NVIDIA Boosts World's Leading Deep Learning Computing Platform, Bringing 10x Performance Gain in Six Months

Gain Driven by New Tesla V100 32GB GPU with 2x the Memory, Revolutionary NVSwitch Fabric, Comprehensive Software Stack; DGX-2 Is First 2 Petaflop Deep Learning System

GPU Technology Conference -- NVIDIA today unveiled a series of important advances to its world-leading deep learning computing platform, which delivers a 10x performance boost on deep learning workloads compared with the previous generation six months ago.

Key advancements to the NVIDIA platform -- which has been adopted by every major cloud-services provider and server maker -- include a 2x memory boost to NVIDIA® Tesla® V100, the most powerful datacenter GPU, and a revolutionary new GPU interconnect fabric called NVIDIA NVSwitch™, which enables up to 16 Tesla V100 GPUs to simultaneously communicate at a record speed of 2.4 terabytes per second. NVIDIA also introduced an updated, fully optimized software stack.

Additionally, NVIDIA launched a major breakthrough in deep learning computing with NVIDIA DGX-2™, the first single server capable of delivering two petaflops of computational power. DGX-2 has the deep learning processing power of 300 servers occupying 15 racks of datacenter space, while being 60x smaller and 18x more power efficient.

"The extraordinary advances of deep learning only hint at what is still to come," said Jensen Huang, NVIDIA founder and CEO, as he unveiled the news at GTC 2018. "Many of these advances stand on NVIDIA's deep learning platform, which has quickly become the world's standard. We are dramatically enhancing our platform's performance at a pace far exceeding Moore's law, enabling breakthroughs that will help revolutionize healthcare, transportation, science exploration and countless other areas."

Tesla V100 Gets Double the Memory

The Tesla V100 GPU, widely adopted by the world's leading researchers, has received a 2x memory boost to handle the most memory-intensive deep learning and high performance computing workloads.

Now equipped with 32GB of memory, Tesla V100 GPUs will help data scientists train deeper and larger deep learning models that are more accurate than ever. They can also improve the performance of memory-constrained HPC applications by up to 50 percent compared with the previous 16GB version.

The Tesla V100 32GB GPU is immediately available across the complete NVIDIA DGX system portfolio. Additionally, major computer manufacturers Cray, Hewlett Packard Enterprise, IBM, Lenovo, Supermicro and Tyan announced they will begin rolling out their new Tesla V100 32GB systems within the second quarter. Oracle Cloud Infrastructure also announced plans to offer Tesla V100 32GB in the cloud in the second half of the year.

NVSwitch: A Revolutionary Interconnect Fabric

NVSwitch offers 5x higher bandwidth than the best PCIe switch, allowing developers to build systems with more GPUs hyperconnected to each other. It will help developers break through previous system limitations and run much larger datasets. It also opens the door to larger, more complex workloads, including modeling parallel training of neural networks.

NVSwitch extends the innovations made available through NVIDIA NVLink™, the first high-speed interconnect technology developed by NVIDIA. NVSwitch allows system designers to build even more advanced systems that can flexibly connect any topology of NVLink-based GPUs.

Advanced GPU-Accelerated Deep Learning and HPC Software Stack

The updates to NVIDIA's deep learning and HPC software stack are available at no charge to its developer community, which now totals more than 820,000 registered users, compared with about 480,000 a year ago.

Among its updates are new versions of NVIDIA CUDA®, TensorRT, NCCL and cuDNN, and a new Isaac software developer kit for robotics. Additionally, through close collaboration with leading cloud service providers, every major deep learning framework is continually optimized to take full advantage of NVIDIA's GPU computing platform.

NVIDIA DGX-2: World's First Two Petaflop System

NVIDIA's new DGX-2 system reached the two petaflop milestone by drawing from a wide range of industry-leading technology advances developed by NVIDIA at all levels of the computing stack.

DGX-2 is the first system to debut NVSwitch, which enables all 16 GPUs in the system to share a unified memory space. Developers now have the deep learning training power to tackle the largest datasets and most complex deep learning models.

Combined with a fully optimized, updated suite of NVIDIA deep learning software, DGX-2 is purpose-built for data scientists pushing the outer limits of deep learning research and computing.

DGX-2 can train FAIRSeq, a state-of-the-art neural machine translation model, in less than two days -- a 10x improvement in performance from the DGX-1 with Volta, introduced in September.

Industry Support for Tesla V100 32GB

"Our collaboration with NVIDIA on AI technologies includes recent breakthroughs in Chinese-to-English translation," said Xuedong Huang, technical fellow and head of speech and language at Microsoft. "With the new Tesla V100 32GB GPUs, we will be able to train larger, more complex AI models faster. This will help extend the accuracy of our models on speech recognition and machine translation reaching human capabilities and enhancing offerings such as Cortana, Bing and Microsoft Translator."

“We evaluated DGX-1 with the new Tesla V100 32GB for our SAP Brand Impact application, which automatically analyzes brand exposure in videos in near real-time," said Michael Kemelmakher, vice president, SAP Innovation Center, Israel. “The additional memory improved our ability to handle higher definition images on a larger ResNet-152 model, reducing error rate by 40 percent on average. This results in accurate, timely and auditable services at scale."

NVIDIA DGX Product Portfolio
DGX-2 is the latest addition to the NVIDIA DGX product portfolio, which consists of three systems designed to help data scientists quickly develop, test, deploy and scale new deep learning models and innovations.

DGX-2, with 16 GPUs, is the top of the lineup. It joins the NVIDIA DGX-1 system, which features eight Tesla V100 GPUs, and DGX Station™, the world's first personal deep learning supercomputer, with four Tesla V100 GPUs in a compact, deskside design. These systems enable data scientists to scale their work from the complex experiments they run at their desks to the largest deep learning problems, allowing them to do their life's work.

Further information, including detailed technical specifications and order forms, is available at https://nvda.ws/2IRilLe.

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 benefits, impact, performance and abilities of the NVIDIA Tesla V100 GPUs, NVIDIA NVSwitch, updated software stack, NVIDIA DGX-2, NVIDIA DGX-1 and NVIDIA DGX Station; the implications, benefits and impact of deep learning advances and the breakthroughs it will enable; major computer manufacturer’s upcoming use of Tesla V100 systems; and availability of the NVIDIA DGX-2 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 reports NVIDIA files with the Securities and Exchange Commission, or SEC, including its Form 10- K for the fiscal period ended January 28, 2018. 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.

Media Contacts

Kristin Bryson
+1 203 241 9190
kbryson@nvidia.com