NVIDIA TensorRT 3 Dramatically Accelerates AI Inference for Hyperscale Data Centers

Alibaba, Baidu, Tencent, JD.com and Hikvision Adopt NVIDIA TensorRT for Programmable Inference Acceleration

GTC China - NVIDIA today unveiled new NVIDIA® TensorRT™ 3 AI inference software that sharply boosts the performance and slashes the cost of inferencing from the cloud to edge devices, including self-driving cars and robots.

The combination of TensorRT 3 with NVIDIA GPUs delivers ultra-fast and efficient inferencing across all frameworks for AI-enabled services -- such as image and speech recognition, natural language processing, visual search and personalized recommendations. TensorRT and NVIDIA Tesla® GPU accelerators are up to 40 times faster than CPUs(1) at one-tenth the cost of CPU-based solutions.(2)

"Internet companies are racing to infuse AI into services used by billions of people. As a result, AI inference workloads are growing exponentially," said NVIDIA founder and CEO Jensen Huang. "NVIDIA TensorRT is the world's first programmable inference accelerator. With CUDA programmability, TensorRT will be able to accelerate the growing diversity and complexity of deep neural networks. And with TensorRT's dramatic speed-up, service providers can affordably deploy these compute intensive AI workloads."

More than 1,200 companies have already begun using NVIDIA's inference platform across a wide spectrum of industries to discover new insights from data and deploy intelligent services to businesses and consumers. Among them are Amazon, Microsoft, Facebook and Google; as well as leading Chinese enterprise companies like Alibaba, Baidu, JD.com, iFLYTEK, Hikvision, Tencent and WeChat.

"NVIDIA's AI platform, using TensorRT software on Tesla GPUs, is an outstanding technology at the forefront of enabling SAP's growing requirements for inferencing," said Juergen Mueller, chief innovation officer at SAP. "TensorRT and NVIDIA GPUs make real-time service delivery possible, with maximum machine learning performance and versatility to meet our customers' needs."

"JD.com relies on NVIDIA GPUs and software for inferencing in our data centers," said Andy Chen, senior director of AI and Big Data at JD. "Using NVIDIA's TensorRT on Tesla GPUs, we can simultaneously inference 1,000 HD video streams in real time, with 20 times fewer servers. NVIDIA's deep learning platform provides outstanding performance and efficiency for JD."

TensorRT 3 is a high-performance optimizing compiler and runtime engine for production deployment of AI applications. It can rapidly optimize, validate and deploy trained neural networks for inference to hyperscale data centers, embedded or automotive GPU platforms.

It offers highly accurate INT8 and FP16 network execution, which can save data center operators tens of millions of dollars in acquisition and annual energy costs. A developer can use it to take a trained neural network and, in just one day, create a deployable inference solution that runs 3-5x faster than their training framework.

To further accelerate AI, NVIDIA introduced additional software, including:

- DeepStream SDK:NVIDIA DeepStream SDK delivers real-time, low-latency video analytics at scale. It helps developers integrate advanced video inference capabilities, including INT8 precision and GPU-accelerated transcoding, to support AI-powered services like object classification and scene understanding for up to 30 HD streams in real time on a single Tesla P4 GPU accelerator.

- CUDA 9: The latest version of CUDA®, NVIDIA's accelerated computing software platform, speeds up HPC and deep learning applications with support for NVIDIA Volta architecture-based GPUs, up to 5x faster libraries, a new programming model for thread management and updates to debugging and profiling tools. CUDA 9 is optimized to deliver maximum performance on Tesla V100 GPU accelerators.

Inference for the Data Center
Data center managers constantly balance performance and efficiency to keep their server fleets at maximum productivity. Tesla GPU accelerated servers can replace over a hundred hyperscale CPU servers for deep learning inference applications and services, freeing up precious rack space, reducing energy and cooling requirements, and reducing cost as much as 90 percent.

NVIDIA Tesla GPU accelerators provide the optimal inference solution -- combining the highest throughput, best efficiency and lowest latency on deep learning inference workloads to power new AI-driven experiences.

Inference for Self-Driving Cars and Embedded Applications
With NVIDIA's unified architecture, deep neural networks on every deep learning framework can be trained on NVIDIA DGX™ systems in the data center, and then deployed into all types of devices -- from robots to autonomous vehicles -- for real-time inferencing at the edge.

TuSimple, a startup developing autonomous trucking technology, increased inferencing performance by 30 percent after TensorRT optimization. In June, the company successfully completed a 170-mile Level 4 test drive from San Diego to Yuma, Arizona, using NVIDIA GPUs and cameras as the primary sensor. The performance gains from TensorRT allow TuSimple to analyze additional camera data, and add new AI algorithms to their autonomous trucks, without sacrificing response time.

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 iNVIDIA'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/.

(1) Performance comparison based on ResNet-50 on NVIDIA Tesla V100 GPU running TensorRT 3 RC vs. Intel Xeon-D 1587 Broadwell-E CPU and Intel DL SDK. Score doubled to comprehend Intel's stated claim of 2x performance improvement on Skylake with AVX512.

(2) Comparison based on cost and ResNet-50 inference performance of an HGX-1 server with 8x NVIDIA Tesla V100, and estimated cost and ResNet-50 performance of a dual socket Intel Skylake scale-out server. Skylake performance estimation comprehends Intel's stated claim of 2x performance improvement on Skylake with AVX512.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact and performance of NVIDIA TensorRT 3 AI inferencing software, NVDLA, NVIDIA DeepStream SDK, CUDA 9, Volta, Tesla GPU accelerators; and the demands for AI computing growing every day 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 30, 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.

© 2017 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, CUDA, DGX, TensorRT and Tesla 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.

Media Contacts

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