



## Supermicro Expands Portfolio with Fully Integrated NVIDIA A100 GPU-Powered Systems Delivering 5 PetaFLOPS of AI Performance in a Single 4U Server

May 14, 2020

*Supermicro Boosts Performance on Industry's Broadest Portfolio of GPU Servers Optimized for AI, Deep Learning, Data Science, and Compute Workloads with Full Support for the New NVIDIA A100 GPUs, Third Generation NVIDIA NVLink, and NVIDIA NVSwitch*

SAN JOSE, Calif., May 14, 2020 /PRNewswire/ -- **Super Micro Computer, Inc.** (Nasdaq: SMCI), a global leader in enterprise computing, storage, networking solutions and green computing technology, today announced two new systems designed for artificial intelligence (AI) deep learning applications that fully leverage the third-generation NVIDIA HGX™ technology with the new NVIDIA A100™ Tensor Core GPUs as well as full support for the new NVIDIA A100 GPUs across the company's broad portfolio of 1U, 2U, 4U and 10U GPU servers. NVIDIA A100 is the first elastic, multi-instance GPU that unifies training, inference, HPC, and analytics.

"Expanding upon our industry-leading portfolio of GPU systems and NVIDIA HGX-2 system technology, Supermicro is introducing a new 2U system implementing the new NVIDIA HGX™ A100 4 GPU board (formerly codenamed Redstone) and a new 4U system based on the new NVIDIA HGX A100 8 GPU board (formerly codenamed Delta) delivering 5 PetaFLOPS of AI performance," said Charles Liang, CEO and president of Supermicro. "As GPU accelerated computing evolves and continues to transform data centers, Supermicro will provide customers the very latest system advancements to help them achieve maximum acceleration at every scale while optimizing GPU utilization. These new systems will significantly boost performance on all accelerated workloads for HPC, data analytics, deep learning training and deep learning inference."

As a balanced data center platform for HPC and AI applications, Supermicro's new 2U system leverages the NVIDIA HGX A100 4 GPU board with four direct-attached NVIDIA A100 Tensor Core GPUs using PCI-E 4.0 for maximum performance and NVIDIA NVLink™ for high-speed GPU-to-GPU interconnects. This advanced GPU system accelerates compute, networking and storage performance with support for one PCI-E 4.0 x8 and up to four PCI-E 4.0 x16 expansion slots for GPUDirect RDMA high-speed network cards and storage such as InfiniBand™ HDR™, which supports up to 200Gb per second bandwidth.

"AI models are exploding in complexity as they take on next-level challenges such as accurate conversational AI, deep recommender systems and personalized medicine," said Ian Buck, general manager and VP of accelerated computing at NVIDIA. "By implementing the NVIDIA HGX A100 platform into their new servers, Supermicro provides customers the powerful performance and massive scalability that enable researchers to train the most complex AI networks at unprecedented speed."

Optimized for AI and machine learning, Supermicro's new 4U system supports eight A100 Tensor Core GPUs. The 4U form factor with eight GPUs is ideal for customers that want to scale their deployment as their processing requirements expand. The new 4U system will have one NVIDIA HGX A100 8 GPU board with eight A100 GPUs all-to-all connected with NVIDIA NVSwitch™ for up to 600GB per second GPU-to-GPU bandwidth and eight expansion slots for GPUDirect RDMA high-speed network cards. Ideal for deep learning training, data centers can use this scale-up platform to create next-gen AI and maximize data scientists' productivity with support for ten x16 expansion slots.

Customers can expect a significant performance boost across Supermicro's extensive portfolio of 1U, 2U, 4U and 10U multi-GPU servers when they are equipped with the new NVIDIA A100 GPUs. For maximum acceleration, Supermicro's new A+ GPU system supports up to eight full-height double-wide (or single-wide) GPUs via direct-attach PCI-E 4.0 x16 CPU-to-GPU lanes without any PCI-E switch for the lowest latency and highest bandwidth. The system also supports up to three additional high-performance PCI-E 4.0 expansion slots for a variety of uses, including high-performance networking connectivity up to 100G. An additional AIOM slot supports a Supermicro AIOM card or an OCP 3.0 mezzanine card.

As the leader in AI system technology, Supermicro offers multi-GPU optimized thermal designs that provide the highest performance and reliability for AI, Deep Learning, and HPC applications. With 1U, 2U, 4U, and 10U rackmount GPU systems; Ultra, BigTwin™, and embedded systems supporting GPUs; as well as GPU blade modules for our 8U SuperBlade®, Supermicro offers the industry's widest and deepest selection of GPU systems to power applications from Edge to Cloud.

To deliver enhanced security and unprecedented performance at the edge, Supermicro plans to add the new NVIDIA EGX™ A100 configuration to its edge server portfolio. The EGX A100 converged accelerator combines an NVIDIA Mellanox SmartNIC with GPUs powered by the new NVIDIA Ampere architecture, so enterprises can run AI at the edge more securely.

For more on Supermicro's new GPU systems, please visit [www.supermicro.com/A100](http://www.supermicro.com/A100).

Follow Supermicro on [LinkedIn](#), [Twitter](#), and [Facebook](#) to receive their latest news and announcements.

### **About Super Micro Computer, Inc.**

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.

Supermicro, Server Building Block Solutions, BigTwin, SuperBlade, and We Keep IT Green are trademarks and/or registered trademarks of Super Micro Computer, Inc.

All other brands, names and trademarks are the property of their respective owners.

SMCI-F

 View original content to download multimedia:<http://www.prnewswire.com/news-releases/supermicro-expands-portfolio-with-fully-integrated-nvidia-a100-gpu-powered-systems-delivering-5-petaflops-of-ai-performance-in-a-single-4u-server-301059071.html>

SOURCE Super Micro Computer, Inc.

Michael Kalodrich, Super Micro Computer, Inc., PR@supermicro.com