PRODUCT BRIEF

SUPERMICRO AND INTEL GAUDI 3 SYSTEMS ADVANCE ENTERPRISE AI INFRASTRUCTURE

High-Bandwidth AI System Using Intel Xeon 6 Processors for Efficient LLM and GenAI Training and Inference Across Enterprise Scales

TABLE OF CONTENTS

Executive Summary

Supermicro introduces the X14 Gaudi 3 AI System (SYS-822GA-NGR3), a high-performance platform designed to revolutionize large-scale AI model training and inferencing. This cutting-edge system combines the power of two Intel Xeon 6 CPUs (6900-Series with P-Cores) and eight Gaudi 3 AI accelerators, creating a powerhouse for demanding AI workloads.

The rapid advancement of Large Language Models (LLMs) and Generative AI (GenAI) has intensified the need for powerful, efficient computing solutions. These AI models, which can contain billions of parameters and require massive datasets for training, have become critical for applications ranging from natural language processing to content generation. However, their computational demands present significant challenges for many organizations.

The Supermicro X14 Gaudi 3 AI System addresses these challenges by providing the necessary computational power while maintaining cost-effectiveness. It enables businesses to train and deploy sophisticated AI models, including LLMs and GenAI applications, without disproportionate increases in infrastructure costs. The system's architecture, featuring high-bandwidth connectivity and scalability, allows organizations to expand their AI capabilities from initial projects to enterprise-wide implementations, adapting to the evolving demands of AI technology.

Supermicro X14 Gaudi 3 Systems – Specifications

Gaudi 3 AI Accelerators: The Supermicro X14 Gaudi 3 AI System is powered by eight Intel Gaudi 3 AI accelerators. These specialized processors are fundamental to the system's AI capabilities, providing the computational power necessary for advanced AI workloads. The Gaudi accelerators' unique features and performance characteristics significantly contribute to the system's overall efficiency in AI tasks.

Processor and Accelerator Configuration: The system harnesses the combined power of two Intel Xeon 6 CPUs (6900-Series with P-Cores) and eight Gaudi 3 AI accelerators. This robust configuration creates a powerhouse capable of handling the most demanding AI workloads, from large-scale model training to efficient inferencing.

Thermal Design and Management: Engineered as an 8U air-cooled system, the SYS-822GA-NGR3 demonstrates impressive thermal management capabilities. It easily handles the 900W TDP per chip, showcasing Supermicro's expertise in designing high-performance, thermally efficient systems for intensive AI computations.

Networking Prowess: One of the standout features is the system's networking capability. With onboard 6x 800GbE OSFP using open industry-standard Ethernet, the SYS-822GA-NGR3 excels in scale-out operations. This high-bandwidth connectivity facilitates rapid data transfer and efficient communication, enabling seamless scaling from eight accelerators to thousands in large-scale AI deployments.

Storage Options: The system is equipped with 8x NVMe hot-swap 2.5" drive bays for local storage needs. While not intended for large-scale data handling, this configuration provides ample space for operating system installation, local caching, and small-scale AI project experimentation.

Power Efficiency: Each Gaudi system consumes an average of 13kW, with the actual consumption varying based on specific components and system configuration. This power profile reflects a balance between high performance and energy efficiency, which is crucial for data centers managing large-scale AI operations.

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 onboard OSFP 800GbE ports for scale-out
Expansion Slots	2 PCIe 5.0 x16 (FHHL) + 2 PCIe 5.0 x8 (FHHL)

Table 1 – Supermicro X14 Gaudi 3 Servers Specifications

The Heart of the System: Intel Gaudi 3 AI Accelerator

The core of the Supermicro X14 Gaudi 3 AI System is the Intel Gaudi 3 AI accelerator (HL-325L), which complements the advanced Intel Xeon 6900P-series processor. This high-performance mezzanine card, built on 5nm process technology, features 8 MME engines, 64 programmable Tensor Processor Cores, 128GB of HBM2E memory, and 96MB of SRAM.

The Gaudi 3 incorporates Intel's fully programmable Tensor Processor Core (TPC) and GEMM Engine, supporting advanced AI data types, including FP8, BF16, FP16, TF32, and FP32. This versatility allows the system to adapt to various AI workloads, from training complex models to running efficient inference tasks.

Memory performance is a critical factor in AI computations, and the Gaudi 3 sets a new standard with its impressive 128GB HBM2E capacity, offering a total throughput of 3.7TB/s. Its advanced HBM controller is optimized for both random and linear access patterns, ensuring efficient data handling across various AI applications and addressing the challenge of data bottlenecks in AI processing.

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 provides unmatched scalability with 9.6 Terabits per second bi-directional networking capacity, allowing the SYS-822GA-NGR3 to use standard Ethernet technology for massive data center scalability. This feature directly addresses the scalability challenges faced by organizations as they grow their AI initiatives.

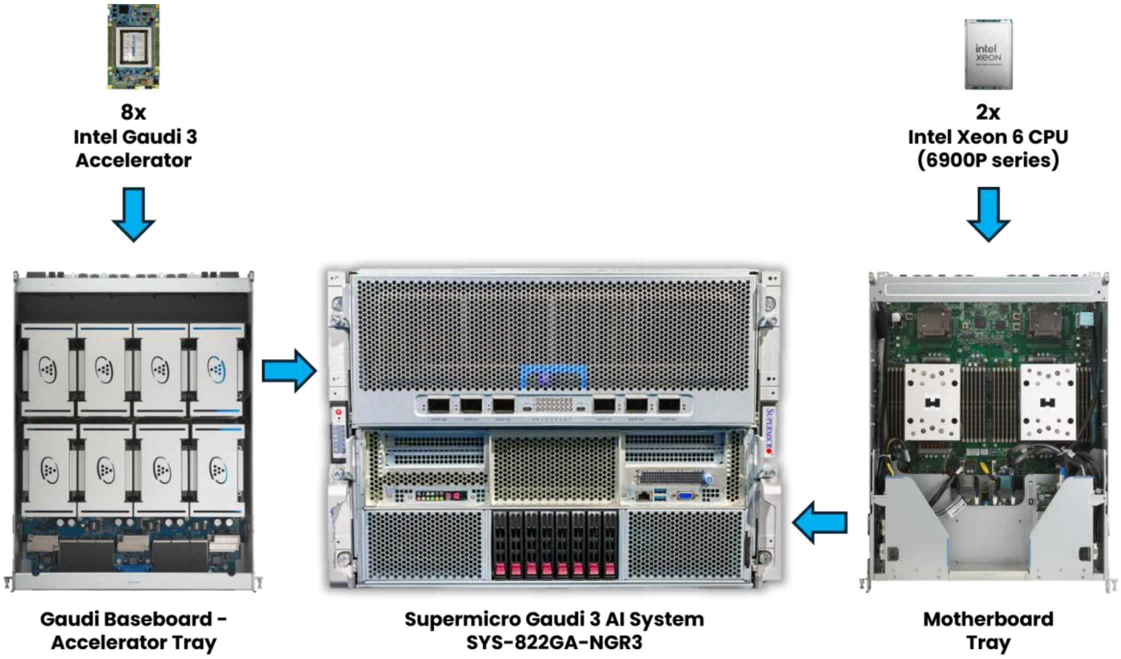

Figure 1 - Supermicro Gaudi 3 Systems: Key Components

The Supermicro Intel Gaudi Advantage

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. This expertise translates into optimized configurations and superior support for customers implementing AI solutions.

Furthermore, Supermicro is the only vendor offering the latest Intel Xeon 6 CPU with Xeon 6900P-series configuration. The advantages of the Xeon 6900P-series processor are particularly beneficial for AI workloads. The significantly higher core count (up to 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, addressing the challenge of handling the massive amounts of data required to train advanced AI models.

Figure 2 - Intel Xeon 6 Series processors with P-Cores

The improved memory capabilities of the Xeon 6900P-series, including faster speeds (DDR5 6400 MHz, MRDIMM 8800MHz), enable quicker data access and transfer. This memory performance is essential for feeding data to GPUs in AI training scenarios without bottlenecks. Likewise, the increased PCIe lane count and CXL 2.0 support enhance connectivity with GPUs and high-speed storage, allowing for more efficient data movement in complex AI systems.

Generational Leap: Gaudi 3's Performance Breakthroughs

As AI adoption accelerates, organizations face the dual challenge of meeting immense computational demands while adhering to budget constraints. The Supermicro X14 Gaudi 3 AI System addresses this challenge head-on, offering a substantial performance leap over its predecessor while maintaining a cost-effective price point. This generational advancement enables organizations to tackle more complex AI workloads without proportionally increasing their infrastructure investments.

The Gaudi 3 accelerator marks a significant evolution in AI compute capabilities, delivering substantial performance improvements over its predecessor across key metrics:

- 4x increase in AI compute performance for BF16 operations.
- 2x increase in AI compute for FP8 operations.
- 1.5x increase in memory bandwidth.
- 2x improvement in networking bandwidth, enabling massive system scale-out.

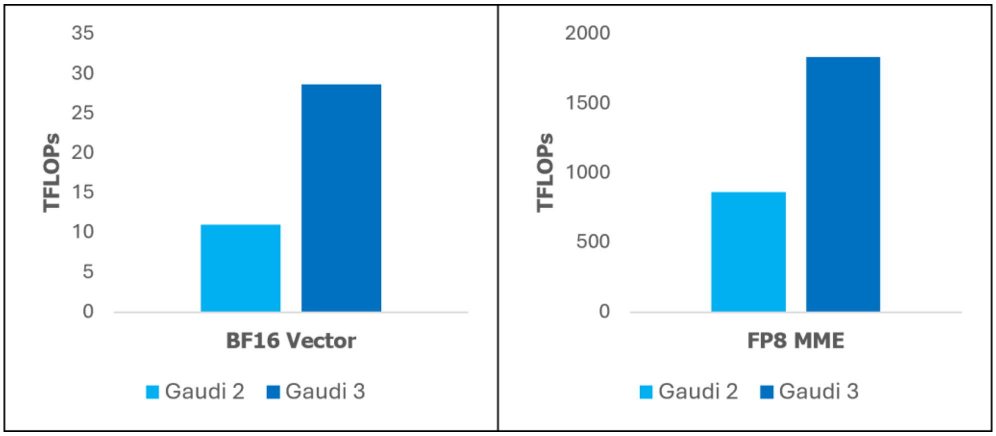

Figure 3 – Gaudi 3 to Gaudi 2 Comparison – Relative Performance

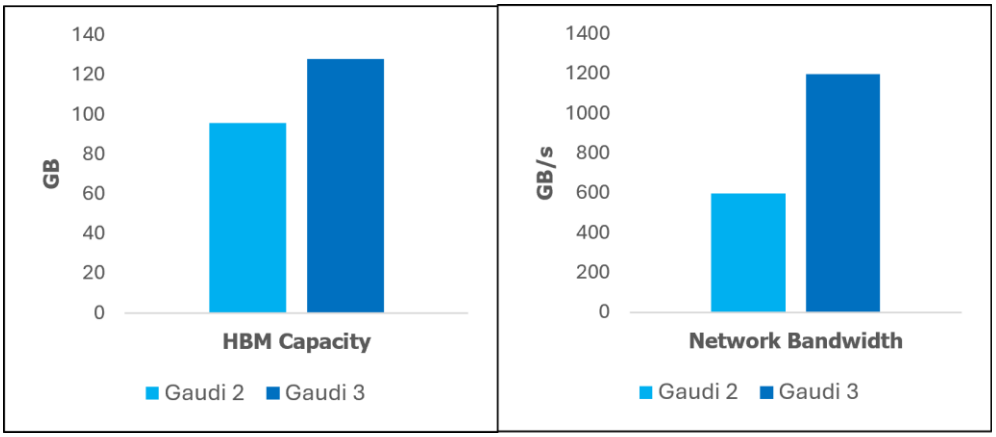

Figure 4 – Gaudi 3 to Gaudi 2 Comparison – Memory and Networking Bandwidth

These improvements translate into tangible benefits for AI workloads, empowering organizations to tackle more complex AI challenges. Particularly in training and inference tasks for popular large language models (LLMs) and multimodal models, organizations can now handle larger models and more extensive datasets with unprecedented efficiency without a proportional increase in infrastructure costs.

The leap in performance also accelerates time-to-insight and opens up new possibilities for AI applications across various industries. By providing enhanced compute power, improved memory bandwidth, and advanced networking capabilities, the Gaudi 3 positions businesses at the forefront of AI innovation, ready to exploit the full potential of cutting-edge AI technologies.

Use Cases and Applications

The Supermicro X14 Gaudi 3 AI System's impact extends across various industries, revolutionizing AI applications and research.

Enhancing Large Language Models (LLMs)

The system's 4x increase in AI compute performance for BF16 operations, and 2x increase for FP8 workloads significantly advance natural language processing and generative AI. This enhanced capability enables efficient training and running of larger, more complex language models. For instance, it can accelerate the development of models like GPT-3, BERT, and their variants, powering more sophisticated chatbots, real-time translation services, and AI-driven content generation tools.

Accelerating AI Research and Development

With its 1.5x increase in memory bandwidth and 2x improvement in networking bandwidth, the system empowers researchers to tackle more complex problems in AI. This faster memory bandwidth translates to faster progress in areas such as climate modeling, drug discovery, and computer vision. For example, researchers can now process larger datasets for more accurate weather predictions or simulate complex molecular interactions for pharmaceutical development at unprecedented speeds.

Transforming Industries with AI Solutions

The Gaudi 3's versatility makes it a powerful tool across multiple sectors. It accelerates diagnostic algorithms in healthcare and enables more sophisticated personalized medicine approaches, potentially improving early disease detection and patient care. The finance sector benefits from enhanced real-time fraud detection and more complex algorithmic trading models, such as those used by high-frequency trading firms. Other industries, from manufacturing to cybersecurity, can leverage the system's capabilities for optimizing operations and improving threat detection.

Summary: Empowering the Future of AI

The Supermicro X14 Gaudi 3 AI System represents a significant leap forward in AI infrastructure. By combining the power of Intel Xeon 6 CPUs with the advanced capabilities of the Gaudi 3 AI accelerators, Supermicro has created a solution that addresses the key challenges of enterprise AI adoption: performance, scalability, and cost-effectiveness.

This system offers the computational power needed to handle complex AI workloads, the scalability to grow with an organization's needs, and the efficiency to optimize the performance-to-cost ratio. Whether an organization is just beginning its AI journey or looking to expand its existing capabilities, the Supermicro X14 Gaudi 3 AI System provides a robust, flexible, and future-proof foundation.

Further Information

For more information on how the Supermicro X14 Gaudi 3 AI System can address your organization's AI infrastructure needs, please contact your Supermicro representative. Our team of experts is ready to help you navigate the complex landscape of AI hardware and software, ensuring you have the optimal solution for your unique requirements.

Supermicro X14 Gaudi 3 Detailed Specs

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

INTEL

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, go to newsroom.intel.com and www.intel.com