Hewlett Packard Enterprise

HPE Private Cloud AI 1.5 Administration Guide

Part Number: 20-PCAI-UG-ED1
Published: September 2025
Edition: 1.5

HPE Private Cloud AI 1.5 Administration Guide

Abstract

This guide is the primary reference for Cloud Administrators, AI Administrators, and AI Users who work with the HPE Private Cloud AI AI solution. The guide describes the solution architecture, components, and configurations, with a focus on HPE AI Essentials - the core engine that powers the private cloud AI infrastructure. It covers essential topics, including user roles and permissions, system management, troubleshooting procedures, and maintenance protocols.

Part Number: 20-PCAI-UG-ED1
Published: September 2025
Edition: 1.5

Notices

Acknowledgments

Document History

Table 1. Document History

Document Version	Date	Private Cloud AI Version	Description
1.5	04 Sept 2025	1.5	AI Essentials 1.9.x support

Table of contents

Overview of the HPE Private Cloud AI engineered system

The Private Cloud AI engineered system by Hewlett Packard Enterprise features explicit support for HPE and NVIDIA AI software, and is designed to meet the growing demand for secure, scalable, and efficient Artificial Intelligence infrastructure. As organizations increasingly rely on AI to drive innovation and competitive advantage, HPE Private Cloud AI offers a powerful platform that combines the benefits of cloud computing with the control and security of on-premises systems.

At its core, HPE Private Cloud AI is a purpose-built infrastructure optimized for AI workloads, particularly generative AI applications. It leverages extensive experience in high-performance computing, data management, and enterprise IT to create a unified environment where businesses can build, deploy and operate AI models with unprecedented ease and efficiency.

HPE AI Essentials, the core engine of HPE Private Cloud AI, offers a comprehensive turnkey solution for Generative AI applications. This platform integrates pre-configured AI and data foundation tools under a unified control plane, providing adaptable solutions tailored to specific business needs. It combines cutting-edge NVIDIA AI computing technology with enterprise-grade security and compliance measures, enabling swift deployment, management, and scaling of AI initiatives.

The architecture of HPE Private Cloud AI is optimized for AI workloads, featuring high-performance NVIDIA GPUs, low-latency networking, and specialized storage systems to handle massive datasets and complex computations for inference, RAG, and fine-tuning tasks. HPE AI Essentials includes built-in data management tools for effective data curation, preparation, and governance. With ongoing enterprise support, trusted AI services for data and model compliance, and features ensuring AI pipeline compliance, explainability, and reproducibility throughout the AI lifecycle, HPE AI Essentials empowers organizations to fully leverage Generative AI while maintaining high performance, security, and trustworthiness.

HPE Private Cloud AI offers a range of deployment options to suit diverse organizational needs. The platform can be implemented on-premises, in a colocation facility, or as a managed service, providing flexibility in how organizations host their AI infrastructure. This adaptability extends beyond deployment models. Once in place, HPE Private Cloud AI allows for independent scaling of compute and storage resources, enabling organizations to fine-tune their infrastructure as demands evolve. This dual-layered flexibility—in initial deployment and ongoing resource management—ensures that HPE Private Cloud AI can accommodate various use cases and growth patterns.

HPE Private Cloud AI, powered by HPE AI Essentials, prioritizes security, privacy, and performance in its design. By keeping AI workloads and sensitive data within a private cloud environment, organizations maintain full control over their intellectual property and ensure regulatory compliance. HPE AI Essentials provides a comprehensive security and observability framework that integrates robust protection measures with advanced performance monitoring. This includes encryption, access controls, and audit logging alongside AI model performance tracking, drift detection, and resource utilization monitoring. This holistic approach enables organizations to safeguard their AI assets, maintain model reliability and effectiveness over time, proactively address potential issues, and optimize resource allocation—all while leveraging the power of AI within a secure private cloud ecosystem.

Benefits of HPE Private Cloud AI

The benefits of HPE Private Cloud AI include:

- **Accelerated AI development**: HPE Private Cloud AI's comprehensive ML/AI/GenAI stack, featuring HPE AI Essentials with NVIDIA NIM™ integration, empowers organizations to swiftly build, deploy, and operate AI/ML models and GenAI applications. NVIDIA NIM™, part of NVIDIA AI Enterprise, provides GPU-accelerated inferencing microservices for pretrained and customized AI models, deployable across clouds, data centers, and workstations with a single command. These microservices, built on pre-optimized inference engines such as NVIDIA® TensorRT™ and TensorRT-LLM, optimize runtime response latency and throughput for each model-GPU combination. With industry-standard APIs, built-in observability, and Kubernetes autoscaling support, NVIDIA NIM significantly streamlines AI integration and scaling. Combined with HPE AI Essentials, this optimized ecosystem accelerates the entire AI lifecycle from development to production, dramatically reducing time-to-value for AI initiatives across the enterprise.
- **Cost optimization**: By providing a dedicated AI infrastructure, HPE Private Cloud AI helps organizations avoid the unpredictable costs associated with public cloud AI services, especially for large-scale or continuous workloads.
- **Enhanced security and compliance**: The private cloud approach, which involves deploying cloud services on dedicated infrastructure either on-premises or through a third-party provider, allows for greater control over data and models, facilitating regulatory compliance and intellectual property protection. This enhanced control enables organizations to implement customized security measures, maintain data isolation, and create detailed audit trails. By utilizing dedicated hardware, organizations can physically separate sensitive information, implement tailored security protocols, and maintain direct oversight of physical and digital security measures. This level of control allows for more effective risk management and makes it easier to demonstrate compliance with regulations such as GDPR or HIPAA. Additionally, keeping proprietary algorithms and models within a controlled environment helps safeguard intellectual property from potential exposure or theft, providing a higher level of security and compliance assurance compared to public cloud solutions where infrastructure is shared and controlled by the cloud provider.

- **Scalability**: HPE Private Cloud AI's modular design allows organizations to start small and expand their AI capabilities as needed, without major disruptions or reinvestments.
- **Simplified management**: HPE AI Essentials provides comprehensive management tools and automation features, reducing the operational complexity of running AI infrastructure.
- **Ecosystem integration**: HPE Private Cloud AI is designed to work seamlessly with popular AI frameworks and development tools, such as PyTorch, TensorFlow and JAX, allowing organizations to leverage their existing investments and skills.
- **Expert support**: HPE and NVIDIA provide end-to-end support for HPE Private Cloud AI, including consulting services to help organizations optimize their AI strategies and implementations. HPE offers comprehensive infrastructure solutions, while NVIDIA provides support for its AI frameworks and GPU technologies, enabling organizations to build and deploy powerful AI systems on-premises.

Example Use Cases

HPE Private Cloud AI offers organizations across various industries the ability to leverage the power of AI while maintaining enhanced security, data privacy, and customization capabilities. The following examples represent a **subset** of potential use cases, demonstrating how customers can develop tailored solutions using to address their specific needs.

Subtopics

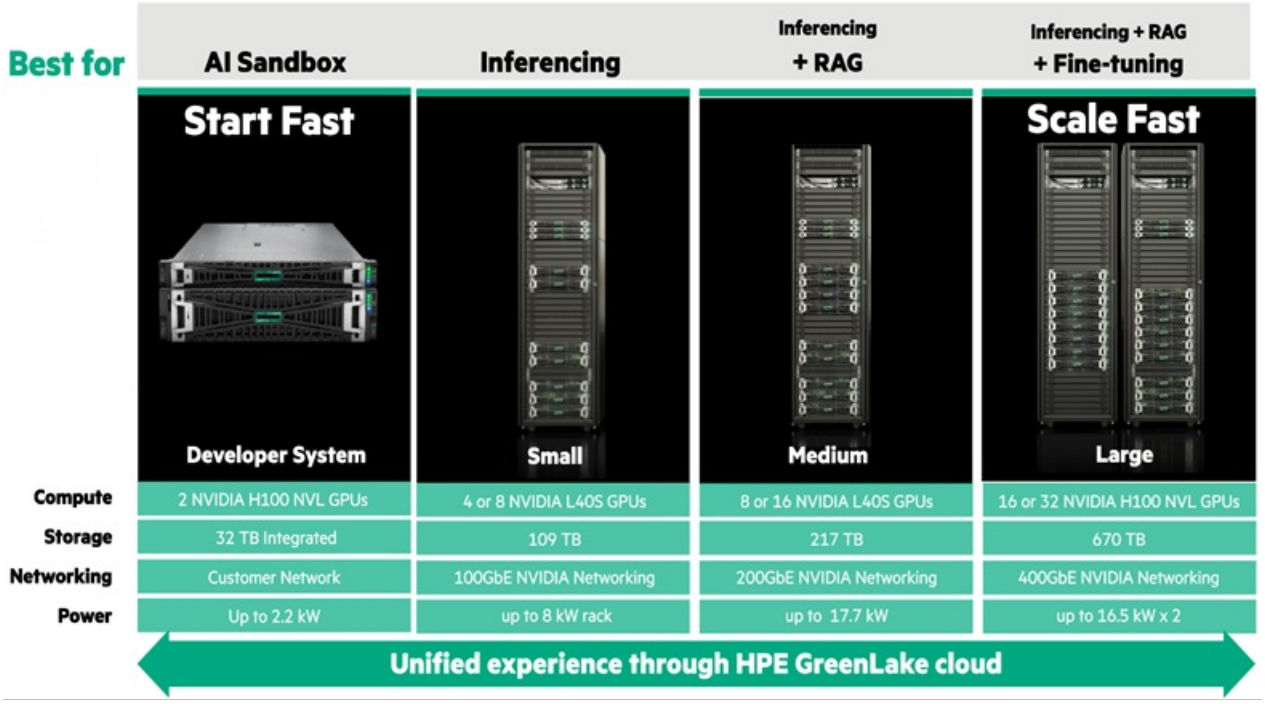
Private Cloud AI Stack

The Private Cloud AI solution stack includes robust hardware, AI-native computing, storage, and networking components, all of which are accessible and manageable through the GreenLake Cloud platform.

The following diagram shows the HPE Private Cloud AI configurations, which include:

- Developer System configuration
- Small configuration
- Medium configuration
- Large configuration

Figure 1. Private Cloud AI Configurations



Developer System

HPE Private Cloud AI 1.4 introduced a new configuration called the **Developer System**. Designed for developers and Data Engineers, the Developer System is a customer-installable, low-cost alternative to the Small, Medium, and Large configurations.

The Developer System includes fewer control nodes and worker nodes, but runs the same HPE and NVIDIA software. Use cases span the entire AI/ML spectrum but are limited in scale. The system is targeted to small-scale inference and broad prototyping. RAG workflows with very large LLMs and large-scale inference are not supported because the Developer System has only two GPUs.

The following table compares the Developer System with the other HPE Private Cloud AI configurations:

Feature	Developer System	Small	Medium	Large
Control Node Qty	1	3	3	3
Worker Node Qty	1	1 or 2	2 or 4	4 or 8
CPU Type / Qty (per node)	2x Xeon 32-core CPUs	2x Xeon 32-core CPUs	2x Xeon 32-core CPUs	2x Xeon 32-core CPUs
GPU Type / Qty	2x H100 NVL	4 or 8x L40S	8 or 16x L40S	16 or 32x H100 NVL
File System	Local NFS	HPE GreenLake for File Storage	HPE GreenLake for File Storage	HPE GreenLake for File Storage
Storage	32TB internal file/object	109 TB GreenLake for File with Object Storage	217 TB GreenLake for File with Object Storage	670 TB GreenLake for File with Object Storage
Networking Switches	N/A	• NVIDIA 4600cM • Aruba 6300M (OOBM)	• NVIDIA 4700M • Aruba 6300M (OOBM)	• NVIDIA 4700M • Aruba 6300M (OOBM)
NIC Speed (AI Network)	200 Gb NICs	100 Gb NICs	200 Gb NICs	400 Gb NICs
Rack / PDU	N/A	1x 42U Rack with PDUs	1x 42U Rack with PDUs	2x 42U Racks with PDUs
Installation Services Included	N/A	Yes	Yes	Yes
OpsRamp Included	No	Yes	Yes	Yes

HPE AI Essentials: Core Concepts

HPE AI Essentials Software is part of the HPE Private Cloud AI software and data foundation layer that accelerates the time from AI pilot to production. HPE AI Essentials Software provides a ready-to-run set of AI and open-source tools to build end-to-end GenAI solutions.

From novice to expert, HPE AI Essentials Software provides every user immediate access to a complete suite of proprietary and open-source tools. The no-code/low-code features and automated accelerators empower beginners to quickly build generative AI applications such as chatbots and productivity tools. Experienced developers benefit from APIs, notebooks, and advanced tools that enable rapid experimentation, iteration, and deployment, eliminating IT bottlenecks for resource acquisition and provisioning.

NOTE

To start using HPE AI Essentials Software, see the **HPE AI Essentials Software Documentation** .

The following sections summarize the key features and capabilities of HPE AI Essentials Software.

Core Components

Open-Source Tools & Frameworks

- Implements a comprehensive managed ecosystem of interconnected open-source tools and frameworks specifically designed for AI workloads.
- Features a unified authentication system through Single Sign-On (SSO), enabling seamless transitions between applications without requiring multiple authentication steps.

Data Engineering and Analytics Tools

- **Apache Spark**: Enterprise-grade analytics engine supporting large-scale data processing and transformation operations. Includes a wizard-driven UI for creating and scheduling Spark jobs, making complex data operations more accessible.
- **Apache Airflow**: Advanced workflow orchestration platform that integrates directly with HPE AI Essentials data sets. Enables creation and management of complex data pipelines with sophisticated scheduling capabilities.
- **EzPresto**: SQL-based query engine optimized for large-scale data analysis. Provides an interactive Query Editor interface for direct data manipulation and exploration of datasets.
- **Superset**: Enterprise-class business intelligence platform offering advanced data visualization capabilities. Enables creation of interactive dashboards and complex data exploration through a user-friendly interface.

Data Science Tools

- **KubeFlow**: Comprehensive machine learning operations (MLOps) platform deployed on Kubernetes, providing end-to-end lifecycle management for ML projects. Includes:
  - Native support for **Jupyter** lab and **VS code** environments with customizable computational resources
  - Advanced pipeline management system for orchestrating complex ML workflows
  - **KServe** integration for robust model serving and inference capabilities
- **MLflow**: Production-grade ML lifecycle management tool providing extensive experiment tracking capabilities and a centralized model registry. Enables version control and reproducibility in ML experiments.
- **Ray**: Distributed computing framework optimized for scaling AI and Python workloads across clusters. Tuned explicitly for efficient deployment of machine learning and deep learning workloads.
- **Feast**: Production-ready feature store designed for machine learning operations. Provides unified storage and serving of features for both batch and real-time inference scenarios.

NVIDIA AI Enterprise Integration

HPE, in collaboration with NVIDIA AI Enterprise, integrates pre-packaged NVIDIA NIM for large language models (LLMs) with HPE AI Essentials Software. **NVIDIA NIM** is a set of easy-to-use microservices that accelerate the deployment of generative AI models. For example, NVIDIA NIM brings the power of state-of-the-art LLMs to enterprise applications, providing unmatched natural language processing and understanding capabilities.

NVIDIA NIM for LLMs leverages NVIDIA's cutting-edge GPU acceleration and scalable deployment to build powerful copilots, chatbots, and

AI assistants for the fastest path to inference with unparalleled performance.

HPE AI Essentials Software currently includes the following pre-packaged NVIDIA NIM:

- NVIDIA NIM for GPU accelerated NVIDIA Retrieval QA **Mistral 7B** Embedding v2 inference
- NVIDIA NIM for GPU accelerated NVIDIA Retrieval **QA E5** Embedding v5 inference
- NVIDIA NIM for GPU accelerated **Llama 3 70B** inference through OpenAI compatible APIs
- NVIDIA NIM for GPU accelerated **Llama 3 8B** inference through OpenAI compatible APIs
- NVIDIA NIM for GPU accelerated NVIDIA Retrieval QA **Mistral 4B Reranking** v3 inference

NOTE

You can access NVIDIA NIM models on the NVIDIA tab under Tools & Frameworks. Clicking View Details on the NVIDIA NIM tile shows the metadata, such as the version number, storage location, and release date.

Empowering Innovation with NVIDIA Blueprints

NVIDIA Blueprints are predefined, customizable AI workflows from NVIDIA that can assist you in creating and deploying generative AI applications. Blueprints are designed for you to download, customize, and incorporate into your existing applications. The full catalogue of NVIDIA Blueprints is available at **NVIDIA Blueprints**.

You can use Blueprints to jumpstart or accelerate the development and deployment of AI use cases. Blueprints cover a wide range of use cases, and new Blueprints are always being added to the NVIDIA Blueprints catalog. Use cases range from designing enterprise RAG pipelines to building AI virtual assistants and digital twins.

Whether you're looking to develop advanced analytics tools, industry-specific AI models, or innovative applications that push the boundaries of what's possible with AI, HPE AI Essentials provides the robust infrastructure and customization capabilities you need. By leveraging its comprehensive components, businesses can transform their unique expertise and data into groundbreaking AI solutions that drive competitive advantage and unlock new opportunities in their respective fields.

Advanced Features

Notebooks Environment

- Provides isolated, secure notebook environments for each user with dedicated computational resources.
- Automatically mounts HPE GreenLake for File Storage volumes under /mnt/datasources for immediate data access.
- Implements a shared directory system for team collaboration while maintaining security boundaries.
- Includes comprehensive tutorials and predictive analysis examples as reference material for model development.

Data Source Connectivity

- Primary storage platform: HPE GreenLake for File Storage, chosen for enterprise-grade reliability and scalability.
- Extensive external connectivity options:
  - Structured data support: Full integration with enterprise databases, including **MySQL**, **Hive**, **Snowflake**, **MSSQL**, plus support for modern data lake formats (**Delta Lake**, **Iceberg**)
  - Unstructured data capabilities: Compatible with various S3-compatible object stores, including **AWS S3**, **MinIO**, and **HPE Ezmeral Data Fabric**
  - Volume data management: Direct integration with HPE Ezmeral Data Fabric File Store and HPE GreenLake for File Storage

Security Architecture

Access Control Implementation

- Implements Role-Based Access Controls (RBACs) inherited from HPE GreenLake Private Cloud.
- Features workspace isolation technology that creates secure, user-designated environments.
- Provides enterprise-grade Single Sign-On (SSO) integration across all platform components.

Security Framework Components

- Keycloak Integration: Enterprise-grade identity and access management system implementing OIDC provider capabilities.
- Oauth2 Proxy: Security gateway for legacy applications without native OIDC support.
- Auth Token System: Sophisticated token management system for secure service-to-service communication.
- Istio Service Mesh: Advanced network security layer providing:
  - Intelligent request routing
  - Policy enforcement
  - JWT validation
  - Service-to-service communication security
- SPIFFE Implementation: The secure service identity framework is managed through Istio and SPIRE integration.

Monitoring and Management Capabilities

Observability Infrastructure

- Comprehensive monitoring system covering applications and core services
- Implements Prometheus for metric collection across all system components
- Supports external telemetry integration through OTEL exporter functionality
- Provides real-time alerting and detailed logging capabilities

Model Monitoring Systems

- Implements dual-layer monitoring approach:
  - Operational monitoring through KServe and MLflow for deployment metrics
  - Functional monitoring via whylogs for model performance and accuracy tracking
- Enables continuous evaluation of model reliability and alignment with business objectives

Administrative Functions

- Supports custom framework import through both UI and API interfaces
- Provides complete application lifecycle management capabilities
- Enables granular user access control and permission management
- Features sophisticated data source connection management with security controls
- Allows administrators to configure and customize included applications through the user interface

Solution components

WARNING

The listed versions are the minimum required; do not downgrade, modify, or replace components outside of HPE-approved update processes.

The Private Cloud AI features the following components:

Entity	Component Details	
	Small, Medium and Large	Developer System
Software	Built on the HPE GreenLake Cloud Platform, featuring HPE AI Essentials with HPE NVIDIA AI Enterprise.	
Operating System	Red Hat Enterprise Linux (RHEL) OS (ver.8.10).	
Virtualization	• 1.5: VM Essentials 8.0.8-1 • 1.4.1: VM Essentials 8.0.6.3 • 1.4: VM Essentials 8.0.3.2 • 1.3: VM Essentials 8.0.3.2 • 1.2: VM Essentials 8.0.3.2 • 1.0 and 1.1.x: VMware vCenter Server (ver.8.0 U2 March2024 - No bare metal support currently) and VMware ESXi (ver.8.0.2)	
Data Service Connectors	• 1.5: Script version 103.0.1.0 • 1.4.1: Script version 102.0.5.3 • 1.4: Script version 102.0.0.1 • 1.3: Script version 102.0.0.1 • 1.2: Script version 102.0.0.1 • 1.0 and 1.1.x: Script version 4.3.5.1	
Storage	HPE GreenLake for File Storage with a minimum capacity of at least 109 TB. The version of the File Storage depends on the Private Cloud AI version.	Local NFS with 32TB of internal file and object storage.
Network	Minimum of 2 HPE SN4600cM 32-port switches (ver.11.2008.3328) with at least 100GbE each.	No switches provided
Infrastructure	Contains at least 3 Control nodes on HPE ProLiant DL325 servers, 2 Aruba 6300M Out-Of-Band (OOB) management interfaces (ver.10.12.1030), and 8kW of power per rack.	1 control node consisting of an HPE ProLiant DL325 server with 2 Xeon 32-core CPUs and up to 2.2 kW of power per rack.
Hardware	Robust HPE ProLiant DL380a Gen11 servers with at least 4 NVIDIA L4OS GPUs and 2 NVIDIA ConnectX-7 400 Gbps Infiniband adapter cards.	One HPE ProLiant DL380a Gen11 servers with 2 NVIDIA H100NVL GPUs.

NOTE

The Developer System is available starting with HPE Private Cloud AI 1.4.

IMPORTANT

For firmware and software compatibility information, refer to the Firmware and Software Compatibility Matrix.

- For Private Cloud AI version 1.0 - Click **here**
- For Private Cloud AI version 1.1 - Click **here**
- For Private Cloud AI version 1.2 - Click **here**
- For Private Cloud AI version 1.3 - Available soon
- For Private Cloud AI version 1.4 - Click **here**
- For Private Cloud AI version 1.4.1 - Click **here**
- For Private Cloud AI version 1.5 - Click **here**
- For Private Cloud AI version 1.5 (Developer System) - Click **here**

Subtopics

Optimum environment

To provide optimum performance with minimum maintenance for your rack environment, you must meet specific requirements for airflow, power, space, temperature, and firewall and ports.

Subtopics

Airflow requirements

HPE rack-mountable products draw in cool air through the front of the rack and expel warm air through the rear. You must ensure:

- The front door has sufficient ventilation to allow ambient room air to enter.
- The rear door has adequate ventilation to let warm air escape.
- You maintain clear, unobstructed ventilation apertures throughout the rack.

Power requirements

When planning power distribution for your rack configuration, you must ensure:

- The power load is evenly balanced across available AC supply branch circuits
- The total system AC load does not exceed 80% of the branch circuit's rated AC
- For installations using a UPS, the load must not exceed 80% of the UPS's marked electrical current rating

Licensed electricians must install this equipment according to local and regional IT installation regulations. The installation must comply with:

- The National Electric Code (ANSI/NFPA-70, 1993)
- The Code for Protection of Electronic Computer/Data Processing Equipment (NFPA-75, 1992)

For electrical power ratings of optional components, consult either:

- The product's rating label
- The user documentation provided with that option

Space requirements

When choosing a location for your rack:

A clearance of at least 1,219 mm (48 inches) is required:

- Around the entire pallet and above the rack to allow for the removal of packing materials
- In front of the rack to ensure the door can fully open

Additionally, you need:

- At least 762 mm (30 inches) of clearance behind the rack for component access
- At least 380 mm (15 inches) of space around the power supply for servicing

Temperature requirements

To maintain safe and reliable operation, you must install and position the rack in a well-ventilated, climate-controlled environment.

Remember that the operating temperature of the rack consistently exceeds the room temperature and varies based on the equipment configuration. Before installation, verify the TMRA (maximum allowable operating temperature) for each piece of equipment.

CAUTION

When installing third-party options, take these precautions to reduce the risk of equipment damage:

- Ensure optional equipment does not restrict airflow around the system or raise the internal rack temperature beyond the maximum allowable limits.
- Keep the temperature within the manufacturer's specified TMRA.

Firewall and port requirements

To understand firewall and port requirements for the HPE Private Cloud AI solution, contact your deployment manager or HPE representative after ordering the solution but before the cluster is deployed.

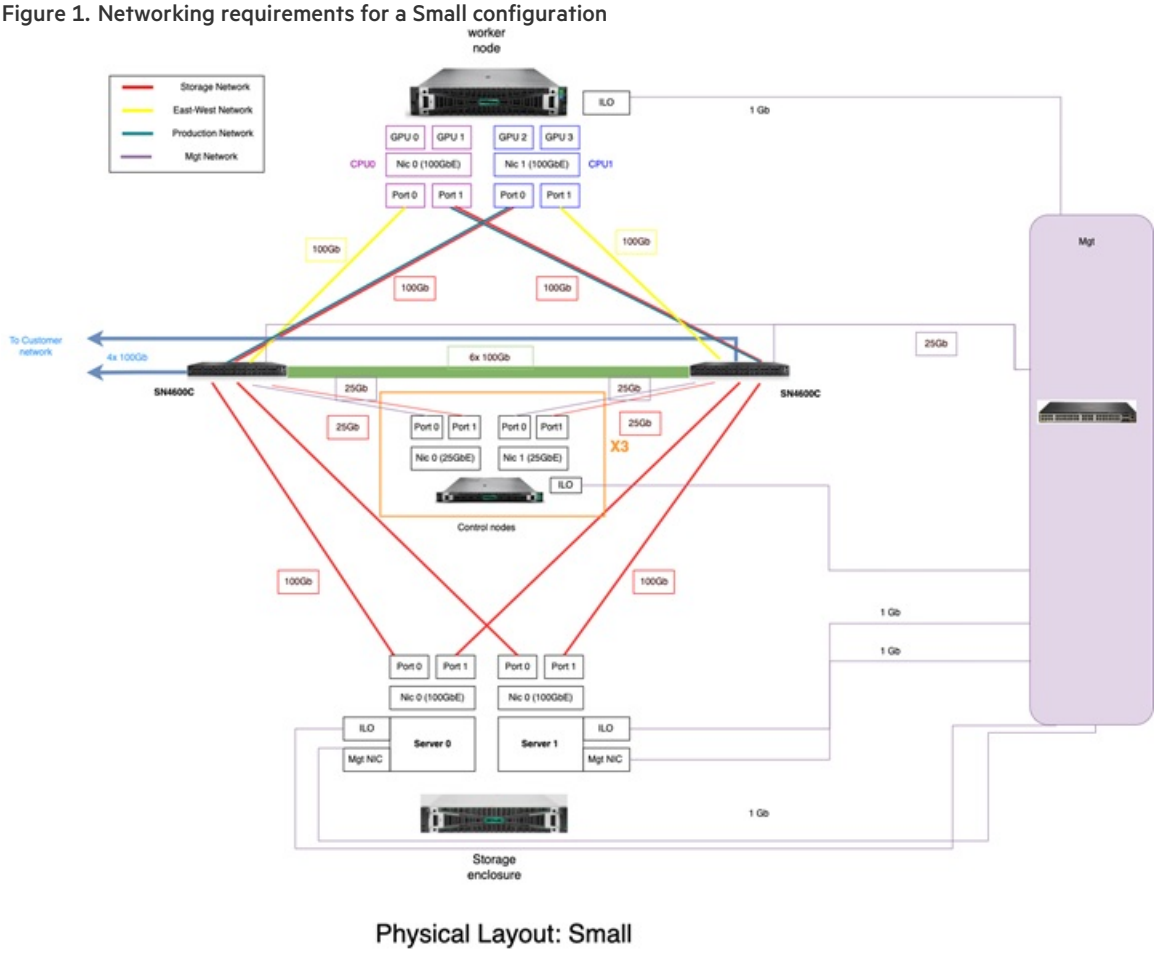
Networking requirements

The following diagrams show networking details for the small, medium, and large configurations of the HPE Private Cloud AI solution.

For Developer System networking information, see the **HPE Private Cloud AI Developer System Installation Guide**.

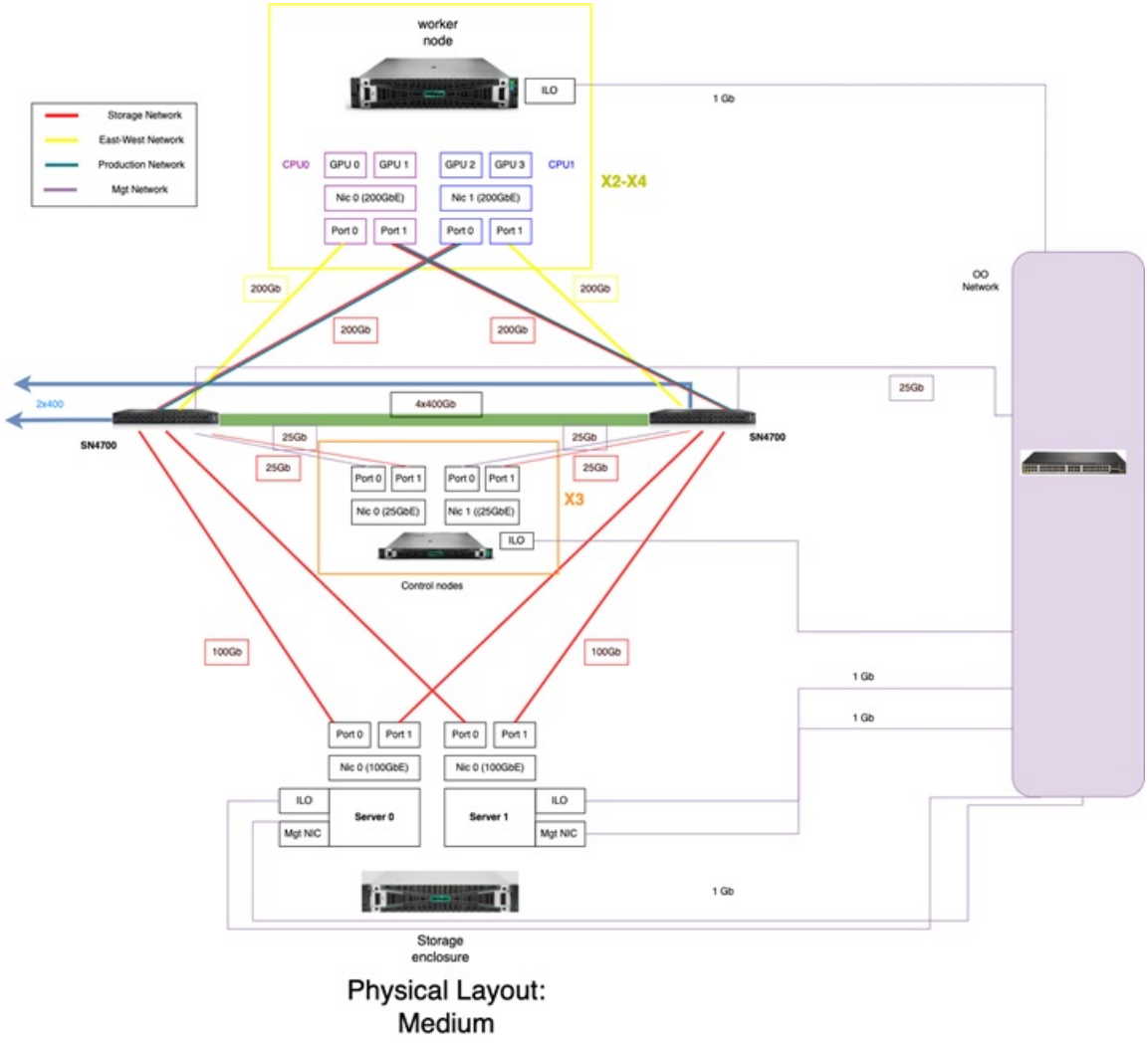
Small configuration

Figure 1. Networking requirements for a Small configuration



Medium configuration

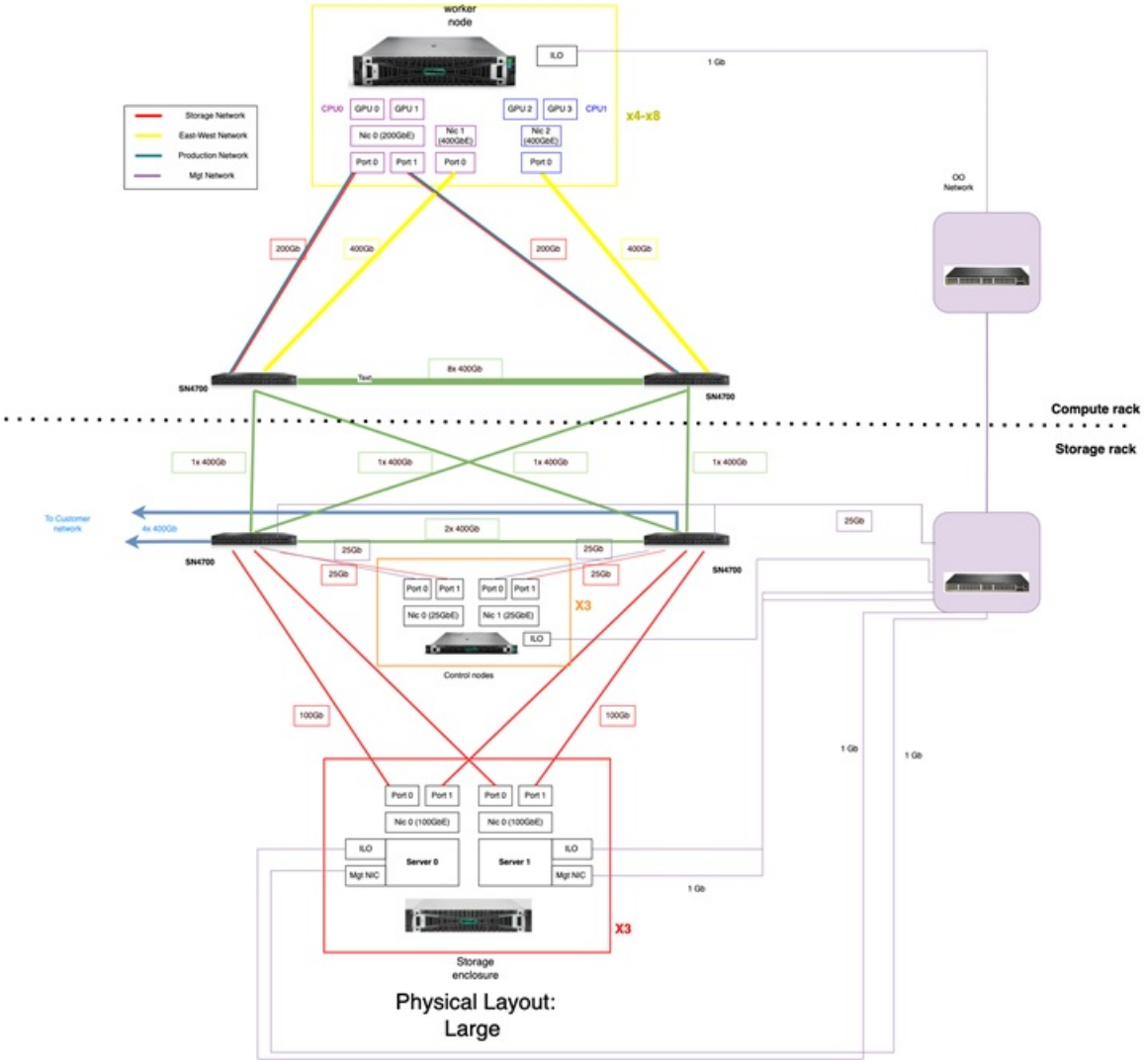
Figure 2. Networking requirements for a Medium configuration



Physical Layout: Medium

Large configuration

Figure 3. Networking requirements for a Large configuration



Network bandwidth and software updates

HPE Private Cloud AI requires a sustained Internet connection of at least 100 Mbps to ensure that software updates finish in hours and not days. For example, at 100 Mbps, the pre-staging step for HPE AI Essentials can take up to eight hours if 100% of the software bits need to be updated. In practice, it generally takes about three or four hours as only a subset of the components usually needs to be updated.

Licensing requirements

The following table describes how you receive and apply license information for the software components that make up the HPE Private Cloud AI solution:

Product	How to receive the license information	How to apply the license information
HPE VM Essentials	HPE sends an email receipt with the order details.	No action is required.

Product	How to receive the license information	How to apply the license information
Red Hat Enterprise Linux (RHEL)	HPE sends an email receipt with the order details and instructions to register the products. The email is sent to the contact on the Purchase Order.	1. The customer activates the purchased products in the My HPE Software Center. 2. HPE sends an email confirmation of successful registration within the My HPE Software Center. 3. The customer completes the assignment online by logging into their account at **https://redhat.com**. After logging in, click Manage Accounts > Subscriptions > Subscription Inventory > Subscription Activation. 4. Enter the Subscription Activation number that was provided in the HPE registration. 5. Service personnel log in to the worker node OS. 6. Run `subscription-manager register`. 7. The customer enters their Red Hat login credentials. 8. Run `subscription-manager attach --auto` or alternatively manually specify a pool.
HPE AI Essentials	HPE sends an email receipt with the order details and instructions to register the products. The email is sent to the contact on the Purchase Order.	After completing the "HPE Private Cloud AI Setup wizard" step for initial deployment: 1. HPE TC launches the AI Essentials application. 2. The first time the application is launched, the platform ID is displayed to the HPE TC. 3. The HPE TC provides the customer with the platform ID either directly or via the IPM. 4. The customer goes to the HPE Software Center and entitles both the Ezmeral vCPU and Ezmeral vGPU licenses. 5. HPE sends an email confirmation to the customer of successful registration within the My HPE Software Center. 6. The customer provides HPE the license key files. 7. The installer drags/drops the file for the Ezmeral vCPU into the application. Then they open the AI Essentials application. 8. Once in the application, they apply the Ezmeral vGPU license by clicking Administration > Settings > Activation Key > Upload Activation Key. Select the file for the Ezmeral vGPU license, and upload it.

Product	How to receive the license information	How to apply the license information
GreenLake for File with Object Storage Enabled	HPE sends an email receipt with the order details and instructions to register the products. The email is sent to the contact on the Purchase Order.	1. The customer activates the purchased products in the My HPE Software Center. 2. HPE sends the customer an email confirmation of successful registration within the My HPE Software Center. 3. The customer must add the subscription license to their HPE GreenLake workspace by using the **Adding a customer-owned device to HPE GreenLake** instructions. 4. They need to use Product Number S1F24A
NVIDIA AI Enterprise	See Activating the NVIDIA software subscription.	See Activating the NVIDIA software subscription.
HPE Private Cloud AI Platform	The subscription is provided through GreenLake Cloud.	No action is required.
OpsRamp Enterprise	The subscription is provided through GreenLake Cloud.	No action is required.

NOTE

If the AI Essentials license expires, existing workloads will continue to run without interruption as long as they remain active. Users with valid login sessions on AIE Keycloak can maintain their current workloads. However, the license check is enforced when attempting to create new workloads for scheduling on worker nodes, when pods are restarted, or when a node is rebooted. In these scenarios, new workload creation will be blocked until a valid license is restored.

End user license agreement

For information about the end user license agreement (EULA) and additional license authorizations (ALAs), see the **HPE End User License Agreement** page.

Activating the NVIDIA software subscription

The following table describes adding the NVIDIA license information for various Private Cloud AI configurations.

Private Cloud AI Configuration	How to receive the license information	How to apply the license information
Small and Medium Configuration (NVIDIA L40S)	HPE sends an email receipt with order details and instructions to register the product. The email is sent to the contact on the Purchase Order when you order Private Cloud AI.	1. The customer activates the purchased products in My HPE Software Center. 2. HPE sends an email confirmation to the customer of successful registration within My HPE Software Center. 3. NVIDIA also sends the customer an email with instructions for how to register. 4. The customer must register on the NVIDIA website to receive NVIDIA support. 5. No additional on-prem licensing is required.
Developer and Large Configuration (NVIDIA H100 NVL)	The customer does not receive any notification.	1. Log into the NGC account at ngc.nvidia.com/signin 2. Navigate to Organization 3. Select Activate Subscription 4. Complete the company information form 5. Enter the GPU serial numbers 6. Click Activate Subscription 7. Review information in the pop-up window 8. Click Request Subscription 9. Wait for approval (up to 48 hours) 10. Access the Enterprise Catalog from the left menu 11. Download the NVIDIA AI Enterprise software 12. Watch for the NVIDIA email (within 24 hours) 13. Create your Enterprise Support account using the email link 14. Access the licensing portal through the email link

Solution configurations

WARNING

The listed versions are the minimum required; do not downgrade, modify, or replace components outside of HPE-approved update processes.

The HPE Private Cloud AI solution is available in the following configurations:

Table 1. Private Cloud AI Configurations

Entity	Developer System	Small	Medium	Large
Use Case	AI Sandbox	Inferencing	Inferencing + RAG (Retrieval-Augmented Generation)	Inferencing + RAG + Fine-Tuning

Entity	Developer System	Small	Medium	Large
Compute	2 **NVIDIA H100 NVL** on a **HPE ProLiant DL380a Gen11 server**	4-8 **NVIDIA L40S** on a **HPE ProLiant DL380a Gen11 server** with service pack **ver.2024.04.00.00**	8-16 **NVIDIA L40S** on a **HPE ProLiant DL380a Gen11 server** with service pack **ver.2024.04.00.00**	16 or 32 **NVIDIA H100 NVL** on a **HPE ProLiant DL380a Gen11 server** with service pack **ver.2024.04.00.00**
Networking	2x 200-GB ports	2 **100GbE CX7, 32-port NVIDIA SN4600M**(ver.11.2008.3328	2 **200GbE CX7, 32-port NVIDIA SN4700N**(ver.11.2008.3328)	2 **400GbE CX7, 32-port NVIDIA SN4700N** (ver.11.2008.3328)
Storage	Local NFS with 32TB of internal file and object storage.	**HPE GreenLake for File Storage (Standard Density aka Razor Crest)** (ver. 3.1) 109-529 TB **Upgrades:** • 109 TB to 248 TB (adding 1x JBOF) • 109 TB to 529 TB (adding 3x JBOFs)	**HPE GreenLake for File Storage (Standard Density aka Razor Crest)** (ver. 3.1) 217 TB-1.088 PB **Upgrades:** • **For 8 GPU configuration:** ◦ 217 TB to 497 TB (adding 1x JBOF) ◦ 217 TB to 1060 TB (adding 3x JBOFs) • **For 16 GPU configuration:** ◦ 217 TB to 390 TB (adding 1xCNode and 2x JBOFs) ◦ 217 TB to 529 TB (adding 1xCNode and 3x JBOFs) • 217 TB to 1088 TB (adding 1xCNode and 7x JBOFs)	**HPE GreenLake for File Storage (Standard Density aka Razor Crest)** (ver. 3.1) 500TB-1PB
Management/Control	1 control node on HPE **ProLiant DL325 servers**	3 control nodes on HPE **ProLiant DL325 servers**, 2 **Aruba 6300M OOB management switches** (ver.10.12.1030)	3 control nodes on HPE **ProLiant DL325 servers**, 2 **Aruba 6300M OOB management switches** (ver.10.12.1030)	3 control nodes on HPE **ProLiant DL325 servers**, 2 **Aruba 6300M OOB management switches** (ver.10.12.1030)
Power per rack	~2.2kW (x 1 rack)	~8kW (x 1 rack)	~18kW (x 1 rack)	~25kW (x 1 rack)
NVIDIA AI Integration	• **Models: Mistral-7B**, **QA E5**, **Llama 3 8B and Llama 3 70B** • **NVIDIA Inference Microservice (NIM)** for embeddings (RAG pipeline support) • NIM **Mistral 4B** for reranking • **LORA adapters for model fine-tuning**			
Community AI Models and Repositories	**Aleph Alpha**, **Hugging Face**			
Platform	Managed through the HPE GreenLake Cloud compromising of HPE AI Essentials with HPE NVIDIA AI Enterprise.			
OS	**Red Hat Enterprise Linux (RHEL) OS** (ver.8.10)			

Entity	Developer System	Small	Medium	Large
Virtualization	VM Essentials			
Data Service Connectors	HPE-DSC VM with equivalent KickStart Scripts			
Data Engineering Tools	• **Apache Airflow** to orchestrate workflows • **Apache Spark** to process data • **EzPresto** to query data • **KServe** to deploy and serve ML models on Kubernetes • **Feast** to define, manage, validate, and serve features for production AI/ML			
Data Analytics Tools	• **Jupyter** notebooks • **Apache Superset** for BI visualization • Custom interface for Spark job submission			
Data Science Tools	• **MLflow** for experiment tracking • **Ray** for distributed computing • **Kubeflow** for notebooks			

Subtopics

Configuration sizes

Private Cloud AI is available in the following configuration sizes. Each configuration provides support for increasing levels of Generative AI Inference, Retrieval Augmented Generation (RAG), and Model Fine tuning use cases.

Developer System

Target Workloads: AI Sandbox

Control Plane
1x DL325 Gen 11 Node

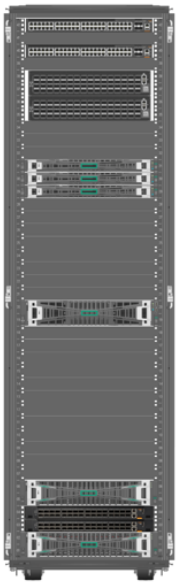
AI Worker Node
1x DL380a Gen 11 AI Optimized Node
- 2x H100 NVL GPUs

Storage
32TB of internal file and object storage
Local NFS support for VM and container storage
Local MinIO for object storage

SW Licensing
HPE AI Essentials with NVIDIA AI Enterprise Software
- 3 Year Subscription

Small

Target Workloads: Generative AI Inference



Small

Network Switches
2 x Nvidia SN4600cM Switches (100GbE data network)
2 x Aruba 6300M (Management & ILO)

Control Plane
3x DL325 Gen 11 Nodes

AI Worker Nodes
1x DL380a Gen 11 AI Optimized Node
- 4x L40s GPUs per node (4 total)

Storage
109TB GreenLake for File Storage
Based on Alletra MP in Standard Density
1c x 1d x 2s configuration with 7.68TB Drives

SW Licensing
HPE AI Essentials with Nvidia AI Enterprise Software
- 3 Year Subscription

Expanded Small

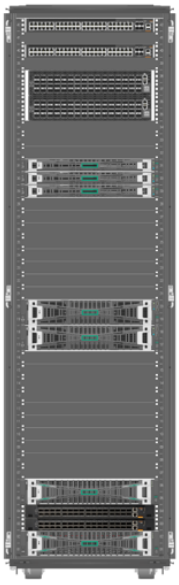
Network Switches
2 x Nvidia SN4600cM Switches (100GbE data network)
2 x Aruba 6300M (Management & ILO)

Control Plane
3x DL325 Gen 11 Nodes

AI Worker Nodes
2 x DL380a Gen 11 AI Optimized Node
- 4x L40s GPUs per node (8 total)

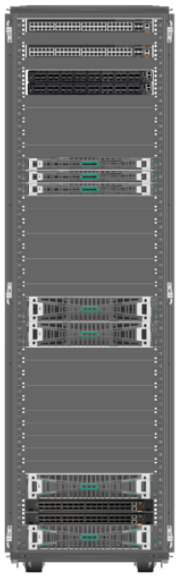
Storage
109TB GreenLake for File Storage
Based on Alletra MP in Standard Density
1c x 1d x 2s configuration with 7.68TB Drives

SW Licensing
HPE AI Essentials with Nvidia AI Enterprise Software
- 3 Year Subscription



Medium

Target Workloads: Gen AI Inference, Retrieval Augmented Generation (RAG)



Medium

Network Switches
2 x Nvidia SN4700M Switches (400GbE data network)
2 x Aruba 6300M (Management & ILO)

Control Plane
3x DL325 Gen 11 Nodes

AI Worker Nodes
2x DL380a Gen 11 AI Optimized Node
- 4x L40s GPUs per node (8 total)

Storage
217 TB GreenLake for File Storage
Based on Alletra MP in Standard Density
1c x 1d x 2s configuration with 15 TB Drives

SW Licensing
HPE AI Essentials with Nvidia AI Enterprise Software
- 3 Year Subscription

Expanded Medium

Network Switches
2 x Nvidia SN4700M Switches (400GbE data network)
2 x Aruba 6300M (Management & ILO)

Control Plane
3x DL325 Gen 11 Nodes

AI Worker Nodes
4 x DL380a Gen 11 AI Optimized Node
- 4x L40s GPUs per node (16 total)

Storage
217 TB GreenLake for File Storage
Based on Alletra MP in Standard Density
1c x 1d x 2s configuration with 15TB Drives

SW Licensing
HPE AI Essentials with Nvidia AI Enterprise Software
- 3 Year Subscription

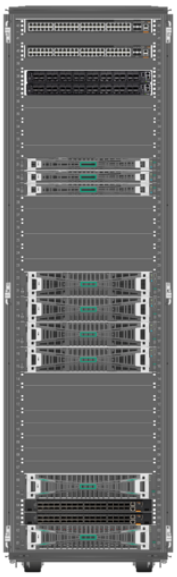
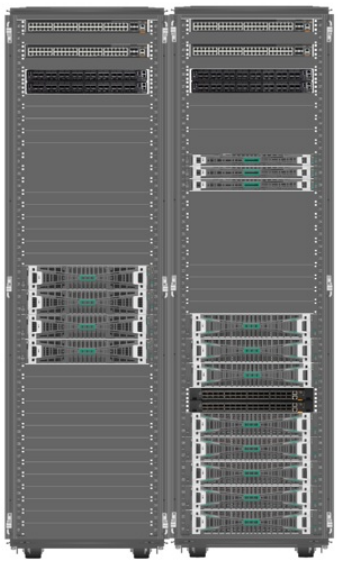


Figure 1. Private Cloud AI Large configuation size

Large

Target Workloads: Gen AI Inference, Retrieval Augmented Generation (RAG), Model Fine tuning



Large

Network Switches
4 x Nvidia SN4700M Switches
(400GbE data network)
4 x Aruba 6300M (Management & ILO)

Control Plane
3x DL325 Gen 11 Nodes

AI Worker Nodes
4x DL380a Gen 11 AI Optimized Node
- 4x H100 NVL GPUs per node (16 total)

Storage
670 TB GreenLake for File Storage
Based on Alletra MP in Standard Density
3c x 5d x 2s configuration with 7.68TB Drives

SW Licensing
HPE AI Essentials with Nvidia AI Enterprise Software
- 3 Year Subscription

Expanded Large

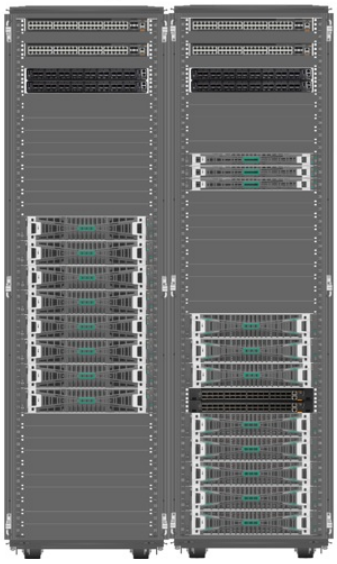
Network Switches
4 x Nvidia SN4700M Switches
(400GbE data network)
4 x Aruba 6300M (Management & ILO)

Control Plane
3x DL325 Gen 11 Nodes

AI Worker Nodes
8 x DL380a Gen 11 AI Optimized Node
- 4x H100 NVL GPUs per node (32 total)

Storage
670 TB GreenLake for File Storage
Based on Alletra MP in Standard Density
3c x 5d x 2s configuration with 7.68TB Drives

SW Licensing
HPE AI Essentials with Nvidia AI Enterprise Software
- 3 Year Subscription



Private Cloud AI User roles

Private Cloud AI supports three configurable user roles: **Private Cloud AI Cloud Administrator**, **Private Cloud AI Administrator**, and **Private Cloud AI User**.

- The **Private Cloud AI Cloud Administrator** performs cloud infrastructure administrative tasks such as:
  - **Infrastructure Administration**
    - Performs initial platform setup and configuration of all core components and services
    - Manages comprehensive user onboarding processes including access provisioning and account setup

        - Continuously monitors system health, performance metrics, and overall resource usage patterns
        - Handles infrastructure expansion through node upgrades and capacity planning
        - Implements and oversees system and software upgrade processes
    - **AI-Specific Administration**
        - Establishes and maintains data Role-Based Access Controls (RBACs), configures data volumes, and sets up data connections
        - Manages the installation and integration of additional AI tools and frameworks
        - Implements GPU resource prioritization strategies to optimize AI/ML workload performance
- The **Private Cloud AI Administrator** performs AI performance and workload management tasks such as:
    - Managing AI Essentials Software resources
    - Managing AI model deployment and versioning
    - Monitoring AI performance and fine-tuning models and GPU resource allocation
    - Ensuring AI system compliance and ethical use
- The **Private Cloud AI User**: develops and deploys AI applications. Users with this role are typically researchers and data scientists who:
    - Maintain self-service access to various AI tools, frameworks, and development environments, such as HPE AI Essentials Software
    - Develop and optimizes Retrieval-Augmented Generation (RAG) pipelines for enhanced AI applications
    - Perform model fine-tuning to adapt pre-trained models for specific use cases and requirements
    - Manage the deployment, scaling, and monitoring of AI models in production environments
    - Collaborate with platform administrators to optimize resource utilization and performance

Signing in to the HPE GreenLake platform and launching Private Cloud AI

Prerequisites

As part of the Private Cloud AI installation process, your HPE Support representative will:

- Create an HPE GreenLake User Account on your behalf (if needed)
- Create an HPE GreenLake workspace
- Add Data Services Cloud Console to your workspace
- Add the Private Cloud AI service to Data Services Cloud Console
- Assign user roles

After the Private Cloud AI solution is installed, users with the Private Cloud AI Cloud Administrator role can also complete these operations. For more information, see Getting started with HPE GreenLake.

Procedure

1. Go to **https://common.cloud.hpe.com/**.
2. Enter your login credentials.

   The Welcome to HPE GreenLake page appears. Here, you can create a workspace or sign in to an existing workspace.
3. Locate the workspace associated with your HPE GreenLake account and click Launch.
4. Click My Services.

5. Locate the Private Cloud AI tile and click Launch.

   The areas of the Private Cloud AI user interface that you can access when signing in will depend on the Private Cloud AI user role you have been assigned.