Hewlett Packard Enterprise

Solution brief

Ultra-scalable GPU acceleration for enterprise AI

HPE ProLiant Compute DL380a Gen12 with NVIDIA RTX PRO™ 6000 Blackwell Server Edition and H200 NVL Tensor Core GPUs

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 GPU support

Fueling the acceleration of generative AI, the HPE ProLiant Compute DL380a Gen12 supports the following two GPUs:

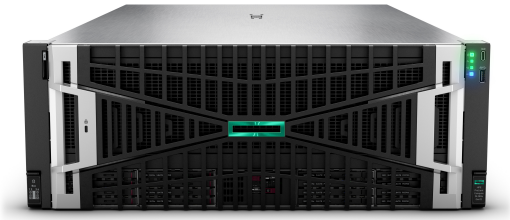
- **NVIDIA RTX PRO™ 6000**
  The NVIDIA RTX PRO™ 6000 Blackwell Server Edition is the ultimate data center GPU for AI and visual computing, delivering breakthrough acceleration for the most demanding enterprise workloads, from agentic and physical AI to scientific computing, graphics, and video applications. Optimized for workloads requiring the compute density and scale that deploying in the data center offers, it features a passively cooled thermal design and 96 GB of ultra-fast GDDR7 memory.
- **NVIDIA H200 NVL**
  The NVIDIA H200 Tensor Core GPU supercharges generative AI and high-performance computing (HPC) workloads with game-changing performance and memory capabilities. As the first GPU with HBM3e, the H200's larger and faster memory fuels the acceleration of generative AI and large language models (LLMs) while advancing scientific computing for HPC workloads.

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 ten double-wide GPUs—GPUs that utilize the world's most powerful GPU architecture for supercharging AI workloads—the HPE ProLiant Compute DL380a Gen12 delivers:

- **Unprecedented performance and efficiency**, with optional Direct Liquid Cooling to enhance energy efficiency and up to either 16 single-wide or 10 double-wide GPUs.
- **Enterprise-grade reliability**, with advanced power management that delivers six dedicated and redundant power supplies for the GPUs for more efficient and reliable performance.
- **Next-level security** with HPE iLO 7 multi-layer silicon root of trust protects servers from manufacturing to end-of-life and provides compliance readiness for future quantum-computing attacks.



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 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.

Advanced management and monitoring

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

- **HPE iLO 7** with silicon root of trust, a robust security foundation embedded within the server hardware, ensures that every layer of firmware and software loads securely and is verified from the moment the server is powered on. This results in providing an unbreakable chain of trust; stated differently, protecting servers against firmware attacks, unauthorized access, and tampering, thereby ensuring the highest level of data integrity and system reliability.
- **HPE OneView** works within the data center to help automate and streamline IT operations by providing a centralized interface to manage and monitor servers, storage, and networking devices. This ensures seamless integration and efficient management of diverse infrastructure environments.
- **HPE Compute Ops Management** ensures smooth enterprise operations with proactive and predictive automation from data center to edge, leveraging a single management solution powered by AI-driven insights. Enable operators to react quicker and gain greater control, from forecasting energy costs to managing a global server footprint. Boost productivity of IT staff by quickly pinpointing problem areas through dashboards, intelligent alerts, and global map view of all servers with status and activity.

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 AI Computing by HPE

Visit **HPE.com**

Chat now (sales)

© Copyright 2025 Hewlett Packard Enterprise Development LP. The information contained herein is subject to change without notice. The only warranties for Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

Intel, Intel Xeon, and the Intel logo are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries. NVIDIA RTX, NVIDIA, and the NVIDIA logo are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. All third-party marks are property of their respective owners.

a00138852ENW, Rev. 2

Hewlett Packard Enterprise