World’s Top System Makers Unveil NVIDIA A100-Powered Servers to Accelerate AI, Data Science and Scientific Computing

Cisco, Dell Technologies, HPE, Inspur, Lenovo, Supermicro Announce Systems Coming This Summer

ISC Digital--NVIDIA and the world's leading server manufacturers today announced NVIDIA A100-powered systems in a variety of designs and configurations to tackle the most complex challenges in AI, data science and scientific computing.

More than 50 A100-powered servers from leading vendors around the world -- including ASUS, Atos, Cisco, Dell Technologies, Fujitsu, GIGABYTE, Hewlett Packard Enterprise, Inspur, Lenovo, One Stop Systems, Quanta/QCT and Supermicro -- are expected following last month's launch of the NVIDIA Ampere architecture and the NVIDIA A100 GPU.

Availability of the servers varies, with 30 systems expected this summer, and over 20 more by the end of the year.

“Adoption of NVIDIA A100 GPUs into leading server manufacturers' offerings is outpacing anything we've previously seen," said Ian Buck, vice president and general manager of Accelerated Computing at NVIDIA. “The sheer breadth of NVIDIA A100 servers coming from our partners ensures that customers can choose the very best options to accelerate their data centers for high utilization and low total cost of ownership."

The first GPU based on the NVIDIA Ampere architecture, the A100 can boost performance by up to 20x over its predecessor -- making it the company's largest leap in GPU performance to date. It features several technical breakthroughs, including a new multi-instance GPU technology enabling a single A100 to be partitioned into as many as seven separate GPUs to handle varying compute jobs; third-generation NVIDIA® NVLink® technology that makes it possible to join several GPUs together to operate as one giant GPU; and new structural sparsity capabilities that can be used to double a GPU's performance.

NVIDIA also unveiled a PCIe form factor for the A100, complementing the four- and eight-way NVIDIA HGX™ A100 configurations launched last month. The addition of a PCIe version enables server makers to provide customers with a diverse set of offerings -- from single A100 GPU systems to servers featuring 10 or more GPUs. These systems accelerate a wide range of compute-intensive workloads, from simulating molecular behavior for drug discovery to building better financial models for mortgage approvals.

Server manufacturers bringing NVIDIA A100-powered systems to their customers include:

- ASUS will offer the ESC4000A-E10, which can be configured with four A100 PCIe GPUs in a single server.
- Atos is offering its BullSequana X2415 system with four NVIDIA A100 Tensor Core GPUs.
- Cisco plans to support NVIDIA A100 Tensor Core GPUs in its Cisco Unified Computing System servers and in its hyperconverged infrastructure system, Cisco HyperFlex.
- Dell Technologies plans to support NVIDIA A100 Tensor Core GPUs across its PowerEdge servers and solutions that accelerate workloads from edge to core to cloud, just as it supports other NVIDIA GPU accelerators, software and technologies in a wide range of offerings.
- Fujitsu is bringing A100 GPUs to its PRIMERGY line of servers.
- GIGABYTE will offer G481-HA0, G492-Z50 and G492-Z51 servers that support up to 10 A100 PCIe GPUs, while the G292-Z40 server supports up to eight.
- HPE will support A100 PCIe GPUs in the HPE ProLiant DL380 Gen10 Server, and for accelerated HPC and AI workloads, in the HPE Apollo 6500 Gen10 System.
- Inspur is releasing eight NVIDIA A100-powered systems, including the NF5468M5, NF5468M6 and NF5468A5 using A100 PCIe GPUs, the NF5488M5-D, NF5488A5, NF5488M6 and NF5688M6 using eight-way NVLink, and the NF5888M6 with 16-way NVLink.
- Lenovo will support A100 PCIe GPUs on select systems, including the Lenovo ThinkSystem SR670 AI-ready server. Lenovo will expand availability across its ThinkSystem and ThinkAgile portfolio in the fall.
- One Stop Systems will offer its OSS 4UV Gen 4 PCIe expansion system with up to eight NVIDIA A100 PCIe GPUs to allow AI and HPC customers to scale out their Gen 4 servers.
- Quanta/QCT will offer several QuantaGrid server systems, including D52BV-2U, D43KQ-2U and D52G-4U that support up to eight NVIDIA A100 PCIe GPUs.
- Supermicro will offer its 4U A+ GPU system, supporting up to eight NVIDIA A100 PCIe GPUs and up to two additional high-performance PCI-E 4.0 expansion slots along with other 1U, 2U and 4U GPU servers.

NVIDIA is expanding its portfolio of NGC-Ready™ certified systems. Working directly with NVIDIA, system vendors can receive NGC-Ready certification for their A100-powered servers. NGC-Ready certification assures customers that systems will deliver the performance required to run AI workloads.

NGC-Ready systems are tested with GPU-optimized AI software from NVIDIA's NGC™ registry, which is available for NVIDIA GPU-powered systems in data centers, the cloud and at the edge.

NVIDIA A100 Optimized Software Now Available
NVIDIA A100 is supported by NVIDIA Ampere-optimized software, including CUDA 11; new versions of more than 50 CUDA-X™ libraries; NVIDIA Jarvis, a multimodal, conversational AI services framework; NVIDIA Merlin, a deep recommender application framework; the RAPIDS™ suite of open source data science software libraries; and the NVIDIA HPC SDK, which includes compilers, libraries and software tools to maximize developer productivity and the performance and portability of HPC applications.

These powerful software tools enable developers to build and accelerate applications in HPC, genomics, 5G, data science, robotics and more.

About NVIDIA
NVIDIA‘s (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics and revolutionized

parallel computing. More recently, GPU deep learning ignited modern AI — the next era of computing — with the GPU acting as the brain of computers, robots and self-driving cars that can perceive and understand the world. More information at http://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the server manufacturers developing NVIDIA A100-powered systems and what they will offer, feature, plan and support; the number of and timing for vendors producing NVIDIA A100 servers; the benefits, abilities and performance of the NVIDIA A100, including accelerating and tackling challenges in AI, data science and scientific computing; the availability of the NVIDIA A100 servers; the pace of the adoption of NVIDIA A100 GPUs; the breadth of offerings using A100 GPUs and ensuring that customers have options and its benefits; the features and technical breakthroughs of the A100 and what it enables; PCIe enabling server makers to provide customers with diverse offerings and it accelerating workloads; NVIDIA expanding its portfolio of NGC-Ready systems; NGC-Ready certification assuring customers that systems will deliver the performance required to run AI workloads; and the availability, benefits and performance of NVIDIA A100-optimized software and it enabling developers to build and acceleration applications 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