PURESTORAGE®

WHITE PAPER

# The Right Storage Architecture for Modern AI and HPC Workloads

Why choosing the right data storage platform to power large-scale AI and HPC workloads is essential.

# Contents

## Introduction to FlashBlade//EXA

Modern AI and high-performance computing (HPC) environments demand storage architectures that scale seamlessly, manage metadata efficiently, and sustain high throughput under extreme workloads. Existing storage approaches—whether traditional parallel file systems, or first-generation disaggregated models—struggle to keep up, leading to GPU underutilization, metadata bottlenecks, and operational complexity.

Traditional parallel file systems have worked well for sequential file access delivering predictable performance. But as AI evolved, the need to handle multimodal datasets has led to performance degradation at scale. Working with multimodal datasets frequently involves small random read/write operations, and during training these datasets often exhibit high-scale, concurrent data operations and have unpredictable IO patterns. Traditional parallel file systems fail to provide architectural flexibility to handle such diverse workloads while training on large language and context-based multimodal datasets. Metadata scaling becomes a critical bottleneck. These parallel architectures also introduce operational complexity to manage and maintain day-to-day operations to support the dynamic and evolving AI workloads. More recently, first-generation disaggregated storage attempted to separate compute from data, but rigid scaling requirements and network congestion between the architectural layers resulted in inconsistent performance.

These storage architectures are fundamentally not designed for AI factories, which are scalable, high-performance infrastructure that can transform raw data into AI-driven insights by integrating compute, storage, and networking to streamline data pipelines for model training and inference. They struggle with:

- Underutilized GPUs due to slow metadata retrieval and data access
- Inflexible scaling that forces metadata and storage growth to occur in lockstep
- High operational overhead due to large proprietary clients and networking complexity that requires ongoing tuning to maintain predictable performance

To power the next generation of AI innovation, a new storage paradigm is required—one that combines the flexibility of disaggregation with the efficiency of a proven scalable metadata core. This is where Pure Storage® FlashBlade//EXA™ comes in.

Unlike traditional storage architectures, FlashBlade//EXA ensures metadata and data are disaggregated without limits. This removes bottlenecks, enables seamless scaling, and ensures that AI and HPC workloads operate at full efficiency—without requiring manual performance tuning. By redefining metadata scalability and providing direct data access, FlashBlade//EXA allows organizations to maximize GPU utilization and reduce infrastructure overhead to accelerate innovation, optimize costs, and maximize infrastructure efficiency, FlashBlade//EXA is more than just another storage system—it is the foundation for AI factories looking to innovate without limits.

## The AI Revolution Is Here, but Can Storage Keep Up?

AI is evolving from experimentation to full-scale deployment, with massive training pipelines, real-time inference, and multimodal data integration. Organizations are rapidly adopting large language models/contexts (LLM/LLCs), AI-driven analytics, and mixed precision training for accuracy with frequent asynchronous checkpoints and restore operations to minimize disruptions in the data pipeline. Existing storage architectures, however, are not keeping up as AI continues to push infrastructure limits. Large foundation models process high-velocity data from internal and external sources, scaling from billions to trillions of parameters. As they shift from pre-training to post-training and inference, they demand storage with extreme throughput and metadata scalability.Without the right architecture GPU underutilization, unpredictable performance, and operational inefficiencies become unavoidable.

The right storage for AI requires:

- **Throughput at scale**: Petabytes of data must be processed instantly to feed GPU clusters.
- **Metadata-aware architecture**: AI workloads rely on billions of tiny files that demand ultra-low-latency access and massive metadata operations.
- **Seamless scalability**: AI factories require storage that expands dynamically without complexity.

Legacy architectures weren't built for this level of data intensity and concurrency. Without storage designed for AI, organizations face inefficiencies, higher costs, and slower innovation.

## Limitations of Existing Storage

As AI workloads scale, traditional storage architectures are breaking down under metadata constraints. AI models are multimodal (ranging from text to video) and often include billions of small files, require frequent fast metadata lookups, and demand real-time access to training datasets. Traditional storage approaches—built for sequential processing—cannot handle the parallel, multimodal, and metadata-intensive nature of AI workflows.

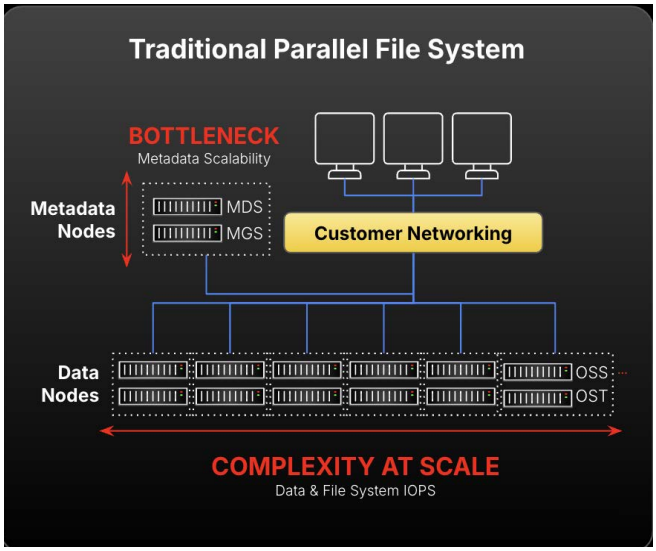
At the same time, metadata demands—ranging from rapid file searches to billions of small file operations per second—are increasing exponentially. Without a high-performance and reliable metadata architecture, storage systems suffer from hotspots, synchronization delays, and inefficient file system operations, leading to GPU idle time and unpredictable performance.

To understand these challenges, let's look at two of the common storage architectural approaches in the market and analyze why they fall short in modern AI and HPC environments.

### Traditional Parallel File Systems: Built for the Past, Not the Future

Parallel file systems were designed for sequential, high-throughput, and large-scale HPC simulations but struggle to support the demands of AI-driven workloads. Their architectural rigidity and reliance on centralized metadata services create significant limitations, particularly around metadata handling, small file performance, and overall complexity.

- **Metadata bottlenecks**: These systems rely on centralized metadata servers, which become hotspots under heavy AI concurrency. Managing metadata efficiently requires ongoing tuning and optimization, leading to operational overhead and inconsistent performance.
- **Small file performance**: AI workloads generate millions to billions of small files, yet parallel file systems are optimized for large, sequential data access. Caching small files becomes ineffective as data is frequently needed, leading to slow access from backend storage. This misalignment results in performance penalties, higher latency, and inefficient resource utilization.
- **Increased complexity**: Parallel file systems demand specialized deployment expertise and ongoing management. High-performance tuning, networking configuration, and filesystem optimizations require dedicated teams to maintain predictable performance at scale.
- **Rigid and heavy client requirements**: These architectures necessitate specialized, resource-intensive client software on data nodes, adding extra complexity and making scalability more cumbersome in large AI environments.

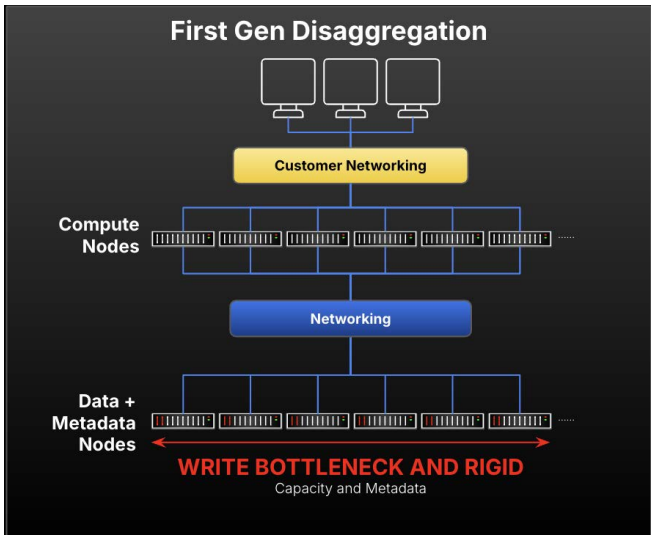



While parallel file systems once served HPC workloads well, their inability to efficiently scale metadata operations and handle AI's multimodal demands makes them unsuitable for today's highly dynamic AI pipelines.

## First-generation Disaggregation: A Step Forward, but Not Enough

To address some of the limitations of traditional parallel file systems, first-generation disaggregated storage architectures attempted to separate compute from data and metadata. While this approach made adoption easier and introduced greater flexibility in some areas, it also created new limitations and inefficiencies that impact AI performance:

- **Variable performance**: These architectures require metadata and data to scale together, forcing organizations to overprovision storage just to accommodate metadata demands. This rigid coupling leads to inefficiencies, unpredictable performance, and unnecessary infrastructure expansion.
- **Slow write performance**: Metadata caching mechanisms were introduced to improve efficiency, but they require write data to be staged in proprietary end-of-life storage technology before being persisted. This impacts write speeds, particularly for high-frequency AI model checkpoints and real-time inference workloads.




- **Network complexity**: Separating compute from metadata and data processing introduces additional networking layers. This results in higher latency, increased management overhead, and additional failure points, especially in distributed AI and HPC environments where ultra-fast, low-latency access is critical.

Although this approach had some advantages over traditional parallel storage architectures, they still suffer from metadata scaling challenges, write inefficiencies, and network complexity that prevent AI and HPC environments from achieving true operational efficiency.

## The Ideal Data Storage Platform for AI and HPC

AI and HPC workloads require a new generation of data storage—one that can seamlessly support massive metadata operations, extreme throughput, and dynamic scaling while maintaining simplicity and future-proof adaptability. The ideal data storage platform must eliminate metadata bottlenecks, scale effortlessly, and ensure predictable high performance across AI pipelines.

### Core Characteristics of an Ideal AI Data Storage Platform

To meet the demands of AI-driven environments, the next-generation data storage platform must offer:

- **Metadata-optimized performance**: Instant access to billions of small files and low-latency metadata access
- **Seamless disaggregation**: The ability to scale metadata and data independently, preventing overprovisioning and performance constraints
- **Extreme throughput**: High-bandwidth architecture capable of sustaining multi-terabyte-per-second workloads
- **Effortless scalability**: A storage platform that grows seamlessly, without requiring complex re-architecting
- **Simplicity at scale**: Eliminating the operational overhead associated with tuning, workload balancing, and manual optimizations
- **Unmatched TCO**: Scaling metadata and data independently and the improved performance density can reduce the total cost of ownership
- **Future-proofing**: A platform designed to evolve with next-generation AI models, foundational architectures, and emerging hardware innovations

### FlashBlade//EXA: The Best of Both Worlds

FlashBlade//EXA delivers the ideal AI data storage platform by integrating the strengths of traditional parallel file systems and first-generation disaggregated storage, while solving their fundamental weaknesses.

From traditional parallel file systems, FlashBlade//EXA incorporates the performance at scale by enabling independent scaling of metadata and data, ensuring high throughput and massive parallelism without constraints. It also maintains consistent write performance by persisting data directly to data nodes, optimizing efficiency for AI training and inference workloads.

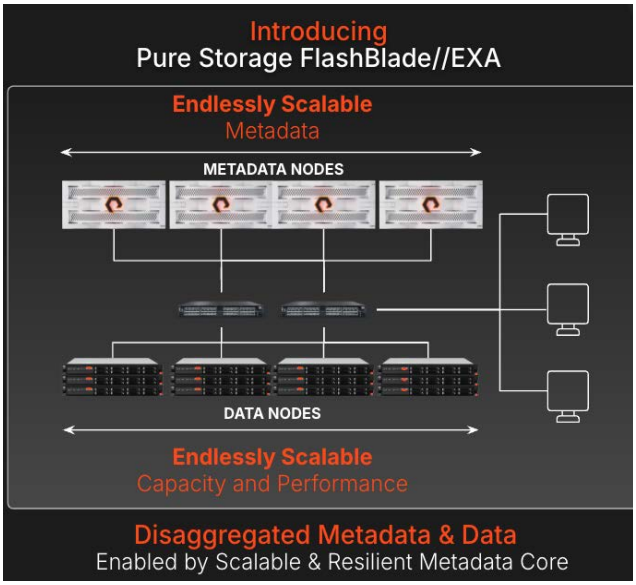
From first-generation disaggregated storage, FlashBlade//EXA retains ease of adoption by leveraging standard protocols and networking, allowing seamless integration into existing AI and HPC environments without specialized software dependencies.

Beyond these advantages, FlashBlade//EXA also integrates proven Pure Storage innovations, enhancing AI storage with a lightweight client architecture that simplifies management while efficiently powering data nodes. It leverages the signature simplicity of Pure Storage, making large-scale AI deployments easier to manage without the complexity of traditional systems. Additionally, its future-proof architecture is designed to scale and evolve alongside next-generation AI workloads, emerging frameworks, and new data models.

## FlashBlade//EXA Core Architectural Tenets

FlashBlade//EXA is built on a foundation of core architectural tenets that redefine AI and HPC storage:

- **Disaggregated metadata and data without compromise**: Unlike legacy approaches, FlashBlade//EXA allows metadata and data to scale independently, eliminating bottlenecks and inefficiencies while ensuring predictable, high-performance access to AI workloads.
- **Proven metadata core**: Built on the Pure Storage battle-tested metadata engine, FlashBlade//EXA delivers predictable, high-performance metadata operations at scale, ensuring uninterrupted AI model training and inference.
- **Unmatched simplicity**: AI storage should be easy to manage at scale—FlashBlade//EXA eliminates the complexity of legacy architectures, making scaling, tuning, and deployment effortless.
- **Future-proof infrastructure**: Designed to support exponentially growing AI models, multi-modal datasets, and evolving AI architectures, FlashBlade//EXA provides a scalable, adaptable platform for the AI-driven future.




## FlashBlade//EXA Architectural Overview

FlashBlade//EXA is built to power large-scale AI and HPC workloads by addressing the limitations of legacy storage architectures. It achieves unmatched performance, scalability, and simplicity through a disaggregated metadata and data architecture, ensuring consistent throughput, low latency metadata access, and operational efficiency at scale.

At its core, FlashBlade//EXA consists of two primary components that work together to deliver high-performance AI storage:

- **Metadata core**: The cluster of FlashBlade//EXA metadata nodes that provides scalable, low-latency metadata access independent of data.
- **Data nodes**: These are off-the-shelf servers in the initial release that enable extreme throughput and efficient storage of data.

## Metadata Core: High-speed, Scalable Metadata

The metadata core is the foundation of FlashBlade//EXA, responsible for managing metadata access and billions of metadata operations per second while ensuring high concurrency, low-latency access, and seamless scalability.

- **Disaggregated metadata access**: Unlike traditional architectures, metadata is handled separately from data, eliminating performance limits and contention issues.
- **Built on the FlashBlade® metadata engine**: Built on the proven massively distributed transactional database and key value store technology, it is optimized for massive parallelism, ensuring AI pipelines remain fully utilized.
- **Direct client communication**: Metadata operations instruct clients on where data is stored, reducing latency and improving efficiency.

## Data Nodes: Optimized for High-performance Storage and Throughput

The data nodes in FlashBlade//EXA are off-the-shelf servers that are validated and capable of handling large-scale AI and HPC workloads, delivering multi-terabyte-per-second performance.

- **Direct data access**: Clients can interact directly with data nodes, ensuring non-blocking, high-speed data retrieval.
- **Efficient write and read operations**: FlashBlade//EXA optimizes writes directly on the data nodes with double the number of reads, ensuring low overhead and predictable performance.
- **Flexible scalability**: Data nodes can be independently scaled, allowing organizations to expand capacity and performance as needed.

## The FlashBlade//EXA Difference

Existing storage architectures—whether traditional parallel file systems or first-generation disaggregated storage—struggle to meet the demands of AI and HPC workloads. Metadata bottlenecks, rigid scaling limitations, and complex management overhead create inefficiencies that slow AI innovation. FlashBlade//EXA is the reimagined data platform for AI and HPC. It removes complexity, scales effortlessly, and ensures that AI models run at full speed, without constraints.

## FlashBlade//EXA vs. Existing Storage Approaches: A Clear Advantage

| Key Factors | Existing Storage Approaches | FlashBlade//EXA |
|---|---|---|
| **Metadata Performance** | Bottlenecks due to centralized metadata servers or shared data and metadata node | Proven metadata core eliminates bottlenecks |
| **Scalability** | Rigid scaling—metadata and data must grow together or independent scaling with limitations | Independent scaling of metadata and data |
| **Throughput** | Inconsistent performance at scale and small file performance penalty | Multi-terabyte-per-second throughput |
| **Write Efficiency** | Caching impacts write performance | Direct data writes for predictable performance |
| **Complexity** | Requires tuning, workload balancing, and specialized clients | Simple to deploy, manage, scale, and adapt |
| **Future-Proofing** | Limited flexibility for evolving AI models | Built to scale with next-gen AI workloads and architectures on-demand. |
| **Operation Expenses** | Upgrades and hardware refreshes are disruptive and cumbersome. | Flexible metadata purchase options and term based licenses for data nodes software |

## Additional Resources

- Learn more about FlashBlade//EXA.
- Discover Pure for AI.





PS2735-01-en 03/25