SOLUTION BRIEF

NETAPP NOVUS: AMBITION IS YOUR ONLY BOUNDARY SCALE BEYOND LIMITS WITH THE FASTEST STORAGE ON THE PLANET

Executive summary

NetApp® Novus is a revolutionary scale-out storage architecture for the next era of AI infrastructure, built for the speed, concurrency, and namespace scale required by neoclouds, sovereign AI providers, hyperscalers, and large enterprises building gigawatt-scale AI factories. GPUs are limited not only by model size or network bandwidth, but by how fast the data infrastructure can keep them fed. NetApp Novus removes that bottleneck with a disaggregated metadata architecture, direct parallel data access, and a common namespace that helps GPU environments stay productive as they scale.

An idle GPU is not just underutilized compute. At AI-factory scale, it is capital, power, and floor space producing heat instead of AI output. NetApp Novus changes the storage equation by letting many ONTAP-based data resources operate as one data environment, while metadata services scale independently to support concurrency and namespace growth. Built for zettabyte-scale file systems and AI infrastructure efficiency, NetApp Novus helps operators keep GPUs fed, reduce data-related stalls, scale without fragmentation, and improve the economics of large-scale AI.

The problem

Picture thousands of GPUs, racked and ready, consuming scarce capital, power, and floor space. They are among the most expensive assets in the modern data center, but they only create value when they are doing useful work: training models, tuning them, serving inference, or advancing agentic AI workflows. At gigawatt-scale AI infrastructure, the limiting factor is not always the model, the GPU, or the network. Increasingly, it is whether the data infrastructure can keep every GPU fed. Every second a GPU waits on data is lost productivity and wasted investment. In many large AI environments, data bottlenecks can leave expensive GPU infrastructure significantly underutilized even as demand for AI output continues to grow. The issue is not simply that teams need "faster storage." The deeper problem is architectural: traditional storage systems were not designed to scale metadata concurrency, bandwidth, capacity, and namespace operations independently for tens of thousands of highly parallel GPU clients.

Every storage system has a ceiling set by its controllers and its interconnect. Push beyond that ceiling with more independent systems, and operators trade one bottleneck for another: more namespaces, more placement decisions, more remounts, more operational silos, and more ways for capacity or performance to become stranded. What should behave like one data environment becomes a collection of storage islands that application teams and infrastructure teams have to stitch together themselves.

That fragmentation becomes an operator-efficiency problem as much as a performance problem. AI teams should not have to become storage engineers to decide where data lives, which cluster to mount, how to rebalance workloads, or how to coordinate checkpoints across fragmented infrastructure. At AI-factory scale, operational complexity becomes a scaling limit of its own.

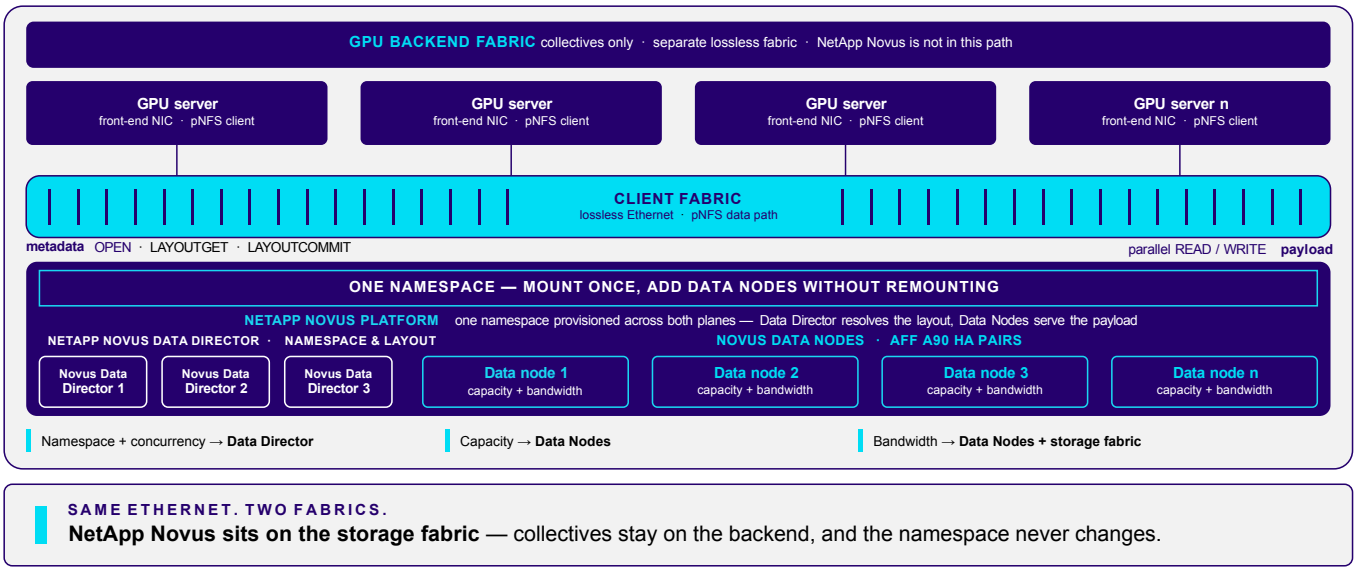
Even if teams engineer their way past one ceiling, they hit another: power. Every watt spent forcing the data infrastructure to keep pace is a watt not available for GPUs doing productive work. Overbuilding storage to compensate for architectural limits can increase power, cooling, footprint, and cost at exactly the moment AI operators are trying to maximize output within a finite power envelope.

When storage cannot keep pace, GPUs sit idle, power is wasted, and racks built to produce intelligence turn scarce capital, power, and floor space into wasted heat instead of AI output. In large-scale AI, utilization is not just a performance metric. It is the economic model.

The solution

NetApp Novus decouples the metadata layer from the storage layer. Metadata tracks data layout across federated storage systems and hands each GPU client its layout, then gets out of the way. The client reads and writes directly to the underlying storage, with no round trip through metadata for every I/O. That's what breaks through the limits of a single storage system. The storage layer itself runs on proven NetApp ONTAP®, serving data at line rate to eliminate that bottleneck. This approach allows you to grow performance granularly and add performance linearly. Add a storage system and you add the performance capacity to the federated pool. Metadata scales on its own path too so that you can tailor your scale independently from your storage layer. This means no remounts and no fragmentation.

None of this requires a new client. NetApp Novus talks to GPU nodes over the same in-kernel NFSv4.2 and pNFS Flex Files stack most fleets already run, with nconnect for multi-connection throughput and GPUDirect Storage for direct GPU-to-storage I/O that skips the CPU entirely. There's nothing proprietary to build, qualify, and roll out across a fleet of thousands of GPU servers — NetApp Novus mounts like NFS, and every client already knows how to talk to it.



NetApp Novus Architecture: NetApp Novus separates metadata coordination from data movement. GPU clients get data layouts from the NetApp Novus Data Director metadata server, then perform I/O directly against the ONTAP-based storage layer, allowing metadata concurrency and storage throughput to scale independently.

How NetApp Novus is delivered

NetApp Novus is delivered as an integrated storage solution that combines a disaggregated metadata layer with ONTAP-based data nodes. In the initial deployment model, ONTAP running on NetApp A90 systems provides the persistent data-serving foundation, while pNFS metadata server services run on qualified server infrastructure. The metadata layer presents the namespace and layout to clients; the ONTAP data nodes serve the data path.

This approach gives customers a curated, high-performance deployment model for AI and GPU-as-a-Service environments while preserving the architectural flexibility of NetApp Novus. The current delivery model is focused on ONTAP-based A90 data nodes and qualified pNFS metadata server infrastructure. Future deployment models are expected to expand the architecture over time, including software-defined options where both metadata services and ONTAP-based data services can run on qualified third-party hardware. Refer to the latest NetApp Novus product documentation for supported configurations, qualification details, and deployment requirements.

Conclusion

AI factory economics live or die on utilization rates. Your competitors have the same GPUs and the same AI models, but what they don't have is a storage layer that keeps those GPUs productive. NetApp Novus removes the bottleneck and federates many systems into zettabyte-scale file system infrastructure, finally letting your biggest infrastructure investment generate tokens and outcomes instead of sitting idle. With NetApp Novus, every GPU stays productive.

Talk to a NetApp Novus architect to map throughput and node count to your GPU target.

KEY BENEFITS

- **Keep every GPU productive.** Idle GPUs are the most expensive line item you can't sell. NetApp Novus removes the storage bottleneck so more compute runs at full capacity, and more of it converts to billable hours.
- **Scale throughput without fragmenting the namespace.** NetApp Novus separates metadata coordination from data movement, so namespace operations do not block the data path. Add ONTAP-based data resources to expand aggregate throughput beneath a common namespace, while scaling the metadata layer independently to support more clients, tenants, and file operations.
- **Improve AI infrastructure efficiency.** NetApp Novus is designed to help operators scale within real constraints for power, floor space, cost, and operations. The ONTAP-based data layer provides an efficient storage foundation, while the software-defined metadata layer lets teams add metadata capacity only where it is needed. Future software-defined data services extend that flexibility so operators can build around the configurations that fit their infrastructure.
- **Secure multi-tenancy and QoS.** Run many tenants on one NetApp Novus deployment, isolated from one another, with tenant-level encryption, policy-based resource allocation, and granular security controls for both people and AI agents.
- **Standards based clients.** No proprietary agents to build, qualify, or patch. NetApp Novus mounts over the in-kernel NFS client already running on your GPU nodes, so a fleet of thousands is production ready without a client rollout project.

This Solution Brief includes forward-looking statements regarding anticipated product features and functionality. These statements reflect our current plans and expectations and are subject to change. Actual product features, functionality and timing may differ from those described, and we undertake no obligation to deliver any specific feature or functionality except as required by applicable law.

Contact Us

NetApp

About NetApp

NetApp is the Intelligent Data Infrastructure company, combining unified data storage, integrated data services, and CloudOps solutions to turn a world of disruption into opportunity for every customer. NetApp creates silo-free infrastructure, harnessing observability and AI to enable the industry's best data management. As the only enterprise-grade storage service natively embedded in the world's biggest clouds, our data storage delivers flexibility across every environment. Our data services create a data advantage through superior cyber resilience, governance, and application agility, while our CloudOps solutions provide continuous optimization of performance and efficiency. No matter the data type, workload, or environment, with NetApp you can transform your data infrastructure to realize your business possibilities. www.netapp.com

© 2026 NetApp, Inc. All Rights Reserved. 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. SB-4532-1026

5-17oct26