EINPRESSWIRE

QCT Expands Its NVIDIA MGX™ and NVIDIA HGX™ System Offerings at COMPUTEX 2024

Company showcases new AI solution for large language models and generative AI, and systems supporting next gen NVIDIA Spectrum™-X networking and NVIDIA NIM™

QCT
NVIDIA
Press Release
QCT Expands Its NVIDIA MGX™ and NVIDIA HGX™ System Offerings at COMPUTEX 2024

QCT NVIDIA COMPUTEX2024 Press Release

TAIPEI, TAIWAN, June 4, 2024 /EINPresswire.com/ -- Quanta Cloud Technology (QCT), a leading provider of data center and AI solutions, is announcing a handful of AI systems for large-scale generative AI at COMPUTEX 2024, including the NVIDIA MGX™ systems powered by the NVIDIA GB200 Grace Blackwell Superchip, and the air-cooled or liquid-cooled NVIDIA HGX™ B100 and B200 platforms. Also featured at the QCT COMPUTEX Booth G0042 is a 72-GPU NVIDIA MGX rack interconnected by fifth-generation NVIDIA® NVLink® and implementing QCT direct-to-chip liquid cooling, and a variety of QCT accelerated platforms with endless flexibility and optimized for diverse AI and HPC workloads.

Among the new accelerated system additions, QuantaGrid D75B-1U and QuantaGrid D75B-2U are NVIDIA MGX systems powered by the NVIDIA GB200 Grace Blackwell Superchip. They are equipped with one Grace CPU and two NVIDIA B200 GPUs on a single Superchip, connected via a 900GB/s bidirectional ultra-low-latency NVIDIA NVLink C2C interconnect. To scale up to the NVIDIA GB200 NVL72 (36x2 and 72x1), eighteen D75B-1U or D75B-2U use the NVIDIA NVLink Switch System with nine NVLink switch trays, and cable cartridges to interconnect the GPUs and switches. QCT's direct-to-chip liquid cooling is also implemented in these servers to not only allow them to cope with the increased thermal design power (TDP) of the latest Superchips, but also enable the D75B-1U and D75B-2U to deliver the full potential of the NVIDIA Blackwell GPUs to navigate the complexities of trillion-parameter AI models with unprecedented ease.

QCT is also announcing four AI systems that adopt the NVIDIA HGX architecture to take computing to the next level. QuantaGrid D75H-7U is a NVIDIA HGX B100 platform optimized for real-time inference performance, featuring drop-in replacement compatibility for existing HGX H100 infrastructure. The air-cooled QuantaGrid D75F-9U and liquid-cooled QuantaGrid D75L-5U

& QuantaGrid D75M-5U are designed for the most demanding AI, data analytics, and high-performance computing (HPC) workloads. With NVIDIA HGX B200, these models are premier accelerated scale-up x86 platforms that deliver up to 15X faster real-time inference performance, 12X lower cost, and 12X less energy, propelling the data center into a new era of accelerating computing.

“QCT sees the NVIDIA MGX and HGX platforms powered by NVIDIA GB200 Grace Blackwell Superchips and Blackwell GPUs as the building block for the future of AI,” said Mike Yang, President of QCT. “Our latest QuantaGrid accelerated systems are built upon QCT’s expertise in server design and NVIDIA's proven MGX and HGX system architectures. Optimized by QCT system-level and rack-level liquid-cooling solutions, these systems aren’t only leveraging NVIDIA technologies to provide the highest application performance, but also achieving better energy efficiency to help customers achieve sustainability goals.”

More GPU-optimized QCT systems will be shown at COMPUTEX 2024 and validated for the latest NVIDIA AI Enterprise software, which adds support for NVIDIA NIM inference microservices. QCT systems will also support NVIDIA Spectrum-X, the world’s first ethernet platform for AI, consisting of the NVIDIA Spectrum™-X800 SN5600 switch and the NVIDIA BlueField®-3 SuperNIC800. Optimized for the NVIDIA Blackwell architecture, Spectrum-X delivers the highest level of networking performance for AI infrastructures with a 1:1 GPU-to-NIC ratio.

Come to QCT Booth G0042 on the 3rd floor of Hall 1, Nangang Exhibition Center from June 4-7, or visit QCT’s NVIDIA product page to learn more.

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. https://www.qct.io

All other brands, names, and trademarks are the property of their respective owners.

Jean Ko
QCT
jean_ko@quantatw.com
Visit us on social media:
Facebook
X
LinkedIn
YouTube