QCT QOOLRACK STAND-ALONE

Advanced Liquid Cooling for NVIDIA GB200 NVL72 Systems

Produced by TCI Media Custom Publishing in conjunction with:

HPCwire

QCT

NVIDIA

Executive Summary

With artificial intelligence (AI) and HPC solutions becoming more ubiquitous, organizations are looking for high-powered computing solutions that also address thermal management concerns.

The NVIDIA GB200 NVL72 is an innovative solution that meets the demands of both AI and HPC workloads. Designed for large-scale AI and HPC workloads, this liquid-cooled server rack has 72 NVIDIA Blackwell GPUs and 36 NVIDIA Grace™ CPUs to offer a large amount of computing power. What's more, the NVIDIA GB200 NVL72 tray has direct-to-chip liquid cooling for even the most intensive thermal management requirements within an AI/HPC environment. The QoolRack Stand-alone solution adds significant thermal management capabilities.

Additionally, the NVIDIA GB200 NVL72 works with the MGX product line. This is a flexible design framework based on NVIDIA's Modular GPU Accelerated architecture that enables flexible and scalable solutions for various AI and HPC workloads. MGX enables different NVIDIA components like CPUs and GPUs to work together in a way that can be easily updated or changed as new technology comes out. The modular architecture also allows users to mix and match different parts depending on what they need, making it easier for organizations to build for specific AI and HPC tasks.

Challenges in High-Performance Hardware

There are many specific concerns that plague organizations working with high-performance hardware as well as in AI environments. A good place to begin this discussion is with thermal management problems. While air cooling has its place within computing environments, it simply isn't capable of performing effectively for AI and HPC workloads – specifically with GPUs and CPUs. The current generation of CPUs and GPUs have significantly higher thermal design power requirements, which translates to more heat that needs to be dissipated. Air cooling is insufficient for these power-heavy components and can lead to thermal throttling and reduced performance.

The rise in AI applications in recent years has coincided with a similar increase in compute demands. AI models – specifically LLMs – are becoming larger and more complex, thereby necessitating a more robust computing infrastructure. Trillion-parameter models are here and in use, with GPT-4 reportedly having nearly 1.8 trillion parameters. Thus, forward-thinking organizations must plan for ever-increasing model size and complexity.

QCT's QoolRack Stand-alone solution is an innovative approach to manage the thermal challenges of next-gen AI and HPC systems.

NVIDIA GB200 NVL72 and QoolRack Stand-alone Solution

The NVIDIA GB200 NVL72 is a remarkable solution tailor-made for AI and HPC workloads. Additionally, the QoolRack Stand-alone solution is built to provide extensive thermal management for this high-powered rack-level solution.

NVIDIA GB200 NVL72 Specs

In total, the NVIDIA GB200 NVL72 contains 72 NVIDIA Blackwell GPUs and 36 NVIDIA Grace™ CPUs. Breaking this down further, each rack contains 18 compute trays. Each compute tray contains two NVIDIA Grace™ CPU and four NVIDIA Blackwell GPUs.

- 18 trays x 4 GPUs per tray = 72 GPUs total
- 18 trays x 2 CPUs per tray = 36 CPUs total

Each rack has three power shelves at the top and bottom of the rack, totaling six power shelves. Additionally, each tray is equipped with eight E1.S SSDs to enable data storage, two NVIDIA® BlueField®-3 data processing units (DPUs), and two NVIDIA ConnectX®-7 NIC slots. Unlike other systems that use a traditional power supply, trays in the NVIDIA GB200 NVL72 rack are connected with a busbar clip to provide power to the entire system. Thus, this system does not have a power supply inside of it. The NVIDIA GB200 NVL72 adopts NVIDIA NVLink™ connectivity and has Blindmate UQD04 LC inlets designed to allow customers to easily put the systems inside the rack and connect to the manifold for cooling.

The NVIDIA GB200 NVL72 trays are equipped with direct-to-chip liquid cooling for advanced performance and thermal management. The maximum liquid inlet temperature is 45° Celsius, while the maximum liquid return temperature is 65° Celsius. Additionally, the flow rate needed in this system goes up to 130 liters per minute (LPM) per rack. While the chips themselves are liquid cooled, it is important to mention that other components like the SSD, M.2 Riser Board, and NIC are air cooled. The compute tray has a fan inside to ensure these components are air cooled properly.

The NVIDIA GB200 NVL72 is a remarkable solution tailor-made for AI and HPC workloads.

QoolRack Stand-alone Solution for Thermal Management

The QoolRack Stand-alone solution that accompanies the NVIDIA GB200 NVL72 also has many interesting details. Overall, the cooling system is capable of removing 75kW of heat while maintaining the coolant temperature at 10° Celsius above the ambient

temperature (Approach Temperature Difference) with a coolant flow rate of 110 LPM. This is capable of cooling down the entire rack of the NVIDIA GB200 NVL72. The system also provides up to three (2 + 1) hot-swappable pumps to provide redundancy.

This solution also features six power supply units (PSUs) in a redundant configuration, where three PSUs are required for normal operation and three serve as backups. The solution contains a data center secure control module (DC-SCM) board to help customers monitor the cooling status of their system, including fan speed and pump speed. This cooling rack has a high/low level sensor, a pressure sensor, two coolant leakage sensors, and a coolant temperature sensor.

The ambient temperature of the system is up to 35° Celsius. The coolant flow of the system is 130 LPM, the air cooling requires 17 kW, and the liquid cooling requires 115 kW.

> ***There are many solutions available for those wishing to upgrade their AI and HPC infrastructures, but QCT stands out as a superior option. QCT works with world-leading hardware and software partners to deliver total infrastructure solutions from cloud to edge.***

MGX Product Line

The MGX product line is a standardized server design platform created by NVIDIA that supports over 100 different system configurations. It is designed for rapid deployment of AI and accelerated computing technologies. Designed to be both modular and scalable, MGX is compatible with various NVIDIA GPUs, CPUs, and DPUs. Specifically, it integrates well with NVIDIA Blackwell GPUs and NVIDIA Grace™ CPUs.

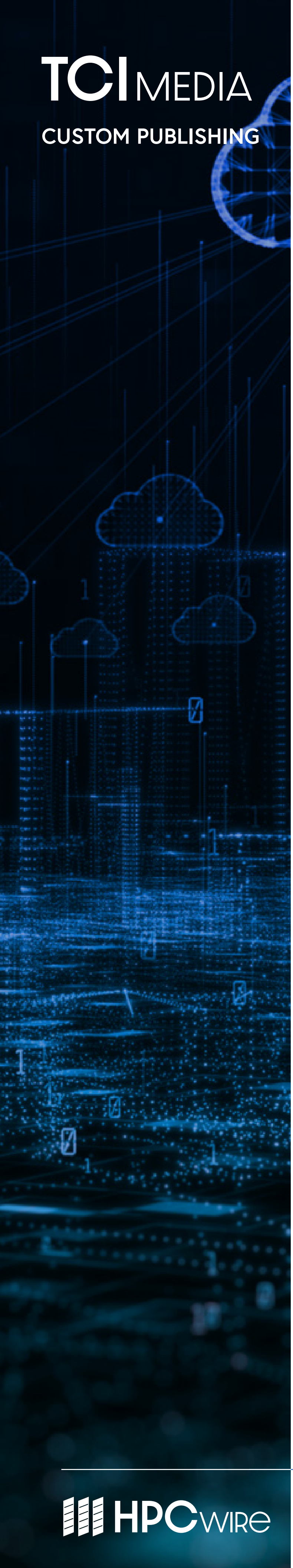
The MGX product line also supports both air and liquid cooing options. As explained, liquid cooling enables more efficient heat dissipation for high-performance configurations with chips. However, air cooling is needed for other components. MGX supporting both air and liquid cooling enables flexibility here.

The QuantaGrid D75E-4U exemplifies the strengths of the MGX product line by providing a high-performance server solution tailored for demanding AI and HPC workloads. The target for this product is a next-gen AI server embracing the latest PCIe GPUs for ultimate flexibility. The QuantaGrid D75E-4U can support up to eight PCIe GPUs, including the latest NVIDIA Blackwell GPU, and each GPU can support up to 600W. The system is designed for optimized cooling, both with air and liquid cooling.

- For air cooling, QCT provides two configurations that are designed for optimal CPU/GPU/Network topology: 2 CPU : 8 GPU : 5 Network cards

  OR

- 2 CPU : 4 GPU : 3 Network cards



For its air-cooled configurations, the QuantaGrid D75E-4U supports two setups, both of which can support PCIe GPUs:

- 2 : 4 : 3 (without PCIe switch)

  OR

- 2 : 8 : 5 (PCIe switch for RDMA)

With exceptional flexibility and computing power, the QuantaGrid D75E-4U supports various applications, including LLM inference, video AI, fine-tuning, and edge video.

Why Choose QCT?

There are many solutions available for those wishing to upgrade their AI and HPC infrastructures, but QCT stands out as a superior option. QCT works with world-leading hardware and software partners to deliver total infrastructure solutions from cloud to edge. Its products are based on a modular design philosophy to not only reduce time to market but also deliver performance, scalability, and future compatibility. As time goes on and technology improves, this allows organizations to easily upgrade their systems without needing a complete overhaul.

QCT also emphasizes serviceability and thermal management in its products. The QuantaGrid D75E-4U features an easy service design with hot-swappable fans and E1.S SSDs, which significantly reduces downtime during maintenance. QCT's thermal solutions also support advanced heatsinks that enhance GPU cooling efficiency, resulting in substantial power savings due to lower fan speeds and improved GPU performance.

Conclusion

The NVIDIA GB200 NVL72 system delivers high performance with a dense configuration of NVIDIA Grace™ CPUs and NVIDIA Blackwell GPUs. Additionally, the MGX platform represents a significant advancement in server design, offering unprecedented flexibility and scalability for AI and accelerated computing. QCT's implementation of MGX demonstrates the platform's versatility, with products like the QuantaGrid D75E-4U, catering to diverse workloads and cooling requirements. What's more, QCT's QoolRack Stand-alone solution is an innovative approach to manage the thermal challenges of next-gen AI and HPC systems.

For more information on how QCT can help your organization, see: QCT Cutting-Edge Infrastructure and Solution Powered by NVIDIA | QCT.