White Paper

Data Center
Artificial Intelligence

Accelerate Deep Learning Training with Habana® Gaudi® AI Processor and DDN AI Storage Solutions

Habana Labs, an Intel company, partners with Supermicro and DataDirect Networks (DDN) to provide end-to-end solutions for highly scalable deep learning training.

Artificial intelligence (AI) is becoming essential as demand for services such as speech and image recognition and natural language processing (NLP) continues to increase. But as the complexity of AI models increases, the time and expense of training these models also increases.

Habana Labs, an Intel company, partners with DataDirect Networks (DDN) and Supermicro to deliver integrated, turnkey deep learning (DL) solutions. These solutions enhance the performance of AI DL workloads with advanced data management and AI-specific storage. To help accelerate DL training workloads, Supermicro combines the capabilities of eight Habana Gaudi AI DL processors in the Supermicro X12 Gaudi AI Training System, a power-efficient server design that also features 3rd Generation Intel® Xeon® Scalable processors. Additionally, the DDN AI400X storage appliance provides capacity and performance that can help DL clusters scale up to hundreds of Supermicro servers.

As a turnkey solution available from and supported by Supermicro, this DL training solution is a reliable, high-performance alternative to general-purpose servers for AI training applications. This paper explores the Habana, Supermicro, and DDN components and how they are integrated. The paper also describes a performance validation that measured the throughput between the Supermicro X12 servers and the DDN AI400X storage appliances in various cluster configurations.

Supermicro simplifies purchasing, installation, and support

Designing, validating, and implementing any size AI training cluster can be challenging for IT teams who might not be familiar with DL training solutions. Supermicro provides all of the components—network, compute, and storage—as a turnkey solution that simplifies purchasing, installation, and support.

Supermicro works with organizations to design a solution that is appropriate for the organization's DL training workload requirements. Once designed, Supermicro assembles, configures, and validates all solution components. These components include the Supermicro X12 servers, DDN storage appliances, and network switches. Once validated, Supermicro then delivers the solution and installs it at the organization's site.

Supermicro also provides one-stop support for all components and software. If an organization requires assistance with any part of the solution, Supermicro provides one number to call, which helps simplify support.

Habana Gaudi processors help accelerate DL workloads

Built from the ground up to accelerate DL training workloads, the Habana Gaudi HL-2000 processor uses an AI purpose-built architecture that provides performance, scalability, power efficiency, and cost savings. When combined with the Habana® SynapseAI® software suite, this architecture also gives developers and data scientists familiar tools for building workloads.

Habana Gaudi processors are based on the fully programmable Tensor Processing Core (TPC) 2.0 architecture designed by Habana. Habana's TPCs accelerate matrix multiplication, which is crucial to AI training performance. In addition to the TPCs, each Gaudi processor incorporates several features on the silicon that help accelerate DL workloads:

- Eight clustered, programmable cores that incorporate static random-access memory (SRAM), which acts as local memory for each individual core
- Four high-bandwidth memory (HBM) devices that provide 32 GB of capacity and one terabyte-per-second of memory bandwidth
- A dedicated General Matrix to Matrix Multiplication (GEMM) engine that lets the Habana Gaudi processor increase the performance of multiplying large matrices

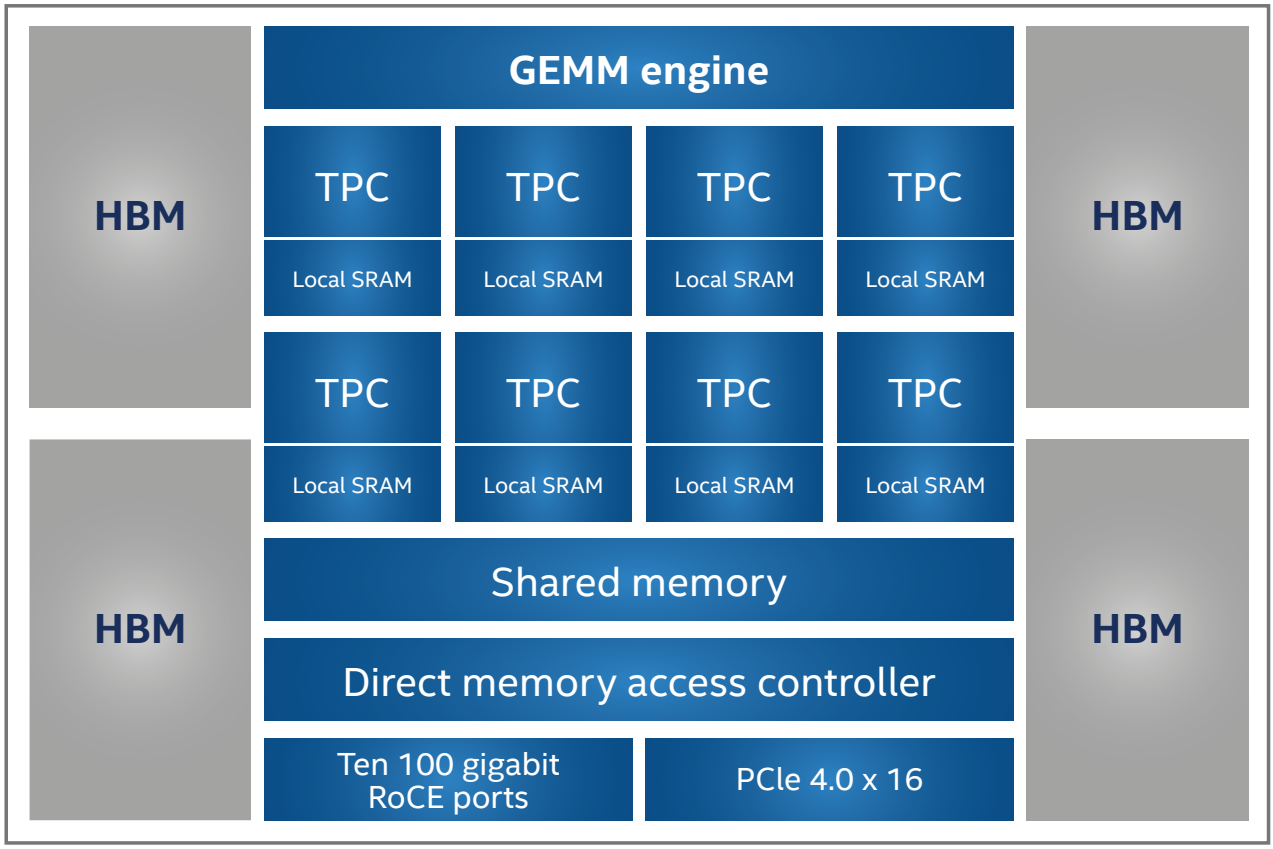

Figure 1. Each Habana Gaudi processor combines eight TPCs with 32 GB of HBM, a PCIe 4.0 interface with 16 lanes, and ten 100 gigabit remote direct memory access (RDMA) over converged Ethernet (RoCE) ports

Habana Gaudi processors are the first DL training processors to integrate ten 100 gigabit integrated remote direct memory access (RDMA) over converged Ethernet (RoCE) ports on the silicon. This networking capability:

- Provides up to 2 terabytes-per-second of bidirectional throughput
- Gives enterprises the ability to scale up in a rack configuration or to scale out across racks
- Uses standard Ethernet to eliminate proprietary interfaces and scale from one to thousands of Habana Gaudi processors
- Provides connections directly between Habana Gaudi processors within a single server, or between Habana Gaudi processors located across multiple servers using standard Ethernet switches
- Reduces communication bottlenecks between Habana Gaudi processors and Habana Gaudi processor–based servers
- Reduces total system cost with integration of network interface controllers (NICs) on the processor, which helps reduce overall component count and cost

The ten 100 gigabit RoCE ports can also be configured as 20 RoCE ports providing 50 or 25 gigabit speeds. This capability gives engineers more options for connecting Habana Gaudi processors to legacy Ethernet switches.

Each Habana Gaudi processor is packaged on a Habana® HL-205 Open Compute Project Accelerator Module (OCP-OAM) mezzanine card that communicates with the host server through a 16-lane PCIe 4.0 interface. OCP-OAM is an open standard that supports combining multiple Habana HL-205 mezzanine cards within server systems, such as the Supermicro X12 server. Additionally, multiple Habana Gaudi processor–based servers can be combined to provide massive, scalable parallelism.

Habana SynapseAI software suite

The Habana SynapseAI software suite is a software stack and set of tools that provides the ability to port existing models or build new models that use Habana Gaudi architecture capabilities. Habana Labs designed the SynapseAI software suite to provide ease of use to both software developers and data scientists. Additionally, Habana Labs built SynapseAI for seamless integration with existing frameworks that define neural networks and manage DL training.

The Habana SynapseAI software suite provides more than 1,400 TPC kernel libraries that open a full suite of Habana Gaudi processor capabilities. These kernel libraries allow for the development of customized TPC kernels that can augment kernels provided by Habana Labs. Additionally, the Habana SynapseAI software suite includes a debugger, a simulator, and a compiler.

Specific features of the Habana SynapseAI software suite include:

- **Integration with TensorFlow and PyTorch:** Both TensorFlow and PyTorch are leading frameworks that enable DL. Integration with these two frameworks means that data scientists and software developers can use existing models and familiar programming.[1]
- **Graph compiler and runtime:** The SynapseAI graph compiler and runtime implements model topologies on the Habana Gaudi using parallel execution of framework graphs. A multi-stream execution environment takes advantage of Habana Gaudi processors' unique combination of compute and networking capabilities. The processor also synchronizes execution of compute, network, and direct memory access (DMA) functions.[2]
- **Habana Communication Library (HCL):** By using the RDMA communications capabilities of the Habana Gaudi architecture, the HCL facilitates communication between Habana Gaudi processors in a single server or across multiple Habana Gaudi processor–based servers.[3]
- **TPC software development kit (TPC SDK):** While the Habana SynapseAI software suite provides more than 1,400 TPC kernel libraries, the TPC SDK gives software developers a compiler, simulator, and debugger that provides a development environment for custom TPC kernels. The TPC SDK uses the TPC-C programming language that supports a wide variety of DL instructions. These instructions include tensor-based memory access, special function acceleration, random number generation, and multiple data types.[4]

Supermicro X12 servers with Habana Gaudi AI processors: A foundation for performance

Powered by Habana Gaudi AI processors and 3rd Generation Intel Xeon Scalable processors, the Supermicro X12 Gaudi AI Training System (SYS-420GH-TNGR) is a 4U server designed specifically to train AI models quickly while helping reduce training costs.[5] The hardware uses the Habana Gaudi processor on-chip RoCE network engines to facilitate high-speed inter-processor communication during DL training, and to provide scale-up and scale-out capabilities.

Each Habana Gaudi processor dedicates seven of the ten integrated 100 gigabit RoCE ports to communication between the Habana Gaudi processors. Using an all-to-all connectivity method, each Habana Gaudi processor communicates with any other Gaudi processor within the system. The three remaining ports per Gaudi processor are available for scaling out across multiple Supermicro X12 servers using standard 100 gigabit switches. The ports can also be reconfigured into 50 or 25 gigabit ports for use with legacy switches.

Each Supermicro X12 server:

- Contains eight Habana HL-205 mezzanine cards for a total of 64 TPCs, dual 3rd Generation Intel Xeon Scalable processors, and up to 8 TB of 3,200 MHz DDR4 memory.
- Provides high compute utilization for GEMM calculations and convolutions.

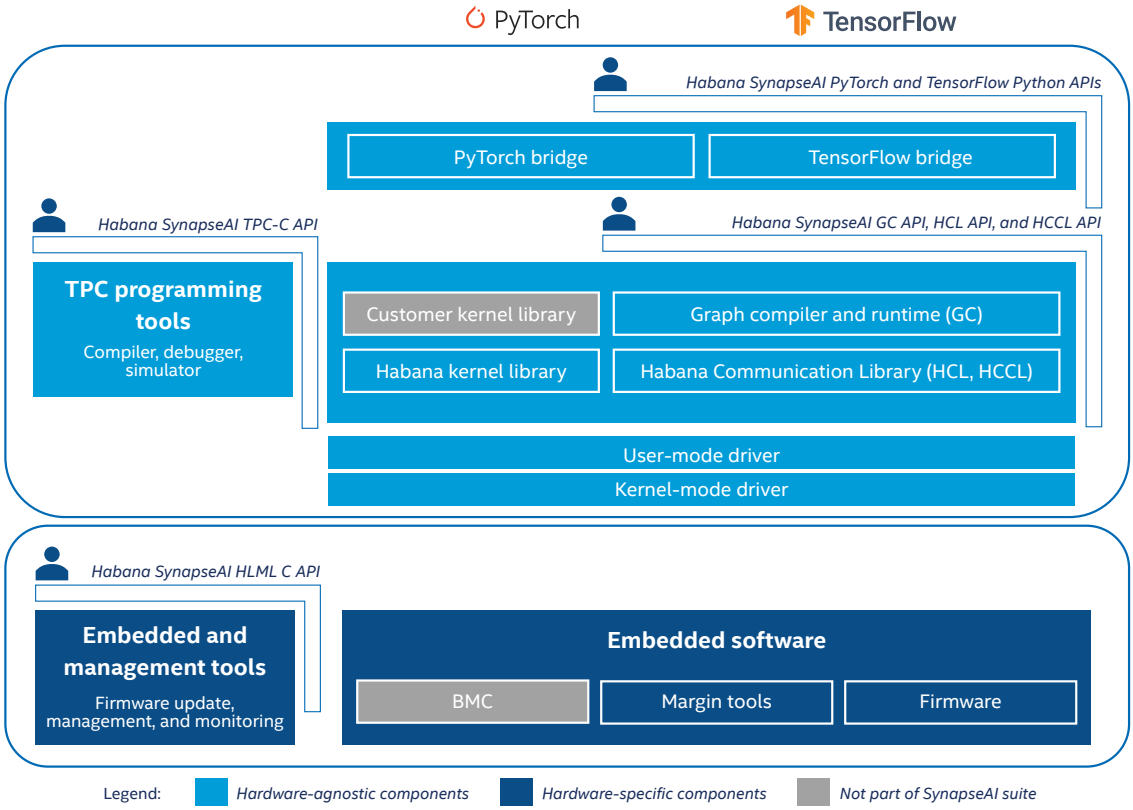

Figure 2. The Habana SynapseAI software suite provides tools for mapping neural network topologies onto Habana Gaudi processors

Supermicro also designed the Supermicro X12 server around a resource-saving architecture that continues a unique tradition of environmentally friendly innovation. This architecture helps lower operational costs by reducing power and cooling requirements while reducing e-waste by allowing components to be replaced as needed.

Each Supermicro X12 server uses a low-power system design that enables high-efficiency AI model training across virtually any AI use case, such as:

- **Computer vision applications:** The Supermicro X12 server can train vision models that are used across a wide variety of industries. These models can include applications that inspect manufactured items for defects, recognize the use of safety equipment in Internet of Things (IoT) camera feeds, or provide continuous improvement of assembly-line operations.
- **Inventory management:** Organizations across all industries that maintain any type of inventory—from grocery chains to healthcare organizations—use complex systems to manage inventory levels, warehouse space, demand forecasting, and customer feedback. The Supermicro X12 server can train models that predict understock or overstock conditions, help optimize warehouse space, and produce insights that are based on customer feedback.
- **Medical imaging:** Medical imaging devices, from X-ray machines to systems such MRI or CT scanners, have revolutionized medical care. The Supermicro X12 server can help researchers develop AI models that assist radiologists in detecting cancers and heart disease, or they can aid surgeons in developing their skills through surgical training platforms.
- **Language applications:** NLP is used across a wide variety of areas. The Supermicro X12 server can train AI models to increase the efficiency of NLP models. These models can help organizations understand customer sentiment on social-media platforms or summarize text by extracting only the most critical information.

Figure 3. The Supermicro X12 Gaudi AI Training System, powered by Habana Gaudi processors and 3rd Generation Intel Xeon Scalable processors, features 64 Habana TPCs and up to 8 TB of DDR4-3200MHz memory

Organizations with large AI training requirements can combine multiple Supermicro X12 servers to scale out to hundreds of Gaudi processors in a single AI cluster. Additionally, the high level of component integration of Supermicro X12 servers can help reduce system complexity and cost at any scale.

DDN A³I storage provides high-performance storage for Habana Gaudi processor–based clusters

While the Supermicro X12 server and Habana Gaudi processors provide an optimal compute environment for DL, a complete solution also requires fast, scalable, managed storage that reduces bottlenecks within the cluster. DDN provides a fully optimized storage solution that helps ensure that the Habana Gaudi processors are fully utilized at any scale.

DDN engineered the DDN Accelerated Any-Scale AI (A³I) solution from the ground up to help accelerate AI training application performance on Habana Gaudi processors. DDN has worked closely with Supermicro, Intel, and Habana Labs to provide predictable performance and capacity to Supermicro X12 servers.

The Voyager supercomputer project

The San Diego Supercomputer Center (SDSC) at the University of San Diego was awarded a National Science Foundation grant to build a unique, AI-focused supercomputer. The supercomputer, called Voyager, consists of 42 Supermicro X12 server Habana Gaudi AI processor–based training systems, in addition to two Supermicro SuperServer inferencing nodes that utilize 2nd Generation Intel Xeon Scalable processors and 16 Habana Goya HL-100 inferencing cards. The system contains a total of 336 Gaudi HL-205 processors.

Supermicro assisted with the Voyager design and assembled and tested all servers and clusters at its Northern California factory. The supercomputer provides data scientists with a unique system dedicated to advancing AI across science and engineering fields.

Shared parallel architecture increases storage performance and resiliency

The DDN AI400X storage appliance, part of the DDN A³I solution, provides a fully integrated shared data platform that delivers more than 50 GB/s and three million input/output operations per second (IOPS) directly to Supermicro X12 servers.[6] The DDN AI400X appliance integrates the DDN A³I shared parallel architecture that provides redundancy and automatic failover capabilities and gives Habana Gaudi processor–based clusters data resiliency. The storage appliance provides low latency with multiple parallel paths between storage and containerized applications running on Supermicro X12 servers. The DDN AI400X appliance also provides redundancy and automatic failover capabilities for high availability, and it enables concurrent and continuous execution of DL training across all Supermicro X12 servers in an AI cluster.

The shared parallel architecture can help accelerate DL by providing concurrent DL workflow execution across multiple Supermicro X12 servers. This concurrency helps complex DL models train faster and more efficiently across any number of Supermicro X12 servers in a Habana Gaudi processor–based AI cluster.

NUMA-optimized DDN A³I client reduces CPU and memory bottlenecks

DDN produces a DDN A³I client that is optimized for Supermicro X12 servers. The client is non-uniform memory access (NUMA)-aware and can automatically pin storage processing threads to specific Supermicro X12 NUMA nodes. This pinning helps ensure input/output (I/O) activity is optimized across the entire Habana Gaudi processor–based environment. By pinning storage processing threads to specific NUMA nodes, the DDN A³I client prioritizes processor and memory access for the threads, which helps accelerate data access from the DDN AI400X storage appliance to the Habana Gaudi processors.

Multirail networking increases bandwidth and redundancy

DDN AI400X storage appliances provide multirail networking that helps increase storage performance and resiliency across the AI cluster. Modern Ethernet networks provide high-bandwidth, low-latency connections between servers and storage appliances. But even when running at 100 gigabits-per-second, network connections can become a bottleneck when data is moving between processors on multiple cluster servers, or between cluster servers and storage. Additionally, cluster servers that rely on single connections to other servers or storage can experience network failures from faulty cabling or other factors.

DDN A³I MultiRail aggregates multiple network interfaces on Supermicro X12 servers with Gaudi processors, which provide the following capabilities:

- **Network traffic load balancing:** DDN A³I MultiRail balances network traffic dynamically across all network interfaces, which helps ensure that each interface is fully utilized and not overloaded.
- **Network link redundancy and failure detection:** If a network interface fails, DDN A³I MultiRail automatically rebalances traffic onto the remaining network interfaces, which helps ensure data availability and network resiliency.
- **Automatic recovery:** In addition to rerouting traffic across available network interfaces should a failure occur, DDN A³I MultiRail automatically routes traffic to network interfaces once they become available again.

When combined with multiple paths through multiple network switches, DDN A³I MultiRail provides a high-performance, resilient network backbone between Supermicro X12 servers and storage.

DDN A³I container client provides direct data access to containers

Habana Labs provides optimized TensorFlow and PyTorch containers that enable rapid development and deployment of DL framework applications. But shared container environments might not provide direct access to network storage from within containers running on a host server. Rather, containers either do not provide data persistence to applications within the container, or the containers and their applications rely on local storage volumes or a host-level connection to storage for persistence.

The DDN A3I container client provides direct, parallelized connections between application containers running on the Supermicro X12 servers and DDN storage appliances. By providing direct access, the DDN A3I client can help overcome data-sharing barriers and storage latency across the AI cluster.

Digital security framework and multitenancy help keep container environments more secure and dynamic

Container environments can be vulnerable to security breaches through privilege escalation attacks and shared data access. Additionally, container environments might not provide adequate multitenancy controls and security to share resources across a large AI cluster environment. DDN A³I client multitenancy provides a comprehensive digital-security framework and multitenant capabilities to help keep containers segregated. This container segregation provides the ability to share Supermicro X12 servers to a large number of users.

The DDN A³I container client enforces data segregation by restricting data access within containers while providing a security framework that prevents data access should a container be compromised. In addition to enforcing container security, the container client also provides multitenancy capabilities that make it simple to quickly provision resources among users. This multitenant capability helps balance loads across multiple Supermicro X12 servers. Multitenancy also reduces unnecessary data movement between storage locations, which can help increase DL performance.

Automatic tiering efficiently manages data storage and performance

The longer data remains on a storage device, the less likely it will be accessed by applications or users. The DDN A³I client provides automatic data tiering to keep frequently accessed data available on high-performance flash-based storage (hot pools) while moving older data to higher capacity, slower hard-drive-based storage (cool pools). Both pools can be scaled independently to help optimize storage costs and increase the performance of frequently used data.

DDN AI400X storage appliances can scale to any size Habana Gaudi processor–based cluster

Each DDN AI400X storage appliance communicates with Supermicro X12 servers using 50 gigabit Ethernet (GbE) and provides three million IOPS directly to the servers.[6] As more Supermicro X12 servers are added to a DL cluster, more DDN AI400X storage appliances can also be added to scale performance linearly.

DDN recommends one DDN AI400X storage appliance to service up to four Supermicro X12 servers with Gaudi AI processors. With this basic metric in mind, engineers can design virtually any size of DL cluster based on workload requirements.

High-performance networking ties compute and storage together

Habana Gaudi processor–based DL clusters typically require three networks for optimal performance: storage and cluster management, Habana Gaudi processor–based communication, and management. Supermicro can provide 1 GbE, 100 GbE, and 400 GbE network equipment to power these networks.

Storage and cluster management network

The storage and cluster management network provides the backbone for storage traffic and management of the Habana Gaudi processor–based DL cluster. The speed, latency, and stability of this network are crucial for overall cluster performance.

Supermicro can provide 100 GbE or higher network switches that can power this network. Modular switches can scale as storage and cluster needs increase.

Habana Gaudi processor–based network

The Habana Gaudi processor–based network provides the ability for the Habana Gaudi processors in each Supermicro X12 server to communicate directly with other Habana Gaudi processors across the cluster. This communication fabric is critical to enabling large DL models that can run in parallel across multiple Supermicro X12 servers.

High-performance, low-latency Ethernet is crucial for this network. Supermicro can provide 400 GbE switches that can scale as more Supermicro X12 servers are added to the DL cluster.

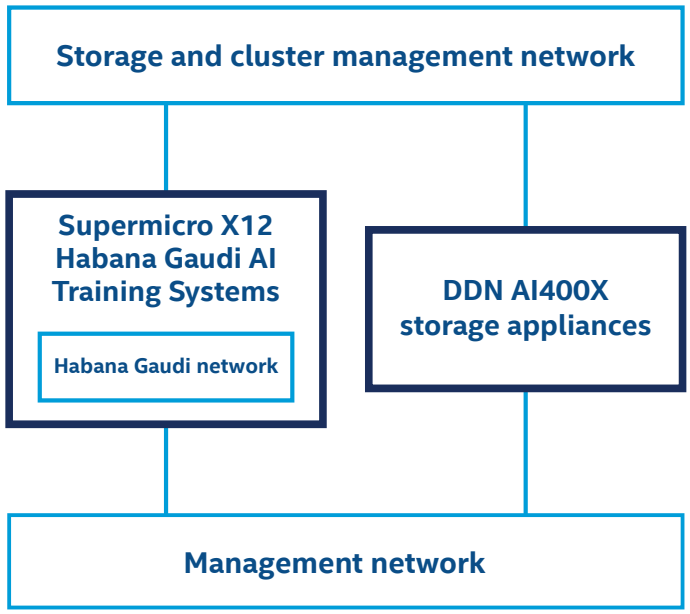

Figure 4. High-performance network switches are crucial for powering Habana Gaudi processor–based DL networks

Management network

A management network within a Habana Gaudi processor–based DL cluster is a lower-speed network used for managing individual cluster components. Management tasks can include monitoring the health of individual Supermicro X12 servers or managing containers. These tasks often don't require high bandwidth, so higher density, cost-efficient 1 GbE switches can be used.

Performance validation

DDN performed a series of tests to validate the performance of a Habana Gaudi processor–based DL cluster solution that included two sets of configurations. The first configuration contained a single DDN AI400X storage appliance and four Supermicro X12 servers. The second configuration contained up to eight DDN AI400X storage appliances and 32 Supermicro X12 servers.

DDN engineers used the open source fio benchmark tool to test the performance between the Supermicro X12 servers and the DDN AI400X storage appliances. This tool simulates a general-purpose workload but does not include any optimizations that enhance performance. Separate tests were run that simulated both 100 percent read and 100 percent write workloads.

Test results

Figure 5 shows that a single DDN AI400X storage appliance can provide more than 50 gigabytes-per-second of sustained read speeds, and more than 30 gigabytes-per-second of sustained write speeds for up to four Supermicro X12 servers. Performance increased as more servers were added to the cluster, with similar performance across the two-, three-, and four-server configurations.

Figure 6 shows the performance results for up to eight DDN AI400X storage appliances and 32 Supermicro X12 servers. Individual tests were run with the storage-appliance-to-server ratios shown in Table 1.

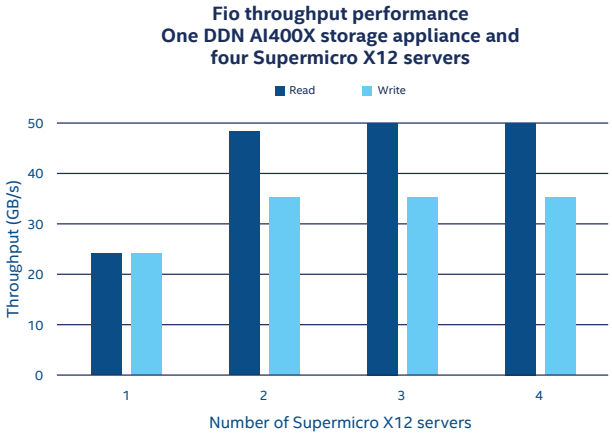

Figure 5. Performance results for one DDN AI400X storage appliance and four Supermicro X12 servers

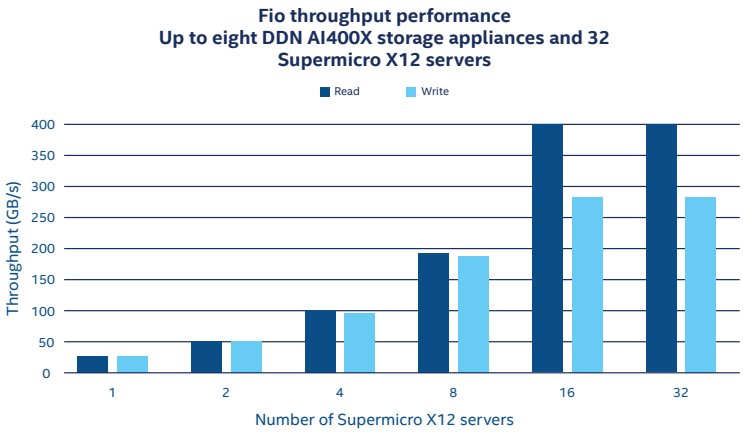

Figure 6. Performance results for up to eight DDN AI400X storage appliances and 32 Supermicro X12 servers

Table 1. Ratio of Supermicro X12 servers to DDN AI400X storage appliances

Number of Supermicro X12 servers	Number of DDN AI400X storage appliances
1	1
2	1
4	1
8	2
16	4
32	8

The total throughput increased as the number of servers and storage appliances increased. A single DDN AI400X storage appliance can provide up to 25 gigabytes-per-second of sustained read and write speeds for up to four Supermicro X12 servers. Performance peaked at 400 gigabytes-per-second of sustained read speed, and more than 250 gigabytes-per-second of sustained write speed, with the higher storage-appliance-to-server ratios.

Habana Labs, Supermicro, and DDN provide scalable capacity for the largest DL infrastructures

Habana Labs and Intel have partnered with DDN and Supermicro to provide an all-in-one DL training solution that helps enterprises overcome DL training cost and timing barriers. The solution is fully integrated, tested, built, and supported by Supermicro, and it can scale as DL models grow, from a single server and storage appliance to hundreds of servers and storage appliances.

For more information, please visit <<future landing page>>.

[1] For more information about Habana SynapseAI software suite integration with TensorFlow and PyTorch, visit https://docs.habana.ai/en/latest/index.html.
[2] For more information about the Habana SynapseAI graph compiler and runtime, visit https://docs.habana.ai/en/latest/Gaudi_Overview/Gaudi_Overview.html.
[3] For more information about the Habana Communications Library (HCL), visit https://docs.habana.ai/en/latest/Gaudi_Overview/Gaudi_Overview.html.
[4] For more information about the Habana SynapseAI TPC SDK, visit https://docs.habana.ai/en/latest/Gaudi_Overview/Gaudi_Overview.html.
[5] For more information about the Supermicro X12 Gaudi AI Training System (SYS-420GH-TNGR), visit supermicro.com/en/products/system/ai/4u/sys-420gh-tngr.
[6] DataDirect Networks. "DDN A^{3}I Solutions with Supermicro X12 Gaudi AI Servers." November 2021. ddn.com/wp-content/uploads/2021/11/A3I-X12-Gaudi-Reference-Architecture.pdf.

Printed in USA 0322/KM/PRW/PDF Please Recycle 350367-001US