QCT GENAI SOLUTIONS

QCT POD with NVIDIA GB200 NVL72: Powering the Future of Large Language Models

Produced by TCI Media Custom Publishing in conjunction with:

HPCwire

QCT NVIDIA

TCI MEDIA
CUSTOM PUBLISHING

Executive Summary

Advances in computing technology have made artificial intelligence (AI) a widely used tool with a broad range of applications. Specifically, Generative AI (GenAI) has become a means of improving efficiency in many fields. However, this increased productivity comes with intensive infrastructure requirements.

To alleviate this concern, QCT offers the NVIDIA GB200 NVL72 architecture as well as its QCT Platform on Demand (POD) solution. The NVIDIA GB200 NVL72 system is suitable for trillion-parameter large language model (LLM) training, as well as running mixture of experts (MoE) machine learning techniques. Additionally, the QCT POD provides a pre-validated and pre-configured rack-level solution that combines compute, storage, networking components, and infrastructure software.

This paper will describe how both the NVIDIA GB200 NVL72 architecture and the QCT POD solution can be used to increase the efficiency and effectiveness of AI workloads.

What is GenAI and Why is it Valuable?

GenAI specifically refers to certain deep-learning models that are capable of creating text, images, or other content based on training data. It's a versatile tool that has applications in software coding, creative writing, graphic design, 3D modeling, and more.

LLMs are fundamental to GenAI work. In a sense, LLMs form the backbone of many GenAI systems by providing a pre-trained foundation for understanding and generating content across diverse contexts and tasks. Despite their usefulness, LLM training is resource-intense, and takes vast amounts of data and months of compute time to train.

These challenges have been directing trends in the GenAI space. HPC systems equipped with powerful GPUs are often necessary for the efficient parallel processing needed for LLM training. Additionally, AI companies are developing larger models to improve performance. Various scaling laws, such as "Chinchilla scaling," come to the conclusion that larger models generally perform better, with performance improving as a function of the number of parameters. This explains the race to larger LLM parameter sizes, such as GPT-4's estimated 1.8 trillion parameters.

While leveraging pre-existing LLMs such as GPT-4 comes with numerous advantages, many organizations are also choosing to train their own LLM models. For instance, organizations dealing with sensitive or regulated data may prefer the data sovereignty and compliance adherence that training their own model would enable. The same can be said for government agencies involved in national security who

desire complete oversight of the model's architecture, training data, and capabilities. Finally, local cultural issues or extremely specific use cases may push organizations toward training their own LLMs.

Challenges in GenAI

While GenAI is clearly a useful tool for many, there are challenges that organizations must address. With the ever-increasing parameter sizes, training these LLMs will consume more compute resources – specifically power. Such demands are pushing organizations to adopt more powerful and power-efficient computing systems.

Additionally, inter-GPU communication is vital for training LLMs as model sizes continue to increase. It directly impacts performance by enabling distributed training and minimizing bottlenecks that can lead to the underutilization of GPU resources. To train these giant models, systems require thousands of interconnected GPUs within a single domain to reduce latency and maintain high bandwidth.

Outside of compute requirements, networking and storage are also equally important in contributing to GenAI workload performance. High-performance networking is required to facilitate fast inter-GPU communication to minimize latency and maximize throughput when thousands of GPUs are interconnected in a single domain. High-performance storage solutions are also necessary to manage the large amounts of data required for training to ensure data can be efficiently accessed and processed without creating bottlenecks.

With so much to consider, putting compute, storage, and networking solutions together to design and deploy the entire cluster is an enormously complex task. Each component has its own set of requirements and potential bottlenecks that must be addressed, such as managing data transfer speeds and ensuring low-latency communication between GPUs. Thus, deep technical knowledge and specialized skills are essential to architect and integrate these different components for an efficient supercomputing environment.

Finally, there are extensive thermal challenges to consider when implementing GenAI solutions. The large and technologically advanced systems needed for GenAI work require significant amounts of power, which thereby require increasingly sophisticated cooling solutions. This challenge is exacerbated by the possibility of extensive downtime or even damage to expensive components due to poor thermal management. Again, deep technical expertise and specialized skills are often required to ensure proper thermal management.

Despite their usefulness, LLM training is resource-intense, and takes vast amounts of data and months of compute time to train.

QCT Works with NVIDIA to Deliver Solutions for GenAI

GenAI challenges are just as widespread and varied as the technology's real-world applications. As such, QCT has both hardware and software solutions that leverage NVIDIA technologies to alleviate certain GenAI problems.

NVIDIA GB200 NVL72 Solution

To help alleviate hardware concerns about GenAI workloads, QCT has worked with NVIDIA to offer the NVIDIA GB200 NVL72 solution. A system designed for trillion-parameter training, the NVIDIA GB200 NVL72 has all the hardware necessary to perform GenAI tasks. This rack-level solution consists of 18x QuantaGrid D75B-2U systems, each equipped with two NVIDIA Grace™ CPUs and four NVIDIA Blackwell GPUs – resulting in a total configuration of 72 NVIDIA Blackwell GPUs interconnected via 5th generation NVIDIA NVLink™ with a single NVLink domain. In essence, this forms a gigantic, singular GPU that is capable of delivering ultra performance.

The NVIDIA GB200 NVL72 leverages the NVIDIA Quantum-2 InfiniBand platform, which offers speeds up to 400 Gb/s Next-Generation Data Rate (NDR) InfiniBand, doubling the speed of previous generations. This platform includes the NVIDIA Quantum-2 InfiniBand switch, ConnectX-7 network adapters, and BlueField-3 data processing units (DPUs), providing a comprehensive networking solution. Compared to traditional implementations, the NVIDIA Quantum-2 platform significantly enhances communication speed, reduces latency, and offers advanced features such as in-network computing acceleration and multi-tenant performance isolation. This architecture enables the NVIDIA GB200 NVL72 to handle the massive data throughput required for next-generation AI workloads and extreme-scale systems. However, a separate ethernet connection is implemented for management and storage purposes, thereby minimizing interference network traffic and optimizing overall system performance.

A system designed for trillion-parameter training, the NVIDIA GB200 NVL72 has all the hardware necessary to perform GenAI tasks.

High-performance storage is vitally important in loading AI models and handling data during both the training and fine-tuning processes. Given that AI workloads have mixed I/O patterns, they require high-performance storage. QCT offers NVMe-based storage servers that can run mainstream high-performance storage solutions designed to handle AI workloads. Additionally, the NVIDIA GB200 NVL72 solution offers object storage as an option for cost-effective warm and cold storage. This enables organizations to manage their data lifecycle and ensure that frequently accessed

data is stored in high-performance environments while less critical data can be stored in more economical ways. QCT is also able to fine-tune the storage performance to deliver optimal output for AI workloads.

In terms of thermal management, the QCT QoolRack Stand-alone solution minimizes facility changes for implementing liquid cooling while providing sufficient cooling capacity – up to 120kW. This solution ensures the system operates within safe temperatures and increases performance by ensuring that optimal operating conditions are met. What's more, QCT's solution is capable of working with other 3rd party liquid-to-liquid cooling systems.

Utility nodes are also an essential system in the NVIDIA GB200 NVL72 solution. They provide services like login-node functionality, cluster management, and hosting Kubernetes clusters. QCT is able to assist here by selecting suitable server models and the right configuration to serve as utility nodes.

QCT POD Infrastructure Solution

The software environment chosen to be a part of AI infrastructure is just as important as the hardware decisions. As such, QCT offers a pre-validated compliance solution for the NVIDIA GB200 NVL72 solution called QCT POD, which is meant to reduce complexity for users. The pre-validation from QCT checks that all components are working together as they are meant to, minimizing integration issues and enabling organizations to spend their time focusing on AI projects rather than system setup. The QCT POD offers pre-installed software to facilitate AI workflow.

Additionally, QCT POD offers a deployment option to set up cloud-native environment for AI workloads. Setting up a container environment like Kubernetes for AI workloads can be challenging due to the complexity of configuring the infrastructure to support GPU acceleration, managing dependencies, and ensuring compatibility with container runtimes. As such, QCT sets up a container environment and configures it for customers with QCT POD.

QCT POD also offers quick access to NVIDIA GPU Cloud (NVIDIA® NGC™), which is a cloud hub for pre-optimized containers, AI models, and SDKs. NVIDIA® NGC™ provides ready-to-use resources for a variety of AI tasks that are optimized for performance on NVIDIA hardware.

This software solution also provides a pre-installed and pre-configured Jupyter Notebook. This is an interactive environment for writing and testing code for developing AI models. It is also integrated with QCT POD user management features for a unified user management experience.

Overall, QCT POD helps customers to spin up AI workloads without having to worry about all the complexities involved, saving them considerable time and effort for setting up their software infrastructure.

QCT POD has other pre-installed AI tools and libraries meant to allow organization to hit the ground running with their AI workloads. These pre-installed software include common AI frameworks such as PyTorch, TensorFlow, and other related libraries and compilers, saving users download time and having necessary software ready. It also resolves the complex software dependency issues that many AI workloads demand.

Overall, QCT POD helps customers to spin up AI workloads without having to worry about all the complexities involved, saving them considerable time and effort for setting up their software infrastructure. QCT provides a set of necessary tools and integrated services that are ready-to-use, thereby streamlining the AI workflow and increasing time-to-value.

NVIDIA AI Enterprise for NVIDIA GB200 NVL72

Customers of these solutions can also choose to run NVIDIA AI Enterprise, which is an enterprise-grade AI platform, on top of NVIDIA GB200 NVL72. This provides enterprise-grade security and support from NVIDIA, enhancing the capabilities of this hardware. The NVIDIA AI Enterprise platform includes pre-trained generative AI models from frameworks such as NVIDIA NeMo™, NVIDIA® Riva, and NVIDIA Megatron-Core. Resources like this streamline the process of building applications for GenAI workloads by reducing the time and computational cost to rain the models from scratch. NVIDIA AI Enterprise also includes tools for data processing, model training, fine-tuning, deployment, and monitoring – making it a complete solution for the GenAI lifestyle.

Conclusion

GenAI tools have proven their staying power, and as such, organizations are working to ensure they are capable of handling these workloads. The rapid development of GenAI – particularly in respect to LLMs – is driving an increased demand for AI infrastructure that can handle trillion-parameter models and complex AI workflows.

The NVIDIA GB200 NVL72 solution addresses the requirement for high performance computing needed to train these trillion-parameter models, and offers a comprehensive rack-level system with powerful GPUs, low-latency networking, and optimized storage. Additionally, QCT POD comes as a pre-validated solution for the NVIDIA GB200 NVL72 system. It simplifies the deployment and management of AI infrastructure by offering pre-installed and pre-configured software stacks, cloud-native environments, and pre-integrated AI development tools. This helps accelerate time-to-value for organizations implementing AI projects.

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