PRODUCT BRIEF

SUPERMICRO X14 INTEL® GAUDI® AI ACCELERATOR CLUSTER REFERENCE DESIGN

Accelerating and Driving Down the Cost of AI Solutions with Supermicro's X14 Intel® Gaudi® 3 Accelerator Based System by Building on Open-Source Software and Industry-Standard Ethernet

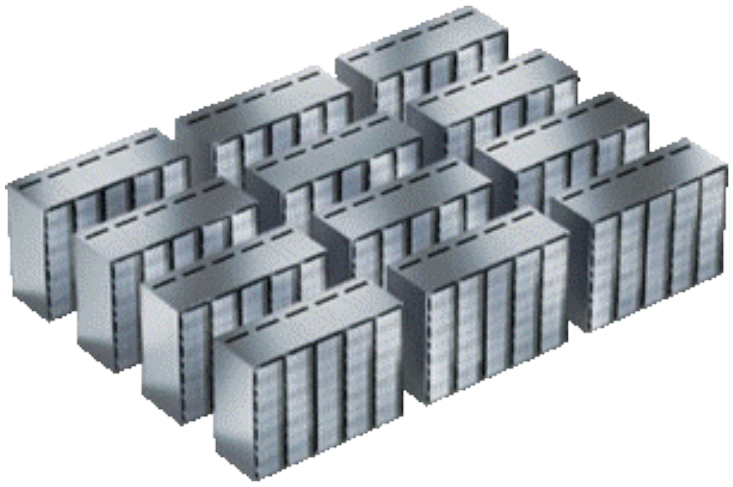

TABLE OF CONTENTS

Executive Summary

Supermicro's X14 Intel® Gaudi® 3 AI Accelerator-based cluster is a revolutionary AI infrastructure solution for large-scale training and inferencing, including LLMs and GenAI applications. Built around Supermicro X14 Gaudi 3 AI Systems, the SYS-822GA-NGR3, this cluster offers unparalleled scale, performance, and flexibility. As AI models grow in complexity, Intel Gaudi AI accelerator-based clusters address critical computational power, scalability, and cost-effectiveness challenges, enabling the efficient deployment of sophisticated AI inferencing solutions.

Optimized for AI inferencing, Intel® Gaudi® AI Accelerator-based clusters deliver low-latency, high-throughput performance for

complex models. Its modular design starts from a single cluster segment of eight SYS-822GA-NGR3 Intel® Gaudi® AI server solutions and can scale to over 512 nodes (4,096 GPUs, 64 cluster segments), all while being adaptive to growing AI workloads. Its flexibility and energy efficiency allow organizations to manage operational costs while staying at the forefront of AI innovation.

Powered by Intel® Gaudi® 3 AI accelerators and Intel® Xeon® 6 processors with Performance-cores, Supermicro is building out clusters that are purpose-built for running deep learning workloads of all sizes across multi-tenant data centers. With Supermicro's already proven expertise in building out data centers at scale, it is exciting to create a solution as a viable alternative to the competition in Generative AI compute capability, pricing, energy efficiency, and market availability.

Advantages of the Supermicro X14 Intel Gaudi 3 Cluster

1. **Unparalleled Gaudi Expertise:** As the sole provider of Intel® Gaudi® 1 and Intel® Gaudi® 2 systems, Supermicro has developed deep technical knowledge in maximizing Intel® Gaudi® AI accelerator performance. This expertise enables optimized configurations and superior support for Intel® Gaudi® 3 systems.

2. **Exclusive High-Performance CPU Configuration:** Supermicro uniquely offers Intel® Gaudi® 3 systems with dual 6th Gen Intel® Xeon® 6 Processors with P-cores, providing superior parallel processing, faster memory access, and AI-specific instructions for enhanced AI workload performance.

3. **Massive, Linear Scalability:** The innovative Intel® Gaudi® 3 cluster architecture, with a theoretical performance of 115.2 PFLOP/s (FP8 and BF16) per cluster segment, scales seamlessly from a single cluster segment of 8 nodes to over 512 nodes (4,096 accelerators, 64 cluster segments). This linear scaling, achieved using standard Ethernet switches, ensures cost-effective expansion while maintaining consistent performance across the entire deployment range.

4. **Advanced Networking Integration:** Incorporating cutting-edge 800GbE technology and Arista switches, this cluster architecture design delivers high-bandwidth, low-latency connectivity that scales with deployment size, crucial for distributed AI workloads.

5. **Optimized AI Architecture:** Purpose-built for diverse AI workloads, Intel's Gaudi® 3 AI Accelerator excels in training and inferencing large language models and complex AI applications, leveraging 64 Intel® Gaudi® 3 AI accelerators per cluster segment for exceptional performance.

6. **End-to-End Setup and Support:** Supermicro provides comprehensive solutions with rack-scale Plug & Play deployment, optimized architecture, and advanced management tools. Offering industry-leading TCO, flexible cooling options, and high-capacity manufacturing, Supermicro ensures efficient setup and operation of rack deployments with single-vendor simplicity.

Key Components

Supermicro X14 Intel® Gaudi® 3 AI System (Compute Node)

At the heart of the Supermicro X14 Intel® Gaudi® 3-based cluster is Supermicro's X14 Intel® Gaudi® 3 AI System, SYS-822GA-NGR3, a powerhouse designed for AI excellence. This system uniquely combines eight Intel® Gaudi® 3 accelerators with dual Intel®

Xeon® 6 Processors with P-Cores, a configuration exclusive to Supermicro. Intel's Gaudi® 3 accelerators deliver exceptional AI performance, while the high-core-count 6th Gen Xeon® Processors provide superior parallel processing capabilities and faster memory access, significantly enhancing overall system performance for AI workloads.

Supermicro's X14 Intel® Gaudi® 3 AI system solution features advanced networking capabilities, with six OSFP ports providing 800GbE connectivity. This high-bandwidth network ensures seamless data flow across the entire infrastructure, which is crucial for distributed training and real-time inference at scale.

Intel's Gaudi® 3 AI Accelerator is one of the driving forces behind Supermicro's X14 Intel Gaudi 3 server's exceptional performance. Each accelerator features 128 GB of HBM2e memory with 3.7TB/s of aggregate memory bandwidth, optimized for the most challenging AI computations. Intel's Gaudi® 3 AI Accelerator provides 1.5X higher bandwidth per accelerator at 900GBps, ensuring efficient data processing for complex AI models.

A key feature of Intel's Gaudi® 3 AI Accelerator is its on-chip RDMA over Converged Ethernet (RoCE v2) capability. This enables efficient scaling without additional networking hardware, which is crucial for building large-scale AI clusters.

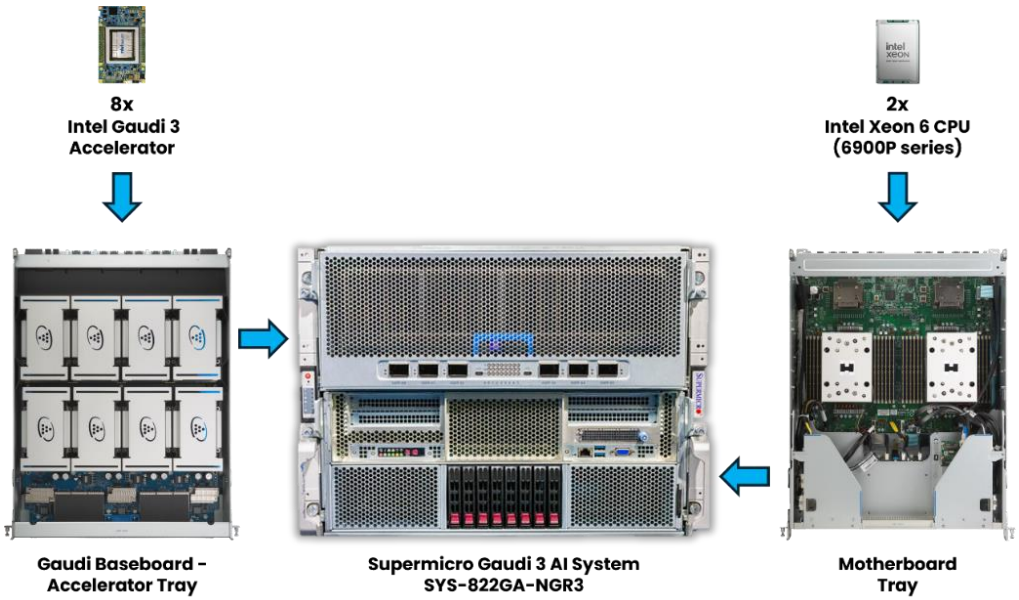

Figure 1 - Supermicro X14 Intel Gaudi 3 System - Key Components

Cluster Networking (Ethernet Switches)

Typically, generative AI clusters have three primary high-speed networks: compute, storage, and control plane servers – and a fourth network for out-of-band management. In an Intel® Gaudi® 3 AI accelerator-based cluster, a dedicated accelerator fabric connects each of Supermicro's X14 SYS-822GA-NGR3 systems. This fabric is an Ethernet network that Intel's Gaudi® 3 AI Accelerators use to directly communicate with other SYS-822GA-NGR3 nodes. It is implemented as a three-ply full Clos Ethernet fabric using OSFP4x 200 Gbps links to provide 800 Gbps per connector.

Knowing this, high-performance Ethernet switches will form the backbone of the cluster network architecture, seamlessly interconnecting its nodes.

Supermicro partners with industry leaders like Arista to provide optimal connectivity, incorporating validated switches that ensure reliable, high-speed communication throughout the cluster. These switches are crucial for maintaining Intel® Gaudi® 3 AI Accelerator-based clusters with impressive scalability and performance, from single-unit deployments to massive, multi-node configurations.

As networking requirements evolve, Supermicro remains committed to working with a growing ecosystem of switch vendors. For instance, some Cisco and Broadcom switches have already been validated, and solutions from other vendors could be supported in the future. Supermicro continues to validate and support additional choices, allowing customers to select the best-fitting networking components for their AI infrastructure.

Storage

Storage is the most customer-dependent aspect of a cluster. For this Reference Design, the storage block can be based on 2U servers with a minimum of 8x NVMe SSDs and 2x 100 Gbps network connections per server. For example, if you were to scale this up to a 32-node cluster, a storage block consisting of thirty-two 2U servers could deliver about 20 GB/s large-block random read performance per server or approximately 0.5 TB/s random read per cluster. Storage capacity is highly customizable and workload-dependent and can be configured by both size and number of SSDs used per storage node.

Each storage node should have two 100 Gbps links to a leaf switch at the top of each storage block rack. Complimenting all installed systems, the storage servers are connected to redundant power. BMC/IPMI is linked to management switches, with the 100 Gbps connections to the storage leafs also linked to management switches.

Remember that the management fabric, or out-of-band fabric, provides isolated access to core data center infrastructure. The three main areas are BMCs, network switches, and PDUs. These devices are connected to 1 Gbps management switches associated with each node and cluster segment. Although there are many ways to configure this network, three VLANs are implemented in the four management switches to isolate BMC, PDU, and switch access.

Control Plane

Control plane servers provide several functions and run the control stack, management fabric backbone, control plane fabric spine, storage spine, service leaf, and console fabric. They also provide highly reliable storage for the control plane.
Configurations can be flexible depending on customer requirements. For smaller clusters, as few as two control plane servers are required. For larger configurations, it can scale up to needing six (or more) control plane servers.

Intel Gaudi Software Suite

The software stack for an Intel® Gaudi® 3 AI accelerator-based cluster is designed to streamline cluster deployment and operation. The deployment and operations include cluster provisioning, configuration, upgrades, monitoring, and visualization; network configuration; performance and system health monitoring; and user management.

Category	Function	Tool
Virtualization, Containerization, and Orchestration	Container foundation	Kubernetes (K8s)
	K8s cluster setup	Kubespray
	K8s device plugin	Gaudi K8s plugin
	K8s virtualization	KubeVirt
Authentication and Key Management	Key storage and management	HashiCorp Vault
	AAA, LDAP, DNS, key management	Feeipa
Network	K8s networking and security	Calico
	Load balancing	MetalB
	Network source of truth / IP address management (IPAM)	NetBox
Storage	Storage	WEKA
Provisioning and Management	Provisioning and configuration management	Ansible
	Network Time Protocol (NTP)	Chrony
	MLOps	Kubeflow
Monitoring	Real-time metrics	Prometheus
	Visualization	Grafana
	Centralized log management	Loki
	Log aggregation	Promtail
Operating System	Libraries, compiler, runtime	Intel® Gaudi® software
	Node OS	Ubuntu Linux

Table 1 - Software Components

Cluster Architecture and Layout

Each cluster segment is built to house eight nodes (see Figure 2 on the previous page). Each group shares three networking switches used as Intel Gaudi AI accelerator interconnect leaf switches. The eight systems are arranged in three racks, in a 3-2-3 configuration, encircling the three leaf switches. This minimizes the length of 800 Gbps active Cu cables between the compute nodes and the networking switches. Racks are configured with either two or three compute nodes per rack. The Intel Gaudi AI accelerator interconnect leaf switches are placed in the two-system rack. Each group of eight systems also has a 1 Gbps management switch for connecting BMCs, PDUs, and switches into a management fabric.

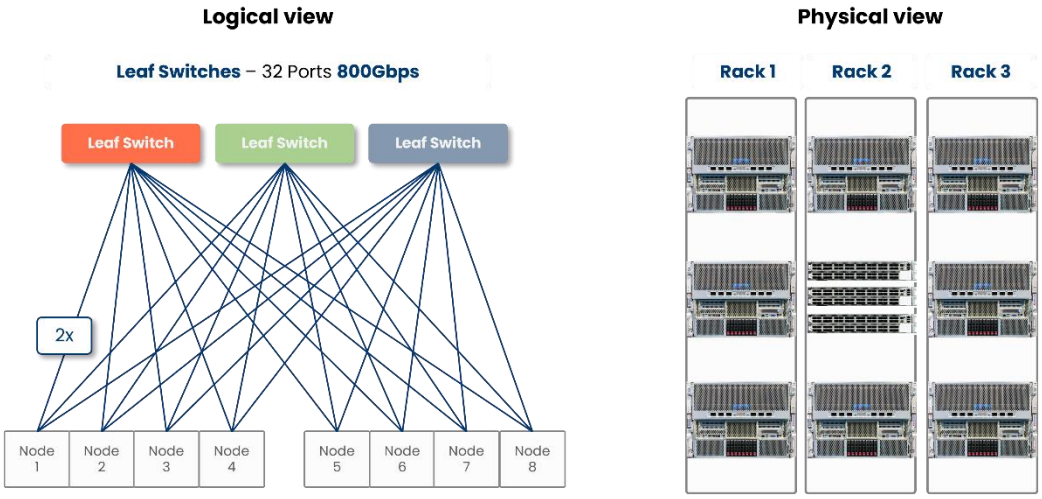

Figure 2 - Compute Rack Configuration Overview: Logical and Physical Views

Scalability

Supermicro's cluster architecture is designed for seamless scale-out, allowing organizations to grow their AI infrastructure as needed. The scalability ranges from a single node to massive deployments of 512 nodes or more. Defining segments, or blocks, that can be quickly and repeatedly represented by an instance and tied together by a core infrastructure is the key to Supermicro's ability to build at scale.

As deployments scale out, the network infrastructure grows proportionally to support increased computational capacity. The network architecture employs a two-tier topology, using 32-port or 64-port Leaf Switches and 64-port Spine Switches. This design enables efficient data flow and inter-node communication across the entire range of configurations.

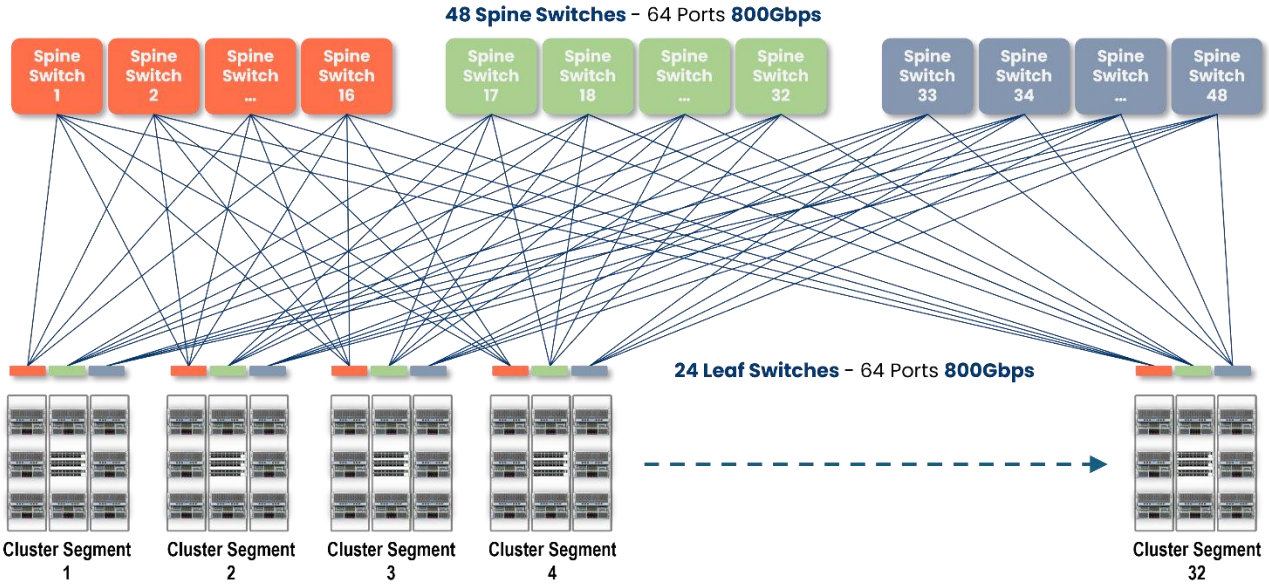

Figure 3 - Leaf - Spine Switch Reference Architecture for 256 Nodes

In larger deployments, such as 32 cluster segments (256 nodes), the network expands to incorporate 24 64-port leaf switches and 48 64-port spine switches. See Table 1 for more information. Further, in even larger clusters with 4K or 8K Intel® Gaudi® 3 AI Accelerators (Table 1 for reference), it is possible to scale by using the industry-standard Ethernet-based accelerator fabric – and there are many ways to achieve this. For example, the simplest way is to increase the radix of spine switches. Cluster build-outs with 16K or more accelerators are possible with commercially available Ethernet switch products. As you may have already deduced, all the fabrics and storage capacity requirements will need to increase proportionally. Additionally, a few control plane servers will most likely, be needed as more demand is placed on operational services. See Table 1 as a scaling table guide for cluster design.

This extensive network maintains full bisection bandwidth throughout the cluster, ensuring optimal performance even at the largest scales.

Cluster Size (Nodes)	Number of Intel® Gaudi® 3 AI Accelerators	FP8 AI Compute*	Number of Spine Switches (64-ports)	Number of Leaf Switches (64 ports)
256	2048	3.76EF	24	48
512	4096	7.52EF	48	96
1024	8192	15EF	96	192

Table 2 - Intel Gaudi 3 AI Compute Switching Network Reference

*Peak projected performance varies by use, configuration, and other factors. Results may vary.

For more information about switching network options, refer to the Supermicro Gaudi 3 System reference architecture in the reference section.

Power Consumption

Supermicro X14 Intel® Gaudi® 3 cluster build-out power consumption is critical in data center planning and operations. Here's a breakdown of the power requirements:

- Individual Intel® Gaudi® 3 Accelerator: Approximately 900W.
- Supermicro X14 Intel® Gaudi® 3 Server (SYS-822GA-NGR3): Approximately 13kW.
- Full 8-node cluster segment: Average 110kW, including 8 servers and 3 switches.

It's important to note that the actual power consumption of a cluster segment may be higher when considering additional elements such as storage systems, cooling infrastructure, and other auxiliary equipment. Data center planners should account for peak power usage that could exceed the average 110kW per cluster segment.

In large-scale deployments, power management becomes increasingly critical. For instance, deploying 256 nodes (32 cluster segments / 2,048 AI accelerators) could theoretically consume up to 3.5MW of power. This underscores the importance of efficient power delivery and cooling solutions in data centers hosting large Intel Gaudi 3 clusters.

AI Applications and Inferencing

Supermicro's cluster architecture excels in a wide range of AI inferencing applications, with the ability to expand from beyond small-to-medium deployments to massive cluster configurations. This flexibility allows the ability to tackle AI workloads of any size, from focused projects to enterprise-wide implementations:

• Large Language Models (LLMs): Enables real-time language processing for chatbots, automated content generation, and translation services.
• Computer Vision: Powers advanced image recognition, real-time video analysis, and object detection at scale, crucial for autonomous vehicles, manufacturing quality control, and medical imaging.
• Natural Language Processing: Facilitates rapid sentiment analysis, text classification, and summarization, enhancing sophisticated language understanding.
• Scientific Computing: Accelerates complex simulations and modeling, potentially advancing fields like drug discovery, climate science, and financial risk analysis.

This cluster architecture is optimized for high throughput inferencing and allows for parallel processing of vast amounts of data across these diverse applications, with performance that grows linearly as more pods are added to the system.

Summary

Supermicro's X14 Intel® Gaudi® 3 AI cluster build outs represent a paradigm shift in AI infrastructure. Its innovative design, combining unparalleled compute power, advanced networking, and modular scalability, provides a future-proof foundation for the most ambitious AI projects.

As Generative AI becomes steadily embedded in every industry, Supermicro's cluster network architecture offers a clear path forward. It can provide the high performance necessary for rapid training and inference at a scale and efficiency desired by companies of all sizes. This is more than just hardware; it's a complete ecosystem designed to accelerate AI innovation and drive business transformation, from small-to-medium deployments all the way up to massive, data-center-scale installations.

This reference design gives an example of simplified cluster deployment by providing an overview of selecting and configuring system components. Using this reference design, small, medium, and large enterprises can obtain an overview of how to take their AI journey to the next level.

To learn more about how Supermicro's X14 Intel® Gaudi® 3 cluster can revolutionize your AI infrastructure and capabilities, please contact your Supermicro representative.

Further Information

Supermicro Gaudi 3 Reference Architecture:
https://www.supermicro.com/white_paper/white_paper_Gaudi3_Reference_Archtecture.pdf

Supermicro Gaudi 3 Product Brief: https://www.supermicro.com/products/brief/Product-Brief-Gaudi3.pdf

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements. Visit www.supermicro.com