




Achieve next-level performance for enterprise AI

HPE ProLiant Compute DL384 Gen12 with NVIDIA GH200 NVL2

Get ready for a transformative AI journey

AI is evolving—fast. To get the most from AI, you need to embrace enterprise AI—where AI is operationalized across your organization and AI technologies are developed, deployed, and managed at scale.

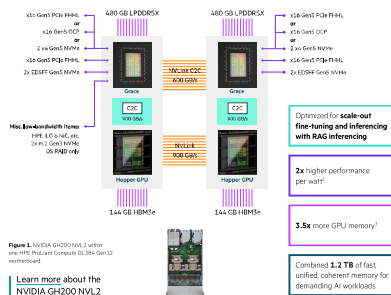


You need a solution that enables you to deploy AI at scale for any destination.

To ensure successful enterprise AI, you must prepare your data center infrastructure for this technology shift.

Introducing HPE ProLiant Compute DL384 Gen12 with NVIDIA® GH200 NVL2, part of the NVIDIA AI Computing by HPE portfolio

Built from the ground up to achieve next-level performance for mixed or memory-intensive workloads such as AI fine-tuning and inferring with RAG,¹ this innovative AI-optimized solution has two NVLink connected NVIDIA GH200 processors on one HPE ProLiant Compute DL384 Gen12 Server, enabling the system to deliver:

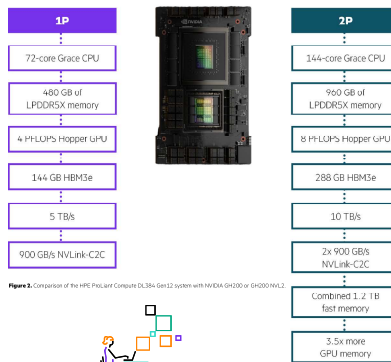
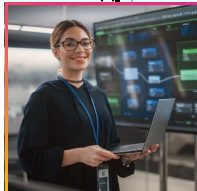


You can use HPE ProLiant Compute DL384 Gen12 with NVIDIA GH200 NVL2 with combined 1.2 TB of fast, unified, coherent memory to:

- Handle scale-out fine-tuning and inferring with RAG
- Increase performance for other (mixed) workloads—such as job scheduling, large-scale simulation, weather forecasting, biomedical, and more—across systems and GPUs
- Protect your data by keeping your IT resources on-premises

1P/2P options

Match the number of NVIDIA GH200 NVL2—single or dual—to your mixed or AI workloads, business needs, and budget.

With this innovative, accelerated compute solution, you can achieve next-level performance for mixed or memory-intensive workloads, enabling you to:

- Create a versatile, scale-out, accelerated computing platform for tackling the challenges of AI model fine-tuning and inferring with RAG
- Accelerate AI success with HPE Private Cloud AI, the industry's first full-stack, runkey private cloud for AI, part of the NVIDIA AI Computing by HPE portfolio.

¹ Retrieval-augmented generation (RAG).
² NVIDIA Grace CPU Supporting NVIDIA accelerated AI. 2024.
³ NVIDIA GH200 NVL2 Gen12 AI-optimized Server Platform for Enterprise AI Computing and Generative AI. NVIDIA Newsroom. August 2023.