Overview

HPE GPU Cloud Service

HPE GPU Cloud Service is the public cloud for HPC and AI: a public cloud service for GPU-accelerated workloads with performance at scale. You can reserve compute clusters with a high-speed network and distributed storage in data centers in the U.S. and Canada. Cluster size and contract duration are flexible to accommodate your AI and high-performance computing (HPC) workloads and scale to your needs. High-performance scratch storage and resilient general-purpose storage are supported for cluster-wide access. Free data transfer lets you predict the cost of the service over the whole contracted time. The solution is built on HPE's industry-leading HPC platforms, exclusively with direct liquid cooling and powered by renewable energy sources.

Cloud Management Platform

You can use our performance-optimized cloud management platform to partition your reservation of compute nodes into Kubernetes or traditional HPC clusters. The platform provides a web application and exposes a REST API covering cluster management, storage partitioning/ assignment, and user management for administrators. Users can access their compute clusters through a Rancher instance or its Kubernetes API. The solution is highly flexible to run cloud-native and traditional HPC workloads in a customized environment using workload managers, access to jump boxes, and infrastructure as code providers such as Terraform. We've developed a thin virtualization layer with pass-through GPU, network, and storage access to balance performance and security requirements in an optimal way.

Workloads and Use Cases

- AI Model development and training
  - Data curation and pipeline development
  - LLM training at scale
  - LLM customization and fine-tuning
  - Advanced Driver Assistance Systems (ADAS)
- AI Application deployment and inference
  - Predictive Analytics
  - Computer Vision
  - Natural Language Processing
  - LLM chat-bots & API services
- Traditional HPC workloads
  - Computer aided engineering (CAE)
  - Finite element analysis (FEA)
  - Electronic design automation (EDA)
  - Scientific simulations

Standard Features

Key Technical Features

- GPU compute nodes with NVIDIA® H100 SXM, B200 SXM in preparation
- 8-way InfiniBand NDR fabric with rail-optimized fat-tree topology
- 100+ Gbit/s Ethernet backbone network
- Scratch storage with HPC-grade performance characteristics
- Home storage with S3 access and no ingress/egress fees
- Performance-optimized cloud platform with GUI and REST API
- Kubernetes clusters with SUSE Rancher GUI and API access
- Traditional HPC cluster experience with login node and Slurm workload manager
- No node sharing
- Direct liquid cooling
- Powered by renewable energy

Notes:

- Traditional HPC clusters will be supported later in 2025.
- High-performance scratch storage will be supported later in 2025

The core feature set includes following cloud management platform features (GUI & API):

- User access through HPE login with single-sign-on
- User management, including user groups.
- Managed cluster creation and deletion
- Storage partitioning, storage assignment
- Observability features through Prometheus / Grafana stack
- Secure Layer-2 network isolation encryption.
- Support for GPUDirect RDMA on all nodes

Software services and solution enablement on HPE GPU Cloud Service:

- Scalable Batch service (beta)
- Function as a service (beta)
- Queue as a service (beta)
- In-memory Python distributed dictionary (beta)
- 3rd party software solutions documented:
  - JupyterHub, Pytorch, HuggingFace, KServe, KubeFlow, Knative
  - Tensorflow, Keras, Dask, Ray, Argo, Airflow

Notes: Some 3rd party software will be supported later in 2025.

Service and Support

HPE GPU Cloud Service Support offers a 24x7 Service Desk via phone, email, and service portal to respond and resolve incidents related to the Support Environment (hardware, management platform). Incident Management involves responding promptly to service disruptions using a tiered support process led by experience professionals. Incidents are addressed based on the severity and impact of the issue in the Supported Environment. HPE GPU Cloud Service will oversee an incident from start to finish, which includes registering, categorizing, prioritizing, investigating, diagnosing, resolving, and closing the incident. Additionally, HPE GPU Cloud Service will provide hardware remediation support.

Service	Support
Technical Support Access	24x7
Service Availability	99%
Initial Response	**Severity 1:** 1 Hour
	Severity 2: 3 Hours
	Severity 3: 1 Business Day
	Severity 4: 3 Business Days
Service Request	3 Business Days
Access to Customer Portal	24x7
Phone	24x7
Product Coverage	Hardware/Storage/Networking/Cloud Platform*

Notes: *Excluded on Bare Metal offerings

Configuration Information

Compute

Access to high-performance GPU instances can be ordered alongside general-purpose compute virtual machines. All performance instances support GPUDirect RDMA via pass-through.

General-purpose compute VMs are run on shared infrastructure and are virtualized with conventional techniques. They are intended to run auxiliary tasks on the tenant Kubernetes cluster, like databases. These nodes are marked “SVC” and can be ordered starting with 4 cores.

For a comprehensive list of available instances and their specification see docs.aicloud.hpe.com under “Clusters/Instances”.

Instance Name	GPU	# of GPUs	CPU	CPU Cores	Memory	Local Storage Capacity
H100-SPR-64-2048	NVIDIA H100 SXM	8	Intel Xeon-Platinum 8462Y+	64	2TB	30.72TB
SVC-SPR-64-1024	None	N/A	Intel Xeon-Platinum 8462Y+	64	1TB	9.6TB

The following durations are available:

12months	18 months	24 months	36 months	60 months

Long durations are rebated. POCs shorter than 6 months are available on request at gpus@hpe.com

The following regions are available:

- que1 (Quebec, Canada)

Storage

Resilient general-purpose (/home) storage with S3 access is available in the following durations:

12months	18 months	24 months	36 months	60 months

The minimum order size 1TB. A general-purpose storage allocation must be ordered and be available before, during and after all compute allocations. Data transfer to and from the general-purpose storage via S3 is free of charge.

High-performance (/scratch) storage is available in the following durations:

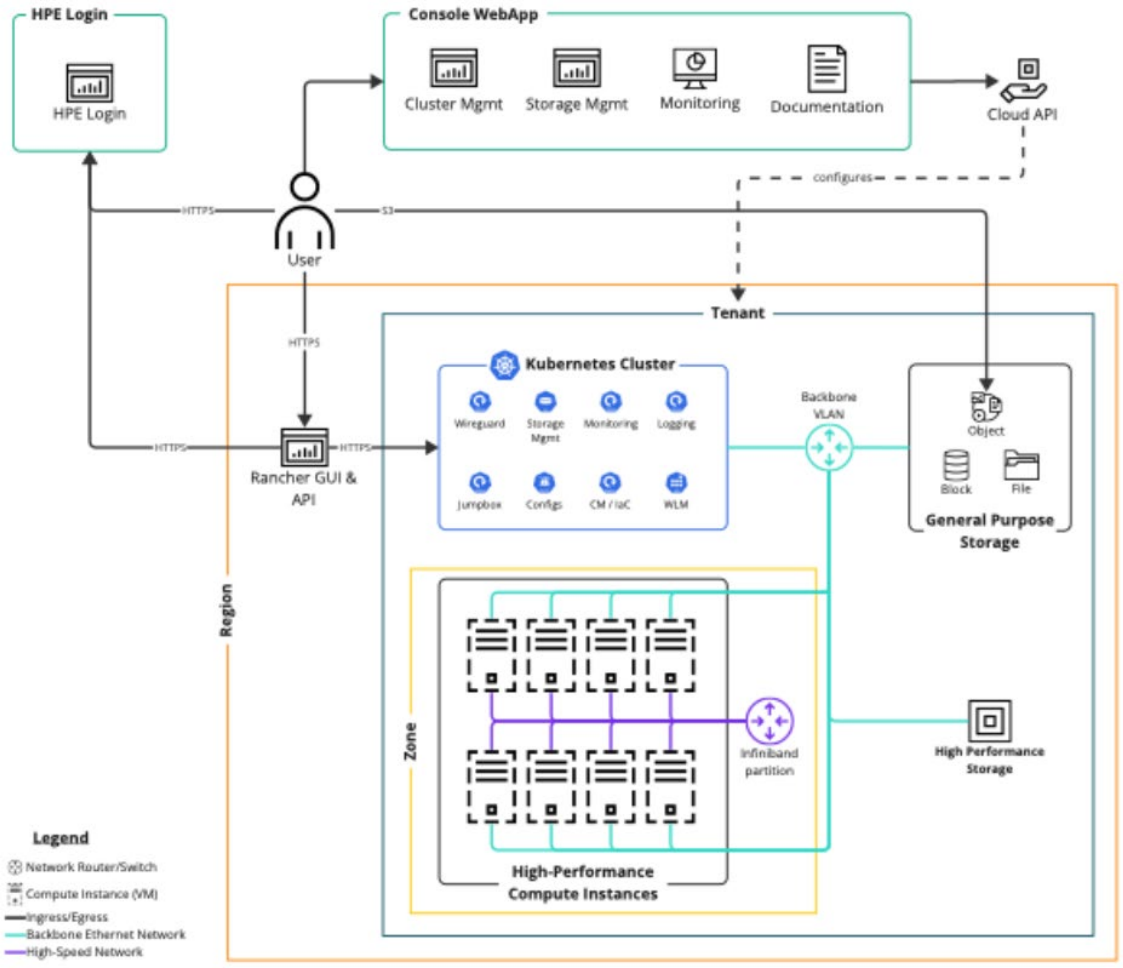
12months	18 months	24 months	36 months	60 months

For technical properties of the storage types see docs.aicloud.hpe.com under “Storage”.

Notes:

- CPU compute nodes and B200-based compute nodes will be available later in 2025.
- High-performance scratch storage will be available later in 2025.

Technical Specifications



The architecture diagram above shows the following components:

- The User interacting with other components remotely through various APIs and web applications
- The HPE Login that users need to use to access the cloud remotely
- The Console Webapp and Cloud API that users and admin us to setup and change clusters, storage allocations, monitor the usage of the system and access inline documentation.
- The Region that represents a (part of) a datacenter (see list of regions in “Configuration Information”)
- The Tenant that represents all allocated resources from one customer at this region. The tenant is configured through the Cloud API.
- The Rancher GUI and API that users access to configure their Kubernetes cluster
- The Kubernetes Cluster run on general-purpose compute to enable custom use cases.
- The High-Performance Compute Instances that can be integrated with the Kubernetes Cluster (cloud-native use case) or external to the Kubernetes cluster (traditional HPC cluster). These instances are GPUDirect RDMA capable and run our performance optimized virtualization layer with pass-through.
- The General-Purpose Storage that can be configured to expose object, file and block storage devices. It is reachable from the outside through the S3 interface. Data transfer is free of charge.
- The High-Performance Storage that can be added to a compute cluster exposing a file interface.

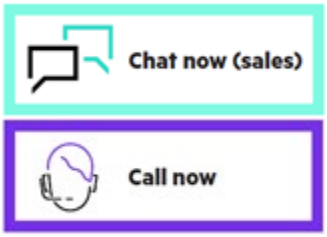
More technical details are available at docs.aicloud.hpe.com.

Summary of Changes

Date	Version History	Action	Description of Change
02-Jun-2025	Version 2	Changed	H200 -> B200, simplify some language
03-Mar-2025	Version 1	New	New QuickSpecs

Copyright

Make the right purchase decision.
Contact our presales specialists.



Get updates

Hewlett Packard Enterprise

The information contained herein is subject to change without notice. The only warranties for Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.
© Copyright 2025 Hewlett Packard Enterprise Development Company, L.P.

a50007001enw - 17111 - Worldwide - V2 - 02-June-2025