NVIDIA Announces Major Updates to Triton Inference Server as 25,000+ Companies Worldwide Deploy NVIDIA AI Inference

Capital One, Microsoft, Samsung Medison, Siemens Energy, Snap Among Industry Leaders Worldwide Using Platform

GTC—NVIDIA today announced major updates to its AI inference platform, which is now being used by Capital One, Microsoft, Samsung Medison, Siemens Energy and Snap, among its 25,000+ customers.

The updates include new capabilities in the open source NVIDIA Triton Inference Server™ software, which provides cross-platform inference on all AI models and frameworks, and NVIDIA TensorRT™, which optimizes AI models and provides a runtime for high-performance inference on NVIDIA GPUs.

The company also introduced the NVIDIA A2 Tensor Core GPU, a low-power, small-footprint accelerator for AI inference at the edge that offers up to 20x more inference performance than CPUs.

“NVIDIA’s AI inference platform is driving breakthroughs across virtually every industry, including healthcare, financial services, retail, manufacturing and supercomputing,” said Ian Buck, vice president and general manager of accelerated computing at NVIDIA. “Whether delivering smarter recommendations, harnessing the power of conversational AI, or advancing scientific discovery, NVIDIA’s platform for inference provides low-latency, high-throughput, versatile performance with the ease of use required to power key new AI applications worldwide.”

Key Software Optimizations
Updates to Triton Inference Server include:

- **Triton Model Analyzer** — This new tool automates a key optimization task by helping select the best configurations for AI models from hundreds of possibilities. It achieves the optimal performance while ensuring quality of service required for applications.
- **Multi-GPU Multinode Functionality** — This new functionality enables Transformer-based large language models, such as Megatron 530B, that no longer fit in a single GPU to be inferenced across multiple GPUs and server nodes and provides real-time inference performance.
- **RAPIDS FIL** — This new backend for GPU or CPU inference of random forest and gradient-boosted decision tree models provides developers a unified deployment engine for both deep learning and traditional machine learning with Triton.
- **Amazon SageMaker Integration** — This seamless integration allows customers to easily deploy multi-framework models with high performance using Triton within SageMaker, AWS’s fully managed AI service.
- **Support for Arm CPUs** — Triton now includes backends to optimize AI inference workloads on Arm CPUs, in addition to NVIDIA GPUs and x86 CPUs

Triton provides AI inference on GPUs and CPUs in the cloud, data center, enterprise edge and embedded, is integrated into AWS, Google Cloud, Microsoft Azure and Alibaba Cloud PAI-EAS, and is included in NVIDIA AI Enterprise.

NVIDIA AI Enterprise is an end-to-end software suite for development and deployment of AI. It is optimized, certified and supported by NVIDIA to enable customers to run AI workloads on mainstream servers in on-prem data centers and private clouds.

In addition to Triton, TensorRT is now integrated with TensorFlow and PyTorch, providing 3x faster performance versus inference in-framework with just one line of code. This provides developers with the power of TensorRT in a vastly simplified workflow.

NVIDIA TensorRT 8.2, the latest version of the SDK, accelerates high-performance, deep learning inference, delivering high throughput and low latency in the cloud, on premises or at the edge. With new optimizations, language models with billions of parameters can be run in real time.

Industry Leaders Embrace NVIDIA AI Platform for Inference
Industry leaders are using the NVIDIA AI inference platform to improve their business operations and offer customers new AI-enabled services.

Microsoft Azure Cognitive Services provides cloud-based APIs to high-quality AI models to create intelligent applications. It is using Triton to run speech-to-text models that provide Microsoft Teams users with accurate live captions and transcriptions.

“Microsoft Teams is an essential tool for communication and collaboration worldwide, with nearly 250 million monthly active users,” said Shalendra Chhabra, principal PM manager for Teams Calling and Meetings and Devices at Microsoft. “AI models like these are incredibly complex, requiring tens of millions of neural network parameters to deliver accurate results across dozens of different languages. The bigger a model is, the harder it is to run cost-effectively in real time. NVIDIA GPUs and Triton Inference Server on Microsoft Azure Cognitive Services are helping boost live captioning and transcription capabilities in a cost-effective way, using 28 languages and dialects, with AI in near real time.”

Samsung Medison, a global medical equipment company and an affiliate of Samsung Electronics, is using NVIDIA TensorRT to provide enhanced medical image quality using Intelligent Assist features for its ultrasound systems. Samsung Medison is dedicated to enhancing patient and healthcare professionals lives by enhancing their comfort, reducing scan time, simplifying workflow and ultimately increasing the system throughput.

“By leveraging NVIDIA TensorRT in the new coming V8 high-end Ultrasound system, we’re able to better support medical experts when reading and diagnosing images,” said Won-Chul Bang, vice president and head of the Customer Experience Team at Samsung Medison. “We are actively introducing AI-based technologies to our ultrasound systems for providing better support for medical professionals, so they can focus on the more important aspects of diagnosis and treatment of patients.”

Siemens Energy, a pure-play energy company with leading energy technology solutions, is using Triton to help its power plant customers manage their facilities with AI.

“The flexibility of NVIDIA Triton Inference Server is enabling highly complicated power plants, often equipped with cameras and sensors but with legacy software systems, to join the autonomous industrial revolution,” said Arik Ott, portfolio manager of autonomous operations at Siemens Energy.

Snap, the global camera and social media company comprising products and services such as Snapchat, Spectacles and Bitmoji, is using NVIDIA technology to improve monetization and lower costs.

“Snap used NVIDIA GPUs and TensorRT to improve machine learning inference cost-efficiency by 50 percent and decrease serving latency by 2x," said Nima Khajehnouri, vice president of engineering for the Mapping and Monetization Group at Snap. "This provides us the compute headroom to experiment and deploy heavier, more accurate ad and content ranking models.”

NVIDIA AI Platform for Inference Includes New NVIDIA-Certified Systems, New A2 GPU
NVIDIA-Certified Systems™ enable customers to identify, acquire and deploy systems for diverse modern AI applications on a high-performance, cost-effective and scalable infrastructure and now includes two new categories for edge AI.

The expanded categories allow NVIDIA’s systems partners to offer customers a complete lineup of NVIDIA-Certified Systems powered by NVIDIA Ampere architecture-based GPUs to handle virtually every workload. This includes the new NVIDIA A2 GPU, an entry-level, low-power, compact accelerator for inference and edge AI in edge servers. With the NVIDIA A30 for mainstream enterprise servers and the NVIDIA A100 for the highest performance AI servers, the addition of NVIDIA A2 delivers comprehensive AI inference acceleration across edge, data center and cloud.

Leading global enterprise system providers such as Atos, Dell Technologies, GIGABYTE, Hewlett Packard Enterprise, Inspur, Lenovo and Supermicro support NVIDIA AI Enterprise on NVIDIA-Certified Systems in their AI systems portfolios.

Additional system providers such as Advantech, ASRock Rack, ASUS, H3C, Nettrix and QCT also offer NVIDIA-Certified Systems for a variety of workloads. The first NVIDIA-Certified Systems to pass certification in the new edge categories will be available soon from leading providers including Advantech, GIGABYTE and Lenovo.

Availability
Triton is available from the NVIDIA NGC™ catalog, a hub for GPU-optimized AI software including frameworks, toolkits, pretrained models and Jupyter Notebooks, and as open source code from the Triton GitHub repository.

TensorRT is available to members of the NVIDIA Developer program from the TensorRT page. The latest versions of plugins, parsers and samples are also available as open source from the TensorRT GitHub repository.

Customers can experience NVIDIA Triton in the NVIDIA AI Enterprise software suite through curated labs available around the world in NVIDIA LaunchPad, announced separately today.

The NVIDIA AI Enterprise software suite is available from worldwide NVIDIA channel partners, including Atea, Axians, Carahsoft Technology Corp., Computacenter, Insight Enterprises, Presidio, Sirius, SoftServe, SVA System Vertrieb Alexander GmbH, TD SYNNEX, Trace3 and World Wide Technology.

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: the benefits, impact, features, and availability of the NVIDIA AI platform, the Triton Inference Server, NVIDIA TensorRT, NVIDIA AI Enterprise, NVIDIA A2 Tensor Core GPU, NVIDIA A30, NVIDIA A100, NVIDIA-Certified Systems and the NVIDIA NGC catalog; NVIDIA's AI inference platform driving breakthroughs across virtually every industry, including healthcare, financial services, retail, manufacturing and supercomputing; and industry leaders embracing the NVIDIA AI platform for inference 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.

Anna Kiachian
PR Manager
NVIDIA Corporation
+1-650-224-9820
akiachian@nvidia.com