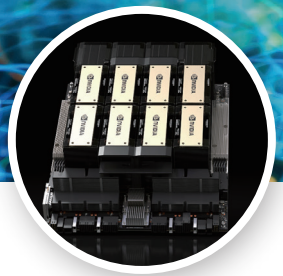
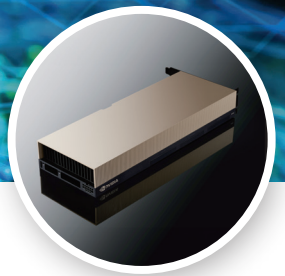
QCT Cutting-Edge Infrastructures Accelerated by NVIDIA

NVIDIA MGX™



NVIDIA HGX™



NVIDIA PCIe GPU

QCT | NVIDIA

QCT's NVIDIA MGX Architecture Product Family

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

QCT's NVIDIA HGX Platform Product Family

| QuantaGrid D75H-10U | QuantaGrid D75L-2U | QuantaGrid D75F-7U | QuantaGrid D74F-7U | QuantaGrid D74H-7U | QuantaGrid D74M-5U |

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

| QuantaGrid D54U-3U | QuantaGrid D55Q-2U | QuantaGrid D55X-1U | QuantaGrid S55R-1U | QuantaGrid S44NL-1U | QuantaGrid D44N-1U | QuantaPlex S25Z-2U | QuantaPlex S45Z-2U | QuantaEdge EGX88D-1U | QuantaEdge EGX77B-1U | QuantaEdge EGX74I-1U |

NVIDIA GB300 NVL72

The NVIDIA GB300 NVL72 brings enhanced compute and memory capabilities to the next generation of AI and accelerated computing with 72 interconnected NVIDIA Blackwell Ultra GPUs acting as one gigantic GPU.

Accelerated by the NVIDIA GB300 Grace Blackwell Ultra Superchip, the NVIDIA GB300 NVL72 offered by QCT comprises 18 QuantaGrid D75U-1U servers. Each server is designed as a single HPM supporting 2 NVIDIA Grace™ CPUs and 4 NVIDIA Blackwell Ultra GPUs, utilizing SXM7 modules for flexibility and serviceability.

Additionally, the NVIDIA GB300 NVL72 uses NVIDIA NVLink™ and showcases energy-efficient liquid cooling. These innovations minimize one's carbon footprint, optimize energy usage, and maximize compute density, all while enhancing GPU communication through high-bandwidth and low-latency integration. With such features, QCT servers accelerated by NVIDIA Blackwell Ultra GPUs deliver exceptional performance for various AI applications.

Gen-to-Gen Comparison

Seamless integration of D75U-1U GB300 NVL72 into existing D75B-1U GB200 NVL72 infrastructure, both built on the NVIDIA GB200 Grace Blackwell Superchip, achieves improved readiness, design excellence, and development efficiency.

Nearly all small boards are shared between GB200 and GB300. Only the PDB is modified to support GB300 higher power consumption. E1.S BP is slightly adjusted for different signal sources.

Mechanical and thermal components remain the same but only adjusting water pressure to accommodate higher power consumption and performance.

	GB300 NVL72	GB200 NVL72
Blackwell GPUs | Grace CPUs	72 | 36	
CPU Cores	2592 Arm Neoverse V2 cores	
Total FP4 Tensor Core	1,080 | 1,440 PFLOPS	720 | 1,440 PFLOPS
Total FP8/FP6 Tensor Core	720 PFLOPS	
Total Fast Memory	Up to 39 TB	30 TB
Total Memory Bandwidth	Up to 576 TB/s	576 TB/s
Total NVLink Bandwidth	130 TB/s	
	Per Blackwell SXM Specifications	
GPU Memory | Bandwidth	Up to 288GB HBM3e | 8 TB/s	186GB HBM3e | 8 TB/s
Max Thermal Design Power (TDP)	Configurable up to 1,400W	Configurable up to 1,200W
Interconnect	5th Generation NVIDIA NVLink™: 1.8TB/s PCIe Gen6: 256GB/s	5th Generation NVIDIA NVLink™: 1.8TB/s PCIe Gen5: 128GB/s

Source: NVIDIA

QuantaGrid D75U-1U (NVIDIA MGX™ Architecture)

The QuantaGrid D75U-1U serves as the compute tray to the NVIDIA GB300 NVL72 produced by QCT. Accelerated by the NVIDIA GB300 Grace Blackwell Ultra Superchips, it features enhanced compute and memory, network upgrades with the latest NVIDIA ConnectX®-8 SuperNIC, and simplified cable routing thanks to the NVIDIA GB300 NVL72's upgraded power distribution design. The D75U-1U-based GB300 NVL72 allows for seamless L11/ data center integration with our existing NVIDIA GB200 NVL72 infrastructure. Moreover, it delivers higher performance at the same power consumption as the NVIDIA GB200 NVL72, and uses efficient liquid cooling to save energy while maximizing compute density and floor space.

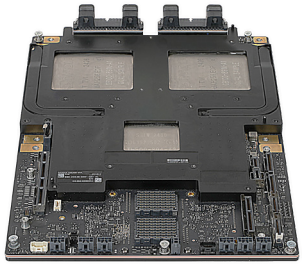
QuantaGrid D75U-1U

Form Factor	SuperChip Number
1U	2

Platform	(2) NVIDIA GB300 Grace Blackwell Ultra Superchip
Processor GPU	(1) NVIDIA Grace™ CPU (2) NVIDIA Blackwell Ultra GPUs
Memory	CPU: Up to 512GB LPDDR5X per CPU GPU: Up to 288GB HBM3e embedded per GPU
Storage	(4) E1.S 15mm PCIe SSDs, (8) slots available
Onboard Storage	(1) PCIe M.2 22110/2280 SSDs
Networking	(1) NVIDIA® BlueField®-3 B3240 400G DPU (4) NVIDIA ConnectX®-8 800Gb OSFP ports (1) FHHL PCIe 5.0 x16 slot
Power	48-54V DC bus bar
Dimensions	(W) 438 x (H) 43.6 x (D) 766mm

NVIDIA GB200 NVL72

Delivering Trillion-Parameter LLM Training and Real-Time Inference

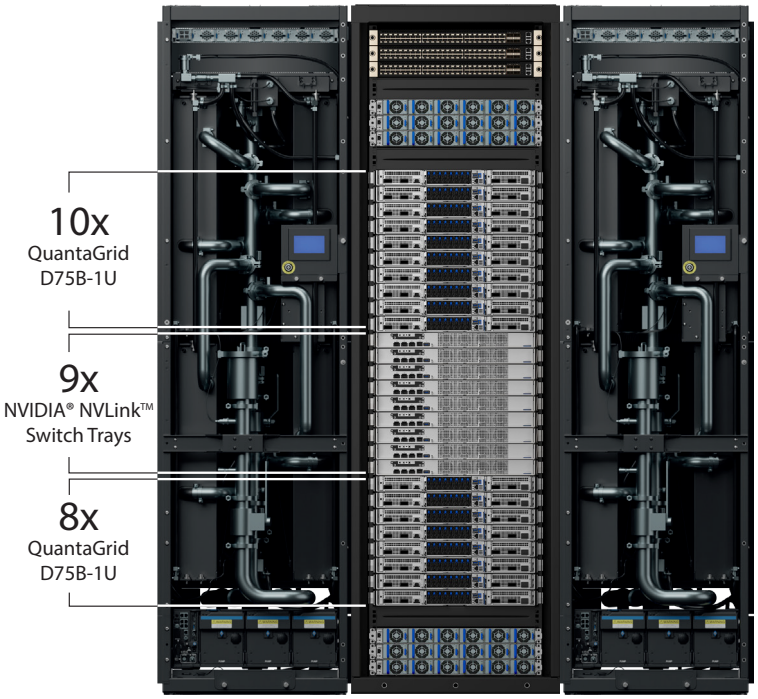


30x Faster LLM Inference compared with NVIDIA H100 GPU

4x Faster LLM Training compared with NVIDIA H100 GPU

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

NVIDIA GB200 NVL72 Populated by QuantaGrid D75B-1U



QCT QoolRack Sidecar heat exchangers for the AI cluster

QuantaGrid D75B-1U (NVIDIA MGX™ Architecture)

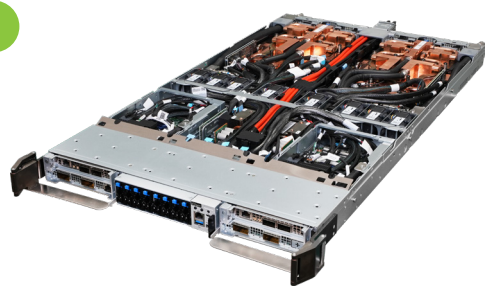
At the heart of the NVIDIA GB200 NVL72 is QCT's QuantaGrid D75B-1U servers. The servers can support up to 2x NVIDIA GB200 Grace Blackwell Superchip in a 1U form factor.

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 NVIDIA Grace™ CPUs and 72 NVIDIA Blackwell, this liquid-cooled, rack-scale system is designed to navigate the complexities of trillion-parameter AI models with unprecedented ease.

QuantaGrid D75B-1U

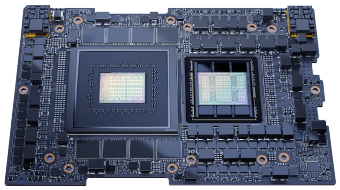
Form Factor	SuperChip Number
1U	2



Platform	(2) NVIDIA GB200 Grace Blackwell Superchip
Processor **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) hot-swappable E1.S 15mm PCIe SSDs
Onboard Storage	(1) PCIe M.2 22110/2280 SSDs
Networking	(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

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 & QuantaEdge EGX77GE-2U (NVIDIA MGX™ Architecture)

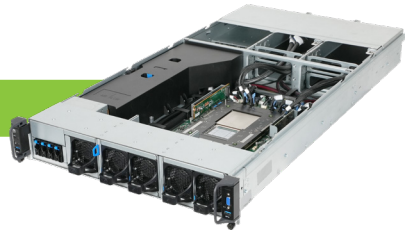
QCT systems accelerated by the NVIDIA MGX™ architecture such as the QuantaGrid S74G-2U, QuantaEdge EGX77GE-2U allow different configurations of GPUs, CPUs and DPUs, shortening the time frame for building future compatible solutions.

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

QuantaGrid S74G-2U



Breakthrough Accelerated Performance for Giant-scale AI-HPC Applications

Form Factor	SuperChip Number
2U	1

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

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

Processor	NVIDIA GH200 Grace Hopper™ Superchip, 1000W TDP
Memory	CPU: Up to 480GB LPDDRX embedded GPU: 144GB HBM3e 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) hot-swappable E1.S NVMe SSDs
Expansion Slot	(3) FHFL DW PCIe 5.0 x16
Dimensions	(W) 438 x (H) 87.5 x (D) 900mm

QuantaEdge EGX77GE-2U

First 400mm Edge Server Adopting NVIDIA MGX Architecture

Form Factor	SuperChip Number
2U	1

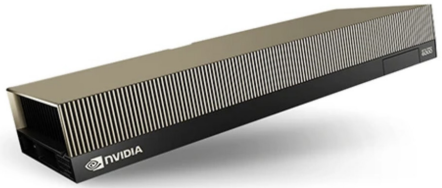
#Smart City #Smart Manufacturing #Real-time Image Analysis #Edge AI

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

Processor	NVIDIA GH200 Grace Hopper™ Superchip, 1000W TDP
Storage	Internal Storage: (2) SATA/NVMe M.2 22110/2280 SSDs External Storage: (2) E1.S SSDs
Expansion Slot	(3) FHFL PCIe 5.0 x16
Dimensions	(W) 447.8 x (H) 86.8 x (D) 400mm

NVIDIA RTX PRO™ 6000 Blackwell Server Edition

Most Efficient Compute for LLM Inference, AI, and HPC



Up to **5.6x** LLM inference performance compared with NVIDIA L40S

Up to **2.2x** Genomics performance compared with H100

Up to **2x** MIG instance compared with NVIDIA L4

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 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.

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
- 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
- Offers infinite flexibility to support any AI/HPC-related workloads

Processor	(2) Intel® Xeon® 6 processors, up to 350W TDP	
Networking	(1) Dedicated 1GbE management port	
Accelerator	Air cooling: 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) FHFL DW PCIe 5.0 x 16 slots for GPU (3) FHFL SW PCIe 5.0 x 16 slots for networking	**Air cooling - (8) DW GPUs:** (8) FHFL DW PCIe 5.0 x 16 slots for GPU (4) FHFL SW PCIe 5.0 x 16 slots for networking (1) FHHL SW PCIe 5.0 x 16 slot for networking (1) HHHL SW 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™ B300 Delivers

11x Faster AI performance for inference (144 petaFLOPS)

4x Speedup for training (72 petaFLOPS)

2x HBM3e memory size (2.3TB HBM3e)

2x Networking bandwidth with onboard NVIDIA ConnectX®-8 SuperNICs

(Compared with NVIDIA Hopper™ platforms)

NVIDIA HGX™ B300 Systems by QCT

- Ideal for accelerating LLM workloads with 72 PFLOPS FP8 for training and 144 PFLOPS FP4 for inference.
- Based on NVIDIA's reference design, with chassis compatible with the NVIDIA MGX™ architecture and traditional enterprise racks.
- Enable hyperscalers to build high-speed, large scale-out GPU clusters

QuantaGrid D75H-10U

Enabling Diverse AI Workloads for Every Enterprise

Form Factor	SuperChip Number	Memory Number
10U	2	32

#AI reasoning #Agentic AI #HPC #Video Inferencing

- Supports NVIDIA HGX™ B300, accelerated by NVIDIA Blackwell Ultra GPUs
- Features PCIe Gen 6 to enable 800G east-west data transfer to build massive-scale GPU clusters
- Supports NVIDIA® BlueField®-3 DPUs for north-south data transfer, minimizing the bottleneck and empowering the most complex AI-HPC workloads

Processor	(2) Intel® Xeon® 6 processors, up to 350W TDP (8) NVIDIA B300 SXM7
Networking	(8) OSFP ports serving (8) single-port NVIDIA ConnectX®-8 VPI (1) Dedicated 1GbE management port
Storage	(8) hot-swappable E1.S SSDs (2) PCIe M.2 2280 SSDs
Expansion Slot	(4) FHHL SW or (2) FHHL DW PCIe 5.0 x16 slots
Dimensions	(W) 447 x (H) 441.75 x (D) 800mm

QuantaGrid D75L-2U

Maximum GPU Density in AI Clusters

Form Factor	SuperChip Number	Memory Number
2U	2	32

#AI reasoning #Agentic AI #HPC #Video Inferencing

- Supports NVIDIA HGX™ B300, accelerated by NVIDIA Blackwell Ultra GPUs
- Features PCIe Gen 6 to enable 800G east-west data transfer to build massive-scale GPU clusters
- Supports NVIDIA® BlueField®-3 DPUs for north-south data transfer, minimizing the bottleneck and empowering the most complex AI-HPC workloads

Processor	(2) Intel® Xeon® 6 processors, up to 350W (8) NVIDIA B300 SXM7
Networking	(8) OSFP ports serving (8) single-port NVIDIA ConnectX®-8 VPI (1) Dedicated 1GbE management port
Storage	(8) hot-swappable E1.S SSDs (2) PCIe M.2 22110 SSDs
Expansion Slot	(4) FHHL SW or (2) FHHL DW PCIe 5.0 x16 slots
Dimensions	(W) 447 x (H) 86.9 x (D) 950mm

NVIDIA HGX Architecture Systems by QCT (Hopper Platform)

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

QuantaGrid D75F-7U & D74F-7U & D74H-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
- Supports flexible expansion options, including OCP or PCIe form factor for north-south & east-west data transfer
- Modularized system design for optimal flexibility and easy serviceability
- Supports NVIDIA® BlueField®-3 DPU and SuperNICs to minimize the bottleneck and empowering the performance for the most complex AI-HPC workloads
- System optimized for generative AI, LLM, and HPC workloads

Processor	(2) 5th/4th Gen Intel® Xeon® Scalable processors, up to 350W TDP (8) NVIDIA H200 SXM 8-GPU, 700W TDP
Networking	(1) Dedicated 1GbE management port
Storage	(18) hot-swappable 2.5" NVMe SSDs
Expansion Slot	**[D75F-7U]:** (2) FHHL DW PCIe 5.0 x16 slots + (8) HHHL SW PCIe 5.0 x16 slots **[D74F-7U]:** (1) FHHL SW PCIe 5.0 x16 slots + (1) OCP 3.0 SFF slot + (10) OCP NIC 3.0 TSFF slots **[D74H-7U]:** (2) OCP 3.0 SFF slots + (10) OCP NIC 3.0 TSFF slots
Dimensions	**[D75F-7U]:** (W) 447.8 x (H) 307.85 x (D) 950mm **[D74F-7U/D74H-7U]:** (W) 447.8 x (H) 307.85 x (D) 886mm

QuantaGrid D75M-5U

Accelerated Performance for Modern AI-HPC Workloads

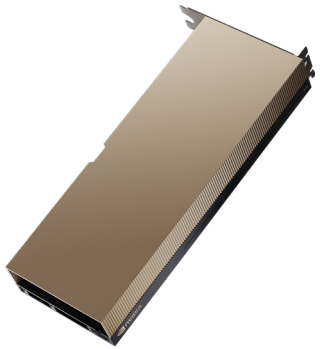
Form Factor	CPU Number	Memory Number
5U	2	24

#AI Training #Natural Language Processing #HPC #Gen AI #Large Language Models

- Accelerated by NVIDIA H200 8-GPU
- Powered by 2x AMD EPYC™ 9004/9005 Series processors
- Liquid cooled by design
- Modularized design for easy serviceability
- Multiple-GPU server for HPC/AI Training

Processor	(2) AMD EPYC™ 9004/9005 Series processors, up to 500W TDP (8) NVIDIA H200 SXM 8-GPU, 700W TDP
Networking	(1) Dedicated 1GbE management port
Storage	(18) hot-swappable 2.5" 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) 950mm

NVIDIA Scalable and Lightweight AI System



4.68x Bandwidth over PCIe Gen 5 for GPU-GPU interconnection with 600 GB/s for NVIDIA 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 #Scientific Computing #Image Classification #Video Streaming #Omniverse

- Powered by dual 5th/4th Gen Intel® Xeon® Scalable processors
- Tool-less GPU module design for easy serviceability
- Flexible GPU configurations, targeting AI inferencing, training and HPC workloads

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) hot-swappable 2.5" SATA/SAS/NVMe SSDs
Expansion Slot	(1) OCP 3.0 PCIe 5.0 x16 slot (2) HHHL PCIe 5.0 x16 slots (1) HHHL PCIe 5.0 x8 slots
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

- Accelerated by NVIDIA GPUs
- Powered by dual Intel® Xeon® 6 processors
- Adopts DC-MHS to promote open standards
- Ultimate compute performance and a workload-driven architecture

Processor	(2) Intel® Xeon® 6 processors, up to 350W TDP
Networking	(1) Dedicated 1GbE management port
Accelerator	NVIDIA H100 GPU, NVIDIA L40S GPU, NVIDIA L4 GPU
Storage	**SKU - #1:** (12) hot-swappable 3.5" SATA/SAS or (12) hot-swappable 2.5" NVMe SSDs **SKU - #2:** (24) hot-swappable 2.5" NVMe/ SATA/ SAS SSDs
Expansion Slot	**Option1:** (4) FHHL PCIe 5.0 x8 slots+ (2) HHHL PCIe 5.0 x16 slots + (1)HHHL PCIe 5.0 x8 slots + (2) OCP 3.0 slots **Option2:** (4) FHHL PCIe 5.0 x8 slots + (2) HHHL PCIe 5.0 x16 slots + (1)HHHL PCIe 5.0 x8 slots + (2) OCP 3.0 slots [supports SW GPU] **Option3:** (3) FHFL PCIe 5.0 x16 slots + (2) HHHL PCIe 5.0 x16 slots + (1)HHHL PCIe 5.0 x8 slots + (2) OCP 3.0 slots [supports DW GPU]
Dimensions	(W) 440 x (H) 87.5 x (D) 780mm

QCT QuantaGrid 1U General Purpose Servers

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

- Accelerated by NVIDIA GPUs
- Powered by Intel® Xeon® 6 processors
- Adopts DC-MHS to promote open standards
- Ultimate compute performance and a workload-driven architecture

Processor	(2) Intel® Xeon® 6 processor, up to 350W TDP
Networking	(1) Dedicated 1GbE management port
Accelerator	NVIDIA L4 GPU, NVIDIA A2 GPU
Storage	**SKU - #1:** (12) hot-swappable 2.5" SSDs **SKU - #2:** (16) hot-swappable E1.S SSDs **SKU - #3:** (20) hot-swappable E3.S SSDs **SKU - #4:** (4) hot-swappable E3.S SSDs + (8) E3.S CXL memory
Expansion Slot	**Option 1** (3) HHHL PCIe 5.0 x16 slots + (2) OCP 3.0 slots **Option 2** (2) FHHL PCIe 5.0 x16 slots + (2) OCP 3.0 slots
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

- Powered by a single Intel® Xeon® 6 processor
- A cost-efficient 1U1P density compute architecture
- Adopts DC-MHS to promote open standards
- System optimized for HPC and cloud computing

Processor	(1) Intel® Xeon® 6 processor, up to 350W TDP
Networking	(1) Dedicated 1GbE management port
Accelerator	NVIDIA L4 GPU, NVIDIA A2 GPU
Storage	(12) hot-swappable 2.5“ NVMe/SATA/SAS SSDs
Expansion Slot	**Option 1** (2) HHHL PCIe 5.0 x16 slots + (1) OCP 3.0 slot **Option 2** (2) FHHL PCIe 5.0 x16 slots + (1) OCP 3.0 slot
Dimensions	(W) 440 x (H) 43.2 x (D) 780mm

QuantaGrid S44NL-1U & D44N-1U

Cost-effective Server with Scalable Expansion Options and Optimized Performance

Form Factor	CPU Number	Memory Number	SW GPU Supported
1U	1or2	24	2

- Featuring up to 3x single-width accelerators to support AI inference workloads
- Powered by AMD EPYC™ 9004/9005 Series processors
- Advanced air-cooling architecture supporting top-bin CPUs
- Liquid-cooled design
- Up to 5x PCIe 5.0 expansion slots and the DC-SCM architecture to meet different configuration needs
- System optimized for cloud, enterprise, AI, HPC, networking, security and IoT workloads

Processor	**S44NL-1U:** (1) AMD EPYC™ 9004/9005 Series processor, up to 500W TDP **D44N-1U:** (2) AMD EPYC™ 9004/9005 Series processors, up to 500W TDP
Networking	(1) Dedicated 1GbE management port
Accelerator	NVIDIA L4 GPU, NVIDIA A2 GPU
Storage	**SKU - #1** (12) hot-swappable 2.5" NVMe/SATA/SAS SSDs **SKU - #2** (12) or (16) hot-swappable E1.S NVMe SSDs
Expansion Slot	**Option 1** (2) FHHL PCIe 5.0 x16 slots (2) OCP 3.0 PCIe 5.0 x16 slots **Option 2** (2) HHHL PCIe 5.0 x16 slots (2) OCP 3.0 PCIe 5.0 x16 slots **Option 3** (3) HHHL PCIe 5.0 x16 slots (2) OCP 3.0 PCIe 5.0 x16 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, 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

- Powered by a single Intel® Xeon® 6 processor
- High density 2U2N architecture
- Optimized serviceability with a front-access, tool-less, hot- swappable design
- Adopts DC-MHS (M-DNO) to promote open standards

*Per node

Processor	(1) Intel® Xeon® 6 processor, up to 350W TDP
Networking	Flexible network options with (1) OCP 3.0 SFF slot and (3) PCIe slots per node
Accelerator	NVIDIA L40S GPU, NVIDIA L4 GPU, NVIDIA A2 GPU
Storage	(3) hot-swappable 2.5" NVMe SSDs
Expansion Slot	Per slot (1) FHFL SW PCIe 5.0 x16 (1) FHHL SW PCIe 5.0 x16 (1) FHFL SW PCIe 5.0 x8 (1) OCP 3.0 SFF PCIe 5.0 x16 Or (1) FHFL DW PCIe 5.0 x16 (1) FHHL SW PCIe 5.0 x16 (1) OCP 3.0 SFF PCIe 5.0 x16
Dimensions	(W) 447.8 x (H) 86.3 x (D) 875mm

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

- Powered by a single Intel® Xeon® 6 processor
- High density 2U4N architecture
- Optimized serviceability with a front-access, tool-less, hot-swappable design
- Adopts DC-MHS (M-DNO) to promote open standards

*Per node

Processor	(1) Intel® Xeon® 6 processor, up to 350W TDP
Networking	Flexible network options with (1) OCP 3.0 SFF slot and (3) PCIe slots per node
Accelerator	NVIDIA L4 GPU, NVIDIA A2 GPU
Storage	(2) hot-swappable E1.S SSDs
Expansion Slot	(1) HHHL SW PCIe 5.0 x16 (1) OCP 3.0 SFF PCIe 5.0 x16
Dimensions	(W) 447.8 x (H) 86.3 x (D) 875mm

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 300mm MEC Server

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

- Powered by a single Intel® Xeon® 6 processor
- 300mm ultra short-depth design
- Easy access for cable management
- Sufficient space for airflow in rack/cabinet
- Fully integrates vRAN acceleration
- Support 24xSFP28 with SyncE LAN ports in a 1U chassis
- Integrates GNSS
- 1+1 redundant AC/DC PSU
- Front access

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

QuantaEdge EGX77B-1U

Carrier-Grade Flexible 300mm MEC Server

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

- Powered by a single 5th/4th Gen Intel® Xeon® Scalable processor
- 300mm ultra short-depth design
- NEBS GR63 Level 3 compliant (GR3108 Class 2 optional)
- Operating temperature between -5°C ~ 55°C (-40°C ~ 65°C optional)
- All LOM support 1588 + SyncE
- Thermal optimization
- PFR function reserved

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

- Powered by a single 5th/4th Gen Intel® Xeon® Scalable processor
- 400mm ultra short-depth design
- NEBS GR63 Level 3 compliant (GR3108 Class 2 optional)
- Operating temperature between -5°C ~ 55°C (-40°C ~ 65°C optional)
- SMA connections reserved
- PFR function reserved

Processor	(1) 5th/4th Gen Intel® Xeon® Scalable processor, up to 250W TDP
Networking	(4) 25GbE SFP28 ports (NCSI) (1) 1GbE RJ45 management port
Accelerator	NVIDIA L4 GPU, NVIDIA A2 GPU
Storage	**SKU - #1** (2) SATA/NVMe M.2 2280 SSDs **SKU - #2** (2) SATA/NVMe M.2 2280 SSDs (2) 2.5" U.2 SSDs
Expansion Slot	**SKU - #1** (2) FH3/4L PCIe 5.0 x16 slots (1) FHHL PCIe 5.0 x16 slot **SKU - #2** (2) FH3/4L PCIe 5.0 x16 slots
Dimensions	(W) 447.8 x (H) 42.8 x (D) 400mm

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

ver May., 2025