ACCELERATING AI COMPUTE WITH SUPERMICRO SERVERS IN THE INTEL® DEVELOPER CLOUD

Supermicro Advanced AI Servers featuring Intel® Xeon® Processors and Intel® Gaudi® 2 AI Accelerators Bring High-Performance, High-efficiency AI Cloud Compute, Training, and Inferencing to Developers and Enterprises

Supermicro SYS-820GH-TNR2

Intel® Gaudi® 2 AI Accelerator

INDUSTRY

Cloud Service Technology Provider

Introduction

Intel is a world leader in producing leading hardware and software technologies that underpin today's data centers, workstations, and personal computers. Within Intel's AI solution portfolio, the company recently developed a new generation of accelerators to address industry demand for high-performance, high-efficiency GenAI training and inference. The Intel® Gaudi® 2 AI accelerator delivers exceptional AI performance with low power usage while performing training and inference tasks.

The Intel® Developer Cloud provides enterprise companies and developers with a flexible, scalable AI infrastructure to accelerate model training and inference workloads. The cloud features systems with CPUs and a choice of accelerators. The Supermicro servers provide on-demand instances based on Intel® Xeon® processors and the Intel® Gaudi® 2 accelerator. They are designed to give developers easy access to develop, optimize, test, and deploy their models and applications in production environments. One of the highlights of this cloud is the large cluster of installed Supermicro servers, which allows developers to experience the advantages of a high-performance AI system.

CHALLENGES

- High Performance Servers
- Lower Power Use Required For AI Training

Challenges

When designing the Intel Developer Cloud, Intel needed to employ server hardware that would allow for the most demanding AI applications to be executed. Intel needed a significant number of servers enabling developers to access AI compute at scale. The systems needed to run demanding GenAI training and inference workloads and be scalable so that multiple systems could work together to train or deploy the largest AI models. The servers required support for up to 8TB of

memory and to utilize Xeon's advanced AI capabilities such as Intel® Advanced Matrix Extensions, Intel® In-Memory Analytics Accelerator, and others.

Solution

Intel selected the Supermicro SYS-820GH-TNR2 servers, which contain dual 4th Gen Intel Xeon Scalable processors and 2TB of memory. Each system features 8 Gaudi2 AI accelerators for training and deploying the largest models. Manufactured in 7nm process technology, each Intel Gaudi 2 accelerator integrates 24 100 Gb Ethernet network ports on the chip. In addition, the Intel Gaudi accelerator compute engine features 24 AI-custom Tensor Processor Cores, dual matrix multiplication engines, 96GB of HBM2E memory, and 48MB of SRAM.

SOLUTION

Supermicro 8U GPU Server

- Dual 4th Gen Intel Xeon Scalable Processors
- 8x Intel® Gaudi® 2 Accelerators
- 2TB DDR4 Memory
- 6x 400 GbE QSFP Connectors

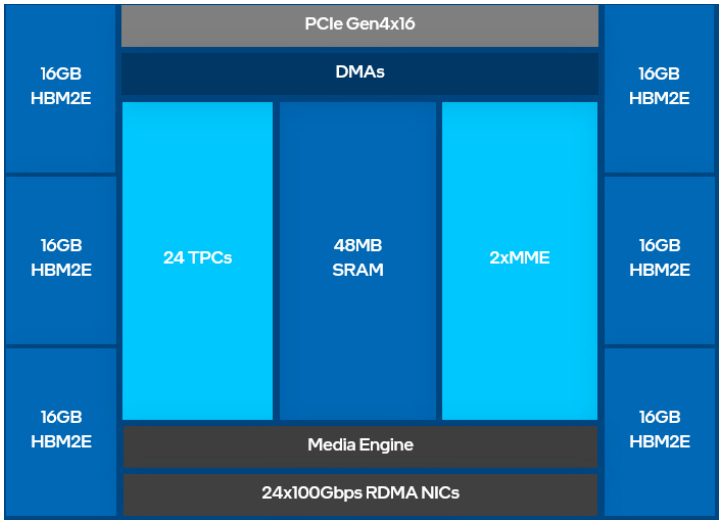

Figure 1 - Intel Gaudi 2 Accelerator Architecture Diagram

Scaling Systems with Intel Gaudi 2 AI Accelerator

Integrating industry-standard Ethernet on the Intel Gaudi 2 accelerator enables flexible and efficient scale-up and scale-out from one node to thousands to meet the scale demands of today's Generative AI systems. Intel integrated the new Supermicro AI servers into its Intel Developer Cloud quickly and efficiently, using open standard 400 Gigabit Ethernet switching throughout the system. The platform's scalability allows developers and production users to scale their AI training to high levels with a simple Ethernet connection.

Supermicro tested the MLPerf v3.0 BERT benchmark using 1 to 8 Gaudi2 based servers. When using up to 8 servers, the performance was very close to linear, demonstrating excellent scaling architecture. Supermicro servers based on Intel Gaudi 2 accelerators demonstrated strong competitive performance on both the MLPerf training and inference benchmarks.

Supermicro Servers with Intel Gaudi 2 Accelerators

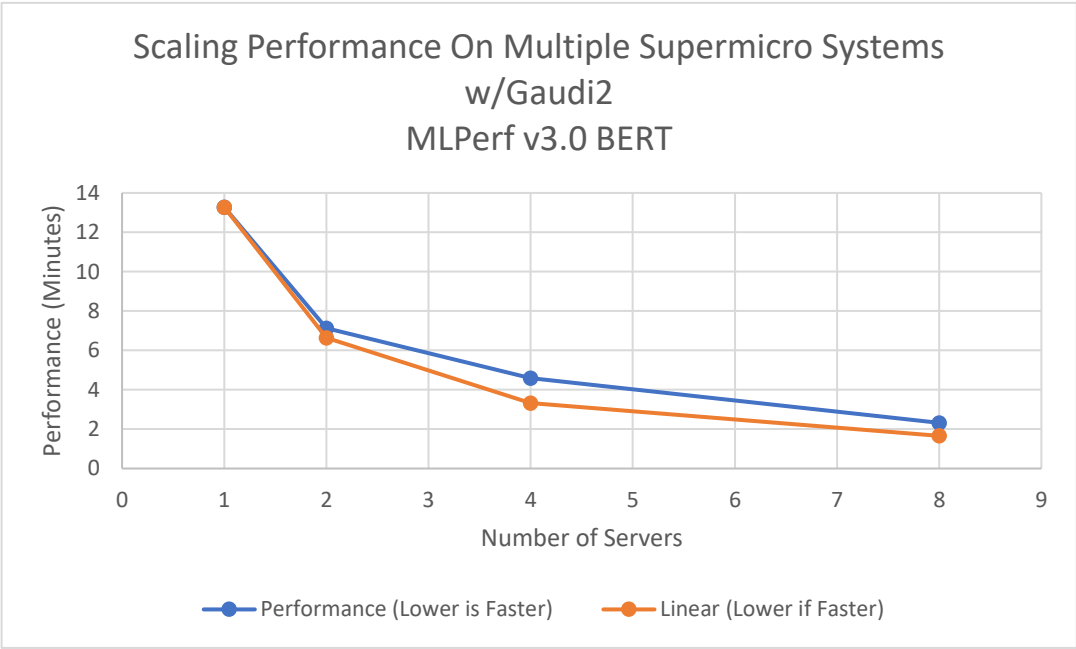

Figure 2 - Performance Gains Using Multiple Servers with Intel Gaudi 2

BENEFITS

- High Performance for time-to-train and inference latency
- Efficient scalability based on industry-standard network fabric
- Development on open source software—PyTorch and Hugging Face

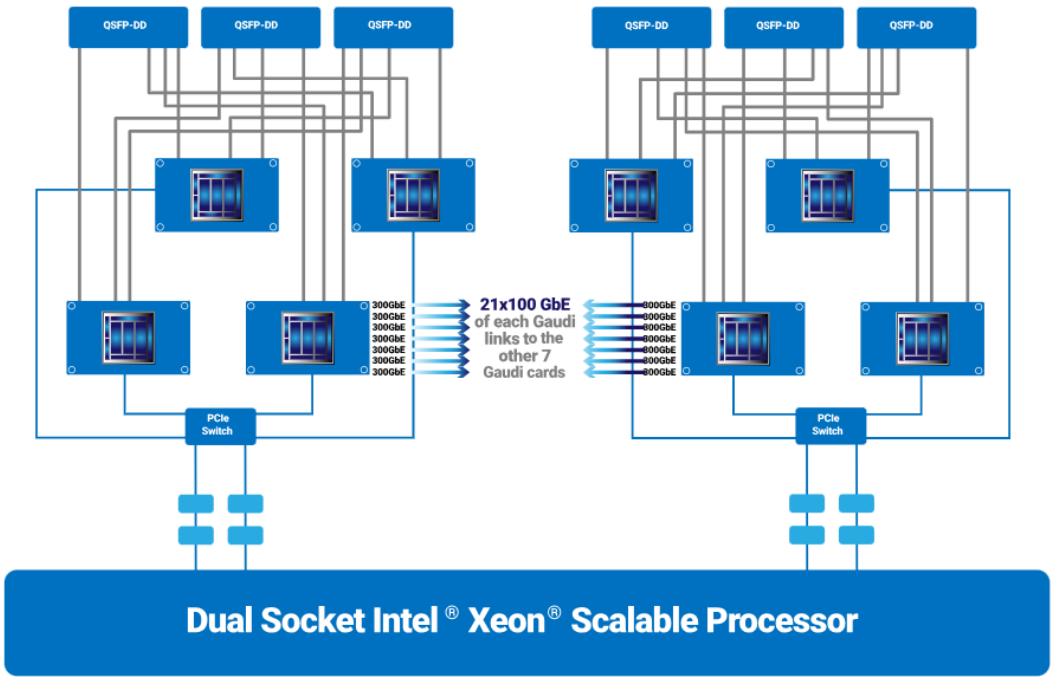

Figure 3 - Intel Gaudi 2 Accelerator Server Diagram

"We chose Supermicro servers based on Xeon processors and Intel Gaudi 2 AI accelerators for the Intel Developer Cloud because of the platform's strong performance, efficiency, and scalability for AI usage. Our users benefit from this state-of-the art platform with ease of development on industry-leading open source models and flexible scale-out. We continue working with Supermicro to offer even larger scaling options and look forward to the next generation of servers for the Intel Developer Cloud."

- Markus Flierl, Corporate Vice President of Intel Developer Cloud

For More Information:

Supermicro GPU Server:
https://www.supermicro.com/en/products/system/ai/8u/sys-820gh-tnr2

Intel Developer Cloud:
https://www.intel.com/content/www/us/en/developer/tools/devcloud/overview.html

SUPERMICRO

Supermicro is a global leader in high performance, green computing server technology and innovation. We provide our global customers with application-optimized servers and workstations customized with blade, storage, and GPU solutions. Our products offer proven reliability, superior design, and one of the industry's broadest array of product configurations, to fit all computational need.

Visit https://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 www.intel.com