior** by measuring throughput and latency under concurrent request scenarios representative of production GenAI and agentic workflows
- **Establish IBM Storage Scale ECE on Supermicro Petascale storage servers as a viable G4-tier KV cache platform** by demonstrating, through benchmark evidence, that a shared parallel file system operating at the G4 storage tier can deliver the bandwidth, latency, and multi-tenant throughput required to serve as a scalable, high-performance KV cache tier for large-scale production inference workloads — including under adverse, noisy-neighbor network conditions

Key findings

- In single request testing measuring TTFT, with KV cache persisted on IBM Storage Scale on Supermicro Petascale Storage Servers, TTFT remains nearly flat across all prompt sizes, delivering a 56x speedup with an input sequence length of 130k tokens and eliminating prompt-length sensitivity for inference latency.
- Under concurrent load, our solution demonstrates throughput increases from 0.19 requests-per-second (RPS) to 4.26 RPS, a 22x improvement. Total processing time for 200 requests drops by 95%, confirming significantly improved GPU utilization and scalability for high-volume inference workloads.
- Under a noisy-neighbor stress test with four concurrent clients generating 200 GB/s of competing network I/O, IBM Storage Scale ECE was able to sustain inference at 3.6 RPS and completed all 200 requests in 55.56 seconds. This result is an 18x throughput improvement over the GPU recompute baseline RPS. This represents only an 18% throughput reduction with noisy-neighbor network traffic compared to the clean-network throughput result, confirming that IBM Storage Scale ECE maintains resilient, high-performance KV cache delivery in shared, multi-tenant production environments and validating G4 shared storage as a scalable and practical KV cache tier.

Customers can accelerate production GenAI and agentic AI inference with this validated reference architecture combining NVIDIA Dynamo for distributed KV cache management, IBM Storage Scale as a shared high-performance KV cache tier, Supermicro Petascale infrastructure servers, and NVIDIA Spectrum-X Ethernet for low-latency, high-bandwidth data transfer. By enabling persistent, shareable KV cache and eliminating re-computation bottlenecks, the solution delivers predictable low-latency inference and higher concurrency, providing a scalable and cost-efficient path to large-scale inference deployment.

Introduction: the AI inference challenge at scale

As large language models transition from single-turn question-and-answer interactions to complex, multi-step agentic workflows, the computational demands of AI inference have fundamentally shifted. Modern deployments now routinely handle context windows spanning hundreds of thousands to millions of tokens, driven by use cases such as multi-turn conversational agents, retrieval-augmented generation pipelines, and long-document reasoning tasks. At this scale, the key-value (KV) cache, the data structure that stores a model's intermediate attention state across all

prior tokens grows proportionally with context length, placing severe and often unsustainable pressure on GPU high-bandwidth memory (HBM) and GPU compute power for calculating prefill. Because HBM capacity is finite and shared across concurrent inference requests, larger context windows directly reduce the number of requests a GPU can serve simultaneously, degrading throughput, increasing tail latency, and driving up the total cost per query.

When the KV cache for an active session cannot fit within available HBM, inference frameworks must either evict and recompute it or stall the decode pipeline entirely; both outcomes represent significant waste of resources. Re-computation is particularly costly: regenerating a KV cache for a long context requires a full forward pass through the model, consuming GPU cycles that could otherwise be spent generating new tokens, effectively increasing the compute cost of serving returning users or resuming agentic sessions.

The NVIDIA KV cache memory tier model

The traditional remedy of simply adding more GPU memory is neither economically sustainable nor architecturally sufficient at data center scale. What is needed instead is a tiered memory architecture that extends the KV cache beyond GPU HBM into progressively larger and more affordable storage layers, without introducing latency penalties that stall the decode phase. This is precisely the problem that NVIDIA Dynamo's distributed KV cache management framework was designed to solve, orchestrating context movement across a hierarchy spanning GPU HBM (G1), system DRAM (G2), local server NVMe flash (G3), a pod-level shared flash tier (G3.5 via NVIDIA CMX), and shared enterprise storage (G4).

G1 tier

GPU HBM memory for KV tensors requiring microsecond-level access.

G2 tier

System DRAM memory within and across nodes, for KV tensor storage that exceeds GPU HBM capacity while maintaining millisecond access speeds within a single node.

G3 tier

Storage on local servers that can hold larger warm KV cache over shorter timescales.

G3.5 tier

NVIDIA’s Inference Context Memory Storage appliance (CMX) to provide a scalable, pooled flash solution for rapid KV block caching and pre-staging, persisting for minutes to hours.

G4 tier

Shared storage like IBM Storage Scale ECE, to provide a high-capacity persistence layer that enables "Context Caching," allowing massive KV cache datasets to be globally accessible across the entire GPU cluster and to persist for days, weeks or months.

Together, these tiers provide a continuum of capacity and latency targets, enabling NVIDIA Dynamo to intelligently place, evict, and reload context across the full storage stack depending on workload access patterns and cost constraints.

Why storage and networking are now core inference infrastructure

For KV cache content that is not latency critical, including inactive multi-turn session state, shared agent context, and historical query artifacts, a high-performance shared storage platform, such as IBM Storage Scale ECE, at the G4 tier is both viable and strategically advantageous. IBM Storage Scale ECE, deployed on Supermicro GPU-optimized servers and interconnected via NVIDIA Spectrum-X Ethernet for low-latency, lossless RDMA fabric, provides a massively parallel, scale-out file system capable of delivering the aggregate bandwidth and capacity required to serve KV cache content to thousands of inference nodes simultaneously. By placing reusable KV cache on IBM Storage Scale ECE rather than recomputing it on expensive GPU resources, organizations can dramatically increase tokens per second, reduce time to first token, improve GPU utilization, and extend the effective inference capacity of their AI infrastructure without proportional increases in GPU spend.

The remainder of this paper grounds these claims in benchmark evidence through a jointly conducted study between IBM, Supermicro, and NVIDIA. The sections that follow document the test environment, real-world performance results, and the recommended reference architecture and sizing options that emerge from those findings.

High-performance KV cache platform

This section introduces the technology components used in our benchmark and the specific role each plays within the validated KV cache reference architecture.

NVIDIA Dynamo

NVIDIA Dynamo™ is a distributed inference-serving framework purpose-built to solve one of the most critical bottlenecks in large-scale AI inference: the efficient management of Key-Value (KV) cache. As LLMs process long-context workloads, multi-turn conversations, agentic workflows, and deep research tasks, KV cache grows linearly and the prefill compute effort grows quadratically with prompt length. This rapidly exhausts expensive GPU memory, forcing costly trade-offs between re-computation, context window limits, and additional GPU provisioning. Dynamo resolves this through three composable KV cache capabilities:

- KV Cache-Aware Routing: mapping KV cache states across potentially thousands of GPUs and routes new requests to the GPU with the best knowledge match, avoiding costly re-computations
- Disaggregated Serving: separating the compute-intensive prefill phase from the memory-bandwidth-bound decode phase for independent scaling
- KV Cache Offloading - Dynamo's KV Block Manager (KVBM) instantly transfers KV cache from GPU memory to cost-efficient storage tiers, such as shared storage, via NIXL (low-latency transfer library) without interrupting inference.

Together, these capabilities deliver higher cache hit rates, reduced time-to-first-token, lower total cost of ownership, more tokens per second and greater concurrency.

Figure. NVIDIA Dynamo - Disaggregated LLM inference with KV cache-aware routing and offloading

IBM Storage Scale

IBM Storage Scale Erasure Code Edition (ECE) is a software-defined storage solution with a high-performance parallel file system (GPFS). It has been validated to work with Supermicro's Petascale storage servers to offer exabyte-scale capacity, ultra-low latency, parallel I/O, and tiered architectures to efficiently support training and inference workloads. It can enable seamless data sharing across three to many hundreds of Supermicro Petascale nodes, providing a cost-optimized, scalable, and resilient platform for AI. With policy-based data management, it automatically tiers data across various storage mediums within Supermicro's Petascale storage system to optimize performance and cost. This makes IBM Scale ECE an ideal solution for organizations seeking the elasticity and manageability of software-defined storage with the performance characteristics of enterprise-class file systems.

Key IBM Storage Scale ECE features (see the following figure) for AI include:

- Data starvation prevention: Eliminate GPU idle time by delivering sustained multi-GB/s throughput per node, ensuring compute resources remain fully utilized throughout training or fine-tuning cycles.
- Checkpoint efficiency: Enable rapid saving and restoration of multi-terabyte model checkpoints without disrupting training pipelines or consuming excessive wall-clock time.
- Multi-tenant performance: Support concurrent access from multiple training and inference tasks and teams without performance degradation or resource contention.
- Dataset versioning and management: Provide efficient snapshot and cloning capabilities for experiment reproducibility and dataset lineage tracking.
- Eliminating Data Transfer Bottlenecks: Support NVIDIA GPUDirect Storage (GDS), enabling direct data transfer between IBM storage and GPU memory while bypassing the CPU. Additionally, integration with NVIDIA Dynamo via the NIXL library presents a global, shared G4 namespace for KV cache cluster wide and maximizing GPU utilization across both training and inference workloads.

- Scalability without redesign: Grow from terabytes to exabytes seamlessly as model sizes and dataset volumes expand, without architectural overhauls.

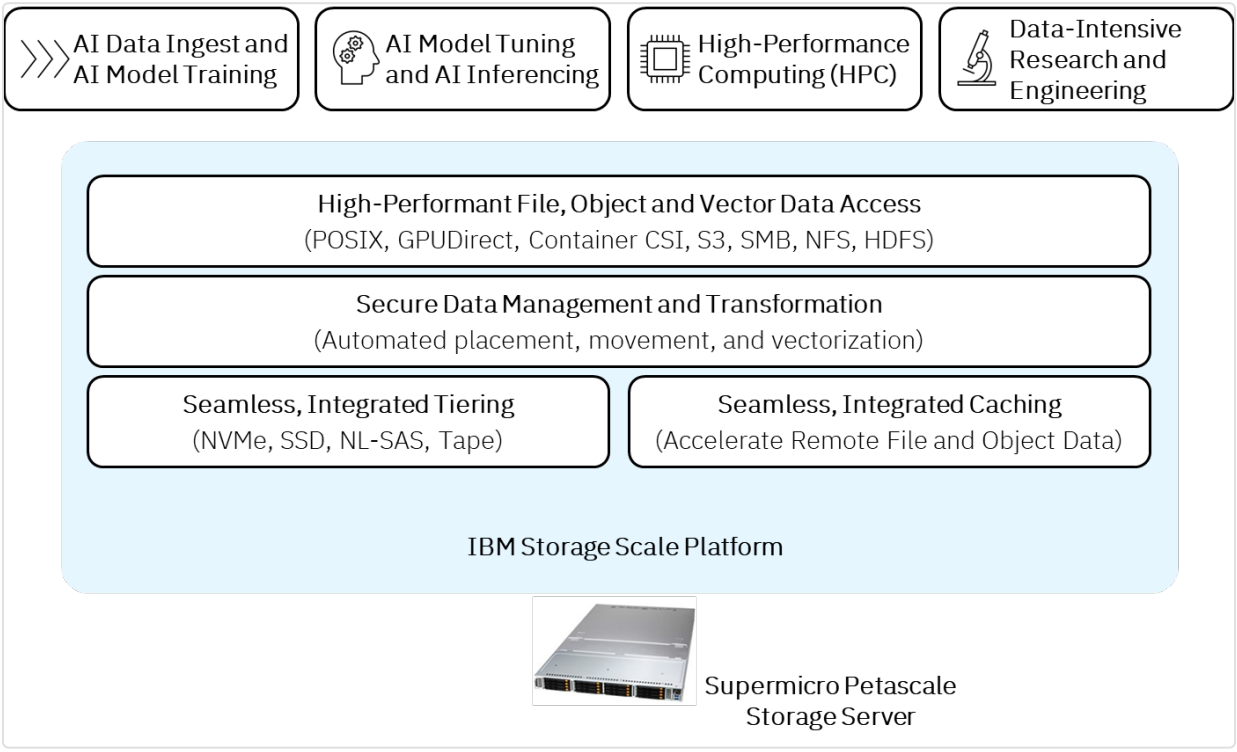


Figure. IBM Storage Scale key features for Supermicro Petascale storage servers

Supermicro Petascale storage servers

This KV Cache benchmark design leverages 1U Supermicro Petascale storage servers ASG-1115S-NE316R (see the following figure) featuring symmetrically balanced I/O architecture. Each node utilizes a single AMD EPYC CPU with 128 lanes of PCIe Gen5, with 64 lanes dedicated to NVMe storage and 64 lanes for network expansion. This design enables efficient end-to-end data flow from storage through the server root complex to the high-performance Spectrum-X Ethernet fabric.

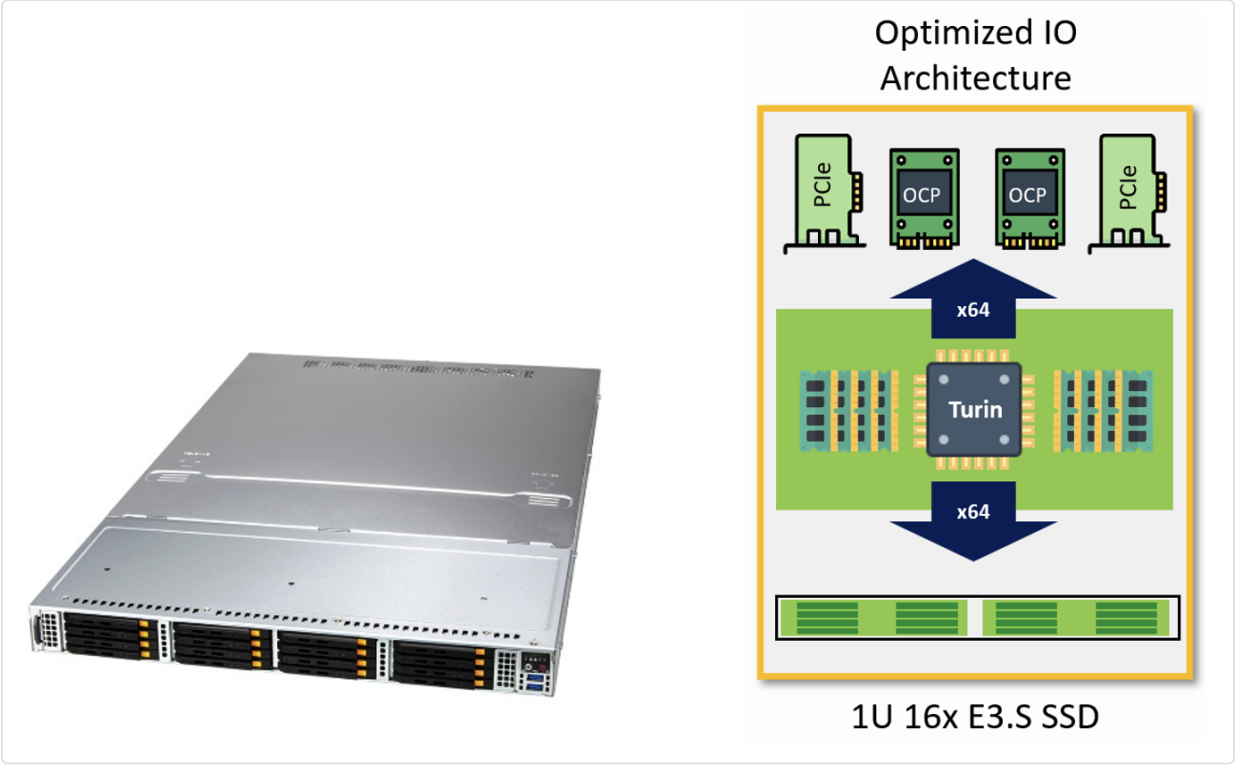


Figure. Optimized IO architecture for high-density 1U E3.S SSD storage with Turin CPU and high-bandwidth PCIe/ OCP connectivity

Supermicro NVIDIA Blackwell system portfolio

Supermicro's NVIDIA Blackwell-powered solutions, developed in close collaboration with NVIDIA, leverage Data Center Building Block Solutions® (DCBBS) to provide unprecedented performance and efficiency through both air-cooled and liquid-cooled architectures. For this specific test project, the SYS-212GB-FNR was utilized (see the following figure). It is a single-socket system based on the NVIDIA MGX™ architecture featuring an Intel® Xeon® 6700 series processor with P-cores. This high-density 2U platform supports up to four double-width PCIe-card GPUs, utilizes high-speed DDR5-6400 memory, and incorporates four front hot-swap E1.S NVMe drive bays powered by redundant 2000W Titanium Level supplies to ensure sustainable, cutting-edge performance for AI innovation.

Figure. Supermicro single socket GPU server SYS-212GB-FNR

NVIDIA Spectrum-X Ethernet

Storage networking performance is critical, and this benchmark design utilizes three NVIDIA Spectrum-X SN5600 Ethernet switches to connect all storage and compute nodes via NVIDIA BlueField-3 DPUs and SuperNICs. The fifth-generation Spectrum SN5000 series delivers port speeds from 10 to 800 Gb/s, specifically designed to accelerate data center fabrics. The end-to-end NVIDIA Spectrum-X Ethernet represents the first Ethernet platform purpose-built for AI workloads, addressing the performance bottlenecks that traditional Ethernet solutions face with modern AI models.

Key technology innovations include integrated adaptive routing, programmable congestion control, and QoS capabilities. These features enable multi-tenant environments to run data-intensive workloads on shared infrastructure while preventing noisy neighbor issues and ensuring consistent performance isolation across concurrent workloads.

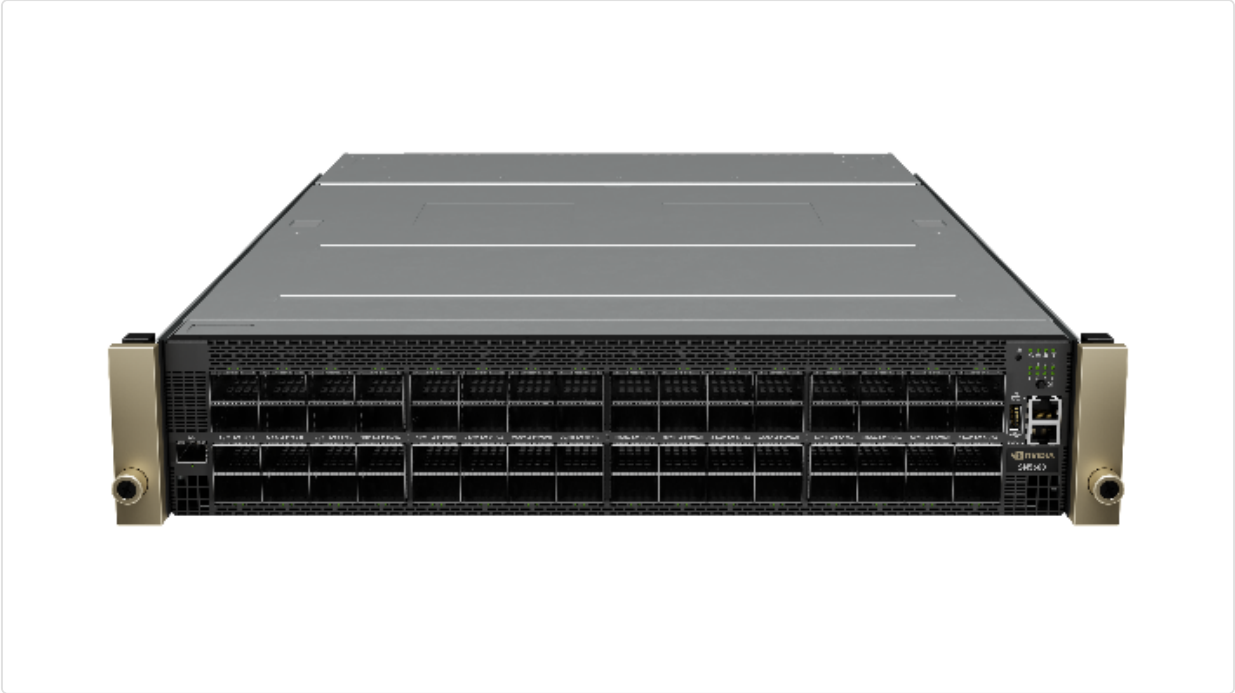


Figure. NVIDIA Spectrum-X SN5600/5610, High-performance Ethernet switching built for AI infrastructure at scale

Benchmark environment and methodology

The benchmark environment utilizes an 8-node Supermicro Petascale storage cluster with IBM Storage Scale ECE as the foundational layer. The network fabric is built on a high-bandwidth NVIDIA Spectrum-X spine-leaf architecture featuring 800 Gb/s uplinks. The IBM Storage Scale nodes leverage 400 Gb/s ConnectX-7 links for communication between storage nodes, while client nodes utilize NVIDIA BlueField-3 DPUs to deliver 400 Gb/s to GPU-accelerated nodes and 200 Gb/s to CPU-only nodes (see the following figures).

All inference benchmark results presented in this paper were generated exclusively from a single GPU client node, the Supermicro SYS-212GB-FNR equipped with four NVIDIA RTX PRO 6000 GPUs. This demonstrates that using a properly tuned network and storage infrastructure stack, one can attain enterprise grade KV Cache performance with NVIDIA RTX 6000 systems. The remaining four client nodes served a distinct purpose: generating sustained network I/O load in the noisy-neighbor benchmark test (Benchmark 3) to simulate real-world multi-tenant network contention conditions.

Real-world networking topology

We used a spine-leaf network topology, with NVIDIA Spectrum-X switches, to replicate the networking conditions found in AI production environments. Other benchmarks often use a flat network where all hardware resides on the same subnet, or a simple east-west architecture where traffic flows laterally between nodes without tiered switching. Our benchmark on spine-leaf network introduces the additional network hops, bandwidth segmentation, and traffic management overhead that are characteristic of real-world deployments. This ensures our benchmark results can match the performance characteristics customers expect in production, rather than an idealized, low-latency single-segment or single-switch environment. With a 3-switch, spine-leaf topology for testing, we can have greater confidence that our solution will perform predictably at large scale AI inference deployments.

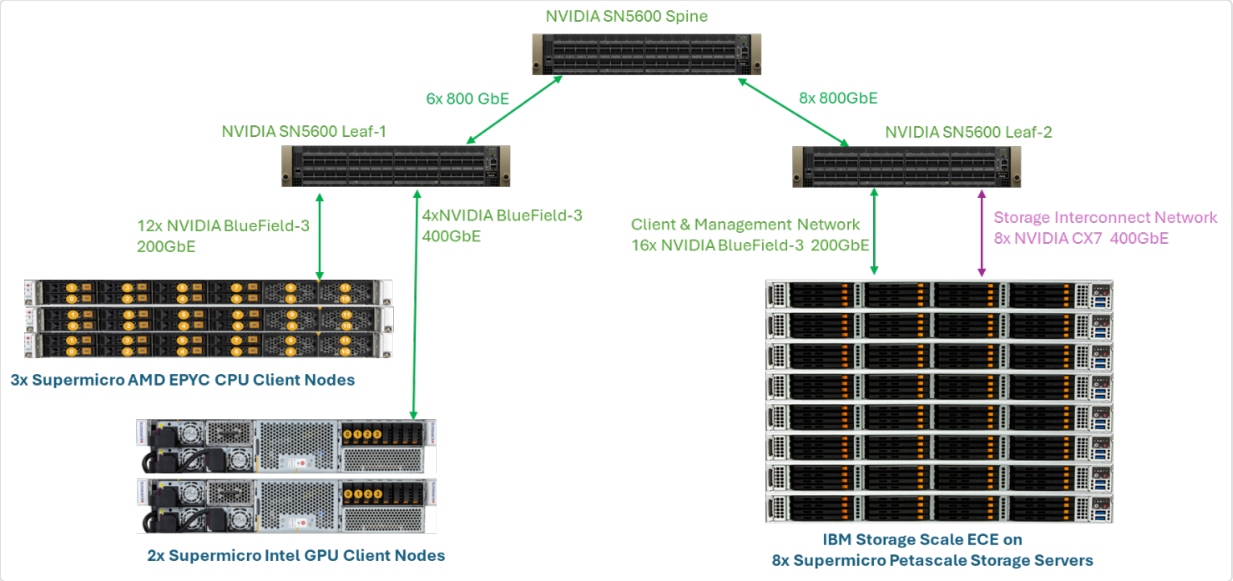


Figure. KV cache benchmark setup with NVIDIA Dynamo, NVIDIA Spectrum-X, IBM Storage Scale, and Supermicro hardware

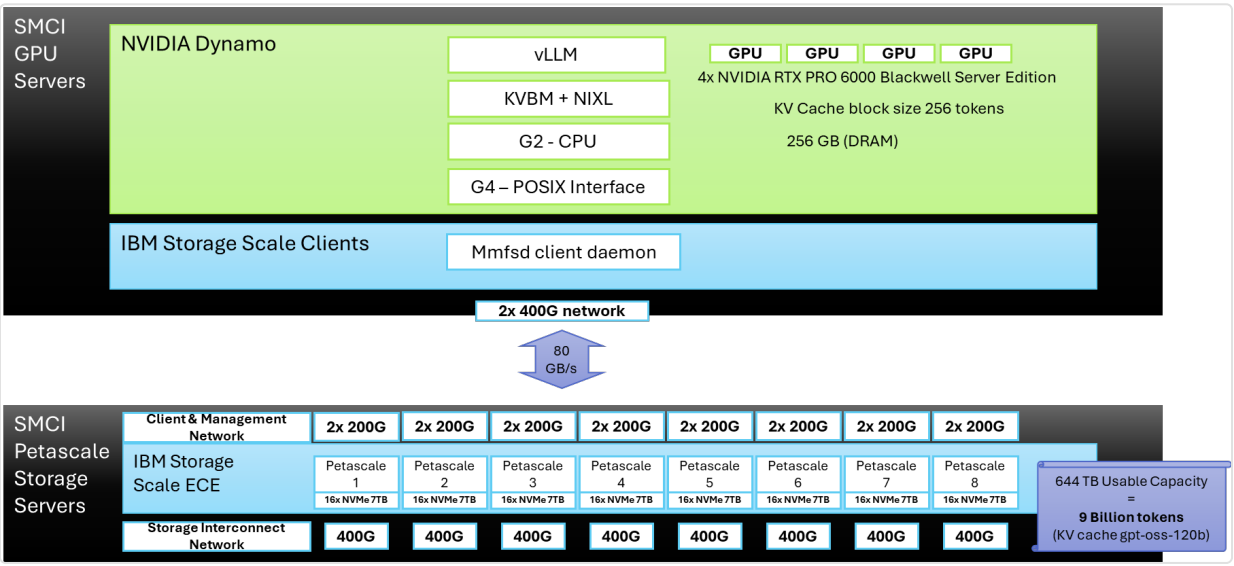


Figure. NVIDIA Dynamo KV cache benchmark setup with IBM Storage Scale ECE

Benchmark configuration: hardware

- 3 NVIDIA Spectrum-X SN5600 switches: Spine/Leaf Configuration
- 8 Supermicro Petascale ASG-1115S-NE316R Storage Nodes
  - 16x Micron E3 7.68TB drives
  - 1 x B3220L E-Series Dual port 200Gb (Client-side)
  - 1x NVIDIA ConnectX-7 400 Gb (Storage Backend)
  - AMD EPYC 9535 CPU

- 5 Heterogeneous client nodes
  - 3x AMD compute client nodes (AMD EPYC 9455/9624), each node has 2x Dual port 200 Gb/s BlueField-3 (B3220) DPU (4 connections per server). 2 of 3 servers have 384GB memory, one server has 768GB memory.
  - 1x Intel 2U GPU System with 2x NVIDIA HGX H100s and 2x NVIDIA H200s with one Intel 8592 CPU, 768GB memory. 2x NVIDIA BlueField-3 B3140H E-series 400Gb/s DPU
  - 1x Intel 2U GPU System with 4x NVIDIA RTX PRO 6000 GPUs with one Intel 8592 CPU, 768GB memory. 2x NVIDIA BlueField-3 B3140H E-series 400Gb/s DPU

Benchmark configuration: software

The benchmark software environment combined IBM Storage Scale ECE v6.0.0.1 with NVIDIA Dynamo to evaluate storage-backed KV cache reuse for long-context LLM inference. IBM Storage Scale ECE was deployed on eight Supermicro Petascale storage nodes running RHEL 9.6, using an 8+2P RAID configuration. Five Supermicro client nodes running Ubuntu 24.04 accessed the shared IBM Storage Scale file system through the IBM Storage Scale client.

NVIDIA Dynamo provided the inference orchestration and KV cache management layer, with inference served through a vLLM v0.14.1 OpenAI-compatible chat endpoint. NVIDIA AIPerf v0.6.0.post1–v0.7.0 was used to generate and measure the inference workload against the openai/gpt-oss-120b model. This software stack enabled comparison of cold-cache execution, warm-cache KV reuse, and performance under noisy-neighbor load.

Software Component	Version / Configuration
IBM Storage Scale ECE	v6.0.0.1
Storage node OS	RHEL 9.6
Client node OS	Ubuntu 24.04
Storage configuration	8 Supermicro Petascale nodes, 8+2P RAID
Client configuration	5 Supermicro client nodes
NVIDIA Dynamo	v0.9.0+
vLLM	v0.14.1
NVIDIA AIPerf	v0.60.post1
Model	openai/gpt-oss-120b

Model Selection and KV Cache Configuration

Running a 120-billion parameter model in native 16-bit precision demands over 240GB of VRAM for weights. We used the Hugging Face GPT-OSS 120B model pre-quantized to MXFP4 to reduce the memory footprint and enable this model to run entirely on a single NVIDIA RTX Pro 6000 Blackwell GPU, keeping the benchmark self-contained and cost-competitive.

The vLLM engine was initialized at 90% GPU memory utilization, with the MXFP4 model weights and CUDA graph pool consuming approximately 60.9 GiB combined, leaving 19.6 GiB free for the KV cache. Given this headroom, we ran the KV cache in standard FP16 precision (kv_cache_dtype=auto), preserving full mathematical accuracy and supporting 1,141,783 tokens stored in GPU memory on a single node.

Storage Tiering and Isolation Strategy

With G1 (GPU HBM) and G2 (System DRAM) cache tiers disabled by setting their capacity to zero in Dynamo's configuration, all 1.4 million tokens were managed exclusively through the G4 storage tier, backed by IBM Storage Scale ECE on Supermicro Petascale storage servers. When the GPU needs cached context, Dynamo's Block Manager retrieves it in two steps: a Remote-to-Host (R2H) transfer pulls the data from external storage into CPU RAM, followed by a Host-to-Device (H2D) transfer that moves it from CPU RAM onto the GPU and transferred in chunks to keep data flowing efficiently. This isolates the benchmark to a clean measurement of G4's ability to sustain high-concurrency, multi-turn workloads, proving that large-scale context persistence is achievable on a single node, eliminating the need to provision additional GPU nodes.

Performance Results and Analysis

NVIDIA AIPerf benchmarking tool was used to perform our benchmark testing. NVIDIA AIPerf is a comprehensive benchmarking tool included as a core component of the NVIDIA Dynamo framework, designed to measure generative AI model performance across all major inference backends including SGLang, TensorRT-LLM, and vLLM. It provides standardized, reproducible measurements across key LLM inference metrics, including Time to First Token (TTFT), enabling comparisons across different hardware configurations, storage architectures, and parallelization strategies. The benchmark results presented here leverage AIPerf within NVIDIA Dynamo to evaluate the impact of IBM Storage Scale on Supermicro as an external KV cache (G4 layer), measuring TTFT and throughput response times across a range of prompt lengths to quantify the performance gains achievable with storage-offloaded caching.

Benchmark 1: TTFT from a single GPU

The benchmark results (see the following figure) demonstrate a dramatic performance advantage when using IBM Storage Scale as a KV cache for LLM inference. Without KV caching, GPU recompute times scale steeply with prompt length, rising from 0.572 seconds at 10k tokens all the way to 32.14 seconds at 130k tokens, reflecting the exponential cost of reprocessing long contexts from scratch on every request.

IBM Storage Scale (G4), however, maintains a nearly flat sub-second TTFT profile across the entire prompt length range (between 0.193 and 0.57 seconds) regardless of context size, providing a 56x speedup at the 130k token mark. As GPU recompute times degrade sharply as prompt length grows, IBM Storage Scale effectively eliminates prompt-length sensitivity by offloading KV cache retrieval to high-performance storage, making inference latency predictable and consistent even at very large context windows. For production deployments handling long documents, multi-turn conversations, or large system prompts, this translates to a more responsive, scalable and cost-efficient inference architecture.

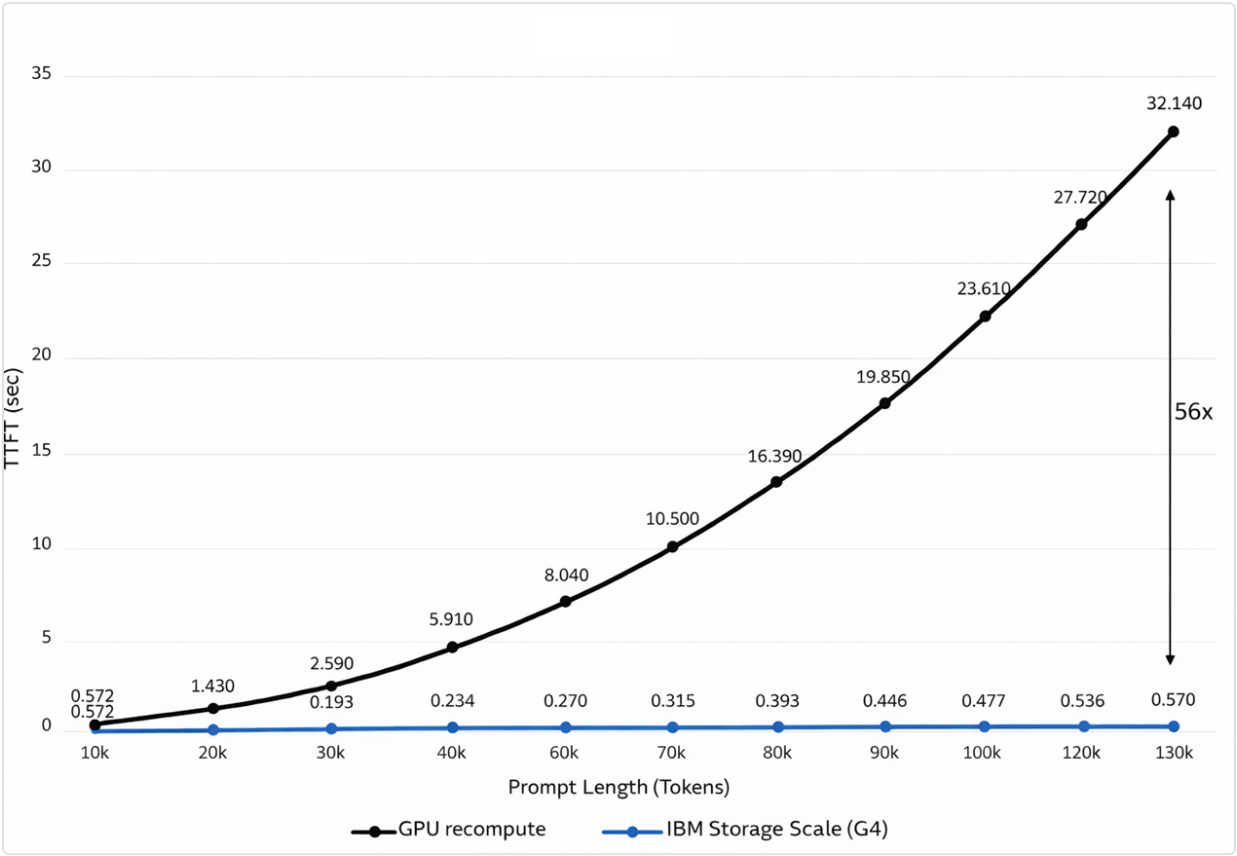


Figure. Cached requests are up to 56x faster on IBM Storage Scale

Benchmark 2: throughput and response time

The throughput acceleration benchmark (see the following figure) ran 200 requests against a 120B parameter model at 28 concurrent connections, drawing randomly from 100 unique prompts representing 24 million tokens and 825 GB of KV cache data. One important methodological detail: the G4 cache started empty at the beginning of the run, meaning early requests were served without any cache benefit. Acceleration compounded progressively as cache hit rates built up across the 200-request sequence. This benchmark contains full run average including a cold-start period.

Despite this cold-start disadvantage, the results still reveal a dramatic 22x efficiency gain when using IBM Storage Scale ECE as the G4 KV cache backend. Requests per second jumped from just 0.19 with GPU recompute to 4.26 RPS with IBM Storage Scale, and total time to process all 200 requests dropped from 1,048 seconds down to just 46.94 seconds, a reduction of over 95%. We believe the 22x efficiency figure is a conservative number. Had the cache been pre-warmed before the run, the acceleration gap ratio would be even greater.

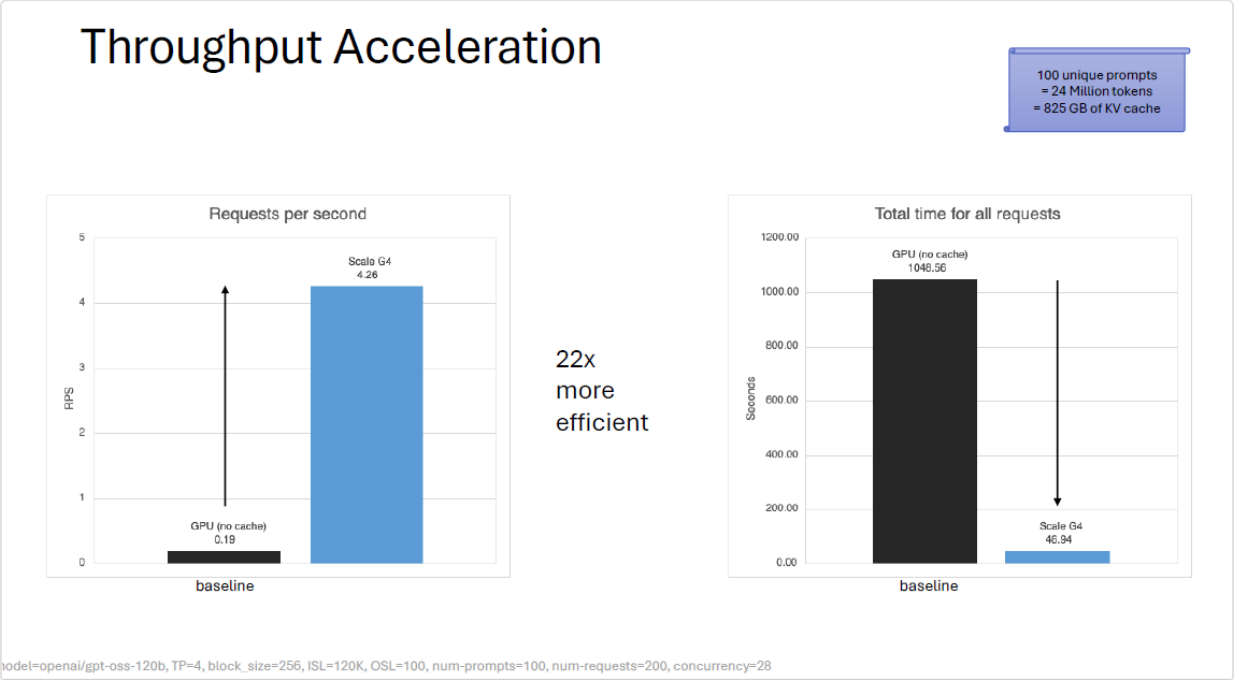


Figure. Response time throughput comparison of 100 prompts

Benchmark 3: noisy neighbor test

This throughput acceleration benchmark extends the previous test by introducing a "noisy neighbor" scenario with four concurrent clients continuously generating 200 GB per second of network I/O traffic to simulate the real-world conditions of a shared, multi-tenant environment. Even under this significant network pressure, IBM Storage Scale delivers 3.6 RPS compared to the GPU recompute baseline of just 0.19 RPS, an 18x efficiency improvement, and completes all 200 requests across 28 concurrent connections in just 55.56 seconds versus 1,048.56 seconds for the no-cache baseline. Compared directly to the clean (no-noise) IBM Storage Scale G4 result of 4.26 RPS and 46.94 seconds total time for all requests, we are only seeing a minimal degradation with 18% reduction in throughput and roughly a 9-second increase in total completion time. This resilience demonstrates that IBM Storage Scale is not only capable of delivering high KV cache performance in ideal conditions, but that its architecture is robust enough to maintain superior inference efficiency even when competing with heavy concurrent network workloads.

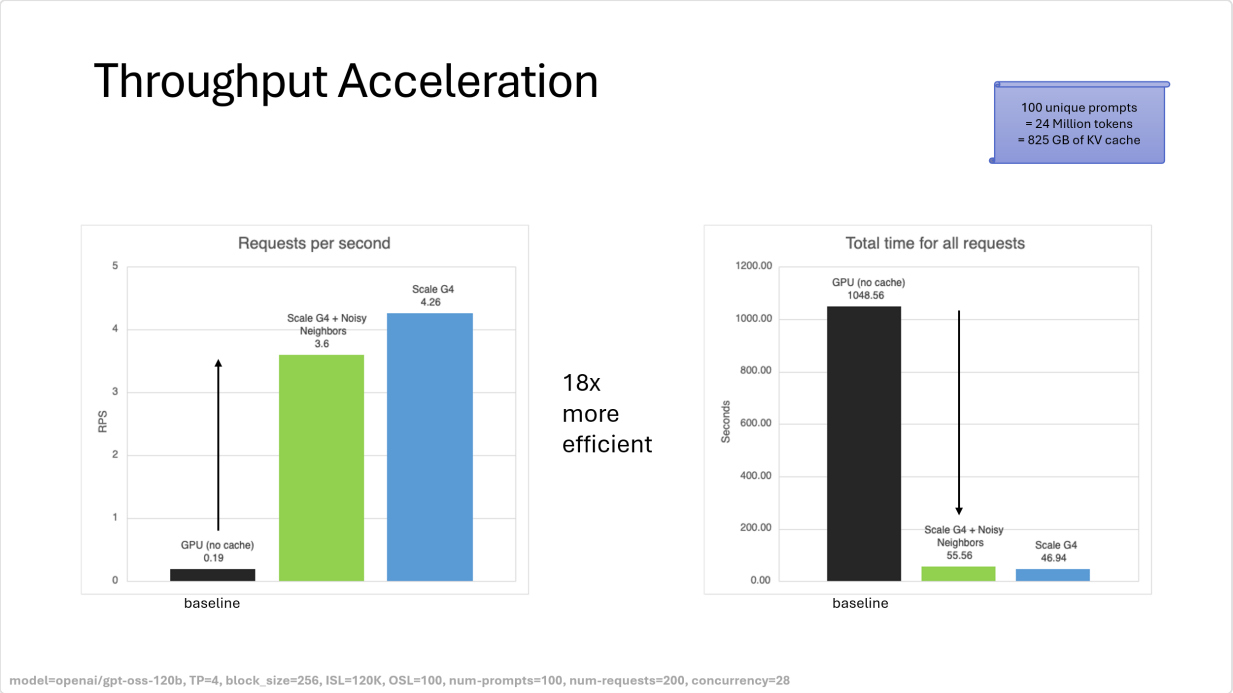


Figure. Throughput acceleration with noisy neighbors

Why hardware matters

The robust KV Cache benchmarking results are attributed not only to the capabilities of IBM Storage Scale but also to the high-performance Supermicro hardware foundation underpinning the solution. To simulate concurrent, high-demand inference loads, the five client nodes driving the NVIDIA Dynamo workload provided the essential CPU and GPU compute density, memory bandwidth, and high-speed networking.

On the storage tier, eight Supermicro Petascale nodes, purpose-built for AI, ensured that data movement never became a bottleneck, supported by the following architectural advantages:

- **Ultra-Low Latency at Scale:** GPUs require constant, high-speed data input to maintain parallel processing efficiency. By utilizing all-flash media and ConnectX network interfaces, the Petascale system delivers millions of IOPS and multi-GB/s throughput, preventing high-latency "hangs" during metadata operations, batch loading, and checkpointing.
- **Parallelism and Scalability:** AI training and inference often span multiple GPUs and nodes, requiring a storage foundation that can support high-performance parallel file systems. The Petascale architecture is a proven industry standard for parallel file system applications, providing the necessary structural foundation for distributed data.
- **Synchronized Bandwidth:** To maintain maximum GPU utilization, storage speeds must match network and GPU throughput. Each Petascale system utilizes 64 PCIe Gen5 lanes for 16 NVMe drives alongside dual 400Gb/s network interfaces to feed GPU servers at line rate.
- **NVIDIA Magnum IO GPUDirect Storage (GDS) Support** : GDS is a critical requirement for modern AI workloads. The Petascale systems support this via high-performance RDMA capabilities, allowing data to bypass the CPU and move directly from storage to GPU memory.

The Supermicro Petascale storage server features a balanced design to eliminate "hot spots" within the system. In KV Cache scenarios, where metadata and variable data sizes must move between memory and storage with extreme velocity, this 1U system supports 16 NVMe drives to provide maximum storage density and performance.

The integration of NVIDIA BlueField-3 DPUs on the client nodes further offloads networking and data movement tasks from the CPU. This ensures that CPU and GPU resources remain dedicated to inference execution. Together, this Supermicro infrastructure ensures that compute, memory, and storage I/O are never limiting factors for IBM Storage Scale.

Scaling Dynamo KV cache results

The benchmark results presented in this section reflect what is achievable today with the current 2x400GbE networking configuration between the client and storage nodes. This setup supports up to 80 GB/sec of client-facing bandwidth. Notably, IBM Storage Scale ECE running on the 8 Supermicro Petascale storage nodes can deliver up to 315 GB/sec of read throughput, nearly 4x what the installed networking was able to take advantage of during this testing. The storage tier has significant performance headroom that has yet to be fully utilized. If more 400GbE client-side networking can be added, KV cache retrieval performance can be expected to improve substantially. Organizations deploying this reference architecture can look forward to even lower TTFT latencies and higher request throughput as the full storage bandwidth potential is fully unlocked.

Additional benefits over other KV cache implementations

While GPU HBM (G1), System DRAM (G2), and local NVMe (G3) support cross-node KV cache movement via east-west fabrics, each tier remains fundamentally bounded by per-node physical capacities. GPU HBM provides the bandwidth required for real-time token generation, but its premium cost and capacity limits make it suboptimal for long-term KV Cache retention. Keeping idle KV caches in HBM directly displaces active computation, severely restricting maximum batch sizes and aggregate throughput.

System DRAM (G2) adds headroom but at significant hardware cost. Local NVMe (G3) expands capacity further but remains ceiling-bound by the storage capacity of a single server. A node failure across any of these tiers means cached KV data is lost. While NVIDIA Dynamo's KV-aware Router and NIXL library enable cross-node sharing, G2 and G3 data still require explicit cross-node transfer. The KV-aware Router must locate the block and NIXL must initiate a point-to-point fetch before inference can proceed. IBM Storage Scale ECE as G4 overcomes this entirely by aggregating capacity across the storage cluster, presenting a single KV cache namespace accessible to multiple GPU clusters simultaneously, and delivering persistent, fault-tolerant storage with benchmark performance closely approaching local NVMe.

Deployment sizing options

IBM Storage Scale ECE deployed on Supermicro Petascale storage servers represents a software-defined storage solution purpose-built for the demanding performance and capacity requirements of KV Cache based inference workloads. This enterprise-grade storage infrastructure can scale efficiently from small KV Cache deployments to large AI factories.

The IBM Storage Scale ECE solution offers flexible deployment options, from as few as 3 storage servers to 256 for large-scale deployments. Sizing a KV cache infrastructure correctly requires balancing several interdependent factors, including usable storage capacity, aggregate read bandwidth, erasure coding overhead, concurrent request density, supported context window size, and the number of GPU clients the storage tier must serve simultaneously. From a KV cache perspective specifically, undersizing any one of these dimensions has compounding consequences:

- Insufficient storage bandwidth starves GPUs of context data and drives up TTFT
- Inadequate capacity forces premature cache eviction and re-computation

- Insufficient concurrency headroom collapses throughput under multi-tenant production loads

All of which wastes expensive GPU resources and degrades the end-user experience. Therefore, ensuring that storage capacity, bandwidth, and concurrency are jointly sized to the target workload profile is critical.

The workload will also greatly influence the storage requirements. For example, aggregate throughput requirements can vary dramatically based on the type of model and workload type. An inference workload may require 1 GB/sec per GPU of throughput, while a training workload on uncompressed high-resolution images may require 4 GB/sec per GPU.

The following table represents estimates of small, medium, and large KV Cache deployment configurations to support a sub-second TTFT response. However, it is important to note that the solution can scale far beyond what is represented below. It is highly recommended to work with Supermicro and IBM to derive a configuration and the appropriate KV Cache storage capacity that will meet your desired performance objectives and to fully utilize powerful and costly GPUs.

	Small	**Medium (tested configuration)**	**Large**
KV Cache Use Cases	Proof-of-concept, dev/test environments, single workload inference with small to mid-sized LLMs (7B–34B parameters)	Production inference, multi-tenant environments, large LLMs (70B–120B parameters), RAG workloads, and long-context multi-turn conversations	Enterprise AI factories, mixed training and inference, multiple concurrent large models, and ultra-high concurrency production deployments
IBM Storage Scale ECE on SMC Petascale node count	4 nodes	8 nodes (configuration used for this testing)	16 nodes+
Recommended Erasure Code	4+2P	8+2P	16+3P
Usable Capacity	67%	80%	84%
Model Parameter Range	7B – 34B	70B – 120B	120B+ / Multi-model
Est. KV Cache per Request (FP16 precision, GQA models)	~3–10 GB per request (MHA models: ~12–80 GB)	~40–80 GB per request (e.g. Llama 3.1 70B @ 128K ≈ 40 GB; 120B @ 130K ≈ 60–80 GB) (MHA: multiply 4–8×)	~80–160 GB+ per request (MHA models: significantly higher)
Context Window Size Support	Dependent on Model	Dependent on Model	Dependent on Model
Estimated GPUs	1–64	100-256	300–1000+
Required Aggregate Storage Bandwidth (Read)	~20–75GB/s	100-300 GB/sec (315 GB/s achieved in previous testing)	300+ GB/s

For more information on data protection and storage utilization in IBM Storage Scale ECE, see Data protection and storage utilization.

Summary

This solution paper demonstrates that a shared storage based KV cache infrastructure can address enterprises operating generative AI and agentic workloads at scale. Through three rigorous benchmarks on a validated reference architecture combining NVIDIA Dynamo, NVIDIA Spectrum-X Ethernet, IBM Storage Scale ECE, and Supermicro Petascale servers, this paper quantifies the business impact of eliminating GPU re-computation through shared G4-tier KV cache storage: a 56x reduction in time-to-first-token for large prompt context lengths, a 22x improvement in request throughput under concurrent production load, and a sustained 18x throughput advantage even under noisy-neighbor network stress. This is all validated on a spine-leaf topology representative of real production environments. For enterprises seeking to maximize the return on their GPU infrastructure investment, this architecture delivers a clear and immediately deployable path to higher throughput, lower latency, greater concurrency, and a fundamentally more cost-efficient inference platform.

The value of this solution paper goes beyond just the benchmarking results. We have integrated, tested, and tuned 4 products (NVIDIA Dynamo, IBM Storage Scale, Supermicro servers, and NVIDIA Spectrum-X Ethernet) together in a real-world environment to ensure interoperability, and the organizational confidence to deploy at scale, all of which this reference architecture delivers.

References

IBM Storage Scale references

- IBM Storage Scale - IBM Documentation

NVIDIA Networking & Software References

- Scale and Serve Generative AI | NVIDIA Dynamo
- NVIDIA Spectrum-X Ethernet Platform for AI Networking
- BlueField Networking Platform | NVIDIA

Supermicro hardware references

- Supermicro Petascale Storage Server (ASG-1115S-NE316R)
- Supermicro NVIDIA Blackwell Portfolio (SYS-212GB-FNR)

Open source software and models

- vLLM Serving Engine
- Hugging Face GPT-OSS-120B Model

Glossary

AI - Artificial Intelligence

Computer systems designed to perform tasks that typically require human intelligence, including learning, reasoning, and problem-solving. IBM® Storage Scale optimizes data access for AI workloads.

CPU - Central Processing Unit

The primary processor in a computer system that executes instructions and manages system operations.

CUDA - Compute Unified Device Architecture

NVIDIA's parallel computing platform and programming model for GPU acceleration.

DDR5 - Double Data Rate 5

Fifth generation of DDR SDRAM memory technology, providing higher bandwidth and capacity than DDR4.

DPU - Data Processing Unit

Specialized processor that offloads networking, storage, and security tasks from the CPU. NVIDIA BlueField-3 DPUs are used in this architecture for high-performance data movement.

DRAM - Dynamic Random-Access Memory

Standard system memory providing fast access but requiring constant refresh. Represents the G2 tier in NVIDIA's memory hierarchy.

ECE - Erasure Code Edition

IBM® Storage Scale variant that uses erasure coding for data protection, providing better storage efficiency than traditional RAID while maintaining high availability.

FP16 - 16-bit Floating Point

Numerical precision format using 16 bits per number, balancing accuracy and memory efficiency for AI workloads.

GB/s - Gigabytes per Second

Data transfer rate measurement indicating billions of bytes transferred per second.

Gb/s - Gigabits per Second

Network speed measurement indicating billions of bits transferred per second.

GDS - GPUDirect Storage

NVIDIA technology enabling direct data transfer between storage and GPU memory, bypassing CPU and system memory for improved performance.

GenAI - Generative AI

AI systems capable of creating new content (text, images, code) based on learned patterns from training data.

GiB - Gibibyte

Binary unit of digital information equal to 1,024 mebibytes (2^30 bytes).

GPFS - General Parallel File System

IBM's high-performance clustered file system, now known as IBM® Storage Scale.

GPU - Graphics Processing Unit

Specialized processor designed for parallel processing, widely used for AI/ML training and inference.

GQA - Grouped Query Attention

Attention mechanism variant that groups queries to reduce KV cache size while maintaining model quality.

H2D - Host-to-Device

Data transfer operation moving data from system memory (CPU) to GPU memory.

HBM - High Bandwidth Memory

High-performance RAM used in GPUs, providing extremely high bandwidth but limited capacity. Represents the G1 tier in NVIDIA's memory hierarchy.

I/O - Input/Output

Operations that transfer data between a computer and external devices or storage systems.

IOPS - Input/Output Operations Per Second

Performance measurement for storage devices indicating the number of read/write operations completed per second.

KV - Key-Value

Data structure format used in LLM inference to cache attention mechanism states. KV cache stores computed attention keys and values to avoid recomputation across tokens.

KVBM - KV Block Manager

NVIDIA Dynamo component responsible for managing KV cache blocks across storage tiers.

LLM - Large Language Model

AI models with billions of parameters trained on vast text datasets to understand and generate human language.

MHA - Multi-Head Attention

Standard attention mechanism in transformer models that uses separate attention heads. Requires more KV cache memory than GQA.

NIXL - NVIDIA Inference eXchange Library

NVIDIA library for low-latency KV cache transfer between storage tiers in the Dynamo framework.

NVMe - Non-Volatile Memory Express

High-performance storage protocol designed for solid-state drives (SSDs) that connects directly to the PCIe bus.

OS - Operating System

System software that manages computer hardware and software resources and provides common services for computer programs.

PCIe - Peripheral Component Interconnect Express

High-speed serial computer expansion bus standard. Gen5 provides up to 128 GB/s bandwidth per x16 slot.

QoS - Quality of Service

Network management technique that prioritizes certain types of traffic to ensure consistent performance for critical workloads.

R2H - Remote-to-Host

Data transfer operation moving KV cache from remote storage (G4) to system memory (G2).

RAG - Retrieval-Augmented Generation

AI technique that enhances LLM responses by retrieving relevant information from external knowledge bases before generating output.

RAID - Redundant Array of Independent Disks

Data storage technology that combines multiple disk drives into a logical unit for redundancy and/or performance.

RDMA - Remote Direct Memory Access

Technology that allows direct memory access from one computer to another without involving the operating system, enabling high-throughput, low-latency networking.

RHEL - Red Hat Enterprise Linux

Commercial Linux distribution developed by Red Hat. Used as the operating system for storage nodes in this architecture.

RPS - Requests Per Second

Throughput metric measuring the number of inference requests completed per second.

TTFT - Time to First Token

Latency metric measuring the time from request submission to generation of the first output token. Critical user experience metric for LLM inference.

vLLM - Versatile Large Language Model

Open-source LLM inference engine optimized for high throughput and efficient memory usage.

VRAM - Video Random-Access Memory

Memory used by graphics cards and GPUs, typically HBM in modern AI accelerators.

Trademarks

IBM, IBM Storage Scale, and the IBM logo are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.

NVIDIA, Dynamo, Spectrum-X, BlueField, ConnectX, RTX, HGX, GPUDirect, CUDA, Blackwell, and MGX are trademarks or registered trademarks of NVIDIA Corporation in the U.S. and other countries.

Supermicro and Petascale are trademarks or registered trademarks of Super Micro Computer, Inc.

Intel and Xeon are trademarks of Intel Corporation or its subsidiaries.

AMD and EPYC are trademarks of Advanced Micro Devices, Inc.

Red Hat and RHEL are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

Ubuntu is a registered trademark of Canonical Ltd.

Micron is a trademark of Micron Technology, Inc.

OpenAI and GPT are trademarks of OpenAI, Inc.

Other company, product, and service names may be trademarks or service marks of others.

AI Attribution

This work was primarily human-created. AI was used to make stylistic edits, such as changes to structure, wording, and clarity. AI was used to make content edits, such as changes to scope, information, and ideas. AI was prompted for its contributions, or AI assistance was enabled. AI-generated content was reviewed and approved. The following model(s) or application(s) were used: IBM Bob.