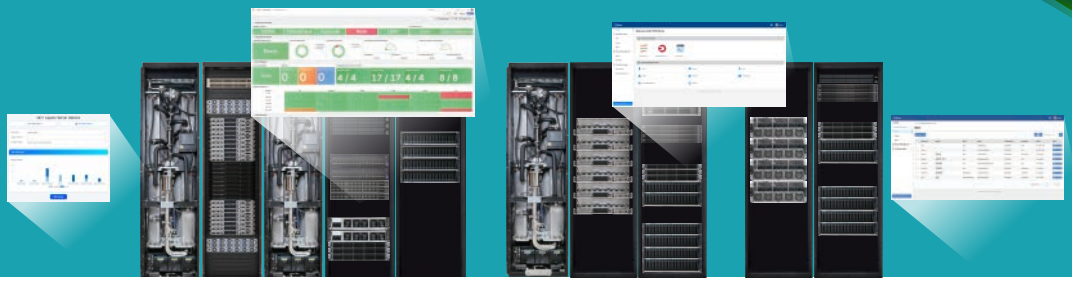
QCT AI POD Solution

AI Ready Platform Powered by Advanced Cluster-Level Solutions



From Complexity to Clarity: The Easy Path to Cluster-Level AI Infrastructure

As AI advances rapidly, enterprises face growing challenges in building cluster-level infrastructure to support diverse workloads. Designing a system that integrates compute, storage, networking, facilities, and software is quite complex and resource-intensive.

To simplify the cluster-level system design, QCT delivers the QCT AI POD solution, a pre-configured system designed to support a wide range of AI workloads. It combines pre-optimized hardware and software with system monitoring, management, and deployment tools, along with a robust AI development environment, helping enterprises reduce complexity and accelerate AI adoption.

- **Pre-Configured Cluster-Level Design**
- **Simplicity and Fast Time-to-Market**
- **Flexibility for Hardware & Software Design**

QCT AI POD Solution- QCT's Rack-level AI Infrastructure + Software Service

QCT AI POD provides end-to-end support, spanning infrastructure to software, effectively addressing a wide range of industrial workload requirements.

Industrial Workloads

Manufacturing | **Healthcare** | **Education & Research** | **Logistics** | **Enterainment** | **Transportation**

Computer Vision	LLM Inference	LLM Fine-tune Training	Trillion-parameter LLM Training
Generative AI Inference with RAG	Deep Learning Training	Agentic AI	AI Reasoning

QCT Software Platforms

Rapid Deployment | **System Management** | **Real-time Monitoring** | **Development Environment** | **User Workspaces**

Bare Metal	Cloud-native

QCT AI Cluster Infrastructure

Compute | **Storage** | **Network** | **Cooling System**

QCT AI POD Software Service: Simplifying AI for IT Admins & Developers

By leveraging open technologies, QCT AI POD delivers comprehensive software services that streamline deployment, simplify management, enhance monitoring, and provide a robust AI development environment to accelerate innovation.

Developers & End users

QCT AI POD provides tailored solutions for cloud-native and bare-metal environments, featuring a web-based workspace that seamlessly integrates development tools and resource scheduling systems:

✓ Multiple User Workspace

Provide multiple user interfaces options based on user development preferences

✓ Comprehensive Development Tools

Offer ready-to-use AI framework, libraries and SDKs to speed up development

✓ Advanced Job Schedulers

Integrate an advanced job scheduler to maximize resource utilization

Development Environment

- User Web UI
- Development Tools
- Resource Management & Job Scheduling
  - BareMetal
  - Container

Cluster Management and Monitoring

- Admin Web UI
- Software Defined Storage
- System Management
- System Monitoring
- Bare Metal
- Container
- Operating System

Offline System Deployment

- Text UI
- Deployment & Provisioning Tools

IT Administrators

QCT AI POD offers rapid deployment and powerful management tools to simplify administrative tasks for IT staff, accelerating time to value:

✓ Rapid System Deployment

#UI Configurator #Pre-defined Playbook #Offline Installation

Empowers users to expedite the system deployment process, reducing setup time for a new cluster from days to hours through a text-based UI configurator

✓ Real-time System Monitoring

#Web-Based Dashboard #System Status Report #Log Mgmt.

Delivers a web-based dashboard for real-time monitoring of system resource utilization

✓ Easy System Management

#Account Mgmt. #Cluster Mgmt. #Storage Mgmt.

Encompasses user authentication, resource management, and container orchestration tools, allowing administrators to effortlessly control user access and resource allocation

Building Blocks of Smarter AI Clusters

To complete the cluster-level system design, the QCT AI POD solution takes below hardware components into consideration:

Compute Node
Provides optimized performance for different AI workloads

Storage Node
Supports configuration with NAS or third-party storage systems

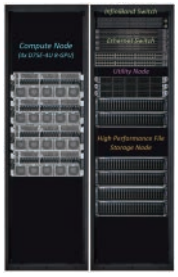
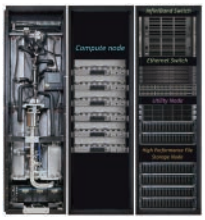
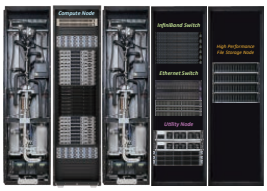
Management Node
Designed for login, administration, deployment, and Kubernetes cluster management

Network Fabric
Built to support east–west traffic, north–south traffic, and out-of-band management

Cooling System
Offers QCT QoolRack Sidecar liquid-to-air cooling options (for liquid-cooled compute node designs)

Choose the Right QCT AI POD for Your Needs

By incorporating compute node, storage node, management node, network fabrics , and even cooling system into the cluster design, QCT AI POD delivers a complete portfolio to support diverse AI workloads:

	QCT 4-GPU AI POD	QCT 8-GPU AI POD			QCT 72-GPU AI POD
	PCIe GPU	PCIe GPU	SXM GPU with Air-cooling	SXM GPU with Liquid-cooling	NVIDIA GB300 NVL72
Cluster Design					
Compute	(2) QuantaGrid D75E-4U	(4) QuantaGrid D75E-4U	(8) QuantaGrid D75H-10U	(6) QuantaGrid D75L-3U	(18) QuantaGrid D75U-1U
Storage	(4) QuantaGrid D55Q-2U with NFS/ Ceph or 3rd-party storage solutions (optional)				
Management	(2) QuantaGrid D55Q-2U + (3) QuantaGrid D55X-1U				
Network	Configured with NVIDIA Quantum InfiniBand or NVIDIA Spectrum™-X Ethernet switches				
Cooling	Air-cooled system			Liquid- cooled system with QCT QoolRack Sidecar Solution	
Key Features	· Offers a variety of PCIe GPUs · Cost-effective solution		· 8x SXM GPUs interconnected by NVIDIA NVLink™ (HGX) in a single node · Options for Air-cooling and Liquid-cooling		· Based upon the NVIDIA MGX™ reference architecture · Interconnected by NVIDIA NVLink™ switches across nodes · Liquid cooled system design
Target Workloads	· Generative AI with RAG · Agentic AI · LLM Inference · Computer Vision, 3D Visualization		· LLM Fine-Tune Training · Deep Learning Training with Multi-GPU		· Trillion Parameter LLM Training · AI Factory

Modern Hardware Platforms for QCT AI POD

Compute · Management · Storage · Cooling

QuantaGrid D75E-4U

- Powered by Intel® Xeon® 6 CPUs
- Up to (8) DW accelerators
- Support 2-way & 4-way NVLink Bridge

QuantaGrid D75H-10U

- Powered by Intel® Xeon® 6 CPUs
- Onboard CX-8 800Gb for E-W traffic
- (8) NVIDIA Blackwell Ultra SXM GPU modules with an HGX baseboard

QuantaGrid D75L-3U

- Powered by Intel® Xeon® 6 CPUs
- Onboard CX-8 800Gb for E-W traffic
- (8) NVIDIA Blackwell Ultra SXM GPU modules with an HGX baseboard
- Liquid cooling

QuantaGrid D55X-1U

- Powered by Intel® Xeon® 6 CPUs
- PCIe Gen5 and DDR5 platform ready
- Multiple storage options including 2.5”, E1.S, E3.S

QuantaGrid D55Q-2U

- Powered by Intel® Xeon® 6 CPUs
- PCIe Gen5 and DDR5 platform ready
- Multiple storage options including 3.5”, 2.5”

QuantaGrid D75U-1U

- Compute tray to the NVIDIA GB300 NVL72
- NVIDIA MGX™ architecture-based system
- Onboard CX-8 800Gb for E-W traffic
- Liquid cooling

QoolRack Sidecar Solution

- Liquid-to-air (L2A) design
- 70KW cooling capacity
- Smart management of cooling status

QCT®

www.QCT.io

Contact Us

All specifications and figures are subject to change without prior notice. Actual products may look different from the photos.
QCT, the QCT logo, Quanta, and the Quanta logo are trademarks or registered trademarks of Quanta Computer Inc.
All trademarks and logos are the properties of their respective holders.
Copyright © 2025 Quanta Computer Inc. All rights reserved.

ver NOV., 2025