SOLUTION BRIEF

Empower AI Workloads with QCT Infrastructures

AI-Ready Enterprise Platform with NVIDIA and VMware

Unleashing AI in any workload, any vertical and anywhere

Artificial intelligence (AI) has become a mainstream technology to boost productivity and lead digital transformation in data-driven enterprises. An increasing number of AI applications have been used in different verticals. For example, medical image recognition leverages AI training and inference models to accelerate medical image interpretation; autonomous car manufacturers can have better product development cycles by harnessing AI and machine learning. Many other AI workloads have also emerged in surveillance, social media, entertainment, education and other areas. Organizations believe that AI will give them a competitive advantage for their businesses.

However, depite the great populatiry of AI applications, enterprises looking to adopt such technologies still face some challenges and pain points.

First is complexity in AI. There are the risks for putting all the AI components together and integrating with existing infrastructures when conducting core tasks like training, testing and inference in AI development.

Second is the underutilization of GPU resources. Since for the physical GPU, the GPU resources cannot be divided according to different workloads, and thus, it may make the resources idle. And if there's an abundance of underutilized GPU, the AI tasks may fail.

Last but not least, there is a scalability issue. Enterprises may not afford the high cost of using AI applications at scale, not to mention the slow deployment and availability issue for an enterprise-grade AI infrastructure.

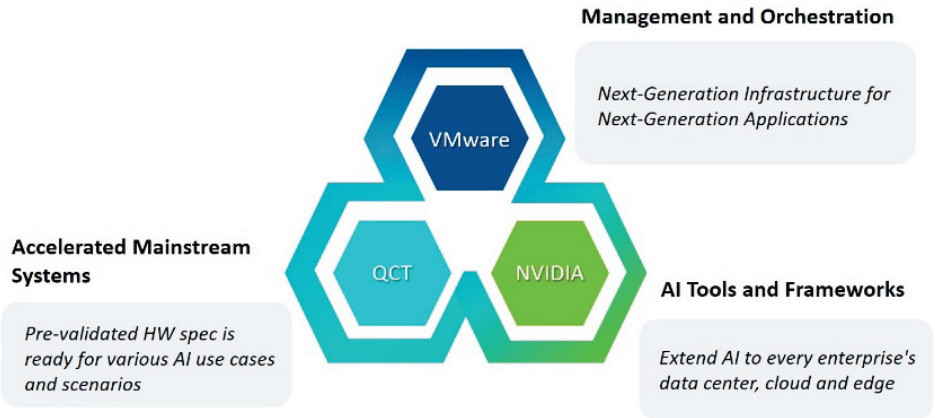
Main Challenges for AI Workloads

- **Complexity in AI**
  There are risks for putting all the AI components together and integrating with existing infrastructure
- **Underutilization of GPU**
  Using a physical GPU may make the resources idle since the GPU resources can't be divided according to different workloads
- **Scalability Issue**
  Enterprise may not be able to afford using AI applications at scale

An AI-Ready Enterprise Platform

To address the above pain points enterprises may face, three global-leading companies- NVIDIA, VMware and QCT, developed an AI solution- NVIDIA AI Enterprise, making it easier to run AI applications and worklaods. The integrated AI solution is composed of NVIDIA AI Enterprise Suite, a leading AI software, VMware vSphere®, the industry' s best-in-class virtualization platform, and QCT' s pre-validated systems.

By delivering an end-to-end enterprise platfrom optimized for AI workloads, NVIDIA AI Enterpise streamlines the development and deployment process and accelerate the speed of building AI infrastructure, delivering a better efficiency for developers.



Key Technology Building Blocks

NVIDIA AI Enterprise brings AI to the exisiting VMware vSphere platform which runs on NVIDIA’ s pre-validated systems. To be more specific, the pre-validated AI solution is composed of three different technology layers:

NVIDIA AI Enterprise Suite

NVIDIA AI Enterprise is an optimized end-to-end software suite of AI and data analytic applications that is exclusively certififed by NVIDIA to run on the VMware vSphere platform, where NVIDIA AI containers and pre-trained models are ready for the virtualized GPU to do compute and AI/ML jobs. It includes key enabling technologies and software for the rapid deployment, management and scaling of AI workloads, streamlining the deployment and development of AI and machine learning infrastructures.

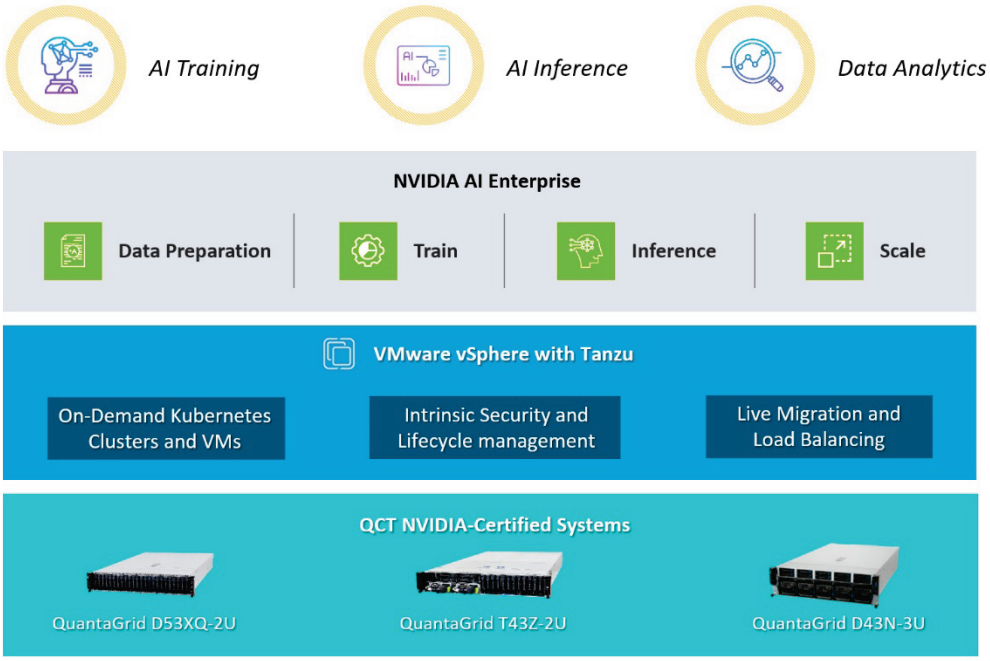
VMware vSphere® with Tanzu™

VMware vSphere® is an industry leading virtualization and cloud platform, providing a powerful, flexible and secure foundation for business agility. By enabling VMware vSphere® with Tanzu™, you can run kubernetes workloads using your existing infrastructure. It transforms bare-metal servers into centrally managed AI and machine learning infrastructure pools that can quickly provision virtual machines and containers on demand. Furthermore, it gives you intrinsic security and lifecycle management for your infrastructure including network service, registry service, storage service and VM service.

VMware vSphere ® has been enhanced to run the NVIDIA AI Enterprise software suite with NVIDIA’s latest Ampere architecture-based GPUs, providing an innovative infrastructure for next- generation applications.

NVIDIA-Certified Systems

NVIDIA-Certified Systems allow businesses to easily and more confidentally deploy software solutions such as AI, ML, and data analytics. The NVIDIA AI Enterprise software is certified to run on NVIDIA-Certified systems, which include NVIDIA Ampere architecture-based GPUs, NVIDIA® ConnectX® smart network interface cards (SmartNICs), and the NVIDIA BlueField® data processing unit (DPU).



QCT NVIDIA AI Enterprise Solution Offerings

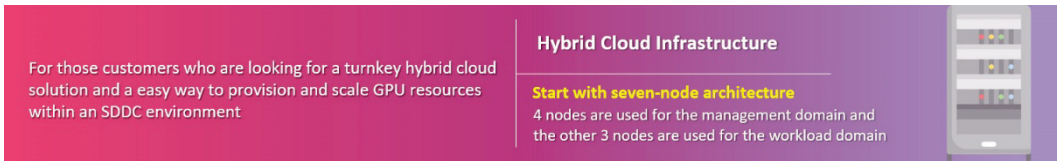
QCT AI-ready platform validated with the NVIDIA AI Enterprise suite and VMware vSphere® with Tanzu ™ delivers a best-in-class enterprise grade AI solution. QCT provides QuantaGrid D43N in this solution, a high performance and optimized accelerated server. Designed with the latest AMD EPYC™ processors in a 3U design, QuantaGrid D43N can support up to 4 NVIDIA A100 GPUs to suit your graphic intensive workloads. By adopting the latest MIG (Multi-Instance GPU) technology, QCT NVAIE solution helps you achieve better efficiency in AI training, inference and data analytics.

Considering there are a wide range of use cases which need different scales of infrastructure, QCT has designed and validated three kinds of solution architectures to fulfill varying customer needs:

Disaggregated Infrastructure

Hyper Converged Infrastructure

Hybrid Cloud Infrastructure



About 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.

Product lines include hyperconverged and software-defined data center solutions as well as servers, storage, switches and integrated racks with a diverse ecosystem of hardware components and software partners. QCT designs, manufactures, integrates and services cutting-edge offerings via its own global network. The parent of QCT is Quanta Computer, Inc., a Fortune Global 500 corporation. http://www.QCT.io

About VMware

VMware is a leading provider of multi-cloud services for all apps, enabling digital innovation with enterprise control. As a trusted foundation to accelerate innovation, VMware software gives businesses the flexibility and choice they need to build the future. Headquartered in Palo Alto, Calif., VMware is committed to building a better future through the company's 2030 Agenda. For more information, please visit www.vmware.com/company.

About NVIDIA

Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics and ignited the era of modern AI. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com/.

United States
QCT LLC., Silicon Valley Office
1010 Rincon Circle, San Jose, CA 95131 TOLL-FREE: 1-855-QCT-MUST
TEL:+1-510-270-6111
FAX: +1-510-270-6161
Support: +1-510-270-6216

China
云达科技,北京办公室 (Quanta Cloud Technology)
北京市朝阳区东大桥路 12 号润诚中心2 号 楼
TEL:+86-10-5920-7600
FAX: +86-10-5981-7958

云达科技,杭州办公室 (Quanta Cloud Technology)
浙江省杭州市西湖区古墩路浙商财富中心 4 号 楼 303 室
TEL: +86-571-2819-8650

Japan
Quanta Cloud Technology Japan 株式會社日本国東京都港区芝大門二丁目五番八号牧田ビル 3 階
TEL:+81-3-5777-0818
FAX: +81-3-5777-0819

Germany
Quanta Cloud Technology Germany GmbH
Rurbenden 48
52353 Düren
TEL: +49-2421-3863400
Fax: +49-2421-3863899

Korea
QCT Korea, Inc. (주식회사 큐씨티코리아)
서울특별시 영등포구 의사당대로 97
교보증권빌딩 10 층, 07327
TEL:+82-10-5397-1412
FAX: +82-2-6336-6710

Other regions
Quanta Cloud Technology
No. 211 Wenhua 2nd Rd., Guishan Dist., Taoyuan City 33377, Taiwan
TEL:+886-3-327-2345
FAX: +886-3-397-4770