NVIDIA’s New Ampere Data Center GPU in Full Production

New NVIDIA A100 GPU Boosts AI Training and Inference up to 20x; NVIDIA’s First Elastic, Multi-Instance GPU Unifies Data Analytics, Training and Inference; Adopted by World’s Top Cloud Providers and Server Makers

GTC 2020 -- NVIDIA today announced that the first GPU based on the NVIDIA® Ampere architecture, the NVIDIA A100, is in full production and shipping to customers worldwide.

The A100 draws on design breakthroughs in the NVIDIA Ampere architecture — offering the company’s largest leap in performance to date within its eight generations of GPUs — to unify AI training and inference and boost performance by up to 20x over its predecessors. A universal workload accelerator, the A100 is also built for data analytics, scientific computing and cloud graphics.

“The powerful trends of cloud computing and AI are driving a tectonic shift in data center designs so that what was once a sea of CPU-only servers is now GPU-accelerated computing,” said Jensen Huang, founder and CEO of NVIDIA. “NVIDIA A100 GPU is a 20x AI performance leap and an end-to-end machine learning accelerator — from data analytics to training to inference. For the first time, scale-up and scale-out workloads can be accelerated on one platform. NVIDIA A100 will simultaneously boost throughput and drive down the cost of data centers.”

New elastic computing technologies built into A100 make it possible to bring right-sized computing power to every job. A multi-instance GPU capability allows each A100 GPU to be partitioned into as many as seven independent instances for inferencing tasks, while third-generation NVIDIA NVLink® interconnect technology allows multiple A100 GPUs to operate as one giant GPU for ever larger training tasks.

The world’s leading cloud service providers and systems builders that expect to incorporate A100 GPUs into their offerings include: Alibaba Cloud, Amazon Web Services (AWS), Atos, Baidu Cloud, Cisco, Dell Technologies, Fujitsu, GIGABYTE, Google Cloud, H3C, Hewlett Packard Enterprise (HPE), Inspur, Lenovo, Microsoft Azure, Oracle, Quanta/QCT, Supermicro and Tencent Cloud.

Immediate Adoption Worldwide
Among the first to tap into the power of NVIDIA A100 GPUs is Microsoft, which will take advantage of their performance and scalability.

“Microsoft trained Turing NLG, the largest language model in the world, using the current generation of NVIDIA GPUs,” said Mikhail Parakhin, corporate vice president at Microsoft. “We will train dramatically bigger AI models using thousands of NVIDIA’s new generation of A100 GPUs in Azure at scale to push the state of the art on language, speech, vision and multi-modality.”

DoorDash, an on-demand food platform serving as a lifeline to restaurants during the pandemic, notes the importance of having a flexible AI infrastructure.

“Modern and complex AI training and inference workloads that require a large amount of data can benefit from state-of-the-art technology like NVIDIA A100 GPUs, which help reduce model training time and speed up the machine learning development process,” said Gary Ren, machine learning engineer at DoorDash. “In addition, using cloud-based GPU clusters gives us newfound flexibility to scale up or down as needed, helping to improve efficiency, simplify our operations and save costs."

Other early adopters include national laboratories and some of the world’s leading higher education and research institutions, each using A100 to power their next-generation supercomputers. They include:

- **Indiana University**, in the U.S., whose Big Red 200 supercomputer is based on HPE’s Cray Shasta system, will support scientific and medical research, and advanced research in AI, machine learning and data analytics.
- **Jülich Supercomputing Centre**, in Germany, whose JUWELS booster system being built by Atos is designed for extreme computing power and AI tasks.
- **Karlsruhe Institute of Technology**, in Germany, which is building its HoreKa supercomputer with Lenovo, will be able to carry out significantly larger multi-scale simulations in the field of materials sciences, earth system sciences, engineering for energy and mobility research, and particle and astroparticle physics.
- **Max Planck Computing and Data Facility**, in Germany, with its next-generation supercomputer Raven built by Lenovo, provides high-level support for the development, optimization, analysis and visualization of high-performance-computing applications to Max Planck Institutes.

- The **U.S. Department of Energy's National Energy Research Scientific Computing Center**, located at Lawrence Berkeley National Laboratory, which is building its next-generation supercomputer Perlmutter based on HPE's Cray Shasta system to support extreme-scale science and develop new energy sources, improve energy efficiency and discover new materials.

Five Breakthroughs of A100
The NVIDIA A100 GPU is a technical design breakthrough fueled by five key innovations:

- **NVIDIA Ampere architecture** — At the heart of A100 is the NVIDIA Ampere GPU architecture, which contains more than 54 billion transistors, making it the world's largest 7-nanometer processor.
- **Third-generation Tensor Cores with TF32** — NVIDIA's widely adopted Tensor Cores are now more flexible, faster and easier to use. Their expanded capabilities include new TF32 for AI, which allows for up to 20x the AI performance of FP32 precision, without any code changes. In addition, Tensor Cores now support FP64, delivering up to 2.5x more compute than the previous generation for HPC applications.
- **Multi-instance GPU** — MIG, a new technical feature, enables a single A100 GPU to be partitioned into as many as seven separate GPUs so it can deliver varying degrees of compute for jobs of different sizes, providing optimal utilization and maximizing return on investment.
- **Third-generation NVIDIA NVLink** — Doubles the high-speed connectivity between GPUs to provide efficient performance scaling in a server.
- **Structural sparsity** — This new efficiency technique harnesses the inherently sparse nature of AI math to double performance.

Together, these new features make the NVIDIA A100 ideal for diverse, demanding workloads, including AI training and inference as well as scientific simulation, conversational AI, recommender systems, genomics, high-performance data analytics, seismic modeling and financial forecasting.

NVIDIA A100 Available in New Systems, Coming to Cloud Soon
The NVIDIA DGX™ A100 system, also announced today, features eight NVIDIA A100 GPUs interconnected with NVIDIA NVLink. It is available immediately from NVIDIA and approved partners.

Alibaba Cloud, AWS, Baidu Cloud, Google Cloud, Microsoft Azure, Oracle and Tencent Cloud are planning to offer A100-based services.

Additionally, a wide range of A100-based servers are expected from the world's leading systems manufacturers, including Atos, Cisco, Dell Technologies, Fujitsu, GIGABYTE, H3C, HPE, Inspur, Lenovo, Quanta/QCT and Supermicro.

To help accelerate development of servers from its partners, NVIDIA has created HGX A100 — a server building block in the form of integrated baseboards in multiple GPU configurations.

The four-GPU HGX A100 offers full interconnection between GPUs with NVLink, while the eight-GPU configuration offers full GPU-to-GPU bandwidth through NVIDIA NVSwitch™. HGX A100, with the new MIG technology, can be configured as 56 small GPUs, each faster than NVIDIA T4, all the way up to a giant eight-GPU server with 10 petaflops of AI performance.

Software Optimizations for A100
NVIDIA also announced several updates to its software stack enabling application developers to take advantage of A100 GPU's innovations. They include new versions of more than 50 CUDA-X™ libraries used to accelerate graphics, simulation and AI; CUDA 11; NVIDIA Jarvis, a multimodal, conversational AI services framework; NVIDIA Merlin, a deep recommender application framework; and the NVIDIA HPC SDK, which includes compilers, libraries and tools that help HPC developers debug and optimize their code for A100.

About NVIDIA
NVIDIA's 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 https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, performance, features and availability of our products and technologies, including NVIDIA A100 and the NVIDIA Ampere GPU architecture, NVIDIA NVLink interconnect technology, cloud-based GPU clusters, Tensor Cores with TF32, multi-instance GPU, structural sparsity, the NVIDIA DGX A100 system, HGX A100, and NVIDIA software optimizations for A100; cloud computing and AI driving a tectonic shift in data center designs; the cloud service providers and systems builders that expect to incorporate A100 GPUs into their offerings; Microsoft's plans with NVIDIA A100 GPUs; and the national laboratories and higher education and research institutes that plan to use A100 to power their next-generation supercomputers and the related impacts 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 Bryson
Enterprise Data Center, AI/DL
+1-203-241-9190
kbryson@nvidia.com