Hewlett Packard Enterprise | NVIDIA

Solution overview

Accelerate enterprise AI

A turnkey AI private cloud, part of the NVIDIA AI Computing by HPE portfolio

Figure 1. Illustration of HPE Private Cloud AI as part of NVIDIA® AI Computing by HPE

Uncover the latest AI breakthroughs

Artificial intelligence (AI) has almost limitless potential to unlock insights, knowledge, and experiences that can reshape businesses and society.

AI adoption is erupting in the enterprise space, enabling companies to enhance business outcomes and innovate faster than competitors.

Most enterprise AI strategies focus on two areas: inference workloads, ranging from applications that utilize large language models (LLMs) to specific industry use cases, and using enterprise data to provide context with techniques such as fine-tuning and retrieval augmented generation (RAG). These advances are opening broader opportunities for generative AI (GenAI) to accelerate everything we do. Although enterprises are experimenting with multiple AI projects, the return on investment of each project is not always clear. This means enterprises need a simple, low-risk way to experiment with transformative technology.

Public cloud solutions (while feature rich) can expose data and models to threats, with potentially significant consequences to IP. This is driving enterprises to private cloud solutions with self-hosted open-source models to save time/resources, enhance data visibility, and increase flexibility. Even with private models, enterprises need the ability to govern data usage and track/monitor model usage to ensure AI models deliver the value they were intended to deliver and avoid introducing new risks. A private-focused approach offers better AI control and security while reducing environmental complexity. This shift has launched a return to the private cloud. 68% of enterprises view hybrid multicloud as key to GenAI strategy, and more than 50% plan to use dedicated private infrastructure.[1]

[1] "Collaboration Insights—IDC's Future Enterprise Resiliency and Spending Survey," IDC, 2023

Bring a flexible cloud experience to AI

IDC predicts that by 2026, 60% of enterprises will underperform on GenAI initiatives by failing to engineer connections between data, AI models, and applications.[2] Only 10% of AI projects will move into production.[3] So, why is this happening? And what can enterprises do to avoid it? Encouraging innovation is key, and enterprises need the right tools and infrastructure in place to increase the success rates of AI projects.

Emerging technologies are increasing IT complexity and require highly specialized skills to manage. For example, while there is a tremendous amount of innovation in the open-source domain, adopting these technologies for the enterprise can be overwhelming. Tuning a variety of software components to work seamlessly with underlying hardware can be expensive and time consuming. Talent gaps cause significant roadblocks to launching new use cases while underperforming technology slows down data pipelines and opens data to security threats.

Enterprises also face risks to reputation, data privacy, customer experience, and business growth due to improper use of AI.

Enterprises are increasingly embracing a hybrid cloud approach to AI with data privacy and control as key factors. A full-stack private cloud simplifies the management of fragmented technologies and boosts the productivity of AI users. This turnkey solution, known as HPE Private Cloud AI, supports the entire AI development lifecycle and provides seamless experiences for both IT/cloud ops and data science / AI teams. Also, HPE Private Cloud AI continues to deliver consistent experiences as AI infrastructure changes over time.

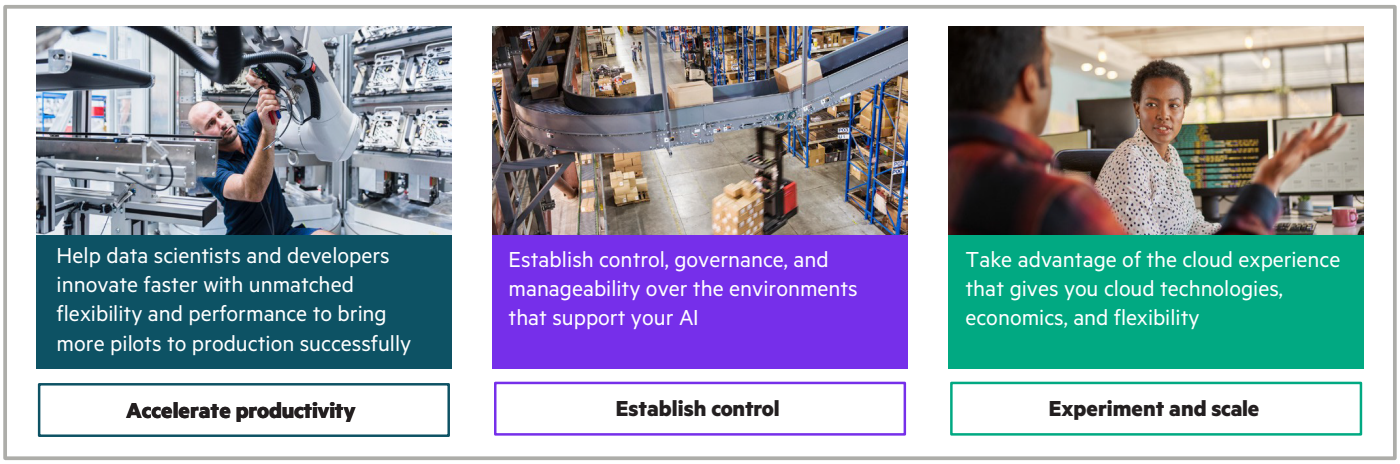

Figure 2. What enterprises need to unlock the power of AI

[2] "Generative AI will drive a foundational shift for companies—IDC," Computer World, 2024

[3] "Reasons Why Generative AI Pilots Fail To Move Into Production," Forbes, 2024

Why HPE Private Cloud AI?

As we scan the market, we see a gap and an opportunity for Hewlett Packard Enterprise to uniquely help enterprises streamline the use of AI. Our answer is HPE Private Cloud AI.

HPE Private Cloud AI is the first coengineered, turnkey solution to come from NVIDIA AI Computing by HPE—a new joint initiative to help enterprises unlock their AI ambitions. NVIDIA AI Computing by HPE brings together people, technology, and economics to accelerate AI deployments, protect against risks, and optimize AI costs long term. The solution is tailored to AI models and designed to scale easily with the growth and utilization of AI use cases. With this proven foundation, HPE Private Cloud AI accelerates data scientist productivity and helps overcome common challenges in operationalizing AI by delivering a flexible, pretested, AI-optimized private cloud.

Many AI solutions today focus on Day 0 to Day 1 challenges (such as integrating the technology stack). This has limited value given that design/setup is a fraction of the AI lifecycle. These solutions typically lack support for Day 2 challenges and beyond. HPE and NVIDIA are looking to change that. We empower your AI and IT teams with a rich ecosystem of proprietary and open-source tools to rapidly deploy AI workloads, simplify infrastructure configuration and management, and gain the freedom to experiment and scale AI projects while keeping your data private and secure.

This is where HPE GreenLake comes in. HPE Private Cloud AI provides a self-service cloud experience enabled by the HPE GreenLake cloud. It offers a single platform-based control plane with a portfolio of cloud services to automate, orchestrate, and manage users and data across hybrid environments. Start as small as a single AI pilot and evolve quickly for multiple use cases or higher throughput in one solution. Plus, you can deploy either on-prem, colocation, or in the cloud while maintaining control over financial risks.

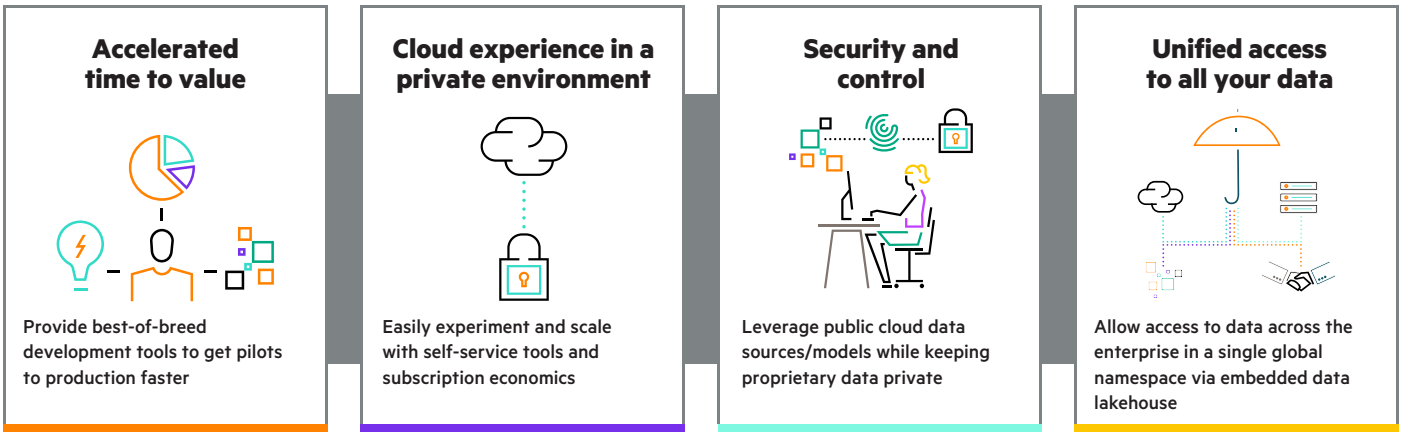

Figure 3. HPE and NVIDIA deliver key capabilities to optimize AI

Accelerate your AI efforts with a turnkey solution

HPE Private Cloud AI brings a fully curated solution to succeed with AI, from purpose-built infrastructure and the right tools for each stage of AI development to a library of models most applicable to your enterprise—all through a common experience and user interface. This flagship offering from NVIDIA AI Computing by HPE can make AI more streamlined than ever before.

The solution combines NVIDIA AI computing, networking, and software with robust HPE ProLiant Gen12 inferencing servers, HPE AI storage, and HPE GreenLake cloud to provide enterprises of every size a fast, flexible path for developing and deploying GenAI applications.

AI-optimized hardware is delivered as a single rack in small or medium configurations. Small configurations are ideal for basic LLM inference while medium configurations can support RAG for LLMs. Additionally, large multirack configurations are available, capable of fine-tuning the most complex models.

The software layer has a specialized set of AI tools leveraging NVIDIA AI Enterprise software to address your long-term AI needs. In a simple configuration, HPE AI Essentials Software provides a curated set of tools from HPE and NVIDIA to expedite data pipelines and use cases. Integration with NVIDIA NIM inference microservices helps you create data pipelines, develop and fine-tune your models, and deploy AI applications faster than before. Enterprise-grade tools support collaboration with role-based access control, data versioning and lineage, and development capabilities for model fine-tuning.

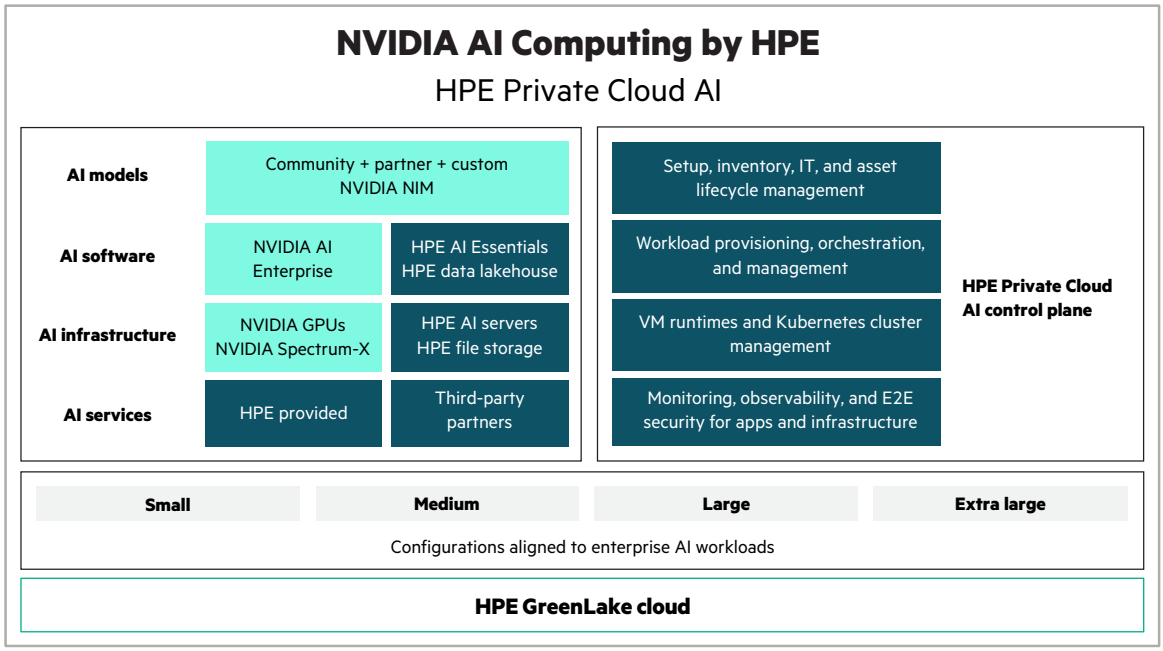

Figure 4. HPE Private Cloud AI architecture

Configurations	Small	Medium	Large	Extra large
Best for	Inferencing	Inferencing +RAG	Inferencing + RAG + fine-tuning	Inferencing + RAG + fine-tuning
Compute	4 or 8 NVIDIA L40S GPUs	8 or 16 NVIDIA L40S GPUs	16 or 32 NVIDIA H100 NVL GPUs	12 or 24 NVIDIA GH200 NVL2
Storage	30 TB to 248 TB	109 TB to 390 TB	670 TB to 1088 TB	670 TB to 1088 TB
Networking	100GbE NVIDIA Networking	200GbE NVIDIA Networking	400GbE NVIDIA Networking	800GbE NVIDIA Networking
Power	Up to 8 kW rack	Up to 17.7 kW	Up to 25 kW x 2	Up to 25 kW x 2

Figure 5. HPE Private Cloud AI infrastructure configurations

Discover the future of AI in private cloud

One thing is certain: AI will continue to transform the way we live and work by making our lives easier and safer and by introducing new challenges and ethical questions we have never encountered before.

These advances will revolutionize the financial services industry, accelerating tasks like reviewing complex financial documents (such as loan applications) while keeping our money more secure with advanced fraud detection and prevention capabilities. Healthcare will see personalized treatments and rapid diagnosis while lifting the burden of administrative work from doctors with powerful virtual assistants that can leverage expansive medical knowledge combined with private patient data. Industries like retail and the public sector will be enabled to improve their customer experiences and efficiency through automation and dynamic forecasting that reduces risk and drives increased customer satisfaction. These applications will be just the beginning.

AI will be a fast-growing part of the IT estate for decades to come. A private cloud environment is key to easing the complexities of AI adoption and mitigating risk as you experiment, innovate, and push the boundaries of what's possible with AI.

HPE and NVIDIA are ready to take you on the journey. HPE Private Cloud AI is unlike anything on the market today, designed with your needs and future needs in mind, whether you are an established AI user or just getting started. HPE AI services are available globally to support your transformation. HPE and NVIDIA experts work with you to plan, launch, and manage your AI environment—from strategy to technology selection, to design and proof of concepts, to deployment and ongoing production and management.

Discover how a private cloud approach can give you better AI control and security, so you can unleash the power of AI.

Learn more at

HPE.com/Private-Cloud-AI

Chat now (sales)

a00141386ENW

Hewlett Packard Enterprise