# QCT Showcases Server Portfolio for Building Co-Designed AI Infrastructures Accelerated by NVIDIA at Computex 2026

*Enabling organizations to build, scale, and optimize agentic AI with NVIDIA DSX AI factory-scale platform*

TAIPEI, TAIWAN, June 2, 2026 /EINPresswire.com/ -- Quanta Cloud Technology (QCT), a leading provider of AI and 5G solutions, today is showcasing its NVIDIA-accelerated server portfolio for co-designed AI Infrastructures at Computex 2026. QCT's systems integrate the latest NVIDIA 3rd generation rack-scale platforms with GPUs, CPUs, and networking technologies to support AI factory design, simulation, and operations. Aligned with NVIDIA's multi-layer AI infrastructure and the NVIDIA DSX AI factory-scale platform announced at GTC 2026, QCT combines hardware, networking, and AI software frameworks to simplify deployment and optimize tokens per watt.

"Our collaborations with NVIDIA push the boundaries of what AI infrastructures can achieve," said Mike Yang, President of QCT. "By combining the NVIDIA AI factory-scale platform with QCT's system designs and integration expertise, we are delivering high-performance, scalable solutions that empower organizations to turn AI ambitions into real-world outcomes."

"AI factories require deep co-design across compute, networking, cooling, power and operations software," said Vladimir Troy, vice president of AI infrastructure at NVIDIA. "QCT's NVIDIA DSX ready systems help customers accelerate the deployment of scalable, energy efficient AI factories optimized for performance, resiliency and lowest token cost."

QCT continues to expand support for NVIDIA's accelerated and computing platforms through extreme co-design of seven chips spanning compute, networking, and storage. The result is five purpose-built rack-scale systems for agentic AI workloads:

- NVIDIA Vera Rubin NVL72 Built by QCT - A next-generation, 3rd generation rack-scale AI

platform with 18x QuantaGrid D76V-1U systems. At its core, it unites 72 NVIDIA Rubin GPUs, 36 NVIDIA Vera CPUs, NVIDIA NVLink 6 Switch for scale-up, NVIDIA ConnectX-9 SuperNICs, and NVIDIA BlueField-4 DPUs  to deliver 10x lower cost per token and 10x higher performance per watt. It features massive compute density and high-bandwidth HBM4 memory. In this rack QCT implements Corning® GlassWorks AI™ solutions, leveraging Corning's expertise in glass science to support the AI era with high-density, reliable fiber, cable, and connectivity solutions. The Corning V-Panel Housing accommodates up to 3,456 fibers in a compact 1RU design and features simplified installation and management. In addition, the MMC Connector delivers high-density connectivity designed for AI and machine learning applications, meeting the increasing bandwidth demands of data centers. Paired with Corning's dependable and widely deployed fiber, these solutions are built on a strong foundation of quality and proven performance. QCT also integrates LITEON's 110kW Power Shelf into this rack to deliver high power density, resilient power delivery, intelligent power smoothing with integrated energy storage, and intelligent load management, advancing high□performance and sustainable data center infrastructure for the AI era.

□ NVIDIA Vera CPU Rack - Equipped with the upcoming QCT QuantaGrid D66Q-2U powered by NVIDIA Vera CPUs, this rack provides industry-leading single-thread performance and energy efficiency for agentic AI.

□ NVIDIA Groq 3 LPX Inference Accelerator Racks - Powered by QCT's upcoming QuantaGrid server supporting NVIDIA Groq 3 LPX and co-designed with NVIDIA Vera Rubin platform, this deterministic, ultra-low-latency inference rack is optimized for AI reasoning, long-context, and agentic inference.

□ NVIDIA BlueField-4 STX AI-Native Storage Rack - Featuring the QuantaGrid D66F-2U at its core to host NVIDIA CMX context memory (KV cache) storage platform, this cluster-scale memory and storage expansion rack is optimized for AI-native data access. It combines the power of NVIDIA BlueField-4 DPU, Vera CPUs and NVIDIA ConnectX-9 SuperNIC to enable high-bandwidth, low-latency data sharing across the AI factory.

□ NVIDIA Spectrum-6 SPX Networking Racks - Powered by NVIDIA Spectrum-6 Ethernet switches with pluggable or co-packaged optics, this design improves efficiency, resiliency, and bandwidth while reducing latency.

□ NVIDIA DSX - QCT is leveraging NVIDIA DSX to design, build and optimize AI Factories across the full stack. By adopting Dassault Systèmes's 3DEXPERIENCE platform, QCT is able to engineer and validate rack-scale AI infrastructure even before its physical construction, supply and deployment. In particular, as facility complexity grows to gigawatt scale, adopting model-based systems engineering (MBSE) from Dassault Systèmes allows QCT to accelerate the deployment of AI factories through simulation. This bridges the gap between concept and deployment, improving quality, shortening time-to-first-token and improving token-per-watt efficiency.

QCT is also showcasing its AI solutions with NVIDIA at COMPUTEX 2026:

 QCT AI POD - A software-defined cluster system for AI workloads that uses NVIDIA and open-source stacks to simplify deployment, monitoring, and management with pre-validated tools. QCT AI POD streamlines AI development and leverages NVIDIA NeMoClaw and NVIDIA Nemotron to automate log management and analysis with LLM-powered agentic AI.

 QCT Dev. Kit for Physical AI - QCT collaborates with Techman Robot to bring physical AI to life using a pre-integrated Dev. Kit built on NVIDIA's robotics stack. It streamlines data generation and model training on QuantaGrid systems, preparing the TM Xplore I humanoid for advanced bimanual industrial tasks with training results.

 QCT AI-RAN Solution - QCT is enabling AI-native network innovations with an expanding AI-accelerated computing portfolio integrated with software stacks from ecosystem partners. Designed to enable AI□RAN deployments across telecom environments, the QuantaEdge EGN77C□2U is built on the NVIDIA Aerial RAN Computer Pro (ARC□Pro) platform, allowing operators to scale compute and networking from centralized data centers to the AI Grid.

To learn more about QCT's servers, solutions and AI factories built on NVIDIA platforms, visit QCT Computex 2026 Booth #G0042 from June 2nd - 5th, 2026.

About QCT
Quanta Cloud Technology (QCT) designs, manufactures, integrates, and services cutting-edge offerings for 5G Telco/Edge, AI/HPC, Cloud, and Enterprise infrastructure via its global network. Product lines include hyper-converged and software-defined data center solutions as well as servers, storage, and network switches from 1U to entire racks with a diverse ecosystem of hardware components and software partners to fit a variety of business verticals and workload parameters.



QCT Marketing Communication
QCT
marcom@qct.io
Visit us on social media:
LinkedIn
Facebook
X