SOLUTION BRIEF

UNLOCKING PETASCALE STORAGE I/O IN MULTI-TENANT AI ENVIRONMENTS WITH NVIDIA SPECTRUM-X™ ETHERNET NETWORKING AND IBM STORAGE SCALE

Supermicro Petascale Server

TABLE OF CONTENTS

Executive Summary

In this solution brief, we will describe a high-performance storage cluster, purpose-built for the most demanding AI training and inference workloads running over an Ethernet network. The key components of the storage architecture include Supermicro’s Petascale server equipped with Micron E3.S NVMe, connected via NVIDIA Spectrum-X Ethernet.

The final enabling element for this solution is IBM Storage Scale Erasure-Code-Edition (ECE), a high-performance software-defined storage (SDS) solution featuring a parallel file system with a proven history in large-scale HPC, offering extreme throughput and low-latency multi-tenant use cases. To best simulate a real-world network topology, a leaf-and-spine switch configuration is used. This is the same switch topology used for rapid scaling of AI environments in production. The performance benchmarks will demonstrate the advantages of using end-

to-end Spectrum-X Ethernet features with support for NVIDIA Ethernet SuperNICs and DPUs, as well as the advanced tools and feature sets NVIDIA offers for optimal application performance in multi-tenant AI environments.

Solution/Target Objectives

- Demonstrate optimal AI storage performance using validated leaf-spine network topology with measured throughput, latency, and IOPS benchmarks for training and inference workloads.

- Validate multi-tenant performance isolation by simulating noisy neighbor scenarios with concurrent workload interference patterns and quantifying the impact on QoS guarantees.

- Deliver reference design with pre-validated configurations, scaling parameters, and deployment guidelines to accelerate customer time-to-production.

Solution Test Summary

- By leveraging the key components and features outlined in this document, the solution achieved up to a 9% throughput gain compared to non–NVIDIA Spectrum-X Ethernet environments — 307.2 GiB/s aggregated read bandwidth versus 281.8 GiB/s. This improvement enables the AI storage solution to reach 98% efficiency relative to the maximum theoretical throughput of 314.8 GiB/s.

- Spectrum-X Ethernet maintains superior performance under multi-tenant contention. In noisy-neighbor testing with up to three concurrent elephant flows competing for ISL bandwidth, Spectrum-X consistently outperformed off-the-shelf Ethernet by 8–17%. This demonstrates stronger traffic management and QoS enforcement even when contention occurs at both the network fabric and client NIC layers—critical for shared AI infrastructures where cross-GPU training and storage workloads compete for resources.

Customers can accelerate their AI infrastructure deployment with this validated solution featuring Supermicro Petascale servers, IBM Storage Scale ECE, and NVIDIA Spectrum-X Ethernet. Leverage this pre-tested leaf-spine reference design that delivers superior multi-tenant performance and throughput while ensuring predictable QoS for your production AI workloads.

High Performance AI Storage Network Infrastructure (Featured Products)

In the race to train larger AI models, organizations are discovering that their most expensive investment—GPU compute power—often sits idle, waiting for data. The culprit isn't inadequate processing capability, but storage infrastructure that can't keep pace with modern AI workloads. From training large language models on massive datasets to managing continuous checkpointing, batch loading, and logging, every aspect of the AI training pipeline depends on storage that can deliver data at GPU speed. When storage lags, the consequences ripple through operations: training times stretch from days to weeks, computational costs spiral upward, and the ability to experiment with new models grinds to a halt. In AI inference scenarios, storage performance directly affects critical workflows such as Retrieval Augmented Generation (RAG) and Key-Value cache efficiency.

Yet achieving fast storage isn't simply a matter of deploying high-performance drives—the network fabric connecting GPU nodes to storage nodes often becomes the ultimate bottleneck. Traditional leaf-spine architectures built on static Equal-Cost

Multi-Path (ECMP) routing struggle with the "elephant flows" that characterize AI data transfers—large, sustained streams moving between storage and compute. These conventional networks achieve only 60% of theoretical throughput, with traffic imbalances creating congestion on some paths while others remain underutilized. The result is performance degradation: increased latency disrupts the continuous data flow that distributed training demands, inter-switch congestion throttles bandwidth when needed most, and in shared networks handling both storage and management traffic, competing demands further erode performance.

To address these critical storage and network challenges, IBM, Supermicro, and NVIDIA collaborated to validate a high-performance reference design integrating four proven technologies to eliminate them and unlock the full potential of GPU-accelerated workloads.

Supermicro Petascale Storage Servers

The reference design leverages 1U Supermicro Petascale servers (Figure 1) featuring symmetrically balanced I/O architecture. Each node uses a single AMD EPYC CPU with 128 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.

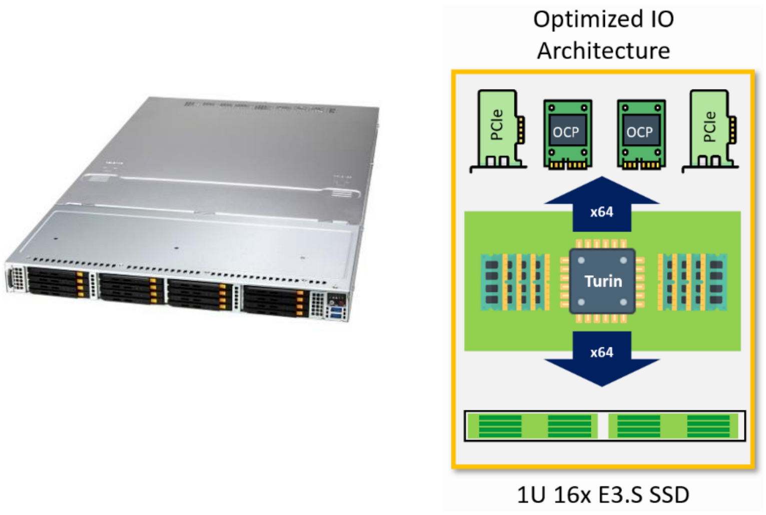

Figure 1 - Supermicro Petascale Storage Server and Internal I/O Architecture

Micron 9550 NVMe Storage

Figure 2 - Micro 9550 PRO NVMe Drives

Eight Supermicro Petascale Storage servers were used, each populated with 16 Micron 9550 NVMe drives in E3.S form factor (Figure 2). Purpose-built for AI and data center workloads, the Micron 9550 delivers industry-leading performance with capacities up to 30.72TB per drive. The drives feature a PCIe Gen5 interface with NVMe protocol and Micron 3D NAND technology, providing high throughput, IOPS, and low latency with consistent quality of service. Enterprise-grade reliability is ensured through cyclic redundancy checks, end-to-end data-path protection, and capacitor-backed power-loss protection. Advanced monitoring capabilities include thermal sensors, SMART attributes for status polling, and SMBus support for out-of-band management.

The Micron 9550 SSD Series offers two endurance classes:

- PRO for read-centric use up to 1 DWPD (Drive Write Per Day)
- MAX for mixed-use workloads up to 3 DWPD

NVIDIA Spectrum-X Ethernet Switches

Storage networking performance is critical, and this design utilizes three NVIDIA Spectrum-X SN5600 Ethernet switches (Figure 3) to connect all storage and compute nodes. The fifth-generation Spectrum SN5000 series delivers port speeds from 10 to 800 Gb/s and is specifically designed to accelerate data center fabrics. Composed of SN5600/5610 Spectrum-X Ethernet switches, BlueField-3 DPUs, and SuperNICs, 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.

Figure 3 - NVIDIA Spectrum-X Ethernet Switch

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.

Each storage server and GPU node was equipped with a single NVIDIA BlueField-3 SuperNIC (Figure 4), designed specifically for AI workloads. The BlueField-3, a key component of the Spectrum-X Ethernet architecture, features dual 200 Gb/s Ethernet ports, 8 ARM cores, and 16 GB of onboard memory, providing accelerated networking optimized for GPU-accelerated systems. Storage nodes are connected via a 400 Gb/s backend. Advanced capabilities include best-in-class RoCE (RDMA over Converged Ethernet) networking, high-speed out-of-order packet reordering, and programmable congestion control, making it the most optimized NIC for AI computing environments.

Figure 4 - NVIDIA BlueField-3 SuperNIC

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, offering 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 hundreds of Supermicro Petascale nodes, providing a cost-optimized, scalable, and resilient platform for AI. With policy-based data management, data is automatically tiered across various storage media 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 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, thereby maximizing GPU utilization for AI training and inference workloads.
- Scalability without redesign: Grow from terabytes to exabytes seamlessly as model sizes and dataset volumes expand, without architectural overhauls

Benchmark Environment

The benchmark environment for testing the performance of this reference design consists of 8 Supermicro Petascale storage nodes running IBM Storage Scale, connected via three NVIDIA Spectrum-X Ethernet switches in a spine-leaf topology, with 400 Gb backend and 200 Gb client-side networking. Five heterogeneous client nodes were used to generate I/O traffic patterns for our testing (Figure 5)

Benchmark Configuration

- 3 NVIDIA Spectrum-X Ethernet SN5600 switches: Spine/Leaf Configuration
- 8 Supermicro Petascale ASG-1115S-NE316R Storage Nodes
    - 16x Micron 9550 E3 7.68TB drives
    - 1x NVIDIA BlueField-3 B3220L E-Series Dual port 200Gb (Client-side)
    - 1x NVIDIA ConnectX-7 NIC 400 Gb (Storage Backend)
    - AMD-EPYC™ 9535 CPU
- 5x Client Heterogeneous client nodes

    - 4x AMD compute client nodes (AMD EPYC 9455/9624)
    - 1x Intel 8U GPU System with NVIDIA HGX H100/H200 8-GPU (Intel 8592 CPUs)
    - 2x Dual port 200 Gb BlueField-3 (B3220) per client (4 connections per server)
- IBM Storage Scale ECE version 5.2.3.1 with RHEL v9.2 or above
    - IBM Storage Scale ECE 5.2.3.1 deployed on all 8 Supermicro Petascale nodes, in 8+2P RAID setting
    - IBM Storage Scale 5.2.3.1 client installed on all 5 Supermicro client nodes

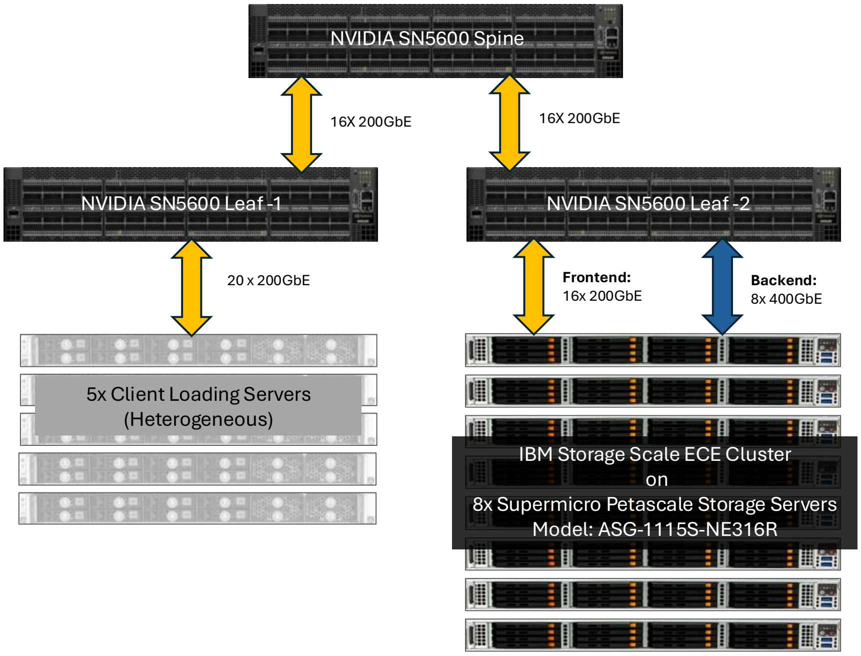

Figure 5 - "Real World" Spine-Leaf Benchmark Setup Topology

Performance Validation Testing & Analysis

To validate the integrated performance of IBM Storage Scale ECE on Supermicro Petascale storage servers connected via NVIDIA Spectrum-X Ethernet, we use FIO (Flexible I/O Tester) configured to emulate realistic AI training workloads. We ran distributed FIO tests from multiple client nodes simultaneously to stress-test the entire stack—this validates not just raw storage performance but also how well Spectrum-X Ethernet handles concurrent high-bandwidth flows and whether IBM Storage Scale ECE on Supermicro Petascale maintains consistent performance under multi-client load.

Preliminary Baseline Test:

Before conducting leaf-spine performance tests, we established a performance baseline by simplifying the network to a single switch configuration for single-hop latency (Figure 6). FIO was configured to perform large-block (16MB) sequential reads to simulate dataset streaming during training and checkpointing epochs. Using five client nodes with incrementally scaled workloads up to 64 jobs per node, the storage solution sustained aggregate throughput of 314.8 GiB/s across all clients, establishing our maximum baseline for subsequent comparisons.

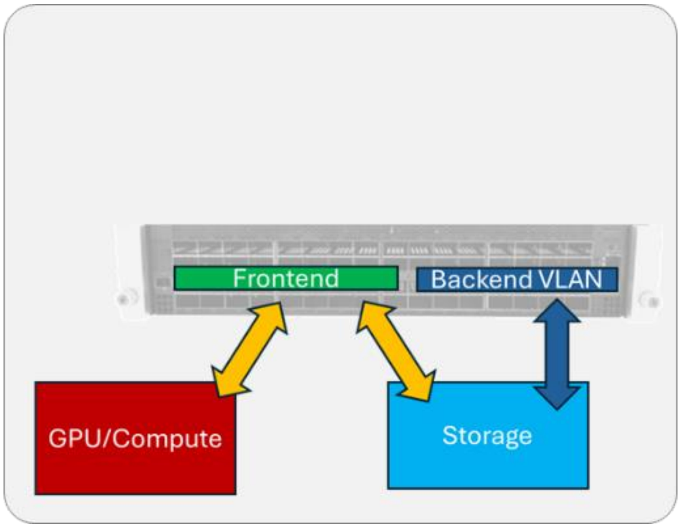

Figure 6 - Baseline Testing with a Single Switch

Real-World Spine-Leaf Testing:

The production testing utilized a spine-leaf network topology (Figure 7) to simulate realistic data center deployments. The storage backend was isolated on its own subnet during initial testing, though retrospective analysis determined this isolation did not contribute to overall performance and is not required in general practice.

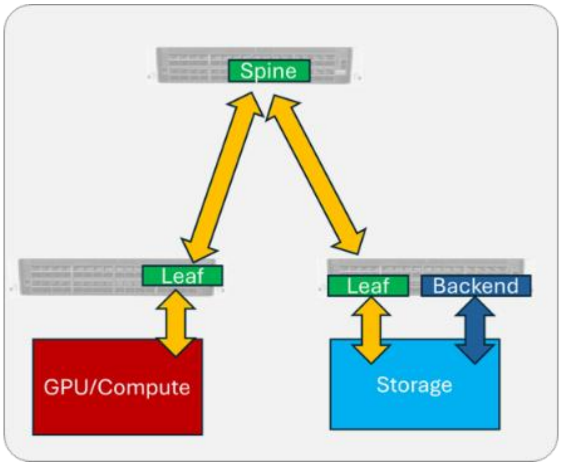

Figure 7 – "Real World Based Testing with Spine and Leaf Switches

Benchmark Test 1: Spectrum-X Ethernet Performance Impact

With the baseline established, we implemented a leaf-spine network topology and tested performance with Spectrum-X Ethernet adaptive routing and congestion control enabled and disabled (Figure 8). The blue dashed line represents the maximum bandwidth achieved when all storage and client nodes are connected to a single switch (314.8 GiB/s baseline). The red line shows the three-switch spine-leaf architecture using standard RoCE over standard off-the-shelf Ethernet without Spectrum-X Ethernet. The green line demonstrates the same three-switch spine-leaf topology with Spectrum-X Ethernet enabled.

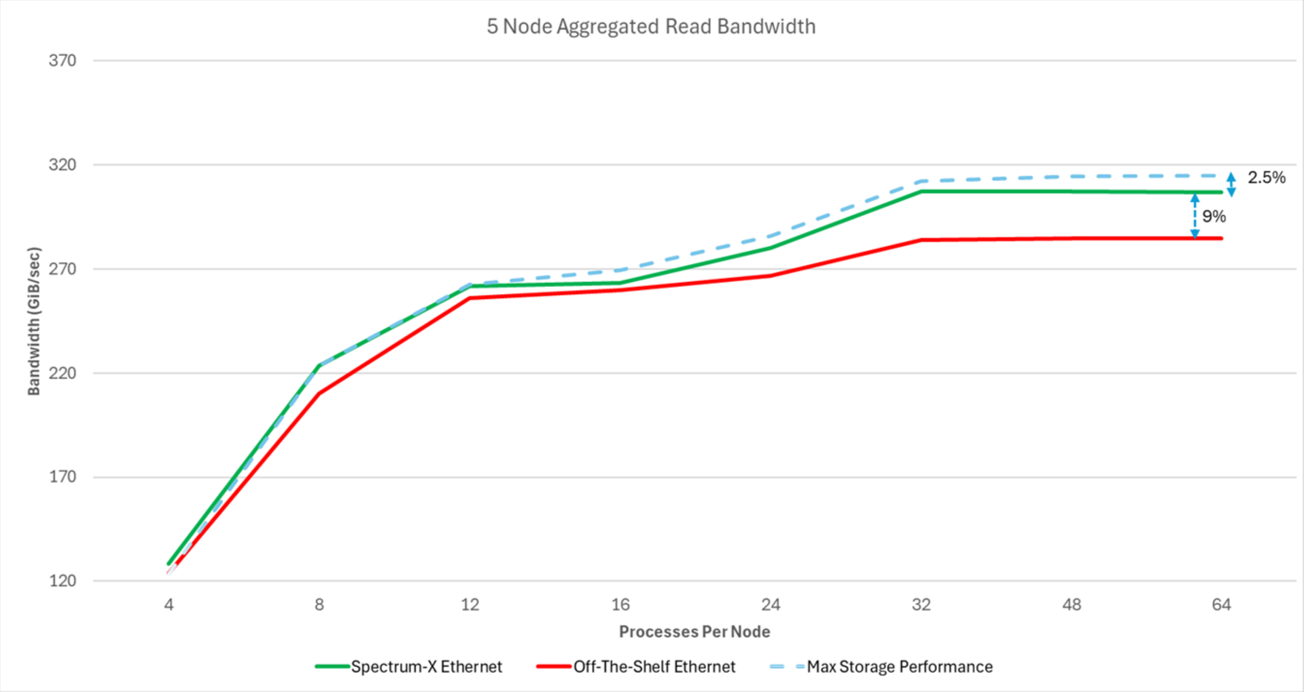

Figure 8 - Aggregated FIO Read Test using 16 MB Black Size

The results show that Spectrum-X Ethernet with adaptive routing brings spine-leaf performance very close to single-switch baseline levels, effectively eliminating the multi-hop penalty typically associated with spine-leaf architectures. Simply enabling Spectrum-X Ethernet in the spine-leaf configuration delivered a 9% improvement in storage performance over RoCE on standard off-the-shelf Ethernet, demonstrating measurable bandwidth gains enabled by Spectrum-X Ethernet's intelligent traffic management and congestion control.

Benchmark Test 2: Multi-Tenant Performance Isolation

This test validates whether IBM Storage Scale's QoS policies and Spectrum-X Ethernet congestion control can isolate performance in multi-tenant scenarios. We ran high-priority sequential read workloads on all clients while simultaneously launching random I/O workloads from other clients, simulating noisy neighbor conditions in which one tenant's workload could degrade another's throughput.

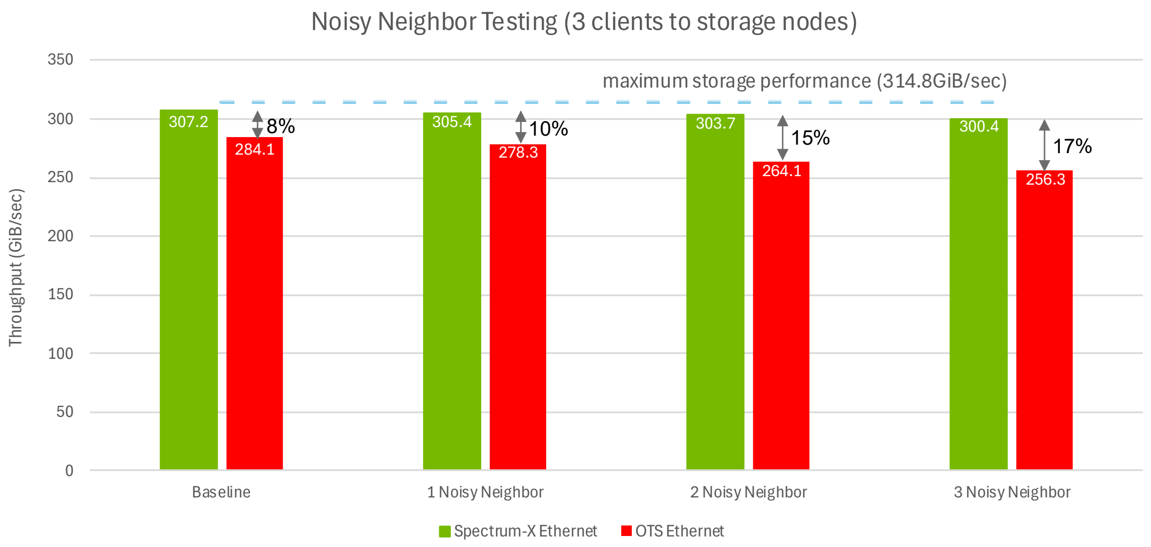

Figure 9 - Noisy Neighbors Testing using 4MB "noise" Packets

The noisy neighbor test (Figure 9) demonstrates the impact of elephant flows—large data transfers running across the same inter-switch links (ISLs) as storage traffic. These flows represent realistic scenarios such as cross-GPU AI training traffic or other workloads sharing the same network fabric. Testing methodology began with a baseline measurement with no additional traffic (matching the peak performance from the previous test), then progressively introduced interference by using 1, 2, and then three client nodes to generate competing traffic flows (noise) consisting of 4 MB packets. This approach created contention not only on the ISLs but also on the client NIC itself, providing a realistic worst-case scenario for performance isolation testing.

Solution Key Insights and findings

- Baseline Performance Advantage: Spectrum-X Ethernet with IBM Scale ECE on Supermicro Petascale storage delivers measurably better performance than standard RoCE configurations even under baseline conditions with no competing traffic, demonstrating inherent efficiency gains from adaptive routing and optimized traffic management.

- Superior Multi-Tenant Isolation: Spectrum-X Ethernet maintains significantly more stable performance under noisy neighbor scenarios, with minimal degradation observed at 2-3 concurrent competing workloads compared to substantial drops versus using off-the-shelf Ethernet configurations. This validates the effectiveness of Spectrum-X Ethernet's congestion control in isolating tenant workloads and maintaining consistent QoS.

- Physical Bandwidth Ceiling: Performance degradation becomes pronounced across all configurations once 4-5 competing data flows are introduced, as this exhausts available ISL bandwidth regardless of congestion control mechanisms. This represents the physical bandwidth limit of the network fabric, beyond which software optimization cannot compensate for saturated links.

Recommended IBM Storage Scale ECE Configuration Selection Guide

	POC	*Small*	*Medium*	*Large*	*x-Large*
Use Cases	Dev & test environments, Proof-of-concept deployments	Small production tier-2 workload system	Mid-sized AI/ML projects, multi-tenant environments for Inference	Large enterprise multi-tenant tier-1 AI production workloads	Large AI factories, training at scale, mixed training, fine-tuning, and inference
IBM Storage Scale ECE on Supermicro Petascale node count	3 nodes	6 nodes	8 nodes (configuration used for this benchmark)	10 nodes+	18 nodes+
Recommended Erasure Code	3 Way Replication	4+2P	4+2P	8+2P	16+2P
Usable Capacity	33%	67%	67%	80%	89%

Note – in addition, a small VM or server is required to host the management node and UI
https://www.ibm.com/docs/en/storage-scale-ece/5.2.3?topic=selection-data-protection-storage-utilization

Recommended configuration for each Supermicro Petascale Storage Server node:

Model	ASG-1115S-NE316R
CPU	PSE-SMPTUR9535-64C300W – AMD Turin CPU
RAM	12x MEM-DR564L-CL01-ER64 - 64GB MICRON Memory 6400 MT/S
STORAGE	16x MICRON 9550 E3.S 7.68 TB - HDS-SMP-MTFDLBQ7T6THA1BK 1x Micron 7450 480GB M.2 Boot Drive
NETWORK	1x BlueField-3 - B3220L with 2x 200G ports for Frontend Network 1x ConnectX-7 NIC - MCX75310AAS-NEAT with 1x 400G port for Backend Network 1x 10G AIOM NIC Card for Management Network

Summary

This solution demonstrates that combining IBM Storage Scale ECE with Supermicro Petascale servers and NVIDIA Spectrum-X Ethernet delivers exceptional performance for demanding, multi-tenant AI workloads. The validated reference architecture achieved 307.2 GiB/s aggregate read bandwidth, reaching 98% of the 314.8 GiB/s theoretical maximum. Spectrum-X Ethernet delivered a consistent 8–17% performance advantage over RoCE, even under challenging noisy-neighbor conditions. Testing confirmed that Spectrum-X adaptive routing effectively eliminates the multi-hop penalty inherent to spine-leaf architectures while maintaining strong QoS isolation across competing tenant workloads.

By leveraging this pre-validated leaf-spine design with proven erasure coding configurations, organizations can eliminate GPU idle time from storage bottlenecks, maintain predictable QoS across multiple teams and projects, and scale seamlessly from development to production without architectural redesign. Ready to accelerate your AI infrastructure deployment? Contact sales@supermicro.com to discuss how this validated solution can be tailored to your specific workload requirements and capacity needs.

References

IBM Storage Scale:

- IBM Storage Scale Product Overview: https://www.ibm.com/products/storage-scale
- IBM Storage Scale ECE Documentation: https://www.ibm.com/docs/en/storage-scale-ece/5.2.3
- Data Protection and Storage Utilization Guide: https://www.ibm.com/docs/en/storage-scale-ece/5.2.3?topic=selection-data-protection-storage-utilization

Supermicro:

- Supermicro Petascale Storage Servers: https://www.supermicro.com/en/products/storage
- Supermicro Storage Partner Solutions: : Supermicro Software-Defined Storage and Memory Solutions

Micron:

- Micron 9550 NVMe SSD Product Page: https://www.micron.com/products/storage/ssd/data-center-ssd/9550-ssd
- Micron Data Center Solutions: https://www.micron.com/products/storage/data-center-storage

NVIDIA:

- NVIDIA Spectrum-X Ethernet Platform: https://www.nvidia.com/en-us/networking/spectrumx/
- NVIDIA BlueField-3 DPU: https://www.nvidia.com/en-us/networking/products/data-processing-unit/
- NVIDIA GPUDirect Storage: https://developer.nvidia.com/gpudirect-storage

Appendix - Key Configuration Settings

This appendix documents the critical configuration parameters used throughout the benchmark testing to ensure reproducibility and provide a reference for future deployment.

Config Parameter	*Recommended Setting*	*Note*
verbsRdma	enable	Enable RDMA on Scale ECE
verbsRdmaSend	yes	Enable RDMA for small messages
verbsRdmaCm	enable	Enable Connection Manager for RoCE
verbsPorts	mlx5_0/1/1 mlx5_1/1/2 mlx5_2/1/3	Configure BlueField-3 and CX7 ports for RDMA

IBM Storage Scale ECE Settings:

- File System Block Size: 16 MB
- Data Protection Setting: 8+2p + 4 spare drives for approximately 78% usable capacity. IBM Storage Scale ECE sizing utilities are available for exact capacity.

IBM Storage Scale ECE Client Settings:

Config Parameter	***Recommended Setting***	***Note***
verbsRdma	enable	Enable RDMA on Scale ECE
verbsRdmaSend	yes	Enable RDMA for small messages
verbsRdmaCm	enable	Enable Connection Manager for RoCE
verbsPorts	mlx5_0/1/1 mlx5_1/1/2 mlx5_2/1/3	Configure BlueField-3 and CX7 ports for RDMA
maxMBpS	80000	Set buffers for the expected network bandwidth
ignorePrefetchLUNCount	yes	Best practice prefetch algorithm for Scale ECE
maxFilesToCache	1M	Increase cached file count
maxStatCache	1M	Increase stat cache
numaMemoryInterleave	yes	Best practices for systems with multiple NUMA domains
nBucketGroups	1024	Performance enhancement
prefetchPct	50	Increase prefetch buffer to 50% of the page pool
fsyncIsGlobal	no	NFS-style fsync
workerThreads	1024	Increase thread count for high-core systems
pagepool	128GB	Increase page pool to 128 GB

Linux Tunables:

The following settings were made across all client and storage nodes for best performance in a low-latency RoCE environment:

Reduce low-power CPU states

- tuned-adm profile latency-performance
- cpupower idle-set -D 0
- cpupower frequency-set -g performance

Sysctl settings for RoCE

- net.ipv4.tcp_ecn=1

RoCE on Off-The-Shelf Ethernet Settings:

The following settings were configured for each adapter for all testing without NVIDIA Spectrum-X to configure standard RoCE, per best practices:

```
/usr/bin/mlnx_qos -i ADAPTER_NAME --pfc=0,0,0,1,0,0,0,0 --trust=dscp
```

```
/usr/sbin/cma_roce_tos –D ADAPTER_RDMA_DEVICE -t 96
```

```
/usr/bin/echo 96 > /sys/class/infiniband/ADAPTER_RDMA_DEVICE/tc/1/traffic_class
```

NVIDIA Spectrum-X Ethernet Settings:

For all NVIDIA Spectrum-X Ethernet testing, NVIDIA provided tools were used to configure hosts with the appropriate settings. Adaptive routing was enabled on all switches (nv set router adaptive-routing enable on).

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements. Visit www.supermicro.com

NVIDIA

NVIDIA (NASDAQ: NVDA) is the world leader in AI and accelerated computing.

Visit www.nvidia.com

IBM

IBM is a leading global hybrid cloud and AI, and business services provider, helping clients in more than 175 countries capitalize on insights from their data, streamline business processes, reduce costs and gain the competitive edge in their industries. Thousands of governments and corporate entities in critical infrastructure areas such as financial services, telecommunications and healthcare rely on IBM's hybrid cloud platform and Red Hat OpenShift to affect their digital transformations quickly, efficiently and securely. IBM's breakthrough innovations in AI, quantum computing, industry-specific cloud solutions and business services deliver open and flexible options to our clients. All of this is backed by IBM's legendary commitment to trust, transparency, responsibility, inclusivity and service. Visit www.ibm.com