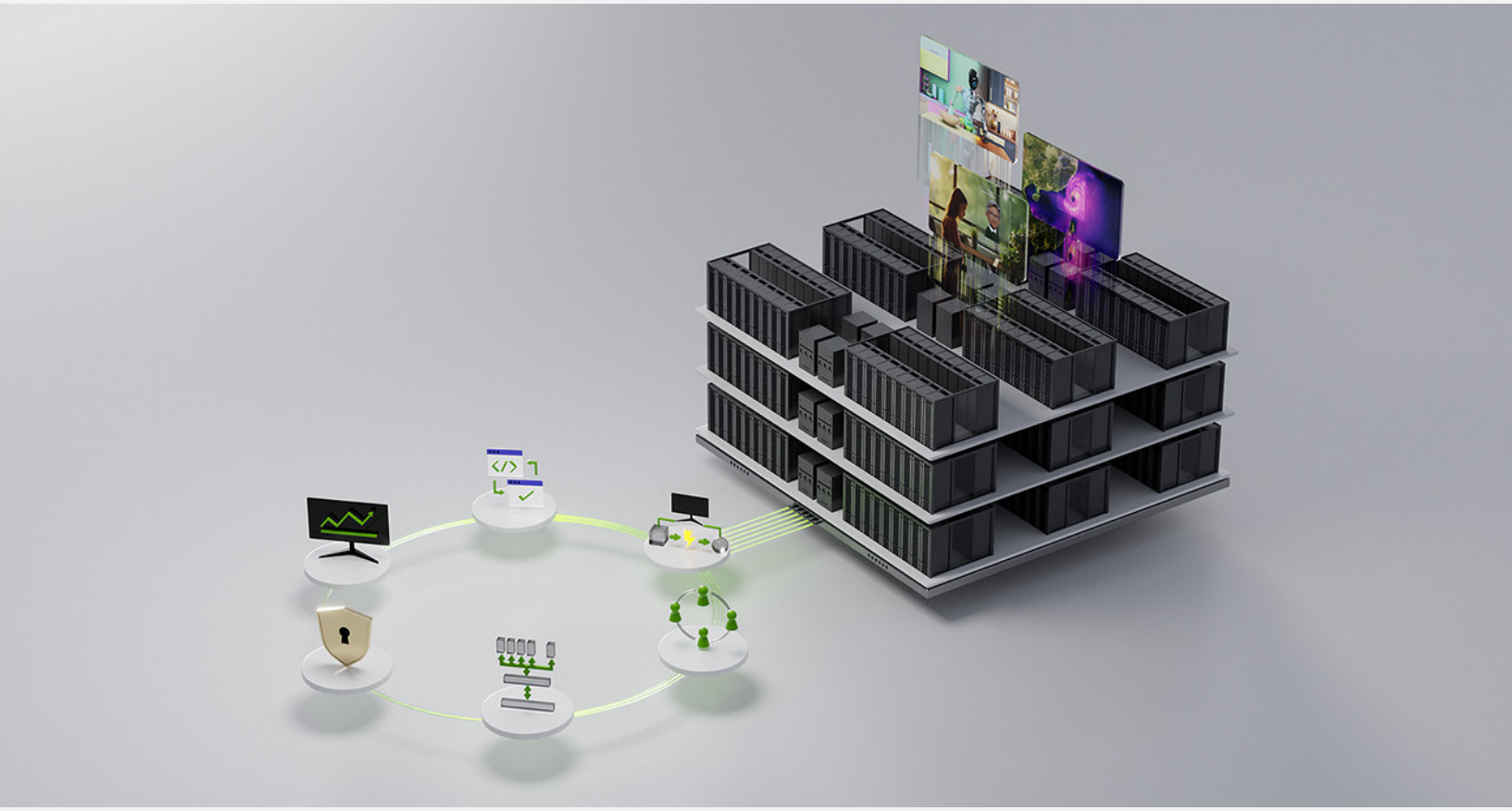


NVIDIA Success Story

DDN & NVIDIA Collaborate to Leverage NVIDIA's DGX SuperPOD™ Reference Architectures for AI Factories

ddn | NVIDIA

NVIDIA has created the AI Factory for the age of AI and provides solutions that deliver breakthrough performance for workloads at any scale, driving business decisions in real-time and resulting in faster time to value.

DDN serves as the certified storage and data-intelligence layer of modern AI Factories, keeping GPUs fed, delivering predictable SLAs, and scaling elastically from terabytes to exabytes. In production deployments this translates to 90–95% GPU utilization (up to 99%) versus the 40–60% that many competitors provide.

Built from the ground up for enterprise AI, the NVIDIA DGX™ platform combines the best of NVIDIA software, infrastructure, and expertise. Pre-validated with NVIDIA AI Enterprise, Blackwell, and GB200 NVL72 designs—and backed by an 8-year co-engineering partnership with NVIDIA—DDN removes POC-to-production delays and ensures day-one readiness for new GPUs and fabrics. By consolidating the power of an entire data center into a single AI Factory, NVIDIA has revolutionized how complex machine learning workflows and AI models are developed and deployed by an enterprise or AI cloud provider.

The Challenge

With the explosive growth of AI applications, an entirely novel approach to the data center was necessary. NVIDIA required a high-capacity, reliable, and easy-to-integrate AI data storage and management solution to not only deliver supercomputing services to meet complex demands from its internal developers but also create the blueprint for deploying turnkey supercomputers for their new breed of AI customers.

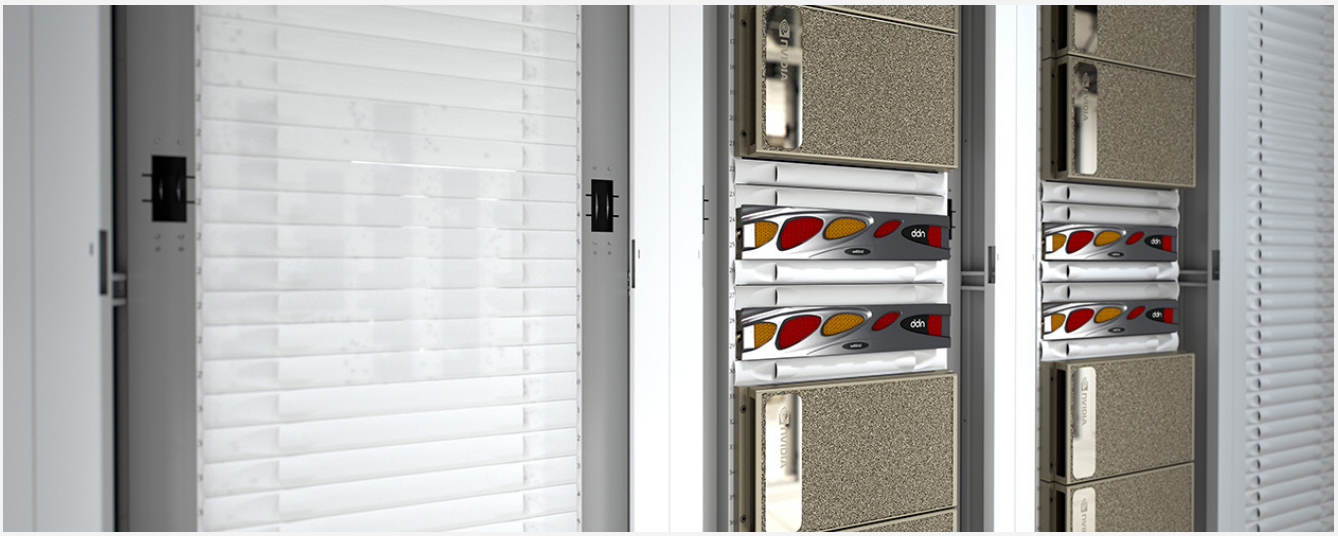
Beginning with the initial supercomputer collaboration, Selene, NVIDIA wanted to build a system powerful enough to train the AI models their colleagues were building for autonomous vehicles and general purpose enough to serve the needs of any deep- learning researcher. As the size and complexity of AI models continued to grow, NVIDIA incorporated new technologies into subsequent systems to fulfill the ongoing goal of creating best-in-class infrastructure for all AI workloads, and they needed storage solutions that could keep up.

NVIDIA required a reliable data storage platform and provider partner that could handle large computational problems distributed across hundreds of systems operating in parallel using a standard set of scalable storage building blocks. To reduce complexity, these storage building blocks needed to supply excellent performance for both reads and writes and scale out without needing to re-architect to accommodate future growth.

"What's needed is data center-scale computing, so AI models and datasets can be processed across many systems in parallel, enabling applications to train in hours instead of weeks."

~ Tony Paikeday

Senior Director of Product Marketing | NVIDIA



The Solution

Since 2018, DDN and NVIDIA have run extensive validation testing and collaborative development projects to create an optimal infrastructure architecture for AI workloads and applications. This has resulted in DDN storage being used for NVIDIA's Selene, Cambridge-1 and Eos AI supercomputers, as well as the creation of reliable and repeatable reference architectures that scale with ease for enterprise AI customers.

Historically, most supercomputers were custom-built one-off designs, but the new breed of enterprise AI customers does not have the experience, expertise or time to build one this way. With the experience building Selene leveraging DDN's A^{3}I appliances, accomplished in 2020 over just three weeks, NVIDIA was able to create the blueprint for AI Factories that came to be known as NVIDIA DGX SuperPOD™. The DGX SuperPOD delivers reduced time-to-outcomes while minimizing the complexity of increasingly diverse AI models, including conversational AI, recommender systems, computer vision workloads, autonomous vehicles and DDN was certified as the first storage solution for this world- class, turnkey AI Factory.

"When we developed Selene, we had a design in mind, to grow from a smaller unit into the full-size supercomputer. We wanted to be able to take on that effort of going through the pain of putting this together and figuring out where the gaps were so that joint customers of ours could go out and take the same architecture for whatever scale that they need. [We are giving them] the confidence of knowing that somebody has done this and that it works, and that expectations can be met."

~ Prethvi Kashinkunti

Senior Data Center Systems Engineer | NVIDIA

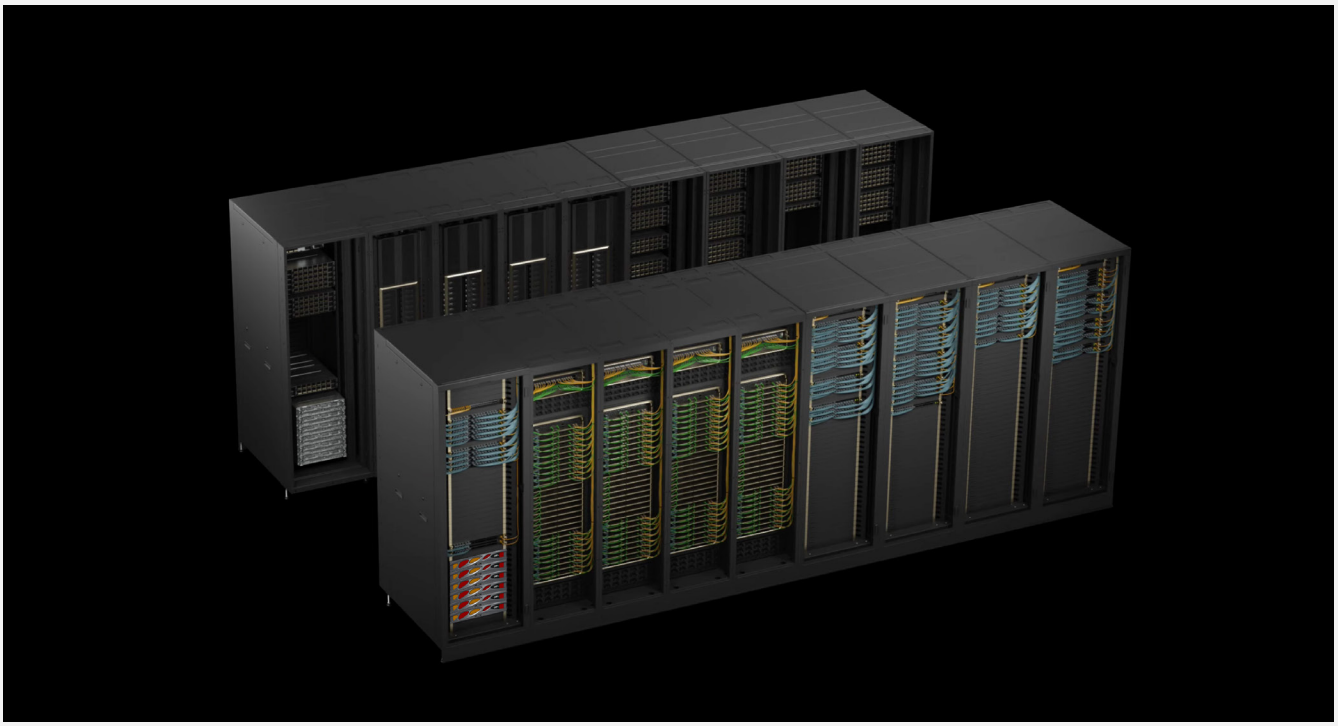
Over time, the increase in the size and complexity of AI models has driven NVIDIA and DDN to collaborate on additional systems to achieve unprecedented performance and predictable uptime, dramatically boosting utilization and productivity and increasing the ROI of NVIDIA's internal systems and customer AI initiatives alike. Most recently, NVIDIA unveiled its Eos system, which is comprised of 576 NVIDIA DGX H100 systems and NVIDIA Quantum-2 InfiniBand networking, where NVIDIA uses DDN's AI400X2 appliances for their storage & data layer.

"There are many important considerations when designing the world's most powerful AI systems. Storage is one that is often overlooked. As the data models get bigger and bigger, and the computation becomes bigger and bigger, more and more data is needed. It's not just about moving that data; it's about moving the data at the same time."

~ Marc Hamilton

VP Solutions Architecture and Engineering | NVIDIA

By utilizing DDN, NVIDIA received a data platform well matched to its DGX systems, with high-performance networking, ample I/O capabilities, and a design that scaled well with its growing data needs and customers' growing demands.



The Benefits

"DDN's performance and scalability are essential to reducing total time to solution, which is king."

~ Michael Houston

Chief Architect, AI Systems | NVIDIA

DDN is proud to be integrated with many NVIDIA AI Factories sold around the world today for shared clouds, generative AI, sovereign AI, and other applications. The flexible and performance optimized solution has allowed customers to get faster ROI with more effective generative AI and LLM training across autonomous vehicles, genomics and biosciences, financial services, robotics, manufacturing, and countless other industries.

Benefits provided by DDN include:

- 30–40% lower TCO
- 74% less power & cooling
- $257M ROI (3 years @ 10K GPUs)

Additionally, DDN's solutions have kept up with NVIDIA's advances in GPU technology. As GPUs get more powerful, they need to stay busy and DDN has increased the performance of its appliances in successive generations by 50% in the same power and rack space requirement.

"Having a partner who stands shoulder-to-shoulder with our engineers to solve the big challenges is where the true value comes from. We're definitely pushing the boundaries of what's possible today while exploring new frontiers for the future."

~ Michael Houston

Chief Architect, AI Systems | NVIDIA

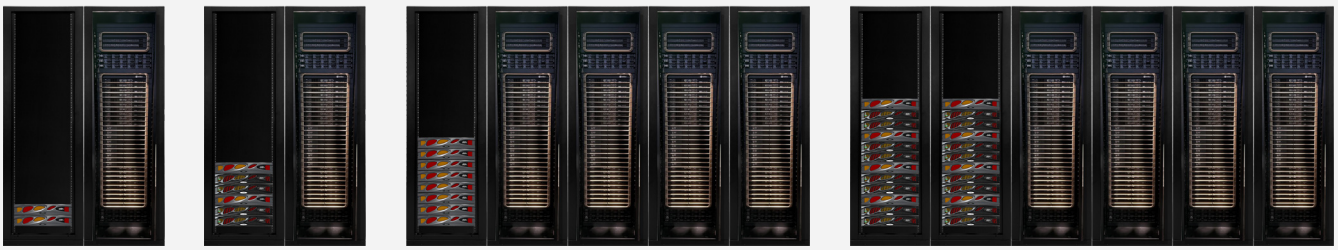
With a good balance of read and write performance, DDN maximizes GPU utilization by minimizing the time it takes to run I/O intensive operations like data load, model load, and checkpoints. Checkpoints, a critical recurrent step in training workloads where models are saved to persistent storage for a variety of reasons, can be a significant bottleneck. Because of DDN's efficient write performance, these checkpoints are significantly faster than alternative storage solutions, reducing wait time and making the entire system more productive.

"Having the storage technology that can provide the appropriate amount of bandwidth both for reads and writes is critical to ensure we maintain that level of efficiency. The DDN technology was the right fit for this type of application."

~ Prethvi Kashinkunti

Senior Data Center Systems Engineer | NVIDIA

Conversely, DDN has delivered complete campus-wide, departmental and cloud storage solutions to hundreds of universities around the world, combining sophisticated technology with an in-depth understanding of the diverse requirements in academic research.



Looking Ahead

By consolidating the power of an entire data center into a single platform, NVIDIA is revolutionizing how complex machine learning workflows and AI models are developed and deployed in an enterprise. With the addition of DDN storage to the advanced AI Factory provided by NVIDIA, they are providing world-class AI solutions for enterprise customers.

"I would say to anybody who is thinking about using DDN, that they would be getting an engineering partner and a team that knows how to support customers that have such a large scale like we do. They have the ability to continue to innovate and provide new solutions for improving performance of future AI applications."

~ Prethvi Kashinkunti

Senior Data Center Systems Engineer | NVIDIA

NVIDIA is also making access to accelerated computing as easy, fast, and flexible as possible. Whether they are deploying in their own data center, as a hosted private solution, or in a public cloud, customers can be confident that providers following the standardized reference architectures will supply an efficient and well-proven solution. With DDN as a key component of these AI Factories, customers can expect higher utilization, lower TCO, and faster time-to-results.

"What I love about DDN is that they're not new to high-performance. They're the de facto name in high- performance computing storage. And now, by working with us on our DGX SuperPOD, they're the de facto name for AI storage in high-performance environments."

~ Marc Hamilton

VP Solutions Architecture and Engineering | NVIDIA