PRODUCT BRIEF

ACCELERATED COMPUTE INFRASTRUCTURE FOR AI AND HPC WORKLOADS WITH SUPERMICRO SOLUTIONS FEATURING THE NVIDIA® H100 GPU

Supermicro Infrastructure for Demanding AI/ML and HPC Environments

TABLE OF CONTENTS

Executive Summary 1

Supermicro Product Lines for AI and HPC 2

NVIDIA H100 Highlights 3

Workload Matching 5

Additional Information and Resources 6

Executive Summary

Artificial Intelligence(AI) affects many aspects of daily life, whether we know it or not. This advancement of AI technologies, both in hardware and software, continues to evolve, with new and more enabling products being released constantly. Underlying and supporting these new AI applications is the sophisticated accelerated compute infrastructure that powers systems, storage, and networking creating new user experiences in every part of the market. Central to the infrastructure is the latest generation of GPUs (the hardware at the core of many AI solutions) which advance what is possible for AI implementations and high-performance computing applications. For example, with the architecture of the latest GPUs, multiple numeric formats are available for specific types of calculations that can be executed in parallel for a tremendous performance increase.

The NVIDIA H100 Tensor Cores GPUs are a significant step forward in the evolution of GPUs. With the performance of up to 30 times[1] over the current NVIDIA A100 GPU, innovative AI applications that were previously not possible will now function perfectly, reducing AI training times significantly. In addition, HPC applications are anticipated to run up to 7X compared to previous generations using the NVIDIA A100 GPUs.

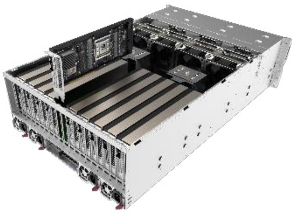
Supermicro Product Lines for AI and HPC Based on NVIDIA H100

Supermicro has a long history of providing the most advanced systems for accelerated compute infrastructure and HPC applications. In addition to incorporating the latest GPU technologies from NVIDIA, Supermicro offers GPUs in a wide range of designs optimized for specific workloads. By incorporating GPUs from the edge units to the very powerful data center GPU servers, applications can be placed closer to where data is generated, reducing latency and optimizing AI training and inferencing applications.

The Supermicro GPU Family of Servers, workstations, and edge units are designed and optimized to address engineering professionals requiring the latest CPU and GPU technologies. These no-compromise systems are designed for the latest CPUs and GPUs, with options for liquid cooling where needed. While many Supermicro servers can house a single or dual GPU per node, Supermicro's NVIDIA-certified GPU servers are designed for up to ten GPUs. These servers require a detailed engineering effort to ensure that the proper amount of cooling is designed into the system to handle the high-end thermal demands of GPUs.

Supermicro is offering the following servers, containing the NVIDIA H100 GPU and containing the latest processors from Intel or AMD.

- 4U 10GPU systems - SYUS-420GP-TNR and SYS-420GP-TNR2 with dual processors. The 10 GPU server is ideal for AI Training, Large Scale Metaverse implementations, and High-Performance Computing applications. The NVIDIA H100 GPUs are connected to the CPUs via the latest generation PCI-E bus.



- 4U Workstation – SYS-740GP-TNR – is ideal for professional workstations users in the areas of Scientific Visualization, High-Performance Computing, Rendering and AI Training, and Deep Learning

Comparison of Supermicro Systems initially available with the NVIDIA H100 GPU

Supermicro Server or WS	Max Memory	Max Number H100 GPUs (PCI-E)
4U 10GPU	2 TB	10
4U 4GPU	2 TB	4
Workstation - SYS-740GP-TNR	1 TB	4

NVIDIA H100 Highlights

The NVIDIA H100 is the latest GPU in a long line of optimized GPUs from NVIDIA. The NVIDIA H100 Tensor Core GPU, powered by the Hopper architecture, delivers a leap forward in accelerating AI and HPC workloads. In addition, the H100 secures data through its NVIDIA Confidential Computing capability. As a result, application developers can distribute and deploy their proprietary AI models at scale on a shared or remote infrastructure. The NVIDIA H100 specifically uses these advanced technologies:

- 80 GB of HBM3 memory, including a 50 MB of L2 cache – more data can be acted on. To fully utilize that compute performance, H100 contains HBM3 memory with a class-leading three terabytes per second (TB/sec) of memory bandwidth, a 50 percent increase over the previous generation. The combination of this faster HBM memory and larger cache provides the capacity to accelerate the most computationally intensive AI models.

- PCI-E-5.0 – Faster communication with the CPUs - H100 is the first GPU to support PCIe Gen5, providing the highest speeds possible at 128GB/s (bi-directional). This fast communication enables optimal connectivity with the highest performing CPUs and NVIDIA ConnectX-7 SmartNICs and BlueField-3 DPUs, allowing up to 400Gb/s Ethernet or NDR 400Gb/s InfiniBand networking acceleration for secure HPC and AI workloads.

- Up to 256 GPUs can be linked together with the NVIDIA NVLink Switch System. This technology accelerates everything from exascale scale workloads with a dedicated Transformer Engine for trillion parameter language models to right-sized Multi-Instance GPU (MIG) partitions.

From NVIDIA, in describing the NVIDIA H100.

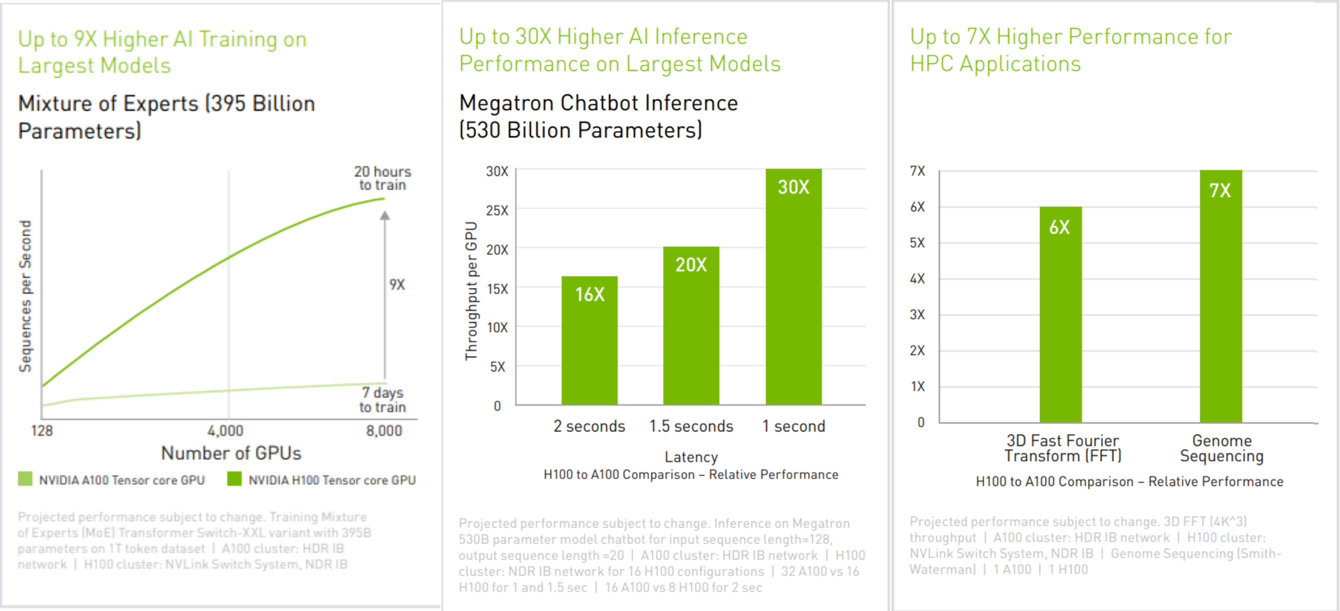
> *"An Order-of-Magnitude Leap for Accelerated Computing The NVIDIA H100 Tensor Core GPU delivers unprecedented performance, scalability, and security for every workload. With NVIDIA® NVLink® Switch System, up to 256 H100 GPUs can be connected to accelerate exascale workloads, while the dedicated Transformer Engine supports trillion parameter language models. H100 uses breakthrough innovations in the NVIDIA Hopper™ architecture to deliver industry-leading conversational AI, speeding up large language models by 30X over the previous generation."*

The following table shows the raw performance of the NVIDIA H100 GPU:

Form Factor	H100 PCI-E
FP64	24 Teraflops
FP64 Tensor Core	48 Teraflops
FP32	48 Teraflops
TF32 Tensor Core	800 Teraflops *
BFLOAT16 Tensor Core	1,600 Teraflops *
FP16 Tensor Core	1,600 Teraflops *
FP8 Tensor Core	2,200 Teraflops *
INT8 Tensor Core	3,200 Teraflops *
GPU Memory	80GB
GPU Memory Bandwidth	2TB/s
Decoders	7 NVDEC , 7 JPEG
Max Thermal Design Power (TDP)	350W
Multi-Instance GPUs	Up to 7 at 10GB/sec
Form Factor	PCI-E Dual-slot Air-Cooled
Interconnect	NVLINK: 600GB/s PCI-E 5.0 128GB/s
Server Options	NVIDIA-Certifed with 1-8 GPUs

* Shown with sparsity. Specifications are one-half lower without sparsity.

From a performance view, the H100 is significantly faster than the NVIDIA A100 in several tasks and applications. The following charts (courtesy of NVIDIA) show how much faster the NVIDIA H100 is, compared to the NVIDIA A100.



Up to 9X Higher AI Training on Largest Models

Mixture of Experts (395 Billion Parameters)

Projected performance subject to change. Training Mixture of Experts (MoE) Transformer Switch-XXL variant with 395B parameters on 1T token dataset | A100 cluster: HDR IB network | H100 cluster: NVLink Switch System, NDR IB

Up to 30X Higher AI Inference Performance on Largest Models

Megatron Chatbot Inference (530 Billion Parameters)

Projected performance subject to change. Inference on Megatron 530B parameter model chatbot for input sequence length=128, output sequence length =20 | A100 cluster: HDR IB network | H100 cluster: NDR IB network for 16 H100 configurations | 32 A100 vs 16 H100 for 1 and 1.5 sec | 16 A100 vs 8 H100 for 2 sec

Up to 7X Higher Performance for HPC Applications

7X
6X
5X
4X
3X
2X
1X
0
6X
7X
3D Fast Fourier Transform (FFT)
Genome Sequencing
H100 to A100 Comparison – Relative Performance

Projected performance subject to change. 3D FFT (4K^3) throughput | A100 cluster: HDR IB network | H100 cluster: NVLink Switch System, NDR IB | Genome Sequencing (Smith-Waterman) | 1 A100 | 1 H100

Figure 1 - AI Training on Various Workloads (H100 vs. A100) – Benchmarks courtesy of NVIDIA

Workload Matching

Several workloads will be optimized with different products from Supermicro. Below is a general matching of AI and HPC workloads that are optimized with the NVIDIA H100 GPU.

Supermicro Server/Workstation	AI Training	AI Inferencing	3D Metaverse Collaboration	HPC
4U 10GPU (PCI-E)	Large Models		Large Models	Large Simulations
4U 4 GPU (PCI-E)	Medium Models	Large Models	Medium Models	Medium Simulations
Workstation - SYS-740GP-TNR	Medium Models	Medium Models	Medium Models	Medium Simulations