WHITE PAPER

SUPERMICRO WITH GAUDI 3 AI DELIVERS SCALABLE PERFORMANCE FOR AI REQUIREMENTS

Range of Optimized Solutions for Data Centers of Any Size and Workloads For New Services and Increased Customer Satisfaction

TABLE OF CONTENTS

Executive Summary

The rapid advancement of Artificial Intelligence (AI) has led to widespread adoption across industries, driving innovation in sectors from healthcare and finance to manufacturing. With global AI revenue projected to reach $1.8 trillion by 2030, the demand for robust and efficient AI infrastructure continues to grow.

Organizations are swiftly transitioning from experimental projects to full-scale implementation of generative AI (GenAI) and Large Language Models (LLMs). This shift presents new challenges, primarily the need for specialized AI compute accelerators. However, it is crucial to consider the entire end-to-end solution and a scale-out architecture that supports growth from development to production. Enterprises must optimize their infrastructure while managing costs and meeting the increasing demand for more powerful AI capabilities.

The Supermicro Gaudi 3 AI System SYS-822GA-NGR3, featuring Intel Gaudi 3 Accelerators and Intel Xeon 6 CPUs, offers a compelling choice in the enterprise AI market. The system, coupled with a proven, scalable AI platform, enables businesses to efficiently expand their AI operations to thousands of AI accelerators. This design guide provides a detailed roadmap for designing a cohesive AI environment that balances performance, scalability, software integration, and infrastructure management.

Introduction

The Challenges of AI Infrastructure

The synergistic combination of LLMs, extensive datasets, and high-performance GPU computing drives the current revolution in Artificial Intelligence. This powerful convergence leads to the widespread adoption of AI technologies across industries. However, it has also unveiled critical pain points in developing and deploying scale-out AI infrastructure.

A primary concern is managing the immense computational power required for training and inference serving of GenAI models, particularly LLMs. Organizations must scale their infrastructure to support increasing workload demands while maintaining high performance. This requirement necessitates solutions offering seamless expansion and integration of additional resources without compromising processing speed or efficiency.

The shift towards new data formats like BF16, FP8, and the upcoming FP4 presents both an opportunity and a challenge. These data types offer the potential for improved performance and reduced memory usage, but they also require hardware and software stacks capable of efficiently managing these new precisions.

Many organizations are also constrained by proprietary software ecosystems, limiting flexibility and interoperability. The lack of standardized, open-source frameworks can hinder innovation and adaptability to evolving AI technologies.

Another set of challenges revolves around data management and system interconnectivity. AI applications generate and process vast amounts of data, requiring robust storage solutions and efficient data management strategies. High-capacity, fast-access storage systems are essential, as is the ability to move data quickly between storage and processing units. This necessitates high-bandwidth, low-latency networking solutions to support seamless data flow and inter-node communication.

As a result, a single AI system, or even a few, is woefully inadequate to harness the full potential of these modern workloads. To truly power AI across industries, a robust infrastructure stack is essential—one that offers open-source flexibility and ensures high-performance interconnectivity and seamless scales to meet the ever-growing demands of artificial intelligence.

Requirements For Deploying AI Infrastructure

Building effective AI infrastructure requires a comprehensive approach that addresses multiple facets of system design, performance, and management. The following points outline the key requirements for a scalable and future-proof AI environment:

High-performance Hardware Acceleration:

- Dense computing power with specialized AI accelerators.
- Latest CPU configuration with higher core count for parallel AI workloads.

• High-speed memory for faster data access and improved system reliability.

Robust and Scalable Network Architecture:
• Optimized chip-to-chip performance between servers.
• Scalable design to accommodate growing computational needs.
• Eliminate vendor lock-in from proprietary networking fabrics.

Efficient and Scalable Storage Solutions:
• High bandwidth and low latency network to accelerate AI data pipelines.
• Diverse system and software solutions designed to fit specific customer's needs.
• Seamlessly integrated with the overall architecture, ensuring interoperability across all components.

Optimized and Open Software Stack:
• Open software ecosystem: no software licensing costs and community-based open-source software stack.
• Ease of use with simplified processes for model testing and inferencing.
• Support for all major open-source AI frameworks and ease of model porting across hardware types.

Effective Deployment and Management Strategies:
• Centralization of compute and storage resources for efficient management.
• Comprehensive cluster management and monitoring capabilities.

These requirements in AI infrastructures underscore the need for high-performance hardware (including specialized AI accelerators and servers), robust and scalable network architecture, and efficient storage solutions. Additionally, there is a need for optimized and open software stacks and management tools.

Supermicro's Solution with Intel Gaudi 3 AI Acceleration Platform

The Supermicro Gaudi platform is built on a foundation of comprehensive infrastructure. At its core is the SYS-822GA-NGR3 system, which provides the essential compute capabilities and hosts the Intel Gaudi 3 Mezzanine Cards (HL-325L) for AI acceleration. This design directly addresses the need for high-performance hardware acceleration in AI infrastructure.

Complementing the computing power, Supermicro's infrastructure includes a range of network switches, from 1Gbps to 800Gbps, for AI compute, management, and storage connectivity. The storage solutions also integrate with the overall infrastructure, offering high-performance storage as an integral part of the platform. These options cater to diverse storage needs, from ultra-fast NVMe drives to scalable all-flash arrays refined for AI workloads. This comprehensive approach fulfills robust, scalable network architecture requirements and efficient, scalable storage solutions.

The hardware foundation is further enhanced by Supermicro's management tools, such as Cloud Composer and SuperCloud Orchestrator, which improve system flexibility and ease of administration across compute, network, and storage resources. These tools support effective deployment and management strategies for AI infrastructure.

The Supermicro Gaudi solution runs on Enterprise Linux with container runtime support, creating an optimal environment for AI operations. This foundation provides isolation, portability, and efficient resource utilization, enabling seamless deployment and scaling of AI applications. This aligns with the requirement for an optimized software stack.

At the heart of the stack, the Intel Gaudi Software Suite provides frameworks, libraries, and orchestration tools specifically optimized for AI workloads on Gaudi hardware. Intel's Gaudi Software Suite utilizes popular frameworks such as PyTorch and DeepSpeed, which use the Gaudi accelerator for high-performance model training and inference. The Intel Embedded Software layer ensures tight integration between the software suite and the underlying hardware. This suite meets the need for support of major open-source AI frameworks and ease of model porting.

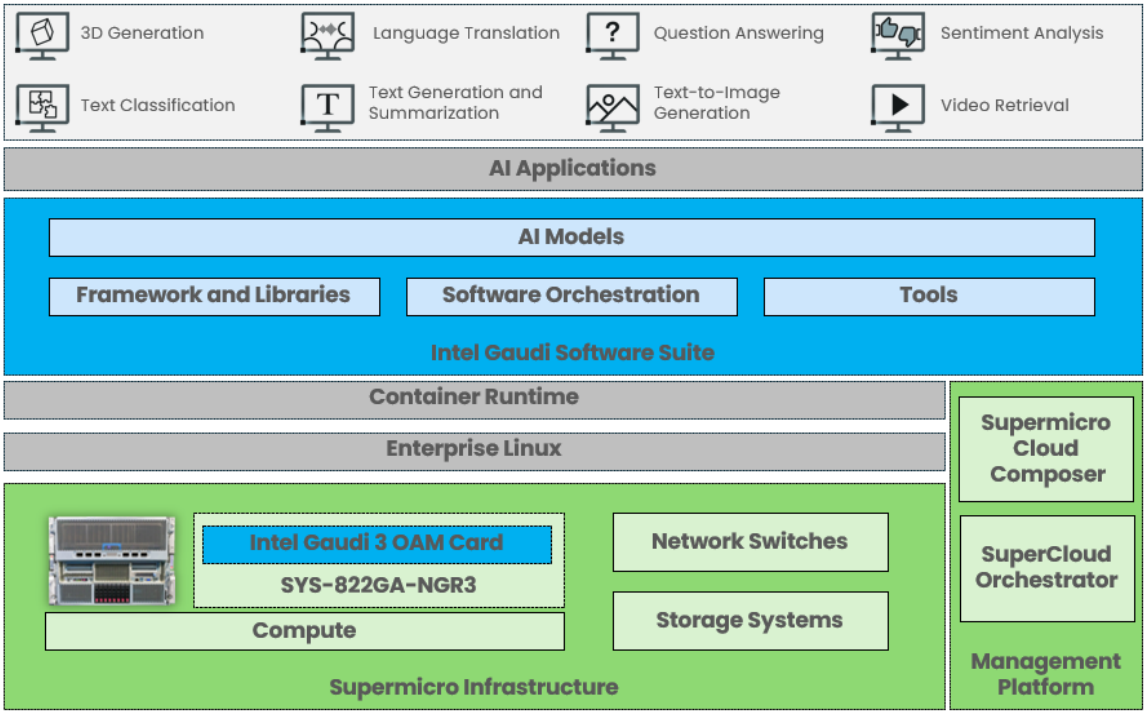

Figure 1 - Supermicro Gaudi Platform

Overall, this software solution enables a wide range of AI applications, from 3D generation and language translation to sentiment analysis and video retrieval. Combined, this offers a versatile, powerful platform for organizations looking to leverage innovative AI technologies across various domains, addressing the comprehensive requirements for deploying AI infrastructure.

Supermicro Gaudi 3 AI System Overview

Supermicro X14 Gaudi 3 AI Training and Inference Platform

Supermicro brings choice to the enterprise AI market with the introduction of the Gaudi 3 AI platform—part of the company's X14 systems generation. The high-performance SYS-822GA-NGR3 system is designed to further increase the efficiency of large-scale AI model training and AI inferencing. It combines the power of two Intel Xeon 6 CPUs (6900-Series with P-Cores) and eight Gaudi 3 AI accelerators, creating a robust platform for demanding AI workloads.

The SYS-822GA-NGR3 is designed as an 8U air-cooled system, capable of easily managing the 900W TDP per chip, showing substantial computational power and thermal management capabilities.

The networking capabilities are particularly noteworthy, featuring on-board 6x 800GbE OSFP, open industry-standard Ethernet, for scale-out operations. This high-bandwidth connectivity ensures rapid data transfer and efficient communication in large-scale AI deployments, from eight accelerators to thousands.

The system includes 8x NVMe hot-swap 2.5" drive bays for local storage. Although this storage is not primarily intended for large-scale data handling, it provides ample space for operating system installation, local caching, or conducting small-scale proof-of-concept AI projects.

Regarding power consumption, each Gaudi system will consume an average of 13kW. It is important to note that this power will vary depending on the components and system configuration:

Supermicro Gaudi 3 AI Server: SYS-822GA-NGR3	
AI Accelerator	8 Gaudi 3 HL-325L (air-cooled) accelerators on OAM 2.0 baseboard
CPU	Dual Intel® Xeon® 6 processors (6900-Series with P-Cores)
Memory	24 DIMMs - up to 6TB memory in 1DPC
Power Supplies	8 3000W high efficiency fully redundant (4+4) Titanium Level
Networking	6 on-board OSFP 800GbE ports for scale-out
Expansion Slots	2 PCIe 5.0 x16 (FHHL) + 2 PCIe 5.0 x8 (FHHL)

Table 1- SYS-822GA-NGR3 Key Specifications

Intel Gaudi 3 AI Accelerator

The Intel Gaudi 3 AI accelerator (HL-325L) is a high-performance mezzanine card designed for large-scale deployment in data centers. Built on 5nm process technology, it features 8 MME engines, 64 programmable Tensor Processor Cores, 128GB of HBM2E memory, and 96MB of SRAM.

Regarding compute technology, the Gaudi 3 brings 64 fully programmable Tensor Processor Cores (TPC) and GEMM Engines. Intel Gaudi 3 supports advanced data types for AI, including FP8, BF16, FP16, TF32, and FP32. The TPC core is specifically designed for deep learning, training, and inference workloads. Memory-wise, the Gaudi 3 incorporates HBM technology, offering an impressive 128GB capacity with a 3.7TB/s total throughput. Its advanced HBM controller is optimized for both random and linear access patterns.

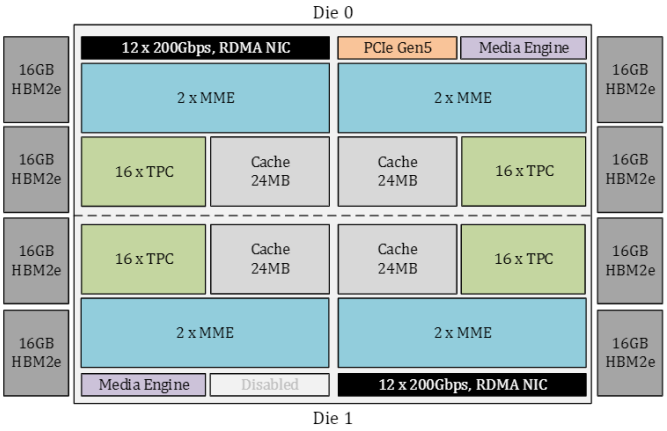

Figure 2 - Intel Gaudi 3 AI Accelerator: Block Diagram (Courtesy of Intel)

A standout feature of the Gaudi 3 is its scale-out capability with integrated RDMA. It is the only AI deep learning processor to integrate on-chip RDMA over converged Ethernet (RoCEv2), interfacing with industry-standard Ethernet networking. The chip interconnect technology is based on forty-eight pairs of 112Gbps, configured as twenty-four ports of 200Gbps Ethernet. Intel's Habana Processing Unit (HPU) provides unmatched scalability with 9.6 Terabits per second bi-directional networking capacity, allowing the SYS-822GA-NGR3 to use standard Ethernet technology and supporting a card with a TDP of up to 900W.

Compared to its predecessor, a single Gaudi 3 accelerator delivers 4x AI compute for BF16, 2x AI compute for FP8, 1.5x increase in memory bandwidth, and 2x networking bandwidth for massive system scale-out. This significant leap in performance and productivity enhances AI training and inference on popular large language models (LLMs) and multimodal models.

Compute Board: Universal Baseboard (UBB)

Intel's Gaudi 3 baseboard (HLB-325) assembles eight Intel Gaudi 3 AI accelerator OAM mezzanine cards, offering a readily integrated module subsystem for Supermicro's SYS-822GA-NGR3.

The baseboard provides direct all-to-all connectivity with 8.4TB/s per second of bi-directional bandwidth between the eight accelerators without requiring a separate switching IC. Additionally, it offers 1.2TB per second of bi-directional scale-out bandwidth through 6 OSFP connectors used by Supermicro's Gaudi 3 Server for massive data center scalability. This baseboard supports a total TDP of 7.6KW.

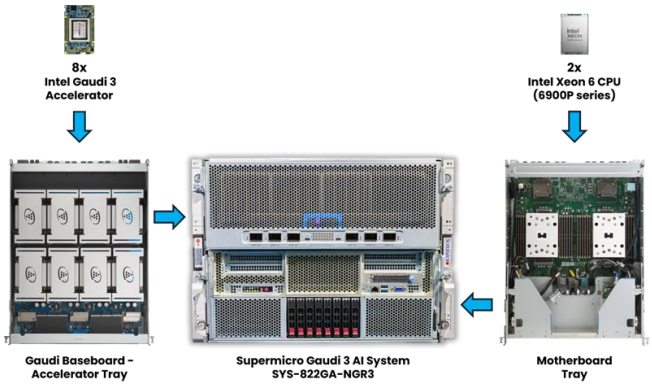

Figure 3 - Supermicro Gaudi 3 System: Key Components

The Supermicro Intel Gaudi Advantage

Supermicro Intel Gaudi Legacy

Supermicro's leadership in offering Intel Gaudi-based systems sets it apart in the AI infrastructure market. As the sole provider of Gaudi 1 and Gaudi 2 systems, Supermicro has developed deep technical knowledge and refined its system designs to maximize the performance of Gaudi accelerators.

Supermicro Gaudi 1
SYS-420GH-TNGR

Supermicro Gaudi 2
SYS-820GH-TNR2

Figure 4 - Supermicro Gaudi 1 and Gaudi 2 Systems

This expertise extends to customer environments, where Supermicro has successfully implemented and supported production deployments of Gaudi-based systems. In addition, our knowledge in conducting numerous proof-of-concept (POC) projects with Gaudi accelerators has provided invaluable insights into diverse AI workloads and customer requirements.

Supermicro is uniquely positioned to offer optimized configurations, provide expert guidance on system architecture, and provide superior support for the new Gaudi 3 systems.

Supermicro X14 Gaudi 3 with Intel Xeon 6 CPUs

Supermicro is the only vendor offering dual Intel Xeon 6900P-series with P-cores configuration, setting it apart from all competition. This unique combination of advanced CPUs and specialized AI accelerators positions Supermicro's Gaudi 3 system as a leader in high-performance AI infrastructure solutions with Intel.

Figure 5 - Intel Xeon 6 series processor with P-Cores

The advantages of the Intel Xeon 6900 series with P-cores CPU are particularly beneficial for AI workloads due to the nature of AI computations and data handling requirements. The significantly higher core count (128 cores) allows for more efficient

parallel processing of AI tasks, such as data preprocessing, feature extraction, and model training. This parallelism is crucial in AI workloads where large datasets must be processed simultaneously.

The improved memory capabilities, including faster DDR5 speeds of up to 6400 MT/s and support for 8800 MT/s MRDIMMs, enable quicker data access and transfer, which is essential for feeding data to GPUs in AI training scenarios without bottlenecks. Likewise, the increased PCIe lane count and CXL 2.0 support enhanced connectivity with GPUs and high-speed storage, allowing for more efficient data movement in complex AI systems.

The range of power options (400 to 500W) and improved scalability also mean that these processors can be more finely tuned to specific AI workloads, optimizing performance and energy efficiency. Moreover, the Intel Xeon 6900 series processors with P-cores add AI-specific instructions like AVX2 with VNNI/INT8 and BFloat16 support, which can dramatically accelerate certain types of AI computations, particularly in inference tasks.

All these factors combined lead to potentially faster training times, more efficient inference, and the ability to handle larger and more complex AI models, making the Supermicro Gaudi 3 System, with Intel Xeon 6900 series with P-cores, particularly well-suited for demanding AI applications.

Supermicro Gaudi 3 Scale-out Network Architecture

Having examined the individual server configuration and its key components, let us focus on the interconnection of multiple systems in Supermicro Gaudi 3 deployments. In these deployments, the AI compute network architecture plays a crucial role in ensuring optimal performance and scalability. Gaudi System's architecture is designed to provide high-performance connectivity for AI applications across a range of deployment sizes. It supports configurations from small clusters to large-scale systems, accommodating the diverse needs of AI researchers and enterprise users.

The optimized AI compute network of the system, featuring the 6x 800GbE OSFP, allows for seamless and cost-effective integration into Ethernet-based networks. This advanced architecture is designed to efficiently support a wide range of GenAI workloads, scaling effortlessly from single-node setups to large-scale deployments spanning thousands of nodes. Scalability is essential for training and deploying the latest generation of LLMs, which often require distributed computing across multiple nodes to achieve reasonable training times and inference performance.

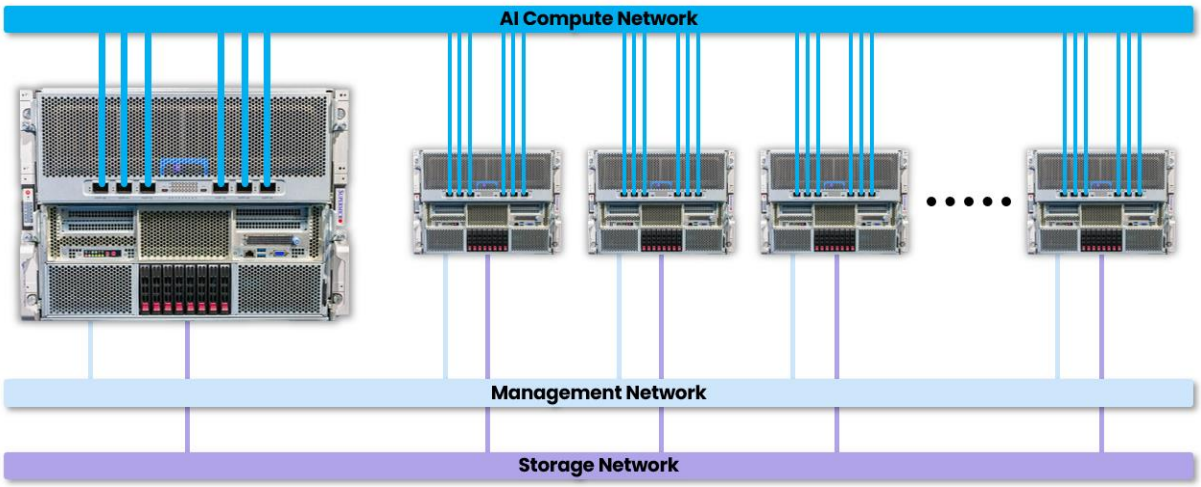

Figure 6 - Network Schematic

Alongside the AI compute network, the Management Network is also required for system administration, monitoring, and remote control of AI servers and associated infrastructure, and the Storage Network also provides dedicated high-speed access to training datasets, model checkpoints, and other large-scale data required for AI workloads.

The scale-out architecture is particularly vital in the era of increasingly large and complex language models (LLMs). As these models increase in size and capability, they demand immense computational resources and high-bandwidth, low-latency communication between nodes and rapid access to vast datasets. The integrated network fabric ensures seamless data movement and model parallelism, while the storage subsystem provides the necessary throughput to feed these data-hungry models efficiently.

Small Setup Reference – POC/EVAL

Small setups are predominantly used for Proof of Concept (POC) or System Evaluation (EVAL) environments. No additional networking switch is required for single-system deployments, which are common in early-stage evaluations. It offers an ideal platform to evaluate the compatibility and performance of various AI frameworks, libraries, and custom applications with the Gaudi 3 architecture.

When scaling to small clusters of two to four systems, which is characteristic of more advanced POCs, the architecture incorporates one switch: usually 32 ports. This configuration is ideal for initiating scale-out scenario testing; it offers the necessary connectivity and bandwidth to support inter-system communication for distributed AI workloads while maintaining a balance in system performance.

For POCs, the built-in local storage of the Gaudi 3 systems typically suffices, eliminating the need for additional or external storage solutions. These setups provide a flexible and manageable environment for experimenting with various AI models and workloads, allowing teams to gather valuable insights before committing to larger-scale deployments.

Supermicro Gaudi 3 MegaPod

For larger setups or production environments, a MegaPod architecture is required. A MegaPod consists of three racks with 8 Gaudi nodes connected to three 800Gbps leaf switches, forming the first level of the network hierarchy. This Lego-like solution encompasses compute and networking resources, providing an efficient and easily manageable AI infrastructure.

To achieve an optimal balance between power and density, a MegaPod is distributed across three racks:
- Two racks house three Gaudi 3 servers each.
- A third rack accommodates the remaining 2 Gaudi servers, 3 leaf switches, and additional systems and equipment.

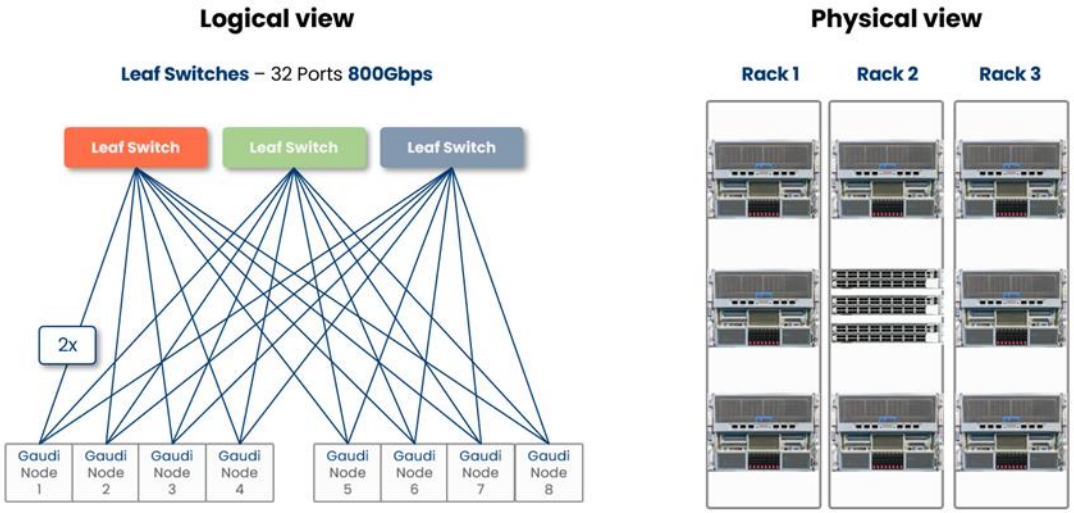

Figure 7 - Supermicro Gaudi MegaPod; Logical and Physical View

For simplicity, the MegaPod diagrams above focus on the Gaudi accelerators and leaf switches within the racks. With this configuration, on average, each MegaPod consumes 110kW. However, a complete architecture would also include essential components such as spine switches, storage systems, and a management network, as required for the business deployment.

Massive Scale-Out As we move into larger deployments, ranging from 32 to 2,048 systems, a transition occurs to leaf/spine network architecture. In this setup, we used the Supermicro MegaPods as building block units to scale out the computing performance.

In large deployments, the spine layer then acts as the second level, interconnecting these MegaPods and facilitating efficient communication across the entire cluster. This architecture allows for impressive scalability, with the number of leaf and spine switches increasing proportionally to the size of the deployment.

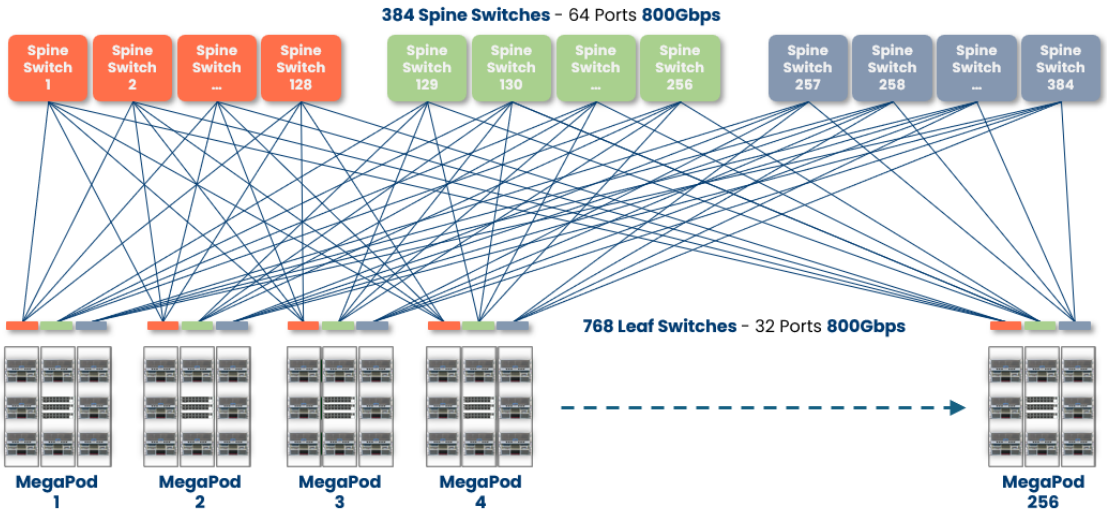

Figure 8 - Leaf/Spine Switch Reference Architecture

As deployments of Gaudi 3 Accelerators scale out, the network infrastructure grows proportionally to support increased computational capacity. The network architecture, employing the two-tier topology, uses 32-port for Leaf Switches and 64-port for Spine Switches. This design enables efficient data flow and inter-node communication across a wide range of configurations. In the largest deployment, featuring 256 Supermicro Gaudi 3 MegaPods, the network expands to incorporate 768 32-port Leaf Switches and 384 64-port Spine Switches.

Compute Unit			Network Unit	
Gaudi 3 Accelerators	Gaudi 3 Servers	Supermicro MegaPods	32-port Leaf Switches	64-port Spine Switches
64	8	1	3	--
128	16	2	6	3
256	32	4	12	6
512	64	8	24	12
1024	128	16	48	24
2048	256	32	96	48
4096	512	64	192	96
8192	1024	128	384	192
16384	2048	256	768	384

Table 2 - Intel Gaudi 3 AI Compute Switching Network Reference

The comprehensive Gaudi network architecture enables organizations to start with smaller Supermicro Gaudi 3 System deployments, commonly as POCs, and seamlessly scale up to massive clusters, all while maintaining high-bandwidth, low-latency connectivity.

Networking Switches

Supermicro offers a range of network switch options to provide flexibility and choice when deploying the Gaudi 3 AI Server platform. Customers can select from Supermicro's in-house switch solutions or validated third-party options, ensuring the networking infrastructure aligns with their specific requirements and preferences.

Two choices stand out for Intel's Gaudi 3 AI compute network: Supermicro's SSE-T8032S and SSE-T8164S/SR switches. Supermicro's SSE-T8032S 32-port 800GbE switch will be an ideal choice for the leaf network, suited for high-performance applications and large-scale cluster requirements. It comes with twin ports of 2x400G with 32 physical OSFP ports, offering a dense 32x800G in just a 1U form factor. On the other hand, the SSE-T8164S/SR offers a 64x800G in a 2U form factor, ideal for the spine networking.

For storage and management, Supermicro also offers a range of switches designed for various use cases in data centers and AI clusters. More about the storage systems will be discussed in the next section.

Network Solution	Supermicro Switch	Ports x Speeds	Ideal For
AI Compute	SSE-T8164S	64x800G OSFP	Spine
	SSE-T8032S	32x800G (64*400G) OSFP	Leaf
Storage	SSE-T8032S	32x800G (64*400G) OSFP	All-Flash
	SSE-F3548S/SR	48x25G,100G	Tiering to Object Storage
Management	SSE-X3348TR	48x10G,40G	OS Management
	SSE-G3748/R-SMIS	48x1G,10G,25G	IPMI/BMC

Table 3 - Supermicro Networking Reference for the Gaudi 3 System Solution

In addition to Supermicro's switch offerings, Supermicro partnered with industry-leading vendors like Arista and qualified their switches for seamless integration with Supermicro's Gaudi 3 AI Server. Arista's 7060DX5-64E and 7060PX5-64E switches are 800Gbps systems tested and validated to provide reliable, high-speed connectivity for Intel Gaudi-based systems.

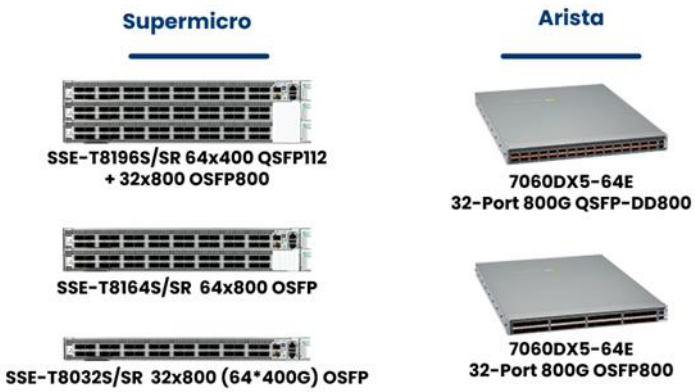

Figure 9 - Switching Options

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 Solution

There is no AI without data. This fundamental truth underscores the critical importance of storage infrastructure in the AI ecosystem, demanding solutions that are not only vast in capacity but also highly performant and intelligently managed to support the data-intensive nature of modern AI workloads. Supermicro's Scale-Out Storage Architecture offers a comprehensive solution for data-intensive applications, particularly those involving AI and machine learning. At scale, the solution is built on a three-tier architecture consisting of:

- Supermicro's Guadi 3 Server: The Application Tier for running workloads and consuming data.
- All-Flash Tier for high-performance, active data (10-20% of total data)
- Object Tier for long-term, capacity-optimized storage (80-90% of total data)

Supermicro provides a range of storage systems and software partners for each tier, allowing organizations to tailor the solution to their specific needs. The Application Tier, which is Supermicro's Intel Gaudi 3 AI Server, offers optimized AI/ML workloads, leveraging Intel accelerators, with high memory capacity and bandwidth for high-performance computing.

Figure 10 - Supermicro Gaudi 3 Storage Subsystems Reference: All-Flash and Object Store

For the All-Flash Tier, options range from 1U to 2U servers, supporting various NVMe drive configurations. This solution is supported by the Weka Data Platform, a distributed parallel file system that connects to the object storage. It supports clusters from eight to thousands of nodes and uses validated system configurations that offer several advantages.

The Object Tier uses high-capacity SuperStorage servers. There are multiple technology partners for Object Tier software that interoperate with the Supermicro architecture, including commercial and open-source options.

Storage Solutions (Tier)	Supermicro System	Software Options
Application	• SYS-822GA-NGR3 • (Supermicro Gaudi 3 AI Server)	• N/A
All-Flash	• SSG-122B-NE316R • SSG-121E-NES24R • SSG-221E-NE324R	• Weka • VAST
Object	• SSG-640SP-DE1CR90	• Quantum • OSNexus • Cloudian • Scality

Table 4 - Supermicro Storage Tiering Architecture Reference for AI Workloads

It is important to note that Supermicro collaborates with numerous technology partners in the storage domain, beyond those previously mentioned. This extensive ecosystem enables Supermicro to offer a comprehensive Software-Defined Storage solution tailored for AI workloads. By integrating diverse, cutting-edge storage technologies, Supermicro provides customers with flexible, scalable options that address the demanding storage requirements of modern AI applications, from high-performance file systems to scalable object storage.

AI infrastructure - Small-size reference architecture

When combining AI compute, networking, and storage, the overall solution is built to scale in terms of both capacity and performance. Organizations can start small and expand as their needs grow. A complete AI infrastructure must be designed for high performance, high bandwidth, and low latency—crucial elements for accelerating AI data pipelines.

Let us consider a typical small-scale AI infrastructure reference architecture comprised of 256 Gaudi 3 accelerators designed to support high-performance AI compute and storage demands.

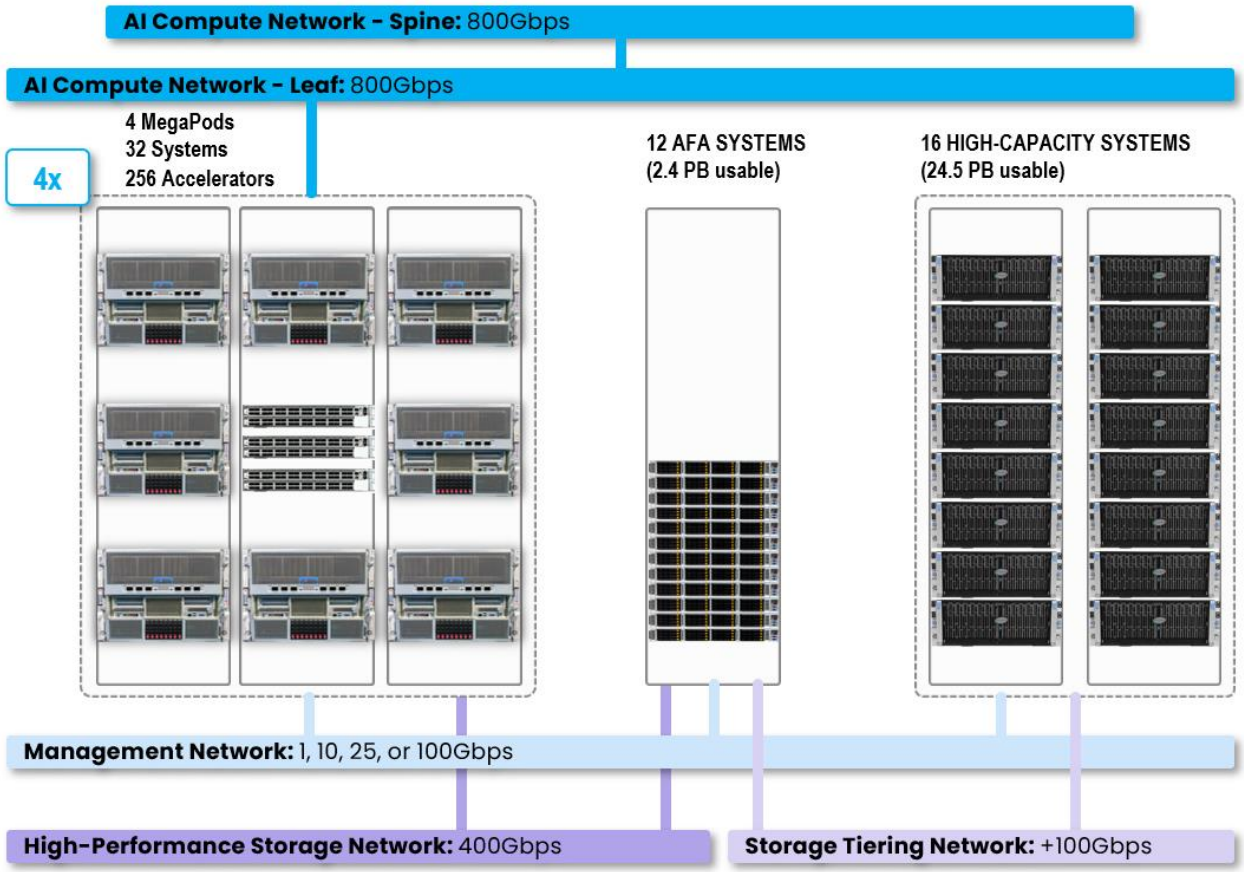

Figure 11 - Entire AI Solution

At the core of this infrastructure is the AI Compute Network, connecting 4 MegaPods for a total of 32 SYS-822GA-NGR3 systems. These MegaPods form the primary compute cluster, connected via an 800Gbps Leaf network using three SSE-T8032S/SR switches per MegaPod. The MegaPods interconnect using an 800Gbps Spine network, utilizing six SSE-T8164S/SR switches for this cluster configuration.

Complementing the MegaPods are the Storage clusters, connected to the compute systems through a separate Ethernet network. The high-performance storage tier consists of twelve SSG-122B-NE316R (All-Flash Array) subsystems, providing 2.4 PB of usable NVMe storage capacity to ensure optimal performance for data-intensive operations. These subsystems are integrated into the infrastructure via SSE-T8032S/SR 400Gbps Ethernet switches. SSG-640SP-DE1CR90 high-capacity systems are

employed for long-term object storage, offering a total of 24.5 PB usable capacity. This storage network utilizes 100Gbps SSE-F3548S/SR switches to facilitate efficient data movement between different storage tiers.

Management Networks further enhances this setup by operating at various speeds: 1Gbps for IPMI/BMC and over 1025Gbps for OS management. This comprehensive network architecture ensures robust connectivity, scalability, and optimized performance across all components of the AI infrastructure.

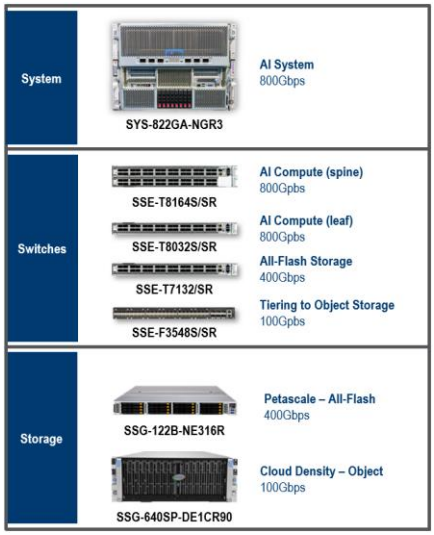

Figure 12 - Entire Hardware Stack

Intel Gaudi Software: Open-Source software stack

The Intel Gaudi Software Suite complements Supermicro's powerful hardware foundation, designed to work perfectly with the Gaudi 3 Systems. This software-hardware synergy is critical to unlocking the full potential of AI solutions, especially when dealing with large-scale models and datasets.

Enterprise Focused AI Software Stack

This software stack offers a comprehensive approach to enterprise-grade AI development and deployment, optimized for integration with high-performance hardware for the Supermicro Gaudi 3 System. The foundation is the embedded software layer, which interfaces advanced AI applications and hardware. It includes essential components such as BMC, Margin Tools, and Firmware, ensuring optimal performance and system stability.

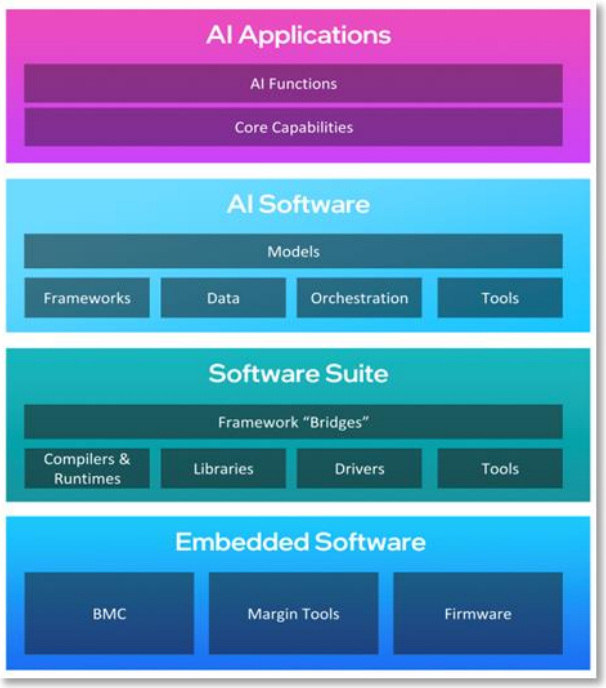

Figure 13 - Enterprise Focused AI Software Stack

The Software Suite layer builds upon this, bridging low-level hardware capabilities with high-level AI frameworks. It comprises framework bridge compilers, runtimes, libraries, drivers, and development tools. This suite efficiently translates complex AI algorithms into optimized instructions for AI accelerators.

Extensive Model & Framework Support

The platform supports a wide array of representative models, ranging from the widely used BERT and GPT-2 to more recent innovations like Llama, Mistral, and BLOOM. This diversity enables researchers and developers to work with state-of-the-art architectures, facilitating innovative AI development across various domains.

In terms of frameworks and libraries, the inclusion of PyTorch, DeepSpeed, and Hugging Face provides a robust foundation for AI development. PyTorch's popularity in research environments, combined with DeepSpeed's capabilities for training large models, offers a powerful toolkit. The integration of Hugging Face simplifies access to pre-trained models and promotes collaboration within the AI community.

The orchestration layer, featuring Kubernetes, Red Hat, and OpenShift, addresses the critical need for scalable and manageable AI deployments. This infrastructure support is essential for enterprises looking to operationalize AI at scale, ensuring efficient resource utilization and streamlined workflows.

TensorBoard, cnvrg.io, and RAY round out the ecosystem by providing essential model visualization, experiment tracking, and distributed computing capabilities. These tools are crucial for optimizing model performance and managing the complexities of large-scale AI projects.

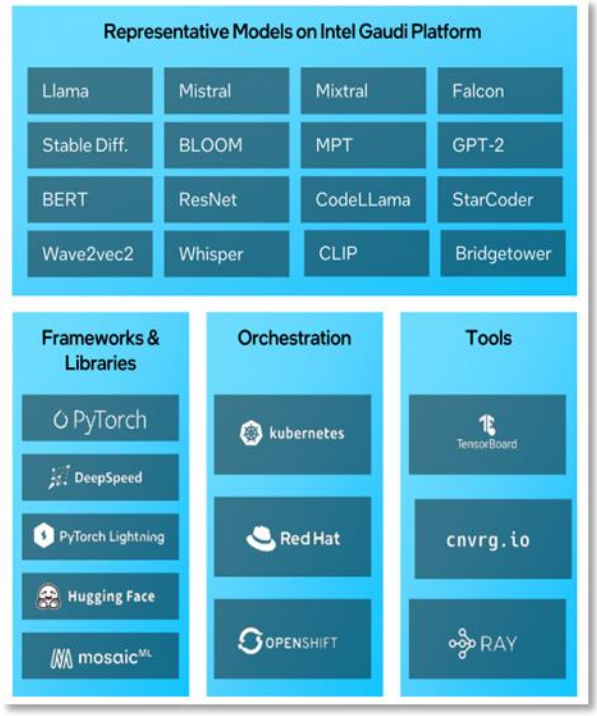

Figure 14 - Models for the Intel Gaudi Platform

Running AI models on the Intel Gaudi System

For Intel Gaudi environments, there are three primary methods to run models. First, Habana model references on GitHub offer a set of validated models with clear instructions for implementation. Second, the Hugging Face Optimum-Habana Library enables any Hugging Face transformer model to run on Gaudi. Lastly, the GPU Migration Toolkit facilitates the transition of existing models from GPU or other architectures to Gaudi.

Habana model references on GitHub

The model references provide a comprehensive library of optimized PyTorch models, serving as a great starting point for running models on Intel Gaudi and fostering innovation. Each model includes a detailed README with explicit instructions on running it, making it easy for users to follow the steps and execute any of the validated models.

The commands to run the PyTorch example model on various configurations, such as 1 HPU, 8 HPU, FP32, BF16, FP8, eager mode, or lazy mode, can be found in the Habana GitHub repository.

Hugging Face Optimum-Habana Library

Intel has collaborated with Hugging Face to create the Optimum Habana library, which enables any transformer model from Hugging Face to run on Gaudi. Customers can start with an existing library or examples, such as the run_glue.py script that runs the GLUE benchmark.

This script fine-tunes a BERT large model with the MRPC dataset and performs both training and evaluation for inference. When using multiple cards, the dataset is split across eight cards, resulting in significantly higher performance and throughput than single-card training.

GPU Migration Toolkit

The GPU Migration Toolkit converts existing models previously running on GPUs or other architectures to Intel Gaudi AI Accelerators. It achieves this by making changes at the Python level to the model code, mapping specific API calls from Python libraries and modules such as CUDA, Apex, and pynvml. The toolkit intercepts these API calls and changes them to run on Gaudi.

The GPU Migration Toolkit is preinstalled in the Habana Docker image, cutting the need for other installations. It also provides logging functionality to show what was changed and modified in the existing code, helping the migration process.

Using the GPU Migration Toolkit requires minimal steps. Follow the example below.

- First, start with the main Python file for the model, like main.py or train.py.
- Then, import the GPU Migration Toolkit, as shown in yellow in Figure 15.
- Finally, import the Habana PyTorch frameworks and then set two mark_steps here after the loss backward and optimizer steps that are part of the training loop.
- When the model is run, the GPU Migration Toolkit will analyze the GPU-based code and convert it into Gaudi-specific code.

```
import torch
import torch.nn as nn
import torch.optim as optim
import torch.nn.functional as F
import torchvision
import torchvision.transforms as transforms
import os

# Import the Intel Gaudi GPU Migration Toolkit Library
import habana_frameworks.torch.gpu_migration
import habana_frameworks.torch.core as htcore

# neural network model
class SimpleModel(nn.Module):
    ...

# training loop
def train(net, criterion, optimizer, trainloader, device):
    ...
    loss.backward()

    # API call to trigger execution
    htcore.mark_step()

    optimizer.step()

    # API call to trigger execution
    htcore.mark_step()
```

Figure 15 - GPU Migration Code Example

Management Infrastructure

Supermicro's infrastructure management capabilities are significantly enhanced by two key solutions: Supermicro Cloud Composer and Supermicro SuperCloud Orchestrator. Cloud Composer provides a unified interface for configuring, monitoring, and maintaining Supermicro Gaudi 3 Systems and storage infrastructure, offering streamlined operations through automated provisioning and real-time monitoring. This comprehensive management solution enables efficient resource allocation and system optimization, which is crucial for the demanding requirements of AI workloads.

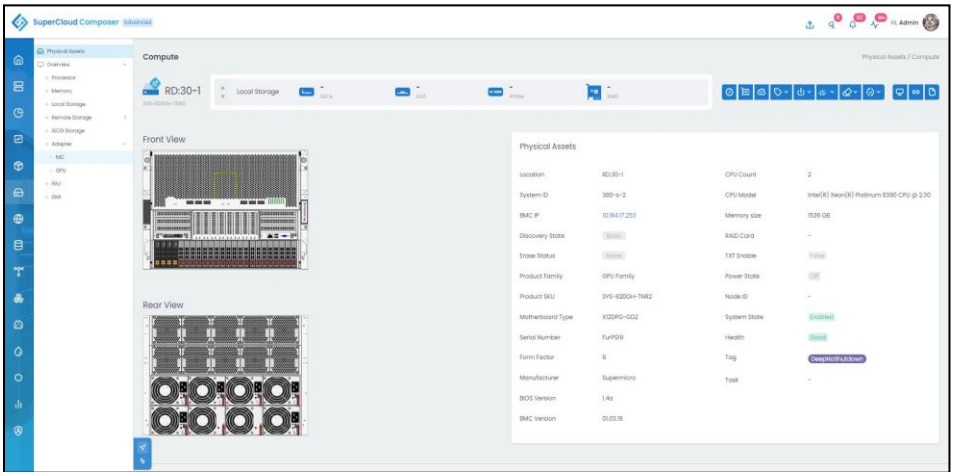

Figure 16 - SuperCloud Composer for Supermicro Gaudi 3 System

Complementing Cloud Composer, SuperCloud Orchestrator focuses on deploying and managing cloud environments, enabling efficient resource allocation and workload optimization across hybrid and multi-cloud setups. These tools offer a comprehensive management ecosystem, allowing IT teams to effectively oversee and optimize their AI infrastructure while maintaining flexibility through integration with industry-standard protocols and existing management tools.

The importance of these management solutions in bringing AI infrastructure from concept to deployment cannot be overstated. They bridge the gap between cutting-edge hardware capabilities and real-world operational requirements, ensuring that organizations can fully leverage the power of their Supermicro Gaudi 3 Systems. This system management solution is crucial for accelerating the deployment of AI solutions, reducing time-to-value, and maintaining optimal performance in production environments.

Conclusion

In response to these market dynamics, Supermicro continues its partnership with Intel to provide a cloud scale system and rack design with Intel Gaudi AI Accelerators, making GenAI and LLMs more accessible and performant. The Supermicro Gaudi 3 AI Server SYS-822GA-NGR3 offers a proven AI environment that combines optimized and high-performance hardware with a validated networking and storage solutions. The integration of this infrastructure enables organizations to rapidly deploy a fully optimized and supported platform for AI with exceptional performance in an open and flexible ecosystem.

References

https://www.supermicro.com/en/accelerators/intel

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

https://www.supermicro.com/en/products/system/ai/8u/sys-822ga-ngr3

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.