NVIDIA Launches Revolutionary Volta GPU Platform, Fueling Next Era of AI and High Performance Computing

Volta-Based Tesla V100 Data Center GPU Shatters Barrier of 120 Teraflops of Deep Learning

NVIDIA today launched Volta -- the world's most powerful GPU computing architecture, created to drive the next wave of advancement in artificial intelligence and high performance computing.

The company also announced its first Volta-based processor, the NVIDIA® Tesla® V100 data center GPU, which brings extraordinary speed and scalability for AI inferencing and training, as well as for accelerating HPC and graphics workloads.

"Artificial intelligence is driving the greatest technology advances in human history," said Jensen Huang, founder and chief executive officer of NVIDIA, who unveiled Volta at his GTC keynote. "It will automate intelligence and spur a wave of social progress unmatched since the industrial revolution.

"Deep learning, a groundbreaking AI approach that creates computer software that learns, has insatiable demand for processing power. Thousands of NVIDIA engineers spent over three years crafting Volta to help meet this need, enabling the industry to realize AI's life-changing potential," he said.

Volta, NVIDIA's seventh-generation GPU architecture, is built with 21 billion transistors and delivers the equivalent performance of 100 CPUs for deep learning.

It provides a 5x improvement over Pascal™, the current-generation NVIDIA GPU architecture, in peak teraflops, and 15x over the Maxwell™ architecture, launched two years ago. This performance surpasses by 4x the improvements that Moore's law would have predicted.

Demand for accelerating AI has never been greater. Developers, data scientists and researchers increasingly rely on neural networks to power their next advances in fighting cancer, making transportation safer with self-driving vehicles, providing new intelligent customer experiences and more.

Data centers need to deliver exponentially greater processing power as these networks become more complex. And they need to efficiently scale to support the rapid adoption of highly accurate AI-based services, such as natural language virtual assistants, and personalized search and recommendation systems.

Volta will become the new standard for high performance computing. It offers a platform for HPC systems to excel at both computational science and data science for discovering insights. By pairing CUDA® cores and the new Volta Tensor Core within a unified architecture, a single server with Tesla V100 GPUs can replace hundreds of commodity CPUs for traditional HPC.

Breakthrough Technologies
The Tesla V100 GPU leapfrogs previous generations of NVIDIA GPUs with groundbreaking technologies that enable it to shatter the 100 teraflops barrier of deep learning performance. They include:

- Tensor Cores designed to speed AI workloads. Equipped with 640 Tensor Cores, V100 delivers 120 teraflops of deep learning performance, equivalent to the performance of 100 CPUs.
- New GPU architecture with over 21 billion transistors. It pairs CUDA cores and Tensor Cores within a unified architecture, providing the performance of an AI supercomputer in a single GPU.
- NVLink™ provides the next generation of high-speed interconnect linking GPUs, and GPUs to CPUs, with up to 2x the throughput of the prior generation NVLink.
- 900 GB/sec HBM2 DRAM, developed in collaboration with Samsung, achieves 50 percent more memory bandwidth than previous generation GPUs, essential to support the extraordinary computing throughput of Volta.
- Volta-optimized software, including CUDA, cuDNN and TensorRT™ software, which leading frameworks and applications can easily tap into to accelerate AI and research.

Ecosystem Support for Volta
Volta has received broad industry support from leading companies and organizations around the world:

"NVIDIA and AWS have worked together for a long time to help customers run compute-intensive AI workloads in the cloud. We launched the first GPU-optimized cloud instance in 2010, and introduced last year the most powerful GPU instance available in the cloud. AWS is home to some of today's most innovative and creative AI applications, and we look forward to helping customers continue to build incredible new applications with the next generation of our general-purpose GPU instance family when Volta becomes available later in the year."
-- Matt Garman, vice president of Compute Services, Amazon Web Services

"We express our congratulations to NVIDIA's latest release of Volta. From Baidu Cloud to Intelligent Driving, Baidu has been strengthening its efforts in building an open AI platform. Together with NVIDIA, we believe we will accelerate the development and application of the global AI technology and create more opportunities for the whole society."
-- Yaqin Zhang, president, Baidu

"NVIDIA and Facebook have been great partners and we are excited about the contributions NVIDIA has made to Facebook's Caffe2 and PyTorch. We look forward to the AI advances NVIDIA's new high-performing Volta graphics architecture will enable."
-- Mike Schroepfer, chief technology officer, Facebook

"NVIDIA's GPUs deliver significant performance boosts for Google Cloud Platform customers. GPUs are an important part of our infrastructure, offering Google and our enterprise customers extra computational power for machine learning or high performance computing and data analysis. Volta's performance

improvements will make GPUs even more powerful and we plan to offer Volta GPUs on GCP."
-- Brad Calder, vice president of Engineering for Google Cloud Platform, Google

"Microsoft and NVIDIA have partnered for years on AI technologies, including Microsoft Azure N-series, Project Olympus and Cognitive Toolkit. The new Volta architecture will unlock extraordinary new capabilities for Microsoft customers."
-- Harry Shum, executive vice president of Microsoft AI and Research Group, Microsoft

"Oak Ridge National Laboratory will begin assembling our next-generation leadership computing system, Summit, this summer. Summit is powered by Volta GPUs and will be the top supercomputer in the U.S. for scientific discovery when completed in 2018. It will keep the U.S. at the forefront of scientific research and help the Department of Energy address complex challenges with computational science and AI-assisted discovery."
-- Jeff Nichols, associate laboratory director of the Computing and Computational Sciences Directorate, Oak Ridge National Laboratory

"A large variety of our products, including voice technology in wechat, photo/video technology in QQ and Qzone, and the deep learning platform based on Tencent Cloud, already rely on AI. We believe Volta will provide unprecedented computing power for our AI developers, and we're excited to open up those capabilities soon from Tencent Cloud to more clients."
-- Dowson Tong, senior executive vice president, Tencent

Keep Current on NVIDIA
Subscribe to the NVIDIA blog, follow us on Facebook, Google+, Twitter, LinkedIn and Instagram, and view NVIDIA videos on YouTube and images on Flickr.

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 impact, performance and benefits of the Volta architecture and the NVIDIA Tesla V100 data center GPU; the impact of artificial intelligence and deep learning; and the demand for accelerating AI 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 29, 2017. 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