HPE Private Cloud AI

What's new

- HPE Private Cloud AI Air-Gapped Large deployment option.
- HPE Private Cloud AI Developer system featuring NVIDIA® RTX Pro 6000 Blackwell Server Edition GPUs.
- HPE Private Cloud AI Large configuration featuring NVIDIA® RTX Pro 6000 Blackwell Server Edition GPUs.
- HPE Private Cloud AI with the HPE Alletra Storage MP X10000 storage solution and support for data intelligence.

Overview

The promise of AI is often blocked by complexity and risk. HPE Private Cloud AI is the strategic answer—a complete AI factory designed to remove the technical risks and performance bottlenecks of traditional solutions. This means you can achieve faster time to value and a predictable ROI. Co-engineered with NVIDIA®, our solution provides the foundation for confident AI innovation:

•Cost predictability: Our on-premises model removes operational bottlenecks, giving you the financial clarity you need to move beyond pilots.

•Streamlined innovation: Pre-validated tools and notebooks standardize workflows and model development, all while providing consistent governance and zero-touch security from day one.

•Future-proof scalability: Seamlessly expand your AI infrastructure across diverse compute and GPU architectures, so you can innovate with confidence and adapt to the AI technologies of the future.

Features

A Turnkey AI Factory

HPE Private Cloud AI is an engineered AI factory built from the ground up to slash complexity and accelerate your path to AI. We help you create and deploy AI models faster, so you can move from pilot to production with speed and confidence.

Replaces a multi-month infrastructure project with our engineered AI factory, and moves from concept to production in days, not months.

Designed to slash complexity, accelerate time to value, and deliver the cost predictability needed to scale AI innovation enterprise wide.

One of the fastest ways to turn AI ambition into reality.

Ready-to-run Comprehensive Suite of AI Tools

Comprehensive, out-of-the-box AI ecosystem reduces the need to build AI stacks from scratch.

Accelerates AI adoption by helping customers move from pilot to production by reducing complexity, time, and risk associated with AI initiatives.

A consistent, pre-integrated software platform removes the friction and complexity that arise from a fragmented toolchain.

Data scientists and developers can seamlessly move between a wide range of popular AI and open-source tools—from JupyterLab to Apache Spark—with a single sign-on, all while maintaining consistent governance and security from day one.

Empowers novice to expert data scientists with features like low-code capabilities, endpoint APIs, JupyterLab, and pre-built Jupyter notebooks.

Unlock All Data for AI

Gain seamless access to data across your heterogeneous storage systems with a federated data lakehouse. This approach allows you to train and fine-tune your AI models on a single, unified view of your data, without the time, cost, and risk of moving it.

The global namespace is the central hub that solves the pain of time-consuming data integration projects. By giving you a single, seamless access point for your data, it dramatically accelerates the time it takes to prepare and use data for AI, speeding up your entire AI lifecycle.

Accelerate insights and AI model development by empowering your teams to use their preferred analytics engines—from SQL to specialized AI/ML tools—on data stored across your data lakes.

Provides consistent, reliable AI models by eliminating the inconsistencies and version control issues that slow down projects. Centralized governance and security mean you have a single source of truth for all your data, from day one.

Flexible Cloud Experience for Continuous AI Innovation

We help you right-size your AI infrastructure from day one. Choose from a selection of validated starting points that are optimized for your specific use cases.

Start small and seamlessly scale up on proven configurations, helping ensure your platform grows with your needs for more users, higher throughputs, and new AI initiatives.

Future-ready your AI with a perpetual platform. HPE Private Cloud AI is regularly refreshed with the latest CPUs, GPUs, and hardware innovations, all while one-click upgrades help ensure your software is current.

Technical specifications	HPE Private Cloud AI
Configuration 1	Large Air-Gapped (16x H200 GPUs) - AI Inference, Retrieval Augmented Generation (RAG), and Model Fine Tuning 2x HPE ProLiant Compute DL380a Gen12 AI Optimized Node (8x NVIDIA® H200 GPUs per node) 3x HPE ProLiant DL325 Gen11 Control Nodes 109 TB GreenLake for File Storage with Object enabled NVIDIA® SN4700M Switches (400 GbE Networking) 42U Rack with PDUs HPE AI Essentials with NVIDIA® AI Enterprise Software 3 or 5-year subscription Can be expanded to 64x H200 GPUs by adding up to three G2 Expansion Racks
Configuration 2	Developer System - AI Inference, Retrieval Augmented Generation (RAG), AI Sandbox Development 1x HPE ProLiant Compute DL380a Gen11 AI Optimized Node (2x NVIDIA® H100NVL GPUs Total) 1x HPE ProLiant DL325 Gen11 Control Node 22TB Integrated File/Object Storage 2x 200 GbE Network Ports HPE AI Essentials with NVIDIA® AI Enterprise Software 3 or 5-year subscription Switches, Racks, and PDUs are not includ
Configuration 3	Medium (8x H200 GPUs) - AI Inference and Retrieval Augmented Generation (RAG) 2x HPE ProLiant Compute DL380a Gen12 AI Optimized Node (4x NVIDIA® H200 GPUs per node) 3x HPE ProLiant DL325 Gen11 Control Nodes 62TB HPE Alletra Storage MP X10000 File and Object Storage NVIDIA® SN4700M Switches (400 GbE Networking) 42U Rack with PDUs HPE AI Essentials with NVIDIA® AI Enterprise Software 3 or 5-year subscription Can be expanded to 24x H200 GPUs by adding one G2 Expansion Rack
Configuration 4	Large (16x H200 GPUs) - AI Inference, Retrieval Augmented Generation (RAG), and Model Fine Tuning 2x HPE ProLiant Compute DL380a Gen12 AI Optimized Node (8x NVIDIA® H200 GPUs per node) 3x HPE ProLiant DL325 Gen11 Control Nodes 124TB HPE Alletra Storage MP X10000 File and Object Storage NVIDIA® SN4700M Switches (400 GbE Networking) 42U Rack with PDUs HPE AI Essentials with NVIDIA® AI Enterprise Software 3 or 5-year subscription Can be expanded to 64x H200 GPUs by adding up to three G2 Expansion Racks

Technical specifications	HPE Private Cloud AI
Configuration 5	Medium Air-Gapped (8x H200 GPUs) - AI Inference and Retrieval Augmented Generation (RAG) 2x HPE ProLiant Compute DL380a Gen12 AI Optimized Node (4x NVIDIA® H200 GPUs per node) 3x HPE ProLiant DL325 Gen11 Control Nodes 109 TB GreenLake for File Storage with Object enabled NVIDIA® SN4700M Switches (400 GbE Networking) 42U Rack with PDUs HPE AI Essentials with NVIDIA® AI Enterprise Software 3 or 5-year subscription Can be expanded to 24x H200 GPUs by adding one G2 Expansion Rack Supports deployments in a disconnected or Air-gapped environment
Configuration 6	Small 1-Node (4x NVIDIA® RTX PRO 6000 Blackwell Server Edition GPUs) - AI Inference, Retrieval Augmented Generation (RAG), Digital Twins, Physical AI 1x HPE ProLiant Compute DL380a Gen12 AI Optimized Node (4x NVIDIA® RTX PRO 6000 Blackwell Server Edition GPUs per node) 3x HPE ProLiant DL325 Gen11 Control Nodes 62TB HPE Alletra Storage MP X10000 File and Object Storage NVIDIA® SN4700M Switches (400 GbE Networking) 42U Rack with PDUs HPE AI Essentials with NVIDIA® AI Enterprise Software 3 or 5-year subscription Can be expanded by 8x NVIDIA® RTX PRO 6000 Blackwell Server Edition GPUs or 16x H200 GPUs by adding one G2 Expansion Rack
Configuration 7	Small 2-Node (8x NVIDIA® RTX PRO 6000 Blackwell Server Edition GPUs) - AI Inference, Retrieval Augmented Generation (RAG), Digital Twins, Physical AI 2x HPE ProLiant Compute DL380a Gen12 AI Optimized Node (4x NVIDIA® RTX PRO 6000 Blackwell Server Edition GPUs per node) 3x HPE ProLiant DL325 Gen11 Control Nodes 62TB HPE Alletra Storage MP X10000 File and Object Storage NVIDIA® SN4700M Switches (400 GbE Networking) 42U Rack with PDUs HPE AI Essentials with NVIDIA® AI Enterprise Software 3 or 5-year subscription Can be expanded by 8x NVIDIA® RTX PRO 6000 Blackwell Server Edition GPUs or 16x H200 GPUs by adding one G2 Expansion Rack

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.

GreenLake

GreenLake is the cloud to run and manage your entire hybrid landscape—private, public, and edge. It helps you to:

— Streamline IT Operations across compute, storage, and networking without the chaos

— Unify and secure data, as you move faster

— Accelerate AI, from pilot to production

The result: greater operational efficiency, lower TCO, and faster AI delivery—all from one unified, intelligent platform built for today's hybrid enterprise.

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

Visit HPE.com

Chat now

© Copyright 2026 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.

Parts and Materials: HPE will provide HPE-supported replacement parts and materials required to maintain the covered hardware.

Parts and components that have reached their maximum supported lifetime and/or the maximum usage limitations as set forth in the manufacturer's operating manual, product quick-specs, or the technical product data sheet will not be provided, repaired, or replaced as part of these services.

NVIDIA is a trademark and/or a registered trademark of NVIDIA Corporation in the U.S. and other countries.
All third-party marks are property of their respective owners.

Image may differ from the actual product.

PSN1014847366SAEN, October, 2026.

HEWLETT PACKARD ENTERPRISE

hpe.com