NVIDIA EGX Edge AI Platform Brings Real-Time AI to Manufacturing, Retail, Telco, Healthcare and Other Industries

Ecosystem Expands with EGX A100 and EGX Jetson Xavier NX Supported by AI-Optimized, Cloud-Native, Secure Software to Power New Wave of 5G and Robotics Applications

GTC 2020 -- NVIDIA today announced two powerful products for its EGX Edge AI platform -- the EGX A100 for larger commercial off-the-shelf servers and the tiny EGX Jetson Xavier NX for micro-edge servers -- delivering high-performance, secure AI processing at the edge.

With the NVIDIA EGX™ Edge AI platform, hospitals, stores, farms and factories can carry out real-time processing and protection of the massive amounts of data streaming from trillions of edge sensors. The platform makes it possible to securely deploy, manage and update fleets of servers remotely.

The EGX A100 converged accelerator and EGX Jetson Xavier NX micro-edge server are created to serve different size, cost and performance needs. Servers powered by the EGX A100 can manage hundreds of cameras in airports, for example, while the EGX Jetson Xavier NX is built to manage a handful of cameras in convenience stores. Cloud-native support ensures the entire EGX lineup can use the same optimized AI software to easily build and deploy AI applications.

“The fusion of IoT and AI has launched the 'smart everything' revolution," said Jensen Huang, founder and CEO of NVIDIA. “Large industries can now offer intelligent connected products and services like the phone industry has with the smartphone. NVIDIA's EGX Edge AI platform transforms a standard server into a mini, cloud-native, secure, AI data center. With our AI application frameworks, companies can build AI services ranging from smart retail to robotic factories to automated call centers."

EGX A100 Powered by NVIDIA Ampere Architecture

The EGX A100 is the first edge AI product based on the NVIDIA Ampere architecture. As AI moves increasingly to the edge, organizations can include EGX A100 in their servers to carry out real-time processing and protection of the massive amounts of streaming data from edge sensors.

It combines the groundbreaking computing performance of the NVIDIA Ampere architecture with the accelerated networking and critical security capabilities of the NVIDIA Mellanox® ConnectX®-6 Dx SmartNIC to transform standard and purpose-built edge servers into secure, cloud-native AI supercomputers.

The NVIDIA Ampere architecture -- the company's eighth-generation GPU architecture -- delivers the largest-ever generational leap in performance for a wide range of compute-intensive workloads, including AI inference and 5G applications running at the edge. This allows the EGX A100 to process high-volume streaming data in real time from cameras and other IoT sensors to drive faster insights and higher business efficiency.

“Data, AI and intelligent cloud-native applications are transforming the enterprise edge in every industry," said Chris Wright, senior vice president and chief technology officer at Red Hat. “NVIDIA's new EGX A100 converged accelerators combined with precompiled drivers for Red Hat Enterprise Linux and certified operators for Red Hat OpenShift simplify deployment and management of the hardware and help our joint customers address some of the most demanding AI, edge and 5G workloads."

With an NVIDIA Mellanox ConnectX-6 Dx network card onboard, the EGX A100 can receive up to 200 Gbps of data and send it directly to the GPU memory for AI or 5G signal processing. With the introduction of NVIDIA Mellanox's time-triggered transport technology for telco (5T for 5G), EGX A100 is a cloud-native, software-defined accelerator that can handle the most latency-sensitive use cases for 5G. This provides the ultimate AI and 5G platform for making intelligent real-time decisions at the points of action -- stores, hospitals and factory floors.

“We've been collaborating with NVIDIA to build Mavenir's high-performance virtualized 5G radio access network and accelerated 5G packet core network," said Pardeep Kohli, president and CEO of Mavenir. “This will enable us to deliver a wide range of new GPU-accelerated 5G services from AI/ML to AR/VR applications."

Small But Mighty EGX Jetson Xavier NX

The EGX Jetson Xavier NX is the world's smallest, most powerful AI supercomputer for microservers and edge AIoT boxes, with more than 20 solutions now available from ecosystem partners. It packs the power of an NVIDIA Xavier SoC into a credit card-size module. EGX Jetson Xavier NX, running the EGX cloud-native software stack, can quickly process streaming data from multiple high-resolution sensors.

The energy-efficient module delivers up to 21 TOPS at 15W, or 14 TOPS at 10W. As a result, EGX Jetson Xavier NX opens the door for embedded edge-computing devices that demand increased performance to support AI workloads but are constrained by size, weight, power budget or cost.

“NVIDIA Jetson and NVIDIA EGX are helping us transform retail, making the self-checkout experience quicker and more secure," said Matt Scott, cofounder and CEO of Malong Technologies. “Through the power of AI, via our RetailAI suite of offerings, it is now possible to accurately recognize hundreds of thousands of products in real time to create more seamless and protected shopping experiences, easily deployable at large scale. We're continuing to explore NVIDIA's powerful lineup to discover new ways to increase customer satisfaction and decrease retail shrink, by bringing more intelligence to the edge."

Fully Optimized, Cloud-Native Software Across the EGX Edge AI Platform

The EGX Edge AI platform's cloud-native architecture allows it to run containerized software to support a range of GPU-accelerated workloads.

NVIDIA application frameworks include Clara for healthcare, Aerial for telcos, Jarvis for conversational AI, Isaac for robotics, and Metropolis for smart cities, retail, transportation and more. They can be used together or individually and open new possibilities for a variety of edge use cases.

With support for cloud-native technologies now available across the entire NVIDIA EGX lineup, manufacturers of intelligent machines and developers of AI applications can build and deploy high-quality, software-defined features on embedded and edge devices targeting robotics, smart cities, healthcare, industrial IoT and more.

Global Support for EGX Ecosystem

Existing edge servers enabled with NVIDIA EGX software are available from global enterprise computing providers Atos, Dell Technologies, Fujitsu, GIGABYTE, Hewlett Packard Enterprise, IBM, Inspur, Lenovo, Quanta/QCT and Supermicro. They are also available from major server and IoT system makers such as ADLINK and Advantech.

These servers along with optimized application frameworks can be used by software vendors such as Whiteboard Coordinator, Deep Vision AI, IronYun, Malong and SAFR by RealNetworks to build and deploy healthcare, retail, manufacturing and smart cities solutions.

Availability

The EGX A100 will be available at the end of the year. Ready-to-deploy micro-edge servers based on the EGX Jetson Xavier NX are available now for companies looking to create high-volume production edge systems.

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 https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, performance, features and availability of our products and technologies, including the EGX A100, EGX Jetson Xavier NX, NVIDIA Mellanox ConnectX-6 Dx SmartNIC, and NVIDIA Ampere architecture; and AI revolutionizing every industry and transforming everything 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.

Kristin Uchiyama
Enterprise and Edge Computing
+1-408-486-2248
kuchiyama@nvidia.com