NVIDIA Unveils GPU-Accelerated AI-on-5G System for Edge AI, 5G and Omniverse Digital Twins

Built with NTT DOCOMO, Fujitsu and others, the all-in-one solution enables telcos to deliver immersive graphics, metaverse applications and computer vision from a single server.

Author: Soma Velayutham

Telcos are seeking industry-standard solutions that can run 5G, AI applications and immersive graphics workloads on the same server — including for computer vision and the metaverse .

To meet this need, NVIDIA is developing a new AI-on-5G solution that combines 5G vRAN, edge AI and digital twin workloads on an all-in-one, hyperconverged and GPU-accelerated system.

The lower cost of ownership enabled by such a system would help telcos drive revenue growth in smart cities, as well as the retail, entertainment and manufacturing industries, to support a multitrillion-dollar, 5G-enabled ecosystem.

The AI-on-5G system consists of:

Fujitsu's virtualized 5G Open RAN product suite, which was developed as part of the 5G Open RAN ecosystem experience (OREX) project promoted by NTT DOCOMO. It also includes Fujitsu's virtualized central unit (vCU) and distributed unit (vDU), plus other virtualized software functions of vRAN from Fujitsu.

The NVIDIA Aerial software development kit for 5G vRAN; NVIDIA Omniverse for building and operating custom 3D pipelines and large-scale simulations; NVIDIA RTX Virtual Workstation (vWS) software; and NVIDIA CloudXR for streaming extended reality.

Hardware includes the NVIDIA A100X and L40 converged accelerators.

OREX has supported performance verification and evaluation tests for this system.

"Fujitsu is delivering a fully virtualized 5G vRAN together with multi-access edge computing on the same high-performance, energy-efficient, versatile and scalable computing infrastructure," said Masaki Taniguchi, senior vice president and head of mobile systems at Fujitsu. "This combination, powered by AI and XR applications, enables telcos to deliver ultra-low latency services, highly optimized TCO and energy-efficient performance."

The announcement is a step toward accomplishing the O-RAN alliance's goal of enabling software-defined, AI-driven, cloud-native, fully programmable, energy-efficient and commercially ready telco-grade 5G Open RAN solutions. It's also consistent with OREX's goal of implementing a widely adopted, high-performance and multi-vendor 5G vRAN for both public and enterprise 5G deployments.

The all-in-one system uses GPUs to accelerate the software-defined 5G vRAN, as well as the edge AI and graphics applications, without bespoke hardware accelerators nor a specific telecom CPU. This ensures that the GPUs can accelerate the vRAN (based on NVIDIA Aerial), AI video analytics (based on NVIDIA Metropolis), streaming immersive extended reality (XR) experiences (based on NVIDIA CloudXR) and digital twins (based on NVIDIA Omniverse).

"Telcos and their customers are exploring new ways to boost productivity, efficiency and creativity through immersive experiences delivered over 5G networks," said Ronnie Vasishta, senior vice president of telecom at NVIDIA. "At Mobile World Congress, we are bringing those visions into reality, showcasing how a single GPU-enabled server can support workloads such as NVIDIA Aerial for 5G, CloudXR for streaming virtual reality and Omniverse for digital twins."

The AI-on-5G system is part of a growing portfolio of 5G solutions from NVIDIA that are driving transformation in the telecommunications industry. Anchored on the NVIDIA Aerial SDK and A100X converged accelerators — combined with BlueField DPUs and a suite of AI frameworks — NVIDIA provides a high-performance, software-defined, cloud-native, AI-enabled 5G for on-premises and telco operators' RAN.

Telcos working with NVIDIA can gain access to thousands of software vendors and applications in the ecosystem, which can help address enterprise needs in smart cities, retail, manufacturing, industrial and mining.

NVIDIA and Fujitsu will demonstrate the new AI-on-5G system at Mobile World Congress in Barcelona, running Feb. 27-March 2, at hall 4, stand 4E20.

Original URL: https://blogs.nvidia.com/blog/2023/02/27/mwc-ai-on-5g-system/