



NEWS RELEASE

NetApp Removes Storage Bottleneck for AI Factories

2026-09-29

New architecture delivers fastest storage on the planet, designed to maximize GPU utilization and architected for up to 100 TB per second aggregate throughput

SAN JOSE, Calif.--(BUSINESS WIRE)-- NetApp® (NASDAQ: NTAP), the Intelligent Data Infrastructure company, today announced NetApp Novus™, the fastest storage on the planet designed for AI infrastructure providers operating large GPU environments. Built to support a zettabyte-scale file system, this next-generation architecture will allow AI factories to keep millions of GPUs continuously fed with data, maximizing the efficiency, utilization, and profitability of their investments.

Neoclouds and GPU-as-a-Service providers are building AI factories at a scale that traditional storage architectures were never designed to support, creating bottlenecks that reduce GPU utilization and increase infrastructure costs. Under traditional architectures, GPU utilization can drop below 30 percent when AI factories cannot feed enough data to their GPUs to generate the results they need and make a return on their investment.

"AI factories struggle and GPU economics collapse when data can't keep up," said Syam Nair, Chief Product Officer at NetApp. "HPC-era storage alone cannot serve AI Factory scale — the industry needs a true architecture built for it from the ground up. AI factories have long been searching for infrastructure that can scale at AI speed. Leveraging the power of the NetApp Platform, Novus delivers 100 TB/s of unlimited scale-out performance, leapfrogging the market to serve as the new high-performance, AI-scale storage architecture that unlocks the performance of millions of GPUs and delivers the return on investment needed to sustain the next wave of AI innovation. We've just delivered the fastest storage on the planet, that is purpose built for the AI Era."

NetApp Novus addresses the challenges faced by AI factories by providing a next generation architecture that is designed to eliminate scale limitations in separating metadata from the data path so performance, capacity, and concurrency can scale independently under a single name space using industry-standard NFS. The initial Novus releases combines NetApp Novus Data Director, the metadata software, running on qualified Supermicro infrastructure with NetApp ONTAP® data services delivered through AFF A90 systems. Together they create a disaggregated architecture that is built to exceed throughput of 100TB per second and support high utilization of hundreds of thousands of GPUs while preserving the enterprise data services customers trust from ONTAP. This architecture is designed to meet the demands of multi-tenant, cloud-scale AI factories, is orderable today, and provides a path to software-defined deployments over time while maintaining the same namespace, operating model, and data foundation.

Customers that adopt the Novus architecture can realize the following benefits:

- **Keep Every GPU Productive:** Novus is designed to maximize GPU utilization by separating metadata services from the data path, reducing storage-related bottlenecks, and keeping AI infrastructure fed with high-performance data.
- **Scale Without Fragmentation:** Novus enables billions of files and massive AI datasets to grow under a single namespace, allowing metadata, bandwidth, and capacity to scale independently without creating storage silos or operational complexity.
- **Improve AI Infrastructure Efficiency:** Novus lets operators scale only the resources required for each workload, helping reduce power, space, and infrastructure overhead while preserving the resiliency, security, and multi-tenant capabilities needed for shared AI environments.

"As AI infrastructure providers scale GPU environments, the economics increasingly depend on keeping compute resources fully utilized," said Mike Leone, Vice President & Principal Analyst at Moor Insights & Strategy. "Storage architectures that can't deliver data at cloud scale can become a limiting factor for training performance, tenant isolation, and return on infrastructure investment. NetApp Novus directly addresses that limit by combining disaggregated metadata management with high-performance data services, giving neocloud and GPU-as-a-service providers a more scalable foundation for AI factory operations."

"Omdia audited NetApp Novus demonstrate near-linear performance scaling as ONTAP clusters were added to a single global namespace," said Tony Palmer, Chief Analyst with Omdia. "Based on the rigorous testing we observed, our modeling projects that NetApp Novus disaggregated metadata and data architecture can scale up and out to 100TBps sequential read throughput with dozens of exabytes of effective capacity, and beyond. We are confident in our projections that the solution can scale to the needs of zettascale deployments to fuel the most demanding and data hungry AI workloads."

"Our longstanding collaboration with NetApp is about bringing compute, networking and data together to make complex infrastructure simpler for customers," said Jeremy Foster, Senior Vice President and General Manager, Cisco Compute. "As AI pushes infrastructure to new levels of scale and speed, the real differentiation isn't the GPU — it's the data and how customers use it. NetApp Novus, together with the rest of the stack, helps customers put that data to work and drive the outcomes that matter to their business."

Statements by NetApp about unreleased offerings and future plans are for informational purposes only, are subject to change without notice, and should not be relied upon for purchasing or other decisions. Such statements do not constitute a commitment, obligation, guarantee, or warranty of any kind by NetApp, including about availability, functionality, pricing, or timing.

Additional Resources

- [NetApp Novus. Keep Every GPU Fed, Productive and Ready to Scale](#)
- [A New Data Architecture for AI Factories at Scale](#)

About NetApp

For more than three decades, NetApp has helped the world's leading organizations navigate change – from the

rise of enterprise storage to the intelligent era defined by data and AI. Today, NetApp is the Intelligent Data Infrastructure company, helping customers turn data into a catalyst for innovation, resilience, and growth.

At the heart of that infrastructure is the NetApp data platform – the unified, enterprise-grade, intelligent foundation that connects, protects, and activates data across every cloud, workload, and environment. Built on the proven power of NetApp ONTAP, our leading data management software and OS, and enhanced by automation through the AI Data Engine and AFX, it delivers observability, resilience, and intelligence at scale.

Disaggregated by design, the NetApp data platform separates storage, services, and control so enterprises can modernize faster, scale efficiently, and innovate without lock-in. As the only enterprise storage platform natively embedded in the world's largest clouds, it gives organizations the freedom to run any workload anywhere with consistent performance, governance, and protection.

With NetApp, data is always ready – ready to defend against threats, ready to power AI, and ready to drive the next breakthrough. That's why the world's most forward-thinking enterprises trust NetApp to turn intelligence into advantage.

Learn more at www.netapp.com or follow us on [X](#), [LinkedIn](#), [Facebook](#), and [Instagram](#).

NETAPP, the NETAPP logo, and the marks listed at www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.

Media Contact:

Kenya Hayes

NetApp

kenya.hayes@netapp.com

Investor Contact:

Kris Newton

NetApp

kris.newton@netapp.com

Source: NetApp