



New NVIDIA Pascal GPUs Accelerate Deep Learning Inference

Tesla P4, P40 Accelerators Deliver 45x Faster AI; TensorRT and DeepStream Software Boost AI for Video Inferencing

GTC China - NVIDIA today unveiled the latest additions to its Pascal™ architecture-based deep learning platform, with new NVIDIA® Tesla® P4 and P40 GPU accelerators and new software that deliver massive leaps in efficiency and speed to accelerate inferencing production workloads for artificial intelligence services.

Modern AI services such as voice-activated assistance, email spam filters, and movie and product recommendation engines are rapidly growing in complexity, requiring up to 10x more compute compared to neural networks from a year ago. Current CPU-based technology isn't capable of delivering real-time responsiveness required for modern AI services, leading to a poor user experience.

The Tesla P4 and P40 are specifically designed for inferencing, which uses trained deep neural networks to recognize speech, images or text in response to queries from users and devices. Based on the [Pascal architecture](#), these GPUs feature specialized inference instructions based on 8-bit (INT8) operations, delivering 45x faster response than CPUs¹ and a 4x improvement over GPU solutions launched less than a year ago.²

The Tesla P4 delivers the highest energy efficiency for data centers. It fits in any server with its small form-factor and low-power design, which starts at 50 watts, helping make it 40x more energy efficient than CPUs for inferencing in production workloads.³ A single server with a single Tesla P4 replaces 13 CPU-only servers for video inferencing workloads,⁴ delivering over 8x savings in total cost of ownership, including server and power costs.

The Tesla P40 delivers maximum throughput for deep learning workloads. With 47 tera-operations per second (TOPS) of inference performance with INT8 instructions, a server with eight Tesla P40 accelerators can replace the performance of more than 140 CPU servers.⁵ At approximately \$5,000 per CPU server, this results in savings of more than \$650,000 in server acquisition cost.

"With the Tesla P100 and now Tesla P4 and P40, NVIDIA offers the only end-to-end deep learning platform for the data center, unlocking the enormous power of AI for a broad range of industries," said Ian Buck, general manager of accelerated computing at NVIDIA. "They slash training time from days to hours. They enable insight to be extracted instantly. And they produce real-time responses for consumers from AI-powered services."

Software Tools for Faster Inferencing

Complementing the Tesla P4 and P40 are two software innovations to accelerate AI inferencing: NVIDIA TensorRT and the NVIDIA DeepStream SDK.

[TensorRT](#) is a library created for optimizing deep learning models for production deployment that delivers instant responsiveness for the most complex networks. It maximizes throughput and efficiency of deep learning applications by taking trained neural nets -- defined with 32-bit or 16-bit operations -- and optimizing them for reduced precision INT8 operations.

[NVIDIA DeepStream SDK](#) taps into the power of a Pascal server to simultaneously decode and analyze up to 93 HD video streams in real time compared with seven streams with dual CPUs.⁶ This addresses one of the grand challenges of AI: understanding video content at-scale for applications such as self-driving cars, interactive robots, filtering and ad placement. Integrating deep learning into video applications allows companies to offer smart, innovative video services that were previously impossible to deliver.

Leap Forward for Customers

NVIDIA customers are delivering increasingly more innovative AI services that require the highest compute performance.

"Delivering simple and responsive experiences to each of our users is very important to us," said Greg Damos, senior researcher at Baidu. "We have deployed NVIDIA GPUs in production to provide AI-powered services such as our Deep Speech 2 system and the use of GPUs enables a level of responsiveness that would not be possible on un-accelerated servers. Pascal with its INT8 capabilities will provide an even bigger leap forward and we look forward to delivering even better experiences to our users."

Specifications

Specifications of the Tesla P4 and P40 GPUs include:

Specification	Tesla P4	Tesla P40
Single Precision TeraFLOPS*	5.5	12
INT8 TOPS* (Tera-Operations Per Second)	22	47
CUDA Cores	2,560	3,840
GPU GDDR5 Memory	8GB	24GB
Memory Bandwidth	192GB/s	346GB/s
Power	50 Watt (or higher)	250 Watt

* With boost clock on

Availability

The NVIDIA Tesla P4 and P40 are planned to be available in November and October, respectively, in qualified servers offered by ODM, OEM and channel partners.

Supporting Resources

- [What's the Difference Between Deep Learning Training and Inference?](#)
- [Tesla P4 data sheet](#) (pdf)
- [Tesla P40 data sheet](#) (pdf)
- [TensorRT product page](#)
- [DeepStream SDK product page](#)
- [NVIDIA data center solutions](#)
- [Deep learning overview](#)

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 Comparing latency using VGG-19 neural network, batch size=4. CPU: Xeon E5-2690v4 using Intel MKL 2017. GPU: Tesla P40 using TensorRT internal version. Intel optimized VGG-19 from https://github.com/intel/caffe/tree/master/models/mkl2017_vgg_19.

2 Comparing img/sec using Caffe GoogLeNet neural network, batch size = 128. GPU server with 8x P40 compared to GPU server with 8x M40. Both using TensorRT internal version.

3 Comparing img/sec/watt using Caffe AlexNet neural network, batch size = 128. CPU: E5-2690v4 using Intel MKL 2017. Using Intel-optimized Caffe and AlexNet from <https://github.com/intel/caffe>. GPU: Tesla P4 measuring GPU power.

4 Using Intel optimized GoogLeNet, dual-socket CPU server, Xeon E5-2650v4 using Intel MKL 2017. GPU server with 1x Tesla P4 using DeepStream SDK. Video streaming at 720p @ 30FPS.

5 Comparing img/sec using GoogLeNet neural network, batch size=128. Dual-socket CPU server, Xeon E5-2690v4 using Intel MKL 2017, 358 images/sec. GPU server with 8x Tesla P40 using TensorRT internal version, 52K images/sec, 145x higher throughput than CPU server.

6 Intel optimized Caffe using dual-socket E5-2650 v4 CPU servers, Intel MKL 2017, based on GoogLeNet optimized by Intel: https://github.com/intel/caffe/tree/master/models/mkl2017_googlenet_v2, running transcode at 720p at 30FPS. GPU: using a single Tesla P4 with dual-socket E5-2650 v4 CPU server.

Certain statements in this press release including, but not limited to, statements as to: the impact, performance, features, benefits and availability of the NVIDIA Tesla P4 and P40 GPU accelerators 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-Q for the fiscal year ended July 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 and Pascal 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