DDN A³I® SOLUTIONS WITH NVIDIA DGX™ A100 SYSTEMS

Fully-integrated and optimized data platforms for accelerated at-scale AI, Analytics and HPC

DDN A³I – Accelerated, Any-Scale AI

Artificial intelligence (AI) and deep learning (DL) are creating the toughest workloads in modern computing history. They pose exceptional challenges to and put significant strain on compute, storage, and network resources. Enterprise file storage architectures and protocols like NFS can starve AI workloads of data, slowing down applications and hindering business innovation. An AI-enabling datacenter must concurrently and efficiently service the entire spectrum of activities involved in DL workflows, including data ingest, data curation, training, inference, validation and simulation.

DDN A³I solutions (Accelerated, Any-Scale AI) break new ground for AI, Analytics, and high-performance computing (HPC). Engineered from the ground up for the AI-enabled datacenter, DDN's A³I solutions accelerate AI applications and streamline DL workflows using the DDN shared parallel architecture. With DDN, flash performance layers can scale-up or scale-out independently from capacity layers, all within a single integrated solution and namespace. DDN A³I solutions provide unmatched flexibility for your organization's AI needs.

DDN A³I Solutions Features and Benefits

DDN A³I solutions are easy to deploy and manage, highly scalable in both performance and capacity, and represent a highly efficient and resilient platform for AI workflows at-scale. Below we summarize the benefits of DDN A³I solutions.

1. Fully Integrated, GPU Optimized Data Platform

DDN A³I solutions are optimized for AI and DL and benefit from the DDN advanced hardware and software technologies. DDN A³I solutions are turn-key and pre-configured, making them easy to deploy, and speed-up your time frame to the most capable scale-out platform for capacity and performance. DDN A³I solutions deliver a fully integrated and optimized data flow from application to flash and disk that ensures peak utilization of GPUs for maximum productivity at any scale.

2. Highest Performance AI Storage with Peak GPU Saturation

DDN A³I solutions are designed for all types of I/O patterns and data layouts. The DDN high-performance shared parallel architecture delivers data to applications with highest bandwidth and lowest latency, ensuring complete GPU resource utilization even with distributed applications running on multiple computing servers. DDN A³I solutions offer best performance, efficiency, and economics for AI infrastructure at any scale. A single DDN AI400X storage appliance can achieve 48GB/s of throughput and over 3M IOPS by extending the massive performance of dual ported NVMe drives all the way to the GPUs and applications.

3. Best Capacity Efficient AI Storage

DDN A^{3}I solutions provide over 370TB of reliable and efficient dual ported NVMe flash in a highest-density 2 RU configuration. The solutions can scale seamlessly to multiple petabytes of flash as workflow needs evolve to provide additional capacity, performance, and capability. Modular and extremely flexible, DDN A^{3}I solutions also deliver a more manageable unified namespace of a high-capacity, high- efficiency hard disk drive storage pool which can scale up to tens or hundreds of petabytes.

4. Highest Resiliency, Reliability, Security at Scale

DDN A^{3}I solutions are engineered to provide highest data availability and maximum system uptime. All hardware and software components are integrated as redundant systems to ensure availability. The multi-dimensional scaling capabilities of the architecture provide extremely reliable expansion options to meet evolving workflow requirements. Additional GPU computing appliances can be seamlessly integrated and provisioned instantly for immediate access by applications. DDN A^{3}I solutions deliver ease of use, best performance efficiency, optimized economics all under a single portfolio.

5. Unified Namespace

DDN A^{3}I solutions integrate a true distributed parallel file system that provides a very simple highly scalable, single namespace structure. This shared parallel filesystem architecture enables concurrent high-performance data access from multiple computing appliances and eliminates unnecessary data movements between storage areas. The architecture allows consolidation of hot training data and warm expanding data libraries into a single platform, providing easy data access from a unified interface.

6. Easy to Manage Multi-Tenancy and Quota Support

DDN A^{3}I solutions can be secured on a per-tenant basis through reliable controls that ensure users and applications can only access the data that they're entitled to, without compromising the high-performance parallel access. Advanced quota controls integrated within the solution provide easy management of filesystem consumption at the user, group, and project level. Fine-grained monitoring tools embedded within all components of the solution provide extensive metrics for comprehensive analysis and optimization of live workloads.

DDN AI200X and AI400X

DDN's AI200X and AI400X are the world's most efficient, reliable, and easy to use data storage appliances for AI and DL applications. The AI200X and AI400X (**Figure 1**) bring new levels of performance, simplicity, and flexibility to help deal with rapid evolution of your AI deployment. With true end-to-end parallelism, the AI200X and AI400X eliminate the bottlenecks associated with NFS-based platforms and deliver the performance of NVMe Flash directly to your AI application.

The DDN AI400X is a fast, dense, scalable, and highly flexible data storage platform. The AI400X delivers up to 48GB/s of filesystem throughput, over 3M IOPS and 370 terabytes of dual ported NVMe flash in just 2RU. Both the AI200X and AI400X can scale horizontally as a single, simple namespace. They integrate tightly with hard disk tiers to help manage economics when data volumes expand. The AI200X and AI400X are specifically optimized to keep GPU computing resources saturated, ensuring maximum efficiency while easily managing tough data operations, from bursty ingest to large data transformations.

	DDN AI200X	DDN AI400X
Performance	24GB/s, 1.5M IOPS	48GB/s, 3M IOPS
Connectivity	(4) HDR100 IB or (4) 100GbE	(8) HDR100 IB or (8) 100GbE
Capacity	4 options: 32TB, 64TB, 128TB, 256TB	

Figure 1. DDN AI200X and AI400X all-NVME storage appliances for AI and DL.

DDN AI200X and DDN AI400X for GPU Accelerated Workloads

DDN's A^{3}I reference architectures are designed in collaboration with NVIDIA to provide highest performance, optimal efficiency, and flexible growth for NVIDIA DGX™ A100 systems.

Wholly integrated into the DDN AI200X and AI400X is a shared parallel filesystem optimized specifically for data delivery to applications running on the GPU compute servers. The AI200X and AI400X can be easily connected to the servers through a high-performance, low-latency network using HDR InfiniBand (HDR IB) or 100 Gbps Ethernet (100 GbE). A complete solution starts with 30 TB of dual ported NVMe capacity that delivers up to 48GB/s in throughput and over 3M IOPS with options for growth. The AI200X and AI400X can scale-out effortlessly by integrating additional AI200X and AI400X storage appliances for more performance and capacity.

The minimal implementation supports a single GPU compute server. A single AI200X or AI400X can saturate multiple DGX systems with substantial headroom for dealing with data ingest, transformation, and movement concurrently.

Extensive testing with widely-used AI and DL applications demonstrates that a single AI200X provides tremendous acceleration for data preparation, neural network training and inference tasks using GPU compute servers. As illustrated in **Figure 2**, Pytorch-based applications running on an NVIDIA DGX-1™ system demonstrate 4X increased image throughput and 4X shorter completion times with AI200X. Testing also demonstrates that the DDN shared parallel architecture maintains linear performance for applications that leverage distributed computing on multiple GPU compute servers, such as Tensorflow and Horovod.

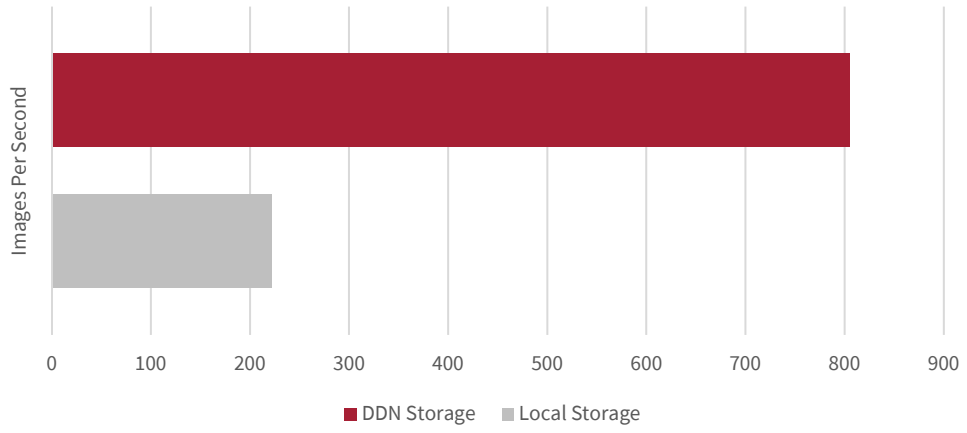


Figure 2. Pytorch application read throughput on a DGX-1 system.

DDN AI7990X

DDN's AI7990X (**Figure 3**) is a high-density, integrated, scalable hybrid flash and hard disk data storage appliance for AI and DL. With true end-to-end parallel access to flash and deeply expandable HDD storage, the AI7990X outperforms NAS platforms and delivers the economics of HDD for your growing data library. Delivering flash performance direct to your AI application, the AI7990X brings new levels of simplicity and flexibility to help deal with unexpected turns and complications in your AI deployment.

DDN's AI7990X is fast, dense, scalable and flexible in deployment with up to 24GB/s of filesystem throughput and over 800K IOPS in 4RU as a single namespace to meet demand. DDN flash and spinning disk storage is integrated tightly within a single unit for up to 1PB in just 4RU. DDN disk tiers to help manage AI economics when data volumes expand. The AI7990X keeps GPU servers saturated with data ensuring absolute maximum utilization whilst also managing tough data operations from bursty ingest to large scale data transformations.

	DDN AI7990X
Performance	24GB/s, 800K IOPS
Connectivity	(4) HDR100 IB or (4) 100GbE
Capacity	3 options: 1PB, 2PB, 4PB

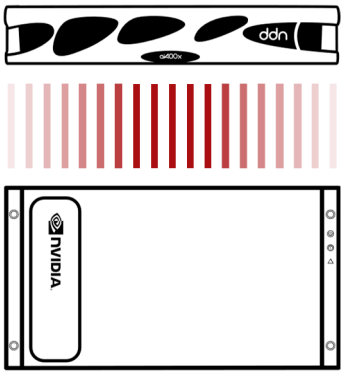
Figure 3. DDN AI7990X hybrid flash and capacity storage appliance for AI and DL

DDN A³I End-To-End Enablement for NVIDIA DGX Systems

DDN A³I solutions are architected to achieve the most from at-scale AI, Analytics and HPC applications running on DGX systems. They are designed to provide extreme amounts of performance, capacity and capability through a tight integration between DDN and NVIDIA systems. Every layer of hardware and software engaged in delivering and storing data is optimized for fast, responsive, and reliable access.

DDN A³I solutions are designed, developed, and optimized in close collaboration with NVIDIA. The deep integration of DDN AI appliances with DGX systems ensures a predictable and reliable experience. DDN A³I solutions are highly configuration for flexible deployment in a wide range of environments and scale seamlessly in capacity and capability to match evolving workload needs. DDN A³I solutions are deployed globally and at all scale, from a single DGX system all the way to the largest NVIDIA DGX SuperPOD™ in operation today.

DDN brings the same advanced technologies used to power the world's largest supercomputers in a fully-integrated package for DGX systems that's easy to deploy and manage. DDN A³I solutions are proven to provide maximum benefits for at-scale AI, Analytics and HPC workloads on DGX systems.



DDN A³I Shared Parallel Architecture

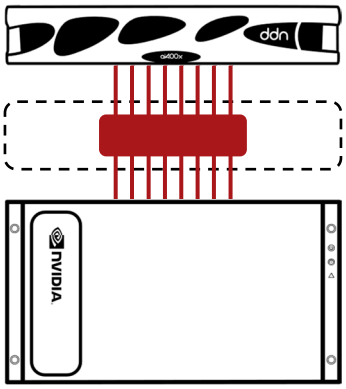
The DDN A³I shared parallel architecture and client protocol provides superior performance, scalability, security, and reliability for DGX systems. Multiple parallel data paths extend from the drives all the way to containerized applications running on the GPUs in the DGX system. With DDN's true end-to-end parallelism, data is delivered with high-throughput, low-latency, and massive concurrency in transactions. This ensures applications achieve the most from DGX servers with all GPU cycles put to productive use. Optimized parallel data-delivery directly translates to increased application performance and faster completion times. The DDN A³I shared parallel architecture also contains redundancy and automatic failover capability to ensure high reliability, resiliency, and data availability in case a network connection or server becomes unavailable.

The A³I DDN shared parallel architecture provides proven enablement and acceleration for AI infrastructure and workloads on DGX systems.

DDN A³I Streamlined Deep Learning

DDN A³I solutions enable and accelerate end-to-end data pipelines for DL workflows of all scale running on DGX servers. The DDN shared parallel architecture enables concurrent and continuous execution of all phases of DL workflows across multiple DGX systems. This eliminates the management overhead and risks of moving data between storage locations. At the application level, data is accessed through a standard highly interoperable file interface, for a familiar and intuitive user experience.

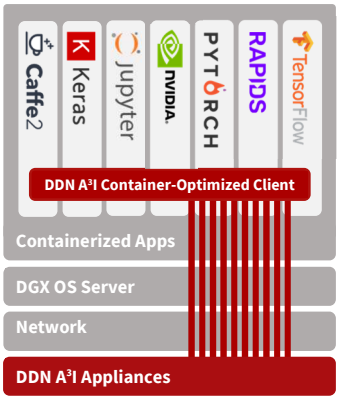
Significant acceleration can be achieved by executing an application across multiple DGX systems simultaneously and engaging parallel training efforts of candidate neural networks variants. These advanced optimizations maximize the potential of DL frameworks. DDN works closely with NVIDIA and its customers to develop solutions and technologies that allow widely-used DL frameworks to run much faster, better, and more reliably on DGX systems.



DDN A³I Multirail Networking

DDN A³I solutions integrate a wide range of networking technologies and topologies to ensure streamlined deployment and optimal performance for AI infrastructure. Latest generation InfiniBand and Ethernet provide both high-bandwidth and low-latency data transfers between applications, compute servers and storage appliances. DDN A³I Multirail greatly simplifies and optimizes DGX server networking for fast, secure, and resilient connectivity.

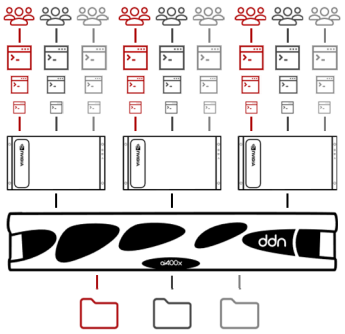
DDN A³I Multirail enables grouping of multiple network interfaces on a DGX sever to achieve faster aggregate data transfer capabilities. The feature balances traffic dynamically across all the interfaces, and actively monitors link health for rapid failure detection and automatic recovery. DDN A³I Multirail makes designing, deploying, and managing high-performance networks very simple, and is proven to deliver complete connectivity for at-scale infrastructure including on the DGX SuperPOD.



DDN A³I Container Client

Containers encapsulate applications and their dependencies to provide simple, reliable, and consistent execution. DDN enables a direct high-performance connection between DGX system application containers and the DDN parallel filesystem. This brings significant application performance benefits by enabling low latency, high-throughput parallel data access directly from a container. Additionally, the limitations of sharing a single host-level connection to storage between multiple containers disappear. The DDN in-container filesystem mounting capability is added at runtime through a universal wrapper that does not require any modification to the application or container.

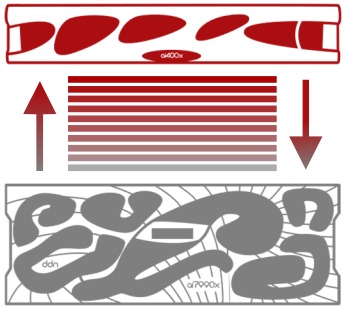
Containerized versions of popular DL frameworks, specially optimized for the DGX systems are available from NVIDIA. They provide a solid foundation that enables data scientists to rapidly develop and deploy applications on the DGX system. In some cases, open-source versions of the containers are available, further enabling access and integration for developers. The DDN A³I container client provides high-performance parallelized data access directly from containerized applications on the DGX system. This provides containerized DL frameworks with the most efficient dataset access possible, eliminating all latencies introduced by other layers of the computing stack.



DDN A³I Multitenancy

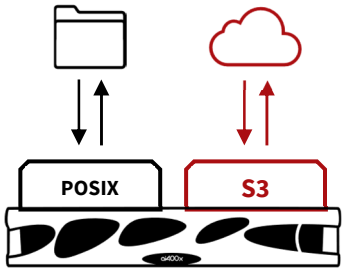
Container clients provide a simple and very solid mechanism to enforce data segregation by restricting data access within a container. DDN A³I makes it very simple to operate a secure multitenant environment at-scale through its native container client and comprehensive digital security framework. DDN A³I multitenancy makes it simple to share DGX systems across a large pool of users and still maintain secure data segregation. Multi-tenancy provides quick, seamless, dynamic DGX system resource provisioning for users. It eliminates resource silos, complex software release management, and unnecessary data movement between data storage locations. DDN A³I brings a very powerful multitenancy capability to DGX systems, and makes it very simple for customers to deliver a secure, shared innovation space, for at-scale data-intensive applications.

Containers bring security challenges and are vulnerable to unauthorized privilege escalation and data access. The DDN A³I digital security framework provides extensive controls, including a global *root_squash* to prevent unauthorized data access or modification from a malicious user, and even if a node or container are compromised.



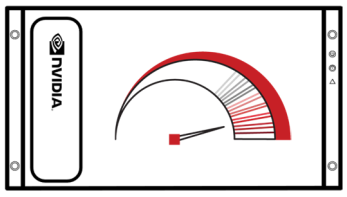
DDN A³I Hot Pools

Hot Pools delivers user transparent automatic migration of files between the Flash tier (Hot Pool) to HDD tier (Cool Pool). Hot Pools is designed for large scale operations, managing data movements natively and in parallel, entirely transparently to users. Based on mature and well tested file level replication technology, Hot Pools allows organizations to optimize their economics – scaling HDD capacity and/or Flash performance tiers independently as they grow.



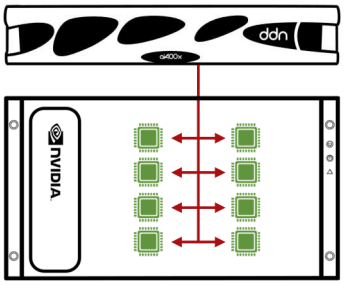
DDN A³I S3 Data Services

DDN S3 Data Services provide hybrid file and object data access to the shared namespace. The multi-protocol access to the unified namespace provides tremendous workflow flexibility and simple end-to-end integration. Data can be captured directly to storage through the S3 interface and accessed immediately by containerized applications on DGX system through a file interface. The shared namespace can also be presented through an S3 interface, for easy collaboration with multisite and multicloud deployments. The DDN S3 Data Services architecture delivers robust performance, scalability, security, and reliability features.



DDN A³I Advanced Optimizations for DGX A100 System Architecture

The DDN A³I client's NUMA-aware capabilities enable strong optimization for DGX systems. It automatically pins threads to ensure I/O activity across the DGX system is optimally localized, reducing latencies and increasing the utilization efficiency of the whole environment. Further enhancements reduce overhead when reclaiming memory pages from page cache to accelerate buffered operations to storage. The DDN A³I client software for DGX A100 systems has been validated at-scale with the largest DGX A100 system deployment currently in operation.



DDN A³I with NVIDIA GPUDirect Storage (GDS)

DDN A³I solutions interface directly with GPU memory for fastest and most efficient I/O operations possible. DDN fully integrates GDS which enables a direct DMA data path between GPU memory and storage, thus avoiding a bounce buffer through the CPU. This direct path increases system bandwidth while decreasing latency and utilization load on the CPU and GPU. The DDN shared parallel architecture combined with GDS enables customers to maximize DGX A100 system I/O capabilities. With GDS, DDN can deliver over 178 GB/s of throughput directly to GPU memory on a single DGX A100 system, fully-saturating the network interfaces on the server, and delivering 50% more throughput than available over standard data paths. This significantly improves AI, Analytics and HPC application performance on DGX A100 systems. GDS is fully implemented in current generation DDN AI storage appliances and validated with all GDS-supported DGX A100 systems and with multi-node deployments.

Testing demonstrates significant read and write throughput performance benefits of DDN A³I with GDS for DGX A100 systems (Figure 4). The graph compares peak read and write throughput from a single client node. With GDS, the client performance increases 1.7X. This test was executed on a single DGX A100 system and four AI400X appliances connected over an HDR200 network.

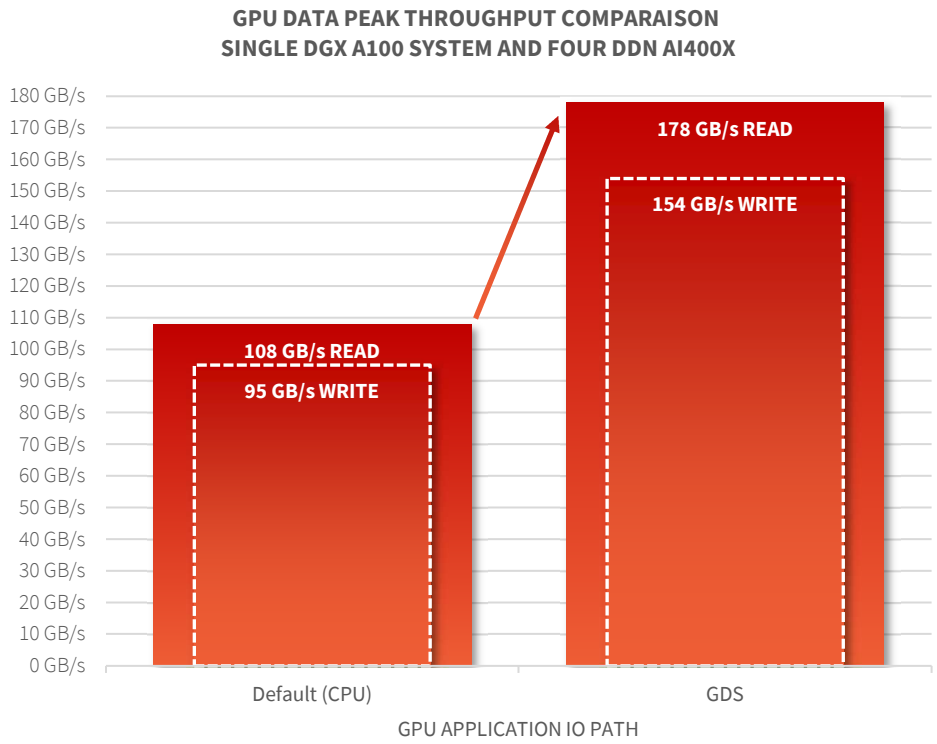


Figure 4. DDN A³I with GDS increased throughput performance on DGX A100 system

The architecture of the DGX A100 system enables individual GPUs to consume up to 24 GB/s of data from the network interface cards located on the same PCIe switch. Testing demonstrates that the DDN shared parallel architecture can fully saturate read throughput for all eight GPUs in the DGX A100 system simultaneously, delivering over 178 GB/s with linear performance scaling as more GPUs are engaged. The results also demonstrate almost full saturation of the eight HDR200 network interface cards simultaneously from a single shared data platform (Figure 5). This test was executed on a single DGX A100 system and four AI400X appliances connected over an HDR200 network.

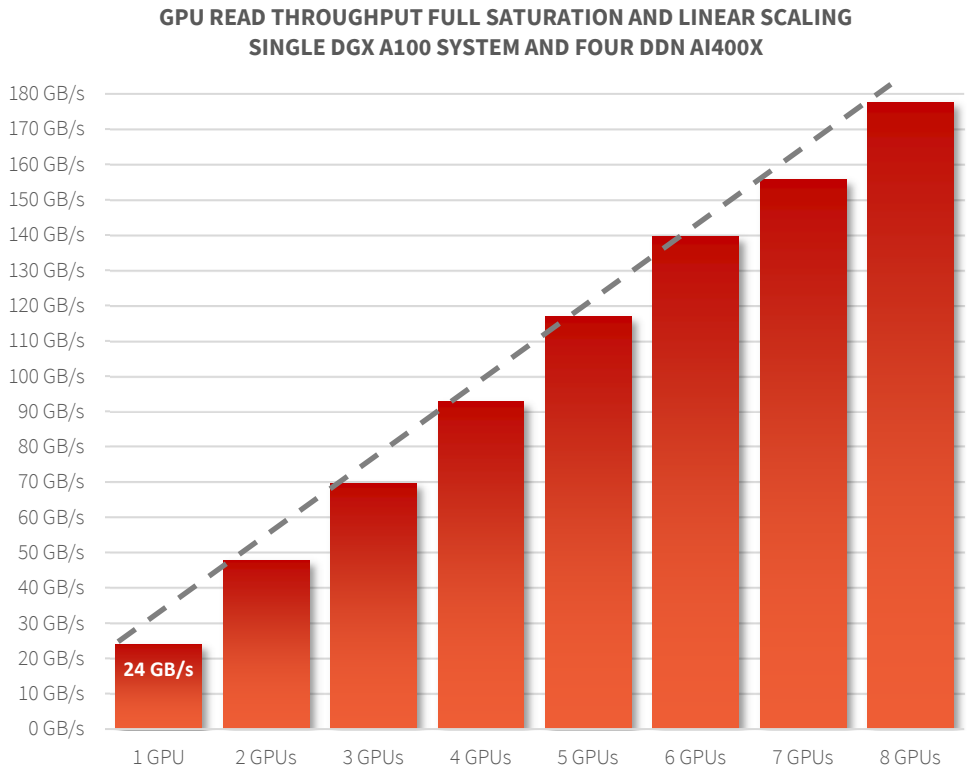


Figure 5. *DDN A³I with GDS GPU throughput scaling on DGX A100 system*

GDS provides significant advantage for applications running on DGX A100 systems. DDN appliances are proven to provide the performance required to maximize the value of GDS, especially with at-scale workloads and multi-node deployments. The capability is fully-integrated and can be enabled seamlessly for customers looking to engage in the ongoing early access program with NVIDIA. Contact DDN for more information.

DDN A³I Solutions with NVIDIA DGX A100 Systems

The DDN A³I scalable architecture integrates DGX A100 systems (Figure 6) with DDN AI shared parallel file storage appliances, and delivers fully-optimized end-to-end AI and DL workflow acceleration. DDN A³I solutions greatly simplify the use of the DGX A100 system, while also delivering performance and efficiency for maximum GPU saturation, and high levels of scalability.

NVIDIA DGX A100 System

Figure 6. NVIDIA DGX A100 system

The NVIDIA DGX A100 system is the universal system for all AI workloads, offering unprecedented compute density, performance, and flexibility in the world's first 5 petaFLOPS AI system. Built on the revolutionary NVIDIA A100 Tensor Core GPU, the DGX A100 system unifies data center AI infrastructure, running training, inference, and analytics workloads simultaneously with ease. More than a server, the DGX A100 system is the foundational building block of AI infrastructure and part of the NVIDIA end-to-end data center solution created from over a decade of AI leadership by NVIDIA. The DGX A100 system integrates exclusive access to a global team of AI-fluent experts that offer prescriptive planning, deployment, and optimization expertise to help fast-track AI transformation.

DDN A³I Reference Architectures for DGX A100 Systems

The following reference architectures are designed and optimized for AI, Analytics and HPC workloads running on DGX A100 systems. They provide guidance for rapidly deploying highest-performance fully-integrated storage, network, and compute infrastructure. Every solution has been validated for DGX A100 systems, and thoroughly tested by DDN technical experts.

The DDN A³I architecture is very flexible and can scale seamlessly in capacity, performance, and capability. Deployments can start with a single DGX A100 systems and a single AI200X, AI400X and AI7990X storage appliance (**Figure 7**). Additional DGX A100 systems and AI200X, AI400X and AI7990X storage appliances can be integrated rapidly to meet evolving workflow and workload requirements. At all scale, the DDN A³I architecture continuously delivers an optimized, extremely cost-effective solution.

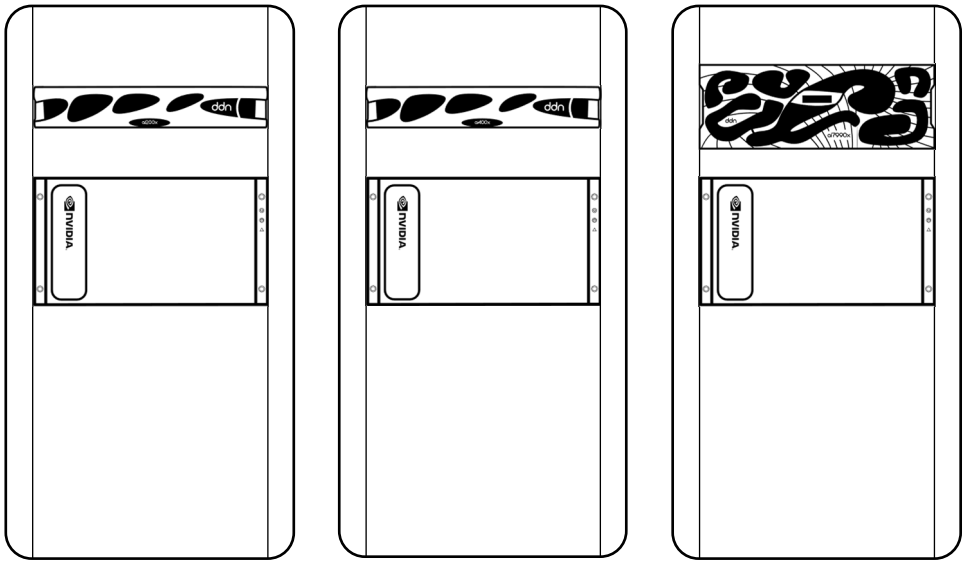
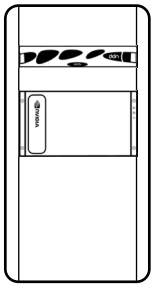


Figure 7. Rack illustrations for DDN A³I reference configurations



Configuration for a single DGX A100 system

Figure 88 illustrates the DDN A^{3}I architecture in a 1:1 configuration in which a single DGX A100 system is connected to an AI200X storage appliance through an HDR IB or 100 GbE network. The AI200X storage appliance connects to a single network switch via four links, and the DGX A100 system via two links.

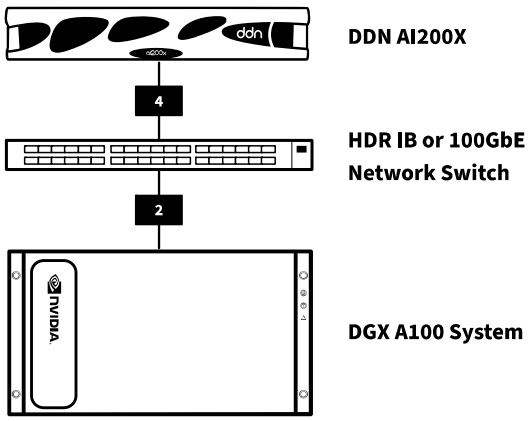
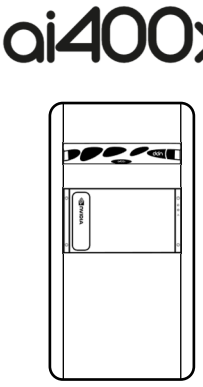


Figure 8. DDN A^{3}I reference architecture for AI200X with a DGX A100 system in a 1:1 configuration



Configuration for a single DGX A100 system

Figure 9 illustrates the DDN A^{3}I architecture in a 1:1 configuration in which a single DGX A100 system is connected to an AI400X storage appliance through an HDR IB or 100 GbE network. The AI400X storage appliance connects to a single network switch via eight links, and the DGX A100 system via two links.

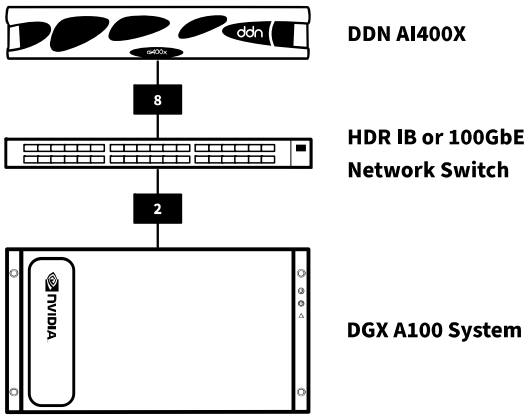
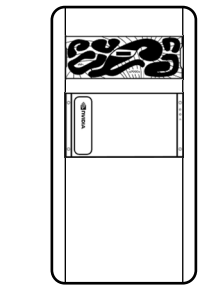


Figure 9. DDN A^{3}I reference architecture for AI400X with a DGX A100 system in a 1:1 configuration



Configuration for a single DGX A100 system

Figure 10 illustrates the DDN A^{3}I architecture in a 1:1 configuration in which a single DGX A100 system is connected to an AI7990X storage appliance through an HDR IB or 100 GbE network. The AI7990X storage appliance connect to a single network switch via four links, and the DGX A100 system via two links.

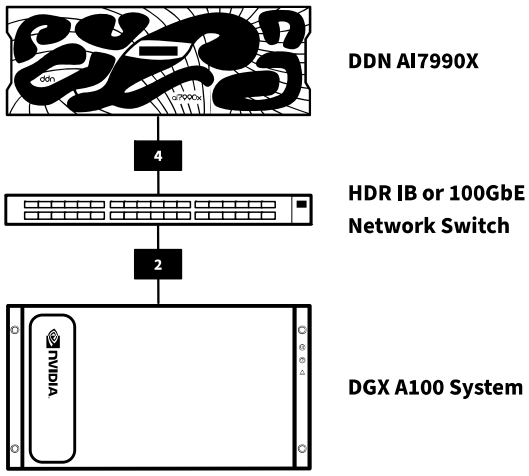


Figure 10. DDN A^{3}I reference architecture for AI7990X with a DGX A100 system in a 1:1 configuration.

POWERFUL & PROVEN DGX POD SOLUTIONS WITH DDN AI400X

DDN A³I Reference Architectures for DGX POD with DGX A100 Systems

DDN proposes the following reference architectures for multi-node DGX POD configurations. DDN A³I solutions are powerful, proven and deployed at-scale., including the systems operated by NVIDIA.

The DDN AI400X is a turnkey appliance for at-scale DGX deployments. DDN recommends the AI400X as the optimal data platform for DGX POD designs with the DGX A100 system. The AI400X delivers maximum GPU performance for every workload and data type in a dense, power efficient 2RU chassis. The AI400X simplifies the design, deployment and management of a DGX POD and provides predictable performance, capacity and scaling. The AI400X arrives fully configured, ready to deploy and installs in minutes. The appliance is designed for seamless integration with DGX systems, and enables customers to move rapidly from test to production. DDN provides complete expert design, deployment, and support services globally and ensures best customer experience. The DDN field engineering organization has already deployed hundreds of solutions for customers based on the A³I reference architectures.

As general guidance, DDN recommends an AI400X for every four DGX A100 systems in a DGX POD (**Figure 11**). These configurations can be adjusted and scaled easily to match specific workload requirements. For the storage network, DDN recommends HDR200 InfiniBand technology in a non-blocking topology, with redundancy to ensure data availability. DDN recommends use of at least two HDR200 connections per DGX A100 system to the storage network.

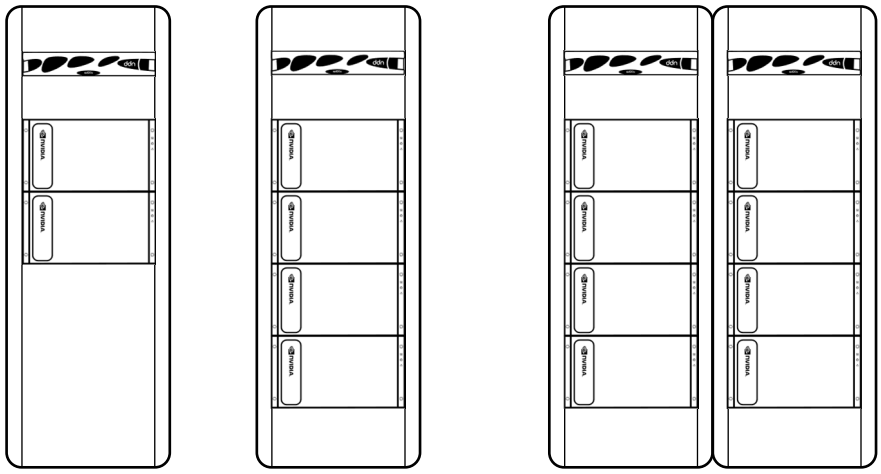
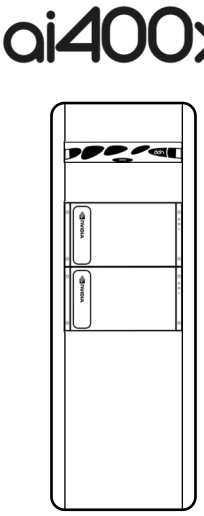


Figure 11. Rack illustrations for DDN A³I reference configurations.



DGX POD with Two DGX A100 systems

Figure illustrates the DDN A^{3}I architecture in a 1:2 configuration in which two DGX A100 system are connected to an AI400X storage appliance through a pair of network switches that are configured for high-availability (HA). Every DGX A100 system connects to each of the network switches via one HDR IB or 100 GbE links. The AI4000X storage appliance connects to each of the network switches via four HDR IB or 100 GbE links. The network switches are interconnected with four dedicated links. This ensures non-blocking data communication between every device connected to the network. The HA design provides full-redundancy and maximum data availability in case of component failure in one of the devices.

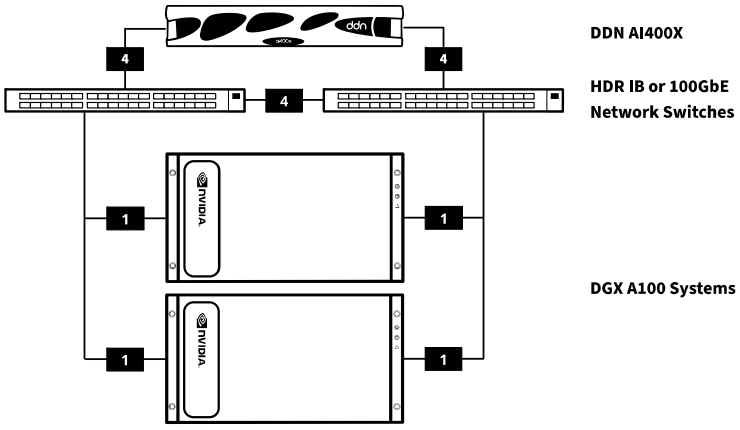
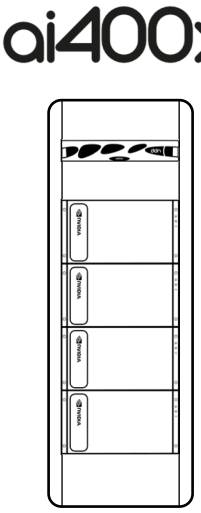


Figure 12. DDN A^{3}I reference architecture for AI400X with DGX A100 systems in a 1:2 configuration



DGX POD with Four DGX A100 systems

Figure 12 illustrates the DDN A^{3}I architecture in a 1:4 configuration in which four DGX A100 systems are connected an AI400X storage appliance through a pair of network switches that are configured for high-availability (HA). Every DGX A100 system connects to each of the network switches via one HDR IB or 100 GbE links. The AI4000X storage appliance connects to each of the network switches via four HDR IB or 100 GbE links. The network switches are interconnected with four dedicated links. This ensures non-blocking data communication between every device connected to the network. The HA design provides full-redundancy and maximum data availability in case of component failure in one of the devices.

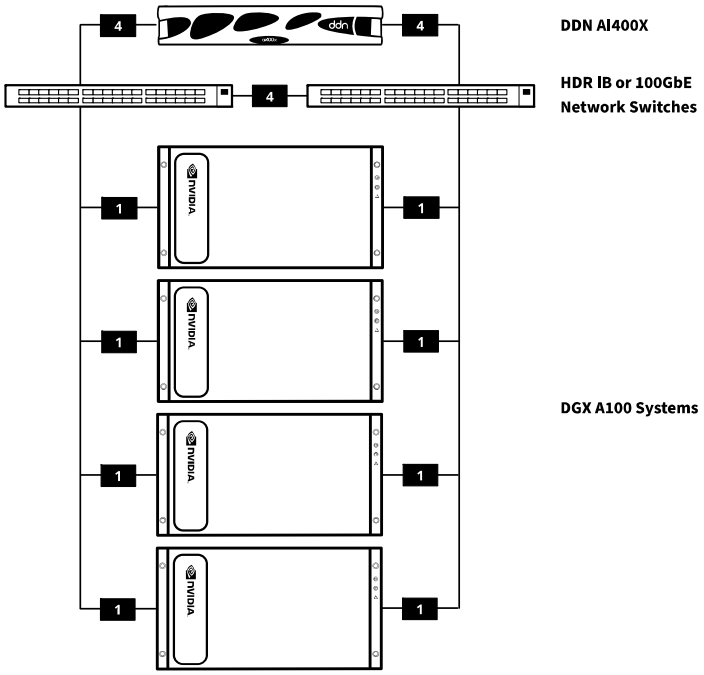
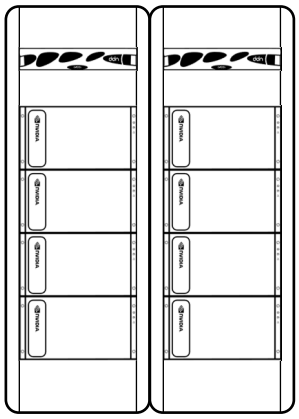


Figure 12. DDN A^{3}I reference architecture for AI400X with DGX A100 systems in a 1:4 configuration.

ai400x



DGX POD with Eight DGX A100 Systems

Figure 13 illustrates the DDN A^{3}I architecture in a 2:8 configuration in which eight DGX A100 systems are connected with two AI400X storage appliances through a pair of network switches that are configured for high-availability (HA). Every DGX A100 system connects to each of the network switches via one HDR IB or 100 GbE links. The AI4000X storage appliance connects to each of the network switches via four HDR IB or 100 GbE links. The network switches are interconnected eight dedicated links. This ensures non-blocking data communication between every device connected to the network. The HA design provides full-redundancy and maximum data availability in case of component failure in one of the devices.

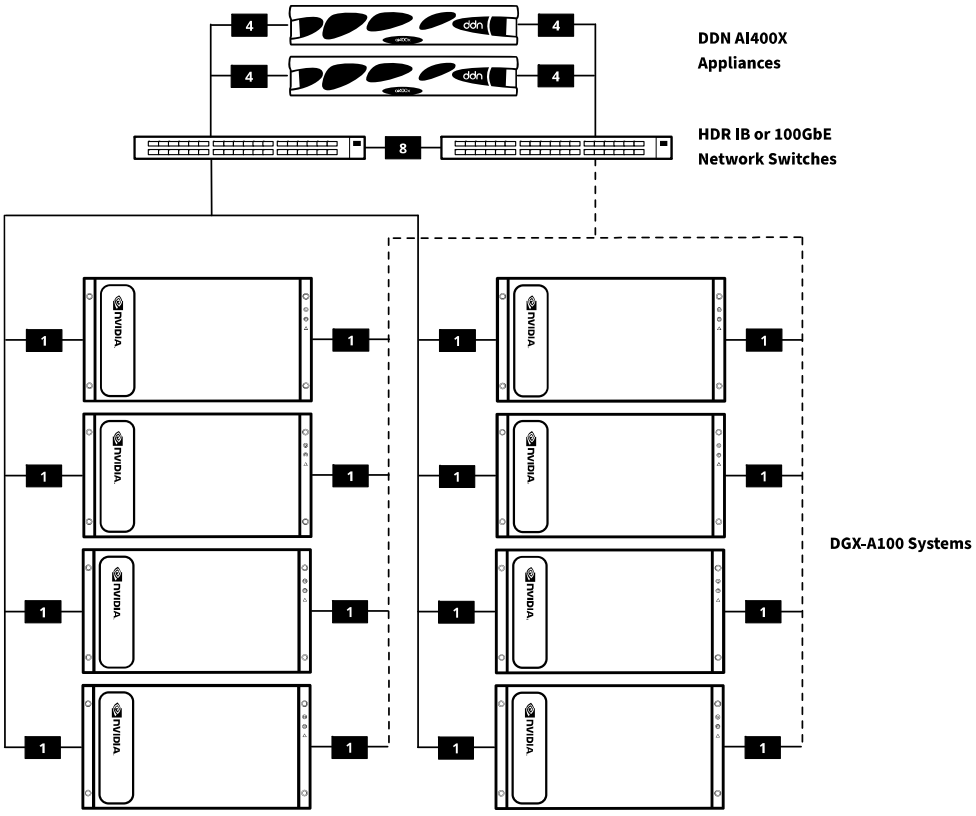


Figure 13. DDN A^{3}I reference architecture for AI400X with DGX A100 systems in a 2:8 configuration.

POWERFUL & PROVEN DGX AT-SCALE SOLUTIONS WITH DDN AI400X

DDN A³I Deployment Architectures for Large DGX A100 System Clusters

DDN proposes the following reference architectures for at-scale deployments of DGX A100 systems clusters. DDN A³I solutions are powerful and proven with the largest DGX A100 systems clusters currently in operation, including the large clusters operated by NVIDIA.

The DDN AI400X is a turnkey fully-integrated and optimized appliance for at-scale DGX deployments. DDN recommends the AI400X as the optimal data platform for large clusters of DGX A100 systems. The AI400X delivers maximum GPU performance for every workload and data type in a dense, power efficient 2RU chassis. The AI400X simplifies the design, deployment and management of a POD and provides predictable performance, capacity, and scaling. The AI400X arrives fully configured, ready to deploy and installs in minutes. The appliance is fully-optimized for seamless integration with DGX systems, and enables customers to move rapidly from test to production. DDN provides complete expert design, deployment, and support services globally and ensures best customer experience. The DDN field engineering organization has already deployed hundreds of solutions for customers based on the A³I reference architectures.

The DDN A³I shared parallel architecture is proven to deliver at-scale performance required by customers of multiple DGX A100 systems. The technology is used to enable and accelerate over 2/3 of the world's largest supercomputers. The DDN A³I shared parallel architecture maximizes at-scale infrastructure performance, streamlines end-to-end workflows, simplifies data management, and enables operators of clustered DGX A100 systems to scale flexibly and in full-confidence. The platform is feature rich and includes advanced capabilities that are ideal for DGX A100 systems cluster operations: full support for container applications, secure multitenancy, intelligent data management, enhanced data security and governance, as well as extensive monitoring, logging and simple management features to ensure optimal system performance and reliability. The DDN A³I shared parallel architecture extends beyond the DGX A100 system clusters with comprehensive multiprotocol and multicloud facilities.

DDN proposes the following reference architectures for DGX A100 systems cluster configurations, based on 20 client node increments. These configurations can be adjusted and scaled easily to match specific workload requirements. For the storage network, DDN recommends HDR200 InfiniBand technology in a non-blocking topology, with redundancy to ensure data availability. DDN recommends use of at least two HDR200 connections per DGX A100 system to the storage network.

DDN A³I All-Flash Configuration for DGX A100 Systems At-Scale

This configuration is designed entirely with all-NVMe AI400X and provides best performance for all workloads and data types, for maximum operational flexibility with DGX A100 systems clusters (Figure 15). The DDN A³I shared parallel architecture enables all nodes in the cluster to access data simultaneously through a single unified name space and extends the full advantages of NVME devices all the way to GPUs and containerized applications across the DGX A100 systems. This unified architecture eliminates data management and movement to local node cache. DDN A³I solutions interfaces easily with file, object and cloud data repositories for data ingest, output and archive. This configuration scales easily, reliably, and predictably using turnkey appliances to match evolving workflow requirements.

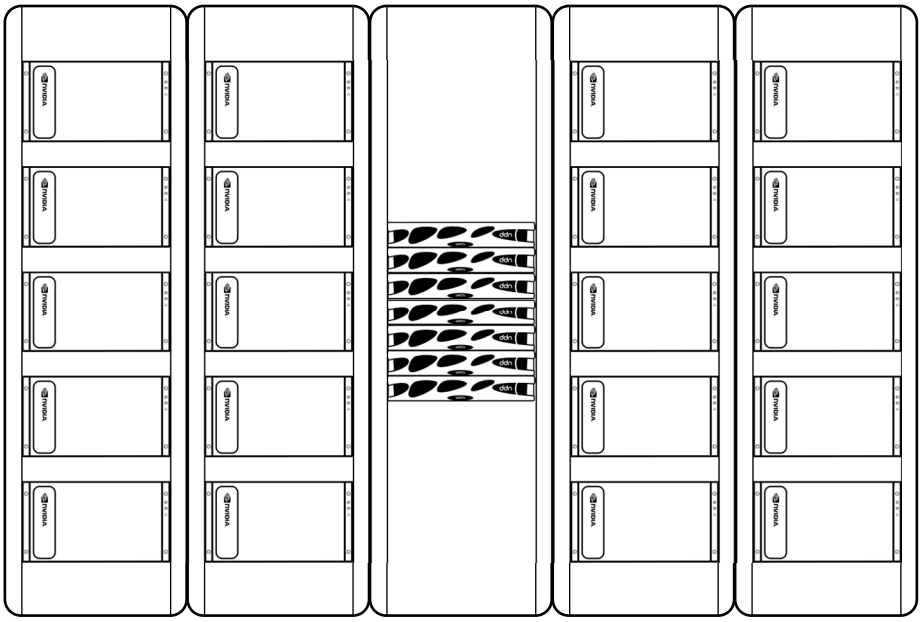


Figure 14. Rack illustrations for DDN A³I reference configurations.

Technical Specifications

Configuration	7 DDN AI400X
Capacity	2 PB (NVME)
Performance	336 GB/s Throughput, 21 M IOPS
Networking	56 HDR100 or 100GbE
Footprint	14 RU, 10.5Kw Nominal

DDN A³I Hybrid Solutions for DGX A100 Systems At-Scale

This hybrid configuration augments the data platform with a large capacity storage repository, ideal to maintain large datasets close to the DGX A100 systems for fast and flexible access (Figure 16). This configuration extends the AI400X with additional capacity disk and delivers both high-performance and deep capacity tiers from the same system This maximizes the physical footprint for the data platform with highest performance/density configuration available. The capacity and performance tiers can scale independently, and interface seamlessly with file, object and cloud-based data repository for simplified at-scale data management. The hybrid configuration greatly simplifies end-to-end workflows, notably DL, with integrated secure data ingest, processing and retention capabilities.

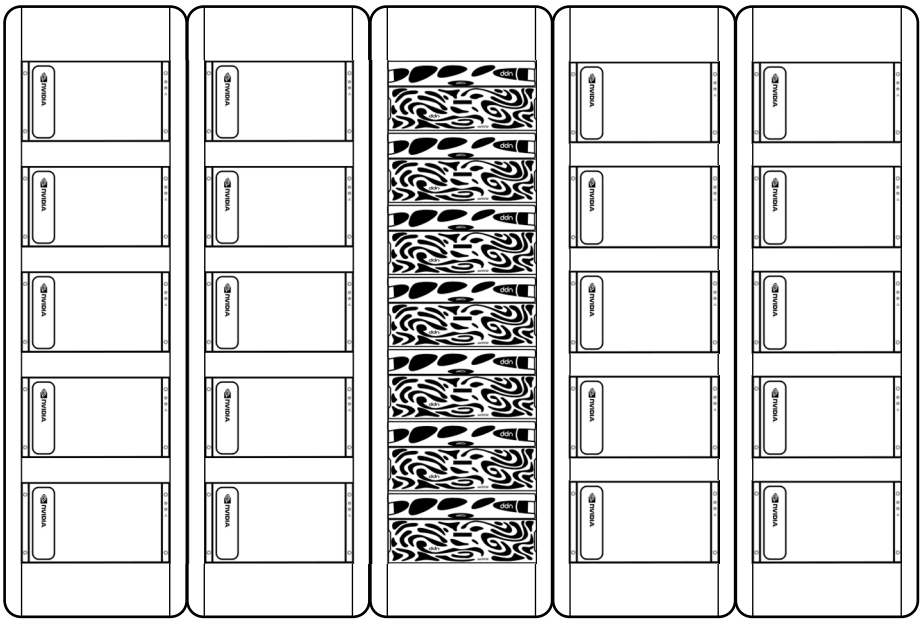


Figure 15. Rack illustrations for DDN A³I reference configurations.

Technical Specifications

Configuration	7 DDN AI400X Hybrid
Capacity	2 PB (NVME), 7PB/14PB (HDD)
Performance	336 GB/s Throughput, 21 M IOPS
Networking	56 HDR100 or 100GbE
Footprint	42 RU, 18.5Kw Nominal

Contact DDN to Unleash the Power of Your AI

DDN has long been a partner of choice for organizations pursuing at-scale data-driven projects. Beyond technology platforms with proven capability, DDN provides significant technical expertise through its global research and development and field technical organizations.

A worldwide team with hundreds of engineers and technical experts can be called upon to optimize every phase of a customer project: initial inception, solution architecture, systems deployment, customer support and future scaling needs.

Strong customer focus coupled with technical excellence and deep field experience ensures that DDN delivers the best possible solution to any challenge. Taking a consultative approach, DDN experts will perform an in-depth evaluation of requirements and provide application-level optimization of data workflows for a project. They will then design and propose an optimized, highly reliable and easy to use solution that best enables and accelerates the customer effort.

Drawing from the company's rich history in successfully deploying large scale projects, DDN experts will create a structured program to define and execute a testing protocol that reflects the customer environment and meet and exceed project objectives. DDN has equipped its laboratories with leading GPU compute platforms to provide unique benchmarking and testing capabilities for AI and DL applications.

Contact DDN today and engage our team of experts to unleash the power of your AI projects.

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.

V6 8/20