QSAN
Official Document

XN1209-4A Performance with GB10

Lab Report

February 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, Teams / 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. OVERVIEW

1.1. Origins and Challenges

GB10 (Grace Blackwell Superchip) is NVIDIA's latest hybrid Superchip, which integrates a high-performance CPU and GPU for native AI computing systems such as desktops and workstations. The key features and applications are

- **High AI Performance**: Designed for LLM (Large Language Model), generative AI, deep learning inference, and fine-tuning.
- **Unified Memory Design**: The CPU and GPU directly share a large amount of high-speed memory, reducing data transfer latency and improving AI training and inference efficiency.
- **Low Power Consumption and Performance**: Compared to traditional professional GPU solutions, it consumes less power and has higher integration, making it suitable for desktops and workstations.

We might wonder, how can such a powerful AI computing system help SMB? If your business needs on-premises AI and a comprehensive knowledge base, then this may be the best option. How is it implemented? The simple structure is as follows.

GB10 System + **LLM** + **RAG**

RAG (Retrieval-Augmented Generation) is an AI technique that combines LLM with a retrieval mechanism from an external knowledge base. This allows responses to be generated based on up-to-date, domain-specific information, rather than relying solely on the model's pre-trained knowledge. This approach makes AI more accurate, reduces hallucinations, and incorporates private or up-to-date information into responses.

RAG requires significant storage space to store results, thus necessitating external storage devices. Addressing the pain point of insufficient native storage capacity (max. 4TB) in GB10 system. And rapidly expand free storage space to support massive data retrieval and consolidation. If the computing power of a single GB10 is insufficient, multiple GB10 systems can be prepared, with external storage becoming shared data across these systems.

GB10 System + External Storage + LLM + RAG

The actual implementation will utilize GB10 System (such as NVIDIA DGX Spark or its compatible products), paired with the QSAN XN1 series tower storage to construct a complete solution. The specific solution is as follows.

Figure 1-1 On-premises AI Solution

In this lab report, we prepared not only a QSAN XN1 storage device but also a control group using the same hardware but with a general-purpose NAS system to test its performance. Through this test, we conducted an empirical analysis of the performance of the QSAN XN1 and a standard Ubuntu NAS under real-world AI workloads.

1.2. Introduction to QSAN XN1209-4A

Redefines agile and compact desktop tower storage by delivering performance once reserved for compact systems — now in a compact tower form factor for professionals. Equipped with a dedicated 2.5" SSD slot and NVMe for tiered or cache acceleration, it transforms hybrid flash into a powerful I/O engine. With PCIe Gen4 x8 expansion, the XN1 Series supports high-speed host cards to tackle demanding workloads like VDI, virtualization, and 4K media editing.

Figure 1-2 QSAN XN1209-4A Appearance

1.2.1. QSAN XN1209-4A Specifications

Please refer to the XN1 Series Data Sheet. Empowers professional workgroups with high-speed connectivity and enterprise-grade efficiency. Featuring onboard 10 GbE ports and PCIe Gen4 x8 expansion, it supports seamless upgrades to 25 GbE connections. Combined with QSM, QSAN's unified storage operating system, the XN1 delivers streamlined file sharing with multi-protocol access, advanced permission controls, and intuitive management, making it the perfect solution for agile teams that demand without delays.

2. PERFORMANCE DATA

2.1. Performance Report

The QSAN XN1209A integrates GB10, providing lower latency performance compared to the same hardware platform running Ubuntu. This test aims to eliminate hardware variables and focus on comparing the true efficiency of storage software stacks when handling AI workloads. We ensured consistency across all key components for the purest performance evaluation.

Test Equipment and Configurations

- Server
  - Model: Altos BrainSphere GB10 F1
    GPU：NVIDIA Grace Blackwell
    CPU：20 core Arm, 10 Cortex-X925 + 10 Cortex-A725 Arm
    Memory: 128GB LPDDR5x, unified system memory
    Storage: 4 TB NVMe M.2 with self-encryption
    NIC: NVIDIA ConnectX®-7 NIC (200G × 2 QSFP)
    OS: NVIDIA DGX™ OS
    FS: EXT4
- Storage 1
  - Model: QSAN XN1209-4A
    Memory: 16 GB DDR5 ECC DIMM
    Host Card: Intel E810 (25 GbE x 2 SFP28)
    Firmware: 4.1.6
  - SSD: 8 x Sandisk SSD 78 GB
  - Pools:
    1 x RAID 5 with 8 SSDs
  - Volumes:
    1 x 500 GB in Pool
    Block Size: 512 Byte
- Storage 2 (control group)
  - Hardware: Same as QSAN XN1209-4A

  - OS: Ubutu 24.04.3 LTS
  - Volumes: Build an 8-disk RAID5 using LVM2
- I/O Pattern
  - Tool: GDSIO (NVIDIA GPUDirect Storage I/O) 1.16
    runtime: 30
    iodepth: 4
    threads: 1
    IO size: 4K or 128K

Test Scenario

We used the GDSIO tool to build a RAID 5 array containing 8 SSDs for performance testing. The first round of testing was conducted on the local SSDs. The second and third rounds of testing were conducted on storage 1 and storage 2 via the NFS over TCP, respectively. The final two rounds of testing were conducted via the NFS over RDMA.

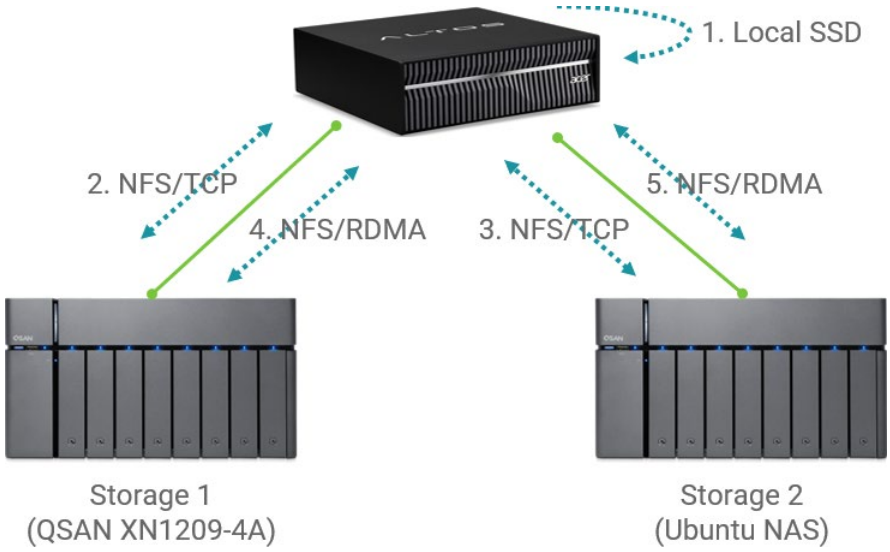


Figure 2-1 Test Architecture Diagram

Performance Results

The chart below reveals the performance report.

Table 2-1 Performance Results

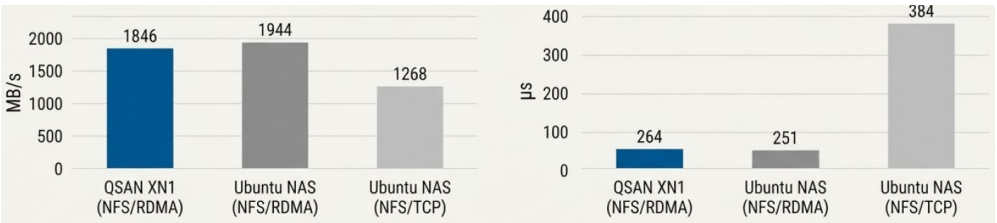
	GB10 LOCAL SSD (EXT4)	STORAGE 1 (QSAN XN1) NFS/TCP	STORAGE 2 (UBUNTU) NFS/TCP	STORAGE 1 (QSAN XN1) NFS/RDMA	STORAGE 2 (UBUNTU) NFS/ RDMA
128K Sequential Read	2,302 MB/s	1,182 MB/s	1,268 MB/s	1,846 MB/s	1,944 MB/s
Latency	212 µs	413 µs	384 µs	264 µs	251 µs
128K Sequential Write	4,996 MB/s	971 MB/s	52 MB/s	537 MB/s	67 MB/s
Latency	97 µs	502 µs	9,237 µs	908 µs	7,262 µs
4K Random Read	1,575K	608K	715K	1,088K	1,201K
Latency	75 µs	194 µs	166 µs	107 µs	98 µs
4K Random Write	4,056K	457K	26K	131K	26K
Latency	29 µs	256 µs	5,208 µs	903 µs	4,742 µs

2.2. Analysis Results

Based on the above test results, the following four analysis and discussion results are presented.

2.2.1. Analysis 1: 128K Sequential Read

In data streaming tasks, the QSAN XN1209A with NFS over RDMA achieved a transfer speed of 1,846 MB/s, while the Ubuntu platform achieved 1,944 MB/s. Both demonstrated excellent performance through RDMA. This also outperformed NFS over TCP at 1,268 MB/s.



Mode	Bandwidth (MB/s)	Latency (µs)
QSAN XN1 – NFS/RDMA	1846	264
Ubuntu NAS – NFS/RDMA	1944	251
Ubuntu NAS – NFS/TCP	1268	384

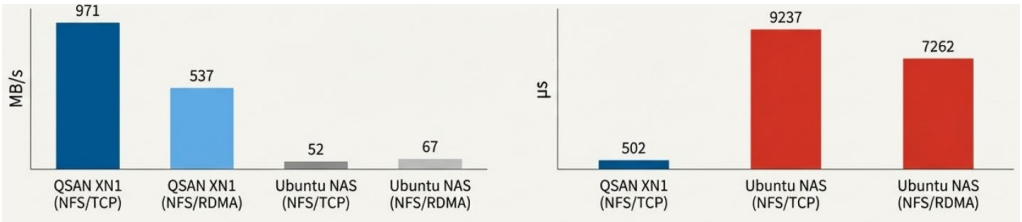
Figure 2-2 128K Sequential Read Analysis Chart

As shown in the figure above, RDMA offers superior performance, while the performance of both is comparable. RDMA delivers a significant improvement, increasing throughput by approximately 45-55% compared to traditional NFS/TCP. The RDMA performance of the QSAN XN1 is very close to the peak performance of the Ubuntu NAS, demonstrating that the network path is optimized and not the bottleneck.

Our Insight: Both are suitable for streaming large AI datasets. This confirms that our test benchmark is fair and objective.

2.2.2. Analysis 2: 128K Sequential Write

In lage-scale write tasks, the QSAN XN1209A with NFS over TCP achieved a transfer speed of 971 MB/s, while the Ubuntu platform achieved only 52 MB/s. A similar disparity also occurs with NFS over DRMA. The performance of general-purpose NAS suffered a catastrophic collapse.



Mode	Bandwidth (MB/s)	Latency (µs)
QSAN XN1 – NFS/TCP	971	502
QSAN XN1 – NFS/RDMA	537	908
Ubuntu NAS – NFS/TCP	52	9237
Ubuntu NAS – NFS/RDMA	67	7262

Figure 2-3 128K Sequential Write Analysis Chart

As shown in the figure above, the write performance differs by 10-18 times, with the QSAN XN1 boasting a write bandwidth of up to 971 MB/s, while the Ubuntu NAS only reaches 50 ~ 70 MB/s. This is not a minor difference, but a significant one. On the other hand, the Ubuntu NAS suffers from latency as high as 7-9 milliseconds, which is disastrous. For LLMs (Large Language Models) that require frequent storage checkpoints, millisecond-level latency can cause GPU pipeline stalls, directly wasting your expensive hardware investment and valuable training time.

Our Insight: The write performance of general-purpose NAS is insufficient for modern AI training. QSAN's optimized storage stack demonstrates a clear advantage here.

2.2.3. Analysis 3: 4K Random Read

In fragmented reading tasks, the QSAN XN1209A with NFS over RDMA achieved a transfer speed of 1,088K IOPS, while the Ubuntu platform achieved 1,201K IOPS. Both parties once again demonstrated excellent IOPS performance when handling the reading of small files and metadata.

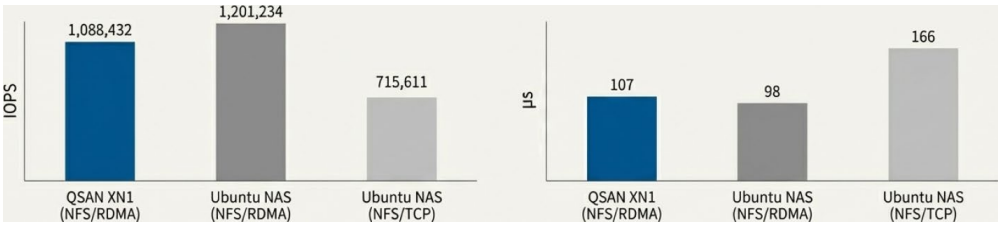


Figure 2-4 4K Random Read Analysis Chart

Our Insight: QSAN XN1 proves that it is not only designed for large file transfers, but also capable of efficiently handling metadata-intensive AI workloads.

2.2.4. Analysis 4: 4K Random Write

In random write tasks, the QSAN XN1209A with NFS over TCP achieved a transfer speed of 457K IOPS, while the Ubuntu platform achieved only 26K IOPS. In the most demanding random write tests, latency issues once again became apparent, directly impacting GPU efficiency. This highlights another fatal weakness of general-purpose NAS.

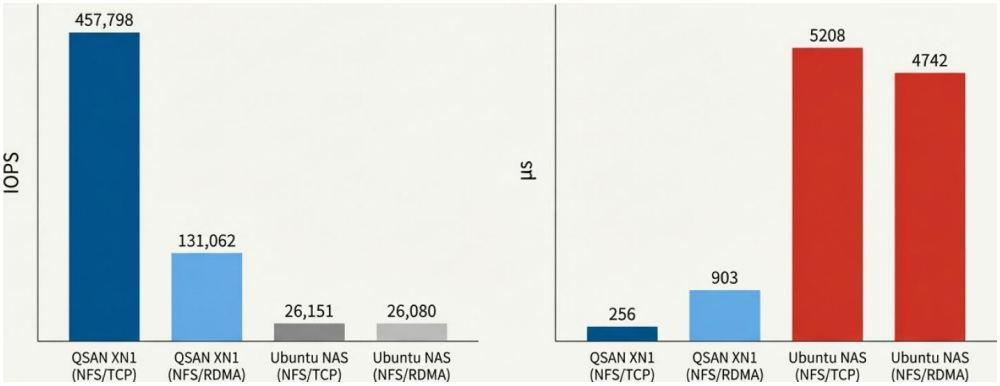


Figure 2-5 4K Random Write Analysis Chart

Latency is the number one killer of GPU performance; latency exceeding 5 milliseconds is unacceptable. The QSAN XN1 demonstrates an overwhelming advantage in random write IOPS, with IOPS performance improvements of up to 5 to 17 times. In contrast, the latency of an Ubuntu NAS can soar to over 5 milliseconds. In AI training, any storage latency exceeding 1 millisecond can cause expensive GPUs to enter an idle state, resulting in a huge waste of computing power.

Our Insight: This perfectly explains why general-purpose Linux NAS performs poorly under AI workloads. Low latency and stability are key to success.

2.3. Performance Differences

Based on the above results, the following table is obtained.

Table 2-2 Final Scorecard of Analysis Results

ANALYSIS 1 SEQUENTIAL READ	ANALYSIS 2 SEQUENTIAL WRITE	ANALYSIS 3 RANDOM READ	ANALYSIS 4 RANDOM WRITE
Draw	QSAN XN1 Win	Draw	QSAN XN1 Win
Both can meet the requirements for loading AI datasets.	General-purpose NAS performance has failed and is not suitable for model checkpoint storage.	Both sides can efficiently handle metadata retrieval.	The extremely high latency of general-purpose NAS can cause the GPU pipeline to stall.

The root cause of performance differences lies in optimized software stacking versus a general-purpose kernel.

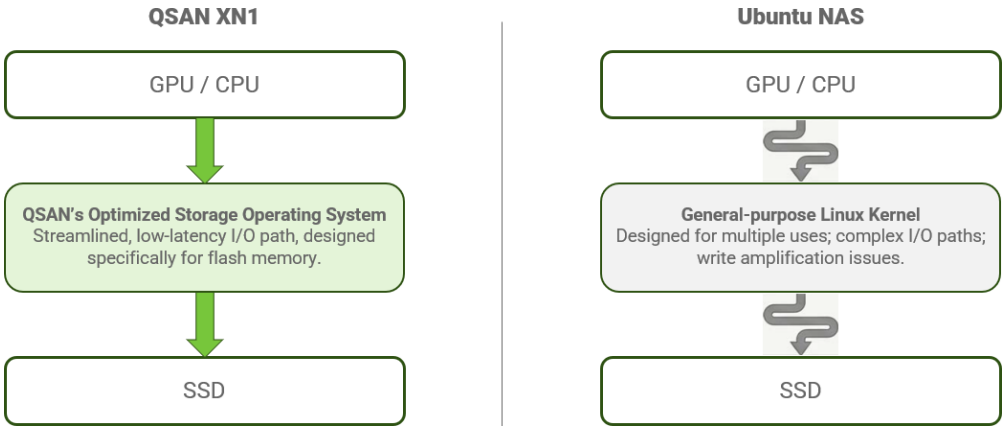


Figure 2-6 Root Cause of Performance Differences

The significant performance difference doesn't stem from the SSD hardware or network, but rather from QSAN's software stack tailored for all-flash storage. It bypasses the bottlenecks inherent in general-purpose operating systems, delivering unparalleled efficiency, particularly when handling write amplification and latency-sensitive operations.

3. CONCLUSION

In AI workloads, write performance is the true bottleneck. For serious AI/machine learning workloads, the key to choosing a storage solution is no longer read speed or capacity, but rather an optimized software stack that delivers stable, low-latency write performance.

QSAN XN1 leverages high-performance I/O with RDMA support to deliver exceptional write performance. In AI projects, it significantly reduces model checkpoint storage time, enabling your team to experiment and fine-tune faster, thus accelerating the innovation cycle. It eliminates GPU idle time caused by storage bottlenecks, ensuring that every bit of hardware investment translates into actual computing power, not unnecessary waiting. Stable, predictable, and low-latency performance means that long-running training tasks will not fail unexpectedly due to storage performance fluctuations, thus ensuring project progress.

This architecture is particularly suitable for enterprise and research environments, such as AI labs, AI teams in the semiconductor and manufacturing industries, and financial and telecommunications institutions developing in-house models. It represents a robust, scalable, and enterprise-grade multi-node LLM fine-tuning infrastructure solution.

4. APPENDIX

4.1. Reference

Product Page

- XN1 Series Product Page

Data Sheet

- XN1 Series Data Sheet
- QSM 4 Data Sheet