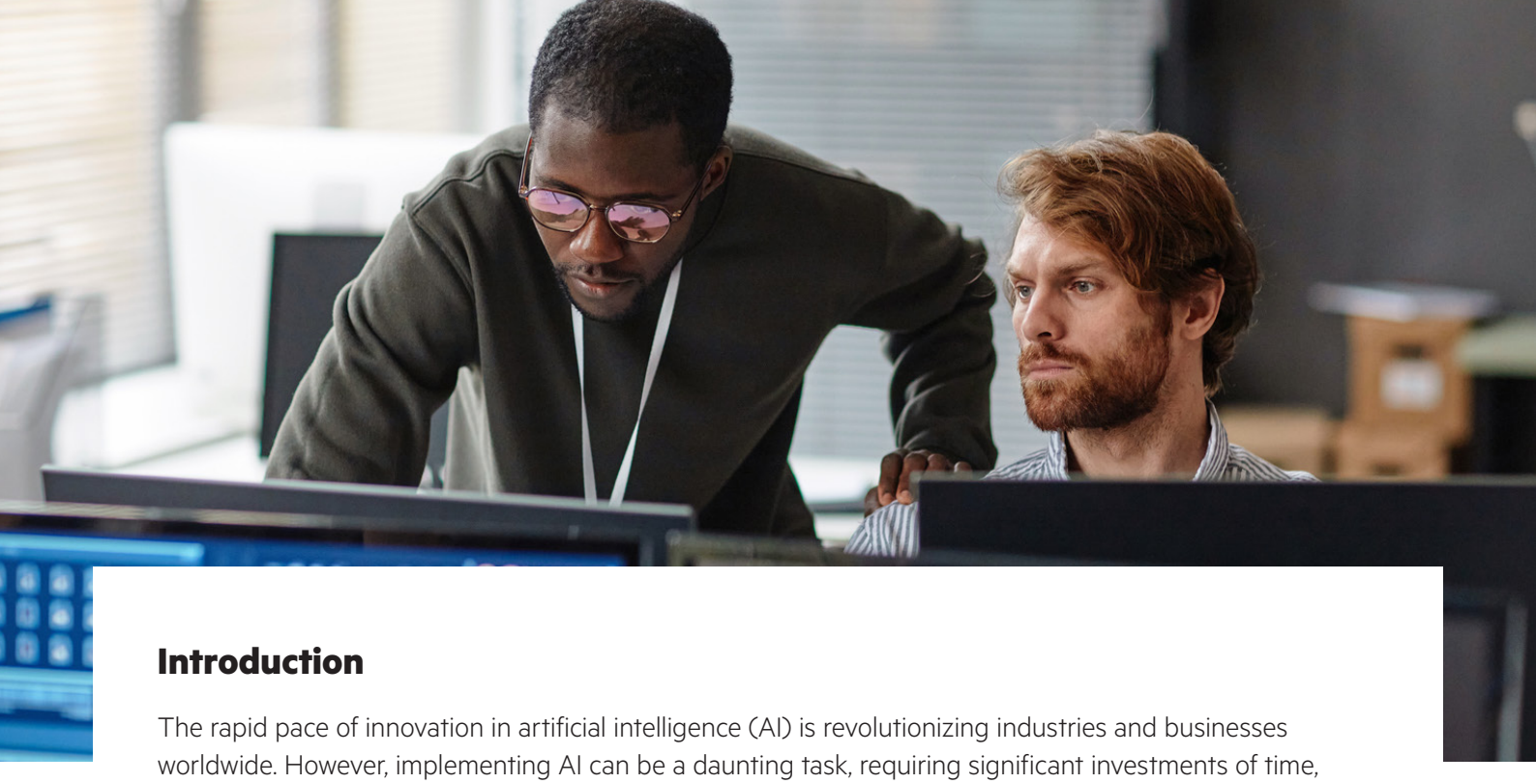
HPE Alletra Storage

HPE Private Cloud AI

Accelerate your AI path with a turnkey AI private cloud

HPE GreenLake



Introduction

The rapid pace of innovation in artificial intelligence (AI) is revolutionizing industries and businesses worldwide. However, implementing AI can be a daunting task, requiring significant investments of time, money, and expertise. That is where HPE Private Cloud AI comes in, a turnkey offering that simplifies the process of deploying AI workloads. When it comes to AI, most organizations face significant challenges. They must integrate multiple components, such as data preparation, model training, and deployment, which can be overwhelming. Moreover, they require specialized expertise, infrastructure, and resources—a tall order for many companies. That is why Hewlett Packard Enterprise has developed HPE Private Cloud AI, which makes AI accessible, manageable, and scalable.

Overview

HPE Private Cloud AI is a comprehensive solution that combines the strengths of HPE Storage expertise with the leadership of NVIDIA® in AI computing. This integrated system provides a secure, scalable, and managed environment for deploying AI workloads. With HPE Private Cloud AI, organizations can easily integrate data from various sources, train and deploy AI models quickly and efficiently, scale their AI infrastructure as needed, and leverage expert services and support.

This solution is designed to meet the needs of three key audiences: data scientists, developers, and IT professionals. Data scientists will appreciate the ease of using popular frameworks to build, train, and deploy AI models. Developers will benefit from the tools and frameworks provided to integrate AI into their applications. IT professionals will find a comprehensive solution for managing and deploying AI workloads.

By deploying HPE Private Cloud AI, organizations can gain a competitive edge in the market. They will be able to accelerate time-to-market with AI-powered applications, improve decision-making with data-driven insights, enhance customer experiences through personalized interactions, and drive innovation and growth through AI-enabled processes. Key benefits of HPE Private Cloud AI include simplified AI deployment and management, scalable infrastructure for growing AI workloads, expert services and support to ensure success, integration with existing IT infrastructure, and enhanced security and compliance features. With HPE Private Cloud AI, organizations can unleash the full potential of AI and drive business success.

T-shirt size configurations

HPE understands that every organization is unique and has different needs when it comes to AI. That is why we offer three T-shirt size configurations—small, medium, and large—each designed to meet specific AI workload requirements. Whether you are a small start-up or a large enterprise, HPE has a solution that fits your needs.

HPE Private Cloud AI solutions come in two granular options for each size: standard and expanded. The main difference between the sizes is the number of AI worker nodes. Table 1 lists the specs for the T-shirt sizes in both standard and expanded options.

Table 1. HPE Private Cloud AI T-shirt size options

T-shirt sizes	Components	Standard	Expanded
Small	Network switches	2x NVIDIA SN4600cM switches (100GbE data network) 2x HPE Aruba Networking 6300M (management and HPE iLO)	2x NVIDIA SN4600cM switches (100GbE data network) 2x HPE Aruba Networking 6300M (management and HPE iLO)
	Control plane	3x HPE ProLiant DL325 Gen11 Nodes	3x HPE ProLiant DL325 Gen11 Nodes
	AI worker nodes	1x HPE ProLiant DL380a Gen11 AI-optimized Node 4x L40s GPUs per node (4 total)	2x HPE ProLiant DL380a Gen11 AI-optimized Node 4x L40s GPUs per node (8 total)
	Storage	109 TB HPE GreenLake for File Storage Based on HPE Alletra MP in standard density 1c x 1d x 2s configuration with 7.68 TB drives	109 TB HPE GreenLake for File Storage Based on HPE Alletra MP in standard density 1c x 1d x 2s configuration with 7.68 TB drives
	Software licensing	HPE AI Essentials with NVIDIA AI Enterprise software 3-year subscription	HPE AI Essentials with NVIDIA AI Enterprise software 3-year subscription
Medium	Network switches	2x NVIDIA SN4600cM switches (100GbE data network) 2x HPE Aruba Networking 6300M (management and HPE iLO)	2x NVIDIA SN4600cM switches (100GbE data network) 2x HPE Aruba Networking 6300M (management and HPE iLO)
	Control plane	3x HPE ProLiant DL325 Gen11 Nodes	3x HPE ProLiant DL325 Gen11 Nodes
	AI worker nodes	2x HPE ProLiant DL380a Gen11 AI-optimized Node 4x L40s GPUs per node (8 total)	4x HPE ProLiant DL380a Gen11 AI-optimized Node 4x L40s GPUs per node (16 total)
	Storage	217 TB HPE GreenLake for File Storage Based on HPE Alletra MP in standard density 1c x 1d x 2s configuration with 15 TB drives	217 TB HPE GreenLake for File Storage Based on HPE Alletra MP in standard density 1c x 1d x 2s configuration with 15 TB drives
	Software licensing	HPE AI Essentials with NVIDIA AI Enterprise software 3-year subscription	HPE AI Essentials with NVIDIA AI Enterprise software 3-year subscription
Large	Network switches	4x NVIDIA SN4600cM switches (100GbE data network) 4x HPE Aruba Networking 6300M (management and HPE iLO)	4x NVIDIA SN4600cM switches (100GbE data network) 4x HPE Aruba Networking 6300M (management and HPE iLO)
	Control plane	3x HPE ProLiant DL325 Gen11 Nodes	3x HPE ProLiant DL325 Gen11 Nodes
	AI worker nodes	4x HPE ProLiant DL380a Gen11 AI-optimized Node 4x H100 NVL GPUs per node (16 total)	8x HPE ProLiant DL380a Gen11 AI-optimized Node 4x H100 NVL GPUs per node (32 total)
	Storage	670 TB HPE GreenLake for File Storage Based on HPE Alletra MP in standard density 3c x 5d x 2s configuration with 7.68 TB drives	670 TB HPE GreenLake for File Storage Based on HPE Alletra MP in standard density 3c x 5d x 2s configuration with 7.68 TB drives
	Software licensing	HPE AI Essentials with NVIDIA AI Enterprise software 3-year subscription	HPE AI Essentials with NVIDIA AI Enterprise software 3-year subscription

This section provides a closer look at the most suitable use cases for each option and describes its corresponding benefits.

Small

The small HPE Private Cloud AI offering is perfect for organizations just starting their AI journey or those looking to test the waters with a proof of concept. This system supports up to 8 GPUs, 109 TB storage, and a 100GbE network, making it an excellent choice for small-scale AI projects. With use cases such as inferencing, prototyping, or small-scale AI projects, this size option is ideal for organizations looking to explore the world of AI without going over budget.

Benefits:

- Quick setup and deployment
- Low up-front costs
- Easy scalability for future growth

Medium

The medium size is suitable for midsized organizations or teams with moderate AI workloads. This system supports twice the AI worker nodes, a larger storage configuration, and a 200GbE network, making it an excellent choice for retrieval-augmented generation (RAG) use cases. With benefits such as improved performance and scalability as compared to small HPE Private Cloud AI, this size option is ideal for organizations that want to take their AI workloads to the next level.

Benefits:

- Improved performance and scalability
- Easy management experience through the HPE GreenLake cloud

Large

The large HPE Private Cloud AI offering is designed for larger-scale AI deployments such as fine-tuning. This system features more powerful GPUs, four times the AI worker nodes, 400 Gbps internal bandwidth switch network, and six times the storage deployment. With use cases such as fine-tuning, data analytics, or other more complex AI workloads, this size option is ideal for organizations that want to tackle demanding AI projects.

Benefits:

- High performance processing capabilities for demanding AI workloads
- Scalable architecture for easy expansion to meet growing demands

Hardware architecture

The building blocks of the HPE Private Cloud AI hardware architecture include management and data networks, control plane servers, storage, and AI worker nodes. This section describes the details of each component, as well as the purposes it serves. Figure 1 provides a high-level view of how all the components are connected.

Management network

The management network is composed of redundant 1 Gbps network switches for management ports and HPE iLO ports connectivity for all the components in HPE Private Cloud AI, including data network switches, control plane servers, AI worker node servers, and storage. The management network switches should be uplinked into the customer's own network for access.

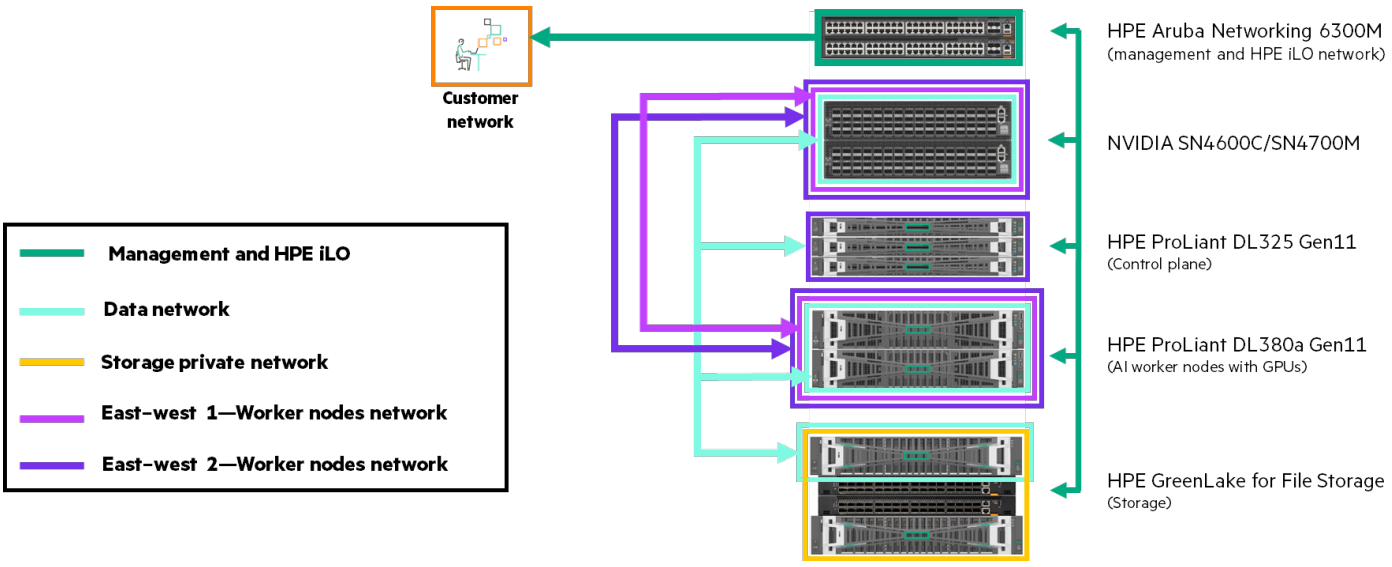


Figure 1. High-level network connections of HPE Private Cloud AI

Data network

Data network switches are pairs of redundant 100GbE network switches that connect the control plane servers and the AI worker nodes to the storage controller nodes for high-bandwidth network access to the data stored on HPE GreenLake for File Storage. Notice that the east–west traffic among multiple AI worker nodes is also supported by the same data network switches, but the traffic for node-to-node communication is on a dedicated network.

Each HPE ProLiant DL380a server comes with a dual-port 100GbE HBA card installed, which enables you to connect two ports to the data network switches. Each HPE ProLiant DL325 Gen11 control plane server also comes with a dual-port 100GbE HBA card installed, which enables you to connect two ports per server to the data network switches.

The storage system, HPE GreenLake for File Storage, is connected to the same data network through the dual-port 100GbE HBAs. Each CBox of the HPE GreenLake for File Storage contains two of the dual-port 100GbE HBAs, which enable four ports on the redundant data network switches. The data network excludes the network traffic between the HPE GreenLake for File Storage controller nodes and just a bunch of flashes (JBOFs). The internal network runs strictly over the internal switch pair (HPE Aruba Networking 8325), which is not connected to the HPE Private Cloud AI data network.

Control plane servers

For the control plane servers, the system uses three HPE ProLiant DL325 Gen11 Servers as the VMware ESXi™ servers supporting a VMware vCenter® cluster. The control plane mainly services the control plane Kubernetes cluster that orchestrates one or more AI workload clusters hosted on the AI worker nodes. It manages common services for the AI workload clusters, such as service mesh, cert-manager, identity broker, log collector, and container registry.

The control plane also provides the two-way tunnel connection to the HPE GreenLake for managing HPE Private Cloud AI through a VM running in the vCenter, the data services connector (DSC) VM.

The vCenter consumes storage space from HPE GreenLake for File Storage through the container storage interface (CSI) driver to provision PersistentVolumeClaims (PVCs) in the Kubernetes cluster and the datastore created from mounted Network File System (NFS) shares.

Storage

Storage of HPE Private Cloud AI is provided by HPE GreenLake for File Storage. The storage system consists of controller nodes in CBoxes, a redundant 100GbE HPE Aruba Networking 8325 switch pair, and JBOFs. HPE GreenLake for File Storage adopts a Disaggregated Shared-Everything architecture that enables the controller nodes to be stateless and connected to the JBOFs through an NVMe-oF network. This architecture isolates the storage system's 100GbE internal network from the HPE Private Cloud AI data network.

HPE GreenLake for File Storage provides storage by including NFS shares and S3 endpoints. NFS storage is consumed by both ESXi and the Kubernetes services to provide NFS datastores for VM hosting and PersistentVolume support through the CSI driver. S3 is consumed predominantly by the Kubernetes service by orchestrating the bucket lifecycle on a precreated S3 endpoint.

AI worker nodes

AI worker nodes are the hosts for the AI workload Kubernetes clusters. Kubernetes clusters are deployed, managed, and destroyed by the HPE Private Cloud AI control plane. Each worker node is equipped with GPUs for parallel computation and the capability for RDMA and GDS connectivity to the storage system.

Software architecture

Since you can deploy HPE Private Cloud AI with an AI application in just three clicks, it's easy to assume that not much is going on behind the scenes. However, a complex software stack is at work, making it all possible. Let's take a closer look at what's happening beneath the surface.

Orchestration

The software stack starts with the orchestration and management piece, which contains three ESXi servers. Within this environment, a Kubernetes cluster instance runs the HPE Private Cloud AI control plane hosted by VMs within the vCenter/ESXi cluster, which is responsible for managing the worker node clusters. In addition, Data Services Cloud Console connectivity to and management of this environment requires separate VMs, called the DSC VMs.

Next, there are the worker nodes, also known as GPU hosts, where the AI workload (Kubernetes) clusters are deployed and managed by the HPE Private Cloud AI control plane. Within these clusters, HPE AI Essentials software packages are deployed and run.

Finally, HPE GreenLake for File Storage ties everything together, providing storage for both the HPE Private Cloud AI control plane and the AI worker nodes. Figure 2 shows a high-level view of the software stack of HPE Private Cloud AI.

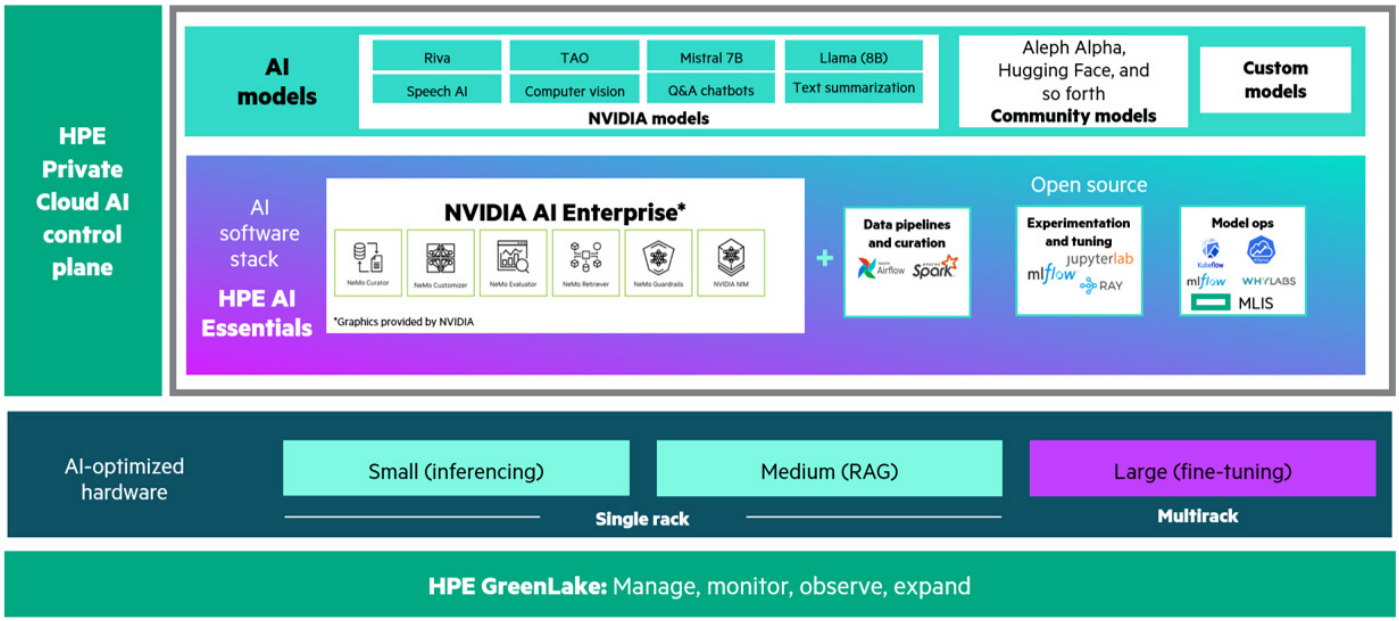


Figure 2. A high-level view of the software stack of HPE Private Cloud AI

HPE AI Essentials

HPE AI Essentials is a set of software tools and frameworks that enable customers to develop and deploy AI models. It provides a cloud-like experience while keeping the data private. These are some of the key features of HPE AI Essentials:

- **NVIDIA AI Enterprise:** Provides prestaged models and NIMs that support pretrained large language models (LLMs) such as Llama (8B and 70B tokens), Text Embeddings Inference service, and reranking services. These prestaged models enable end users to simply deploy an LLM or a RAG instance with a few clicks, without the steep learning curve of setting up the environment
- **Model development:** Helps developers build, train, and deploy AI models by using a variety of programming languages and tools. Data scientists or engineers can easily browse through the catalog and find tools such as MLflow, Ray, Feast, HPE Machine Learning Development Environment Software, and Spark Operator for developing or fine-tuning AI models. In contrast to traditional ways of deployment, end users do not have to learn how to deploy applications, for example, deploying Spark applications from the CLI; HPE AI Essentials simplifies the process by providing an abstraction layer or a UX that enables them to deploy the applications while saving time and effort on AI development
- **Data engineering:** Provides data management capabilities for storing, processing, and analyzing large datasets. You can connect to existing external data sources or choose to create new ones. Data engineering provides support for various unstructured data sources. These might be in the forms of S3-compatible data sources and files or directories that are provisioned through the CSI driver. It also supports structured data sources. You can browse, query, and edit database items such as SQL databases or S3 objects directly from the HPE AI Essential GUI
- **Observability:** Provides auditing, alerting, logging, and metering capabilities to track model performance and detect issues
- **Security:** Includes security features such as SSO integration, user management, access controls, and auditing. It also supports integrated Jupyter Notebook for securely accessing other services and data pipelines

Setup and manageability

The HPE Private Cloud AI solution offers a turnkey process for customers who want to simplify the overall setup experience at their site.

Factory setup

This process involves installing ESXi on three HPE ProLiant DL325 nodes, similar to the HPE GreenLake for Private Cloud Business Edition process familiar to those with previous setup experience. As part of this process, the DSC VM is deployed on each node, and a Rocky Linux® Open Virtual Appliance (OVA) image is uploaded to deploy the HPE Private Cloud AI control plane. This marks the first stage of the setup process.

Factory express and integration center

The stack then moves to the factory express and integration center for further processing. At this stage, Red Hat® OS is installed on the AI worker nodes (GPU hosts) on the HPE ProLiant DL380a. In addition, the HPE Private Cloud AI control plane OVA is uploaded onto HPE GreenLake for File Storage. The switches, servers, and default switch configurations are applied to the Mellanox and HPE Aruba Networking switches.

The power distribution unit (PDU) is also configured because the rack is assembled and cabled like a full system. The HPE GreenLake for File Storage system is racked, stacked, and cabled, with some preconfiguration as well.

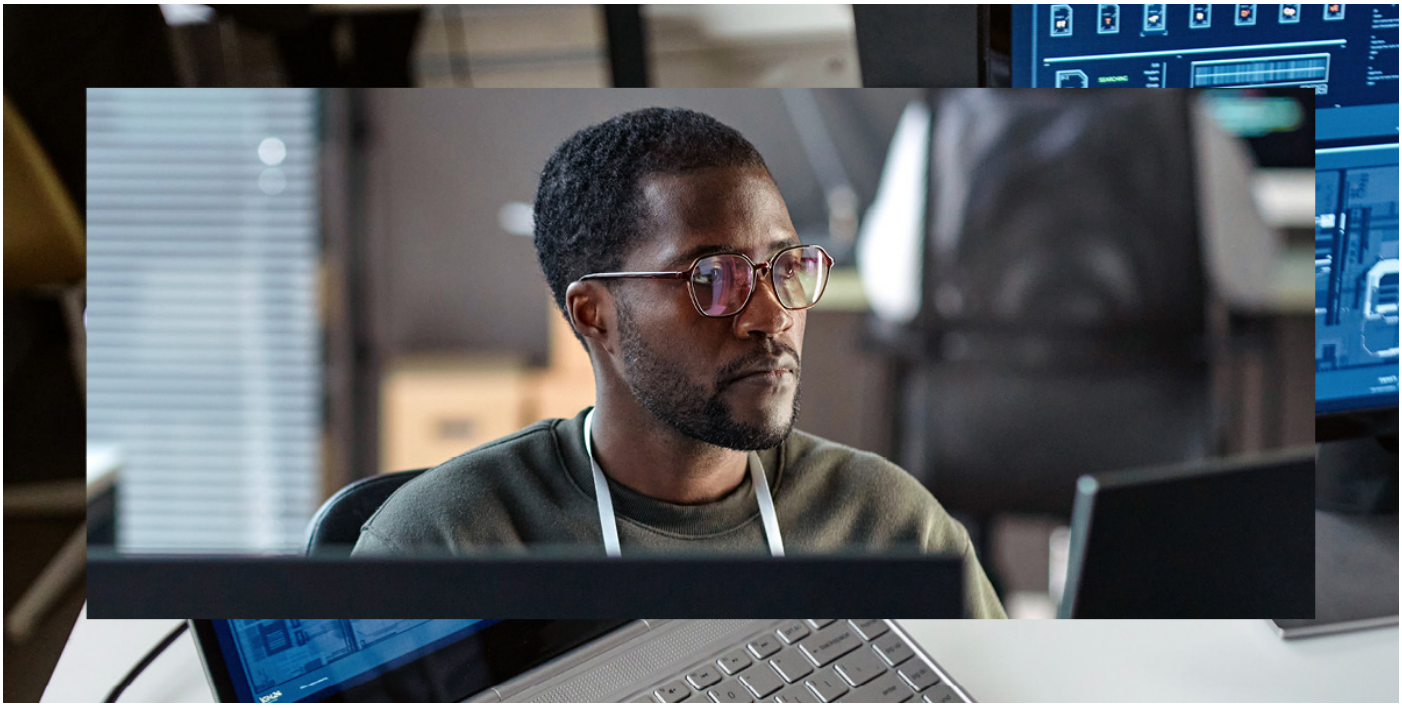
Day 0 setup

Upon receipt of the rack, HPE Service personnel unbox it, move it to its designated location, power it on, and apply prerequisite configurations before setting up the system. One such prerequisite is to help ensure that the Mellanox switches are interconnected with the customer's network, enabling outbound connectivity to HPE GreenLake endpoints.

The setup process consists of three stages:

1. **Initial setup:** This stage involves connecting DSC VM and HPE GreenLake for File Storage to connect to the HPE GreenLake cloud, initializing the DSC VM, and performing a few necessary configurations.
2. **HPE Private Cloud AI setup service:** This service guides customers and HPE Services teams through a full deployment of the stack, enabling them to start consuming HPE Private Cloud AI both from the cloud and on-premises.
   2.1. Infrastructure setup: This stage involves deploying vCenter; creating a VMware vSphere® cluster; configuring the ESXi vSwitch, the VMK, and the HPE iLO IP address; and creating an NFS datastore for the content library containing HPE Private Cloud AI control plane OVAs.
   2.2. AI worker node cluster setup: During this process, HPE Private Cloud AI control plane VMs are deployed, and configurations are applied for static IPs, HPE iLO, logins, and SSH keys. The AI worker nodes are then configured similarly, and the worker node cluster is deployed to orchestrate AI / machine learning (ML) tools and frameworks.
3. **Completion:** The system is ready for AI application deployment and AI workloads.

At this point, customers can deploy AI applications and begin using their HPE Private Cloud AI solution. On Day 1 and beyond, they will be able to manage their HPE Private Cloud AI infrastructure from an administrative standpoint by accessing the platform of HPE GreenLake through the HPE Private Cloud AI tile.



Conclusion

HPE Private Cloud AI is a game-changing solution that simplifies the process of deploying AI workloads. By providing a secure, scalable, and managed environment for AI deployment, HPE is empowering organizations to unleash the full potential of AI.

Learn more at

HPE Private Cloud AI

Visit **HPE GreenLake**

Chat now (sales)

a00143345ENW

Hewlett Packard Enterprise