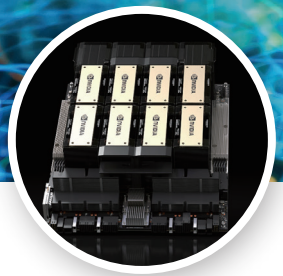
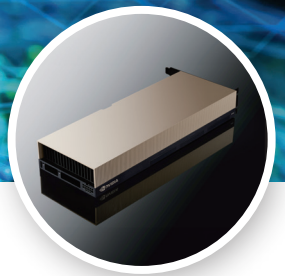
QCT Cutting-Edge Infrastructures Accelerated by NVIDIA

NVIDIA MGX™



NVIDIA HGX™



NVIDIA GPU

QCT | NVIDIA

QCT's NVIDIA MGX Architecture Product Family

| QuantaGrid D75U-1U | QuantaGrid D75B-2U | QuantaGrid D75B-1U | QuantaGrid D75E-4U | QuantaGrid S74G-2U | QuantaEdge EGX77GE-2U |

QCT's NVIDIA HGX Platform Product Family

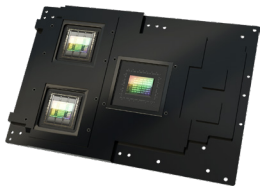
| QuantaGrid D74F-7U | QuantaGrid D75F-7U | QuantaGrid D75M-5U | QuantaGrid D74H-7U | QuantaGrid D74A-7U |

QCT's QuantaGrid/QuantaPlex /QuantaEdge NVIDIA PCIe GPU Server Family

| QuantaGrid D54U-3U | QuantaGrid D55Q-2U | QuantaGrid D55X-1U | QuantaGrid S44NL-1U | QuantaGrid D44N-1U | QuantaGrid S55R-1U | QuantaPlex S25Z-2U | QuantaPlex S45Z-2U | QuantaEdge EGX44I-1U | QuantaEdge EGX88D-1U | QuantaEdge EGX77B-1U |

NVIDIA GB200 NVL72

Delivering Trillion-Parameter LLM Training and Real-Time Inference

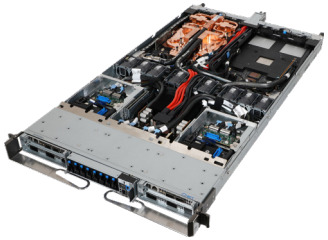


30x Faster LLM Inference compared with NVIDIA H100 GPU

4x Faster LLM Training compared with NVIDIA H100 GPU

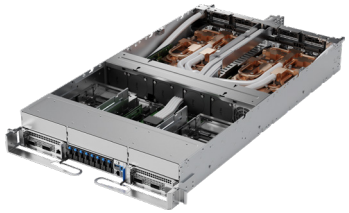
25x Lower Total Cost of Ownership (TCO), representing a leap in sustainability and cost-effectiveness in AI operations

QuantaGrid D75B-1U



Platform	(2) NVIDIA GB200 Grace™ Blackwell Superchips
CPU **GPU**	(2) NVIDIA Grace™ CPUs (4) NVIDIA Blackwell GPUs
Memory	CPU: Up to 480GB LPDDRX embedded per CPU GPU: Up to 186GB HBM3e embedded per GPU
Storage	(8) E1.S 15mm PCIe SSDs
Onboard Storage	(1) PCIe M.2 22110/2280
Network	(2) NVIDIA BlueField®-3 B3240 dual port 400G DPUs (4) NVIDIA ConnectX®-7 400Gb OSFP ports
Cooling	CPU/GPU: Liquid cooling cold plate Peripheral: (8) 4056 dual rotor fans
Power	48-54V DC bus bar clip
Dimensions	(W) 438 x (H) 43.6 x (D) 766mm

QuantaGrid D75B-2U



Platform	(2) NVIDIA GB200 Grace™ Blackwell Superchips
CPU **GPU**	(2) NVIDIA Grace™ CPUs (4) NVIDIA Blackwell GPUs
Memory	CPU: Up to 480GB LPDDRX embedded GPU: 144GB HBM3e embedded
Storage	(8) E1.S 15mm PCIe SSDs
Onboard Storage	(1) PCIe M.2 22110/2280
Network	(2) NVIDIA BlueField®-3 B3240 dual port 400G DPUs (4) NVIDIA ConnectX®-7 400Gb OSFP ports
Cooling	CPU/GPU: Liquid cooling cold plate Peripheral: (6) 6056 dual rotor fans
Power	48-54V DC bus bar clip
Dimensions	(W) 438 x (H) 87 x (D) 766mm

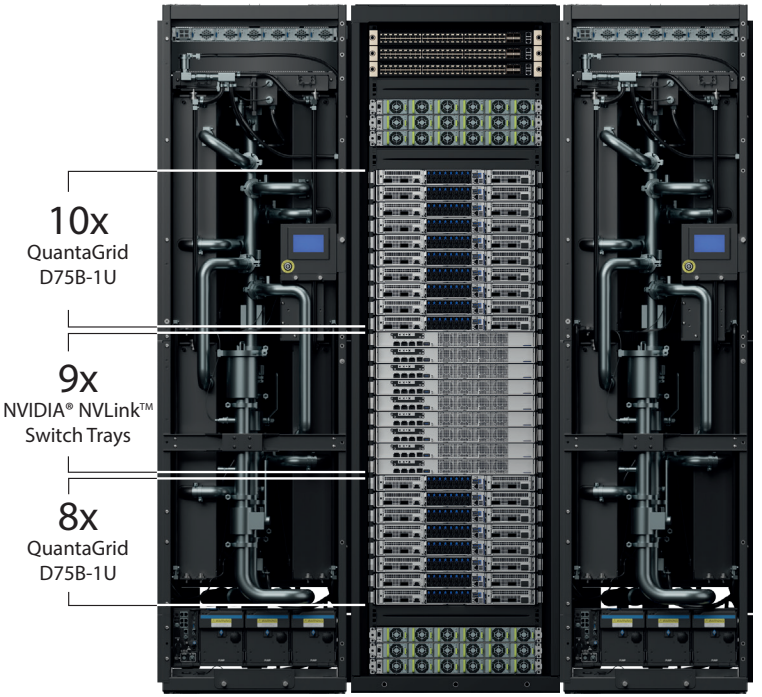
QuantaGrid D75B-2U (NVIDIA MGX Architecture)

At the heart of the NVIDIA GB200 NVL72 are QCT's QuantaGrid D75B-2U and QuantaGrid D75B-1Uservers. Each of these servers can support up to 2x NVIDIA GB200 Grace™ Blackwell Superchips in a 2U or 1U form factor.

The NVIDIA GB200 Grace™ Blackwell Superchip is a key component in the NVIDIA GB200 NVL72, connecting two high-performance NVIDIA Blackwell GPUs and an NVIDIA Grace™ CPU using the NVIDIA® NVLink®-C2C interconnect.

This can be scaled up to the NVIDIA GB200 NVL72 configuration: a powerhouse that functions as a single, colossal GPU. Connecting 36 Grace CPUs and 72 Blackwell GPUs, this liquid-cooled, rack-scale system is designed to navigate the complexities of trillion-parameter AI models with unprecedented ease.

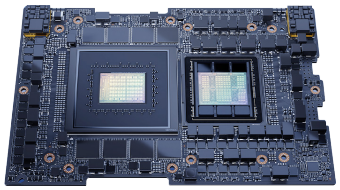
NVIDIA GB200 NVL72 Populated by QuantaGrid D75B-1U



QCT QoolRack, side car heat exchangers for the IT racks

NVIDIA GH200 Grace Hopper Superchip

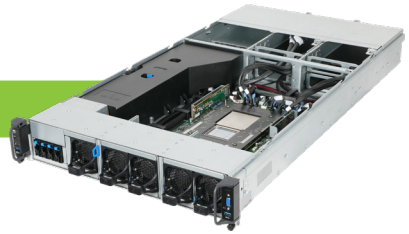
Delivering Higher Performance, Faster Memory, and Massive Bandwidth for Compute Efficiency



7x Faster than PCIe 5.0 by using the NVIDIA® NVLink®-C2C CPU-GPU coherent interface

624GB Shared memory for fast-access memory, utilizing HBM3 and HBM3e memory

QuantaGrid S74G-2U



Breakthrough Accelerated Performance for Giant-scale AI-HPC Applications

Form Factor	SuperChip Number	Memory Number
2U	1	576GB

#AI Training *#Scientific Computing* *#HPC* *#Data Visualization* *#3D Modeling*

- First gen NVIDIA MGX™ architecture with a modular design
- Powered by the NVIDIA Grace Hopper™ Superchip
- Optimized for memory-intensive inference and HPC workloads

Processor	NVIDIA GH200 Grace Hopper™ Superchip, 1000W TDP
Memory	CPU: Up to 480GB LPDDRX embedded GPU: 96GB HBM3 memory Coherent memory between CPU and GPU with NVIDIA® NVLink®-C2C interconnect with a speed of 900GB/s
Networking	(1) Dedicated 1GbE management port
Storage	(4) E1.S NVMe SSDs
Expansion Slot	(3) PCIe 5.0 x16 FHFL dual-width slots
Dimensions	(W) 438 x (H) 87.5 x (D) 900mm

NVIDIA MGX Architecture Systems

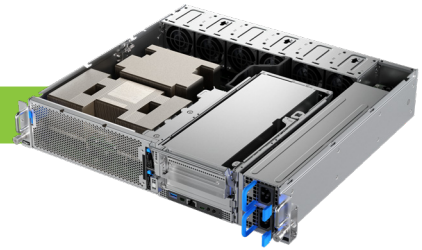
QCT systems accelerated by the NVIDIA MGX™ architecture such as the QuantaGrid S74G-2U, QuantaEdge EGX77GE-2U allow different configurations of GPUs, CPUs and DPUs, shortening the time frame for building future compatible solutions.

Based on the modular reference design, these configurations can not only support future accelerators, but also meet the requirements of diverse workloads, including those that incorporate liquid cooling, to shorten the development journey and reduce time to market.

Key Features and Benefits

- Accelerates time-to-market
- Flexible
- Runs full NVIDIA software stack

QuantaEdge EGX77GE-2U



First 400mm Edge Server Adopting NVIDIA MGX Architecture

Form Factor	SuperChip Number	Memory Number
2U	1	512GB

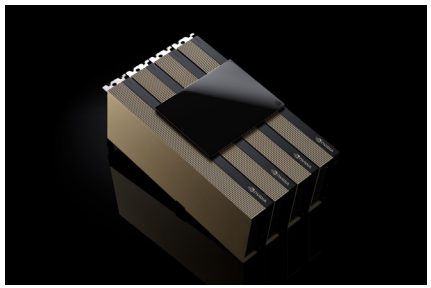
#Smart City *#Smart Manufacturing* *#Real-time Image Analysis* *#EdgeAI*

- Modular infrastructure based on the NVIDIA MGX™ architecture
- First QCT edge server to be powered by the NVIDIA Grace Hopper™ Superchip
- NVIDIA® NVLink®-C2C high-bandwidth low-latency interconnect
- 400 mm ultra short-depth design server

Processor	NVIDIA GH200 Grace Hopper™ Superchip, 1000W TDP
Storge	Internal Storage: (2) SATA/NVMe M.2 2280/22110 External Storage: (2) E1.S
Expansion Slot	(3) FHFL PCIe 5.0 x16
Dimensions	(W) 447.8 x (H) 42.8 x (D) 400mm

NVIDIA H200 GPU

Most Efficient Compute for LLM Inference, AI, and HPC



7x Faster than PCIe 5.0 NVIDIA® NVLink™ for GPU to GPU communication

1.7x Fine-tune LLMs inference performance compared with NVIDIA H100 NVL

1.5x GPU memory increase compared with NVIDIA H100 NVL

QuantaGrid D75E-4U

Tailored for Every Customer Need to Unlock Infinite AI Possibilities

Form Factor	SuperChip Number
4U	2

- Supports NVIDIA next-gen PCIe GPUs, up to 8x DW AC 600W
- Offers infinite flexibility to support any AI/HPC-related workloads
- All PCIe 5.0 expansion slots are designed to support up to 150W
- Remote heatsink solution for improved thermal performance
- Enhanced serviceability with tool-less, hot-pluggable designs

QuantaGrid D75E-4U (NVIDIA MGX Architecture)

The QuantaGrid D75E-4U is more than just an x86-based system built on the Intel® Xeon® platform. It adheres to the NVIDIA MGX™ architecture, offering a modular design that meets diverse AI applications and customer demands. This system is compatible with a full range of NVIDIA data center PCIe GPUs, including the NVIDIA Blackwell Ultra GPUs, NVIDIA RTX PRO™ 6000 Blackwell Server Edition, NVIDIA H200 NVL, NVIDIA H100 NVL, NVIDIA L40S GPU, NVIDIA L4 GPU, NVIDIA A2 GPU, NVIDIA A10 GPU, and NVIDIA A16 GPU, enabling unparalleled flexibility and performance.

The NVIDIA H200 NVL is particularly suited for organizations with data centers seeking low-power, air-cooled enterprise rack designs. It delivers versatile acceleration for AI and HPC workloads of all sizes, making it an ideal choice for enterprises prioritizing efficiency and scalability.

With the QuantaGrid D75E-4U, customers can maximize computing power in compact spaces. The system supports flexible GPU configurations—1, 2, 4, or 8 GPUs—allowing companies to optimize their existing rack infrastructure and tailor performance to their specific requirements.

Processor	(2) Intel® Xeon® 6 processors, up to TDP 350W	
Networking	(1) Dedicated 1GbE management port	
Accelerator	Air cooling: NVIDIA Blackwell Ultra GPUs, NVIDIA RTX PRO™ 6000 Blackwell Server Edition, NVIDIA H200 NVL, NVIDIA H100 NVL, NVIDIA L40S GPU, NVIDIA L4 GPU, NVIDIA A2 GPU, NVIDIA A10 GPU, NVIDIA A16 GPU	
Memory	(32) DDR5 RDIMM up to 6,400 MHz, (16) MRDIMM up to 8,000 MHz	
Storage	**Air cooling - (4) DW GPUs:** (12) Hot-swappable E1.S SSDs	**Air cooling - (8) DW GPUs:** (24) Hot-swappable E1.S SSDs
Expansion Slot	**Air cooling - (4) DW GPUs:** (4) DW FHFL PCIe 5.0 x 16 slots for GPU (3) SW FHFL PCIe 5.0 x 16 slots for networking	**Air cooling - (8) DW GPUs:** (8) DW FHFL PCIe 5.0 x 16 slots for GPU (4) SW FHFL PCIe 5.0 x 16 slots for networking (1) SW FHHL PCIe 5.0 x 16 slot for networking (1) SW HHHL PCIe 5.0 x 16 slot (reserved)
Cooling	Air cooling (design reserved for liquid cooling)	
Power	3+1 2700W/3200W CRPS titanium PSUs	
Dimensions	(W) 438 x (H) 176 x (D) 800mm	

NVIDIA HGX Architecture Systems

- Support NVIDIA HGX™ H100/H200 8-GPU
- Future-proof design for NVIDIA HGX™ B200 8-GPU
- Modular motherboard tray design supporting both Intel and AMD platforms
- Liquid cooling available for both Intel and AMD platforms

QuantaGrid D75F-7U

Delivering Advanced Performance for the Most Extreme AI-HPC Workloads

Form Factor	CPU Number	Memory Number
7U	2	32

#AI Training #Scientific Computing #HPC #Data Visualization #3D Modeling

- Supports the NVIDIA HGX™ H200 8-GPU
- Modularized system design for optimal flexibility and easy serviceability
- Support NVIDIA DPU and NVIDIA BlueField®-3 SuperNICs to do the N-S & E-W data transfer, minimizing the bottleneck and empowering the performance for the most complex AI-HPC workloads

Processor	(2) Intel® Xeon® 4th/5th Gen CPUs, up to 350W (8) NVIDIA H200 SXM5 8-GPU, 700W TDP
Networking	(1) Dedicated 1GbE management port
Storage	Front (18) hot-swap 2.5" PCIe 5.0 NVMe SSDs
Expansion Slot	(2) FHHL DW PCIe 5.0 x16 slots (8) HHHL SW PCIe 5.0 x16 slots
Dimensions	(W) 447.8 x (H) 307.85 x (D) 886 mm

QuantaGrid D75M-5U

Accelerated Performance for Modern AI-HPC Workloads

Form Factor	CPU Number	Memory Number
5U	2	24

#AI Training #Natural Language Processing (NLP) #HPC #GenAI #Large Language Models (LLMs)

- Multiple-GPU server for HPC/AI Training
- Powered by 2x AMD EPYC™ 9005 Series processors; designed with the NVIDIA H200 Architecture
- Liquid cooling by design
- Modularized design for easy serviceability

Processor	(2) AMD EPYC™ 9005 Series processors, up to 500W (8) NVIDIA H200 SXM5 8-GPU, 700W TDP
Networking	(1) Dedicated 1GbE management port
Storage	(18) hot-swappable 2.5" PCIe 5.0 NVMe SSDs
Expansion Slot	(2) FHHL DW PCIe 5.0 x16 slots (8) HHHL SW PCIe 5.0 x16 slots
Dimensions	(W) 447.8 x (H) 219.5 x (D) 950 mm

QuantaGrid D74F-7U

Modularized System Design for Optimal Flexibility

Form Factor	CPU Number	Memory Number
7U	2	32

#AI Training #Scientific Computing
#HPC #Data Visualization #3D Modeling

- Supports the NVIDIA HGX™ H200 8-GPU
- Modularized system design for easy serviceability
- Supports NVIDIA DPU to do N-S data transfer, minimizing the bottleneck and empowering the performance for the most complex AI-HPC workloads

Processor	(2) 5th/4th Gen Intel® Xeon® Scalable processors, up to 350W (8) NVIDIA H200 SXM5 8-GPU, 700W TDP
Networking	(1) Dedicated 1GbE management port
Storage	Front (18) hot-swap 2.5" PCIe 5.0 NVMe SSDs
Expansion Slot	(1) FHHL SW PCIe 5.0 x16 slot (1) OCP NIC 3.0 SFF PCIe 5.0 x16 slot (10) OCP NIC 3.0 TSFF PCIe 5.0 x16 slots
Dimensions	(W) 447.8 x (H) 307.85 x (D) 886 mm

QuantaGrid D74H-7U & D74A-7U

Advanced Performance for the Most Extreme AI-HPC Workloads

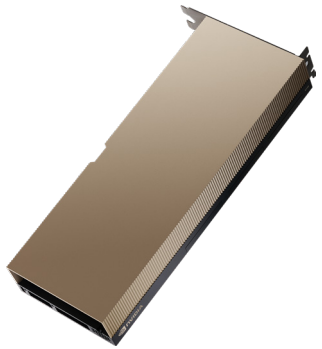
Form Factor	CPU Number	Memory Number
7U	2	32

#Generative Adversarial Networks (GAN)
#Natural Language Processing (NLP)
#Large Language Models (LLMs) #GenAI

- System optimized for Generative AI, Large Language Models (LLMs) and HPC workloads
- Modularized system design for optimal flexibility and easy serviceability

Processor	**[D74H-7U]** (2) 5th/4th Gen Intel® Xeon® Scalable processors, up to 350W TDP (8) NVIDIA H200 SXM5 8-GPU, 700W TDP **[D74A-7U]** (2) 4th Gen AMD EPYC™ processors, up to 400W TDP (8) NVIDIA H200 SXM5 8 GPU, 700W TDP
Networking	(1) Dedicated 1GbE management port
Storage	Front (18) 2.5" hot-pluggable NVMe drives
Expansion Slot	(2) PCIe 5.0 x16 OCP 3.0 SFF slots (10) PCIe 5.0 x16 OCP 3.0 TSFF slots
Dimensions	(W) 447.8 x (H) 307.85 x (D) 886mm

NVIDIA® PCIe GPU Features



4.68x Bandwidth over PCIe Gen 5 for GPU-GPU interconnection with 600 GB/s for NVLink Bridge

QCT has adopted a rich portfolio of NVIDIA cutting-edge GPUs to accelerate some of the world's most demanding workloads including AI, HPC and data analytics, pushing the boundaries of innovation from cloud to edge

AI | NVIDIA Omniverse™ | HPC & Data Analytics | Video Analytics | Virtual Workstation | 3D Modeling | Scientific Computing

QuantaGrid D54U-3U

Endless Flexibility Optimized for Diverse AI and HPC Workloads

Form Factor	CPU Number	Memory Number	SW GPU Supported	DW GPU Supported
3U	2	32	8	4

#3D Modeling #Cloud Gaming #AI Inferencing #NVIDIA Omniverse™ #Scientific Computing #Image Classification #Video Streaming

- Flexible GPU configurations, targeting AI inferencing, training and HPC workloads
- Tool-less GPU module design for easy serviceability

Processor	(2) 5th/4th Gen Intel® Xeon® Scalable processors, up to 350W TDP
Networking	(1) Dedicated 1GbE management port
Accelerator	NVIDIA H100 GPU, NVIDIA L40S GPU, NVIDIA L40 GPU, NVIDIA A16 GPU, NVIDIA A2 GPU
Storage	(10) 2.5" hot-pluggable SATA/SAS/NVMe drives
Expansion Slot	(1) PCIe 5.0 x16 OCP 3.0 slot (2) PCIe 5.0 x16 HHHL slots (1) PCIe 5.0 x8 HHHL slot
Dimensions	(W) 438 x (H) 131.6 x (D) 760mm

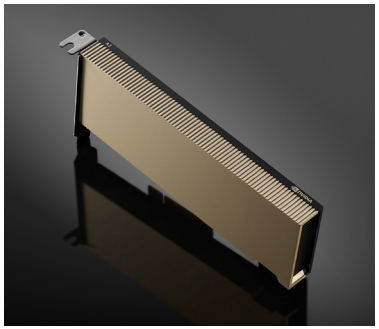
QuantaGrid D55Q-2U

Scale Ahead for the Workload of Tomorrow

Form Factor	CPU Number	Memory Number	SW GPU Supported	DW GPU Supported
2U	2	32	4	3

#AI Inferencing #HPC #Enterprise

Processor	(2) Intel® Xeon® 6 processors, up to 330W/350W
Networking	(1) Dedicated 1GbE management port
Accelerator	NVIDIA H100 GPU, NVIDIA L40S GPU, NVIDIA L4 GPU
Storage	**[SKU -1]** (12) 3.5" hot-pluggable SATA/SAS or (12) 2.5" hot-pluggable NVMe drives **[SKU -2]** (24) 2.5" hot-plug NVMe/ SATA/ SAS drives
Expansion Slot	**Option 1** (4) FHHL + (3) HHHL + (2) OCP 3.0 **Option 2** (4) FHFL + (3) HHHL + (2) OCP 3.0 (supports SW GPU) **Option 3** (3) FHFL + (3) HHHL + (2) OCP 3.0 (supports DW GPU)
Dimensions	(W) 440 x (H) 87.5 x (D) 780mm



Available with Intel CPU platforms, these servers can support up to 2 or 3 single-width GPUs like the NVIDIA L4 GPU or the NVIDIA A2 GPU in a 1U form factor to run AI and HPC workloads.

QuantaGrid D55X-1U

Ultimate Compute Performance and Security Architecture

Form Factor	CPU Number	Memory Number	SW GPU Supported
1U	2	32	3

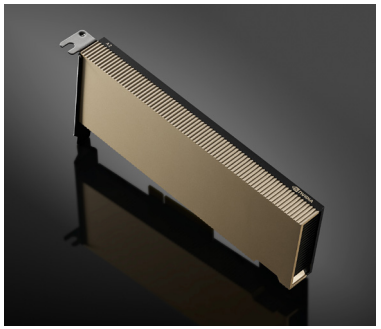
Processor	(2) Intel® Xeon® 6 processor, up to 330W/350W
Networking	(1) Dedicated 1GbE management port
Accelerator	NVIDIA L4 GPU, NVIDIA A2 GPU
Storage	**Option 1:** (12) 2.5" drives **Option 2:** (16) E1.S drives **Option 3:** (20) E3.S drives **Option 4:** (4) E3.S drives + (8) E3.S CXL memory
Expansion Slot	**Option 1** (3) HHHL + (2) OCP 3.0 **Option 2** (2) FHHL + (2) OCP 3.0
Dimensions	(W) 440 x (H) 43.2 x (D) 780mm

QuantaGrid S55R-1U

Your Gateway to Cost-efficient, Enterprise-grade Computing Excellence

Form Factor	CPU Number	Memory Number	SW GPU Supported
1U	1	16	2

Processor	(1) Intel® Xeon® 6 processor, up to 350W TDP
Networking	(1) Dedicated 1GbE management port
Accelerator	NVIDIA L4 GPU, NVIDIA A2 GPU
Storage	**[All 2.5" U.2]** (12) 2.5" NVMe SSDs (12) 2.5" SATA/SAS SSDs (4) 2.5" NVMe RAID + (8) NVMe SSDs
Expansion Slot	**Option 1** (2) PCIe 5.0 x16 HHHL slots + (1) OCP 3.0 **Option 2** (2) PCIe 5.0 x16 FHHL slots + (1) OCP 3.0
Dimensions	(W) 440 x (H) 43.2 x (D) 780mm



Available with AMD CPU platforms, these servers can support up to 2 single-width GPUs like the NVIDIA L4 GPU or the NVIDIA A2 GPU in a 1U form factor to run AI and HPC workloads.

QuantaGrid S44NL-1U

Cost-effective Server with Scalable Expansion Options and Optimized Performance

Form Factor	CPU Number	Memory Number	SW GPU Supported
1U	1	24	2

Processor	(1) AMD EPYC™ 9005 Series processor, up to 500W TDP
Networking	(1) Dedicated 1GbE management port
Accelerator	NVIDIA L4 GPU, NVIDIA A2 GPU
Storage	**[All U.2]** Front (12) 2.5" hot-pluggable NVMe/SATA/SAS drives **[All E1.S]** Front (12) or (16) E1.S hot-pluggable NVMe drives
Expansion Slot	**Option 1** (2) PCIe 5.0 x16 FHHL slots (2) PCIe 5.0 x16 OCP 3.0 slots **Option 2** (2) PCIe 5.0 x16 HHHL slots (2) PCIe 5.0 x16 OCP 3.0 slots **Option 3** (3) PCIe 5.0 x16 HHHL slots (2) PCIe 5.0 x16 OCP 3.0 slots
Dimensions	(W) 440x (H) 43.2 x (D) 780mm

QuantaGrid D44N-1U

Ultimate 1U Server with CPU Performance Breakthrough

Form Factor	CPU Number	Memory Number	SW GPU Supported
1U	2	24	2

Processor	(2) AMD EPYC™ 9005 Series processors, up to 500W TDP
Networking	(1) Dedicated 1GbE management port
Accelerator	NVIDIA L4 GPU, NVIDIA A2 GPU
Storage	**[All U.2]** Front (12) 2.5" hot-pluggable NVMe/SATA/SAS drives **[All E1.S]** Front (12) or (16) E1.S hot-pluggable NVMe drives
Expansion Slot	**Option 1** (2) PCIe 5.0 x16 FHHL slots (2) PCIe 5.0 x16 OCP 3.0 slots **Option 2** (2) PCIe 5.0 x16 HHHL slots (2) PCIe 5.0 x16 OCP 3.0 slots **Option 3** (3) PCIe 5.0 x16 HHHL slots (2) PCIe 5.0 x16 OCP 3.0 slots
Dimensions	(W) 440 x (H) 43.2 x (D) 780mm

QCT QuantaPlex Multi-node Servers

Compact multi-node servers designed for seamless scalability and high-performance computing. Flexible to support both PCIe GPUs and the NVIDIA BlueField®-3 DPU, these systems are ideal for demanding AI, ML, and HPC workloads in a dense, energy-efficient form factor.

QuantaPlex S25Z-2U

2U2N Multi-node Server

Form Factor	CPU Number	Memory Number	SW GPU Supported	DW GPU Supported
2U	2	16	2 per node	1 per node

#HPC #Cloud Gaming #Scale-Out Object Storage
#Multi-Purpose CDN # IoT Edge Computing
#Deep Learning/AI/Machine Learning

- DC-MHS (M-DNO) reference architecture
- Optimized serviceability with front-access, tool-less, hot-pluggable design

*Per node

Processor	(1) Intel® Xeon® 6 processor
Networking	Flexible network options with (1) OCP 3.0 SFF slot and (3) PCIe slots per node
Accelerator	NVIDIA L40S GPU, NVIDIA L4 GPU
Storage	(3) 2.5" NVMe SSDs
Expansion Slot	Per slot (1) PCIe 5.0 x16 FHFL SW (1)PCIe 5.0 x16 FHHL SW (1) PCIe 5.0 x8 FHFL SW (1) PCIe 5.0 x16 OCP 3.0 SFF Or (1) PCIe 5.0 x16 FHFL DW (1)PCIe 5.0 x16 FHHL SW (1) PCIe 5.0 x16 OCP 3.0 SFF
Dimensions	(W) 447.8 x (H) 86.3 x (D) 875 mm

QuantaPlex S45Z-2U

2U4N Multi-node Expansion Server

Form Factor	CPU Number	Memory Number	SW GPU Supported
2U	4	16	1 per node

#HPC #Cloud Gaming #Scale-Out Object Storage
#Multi-Purpose CDN # IoT Edge Computing
#Deep Learning/AI/Machine Learning

- DC-MHS (M-DNO) reference architecture
- Optimized serviceability with front-access, tool-less, hot-pluggable design.

*Per node

Processor	(1) Intel® Xeon® 6 processor
Networking	Flexible network options with (1) OCP 3.0 SFF slot and (3) PCIe slots per node
Accelerator	NVIDIA L4 GPU
Storage	(2) E1.S SSDs
Expansion Slot	(1) PCIe 5.0 x16 HHHL SW (1) PCIe 5.0 x16 OCP 3.0 SFF
Dimensions	(W) 447.8 x (H) 86.3 x (D) 875 mm

QCT QuantaEdge Edge Servers

These servers are available for NVIDIA GPU expansion to bring AI to the edge. The QuantaEdge EGX88D-1U, QuantaEdge EGX77B-1U, and QuantaEdge EGX74I-1U can support single- width GPUs to deliver faster time-to-results and higher-accuracy for image classification and computer vision workloads.

QuantaEdge EGX88D-1U

Carrier-Grade Flexible 300 mm MEC Server

Form Factor	CPU Number	Memory Number	SW GPU Supported
1U	1	8	1

- 300 mm ultra short-depth chassis, easy access for cable management
- Sufficient space for airflow in rack/cabinet
- vRAN acceleration fully integrated

Processor	(1) Intel® Xeon® 6 processor, up to TDP 325W
Networking	**[Density SKU]** (24) 25GbE SFP28 (LoM & QCT LAN BD) **[Expansion SKU]** (16) 25GbE SFP28 (LoM) **[OCP3.0 D-SFF SKU]** (24) 25GbE SFP28 (LoM & Intel Carter Flat)
Accelerator	NVIDIA L4 GPU, NVIDIA A2 GPU
Storage	(2) SATA/NVMe M.2 2280/22110
Expansion Slot	**[Expansion SKU]** (1) FHHL PCIe 5.0 **[OCP3.0 D-SFF SKU]** (1) OCP 3.0 PCIe 5.0 for Intel Carter Flat
Dimensions	(W) 447.8 x (H) 42.8 x (D) 300.65 mm (ear to rear wall)

QuantaEdge EGX77B-1U

Carrier-Grade Flexible 300 mm MEC Server

Form Factor	CPU Number	Memory Number	SW GPU Supported
1U	1	8	1

- 300 mm ultra short-depth server
- NEBS GR63 Level 3 compliant (GR3108 Class 2 optional)
- All LOM support 1588 + SyncE
- Thermal optimization

Processor	(1) 5th/4th Gen Intel® Xeon® Scalable processor, up to 250W TDP
Networking	**[Config 1]** (8) 25GbE w/ Sync-E, NCSI **[Config 2]** (4) 25GbE and (8) 10GbE w/ Sync-E, NCSI
Accelerator	NVIDIA L4 GPU, NVIDIA A2 GPU
Storage	(2) SATA/NVMe M.2 2280
Expansion Slot	(1) FHHL PCIe 5.0 x16
Dimensions	(W) 447.8 x (H) 42.8 x (D) 300.65 mm (ear to rear wall)

QuantaEdge EGX74I-1U

Carrier-Grade Flexible 400mm MEC Server

Form Factor	CPU Number	Memory Number	SW GPU Supported
1U	1	8	3

- Embedded FEC accelerator for accelerating high-performance, energy-efficient vRAN developments
- 400 mm ultra short-depth form factor

Processor	(1) 5th/4th Gen Intel® Xeon® Scalable processor, up to 250W TDP
Networking	(4) 25GbE SFP28 ports (NCSI) (1) Dedicated 1GbE management port
Accelerator	NVIDIA L4 GPU, NVIDIA A2 GPU
Storage	**[Expansion SKU]** (2) SATA/NVMe M.2 2280 drives **[Storage SKU]** (2) SATA/NVMe M.2 2280 drives (2) 2.5" U.2 SSDs
Expansion Slot	**[Expansion SKU]** (2) FH3/4L PCIe 5.0 x16 slots (1) FHHL PCIe 5.0 x16 slot **[Storage SKU]** (2) FH3/4L PCIe 5.0 x16 slots
Dimensions	(W) 447.8 x (H) 42.8 x (D) 400 mm

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 component 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. For more information, visit the QCT website at www.QCT.io.

QCT Authorized Partner

www.QCT.io

Contact Us

© 2025 Quanta Computer Inc. All rights reserved. Specifications and figures are subject to change without prior notice.

All other brand trademarks, logos, and names are the property of their respective owners. All campaign statements and product images contained herein are copyrighted and may not be reprinted and/or reproduced, in whole or in part without the written consent of Quanta Computer Inc.

QCT, the QCT logo, Quanta, and the Quanta logo are trademarks or registered trademarks of Quanta Computer Inc. QCT shall not be liable for technical or editorial errors or omissions contained herein.

NVIDIA, the NVIDIA logo, NVIDIA Grace Hopper, NVIDIA MGX, NVIDIA HGX and NVLink are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries.

ver March., 2025