

HPE ProLiant Compute DL384 Gen12



What's new

- The first HPE ProLiant rack mount server with the NVIDIA GH200 Grace Hopper™ Superchip.
- Support for dual NVIDIA GH200 Grace Hopper Superchips with NVIDIA GH200 NVL2.
- The GH200 NVL2 fuses two Grace CPUs and two Hopper GPUs in an innovative architecture to deliver 8 petaflops of AI performance into a single node.
- The GH200 NVL2 offers up to 3.5X more GPU memory capacity and 3X more bandwidth than the NVIDIA H100 Tensor Core GPU in a single server for compute- and memory-intensive workloads. [1]
- Train, tune, and deploy generative AI models with confidence.

Overview

Does your enterprise need a server to dramatically enhance performance for AI inferencing workloads in the era of accelerated computing and generative AI with the modern NVIDIA GH200 Grace Hopper™ Superchip?

The HPE ProLiant Compute DL384 Gen12 provides outstanding coherent memory and memory bandwidth with up to two superchips per server with NVIDIA GH200 NVL2, allowing teams to run larger models faster and more cost-effectively. By partnering with NVIDIA, Hewlett Packard Enterprise continues to innovate with impressive infrastructure enhancements to help customers unlock scaled-out and accelerated computing for generative AI. With superchip performance in your AI factory, HPE ProLiant Compute DL384 Gen12 delivers superb performance per GPU in the HPE ProLiant portfolio. Simplify your approach to scaling with data demands and integrate accelerated compute to reliably prepare your data center for mixed workloads and retrieval-augmented generation (RAG) in the modern hybrid reality.



Technical specifications	HPE ProLiant Compute DL384 Gen12
Processor family	NVIDIA Grace CPU and Hopper GPU
Drive supported	Up to 8 EDSFF NVMe Gen5 drives
Infrastructure management	HPE iLO Standard with intelligent provisioning (embedded), HPE iLO Advanced
Power supply type	Up to 4 HPE 1800W-2200W Flex Slot Titanium Hot Plug Power Supply Kits For a single GH200: 2+0, 2+1, or 2+2 redundancy. Minimum of two power supplies. For a dual GH200 with NVL2: 3+1 redundancy, and 4 power supplies required, no other choices supported.
Expansion slots	Up to four high speed PCIe Gen5 x16 devices, Up to four PCIe Full Height Half Length (FHHL) slots, or Up to two PCIe FHHL and two OCP 3.0
Storage controller	HW RAID controllers are not supported this generation for either the M.2 or EDSFF NVMe drives
System fan features	6 hot plug redundant fans
Form factor	2U Air cooled 19-inch rack server with 31.75 in (806 mm) depth

