



NVIDIA Delivers Massive Performance Leap for Deep Learning, HPC Applications With NVIDIA Tesla P100 Accelerators

Five Architectural Breakthroughs Enable Servers to Deliver Over 12x Greater Performance Than Previous Architecture

GPU Technology Conference 2016 -- NVIDIA today introduced the NVIDIA® Tesla® P100 GPU, the most advanced hyperscale data center accelerator ever built.

The latest addition to the [NVIDIA Tesla Accelerated Computing Platform](#), the Tesla P100 enables a new class of servers that can deliver the performance of hundreds of CPU server nodes. Today's data centers -- vast network infrastructures with numerous interconnected commodity CPU servers -- process large numbers of transactional workloads, such as web services. But they are inefficient at next-generation artificial intelligence and scientific applications, which require ultra-efficient, lightning-fast server nodes.

Based on the new NVIDIA Pascal™ GPU architecture with five breakthrough technologies, the Tesla P100 delivers unmatched performance and efficiency to power the most computationally demanding applications.

"Our greatest scientific and technical challenges -- finding cures for cancer, understanding climate change, building intelligent machines -- require a near-infinite amount of computing performance," said Jen-Hsun Huang, CEO and co-founder, NVIDIA. "We designed the Pascal GPU architecture from the ground up with innovation at every level. It represents a massive leap forward in computing performance and efficiency, and will help some of the smartest minds drive tomorrow's advances."

Dr. John Kelly III, senior vice president, Cognitive Solutions and IBM Research, said: "As we enter this new era of computing, entirely new approaches to the underlying technologies will be required to fully realize the benefits of AI and cognitive. The combination of NVIDIA GPUs and OpenPOWER technology is already accelerating Watson's learning of new skills. Together, IBM's Power architecture and NVIDIA's Pascal architecture with NVLink will further accelerate cognitive workload performance and advance the artificial intelligence industry."

Five Architectural Breakthroughs

The Tesla P100 delivers its unprecedented performance, scalability and programming efficiency based on five breakthroughs:

- **NVIDIA Pascal architecture for exponential performance leap** -- A Pascal-based Tesla P100 solution delivers over a 12x increase in neural network training performance compared with a previous-generation NVIDIA Maxwell™-based solution.
- **NVIDIA NVLink for maximum application scalability** -- The [NVIDIA NVLink™ high-speed GPU interconnect](#) scales applications across multiple GPUs, delivering a 5x acceleration in bandwidth compared to today's best-in-class solution¹. Up to eight Tesla P100 GPUs can be interconnected with NVLink to maximize application performance in a single node, and IBM has implemented NVLink on its POWER8 CPUs for fast CPU-to-GPU communication.
- **16nm FinFET for unprecedented energy efficiency** -- With 15.3 billion transistors built on 16 nanometer FinFET fabrication technology, the Pascal GPU is the world's largest FinFET chip ever built². It is engineered to deliver the fastest performance and best energy efficiency for workloads with near-infinite computing needs.
- **CoWoS with HBM2 for big data workloads** -- The Pascal architecture unifies processor and data into a single package to deliver unprecedented compute efficiency. An innovative approach to memory design, Chip on Wafer on Substrate (CoWoS) with HBM2, provides a 3x boost in memory bandwidth performance, or 720GB/sec, compared to the Maxwell architecture.
- **New AI algorithms for peak performance** -- New half-precision instructions deliver more than 21 teraflops of peak performance for deep learning.

The Tesla P100 GPU accelerator delivers a new level of performance for a range of HPC and deep learning applications, including the AMBER molecular dynamics code, which runs faster on a single server node with Tesla P100 GPUs than on 48 dual-socket CPU server nodes³. Training the popular AlexNet deep neural network would take 250 dual-socket CPU server nodes to match the performance of eight Tesla P100 GPUs⁴. And the widely used weather forecasting application, COSMO, runs faster on eight Tesla P100 GPUs than on 27 dual-socket CPU servers⁵.

The first accelerator to deliver more than 5 and 10 teraflops of double-precision and single-precision performance, respectively, the Tesla P100 provides a giant leap in processing capabilities and time-to-discovery for research across a

broad spectrum of domains.

Updates to the NVIDIA SDK

NVIDIA also announced a host of updates to the [NVIDIA SDK](#), the world's most powerful development platform for GPU computing.

These updates include [NVIDIA CUDA® 8](#). The latest version of NVIDIA's parallel computing platform gives developers direct access to powerful new Pascal features, including unified memory and NVLink. Also included in this release is a new graph analytics library -- nvGRAPH -- which can be used for robotic path planning, cyber security and logistics analysis, expanding the application of GPU acceleration into the realm of big data analytics.

NVIDIA also announced [cuDNN version 5](#), a GPU-accelerated library of primitives for deep neural networks. cuDNN 5 includes Pascal GPU support; acceleration of recurrent neural networks, which are used for video and other sequential data; and additional enhancements used in medical, oil and gas, and other industries. cuDNN accelerates leading deep learning frameworks, including Google's TensorFlow, UC Berkeley's Caffe, University of Montreal's Theano and NYU's Torch. These, in turn, power deep learning solutions used by Amazon, Facebook, Google and others.

Tesla P100 Specifications

Specifications of the Tesla P100 GPU accelerator include:

- 5.3 teraflops double-precision performance, 10.6 teraflops single-precision performance and 21.2 teraflops half-precision performance with NVIDIA GPU BOOST™ technology
- 160GB/sec bi-directional interconnect bandwidth with NVIDIA NVLink
- 16GB of CoWoS HBM2 stacked memory
- 720GB/sec memory bandwidth with CoWoS HBM2 stacked memory
- Enhanced programmability with page migration engine and unified memory
- ECC protection for increased reliability
- Server-optimized for highest data center throughput and reliability

Availability

General availability for the Pascal-based NVIDIA Tesla P100 GPU accelerator in the new NVIDIA DGX-1™ deep learning system is in June. It is also expected to be available beginning in early 2017 from leading server manufacturers.

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) NVLink delivers 160GB/sec of bi-directional interconnect bandwidth, compared to PCIe x16 Gen3 that delivers 31.5GB/sec of bi-directional bandwidth.

(2) NVIDIA Tesla P100 GPU has 15.3 billion 16nm FinFET transistors.

(3) CPU system: 48 nodes, each node with 2x Intel E5-2680v3 12 core, 128GB DDR4, FDR IB interconnect. GPU system: Single node, 2x Intel E5-2698 v3 16 core, 512GB DDR4, 4x Tesla P100, NVLink interconnect.

(4) 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>.

(5) CPU system: 2x Intel E5-2698 v3 16 core, 256GB DDR4. GPU system: Single node, 2x Intel E5-2698 v3 16 core, 512GB DDR4, 8x Tesla P100, NVLink interconnect.

Certain statements in this press release including, but not limited to, statements as to: the impact, performance, benefits and availability of the NVIDIA Tesla P100 GPU, the NVIDIA SDK and 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.

© 2016 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, Tesla, Pascal, Maxwell, NVIDIA NVLink, CUDA, NVIDIA GPU BOOST, and DGX-1 are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. Features, pricing, availability and specifications are subject to change without notice.

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