Accelerating AI Innovation for Datacentric Enterprises

Manufacturing
Global

- Machine & Deep Learning Workflows
- Data Analytics
- Supercomputing

Solution

DDN's AI400X delivers scalable petabytes of powerful storage for NVIDIA Selene and DGX SuperPOD deployments for enterprises.

Michael Houston

Chief Architect, AI Systems

NVIDIA

World's most advanced AI system ranked as the fastest industrial system in the U.S. drives AI at-scale discovery, development and deployment breakthroughs

Since NVIDIA's invention of the GPU in 1999, the company has set new standards in computing innovation, deep-learning and data analytics. With the May 2020 introduction of the NVIDIA DGX ™ A100, NVIDIA raised the bar once more.

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. This computing marvel also led to the creation of Selene, the world's seventh fastest computer in total performance and the world's fastest commercially-available system - NVIDIA DGX SuperPOD Solution. DGX A100 systems are being used to fight COVID-19, fuel autonomous vehicles, develop superhuman language understanding, and transform almost every facet of business with AI.

Exhaustive internal research and customer input inform product development. "NVIDIA devotes a lot of effort to envisioning what state-of-the-art looks like and how next-generation architectures should perform," explains Charlie Boyle, Vice President and General Manager of DGX systems at NVIDIA. "We measure ourselves in terms of what's theoretically possible and how to accomplish it at the speed of light."

The Challenge

- Extremely fast storage needed to fuel the exponential growth of AI models, drive increasing demand for higher data rates per GPU and I/O bandwidth
- High degrees of parallel access to small and large datasets necessitated balanced storage performance
- Storage architecture flexibility with easy horizontal scalability required for incremental and on-demand data center expansion

In building the world's most advanced AI system, NVIDIA sought to reduce the complexity of increasingly diverse AI models, including conversational AI, recommender systems, computer vision workloads and autonomous vehicles. "What's needed is data center-scale computing, so AI models and datasets can be processed across many systems in parallel, enabling you to train them in hours, instead of weeks," says Tony Paikeday, Senior Director, Product Marketing at NVIDIA.

To reduce the time-to-solution challenge, NVIDIA needed to be able to distribute such large computational problems across hundreds of systems operating in parallel using an integrated set of compute, network and storage building blocks. This modular architecture lets enterprises scale resources in response to business demand to meet evolving AI requirements – introducing AI at-scale. This approach requires extremely fast storage performance to support intensive processing demands.

"We're pushing the boundaries of what's possible and that means unifying the most powerful compute with ultra-high-speed, low-latency networking and extremely fast storage," Paikeday adds. High degrees of parallel access to small and large files was deemed critical, along with seamless horizontal scaling to ensure incremental expansion of DGX SuperPOD's modular architecture.

Performance

Accelerated AI workloads achieve 1TB/s read and 690GB/s write for 20 appliances for DL and DA workloads

Scale

Horizontal scaling simplifies on-demand expansion

Flexibility

Supports diverse workloads and speeds deployments

Experience

Trusted partnership pushed technology boundaries

"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 with DDN while exploring new frontiers for the future."

Michael Houston

Chief Architect, AI Systems

The Solution

- Extremely fast storage needed to fuel the exponential growth of AI models, drive increasing demand for higher data rates per GPU and I/O bandwidth
- High degrees of parallel access to small and large datasets necessitated balanced storage performance

In selecting the ideal storage for Selene, NVIDIA turned to DDN as a long-time partner with a strong track record of maximizing data storage acceleration to complement the power of NVIDIA's DGX systems. Part of the company's A³I storage family, DDN's AI400X provides all-flash and hybrid storage for handling high-performance and large-scale capacity demands. The storage also integrates seamlessly with NVIDIA's Mellanox InfiniBand switches, which was critical to ensuring seamless DGX A100 operation and simple scalability.

"As a trusted partner, DDN's team worked with us to get the most out of their storage platform," says Michael Houston, Chief Architect, AI Systems, at NVIDIA. "The deep engineering engagement allowed us to push performance and scalability of the DDN platform."

Rapid response during both synthetic and real-world testing facilitated fast issue resolution. "A lot of collaboration between Mike's team and DDN made sure everything was correctly sized, assembled and usable in a very short amount of time," says Boyle. "Bringing up Selene in less than a month was a huge undertaking, but great planning and strong partnerships brought it together."

The Benefits

- Dramatic increase in AI workload performance enables organizations to iterate faster and boost data science productivity
- Modular platform ensures AI infrastructure scalability with greater speed and cost efficiency
- Fully integrated AI reference architectures democratize AI infrastructure for diverse workloads with streamlined deployment and operation

NVIDIA's Selene is a living proof-point of the power of the DGX SuperPOD reference architecture on which it's based. With DGX SuperPOD, NVIDIA empowers global organizations to create AI centers of excellence and DDN's scalable, high-performance storage plays a major role in fast-tracking those initiatives. Selene is backed by 7PBs of DDN A³I storage, deployed alongside NVIDIA's DGX A100 systems to power GPU-accelerated workloads found across every industry.

As part of NVIDIA's DGX SuperPOD reference architecture, which is available in cluster sizes ranging from 20 to 140 individual DGX A100 systems, scalable DDN storage condenses a complex supercomputing environment. "What DDN brings to the table is the ability to shrink time and distance between where the data lives and where the work is done on that data," says Paikeday.

DDN's modular architecture also allows high-performance computing deployments to be completed in weeks, not months or even years matching the modular build-out philosophy of DGX SuperPOD. . Additionally, the performance benefits of DGX A100 systems with DDN storage go beyond pure speed to change how information is processed. "DDN's performance and scalability are essential to reducing total time to solution, which is king," Houston adds.

DDN's high marks extend to energy efficiency, which helped NVIDIA clinch the No. 2 spot on the Green500 list of the most powerful commercially available systems based on energy efficiency. "DDN was the lowest power solution for the performance and capacity requirements we had," adds Houston.

Future Challenges

DDN's and NVIDIA's trusted partnership sets the stage for continued collaboration, including the addition of new capabilities and feature enhancements. Among the short list is tiering to object storage as well as Persistent Client Cache (PCC) to keep pace with continued data growth and ever evolving AI imperatives.

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

About DDN

DataDirect Networks (DDN) is the world's leading big data storage supplier to data-intensive, global organizations. DDN has designed, developed, deployed, and optimized systems, software, and solutions that enable enterprises, service providers, research facilities, and government agencies to generate more value and to accelerate time to insight from their data and information, on premise and in the cloud.

v1 (11/20)