Enabling AI at-Scale for the Datacentric Enterprise

DDN A³I solutions make AI-powered innovation easier, with faster performance, effortless scale, and simplified operations.

Artificial Intelligence (AI) and Deep Learning (DL) are transforming every industry around the world. They create the toughest workloads in modern computing history, and pose design challenges to compute, storage, and networking infrastructure that can be complex and time consuming to solve. DDN A³I solutions fully-integrate DDN storage appliances and NVIDIA DGX™ A100 servers for rapid and simple deployment of an AI-optimized infrastructure. DDN A³I solutions deliver the fastest application performance, and scale seamlessly to meet the demands of evolving AI workloads.

DDN A³I with NVIDIA DGX A100: Accelerated AI-Powered Insights

DDN A³I (Accelerated, Any-Scale AI) solutions break new ground for AI and DL. Engineered from the ground up for the AI-enabled data center, DDN A³I solutions with NVIDIA DGX A100 servers accelerate end-to-end data pipelines for AI and DL workloads of any scale. They are designed to provide extreme amounts of performance and capacity backed by a jointly engineered, validated architecture. DDN A³I with NVIDIA DGX A100 is optimized at every layer of hardware and software to ensure data delivery and storage is fast, responsive and reliable.

The DDN shared parallel architecture enables full utilization of the DGX A100 GPU resources and maximizes developer productivity from every compute cycle on the DGX A100 server. DDN A³I solutions streamline DL by enabling all phases of the workflow to happen concurrently and continuously—including data ingest, curation, training, validation, inference, and simulation. With DDN A³I solutions, data is accessible simultaneously by multiple DGX A100 servers through a fully optimized and unified interface that is easy to use directly from containerized applications.

AI Infrastructure That's Easy to Deploy, Manage and Use

DDN A³I with NVIDIA DGX A100 offers a turnkey and pre-configured solution, that is easy to deploy, shortening the timeline from AI concepts to business insights in a production setting. Based on comprehensive DDN A³I reference architectures, these solutions eliminate design guesswork, and have been validated in collaboration with NVIDIA to ensure the highest performance, optimal efficiency, and flexible growth for DGX A100 servers.

Built on the NVIDIA DGX A100 AI supercomputer as it's compute foundation, this powerful solution delivers over one petaFLOPS of DL training performance, leveraging eight NVIDIA Tesla V100 Tensor Core GPUs, configured in a hybrid cube-mesh topology using NVIDIA NVLink for an ultra-high-bandwidth, low-latency inter-GPU communications fabric. The DGX A100 architecture overcomes the performance bottlenecks of traditional GPU interconnects and offers linearly predictable performance across multiple GPUs. DGX A100 is powered by the NVIDIA DGX Software Stack which is optimized at every layer, including the most popular DL frameworks and the supporting libraries and drivers, DDN A^3I with NVIDIA DGX A100 delivers unmatched multi-GPU and multi-system AI application performance.

To meet the requirements of a variety of workloads, DDN A^3I with NVIDIA DGX A100 leverages the DDN AI200X, AI400X and AI7990X storage appliances. The AI200X and AI400X are all-NVME, fully-integrated parallel file storage appliances that deliver up to 48GB/s of throughput and over 3M IOPS to applications, accelerating even the most I/O intensive workloads. The AI200X and AI400X are specifically optimized to keep GPU computing resources fully utilized, ensuring maximum efficiency while easily managing tough data operations. The AI7990X is a hybrid, parallel file storage appliance that integrates both flash and deeply expandable capacity disk in a unified system for simplicity and flexibility. This integration makes it easy to collocate both hot training data and large libraries while maintaining optimal system efficiency. The AI7990X outperforms competing platforms and delivers the economics of capacity disk for your growing data library.

DDN storage appliances are fully-integrated and optimized for AI and DL workloads.

The DDN architecture allows for instant provisioning of new resources to applications, and easy deployment of additional DGX A100 servers. Advanced monitoring tools embedded within all components of the solution provide easy data management capabilities and extensive metrics for comprehensive optimization of live workloads. A robust digital security framework makes it simple to deploy secure multitenancy and share the DGX A100 server capability with a broad group of users while ensuring data access and governance compliance.

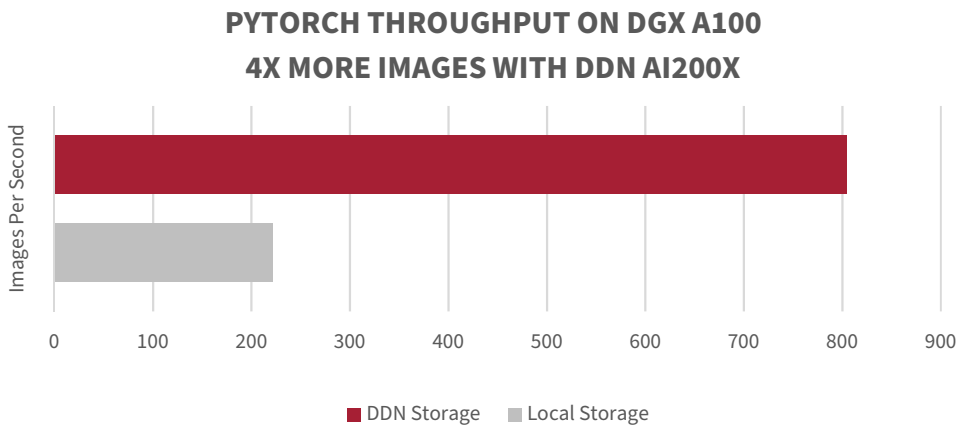
High-Performance AI at-Scale

DDN A^{3}I solutions can scale seamlessly supporting additional DGX A100 servers and AI200X, AI400X, and AI7990X appliances for more performance and capacity through a single unified namespace. Storage and compute are fully-interconnected through a high-performance, low-latency network using HDR and EDR InfiniBand™ or 100Gbps Ethernet.

Fast and flexible access to a large and diverse data set is key to successful AI and DL efforts. The DDN high-performance parallel architecture delivers data to AI workloads with the highest bandwidth, lowest latency and maximum concurrency, ensuring full GPU resource utilization even for distributed DL training running across multiple DGX A100 servers simultaneously.

As part of the solution, DDN has developed an intelligent client for DGX A100 server containers that engages multiple high-speed data paths to the storage and delivers the full performance of NVMe flash directly to the application. The DDN shared parallel architecture provides predictable and linear scaling of performance for training across multiple GPUs on a single DGX A100 server or multiple DGX A100 servers simultaneously. With true end-to-end parallelism, DDN A^{3}I with NVIDIA DGX A100 eliminates the bottlenecks associated with legacy platforms.

Extensive performance and interoperability testing on widely-used DL frameworks demonstrate that containerized applications can now engage the full capabilities of the data infrastructure. All applications demonstrate significantly higher performance and shorter run times with DDN A^{3}I than with other data platforms.



Comprehensive performance and interoperability test results are available in the DDN A^{3}I Scalable Architecture for Artificial Intelligence and Deep Learning with NVIDIA DGX A100.

Backed by Deep AI Expertise from DDN and NVIDIA

For 20 years, DDN has designed, developed, deployed and optimized solutions that enable organizations to generate value by accelerating time to insight from their data, both on-premises and in the cloud. In combination with NVIDIA's leadership and expertise in AI and DL, DDN has successfully deployed data-at-scale solutions for customers in all industries seeking to accelerate their business with AI and DL.

Fully-integrated and optimized, DDN A^{3}I solutions with NVIDIA are delivered and supported by expert partners worldwide, certified to provide comprehensive value-add services for enterprise.

DDN A^{3}I solutions with NVIDIA enable users and applications to easily harness the massive AI power of the DGX A100 for the most efficient utilization of GPU resources, and the fastest time-to-insights. DDN A^{3}I solutions make AI-powered innovation easier, with faster performance, effortless scale, and simplified operations, backed by the AI infrastructure experts.

About NVIDIA

NVIDIA's (NASDAQ:NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics and revolutionized parallel computing. More recently, GPU deep learning ignited modern AI — the next era of computing — with the GPU acting as the brain of computers, robots and self-driving cars that can perceive and understand the world. More information at http://www.nvidia.com/dgx.