



HPE Services

The HPE Private Cloud AI experience



Hewlett Packard
Enterprise

HPE Private Cloud AI is a purpose-built solution designed to provide fast and easy deployment of private AI applications with a focus on inferencing, Retrieval-Augmented Generation (RAG), and fine-tuning. HPE Private Cloud AI is a codeveloped Hewlett Packard Enterprise and NVIDIA enterprise purpose-built solution, including a completed infrastructure, software portfolio, and AI model library managed through HPE GreenLake to offer a **private AI in a box**. HPE Private Cloud AI offers enterprise customers the ability to leverage NVIDIA® AI Enterprise (NVAIE) portfolio, including NVIDIA Inferencing Microservices (NIM), and HPE portfolio of curated market-adopted open-source AI tools and platforms with full private control of their data.

HPE's comprehensive solution provides instant AI productivity, enterprise-grade confidence and control, secure and unified data access, and a cloud experience that keeps data private

- **HPE Tech Care Service**—This service goes beyond break-fix, to combine HPE expertise with leading-edge technology to help you get the most out of your IT investments and empower your business to thrive in the digital age:
 - When speed and reliability matter, trusted HPE experts deliver fast and accurate results
 - HPE Support Center wields the power of AI to simplify and optimize IT management
 - HPE Support Center helps safeguard systems from potential threats and business risk
 - Personalized HPE support delivers tangible business results to your bottom line
- **OpsRamp, a Hewlett Packard Enterprise company allows you to take control of your environment:** Observability to monitor complexity, provide alert management, remediation, and automation
 - Flexible and seamless vendor-agnostic integration
 - High-speed, precise multidimensional monitoring of GPUs
 - Proactive automation of unpredictable, parallel, and resource-hungry GPU workloads
 - Quick decoding and fixing of GPU-specific failure modes
 - Data-driven optimization for high GPU energy consumption

HPE's innovative approach to private cloud AI provides instant AI productivity to offer several key benefits, especially for organizations looking to harness the power of artificial intelligence while maintaining control, security, and scalability.

HPE turns your AI initiatives into competitive advantages.



Embedded services benefits

Technical challenges

Data processing and scalability issues

High GPU systems require meticulous resource oversight, including computational, storage, and networking elements. OpsRamp delivers advanced monitoring, beyond basic checks, uses GPU-specific telemetry and tools such as NVIDIA's DCGM with real-time dashboards, tracking thermal limits and data bottlenecks.

Managing GPU clusters

GPU-intensive environments involve dynamically allocating resources across hundreds or thousands of cores to optimize parallel workloads, such as training large language models or running real-time inference.

HPE Complete Care Service: HPC performance optimization provides all the capabilities needed to keep your investments at peak performance and under control, delivering the expected business outcomes.

Operational challenges

Integration and optimization

Integrating GPU-powered solutions is complex, especially with conflicting legacy systems.

OpsRamp, with AIOps-driven root cause analysis, correlating GPU telemetry, logs, and events, is essential to identify issues quickly, reducing mean time to resolution across the technology stack.

Traditional break-fix obsolescence

GPU-intensive systems are distributed, making error tracing tough with traditional tools.

HPE Complete Care Service enables environment-wide proactive support with predictive analytics and forecast disruptions, addressing issues before they hit workloads, reducing downtime with proactive, precise insights.

Financial challenges

Initial setup costs

Deploying high GPU solutions involves significant costs for compute, storage, and cooling.

OpsRamp provides GPU observability, integrated with NVIDIA's framework and telemetry into a unified dashboard, tracking metrics to control costs and enable chargebacks to business units effectively.

Ongoing maintenance and upgrade expenses

Maintaining and scaling GPU resource-intensive systems can be very expensive. Updates to AI applications, firmware, and software are necessary to ensure optimal performance and security.

HPE Complete Care Service offers full support for all components in the GPU environment, making sure they are properly updated and handled. Keeping a stable and reliable solution that can be consumed aligned with the business demands.



Recommended services expansion

HPE Services engagement framework offers a comprehensive stack of solutions for day 2 operational services for AI initiatives. These services enable IT operations to transform how high GPU environments are managed, optimized, and proactively supported across multiple AI workloads and hybrid environments.

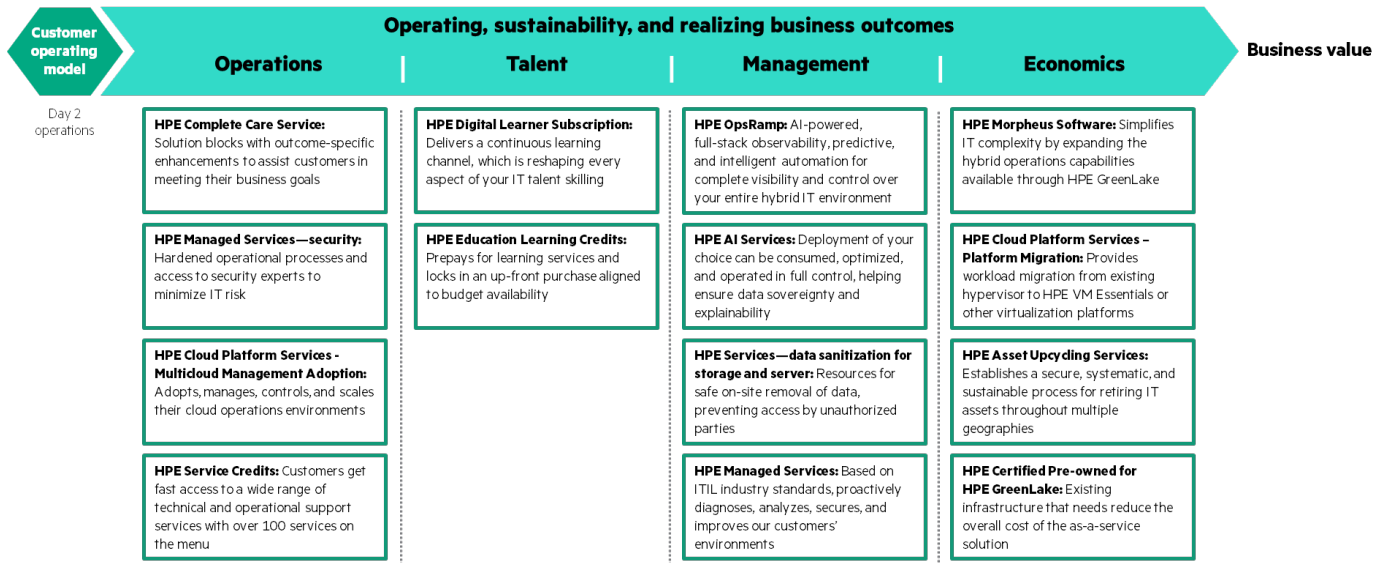


Figure 1. HPE Services engagement framework recommended expansion

Services expansion journey

HPE Services engagement framework: Solution road map for day 2 operational services VDI initiatives, improving IT operations over time.

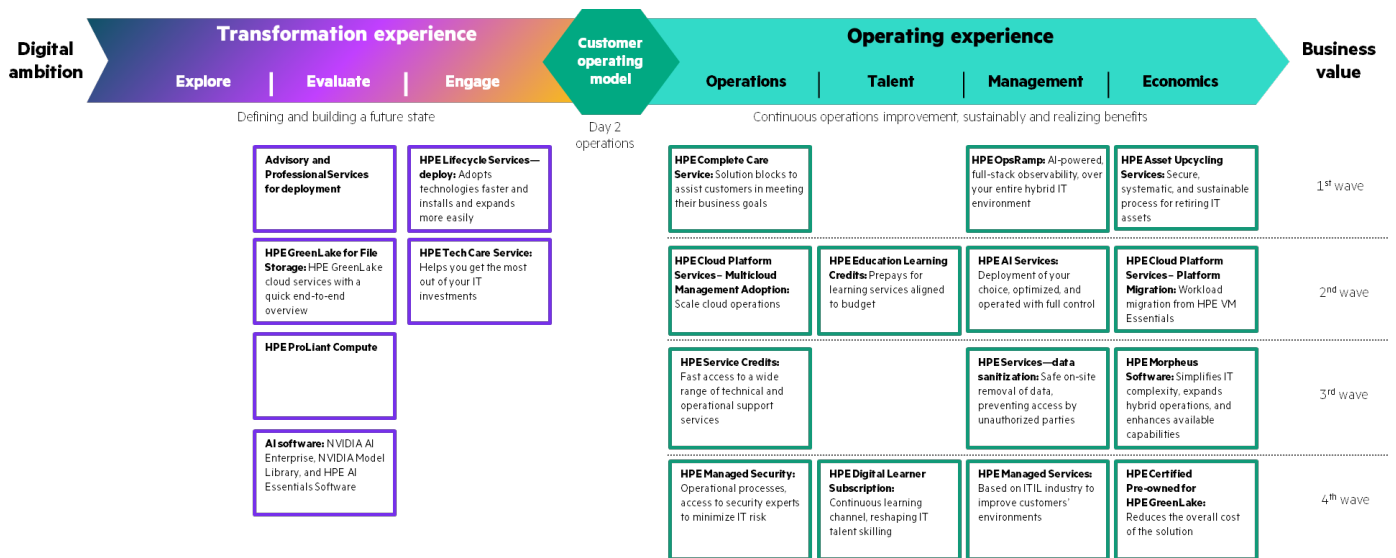


Figure 2. Customer operating model from digital ambition to business value

Learn more at

HPE.com/Services

Visit HPE.com

Chat now (sales)

Hewlett Packard Enterprise

© Copyright 2025 Hewlett Packard Enterprise Development LP. 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.

NVIDIA is a trademark and/or registered trademark of NVIDIA Corporation in the U.S. and other countries. All third-party marks are property of their respective owners.

a50013140ENW