Hewlett Packard Enterprise

HPE Private Cloud AI Administration Guide

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

HPE Private Cloud AI 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

Table of contents

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

Private Cloud AI Release Notes

Read the **Private Cloud AI release notes** to learn about what's new in this release, including new features, enhancements, bug fixes, known issues, workarounds, performance improvements, breaking changes, and security updates.

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 (PCAI), 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

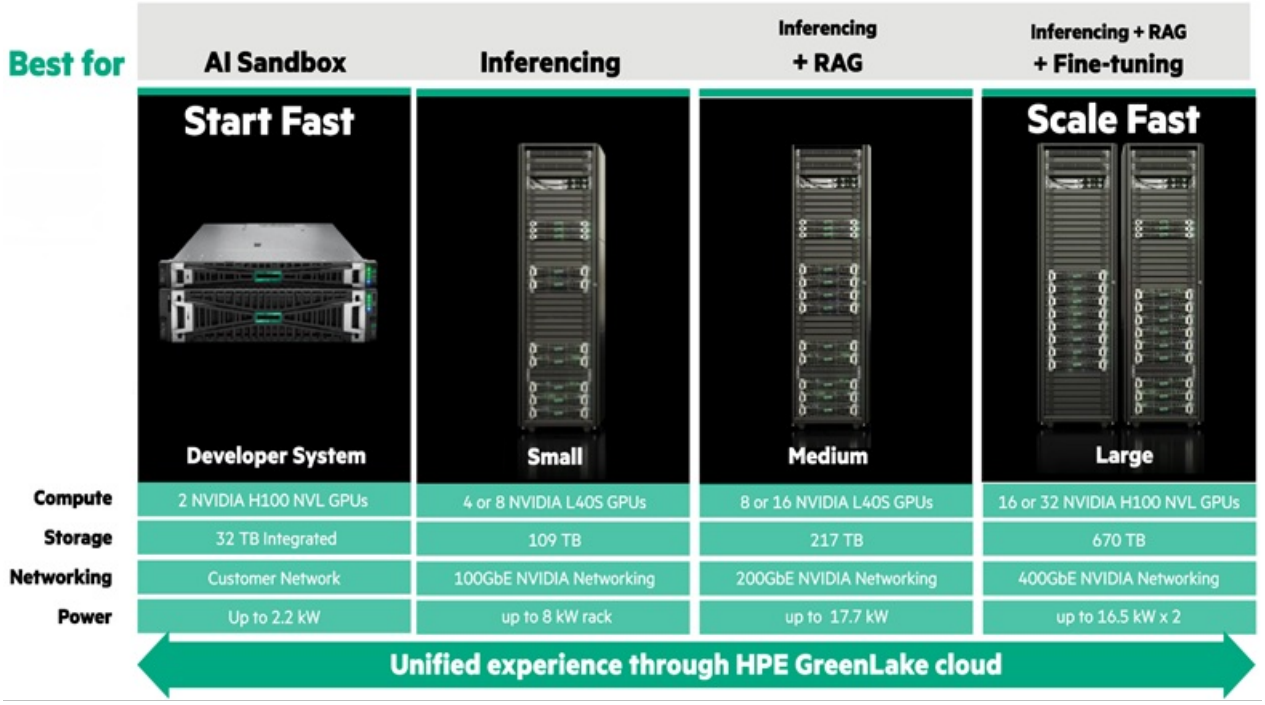
The Private Cloud AI solution stack includes robust hardware, AI-native computing, storage, and networking components, all of which are accessible and manageable through the GreenLake Cloud platform.

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

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 version 1.4.1 - Click **here**
- For Private Cloud version 1.5 - Available soon

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.

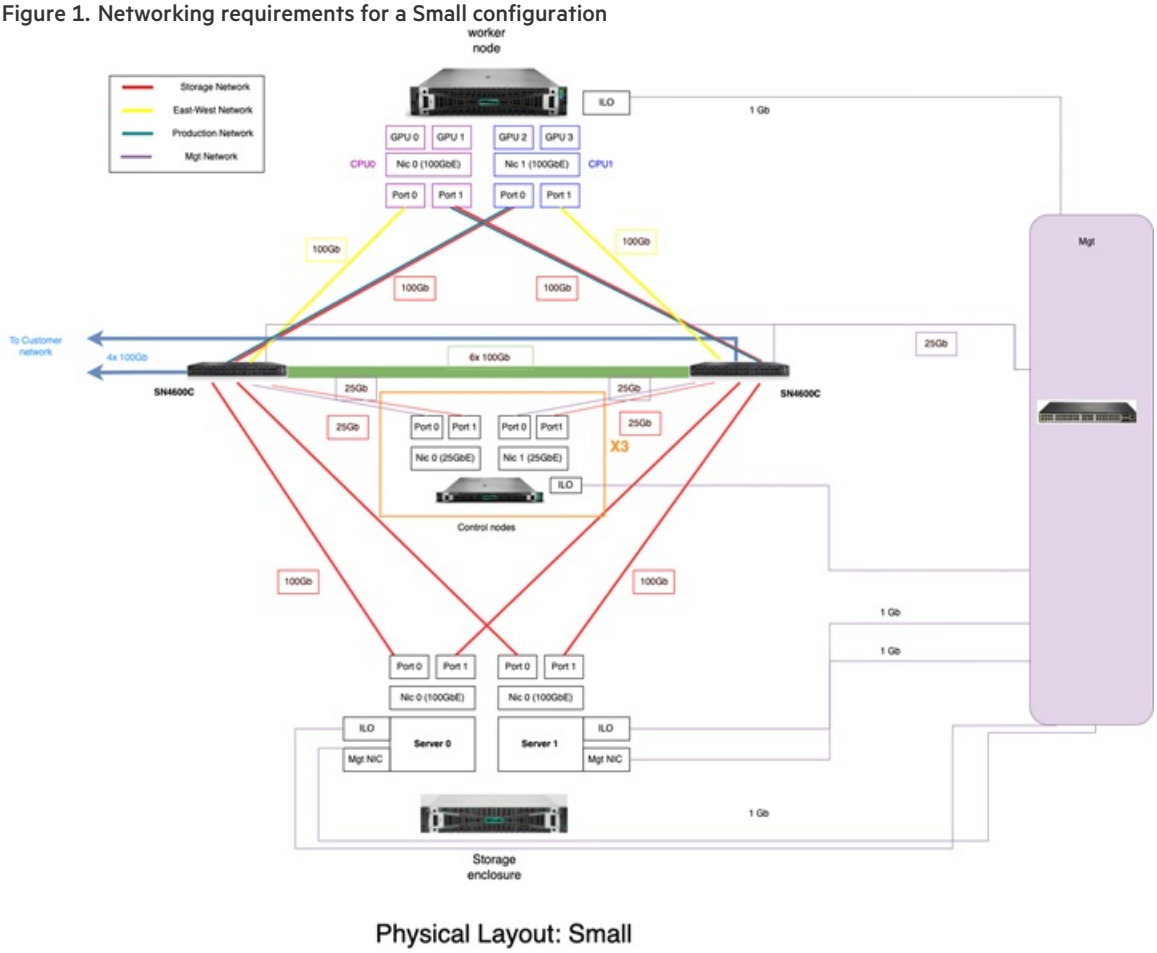
Networking requirements

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**.

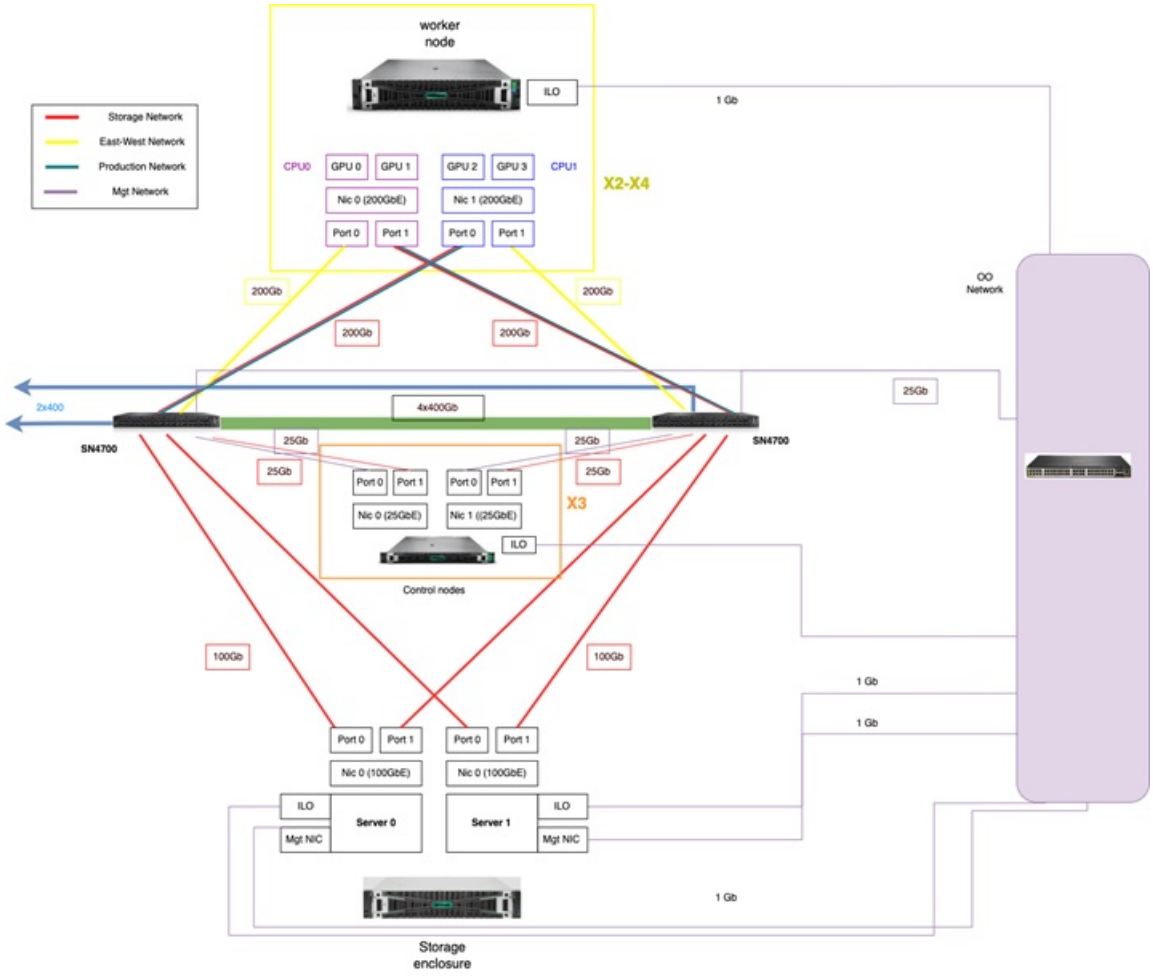
Small configuration

Figure 1. Networking requirements for a Small configuration



Medium configuration

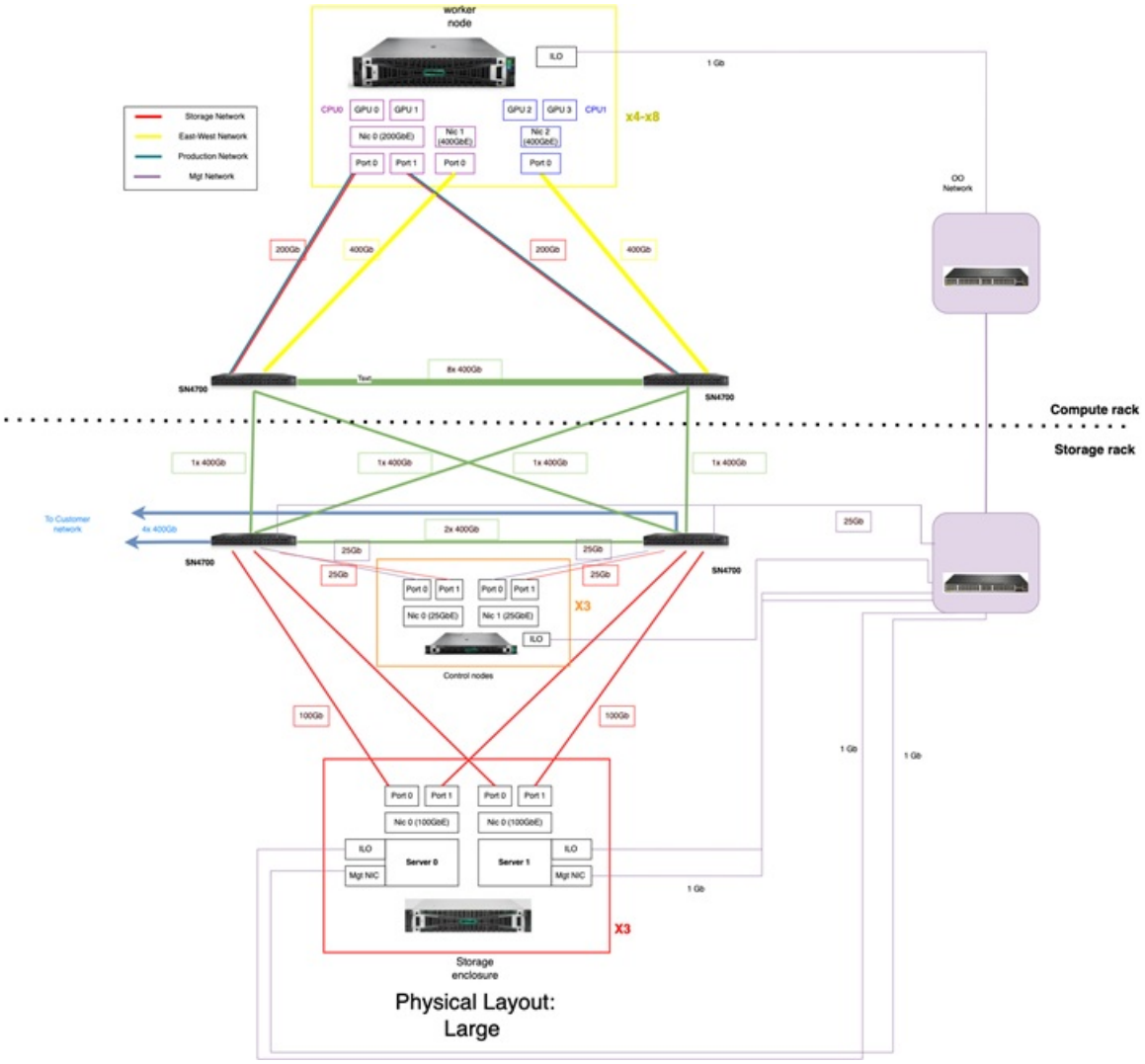
Figure 2. Networking requirements for a Medium configuration



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.

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

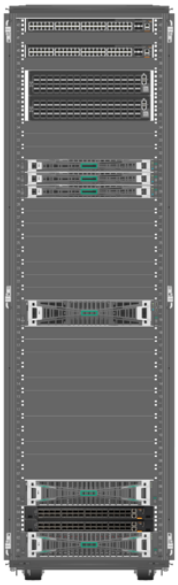
AI Worker Node
1x DL380a Gen 11 AI Optimized Node
- 2x H100 NVL GPUs

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

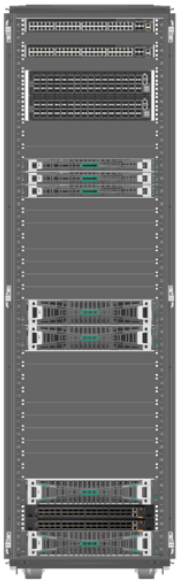
Network Switches
2 x Nvidia SN4600cM Switches (100GbE data network)
2 x Aruba 6300M (Management & ILO)

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)

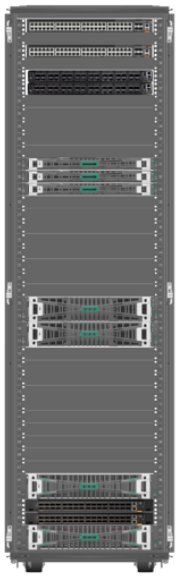
Storage
109TB GreenLake for File Storage
Based on Alletra MP in Standard Density
1c x 1d x 2s configuration with 7.68TB Drives

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

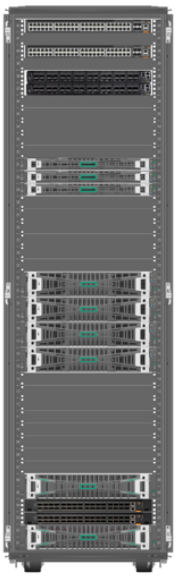
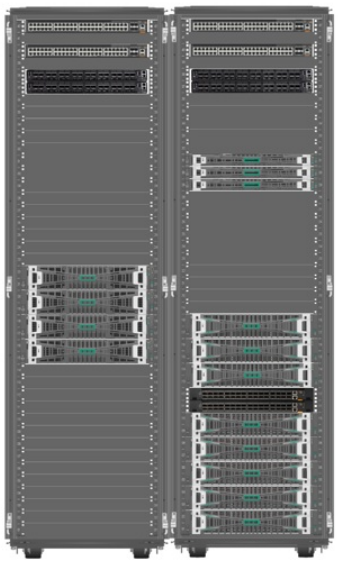


Figure 1. Private Cloud AI Large configuation size

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

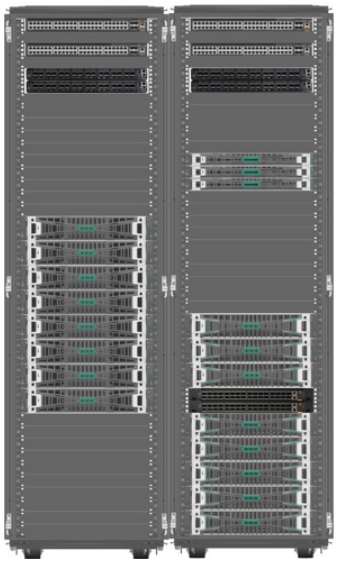
Network Switches
4 x Nvidia SN4700M Switches
(400GbE data network)
4 x Aruba 6300M (Management & ILO)

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.

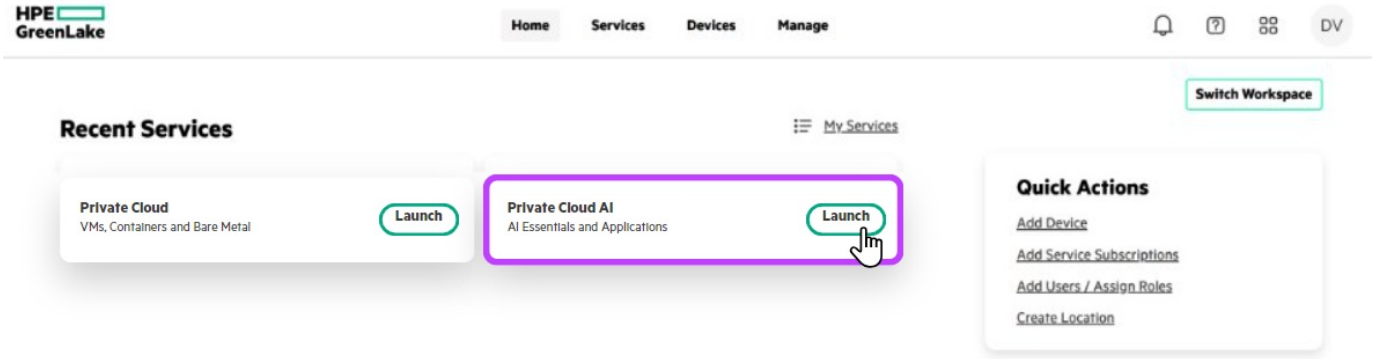
Procedure

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.



AI User: Developing AI Applications

Private Cloud AI dashboard enables users to implement various AI capabilities across multiple domains, including data processing, computer vision, NLP, speech technologies, and financial services. AI users typically engage in tasks such as data analysis and preparation, AI/ML experimentation, NVIDIA NIM integration, model deployment, and monitoring. The platform supports MLOps integration with CI/CD pipelines, artifact management, and automated testing for comprehensive AI development and deployment.

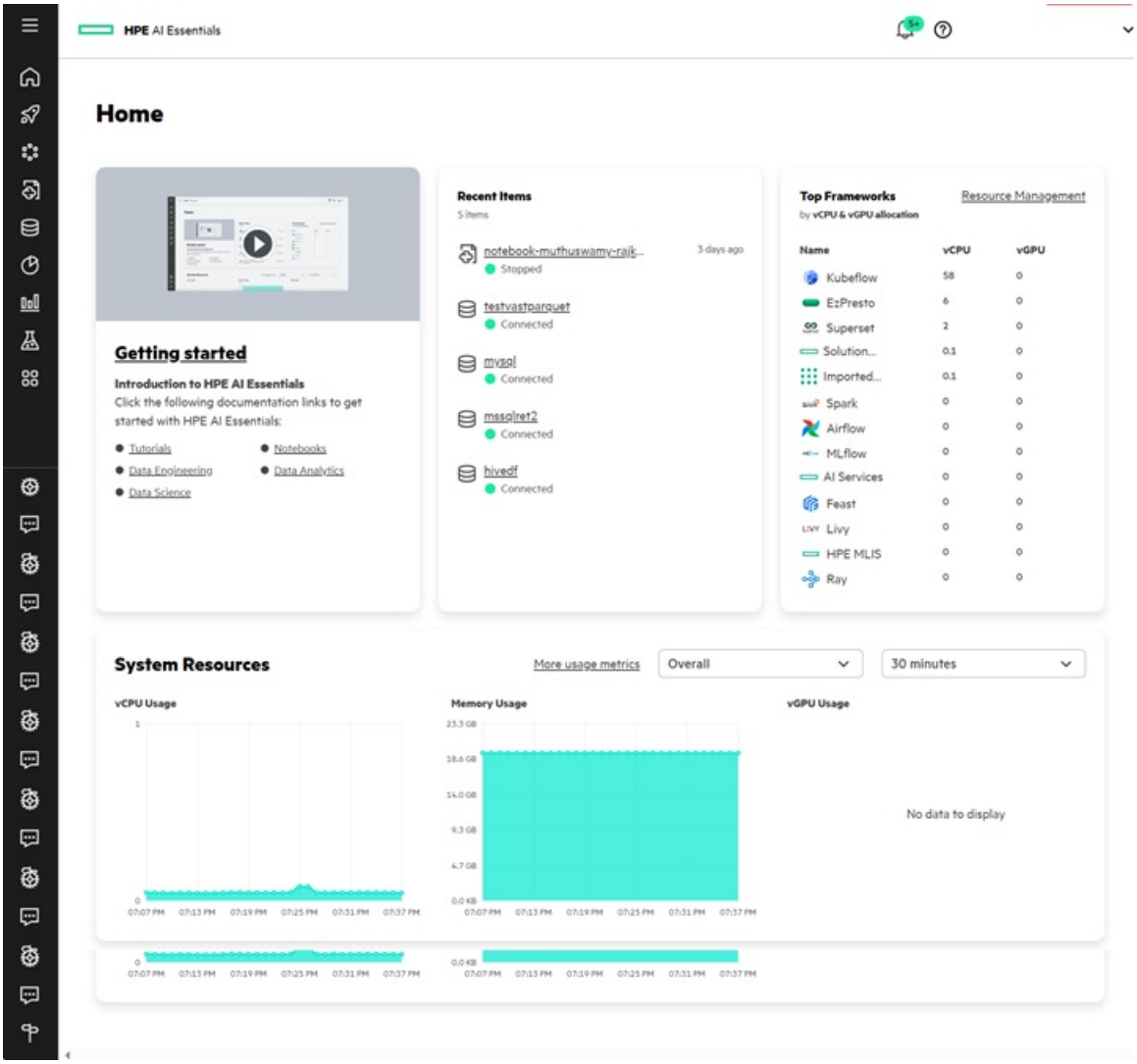
Subtopics

HPE AI Essentials Software Home Page
Additional resources for developing AI applications

HPE AI Essentials Software Home Page

After signing in as an AI user, and selecting an environment to work on, the system displays the HPE AI Essentials Software home page as follows:

Figure 1. AI Essentials home page



The **upper pane** displays three information tiles in real-time.

- The **Getting Started** tile provides an overview video and contains links to articles about Private Cloud AI usage. Click the play icon to watch the video..
- The **Recent Items** tile lists the status of the five most recent AI tasks and the time they were started.
- The **Top Frameworks** tile lists the top five running frameworks by their vCPU consumption.

The **lower pane** presents a live display of the system resources used by the running cluster applications. Click the **More Usage Metrics** link to select additional metrics for display. Alternatively, select the metric from the dropdown. The default refresh time is 30 minutes, but you can change it from the drop-down.

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

Subtopics

Notebook servers

Click the **Notebooks** icon to display the Notebook servers.

Notebooks provide an interactive computing environment where you can develop and run your data science applications. These files, which use the `.ipynb` extension, combine several powerful capabilities in one place:

- Write and execute code snippets
- View results immediately
- Document your methodology and findings
- Create data visualizations
- Save your work within HPE AI Essentials

The Kubeflow Notebook interface lets you run commands, analyze data, and view results in real-time. You have full control over your notebook environment - you can add new notebook servers when you need more computing resources, or remove them when they're no longer required.

What makes notebooks particularly valuable is their all-in-one nature. In a single file, you can:

- Modify your code and parameters
- Document your approach
- Display results and visualizations
- Record your conclusions and insights
- Save everything for future reference or sharing

This combination of code execution, documentation, and visualization makes notebooks an essential tool for data science work.

To create and use Notebooks, see the **HPE AI Essentials Software 1.6.0 Documentation**.

Data Engineering

Click the **Data Engineering** drop down for access to multiple data source connectors that enable seamless data virtualization. This creates a unified, controlled access point for distributed data, regardless of the underlying compute engine.

You can leverage powerful open-source tools like Apache Spark and Apache Airflow to:

- Extract data from various sources
- Transform datasets for consumption
- Manage data workflows efficiently

Here's a practical example of the workflow:

1. Use Spark to transfer data from a source such as Snowflake to HPE Ezmeral Data Fabric
2. Connect HPE AI Essentials Software to your HPE Ezmeral Data Fabric instance
3. Work with your data directly - joining, transforming, and preparing it
4. Create consumable models that users and applications can readily access

This approach simplifies data management by providing a central point of control while maintaining flexibility in how you process and transform your data. The system's architecture ensures that regardless of where your data resides, you can access and manipulate it through a consistent interface.

For more information, see the **HPE AI Essentials Software Documentation**.

Data Analytics

HPE AI Essentials Software serves as a unified platform where both data engineers and data scientists can manage their analytical workloads through:

- Apache Spark Operator for processing tasks
- Apache Livy for interactive sessions
- Apache Airflow for job scheduling

The platform includes built-in support for ACID transactions (Atomicity, Consistency, Isolation, and Durability) through Delta Lake integration. This integration leverages the Delta Transaction Protocol, an open protocol that ensures reliable transaction handling for Apache Spark applications.

Key features of Delta Lake integration:

- Compatible with all Apache Spark APIs for reading and writing data
- Stores data in Parquet format
- Maintains versioned Parquet files for data tracking
- Ensures data consistency and reliability

This architecture provides a robust foundation for data operations, combining the flexibility of Spark with the reliability of ACID-compliant transactions in a single, integrated environment.

For more information, see the **HPE AI Essentials Software Documentation** .

BI Reporting

HPE AI Essentials Software provides a unified platform where data engineers and business analysts can manage their analytical workloads and create powerful visualizations through Apache Superset. The platform supports:

- Interactive dashboards for business intelligence
- Real-time data exploration
- Scheduled report generation

The platform includes built-in data consistency and reliability support through Delta Lake integration, ensuring your BI reports always use accurate, up-to-date data. The Delta Transaction Protocol, an open protocol, guarantees data integrity for all reporting applications.

Critical features for BI reporting:

- Create interactive visualizations using Superset's intuitive interface
- Connect to multiple data sources seamlessly
- Schedule automated report updates
- Share dashboards across teams
- Version control for report templates and data sources

This architecture provides a robust foundation for business intelligence operations, combining the power of modern data storage with user-friendly visualization tools in a single, integrated environment. Teams can easily:

- Build comprehensive dashboards
- Generate regular reports
- Perform ad-hoc analysis
- Share insights across the organization
- Maintain data governance and security

For more information, see the **HPE AI Essentials Software Documentation**.

Data Science

Using performance-optimized open-source tools, data scientists can leverage multiple programming languages (Python, R, Java, and SQL) to develop, train, and deploy machine learning models. The platform provides a comprehensive environment for the entire machine-learning lifecycle:

Model Development and Training

- Conduct exploratory data analysis using interactive Notebooks
- Perform feature engineering and labeling
- Build and train models using popular frameworks such as TensorFlow, Ray, or PyTorch
- Create automated pipelines for repetitive tasks

Distributed Computing and Scaling

- Execute jobs across distributed clusters
- Scale to cloud environments using Kubeflow APIs
- Optimize model performance with AutoML using Katib and MLflow
- Select and tune hyperparameters automatically

Model Deployment and Monitoring

- Package models into containers
- Register models for serving through KServe
- Monitor for data drift, bias, and robustness
- Evaluate model performance continuously

Model Optimization

- Compare model versions
- Replace underperforming models
- Retrain models for improved accuracy
- Deploy updated models seamlessly

This integrated environment streamlines the machine learning workflow, enabling data scientists to focus on developing effective models while automating routine tasks and maintaining model quality through continuous evaluation and improvement.

For more information, see the **HPE AI Essentials Software Documentation**.

Tools and Frameworks

This page showcases a user-friendly interface that neatly organizes the various tools and frameworks for Data Engineering, Analytics, and Data Science. Each tool is presented in a card format, displaying key information such as version numbers, readiness status, brief descriptions, and endpoints.

The interface is designed for quick access to tools, with Open buttons for immediate use and a dedicated Import Frameworks option at the top right for importing additional frameworks. This layout provides a clear overview of the available data infrastructure tools while ensuring easy access to essential information about each service.

For more information, see the **HPE AI Essentials Software Documentation**.

Additional resources for developing AI applications

Here are the direct links to the relevant documentation for NVIDIA and HPE Private Cloud AI platforms:

HPE Private Cloud AI Platform

- **HPE Artificial Intelligence Solutions Overview**: **This page** provides an overview of HPE's AI solutions, covering topics such as AI infrastructure, data management, and industry-specific applications. It also highlights HPE's approach to accelerating AI adoption with integrated hardware and software solutions designed to scale AI workloads efficiently.
- **HPE AI Essentials Software Documentation** : This site offers extensive resources for getting started, managing clusters, data engineering, data analytics, and more. Explore the **HPE AI Essentials Software Documentation**.

NVIDIA AI Platform

- **NVIDIA AI Enterprise Documentation**: Access comprehensive guides, deployment instructions, and workflows for NVIDIA AI Enterprise, including installation and configuration on various platforms such as VMware vSphere, bare metal, and Red Hat OpenShift. You can explore it at **NVIDIA AI Enterprise** and read the **Quick Start Guide**.

AI Administrator: Managing the solution

An AI Administrator plays a crucial role in managing and maintaining the infrastructure, systems, and processes that support artificial intelligence operations within an organization. This multifaceted position requires a blend of technical expertise, strategic thinking, and problem-solving skills to ensure the smooth functioning of AI systems.

The AI Administrator is responsible for overseeing the entire lifecycle of AI projects, from infrastructure setup and data management to performance optimization and troubleshooting. They work at the intersection of cloud computing, data science, and IT operations, bridging the gap between AI development teams and the technical infrastructure that supports their work.

The following sections outline the key responsibilities of an AI Administrator, providing a comprehensive overview of the tasks and challenges associated with this role.

Infrastructure Management

- Implement cloud-based solutions for scalable AI operations.
- Integrate hybrid cloud and on-premise architectures for optimal resource utilization.

Data Management

- Design and maintain efficient data pipelines to support AI model training and inference.
- Enforce data quality checks to ensure the integrity of AI model inputs and outputs.

Security and Compliance

- Deploy robust security measures to protect sensitive data and AI assets.
- Implement granular access controls for AI systems and data repositories.
- Ensure compliance with relevant data protection regulations and industry standards.

Performance Optimization

- Optimize AI system performance through continuous monitoring and tuning.

- Identify and resolve performance bottlenecks in AI infrastructure.
- Scale resources dynamically to accommodate growing AI workloads.

Monitoring and Alerting

- Configure comprehensive monitoring systems for AI infrastructure and applications.
- Set up alerting mechanisms to promptly detect and respond to system anomalies.
- Track and report on key performance indicators (KPIs) for AI operations.

Deployment and Version Control

- Manage AI model versions to maintain consistency across deployments.
- Implement rollback procedures for quick recovery from failed deployments.

Troubleshooting and Resolution

- Conduct proactive system health checks to prevent potential issues.
- Perform root cause analysis on complex AI system problems.
- Develop and implement solutions to ensure continuous AI system availability.

The multifaceted role of an AI Administrator requires not only technical expertise but also the ability to maintain a holistic view of the entire AI infrastructure and its performance. This is where the concept of a centralized dashboard becomes crucial.

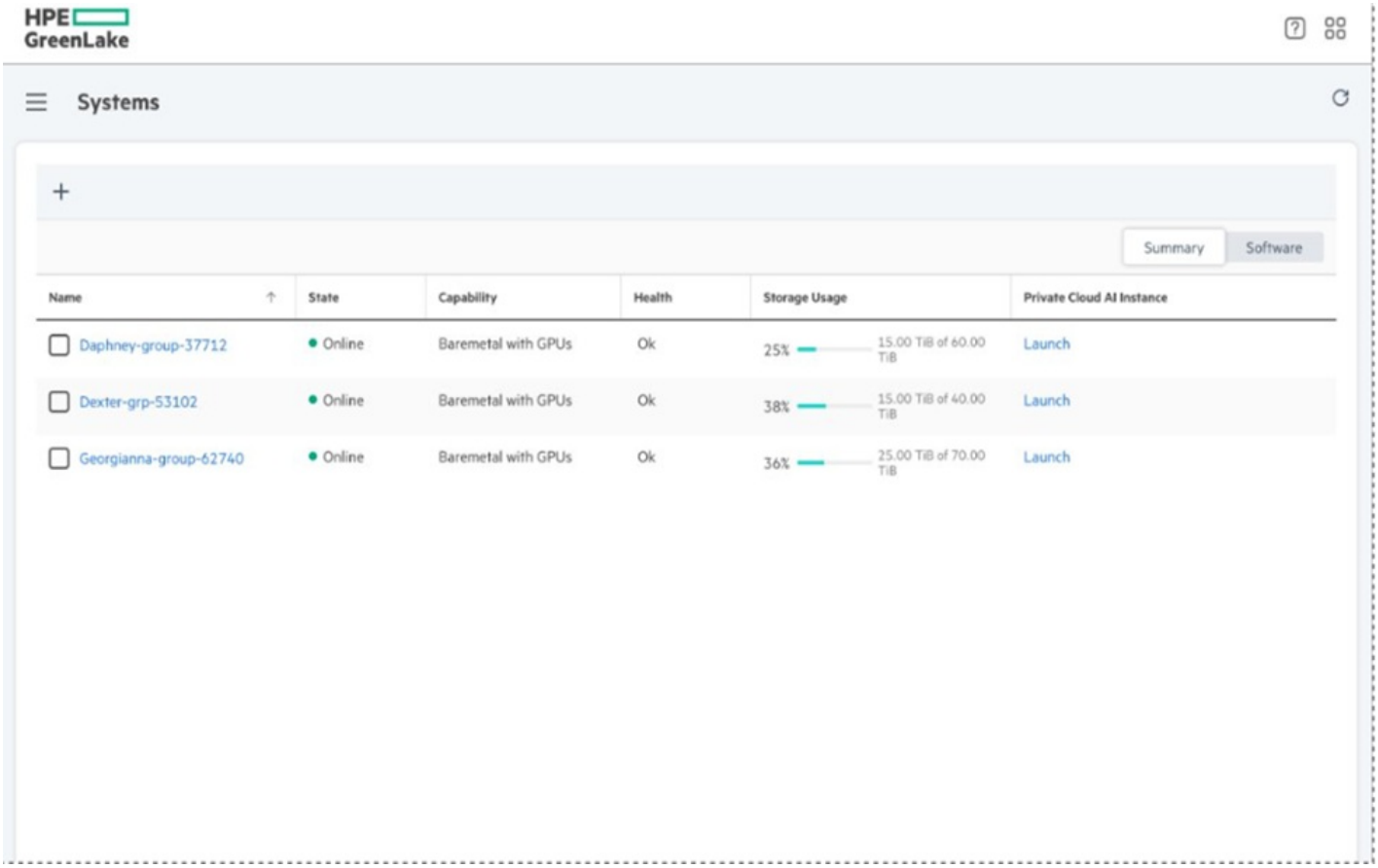
The Private Cloud AI dashboard serves as a nerve center for AI operations, enabling efficient management of complex systems and facilitating data-driven decision-making. By mastering the use of this powerful tool, AI Administrators can elevate their role from reactive troubleshooters to proactive strategists, driving the success and reliability of AI initiatives within their organizations.

Subtopics

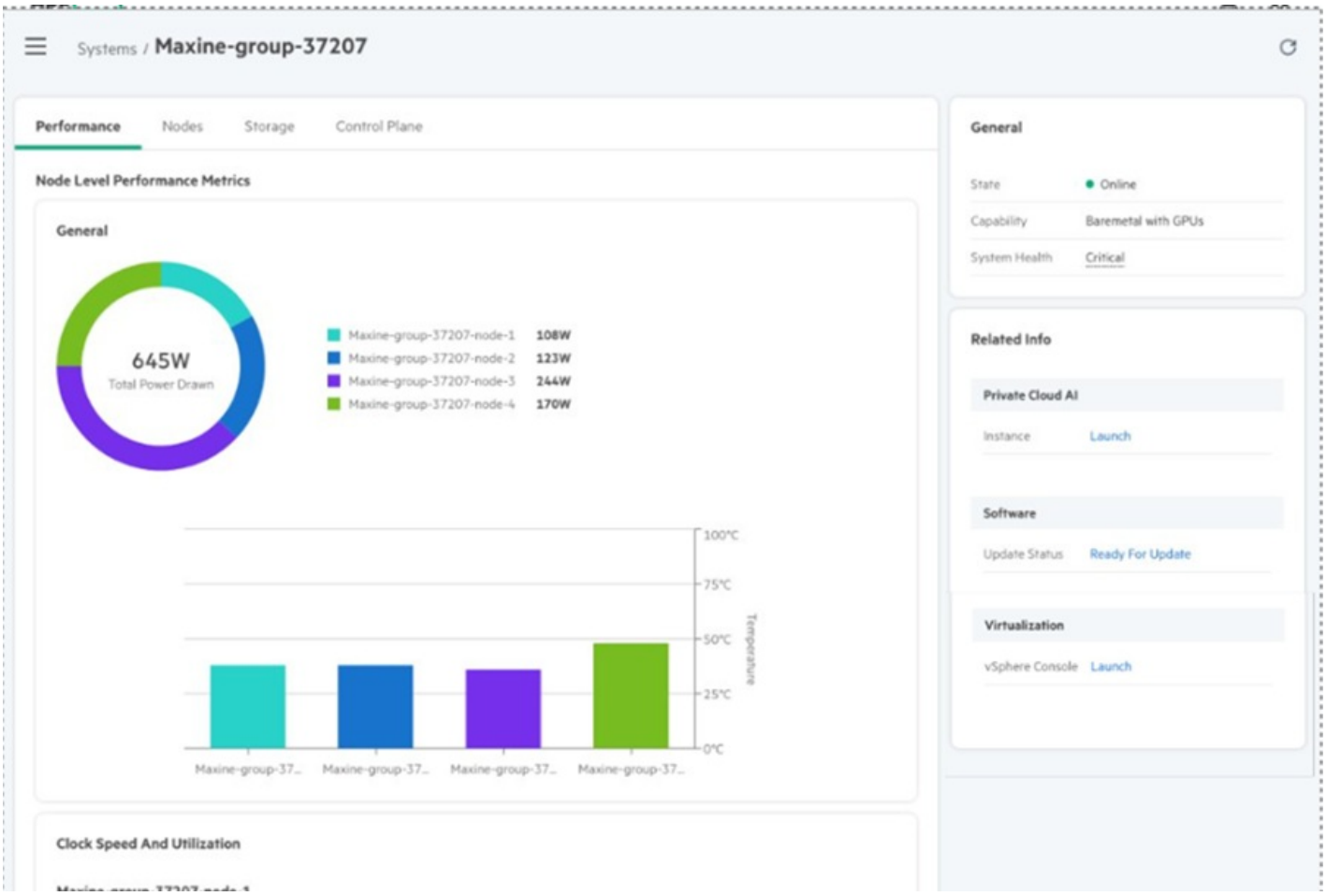
AI Systems page

Select AI Systems from the main menu to list the AI systems that make up the cluster, along with their most critical metrics.

Figure 1. AI Systems List



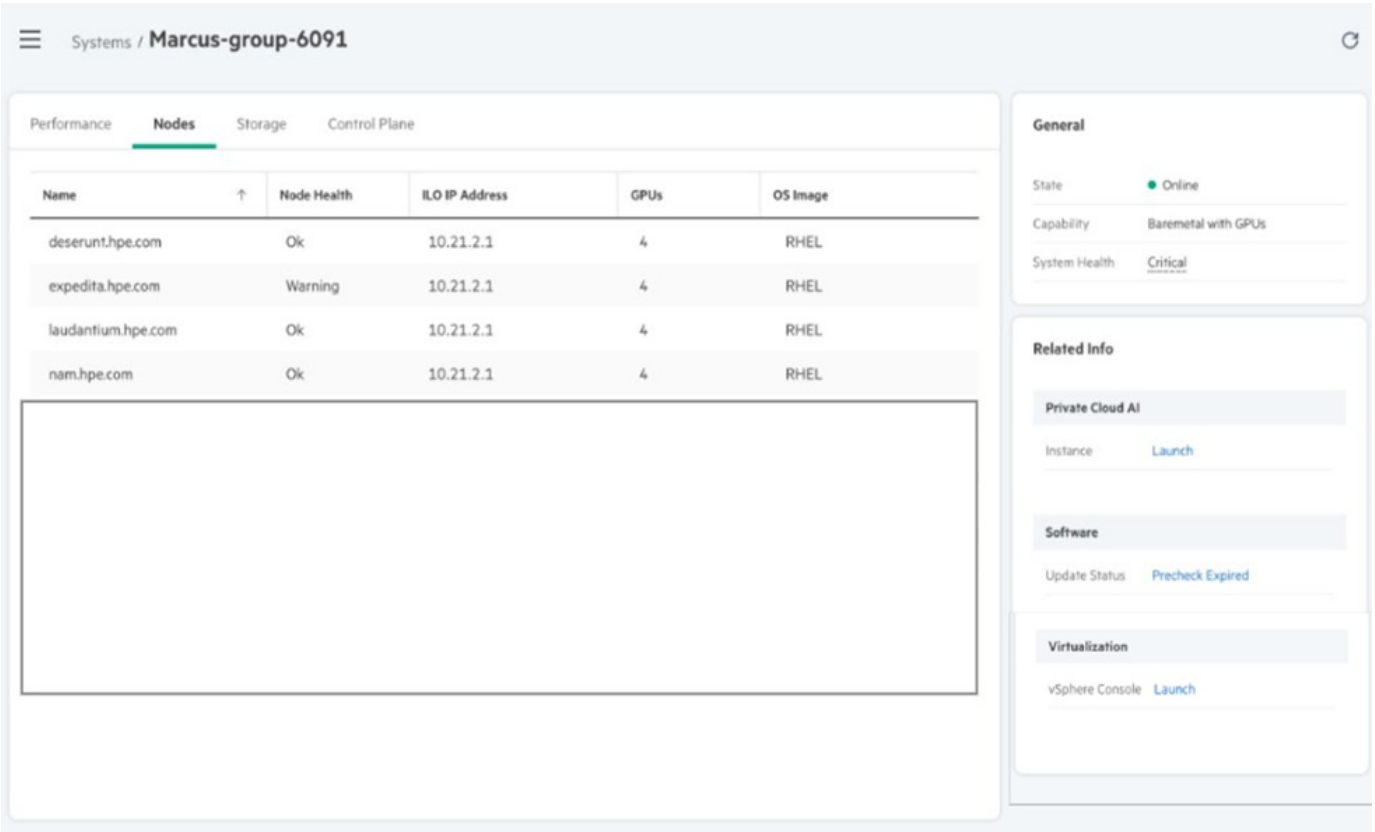
Click an AI system to display its metrics.



The metrics are arranged in the following tabs:

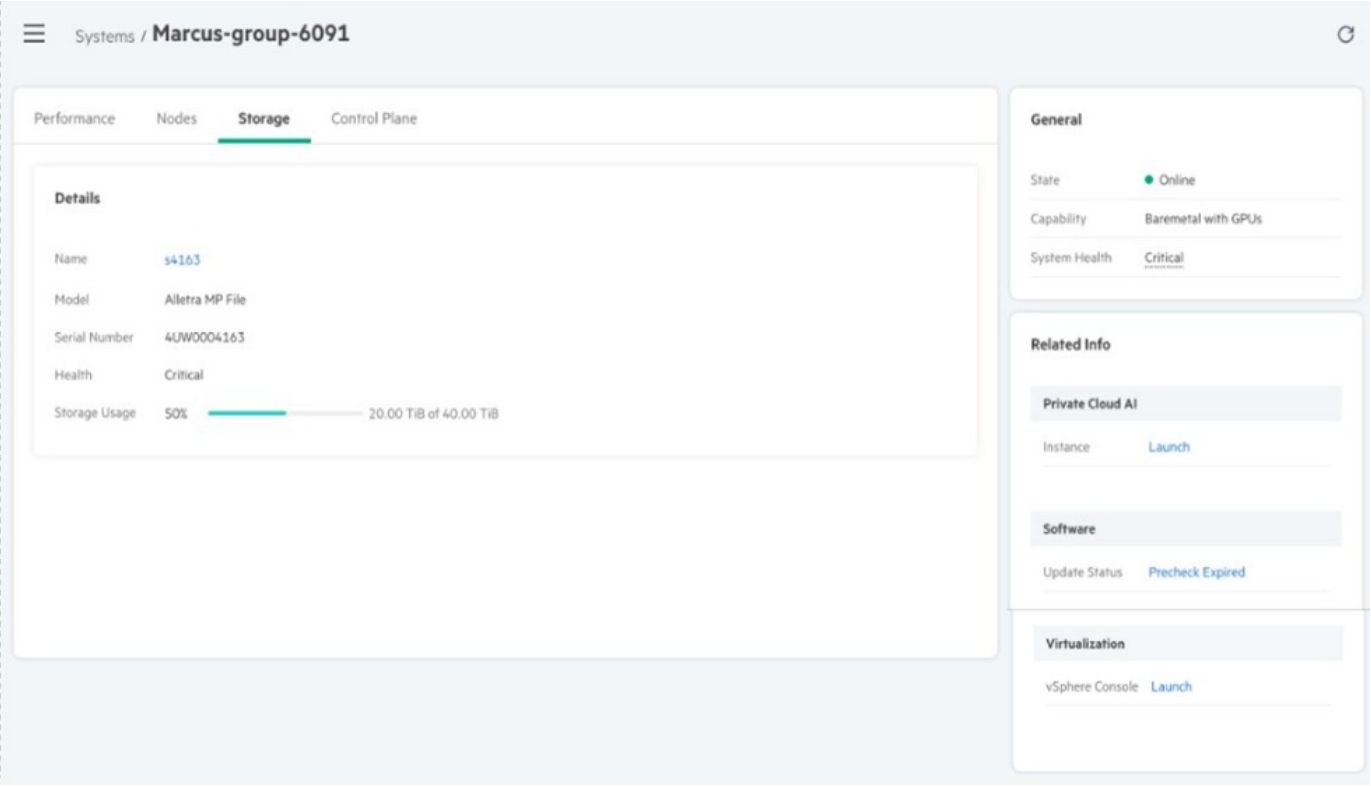
- **Performance**: This tab displays the hardware metrics of this host. Metrics such as power consumption, fan speed, temperature, clock speed, GPU, and memory utilization for all nodes are plotted graphically in real-time.
- **Nodes**: The page displays the nodes and their health status:

  Figure 2. Nodes Tab



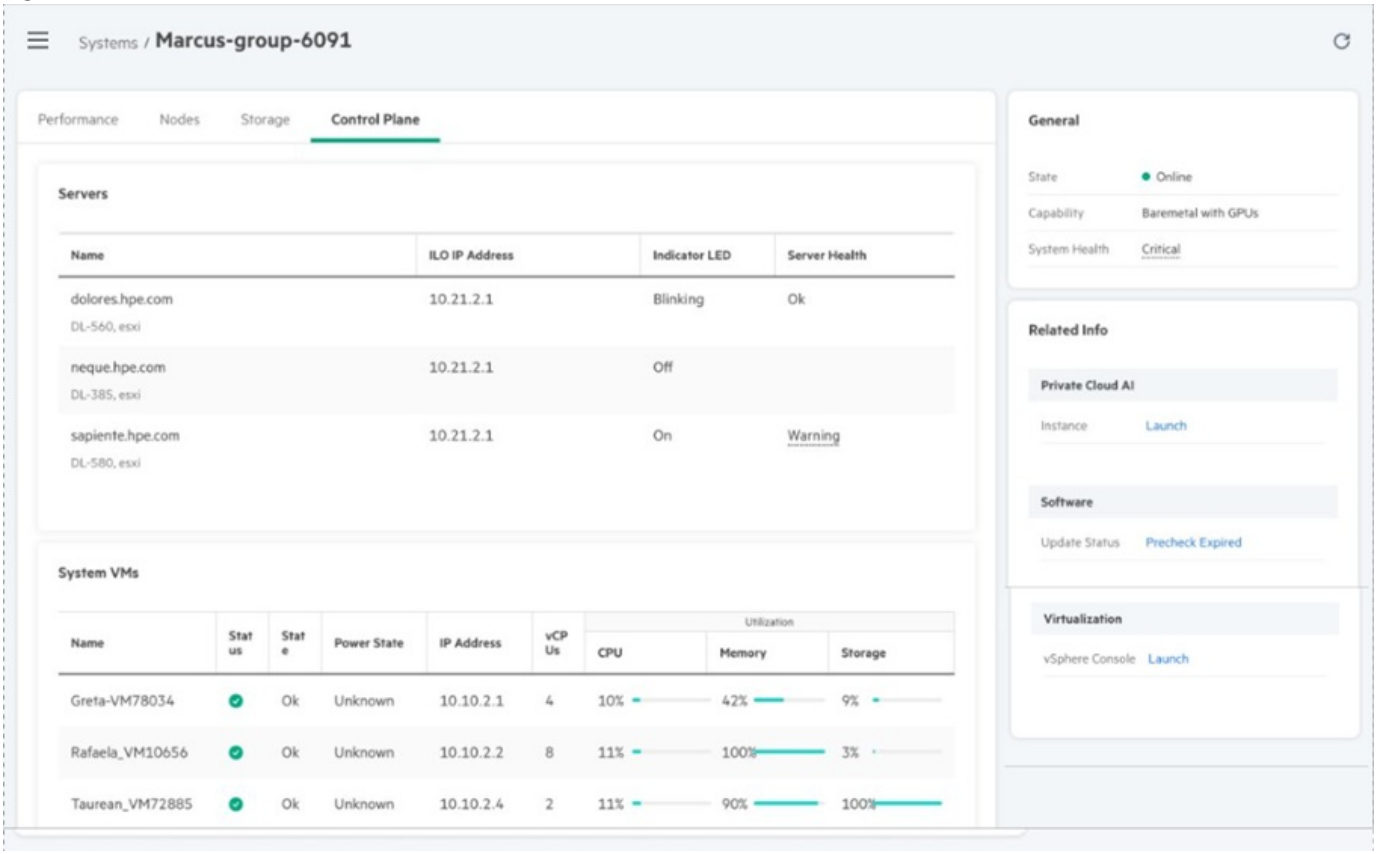
- **Storage**: The tab shows the Alletra MP storage system's serial number, health status and utilization.

  Figure 3. Storage Tab



- **Control Plane**: This tab displays the servers and the virtual machines hosting the nodes, along with their network details, resource usage and health status.

Figure 4. Control Plane Tab



The standard panel on the right displays the cluster's current state, health, the link to launch the Private Cloud AI instance, the update status of the cluster, and the link to launch the virtualization console.

Cloud Administrator: Maintaining the solution hardware and software

The Private Cloud AI Cloud Administrator oversees the overall health and maintenance of the solution hardware and software components in addition to performing common GreenLake administrative functions. For more information about GreenLake administrative tasks, see Getting started with HPE GreenLake

Customer Self Replaceable (CSR) components play a significant role in cloud administration, particularly in hybrid and on-premises cloud environments. For cloud administrators managing physical infrastructure, CSR components offer several advantages:

1. **Reduced Downtime**: Cloud admins can quickly replace faulty components without waiting for on-site technicians, minimizing service interruptions.
2. **Cost Efficiency**: By handling simple replacements in-house, organizations can reduce maintenance costs and avoid expensive service contracts.
3. **Scalability**: CSR components allow for easier hardware upgrades, enabling cloud admins to scale resources as needed without extensive downtime or external support.
4. **Inventory Management**: Cloud admins can maintain a stock of common CSR components, enabling rapid response to hardware issues.
5. **Remote Management**: In distributed cloud environments, local staff can replace CSR components following instructions from central cloud admin teams, enhancing remote management capabilities.
6. **Compliance and Security**: For sensitive environments, CSR components allow internal staff to maintain hardware without external technicians accessing secure areas.
7. **Training and Documentation**: Cloud admins can develop internal knowledge bases and training programs around CSR components, improving overall team capabilities and response times.

By leveraging CSR components effectively, cloud administrators can enhance the reliability, efficiency, and security of their physical infrastructure, complementing the flexibility and scalability of cloud services.

Elevate your AI operations with HPE Private Cloud AI, featuring user-friendly, replaceable components that put you in control of rapid recovery from hardware and software issues. Experience the next level of AI infrastructure where you have the power to minimize downtime, maximize productivity, and keep your business running smoothly with easy-to-manage technology designed for swift, hands-on problem-solving.

For more information about replaceable components, see Private Cloud AI Customer Self Repair (CSR) components .

Subtopics

Getting started with HPE GreenLake
Private Cloud AI Dashboard
Updating Private Cloud AI software
Ingress Gateway SSL Certificate
Describes the CA and the SSL certificate, that the CA signs, for the ingress gateway to use. Also describes how to configure your browser to trust the CA certificate and how to replace the certificate.
Private Cloud AI Customer Self Repair (CSR) components
OpsRamp Overview

Getting started with HPE GreenLake

This section describes the basic steps to access and use the HPE GreenLake platform. These include:

1. Creating an HPE GreenLake User Account
2. Create an HPE GreenLake Workspace
3. Deploy Data Services Cloud Console to Your HPE GreenLake Workspace

4. Assigning user roles

In order to perform these functions, you must be in the Private Cloud AI Cloud Administrator role.

Subtopics

Assigning user roles
Users must be assigned to the appropriate role to access and use services
Removing user roles
Accessing console help and documentation
You can access help for the services available in Data Services Cloud Console from the article widget in the UI. To access the article widget, click the book icon in the upper right corner of the UI.

Assigning user roles

Users must be assigned to the appropriate role to access and use services

Prerequisites

You must be in the HPE GreenLake Account Administrator role to assign users to roles.

Procedure

1. On your workspace Home page, click Manage Workspace in the Quick Links section.
2. On the Manage Workspace page, click Workspace identity & access .
3. On the Identity & Access page, click Assign a Role.
4. On the Assign Role dialog:
   a. Select the User.
   b. Select Data Services as the Service.
   c. Select a **Role**. For access to HPE AI Essentials Software, you must assign one of the following built-in roles (or an equivalent custom role):
      - Private Cloud AI AI Administrator: This role can view and manage AI Essentials resources.
      - Private Cloud AI User: This role can view AI Essentials resources.
   d. Leave the Limit Resource Access toggle off.
5. Back on the Identity & Access page, click the Users card.
6. Find the user, and select View Details from the corresponding more (...) menu. Verify that the user has the expected permissions.

Subtopics

Creating custom roles for Data Services Connector tasks
To perform Data Services Connector tasks, the user must be in one of the built-in roles or be assigned to a custom role with specific permissions.
Permissions and built-in roles required for Data Services Connector
Users must have valid permissions to perform Data Services Connectors operations.

Creating custom roles for Data Services Connector tasks

Creating custom roles for Data Services Connector tasks

To perform Data Services Connector tasks, the user must be in one of the built-in roles or be assigned to a custom role with specific permissions.

Prerequisites

You must be in the HPE GreenLake Account Administrator role to assign users to roles.

Procedure

1. On your workspace home page, click Manage Workspace in the Getting Started section.
2. On the Manage workspace page, click Identity & Access.
3. On the Workspace identity & access page, click Roles & Permissions.
4. Click Create Role, and then select Create new role.
5. Select Data Services from the service manager drop-down, and then click Next.
6. Add a meaningful Role Name and Description, and then click Next.
7. Click Add Permissions.
8. Select Data Services Connector resource, and then select the following permissions:
   - Register Data Services Connector
   - Unregister Data Services Connector
   - View Data Services Connector
   - Modify Data Services Connector
9. Select Hypervisor Manager resource, and then select the following permissions:
   - View Hypervisor Managers
   - Register Hypervisor Managers
   - Unregister Hypervisor Managers
   - Edit Hypervisor Managers
10. Click Add.
11. Review the details, and then click Finish.

Permissions and built-in roles required for Data Services Connector

Users must have valid permissions to perform Data Services Connectors operations.

To perform Data Services Connector tasks, the user must be in one of these roles or have the permissions included in a custom role.

Task	Permission	Roles that include these permissions by default
List the available Data Services Connectors	data-services.data-services-connector.read	• Private Cloud AI Cloud Administrator • Private Cloud Business Edition Administrator • Private Cloud Business Edition Operator • Backup and Recovery Administrator • Backup and Recovery Operator
Download and activate the Data Services Connector in Data Services Cloud Console.	data-services.data-services-connector.create	• Private Cloud AI Cloud Administrator • Private Cloud Business Edition Administrator • Backup and Recovery Administrator
• Update the Data Services Connector network configurations. • Upgrade the Data Services Connector. • Generate the one time password, enable and disable remote support, and collect the support dump.	data-services.data-services-connector.update	• Private Cloud AI Cloud Administrator • Private Cloud Business Edition Administrator • Backup and Recovery Administrator
Unregister the Data Services Connector from Data Services Cloud Console.	data-services.data-services-connector.delete	• Private Cloud AI Cloud Administrator • Private Cloud Business Edition Administrator • Backup and Recovery Administrator
List the associated hypervisor managers (vCenters) for a Data Services Connector.	data-services.hypervisor-manager.read	• Private Cloud AI Cloud Administrator • Private Cloud Business Edition Administrator • Private Cloud Business Edition Operator • Backup and Recovery Administrator • Backup and Recovery Operator
Register a hypervisor manager (vCenter)	data-services.hypervisor-manager.register	• Private Cloud AI Cloud Administrator • Private Cloud Business Edition Administrator • Backup and Recovery Administrator
Update the credentials of the hypervisor manager.	data-services.hypervisor-manager.update	• Private Cloud AI Cloud Administrator • Private Cloud Business Edition Administrator • Backup and Recovery Administrator
Unregister the associated hypervisor manager of a Data Services Connector from Data Services Cloud Console.	data-services.hypervisor-manager.unregister	• Private Cloud AI Cloud Administrator • Private Cloud Business Edition Administrator • Backup and Recovery Administrator

Removing user roles

Prerequisites

> **NOTE**
>
> If, instead of removing a role assignment from a user, you need to remove the user, see **Deleting a user**.

You must be in the HPE GreenLake Account Administrator role to remove users from their roles.

Procedure

1. On the HPE GreenLake cloud header, click the workspace menu and then select Manage Workspace.
2. On the Manage Workspace page, click Workspace identity & access.
3. On the Workspace identity & access page, click the Users card.
4. On the Users page, click the name of a user to view their role assignments in the user details screen.
5. To modify existing role assignments, click More (...) in the row of the role you want to modify. Options include:
   - **Edit Resource Access**—use this option to limit resource access for this role.
   - **Remove Role**—use this option to remove the role assignment for the user.

Accessing console help and documentation

You can access help for the services available in Data Services Cloud Console from the article widget in the UI. To access the article widget, click the book icon in the upper right corner of the UI.

Here are some useful points about the widget:

- All articles and release notes related to Data Services Cloud Console are available in the article widget.
- When you open the widget, articles and release notes relevant to the service that is active in the UI are listed.
- When you navigate to a new screen in Data Services Cloud Console, articles and release notes relevant to the new screen are listed.
- If you click View all documentation, a new window opens in the HPE Support Center.

For more information about documentation, see **Related documentation**.

Private Cloud AI Dashboard

> **NOTE**
>
> The dashboard only displays if there is at least one system that is installed. Else, you are prompted to activate your first system with the activation code provided at the end of the HPE Private Cloud AI On Premise Setup wizard.

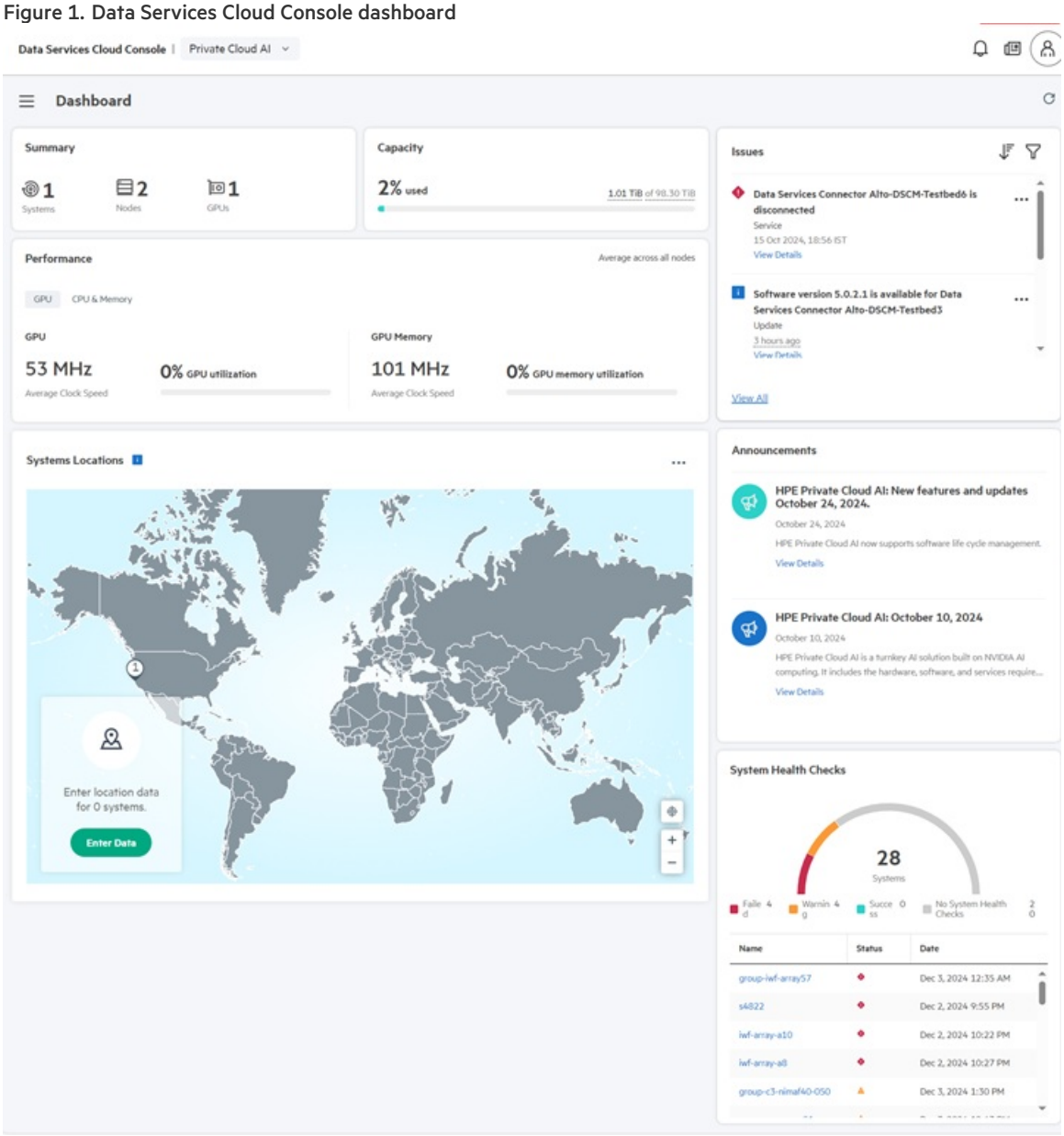
Log in to the HPE GreenLake Edge to Cloud Platform to access the Private Cloud AI Dashboard. The dashboard displays vital system metrics such as the used and remaining File Storage capacity, Memory usage, GPU performance, and a visual display of your systems' locations.

> **NOTE**
>
> The data in the dashboard does not refresh automatically. Manually refresh the page to view the latest data.

NOTE

The dashboard displays data for a maximum of five (5) systems. If there are more, a View All link displays. Click that link to view the Systems page.

Figure 1. Data Services Cloud Console dashboard



The dashboard consists of the following sections:

- **Summary**: Lists the number of AI clusters, the number of nodes that make up the clusters, and the total number of GPUs for the clusters.
- **Capacity**: Displays a listing of the individual file stores and their used capacity. Click a file store to view its details.
- **Performance**: Presents the average clock speed of the GPUs and Memory, and their average utilization in % for all the nodes in each cluster.
- **System Locations**: Presents a world map marked with the locations of the clusters.
- **Issues**: Lists the issues with the clusters. Click the **View Details** link to read more about the problem.
- **Announcements**: Displays announcements from HPE.
- **System Health Checks**: Displays the health status of each system in the cluster. Click a system link to view the details of that system.

Subtopics

Systems

The systems page provides comprehensive monitoring and management capabilities across your entire infrastructure. At a glance, view critical metrics including:

- System health status and real-time alerts
- Storage capacity and utilization trends
- Performance metrics like CPU, memory, and network usage
- Software versions and pending updates
- System capabilities and resource allocation

Select any system to access detailed diagnostics, historical data, and granular configuration options. The intuitive interface lets you quickly identify potential issues, optimize performance, and maintain system health across your infrastructure.

For more information, see **Using Private Cloud AI**.

Data Services Connectors

Data Services Connectors enable Private Cloud AI systems management through dedicated virtual machines.

Use the Data Services Connectors page to monitor and manage your active connectors, including configuration settings, performance metrics, and system status.

For more information, see **Using Private Cloud AI**.

Tasks

The Tasks page shows you a comprehensive overview of all cluster activity, including:

- Task status
- Associated service information
- Task initiator details
- Start timestamps

Select any task to view its detailed execution log and additional information

For more information, see **Using Private Cloud AI**.

Updating Private Cloud AI software

Prerequisites

NOTE
Upgrades from HPE Private Cloud AI 1.4.x to 1.5 are not currently supported.

Before installing software, review these considerations:

- You are assigned the Private Cloud AI Administrator role.
- Software updates require a sustained Internet connection of at least 100 Mbps. For more information about networking requirements, see **Networking requirements**.
- You can select multiple systems and update them in parallel.
- The Private Cloud AI software update feature updates components **one software catalog at a time**. Depending on the version of the specific component you need, a series of software catalog updates might be necessary.
  For example, if your system is running AI Essentials 1.6.1 and you need to update to 1.8.0, you must update from 1.6.1 to 1.7.0 first and then trigger another update to 1.8.0
- If you must perform multiple software catalog updates on the same system, wait **at least one hour** before triggering each new update. This wait period gives the system time to clean up old registry images.
- Before updating, be sure to review the release notes for the AI Essentials version to which you are updating. See the **HPE Private Cloud AI Release Notes**.

If the HPE AI Essentials Software cluster update fails, click **here** for troubleshooting steps.

Procedure

Follow the steps in **Managing software updates** to:

- Refresh software versions
- Run Precheck
- Download software updates
- Update the system software

You can access **Managing software updates** within the in-application Help or in the HPE Support Center under **Using Private Cloud AI for Cloud Administrators**.

Ingress Gateway SSL Certificate

Describes the CA and the SSL certificate, that the CA signs, for the ingress gateway to use. Also describes how to configure your browser to trust the CA certificate and how to replace the certificate.

The ingress gateway in HPE Private Cloud AI securely handles all traffic coming into the cluster with a certificate authority (CA). The CA validates the authenticity of the server certificate being used by the ingress gateway.

By default, the ingress gateway uses a secure socket layer (SSL) certificate signed by a self-signed CA for access. You can configure your browser to trust the CA included with HPE Private Cloud AI, or you can replace the ingress gateway certificate with a server certificate, the related CA root certificate, and intermediate certificates.

Configuring Your Browser to Trust the CA Certificate

You can configure your browser to trust the HPE Private Cloud AI CA certificate. Trusting the CA minimizes the number of security warnings that you have to accept when using the HPE Private Cloud AI. If you do not trust the CA, a security warning is raised every time you use a service or framework within HPE Private Cloud AI.

To configure your browser to trust the included CA, see Configuring Your Browser to Trust the CA Certificate .

Replacing the Ingress Gateway Certificate

Replacing the ingress gateway certificate allows you to set up automated certificate renewal on your IT organization's schedule, if your CA

works with cert-manager.

The generated certificate must cover the DNS domain used by the externally exposed services (web UIs and API endpoints) of the HPE Private Cloud AI cluster and have a subject or SAN that uses a wildcard (*) to cover the domain. For example, ***.my-pcai.com**.

> **IMPORTANT**
> Do not use this domain for any other purpose.

You can replace the CA certificate using either of the following methods:
Cert-Manager

Generate and replace the certificate artifacts using the cluster's cert-manager. You can also use cert-manager to renew certificates automatically. See Using cert-manager to Generate and Replace Certificates .

Manually

Generate and replace the certificate artifacts manually using kubectl or another Kubernetes API client. See Manually Generating and Replacing Certificates.

Subtopics

Using cert-manager to Generate and Replace Certificates
Describes how to use cert-manager to generate and replace or renew certificates for the ingress gateway in HPE Private Cloud AI.
Manually Generating and Replacing Certificates
Describes how to manually generate and replace the ingress gateway certificates in HPE Private Cloud AI.
Configuring Your Browser to Trust the CA Certificate
Describes how to configure your browser to trust the HPE Private Cloud AI certificate authority (CA).

Using cert-manager to Generate and Replace Certificates

Describes how to use cert-manager to generate and replace or renew certificates for the ingress gateway in HPE Private Cloud AI.

> **IMPORTANT**
> - This document describes how to use cert-manager with Let's Encrypt for the certificate authority (CA) and Google Cloud as the DNS service. However, Let's Encrypt and Google Cloud are not a requirement. You can use your preferred CA to generate server certificates and service to manage your DNS domain.
> - The Let's Encrypt server certificates are always generated with a 90 day lifetime. A 90 day lifetime is not appropriate for a "real" long-lived production installation unless you set up automatic renewal using cert-manager.

To generate and replace existing certificates using cert-manager, you must have permission to delete, add, and modify the following namespaces in the HPE Private Cloud AI cluster:

- `cert-manager`
- `istio-system`

To generate and replace certificates, complete the following steps:

1. Configure the CA and/or DNS service.
2. Create a secret and ClusterIssuer.
3. Use YAML to create a certificate and replace the existing ingress gateway certificate.

After you complete these steps, you can use a web browser to visit any of your cluster domain's URLs to verify that the server certificate is being presented in a certificate chain that ends with a trusted root certificate from the appropriate CA.

Configuring the CA and/or DNS Services

Configure the CA and/or DNS service(s) to allow cert-manager to authenticate itself to the CA server or prove that it has permission to access the certificate.

For Let's Encrypt and other ACME-supporting CAs, cert-manager uses the ACME v2 protocol to complete a DNS-01 type challenge to prove that it manages the resources covered by the certificate. The **cert-manager documentation** describes configurations for executing challenges with various specific DNS services as well as with other services that support dynamic DNS.

The following steps summarize how to use Google Cloud as the DNS service. For complete details, see the **cert-manager Google CloudDNS documentation**.

1. Create a Google Cloud service account to manage the domain for the HPE Private Cloud AI cluster.
2. Grant the service account privileges. The `dns.admin` permission works, but is likely too permissive for production environments.
3. Download a key file (`key.json`) for the service account.

Creating the Issuer in the Cluster

This section describes issuers, provides links to reference documentation, and provides the steps to create a ClusterIssuer and its secret.
Issuer Concepts

The issuer generates and renews the certificate. Configuring the issuer is a key aspect of using cert-manager. The **cert-manager Issuer Configuration documentation** covers a variety of integration possibilities, including:

- Home-grown CAs
- Vault
- Venafi
- Public CA services, such as Let's Encrypt that support the **ACME protocol**
- Other **cert-manager integrations**, such as **Google Certificate Authority Service** and **AWS Private CA**

Alternatively, some organizations develop a custom certificate manager plugin for corporate server certificates.

Creating the Secret and ClusterIssuer

A ClusterIssuer can react to any certificate request in any namespace, making it the appropriate type of issuer to use here.
Create a ClusterIssuer and its related secret in the `cert-manager` namespace. The secret holds the service account key file (`key.json`). The secret is referenced in the issuer resource.

To create the secret and ClusterIssuer, complete the following steps:

1. To create the secret in the `cert-manager` namespace, run the following command:

   ```
   kubectl -n cert-manager create secret generic clouddns-dns01-solver-svc-acct --from-file=key.json
   ```

   The secret name in this example is `clouddns-dns01-solver-svc-acct`; however, you can use a different name.

2. To create the ClusterIssuer, use the following YAML, replacing attribute values with values specific to your environment:

   NOTE

   This YAML is specifically for Let's Encrypt and Google Cloud DNS. The YAML shows the form of an issuer with an ACME DNS-01 challenge configured.

   ```
   apiVersion: cert-manager.io/v1
   kind: ClusterIssuer
   metadata:
   name: letsencrypt
   namespace: cert-manager
   spec:
   acme:
   server: https://acme-v02.api.letsencrypt.org/directory
   email: appropriate.person@their.domain.com
   privateKeySecretRef:
   name: letsencrypt
   ```

```
    solvers:
    - dns01:
    cloudDNS:
    project: my-gcloud-project
    hostedZoneName: my-ezaf
    serviceAccountSecretRef:
    name: clouddns-dns01-solver-svc-acct
    key: key.json
```

The following tables describes some of the YAML attributes:

Attribute	Value Description
server	ACME v2 endpoint for the CA. For Let's Encrypt, you can use their staging endpoint for testing: `acme-staging-v02.api.letsencrypt.org/directory` The staging endpoint does not generate a certificate with a trusted root certificate, but it does allow you to verify that things are working and it is not rate-limited like the real endpoint. If you use the staging endpoint, you must switch to the real endpoint for production.
email	This email address receives notifications, such as when certificate renewal is approaching.
project hostedZoneName	The project and hostedZoneName are specific to the Google Cloud project hosting the Google Cloud service account and the zone name in Google Cloud DNS.
serviceAccountSecretRef	The serviceAccountSecretRef name is the name of the secret, for example `clouddns-dns01-solver-svc-acct` and the name of the file used to create it (`key.json`).

Creating the Certificate

To create a certificate, complete the following steps:

1. Delete the existing ingress gateway certificate:

   ```
   kubectl -n istio-system delete certificate ingress-cert
   ```

2. Delete the certificate resource that was used to create the initial self-signed certificate:

   ```
   kubectl -n istio-system delete secret ingress-cert
   ```

3. Use the following YAML to create a certificate resource, replacing attribute values with values specific to your environment. The certificate resource generates a certificate and replaces the ingress gateway certificate.

   ```
   apiVersion: cert-manager.io/v1
       kind: Certificate
       metadata:
       name: ingress-cert
       namespace: istio-system
       spec:
       secretName: ingress-cert
       issuerRef:
       name: letsencrypt
       kind: ClusterIssuer
       group: cert-manager.io
       commonName: "*.my-ezaf.com"
       dnsNames:
       - '*.my-ezaf.com'
       usages:
       - server auth
       - client auth
       renewBefore: 360h
   ```

The following table describes some of the attributes in the YAML:

Attribute	Value Description
issuerRef	Specify the ClusterIssuer that you created.
commonName dnsNames	Specify the wildcard domain for the HPE Private Cloud AI cluster.
renewBefore	Specify when to expire the existing certificate and generate a new one.

NOTE

For additional optional attributes, see **Certificate resource**. Supported attributes depend on the CA used.

4. View the cert-manager logs to monitor the certificate resource that you created:

```
kubectl -n cert-manager logs -f -l app=cert-manager
```

Reference Documentation

The following documentation provides additional information related to certificates:

- **Issuer Configuration**
- **ACME Certificate Authorities**
- **Issuers**
- **Google Certificate Authority Service**
- **Secure Kubernetes with AWS Private CA**

Manually Generating and Replacing Certificates

Describes how to manually generate and replace the ingress gateway certificates in HPE Private Cloud AI.

IMPORTANT

- This document describes how to use Certbot with Let's Encrypt for the certificate authority (CA) and Google Cloud as the DNS service. However, Let's Encrypt, Certbot, and Google Cloud are not a requirement. You can use your preferred CA to generate server certificates and service to manage your DNS domain.
- The Let's Encrypt server certificates are always generated with a 90 day lifetime. A 90 day lifetime is not appropriate for a "real" long-lived production installation unless you set up automatic renewal using cert-manager.

To replace certificates, you must have permission to read and overwrite the `ingress-cert` secret in the `istio-system` namespace. The administrative `kubeconfig` has these privileges. Alternatively, you can set up and use a service account that has role-based access control (RBAC) constrained for this purpose.

To manually generate and replace certificates, complete the following steps:

1. Generate certificates. (This step is not required if your IT organization provides you with the certificates.)
2. Collect the certificate artifacts.
3. Replace the existing certificate.

Manually Generating Certificates

This section describes how to generate a Let's Encrypt certificate using the DNS challenge method.

> **NOTE**
> Let's Encrypt and Google Cloud are not a requirement. You can use your preferred CA and service to generate server certificates and manage your DNS domain.

Using the DNS challenge method confirms to Let's Encrypt that you manage the domain. The following steps summarize the process:

1. Start the Let's Encrypt process through Certbot.
2. Provide the TXT record to add to your DNS domain information, and verify that the TXT record has propagated.
3. Let's Encrypt generates the certificate.

Prepare

- Verify that you have the information and credentials required to manage the DNS domain.
- To interact with Let's Encrypt using the Certbot container image, set up local directories for Certbot to use:

```
LE_BASE=$HOME/Temp/letsencrypt
                mkdir -p "$LE_BASE/etc/letsencrypt"
                mkdir -p "$LE_BASE/var/lib/letsencrypt"
                mkdir -p "$LE_BASE/var/log/letsencrypt"
```

  Verify that `$HOME` is set to a valid directory. Alternatively, set `LE_BASE` to another location if you prefer to store the files elsewhere.

1 - Start the Let's Encrypt process

To start the Let's Encrypt process, run the following command:

```
UA_DOMAIN=my-ezaf.com
            CONTACT_EMAIL=appropriate.person@their.domain.com
            docker run -it --rm --name certbot \
            -v "$LE_BASE/etc/letsencrypt:/etc/letsencrypt" \
            -v "$LE_BASE/var/lib/letsencrypt:/var/lib/letsencrypt" \
            -v "$LE_BASE/var/log/letsencrypt:/var/log/letsencrypt" \
            certbot/certbot \
            certonly --manual --preferred-challenges=dns --email $CONTACT_EMAIL --
agree-tos -d "*.$UA_DOMAIN"
```

> **IMPORTANT**
> - Always use a valid email address as Let's Encrypt sends certification expiration notifications to this address.
> - If you run this in an environment that requires a proxy for internet access, the HTTP_PROXY and HTTPS_PROXY values must be communicated to the Docker Container. For additional details, see **Use a proxy server with the Docker CLI**.
> - For additional command-line options, refer to the **Certbot documentation**.

2 - Provide the TXT record to your DNS

When prompted, add the TXT record to your DNS domain and then use the link provided to verify that the change has propagated.

> **NOTE**
> Steps may vary based on the DNS provider. The following steps are based on Google Cloud DNS.

To provide the TXT record to your DNS, complete the following steps:

1. Select the relevant zone name in your Cloud DNS dashboard.
2. Create a new record set with **ADD STANDARD**.
3. Enter the DNS name returned by the docker `run` command.

4. Choose **TXT** as the record type.
5. For **TXT data**, enter the value returned by the docker `run` command.
6. Click Create.

3 - Locate the artifacts

After the process completes, locate the output (certificate artifacts) in the following directory:

```
$LE_BASE/etc/letsencrypt/live/$UA_DOMAIN
```

For additional information, see **Where are my certificates?**

Collecting Artifacts

You must have the following artifacts to replace the ingress gateway self-signed certificate:

Cert/File	Description
server cert	The concatenation of the server cert and any intermediate certs in the trust chain, up to but not including the trusted root cert.
trusted root cert	The trusted root cert from the certificate authority (CA).
private key	The private key for the server cert.

You can use the `fullchain.pem` and `privkey.pem` as the server certificate and private key. Download the root certificate referenced by `fullchain.pem` and use the root certificate as the CA certificate.

If you need to see which root certificate is referenced by `fullchain.pem`, complete the following steps:

1. Create temporary, separate files from the individual certificates contained in `fullchain.pem`. For example, if there are three different certificate sections delimited in `fullchain.pem`, you can save each of them as separate files (`cert1.pem`, `cert2.pem`, `cert3.pem`).
2. Get the subject and issuer of each certificate. For example, for the `cert1.pem` file:

   ```
   openssl x509 -in cert1.pem -text | grep '\(Issuer:\|Subject:\)'
   ```

   One certificate should have a subject matching the HPE Private Cloud AI cluster domain, for example `"CN = *.my-ezaf.com"` and some issuer. Another certificate should have a subject matching the issuer of the first certificate and its own issuer, with this pattern repeated down the chain.

   The last certificate in the chain should specify an issuer that is missing from the chain, for example:

   ```
   "C = US, O = Internet Security Research Group, CN = ISRG Root X1"
   ```

   In that case, go to the **Let's Encrypt root certs page** and download the self-signed `.pem` format for the `"ISRG Root X1"` cert, as this is your CA certificate artifact (`ca.pem` file).
3. (Optional) Delete the temporary individual certificate files. You only need the three artifacts (`fullchain.pem`, `privkey.pem`, and `ca.pem`) to replace the existing ingress gateway certificate.

Manually Replacing Certificates

To replace the ingress gateway certificate (including CA certificate and private key), complete the following steps:

1. Delete the existing ingress gateway certificate:

   ```
   kubectl -n istio-system delete secret ingress-cert
   ```
2. Back up the existing certificate information:

   ```
   kubectl -n istio-system get secret ingress-cert -o yaml > old-secret-ingress-cert.yaml
   ```
3. Replace the ingress gateway certificate:

   ```
   kubectl -n istio-system create secret generic ingress-cert --from-
   file=tls.crt=fullchain.pem --from-file=tls.key=privkey.pem --from-file=ca.crt=ca.pem
   ```

```
--dry-run=client -o yaml | kubectl apply -f -
```

Configuring Your Browser to Trust the CA Certificate

Describes how to configure your browser to trust the HPE Private Cloud AI certificate authority (CA).

Currently, HPE Private Cloud AI creates its own certificate authority (CA) that self-signs an SSL certificate. This method of verification requires you to accept the SSL certificate every time you access the top-level HPE Ezmeral Unified Analytics domain and any subdomains.

For example, when you sign in to HPE Ezmeral Unified Analytics, you sign in through the top-level domain and must accept the certificate for access to HPE Ezmeral Unified Analytics. For any services or frameworks that you access within HPE Ezmeral Unified Analytics, such as Kubeflow, you must also accept the certificate.

If you configure your browser to trust the HPE Private Cloud AI certificate authority (CA), you will not receive multiple prompts to accept the certificate when using services and frameworks within HPE Ezmeral Unified Analytics.

Configuring your browser to trust the SSL certificate signed by the HPE Private Cloud AI CA

To configure your browser to trust the SSL certificate signed by the HPE Private Cloud AI CA, you must configure the browser to trust the HPE Private Cloud AI CA.
To configure your browser to trust the HPE Private Cloud AI CA:

1. Go to **https://home.your-pcai-domain/ca** and allow your browser to automatically download the certificate of the HPE Private Cloud AI CA.
2. Change the browser settings to trust the certificate. The steps vary by browser. The following steps apply to Chrome:
   a. Go to **chrome://certificate-manager**.
   b. Under **Custom**, click **Installed by you**.
   c. Next to **Trusted Certificates**, select **Import**.
   d. Browse to the location where you downloaded the HPE Ezmeral Unified Analytics certificate and select the certificate.
3. Restart Chrome for the changes to take effect immediately.

(Optional) Update your operating system to trust the CA

Steps vary by operating system. The following steps apply to RHEL. RHEL requires the `ca-certificates` package and a properly configured CA trust store.
To update your operating system to trust the CA:

1. Install the `ca-certificates` package:

   ```
   sudo yum install ca-certificates
   ```

2. If not already enabled, enable dynamic CA:

   ```
   sudo update-ca-trust enable
   ```

3. Copy the certificate to the appropriate directory for the system to recognize it. Common locations for CA certificates include:
   - `/etc/pki/ca-trust/source/anchors/`
   - `/usr/share/pki/ca-trust-source/anchors/`

   ```
   sudo cp /path/to/your/ca.pem /etc/pki/ca-trust/source/anchors/
   ```

4. To update the CA trust store, run the `update-ca-trust` command to refresh the system's trust store:

   ```
   sudo update-ca-trust
   ```

5. To verify that the certificates were added to the trust store, run the following command and check the output:

   ```
   cat /etc/pki/ca-trust/extracted/pem/tls-ca-bundle.pem | grep "Your CA name"
   ```

Alternatively, you can list the trust store:

```
trust list
```

Private Cloud AI Customer Self Repair (CSR) components

The Private Cloud AI features the following components you can repair:

> **NOTE**
> Please do not downgrade the hardware or software components. Only modify or replace components through the approved update process.

Hardware Components

Component	Manuals
HPE ProLiant DL325 and ProLiant DL380a Gen11 servers	• HPE ProLiantDL325 Gen11 Server Maintenance and Service Guide • HPEProLiantDL325 Gen11 Manuals • HPE ProLiant DL325 Gen11 Server User Guide • HPE ProLiant DL380a Gen11 Server User Guide • HPE ProLiant DL380a Gen11 Server Maintenance and Service Guide • HPE ProLiant DL380a Gen11 manuals
NVIDIA GPUs	• L40S Product Brief • NVIDIA H100 NVL 94GB PCIe Accelerator for HPE (S2D86A) • NVIDIA H100 PCIe GPU Product Brief
NVIDIA Networking	NVIDIA Spectrum-3 SN4000 1U and 2U Switch Systems Hardware User Manual
HPE Aruba Networking CX 6300 Switch Series	Aruba CX 6300 Manuals

Software Components

Table 1. Software Components

Component	Manuals
NVIDIA AI Enterprise	Deployment Guides
RedHat Enterprise Linux OS	• Cheatsheet for System Administrators • Documentation
VMware vCenter Server	vCenter Documentation

OpsRamp Overview

OpsRamp is an operations management platform that can help you monitor, manage, and automate your IT infrastructure.

All OpsRamp features are available to users of HPE Private Cloud AI.

To learn more about OpsRamp, see the **OpsRamp Documentation**.

To get started using OpsRamp, see **Getting Started with OpsRamp Service On HPE GreenLake Cloud Platform**.

Backing up and restoring Private Cloud AI configuration

This section outlines the procedures for backing up and restoring the Private Cloud AI solution configuration, ensuring business continuity during hardware failures or system disruptions. By implementing these practices, we aim to minimize downtime and maintain the integrity of our solution environment. Private Cloud AI is available in four configuration sizes: Developer, Small, Medium, and Large. The information in this section applies to Small, Medium, and Large configurations unless otherwise indicated.

NOTE
Backup and restore are not supported for HPE Morpheus VM Essentials Software.

Subtopics

Private Cloud AI configuration data backup strategies

To ensure timely recovery of the Private Cloud AI setup in the event of hardware failure, a robust backup strategy is essential. The Private Cloud AI environment consists of Control Nodes, AI Worker Nodes, GreenLake for File Storage (GLFS), and Network Switches, each requiring specific backup considerations. On-Premises Backup to HPE GLFS leverages the inherent high availability and redundancy of the GLFS system for rapid local recovery. This method is ideal for minimizing downtime within the same data center, though it does not protect against logical data corruption or data center-wide disasters. Customers can supplement GLFS backups with external storage solutions, using VM Essentials to deploy a VM for integrating an external backup solution.

1. On-Premises Backup to HPE GLFS (Highly Available and Redundant storage):

- **Description:**
- This strategy involves backing up Private Cloud AI configuration data directly to the HPE GLFS storage system.
- GLFS provides inherent high availability and redundancy, minimizing the risk of data loss due to hardware failures within the GLFS cluster.
- This method is suitable for rapid recovery within the same data center.
- **Benefits:**
- Leverages existing GLFS infrastructure.
- Fast recovery times due to local storage.
- Simplified backup management.
- **Considerations:**
- Does not protect against logical data corruption or accidental overwrites within GLFS.

- Does not provide protection against data center-wide disasters.

2. Backup to External storage (Configuration and Application Data Backup)

Private Cloud AI is a customer-managed solution. Users have the flexibility to implement their preferred backup strategies for configuration and application data. Use VM Essentials to deploy a VM for the purpose of integrating an external backup solution.

GLFS backup directory structure

For Private Cloud AI environments, the location of backup data storage is crucial for efficient recovery. You must store all Private Cloud AI configuration data on GLFS. Therefore, a dedicated storage space is required to ensure organized and efficient backup management. Control nodes will have a GLFS mount, /vmfs/volumes/glfs-nfs-datastore (ESXi-based solution) or /mnt/<config id> (VME-based solution), as shown in the following images. These are example mount points, and the actual mount points can vary.

Figure 1. GLFS mount on ESXi based control nodes

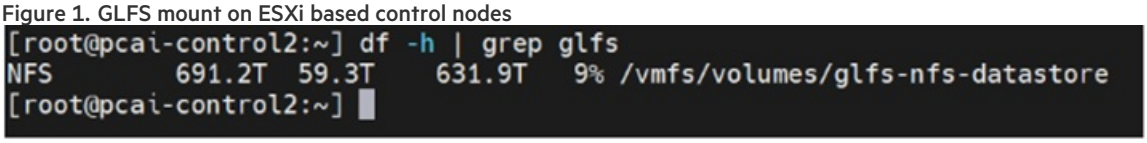
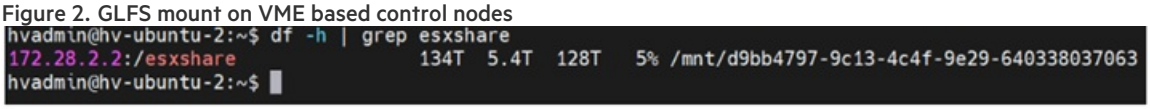


Figure 2. GLFS mount on VME based control nodes



To maintain a well-organized backup structure, create the following directories on the GLFS share: /<GLFS mount path>/PCAI_Backup/Worker, /<GLFS mount path>/PCAI_Backup/Control_Plane, and /<GLFS mount path>/PCAI_Backup/Network. These directories will be used to store the configuration data for worker nodes, control plane nodes, and network switches, respectively. This separation enhances organization and simplifies the recovery process for individual PCAI components. Consistent naming conventions are highly recommended.

NOTE

After you complete each of the following backup steps, you must copy the backup files to the shared folders on the GLFS storage. Doing so ensures your backups are safely stored and easy to find if you need to recover anything.

Control Plane VM snapshot

This section outlines the process for creating and restoring Control Plane VM snapshots. VM snapshots capture a point-in-time image of a virtual machine's disk, memory, and configuration, enabling quick rollback to a previous state. While useful for testing, software installations, and troubleshooting, snapshots are not backups and should not be used for long-term data protection. This section covers CLI methods for creating and restoring snapshots, emphasizing best practices and considerations for minimizing performance impact and ensuring data integrity. It is crucial to understand that snapshots are only for short-term rollback purposes, and a separate backup strategy is essential for disaster recovery.

Procedure to take a snapshot of the VMs running on ESXi

1. Connect to the ESXi host using an SSH client.
2. Identify the VMID of the virtual machine you want to snapshot using the following command:

   ```
   vim-cmd vmsvc/getallvms
   ```

   Figure 1. Identify VMID

```
[root@pcai-control3:~] vim-cmd vmsvc/getallvms
Vmid                    Name                                                                    File
                                Guest OS      Version

                                             Annotation

1       vCLS-9dc4486a-4573-4eb2-b184-128bcb776e95   [datastore1] vCLS-9dc4486a-4573-4eb2-b184-128bcb776e95/vCLS-9dc4486a-
4573-4eb2-b184-128bcb776e95.vmx   otherGuest     vmx-11    vSphere Cluster Service VM is deployed from an OVA with a mi
nimal installed profile of PhotonOS. vSphere Cluster Service manages the resources, power state and availability of thes
e VMs. vSphere Cluster Service VMs are required for maintaining the health and availability of vSphere Cluster Service.
Any impact on the power state or resources of these VMs might degrade the health of the vSphere Cluster Service and caus
e vSphere DRS to cease operation for the cluster.
4       ezfabmaster                          [glfs-nfs-datastore] ezfabmaster/ezfabmaster.vmx
                                rhel8_64Guest   vmx-19



5       HPEVME-Host2                         [datastore1] HPEVME-Host3/HPEVME-Host3.vmx
                                ubuntu64Guest   vmx-21
```

3. Use the following command to create the snapshot:

```
vim-cmd vmsvc/snapshot.create <vmid> <snapshot_name>
<snapshot_description> <include_memory> <quiesced>
```

<include_memory>

This setting determines whether the VM's memory state is included in the backup.

- **1 (Include):** The memory contents of the VM are captured as part of the backup. When your backup includes memory contents, it saves the exact state of your virtual machine at backup time. The saved state allows for faster recovery since the VM can start up right where it left off.
- **0 (Exclude):** The memory contents are not included. The backup only captures the VM's disk data. Smaller backup files result, but the VM must boot up from scratch after a restore.

<quiesced>

This setting controls whether the VM's file system is quiesced (temporarily frozen in a consistent state) before the backup.

- **1 (Quiesce):** The file system is brought to a consistent state before the backup. Such a state is crucial for applications that maintain transactional data (e.g., databases). Quiescing ensures that all pending writes are flushed to disk, preventing data corruption.
- **0 (Do Not Quiesce):** The file system is not quiesced, and can result in inconsistent backups if the VM is running applications that are actively writing data.

Example

Use the following command to create a snapshot named "MySnapshot" with a description "Test Snapshot" for VMID 5, including memory and quiescing the file system:

```
vim-cmd vmsvc/snapshot.create 5 "MySnapshot" "Test Snapshot" 1 1
```

Figure 2. Create a Snapshot

```
[root@pcai-control3:~] vim-cmd vmsvc/snapshot.create 5 Test_Snapshot est_snapshot 1 1
Create Snapshot:
[root@pcai-control3:~] vim-cmd vmsvc/snapshot.get 5
Get Snapshot:
|-ROOT
--Snapshot Name        : Test_Snapshot
--Snapshot Id        : 3
--Snapshot Desciption  : est_snapshot
--Snapshot Created On  : 3/7/2025 11:13:16
--Snapshot State       : powered on
[root@pcai-control3:~]
```

Procedure to restore a snapshot to the VMs running on ESXi

1. Connect to the ESXi host using an SSH client.
2. Power Off the VM (Recommended): While technically you can restore a snapshot while the VM is running, it's strongly recommended to power it off to prevent data corruption. Use the following command, replacing <vmid> with the VM's ID:

   ```
   vim-cmd vmsvc/power.off <vmid>
   ```

3. Use the following command to list all the snapshots with their IDs, names, and creation times.

   ```
   vim-cmd vmsvc/snapshot.get <vmid>
   ```

4. Restore the snapshot: Use the `vmkfstool` command with the `--restore-snapshot` option.

   ```
   vmkfstool --restore-snapshot "[<snapshot_id>]" <vmx_path>
   ```

   NOTE

   Replace `<snapshot_id>` with the numerical ID of the snapshot (including the square brackets) and `<vmx_path>` with the full path to the VM's .vmx file. The .vmx path is usually something like /vmfs/volumes/<datastore_name>/<vm_name>/<vm_name>.vmx.

   You can find this path in the vSphere Client or by browsing the datastore on the ESXi host.

   Example

   ```
   vmkfstool --restore-snapshot "[3]" /vmfs/volumes/datastore1/MyVM/MyVM.vmx
   ```

5. Power on the VM.

   ```
   vim-cmd vmsvc/power.on <vmid>
   ```

VMWare ESXi configuration backup and restore

Maintaining a robust backup strategy for your VMware ESXi host configurations is paramount for ensuring rapid recovery and minimizing downtime in a critical virtualized environment. This section details the procedures for backing up essential ESXi settings, including network configurations, storage mappings, security policies, and advanced system parameters. By capturing these configurations, we establish a reliable recovery point, enabling swift restoration of virtual infrastructure in the event of hardware failures, accidental modifications, or other unforeseen disruptions.

Backup procedure

1. **Establish Connectivity:** Connect to the ESXi host via SSH using appropriate credentials.
2. **Synchronize Configuration**: Use the following command to ensure the running configuration is saved to the persistent storage.

   ```
   vim-cmd hostsvc/firmware/sync_config
   ```

3. **Backup Configuration**: Use the following command to generate the .tgz backup file.

   ```
   vim-cmd hostsvc/firmware/backup_config
   ```

   Figure 1. Generate the backup configuration

   IMPORTANT

   Note the path where the backup is created.

4. Copy the Backup File to GLFS share: Use the `cp` command to copy the file to a GLFS share. For example:

```
cp /usr/lib/vmware/hostd/docroot/downloads/52adbeff-170e-93b6-3123-
8a7106579a96/configBundle-pcai-control1.vcfmr.local.tgz /vmfs/volumes/glfs-nfs-
datastore/PCAI_Backup/configBundle-pcai-control1.vcfmr.local.tgz
```

Restore procedure

Restore the Configuration:

1. Copy the backup to /tmp/.

```
cp volumes/glfs-nfs-datastore/PCAI_Backup/configBundle-pcai-
control1.vcfmr.local.tgz /tmp/
```

2. Run the following command to restore the configuration.

```
vim-cmd hostsvc/firmware/restore_config /tmp/configBundle-pcai-
control1.vcfmr.local.tgz
```

3. Reboot the ESXi Host:

   The command `restore_config` reboots the server. If it does not, use the following command to reboot the server.

```
reboot
```

> **CAUTION**
>
> Rebooting will cause downtime. Plan accordingly.

4. Verify the Configuration: After the ESXi host reboots, verify that the configuration has been restored correctly. Check network settings, storage configurations, and other relevant parameters using the vSphere Client or command-line tools.

Automate backups

Create a script to automate the backup process. Use `cron` (on Linux) to schedule regular backups.

Kubernetes etcd backup and restore

etcd serves as the critical data store for our Kubernetes cluster, holding all configuration and state information. This section of the guide provides detailed procedures for backing up and restoring etcd, ensuring the resilience and continuity of our production environment. Implementing these steps is essential for rapid recovery from data corruption or hardware failures, minimizing downtime and maintaining service availability. To ensure proper backup and restore of the Kubernetes cluster's configuration, run these procedures on the etcd leader node, one of the Master VMs.

Backup procedure

1. **Identify etcd endpoints**

   - You can find the etcd endpoints by checking the Kubernetes API server configuration, which is located at /etc/kubernetes/manifests/kube-apiserver.yaml.
   - Etcd requires authentication. You'll need the client certificates and keys, which are found at /etc/kubernetes/pki/etcd/

     Figure 1. Identtify etcd endpoints

```
[root@workloadmaster ~]# ls -ltr /etc/kubernetes/pki/etcd/
total 32
-rw-r-----. 1 root etcd 1679 Nov 28 04:02 server.key
-rw-r-----. 1 root etcd 1257 Nov 28 04:02 server.crt
-rw-r-----. 1 root etcd 1675 Nov 28 04:02 peer.key
-rw-r-----. 1 root etcd 1257 Nov 28 04:02 peer.crt
-rw-r-----. 1 root etcd 1679 Nov 28 04:02 healthcheck-client.key
-rw-r-----. 1 root etcd 1127 Nov 28 04:02 healthcheck-client.crt
-rw-r-----. 1 root etcd 1679 Nov 28 04:02 ca.key
-rw-r-----. 1 root etcd 1070 Nov 28 04:02 ca.crt
```

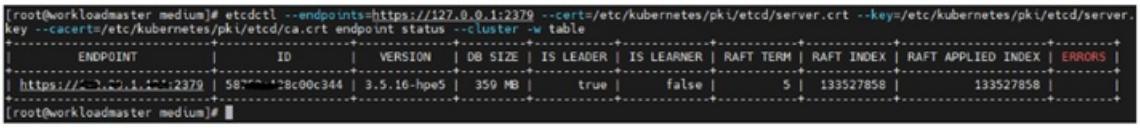
2. **Check the cluster health**

   `etcdctl endpoint status --cluster`

   For example:

   ```
   etcdctl --endpoints=https://<etcd-endpoints> --cert=<client-certificate-file> --key=
   <client-key-file> --cacert=<ca-certificate-file> endpoint status --cluster -w table
   ```

   Figure 2. Check the cluster health

   - Ensure that the **Is Leader** column shows true for one of the endpoints.
   - Verify that the ERRORS column is empty for all endpoints.
   - Check that the DB SIZE is within expected limits.

3. **Create an etcd Snapshot**

   - Use the `etcdctl snapshot save` command to create a snapshot.

     ```
     ETCDCTL_API=3 etcdctl --endpoints=<etcd-endpoints> \
      --cacert=<ca-certificate-file> --cert=<client-certificate-file> --key=<client-key-file>
     \
      snapshot save <snapshot-file-path>
     ```

     NOTE

     Replace `<etcd-endpoints>`, <ca-certificate-file>, <client-certificate-file>, <client-key-file> and <snapshot-file-path> with your actual values.

     Figure 3. Create an etcd snapshot

   - Copy or transfer the snapshot to the GLFS share on the control node.

4. **Verify the Snapshot:** Use the `etcdctl snapshot status` command to verify the snapshot. For example:

   `ETCDCTL_API=3 etcdutl snapshot status <snapshot-file-path>`

5. **Schedule Backups**
   - Automate etcd backups using cron jobs or systemd timers.
   - Perform backups regularly (e.g., hourly or daily), depending on the rate of change in your cluster.

Restore procedure

1. **Identify the etcd Data Directory**
   - Use `systemctl cat etcd` to view the etcd systemd unit file.
   - Look for the `--data-dir` flag in the `ExecStart` line.
2. **Stop the etcd Service:** On each node that runs etcd, stop the etcd service using the command:

   ```
   systemctl stop etcd
   ```
3. **Move the Existing etcd Data Directory (Backup):**
   - Navigate to the Data Directory: Go to the etcd data directory.
   - Rename the Directory: Rename the existing data directory to create a backup. For example:

     ```
     mv <etcd-data-dir> <etcd-data-dir>.bak
     ```

     NOTE

     Replace <etcd-data-dir> with the actual path to your etcd data directory
4. **Restore etcd from Snapshot:**
   - Copy the etcd backup from the GLFS share on control node to the `/tmp/` directory
   - Use the `etcdctl snapshot restore` command to restore the etcd snapshot. For example:

     ```
     ETCDCTL_API=3 etcdutl snapshot restore <snapshot-file-path> \
     --data-dir=<new-etcd-data-dir> \
     --initial-cluster=<initial-cluster-configuration> \
     --initial-cluster-token=<cluster-token> \
     --initial-advertise-peer-urls=<advertise-peer-urls>
     ```

     NOTE

     Replace:
     - <snapshot-file-path>: With the path to your etcd snapshot file.
     - <new-etcd-data-dir>: With the path to the new etcd data directory (e.g., the original directory name).
     - <cluster-token>: With the cluster token (if used).
     - <advertise-peer-urls>: With the advertise peer URLs (e.g., https://<etcd1-advertise-peer-url>:2380).
     - The <initial-cluster-configuration>, <cluster-token>, and <advertise-peer-urls> parameters are typically found in the etcd configuration files or the kube-apiserver pod manifest.
5. **Start the etcd Service:** Start the etcd service on each node using the command:

   ```
   systemctl start etcd
   ```
6. **Verify etcd Health:** Follow step 2 of Etcd Backup to verify etcd health .

Worker Node system configuration backup and restore

To ensure a comprehensive and reliable system recovery, it is essential to capture not only core configurations and data but also the dynamic hardware configurations managed by udev. This section outlines a detailed procedure for backing up crucial system files, including udev rules, logs, network settings, user data, and installed packages. By capturing these elements, we create a robust recovery point, enabling rapid restoration and minimizing downtime.

Backup procedure

1. **Critical data to backup**
   Configuration files

   /etc/
   /etc/sysconfig/
   /etc/ssh/
   /etc/fstab
   /etc/hosts
   /etc/resolv.conf
   /etc/yum.repos.d/

   Logs

   /var/log

   Users and groups

   /etc/passwd
   /etc/shadow
   /etc/group
   /home/

   Network configuration

   /etc/sysconfig/network-scripts/
   `ip route show`
   `ip addr show`
   `iptables-save`

   Cron jobs

   /var/spool/cron/
   /etc/crontab

   udev rules

   /etc/udev/rules.d/: Custom udev rules
   /lib/udev/rules.d/: System default udev rules

2. **Create a backup directory**

   ```
   mkdir /backups/system_backup
   ```

3. **Use tar for file and directory backups**

   ```
   tar -czvf /backups/system_backup/etc_backup.tar.gz /etc/

   tar -czvf /backups/system_backup/var_log_backup.tar.gz /var/log/

   tar -czvf /backups/system_backup/home_backup.tar.gz /home/

   tar -czvf /backups/system_backup/udev_rules.tar.gz /etc/udev/rules.d/

   tar -czvf /backups/system_backup/etc_shadow.tar.gz /etc/shadow
   ```

4. **Capture network configuration**

```
ip route show > /backups/system_backup/routes.txt

ip addr show > /backups/system_backup/ip_addresses.txt

iptables-save > /backups/system_backup/iptables.rules

tar -czvf /backups/system_backup/network-scripts.tar.gz /etc/sysconfig/network-
scripts/
```

5. **Backup installed packages**

```
rpm -qa > /backups/system_backup/installed_packages.txt
```

Automate backups

To maintain the integrity and timeliness of our system backups, it is crucial to automate the process using scripts and schedule them with cron or systemd timers. This ensures that backups are consistently executed, capturing the latest system configurations, critical logs, and user data.

ILO configuration backup and restore

Backing up your iLO configuration is crucial for ensuring business continuity and simplifying system recovery. The backup file contains essential settings like network configuration, user accounts, security settings, and firmware version, allowing you to quickly restore iLO to a known working state in case of hardware failure, misconfiguration, or firmware upgrades gone wrong. This proactive measure minimizes downtime, reduces troubleshooting efforts, and helps maintain a consistent and secure iLO environment.

Prerequisites:

- Access to the iLO web interface (either directly or through a network connection).
- An iLO account with administrator privileges.
- A location to save the backup file (GLFS path).

Backup procedure

1. **Access the iLO web interface:** Open a web browser and enter the iLO IP address or hostname in the address bar.
   Log in using an iLO account with administrator privileges.
2. **Navigate to Lifecycle Management:** In the left-hand navigation menu, click Lifecycle Management.
3. **Select Backup & Restore:** Click the Backup & Restore tab.
4. **Download ILO Configuration:** Click the Download button to download the configuration file to your local system. Then, transfer this file to the PCAI_Backup directory on the GLFS share mounted on the Control node.

Restore procedure

> **NOTE**
> Before starting the restore process, ensure you have copied the iLO backup file to the local system from which you are accessing the iLO web interface.

1. **Access the iLO Web Interface:**
   Open a web browser and enter the iLO IP address or hostname in the address bar.

   Log in using an iLO account with administrator privileges.
2. **Navigate to Lifecycle Management:** In the left-hand navigation menu, click Lifecycle Management.

3. **Select Backup & Restore:**
   Click the Backup & Restore tab and then click Restore.
4. **Choose the Backup File:**
   An **Open** dialog box will appear, allowing you to navigate to the location where you saved the backup file.

   Select the backup file and click Open.
5. **Confirm Restore:**
   Click Upload and Restore.

   A confirmation message will appear, warning you that the restore operation will overwrite the current iLO configuration.

   Click
   Yes
   to proceed with the restore.

   Wait for the Restore to Complete.

   The iLO web interface may become temporarily unavailable during the restore process.

   Once the restore is complete, the iLO web interface will be refreshed with the restored configuration.

Network Switch configuration backup and restore

Maintaining the stability and resilience of our network infrastructure is paramount. This section outlines the procedures for backing up the configurations of our critical network switches, including Nvidia SN series and Aruba 8325/6300M models. These backups provide a crucial recovery point, enabling rapid restoration of network services in the event of configuration errors or hardware failures.

To ensure recoverability, store a backup of the network configuration outside the Private Cloud AI environment. This protects against scenarios where an OOBM switch failure prevents access to the GLFS share and its stored backups.

Backup and restore procedures for each switch type

The GLFS share, mounted on the control node, will be used to store network switch configuration backups. Please ensure the PCAI_Backup directory exists on the GLFS share at /vmfs/volumes/glfs-nfs-datastore.

Subtopics

NVIDIA SN4600, SN4700, SN5600 switches

Backup Procedure

This procedure outlines how to backup the configuration of NVIDIA SN4600, SN4700, and SN5600 switches to the GLFS share.

On NVIDIA switches, the configuration is saved in the file /etc/nvue.d/startup.yaml. We will directly back up this file.

1. **Access the Switch:** Connect to the switch using appropriate credentials.
2. **Transfer the Backup to the GLFS share:** Run the following command to backup the configuration file on GLFS through the control node.

```
scp /etc/nvue.d/startup.yaml root@<control node>:/<GLFS mount
path>/PCAI_Backup/Network/<backup file name>.yaml
```

> **NOTE**
>
> Replace `<control node>` with the IP address of the control node and `<backup file name>` with the appropriate file name, following the naming convention `SwitchType_Model_Date_Backup.yaml`

For example:

```
scp /etc/nvue.d/startup.yaml root@10.10.10.11:/vmfs/volumes/glfs-nfs-datastore/PCAI_Backup/NvidiaMSN4700_Backup.yaml
```

Restore procedure

This procedure outlines how to restore the configuration of Nvidia SN4600, SN4700, and SN5600 switches from a backup stored on the GLFS share.

1. **Access the Control Node:** Connect to the control node via SSH using appropriate credentials.
2. **Transfer the Backup to the NVIDIA switch:** Use the `scp` command to transfer the backup configuration file from the GLFS share on your control node to the Nvidia switch's /etc/nvue.d/ directory: .

   ```
   scp /<GLFS mount path>/PCAI_Backup/Network/<backup file name>. yaml <switch user>@<switch IP address>:/etc/nvue.d/startup.yaml
   ```

   > **NOTE**
   >
   > Replace `<switch user>` with the appropriate username for the NVIDIA switch, `<switch IP address>` with the switch's IP address, and <backup file name>. yaml with the actual name of your backup file.

3. **Verify File Transfer:**
   Connect to the NVIDIA switch via SSH.

   Verify that the startup.yaml file exists in the /etc/nvue.d/ directory and that its contents match the backed-up configuration.
4. **Reload the Configuration:** On the NVIDIA switch, run the following command to reload the configuration from the startup.yaml file:

   ```
   nv set load
   ```

5. **Verify the Configuration:** Use the appropriate NVIDIA switch commands to verify that the configuration has been restored correctly. For example, you can use nv show running config to compare the running configuration with the restored configuration.

Aruba 6300M and Aruba 8325 switches

Backup Procedure

This procedure outlines how to backup the configuration of Aruba 6300M and Aruba8325 switches to the GLFS share.

1. **Access the Switch:** Connect to the Aruba Switch via SSH using appropriate credentials.
2. **Transfer the Backup to the GLFS share:** Run the following command to generate and transfer the switch configuration to the GLFS share.

   ```
   copy running-config scp://root@<control node>:/<GLFS mount path>/PCAI_Backup/Network/<backup file name> cli vrf default
   ```

> **NOTE**
>
> Replace `<control node>` with the IP address of the control node and `<backup file name>` with the appropriate file name, following the naming convention `SwitchType_Model_Date_Backup`.

For example:

```
copy running-config scp://root@172.28.1.119:/vmfs/volumes/glfs-nfs-datastore/PCAI_Backup/Aruba_config cli vrf default
```

Restore procedure

This procedure outlines how to restore the configuration of Aruba 6300M and Aruba8325 switches from a backup stored on the GLFS share.

1. **Access the Control Node:** Connect to the Aruba Switch via SSH using appropriate credentials.
2. **Copy the Backup to the Aruba Switch:** Use the copy scp command to transfer the backup configuration file from the GLFS share on your control node to the Aruba switch.

   ```
   copy scp://root@<control node>:/<GLFS mount path>/PCAI_Backup/Network /<backup file name> running-config cli vrf default
   ```

   > **NOTE**
   >
   > Replace `<control node>` with the IP address of the control node and `<backup file name>` with the backed-up configuration file.

   For example:

   ```
   copy scp://root@172.28.1.119/vmfs/volumes/glfs-nfs-datastore/PCAI_Backup/Aruba_config running-config cli vrf default
   ```

3. **Save the Configuration (Make It Persistent) :**
   Run the following command to save the configuration.

   ```
   copy running-config startup-config
   ```

4. **Verify the Configuration:** Use the `show running-config` command to verify the restored configuration.
5. **Reboot the Switch (If Necessary):** In some cases, a reboot might be required for all configuration changes to take effect.

   ```
   reload
   ```

   > **NOTE**
   >
   > Rebooting will cause network downtime. Plan accordingly.

PDU replacement procedure

The Power Distribution Unit (PDU) does not store data requiring backup. To replace the PDU, please adhere to the instructions outlined in the HPE support document found at: **HPE Basic Power Distribution Unit Installation Guide**.

After the physical replacement of the PDU, perform the following configuration steps:

- Set the administrator password for the PDU.
- Enable the REST API for remote management.
- Configure the IPv4 address to ensure network connectivity.

- If specified by the solution's recipe, install the required firmware (FW) version.

FRU replacement procedure

When determining hardware replacement procedures, the key factors are the sales agreement terms and the component's warranty status. These elements dictate whether HPE or the customer handles the replacement process.

Warranty-covered hardware replacement

1. **Initiating a Service Request:** In the event of a hardware component failure under warranty, initiate a service request by contacting HPE Support. This action is the first step in the hardware replacement process.
2. **Creating a Support Ticket:** A HPE support representative will assist in creating a detailed support ticket. Provide comprehensive information regarding the hardware issue and the required replacement part. This documentation is essential for accurate and timely service.
3. **Field Engineer Dispatch:** After submitting the support ticket, HPE will schedule a qualified field engineer to perform the hardware replacement at your location.
4. **Verification and System Restoration:** Following the hardware replacement, the field engineer will verify the successful installation and functionality of the new hardware. Ensure that the system is fully restored to its operational state before the support ticket is closed.

Out-of-warranty hardware replacement

1. **Hardware Procurement:** For hardware components outside the warranty period, the customer is responsible for procuring the replacement part. This step is essential before proceeding with the replacement.
2. **Hardware Replacement (Self-Service):** Customers are required to perform the hardware replacement independently. Follow all safety guidelines and manufacturer instructions during this process.
3. **Firmware and Software Compatibility Verification:** After the hardware replacement, verify firmware and software compatibility by consulting the HPE Private Cloud AI Firmware and Software Compatibility Matrix for your specific release. This ensures the installation of compatible drivers and firmware.
   Access a **sample matrix**.
4. **Verification and Functionality Testing:** Verify the proper functionality of the newly installed hardware by conducting thorough testing. This step confirms the successful completion of the replacement and ensures system stability.

Troubleshooting HPE AI Essentials Software

EZKF-1994 Failed User Onboarding

The first sign-in to HPE AI Essentials Software onboards a new user. If a new user signs in to HPE AI Essentials Software for the first time and sees the following message, the onboarding process failed:

`Unfortunately, we couldn't get you on board.`

A cluster administrator can manually resolve this issue either by removing an `ezuserconfig` resource associated with the user's email address or by removing the user entry in the cluster's local Keycloak instance.

To resolve this issue, the administrator must have access to the Kubernetes API of the on-premises cluster and `kubectl` to run commands. The `kubectl config` file must provide the necessary administrative credentials. A Private Cloud AI Cloud Administrator can download the `kubeconfig` file (provided at the end of installation) and then use the `kubeconfig` to run the `kubectl` commands. Alternatively, the Private Cloud AI Cloud Administrator can share the `kubeconfig` with the Private Cloud AI Administrator and the Private Cloud AI Administrator can run the `kubectl` commands.

Before completing any steps to resolve the issue, check to see if an `ezuserconfig` resource is associated with the user's email address.

To see if an `ezuserconfig` resource is associated with the user's email address:

1. Specify the email address. For example, if the email address is `user-email@foo.com`, specify the email address as:

   ```
   EMAIL=user-email@foo.com
   ```

2. Locate any `ezuserconfig` resource associated with the email address:

   ```
   kubectl get ezuserconfig --no-headers | awk "\$5 == \"$EMAIL\""
   ```

   - If the email address is **not** associated with an `ezuserconfig` resource, complete the steps listed in **Remove a User from Keycloak**.
   - If the email address is associated with an `ezuserconfig` resources, complete the steps listed in **Remove an `ezuserconfig` Resource**.

Remove an `ezuserconfig` Resource

To remove an `ezuserconfig` resource associated with a user's email:

1. Specify the email address. For example, if the email address is `user-email@foo.com`, you would specify the email address as:

   ```
   EMAIL=user-email@foo.com
   ```

2. Display any user configurations associated with the email:

   ```
   kubectl get ezuserconfig --no-headers | awk "\$5 == \"$EMAIL\""
   ```

3. Capture the `ezuserconfig` resource name:

   ```
   USERCFG=$(kubectl get ezuserconfig --no-headers | awk "\$5 == \"$EMAIL\" { print
   \$1 }")
   ```

4. Display the resource name and verify that it matches:

   ```
   echo $USERCFG
   ```

5. Delete the resource:

   ```
   kubectl delete ezuserconfig $USERCFG
   ```

   NOTE

   After you remove an `ezuserconfig` resource associated with a user's email, the user can sign in again to re-trigger the onboarding process. Do **not** perform the steps in the *Remove a User from Keycloak* section. Deleting the `ezuserconfig` resource updates Keycloak.

6. Instruct the user to sign in again to re-trigger the onboarding process.

Remove a User from Keycloak

If the user's email is not associated with an `ezuserconfig` resource, delete the user entry from the cluster's local Keycloak instance. To remove the user entry from the Keycloak instance:

1. Retrieve the Keycloak administrator credentials:

   ```
   kubectl -n keycloak exec keycloak-0 -- bash -c 'echo $KEYCLOAK_ADMIN_PASSWORD
   '
   ```

2. Point a web browser at the following address:

   ```
   https://keycloak.<your-cluster-domain.com
   ```

   Where `<your-cluster-domain.com>` is the DNS subdomain used for all of the on-premises cluster services. In some versions of Keycloak, your browser immediately presents you with the login form for the Keycloak admin console. If not, select the admin console from the choice tiles on the first page you see.

3. Log in to the admin console with the username `admin` and the password that you retrieved in step 1.

4. In the dropdown menu (in the upper-left corner of the page), select the **UA** or **Unified Analytics** realm.

5. On the sidebar, select **Users**.

6. In the search field on the **Users** page, enter the email address of the user.
7. Select the checkbox to the left of the username to select the user and then click the **Delete user** link above the user table.
8. Instruct the user to sign in again to re-trigger the onboarding process.

NOTE

If you followed the steps provided and the issue persists, contact HPE Support.

EZAF-5899 Failed CTAS query on Hive Metastore in HPE Ezmeral Data Fabric

For Hive connections that authenticate to HPE Ezmeral Data Fabric through MAPRSASL, running a CREATE TABLE AS query against HPE Ezmeral Data Fabric returns the following error:

```
Database 'pa' location does not exist:<file_path>
```

To resolve this issue, create and upload a configuration file that points to the HPE Ezmeral Data Fabric cluster, as described in **Using MAPRSASL to Authenticate to Hive Metastore on HPE Ezmeral Data Fabric**.

EZKF-2166 HPE AI Essentials Software cluster update fails

The upgrade can fail due to the following issues:

- Intermittent failures during HPE AI Essentials Software component updates
- Worker node scheduling issues after Kubernetes upgrade, where the `SchedulingDisabled` label remains on the node

Review the system error message to identify the specific issue and perform the corresponding resolution step as shown in the following table:

Issue	Error Message	Resolution
Component update intermittent failure	Upgrade Workload container posted an error at one of the upgrade components (generic, infra, apps)	This is typically an intermittent issue that can resolve itself. Click Resume Updates to retry. If the problem persists, contact HPE Support.
The `SchedulingDisabled` label remains on the worker node	Failed to upgrade k8s version due to relay CR failure (or) AI Essentials cluster update timed out waiting for events	Contact HPE Support.

Related documentation

This section describes the end-user documentation for the HPE Private Cloud AI engineered system.

A comprehensive list of end-user documentation for all HPE Private Cloud AI versions is available at **https://www.hpe.com/support/Private-Cloud-AI-Docs**.

Title	See this title if you are a . . .	Summary/Abstract
HPE Private Cloud AI Release Notes	• Cloud administrator • Private Cloud AI user	This document describes new features, known issues, resolved issues, and important notes for administering or using the Private Cloud AI system.
HPE Private Cloud AI Developer System Installation Guide	• System installer • Cloud administrator • Security administrator	This guide describes how to install the HPE Private Cloud AI Developer System configuration. The guide covers preparing, installing, configuring, and powering up the Developer System, as well as connecting to the Private Cloud AI service from Data Services Cloud Console. The guide is intended for use by HPE Private Cloud AI customers.

Title	See this title if you are a . . .	Summary/Abstract
HPE Private Cloud AI Administration Guide	• Cloud administrator • Private Cloud AI user • Security administrator	This guide is the primary reference for Cloud Administrators, AI Administrators, and AI Users who work with HPE Private Cloud AI. The guide describes the architecture, components, and configurations, as well as how HPE AI Essentials users can use the engineered system. It covers essential topics, including user roles and permissions, system management, troubleshooting procedures, and maintenance protocols. Earlier versions of this guide were published as the HPE Private Cloud AI User Guide.
Using Private Cloud AI for Cloud Administrators	• Cloud administrator	This documentation is intended for cloud administrators who are using the Data Services Cloud Console UI to set up, upgrade, and monitor the Private Cloud AI infrastructure. This content is available from within the Private Cloud AI UI as context-sensitive help and in HPE Support Center as a collection of topics.
HPE AI Essentials Documentation	• Data engineer • Data scientist • Private Cloud AI user	This documentation is intended for Private Cloud AI Administrators and Private Cloud AI Users who want to use HPE AI Essentials Software to build end-to-end Generative AI solutions. For Private Cloud AI Administrators, this documentation describes how to connect external data sources, perform data, user, and resource management tasks, deploy pre-built Gen AI solution accelerators, monitor the cluster, and configure tools and frameworks. For Private Cloud AI Users, including Data Engineers, Data Scientists, and Data Analysts, this documentation describes how to use the included tools and frameworks to perform ML and AI tasks, such as designing and implementing data pipelines, managing models, building and deploying Retrieval-Augmented Generation (RAG) solutions with NVIDIA NIM, and use the Playground to fine-tune and interact with AI solutions.
HPE AI Essentials Overview Video	• Data engineer • Data scientist • Private Cloud AI user	The HPE AI Essentials Software video provides a brief tour of the HPE AI Essentials interface, highlighting the AI and data-foundation tools within the unified control plane, including NVIDIA NIM, Knowledge Base, and pre-built Solution Accelerators; and provides a look at some of the administrative features, including data access management.
HPE AI Essentials Guided Tours	• Data engineer • Data scientist • Private Cloud AI user	Explore HPE AI Essentials using interactive guided tours. Learn the platform's core functionality, complete key tasks, and understand platform utilization by roles such as Data Scientists and Data Engineers. Step-by-step guidance is available directly within the platform interface for all users. Just click the guided tours icon in the menu to get started.

Title	See this title if you are a . . .	Summary/Abstract
HPE Private Cloud AI Firmware and Software Compatibility Matrix	• Data engineer • Private Cloud AI Cloud Administrator	This document outlines the software and firmware versions for all components based on the recipe version used in the HPE Private Cloud AI engineered system deployment. It is intended for Private Cloud AI users, HPE Integration Center (factory) engineers, cloud administrators, and data engineers. The document is available on the HPE Support Center.

Support and other resources

Subtopics

Accessing Hewlett Packard Enterprise Support

- For live assistance, go to the Contact Hewlett Packard Enterprise Worldwide website:

  https://www.hpe.com/info/assistance

- To access documentation and support services, go to the Hewlett Packard Enterprise Support Center website:

  https://www.hpe.com/support/hpesc

Information to collect

- Technical support registration number (if applicable)
- Product name, model or version, and serial number
- Operating system name and version
- Firmware version
- Error messages
- Product-specific reports and logs
- Add-on products or components
- Third-party products or components

HPE product registration

To gain the full benefits of the Hewlett Packard Enterprise Support Center and your purchased support services, add your contracts and

products to your account on the HPESC.

- When you add your contracts and products, you receive enhanced personalization, workspace alerts, insights through the dashboards, and easier management of your environment.
- You will also receive recommendations and tailored product knowledge to self-solve any issues, as well as streamlined case creation for faster time to resolution when you must create a case.

To learn how to add your contracts and products, see **https://www.hpe.com/info/add-products-contracts**.

Accessing updates

- Some software products provide a mechanism for accessing software updates through the product interface. Review your product documentation to identify the recommended software update method.
- To download product updates:

  Hewlett Packard Enterprise Support Center

  https://www.hpe.com/support/hpesc

  My HPE Software Center

  https://www.hpe.com/software/hpesoftwarecenter

- To subscribe to eNewsletters and alerts:

  https://www.hpe.com/support/e-updates

- To view and update your entitlements, and to link your contracts and warranties with your profile, go to the Hewlett Packard Enterprise Support Center More Information on Access to Support Materials page:

  https://www.hpe.com/support/AccessToSupportMaterials

> **IMPORTANT**
> Access to some updates might require product entitlement when accessed through the Hewlett Packard Enterprise Support Center. You must have an HPE Account set up with relevant entitlements.

Remote support

Remote support is available with supported devices as part of your warranty or contractual support agreement. It provides intelligent event diagnosis, and automatic, secure submission of hardware event notifications to Hewlett Packard Enterprise, which initiates a fast and accurate resolution based on the service level of your product. Hewlett Packard Enterprise strongly recommends that you register your device for remote support.

If your product includes additional remote support details, use search to locate that information.

HPE Get Connected

https://www.hpe.com/services/getconnected

HPE Tech Care Service

https://www.hpe.com/services/techcare

HPE Complete Care Service

https://www.hpe.com/services/completecare

Warranty information

To view the warranty information for your product, see the **warranty check tool**.

Regulatory information

To view the regulatory information for your product, view the Safety and Compliance Information for Server, Storage, Power, Networking, and Rack Products, available at the Hewlett Packard Enterprise Support Center:

https://www.hpe.com/support/Safety-Compliance-EnterpriseProducts

Additional regulatory information

Hewlett Packard Enterprise is committed to providing our customers with information about the chemical substances in our products as needed to comply with legal requirements such as REACH (Regulation EC No 1907/2006 of the European Parliament and the Council). A chemical information report for this product can be found at:

https://www.hpe.com/info/reach

For Hewlett Packard Enterprise product environmental and safety information and compliance data, including RoHS and REACH, see:

https://www.hpe.com/info/ecodata

For Hewlett Packard Enterprise environmental information, including company programs, product recycling, and energy efficiency, see:

https://www.hpe.com/info/environment

Documentation feedback

Hewlett Packard Enterprise is committed to providing documentation that meets your needs. To help us improve the documentation, use the Feedback button and icons (at the bottom of an opened document) on the Hewlett Packard Enterprise Support Center portal (**https://www.hpe.com/support/hpesc**) to send any errors, suggestions, or comments. This process captures all document information.