Supermicro's New Scale-Up Artificial Intelligence and Machine Learning Systems with 8 NVIDIA Tesla V100 with NVLink GPUs Deliver Superior Performance and System Density

March 27, 2018

On display at GTC 2018, Supermicro GPU-optimized systems address market demand for 10x growth in deep learning, AI, and big data analytic applications with best-in-class features including NVIDIA Tesla V100 32GB with NVLink and maximum GPU density

SAN JOSE, Calif., March 27, 2018 /PRNewswire/ -- **Super Micro Computer, Inc.** (NASDAQ: SMCI), a global leader in enterprise computing, storage, networking solutions and green computing technology, today is showcasing the industry's broadest selection of GPU server platforms that support NVIDIA® Tesla® V100 PCI-E and V100 SXM2 GPU accelerators at the GPU Technology Conference in the San Jose McEnery Convention Center, booth 215, through March 29.

SUPERMICRO
Performance Maximized Machine Learning
4U 8 NVIDIA® Tesla® V100 32GB SXM2 GPUs
4U/Tower 4 NVIDIA® Tesla® V100 32GB PCI-E GPUs
1U 4 NVIDIA® Tesla® V100 32GB SXM2 GPUs

For maximum acceleration of highly parallel applications like artificial intelligence (AI), deep learning, self-driving cars, smart cities, health care, big data, HPC, virtual reality and more, Supermicro's new 4U system with next-generation NVIDIA NVLink™ interconnect technology is optimized for maximum performance. SuperServer 4029GP-TVRT supports eight NVIDIA Tesla V100 32GB SXM2 GPU accelerators with maximum GPU-to-GPU bandwidth for cluster and hyper-scale applications. Incorporating the latest NVIDIA NVLink technology with over five times the bandwidth of PCI-E 3.0, this system features independent GPU and CPU thermal zones to ensure uncompromised performance and stability under the most demanding workloads.

"On initial internal benchmark tests, our 4029GP-TVRT system was able to achieve 5,188 images per second on ResNet-50 and 3,709 images per second on InceptionV3 workloads," said Charles Liang, President and CEO of Supermicro. "We also see very impressive, almost linear performance increases when scaling to multiple systems using GPU Direct RDMA. With our latest innovations incorporating the new NVIDIA V100 32GB PCI-E and V100 32GB SXM2 GPUs with 2X memory in performance-optimized 1U and 4U systems with next-generation NVLink, our customers can accelerate their applications and innovations to help solve the world's most complex and challenging problems."

"Enterprise customers will benefit from a new level of computing efficiency with Supermicro's high-density servers optimized for NVIDIA Tesla V100 32GB data center GPUs," said Ian Buck, vice president and general manager of accelerated computing at NVIDIA. "Twice the memory with V100 drives up to 50 percent faster results on complex deep learning and scientific applications and improves developer productivity by reducing the need to optimize for memory."

"At Preferred Networks, we continue to leverage Supermicro's high-performance 4U GPU servers to successfully power our private supercomputers," said Ryosuke Okuta, CTO of Preferred Networks. "These state-of-the-art systems are already powering our current supercomputer applications, and we have already begun the process of deploying Supermicro's optimized new 4U GPU systems loaded with NVIDIA Tesla V100 32GB GPUs to drive our upcoming new private supercomputers."

Supermicro is also demonstrating the performance-optimized 4U SuperServer 4029GR-TRT2 system that can support up to 10 PCI-E NVIDIA Tesla V100 accelerators with Supermicro's innovative and GPU-optimized single root complex PCI-E design, which dramatically improves GPU peer-to-peer communication performance. For even greater density, the SuperServer 1029GQ-TRT supports up to four NVIDIA Tesla V100 PCI-E GPU accelerators in only 1U of rack space and the new SuperServer 1029GQ-TVRT supports four NVIDIA Tesla V100 SXM2 32GB GPU accelerators in 1U.

With the convergence of big data analytics and machine learning, the latest NVIDIA GPU architectures, and improved machine learning algorithms, deep learning applications require the processing power of multiple GPUs that must communicate efficiently and effectively to expand the GPU network. Supermicro's single-root GPU system allows multiple NVIDIA GPUs to communicate efficiently to minimize latency and maximize throughput as measured by the NCCL P2PBandwidthTest.

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

View original content with multimedia:http://www.prnewswire.com/news-releases/supermicros-new-scale-up-artificial-intelligence-and-machine-learning-systems-with-8-nvidia-tesla-v100-with-nvlink-gpus-deliver-superior-performance-and-system-density-300619830.html

SOURCE Super Micro Computer, Inc.

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