New NVIDIA Hyperscale Accelerators Boost Machine Learning Throughput for Web Data Centers

Web-Services Giants Use Artificial Intelligence to Make Smarter Applications, Driving Explosive Growth in Machine Learning Workloads

NVIDIA today announced an end-to-end hyperscale data center platform that lets web-services companies accelerate their huge machine learning workloads.

The NVIDIA hyperscale accelerator line consists of two accelerators. One lets researchers more quickly innovate and design new deep neural networks for each of the increasing number of applications they want to power with artificial intelligence (AI). Another is a low-power accelerator designed to deploy these networks across the data center. The line also includes a suite of GPU-accelerated libraries.

Together, they enable developers to use the powerful Tesla Accelerated Computing Platform to drive machine learning in hyperscale data centers and create unprecedented AI-based applications.

"The artificial intelligence race is on," said Jen-Hsun Huang, co-founder and CEO of NVIDIA. "Machine learning is unquestionably one of the most important developments in computing today, on the scale of the PC, the internet and cloud computing. Industries ranging from consumer cloud services, automotive and health care are being revolutionized as we speak.

"Machine learning is the grand computational challenge of our generation. We created the Tesla hyperscale accelerator line to give machine learning a 10X boost. The time and cost savings to data centers will be significant," he said.

These new hardware and software products are designed specifically to accelerate the flood of web applications that are racing to incorporate AI capabilities. Ground-breaking advances in machine learning have made it possible to use AI techniques to create smarter applications and services.

Machine learning is being used to make voice recognition more accurate. It enables automatic object and scene recognition in video or photos with the ability to tag for later search. It makes possible facial recognition in videos or photos, even when the face is partially obscured. And it powers services that are aware of individual tastes and interests, which can organize schedules, deliver relevant news stories and respond to voice commands accurately and in a conversational tone.

The magic is made possible by machine learning. The challenge is obtaining the daunting amount of supercomputing power needed to innovate and train the growing number of deep neural networks, and the processing to instantly respond to the billions of queries from consumers using the services. The NVIDIA hyperscale accelerator line was created to accelerate these workloads and dramatically increase the throughput of data centers.

These new additions to the NVIDIA Tesla platform include:

- NVIDIA® Tesla® M40 GPU - the most powerful accelerator designed for training deep neural networks
- NVIDIA Tesla M4 GPU - a low-power, small form-factor accelerator for machine learning inference, as well as streaming image and video processing
- NVIDIA Hyperscale Suite - a rich suite of software optimized for machine learning and video processing

NVIDIA Tesla M40 GPU Accelerator
The NVIDIA Tesla M40 GPU accelerator allows data scientists to save days, even weeks, of time while training their deep neural networks against massive amounts of data for higher overall accuracy. Key features include:

- Optimized for Machine Learning - Reduces training time by 8X compared with CPUs (1.2 days vs. 10 days for a typical AlexNet training).
- Built for 24/7 reliability - Designed and tested for high reliability in data center environments.
- Scale-out performance - Support for NVIDIA GPUDirect allowing fast multi-node neural network training.

NVIDIA Tesla M4 GPU Accelerator
The NVIDIA Tesla M4 accelerator is a low-power GPU purpose-built for hyperscale environments and optimized for demanding, high-growth web services applications, including video transcoding, image and video processing, and machine learning inference. Key features include:

- Higher throughput - Transcodes, enhances and analyzes up to 5X more simultaneous video streams compared with CPUs.
- Low power consumption - With a user-selectable power profile, the Tesla M4 consumes 50-75 watts of power, and delivers up to 10X better energy efficiency than a CPU for video processing and machine learning algorithms.
- Small form factor - Low-profile PCIe design fits into enclosures required for hyperscale data center systems.

NVIDIA Hyperscale Suite
The new NVIDIA Hyperscale Suite includes tools for both developers and data center managers, specifically designed for web services deployments, including:

- cuDNN - the industry's most popular algorithm software for processing deep convolutional neural networks used for AI applications.
- GPU-accelerated FFmpeg multimedia software - Harnesses widely used FFmpeg software to accelerate video transcoding and video processing.
- NVIDIA GPU REST Engine - Enables the easy creation and deployment of high-throughput, low-latency accelerated web services spanning dynamic image resizing, search acceleration, image classification and other tasks.
- NVIDIA Image Compute Engine - GPU-accelerated service with REST API that provides image resizing 5 times faster compared to a CPU.

Mesosphere Support
In the latest showing of industry support for the Tesla Accelerated Computing Platform, Mesosphere announced that it is collaborating with NVIDIA to add support for GPU technology to Apache Mesos and the Mesosphere Datacenter Operating System (DCOS). The move will make it easier for web-services companies to

build and deploy accelerated data centers for their next-generation applications.

Availability
The Tesla M40 GPU accelerator and Hyperscale Suite software will be available later this year. The Tesla M4 GPU will be available in the first quarter of 2016. For more information visit the NVIDIA Tesla website.

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

Since 1993, NVIDIA (NASDAQ: NVDA) has pioneered the art and science of visual computing. The company's technologies are transforming a world of displays into a world of interactive discovery -- for everyone from gamers to scientists, and consumers to enterprise customers. More information at http://nvidianews.nvidia.com/ and http://blogs.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the features, benefits, impact and availability of the NVIDIA hyperscale accelerator line, including the NVIDIA Tesla M40 GPU accelerator and the NVIDIA Tesla M4 GPU accelerator, and the NVIDIA Hyperscale Suite; machine learning as one of the most important developments in computing; and the benefits and impact of machine learning 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 period ended July 26, 2015. 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.

About NVIDIA

Since 1993, NVIDIA (NASDAQ : NVDA) has pioneered the art and science of visual computing. The company's technologies are transforming a world of displays into a world of interactive discovery — for everyone from gamers to scientists, and consumers to enterprise customers. More information at http://nvidianews.nvidia.com/ and http://blogs.nvidia.com/.

Media Contacts

George Millington
+1 408 562 7226
gmillington@nvidia.com