NVIDIA MGX Gives System Makers Modular Architecture to Meet Diverse Accelerated Computing Needs of World's Data Centers

QCT and Supermicro Among First to Use Server Spec Enabling 100+ System Configurations to Accelerate AI, HPC, Omniverse Workloads

COMPUTEX—To meet the diverse accelerated computing needs of the world's data centers, NVIDIA today unveiled the NVIDIA MGX™ server specification, which provides system manufacturers with a modular reference architecture to quickly and cost-effectively build more than 100 server variations to suit a wide range of AI, high performance computing and Omniverse applications.

ASRock Rack, ASUS, GIGABYTE, Pegatron, QCT and Supermicro will adopt MGX, which can slash development costs by up to three-quarters and reduce development time by two-thirds to just six months.

"Enterprises are seeking more accelerated computing options when architecting data centers that meet their specific business and application needs," said Kaustubh Sanghani, vice president of GPU products at NVIDIA. "We created MGX to help organizations bootstrap enterprise AI, while saving them significant amounts of time and money."

With MGX, manufacturers start with a basic system architecture optimized for accelerated computing for their server chassis, and then select their GPU, DPU and CPU. Design variations can address unique workloads, such as HPC, data science, large language models, edge computing, graphics and video, enterprise AI, and design and simulation. Multiple tasks like AI training and 5G can be handled on a single machine, while upgrades to future hardware generations can be frictionless. MGX can also be easily integrated into cloud and enterprise data centers.

Collaboration With Industry Leaders
QCT and Supermicro will be the first to market, with MGX designs appearing in August. Supermicro's ARS-221GL-NR system, announced today, will include the NVIDIA Grace™ CPU Superchip, while QCT's S74G-2U system, also announced today, will use the NVIDIA GH200 Grace Hopper Superchip.

Additionally, SoftBank Corp. plans to roll out multiple hyperscale data centers across Japan and use MGX to dynamically allocate GPU resources between generative AI and 5G applications.

"As generative AI permeates across business and consumer lifestyles, building the right infrastructure for the right cost is one of network operators' greatest challenges," said Junichi Miyakawa, president and CEO at SoftBank Corp. "We expect that NVIDIA MGX can tackle such challenges and allow for multi-use AI, 5G and more depending on real-time workload requirements."

Different Designs for Different Needs
Data centers increasingly need to meet requirements for both growing compute capabilities and decreasing carbon emissions to combat climate change, while also keeping costs down.

Accelerated computing servers from NVIDIA have long provided exceptional computing performance and energy efficiency. Now, the modular design of MGX gives system manufacturers the ability to more effectively meet each customer's unique budget, power delivery, thermal design and mechanical requirements.

Multiple Form Factors Offer Maximum Flexibility
MGX works with different form factors and is compatible with current and future generations of NVIDIA hardware, including:

- Chassis: 1U, 2U, 4U (air or liquid cooled)
- GPUs: Full NVIDIA GPU portfolio including the latest H100, L40, L4
- CPUs: NVIDIA Grace CPU Superchip, GH200 Grace Hopper Superchip, x86 CPUs
- Networking: NVIDIA BlueField®-3 DPU, ConnectX®-7 network adapters

MGX differs from NVIDIA HGX™ in that it offers flexible, multi-generational compatibility with NVIDIA products to ensure that system builders can reuse existing designs and easily adopt next-generation products without expensive redesigns. In contrast, HGX is based on an NVLink®-connected, multi-GPU baseboard tailored to scale to create the ultimate in AI and HPC systems.

Software to Drive Acceleration Further
In addition to hardware, MGX is supported by NVIDIA's full software stack, which enables developers and enterprises to

build and accelerate AI, HPC and other applications. This includes NVIDIA AI Enterprise, the software layer of the NVIDIA AI platform, which features over 100 frameworks, pretrained models and development tools to accelerate AI and data science for fully supported enterprise AI development and deployment.

MGX is compatible with the Open Compute Project and Electronic Industries Alliance server racks, for quick integration into enterprise and cloud data centers.

Watch NVIDIA founder and CEO Jensen Huang discuss the MGX server specification in his keynote address at COMPUTEX and learn more in the NVIDIA MGX architecture white paper.

About NVIDIA
Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI and is fueling the creation of the industrial metaverse. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, performance, features and availability of our products, collaborations, services and technologies, including the NVIDIA MGX server specification, Omniverse, NVIDIA GPUs including H100, L40 and L4, NVIDIA DPUs including BlueField-3, NVIDIA CPUs including x86 CPUs, NVIDIA HPCs, large language models, and edge computing, Grace CPU Superchip, GH200 Grace Hopper Superchip, chassis, ConnectX-7 network adapters, NVIDIA HGX, NVLink, NVIDIA AI Enterprise, and the NVIDIA AI platform; our collaborations with QCT, Supermicro, ASRock Rack, ASUS, GIGABYTE, Pegatron and SoftBank Corp., and the benefits, impact, performance and availability thereof; enterprises seeking more accelerated computing options when architecting data centers to meet their specific business and application needs; and data centers increasingly needing to meet requirements for growing compute capabilities and decreasing carbon emissions, while also keeping costs down are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

Kristin Uchiyama
Enterprise and Edge Computing
+1-408-486-2248
kuchiyama@nvidia.com