Hewlett Packard Enterprise | NVIDIA | intel

Solution brief

Ultra-scalable GPU acceleration for enterprise AI

HPE ProLiant Compute DL380a Gen12 with NVIDIA H200 NVL Tensor Core GPUs, part of the NVIDIA AI Computing by HPE portfolio

Embracing enterprise scale for AI

The emergence of artificial intelligence (AI) is expected to fundamentally change every industry. Enterprise organizations need specialized environments where AI applications are developed, deployed, and managed at scale—and represents a paradigm shift in how these technologies are conceptualized, developed, and operationalized within organizations. By embracing an enterprise approach to AI development and management, organizations can unlock new opportunities, but deploying AI at scale involves several considerations to ensure optimal business value.

Scaling computational resources to match your AI lifecycle stage

Logically, computational requirements can vary significantly across different stages of the AI lifecycle, including model training, tuning, and inferencing. Model training involves building an AI model from scratch, which requires substantial data and computational resources. Fine-tuning, on the other hand requires fewer resources because it adapts existing models to specific tasks and demands. Finally, AI inferencing is where AI models go to work and involves leveraging models to provide valuable insights. Inferencing requires less computational resources but organizations looking to deploy need to factor in model size, response time and concurrent users to ensure optimal business performance.

Enterprise can enhance the performance of Large Language Models (LLMs) with Retrieval Augmented Generation (RAG) by combining an information retrieval component with text generation capabilities. It fetches relevant data, providing additional context for better input understanding and more accurate response generation. The separation of retrieval and generation components allows RAG to scale effectively for large datasets and complex queries.

NVIDIA H200 NVL

In the ever-evolving landscape of AI, businesses rely on large language models (LLMs) to address a diverse range of inference needs. An AI inference accelerator must deliver the highest throughput at the lowest TCO when deployed at scale. Fueling the acceleration of generative AI and LLMs while advancing scientific computing for HPC workloads, the NVIDIA H200 NVL:

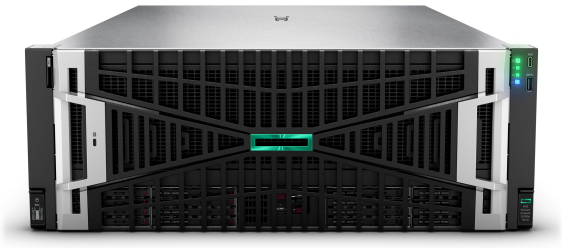
- Boosts inference speeds up to 2x (when handling LLMs like Llama) with NVIDIA H200 Tensor Core using 4-way NVLink Bridges compared to H100 NVL using 2-way NVLink Bridges
- Is the first GPU with HBM3E,[1, 2] thus providing faster memory with 4.8 TB/sec of memory bandwidth
- Delivers an unprecedented performance leap with up to 18x performance versus the NVIDIA A100

HPE ProLiant Compute DL380a Gen12—Ultra-scalable GPU acceleration to achieve best fine-tuning and inference performance

For organizations looking to provide the best performance across fine-tuning and inference with RAG to enhance output of LLMs, the latest addition to the HPE ProLiant family of servers is ideal for harnessing the power of GenAI for enterprise.

The HPE ProLiant Compute DL380a Gen12, part of the NVIDIA AI Computing by HPE portfolio, is engineered with an ultra-scalable architecture to deliver next-gen AI performance for your enterprise needs. With up to two Intel® Xeon® 6 processors and up to eight NVIDIA® H200 NVL Tensor Core GPUs[3]—GPUs that utilize the world's most powerful GPU architecture for supercharging AI and high-performance computing (HPC) workloads— the HPE ProLiant Compute DL380a Gen12 delivers:

- **Unprecedented performance**, with 18x performance versus A100 and up to 2x inferencing speeds, compared to NVIDIA H100 NVL for inferencing
- **Enterprise-grade reliability**, with advanced power management that delivers six dedicated and redundant power supplies for the GPUs for more efficient and reliable performance
- **Industry-leading security innovation** to protect your infrastructure, workloads, and data



HPE ProLiant Compute DL380a Gen12

Improve core density and performance per watt while driving high throughput.

Intel Xeon 6 processors offer a new class of Efficient cores (E-cores) with high-core density, offering distinct advantages for cloud-scale workloads. Using these latest-generation processors from Intel® lowers your energy costs, drives sustainability, and improves your rack density, allowing you to get more from your data center infrastructure—all while adding capacity for new workloads. Built-in accelerators give an additional boost to targeted workloads for even greater performance and efficiency.

[1] "NVIDIA Supercharges Hopper, the World's Leading AI Computing Platform," NVIDIA press release, November 2023.

[2] HBM3E (High Bandwidth Memory 3E) is high-bandwidth memory designed to advance generative AI innovation.

[3] HPE will be time to market to support the compatible NVIDIA Blackwell and Rubin offerings.

Advanced management and monitoring

To simplify management of the HPE ProLiant Compute DL380a Gen12, advanced management and monitoring capabilities are included:

- **HPE iLO 6** allows you to securely view, configure, and update your servers from anywhere in the world.
- **HPE OneView** infrastructure management works within the data center to help automate and streamline IT operations across servers, storage, and networking.
- **HPE GreenLake for Compute Ops Management** provides an intuitive cloud operating experience to seamlessly monitor, manage, and gain visibility of your distributed compute environment, no matter where your compute infrastructure lives.

Fast-track AI production with HPE Private Cloud AI

Accelerate AI success with HPE Private Cloud AI, the industry's first full-stack, turnkey private cloud for AI, part of the NVIDIA AI Computing by HPE portfolio. It gives AI and IT teams powerful tools to experiment and operationalize AI while keeping your data private and secure and leverages 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.

Learn more at

HPE.com/ProLiant/DL380a-gen12

NVIDIA.com/en-us/data-center/H200/

Intel.com/content/www/us/en/products/details/processors/Xeon.html

Visit **HPE.com**

Chat now (sales)

a00138852ENW, Rev. 1

Hewlett Packard Enterprise