DDN A³I® SOLUTIONS WITH SUPERMICRO X12 GAUDI AI SERVERS

Fully-integrated and optimized infrastructure solutions for accelerated at-scale AI with Habana Gaudi AI processors

EXECUTIVE SUMMARY

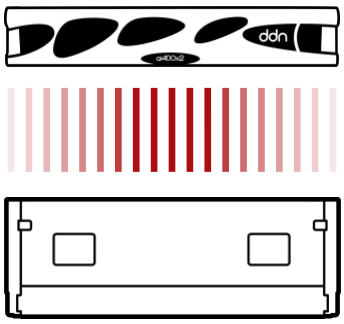
DDN A³I Solutions are proven at-scale to deliver optimal data performance for Artificial Intelligence (AI) training applications running on Habana Gaudi AI Processors. This document describes fully validated reference architectures for scalable configurations. The solutions integrate DDN AI400X2 appliances with Supermicro X12 Gaudi AI servers and Arista Ethernet network switches.

1. DDN A³I End-To-End Enablement for Supermicro X12 Gaudi AI Servers

DDN A³I solutions (Accelerated, Any-Scale AI) are architected to achieve the most from at-scale AI training applications running on Habana Gaudi AI processors. They provide predictable performance, capacity, and capability through a tight integration between DDN AI400X2 appliances and Supermicro X12 Gaudi AI servers. 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 Supermicro and Habana Labs, an Intel company. The deep integration of DDN AI appliances with Supermicro X12 Gaudi AI servers ensures a reliable experience. DDN A³I solutions are highly configurable 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 AI training system all the way to the largest AI infrastructure in operation today.

DDN brings the same advanced technologies used to power the world's largest supercomputers in a fully integrated package for X12 Gaudi AI servers that's easy to deploy and manage. DDN A³I solutions are proven to maximum benefits for at-scale AI. workloads on Habana Gaudi AI processors. This section describes the advanced features of DDN A³I Solutions for Supermicro X12 Gaudi AI servers.



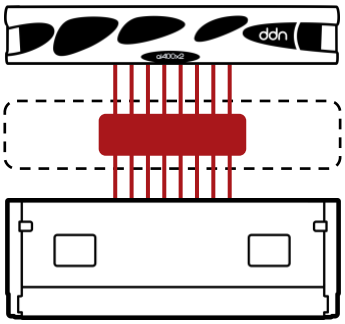
DDN A³I Shared Parallel Architecture

The DDN A³I shared parallel architecture and client protocol ensures high levels of performance, scalability, security, and reliability for X12 Gaudi AI servers. Multiple parallel data paths extend from the drives all the way to containerized applications running on the Habana Gaudi processors in the X12 Gaudi AI server. 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 X12 Gaudi AI servers with all Habana Gaudi AI processor 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 becomes unavailable.

DDN A³I Streamlined Deep Learning

DDN A³I solutions enable and accelerate end-to-end data pipelines for deep learning (DL) workflows of all scale running on X12 Gaudi AI servers. The DDN shared parallel architecture enables concurrent and continuous execution of all phases of DL workflows across multiple X12 Gaudi AI servers. 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.

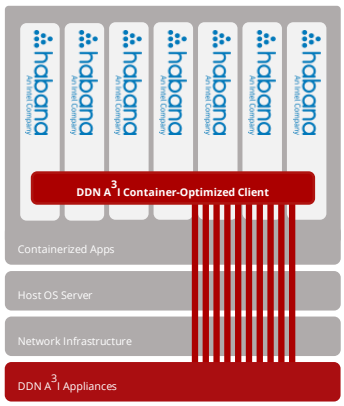
Significant acceleration can be achieved by executing an application across multiple X12 Gaudi AI servers simultaneously and engaging parallel training efforts of candidate neural networks variants. These advanced optimizations maximize the potential of DL frameworks. DDN works closely with Supermicro, Habana and its customers to develop solutions and technologies that allow widely-used DL frameworks to run reliably on X12 Gaudi AI servers.



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 Ethernet provides both high-bandwidth and low-latency data transfers between applications, compute servers and storage appliances. For Habana Gaudi AI processor solutions, DDN recommends an Ethernet network using Arista network switches. DDN A³I Multirail greatly simplifies and optimizes X12 Gaudi AI server networking for fast, secure, and resilient connectivity.

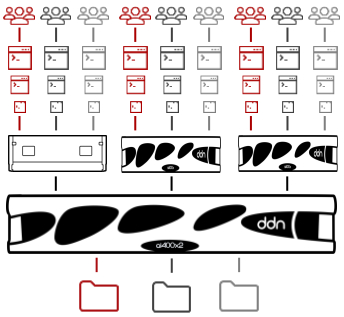
DDN A³I Multirail enables grouping of multiple network interfaces on a X12 Gaudi AI server to achieve faster aggregate data transfer capabilities. The feature balances traffic dynamically across all the interfaces, and actively monitors link health for rapid failure detection and automatic recovery. 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 X12 Gaudi AI server deployments.



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 X12 Gaudi AI server 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.

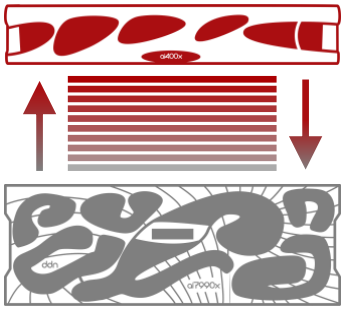
Containerized versions of popular DL frameworks—TensorFlow and PyTorch--specially optimized for the Habana Gaudi AI processor are available from Habana. They provide a solid foundation that enables data scientists to rapidly develop and deploy applications on the X12 Gaudi AI server. The DDN A³I container client provides high-performance parallelized data access directly from containerized applications on the X12 Gaudi AI server. 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

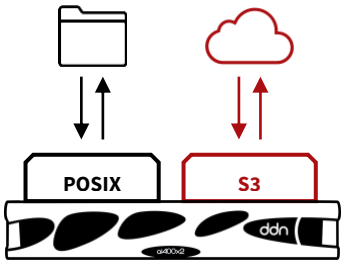
Container clients provide a simple and very solid mechanism to enforce data segregation by restricting data access within a container. DDN A³I makes it very simple to operate a secure multitenant environment at-scale through its native container client and comprehensive digital security framework. DDN A³I multitenancy makes it simple to share X12 Gaudi AI servers across a large pool of users and still maintain secure data segregation. Multi-tenancy provides quick, seamless, dynamic X12 Gaudi AI server 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 X12 Gaudi AI servers 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 applications on a X12 Gaudi AI server 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 Supermicro X12 Gaudi AI Servers

The DDN A³I client's NUMA-aware capabilities enable strong optimization for X12 Gaudi AI servers. It automatically pins threads to ensure I/O activity across the X12 Gaudi AI server 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.

2. DDN A³I Solutions with Supermicro X12 Gaudi AI Servers

The DDN A³I scalable architecture integrates X12 Gaudi AI servers with DDN AI shared parallel file storage appliances and delivers fully optimized end-to-end AI acceleration on Habana Gaudi AI processors. DDN A³I solutions greatly simplify the deployment of X12 Gaudi AI servers in single server and multi-server configurations, while also delivering performance and efficiency for maximum Habana Gaudi AI processors saturation, and high levels of scalability.

This section describes the components integrated in DDN A³I Solutions with Supermicro X12 Gaudi AI servers.

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 throughput and 3M IOPS directly to X12 Gaudi AI servers. Shared performance scales linearly as additional AI400X2 appliances are integrated to the AI infrastructure. The all-NVMe configuration provides optimal performance for a wide variety of workload and data types and ensures that X12 Gaudi AI server operators can achieve the most from at-scale AI 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 X12 Gaudi AI server operators to go beyond basic infrastructure and implement complete data governance pipelines at-scale.

The AI400X2 appliance integrates with X12 Gaudi AI servers over Ethernet and RoCE. It is available in 250 TB 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 Supermicro X12 Gaudi AI Server

The Supermicro X12 Gaudi AI server (SYS-420GH-TNGR), powered by Habana Gaudi Deep Learning Processors, pushes the boundaries of deep learning training and can scale up to hundreds of Gaudi processors in one AI cluster. Supermicro has partnered with Habana Labs and Intel to deliver a high performance and cost-effective server for AI training, coupled with the 3rd Gen Intel Xeon Scalable Processor. The server enables high-performance training at reasonable prices and makes AI training more accessible to a wide range of industries. The Supermicro X12 Gaudi AI server provides superior scalability and has been substantiated by demanding AI customers.

Figure 2. The Supermicro X12 Gaudi AI server

Gaudi is the first DL training processor to integrate ten ports of 100 GbE integrated RDMA over Converged Ethernet (RoCE v2) engines on-chip. With bi-directional throughput of up to 2 Tb/s, these engines play a critical role in the inter-processor communication needed during the training process. This native integration of RoCE allows customers to use the same scaling technology, both inside the server and rack (scale-up) and across racks (scale-out). These can be connected directly between Gaudi processors or through any number of standard Ethernet switches.

The 420GH-TNGR system contains eight Gaudi HL-205 mezzanine cards, dual 3rd Gen Intel® Xeon® Scalable processors, two PCIe Gen 4 switches, four hot swappable NVMe/SATA hybrid hard drives, fully redundant power supplies, and 24 x 100GbE RoCE ports (6 QSFP-DDs) for unprecedented scale-out system bandwidth. This system contains up to 8TB of DDR4-3200MHz memory, unlocking the Gaudi processors' full potential. The HL-205 is OCP-OAM (Open Compute Project Accelerator Module) specification compliant. Each card incorporates the Gaudi HL-2000 processors with 32GB HBM2 memory and ten natively integrated ports of 100GbE RoCE v2.

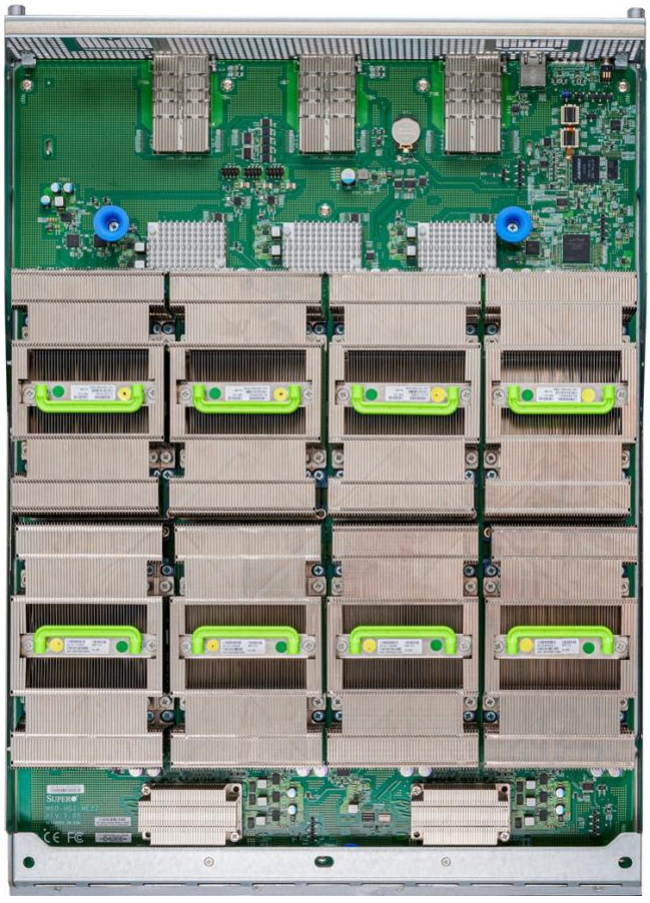


Figure 3. Eight Gaudi HL-205 mezzanine cards in the X12 Gaudi AI server

Each of the Gaudi processors dedicates seven of its ten 100GbE RoCE ports to all-to-all connectivity within the system, with three ports per Gaudi available for scaling out externally for a total of 24 x100GbE RoCE ports per 8-card system. This allows end customers to scale their deployment using standard 100GbE switches. The high throughput of RoCE bandwidth inside and outside the box and the unified standard protocol used for scale-out make the solution easily scalable and cost-effective. Figure 4 shows the Gaudi HL-205 processors and the communication paths between processors and the server CPUs.

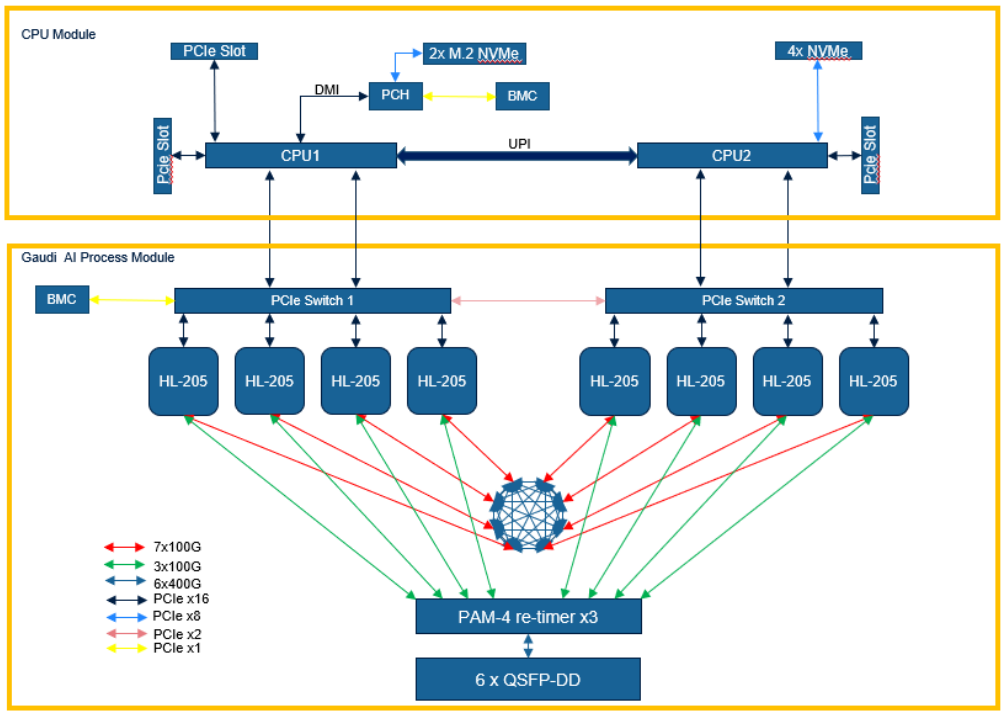


Figure 4. High-performance interconnects within the X12 Gaudi AI server

The X12 Gaudi AI server provides a turnkey computing platform and delivers superior performance for AI workloads at all scale.

2.3 Arista Network Switches

Arista network switches provide optimal interconnect for Habana Gaudi AI processors. DDN recommends Ethernet technology for data-intensive compute and storage networks with X12 Gaudi AI servers.

Gaudi Network Switches

The Arista DCS-7060DX4-32 400 Gb/s Ethernet Switch provides 32 ports of connectivity in a 1U form factor. Every switch is capable of up to 25.6 Tb/s of non-blocking bandwidth with sub 700ns latency. The DCS-7060DX4-32 is the ideal modular network unit to architect scalable solutions with X12 Gaudi AI servers.

Figure 5. Arista DCS-7060DX4-32-F 400 Gbps Ethernet Switch

Storage and Cluster Management Network Switches

The Arista 7170-32C 100 Gb/s Ethernet Switch provides 32 ports of connectivity in a 1U form factor. Every switch is capable of up to 6.4 Tb/s of non-blocking bandwidth with sub 800ns latency. The 7170-32C is the ideal modular network unit to architect scalable, high-performance network for data transfer between storage and compute nodes.

Figure 6. Arista 7170-32C 100 Gbps Ethernet Switch

Management Network Switches

The Arista 7010T Ethernet Switch is a Gigabit Ethernet Layer 3 switch family featuring 54 ports with 48 10/100/1000BASE-T ports, 4 x 10G SFP+ uplink ports. It provides robust capabilities for critical low-intensity traffic like component management.

Figure 7. Arista 7010T 1 Gbps Ethernet Switch

2.4 Gaudi Platform from Habana Labs, an Intel company

Gaudi processors from Habana Labs, an Intel company, are custom-designed and purpose-built for training models for deep learning and machine learning workloads in the data center and in the cloud.

According to the IDC 2020 Semiannual Artificial Intelligence Tracker, there's an explosive demand for AI training driven by:

- Dramatically increasing number of AI applications
- More complex AI data sets and models
- Increasing number of iterations being run on a single model to improve accuracy and currency:
    - 74% of IDC respondents indicate running 5 to 10 iterations to train a model
    - 50% of respondents rebuild models weekly or more frequently
    - 26% report rebuilding models daily or even hourly

In addition, 56% of prospective AI/ML customers surveyed in the IDC study cited the high cost of training as the most significant challenge to implementing AI/ML solutions.

Gaudi processors were designed to drive improved efficiency (better price performance) in the AI data center and cloud, thus enabling end-customers to train more and spend less. In addition, the solutions have been designed for usability, making it easy for customers to build new and migrate existing AI models to Gaudi-based models.

Training on Gaudi AI processors provides:

1. Efficiency – Gaudi's implementation in the Supermicro X12 AI Training Server offers up to 40% better price performance than existing traditional AI compute solutions. Its cost efficiency enables you to train models on larger datasets and with greater frequency at lower cost.

2. Scalability – Every Gaudi AI processor integrates ten 100-Gigabit RDMA over Converged Ethernet (RoCE) ports for flexible and massive scale-up and scale-out capacity. Based on industry standard Ethernet, Gaudi systems help its customers avoid lock-in with vendors who offer only proprietary connectivity and enable use of a wide-array of standard Ethernet switch solutions to build out systems, thus lowering overall system costs.

3. Usability – Gaudi is supported with the Habana SynapseAI® software stack and tools that simplify building new models or migrating existing models to the Gaudi platform. This means that data center operations can reduce the cost of model training without expending significant time, resource, or effort to deploy Gaudi systems.

Gaudi efficiency is based on the processor's architecture--both hardware and software.

Gaudi's hardware architecture features:

- Heterogeneous compute architecture to maximize training efficiency
    - Eight fully programmable, AI-customized Tensor Processor Cores
    - Configurable centralized GEMM engine (matrix multiplication engine)
- Software managed memory architecture with 32 GB of HBM2 memory
- Native integration of 10 x 100 GbE RoCE for flexible scaling

Designed for Scaling Efficiency

The industry's ONLY native integration of 10 x 100 Gigabit RoCE ports onto every Gaudi

- Eliminates network bottlenecks
- Standard Ethernet inside the server and across nodes can scale from one to thousands of Gaudis
- Eliminates lock-in with proprietary interfaces
- Lowers total system cost and power by reducing discrete components

Please see section 2.2 for more detail on how the Supermicro X12 server leverages the ten RoCE ports per Gaudi to eliminate networking bottlenecks and maximize scale-up and scale-out of systems.

Gaudi Usability with SynapseAI™ Software Stack and Development Content and Tools

The SynapseAI® software stack is optimized for the Gaudi hardware architecture and designed for ease of use. It was created with the needs of developers and data scientists in mind, providing versatility and ease of programming to address end-users' unique needs, while allowing for simple and seamless building of new models and porting of existing models to Gaudi. SynapseAI software facilitates developers' ability to customize Gaudi systems, enabling them to address their specific requirements and create their own custom innovations.

The SynapseAI software platform features:

1. Integrated TensorFlow and PyTorch frameworks
2. Support for popular computer vision, NLP and recommendation models
3. TPC programming tools: compiler, debugger and simulator
4. Extensive Habana kernel library and library for Customer Kernel development Habana Communication Libraries (HCL and HCCL)

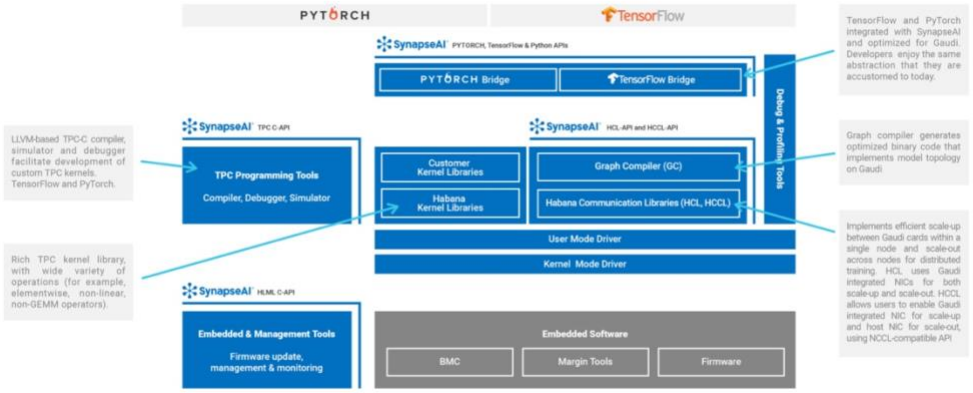


Figure 8. SynapseAI Software Platform Diagram

3. DDN A³I Reference Architectures for Supermicro X12 Gaudi AI Servers

DDN proposes the following reference architectures for configurations with single and multiple X12 Gaudi AI servers. DDN A³I solutions are fully validated with Intel and Supermicro, and already deployed with at-scale AI customers worldwide.

The DDN AI400X2 appliance is a turnkey appliance for at-scale X12 Gaudi AI server deployments. DDN recommends the AI400X2 appliance as the optimal data platform for single and multiple server configurations. The AI400X2 appliances delivers optimal performance for every workload and data type in a dense, power efficient 2RU chassis. The AI400X2 appliance simplifies the design, deployment, and management of X12 Gaudi AI servers and provides predictable performance, capacity, and scaling. The AI400X2 appliance arrives fully configured, ready to deploy and installs rapidly. The appliance is designed for seamless integration with X12 Gaudi AI servers 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.

As general guidance, DDN recommends an AI400X2 appliance for every four X12 Gaudi AI servers (Figure 9). These configurations can be adjusted and scaled easily to match specific workload requirements. For the storage & cluster management network, DDN recommends Ethernet technology in a non-blocking network topology, with dual paths to ensure data availability. DDN recommends use of at least two 100 GbE connections per X12 Gaudi AI server to the storage & cluster management network.

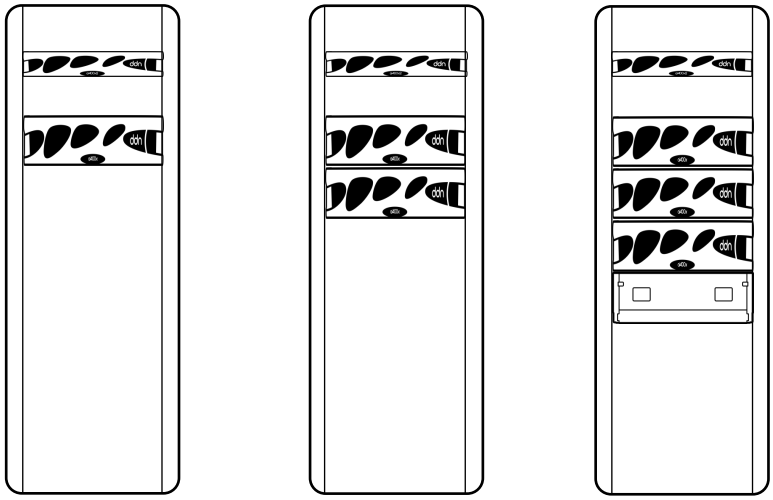


Figure 9. Rack illustrations for DDN A³I reference configurations with X12 Gaudi AI servers (network switches not shown).

3.1 Network Architecture Overview

The reference design includes three networks:

Storage and cluster management network. Provides data transfer between the AI400X2 appliance and the compute nodes. Connects eight ports from each AI400X2 appliance. Connects two ports from each X12 Gaudi AI server, one each from two dual-port 100 GbE interfaces. DDN recommends this for optimal performance and efficiency. This network is also used for inter-host workload communication and external communications from the hosts.

Gaudi network. Provides inter-Gaudi communication. Connects the six 400 GbE interfaces on each of the X12 Gaudi AI servers. Each Gaudi training accelerator includes 10 x 100GbE RoCE on-chip engines. The Gaudi network inter-connects the Gaudi 100Gb Ethernet ports directly for Gaudi to Gaudi communication (HCL). This communication avoids the host PCIe bus and CPU complex.

Management Network. Provides management and monitoring for all components. Connects the management port from each X12 Gaudi AI server and each AI400X2 appliance controller to an Ethernet switch.

An overview of the network architecture is shown in Figure 10, recommended network connections for each X12 Gaudi AI server on Figure 11, and recommended network connections for each DDN AI400X2 appliance on Figure 12.

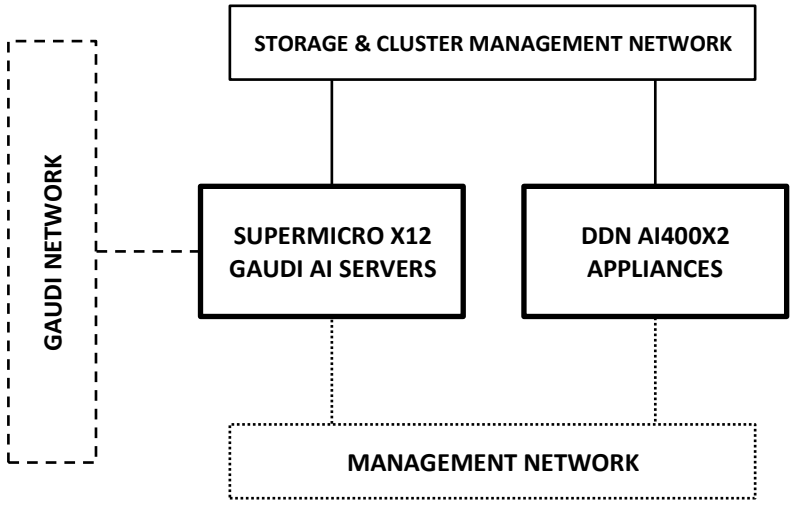


Figure 10. Overview of the network architecture

2.2. Supermicro X12 Gaudi AI Server Network Connectivity

DDN recommends ports 1 to 6 on the X12 Gaudi AI servers be connected to the Gaudi network. DDN recommends that all X12 Gaudi AI servers be equipped with two dual ported 100 GbE cards. Ports 7 and 9 should be connected to the storage & cluster management network. As well, the BMC (“B”) port should be connected to the management network.

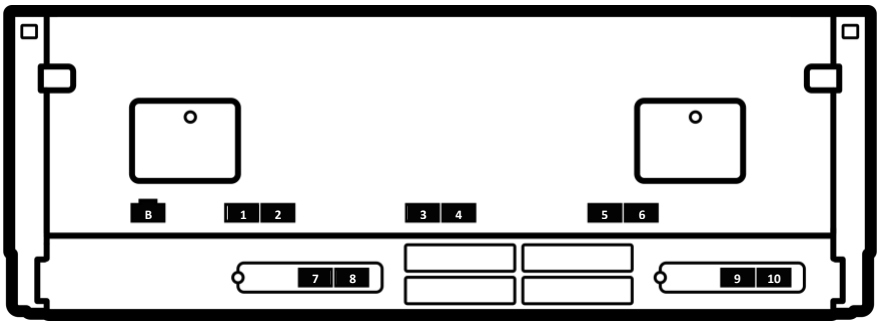


Figure 11. Recommended Supermicro X12 Gaudi AI server network port connections

AI400X2 Appliance Network Connectivity

DDN recommends ports 1 to 8 on the AI400X2 appliance be connected to the storage & cluster management network. As well, the management (“M”) and BMC (“B”) ports for both controllers should be connected to the management network. Note that each AI400X2 appliance requires one inter-controller network port connections (“I”) using the short ethernet cable supplied.

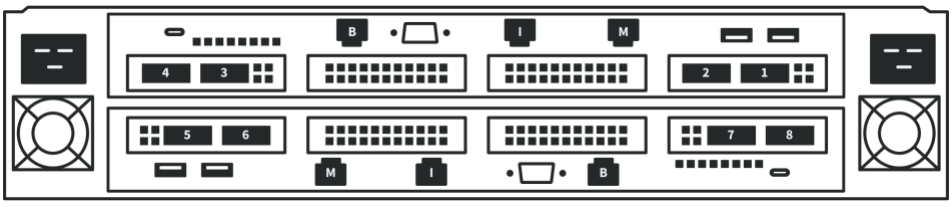
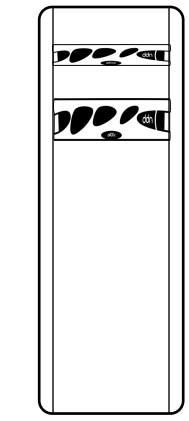


Figure 12. Recommended AI400X2 appliance network port connections



3.2 Architecture with One Supermicro X12 Gaudi AI Server

Figure 13 illustrates the DDN A^{3}I architecture single server configuration. An X12 Gaudi AI server is connected to an AI400X2 appliance through a network switch. The X12 Gaudi AI server connects to the storage & cluster management network switch via two 100 GbE links. The AI400X2 appliance connects to the storage & cluster management network switches via eight 100 GbE links. This ensures non-blocking data communication between every device connected to the network. The multi-path design provides full-redundancy and maximum data availability in case a link becomes unavailable.

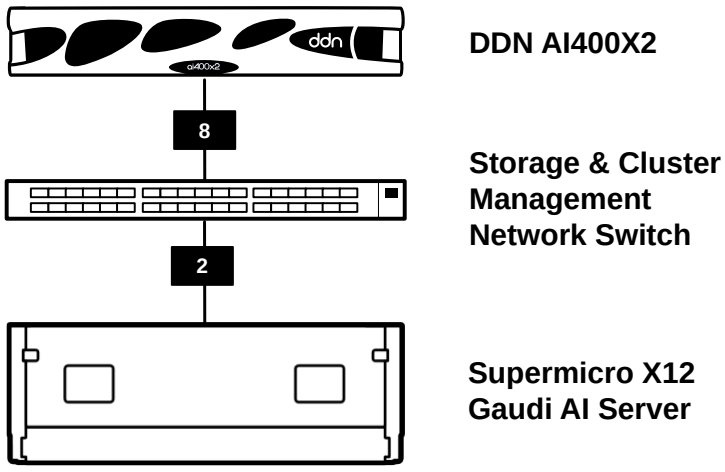


Figure 13. DDN A^{3}I reference architecture with one X12 Gaudi AI server – high-performance networks

Figure 14 illustrates the management network configuration for a single node configuration. All components in the solution are connected to a management network switch with 1 GbE links.

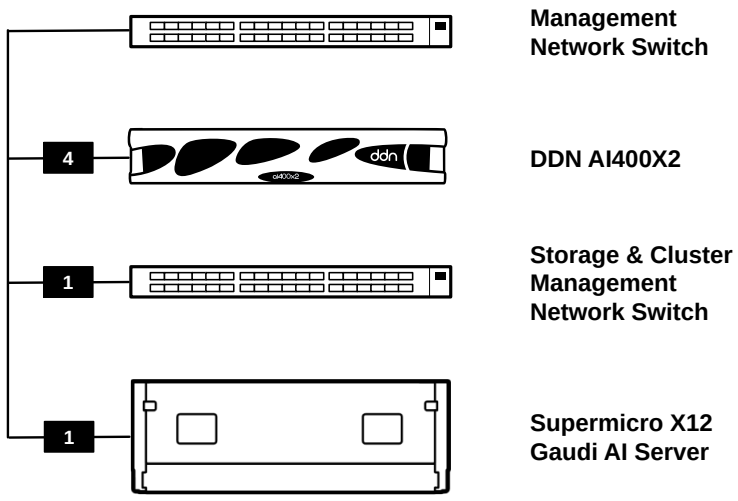
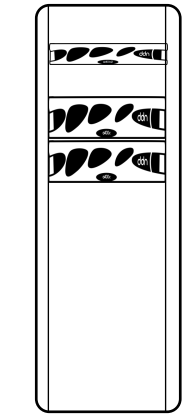


Figure 14. DDN A³I reference architecture with one X12 Gaudi AI server – management network



3.3 Architecture with Two Supermicro X12 Gaudi AI Server

Figure 15 illustrates the DDN A³I architecture dual server configuration. Two X12 Gaudi AI servers are connected to an AI400X2 appliance through a network switch. Every X12 Gaudi AI server connects to the storage & cluster management network switch via two 100 GbE links. The AI400X2 appliance connects to the storage & cluster management network switches via eight 100 GbE links. This ensures non-blocking data communication between every device connected to the network. The multi-path design provides full-redundancy and maximum data availability in case a link becomes unavailable.

Additionally, the X12 Gaudi AI servers are connected through a network switch for Gaudi communication. Every X12 Gaudi AI server connects to the Gaudi network switch via six 400 GbE links.

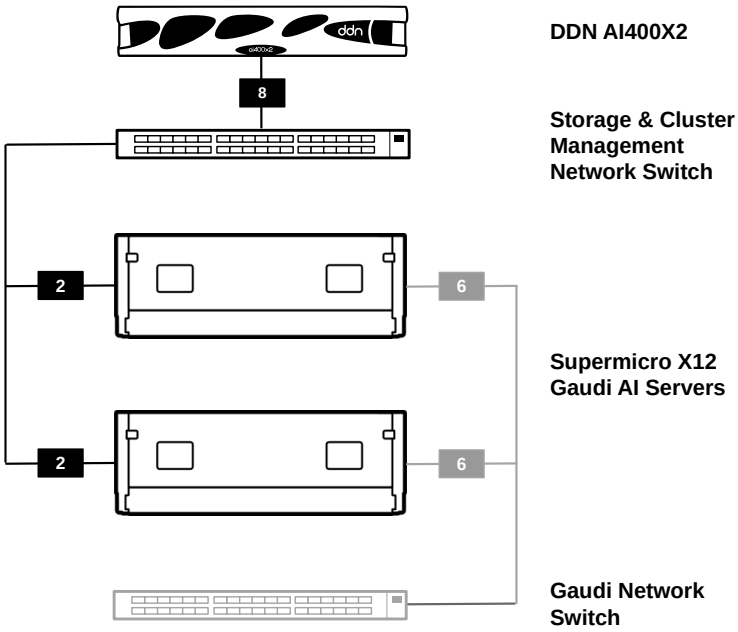


Figure 15. DDN A³I reference architecture with two X12 Gaudi AI servers – high-performance networks

Figure 16 illustrates the management network configuration for a two node configuration. All components in the solution are connected to a management network switch with 1 GbE links.

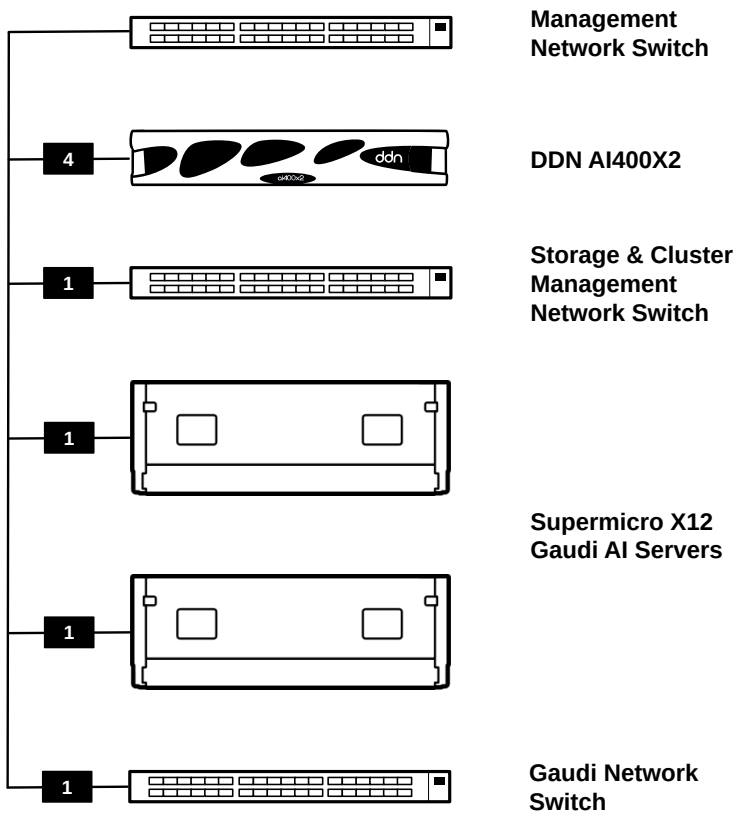
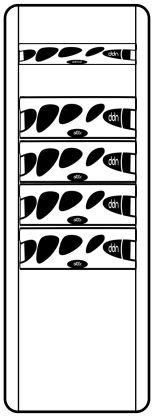


Figure 16. DDN A³I reference architecture with two X12 Gaudi AI servers – management network



3.4 Architecture with Four Supermicro X12 Gaudi AI Servers

Figure 17 illustrates the DDN A³I architecture quad server configuration. Four X12 Gaudi AI servers are connected to an AI400X2 appliance through a network switch. Every X12 Gaudi AI server connects to the storage & cluster management network switch via two 100 GbE links. The AI400X2 appliance connects to the storage & cluster management network switches via eight 100 GbE links. This ensures non-blocking data communication between every device connected to the network. The multi-path design provides full-redundancy and maximum data availability in case a link becomes unavailable.

Additionally, the X12 Gaudi AI servers are connected through a network switch for Gaudi communication. Every X12 Gaudi AI server connects to the Gaudi network switch via six 400 GbE links.

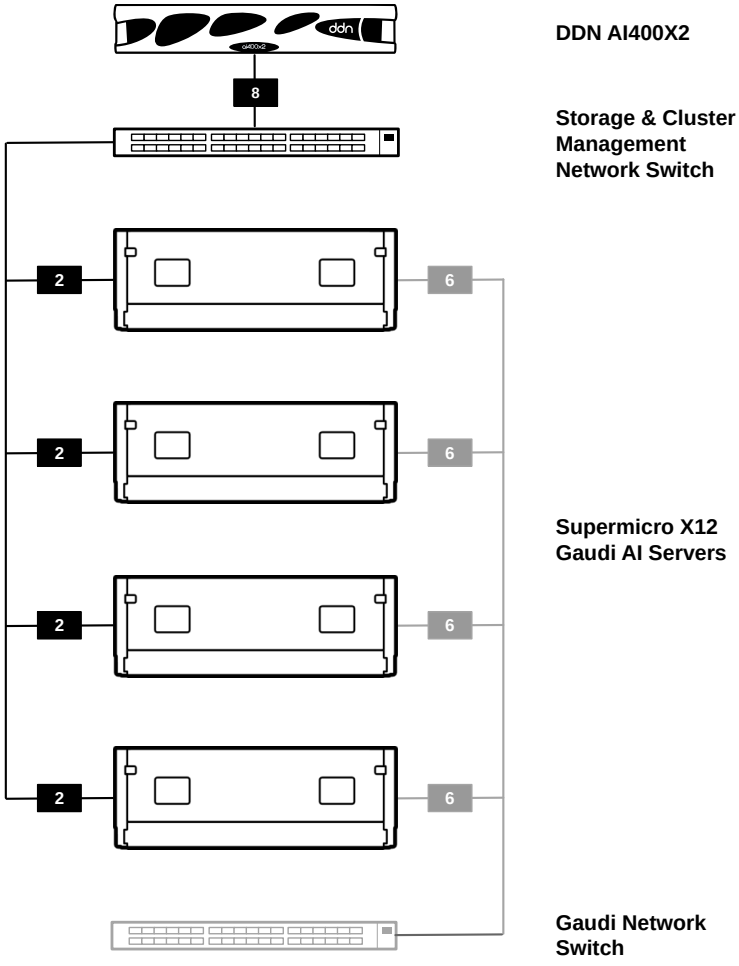


Figure 17. DDN A³I reference architecture with four X12 Gaudi AI servers – high-performance networks

Figure 18 illustrates the management network configuration for a quad node configuration. All components in the solution are connected to a management network switch with 1 GbE links.

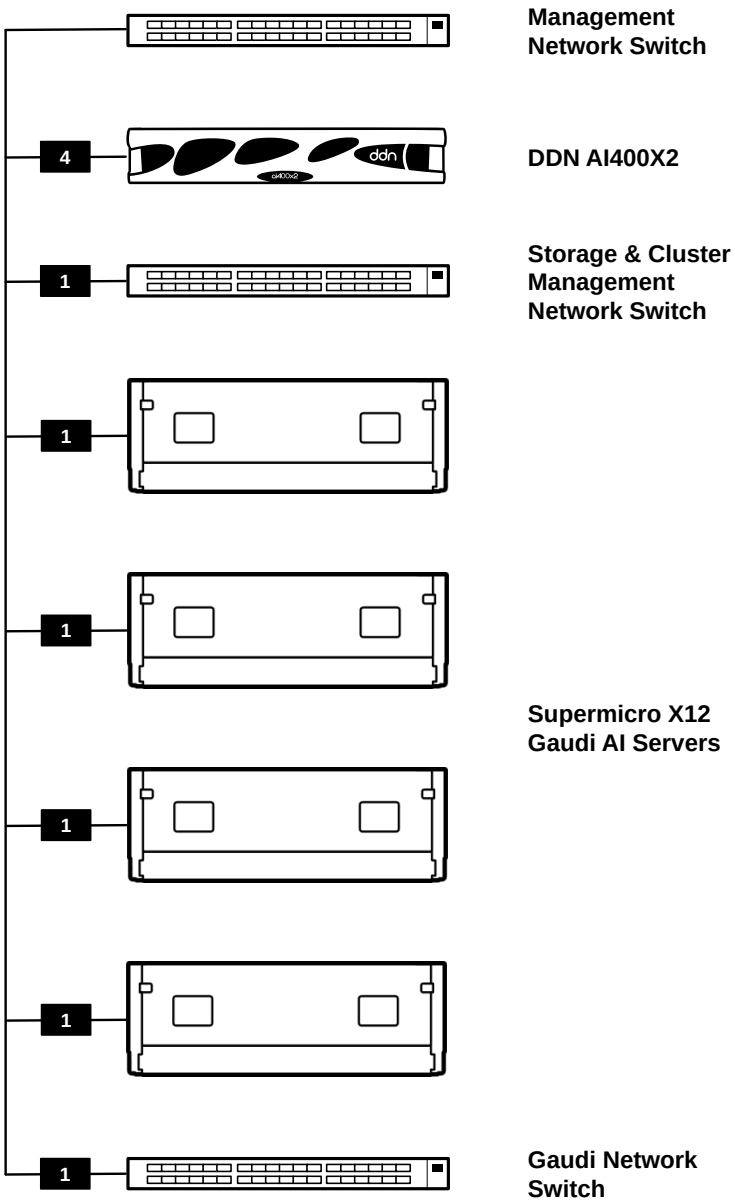


Figure 18. DDN A³I reference architecture with four X12 Gaudi AI servers – management network

3.5 DDN Insight Analytics Server

Optimizing scalable applications for AI and Analytics requires insight into the entire infrastructure stack to diagnose bottlenecks and areas with potential for optimization. DDN Insight provides comprehensive monitoring into your workloads from a single intuitive interface, with essential information on the status of storage systems and performance insights from disk to network to client application.

Typically, storage networks are made up of local block-based arrays, file systems, business analytics repositories, and distributed object stores, each with their own unique management requirements. DDN Insight is a single application that monitors and manages all types of storage resources. It is available as a software-installable package on customer-supplied management servers, designated as the DDN Insight head node within the storage network. Multiple head nodes can be used to provide redundancy in case of failure.

DDN Insight is ideally suited for integrated DDN A^{3}I Solutions with X12 Gaudi AI servers and simplifies management of your storage infrastructure as AI data continues to grow exponentially.

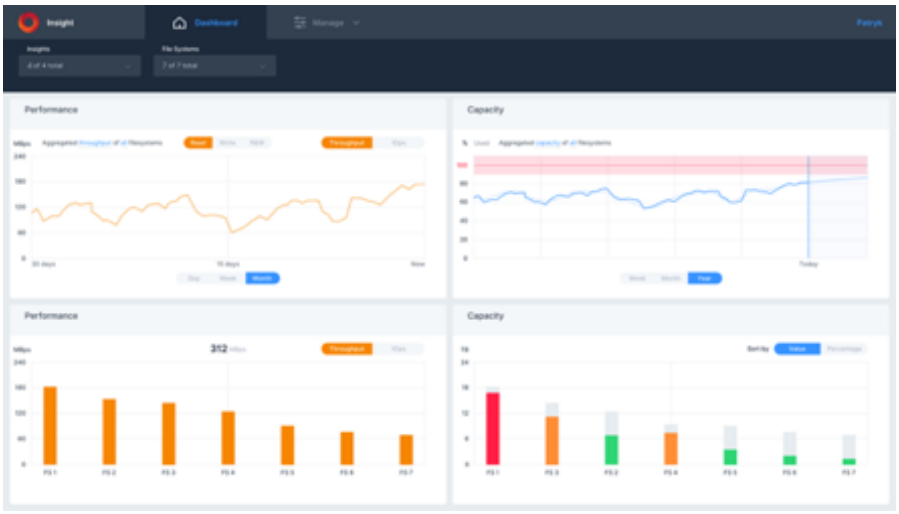


Figure 19. DDN Insight intuitive GUI for AI data solution monitoring and optimization

4. DDN A³I Solutions with X12 Gaudi AI Servers Validation

DDN conducts extensive engineering integration, optimization, and validation efforts in close collaboration with Supermicro and Intel Habana to ensure 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 configurations with X12 Gaudi AI servers.

Performance testing on the DDN A³I architecture has been 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, containerized applications can engage the full capabilities of the data infrastructure and the X12 Gaudi AI servers. Performance is distributed evenly across all the X12 Gaudi AI servers, and scales linearly as more X12 Gaudi AI servers are engaged.

This section details some of the results from recent at-scale testing integrating AI400X2 appliances with up to four X12 Gaudi AI servers.

The tests described in this section were executed at a Supermicro data center on X12 Gaudi AI servers equipped with eight Habana Gaudi AI processors. The two AI400X2 appliances are running DDN EXAScaler v6.0.0.

For the storage & cluster management network, all four X12 Gaudi AI servers are connected to an Arista 7170-32C switch with two 100 GbE links, one per dual-ported adapter (see recommendation in section 3.1). The AI400X2 is connected to the same network with eight 100 GbE links each. For the Gaudi network, all four X12 Gaudi AI servers are connected to an Arista DCS-7060DX4-32 switch. All six 400 GbE interfaces on the X12 Gaudi AI servers are connected to the Gaudi network.

This test environment allows us to demonstrate performance with the largest configuration.

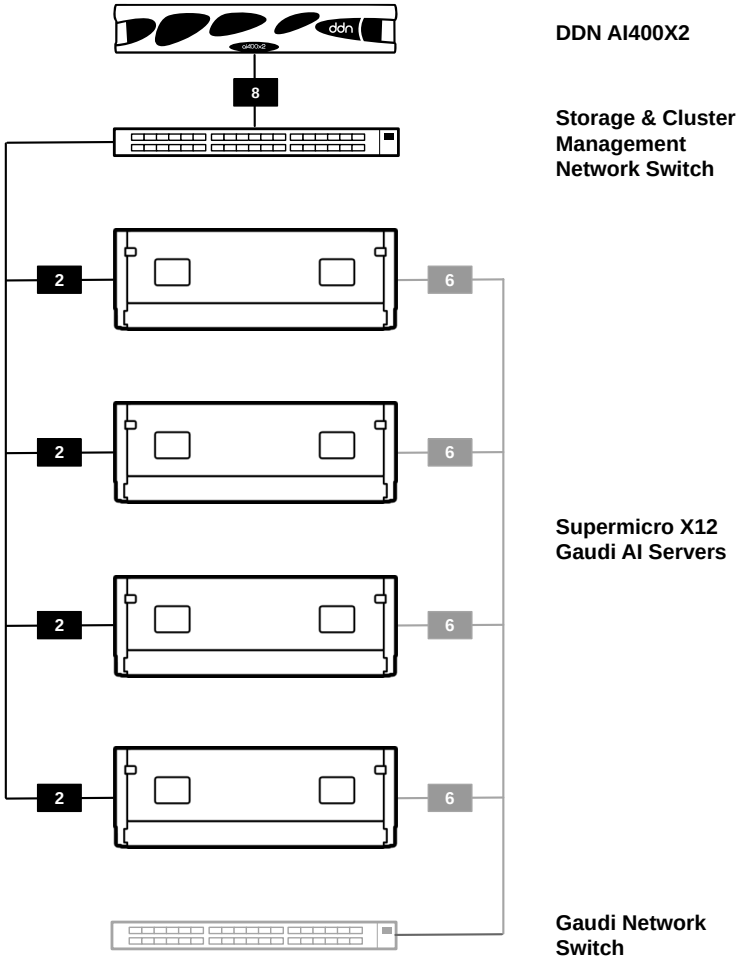


Figure 20. Test environment with four X12 Gaudi AI servers and an AI400X2 (management network not shown)

4.1 AI Infrastructure Performance Validation

This series of tests demonstrate the peak performance of the scalable 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.

The AI400X2 appliance provides predictable, scalable performance. This test demonstrates the architecture's ability to deliver full throughput performance to a small number of clients and distribute the full performance of the DDN solution evenly as a large number of X12 Gaudi AI servers are engaged.

In Figure 21, test results demonstrate that DDN solution can deliver over 25 GB/s of read and write throughput to a single X12 Gaudi AI server, and evenly distribute the full read and write performance of the AI400X2 appliance with up to four X12 Gaudi AI servers engaged simultaneously. The DDN solution can fully saturate both network links on every X12 Gaudi AI servers, ensuring optimal performance for a very wide range of data access patterns and data types for applications running on the systems.

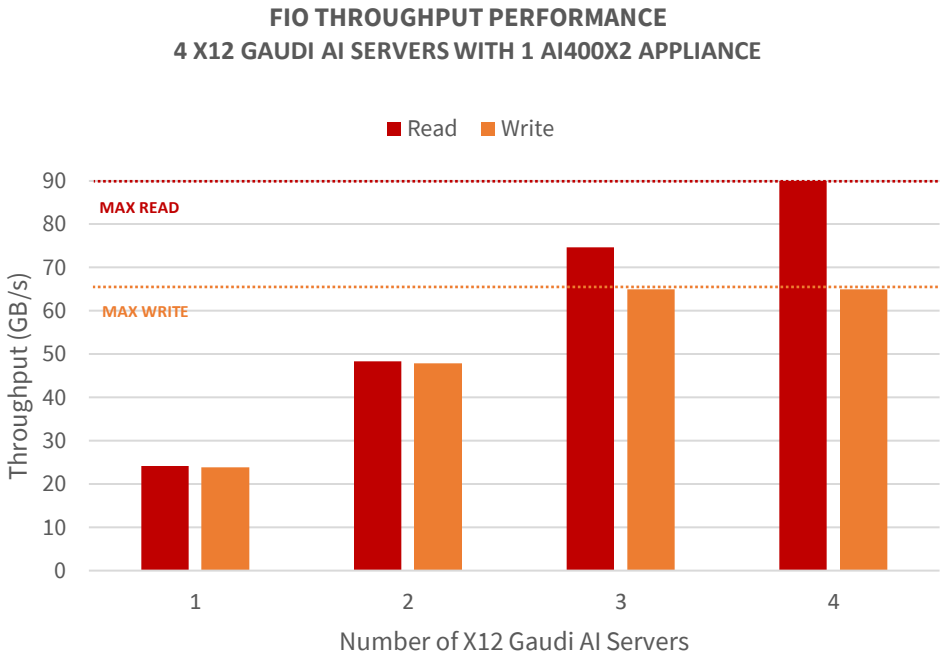


Figure 21. FIO throughput with scaled X12 Gaudi AI server configurations

5. Scaling DDN A³I Reference Architectures for X12 Gaudi AI Servers

The DDN A³I Reference Architectures for X12 Gaudi AI servers are designed to deliver an optimal balance of technical and economic benefits for a wide range of common use cases for AI. Using the AI400X2 appliance as a building block, solutions can scale linearly, predictably, and reliably in performance, capacity, and capability. For configurations with requirements beyond the base reference architecture, it's simple to scale the data platform with additional AI400X2 appliances.

The same AI400X2 appliance and shared parallel architecture used in the DDN A³I Reference Architectures for X12 Gaudi AI servers are also deployed with very large systems. The AI400X2 appliance has been validated to operate properly with up to 560 AI accelerator servers simultaneously.

In figure 24, we show an fio throughput test performed by engineers like the one presented in section 4.1. In this example, up to 32 X12 Gaudi AI servers are engaged simultaneously with 5 AI400X2 appliances. The results of the test demonstrate that the DDN shared parallel architecture scales linearly and fully achieves the capabilities of the five AI400X2 appliances, 450 GB/s throughput for read and 325 GB/s throughput for write, with multipleX12 Gaudi AI servers engaged. This performance is maintained and balanced evenly with up to 32 X12 Gaudi AI servers simultaneously.

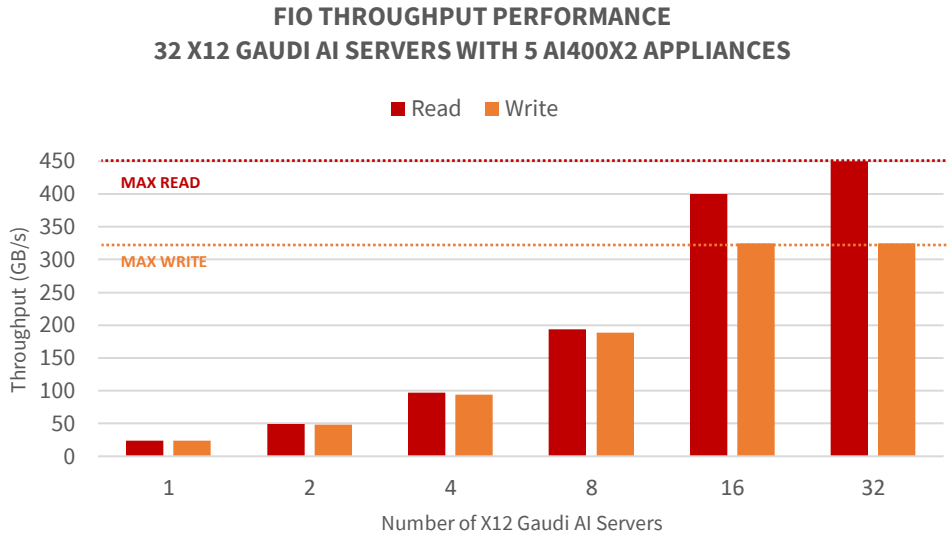


Figure 24. FIO throughput scaling with a very large number of X12 Gaudi AI servers

For more information on at-scale performance and validation with X12 Gaudi AI servers and AI400X2 appliances, contact DDN Sales.

6. Contact DDN to Unleash the Power of Your Habana Gaudi AI Processors

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 processing 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.