Hewlett Packard Enterprise | NVIDIA

Solution brief

Next-level performance for enterprise AI

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

Accelerating the shift to generative AI

Enterprises are increasingly leveraging artificial intelligence (AI), particularly large language models (LLMs), to power new generative AI (GenAI) applications such as text generation, language translation, coding, visual content, drug discovery, and many more.

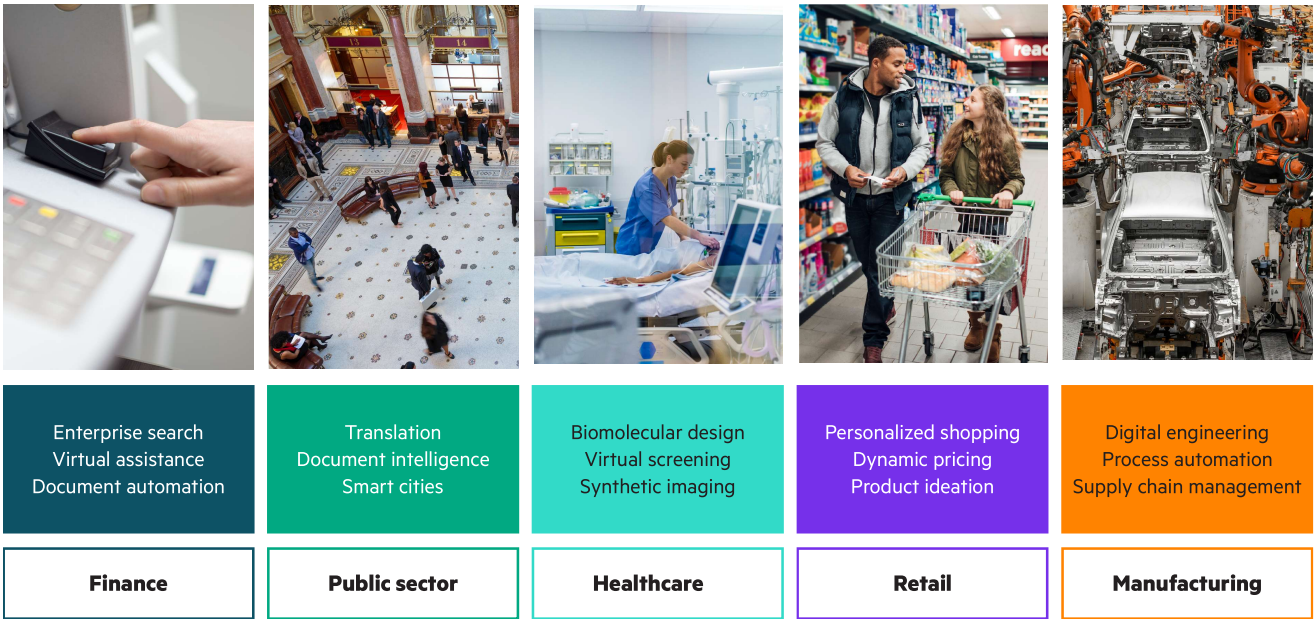

Figure 1. Image illustrating how GenAI can affect every industry, including finance, public sector, healthcare, retail, and manufacturing.

While GenAI promises to deliver game-changing benefits in terms of efficiency, productivity, and innovation, the road to deploying GenAI contains significant hurdles:

- Ensuring data privacy during model tuning, and while augmenting the model with enterprise data for inference
- Scaling computational resources
- Securing enterprise data
- Maintaining control over the AI model's output

As the scale of data increases, an AI model's ability to learn and generate more accurate and diverse responses can improve. More data, however, places greater demands on computational resources. To meet the ever-growing demands for resources, traditional data centers need a simpler approach to scaling and integrating accelerated compute in the data center.

In today's hybrid reality, where an increasing number of processes are AI-supported and data-driven, organizations need to embrace enterprise AI—where AI is operationalized across the organization, and AI technologies are developed, deployed, and managed at scale. A critical success factor of enterprise AI is ensuring the data center infrastructure is prepared for this technology shift.

Generative AI is transforming business

Enterprises that adopt next-generation AI like LLMs and generative AI are 2.6x more likely to increase revenue by 10% or more, but must invest in their AI infrastructure to fully reap the benefits.[1]

Understanding generative AI and AI inferencing vs. tuning and large language models

AI inferencing is the process of running live data through a trained and tuned AI large language model (LLM) to make a prediction or solve a task. "A large language model is a type of AI algorithm that uses deep learning techniques and massively large data sets to understand, summarize, generate and predict new content. The term generative AI is also closely connected with LLMs, which are, in fact, a type of generative AI that has been specifically architected to help generate text-based content."[2]

[1] Accenture Research. Breakthrough Innovation: Is your organization equipped for breakthrough innovation? WEF 2023

[2] techtarget.com/whatis/definition/large-language-model-LLM

Deploy at scale, using rack-based solutions for any AI destination

To help enterprises unlock scale-out accelerated computing for GenAI, HPE and NVIDIA® deliver HPE ProLiant Compute DL384 Gen12 with NVIDIA GH200 NVL2, part of the NVIDIA AI Computing by HPE portfolio. This next-generation 2P server provides next-level performance for enterprise AI—enabling a new era of AI. With this versatile system, enterprises can:

- Optimize for scale-out fine-tuning and AI inferencing with Retrieval Augmented Generation (RAG)
- Maximize data center utilization by providing next-level performance
- Use NVIDIA GH200 NVL2 for 1.2 TB of fast, unified, coherent memory for compute- and memory-intensive workloads
- Benefit from 3.5x capacity and 3x bandwidth[3]
- Get 2x higher inference performance[4]

AI computing from NVIDIA and HPE

To tackle the largest problems in the new era of AI, HPE ProLiant Compute DL384 Gen12 with NVIDIA GH200 NVL2 with combined 1.2 TB fast, unified, coherent memory provides the advanced features and functions you need to support enterprise AI.

Enhanced flexibility. HPE ProLiant Compute DL384 Gen12 with NVIDIA GH200 NVL2 leverages a flexible compute architecture designed to achieve next-level performance for mixed or memory-intensive AI workloads such as AI inferencing and fine-tuning a Retrieval Augmented Generation (RAG) model.

Accelerated computing. With up to 1.2 TB of fast, unified, coherent memory, this versatile scale-out accelerated computing platform is designed to power large language models for inferencing, with up to 2x higher inferencing performance, compared to NVIDIA H100.

Expert customization, service, and support. HPE AI experts will work with you to build and deploy a unique solution that precisely matches your intended purposes, as well as integrate and enhance the ecosystem offerings.

Simplified management. HPE GreenLake Flex Solutions offers you a flexible and scalable approach to managing your IT infrastructure, including your AI environment. HPE GreenLake Flex Solutions combine hardware, software, and services into a single pay-per-use* solution—providing you with the agility and cost savings of a cloud-based model, while keeping your IT resources on-premises.

[3], [4] Compared to NVIDIA H100 accelerators

* May be subject to minimums or reserve capacity may apply

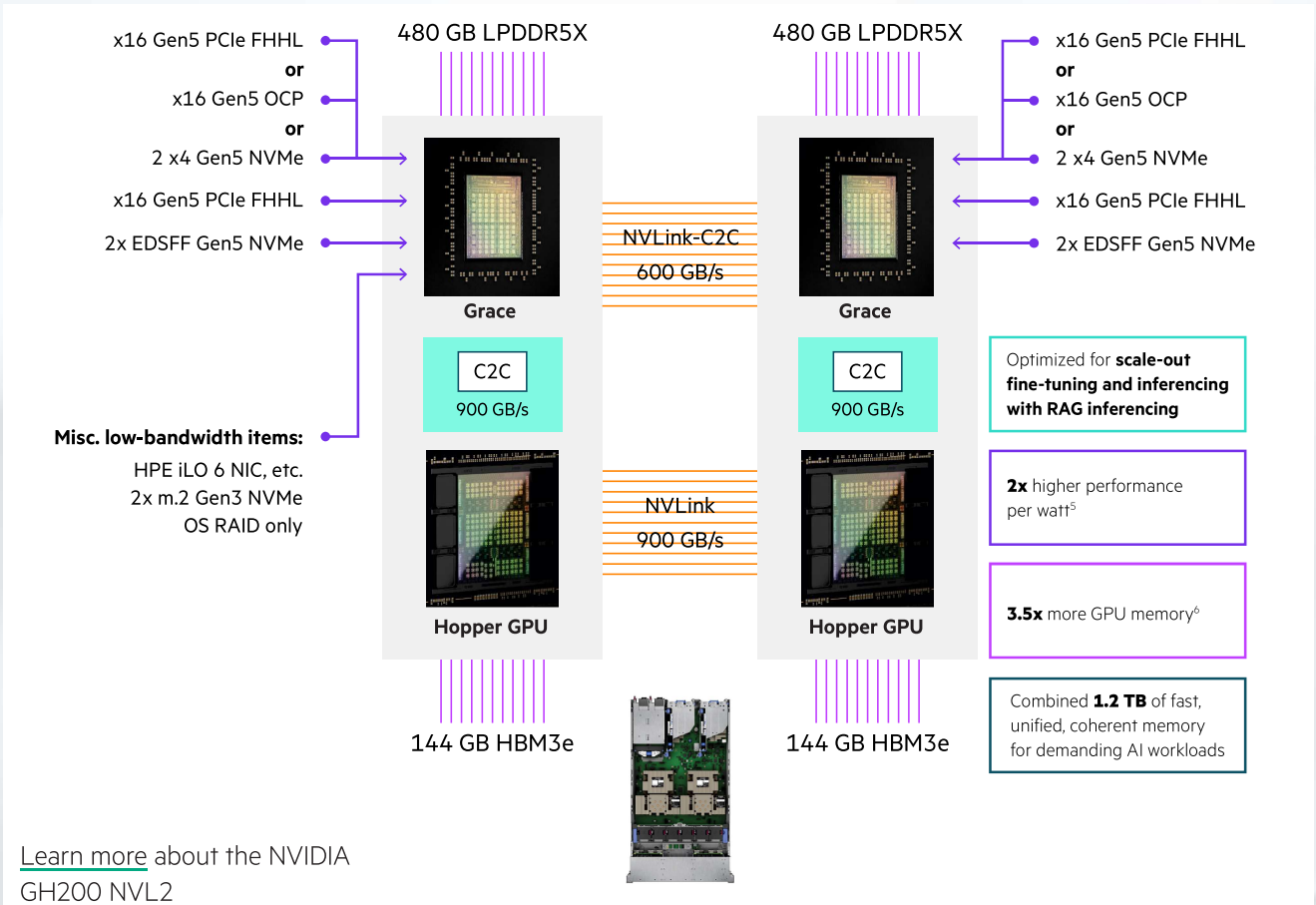

Figure 2. NVIDIA GH200 NVL2 within one HPE ProLiant Compute DL384 Gen12 motherboard

[5] "NVIDIA Grace CPU Superchip," NVIDIA, accessed May 2024.

[6] "NVIDIA Unveils Next-Generation GH200 Grace Hopper Superchip Platform for Era of Accelerated Computing and Generative AI," NVIDIA Newsroom, August 2023.

Choose the right number of processors

Match the number of processors to your workloads, business needs, and budget

To meet a wide range of business needs, HPE ProLiant Compute DL384 Gen12 is available with NVIDIA GH200 or GH200 NVL2. The following image illustrates the differences between the 1P and 2P options.

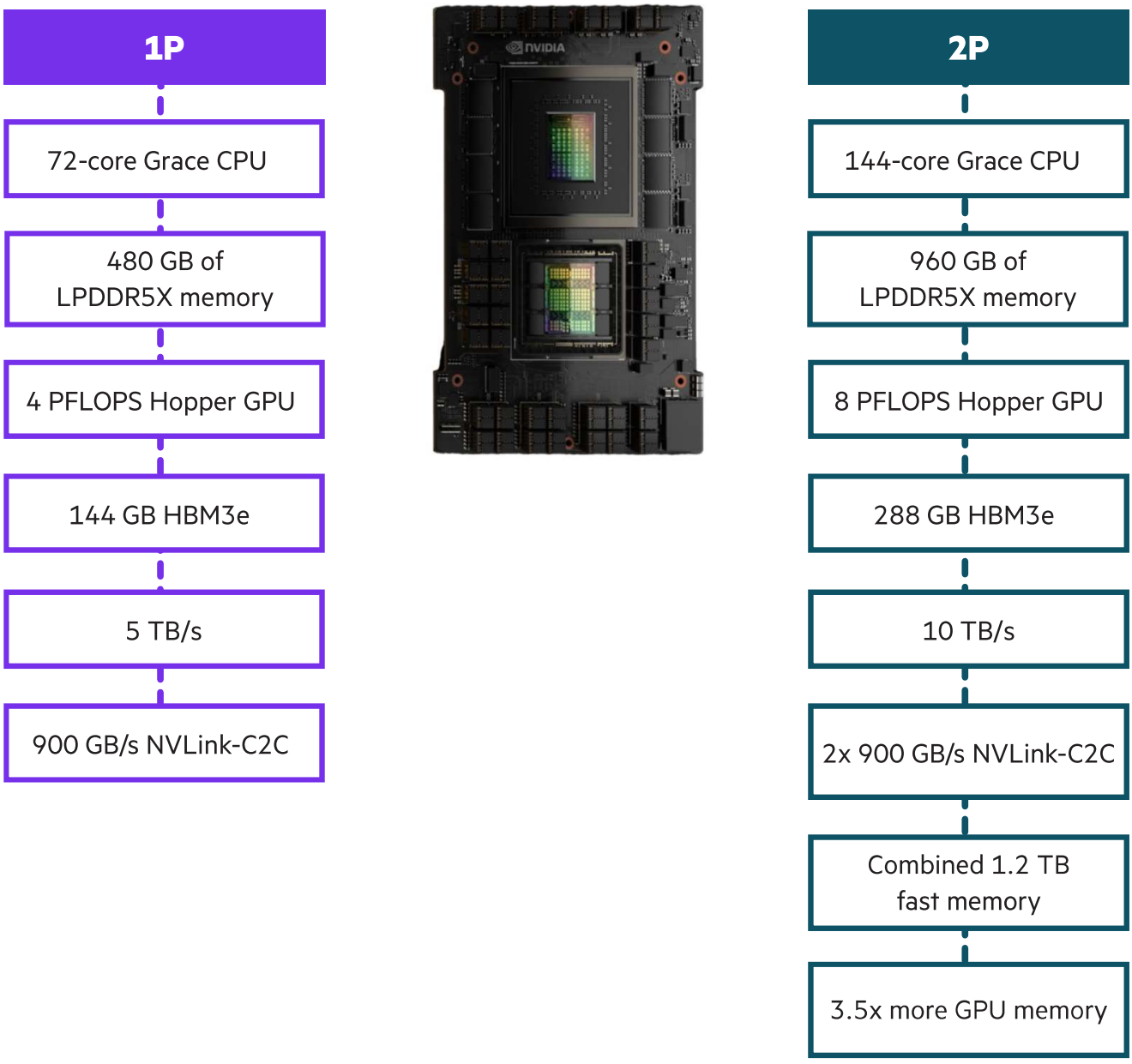

Figure 3. Comparison of the HPE ProLiant Compute DL384 Gen12 system with NVIDIA GH200 or GH200 NVL2.

Fast-track AI production with HPE Private Cloud AI

Accelerate your path to production AI with a scalable, tested, AI-optimized private cloud. HPE Private Cloud AI, part of the NVIDIA AI Computing by HPE portfolio, is an industry-first turnkey private cloud for enterprise AI, codeveloped with NVIDIA. It gives AI and IT teams powerful tools to experiment and operationalize AI while keeping your data private and secure while leveraging market adopted NVIDIA, HPE, and open-source software tools.

Delivered on HPE GreenLake cloud, HPE Private Cloud AI is built on validated designs powered by AI optimized compute, storage, and networking from HPE and NVIDIA. Start as small as a single small-model inferencing pilot and scale to multiple use cases, higher throughputs, RAG or LLM fine-tuning in one solution. Simply expand your infrastructure without new software, integration work, and with less reliance on specialized skills.

HPE Private Cloud AI delivers what organizations love about the cloud experience—self-service, modern development tools, rapid scale and subscription economics—in your own private environment. You can start small and seamlessly scale your tech and investment as your use cases evolve. And with expert services, we can help you pinpoint where to get started.

Capitalize on generative AI

Contact your Hewlett Packard Enterprise representative today to find out how your organization can benefit from **next-level** performance for enterprise AI. Learn how HPE ProLiant Compute DL384 Gen12 with NVIDIA GH200 NVL2 can help you:

- Boost performance per GPU with 1.2 TB coherent memory
- Increase performance for AI and other workloads, such as job scheduling, across systems and GPUs
- Optimize bandwidth from CPU to GPU to handle demanding AI workloads such as large-scale simulation, weather forecasting, and more
- Increase flexibility to tackle the challenges of AI, model fine tuning and inference with RAG
- Create a versatile, scale-out, accelerated computing platform to power the latest LLMs
- Work with HPE AI experts to build and deploy a custom-tailored AI solution
- Simplify management of your IT infrastructure by choosing flexible, scalable HPE GreenLake Flex Solutions

Learn more at

HPE.com/ProLiant/DL384-gen12

Chat now (sales)

Hewlett Packard Enterprise

a50010636ENW