NVIDIA Launches World's First Deep Learning Supercomputer

NVIDIA DGX-1 Delivers Deep Learning Throughput of 250 Servers to Meet Massive Computing Demands of Artificial Intelligence

GPU Technology Conference 2016 -- NVIDIA today unveiled the NVIDIA® DGX-1™, the world's first deep learning supercomputer to meet the unlimited computing demands of artificial intelligence.

The NVIDIA DGX-1 is the first system designed specifically for deep learning -- it comes fully integrated with hardware, deep learning software and development tools for quick, easy deployment. It is a turnkey system that contains a new generation of GPU accelerators, delivering the equivalent throughput of 250 x86 servers.[1]

The DGX-1 deep learning system enables researchers and data scientists to easily harness the power of GPU-accelerated computing to create a new class of intelligent machines that learn, see and perceive the world as humans do. It delivers unprecedented levels of computing power to drive next-generation AI applications, allowing researchers to dramatically reduce the time to train larger, more sophisticated deep neural networks.

NVIDIA designed the DGX-1 for a new computing model to power the AI revolution that is sweeping across science, enterprises and increasingly all aspects of daily life. Powerful deep neural networks are driving a new kind of software created with massive amounts of data, which require considerably higher levels of computational performance.

"Artificial intelligence is the most far-reaching technological advancement in our lifetime," said Jen-Hsun Huang, CEO and co-founder of NVIDIA. "It changes every industry, every company, everything. It will open up markets to benefit everyone. Data scientists and AI researchers today spend far too much time on home-brewed high performance computing solutions. The DGX-1 is easy to deploy and was created for one purpose: to unlock the powers of superhuman capabilities and apply them to problems that were once unsolvable."

Powered by Five Breakthroughs
The NVIDIA DGX-1 deep learning system is built on NVIDIA Tesla® P100 GPUs, based on the new NVIDIA Pascal™ GPU architecture. It provides the throughput of 250 CPU-based servers, networking, cables and racks -- all in a single box.

The DGX-1 features four other breakthrough technologies that maximize performance and ease of use. These include the NVIDIA NVLink™ high-speed interconnect for maximum application scalability; 16nm FinFET fabrication technology for unprecedented energy efficiency; Chip on Wafer on Substrate with HBM2 for big data workloads; and new half-precision instructions to deliver more than 21 teraflops of peak performance for deep learning.

Together, these major technological advancements enable DGX-1 systems equipped with Tesla P100 GPUs to deliver over 12x faster training than four-way NVIDIA Maxwell™ architecture-based solutions from just one year ago.

The Pascal architecture has strong support from the artificial intelligence ecosystem.

"NVIDIA GPU is accelerating progress in AI. As neural nets become larger and larger, we not only need faster GPUs with larger and faster memory, but also much faster GPU-to-GPU communication, as well as hardware that can take advantage of reduced-precision arithmetic. This is precisely what Pascal delivers," said Yann LeCun, director of AI Research at Facebook.

Andrew Ng, chief scientist at Baidu, said: "AI computers are like space rockets: The bigger the better. Pascal's throughput and interconnect will make the biggest rocket we've seen yet."

"Microsoft is developing super deep neural networks that are more than 1,000 layers," said Xuedong Huang, chief speech scientist at Microsoft Research. "NVIDIA Tesla P100's impressive horsepower will enable Microsoft's CNTK to accelerate AI breakthroughs."

Comprehensive Deep Learning Software Suite
The NVIDIA DGX-1 system includes a complete suite of optimized deep learning software that allows researchers and data scientists to quickly and easily train deep neural networks.

The DGX-1 software includes the NVIDIA Deep Learning GPU Training System (DIGITS™), a complete, interactive system for designing deep neural networks (DNNs). It also includes the newly released NVIDIA CUDA® Deep Neural Network library (cuDNN) version 5, a GPU-accelerated library of primitives for designing DNNs.

It also includes optimized versions of several widely used deep learning frameworks -- Caffe, Theano and Torch. The DGX-1 additionally provides access to cloud management tools, software updates and a repository for containerized applications.

System Specifications
The NVIDIA DGX-1 system specifications include:

- Up to 170 teraflops of half-precision (FP16) peak performance
- Eight Tesla P100 GPU accelerators, 16GB memory per GPU
- NVLink Hybrid Cube Mesh
- 7TB SSD DL Cache
- Dual 10GbE, Quad InfiniBand 100Gb networking
- 3U - 3200W

Optional support services for the NVIDIA DGX-1 improve productivity and reduce downtime for production systems. Hardware and software support provides access to NVIDIA deep learning expertise, and includes cloud management services, software upgrades and updates, and priority resolution of critical issues. More information is available at www.nvidia.com/object/dgxsystems-support.

Availability
General availability for the NVIDIA DGX-1 deep learning system in the United States is in June, and in other regions beginning in the third quarter direct from NVIDIA and select systems integrators.

Supporting Resources

- Deep learning video

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 (NASDAQ: NVDA) is a computer technology company that has pioneered GPU-accelerated computing. It targets the world's most demanding users -- gamers, designers and scientists -- with products, services and software that power amazing experiences in virtual reality, artificial intelligence, professional visualization and autonomous cars. More information at http://nvidianews.nvidia.com/.

(1) Compared to Caffe/AlexNet time to train ILSVRC-2012 dataset on cluster of two-socket Intel Xeon E5-2697 v3 processor-based systems with InfiniBand interconnect. 250-node performance estimated using
source: https://software.intel.com/en-us/articles/caffe-training-on-multi-node-distributed-memory-systems-based-on-intel-xeon-processor-e5.

Certain statements in this press release including, but not limited to, statements as to: the impact, performance, benefits and availability of the NVIDIA DGX-1 deep learning system 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 year ended January 31, 2016. 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

Ken Brown
+1 408 486 2626
kebrown@nvidia.com