Overview

HPE Machine Learning Inference Software

HPE Machine Learning Inference Software is a streamlined solution designed to expedite the deployment and scaling of AI models and applications and is particularly tailored for generative AI. By simplifying the process of model deployment, management, and monitoring, HPE Machine Learning Inference Software bridges the gap between AI model development and productization, enabling ML teams to execute and achieve value faster.

The new HPE Machine Learning Inference Software offers a simplified, user-friendly approach for managing, and monitoring machine learning deployments, with a particular strength in handling Large Language Models (LLMs). Designed for both ML practitioners and IT infrastructure engineers, this software minimizes the complexity and specialized expertise needed to efficiently launch and update ML models. It also provides robust tools for monitoring the performance of these models, ensuring they operate at scale.

With this software, Machine Learning Engineers can deploy models for experimental purposes enabling rapid iteration, which is essential for understanding model behavior, visualization, and debugging. Due to the intuitive interface users do not need extensive Kubernetes experience, creating a simplified path to production model deployments for everyone. ITOps/MLOps can take the models developed by the MLE team and deploy them quickly at scale. HPE Machine learning Inference Software has a broad range of supported frameworks, automation to help you deploy swiftly, and is suitable for both development and production environments. Users can manage their model deployments with efficient load balancing for better scalability, consistent performance, and robust error handling with automated testing.

HPE Machine Learning Inference Software is the third and final component completing the vision "…to build the world's best platform for developing and deploying AI applications at scale", along with **HPE Machine Learning Data Management Software** and **HPE Machine Learning Development Environment Software**. HPE Machine Learning Inference Software is differentiated by embracing flexibility and cost efficiency and shunning proprietary tools and mitigating escalating costs. It avoids vendor lock-in, fostering seamless multi-cloud compatibility.

HPE Machine Learning Inference Software does not require extensive education as the platform is intuitive and easy to learn. No need to assemble multiple tools to start your AI journey – HPE Machine Learning Inference Software provides comprehensive solutions tailored to customer needs and, robust integrations with open source and proprietary tools ensure resilient and secure operations, while our adaptive approach effortlessly accommodates technological advancements.

With HPE Machine Learning Inference Software, you're not confined to a single focus; it offers versatile solutions for diverse model deployment requirements and integration with NVIDIA Inference Microservices (NIM) to provide optimized inference for more than two dozen popular downloadable AI models from NVIDIA and its partner ecosystem.

Product Benefits

- Deploy AI/ML models to the datacenter or cloud
- Enterprise-grade security with RBAC and secure authentication for endpoints.
- Seamless integration with NVIDIA NIM
- Diverse framework compatibility for models you create or import.
- Workload optimization for underlying hardware, eliminating manual tuning
- Comprehensive monitoring and management for AI model health and performance
- Streamlined setup via industry-standard Helm charts

Standard Features

Seamless integration with NVIDIA Inference Microservices (NIM)

Leverages NVIDIA'S optimized inference microservices for efficient model deployment at scale, offering unparalleled inference speeds, reducing time to insights. Pre-trained models from NVIDIA and their partner ecosystem can directly imported and used without further optimization.

Diverse Model Framework Compatibility

Facilitates using models from diverse frameworks such as TensorFlow, PyTorch, scikit-learn, and XGBoost, accommodating a broad range of pre-trained and custom models.

Model Packaging

Streamlined integration with Hugging Face and NVIDIA Foundation Models offers a zero-coding deployment experience for large language models (LLMs) directly from Hugging Face and NVIDIA NGC.

Customized models can utilize built-in containerization tools to ensure a streamlined, consistent, and version-controlled deployment process.

Monitoring and Management

Benefit from integrated monitoring and logging for tracking model performance, usage metrics, and system health, facilitating proactive optimization.

Security

Execute workloads within your preferred environment, thus ensuring the security of your models, code, and data. Implement Role-Based Access Controls (RBAC) to securely manage how development and MLOps teams access and share machine learning resources and artifacts.

Enable authentication on your model endpoints to manage access to your inference services including OIDC and OAuth 2.0 with flexible integration options.

Versatile Deployment

Compatible with Kubernetes-supported environments, including major cloud services like GreenLake, AWS, Azure, Google Cloud, and on-premises setups, ensuring flexible integration across diverse infrastructures.

Seamless Scaling

Handle deployment scaling based on the requested load to meet your needs, utilizing built-in mechanisms to automatically scale containers and manage traffic.

Configuration Information

Product Licenses

HPE Machine Learning Inference Software must be licensed to allow software download and support from HPE.

Licensing and Renewals follow a two-tier model:

- Base License: Licenses the installation of HPE Machine Learning Inference Software on an OS instance (physical, virtual, or cloud based) and includes 1 Performance license.
- Performance License: Licenses GPU/Accelerator for use with HPE Machine Learning Inference Software

If the licensing rules are unclear for your specific usage, please contact the local HPE or authorized partner sales representative before ordering.

Step 1 – Calculate the number of Base licenses:

- One base license is required per OS instance (a single running install of an OS, whether physical, virtual, or cloud based)
- Select the appropriate tenure of subscription for that enterprise license (1-, 3-, 4- and 5-year terms are available).
- Customers can acquire their entitlements certificates from the HPE Software Portal.
- Base licenses include 1 performance license

Description	SKU
Base License (1 per OS instance)	
HPE ML Inference SW Base 1yr E-RTU	S3R04AAE
HPE ML Inference SW Base 3yr E-RTU	S3R06AAE
HPE Machine Learning Inference Software Base 4-year E-RTU	S3W26AAE
HPE Machine Learning Inference Software Base 5-year E-RTU	S3W28AAE

Step 2 – Calculate the number of Performance Licenses needed:

- HPE Machine Learning Inference Software licenses GPU/Accelerators for use with HPE Machine Learning Inference Software. Utilize the chart below to determine the number of licenses necessary per accelerator.
- Select the appropriate tenure of license (1-, 3-, 4- and 5-year terms are available).

Description	SKU
HPE ML Inference SW Perf 1yr E-RTU	S3R05AAE
HPE ML Inference SW Perf 3yr E-RTU	S3R07AAE
HPE Machine Learning Inference Software Performance 4-year E-RTU	S3W27AAE
HPE Machine Learning Inference Software Performance 5-year E-RTU	S3W29AAE

Step 3 (optional) Purchase NVIDIA AI Enterprise licenses

- HPE recommends integrating with NVIDIA AI Enterprise, including NVIDIA Inference Microservices (NIM), which optimizes inference for dozens of popular AI models from NVIDIA and its partner ecosystem. Some pre-built models are optimized for specific GPUs. Please reference NVIDIAs **website** for further details.

Description	**SKU**
NVIDIA AI Enterprise Essentials per GPU 1-year Subscription 9x5 Support E-LTU	S2S16AAE
NVIDIA AI Enterprise Essentials per GPU 3-year Subscription 9x5 Support E-LTU	S2S17AAE
NVIDIA AI Enterprise Essentials per GPU 5-year Subscription 9x5 Support E-LTU	S2S18AAE
NVIDIA AI Enterprise Essentials per GPU 1-year 24x7 Support E-LTU	S2S28AAE
NVIDIA AI Enterprise Essentials per GPU 3-year 24x7 Support E-LTU	S2S29AAE
NVIDIA AI Enterprise Essentials per GPU 5-year 24x7 Support E-LTU	S2S30AAE

Service and Support

HPE Support

HPE Tech Care Service enables direct access to product-specific specialists and provides general technical guidance to help customers not only reduce risk but also find ways to do things more efficiently. HPE Tech Care Service Customers can access support through multiple channels that include telephone, a real-time chat facility, automated incident logging, and HPE moderated forums with defined response times.

HPE 1Y Tech Care Basic Service

- Expert access: 9x5, 2hr response
- Outage SLA: Standard elevation

HPE Services

No matter where you are in your digital transformation journey, you can count on HPE Services to deliver the expertise you need when, where and how you need it. From planning to deployment, ongoing operations and beyond, our experts can help you realize your digital ambitions.

https://www.hpe.com/services

Consulting Services

No matter where you are in your journey to hybrid cloud, experts can help you map out your next steps. From determining what workloads should live where, to handling governance and compliance, to managing costs, our experts can help you optimize your operations.

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

HPE Managed Services

HPE runs your IT operations, providing services that monitor, operate, and optimize your infrastructure and applications, delivered consistently and globally to give you unified control and let you focus on innovation.

HPE Managed Services | HPE

Operational services

Optimize your entire IT environment and drive innovation. Manage day-to-day IT operational tasks while freeing up valuable time and resources. Meet service-level targets and business objectives with features designed to drive better business outcomes.

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

HPE Lifecycle Services

HPE Lifecycle Services provide a variety of options to help maintain your HPE systems and solutions at all stages of the product lifecycle. A few popular examples include:

- Lifecycle Install and Startup Services: Various levels for physical installation and power on, remote access setup, installation and startup, and enhanced installation services with the operating system.
- HPE Firmware Update Analysis Service: Recommendations for firmware revision levels for selected HPE products, taking into account the relevant revision dependencies within your IT environment.
- HPE Firmware Update Implementation Service: Implementation of firmware updates for selected HPE server, storage, and solution products, taking into account the relevant revision dependencies within your IT environment.
- Implementation assistance services: Highly trained technical service specialists to assist you with a variety of activities, ranging from design, implementation, and platform deployment to consolidation, migration, project management, and onsite technical forums.
- HPE Service Credits: Access to prepaid services for flexibility to choose from a variety of specialized service activities, including assessments, performance maintenance reviews, firmware management, professional services, and operational best practices.

Notes: To review the list of Lifecycle Services available for your product go to:

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

For a list of the most frequently purchased services using service credits, see the **HPE Service Credits Menu**

Service and Support

Other Related Services from HPE Services:

HPE Education Services

Training and certification designed for IT and business professionals across all industries. Broad catalogue of course offerings to expand skills and proficiencies in topics ranging from cloud and cybersecurity to AI and DevOps. Create learning paths to expand proficiency in a specific subject. Schedule training in a way that works best for your business with flexible continuous learning options. **https://www.hpe.com/services/training**

How to Purchase Services

Services are sold by Hewlett Packard Enterprise and Hewlett Packard Enterprise Authorized Service Partners:

- Services for customers purchasing from HPE or an enterprise reseller are quoted using HPE order configuration tools.
- Customers purchasing from a commercial reseller can find services at **https://ssc.hpe.com/portal/site/ssc/**

AI Powered and Digitally Enabled Support Experience

Achieve faster time to resolution with access to product-specific resources and expertise through a digital and data driven customer experience

Sign into the HPE Support Center experience, featuring streamlined self-serve case creation and management capabilities with inline knowledge recommendations. You will also find personalized task alerts and powerful troubleshooting support through an intelligent virtual agent with seamless transition when needed to a live support agent.
https://support.hpe.com/hpesc/public/home/signin

Consume IT On Your Terms

HPE GreenLake edge-to-cloud platform brings the cloud experience directly to your apps and data wherever they are—the edge, colocations, or your data center. It delivers cloud services for on-premises IT infrastructure specifically tailored to your most demanding workloads. With a pay-per-use, scalable, point-and-click self-service experience that is managed for you, HPE GreenLake edge-to-cloud platform accelerates digital transformation in a distributed, edge-to-cloud world.

- Get faster time to market
- Save on TCO, align costs to business
- Scale quickly, meet unpredictable demand
- Simplify IT operations across your data centers and clouds

To learn more about HPE Services, please contact your Hewlett Packard Enterprise sales representative or Hewlett Packard Enterprise Authorized Channel Partner. Contact information for a representative in your area can be found at "Contact HPE"
https://www.hpe.com/us/en/contact-hpe.html

For more information
http://www.hpe.com/services

Technical Specifications

Supported NVIDIA GPUs and required Performance licenses per accelerator.

Accelerator	# of Performance Licenses required
NVIDIA V100 Tensor core GPU	1
NVIDIA V100S Tensor core GPU	1
NVIDIA T4 Tensor core GPU	1
NVIDIA A2 Tensor core GPU	1
NVIDIA A16 GPU Accelerator	1
NVIDIA L4 Tensor core GPU	2
NVIDIA L40 GPU Accelerator	2
NVIDIA A10 Tensor core GPU	2
NVIDIA A30 Tensor core GPU	2
NVIDIA A30x Tensor core GPU	2
NVIDIA A40 GPU Accelerator	2
NVIDIA L40S GPU Accelerator	4
NVIDIA A100 PCI Tensor core GPU	4
NVIDIA A100 SXM Tensor core GPU	4
NVIDIA H100 94GB PCIE Tensor core GPU	8
NVIDIA H100 80GB PCIE Tensor core GPU	8
NVIDIA H100 94GB SXM Tensor core GPU	10
NVIDIA H100 80GB SXM Tensor core GPU	10
NVIDIA H100 64GB SXM Tensor core GPU	10
NVIDIA H200 Tensor core GPU	10

Notes: Performance licenses are calculated using a GPUs FP16 score, sparsity disabled, rounded up and divided by 100(subject to change as technology evolves)

Supported Environment

HPE Machine Learning Inference Software supports any Kubernetes-compatible platform on physical, virtual, or cloud-based instances, including **HPE Ezmeral**, **Red Hat OpenShift**, **SUSE Rancher, Amazon Web Services Elastic Kubernetes Service(EKS)**, **Microsoft Azure Managed Kubernetes Service(AKS),** and **Google Kubernetes Engine (GKE)**. The minimum dependencies are **Kubernetes** cluster 1.2 or newer, **Helm** 3.0 or newer, and **KServe** 0.11. HPE Machine Learning Inference Software supports NVIDIA accelerators for the Ampere generation and newer

Summary of Changes

Date	Version History	Action	Description of Change
18-Jun-2024	Version 1	New	New QuickSpecs

Copyright

Make the right purchase decision.
Contact our presales specialists.

Get updates

a50009204enw - 17236 - Worldwide - V1 - 18-June-2024

Hewlett Packard Enterprise