HPE Private Cloud AI

HPE Private Cloud AI is a purpose-built solution designed to provide fast and easy deployment of private AI applications with a focus on inferencing, Retrieval-Augmented Generation (RAG), and fine-tuning. HPE Private Cloud AI is a co-developed HPE and NVIDIA enterprise purpose-built solution including in a completed infrastructure, software portfolio

At A Glance

HPE Private Cloud AI delivers a unique cloud experience designed to accelerate data science productivity and time to business value. It delivers instant AI productivity by arriving at a customer's location ready to deploy in 3 clicks. Once available, multiple persona have self-service access to a diverse set of NVIDIA technologies and open-source tools and models to increase productivity by 90% through an evergreen, cloud-managed experience[1]

Notes: [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

AI teams can innovate faster with built-in compliance and explainability to foster model trust, quickly detect model bias, diagnose and improve model performance, and remain compliant with industry regulations.

Built with enterprise-grade controls means organizations can fearlessly innovate, with a scalable platform - all controlled from a unified dashboard.

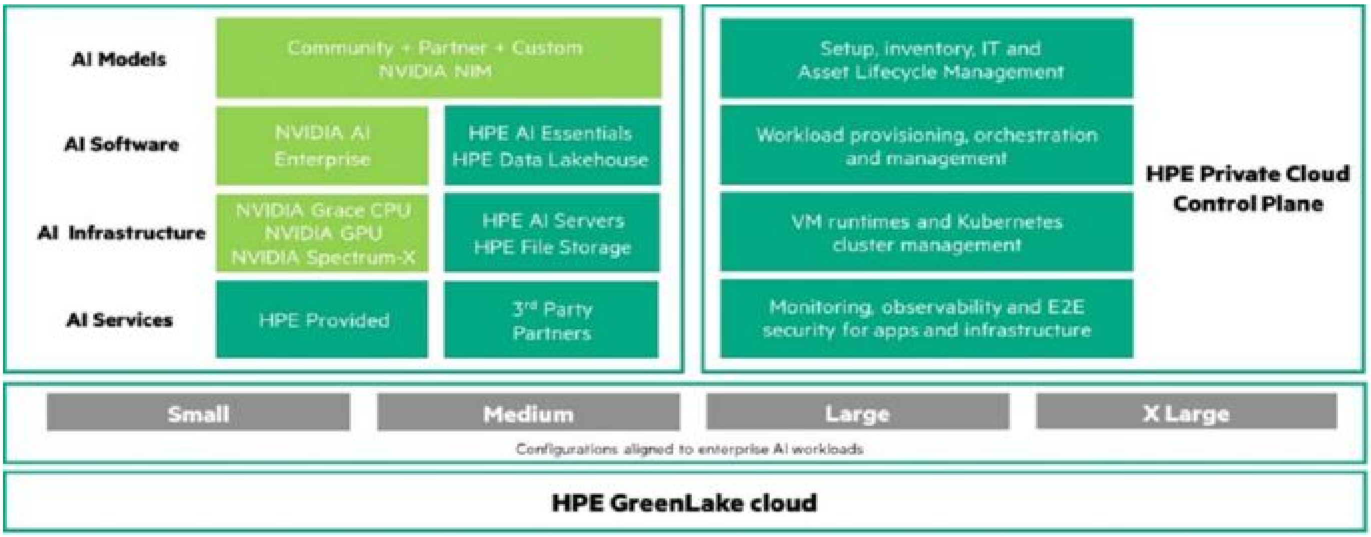
Future-proof your AI journey with HPE. Launch small, scale seamlessly, and invest confidently with our co-developed NVIDIA + HPE solution. One modular architecture protects customer's investment by ensuring compatibility with future innovations from NVIDIA, HPE, and the open-source world.

- Ready to use in three clicks
- Evergreen cloud experience with NVIDIA technologies and a rich ecosystem of open-source tools and models
- Automated AI pipelines with clear data lineage and verifiable changes empower efficient, accountable development
- Robust security, on-demand scalability, and compliance for data and AI models - all managed from a single dashboard
- Start small, then seamlessly expand for future needs while maintaining a consistent cloud experience

One of the challenges businesses face is getting AI pilots to production faster.

HPE Private Cloud AI delivers instant AI productivity with a unique, private cloud experience that accelerates the productivity of data science teams and time to business value with NVIDIA AI Computing.

Overview


AI Models
Community + Partner + Custom
NVIDIA NIM
AI Software
NVIDIA AI Enterprise
HPE AI Essentials
HPE Data Lakehouse
AI Infrastructure
NVIDIA Grace CPU
NVIDIA GPU
NVIDIA Spectrum-X
HPE AI Servers
HPE File Storage
AI Services
HPE Provided
3rd Party Partners
Setup, inventory, IT and Asset Lifecycle Management
Workload provisioning, orchestration and management
VM runtimes and Kubernetes cluster management
Monitoring, observability and E2E security for apps and infrastructure
HPE Private Cloud Control Plane
Small
Medium
Large
X Large
Configurations aligned to enterprise AI workloads
HPE GreenLake cloud

HPE Private Cloud for AI offers enterprise customers the ability to leverage NVIDIA AI Enterprise (NVAIE) portfolio, including NVIDIA Inferencing Microservices (NIM), and HPE portfolio of curated market adopted open-source AI tools and platforms with full private control of their data. The solution will enable enterprises to expedite their Machine Learning and AI initiatives starting from creating their private data lakehouses, to data pipeline, model development and fine-tuning, to operationalizing their GenAI workflows.

What's New

- **HPE Private Cloud AI**

HPE Private Cloud AI delivers instant AI productivity with a unique, private cloud experience that accelerates the productivity of data science teams and time to business value with NVIDIA AI Computing.

Category	Description
Platform	• Server support: HPE ProLiant Gen11 servers • AMD-based HPE ProLiant DL325 Gen 11 Control Nodes • Intel-based HPE ProLiant DL380a Gen 11 AI Worker Nodes • Storage support: HPE GreenLake for File with Object Storage enabled
Manageability	• Cloud-based setup and lifecycle management (single-click upgrades)
Analytics & Monitoring	• Cluster and VM capacity and performance, storage health status information
Support	• One call support experience with HPE Services

- **HPE Private Cloud AI Smart templates**

    – Availability of pre-configured Smart Templates with HPE ProLiant Gen 11 Servers

Key Features and Benefits

HPE Private Cloud AI is turnkey, deployed in minutes, cloud-managed, and ready to use by AI personas and IT operations teams and provides rapid productivity for AI initiatives while protecting data and IP. The key value proposition aligned to customer problems are:

The core feature set includes:

- Instant AI productivity: HPE Private Cloud AI provides a unique, private cloud experience that accelerates data science productivity and time to business value with NVIDIA AI Computing. The solution is pre-integrated and ready to run out of the box in minutes. It is not a reference architecture like other solutions in the market.
- Unify access to all your data: 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. Enterprise-grade confidence and control: HPE Private Cloud AI is managed through a simple control plane on HPE GreenLake. Users can easily provision, orchestrate, manage and monitor the private cloud environment and the hybrid cloud landscape it exists within. Comprehensive, multi-layered controls protect sensitive data and models and maintain high performance, reliability and utilization of AI infrastructure.

Cloud experience that keeps data private: HPE Private Cloud AI delivers a true cloud experience through HPE GreenLake. Deployed on-premises and designed for hybrid, HPE Private Cloud Ai provides flexible and modular choices to expand and grow with AI demand. As business needs change, it's easy for customers to grow the solution. And monthly subscription pricing allows customers to start small financially and grow as their projects prove ROI.

Service and Support

Support is included as part of the subscription for HPE Private Cloud AI. Included with the support is 24x7 telephone and email support for the arrays and hardware components for the chosen subscription term. Refer to the HPE Private Cloud AI Data sheet **https://www.hpe.com/psnow/doc/a50009418ENW?ver=2** for the service deliverables and the shared responsibility model as part of the subscription.

Easy Configuration through Smart templates

There are pre-defined smart templates available that allow for quick and easy ways to quote:

1. HPE Private Cloud AI

Smart templates can be customized for additional options.

There are pre-defined smart templates that allow for quick and easy way to quote HPE Private Cloud AI.

Here is an example of a HPE Private Cloud AI Smart template:

Config Name: PrivateCloudAI-Small-1Svr/4xL40S GPU-109TB File/Object-3Phase/NA-Jpn-PDU-1Rack-3yr
Description: HPE Private Cloud AI Small Single Node-4GPU Solution for AI Inference. 109TB File/Object Storage, 100GbE Networking, Single Rack and 3Phase PDU.

The Smart templates contain the following attributes to choose,

1. **T-Shirt Sizing** - Small, Medium, or Large Configurations
2. **Workload Tier** -
    a. AI Inference
    b. Retrieval Augmented Generation (RAG)
    c. Model Fine Tuning

T-Shirt Size	Entry	Expanded
Small	4x L40s GPUs and 109TB Storage for AI Inference	8x L40s GPUs and 109TB Storage for AI Inference
Medium	8x L40s GPUs and 217TB Storage for AI Inference and RAG	16x L40s GPUs and 217TB Storage for AI Inference and RAG
Large	16x H100NVL GPUs and 670TB Storage for AI Inference, RAG, and Fine Tuning	32x H100NVL GPUs and 670TB Storage for AI Inference, RAG and Fine Tuning

Notes: Storage amounts shown are usable capacity

3. **Network Configuration**

Network Configuration	Detail
Networking equipment included	Two top-of-rack switches and out of band management switches are included along with all transceivers and signal cabling required for the full solution

Notes: The deployment and startup services included with HPE Private Cloud AI will include the setup and configuration of top-of-rack switches.

4. Rack and power Configuration

Rack Configuration	Detail
Rack included	The solution will include a 42U with integrated PDUs for HPE Private Cloud AI.
	Rack Dimensions: 600mm (W), 1200mm (D)

Resources and additional links

- The networking requirements, best practices, supported technologies, and supported network topologies for HPE Private Cloud AI
  https://psnow.ext.hpe.com/doc/a00114771enw

Shared Responsibility Model (SRM)

HPE Private Cloud AI subscription includes the necessary hardware, software, and services to deliver the service level specified. The service levels offered are based on a foundational shared responsibility model (SRM) depicted below:

Customer	HPE
Responsible for the connectivity to GreenLake Cloud Platform (GLCP), the administration, and the management of the data/ objects	Responsible for the functionality of the infrastructure providing the service
Site Readiness including datacenter facilities and internet connectivity	Installation & activation of device
Maintain connectivity to GreenLake Cloud Platform Volume Creation and administration Data resilience and remote replication Data backup	Customer Orientation Access to software, firmware, and documentation updates Onsite hardware support
Applying recommended software updates & security patches Data Monitoring	Proactive support and operational guidance* Test volume Creation Operational Insights and Dashboard* Proactive capacity planning*
Initiating the order of additional capacity beyond total available capacity	Proactive incident alerting*
	Communicate security incident & remediation

Notes: *Proactive communication is delivered using HPE Infosight (included as part of subscription)

Pre-requisite for HPE Private Cloud AI

As part of the shared responsibility model, the customer is expected to make appropriate decisions including but not limited to:

Configuration Information

- Rack Infrastructure
  - Space
  - Rails
- Power Infrastructure
  - PDU - Cables

Summary of Changes

Date	Version History	Action	Description of Change
03-Sep-2024	Version 1	New	New QuickSpecs

Copyright

Make the right purchase decision. Contact our presales specialists.

Chat now (sales)

Call now

Get updates

Hewlett Packard
Enterprise

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.

a50009216enw - 17248 - Worldwide - V1 - 03-September-2024