The New Economics of AI Factory Efficiency

Solidigm, MinIO and Intel Put Object Storage Scaling to the Test

AI infrastructure is at an inflection point as models grow from billions to trillions of parameters, as inference workload multiply through agentic systems, and as enterprises race to build their own AI factories. What is the one bottleneck that links all of these trends? **Storage**. Not just storage capacity, but the ability to serve massive datasets at the throughput modern GPU clusters demand without blowing up the data center power budget or floor space in the process.

That's why Solidigm, MinIO, and Intel partnered to design and benchmark S3-compatible object storage, with a focus on performance and concurrency scaling. This effort leverages Solidigm 122TB QLC NVMe drives, MinIO AIStor software, and Intel® Xeon® 6 processors, showcasing how a tightly integrated stack can deliver the throughput and efficiency required for modern AI workloads at scale.

The AI storage Problem Nobody Wants to Talk About

Inference pipelines, especially those serving retrieval-augmented generation (RAG) and agentic workflows, require low-latency random reads across billions of objects. The traditional answer has been to throw more servers at the problem. But that approach collides with hard physical constraints: power capacity, rack space, cooling, and operational complexity. When your storage tier requires hundreds of nodes to hold your AI model, you've introduced a sprawl problem that undermines the economics of your entire AI factory.

What the industry needs is density without compromise:

- Massive capacity per node
- High throughput per drive
- Software that can extract every bit of performance from the hardware without adding operational overhead

Inside the Build

The cluster we tested is extremely dense. Each storage node delivers approximately **3PB of raw capacity** in a single 2U chassis. Our test was set up to measure S3 object storage GET & PUT throughput as client concurrency increases against the full 8-node storage cluster (~24PB). The baseline was a single client targeting all 8 storage nodes (~24PB), and we scaled up to 8 clients to saturate the cluster.

- **The client servers** featured 16× Solidigm™ D7-PS1010, 7.68TB TLC NVMe drives
- **The storage nodes** featured 24× Solidigm™ D5-P5336, 122TB QLC NVMe drives
- **Client servers** each used 1x400Gb NIC
- **Storage servers** each used 2x400Gb NICs

The client nodes were further over-provisioned with 1TB of DRAM, to ensure the load generator was never the bottleneck.

The storage servers are the fundamental building block of MinIO's ExaPOD reference architecture, which scales linearly from petabytes to exabytes (EB) using this same dense node design. Where ExaPOD defines the blueprint for exascale, this benchmark validates the performance reality at the single-pod level.

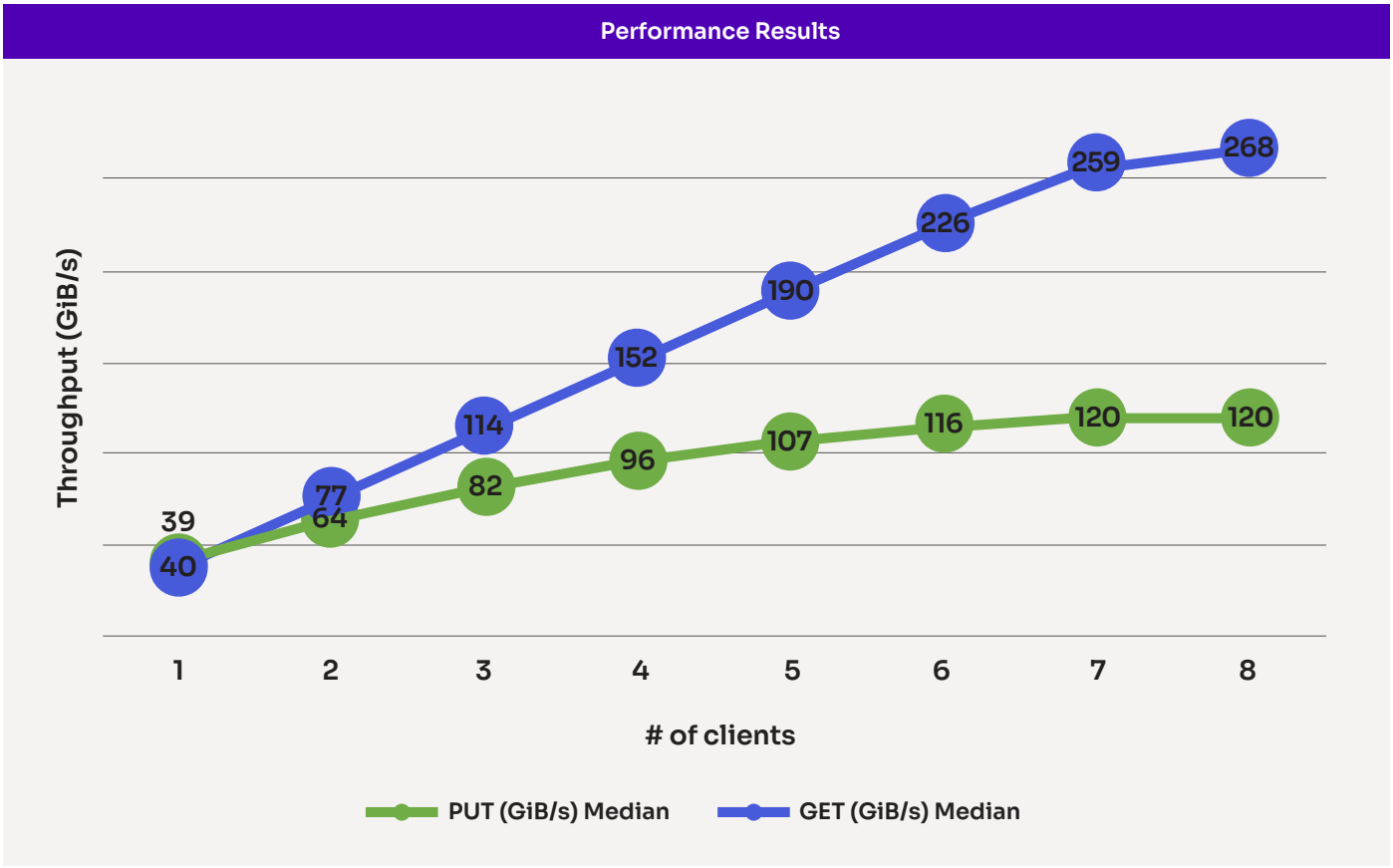
Reaching these numbers required deliberate co-tuning of the full stack, storage software, host OS and network stack, NIC offloads, CPU and IRQ topology, and switch fabric. Each solution brings its own tuning challenges for optimal performance, as evidenced by these performance-tuning results, which achieve 3x performance over hardware and software defaults. The types of adjustments made are listed at the end of this paper.

Performance Results

We ran the MinIO Warp benchmark to measure how S3 throughput scales with increasing client concurrency against a fixed 8-node storage cluster (~24PB). Each run lasted 15 minutes per phase (PUT and GET separately). Object size was set to 256 MiB with 32 concurrent connections per client.

Key Results

120 GiB/s	268 GiB/s	~24 PB	0
PEAK PUT	PEAK GET	CLUSTER CAPACITY 8 × 24 × 122 TB	ERRORS all 16 tests



Data Table

Clients	Total Concurrency	PUT (GiB/s)	GET (GiB/s)
1	32	40	39
2	64	64	77
3	96	82	114
4	128	96	152
5	160	107	190
6	192	116	226
7	224	120	259
8	256	120	268

Key Observations

- **GET throughput scales through 8 clients with headroom remaining.** GET scales smoothly from 38.6 GiB/s (1 client) to a peak median of 267.8 GiB/s at 8 clients, with the fastest 1-second bucket reaching 272.4 GiB/s. Per-client efficiency holds at ~37 to 38 GiB/s per client through 7 clients, then tapers only slightly at full saturation. Notably, the curve continues to rise through the 7 to 8 transition rather than flattening; the cluster is not yet read-bound at this configuration.
- **PUT scales smoothly and saturates near 120 GiB/s.** PUT throughput climbs cleanly with concurrency 40, 64, 82, 96, 107, 116, 120, 120 GiB/s, with no inflection point or stall. The write ceiling is reached at 7 clients and holds flat at 8. Low concurrency write performance is particularly strong. A single client sustains 40 GiB/s of PUT throughput, demonstrating that the cluster does not require high parallelism to deliver useful write bandwidth.
- **8 storage nodes deliver ~24PB in a single cluster.** Each node contributes ~3PB of raw capacity using 24× Solidigm D5-P5336 122 TB QLC NVMe drives. At 267.8 GiB/s GET, the cluster delivers over 11 GiB/s per petabyte of raw capacity, demonstrating that density does not come at the expense of throughput.
- **Zero errors across all 16 tests.** All 16 phases (8 PUT + 8 GET) completed cleanly across 240 minutes of continuous I/O with no object verification failures, critical for production confidence when deploying at this density. First-byte latency stayed at 3 to 5 ms median across the full sweep, with the worst 1-second bucket at 8-client GET still within ~11% of the median, indicating well-behaved tail latency under saturation.

From Benchmark to Blueprint

These results aren't academic. They map directly to the MinIO ExaPOD reference architecture, which scales to 1EB usable across 640 servers in 32 racks, delivering up to 19.2 TB/s aggregate throughput. The current ~24PB, 8-node deployment provides a representative small-scale instantiation of the ExaPOD design, demonstrating the same architectural principles, density, and core performance characteristics in a more compact footprint.

For organizations planning their AI infrastructure, this benchmark answers a critical question. Can you achieve exascale economics at petascale entry points? The answer is yes. The same hardware, the same software, the same operational model.

What This Means for AI Builders

The compute announcements across the industry are impressive. But compute without storage is a GPU waiting on data. As you plan your AI factory, consider the storage math:

- **Fewer nodes** mean less power, cooling, and rack space; critical when data center capacity is constrained
- **S3-native** means your training frameworks, inference pipelines, and lakehouse analytics all speak the same protocol
- **Linear scaling** means you buy what you need today and expand predictably tomorrow
- **QLC economics** means the cost-per-terabyte curve finally bends in your favor at the capacity points that matter

Solidigm and MinIO are proving that the densest storage doesn't have to be the slowest. It can be the foundation of your entire AI data platform.

More Information

The World's Highest Capacity PCIe SSD

At the heart of this density story is the Solidigm™ D5-P5336,a hyper-scalable, cost-effective solution for AI and data-intensive workloads. With industry-leading capacity up to 122.88TB, the D5-P5336 is architected to efficiently accelerate and scale with the increasingly massive datasets found in widely-deployed, modern read-intensive workloads.

This testing was conducted in the Solidigm AI Central Lab, a purpose-built facility that brings together storage and AI capabilities to perform cutting-edge research and improve bottom-line results. The lab features high-performance GPUs, 800 Gbps Ethernet networking, and extensive Solidigm SSD infrastructure, all designed on reference architectures that mirror what hyperscalers and enterprises are deploying in data centers worldwide, making the findings broadly applicable to customer environments.

The AI Central Lab hosts what Solidigm believes to be the densest storage test cluster ever built: 192 Solidigm D5-P5336 SSDs packing 23.4PB into just 16U of rack space. The lab also has the capability to collect telemetry, creating a detailed picture of how system resources are used and where bottlenecks exist, allowing Solidigm and its collaborators to recommend optimizations to improve performance and power efficiency.

Exascale AI data store

MinIO AIStor delivers unmatched enterprise performance, scale, agility, and economics for AI data, agentic computing, and analytics. For technical details on the ExaPOD reference architecture, read MinIO's full announcement: Introducing MinIO ExaPOD: The Reference Architecture for Exascale AI

Intel® Xeon® Processors

Intel® Xeon® processors provide the foundational compute layer that enables this high-density object storage architecture to operate at scale. In this reference design, the Intel 6781P was utilized, which is optimized for I/O-intensive workloads (136 PCIe5 lanes), and delivers the core count, memory bandwidth, and PCIe lane capacity required to efficiently drive dozens of high-capacity NVMe SSDs per node while sustaining line-rate networking across 400GbE interfaces. Advanced features such as Intel® Advanced Vector Extensions (AVX-512), large last-level caches, and integrated acceleration for data movement and compression help minimize CPU overhead during S3 operations, allowing more cycles to be dedicated to application-level throughput. In this configuration, Xeon processors ensure that storage performance scales linearly with concurrency, eliminating CPU bottlenecks and enabling the full potential of Solidigm QLC media and MinIO AIStor software.

Test System Specifications

Storage Nodes (8): Supermicro SYS-212H-TN, single Intel Xeon 6781P (80C), 24× Solidigm D5-P5336 122.88 TB QLC NVMe (~2.95 PB raw per node), 504 GB DDR5, 2× 400G NICs

Clients (8): Supermicro SYS-212H-TN, single Intel Xeon 6781P (80C), 16× Solidigm D7-PS1010 7.68 TB TLC NVMe, 16×64 GB DDR5-6400 RDIMM (1 TB), 2×400G NICs

Benchmark: MinIO Warp, 256 MiB objects, 32 concurrent connections per client, S3 PUT/GET, 15 min per phase, TCP transport (RDMA disabled)

Tuning / Configuration:

CPU/IRQ: SMT disabled, pinned NIC interrupts, irqbalance masked, p-states / deep C-states disabled.

NIC: 8k RX/TX buffers, 63 channels per port, hardware GRO, CQE compression, NAPI hard-IRQ deferral.

Network stack: BBR congestion control, 64 MiB socket buffers, fq qdisc enabled on xmit queues, deep netdev backlog, TCP timestamps disabled.

Switch: Leaf switches in lossy mode, ECN marking, 70/30 DWRR scheduling.

MinIO stack: 24 erasure sets, 8 drives each (EC:4 - four data / four parity per object), background scanner set to slowest mode, CRC32C bitrot detection enabled.

Test facility: Solidigm AI Central Lab, Rancho Cordova, CA

Disclaimers

All product plans, roadmaps, specifications, and product descriptions are subject to change without notice.

Nothing herein is intended to create any express or implied warranty, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, or any warranty arising from course of performance, course of dealing, or usage in trade.

The products described in this document may contain design defects or errors known as "errata," which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your Solidigm representative or your distributor to obtain the latest specifications before placing your product order.

For copies of this document, documents that are referenced within, or other Solidigm literature, please contact your Solidigm representative.

All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase.

Performance results are based on testing as of dates shown in the configurations and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure.

Solidigm or Intel optimizations, for Solidigm or Intel compilers or other products, may not provide optimized performance to the same degree for non-Solidigm or Intel products. Solidigm or Intel technologies may require enabled hardware, software, or service activation.

Your costs and results may vary.

Solidigm does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Some results have been estimated or simulated using internal Solidigm analysis or architecture simulation or modeling, and provided to you for information purposes only. Any differences in your system hardware, software or configuration may affect your actual performance.

© Solidigm. SOLIDIGM and the Solidigm "S" logo are trademarks of Solidigm Inc., registered in the United States, People's Republic of China, Singapore, the European Union, the United Kingdom, Mexico, and other countries. Other names and brands may be claimed as the property of others.