QSAN Official Document

XN5226D Performance in AI Infrastructure

Lab Report

August 2026

ANNOUNCEMENT

Copyright

QSAN believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

Trademarks

- QSAN, the QSAN logo, and QSAN.com are trademarks or registered trademarks of QSAN Technology, Inc.
- Microsoft, Windows, Windows Server, and Hyper-V are trademarks or registered trademarks of Microsoft Corporation in the United States and/or other countries.
- Linux is a trademark of Linus Torvalds in the United States and/or other countries.
- UNIX is a registered trademark of The Open Group in the United States and other countries.
- Mac and OS X are trademarks of Apple Inc., registered in the U.S. and other countries.
- Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
- VMware, ESXi, and vSphere are registered trademarks or trademarks of VMware, Inc. in the United States and/or other countries.
- Citrix and Xen are registered trademarks or trademarks of Citrix Systems, Inc. in the United States and/or other countries.
- Other trademarks and trade names used in this document to refer to either the entities claiming the marks and names or their products are the property of their respective owners.

TABLE OF CONTENTS

FIGURES

TABLES

NOTICES

Information contained in this document has been reviewed for accuracy. But it could include typographical errors or technical inaccuracies. Changes are made to the document periodically. These changes will be incorporated in new editions of the publication. QSAN may make improvements or changes in the products. All features, functionality, and product specifications are subject to change without prior notice or obligation. All statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

PREFACE

Technical Support

Do you have any questions or need help trouble-shooting a problem? Please contact QSAN Support, we will reply to you as soon as possible.

- Via the Web: https://www.qsan.com/technical_support
- Via Telephone: +886-2-77206355
- (Service hours: 09:30 - 18:00, Monday - Friday, UTC+8)
- Via Skype Chat, Skype ID: qsan.support
- (Service hours: 09:30 - 02:00, Monday - Friday, UTC+8, Summer time: 09:30 - 01:00)
- Via Email: support@qsan.com

Information, Tip, and Caution

This document uses the following symbols to draw attention to important safety and operational information.

INFORMATION

INFORMATION provides useful knowledge, definition, or terminology for reference.

TIP

TIP provides helpful suggestions for performing tasks more effectively.

CAUTION

CAUTION indicates that failure to take a specified action could result in damage to the system.

1. INTRODUCTION

Modern AI infrastructure faces structural challenges. While GPU servers are optimized for computation, they are often required to act as storage nodes, managing local NVMe SSDs, running NFS servers, calculating RAID parity, and providing data services to other users. AI servers, handling both tasks simultaneously, often perform poorly in both areas.

This lab report validates a simpler architecture where the AI server is dedicated to computation, while the QSAN XN5226D handles all data processing. The two systems are connected via four 100G direct-connect RDMA links, providing sufficient bandwidth to ensure transparent NAS operation during training while delivering faster performance than local systems for checkpointing and inference workloads.

The QSAN XN5226D is a 2U 26-bay all-NVMe unified storage system equipped with a dual-active controller. It simultaneously supports NVMe-oF over RDMA (RoCE), NFS-RDMA, iSCSI, SMB, and S3. serving as a single data center for AI workloads, other departments, and backups, eliminating the need for separate storage infrastructure configurations for each use case. Testing was conducted on an Altos BrainSphere R785 F6 AI server equipped with a single-block NVIDIA RTX PRO 6000 Blackwell Server Edition GPU. Benchmarks included GPUDirect Storage Validation (GDSIO), AI Training (MLPerf Storage DLIO), Model Checkpointing (MLPerf Storage Checkpoint), LLM Inference Launch (vLLM), and Concurrent Multiprotocol Throughput (qs3d).

2. PERFORMANCE DATA

2.1. Environment and Topology

Test Environment

- AI Server
  - Model: Altos BrainSphere R785 F6
    CPU: 2 × AMD EPYC 9335 (128 cores / 256 threads total)
    Memory: 4 × Samsung 6400 32GB DDR5 (128 GB total)
    GPU: 1 x NVIDIA RTX PRO 6000 Blackwell Server Edition
    GPU VRAM: 95.6 GiB
    GPU TDP: 600W
    GPU Driver: 610.43.02 / CUDA 13.3
    Host NIC: 2 × NVIDIA ConnectX-7 200G
    Local HDD: 1 x Samsung MZ1L21T9HCLS-00A07
    OS: Ubuntu 22.04
- Storage
  - Model: XN5226D-12C
    CPU: 2 x Intel Xeon 12-core
    Memory: 16 GB per controller
    Disk: 26 x QSAN SD4 1.6 TB NVMe SSD
  - Pool: RAID 6 with 26 x NVMe SSD
    Host Card: 4 x 100 GbE with RDMA
    Protocol: NFSv4.2 / RDMA, 4 paths aggregated
- Software
  - GPUDirect Storage: Enabled, cuFile SDK (XferType=GPUD validated)
    vLLM Version: 24.0
    MLPerf Storage: 3.0.26 (DLIO benchmark)
    CUDA: 13.3
    PyTorch: Latest compatible with CUDA 13.3

Test Topology

The AI server and XN5226D are connected via four 100G direct-connect RDMA links, with no switches in the data path. Each NVIDIA ConnectX-7 200G network card provides two 100G QSFP28 ports, forming four independent transmission paths. Both NVMe-oF / RDMA (NVMe namespace access) and NFS-RDMA (NFSv4.2 file system access) can operate simultaneously on this architecture.

Figure 2-1 Test Rack

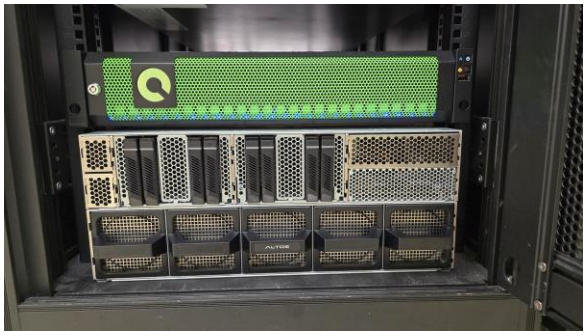

Figure 2-2 XN5226D Front View

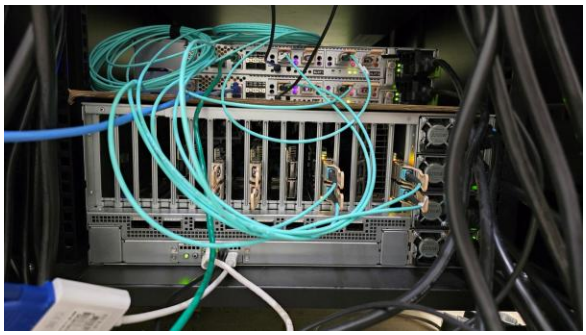

Figure 2-3 XN5226D Rear View and Wiring Diagram

2.2. Test Methodology

2.2.1. GPUDirect Storage Validation

GDSIO is a proprietary benchmarking tool developed by NVIDIA, specifically designed to evaluate and measure storage performance in GDS (GPUDirect Storage) environments. GDSIO benchmarks NAS paths with GPUDirect Storage enabled using NVIDIA's cuFile SDK. A mandatory validation criterion is XferType=GPUD for each path to confirm that data is transferred directly from the NAS to the GPU memory without CPU buffering. Each path uses one large file during testing. Sequential throughput tests use 1M I/O blocks; random IOPS tests use 4K blocks. All four NFS-RDMA paths are run simultaneously, and the results are aggregated.

2.2.2. AI Training

MLPerf Storage DLIO benchmarks the pipeline for loading stored data during deep learning training. The primary metric is AU (Accelerator Utilization), which is the ratio of GPU computation time to data waiting time. The AU ≥ 99% indicates that storage is not a bottleneck.

- Test Tool
  - MLPerf Storage 3.0.26 (DLIO)
    Model: RetinaNet
    Dataset 1,828,596 JPEG files, 550.88 GiB total
    Epochs: 1
    Host Memory Limit: 32 GiB (determines minimum file count)

Accelerator: 1 × RTX PRO 6000 Blackwell SE

2.2.3. Model Checkpoint

MLPerf storage checkpoint benchmark tests the save and restore times of checkpoints. Write (save) and read (restore) throughput are measured against the same 113.9 GiB payload.

- Test Tool
  - MLPerf Storage 3.0.26 (Checkpoint)
    Model: Llama3-70B (subset checkpoint)
    Checkpoint Size: ≈ 113.9 GiB (model weights + optimizer state)
    Ranks: 8

2.2.4. LLM Inference Startup

A vLLM host cold start clears the Linux page cache and forces a full model to reload from storage before each server startup. We measure model load time (safetensor weight loading) and server readiness time (model + CUDA graph compilation). Additionally, we measure post-load service throughput as a comparison to confirm that storage does not impact inference performance.

- Test Tool
  - vLLM 0.24.0
    Model: Qwen2.5-32B-Instruct (BF16)
    Model Size: ≈ 61 GiB (17 safetensor shards)
    GPU Memory Utilization: 85% (configured limit)
    Tensor Parallelism: TP=1 (single GPU)
    Measurement: Host-cold startup (page cache dropped before each run)
    Runs per Path: 3 (median reported)

2.2.5. Concurrent Multiprotocol Throughput

qs3d benchmarks simultaneous S3 + NFS access on a single XN5226D system to verify the effectiveness of the multiprotocol hub model under real-world concurrent loads from various client types.

2.3. Benchmark Results

2.3.1. GDSIO Result

GPUDirect Storage is active and all GDSIO paths returned XferType=GPUD. Data moves directly from XN5226D to GPU VRAM without CPU involvement. The following are the test results.

Sequential Throughput (1M I/O, NFS-RDMA, 4 paths aggregated)

Table 2-1 GDSIO Sequential Throughput

PATTERN	THREADS / PATH	AGGREGATE GIB/S	GDS STATUS
Seq Read, 1 thread	1	4.256	PASS (GPUD)
Seq Read, 2 threads	2	8.257	PASS (GPUD)
Seq Read, 4 threads (peak)	4	11.698	PASS (GPUD)
Seq Write, 2 threads	2	3.511	PASS (GPUD)
Seq Write, 4 threads (peak)	4	5.602	PASS (GPUD)

Summary

GDSIO sequential throughput: NFS-RDMA, 4 paths aggregated. Peak: 11.698 GiB/s read, 5.602 GiB/s

Random IOPS (4K I/O, NFS-RDMA, 4 paths aggregated)

Table 2-2 GDSIO Random IOPS

PATTERN	THREADS / PATH	AGGREGATE GIB/S	GDS STATUS
Rand Read, 32 threads	32	140,903	PASS (GPUD)
Rand Read, 64 threads (peak)	64	234,480	PASS (GPUD)
Rand Write, 32 threads	32	80,503	PASS (GPUD)
Rand Write, 64 threads (peak)	64	105,086	PASS (GPUD)

Summary

GDSIO random IOPS: NFS-RDMA, 4 paths aggregated. Peak: 234,480 IOPS read, 105,086 IOPS write

2.3.2. AI Training Test Result

The accelerator utilization was statistically identical across all three storage paths, confirming that using the XN5226D via RDMA is not a training bottleneck. GPU computation time accounted for over 99.79% of all configurations.

MLPerf Storage DLIO Training Results

Table 2-3 AI Training Test Result

METRIC	LOCAL NVME	NVME-OF / RDMA	NFS-RDMA
Accelerator Utilization (AU)	99.79%	99.80%	99.83%
Training Throughput	500.81 samp/s	501.04 samp/s	502.61 samp/s

Summary

MLPerf Storage DLIO training results: RetinaNet, 550.88 GiB dataset, 1 epoch. All paths: AU > 99.79%. Differences within measurement noise

2.3.3. Checkpoint Test Result

Checkpoint Write (Save) with 113.9 GiB, 8 Ranks

Table 2-4 Checkpoint Test Result

METRIC	LOCAL NVME	NVME-OF / RDMA	NFS-RDMA
Duration	69.632 s	23.121 s ★	41.743 s
Throughput	1.636 GiB/s	4.926 GiB/s ★	2.729 GiB/s
vs. Local	—	3.0× faster	1.7× faster

Summary

MLPerf Checkpoint write (save): 2026-07-01. ★ NVMe-oF/RDMA saves 3× faster than local NVMe

NVMe-oF / RDMA completed a 113.9 GiB checkpoint save in just 23.1 seconds, three times faster than the 69.6 seconds of local NVMe. This seemingly counterintuitive result can be explained by the RAID6 write overhead on the local path: the host serializes parity calculations and write operations onto a single NVMe drive. NVMe-oF/RDMA offloads this operation to the XN5226D's dual controllers, distributing write jobs across 26 drives via RDMA, thus eliminating the host-side parity bottleneck.

Checkpoint Read (Restore) with 113.9 GiB, 8 Ranks

Table 2-5 AI Training Test Result

METRIC	LOCAL NVME	NVME-OF / RDMA	NFS-RDMA
Duration	69.318 s	10.819 s ★	10.743 s ★
Throughput	1.643 GiB/s	10.528 GiB/s	10.603 GiB/s
vs. Local	—	6.4× faster	6.5× faster

Summary

MLPerf Checkpoint read (restore): 2026-07-01. Both NAS paths restore 6× faster than local NVMe

Both NVMe-oF / RDMA and NFS-RDMA can recover checkpoints in approximately 10.8 seconds (approximately 10.5 GiB/s), more than six times faster than the 69.3 seconds of native NVMe. The XN5226D's 26 drives are distributed across dual active controllers, providing aggregated sequential read bandwidth far exceeding that of a single native NVMe SSD, thereby directly reducing latency for training restarts after a failure.

2.3.4. LLM Inference Test Result

Checkpoint Write (Save) with 113.9 GiB, 8 Ranks

The host cold start result measures the actual cold start model loading process from storage to GPU VRAM, with the page cache being cleared before each run.

Table 2-6 LLM Inference Test Result

ACCESS PATH	MODEL LOAD	VS. LOCAL	SERVER READY	VS. LOCAL	1ST REQ.
Local NVMe	38.346 s	—	69.417 s	—	0.374 s
NVMe-oF/RDMA	16.370 s	−57.3%	50.343 s	−27.5%	0.362 s
NFS-RDMA ★	13.386 s	−65.1%	44.213 s	−36.3%	0.363 s

Summary

vLLM host-cold startup — Qwen2.5-32B-Instruct, BF16, TP=1. Median of 3 runs. ★ NFS-RDMA loads 65% faster than local storage

Loading a 32-bit tuple model (61 GiB) using NFS-RDMA takes only 13.4 seconds, 65% faster than native storage. With the model residing in GPU memory, the inference throughput is the same for all paths (approximately 85 tokens/s at concurrency of 4), confirming that storage performance only affects deployment speed and not inference quality.

2.3.5. Concurrent Multi-Protocol Throughput Test Result

Checkpoint Write (Save) with 113.9 GiB, 8 Ranks

Table 2-7 Concurrent Multi-Protocol Throughput Test Result

TEST	RESULT
S3 + NFS Concurrent Peak Throughput	1,697.51 MiB/s
Protocols Active	S3 + NFSv4.2/RDMA (simultaneous)

Summary

qs3d concurrent S3+NFS peak throughput — XN5226D serving multiple protocol clients simultaneously.

The XN5226D delivers an aggregate throughput of 1,697 MiB/s for both S3 and NFS clients simultaneously, validating the effectiveness of the multi-protocol hub model. In this model, AI pipelines, analytics teams, and backup systems can access the same data simultaneously without dedicated protocol-specific infrastructure.

2.4. Technical Analysis

Why Training is Storage-Independent

MLPerf storage DLIO training workloads use asynchronous prefetching: dataset batches are loaded into host memory in the background, while the GPU computes the current batch. As long as the storage supports the prefetch pipeline, the GPU doesn't wait—this is confirmed by AU > 99.79% across all paths.

4 × 100G NVMe-oF/RDMA maintains a peak rate of 11.7 GiB/s (from GDSIO), significantly exceeding the training data rate. A dataset containing 1.8 million files validates the XN5226D's large-scale metadata processing capabilities. Training results match expected benchmark results—more impactful findings are seen in checkpointing and inference.

Why NVMe-oF / RDMA Checkpoint Write Performance Outperforms Local NVMe

The result of NAS write speeds being 3 times faster than local NVMe may seem counterintuitive, but the mechanism is clear. The native NVMe drive operates in RAID 6 mode. Writing 113.9 GiB of data from 8 ranks of a single native SSD requires the host CPU to calculate RAID 6 parity and serialize the write operation. At this point, the disk and parity calculations become the bottleneck.

NVMe-oF / RDMA transfers data to the XN5226D's dual controllers via RDMA (a CPU-invisible transfer method). The XN5226D's controllers handle RAID 6 parity at the hardware level and distribute the write job in parallel across all 26 disks. The result is that the NAS completes writes faster than the native path, which is limited by host performance, while the AI server CPU remains idle throughout the process.

Why NFS-RDMA Loads LLM Models Faster

Loading 61 GiB of data sequentially from a single local NVMe drive is limited by the actual sequential read bandwidth of that hard drive (in real-world conditions, this transfer size is approximately 4–5 GB/s). The XN5226D performs the same read operation across 26 hard drives via dual controllers, with bandwidth provided by four 100G RDMA connections. Its throughput exceeds that of a single hard drive.

In terms of operation: Switching a 3.2 billion parameter model takes only 13 seconds, instead of the previous 38 seconds. In multi-model or multi-user inference environments with frequent model loading, a 65% reduction in cold start time significantly accelerates deployment cycles.

Centralized Data Center Architecture

Traditional AI server architectures deploy computation and storage in the same location: the AI server owns all local NVMe hard drives, runs NFS servers for other users, and manages RAID. This wastes GPU server CPU resources on I/O services, creating data silos on each AI server and making data access dependent on the AI server's operational status.

Using the XN5226D as a central storage hub: AI-generated output (checkpoints, models, fine-tuned weights, inference logs) is directly stored on the XN5226D via NVMe-oF or NFS-RDMA. Other teams (data scientists, maintenance personnel, backup systems) can access the same data at full speed over NFS/SMB/S3 networks without the AI server's involvement. The AI server then becomes a pure computing node, easier to manage, more available, and more efficient.

2.5. XN5226D Benefits in AI Infrastructure

Benchmark results demonstrate that the XN5226D is a fully functional AI storage center that supports multiple infrastructure roles simultaneously:

- AI Training Dataset Pool: Shared high-throughput NVMe storage for training datasets. Multiple GPU servers can simultaneously access the same 550+ GiB dataset via NVMe-oF/RDMA, eliminating the need for redundant dataset storage between servers.
- Model Checkpoint Storage: Compared to native NVMe under RAID6, checkpoint write speeds are 3x faster and recovery speeds are 6x faster. This directly reduces AI training downtime after failures and lowers checkpoint overhead during long training runs.
- LLM Inference Model Storage: Cold-start model loading speed is 65% faster than local storage. Multiple vLLM or TGI inference nodes can independently load models from the same XN5226D via NFS-RDMA, eliminating the need for an AI server as an intermediary node.
- Multi-protocol Shared Data Center: S3+NFS concurrent access speeds up to 1.7+ GiB/s enable AI teams, data science, analytics, and backup to share a single storage system over a full-speed network via native protocols.
- GPUDirect Storage Architecture: All NFS-RDMA paths are verified with XferType=GPUD. Fully supports zero-copy NAS-to-GPU pipelines using cuFile SDK, enabling future AI storage workflows without hardware changes.

- Centralized Data Protection: RAID6 + dual-active controllers + snapshots + replication mechanisms protect all AI outputs with enterprise-grade reliability. AI servers no longer manage any storage protection.

3. CONCLUSION

This lab report demonstrates that the QSAN XN5226D can achieve an ideal AI infrastructure architecture: the AI server handles computation, and the XN5226D processes all data.

These three benchmark results fully demonstrate the performance of the XN5226D. GPU utilization for AI training reached ≥ 99.79% across all storage paths, with the XN5226D remaining completely transparent during the training process. NVMe-oF / RDMA checkpoint write speeds were 3x faster than local NVMe, and checkpoint recovery speeds were 6x faster, accelerating the most storage-intensive operations. NFS-RDMA loading of 32B LLMs was 65% faster than local storage, with the XN5226D not only accelerating data throughput but also speeding up deployment cycles.

GPUDirect Storage was validated on all NFS-RDMA paths (XferType=GPUD), confirming zero-copy functionality from NAS to GPU. Simultaneous S3+NFS access speeds of 1.7 GiB/s enable the XN5226D to serve as a true multi-protocol data center, serving the entire AI environment, from training pipelines to analytics teams to backup systems, all integrated on a single 2U platform.

The QSAN XN5226D allows AI servers to focus on their true computing core function.

4. APPENDIX

4.1. Reference

Product Page

- XN5 Series Product Page

Data Sheet

- XN5 Series Data Sheet
- SD4 Series Data Sheet
- QSM 4 Data Sheet

Document

- QSM 4 Software Manual