DDN A³I® SOLUTIONS WITH NVIDIA DGX SUPERPOD™

Fully-integrated and optimized infrastructure solutions for accelerated at-scale AI, Analytics and HPC featuring NVIDIA DGX H100 Systems

Executive Summary

DDN A^{3}I Solutions are proven at-scale to deliver optimal data performance for Artificial Intelligence (AI), Data Analytics and High-Performance Computing (HPC) applications running on GPUs in NVIDIA DGX H100™ systems. DDN data platforms power nearly all NVIDIA DGX SuperPOD™ systems currently in operation, including NVIDIA SELENE, Cambridge-1 and many others.

This document describes fully validated reference architectures for scalable DGX SuperPOD configurations. The solutions integrate DDN AI400X2 appliances and DDN Insight software with DGX SuperPOD.

1. DDN A³I END-TO-END ENABLEMENT FOR NVIDIA DGX SUPERPOD

DDN A³I Solutions are proven at-scale to deliver optimal data performance for Artificial Intelligence (AI), Data Analytics and High-Performance Computing (HPC) applications running on GPUs in NVIDIA DGX H100™ systems. DDN data platforms power nearly all NVIDIA DGX SuperPOD™ systems currently in operation, including NVIDIA SELENE, Cambridge-1 and many others.

This document describes fully validated reference architectures for scalable DGX SuperPOD configurations. The solutions integrate DDN AI400X2 appliances and DDN Insight software with DGX SuperPOD.

1.1 DDN A³I SHARED PARALLEL ARCHITECTURE

The DDN A³I shared parallel architecture and client protocol ensures high levels of 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 systems 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.

1.2 DDN A³I STREAMLINED DEEP LEARNING WORKFLOWS

DDN A3I solutions enable and accelerate end-to-end data pipelines for deep learning (DL) workflows of all scale running on DGX systems. 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 in a DGX SuperPOD 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 reliably on DGX systems.

1.3 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. The latest generation NVIDIA Quantum InfiniBand (IB) and Spectrum Ethernet technology provide both high-bandwidth and low-latency data transfers between applications, compute servers and storage appliances.

DDN A³I Multirail enables grouping of multiple network interfaces on a DGX system 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 for DGX SuperPOD deployments.

1.4 DDN A³I ADVANCED OPTIMIZATIONS FOR DGX H100 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 H100 systems has been validated at-scale with the largest DGX SuperPOD with DGX A100 and H100 systems deployments currently in operation.

1.5 DDN A³I HOT NODES

DDN Hot Nodes is a powerful software enhancement that enables the use of the NVME devices in a DGX system as a local cache for read-only operations. This method significantly improves the performance of applications if a data set is accessed multiple times during a particular workflow.

This is typical with DL training, where the same input data set or portions of the same input data set are accessed repeatedly over multiple training iterations. Traditionally, the application on the DGX system reads the input data set from shared storage directly, thereby continuously consuming shared storage resources. With Hot Nodes, as the input data is read during the first training iteration, the DDN software automatically writes a copy of the data on the local NVME devices. During subsequent reads, data is delivered to the application from the local cache rather than the shared storage. This entire process is managed by the DDN client software running on the DGX system. Data access is seamless and the cache is fully transparent to users and applications. The use of the local cache eliminates network traffic and reduces the load on the shared storage system. This allows other critical DL training operations like checkpointing to complete faster by engaging the full capabilities of the shared storage system.

DDN Hot Nodes includes extensive data management tools and performance monitoring facilities. These tools enable user-driven local cache management, and make integration simple with task schedulers. For example, training input data can be loaded to the local cache on a DGX system as a pre-flight task before the AI training application is engaged. As well, the metrics expose insightful information about cache utilization and performance, enabling system administrators to further optimize their data loading and maximize application and infrastructure efficiency gains.

1.6 DDN A³I MULTITENANCY

Through its built-in digital security framework, DDN A³I software makes it very easy to operate a DGX SuperPOD as a secure multitenant environment. 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.

1.7 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 the application containers on the DGX H100 system 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 DGX systems are available from NVIDIA. They provide a solid foundation that enables data scientists to rapidly develop and deploy applications on DGX systems. 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 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.

1.8 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 a 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.

1.9 DDN A³I CSI DRIVER

The DDN CSI driver enables optimized data access for workloads managed by Kubernetes within DGX SuperPOD. The driver provides direct control facilities for the container orchestrator for fully automated storage management. Several types of volumes and data access modes are supported. The driver also enables Kubernetes workloads to dynamically provision and deprovision storage resources based on the workload's requirements. The deep integration between DDN software and the orchestrator stack ensures most efficient management and utilization of storage resources. DDN engages in continuous software validation with CSI standards to ensure ongoing compatibility and integration of new capabilities as the standard evolves.

2. DDN A³I SOLUTIONS WITH NVIDIA DGX H100 SYSTEMS

The DDN A³I scalable architecture integrates DGX H100 systems with DDN AI shared parallel file storage appliances and delivers fully-optimized end-to-end AI, Analytics and HPC workflow acceleration on GPUs. DDN A³I solutions greatly simplify the deployment of DGX H100 systems in DGX SuperPOD configurations, while also delivering performance and efficiency for maximum GPU saturation, and high levels of scalability.

This section describes the components integrated in DDN A3I Solutions for DGX SuperPOD.

2.1 DDN AI400X2 APPLIANCE

The AI400X2 appliance is a fully integrated and optimized shared data platform with predictable capacity, capability, and performance. Every AI400X2 appliance delivers over 90 GB/s and 3M IOPS directly to DGX H100 systems in DGX SuperPOD. Shared performance scales linearly as additional AI400X2 appliances are integrated to DGX SuperPOD. The all-NVMe configuration provides optimal performance for a wide variety of workload and data types and ensures that DGX SuperPOD operators achieve the most from at-scale GPU applications, while maintaining a single, shared, centralized data platform.

The AI400X2 appliance integrates the DDN A³I shared parallel architecture and includes a wide range of capabilities described in section 1, including automated data management, digital security, and data protection, as well as extensive monitoring. The AI400X2 appliances enables DGX SuperPOD operators to go beyond basic infrastructure and implement complete data governance pipelines at-scale.

The AI400X2 appliance integrates with DGX SuperPOD over InfiniBand, Ethernet and RoCE. It is available in 60, 120, 250 and 500 TB all-NVMe capacity configurations. Optional hybrid configurations with integrated HDDs are also available for deployments requiring high-density deep capacity storage. Contact DDN Sales for more information.

Figure 1. DDN AI400X2 all-NVME storage appliance.

2.2 DDN INSIGHT SOFTWARE

DDN Insight is a centralized management and monitoring software suite for AI400X2 appliances. It provides extensive performance and health monitoring of all DDN storage systems connected to DGX SuperPOD from a single web-based user interface. DDN Insight greatly simplifies IT operations and enables automated and proactive storage platform management guided by analytics and intelligent software.

Performance monitoring is an important aspect of operating a DGX SuperPOD efficiently. Provided the several variables that affect data I/O performance, the identification of bottlenecks and degradation is crucial while production workloads are engaged. DDN Insight provides deep real-time analysis across the entire DGX SuperPOD, tracking I/O transactions from applications running on DGX nodes all the way through individual drives in the AI400X2 appliances. The embedded analytics engine makes it simple for DGX SuperPOD operators to visualize I/O performance across their entire infrastructure through intuitive user interfaces. These include extensive logging, trending, and comparison tools, for analyzing I/O performance of specific applications and users over time. As well, the open backend database makes it simple to extend the benefits of DDN Insight and integrate other AI infrastructure components within the engine, or export data to third party monitoring systems.

DDN Insight is available a software-installable package on customer-supplied management servers, and as a turnkey server appliance from DDN.
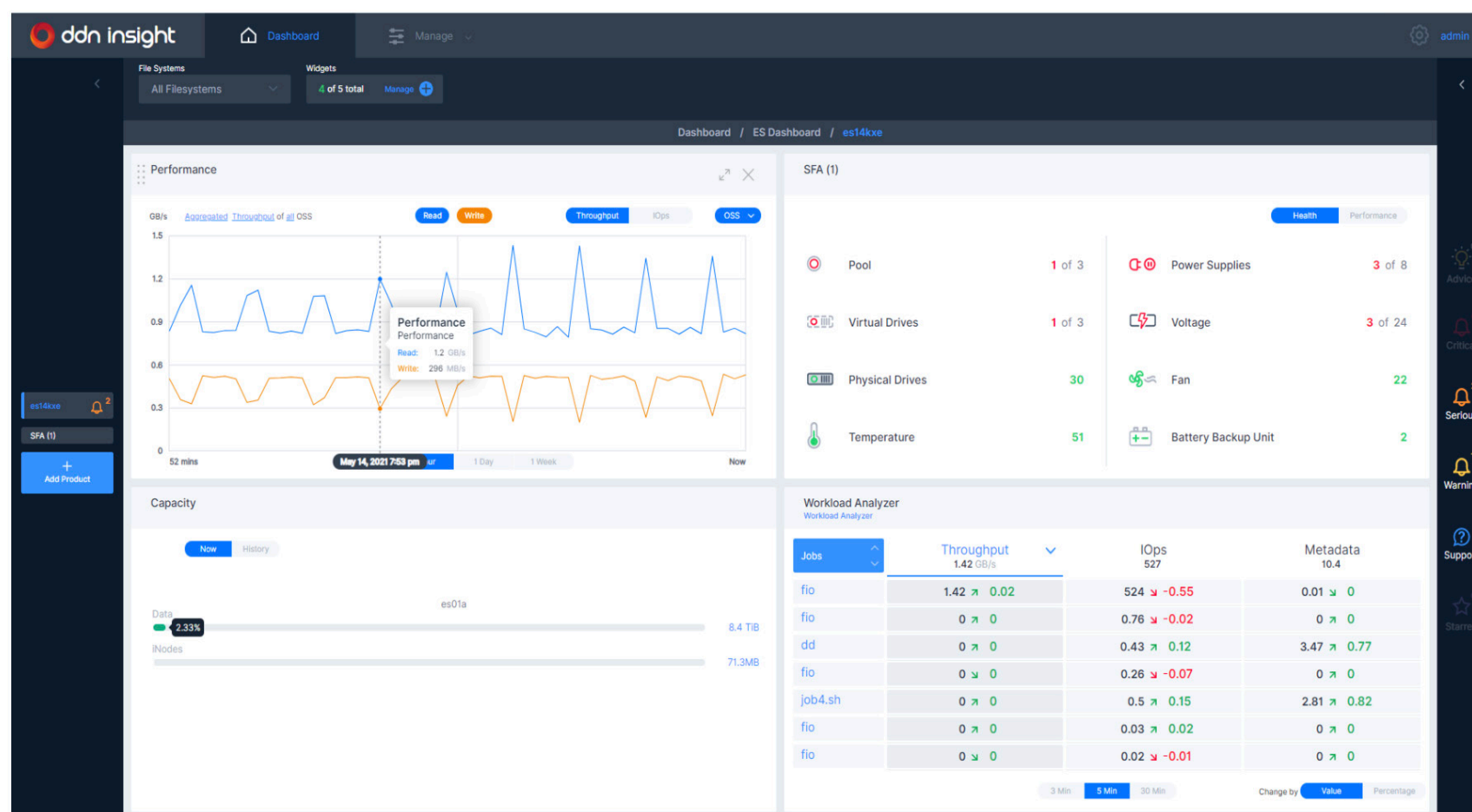

Figure 2. DDN Insight Workload Analyzer Tool.

3. DDN A³I RECOMMENDATIONS FOR DGX SUPERPOD

DDN proposes the following recommended architectures for DGX H100 SuperPOD configurations. DDN A3I solutions are fully validated with NVIDIA and already deployed with several DGX BasePOD and DGX SuperPOD customers worldwide.

The DDN AI400X2 appliance is a turnkey appliance fully-validated and proven with DGX H100 SuperPOD. The AI400X2 appliances delivers optimal GPU performance for every workload and data type in a dense, power efficient 2RU chassis. The AI400X2 appliance simplifies the design, deployment, and management of a DGX SuperPOD and provides predictable performance, capacity, and scaling. The appliance is designed for seamless integration with DGX systems and enables customers to move rapidly from test to production. As well, DDN provides complete expert design, deployment, and support services globally. The DDN field engineering organization has already deployed dozens of solutions for customers based on the A³I reference architectures.

RECOMMENDED STORAGE SIZING

As general guidance, DDN recommends the shared storage be sized to ensure at least 1 GB/s per second of read and write throughput for every H100 GPU in a DGX SuperPOD (Table 1). This ensures minimum performance required to operate the GPU infrastructure. All four configurations are illustrated in this guide.

DGX SuperPOD configuration	32 DGX H100	64 DGX H100	96 DGX H100	128 DGX H100
Recommended DDN storage	4 AI400X2	8 AI400X2	12 AI400X2	16 AI400X2
Shared read throughput	360 GB/s	720 GB/s	1 TB/s	1.4 TB/s
Shared write throughput	260 GB/s	520 GB/s	780 GB/s	1 TB/s
Per GPU read throughput	1.4 GB/s	1.4 GB/s	1.4 GB/s	1.4 GB/s
Per GPU write throughput	1 GB/s	1 GB/s	1 GB/s	1 GB/s

Table 1. Recommended storage sizing for DGX SuperPOD.

3.1 STORAGE NETWORK FOR DGX SUPERPOD

The DGX SuperPOD reference design includes several networks, one of which is dedicated to storage traffic. The storage network provides connectivity between the AI400X2 appliances, the compute nodes and management nodes. This network is designed to meet the high-throughput, low-latency, and scalability requirements of DGX SuperPOD.

The NVIDIA QM9700 InfiniBand Switch is recommended for DGX SuperPOD storage connectivity. It provides 64 ports of NDR over 32 OSFP ports in a 1RU form factor. DDN recommends the QM9700 switch to deploy DGX SuperPOD. Validated cabling configurations are detailed in section 3.1.3.

Figure 3. NVIDIA QM9700 InfiniBand Switch.

An overview of the DGX SuperPOD storage network architecture is shown in Figure 4, recommended network connections for each DGX H100 system on Figure 5, and recommended network connections for each AI400X2 appliance on Figure 6.
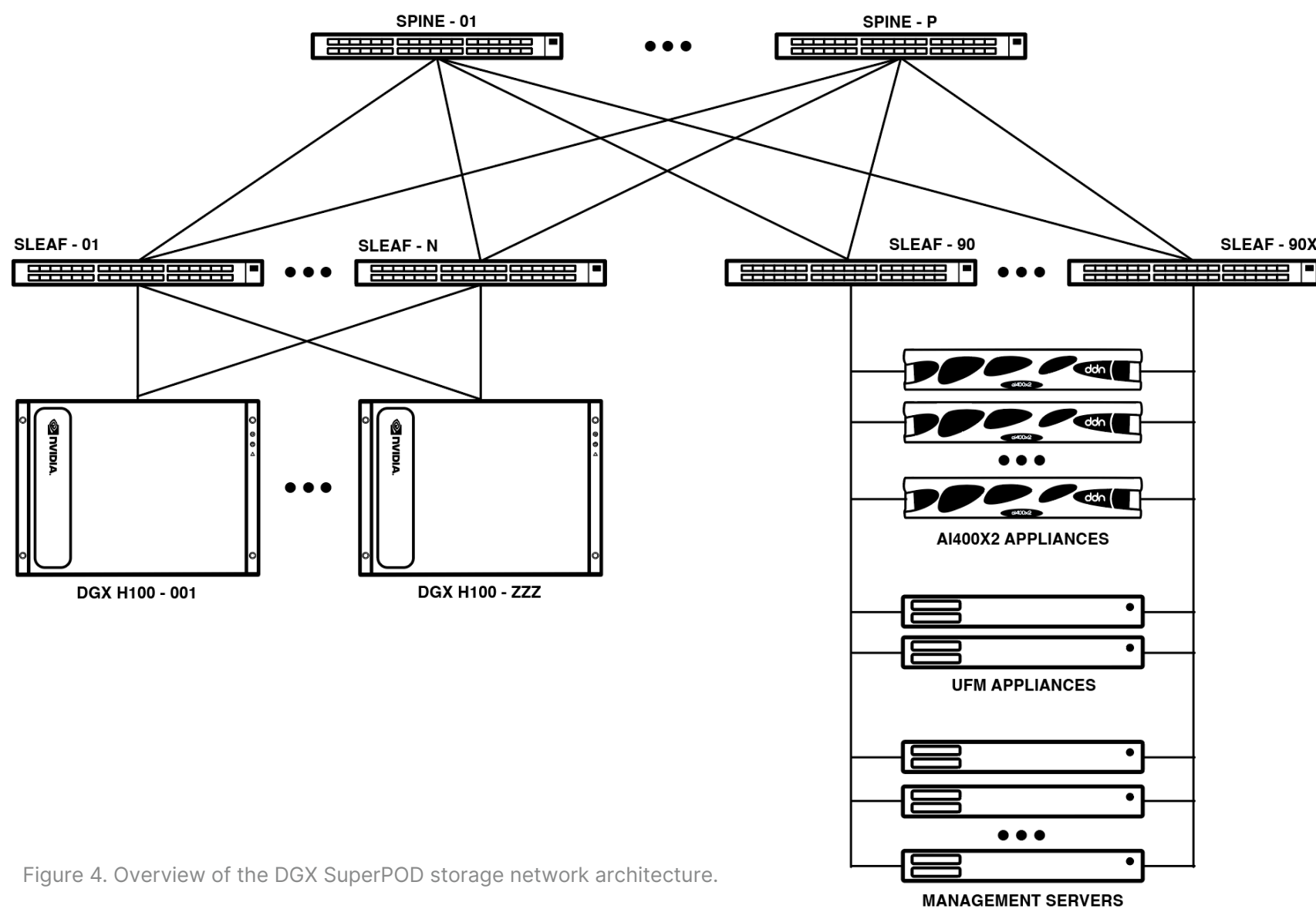

Figure 4. Overview of the DGX SuperPOD storage network architecture.

3.1.1 DGX H100 SYSTEMS NETWORK CONNECTIVITY

For DGX SuperPOD, DDN recommends ports 1 to 4 on the DGX H100 systems be connected to the compute network. Ports 5 and 7 should be connected to the in-band management network, which also serves as the storage network for Ethernet storage configurations. For InfiniBand storage configurations, ports 6 and 8 should be connected to the InfiniBand storage network. As well, the management BMC ("B") port should be connected to the out-of-band management network.
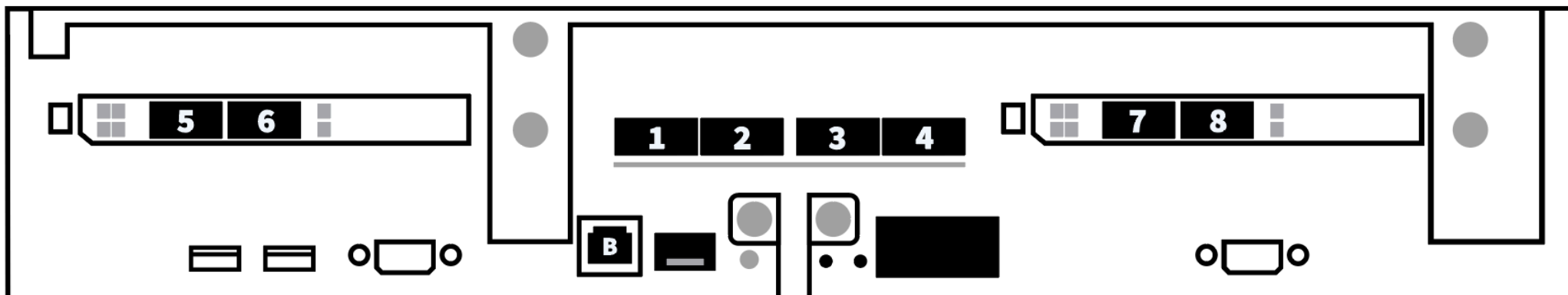

Figure 5. Recommended DGX H100 system network port connections.

3.1.2 AI400X2 APPLIANCE NETWORK CONNECTIVITY

For DGX SuperPOD, DDN recommends ports 1 to 8 on the AI400X2 appliance be connected to the storage network. As well, the management ("M") and BMC ("B") ports for both controllers should be connected to the out-of-band management network. Note that each AI400X2 appliance requires one inter-controller network port connection ("I") using short ethernet cable supplied.
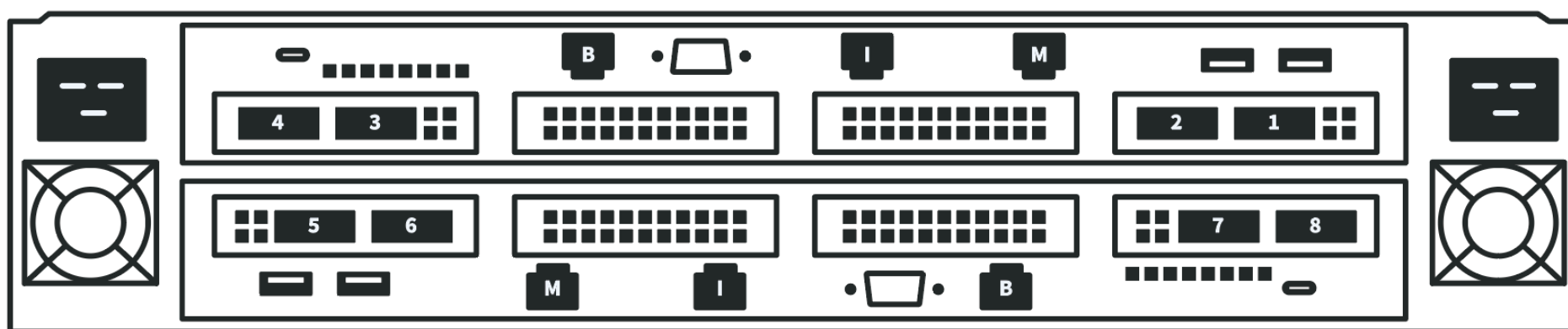

Figure 6. Recommended AI400X2 appliance network port connections.

3.1.3 AI400X2 APPLIANCE CABLING WITH NDR 400GBPS INFINIBAND SWITCHES

The AI400X2 appliance connects to the DGX SuperPOD storage network with 8 200 Gbps InfiniBand interfaces. Particular attention must be given to the cabling selection to ensure compatibility between different InfiniBand connectivity and data rates. DDN and NVIDIA have validated the following cables to connect AI400X2 appliances with QM9700 switches. The use of splitter cables ensures most efficient use of switch ports.

Direct Attached Passive Copper Cables

MCP7Y60-Hxxx 980-9I46K-00Hxxx	NVIDIA DAC 1:2 splitter, InfiniBand 400Gb/s to 2x 200Gb/s, OSFP to 2x QSFP56 xxx indicates length in meters: 001, 01A (1.5m), 002

Requires 4 cables per AI400X2 appliance.

Active Optical Cables

MFA7U10-Hxxx 980-9I117-00Hxxx	NVIDIA AOC 1:2 splitter, InfiniBand 400Gb/s to 2x 200Gb/s, OSFP to 2x QSFP56 xxx indicates length in meters: 003, 005, 010, 015, 020, 030

Requires 4 cables per AI400X2 appliance.

3.2 DGX SUPERPOD WITH 32 DGX H100 SYSTEMS

Figure 7 illustrates the DDN A^{3}I architecture for DGX SuperPOD with 32 DGX H100 systems, 4 DDN AI400X2 appliances and a DDN Insight server. Every DGX H100 system connects to the storage network with two NDR 400Gb/s InfiniBand links. Each AI400X2 appliance connects to the storage network with 8 InfiniBand links using the appropriate cable type. The DDN Insight server connects to the AI400X2 appliances over the 1GbE out-of-band management network. It does not require a connection to the InfiniBand storage network.
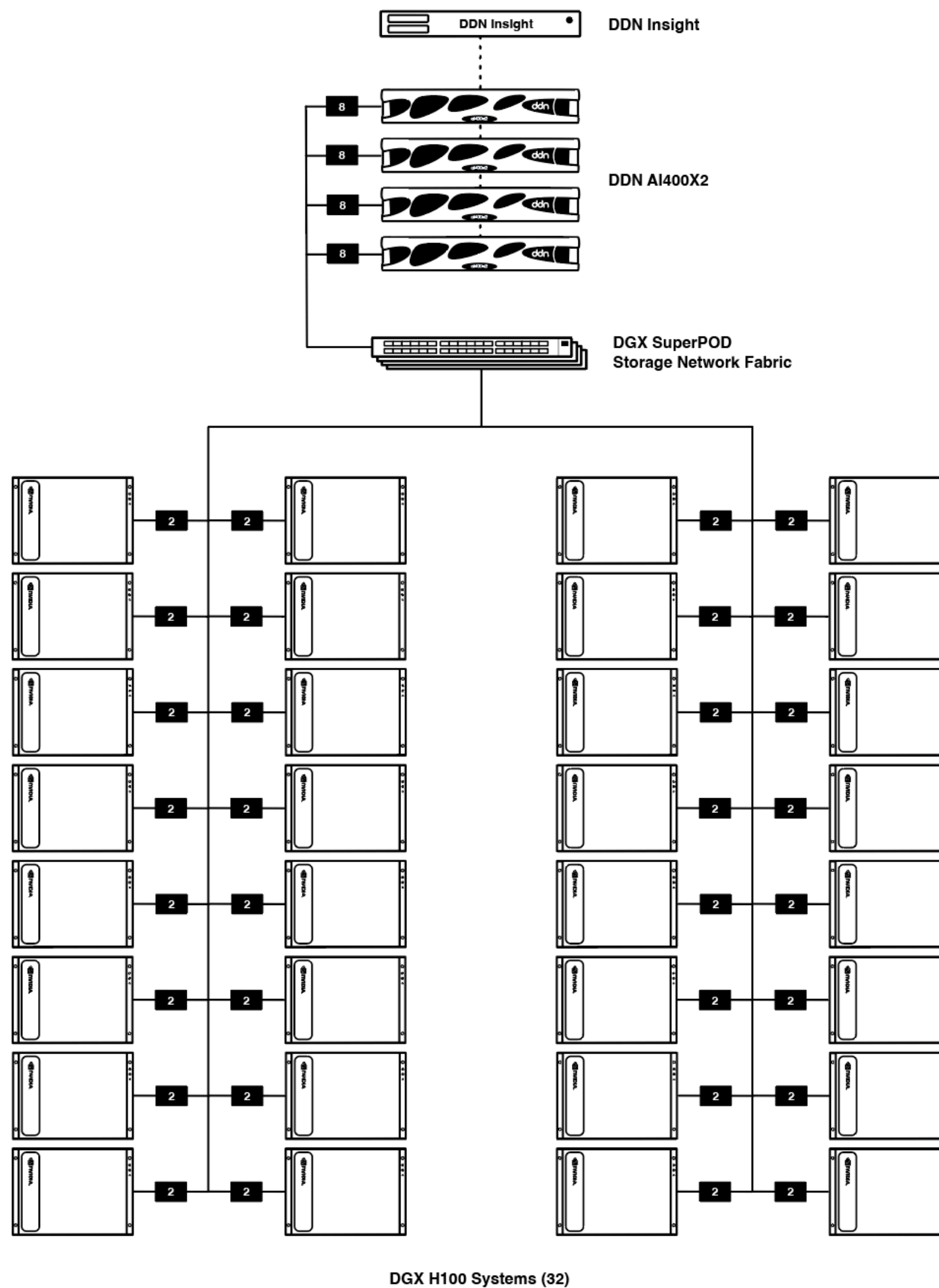

Figure 7. DDN A^{3}I SuperPOD reference architecture with 32 DGX H100 systems.

3.3 DGX SUPERPOD WITH 64 DGX H100 SYSTEMS

Figure 8 illustrates the DDN A³I architecture for DGX SuperPOD with 64 DGX H100 systems, 8 DDN AI400X2 appliances and a DDN Insight server. Every DGX H100 system connects to the storage network with two NDR 400Gb/s InfiniBand links. Each AI400X2 appliance connects to the storage network with 8 InfiniBand links using the appropriate cable type. The DDN Insight server connects to the AI400X2 appliances over the 1GbE out-of-band management network. It does not require a connection to the InfiniBand storage network.
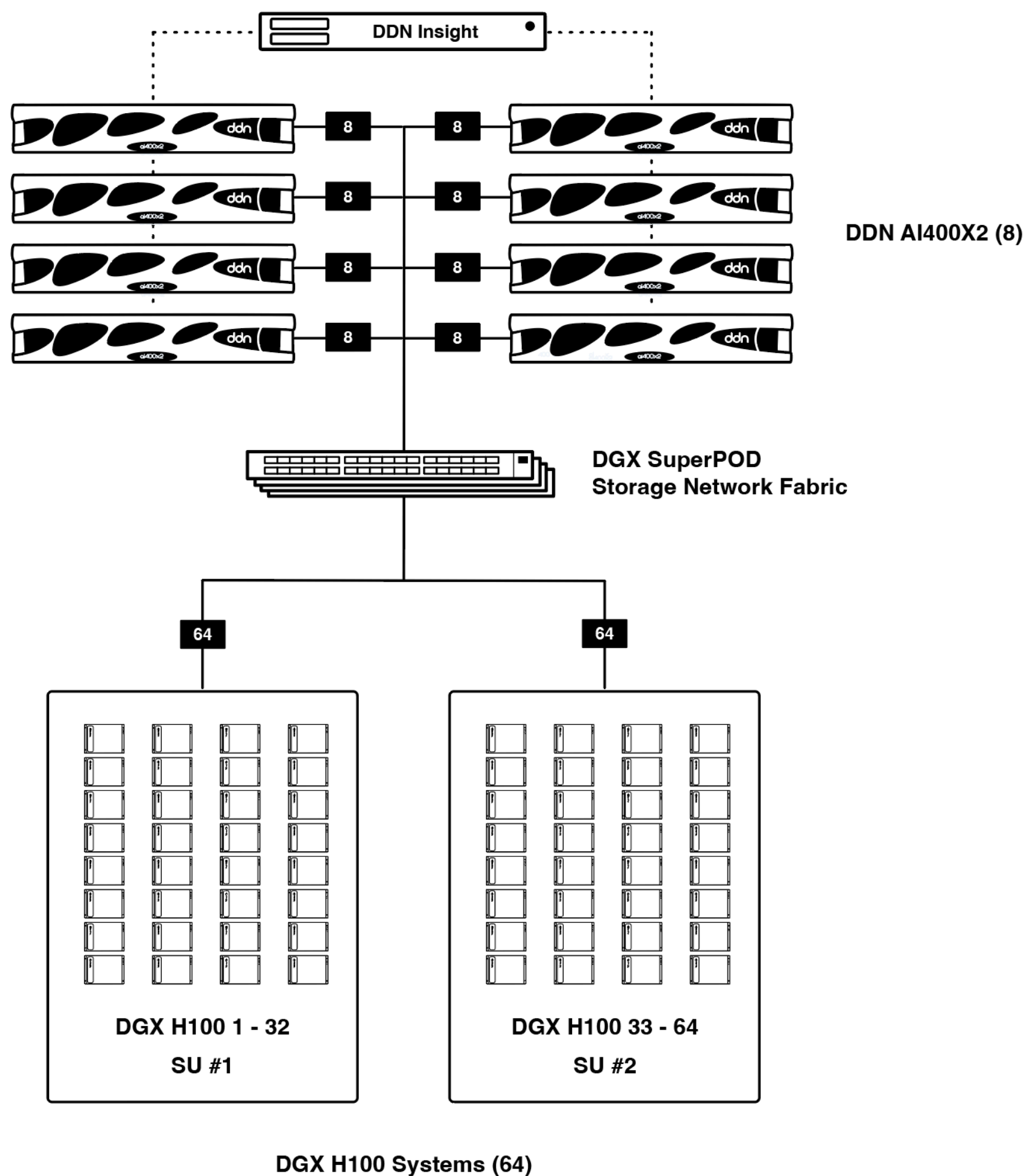

Figure 8. DDN A³I SuperPOD reference architecture with 64 DGX H100 systems.

3.4 DGX SUPERPOD WITH 96 DGX H100 SYSTEMS

Figure 9 illustrates the DDN A³I architecture for DGX SuperPOD with 96 DGX H100 systems, 12 DDN AI400X2 appliances and a DDN Insight server. Every DGX H100 system connects to the storage network with two NDR 400Gb/s InfiniBand links. Each AI400X2 appliance connects to the storage network with 8 InfiniBand links using the appropriate cable type. The DDN Insight server connects to the AI400X2 appliances over the 1GbE out-of-band management network. It does not require a connection to the InfiniBand storage network.
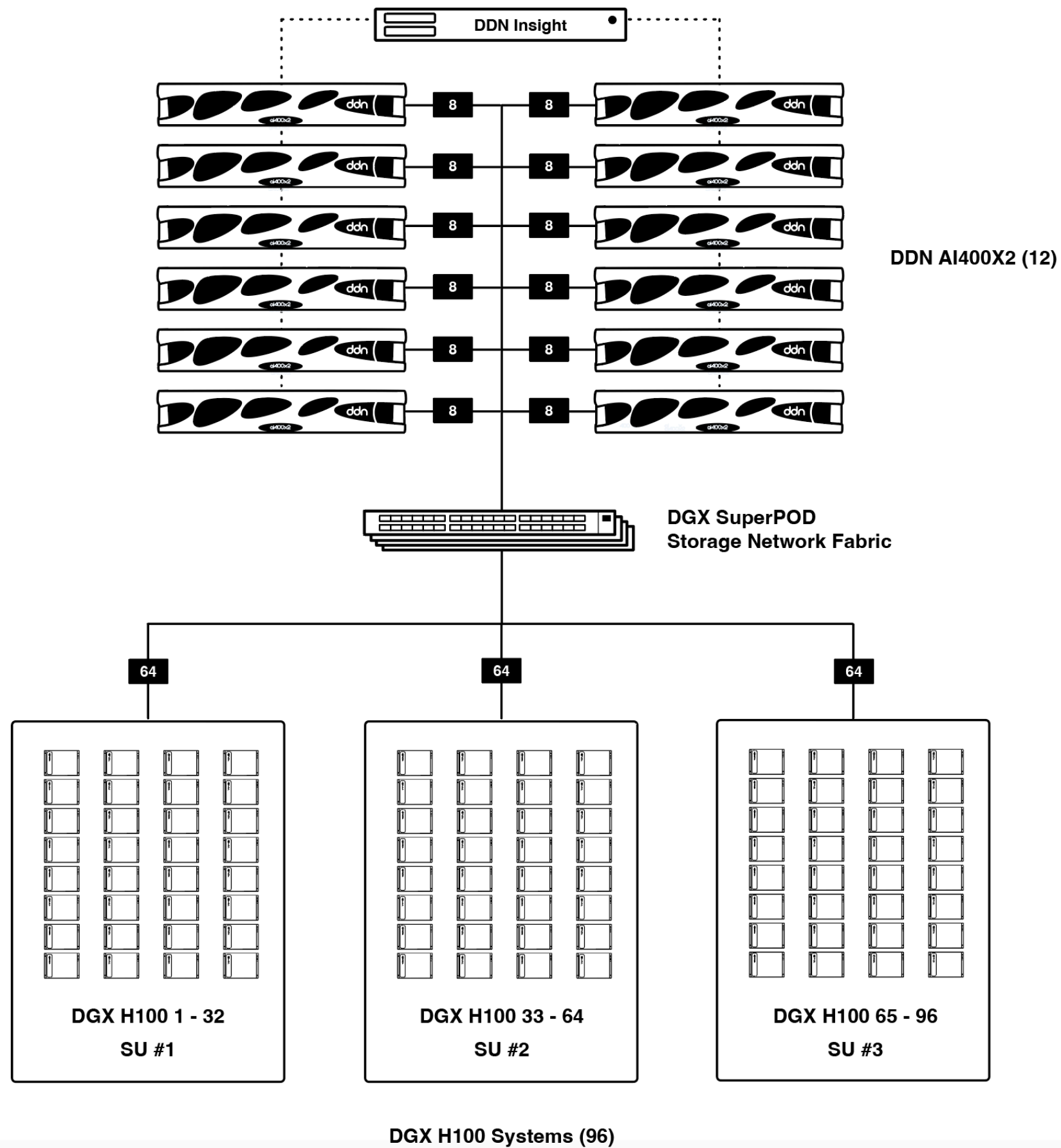

Figure 9. DDN A³I SuperPOD reference architecture with 96 DGX H100 systems.

3.5 DGX SUPERPOD WITH 128 DGX H100 SYSTEMS

Figure 10 illustrates the DDN A³I architecture for DGX SuperPOD with 128 DGX H100 systems, 16 DDN AI400X2 appliances and a DDN Insight server. Every DGX H100 system connects to the storage network with two NDR 400Gb/s InfiniBand links. Each AI400X2 appliance connects to the storage network with 8 InfiniBand links using the appropriate cable type. The DDN Insight server connects to the AI400X2 appliances over the 1GbE out-of-band management network. It does not require a connection to the InfiniBand storage network.
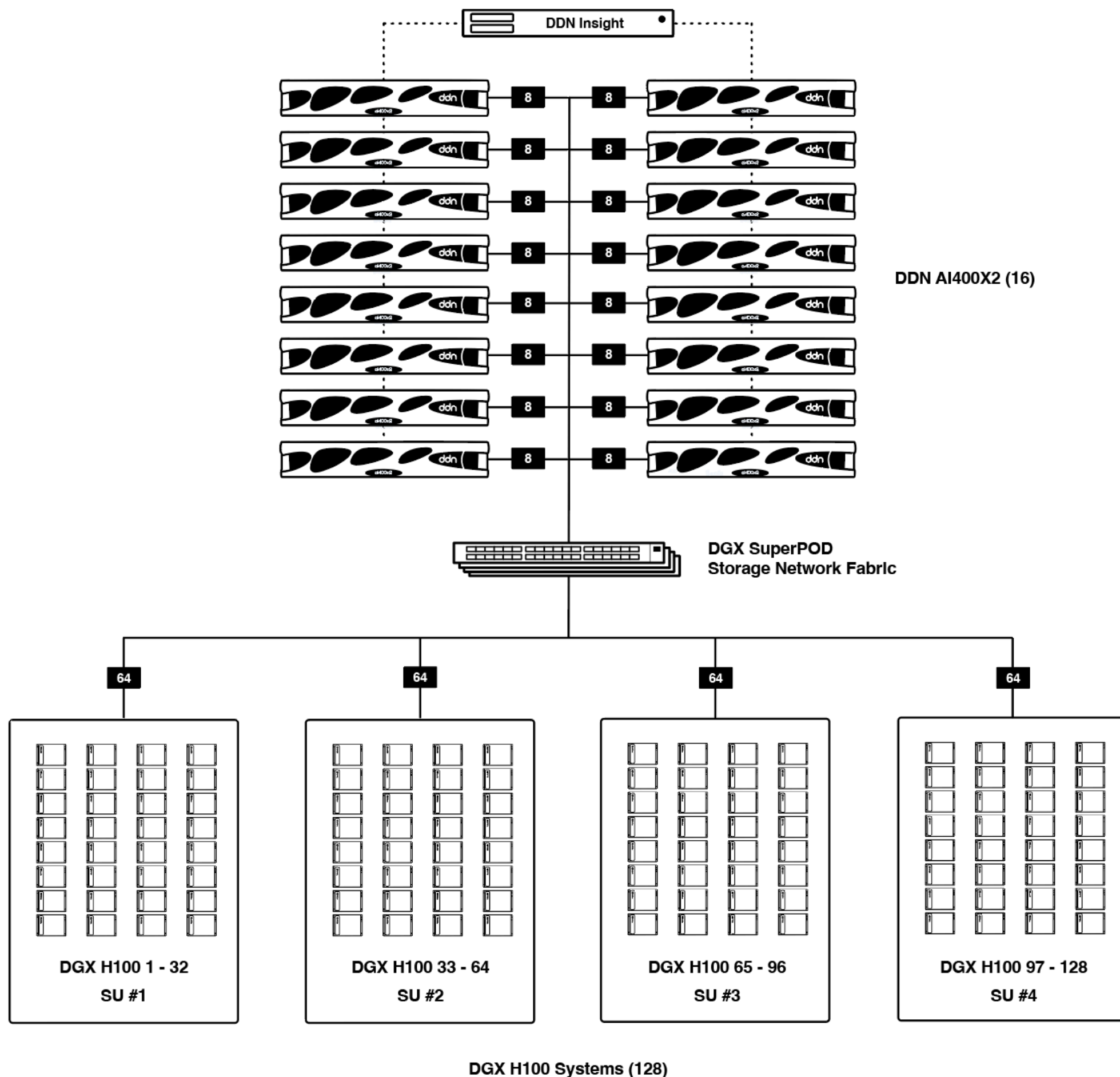

Figure 10. DDN A³I SuperPOD reference architecture with 128 DGX H100 systems.

4. DDN A³I SOLUTIONS WITH NVIDIA DGX SUPERPOD VALIDATION

DDN conducts extensive engineering integration, optimization, and validation efforts in close collaboration with NVIDIA and customers to ensure the best possible end-user experience using the reference designs in this document. The joint validation confirms functional integration, and optimal performance out-of-the-box for DGX SuperPOD configurations.

Performance testing on the DDN A³I architecture is conducted with industry standard synthetic throughput and IOPS applications, as well as widely used DL frameworks and data types. The results demonstrate that with the DDN A³I shared parallel architecture, GPU-accelerated applications can engage the full capabilities of the data infrastructure and the DGX H100 systems. Performance is distributed evenly across all the DGX H100 systems in the DGX SuperPOD, and scales linearly as more DGX H100 systems are engaged.

This section details some of the results from recent at-scale testing integrating AI400X2 appliances with DGX H100 systems in a SuperPOD deployment.

4.1 DGX SUPERPOD SYSTEM PERFORMANCE VALIDATION

The tests described in this section were executed on thirty-two DGX H100 systems each equipped with eight H100 GPUs running DGX OS Server Software 6.0.11 and four AI400X2 appliances running DDN EXAScaler 6.2.0.

For the storage network, all thirty-two DGX H100 systems are connected to NVIDIA Quantum-2 QM9700 switches with two NDR 400Gb/s InfiniBand links, one per dual-ported adapter (see recommendation in section 3.1). The AI400X2 appliances are connected to the same network with eight HDR 200Gb/s InfiniBand links each.

This test environment allows us to demonstrate storage performance with a DGX H100 SuperPOD scalable unit configuration.
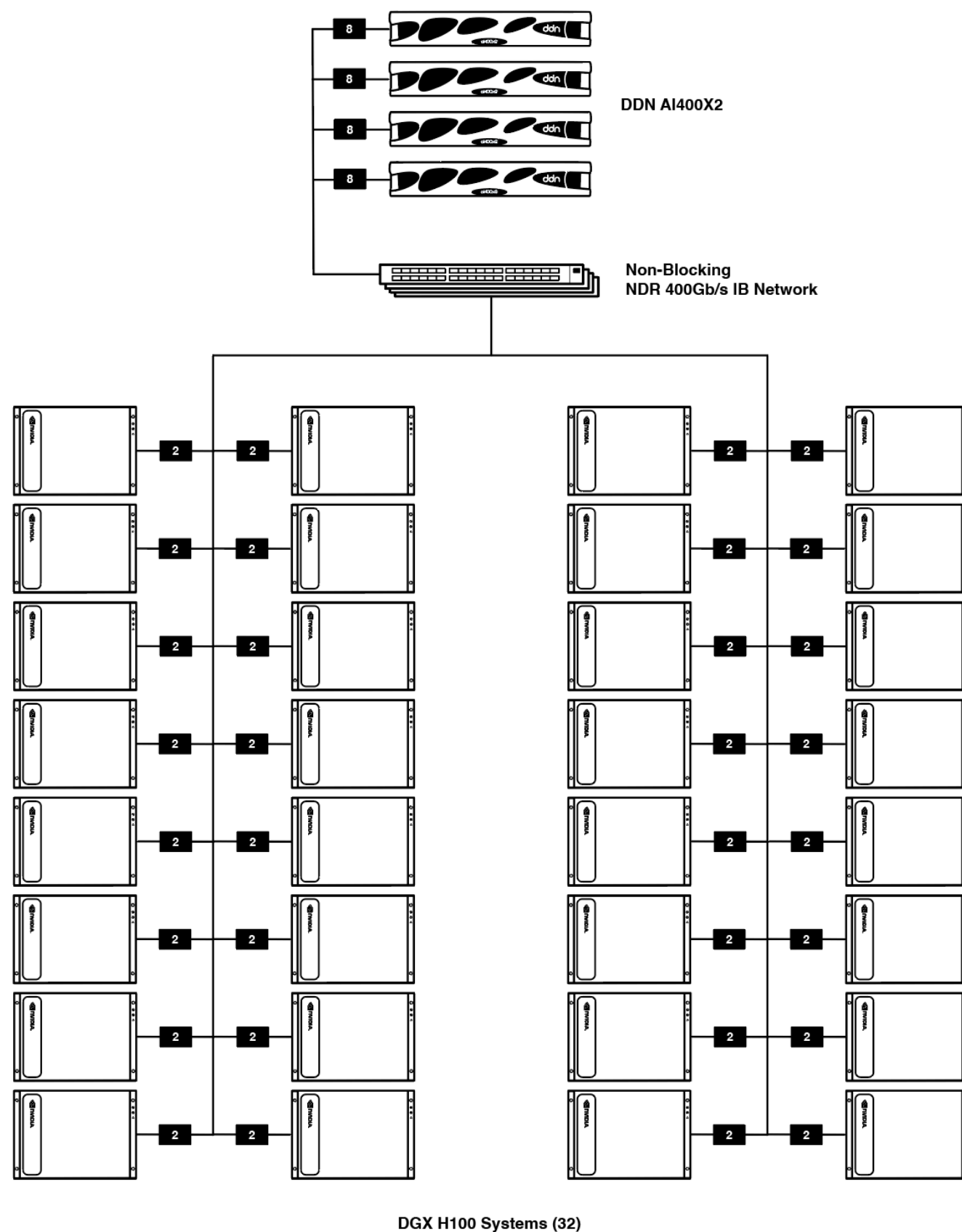

Figure 20. Test environment with 32 DGX H100 systems and 4 AI400X2 appliances.

This test demonstrates the peak performance of the DDN SuperPOD reference architecture using the fio open-source synthetic benchmark tool. The tool is set to simulate a general-purpose workload without any performance-enhancing optimizations. Separate tests were run to measure both 100% read and 100% write workload scenarios.

Below are the specific FIO configuration parameters used for these tests:

- blocksize = 1024k
- direct = 1
- iodepth= 128

- ioengine = posixaio
- bw-threads = 255

In Figure 21, test results demonstrate that the DDN solution can deliver over 99 GB/s of read throughput and over 90 GB/s of write throughput to a single DGX H100 system. In this test, data is accessed through a single posix mount provided by the DDN shared parallel filesystem client installed on the DGX H100 system. The test demonstrates that the DGX H100 system can utilize the full network capabilities on the DGX SuperPOD storage network.
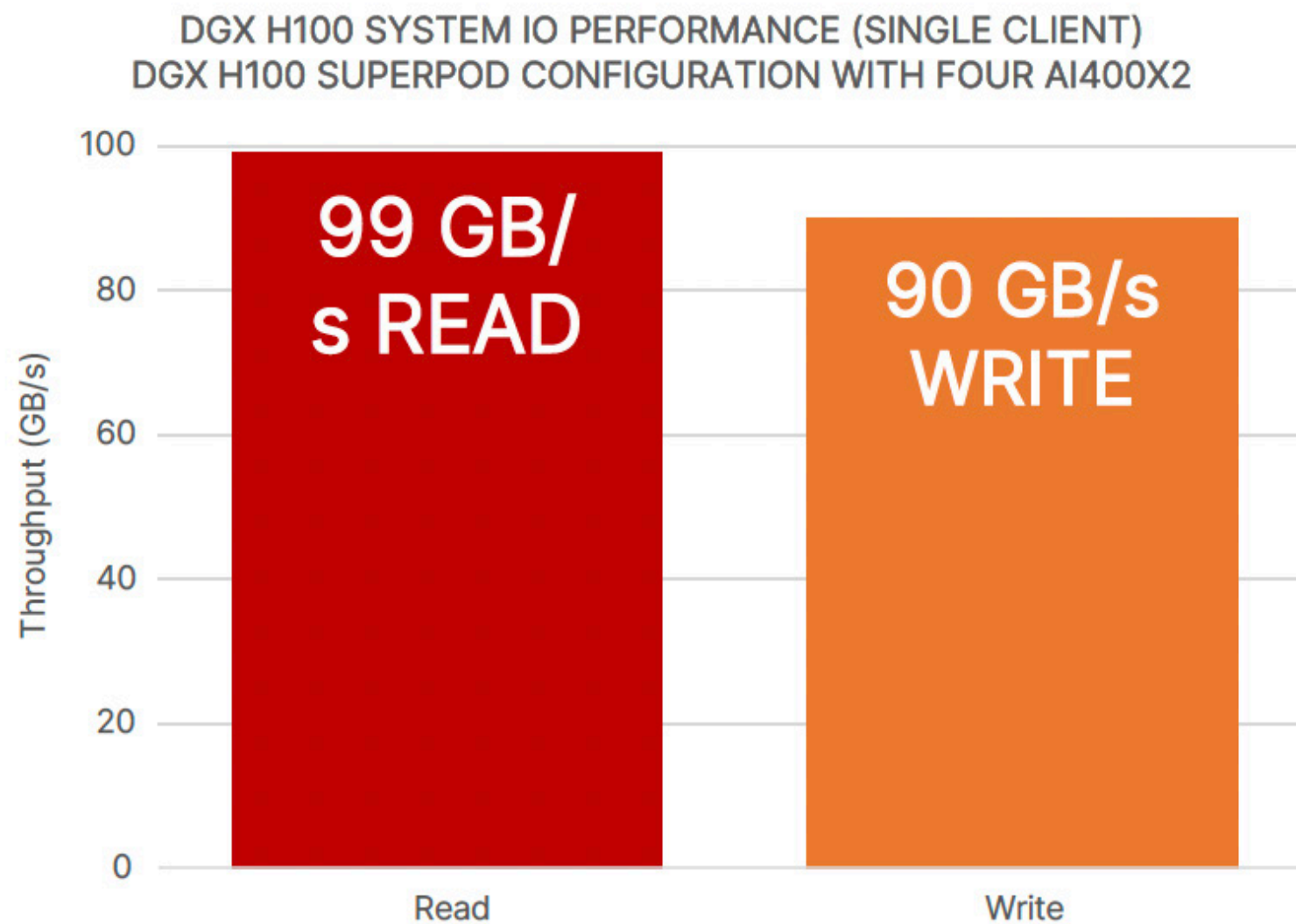

Figure 21. FIO throughput with DGX SuperPOD configuration using a single DGX H100 system.

Figure 22 demonstrates that the DDN software evenly distributes the full read and write performance of the AI400X2 appliances with all thirty-two DGX H100 systems engaged simultaneously. The DDN solution utilizes both network links on every DGX H100 system, ensuring optimal performance for a very wide range of data access patterns and data types.

This test demonstrates that the storage solution is provisioned to deliver at least 1.4 GB/s of read and 1 GB/s of write throughput for every GPU in the SuperPOD.
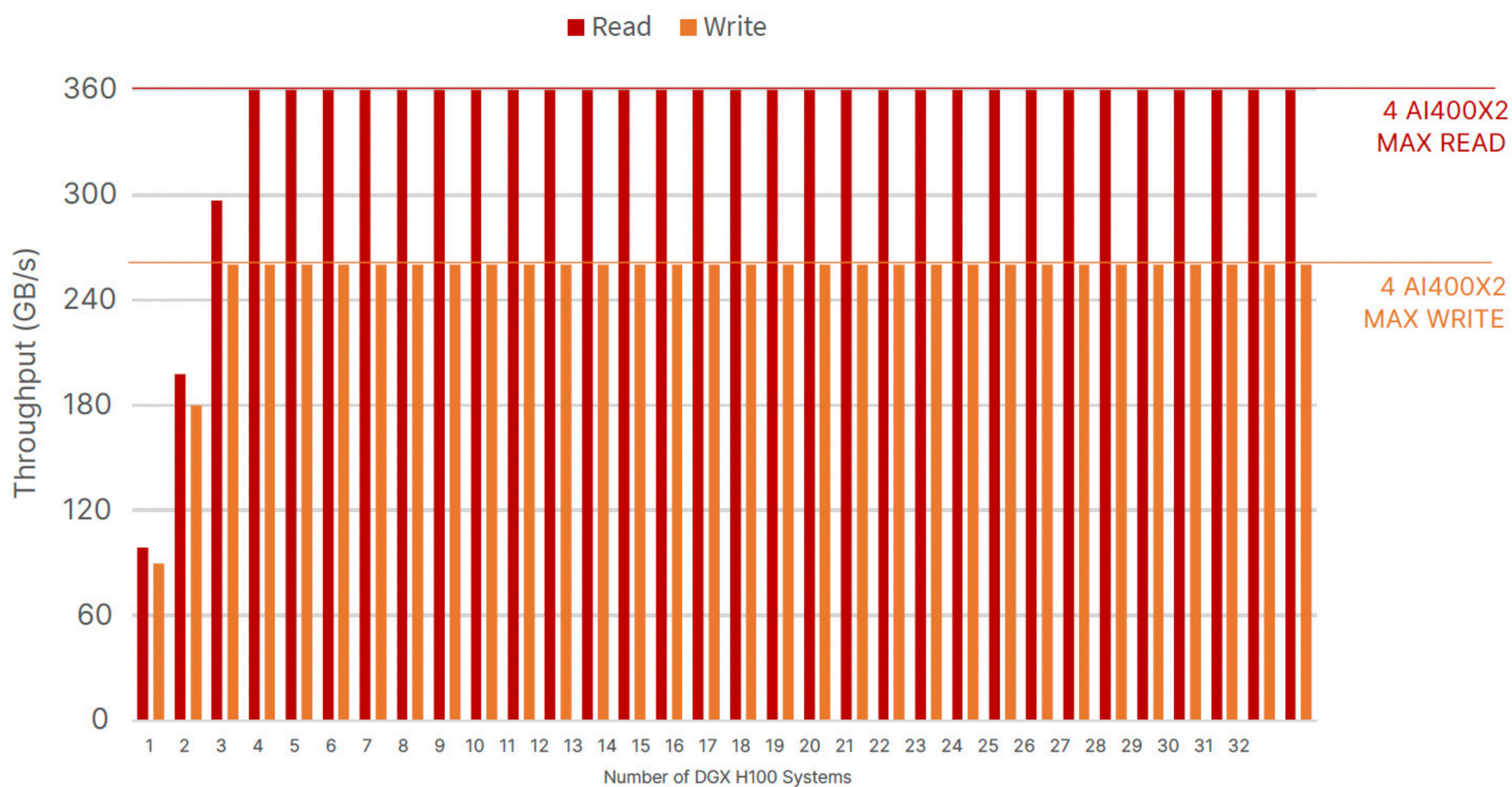

Figure 22. FIO throughput with DGX SuperPOD configuration using 32 DGX H100 systems.

6. CONTACT DDN TO UNLEASH THE POWER OF YOUR DGX SUPERPOD

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.