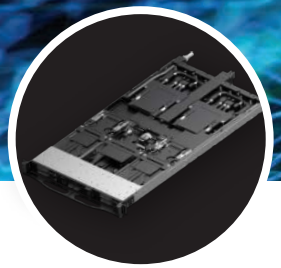
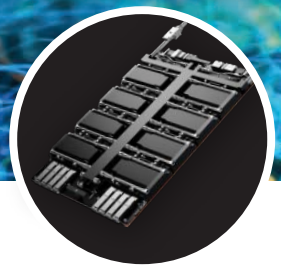
QCT Cutting-Edge Infrastructures Accelerated by NVIDIA





NVIDIA MGX™



NVIDIA HGX™



NVIDIA PCIe GPU

QCT's NVIDIA MGX Architecture Product Family

| QuantaGrid D76V-1U | QuantaGrid D66Q-2U | QuantaGrid T46X-1U | QuantaPlex S26F-2U | QuantaGrid D75U-1U |
| QuantaGrid D75B-1U | QuantaGrid D75E-4U | QuantaEdge EGN77C-2U | QuantaGrid S74G-2U |

QCT's NVIDIA HGX Platform Product Family

| QuantaGrid D76T-2U | QuantaGrid D76W-2U | QuantaGrid D75H-10U | QuantaGrid D75L-3U |
| QuantaGrid D75F-7U | QuantaGrid D74F-7U | QuantaGrid D74H-7U |

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 |

QCT® | NVIDIA

NVIDIA Vera Rubin NVL72



NVIDIA Vera Rubin NVL72 is a next-generation, rack-scale AI platform unifying 72 NVIDIA Rubin GPUs and 36 NVIDIA Vera CPUs to deliver massive compute density and up to 20.7 TB of HBM4 memory. It leverages NVIDIA NVLink™ 6 switches to interconnect GPUs for scale-up intelligence, integrating the NVIDIA ConnectX®-9 SuperNIC™ cards and NVIDIA® BlueField®-4 DPU for elevated networking capabilities.

Designed for expanded scale-out deployments, NVIDIA Vera Rubin NVL72 supports NVIDIA Quantum-X800 InfiniBand and NVIDIA Spectrum™-X Ethernet to deliver breakthrough performance for LLM, AI reasoning, agentic AI and video inferencing applications.

	NVIDIA Vera Rubin NVL72	NVIDIA GB300 NVL72
GPU | CPU	72 NVIDIA Rubin GPUs 36 NVIDIA Vera CPUs	72 NVIDIA Blackwell Ultra GPUs 36 NVIDIA Grace™ CPUs
CPU Cores	3,168 NVIDIA Olympus Cores	2592 Arm Neoverse V2 cores
Total NVFP4 Inference	3,600 PFLOPS	1,440 PFLOPS
GPU Memory | Bandwidth	Up to 20.7TB | Up to 1,400TB/s	Up to 20TB | Up to 576TB/s
CPU Memory	Up to 54 TB	Up to 17TB LPDDR5X | Up to 14TB/s
Total NVLink Bandwidth	Up to 216 TB/s	130TB/s
Total GPU subsystem power (TGP)	Max Q: 1800W Max P: 2300W	1,400W

Source: NVIDIA

QuantaGrid D76V-1U (NVIDIA MGX™ Architecture)

Enhanced Compute & HBM4 Bandwidth: Optimized compute and 288GB HBM4 memory per GPU to increase AI inferencing and training, integrated with almost 3x HBM4 bandwidth and 2x NVIDIA NVLink™ 6 bandwidth to elevate GPU-to-GPU communication efficiency with low latency.

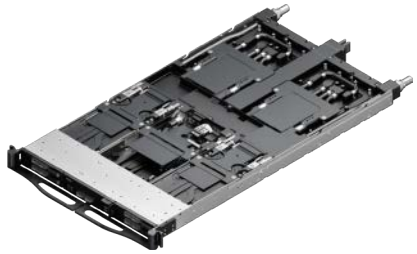
Network Upgrades: New NVIDIA ConnectX®-9 SuperNIC™ cards deliver 1.6 Tb/s connectivity per GPU in NVIDIA Vera Rubin systems, with enhanced networking platforms for peak AI workload efficiency.

Optimized Serviceability: The NVIDIA Vera Rubin system with cableless design provides board to board connection through PHD2 connectors to deliver high speed signals and reduce cable routing effort. Along with modularized demonstration in chassis, NVIDIA Vera Rubin simplifies assembly and serviceability complexity and exerts outstanding mechanical technique to the tailor-made system.

Efficient Cooling: Liquid-cooled NVIDIA Vera Rubin NVL72 racks maximize compute density and floor space utilization to align with carbon-free footprint and ESG goals, fulfilling 100% liquid cooling for the Vera Rubin NVL72 compute and switch trays.

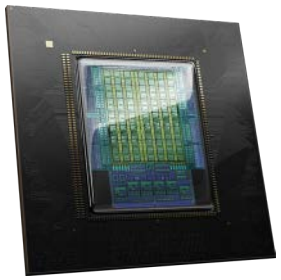
QuantaGrid D76V-1U

Form Factor	SuperChip Number
1U	*2*



Platform	(2) NVIDIA Vera Rubin Superchips
CPU **GPU**	(2) NVIDIA Vera CPUs (4) NVIDIA Rubin GPUs
Memory	**CPU:** Up to 1.5 TB LPDDR5X per CPU **GPU:** Up to 288GB HBM4 per GPU
Storage	(4) E1.S 9.5mm PCIe local SSDs (1) E1.S 9.5mm PCIe boot SSD
Networking	(1) NVIDIA® BlueField®-4 B4240V with dual 400Gb/s QSFP112 ports (800Gb/s total bandwidth) (8) NVIDIA ConnectX®-9 800G OSFP ports
Power	48-54V DC busbar clip
Dimensions	(W) 438 x (H) 43.6 x (D) 766mm

NVIDIA Vera CPU



2x Higher energy efficiency compared to traditional CPUs

3x Capacity and efficiency compared to prior generation (1.5 TB)

2x Memory Bandwidth at half the power compared to traditional CPUs (1.2 TB/s)

QuantaPlex T46X-1U



- Powered by NVIDIA Vera CPUs, delivering maximum single-threaded performance at scale, purpose-built for Agentic AI
- Compact, full-liquid cooled 1U4N system with board-to-board design
- Scalable to 32 trays and 256 Vera CPUs per rack for ultra-high compute density
- Optimized for Reinforcement Learning and Agentic AI, supporting 22,500+ parallel CPU sandboxes per rack to scale with massive GPU training workloads

Chassis Spec	
Form factor	1U4N, (4) HPMs in 1U system
Cooling	Full liquid cooled
Power	48-54V DC bus bar
Storage	(8) E1.S 9.5mm SSDs
Network	(4) NVIDIA® BlueField-4 DPU
BMC	(4) Aspeed AST2600
Dimensions	1RU (W) 438 x (H) 43.6 x (D) 766mm (W) 17.24" x (H) 1.72" x (D) 30.16"
Per HPM Spec	
CPU	(2) NVIDIA® Vera™ CPU soldered down
Memory	(16) SOCAMM modules

QuantaGrid D66Q-2U



Form Factor	CPU Number	Memory Number
2U	2	16

#Agentic AI *#CPU Sandboxes*
#HPC *#Reinforcement Learning*

- Supports NVIDIA Vera CPUs, combining leading per-core performance, massive memory bandwidth, and exceptional energy efficiency
- System optimized for high-concurrency AI execution environments
- Utilizing NVIDIA BlueField®-3 DPUs and NVIDIA ConnectX®-8 SuperNIC™ cards, featuring PCIe Gen 6 slots to enable ultra-fast data movement.

Processor	(2) NVIDIA Vera CPU
Memory	(16) SOCAMM, up to 3TB
Networking	(1) Dedicated LAN port (RJ45) for BMC management (2) HHFL SW PCIe 6.0 x16 slot for NVIDIA® ConnectX-8®
Storage	(8) hot-swappable E1.S SSDs
Expansion Slot	(1) FHFL DW PCIe 6.0 x16 slot for NVIDIA® BlueField®-3
Cooling	Air cooled
Power	(1+1) or (2+2) CRPS PSU
Form factor	2U rack mount
Dimensions	(W) 438 x (H) 87 x (D) 900mm

NVIDIA® BlueField®-4 STX Storage Platform using CMX

First NVIDIA partner to deliver the NVIDIA CMX™ Context Memory Storage Platform

16x
QuantaPlex
S26F-2U



The platform introduces a specialized G3.5 context memory tier between node-local SSDs and shared storage, extending effective GPU KV-cache capacity across the pod. By treating KV cache as a first-class AI data type and optimizing pod-level context management, it delivers up to 5x higher tokens-per-second and up to 5x better power efficiency for agentic AI workloads.

This AI-native architecture uses the NVIDIA® B4480SP Dual BF4 Storage Processor, NVIDIA DOCA™ Memos, NVIDIA, Dynamo, and NVIDIA Spectrum-X™ Ethernet to enable seamless, low-latency context sharing across nodes. By rapidly pre-staging KV-cache context back to the GPU during prefill, it minimizes inference stalls and maximizes end-to-end throughput. This ensures that large-scale AI factories remain both high-performing and power- and cost-effective.

Key Features and Benefits

- Adds a dedicated, RDMA-attached storage tier for KV cache between GPU HBM and enterprise storage.
- Enables scalable, shared context reuse across nodes to support long-context, multi-agent inference and agentic AI workloads.
- **5x** tokens per second compared with traditional storage, **4x** higher energy efficiency compared with traditional CPU architectures for high-performance storage, and **2x** more ingestion of pages per second for enterprise AI data.

QuantaPlex S26F-2U (NVIDIA MGX™ Architecture)

Pod-Level Context Memory Storage: A pod-level inference context memory storage platform designed to add a specialized G3.5 context memory between node-local SSDs and shared storage, extending effective GPU KV-cache capacity across the pod.

High-Density NVMe Storage: Supports up to 24 front-accessible, hot-swappable E3.S 1T or E3.L 1T NVMe drive bays and Ethernet for ultra-fast north-south and east-west data movement across AI clusters.

Ultra-High-Speed Networking: Features NVIDIA ConnectX®-9 SuperNIC™ for up to 800Gb connectivity, optimizing network efficiency for large-scale AI deployments.

QuantaPlex S26F-2U

Form Factor	CPU Number
2U	2



Chassis Spec	
Form factor	2U2N
Cooling	Fully liquid-cooled
Power	48-54V DC Busbar
Dimensions	(W) 438 x (H) 87.5 x (D)766mm
Per Node Spec	
Processor	(1) NVIDIA Vera CPU
Memory	(8) SOCAMM DIMM slots, populate (4) SOCAMM for CMX
Onboard Storage	(1) PCIe M.2 2280
Storage	(12) E3.S/E3.L NVMe data drives (PCIe Gen5/Gen6)

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.

Additionally, NVIDIA GB300 NVl72 uses fifth-generation 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.

Built on the NVIDIA MGX™ Architecture: Ready, Reliable, Efficient

The NVIDIA GB300 NVL72 with QuantaGrid D75U-1U operates within a similar thermal, electrical, and mechanical envelope of existing GB200 NVL72 racks and datacenter designs.

The NVIDIA GB300 NVL72 leverages most of the mechanical and thermal components from its predecessor, with only minor water-flow adjustments to support higher power and performance. The high commonality between NVIDIA GB200 NVL72 and NVIDIA GB300 NVL72 system boards ensures strong design maturity and proven reliability.

	GB300 NVL72	GB200 NVL72
Blackwell GPUs | Grace CPUs	72 | 36	
CPU Cores	2592 Arm Neoverse V2 cores	
Total FP4 Tensor Core	1,440 PFLOPS | 1,080 PFLOPS (dense)	1,440 PFLOPS | 720 PFLOPS (dense)
Total FP8/FP6 Tensor Core	720 PFLOPS	
GPU Memory | Bandwidth	Up to 20TB | Up to 576TB/s	Up to 13.4TB HBM3e | 576TB/s
CPU Memory | Bandwidth	Up to 17TB LPDDR5X | Up to 14TB/s	
Total NVLink Bandwidth	130TB/s	
Total GPU subsystem power (TGP)	1,400W	1,200W

Source: NVIDIA

QuantaGrid D75U-1U (NVIDIA MGX™ Architecture)

The QuantaGrid D75U-1U serves as the compute tray for the NVIDIA GB300 NVL72 produced by QCT. Accelerated by the NVIDIA GB300 Grace Blackwell Ultra Superchips, it delivers enhanced compute and memory performance, along with network upgrades enabled by the latest NVIDIA ConnectX®-8 SuperNIC™ cards. With NVIDIA ConnectX®-8 SuperNIC™ cards supporting PCIe switch functionality, the need for IPEX high-speed connector boards and cables is eliminated, ensuring a simplified system design. The D75U-1U unifies the compute tray design for both NVIDIA GB200 NVL72 and NVIDIA GB300 NVL72 under the NVIDIA Blackwell architecture platform, leveraging the NVIDIA MGX™ modular design for seamless L11 and data center integration with existing NVIDIA GB200 NVL72 infrastructure.

QuantaGrid D75U-1U

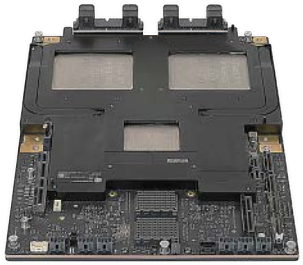
Form Factor	SuperChip Number
1U	*2*



Platform	(2) NVIDIA GB300 Grace Blackwell Ultra Superchip
Processor GPU	(2) NVIDIA Grace™ CPUs (4) NVIDIA Blackwell Ultra GPUs
Memory	**CPU:** Up to 480GB LPDDR5X per CPU **GPU:** Up to 279GB HBM3e 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
Power	48-54V DC busbar
Dimensions	(W) 438 x (H) 43.6 x (D) 766mm

NVIDIA GB200 NVL72

Delivering Trillion-Parameter LLM Training and Real-Time Inference

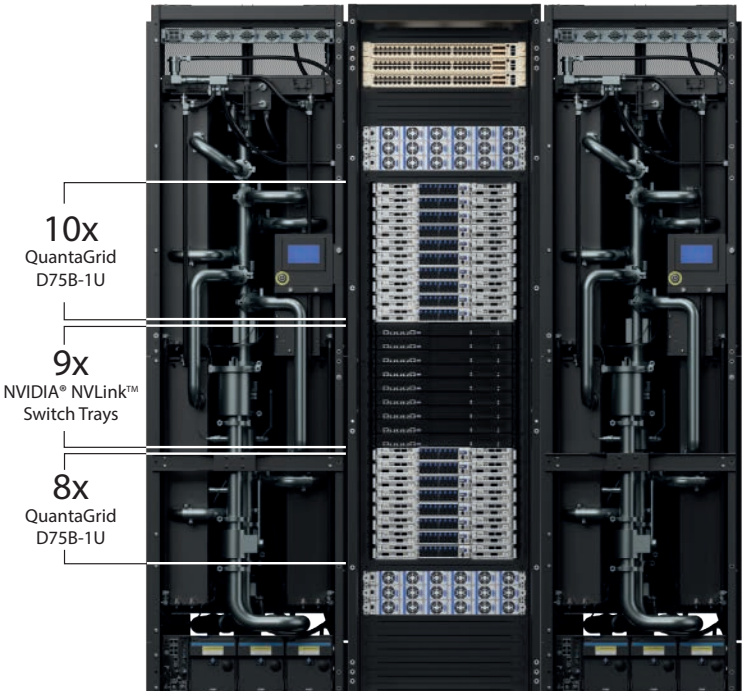


30x Faster LLM Inference compared with NVIDIA H100 GPU

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 Superchips 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 GPUs, this liquid-cooled, rack-scale system is designed to navigate the complexities of trillion-parameter AI models with unprecedented ease.

QuantaGrid D75B-1U

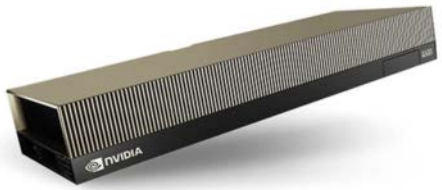
Form Factor	SuperChip Number
1U	2



Platform	(2) NVIDIA GB200 Grace Blackwell Superchips
Processor **GPU**	(2) NVIDIA Grace™ CPUs (4) NVIDIA Blackwell GPUs
Memory	**CPU:** Up to 480GB LPDDR5X per CPU **GPU:** Up to 186GB HBM3e per GPU
Storage	(8) hot-swappable E1.S 15mm PCIe SSDs
Onboard Storage	(1) PCIe M.2 22110/2280 SSD
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 busbar clip
Dimensions	(W) 438 x (H) 43.6 x (D) 766mm

NVIDIA RTX PRO™ 6000 Blackwell Server Edition

Most Efficient for Visual Computing, 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, and NVIDIA L40S 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 to Unlock Infinite AI Possibilities

Form Factor	CPU 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	
Memory	(32) DDR5 RDIMM up to 6,400 MHz	
Networking	(1) Dedicated 1GbE management port	
Storage	**SKU1 - (4) DW GPUs** (12) hot-swappable E1.S SSDs	**SKU2 – (8) DW GPUs** (24) hot-swappable E1.S SSDs
Expansion Slot	**SKU1 - (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	**SKU2 – (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
GPU Expansion	SKU1 - (4) DW GPUs (4) NVIDIA RTX PRO™ 6000, NVIDIA H200, NVIDIA L40S	SKU2 – (8) DW GPUs (8) NVIDIA RTX PRO™ 6000, NVIDIA H200, NVIDIA L40S
Cooling	Air cooling	
Power	(3+1) 2700W/3200W CRPS Titanium PSUs	
Form Factor	4U Rackmount Server	
Dimensions	(W) 438 x (H) 176 x (D) 800mm	

NVIDIA RTX PRO™ 4500 Blackwell Server Edition



Up to **2x isolated instances** enable concurrent execution of multiple workloads

Up to **2x the performance** of the previous generation with FP4 precision for AI model processing

QuantaEdge EGN77C-2U (NVIDIA MGX™ Architecture)

QuantaEdge EGN77C-2U is a carrier-grade server system derived from NVIDIA Aerial RAN Computer (ARC), an AI-RAN platform for software-defined, AI-native 5G and 6G networks.

It integrates NVIDIA RTX PRO™ 4500 Blackwell Server Edition GPU, Grace CPUs, and embedded NVIDIA ConnectX®-8 networking to support time-sync and RAN Layer 1–3 workloads. Optimized for serviceability, reliability, and telco deployments, EGN77C-2U enables operators to evolve from traditional RAN to a scalable, software-defined AI-RAN architecture.

QuantaEdge EGN77C-2U



Form Factor	CPU Number	Memory Number
2U	1 Per node	16

- NVIDIA ARC MGX platform for AI-RAN
- Short-depth, carrier-grade server for the telco edge scenario
- 2U2N form factor optimized for high availability
- Networking and SyncE powered by NVIDIA ConnectX®-8 SuperNIC™
- Accelerated by NVIDIA RTX PRO™ 4500 Blackwell Server Edition

Processor	(1) NVIDIA Grace™ CPU per node
Memory	(16) LPDDR5X chip-down 256GB per node
Networking	(1) Embedded ConnectX-8 for (2) QSFP112 (2) Embedded ConnectX-8 for (16) SFP28
Storage	(2) 2280 M.2 for OS boot and data storage
Expansion Slots	(2) FHFL single-slot GPU per node
Form Factor	2U2N NVIDIA MGX server
Dimension	(W) 438 x (H) 87 x (D) 420mm

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 Utilizing HBM3E GPU memory and LPDDR5X CPU memory

QuantaGrid S74G-2U (NVIDIA MGX™ Architecture)

QCT systems accelerated by the NVIDIA MGX™ architecture such as the QuantaGrid S74G-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*

- Accelerated 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

NVIDIA HGX™ Rubin NVL8



2.7x **AI Inference Performance** (400 PetaFLOPS NVFP4)

2.6x **AI Training Performance** (280 PetaFLOPS NVFP4)

2.8x **HBM4 Bandwidth** (176TB/s HBM4)

2x **NVIDIA NVLink™ Bandwidth** (28.8TB/s)

(Compared with NVIDIA Blackwell platform)

	HGX Rubin NVL8 Baseboard	HGX B300 Baseboard
Total GPU Memory Capacity	2.3TB HBM4	2.1TB HBM3E
Total GPU Memory Bandwidth	176TB/s	62TB/s
Total NVLink Bandwidth	28.8TB/s NVLink 6	14.4TB/s NVLink 5
FP4 Tensor Core	400 PFLOPS (NVFP4 Inference) 280 PFLOPS (NVFP4 Training)	144 PFLOPS*
FP8/FP6 Tensor Core	140 PFLOPS	72 PFLOPS*
Total GPU Power (TGP)	2300W (MaxP) or 1800W (MaxQ)	1100W

*Note: *Sparse*

QuantaGrid D76T-2U & QuantaGrid D76W-2U



Form Factor	CPU Number	Memory Number
2U	2	32

#Agentic AI #AI Reasoning
#AI Factories for Enterprise

- Supports NVIDIA Rubin GPUs, delivering over 2× AI performance to power the most demanding agentic AI workloads.
- Next-gen networking with NVIDIA® BlueField®-4 DPUs and NVIDIA ConnectX®-9 SuperNIC™ cards, enabling up to 800 Gb/s InfiniBand or Ethernet for ultra-fast north-south and east-west data movement across AI clusters.
- A compact, fully liquid-cooled 2U busbar system based on NVIDIA reference designs; can be scaled to a L11 solution fully compatible with the NVIDIA MGX rack.

	QuantaGrid D76T-2U	QuantaGrid D76W-2U
Processor	(2) Intel® Xeon® 6 processors, up to 350W TDP	(2) AMD EPYC™ 9006 series Server CPUs
Memory	(32) DDR5 RDIMM up to 6,400 MTs, (16) MRDIMM up to 8,000 MTs	Supports industry-standard DDR5 RDIMM and MRDIMMs based memory
Networking	(8) 800Gb/s OSFP ports via NVIDIA ConnectX®-9 SuperNIC™ cards (1) Dedicated LAN port (RJ45) for BMC management (1) 1GbE LAN port (RJ45)	
Accelerator	(8) NVIDIA Rubin SXM8 GPUs	
Storage	(8) Hot-swappable E1.S SSDs (2) PCIe M.2 2280 SSDs	(8) Hot-swappable PCIe 6.0 E1.S SSDs (2) PCIe M.2 2280 SSDs
Expansion Slot	(1) FHFL SW PCIe 6.0* x16 slot for NVIDIA® BlueField®-4	
Cooling	Fully liquid cooled	
Power	54V DC busbar	
Dimensions	(W) 448 x (H) 87 x (D) 800mm	

9x
QuantaGrid D76T-2U or QuantaGrid D76W-2U



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 SuperNIC™ cards

(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	CPU 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, utilizing NVIDIA ConnectX®-8 SuperNIC™ cards 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
Networking	(1) Dedicated LAN port (RJ45) for BMC management (1) 1GbE LAN port (RJ45) (8) OSFP ports serving (8) single-port NVIDIA ConnectX®-8 SuperNIC™ cards
Accelerator	(8) NVIDIA Blackwell Ultra GPUs
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 (1) OCP 3.0 SFF PCIe 5.0 x8 slots
Dimensions	(W) 447 x (H) 441.75 x (D) 800mm

QuantaGrid D75L-3U



Maximum GPU Density in AI Clusters

Form Factor	CPU Number	Memory Number
3U	2	32

#AI reasoning #Agentic AI #HPC #Video Inferencing

- Compact 3U chassis designed for ultra-dense, high-speed GPU clusters
- Fully compatible with NVIDIA MGX™ 1.1 rack standards to be seamlessly integrated into standard data center rack layouts
- NVIDIA HGX™ B300 8-GPU with NVIDIA NVLink™ 5 (1.8TB/s), 2.3TB of HBM3e GPU memory per system
- 8x NVIDIA ConnectX®-8 SuperNIC™ cards, up to 800Gb/s, up to three NICs for North/South traffic

Processor	(2) Intel® Xeon® 6 processors, up to 350W
Networking	(1) Dedicated LAN port (RJ45) for BMC management (1) 1GbE LAN port (RJ45) (8) OSFP ports serving (8) single-port NVIDIA ConnectX®-8 SuperNIC™ cards
Accelerator	(8) NVIDIA Blackwell Ultra GPUs
Storage	(8) hot-swappable E1.S SSDs (2) PCIe M.2 22110 SSDs
Expansion Slot	(3) FHHL SW PCIe 5.0 x16 slots
Dimensions	(W) 438 x (H) 131.7 x (D) 864.5mm

NVIDIA HGX™ Architecture Systems by QCT (Hopper Platform)

- Support the NVIDIA HGX™ H200/H100 8-GPU
- Modular motherboard tray design supporting Intel 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
- 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 NVIDIA ConnectX®-7 SuperNIC™ cards 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
Networking	(1) Dedicated 1GbE management port
Accelerator	(8) NVIDIA H200 SXM5 GPUs
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

NVIDIA Scalable and Lightweight AI System



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 H200 NVL GPU, NVIDIA L40S GPU, NVIDIA A100 GPU, NVIDIA A30 GPU, NVIDIA L4 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
Memory	Up to 8TB memory capacity with (32) DDR5 RDIMM or (16) MCRDIMM
Accelerator	NVIDIA H100 GPU, NVIDIA L40S GPU, NVIDIA L4 GPU
Storage	**SKU 1:** (12) hot-swappable 3.5" SATA/SAS HDDs **SKU 2:** (4) hot-swappable 3.5" SATA/SAS HDDs + (8) 2.5" hot-swappable NVMe SSDs **SKU 3:** (24) hot-swappable 2.5" NVMe/ SATA/ SAS SSDs
Networking	(1) Dedicated 1GbE management port
Expansion Slot	**Option1: [General SKU]** (4) FHHL PCIe 5.0 x8 slots+ (2) HHHL PCIe 5.0 x16 slots + (1) HHHL PCIe 5.0 x8 slot + (2) OCP 3.0 slots **Option2: [SW GPU SKU]** (4) FHHL PCIe 5.0 x8 slots + (2) HHHL PCIe 5.0 x16 slots + (1) HHHL PCIe 5.0 x8 slot + (2) OCP 3.0 slots **Option3: [DW GPU SKU]** (3) FHFL PCIe 5.0 x16 slots + (2) HHHL PCIe 5.0 x16 slots + (1) HHHL PCIe 5.0 x8 slot + (2) OCP 3.0 slots
Form Factor	2U Rackmount
Dimensions	(W) 447 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 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 processors, up to 350W TDP
Memory	Up to 8TB memory capacity with (32) DDR5 RDIMM or (16) MCRDIMM
Accelerator	NVIDIA L4 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
Networking	(1) Dedicated 1GbE management port
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
Form Factor	1U Rackmount
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 rich IO 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
Memory	Up to 4T memory capacity with (16) DDR5 RDIMM or (8) MCRDIMM
Networking	(1) Dedicated 1GbE management port
Accelerator	NVIDIA L4 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
Memory	**S44NL-1U:** Up to 6TB memory capacity with (24) DDR5 RDIMM up to 5200 MHz **D44N-1U:** Up to 6TB memory capacity with (24) DDR5 RDIMM up to 6400 MHz
Networking	(1) Dedicated 1GbE management port
Accelerator	NVIDIA L4 GPU
Storage	**SKU - #1** (12) hot-swappable 2.5" NVMe/SATA/SAS SSDs **SKU - #2** (16) hot-swappable E1.S SSDs
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
Form Factor	1U Rackmount
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

- 2U2N high density front access multi-node server
- Powered by (1) Intel® Xeon® 6 CPU per node
- Supports SW or DW PCIe GPUs
- Up to (3) PCIe 5.0 expansion slots per node for diverse device support

*Per node

Processor	(1) Intel® Xeon® 6 processor
Networking	(1) OCP 3.0 SFF slot
Accelerator	NVIDIA L40S GPU, NVIDIA L4 GPU
Storage	(3) hot-swappable 2.5" NVMe SSDs
Expansion Slot	**Option 1** (1) FHFL SW PCIe 5.0 x16 slot (1) FHFL SW PCIe 5.0 x8 slot (1) FHHL SW PCIe 5.0 x16 slot **Option 2** (1) FHFL DW PCIe 5.0 x16 slot (1) FHHL SW PCIe 5.0 x16 slot
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

- 2U4N high density front access multi-node server
- Powered by (1) Intel® Xeon 6® CPU per node
- Supports (1) SW PCIe GPU per node

*Per node

Processor	(1) Intel® Xeon® 6 processor
Networking	(1) OCP 3.0 SFF slot
Accelerator	NVIDIA L4 GPU
Storage	(2) hot-swappable E1.S SSDs
Expansion Slot	(1) HHHL PCIe Gen5 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 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	**Expansion SKU** (16) 25GbE SFP28 (LoM) **OCP3.0 D-SFF SKU** (24) 25GbE SFP28 (LoM & Intel® Ethernet Network Adapter E830-XXVDA8F for OCP 3.0)
Accelerator	NVIDIA L4 GPU
Storage	(2) SATA/NVMe M.2 22110/2280 SSDs
Expansion Slot	**Expansion SKU** (1) FHHL PCIe 5.0 slot **OCP3.0 D-SFF SKU** (1) OCP 3.0 PCIe 5.0 slot
Dimensions	(W) 447.8 x (H) 42.8 x (D) 300.65mm (ear to rear wall)

QuantaEdge EGX77B-1U

Carrier-Grade Flexible 300mm 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
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 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) 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
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

© 2026 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 Vera Rubin, 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 Oct. 2026