AI
BUYER'S GUIDE

INDEX

INTRODUCTION

AI workloads demand ultra-fast data access, high throughput, and flexible deployment. As organisations scale model training, real-time analytics, and distributed operations, storage becomes the core factor shaping performance and responsiveness.

NGX ExaScale is built for this new landscape, delivering sub-millisecond latency, high throughput, NVMe-oF, seamless use on standard networks, and easy implementation into existing infrastructures. With full hardware freedom, ExaScale removes complexity while enabling true acceleration.

This guide highlights the essential storage capabilities required to power modern AI environments.

TODAY'S AI LANDSCAPE

Modern AI workloads generate massive data flows, real-time processing demands, and rapid iteration cycles. As models grow larger and more distributed, storage becomes critical to keeping training, inference, and pipelines fast, stable, and scalable.

High throughput, sub-millisecond latency, and seamless multi-node access are now essential. Legacy infrastructure often struggles to keep up, slowing model development and creating bottlenecks across data pipelines.

To support modern AI, organisations need storage that delivers predictable performance, effortless scaling, and consistent acceleration across training clusters.

Key Challenges

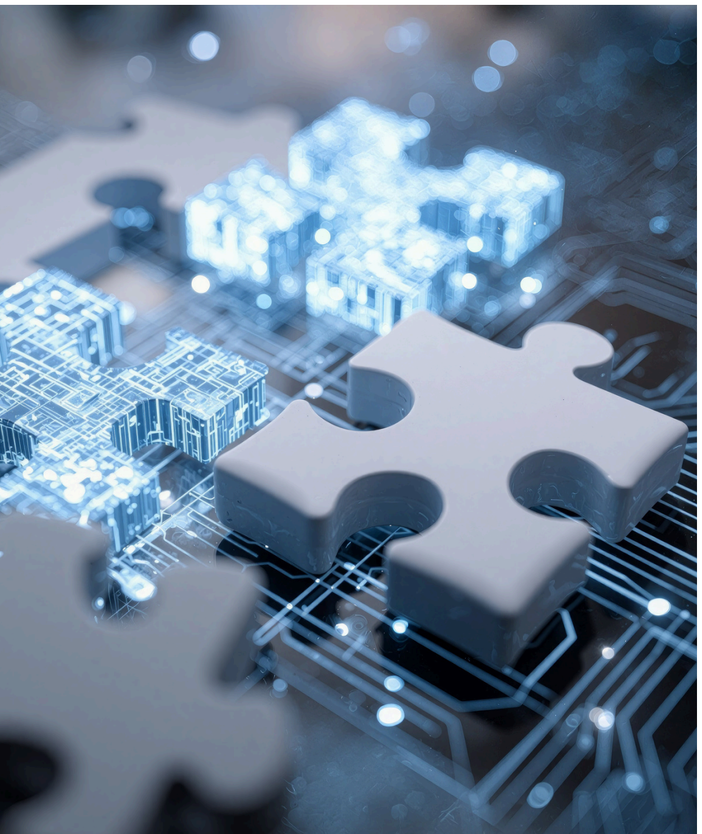
As AI workloads expand, common challenges include:

- Storage latency slowing training and inference
- Bandwidth limits that restrict multi-node scaling
- Inconsistent performance across GPU clusters
- Slow data access delaying model iteration cycles
- Legacy arrays unable to support NVMe-oF or high throughput
- Difficulty integrating AI pipelines with existing infrastructure

These challenges highlight the need for storage engineered for sub-millisecond latency, high throughput, AI performance at scale.

WHAT TO LOOK FOR

WHEN PICKING THE RIGHT PLATFORM



Before comparing AI infrastructure options, focus on storage capabilities that directly impact training speed, inference performance, and scalable data pipelines:

WHAT TO LOOK FOR

- **Sub-millisecond latency:** Essential for fast model training, real-time inference, and GPU efficiency.
- **High throughput:** Keeps multi-node training pipelines fed without bottlenecks.
- **NVMe-oF support:** Unlocks maximum parallelism and low-latency access across AI clusters.
- **Standard network compatibility:** Works seamlessly with existing Ethernet Fabrics.
- **Easy implementation:** Rapid deployment into current GPU, compute, and data environments.
- **Hardware freedom:** Avoid vendor lock-in with flexible, commodity hardware support.
- **Scalable performance:** Maintain consistent speed as datasets, models, and nodes grow.

These fundamentals ensure your AI workloads run fast, scale smoothly, and stay operational across hybrid and distributed environments.

WHAT TO MEASURE FOR PERFORMANCE

To evaluate storage readiness for AI workloads, track the metrics that influence training speed, inference efficiency, data movement, and GPU utilisation:

- **Latency:** Sub-ms responsiveness to keep GPUs fully utilised.
- **Throughput:** High bandwidth for large datasets and parallel reads.
- **IOPS:** Fast small-block performance for metadata and checkpoints.
- **Scalability:** Stable performance as datasets and nodes expand.
- **Snapshot Reliability:** Safe, corruption-free checkpoints and versions.
- **Network Flexibility:** NVMe-oF and standard network support.

These indicators show how well your storage accelerates AI training, supports real-time inference, and maintains performance as workloads scale.

WHY AI PERFORMANCE MATTERS

AI performance depends on how effectively the storage platform feeds GPUs, moves large datasets, and maintains consistency across training, inference, and distributed pipelines. Fast, predictable storage reduces bottlenecks, accelerates experimentation, and keeps AI workflows running without interruption.

PERFORMANCE-READY ARCHITECTURE

NGX platforms use low-latency flash, sub-millisecond responsiveness, and high-throughput design to keep GPUs fully utilised during training and inference.

This architecture supports large datasets, rapid checkpointing, fast parallel reads, and consistent performance as models, nodes, and data volumes grow.

OFFLOAD & SCALABILITY FEATURES

Modern AI storage should support NVMe-oF, standard network fabrics to move datasets across clusters with minimal overhead.

These capabilities eliminate GPU idle time, speed up model iteration, simplify distributed training, and ensure seamless scale-out across heterogeneous environments.

NGX PRODUCTS OVERVIEW

NGX offers a product **-NGX ExaScale-** engineered to accelerate AI workloads with sub-millisecond latency, high throughput, scalable dataset delivery, efficient checkpointing, and seamless integration with existing GPU infrastructure.

NGX PRODUCTS OVERVIEW

NGX EXASCALE: NGX provides a next-generation storage platform purpose-built for AI, engineered to deliver sub-millisecond latency, high throughput, and seamless scaling required for modern GPU-accelerated workloads.

ExaScale supports everything from high-speed data pipelines and distributed training to real-time inference, vector databases, and massive model datasets — all on standard network infrastructure with freedom of hardware choice.

For High-Performance AI Workloads

NGX ExaScale

- Sub-millisecond latency
- High throughput for large datasets
- NVMe-oF
- Works with standard network infrastructure
- Easy implementation into existing environments
- Hardware-agnostic deployment

NGX STORAGE

CHOOSING THE RIGHT OPTION

As AI initiatives expand, organisations typically follow one of three strategic paths.
Your priorities (performance, scalability, or operational efficiency) determine the best direction.

1. **Maximise Training Performance:** Sub-ms latency and high throughput to keep GPUs busy.
2. **Optimise Inference:** Fast, predictable responses with low-latency access to model assets.
3. **Scale Multi-Node AI:** Consistent performance for clustered training and distributed datasets.

The best approach depends on your model size, dataset growth, GPU strategy, and long-term AI infrastructure roadmap.

NEXT STEPS

01 Review AI Performance Gaps

Identify GPU idle time, slow data delivery, metadata bottlenecks, or scaling issues during training or inference.

02 Define Your Priorities

Clarify whether your focus is faster training, low-latency inference, multi-node scalability, or efficient checkpoints/versioning.

03 Assess Your Requirements

Match your priorities with what an AI-ready storage layer must deliver—sub-ms latency, high throughput, snapshot reliability, and sync/async dataset distribution.

04 Compare Platform Capabilities

Evaluate options based on real training throughput, inference latency, checkpoint speed, scalability, and NVMe-oF network integration.

05 Validate With a PoC

Test real training, inference, and dataset workflows to confirm GPU utilisation, pipeline stability, and consistent performance under load.

BEFORE YOU MOVE FORWARD

A strong AI strategy begins with storage built for speed, scale, and reliability. The essentials remain the same: sub-millisecond latency to keep GPUs fed, high throughput for large datasets, fast and reliable checkpoints, and seamless multi-node scalability. NGX delivers all of these with predictable performance, stable dataset delivery, and efficient recovery across demanding AI pipelines.

With NGX underpinning your AI infrastructure, accelerated training, smooth multi-node scaling, and protected model assets become a given, not a challenge.

BULLETPROOF YOUR STORAGE WITH

NGX STORAGE

+90 312 227 04 74

info@ngxstorage.com

Hacettepe Teknokent, Safir C Blok

No:31 Ankara / Turkey