# QCT Builds the Next Frontier of AI With NVIDIA at GTC 2026

*Powering the AI Grid across data center and edge sites with QCT Application-Ready Solutions*

SAN JOSE, CA, UNITED STATES, March 16, 2026 /EINPresswire.com/ -- Quanta Cloud Technology (QCT), a leading provider of AI solutions, today unveiled its NVIDIA Vera CPU and NVIDIA Rubin GPU accelerated servers and integrated solutions for data center and edge at NVIDIA GTC 2026 (GTC26). Expanding upon QCT's years of collaboration with NVIDIA, QCT has applied its Application-Ready Solution approach for faster, cross-industry deployment across multiple sites in the AI Grid. For this purpose. Under the theme, "AI: From Data Center to Edge to End Devices," QCT leverages the latest NVIDIA platforms into their Application-Ready Solutions by integrating software, high-performance hardware, advanced networking, and physical AI technologies into scalable cloud-to-edge infrastructures.

> Our innovations with NVIDIA underscore a shared mission to make AI infrastructure more powerful, more efficient, and more accessible."
>
> *Mike Yang, President of QCT*

"Our innovations with NVIDIA underscore a shared mission to make AI infrastructure more powerful, more efficient, and more accessible," said Mike Yang, President of QCT. "NVIDIA Vera Rubin NVL72 platform marks a major milestone toward the next era of AI infrastructures —one defined by extreme scalability, that will shape the future of AI."

Support for NVIDIA Vera Rubin NVL72

QCT's upcoming QuantaGrid D76V-1U will support the NVIDIA Vera Rubin NVL72 platform and is engineered for extreme AI performance, powered by 3,168 custom NVIDIA Olympus cores and 72 Rubin GPUs delivering 3,600 PFLOPS of NVFP4 inference compute. Each GPU features 288GB of HBM4 with higher bandwidth and faster NVIDIA NVLink 6 interconnect compared to the

previous generation, enabling unprecedented throughput for large-scale AI model training and inference. With NVIDIA ConnectX-9 SuperNIC delivering up to 1.6 Tb/s of networking bandwidth per GPU, the system ensures optimal data flow across massive AI clusters. Its 100% liquid-cooled NVIDIA Vera Rubin NVL72 architecture maximizes compute density and energy efficiency, supporting ESG-driven sustainability goals for high-power AI infrastructure deployments.

“AI infrastructure and systems span cloud data centers, enterprise and the far edge,” said Kaustubh Sanghani, vice president of product management at NVIDIA. “NVIDIA Vera CPUs and Rubin GPUs, combined with our advanced networking and AI software platforms, provide the foundation for scalable AI infrastructure. Working with QCT, we’re enabling organizations to deploy AI systems that deliver powerful intelligence.”

New QCT Multi-node Server Powered by NVIDIA Vera CPU

The QuantaPlex S26F-2U, powered by NVIDIA Vera CPU, is designed for agentic reasoning. Equipped with NVIDIA BlueField-4 data processor, the system supports NVIDIA Context Memory Storage Platform (CMX), delivering a new class of AI-native storage infrastructure for gigascale inference by adding a dedicated, pod-level context memory tier between GPU memory and traditional shared storage. QCT has equipped this new multi-node server with 24 front-accessible, hot-swappable E3.S 1T NVMe bays, delivering fast and dense flash capacity for AI workloads.

Growing AI Server Portfolio based on NVIDIA HGX

The QuantaGrid D76T-2U is a compact 8-GPU system optimized for AI and HPC workloads that will also support NVIDIA HGX Rubin NVL8. Boasting over 2x performance for the most demanding agentic AI workloads, the D76T-2U also supports NVIDIA BlueField-4 DPUs and NVIDIA ConnectX-9 networking, enabling up to 800 Gb/s NVIDIA Quantum-X800 InfiniBand or NVIDIA Spectrum-X Ethernet, enabling ultra-fast cluster-scale data movement.

Simplifying Data Center Buildouts with QCT AI POD

QCT AI POD integrates QCT’s rack-scale data center infrastructure with either NVIDIA AI Enterprise or open-source software to simplify AI deployment. The solution supports NVIDIA platforms including NVIDIA Vera Rubin NVL72, NVIDIA GB300 NVL72, NVIDIA HGX Rubin NVL8, NVIDIA HGX B300, and 4U NVIDIA RTX PRO™ 6000 Blackwell Server Edition based on NVIDIA MGX, along with pre-validated tools for provisioning and workload automation.

QCT AI POD supports NVIDIA NIM inference microservices and NVIDIA NeMo Agent Toolkit, providing production-ready inference services and emerging agentic AI environments for administrators and developers. The pre-configured platform simplifies infrastructure operations while providing streamlined development environments for building, testing, and deploying AI applications. At GTC, QCT is showcasing live demonstrations of streamlined workflows and

accelerated development cycles.

Beyond infrastructure, QCT also delivers end-to-end AI cluster capabilities through close collaboration with system integrator, ITOCHU Techno-Solutions Corporation (CTC), enabling our mutual customers to confidently consult, design, deploy and operate production-grade AI clusters from day zero.

Advancing Edge AI with New NVIDIA ARC-Pro Platform

At GTC, QCT is also highlighting its latest QuantaEdge EGN77C-2U, accelerated by the NVIDIA RTX PRO™ 4500 Blackwell Server Edition GPU and acting as a foundation for AI-RAN deployments across diverse telecom operator environments. Built on NVIDIA Aerial RAN Computer Pro (NVIDIA ARC-Pro) platform and aligned with NVIDIA AI Aerial ecosystem, QCT's QuantaEdge EGN77C-2U enables operators to dynamically scale compute, networking, and AI capabilities from centralized sites to the far edge. By supporting AI-RAN workloads on the NVIDIA ARC-Pro platform, QCT helps operators maximize infrastructure utilization, improve energy efficiency, and unlock new service opportunities—accelerating the path from advanced 5G toward AI-native network architectures designed for the 6G era. Additionally, QCT previously validated AI-RAN on existing NVIDIA MGX architectures with its ecosystem partners for enterprise deployments, combining AI-native RAN software with QCT servers to enable scalable AI-powered networks.

Empowering Physical AI Innovations at the Edge

QCT realizes its Application-Ready Solutions to accelerate time-to-market, enabling faster deployment of industrial robotics. Building on this approach, QCT and Techman Robot are also demonstrating a physical AI development workflow using the Techman Robot TM Xplore I humanoid robot integrated with QCT's Dev Kit for physical AI and GPU servers. The live demo showcases the robot performing bimanual manipulation tasks, powered by NVIDIA Cosmos and NVIDIA Isaac GR00T open foundation models, and NVIDIA Omniverse libraries, then deployed to the physical robot accelerated by NVIDIA robotics.
QCT continues to advance AI infrastructures through alignment with a strong partner ecosystem. Visit Booth #1331 during NVIDIA GTC26 to learn more about these technologies.

About Quanta Cloud Technology (QCT)

Quanta Cloud Technology (QCT) is a global data center solution provider. We combine the efficiency of hyperscale hardware with infrastructure software from a diversity of industry leaders to solve next-generation data center design and operation challenges. QCT serves cloud service providers, telecoms, and enterprises running public, hybrid, and private clouds.



Jean Ko

QCT
+886 912025348
jean_ko@quantatw.com
Visit us on social media:
LinkedIn
Facebook
X

---