NVIDIA Launches Edge Computing Platform to Bring Real-Time AI to Global Industries

Leading Computer Makers Adopt NVIDIA EGX Platform, Offering GPU Edge Servers for Instant AI on Real-Time Streaming Data in Telecom, Healthcare, Manufacturing, Retail, Transportation

Computex -- NVIDIA today announced NVIDIA EGX, an accelerated computing platform that enables companies to perform low-latency AI at the edge -- to perceive, understand and act in real time on continuous streaming data between 5G base stations, warehouses, retail stores, factories and beyond.

NVIDIA EGX was created to meet the growing demand to perform instantaneous, high-throughput AI at the edge -- where data is created - with guaranteed response times, while reducing the amount of data that must be sent to the cloud.

By 2025, 150 billion machine sensors and IoT devices will stream continuous data that will need to be processed[(1)] -- orders of magnitude more than produced today by individuals using smartphones. Edge servers like those in the NVIDIA EGX platform will be distributed throughout the world to process data in real time from these sensors.

“Enterprises demand more powerful computing at the edge to process their oceans of raw data -- streaming in from countless interactions with customers and facilities -- to make rapid, AI-enhanced decisions that can drive their business," said Bob Pette, vice president and general manager of Enterprise and Edge Computing at NVIDIA. “A scalable platform like NVIDIA EGX allows them to easily deploy systems to meet their needs on premises, in the cloud or both."

Scalability
EGX starts with the tiny NVIDIA Jetson Nano™, which in a few watts can provide one-half trillion operations per second (TOPS) of processing for tasks such as image recognition. And it spans all the way to a full rack of NVIDIA T4 servers, delivering more than 10,000 TOPS for real-time speech recognition and other real-time AI tasks.

Enterprise-Grade
NVIDIA has partnered with Red Hat to integrate and optimize NVIDIA Edge Stack with OpenShift, the leading enterprise-grade Kubernetes container orchestration platform.

NVIDIA Edge Stack is optimized software that includes NVIDIA drivers, a CUDA® Kubernetes plugin, a CUDA container runtime, CUDA-X™ libraries and containerized AI frameworks and applications, including TensorRT™, TensorRT Inference Server and DeepStream. NVIDIA Edge Stack is optimized for certified servers and downloadable from the NVIDIA NGC™ registry.

“Red Hat is committed to providing a consistent experience for any workload, footprint and location, from the hybrid cloud to the edge," said Chris Wright, chief technology officer at Red Hat. “By combining Red Hat OpenShift and NVIDIA EGX-enabled platforms, customers can better optimize their distributed operations with a consistent, high-performance, container-centric environment."

An “On-Prem AI Cloud-in-a-Box"
EGX combines the full range of NVIDIA AI computing technologies with Red Hat OpenShift and NVIDIA Edge Stack together with Mellanox and Cisco security, networking and storage technologies. This enables companies in the largest industries -- telecom, manufacturing, retail, healthcare and transportation -- to quickly stand up state-of-the-art, secure, enterprise-grade AI infrastructures.

“Mellanox Smart NICs and switches provide the ideal I/O connectivity for data access that scale from the edge to hyperscale data centers," said Michael Kagan, chief technology officer at Mellanox Technologies. “The combination of high-performance, low-latency and accelerated networking provides a new infrastructure tier of computing that is critical to efficiently access and supply the data needed to fuel the next generation of advanced AI solutions on edge platforms such as NVIDIA EGX."

“Cisco is excited to collaborate with NVIDIA to provide edge-to-core full stack solutions for our customers, leveraging Cisco's EGX-enabled platforms with Cisco compute, fabric, storage, and management software and our leading Ethernet and IP-based networking technologies," said Kaustubh Das, vice president of Cisco Computing Systems.

Enables Hybrid-Cloud and Multi-Cloud IoT
NVIDIA AI computing is offered by major clouds and is architecturally compatible with NVIDIA EGX. AI applications developed in the cloud can run on NVIDIA EGX and vice versa. NVIDIA Edge Stack connects to major cloud IoT services, and customers can remotely manage their service from AWS IoT Greengrass and Microsoft Azure IoT Edge.

“Azure IoT Edge helps customers deploy cloud service to their IoT devices quickly and securely," said Sam George, director of Azure IoT Edge. “We look forward to supporting NVIDIA's EGX edge platform on Azure IoT Edge devices so that customers can deploy AI workloads targeting EGX-compatible hardware."

Widespread Developer Support
NVIDIA EGX is optimizing AI at the edge for a growing ecosystem of software solutions.

These include video analytics applications, which are ideal for large retail chains and smart cities, from software vendors such as AnyVision, DeepVision, IronYun and Malong Technologies, as well as healthcare-specific software offerings from 12 Sigma, Infervision, Quantib and Subtle Medical.

Adoption by World's Top Computer Makers
EGX servers are available from global enterprise computing providers ATOS, Cisco, Dell EMC, Fujitsu, Hewlett Packard Enterprise, Inspur and Lenovo. They are also available from major server and IoT system makers Abaco, Acer, ADLINK, Advantech, ASRock Rack, ASUS, AverMedia, Cloudian, Connect Tech, Curtiss-Wright, GIGABYTE, Leetop, MiiVii, Musashi Seimitsu, QCT, Sugon, Supermicro, Tyan, WiBase and Wiwynn.

NVIDIA EGX servers are tuned for NVIDIA Edge Stack and NGC-Ready validated for CUDA-accelerated containers.

Support from 40+ Companies, Organizations
Early adopters include more than 40 industry-leading companies and organizations.

Among them is BMW Group Logistics. Drawing from NVIDIA's EGX edge computing and Isaac robotic platforms, they are able to bring the power of AI directly to the edge of its logistics processes and handle increasingly complex logistics with real-time efficiency.

Other industry leaders adopting EGX include:

“Foxconn PC production lines are limited by the speed of inspection because it currently requires four seconds to manually inspect each part. Our goal is to increase the throughput of the PC production line by over 40 percent using the NVIDIA EGX platform for real-time intelligent decision-making at the edge. Our model detects and classifies 16 defect types and locations simultaneously using fast neural networks running on NVIDIA GPUs, achieving 98 percent accuracy at a superhuman throughput rate."
-- Mark Chien, general manager, Foxconn D Group

“AI is fundamental to achieving precision health and must be pervasively available from the cloud to the edge and directly on medical devices. NVIDIA's EGX enables GE Healthcare to deliver rapid MR acquisition times, improves image quality and reduces variability by embedding NVIDIA T4 GPUs directly into our medical devices -- all to further our goal of improving patient outcomes. Real-time, critical-care use cases demand AI at the edge. This is why we created our Edison intelligence offering and partnered with NVIDIA to bring AI into our medical devices and Edison edge appliances -- and why we are working with ACR AI-LAB to democratize AI."
-- Jason Polzin, Ph.D., general manager of MR Applications, GE Healthcare

“Hospitals are increasingly using AI to predict adverse patient events, support clinical decision-making and operate more efficiently. However, these AI applications rely on patient data. NVIDIA's EGX AI edge computing platform provides hospitals easy AI infrastructure to keep patient data secure, deliver real-time AI and scale to thousands of AI applications that are needed to improve patient care and reduce the cost of care delivery."
-- Keith Dreyer, D.O., Ph.D., chief data science officer at Partners Healthcare and associate professor of radiology, Harvard Medical School

“At Seagate we have deployed an intelligent edge GPU-based vision solution in our manufacturing plants to inspect the quality of our hard disk read-and-write heads. The NVIDIA EGX platform dramatically accelerates inference at the edge, allowing us to see subtle defects that human operators haven't been able to see in the past. We expect to realize up to a 10 percent improvement in manufacturing throughput and up to 300 percent ROI from improved efficiency and better quality."
-- Bruce King, senior principal data scientist, Seagate 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/.

1. IDC white paper, sponsored by Seagate, “Data Age 2025: The Digitization of the World from Edge to Core,” November 2018.

Certain statements in this press release including, but not limited to, statements as to: NVIDIA launching an edge computing platform to bring real-time AI to global industries; leading computer makers adopting the NVIDIA EGX platform, and it offering GPU Edge servers for instant AI on real-time streaming data in industries; the benefits, performance, features, abilities and impact of NVIDIA EGX, NVIDIA Jetson Nano, NVIDIA AI computing and NVIDIA T4 servers; NVIDIA EGX enabling companies to perform low-latency AI at the edge and it being able to perceive, understand and act in real time on continuous data streaming; NVIDIA EGX meeting the growing demand to perform AI at the edge and its abilities; the year by which 150 billion machine sensors and IoT devices streaming continuous data that will need to be processed; edge servers being distributed throughout the world to process data; enterprises demanding more powerful computing at the edge to process data and drive their business and NVIDIA EGX allowing them to deploy systems to meet their needs; the ability to use EGX to deploy AI quickly and securely from edge to cloud; the benefits of combining NVIDIA EGX, Red Hat OpenShift and NVIDIA Edge Stack and its performance for customers; EGX, Red Hat OpenShift, NVIDIA Edge Stack, and Mellanox and Cisco security enabling companies to stand up AI infrastructures; NVIDIA EGX working with Mellanox technologies to fuel the next generation of advanced AI solutions on the edge; Cisco’s excitement to collaborate with NVIDIA and its impacts; Azure IoT looking forward to supporting NVIDIA’s EGX platform on its devices so customers can deploy AI workloads targeting EGX-compatible hardware; the software solutions enabling AI at the edge; the top computer makers adopting NVIDIA EGX and its availability; the goals, impact and support of customers planning to use NVIDIA EGX; AI being fundamental to achieving precision health and needing to be pervasively available from the cloud to the edge and directly on medical devices; NVIDIA EGX enabling the delivery of rapid MR acquisition times, improving image quality and reducing variability by embedding NVIDIA T4 GPUs directly into medical devices; critical-care use cases demanding AI at the edge; NVIDIA and GE Healthcare partnering to bring AI to devices and working to democratize AI; high-performance, low-latency and accelerated networking providing a new infrastructure tier of computing that is critical to efficiently access and supply the data needed for the next generation of advanced AI solutions; hospitals increasingly using AI to predict adverse patient events, support clinical decision-making and operate more efficiently; NVIDIA EGX AI providing hospitals easy AI infrastructure to keep patient data secure, deliver real-time AI and scale to thousands of AI applications; and NVIDIA EGX accelerating inference at the edge and allowing the detection of defects humans haven’t been able to see in the past 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