Supermicro Introduces AI Inference-optimized New GPU Server with up to 20 NVIDIA Tesla T4 Accelerators in 4U

September 20, 2018

Inference-optimized system extends Supermicro's leading portfolio of GPU Servers to offer customers an unparalleled selection of AI solutions for Inference, Training, and Deep Learning including Singe-Root, Dual-Root, Scale-up and Scale Out designs

SAN JOSE, Calif., Sept. 20, 2018 /PRNewswire/ -- **Super Micro Computer, Inc.** (NASDAQ: SMCI), a global leader in enterprise computing, storage, networking solutions and green computing technology, today is introducing the latest additions to its extensive line of GPU-optimized servers.

Artificial intelligence (AI) is quickly becoming one of the most crucial components to business success now and in the foreseeable future. Today, the necessity of deploying powerful computing platforms that can accelerate and cost-effectively scale their AI-based products and services has become vital for successful enterprises.

Supermicro's new SuperServer 6049GP-TRT provides the superior performance required to accelerate the diverse applications of modern AI. For maximum GPU density and performance, this 4U server supports up to 20 NVIDIA® Tesla® T4 Tensor Core GPUs, three terabytes of memory, and 24 hot-swappable 3.5" drives. This system also features four 2000-watt Titanium level efficiency (2+2) redundant power supplies to help optimize the power efficiency, uptime and serviceability.

"Supermicro is innovating to address the rapidly emerging high-throughput inference market driven by technologies such as 5G, Smart Cities and IOT devices, which are generating huge amounts of data and require real-time decision making," said Charles Liang, president and CEO of Supermicro. "We see the combination of NVIDIA TensorRT and the new Turing architecture based T4 GPU Accelerator as the ideal combination for these new demanding and latency-sensitive workloads and are aggressively leveraging them in our GPU system product line."

"Enterprise customers will benefit from a dramatic boost in throughput and power efficiency from the NVIDIA Tesla T4 GPUs in Supermicro's new high-density servers," said Ian Buck, vice president and general manager of Accelerated Computing at NVIDIA. "With AI inference constituting an increasingly large portion of data center workloads, these Tesla T4 GPU platforms provide incredibly efficient real-time and batch inference."

Supermicro's performance-optimized 4U SuperServer 6049GP-TRT system can support up to 20 PCI-E NVIDIA Tesla T4 GPU accelerators, which dramatically increases the density of GPU server platforms for wide data center deployment supporting deep learning, inference applications. As more and more industries deploy artificial intelligence, they will be looking for high density servers optimized for inference. The 6049GP-TRT is the optimal platform to lead the transition from training deep learning, neural networks to deploying artificial intelligence into real world applications such as facial recognition and language translation.

Supermicro has an entire family of 4U GPU systems that support the ultra-efficient Tesla T4, which is designed to accelerate inference workloads in any scale-out server. The hardware accelerated transcode engine in Tesla T4 delivers multiple HD video streams in real-time and allows integrating deep learning into the video transcoding pipeline to enable a new class of smart video applications. As deep learning shapes our world like no other computing model in history, deeper and more complex neural networks are trained on exponentially larger volumes of data. To achieve responsiveness, these models are deployed on powerful Supermicro GPU servers to deliver maximum throughput for inference workloads.

For comprehensive information on Supermicro NVIDIA GPU system product lines, please go to https://www.supermicro.com/products/nfo/gpu.cfm.

Follow Supermicro on Facebook and Twitter to receive their latest news and announcements.

About Super Micro Computer, Inc. (NASDAQ: SMCI)
Supermicro® (NASDAQ: SMCI), the leading innovator in high-performance, high-efficiency server technology is a premier provider of advanced Server Building Block Solutions® for Data Center, Cloud Computing, Enterprise IT, Hadoop/Big Data, HPC and Embedded Systems worldwide. Supermicro is committed to protecting the environment through its "We Keep IT Green®" initiative and provides customers with the most energy-efficient, environmentally-friendly solutions available on the market.

SMCI-F

SOURCE Super Micro Computer, Inc.

Michael Kalodrich; Super Micro Computer, Inc.; michaelkalodrich@supermicro.com