



Lenovo Storage Guidance for Hybrid AI

Last update: **29 September 2026**
Version 1.0

General guidance of High-Performance Storage for different workloads

Lenovo High-Performance Storage solutions and AI workloads fit

Storage layer solutioning for different workload profiles and KPI mapping

Mapping of Lenovo Hybrid AI Factory platforms with ThinkSystem storage portfolio

Sinan Atan
Conor Elrick
Aurelien Ortiz
Hapsara Sukasdadi



Table of Contents

Abstract	1
Executive Summary	2
1 Introduction.....	3
2 Lenovo Hybrid AI Factory Platforms	4
3 Understanding Storage Requirements for AI	5
3.1 Why AI Workloads Require High-Performance Storage.....	5
3.2 General Storage Performance Guidance.....	6
3.3 Guidance by Workload.....	8
4 Lenovo AI Storage Solutions	12
4.1 Lenovo ThinkSystem DM and DG Series.....	12
4.2 Lenovo ThinkSystem DE Series and Lenovo Partner Storage Solutions	13
5 Selecting Storage for Lenovo Hybrid AI Platforms	15
5.1 Storage Sizing Considerations: Understanding the Workload.....	15
5.2 Mapping Lenovo Storage Platforms to Throughput Requirements	16
5.3 Storage Selection Criteria	20
5.4 Recommended Storage Starting Points for Hybrid AI Deployments	21
6 Conclusions and Future Studies	24
Abbreviations	25
References.....	26
Authors	27
Document History	28

Trademarks and Special Notices 29

Units

This document uses the following notation:

- **b** denotes bits.
- **B** denotes bytes; one byte equals 8 bits.
- **GB** denotes gigabytes, equivalent to 10^9 or 1000^3 bytes.
- **GiB** denotes gibibytes, equivalent to 1024^3 bytes.

Abstract

This paper provides practical guidance for selecting and positioning storage solutions for Lenovo Hybrid AI Factory deployments. It examines how workload characteristics, deployment scale, capacity, data-access protocols, read and write performance, data services, and scalability influence storage architecture decisions. The paper maps these requirements to common enterprise AI workloads, including inference, retrieval-augmented generation, agentic AI, computer vision, fine-tuning, and model training, and considers Lenovo ThinkSystem DM, DG, and DE Series storage solutions alongside scale-out high-performance storage solutions available through the Lenovo partner ecosystem.

The intended audience includes Lenovo sellers, solution architects, business partners, and customers planning enterprise AI infrastructure. The guidance provides a structured approach for identifying an appropriate Lenovo storage starting point and determining when workload scale, performance, data-management, or operational requirements call for a more specialized architecture. Future work will deepen the analysis of Lenovo partner ecosystem solutions, including their positioning, scaling characteristics, data-management capabilities, and applicability to larger and more demanding AI environments.

Executive Summary

Enterprise AI solutions depend on timely, reliable access to growing volumes of structured and unstructured data. Although GPU selection often receives the greatest attention during infrastructure planning, an undersized or poorly matched storage solution can increase model-loading times, delay data-intensive operations, and reduce the effective utilization of valuable compute resources. Storage selection for an AI environment must therefore consider more than capacity. It must account for workload access patterns, aggregate read and write performance, concurrency, metadata operations, data-access protocols, resilience, network limitations, and anticipated growth.

Storage requirements can differ substantially across AI workloads. Inference and retrieval-augmented generation (RAG) often emphasize model and data access, shared data services, metadata operations, and predictable latency. Computer vision and other data-intensive workloads can require substantially higher sequential read throughput. Training and fine-tuning introduce additional requirements for shared dataset access and sufficient write performance to complete checkpoint operations within the target time.

This paper provides a workload-based framework for positioning Lenovo ThinkSystem DM, DG, and DE Series storage, together with scale-out high-performance storage solutions available through the Lenovo partner ecosystem, for Lenovo Hybrid AI Factory deployments. It helps sellers and solution architects identify an appropriate storage starting point for inference, RAG, agentic AI, computer vision, fine-tuning, training, and mixed workloads, and then validate the design against performance, capacity, data services, networking, scalability, and operational requirements.

The recommended solutioning approach begins with the customer's business and operational objectives, Hybrid AI platform, workload mix, deployment scale, and need for shared data. The design should then be validated against usable capacity, read and write performance, metadata and IOPS requirements, data-access protocols, resilience, end-to-end network capacity, expansion needs, and commercial constraints. The objective is not to prescribe a single storage product for every Hybrid AI configuration, but to identify the simplest architecture that meets the customer's requirements, provides a practical growth path, and is recognized when a Lenovo partner ecosystem solution may be more appropriate.

1 Introduction

Lenovo Hybrid AI Factory solutions provide integrated infrastructure platforms designed to support enterprise AI workloads across a range of deployment scales. The Lenovo Hybrid AI 221 and 241 platforms address compact deployments for workloads such as retrieval-augmented generation (RAG), computer vision, and enterprise inference, while the Lenovo Hybrid AI 285 and 289 platforms support larger environments requiring multi-node scalability and high-performance east-west networking for enterprise inference, agentic AI, fine-tuning, and model training.

An efficient storage layer is required to provide the compute plane with timely access to data and help keep accelerated resources productively utilized. The demands placed on this layer increase as deployments add users, concurrent workloads, compute nodes, and Scalable Units (SUs).

An effective AI storage solution must provide the capacity, performance, connectivity, resilience, and data services required by the intended workloads. Inference and RAG may emphasize shared access to models and data, metadata performance, and predictable latency, while computer vision and other data-intensive workloads can require higher sequential read throughput. Training and fine-tuning introduce additional requirements for shared dataset access and sufficient write performance to complete checkpoint operations within the target time. Where shared storage is required, its architecture must be aligned with the scale and connectivity of the Hybrid AI platform. The design must also account for workload mix, concurrency, dataset characteristics, performance objectives, access protocols, network limitations, and expected growth.

This paper provides guidance for positioning Lenovo ThinkSystem DM, DG, and DE Series storage, together with scale-out high-performance storage solutions available through the Lenovo partner ecosystem, for Lenovo Hybrid AI deployments. It explains the role of storage in data-driven AI environments, provides workload-based read and write performance guidance, and maps storage architectures to the requirements of the Hybrid AI Factory platforms. The objective is to help sellers and solution architects identify an appropriate storage starting point and recognize when workload demands, deployment scale, or data-service requirements call for a more specialized high-performance storage architecture.

The paper is organized as follows. Section 2 introduces the Lenovo Hybrid AI Factory platforms and their intended workload focus. Section 3 examines how AI workload characteristics influence storage requirements and provides general guidance for common AI workloads. Section 4 presents the Lenovo AI storage portfolio, including Lenovo ThinkSystem DM, DG, and DE Series storage and solutions available through the Lenovo partner ecosystem. Section 5 provides workload-based sizing guidance, maps storage architectures to the Lenovo Hybrid AI Factory platforms and outlines the principal criteria for selecting an appropriate storage starting point. Finally, Section 6 summarizes the guidance and identifies areas for further study.

2 Lenovo Hybrid AI Factory Platforms

Lenovo Hybrid AI Factory platforms bring together compute, networking, storage, system management, services, and AI software through an integrated infrastructure approach. The platforms provide different architectural starting points based on workload requirements and deployment scale, rather than prescribing a single infrastructure configuration for every enterprise AI environment.

The Lenovo Hybrid AI 221 and 241 platforms [1] are positioned for compact deployments supporting workloads such as enterprise inference, RAG, and computer vision. The Lenovo Hybrid AI 285 platform [2] supports scalable, multi-node environments for enterprise inference, RAG, agentic AI, fine-tuning, and mixed AI workloads. The Lenovo Hybrid AI 289 platform [3] addresses larger and more demanding environments focused on model training, fine-tuning, and large-scale enterprise inference. Table 1 summarizes the reference infrastructure configuration and intended workload focus of each platform.

Lenovo Hybrid AI Platform	Reference Server and Form Factor	Reference GPU Configuration	Intended Workload Focus
221	ThinkSystem SR650a V4, 2U	2× PCIe GPUs, RTX PRO 6000 Blackwell Server Edition	Enterprise inference, RAG, computer vision
241	ThinkSystem SR650a V4 or SR665 V4, 2U	4× PCIe GPUs, RTX PRO 4500 Blackwell Server Edition	Enterprise inference, RAG, computer vision
285	ThinkSystem SR675i V3, 3U	8× PCIe GPUs, RTX PRO 6000 Blackwell Server Edition	Enterprise inference, RAG, agentic AI, fine-tuning, mixed workloads
289	ThinkSystem SR680a V4, 8U	8× NVIDIA HGX GPUs, B300	Model training, fine-tuning, and large-scale enterprise inference

Table 1 – Lenovo Hybrid AI Factory Platform Overview

As Hybrid AI Factory deployments expand from compact, primarily node-local configurations to shared and multi-node environments, their storage requirements also evolve. The appropriate storage architecture depends on workload characteristics, data volume and type, concurrency, required data services, access protocols, and deployment scale. The platform definitions introduced in this section provide the basis for determining where node-local storage, Lenovo ThinkSystem shared storage, and scale-out high-performance storage architectures available through the Lenovo partner ecosystem fit within an enterprise AI solution. The next section examines how these workloads and deployment characteristics translate into storage requirements.

3 Understanding Storage Requirements for AI

Artificial intelligence is rapidly evolving from primarily generative applications such as conversational chatbots toward more dynamic workflows involving RAG, multimodal AI, and agentic AI. The requirements for running these applications extend beyond the capabilities of the underlying models. They increasingly depend on continuous access to enterprise data throughout the AI lifecycle, including data ingestion and preparation, model training and fine-tuning, retrieval and inference, and the generation of new data and model artifacts. As AI becomes more deeply integrated into enterprise processes, the architecture used to store, protect, share, and access data effectively and efficiently becomes an increasingly vital part of the overall AI solution.

This requirement makes AI infrastructure inherently data-centric. Model quality and usefulness depend heavily on the quality, relevance, and availability of the data used by the AI system. At the same time, organizations must maintain the resilience, security, governance, and integrity of that data. Designing a storage solution based on capacity alone is therefore not sufficient. Data must be available to the right workloads with the required performance and reliability. This is particularly important in GPU-accelerated environments, where inadequate data access can leave valuable compute resources waiting for data rather than performing useful work.

Storage architectures for AI must consequently balance capacity and data services with performance characteristics such as throughput, latency, metadata operations, and concurrent access. These requirements vary significantly by workload and scale, from enterprise RAG and inference to data-intensive computer vision and distributed model training. As workload scale and performance demands increase, the role of the high-performance storage becomes increasingly critical within the AI infrastructure, providing the shared data foundation required to keep accelerated compute resources productively utilized.

3.1 Why AI Workloads Require High-Performance Storage

High-performance storage (HPS) refers to storage architectures designed to provide the bandwidth, latency, concurrency, and scalability required by data-intensive workloads. HPS evolved alongside high-performance computing (HPC), where large numbers of compute nodes require concurrent access to shared datasets and storage resources. Rather than being defined by a specific storage technology or protocol, HPS is characterized by its ability to deliver sustained performance to many concurrent workloads accessing shared data.

AI introduces many of the same challenges, but with different I/O patterns depending on the workload. Training may require sustained access to large datasets and sufficient write performance for model checkpointing, while inference and RAG commonly require efficient access to models, source data, and other shared AI assets. Computer vision use-cases combined with very low-latency KPIs such as autonomous driving or incident prevention requires very high throughput to process high resolution images. Local memory and NVMe storage can provide very high performance for data already resident on a compute node, while shared high-performance storage becomes important when data must be accessed across multiple nodes, workloads, users, or when additional capacity is required for datasets, checkpoints, model artifacts, and inference-state offloading.

As GPU performance improves and Hybrid AI Factory deployments scale, the volume of datasets, models, checkpoints, and AI artifacts that must be shared between users, workloads, and compute resources also

increases. Storage therefore becomes a critical part of the end-to-end AI data path, connecting enterprise datasets with accelerated compute while providing the performance, availability, and data services required by the environment. HPS helps address these challenges by delivering the bandwidth, scalability, and concurrent data access needed to efficiently support modern AI workloads. Because the necessary performance remains workload dependent, the following sections provide general sizing guidance for mixed environments and then examine the storage characteristics of individual AI workloads.

Capacity requirements for HPS vary by customer, depending on factors such as dataset size, number of users, retention policies, data-protection requirements, and expected growth. This section therefore focuses primarily on storage throughput, which directly affects the ability to keep GPUs supplied with data and minimize time spent waiting for I/O.

Read requirements are expressed primarily as throughput per concurrently active GPU, allowing the guidance to scale across different deployment sizes. Write requirements for training are evaluated separately because checkpoint bandwidth is driven mainly by model size, precision, optimizer state, checkpoint duration, and the number of training jobs checkpointing concurrently.

These calculations provide initial sizing targets rather than complete storage specifications. The sizing methodology presented in this paper focuses primarily on storage-throughput requirements. Other storage characteristics, including latency, IOPS, metadata performance, protocol efficiency, access patterns, usable capacity, resilience, and data services, can also influence workload behaviour and storage selection and should therefore be evaluated as part of the overall solution-design process.

3.2 General Storage Performance Guidance

Storage performance for an AI environment should be determined from the workloads expected to run concurrently rather than from the GPU platform alone. Different workloads generate different access patterns, and no single throughput value is appropriate for every deployment. Read performance is relevant across inference, RAG, computer vision, fine-tuning, and training. Sustained write performance is particularly important for training because model checkpoints can generate large, blocking write operations.

For initial sizing, NVIDIA defines **Standard** and **Enhanced** storage-performance reference points [4]. The term “Standard” represents compute-dominant workloads for which data I/O has a limited effect on overall execution time and much of the working dataset can be retained in local memory or cache. “Enhanced” on the other hand, represents more I/O-sensitive workloads using larger or multimodal datasets that require more frequent access to shared storage.

The reference values used in this paper are:

- **Standard:** approximately 0.16 GB/s of read throughput and 0.08 GB/s of write throughput per GPU.
- **Enhanced:** approximately 0.5 GB/s of read throughput and 0.25 GB/s of write throughput per GPU.

These values provide reference performance envelopes rather than universal workload requirements. Those can easily go beyond the metrics stated here; for example, use cases in automotive and other computer vision related tasks needs ~4GBps per GPU read performance. Final sizing must reflect the actual workload, cache effectiveness, dataset characteristics, concurrency, checkpoint behavior, and performance objectives.

This paper assumes that read and write performance should be sized differently. Read performance applies across all AI workloads and is generally estimated from the number of GPUs expected to access shared

storage concurrently. Write performance is particularly important for model training and fine-tuning because checkpoint operations can generate large, synchronous writes that temporarily interrupt computation.

Read performance for all workloads

For shared storage supporting RAG, LLM serving, fine-tuning, training, or mixed workloads, this paper uses the Enhanced read-performance target as a conservative starting point for conventional LLM workloads. Lower-throughput environments may be adequately served by Standard guidance, particularly when the working dataset is cached locally or I/O is not a significant contributor to execution time. Conversely, data-intensive workloads such as high-resolution computer vision can require substantially greater sequential read throughput and should be sized from workload-specific measurements or the higher guidance given in this section.

For both homogenous and mixed-workload deployments, the aggregate storage requirement can be estimated as follows:

$$\text{Aggregate read throughput} = \text{active GPU count} \times \text{read throughput target per GPU}$$

The active GPU count should represent GPUs expected to access shared storage concurrently, not necessarily the total number installed in the solution. For example, a design containing several independent workloads may not reach maximum theoretical demand if computation, data loading, inference, and checkpoint operations occur offset to each other with respect to the other workloads. However, environments requiring predictable performance under maximum concurrency should be sized for the combined peak demand of all active workloads. This calculation provides an initial throughput target and does not replace validation of metadata performance, IOPS, latency, usable capacity, protocols, resilience, or end-to-end network bandwidth.

Write performance for training

Training write performance should be calculated primarily from the model size, numerical precision, optimizer and training state, required checkpoint completion time, and number of training jobs expected to checkpoint concurrently. Unlike read-throughput sizing, checkpoint write bandwidth should not be derived directly from the number of GPUs in the solution.

Network limitation

After determining the workload requirement, the resulting throughput target must be compared with the usable bandwidth of the complete network path. In this paper, network bandwidth ceiling represents the maximum storage bandwidth that the referenced platform network can carry under full utilization. It is an architectural limit, not a workload-sizing and not an indication of GPU performance. A storage system capable of exceeding this value may still be useful for future expansion, but the AI compute nodes cannot consume the additional bandwidth through the specified network path. Changes to adapter, switch, and uplink configurations can change this boundary.

The reference network configurations establish the following theoretical per-GPU limits at full concurrent

utilization:

- Hybrid AI 221: up to 12.5 GB/s per GPU, based on 200 Gb/s of node bandwidth shared by two GPUs.
- Hybrid AI 241: up to 6.25 GB/s per GPU, based on the same node bandwidth shared by four GPUs.
- Hybrid AI 285 and 289: up to 3.125 GB/s per GPU, based on 800 Gb/s of storage-facing bandwidth per Scalable Unit shared by 32 GPUs.

These values are network limits per reference designs for the documented configurations, not guidelines to provision every workload at those rates. Detailed information about sizing per workload profile is given in Section 3.3. They also explain why the smaller Hybrid AI 221 and 241 platforms can expose more network bandwidth per GPU: their available node bandwidth is divided among fewer GPUs. This should not be interpreted as indicating that these platforms require more storage performance or are inherently better suited to every high-throughput workload. The appropriate target remains workload dependent.

For mixed workloads, a practical design approach is therefore to identify the highest expected read demand, estimate the number of GPUs accessing shared storage concurrently, add any simultaneous checkpoint write requirements, and compare the resulting aggregate demand with both the sustainable storage-system performance and the end-to-end network limit. Where precise workload data is unavailable, the Standard and Enhanced values provide initial reference points, but the design should be validated through workload characterization before final configuration.

3.3 Guidance by Workload

AI workloads create different storage-access patterns, so no single per-GPU throughput value is appropriate for every deployment. Read performance is relevant across inference, RAG, computer vision, fine-tuning, and training, while sustained write performance is particularly important for model training because checkpoints can temporarily generate large, blocking write operations. The following guidance provides initial sizing targets that must be multiplied by the expected concurrency and validated against the platform network.

LLM training and fine-tuning

AI training generates both storage-read and storage-write requirements. Read performance is required to load datasets and continuously stream training data to the GPUs, while write performance is needed for checkpointing, logging, and model persistence. The relative importance of these requirements depends on the workload, dataset size, storage architecture, caching strategy, and data pipeline implementation. In many environments, dataset-read performance can be significantly influenced by memory capacity, caching, prefetching, and data staging techniques, making it difficult to define a universal read-performance requirement.

By contrast, checkpointing requirements can be derived directly from measurable parameters such as model size, numerical precision, optimizer and training state, checkpoint frequency, target checkpoint completion time, and number of jobs checkpointing concurrently. The checkpoint throughput requirement is theoretically independent of the number of GPUs used for training because it is primarily determined by the amount of model and optimizer state that must be written to storage. Regardless of the training algorithm employed, a lower bound exists on the checkpoint size based on the number of model parameters alone. In practice, however, the total amount of data written during checkpointing depends heavily on the training framework, optimizer implementation, and checkpointing methodology. Optimizers introduce additional memory and I/O

overhead because they maintain extra state information beyond the model parameters themselves.

Starting with the numerical precision used during training (FP4, FP8, FP16, and so on), it is possible to estimate the minimum size of a model checkpoint from the number of model parameters. For example, a 13-billion-parameter model quantized into FP16 precision requires 2 bytes per parameter, resulting in approximately 26GB of model data. This value represents the minimum amount of data that must be written regardless of the training algorithm used.

In practice, training checkpoints often contain significantly more information than the model weights alone. A common example is the Adam [5] or AdamW [6] optimizer, which is widely used in AI training frameworks such as PyTorch [7] and in the training of the Meta Llama family of models [8]. In addition to the model weights, these optimizers maintain optimizer state and master weights that must also be checkpointed. For BF16 or FP16 training, these additional data structures can increase the checkpoint footprint to approximately 16 bytes per model parameter. Using the same 13-billion-parameter example, the total checkpoint size increases from approximately 26GB to approximately 208GB.

Although newer optimizers can reduce this overhead, Adam and AdamW remain common reference implementations in AI training environments. For this reason, this paper uses the Adam/AdamW footprint as a conservative sizing estimate for Lenovo Hybrid AI deployments.

Because GPUs represent a significant infrastructure investment, checkpoint operations should complete quickly enough to minimize the time training jobs spend waiting on storage I/O. A target checkpoint duration therefore provides a practical method for translating checkpoint size into a storage-bandwidth requirement. For the sizing methodology used throughout this paper, checkpoint bandwidth is calculated assuming a checkpoint footprint of 16 bytes per model parameter for FP16 or BF16 training using Adam or AdamW optimizer state, together with a target checkpoint completion time of 300 seconds. Using the previous example of a 13-billion-parameter model, this results in a checkpoint size of approximately 208 GB and a required write throughput of approximately 0.7 GB/s to write a checkpoint in 300 seconds.

This produces the following approximate aggregate write-throughput requirements for a single training job:

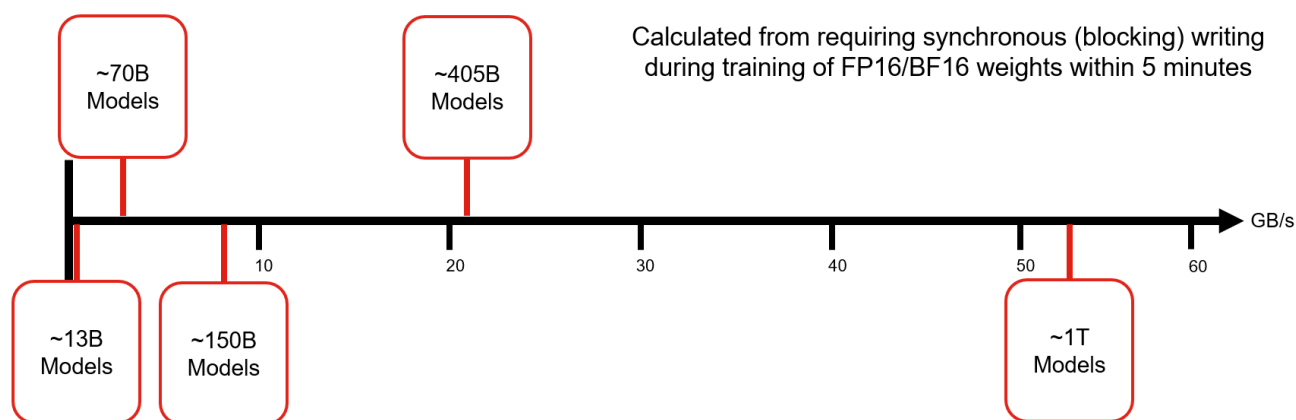


Figure 1 – Aggregate storage write-performance requirements for AI training by model size

If several training jobs may checkpoint simultaneously, their write-bandwidth requirements must be added. A shorter checkpoint target, a larger model state, greater precision, or additional optimizer state increases the

required bandwidth. A longer checkpoint interval, lower precision, reduced checkpoint content, or staggered checkpoint scheduling can lower the peak aggregate requirement. For fine-tuning, these figures should be treated as an upper bound because the amount of state written can be smaller depending on the fine-tuning method and checkpoint implementation.

RAG and LLM serving

RAG and LLM-serving workloads typically require access to models, vector database persistence, source documents, embeddings, and other shared AI assets. Their storage behaviour can include model loading, model swapping, metadata-intensive access, and repeated retrieval of relatively small objects. Consequently, latency, metadata performance, and IOPS can be as important as sequential bandwidth.

For external shared-storage sizing, this paper recommends an initial read target of approximately 0.5 GB/s per concurrently active GPU, corresponding to the Enhanced NVIDIA reference point. Deployments with working data retained effectively in local memory or local storage may operate closer to the Standard reference of approximately 0.16 GB/s per GPU. In later sections, workloads with throughput requirements closer to the lower reference target are categorized as “Low-IO RAG/Inferencing”. The appropriate choice of value depends on model-loading frequency, cache effectiveness, context size, dataset characteristics, concurrency, and the share of the workload served from local rather than shared storage.

Computer vision

Computer-vision workloads can generate sustained sequential reads from large image or video datasets. High-resolution automotive and computer-vision workloads represent the most demanding read case in the current guidance, with a target of approximately 4 GB/s per active GPU. This is a workload-specific upper sizing point and should not be generalized to every computer-vision deployment. Actual demand depends on image resolution, encoding, preprocessing, sampling rate, augmentation, cache effectiveness, and the number of data-loader processes operating concurrently.

Computer-vision workloads may also be more sensitive to unpredictable I/O contention because data access can occur continuously rather than at scheduled checkpoint intervals. Designs supporting high-resolution image or video pipelines should therefore validate not only peak throughput but also sustained performance, concurrency, and the behaviour of the complete data path.

On Hybrid AI 285 and 289 configurations, this 4 GB/s workload target exceeds the referenced network ceiling value of 3.125 GB/s per GPU when all GPUs perform I/O concurrently. The achievable rate would therefore be constrained by the network configuration before reaching the workload target. This does not prevent computer-vision workloads from running, but maximum concurrency may increase contention and data-loading latency. Caching, workload staggering, or changes to the end-to-end network design may be required where sustained performance near the workload target is necessary.

Agentic AI, multimodal, and blended workloads

Agentic AI, multimodal pipelines, and mixed AI environments do not have a single representative I/O profile. These environments often combine model serving, retrieval, vector database operations, image and video processing, data ingestion, tool usage, fine-tuning, and training workloads, each with different storage access characteristics.

As a result, storage performance requirements should be evaluated against the expected range of data

access patterns within the workload mix. At the lower end, deployments dominated by cached inference, RAG, or lightweight model serving may require relatively modest storage throughput. At the upper end, workloads involving large-scale training, fine-tuning, computer vision, high-resolution automotive AI, or other data-intensive processing pipelines can generate substantially higher read and write demands. Mixed environments typically fall between these bounds, and the appropriate storage architecture should be sized according to the dominant workload characteristics and expected concurrency levels.

Figure 2 gives an overview of read requirements for various AI applications on storage plane.

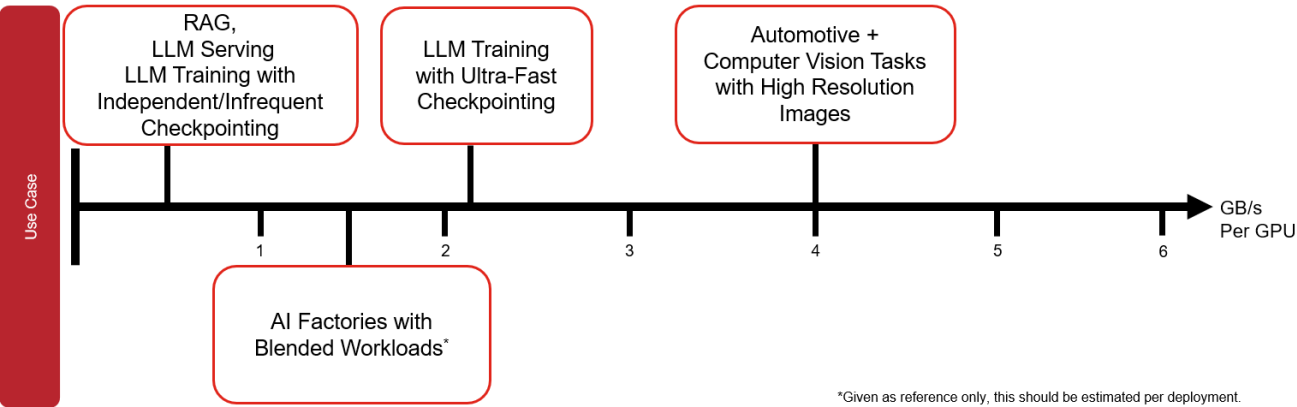


Figure 2 – Use Cases Storage read-performance requirements per GPU

4 Lenovo AI Storage Solutions

Lenovo offers a broad portfolio of storage solutions for enterprise AI, ranging from Lenovo ThinkSystem DM, DG, and DE Series storage to high-performance software-defined storage (SDS) solutions delivered with Lenovo technology partners. Together, these offerings provide options for different workload, capacity, performance, data-service, and scalability requirements across Lenovo Hybrid AI Factory deployments.

The primary focus of this paper is the Lenovo ThinkSystem DM and DG Series, which represent the recommended storage foundation for Lenovo Hybrid AI platforms. Alternative SDS architectures are also considered through the use of Lenovo ThinkSystem DE Series block storage combined with shared file-system software, as well as partner solutions from DDN, WEKA, VAST, IBM Storage Scale, and Cloudian. The following sections introduce these storage platforms and describe their roles within the Lenovo AI storage portfolio.

4.1 Lenovo ThinkSystem DM and DG Series

The Lenovo ThinkSystem DM Series [9] and DG Series [10] are unified ONTAP-based storage platforms designed to provide enterprise data services, high availability, and multiprotocol access for AI workloads. Both platforms support file, block, and object access through protocols including NFS, SMB, S3, iSCSI, NVMe/TCP, Fibre Channel, and NVMe/FC, while providing enterprise data-management capabilities such as snapshots, cloning, quality of service (QoS), compression, deduplication, encryption, backup, replication, and scale-out clustering. By combining dual active controllers with a common ONTAP software foundation, the DM and DG Series provide a resilient and flexible storage architecture capable of supporting a wide range of AI data-access patterns and workload requirements.

While both platforms share the same software capabilities and management framework, they are optimized for different priorities. The Lenovo ThinkSystem DM3200F, DM5200F, and DM7200F are all-NVMe flash arrays designed for performance-oriented deployments requiring high throughput, predictable latency, metadata performance, and scalability. In contrast, the Lenovo ThinkSystem DG5200 and DG7200 utilize high-capacity QLC NVMe flash technology to maximize storage density and cost efficiency while still delivering the benefits of an all-flash architecture.

The DM Series provides a high-performance shared-data foundation for workloads that include model repositories, vector database persistence, training datasets, model checkpoints, enterprise RAG, model serving, fine-tuning, training, and mixed-workload AI environments. The DM Series is therefore the preferred choice when performance, responsiveness, and storage scalability are primary design considerations.

The DG Series provides a shared and resilient storage foundation for model repositories, RAG source data, image and video collections, data lakes, and other large enterprise datasets. Its high-capacity QLC drives make it particularly well suited to deployments where storage density, capacity growth, and cost efficiency take priority over achieving the highest available storage throughput. Within Lenovo Hybrid AI Factory solutions, the DG Series serves as a capacity-focused shared-storage option for inference, RAG, agentic AI, computer vision data repositories, and other workloads whose capacity and performance requirements can be met by the selected configuration.

Compared with TLC flash media used with the DM Series, the QLC technology used by the DG Series provides greater storage density and a lower cost per terabyte. Lenovo internal pricing analysis indicates that

DG-compatible QLC drives maintain a relatively consistent cost per terabyte across the available drive capacities while offering a lower cost per terabyte than comparable TLC-based alternatives. In contrast, the cost per terabyte of TLC-based DM Series drives generally increases as drive capacity decreases, making smaller-capacity configurations comparatively less cost-efficient.

Table 2 compares the principal configuration, scalability, capacity, and performance characteristics of the Lenovo ThinkSystem DM and DG all-flash storage systems. The comparison highlights the different scaling profiles across the portfolio, including supported high-availability pairs, expansion enclosures, maximum drive counts, and estimated read throughput. These characteristics provide an initial indication of how each system may fit different AI deployment sizes and performance requirements. However, the figures should be treated as architectural reference points rather than stand-alone sizing recommendations, because effective performance and usable capacity depend on the selected configuration, access protocols, network design, workload characteristics, and data-protection requirements.

Lenovo ThinkSystem storage performance comparison

DM and DG Series | Configuration, capacity and performance

METRIC	DM3200F All-NVMe flash	DM5200F All-NVMe flash	DM7200F All-NVMe flash	DG5200 All-QLC NVMe flash	DG7200 All-QLC NVMe flash
System memory per HA pair (GB)	 128	 128	 256	 128	 128
NAS cluster limit (HA pairs) ¹	 3	 4	 12	 4	 4
Max expansion shelves ²	 1	 2	 4	 2	 4
Max NVMe SSDs per HA pair	 48	 72	 120	 72	 120
Max raw capacity per HA pair (TB) ¹	 737	 1,106	 1,843	 2,212	 7,374
Write throughput per HA pair (GB/s) ¹	 2.43	 4.34	 8.47	 2.25	 2.45
Read throughput per HA pair (GB/s) ¹	 5.72	 11.97	 20.16	 5.05	 6.95
4 KB random write IOPS per HA pair ¹	 115,486	 275,106	 426,893	 130,548	 162,391
4 KB random read IOPS per HA pair	 414,337	 890,101	 1,341,573	 314,941	 409,126

¹ For NAS configurations. DM7200F SAN cluster limit: 6 HA pairs.

² DM uses DM242N shelves; DG uses DG242N shelves. Four shelves on DM7200F or DG7200 require supported switch connectivity. Capacity and SSD limits include expansion shelves. Raw capacity is before RAID and system overhead.

Filled bar = value relative to the highest model in that row; pale track = 100% of row maximum.

Table 2 – Lenovo DM and DG Storage Solutions Compared

4.2 Lenovo ThinkSystem DE Series and Lenovo Partner Storage Solutions

For data-intensive AI environments, Lenovo supports two complementary approaches to high-performance, scale-out storage:

1. A customizable architecture combining Lenovo ThinkSystem DE Series block storage with a parallel file system, such as IBM Storage Scale, BeeGFS, or Lustre.

2. An integrated software-defined storage or appliance-based solution from Lenovo ecosystem partners, including IBM Storage Scale, WEKA, VAST, DDN, and Cloudian.

The first approach uses Lenovo ThinkSystem DE Series arrays as the block-storage foundation of a customized architecture. The DE Series portfolio includes hybrid H Series systems, which combine different storage media to balance performance and capacity, and all-flash F Series systems for workloads requiring higher performance and lower latency. These systems offer flexible drive configurations, enterprise storage-management capabilities, high availability, and a range of host-connectivity options.

In AI environments, DE Series storage can support large datasets, model checkpoints, and other workloads requiring high aggregate read and write throughput. Because the DE Series provides block storage, shared file access across multiple AI compute nodes requires a parallel file system or data-management layer, together with the supporting servers, networking, and validated architecture. Technologies such as IBM Storage Scale, BeeGFS, and Lustre can provide this shared file system layer. This modular approach gives customers flexibility to design and optimize the architecture around their specific performance, capacity, scalability, and integration requirements.

The second approach uses integrated SDS or appliance-based solutions from Lenovo ecosystem partners, including IBM Storage Scale, WEKA, VAST, DDN, and Cloudian. It builds on Lenovo's experience delivering HPC storage architectures with these partners by combining their technologies with Lenovo compute, networking, integration, and services. Depending on the selected platform, these solutions may provide capabilities such as scale-out file and object storage, large shared namespaces, and specialized AI data-management services. Their more integrated design may also simplify deployment and provide an operational model aligned with the customer's existing software stack and data strategy.

Both approaches can support customized, data-intensive environments where performance, scalability, or specialized data services are primary design considerations. They can address requirements such as high aggregate throughput, shared access across multiple compute nodes and Scalable Units (SUs), high-performance training and checkpointing, object-based data lakes, and specialized AI data-management capabilities.

These approaches represent different architectural models rather than different tiers of the same portfolio. A DE Series-based solution provides a modular architecture that combines block storage with a parallel file system, servers, and networking. An integrated SDS or appliance-based solution delivers storage and data services through a more unified platform.

The appropriate architecture depends on the customer's workload, capacity and performance target, deployment scale, access protocols, resilience requirements, software stack, data-management needs, and operational model. Detailed sizing, architecture validation, and partner selection should therefore be based on the customer's specific requirements.

5 Selecting Storage for Lenovo Hybrid AI Platforms

This section brings together the AI workload and storage requirements presented in Section 3 with the AI positioned storage offerings from Lenovo and Lenovo Partners introduced in Section 4. It provides guidance for selecting an appropriate storage starting point based on workload characteristics, deployment scale, throughput, capacity, concurrency, data services, and operational requirements. These considerations are then mapped to the Lenovo Hybrid AI Factory platforms to identify when Lenovo ThinkSystem DM, DG, or DE Series storage, or a high-performance storage solution from the Lenovo partner ecosystem, may be appropriate.

5.1 Storage Sizing Considerations: Understanding the Workload

Storage selection should be based on the workloads and operational requirements of the deployment, rather than on the compute platform and network configuration alone. Section 3.3 provides initial throughput targets for different AI workload profiles. Conventional RAG and LLM-serving workloads use lower reference targets (i.e., less than a Gigabyte per second per concurrently active GPU), while data-intensive workloads such as high-resolution computer vision can require substantially greater read throughput (i.e., approximately 4 GB/s per concurrently active GPU). In mixed-workload environments, the aggregate requirement depends on the combination of workloads expected to access shared storage concurrently and will typically fall within the range defined by the lowest and highest workload-specific throughput targets. Therefore, alongside the workload specifics, initial sizing should also consider the number of concurrent jobs and users, and the number of GPUs expected to perform storage I/O at the same time.

Training and fine-tuning require separate consideration of checkpoint write performance. The required bandwidth depends primarily on the amount of model, optimizer, and training state written during each checkpoint, the target checkpoint completion time, and the number of jobs that may checkpoint concurrently. A single large-model training workload can generate significant checkpoint-write requirements, while environments supporting many concurrent training jobs must consider the aggregate impact of multiple checkpoint operations. In practice, peak write demand may vary depending on the timing of checkpoint operations across workloads, as workloads are often not checkpointing at the same time. Storage sizing should nevertheless reflect the level of concurrency for which predictable performance is required.

Additional storage-sizing considerations include:

- **Usable capacity and growth requirements:** What is the expected size of datasets, models, checkpoints, and other retained data, and how rapidly are these expected to grow? Consider whether older data can be moved to a separate larger “cold” tier for long term storage.
- **Resilience and data protection requirements:** Different environments may require different levels of availability, redundancy, backup, replication, and recovery. The importance of the data stored on the active storage tier should be considered when selecting an architecture.
- **Access protocols:** Determine whether the workload requires file, block, or object access, such as a shared file system or S3-compatible object storage.

- **Operational and software requirements:** Some environments may require features provided by software-defined or appliance-based solutions, such as large shared namespaces, advanced data-management capabilities, data tiering, or integration with an existing storage strategy.

The following subsections apply these considerations to the Lenovo Hybrid AI Factory platforms and examine how Lenovo ThinkSystem and partner ecosystem storage architectures align with different workload, performance, capacity, scalability, and operation requirements.

5.2 Mapping Lenovo Storage Platforms to Throughput Requirements

Lenovo Hybrid AI 221 and 241 Platform Storage Options

For the Lenovo Hybrid AI 221 and 241 platforms, the scalable unit¹ consists of a single server with two or four GPUs, respectively. Compared to the Lenovo Hybrid AI 285 and 289 platforms, the number of GPUs concurrently accessing storage is therefore significantly lower. As a result, local NVMe storage may be sufficient for many inference, RAG, and computer-vision deployments.

The Lenovo Hybrid AI 221 and 241 platforms, based on the ThinkSystem SR650a V4 or SR675i V3, support configurations with up to eight local drives. The Lenovo Hybrid AI 241 platform, based on the ThinkSystem SR665 V3, supports configurations with up to sixteen local drives. These configurations can provide substantial local-storage capacity and throughput and may satisfy the requirements of many smaller-scale AI deployments without the need for external shared storage.

Where external storage is required, such as to share data across multiple systems, provide additional resilience, support alternative access protocols, or deliver centralized data management, Lenovo ThinkSystem DM and DG Series storage can provide a suitable shared-storage foundation. The storage-positioning guidance presented for the Lenovo Hybrid AI 285 and 289 platforms in the following sections can also be applied to the 221 and 241 platforms. However, because the 221 and 241 platforms contain significantly fewer GPUs per scalable unit, their aggregate storage-throughput requirements are typically lower. In these environments, the DG Series may remain an attractive option for deployments where capacity and cost efficiency are higher priorities than maximum storage throughput.

Lenovo ThinkSystem DG and DM Series for Lenovo Hybrid AI 285 and 289 Platforms

For the Lenovo Hybrid AI 285 and 289 platforms, a scalable unit (SU) consists of four servers interconnected through a high-speed fabric, with each server containing eight GPUs. This architecture enables AI workloads to scale across multiple nodes and large GPU counts while accessing shared datasets, model artifacts, and checkpoints. As a result, external shared storage generally becomes an important part of the storage architecture, particularly for deployments requiring shared data access across nodes, users, workloads, or scalable units.

¹ For the Lenovo Hybrid AI 221 and 241 platforms, the units do not “scale” in the same sense as the 285 or 289 platforms since they do not have the East-West network integration for multi-node GPU communication, but this paper keeps the SU terminology for consistency.

Figure 3, Figure 4, and Figure 5 compare the read-throughput capabilities of the Lenovo ThinkSystem DG5200F, DM5200F, and DM7200F against the workload guidance presented in Section 3.3. The figures illustrate how the supported throughput range of each storage platform maps to different workload profiles and Hybrid AI deployment sizes, with the scaling factor defined as the scalable unit. Figure 3 covers throughput requirements up to 100 GB/s, Figure 4 extends the range to 400 GB/s, and Figure 5 extends it beyond 800 GB/s, where the network limitations discussed in Section 3.2 become a factor. For throughput requirements beyond the practical range of the Lenovo ThinkSystem solutions, the figures also indicate where Lenovo certified partner solutions may be appropriate.

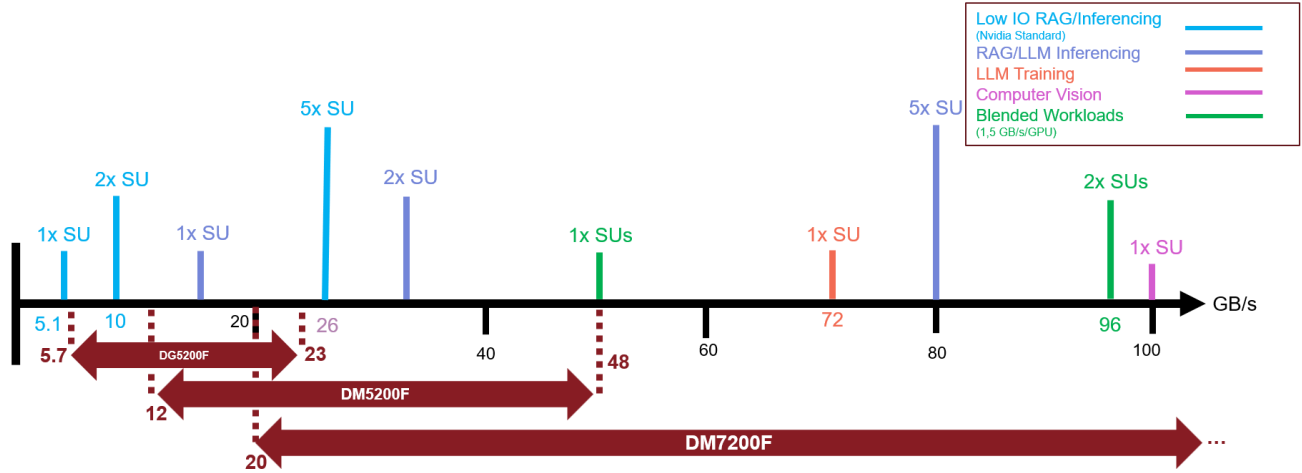


Figure 3 – HPS Read Throughput Guidance for Multiple 285/289 SUs up to 100GB/s

The scalable-unit counts shown in the figures are derived from the workload-specific throughput guidance introduced in Section 3.3. For the Hybrid AI 285 and 289 platforms, each scalable unit contains 32 GPUs, allowing the per-GPU throughput guidance to be translated into aggregate storage-throughput requirements and corresponding deployment sizes (i.e., $32 \times n \times \text{throughput per GPU}$). The figures assume a maximum-concurrency scenario in which all relevant workloads access shared storage simultaneously. In practice, concurrency is often lower because data-loading, inference, and checkpoint operations do not always occur at the same time. Consequently, actual storage-throughput requirements may be lower than the values shown in the figures.

Figure 3 shows that the DG5200F can support many Low IO RAG and inference deployments across a significant portion of the supported Hybrid AI 285 deployment range. In addition to satisfying the throughput requirements of these workloads, the DG Series provides a cost-efficient approach to scaling capacity through the use of high-density QLC flash media. This makes it particularly attractive when storage capacity and cost per terabyte are higher priorities than maximizing throughput.

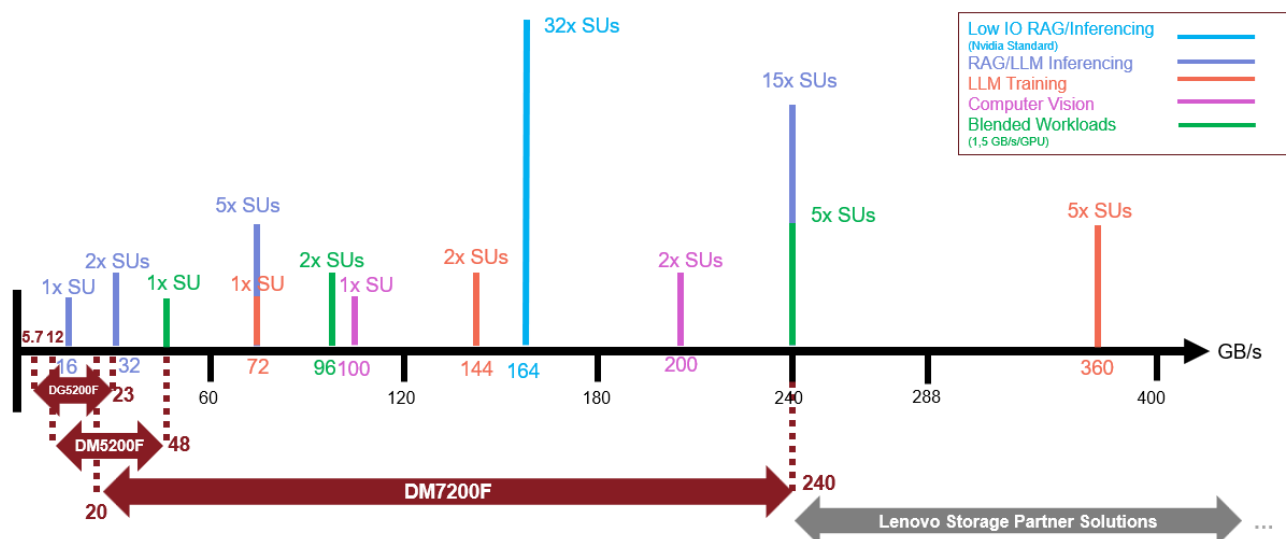


Figure 4 – HPS Read Throughput Guidance for Multiple 285/289 SUs up to 400GB/s

Figure 3 and Figure 4 illustrate that the DM7200F provides a substantially larger performance-scaling envelope than either the DM5200F or DG5200F due to its support for larger scale-out configurations (i.e., up to 12 HA pairs). As a result, the DM7200F is generally the preferred shared-storage starting point for larger Hybrid AI deployments and for workloads with higher throughput requirements. For Hybrid AI 285 deployments which scales up to 5 SU, the DM7200F is a suitable building block to support the throughput for Low IO RAG/Inferencing, [Higher IO] RAG/Inferencing, and Blended workload profiles across the full supported deployment range, subject to the sizing assumptions used throughout this paper.

The DM5200F occupies the space between the DG5200F and DM7200F. For deployments requiring less aggregate throughput or fewer HA pairs, it may provide a more economical alternative while still delivering the performance, data services, and scalability required by the workload.

Although the DG Series can also be used with Hybrid AI 285 and 289 deployments, its more limited throughput scalability should be carefully evaluated against both current workload requirements and anticipated future expansion. It is therefore best suited to environments where capacity efficiency is the primary design consideration and throughput requirements remain within the supported range.

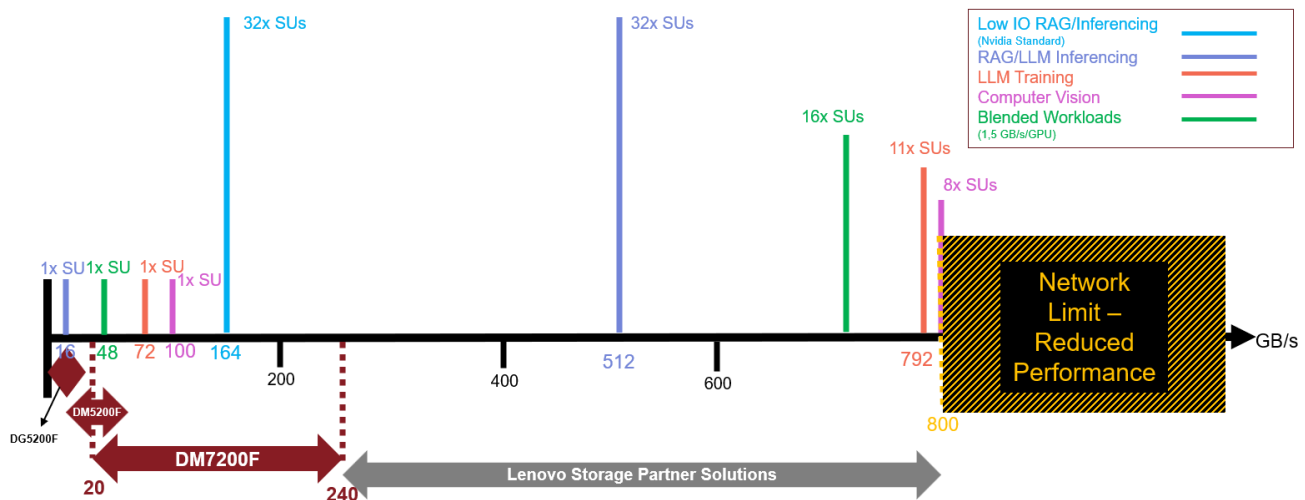


Figure 5 – HPS Read Throughput Guidance for Multiple 285/289 SUs up to 800GB/s and Beyond

Figure 5 highlights the network limits discussed in Section 3.2, which become visible as aggregate throughput approaches approximately 800 GB/s. Beyond this point, the network infrastructure, rather than the storage system itself, becomes the primary limiting factor. Storage platforms capable of delivering additional throughput may still provide value for future expansion, but the referenced Hybrid AI 285 and 289 network architectures cannot sustain higher aggregate bandwidth under full-concurrency conditions.

These limits represent a theoretical maximum-concurrency scenario in which all GPUs access shared storage simultaneously. In practice, workload scheduling, caching, and the natural staggering of I/O-intensive phases often reduce peak demand, allowing higher deployment density before network saturation becomes a practical concern.

When considering the maximum supported 32-SU Hybrid AI 289 deployment, the DM7200F remains an appropriate storage platform for Low IO RAG and inference workloads, while higher-throughput inference environments may require scale-out architectures provided through the Lenovo storage partner ecosystem. For blended workloads, training, and computer-vision use cases, the figures indicate the SU counts at which network limitations may begin to affect maximum-concurrency operation. Actual deployment limits will depend on workload behaviour, concurrency, and the effectiveness of caching and workload scheduling.

For highly data-intensive computer-vision workloads, the Hybrid AI 221 and 241 platforms may provide higher available network bandwidth per GPU than the Hybrid AI 285 and 289 platforms. As discussed in Section 3.2, this characteristic can make them attractive for specialized deployments requiring unusually high per-GPU storage throughput.

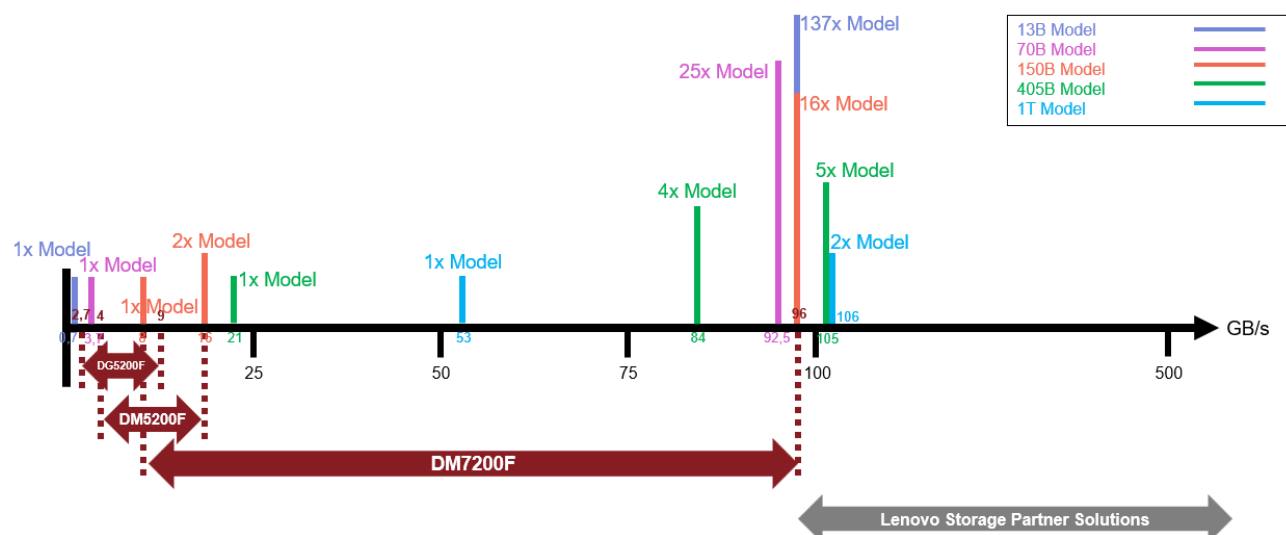


Figure 6 – HPS Write Throughput Guidance for 300 Sec Checkpointing with a 16-bit Precision Model

Figure 6 applies the checkpoint-sizing methodology introduced in Section 3.3 to representative model sizes ranging from 13 billion to 1 trillion parameters. The calculation assumes a checkpoint footprint of 16 bytes per parameter, consistent with the optimizer assumptions used throughout this paper, and a target checkpoint completion time of 300 seconds. The resulting write-throughput requirements are compared with the write-performance capabilities of the ThinkSystem DG5200F, DM5200F, and DM7200F to illustrate how these platforms align with different training scales and levels of checkpoint concurrency.

As with read-throughput sizing, actual write-throughput requirements may be reduced when checkpoint operations are naturally staggered across multiple training jobs. However, environments requiring predictable performance under peak concurrency should be sized according to the aggregate checkpoint demand.

The DM7200F can support training workloads ranging from smaller models at higher concurrency to larger models approaching one trillion parameters at lower concurrency levels. The DM5200F and DG5200F are more appropriate for smaller-scale training workloads with lower checkpoint-bandwidth requirements. For larger models, higher checkpoint concurrency, or future growth beyond the ranges illustrated in Figure 6, Lenovo partner storage solutions become increasingly relevant.

5.3 Storage Selection Criteria

The workload guidance and storage-positioning recommendations presented in this paper are intended to provide a storage starting point rather than prescribe a single solution. After identifying the throughput requirements of the workload, the selected architecture should be validated against capacity, scalability, data-management, operational, and resiliency requirements. The following criteria provide a practical framework for selecting an appropriate storage architecture for a Lenovo Hybrid AI deployment.

Throughput

Use the workload guidance presented in Section 3 and the sizing figures in Section 5.2 to identify solutions capable of meeting the required read and write throughput targets. Storage throughput should be evaluated

based on workload type, deployment size, expected concurrency, and checkpoint requirements rather than GPU count alone. Table 2 can then be used to compare the performance characteristics of the candidate storage platforms.

Capacity and Scalability

Evaluate both current and future capacity requirements, including datasets, model repositories, checkpoints, vector databases, and retained AI artifacts. The selected architecture should provide sufficient usable capacity together with an appropriate path for scaling capacity, performance, and shared-data access as the deployment grows. Multi-SU deployments, large training environments, and highly concurrent workloads may require architectures designed for large-scale operation. Table 2 can be used to compare the scalability characteristics of the candidate storage platforms and verify that the selected architecture provides sufficient capacity and performance headroom for future growth.

Data Services and Access Requirements

Consider whether the workload requires specific storage protocols or data services, such as file, block, or object access, snapshots, cloning, replication, centralized management, shared namespaces, tiering, or integration with an existing storage strategy. These requirements may influence whether a Lenovo ThinkSystem DM, DG, DE-based architecture, or a Lenovo partner storage solution is the most appropriate choice.

Operational and Resiliency Considerations

Consider the availability, recovery, and data-protection requirements of the deployment. AI environments may require capabilities such as snapshots, replication, backup integration, encryption, disaster recovery, and high availability to protect datasets, model artifacts, vector databases, and checkpoints. The importance of these requirements varies across environments. While some deployments may prioritize maximizing performance and capacity, others may place greater emphasis on business continuity, recovery objectives, and operational risk. These considerations should be evaluated alongside throughput and capacity requirements when selecting a storage architecture.

5.4 Recommended Storage Starting Points for Hybrid AI Deployments

The recommendations in this section are intended to provide an initial storage-positioning framework for common Hybrid AI deployment scenarios. They should not replace workload characterization or detailed sizing. The most appropriate storage architecture remains dependent on workload throughput requirements, capacity, data services, access protocols, resiliency objectives, operational requirements, and expected growth.

Based on the conclusions of the previous sections, we present general guidance on the positioning of Lenovo Storage solutions as part of a complete Hybrid AI solution. The guidance presented is chosen with a goal to be relevant to a broad range of deployments, with additional notes for the specialized cases relevant only to a smaller range of customers.

Customer Application	Hybrid AI Factory Offering	Recommended Storage Starting Point	Conditions Requiring Adaptation
Single to few node inference, RAG, and computer vision	Hybrid AI 221 or 241	Local NVMe storage	Consider DG5200F or DM5200F when shared storage, centralized data management, object storage, additional resilience, or larger datasets are required.
Scale out inference, RAG, and blended AI workloads	Hybrid AI 285	DM7200F shared storage	Consider DG Series storage when capacity requirements grow faster than throughput requirements. Consider multiple DM7200F systems or a scale-out storage architecture when throughput, concurrency, or future growth requirements exceed a single-system configuration.
Large-scale inference and blended AI workloads	Hybrid AI 289	Multiple DM7200F systems	Consider Lenovo partner storage solutions when throughput, namespace scale, multi-SU access requirements, or specialized data services exceed the practical range of the ThinkSystem DM Series.
Large-model training and training-intensive environments	Hybrid AI 289	DE Series-based architecture	Consider Lenovo partner storage solutions when checkpoint throughput, deployment scale, concurrency, or specialized data-management requirements exceed the practical range of a DE Series-based architecture.

Table 3 – General Storage Starting Points by Customer Application and Hybrid AI Factory Offering

In general, Lenovo ThinkSystem DM Series storage represents the preferred shared-storage starting point for most Hybrid AI deployments requiring enterprise data services and balanced performance. The DG Series provides a capacity-optimized alternative for deployments where cost efficiency and storage density are primary considerations. For large-scale training environments, highly concurrent workloads, or deployments requiring specialized data-management capabilities, DE Series-based architectures and Lenovo partner storage solutions may provide a more appropriate foundation.

The recommendations in Table 3 are intended as storage starting points rather than prescriptive designs. Specific configurations, including the number of HA pairs required, depend on the workload, deployment size, throughput requirements, capacity targets, and expected growth. The throughput-sizing guidance presented in

Section 5.2 and the capacity and scalability characteristics summarized in Table 2 should therefore be used to validate the final design. The adaptation criteria included in Table 3 highlight common situations where greater performance, capacity, resiliency, scalability, or specialized software capabilities may justify a different storage architecture. While no single recommendation is appropriate for every deployment, this framework is intended to streamline the storage-selection process and help identify suitable storage architectures for a broad range of Lenovo Hybrid AI environments.

6 Conclusions and Future Studies

Storage is a critical part of the end-to-end Lenovo Hybrid AI Factory architecture. Inference, RAG, agentic AI, computer vision, fine-tuning, and model training create different demands for capacity, read throughput, checkpoint write performance, latency, metadata operations, data services, and shared access. These requirements also change as the number of users, concurrent workloads, compute nodes, and Scalable Units increases. For this reason, storage should be selected and sized according to the customer's workload and operational objectives to ensure efficient utilization of infrastructure resources and enable future growth without architectural changes.

For Lenovo Hybrid AI Factory deployments, local NVMe storage may be sufficient for many smaller-scale Hybrid AI 221 and 241 environments. Where shared storage is required, the ThinkSystem DM Series provides the preferred starting point for enterprise inference, RAG, agentic AI, fine-tuning, and mixed workloads that require balanced performance and enterprise data services. The ThinkSystem DG Series provides a capacity-optimized alternative where storage density and cost efficiency are primary considerations. For large-model training and training-intensive Hybrid AI 289 deployments, the ThinkSystem DE Series can provide a high-performance block-storage foundation when combined with an appropriate shared file-system layer. As throughput, concurrency, deployment scale, or specialized data-management requirements increase beyond the practical capabilities of these architectures, Lenovo partner storage solutions provide additional options for scale-out high-performance storage.

Future studies will extend this guidance in three principal areas. First, additional work will examine the positioning and implementation of software-defined storage as Lenovo Hybrid AI Factory environments scale across multiple Scalable Units, including performance scaling, shared-file-system architectures, data-tiering strategies, and operational considerations. Second, future guidance will address AI data management across the data lifecycle, including ingestion, preparation, governance, cataloguing, indexing, vectorization, retrieval, and integration with RAG and agentic AI workflows. Finally, future studies will extend the workload analysis beyond throughput-oriented sizing considerations to include latency, IOPS, metadata performance, and other storage characteristics that can influence AI workload behaviour. These studies will complement the storage foundation presented in this paper and support the development of more complete, scalable AI data-platform architectures.

Abbreviations

AI	: Artificial Intelligence
FP	: Floating Point
BF	: Brain Floating Point
FC	: Fiber Connection
GPU	: Graphical Processing Unit
HPS	: High-Performance Storage
I/O	: Input-Output
IOPS	: Input-Output per Second
iSCSI	: Internet Small Computer System Interface
KPI	: Key Performance Indicator
LLM	: Large Language Model
NFS	: Network File System
NVMe	: Non-Volatile Memory Express
QLC	: Quad-Level Cell
QoS	: Quality of Service
RAG	: Retrieval Augmented Generation
S3	: Simple Storage Service
ONTAP	: Open Network Technology for Appliance Products
SCSI	: Small Computer System Interface
SDS	: Software Defined Storage
SMB	: Server Message Block
SU	: Scalable Unit
TCP	: Transmission Control Protocol
TLC	: Triple-Level Cell

References

- [1] Lenovo, “Lenovo Hybrid AI 221 and 241 Platforms Guide (Product Guide),” [Online]. Available: <https://lenovopress.lenovo.com/lp2313-lenovo-hybrid-ai-221-241-platforms-guide>.
- [2] Lenovo, “Lenovo Hybrid AI 285 Platform Guide (Product Guide),” [Online]. Available: <https://lenovopress.lenovo.com/lp2181-lenovo-hybrid-ai-285>.
- [3] Lenovo, “Lenovo Hybrid AI 289 Platform Guide (Product Guide),” [Online]. Available: <https://lenovopress.lenovo.com/lp2286-lenovo-hybrid-ai-289-platform-guide>.
- [4] NVIDIA, “HPS Architecture for SuperPOD DGX B300 Systems,” [Online]. Available: <https://docs.nvidia.com/dgx-superpod/reference-architecture/scalable-infrastructure-b300/latest/storage-architecture.html>.
- [5] J. B. Diederik P. Kingma, “Adam: A Method for Stochastic Optimization,” *arXiv:1412.6980v9*, 2014.
- [6] F. H. Ilya Loshchilov, “Decoupled Weight Decay Regularization,” *arXiv:1711.05101v3*, 2017.
- [7] The Linux Foundation, “PyTorch - User Guide,” 2026. [Online]. Available: <https://docs.pytorch.org/docs/2.14/generated/torch.optim.Adam.html>. [Accessed 22 09 2026].
- [8] Meta, “The Llama 3 Herd of Models,” *arXiv:2407.21783v3*, p. 14, 2024.
- [9] Lenovo, “Lenovo ThinkSystem DM3200F, DM5200F and DM7200F Unified Storage Arrays (Product Guide),” [Online]. Available: <https://lenovopress.lenovo.com/lp2075-lenovo-thinksystem-dm3200f-dm5200f-and-dm7200f-unified-storage-arrays>.
- [10] Lenovo, “Lenovo ThinkSystem DG5200 and DG7200 Unified Storage Arrays (Product Guide),” [Online]. Available: <https://lenovopress.lenovo.com/lp2074-lenovo-thinksystem-dg5200-and-dg7200-unified-storage-arrays>.

Authors

Sinan Atan is Senior AI Solution Architect at Lenovo. With over 15 years of international experience, he has established himself as a Digital Transformation expert, specializing in consulting, engagement, and the delivery of cutting-edge technologies through complex, multi-year programs. His career spans a wide range of domains and technologies, providing him with a comprehensive, end-to-end expertise in digitalization and AI technologies. Sinan holds both BSc and MSc degrees in Telecommunications and was previously a PhD candidate in the same field. Before joining Lenovo, he held different roles at Ericsson and Accenture, where he contributed to innovative solutions and strategic growth.

Conor Elrick is an HPC/AI Solutions Architect at Lenovo, specializing in the design, benchmarking, and optimisation of HPC and AI infrastructure. His work focuses on delivering high-performance solutions spanning compute, storage, networking, and accelerated AI systems. He holds a PhD in Theoretical Particle Physics from the University of Edinburgh, with a background in computational physics and large-scale simulation.

Aurelien Ortiz is an HPC/AI Software Architect working as part of the Solutions Architect Team in the ISG Offerings Group at Lenovo. Aurelien helps customers, partners, and internal teams design, deploy, and optimize software stacks for HPC and AI environments, covering infrastructure management, orchestration, storage, networking, and AI platforms. He has over 15 years of experience working with HPC clusters, large-scale storage systems, and AI infrastructure. He holds a PhD in Computer Science with a specialization in Distributed Systems from the University of Toulouse, France.

Hapsara Sukasdadi is a Senior Solutions Architect, currently driving Lenovo's AI technical engagements, focusing on architecting solutions in AI infrastructures. In this role, Hapsara collaborates closely with partners and ecosystem providers. Hapsara's primary mission is to deliver comprehensive solutions, encompassing design, planning, and integration across a spectrum of critical areas.

Additionally, the authors would like to thank the following people for their input and insights:

- Pierce Beary
- Kevin Dean

Document History

Version 1.0 : Initial publication with the scope of ThinkSystem DM/DG/DE Series mapping with different workload profiles.

Trademarks and Special Notices

© Copyright Lenovo 2026.

References in this document to Lenovo products or services do not imply that Lenovo intends to make them available in every country.

Lenovo, the Lenovo logo, ThinkCentre, ThinkVision, ThinkVantage, ThinkPlus and Rescue and Recovery are trademarks of Lenovo.

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel Inside (logos), MMX, and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

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

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used Lenovo products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-Lenovo products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by Lenovo. Sources for non-Lenovo list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. Lenovo has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-Lenovo products. Questions on the capability of non-Lenovo products should be addressed to the supplier of those products.

All statements regarding Lenovo future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local Lenovo office or Lenovo authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in Lenovo product announcements. The information is presented here to communicate Lenovo's current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard Lenovo benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.

Any references in this information to non-Lenovo websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this Lenovo product and use of those websites is at your own risk.