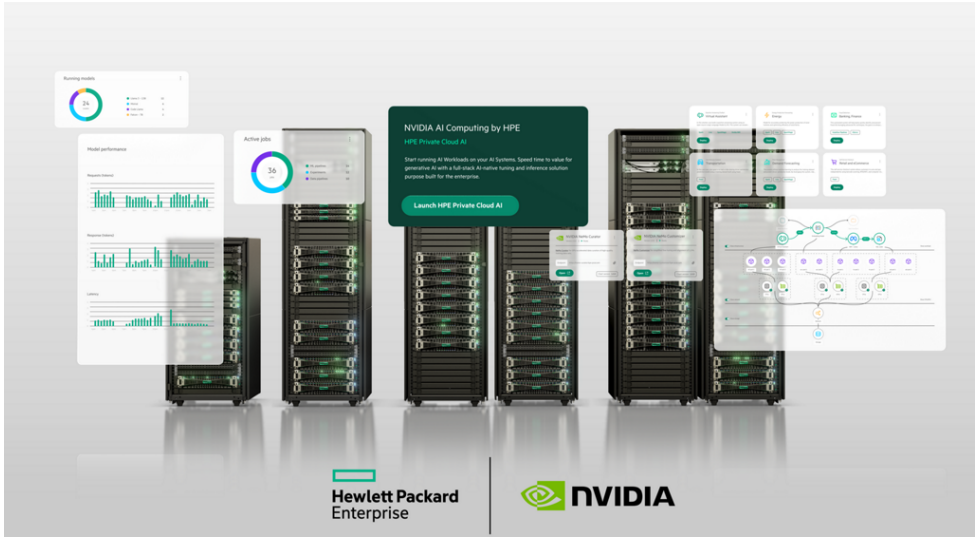
HPE Private Cloud AI



What's new

- HPE Private Cloud AI is now available with a lower cost developer system to accelerate the development and delivery of agentic AI across industries.
- Unified data access: Unify access to your entire data estate while ensuring privacy and security.
- Confidence and control: Comprehensive, controls protect your company's IP, sensitive data and customized models while maintaining high performance, reliability, and utilization.
- Flexible and evergreen cloud experience: Scale up while maintaining performance and support for the latest advancements in AI through an advanced cloud management console.
- Accelerate time to value with self-service

Overview

HPE Private Cloud AI is the first co-developed, turnkey solution to come from NVIDIA AI Computing by HPE—a new joint initiative to help enterprises unlock their AI ambitions. NVIDIA AI Computing by HPE brings together people, technology, and economics to accelerate AI deployments, protect against risks, and optimize AI costs long term. The solution is tailored to AI models and designed to scale easily with the growth and utilization of AI use cases. With this proven foundation, HPE Private Cloud AI accelerates data scientist productivity and helps overcome common challenges in operationalizing AI by delivering a flexible, pretested, AI-optimized private cloud. HPE Private Cloud AI provides a self-service cloud experience enabled by the HPE GreenLake cloud. Start small as a single AI pilot and evolve quickly for multiple use cases or higher throughput in one solution. Plus, you can deploy either on-prem, colocation, or in the cloud while maintaining control over financial risks.

access to NVIDIA AI microservices and other essential AI tools to speed developer productivity

Features

Instant AI Productivity

Ready to run out of the box: Pre-integrated with AI models, software, infrastructure, and services. HPE Private Cloud AI is deployed by HPE in your environment or your chosen colocation in just a few hours.

Optimize performance and utilization: Includes AI optimized infrastructure, networking, software, and models that accelerate the entire AI workflow, so projects reach production faster, with higher accuracy, efficiency, and infrastructure performance at a lower overall cost.[1]

End-to-end AI software platform: NVIDIA AI Enterprise Software and HPE AI Essentials software enable a broad range enterprise grade AI use cases with streamlined solution management and enable rapid deployment of popular generative AI applications.

Hybrid Cloud - Unified Data Access

Secure and unified access to all your data: HPE simplifies data management and reduces cost and complexity by integrating, organizing, and governing enterprise data for seamless access, data integrity and compliance.

Enable enterprise-wide collaboration: Supports organizational collaboration with end-to-end experiment tracking and comparison so teams can iterate quickly and build on each other's work.

Improve AI performance and innovation: Hybrid storage supports the ability to train large models in the cloud and then bring the data on-prem for retrieval augmented generation (RAG) or fine tuning use cases where integrating sensitive or proprietary customer data is preferred.

Enterprise Grade Confidence and Control

Protect private data and models: Maintain data integrity and quality while providing privacy and protection of enterprise data and AI models.

Single-source of world class support: Global HPE and partner services experts provide full-stack support and SLA response times for the complete NVIDIA and HPE software and hardware solution reducing risk for both your solution investment and privacy.

Speed innovation with built-in AI compliance and explainability: Automates AI pipelines with complete lineage, traceability, and auditability for the entire AI lifecycle.

Flexible and Evergreen Cloud Experience

Choice of scalable validated starting points: Optimized for your use cases today and tomorrow, HPE will right size your infrastructure needs so you can start small and then scale up on proven configurations to address multiple use cases, higher throughputs, or more users.

Advanced capabilities and control with HPE GreenLake cloud: Advanced monitoring and management tools built into the HPE GreenLake cloud provide real-time insights into the performance, sustainability metrics, and data governance.

Future-proof your AI investments: Automatically scale out across new hardware, load balance across millions of requests, and monitor all your AI services from an evergreen software platform that is always up to date with the latest features, security patches, and bug fixes.

Technical specifications	HPE Private Cloud AI
Configuration 1	Small - AI Inference 1 x HPE ProLiant DL380a Gen11 AI Optimized Node (4 x NVIDIA L40s GPUs Total) 3 x HPE ProLiant DL325 Gen11 Control Nodes 109 TB HPE GreenLake for File Storage with Object enabled NVIDIA SN4600cM Switches (100GbE Networking) 42U Rack with PDUs HPE AI Essentials with NVIDIA AI Enterprise Software 3 or 5-year subscription
Configuration 2	Expanded Small - AI Inference 2 x HPE ProLiant DL380a Gen11 AI Optimized Node (8 x NVIDIA L40s GPUs Total) 3 x HPE ProLiant DL325 Gen11 Control Nodes 109 TB HPE GreenLake for File Storage with Object enabled NVIDIA SN4600cM Switches (100GbE Networking) 42U Rack with PDUs HPE AI Essentials with NVIDIA AI Enterprise Software 3 or 5-year subscription
Configuration 3	Medium - AI Inference and Retrieval Augmented Generation (RAG) 2 x HPE ProLiant DL380a Gen11 AI Optimized Node (8 x NVIDIA L40s GPUs Total) 3 x HPE ProLiant DL325 Gen11 Control Nodes 217 TB HPE GreenLake for File Storage with Object enabled NVIDIA SN4700M Switches (200GbE Networking) 42U Rack with PDUs HPE AI Essentials with NVIDIA AI Enterprise Software 3 or 5-year subscription
Configuration 4	Expanded Medium - AI Inference and Retrieval Augmented Generation (RAG) 4 x HPE ProLiant DL380a Gen11 AI Optimized Node (16 x NVIDIA L40s GPUs Total) 3 x HPE ProLiant DL325 Gen11 Control Nodes 217 TB HPE GreenLake for File Storage with Object enabled NVIDIA SN4700M Switches (200GbE Networking) 42U Rack with PDUs HPE AI Essentials with NVIDIA AI Enterprise Software 3 or 5-year subscription
Configuration 5	Large - AI Inference, Retrieval Augmented Generation (RAG), and Model Fine Tuning 4 x HPE ProLiant DL380a Gen11 AI Optimized Node (16 x NVIDIA H100NVL GPUs Total) 3 x HPE ProLiant DL325 Gen11 Control Nodes 670 TB HPE GreenLake for File Storage with Object enabled NVIDIA SN4700M Switches (400GbE Networking) 2 x 42U Racks with PDUs HPE AI Essentials with NVIDIA AI Enterprise Software 3 or 5-year subscription

Configuration 6	Expanded Large - AI Inference, Retrieval Augmented Generation (RAG), and Model Fine Tuning 8 x HPE ProLiant DL380a Gen11 AI Optimized Node (32 x NVIDIA H100NVL GPUs Total) 3 x HPE ProLiant DL325 Gen11 Control Nodes 670 TB HPE GreenLake for File Storage with Object enabled NVIDIA SN4700M Switches (400GbE Networking) 2 x 42U Racks with PDUs HPE AI Essentials with NVIDIA AI Enterprise Software 3 or 5-year subscription
Configuration 7	Developer System - AI Inference, Retrieval Augmented Generation (RAG), AI Sandbox Development 1 x HPE ProLiant DL380a Gen11 AI Optimized Node (2 x NVIDIA H100NVL GPUs Total) 1 x HPE ProLiant DL325 Gen11 Control Node 32 TB Integrated File/Object Storage 2x 200GbE Network Ports HPE AI Essentials with NVIDIA AI Enterprise Software 3 or 5-year subscription Switches, Racks, and PDUs are NOT included

[1] Source: HPE internal reports. Comparison between using GPT-4 via OpenAI API vs. self-hosted Llama3, assuming an enterprise account with 5,000 users, 5 chat sessions per day, 8,000 tokens per chat

For additional technical information, available models and options, please reference the QuickSpecs

Make the right purchase decision. Contact our presales specialists.

Call for availability

Chat now (sales)

Call now

Buy now

Share now

Get updates

HPE Services

No matter where you are in your transformation journey, you can count on HPE Services to deliver the expertise you need when, where and how you need it. From strategy and planning to deployment, ongoing operations and beyond, our experts can help you realize your digital ambitions.

Advisory & Professional services

Experts can help you map out your path to hybrid cloud and optimize your operations.

Managed services

HPE runs your IT operations, giving you unified control, so can focus on innovation.

Support services

Optimize your entire IT environment and drive innovation. Manage day-to-day IT operational tasks while freeing up valuable time and resources.

- **HPE Complete Care Service:** a modular service designed to help optimize your entire IT environment and achieve agreed upon IT outcomes and business goals. All delivered by an assigned team of HPE experts.
- **HPE Tech Care Service:** the operational service experience for HPE products. The service provides access to product specific experts, an AI driven digital experience, and general technical guidance to help reduce risk and search for ways to do things better.
- **HPE Multivendor Services:** Single point of accountability for managing on-site hardware and software support for multivendor products. HPE experts help manage your IT across technologies and platforms for HPE and non-HPE technologies, acting as the single point of contact for your IT operational needs.

Lifecycle Services

Address your specific IT deployment project needs with tailored project management and deployment services.

HPE Education Services

Training and certification designed for IT and business professionals across all industries. Create learning paths to expand proficiency in a specific subject. Schedule training in a way that works best for your business with flexible continuous learning options

Defective Media Retention is optional and allows you to retain Disk or eligible SSD/Flash Drives replaced by HPE due to malfunction.

HPE GreenLake

HPE GreenLake edge-to-cloud platform is HPE's market-leading as-a-Service offering that brings the cloud experience to apps and data everywhere – data centers, multi-clouds, and edges – with one unified operating model, on premises, fully managed in a pay per use model.

If you are looking for more services, like **IT financing solutions**, please explore them here.

Visit **HPE.com**

PSN1014847366CZEN, May, 2025.

Hewlett Packard Enterprise