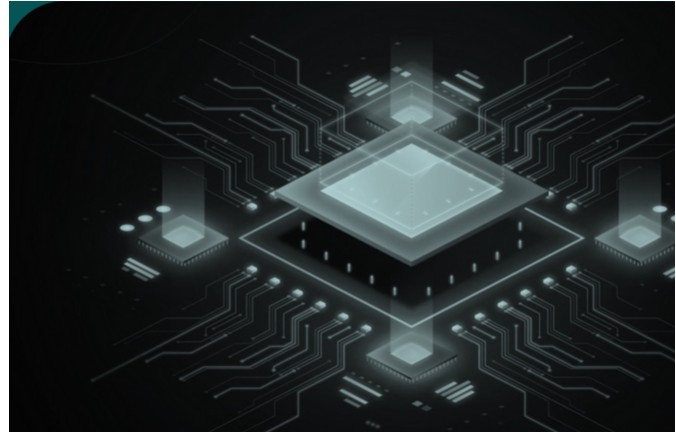


NVIDIA GB200 NVL72

Liquid-cooled exascale compute in a rack with 72 NVIDIA Blackwell GPUs



Scalable compute unit built for trillion-parameter AI models

- **72 NVIDIA Blackwell GPUs:** acting as one GPU with a massive pool of HBM3e memory to deliver the most efficient exascale computing in a rack
- **Pioneers in liquid cooling:** total liquid-cooling solution with up to 40% reduction in electricity cost for the data centre
- **Unmatched manufacturing scale:** the largest liquid-cooling rack-level manufacturing capacity ensures timely and high-quality deployment of the GB200 NVL72, supported by production facilities in the USA, Europe and Asia
- **Comprehensive service offering:** from proof of concept to full-scale deployment, one-stop provision of all necessary parts, networking solutions and on-site installation services
- **Advanced networking ready:** NVIDIA BlueField®-3 SuperNIC, Spectrum™-X, Quantum-2 and next-generation 800Gb/s networking platforms

An exascale supercomputer in a rack

The industry's transition to liquid-cooled data centres with NVIDIA Blackwell delivers a new paradigm of energy efficiency for the rapidly heightened energy demand of AI infrastructure. Built on extensive experience deploying large-scale direct-to-chip (DLC) liquid-cooled AI systems, the leading liquid-cooling technology powers the NVIDIA GB200 NVL72, exascale computing in a single rack, providing up to 25 times more energy efficiency than the previous generation.

Powered by an end-to-end liquid-cooling solution

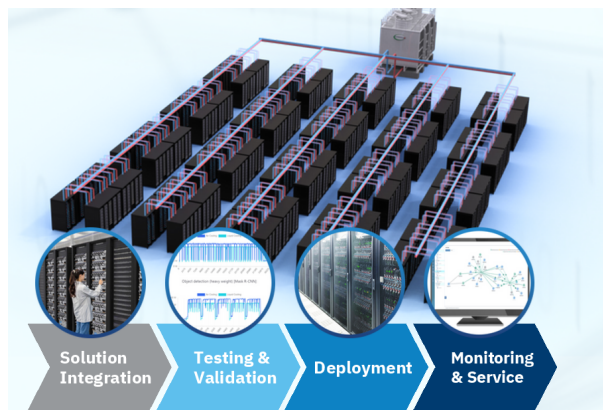
The NVIDIA GB200 NVL72 SuperCluster features the new advanced in-rack or in-row cooling distribution unit (CDU) and custom cold plates designed for the compute trays housing the NVIDIA GB200 Grace™ Blackwell Superchips. The NVIDIA GB200 NVL72 delivers exascale computing capabilities in a single rack with fully integrated liquid cooling. It incorporates 72 NVIDIA Blackwell GPUs and 36 Grace CPUs interconnected by NVIDIA's largest NVLink™ network to date. The NVLink Switch System facilitates 130 terabytes per second (TB/s) of total GPU communications with low latency, enhancing performance for AI and high-performance computing (HPC) workloads.

End-to-end on-site deployment services

From proof-of-concept (PoC) to full-scale deployment, SuperServer.ai provides all necessary parts, liquid cooling, networking solutions, management software and on-site installation services. The comprehensive in-house liquid-cooling ecosystem encompasses custom-designed cold plates optimised for various GPUs, CPUs and memory modules, along with multiple coolant distribution unit form factors and capacities, manifolds, hoses, connectors, cooling towers, and monitoring and management software. This end-to-end solution seamlessly integrates into rack-level configurations, significantly boosting system efficiency, mitigating thermal throttling, and reducing both the total cost of ownership (TCO) and environmental impact of data centre operations for the era of AI.

Liquid-cooled data centre management

A comprehensive data centre management platform provides powerful tools to monitor vital information on liquid-cooled systems and racks, coolant distribution units and cooling towers, including pressure, humidity, pump and valve conditions and more. The liquid-cooling consult module optimises the operational cost and manages the integrity of liquid-cooled data centres.



Rack scale design close-up

Management networking

- In-band management switch
- Out-of-band management switch

10 compute trays

- 4x NVIDIA Blackwell GPUs per tray
- 2x NVIDIA Grace CPUs per tray

Compute interconnect

- 9x NVLink switches
- 72 GPUs and 36 CPUs interconnected at 1.8TB/s

8 compute trays

- 4x NVIDIA Blackwell GPUs per tray
- 2x NVIDIA Grace CPUs per tray

Liquid-cooling options

- 250kW capacity coolant distribution unit (CDU) with redundant PSU and dual hot-swap pumps
- 240kW or 180kW capacity liquid-to-air solution (no facility water required)



72-GPU scalable unit

SRS-GB200-NVL72-M1

GPUs	72x NVIDIA Blackwell B200 GPUs
CPUs	36x NVIDIA 72-core Grace Arm Neoverse V2
Compute trays	18x 1U ARS-121GL-NBO
NVLink switch trays	9x NVLink switch, 4 ports per compute tray connecting 72 GPUs to provide 1.8TB/s GPU-to-GPU interconnect
Power shelves	8x 1U 33kW (6x 5.5kW PSUs), total power 132kW
Rack dimensions (mm)	W 600 x D 1068 x H 2236
Liquid-cooling options	1x in-rack 250kW capacity CDU with redundant PSU and dual hot-swap pumps 1.3MW capacity in-row CDU 180kW/240kW capacity liquid-to-air solutions for facilities without cooling tower and water supply

Subject to change

Compute tray

ARS-121GL-NBO

Overview	1U liquid-cooled system with 2x NVIDIA GB200 Grace Blackwell Superchips
CPU and GPU	2x 72-core NVIDIA Grace Arm Neoverse V2 CPU 4x NVIDIA Blackwell B200 per Superchip
GPU memory	Up to 384GB HBM3e per Superchip (768GB per tray)
CPU memory	Up to 480GB LPDDR5X per Superchip (960GB per tray)
Networking	4x NVIDIA NVLink switch ports
Storage	8x E1.S PCIe 5.0 drives
Power supply	Shared power through 4+4 rack power shelves

Subject to change

SuperServer.ai delivers this system to the UK and Europe in 4 to 5 weeks, against an industry norm of two months or more. Every order starts with a personal consultation to align configuration, lead time and price with your workload.