New NVIDIA Data Center Inference Platform to Fuel Next Wave of AI-Powered Services

Tesla T4 GPU and New TensorRT Software Enable Intelligent Voice, Video, Image and Recommendation Services

GTC Japan -- Fueling the growth of AI services worldwide, NVIDIA today launched an AI data center platform that delivers the industry's most advanced inference acceleration for voice, video, image and recommendation services.

The NVIDIA TensorRT™ Hyperscale Inference Platform features NVIDIA® Tesla® T4 GPUs based on the company's breakthrough NVIDIA Turing™ architecture and a comprehensive set of new inference software.

Delivering the fastest performance with lower latency for end-to-end applications, the platform enables hyperscale data centers to offer new services, such as enhanced natural language interactions and direct answers to search queries rather than a list of possible results.

“Our customers are racing toward a future where every product and service will be touched and improved by AI," said Ian Buck, vice president and general manager of Accelerated Business at NVIDIA. “The NVIDIA TensorRT Hyperscale Platform has been built to bring this to reality -- faster and more efficiently than had been previously thought possible."

Every day, massive data centers process billions of voice queries, translations, images, videos, recommendations and social media interactions. Each of these applications requires a different type of neural network residing on the server where the processing takes place.

To optimize the data center for maximum throughput and server utilization, the NVIDIA TensorRT Hyperscale Platform includes both real-time inference software and Tesla T4 GPUs, which process queries up to 40x faster than CPUs alone.

NVIDIA estimates that the AI inference industry is poised to grow in the next five years into a $20 billion market.

Industry's Most Advanced AI Inference Platform
The NVIDIA TensorRT Hyperscale Platform includes a comprehensive set of hardware and software offerings optimized for powerful, highly efficient inference. Key elements include:

- NVIDIA Tesla T4 GPU - Featuring 320 Turing Tensor Cores and 2,560 CUDA® cores, this new GPU provides breakthrough performance with flexible, multi-precision capabilities, from FP32 to FP16 to INT8, as well as INT4. Packaged in an energy-efficient, 75-watt, small PCIe form factor that easily fits into most servers, it offers 65 teraflops of peak performance for FP16, 130 TOPS for INT8 and 260 TOPS for INT4.
- NVIDIA TensorRT 5 - An inference optimizer and runtime engine, NVIDIA TensorRT 5 supports Turing Tensor Cores and expands the set of neural network optimizations for multi-precision workloads.
- NVIDIA TensorRT inference server - This containerized microservice software enables applications to use AI models in data center production. Freely available from the NVIDIA GPU Cloud container registry, it maximizes data center throughput and GPU utilization, supports all popular AI models and frameworks, and integrates with Kubernetes and Docker.

Supported by Technology Leaders Worldwide
Support for NVIDIA's new inference platform comes from leading consumer and business technology companies around the world.

“We are working hard at Microsoft to deliver the most innovative AI-powered services to our customers," said Jordi Ribas, corporate vice president for Bing and AI Products at Microsoft. “Using NVIDIA GPUs in real-time inference workloads has improved Bing's advanced search offerings, enabling us to reduce object detection latency for images. We look forward to working with NVIDIA's next-generation inference hardware and software to expand the way people benefit from AI products and services."

Chris Kleban, product manager at Google Cloud, said: “AI is becoming increasingly pervasive, and inference is a critical capability customers need to successfully deploy their AI models, so we're excited to support NVIDIA's Turing Tesla T4 GPUs on Google Cloud Platform soon."

More information, including details on how to request early access to T4 GPUs on Google Cloud Platform, is available here.

Additional companies, including all major server manufacturers, voicing support for the NVIDIA TensorRT Hyperscale Platform include:

“Cisco's UCS portfolio delivers policy-driven, GPU-accelerated systems and solutions to power every phase of the AI lifecycle. With the NVIDIA Tesla T4 GPU based on the NVIDIA Turing architecture, Cisco customers will have access to the most efficient accelerator for AI inference workloads -- gaining insights faster and accelerating time to action."
-- Kaustubh Das, vice president of product management, Data Center Group, Cisco

“Dell EMC is focused on helping customers transform their IT while benefiting from advancements such as artificial intelligence. As the world's leading provider of server systems, Dell EMC continues to enhance the PowerEdge server portfolio to help our customers ultimately achieve their goals. Our close collaboration with NVIDIA and historical adoption of the latest GPU accelerators available from their Tesla portfolio play a vital role in helping our customers stay ahead of the curve in AI training and inference."
-- Ravi Pendekanti, senior vice president of product management and marketing, Servers & Infrastructure Systems, Dell EMC

“Fujitsu plans to incorporate NVIDIA's Tesla T4 GPUs into our global Fujitsu Server PRIMERGY systems lineup. Leveraging this latest, high-efficiency GPU accelerator from NVIDIA, we will provide our customers around the world with servers highly optimized for their growing AI needs."
-- Hideaki Maeda, vice president of the Products Division, Data Center Platform Business Unit, Fujitsu Ltd.

“At HPE, we are committed to driving intelligence at the edge for faster insight and improved experiences. With the NVIDIA Tesla T4 GPU, based on the NVIDIA Turing architecture, we are continuing to modernize and accelerate the data center to enable inference at the edge."
-- Bill Mannel, vice president and general manager, HPC and AI Group, Hewlett Packard Enterprise

“IBM Cognitive Systems is able to deliver 4x faster deep learning training times as a result of a co-optimized hardware and software on a simplified AI platform with PowerAI, our deep learning training and inference software, and IBM Power Systems AC922 accelerated servers. We have a history of partnership and innovation with NVIDIA, and together we co-developed the industry's only CPU-to-GPU NVIDIA NVLink connection on IBM Power processors, and we are excited to explore the new NVIDIA T4 GPU accelerator to extend this state of the art leadership for inference workloads."
-- Steve Sibley, vice president of Power Systems Offering Management, IBM

“We are excited to see NVIDIA bring GPU inference to Kubernetes with the NVIDIA TensorRT inference server, and look forward to integrating it with Kubeflow to provide users with a simple, portable and scalable way to deploy AI inference across diverse infrastructures."
-- David Aronchick, co-founder and product manager of Kubeflow

“Open source cross-framework inference is vital to production deployments of machine learning models. We are excited to see how the NVIDIA TensorRT inference server, which brings a powerful solution for both GPU and CPU inference serving at scale, enables faster deployment of AI applications and improves infrastructure utilization."
-- Kash Iftikhar, vice president of product development, Oracle Cloud Infrastructure

“Supermicro is innovating to address the rapidly emerging high-throughput inference market driven by technologies such as 5G, Smart Cities and IOT devices, which are generating huge amounts of data and require real-time decision making. We see the combination of NVIDIA TensorRT and the new Turing architecture-based T4 GPU accelerator as the ideal combination for these new, demanding and latency-sensitive workloads and plan to aggressively leverage them in our GPU system product line."
-- Charles Liang, president and CEO, Supermicro

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 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/.

Certain statements in this press release including, but not limited to, statements as to: Tesla T4 GPU and TensorRT software enabling intelligent voice, video, image and recommender services; NVIDIA’s AI data center platform delivering the industry’s most advanced inference acceleration for voice, video, image and recommendation services; the benefits, performance and abilities of the NVIDIA TensorRT Hyperscale Inference Platform, including Tesla T4 GPUs based on Turing architecture and new inference software, its ability to deliver faster performance at lower latency than other offerings, and enabling hyperscale data centers to offer new services; customers racing toward a future where every product and service will be touched and improved by AI and the Tensor RT Hyperscale Platform being built to bring this to a reality faster and more efficiently than previously thought possible; the value the estimated AI inference industry will grow to in the next five years; the performance and features of Tesla T4 GPUs; NVIDIA TensorRT 5 expanding the set of neural network optimizations for mixed precision workloads; NVIDIA TensorRT inference server enabling applications to use AI models, its availability from the NVIDIA GPU Cloud container registry and its ability to maximize GPU utilization; NVIDIA GPUs enabling Microsoft to reduce object detection latency for images and Microsoft looking forward to working with NVIDIA’s next-generation inference hardware and software to expand the way people benefit from AI products and services; Google Cloud planning to add support for Tesla T4 GPUs on the Google Cloud Platform soon; AI becoming increasingly pervasive, and inference being a critical capability customers need to deploy AI models; major server manufacturers voicing their support for the NVIDIA TensorRT Hyperscale Platform; NVIDIA Tesla T4 GPUs giving Cisco customers access to the most efficient accelerator for AI inference workloads; Dell EMC enhancing the PowerEdge server portfolio to help customers and its collaboration with NVIDIA playing a vital role in helping its customers; Fujitsu’s plan to incorporate Tesla T4 GPUs into its systems lineup and providing its customers with servers optimized for their growing AI needs; HPE using Tesla T4 GPUs to continue to modernize and accelerate the data center to enable inference at the edge; IBM’s plans to explore the Tesla T4 GPU accelerator to extend its state of the art leadership for inference workloads; Kubernetes integrating NVIDIA products with Kubeflow and providing ways to deploy AI inference across diverse infrastructures; NVIDIA TensorRT inference server features enabling faster deployment of AI applications and improving infrastructure utilization; Supermicro innovating in markets which generate data and require real-time decision making and their plans to leverage NVIDIA products in their GPU system product line 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 most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. 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

Kristin Bryson
+1 203 241 9190
kbryson@nvidia.com