NVIDIA AI Delivers Major Advances in Speech, Recommender System and Hyperscale Inference

Amazon, Microsoft, Snap, NTT Communications Deploy NVIDIA AI; NVIDIA Riva and Merlin Enter General Availability; NVIDIA AI Enterprise 2.0 Supports All Major Data Center and Cloud Platforms; NVIDIA AI Accelerated Program Launches

GTC—NVIDIA today announced major updates to its NVIDIA AI platform, a suite of software for advancing such workloads as speech, recommender system, hyperscale inference and more, which has been adopted by global industry leaders such as Amazon, Microsoft, Snap and NTT Communications.

The company also announced the NVIDIA® AI Accelerated program, which helps to ensure performance and reliability of AI applications developed by NVIDIA's software and solution partners. The program increases visibility to a wide range of proven AI-accelerated applications, enabling enterprise customers to deploy with confidence on the NVIDIA AI platform. Adobe, Red Hat and VMware are among the more than 100 partners participating at launch.

"NVIDIA AI is the software toolbox of the world's AI community — from AI researchers and data scientists, to data and machine learning operations teams," said Jensen Huang, founder and CEO of NVIDIA. "Our GTC 2022 release is massive. Whether it's creating more engaging chatbots and virtual assistants, building smarter recommendations to help consumers make better purchasing decisions, or orchestrating AI services at the largest scales, your superpowered gem is in NVIDIA AI."

Freely available for developers, NVIDIA AI includes NVIDIA Riva for speech AI and NVIDIA Merlin™ for smart recommendations, now both generally available. Updates have also been made across the software suite, including tools such as the NVIDIA Triton, NeMo, Maxine and TAO Toolkit.

Additionally, NVIDIA AI Enterprise 2.0 is now optimized, certified and supported across every major data center and cloud platform, including bare-metal servers, virtualized infrastructure and CPU-only systems. The suite is now supported with Red Hat OpenShift and VMware vSphere with Tanzu.

Software Tools to Build Industry-Leading AI Applications
NVIDIA AI is comprised of key enabling SDKs and tools for rapid deployment, management and scaling of AI workloads across multiple nodes to power complex training and machine learning workloads. These include:

- **NVIDIA Triton™** - NVIDIA Triton is a versatile open-source hyperscale model inference solution. The latest release contains three key updates: a Model Navigator for accelerated deployment of optimized models, Management Service for efficient scaling in Kubernetes, and Forest Inference Library enabling inference on tree-based models with explainability for fast, optimized and scalable AI in every application.
- **NVIDIA Riva 2.0** - A world-class speech AI SDK that includes models pretrained with leading recognition rates, enabling developers to customize real-time speech AI applications for their industry with 2x better accuracy over generic services. Riva 2.0 includes speech recognition in seven languages, human-like deep learning-based text-to-speech with both male and female voices, as well as custom tuning with NVIDIA TAO Toolkit. NVIDIA also announced NVIDIA Riva Enterprise, a paid offering that includes enterprise support from NVIDIA.
- **NVIDIA NeMo Megatron 0.9** - A framework for training large language models (LLMs), NeMo Megatron enables researchers and enterprises to train any model to convergence and scale to trillions of parameters for applications such as conversational AI, recommenders and genomics. The latest version includes new optimizations and recipes that shorten end-to-end development and training time, and adds support for training in the cloud. Early users of LLMs on NVIDIA accelerated computing include JD.com, AI Sweden, Naver and the University of Florida.
- **NVIDIA Merlin 1.0** - An accelerated, end-to-end recommender AI framework to build high-performing recommenders at scale, which includes two new libraries: Merlin Models and Merlin Systems. These allow data scientists and machine learning engineers to determine which features and models are the best fit for their use case and deploy recommender pipelines as microservices.
- **NVIDIA Maxine** - An audio and video quality enhancement SDK that reinvents real-time communications with AI, and introduces acoustic echo cancellation and audio super resolution. The two new features enhance audio quality leading to a clearer communication experience.

Customers Embrace NVIDIA AI
Industry leaders are using NVIDIA AI to improve cost-efficiency, create more engaging customer experiences and optimize AI application capabilities.

“On Snapchat, our community plays with Lenses over 6 billion times per day,” said Alan Bekker, head of Conversational AI at Snap. “Snap is using NVIDIA Riva to optimize our AI-based speech capabilities and offer them to Lens Studio creators to build a new generation of compelling AR experiences.”

“The document translation feature within Translator, a Microsoft Azure Cognitive Service, enables efficient business to customer interactions by translating documents retaining format and structure as in source document,” said Xuedong Huang, Microsoft Technical Fellow and Azure AI CTO. “Using NVIDIA Triton, we’re able to deploy the latest Z-Code model to achieve significantly improved document translation quality with the low latency, providing our users with unmatched translation services.”

NVIDIA AI Enterprise Supports Containerized AI Across Data Centers and Cloud
An end-to-end, cloud-native suite of AI and data analytics tools and frameworks, the NVIDIA AI Enterprise 2.0 software suite accelerates AI development and deployment for industries. Certification for Red Hat OpenShift, the industry’s leading enterprise Kubernetes platform, enables customers to use containerized machine learning tools to more easily build, scale and share their models on bare-metal, or virtualized systems, with VMware vSphere.

“The certification of Red Hat OpenShift for NVIDIA AI Enterprise, and the availability of OpenShift on NVIDIA LaunchPad, unites top tools for AI development with a consistent hybrid cloud foundation,” said Stefanie Chiras, senior vice president of Partner Ecosystem Success at Red Hat. “Now, IT teams and data scientists can build and manage NVIDIA AI on Red Hat OpenShift, helping enterprises accelerate the delivery of intelligent applications in production.”

NVIDIA AI Enterprise 2.0 also introduces more NVIDIA AI software containers to support training and inference. Support for the NVIDIA TAO Toolkit allows enterprise developers to fine-tune and optimize NVIDIA pretrained AI models, simplifying the creation of custom, production-ready models, without AI expertise or large training data. The latest release of NVIDIA Triton Inference Server is also included in the software suite.

NTT Communications, the leading global provider of information and communications technology solutions within NTT Group, has adopted NVIDIA AI Enterprise to accelerate its research and development teams building NLP and intelligent video analytics applications.

“Many of our application developers now use accelerated computing, and are in need of an internal infrastructure that provides an easy-to-use, cost-effective GPU-enabled environment,” said Shoichiro Henmi, director of Technology Division, Innovation Center, NTT Communications. “We are confident that NVIDIA AI Enterprise will provide an ideal solution as an AI-enabled platform to support large-scale development in our VMware vSphere, Kubernetes and cloud infrastructure.”

Customers can license NVIDIA AI Enterprise to run on NVIDIA-Certified Systems™, or on the same server models without NVIDIA GPUs, from leading manufacturers including Cisco, Dell Technologies, H3C, Hewlett Packard Enterprise, Inspur, Lenovo, Nettrix and Supermicro.

Enterprises can also choose to deploy on servers hosted at Equinix International Business Exchange™ (IBX®) data centers worldwide. Cloud instances from Amazon Web Services, Google Cloud and Microsoft Azure are also now supported. NVIDIA AI Enterprise is included with NVIDIA DGX systems.

Availability
NVIDIA’s collection of AI software is available to developers as part of the NVIDIA Developer Program.

Enterprises can experience NVIDIA AI software in curated labs for IT teams and AI developers on NVIDIA LaunchPad, a hosted program powered by the Equinix Metal service in nine Equinix IBX data centers worldwide, available at no charge. New LaunchPad labs include a speech AI lab for Riva, as well as NVIDIA AI Enterprise labs featuring Red Hat OpenShift, VMware vSphere with Tanzu, TAO Toolkit and Triton Inference Server with FIL backend.

To learn more about NVIDIA AI, watch Huang’s GTC 2022 keynote. Register for GTC for free to attend sessions with NVIDIA and industry leaders.

About NVIDIA
NVIDIA’s (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market and has redefined modern computer graphics, high performance computing and artificial intelligence. The company’s pioneering work in accelerated computing and AI is reshaping trillion-dollar industries, such as transportation, healthcare and manufacturing, and fueling the growth of many others. More information at https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, performance and availability of the NVIDIA AI platform, including NVIDIA Triton Inference Server, NVIDIA Riva 2.0, NVIDIA NeMo Megatron 0.9, NVIDIA Merlin 1.0, NVIDIA Maxine, NVIDIA AI Enterprise 2.0 and the NVIDIA AI Accelerated program; NVIDIA AI giving developers and enterprises the tools they need to build applications to transform nearly every industry; and our software improving business operations and enabling customers to offer new AI-enabled services 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.

Features, pricing, availability, and specifications are subject to change without notice.

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