Shape the Future of QuickSpecs - Your Input Matters

HPE Private Cloud AI QuickSpecs

HPE Private Cloud AI is a purpose-built solution designed to provide fast and easy deployment of On-Prem 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 solution that includes a complete infrastructure and software portfolio.

Overview

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 three clicks after hardware installation and software onboarding are complete. Once available, multiple personas 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 customers' investment by ensuring compatibility with future innovations from NVIDIA, HPE, and the open-source world.

- 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

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.

HPE Private Cloud 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 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

As agentic AI simplifies automation and provides consistent profitability from AI workloads, the importance of rapid productivity, reliability, and data protection has never been greater. To accelerate the development and delivery of agentic AI across industries, Hewlett Packard Enterprise and NVIDIA® are introducing the Private Cloud AI Developer and Large configurations featuring the NVIDIA RTX Pro 6000 Blackwell Server Edition GPUs to their end-to-end enterprise AI platform with support for Physical AI and Visual Computing workloads

HPE is also supporting secure deployments with the Private Cloud AI Large air-gapped solution. Ensure the protection of sensitive data with private AI model customization by eliminating exposure to external networks.

Category	Description
Platform	– Server support: HPE ProLiant Compute Gen11 and Gen12 servers – AMD-based HPE ProLiant Compute DL325 Gen 11 Control Nodes – Intel-based HPE ProLiant Compute DL380a Gen 12 AI Worker Nodes – Storage support: HPE GreenLake for File with Object Storage enabled for Small/Medium/Large T-shirt size
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 Compute Gen 12 Servers

HPE Private Cloud AI Family

Feature	Developer System	Small	Medium	Large
Control Node Qty	1	3	3	3
Worker Node Qty	1	1 or 2	2	2
Worker Node Generation	HPE ProLiant Compute Gen12	HPE ProLiant Gen12	HPE ProLiant Gen12	HPE ProLiant Gen12
CPU Type / Qty (per node)	2x Xeon 32 Core CPUs	2x Xeon 86 Core CPUs	2x Xeon 86 Core CPUs	2x Xeon 86 Core CPUs
GPU Type / Qty	2x RTX Pro 6000	4 or 8x RTX Pro 6000	8x H200	16x H200 or RTX Pro 6000
Storage	22 TB Internal File/Object	109 TB GreenLake for File with Object Storage	109 TB GreenLake for File with Object Storage	217T B GreenLake for File with Object Storage
Networking Switches	N/A	NVIDIA 4700M Aruba 6300M (oobm)	NVIDIA 4700M Aruba 6300M (oobm)	NVIDIA 4700M Aruba 6300M (oobm)
NIC Speed (AI Network)	200 Gb NICs	400 Gb NICs	400 Gb NICs	400 Gb NICs
Rack / PDU	N/A	1x 42U Rack with PDUs	1x 42U Rack with PDUs	1x 42U Rack with PDUs
Install Services Included	N/A	Yes	Yes	Yes
Sales Motion	Traditional or GreenLake	Traditional or GreenLake	Traditional or GreenLake	Traditional or GreenLake
Air-Gapped Deployment Option	No	No	Yes	Yes
Expansion Rack Support	No	Yes	Yes	Yes

HPE Private Cloud AI Expansion Racks

Feature	RTX Expansion Rack	H200 Expansion Rack
Control Node Qty	0	0
Worker Node QTY	2	2
Worker Node Generation	HPE ProLiant Gen12	HPE ProLiant Gen12
CPU Type / QTY (per node)	2x Xeon 86 Core CPUs	2x Xeon 86 Core CPUs
GPU Type / QTY (per solution)	8x RTX Pro 6000 (Small) 16x RTX Pro 6000 (Large)	16x H200 (Medium or Large)
Expansion Racks Supported	N/A	N/A
Storage	N/A	N/A
Networking Switches	N/A	N/A
NIC Speed (AI Network)	400Gb NICs	400Gb NICs
Rack / PDU	1x 42U Rack with PDUs	1x 42U Rack with PDUs
Install Services Included	Yes	Yes
Sales Motion	Traditional or GreenLake	Traditional or GreenLake
Air-Gapped Deployment Option	No	Yes

Key Features and Benefits

HPE Private Cloud AI is turnkey, deployed in a few hours, 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 propositions 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 hours. 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/a50010051enw?section=Document%20Types 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**

There are pre-defined smart templates that allow for quick and easy ways 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** – Developer System, Small, Medium, or Large Configurations
2. **Workload Tier –**
    a. AI Inference
    b. Retrieval Augmented Generation (RAG)
    c. Model Fine-tuning
    d. Visual / Physical AI

PCAI T-Shirt Size	Base	Expanded
Developer System	2x RTX Pro 6000 GPUs and 22TB of integrated Storage for developing low-mid parameter model AI applications	N/A
Small	4 or 8x RTX Pro 6000 GPUs and 109TB Storage for AI Inference and Visual Computing	8 or 12 RTX Pro 6000 GPUs via 1x Expansion Rack
Medium	8x H200 GPUs and 109TB Storage for AI Inference and RAG	24x H200 GPUs via 1x Expansion Rack
Large	16x H200 or RTX Pro 6000 GPUs and 217TB Storage for AI Inference, RAG, Fine-tuning, and Visual / Physical AI	Up to 64x H200 or RTX Pro 6000 GPUs via 3x Expansion Racks

Notes: Storage amounts shown reflect usable capacity.

Configuration Information

3. Network Configuration

Network Configuration (Small/Medium/Large)	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 l in-rack solution (Customers are responsible for transceivers to connect to core network switches)

Network Configuration (developer system)	Detail
Networking equipment included	None- Requires customer furnished networking (100GbE or 200GbE recommended)

Notes: The deployment and startup services included with HPE Private Cloud AI will include the setup and configuration of top-of-rack switches for Small, Medium, and Large configurations only.

4. Rack and power Configuration

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

Rack Configuration (developer system)	Detail
Rack included	None- Requires customer furnished rack and PDUs (~2400W Recommended)

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:

HPE Private Cloud AI Small/Medium/Large Configurations

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 service
Site Readiness including datacenter facilities and internet connectivity	Installation of hardware systems & activation of Service
Maintain connectivity to GreenLake Cloud Platform 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	
Initiating the order of additional capacity beyond total available capacity	Operational guidance through the platform
Red Hat Linux and Rocky Linux OS lifecycle management	
NVIDIA and Aruba Switch OS/Firmware lifecycle management	Operational insights and dashboard through the platform
HPE PDU Firmware lifecycle management	
NVIDIA GPU Firmware lifecycle management	

HPE Private Cloud AI Medium Air-Gapped Configuration

Customer	HPE
Site Readiness including datacenter facilities	Installation of hardware systems & activation of Service
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	
Initiating the order of additional capacity beyond total available capacity	
NVIDIA and Aruba Switch OS/Firmware lifecycle management	
HPE PDU Firmware lifecycle management	
NVIDIA GPU Firmware lifecycle management	

Configuration Information

HPE Private Cloud AI developer system

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 service
Site Readiness including datacenter facilities and internet connectivity	
Maintain connectivity to GreenLake Cloud Platform Data resilience and remote replication Data backup	Access to software, firmware, and documentation updates Onsite hardware support
Applying recommended software updates & security patches Data Monitoring	
Initiating the order of additional capacity beyond total available capacity	Operational guidance through the platform
Red Hat Linux and Rocky Linux OS lifecycle management	
Installation of hardware systems & activation of Service	Operational insights and dashboard through the platform
NVIDIA GPU Firmware lifecycle management	

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:

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

Summary of Changes

Date	Version History	Action	Description of Change
02-Mar-2026	Version 9	Changed	The current version of HPE Private Cloud AI QuickSpecs reflects a major generational update focused on next-generation NVIDIA Blackwell GPUs, simplification of configurations, and expanded secure (air-gapped) deployment options. Legacy G1 configurations have been removed, while new Developer, Large, and disconnected (air-gapped) options have been introduced to support modern GenAI, RAG, fine-tuning, and Physical/Visual AI workloads. – Removed G1 Configurations. Added New Developer System, Large with RTX and Large Disconnected Configurations
06-Oct-2025	Version 8	Changed	HPE Rebranding FY25
02-Sep-2025	Version 7	Changed	Overview and Configuration Information sections were updated. - AI G2 Small configuration featuring NVIDIA® RTX Pro 6000 Blackwell Server Edition GPUs information was added.
04-Aug-2025	Version 6	Changed	Overview and Configuration Information sections were updated. - Added G2 Medium Air-Gapped configurations
21-Jul-2025	Version 5	Changed	Survey link updated.
02-Jun-2025	Version 4	Changed	Overview and Configuration Information sections were updated.
07-Apr-2025	Version 3	Changed	Overview and Configuration Information sections were updated. Updates for developer system
03-Mar-2025	Version 2	Changed	Overview section was updated
03-Sep-2024	Version 1	New	New QuickSpecs

Copyright

Shape the Future of QuickSpecs - Your Input Matters

Chat now

a50009216enw - 17248 - Worldwide - V9 - 02-March-2026
HEWLETT PACKARD ENTERPRISE
Hpe.com