NVIDIA AI Computing by HPE

STEVE MCDOWELL, CHIEF ANALYST
6/22/24

CONTEXT

At its HPE Discover event in Las Vegas, Hewlett Packard Enterprise and NVIDIA announced a collaborative effort to accelerate the adoption of generative AI across enterprises.

The collaboration, branded as NVIDIA AI Computing by HPE, introduces a portfolio of co-developed AI solutions and joint go-to-market integrations to address the complexities and barriers of large-scale AI adoption.

HPE PRIVATE CLOUD AI

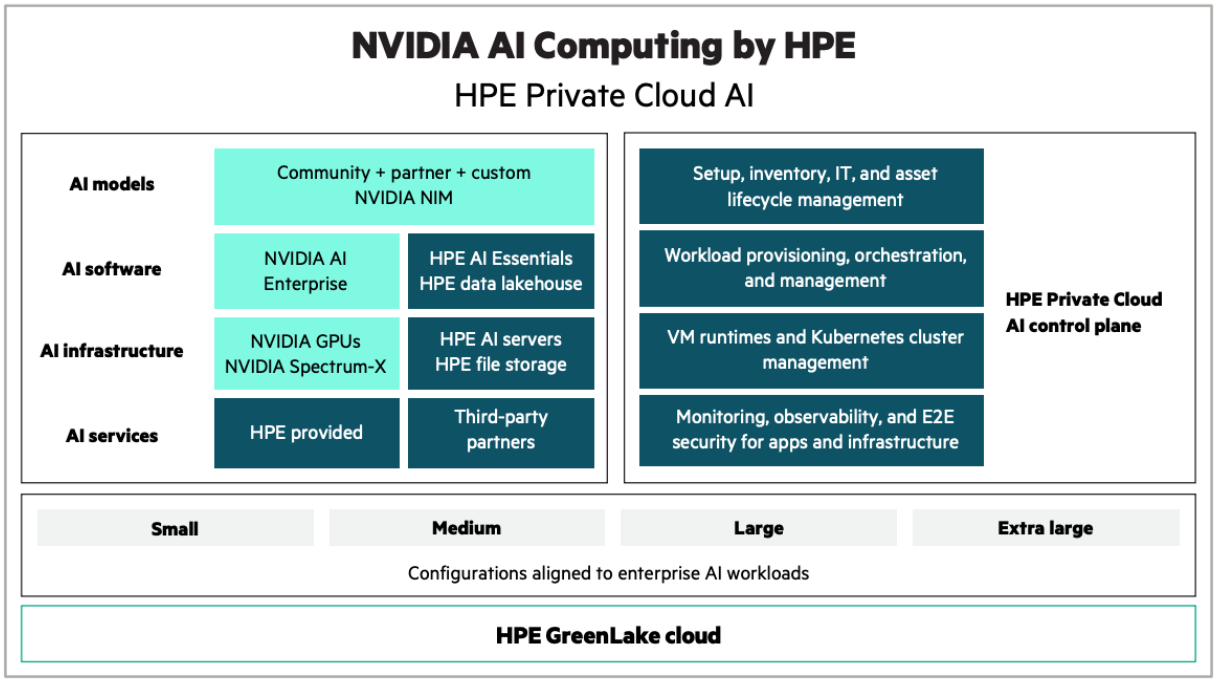
The cornerstone of the new offering is HPE Private Cloud AI, a pioneering solution that deeply integrates NVIDIA's AI computing, networking, and software with HPE's storage, compute, and cloud infrastructure. This solution offers a comprehensive, energy-efficient, and flexible platform for developing and deploying generative AI applications.

HPE Private Cloud AI provides a cloud-based experience designed to accelerate innovation and return on investment while managing enterprise AI risks. It offers:

- **Inference, Fine-tuning, and RAG AI Workloads**: Supports various AI workloads using proprietary data.
- **Enterprise Control**: Ensures data privacy, security, transparency, and governance.
- **Self-Service Cloud Experience**: Provides ITOps and AIOps capabilities for increased productivity.
- **Flexible Consumption**: Allows enterprises to adapt to future AI opportunities and growth.

© Copyright 2024 – NAND Research, All Rights Reserved

The solution includes four right-sized configurations to support a broad range of AI workloads and use cases, making it accessible to enterprises of all sizes.



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

INTEGRATION WITH NVIDIA AI ENTERPRISE

The foundation of HPE Private Cloud AI is the NVIDIA AI Enterprise software platform, which accelerates data science pipelines and streamlines the development and deployment of production-grade AI applications.

Key components include:

- **NVIDIA NIM**: Inference microservices that offer optimized AI model inferencing, facilitating a smooth transition from prototype to secure deployment.

© Copyright 2024 – NAND Research, All Rights Reserved

- **HPE AI Essentials Software**: Provides curated AI and data foundation tools with a unified control plane, ensuring AI pipelines are compliant, explainable, and reproducible.

ENHANCED INFRASTRUCTURE & NETWORKING

HPE Private Cloud AI features a fully integrated AI infrastructure stack, including:

- **NVIDIA Spectrum-X Ethernet Networking**: Enhances networking capabilities for AI workloads.
- **HPE GreenLake for File Storage**: Offers robust storage solutions tailored for AI needs.
- **HPE ProLiant Servers**: These servers support NVIDIA L40S, NVIDIA H100 NVL Tensor Core GPUs, and the NVIDIA GH200 NVL2 platform, ensuring high performance and scalability.

HPE OPSRAMP AI COPILOT & OBSERVABILITY

OpsRamp's IT operations integrated with HPE GreenLake provide enhanced observability and AIOps capabilities. This integration includes:

- **End-to-End NVIDIA Stack Observability**: Monitors NVIDIA NIM, AI software, Tensor Core GPUs, AI clusters, and networking components.
- **Conversational Assistant**: Uses NVIDIA's accelerated computing platform to analyze large datasets, boosting productivity for operations management.
- **Security Integration**: Integrates with CrowdStrike APIs to offer a unified service map view of endpoint security across the infrastructure.

GSI COLLABORATION

Global system integrators such as Deloitte, HCLTech, Infosys, TCS, and Wipro support the NVIDIA AI Computing by HPE portfolio to expedite the development of industry-focused AI solutions. These partnerships aim to reduce the time to value for AI initiatives by providing clear business benefits and robust AI solutions.

OTHER HPE/NVIDIA NEWS

Beyond the introduction of the new NVIDIA AI Compute by HPE, HPE also added support for additional NVIDIA GPUs, CPUs, and Superchips, including:

© Copyright 2024 – NAND Research, All Rights Reserved

- **HPE Cray XD670**: Supports eight NVIDIA H200 NVL Tensor Core GPUs.
- **HPE ProLiant DL384 Gen12 Server**: Supports dual NVIDIA GH200 NVL2.
- **HPE ProLiant DL380a Gen12 Server**: Supports up to eight NVIDIA H200 NVL Tensor Core GPUs.

HPE GreenLake for File Storage has also achieved NVIDIA DGX BasePOD certification and NVIDIA OVX storage validation, providing a proven enterprise file storage solution for AI workloads.

ANALYSIS

One of the most significant barriers to AI adoption has been the complexity and risk of managing fragmented AI technologies. NVIDIA AI Computing by HPE effectively mitigates these issues by delivering a turnkey private cloud solution for AI. This enables enterprises to focus on developing AI use cases without the burden of extensive infrastructure management.

Enterprise AI infrastructure is currently one of the most competitive technology markets. OEMs like Dell Technologies and Lenovo continue to fight public cloud providers for mindshare and market share, with the CSPs having an edge due to the cost and complexity of AI training solutions.

HPE's private cloud approach is a compelling middle-ground. With HPE GreenLake's flexible, as-a-service cloud experience, enterprises can efficiently manage and scale AI infrastructure. HPE's model allows businesses to consume AI capabilities on-demand, optimizing costs and ensuring scalability without significant upfront investments.

At the same time, partnerships with major system integrators like Deloitte, HCLTech, Infosys, TCS, and Wipro enable rapid deployment and customization of AI solutions across various industries. This is a competitive differentiator for HPE that helps enterprises leverage industry-specific expertise and achieve faster time-to-value for their AI investments.

HPE delivers a compelling solution with its NVIDIA AI Computing by HPE offering. It brings enterprises a robust, integrated platform that simplifies AI deployment and management. By addressing key challenges and providing a flexible, high-performance solution, HPE and NVIDIA offer a compelling solution stack that provides cloud-like agility without the cost and complexity of competing on-prem-only solutions.

© Copyright 2024 – NAND Research, All Rights Reserved

© Copyright NAND Research.

NAND Research is a registered trademark of NAND Research LLC, All Rights Reserved.

This document may not be reproduced, distributed, or modified, in physical or electronic form, without the express written consent of NAND Research. Questions about licensing or use of this document should be directed to info@nand-research.com.

The information contained within this document was believed by NAND Research to be reliable and is provided for informational purposes only. The content may contain technical inaccuracies, omissions, or typographical errors. This document reflects the opinions of NAND Research, which is subject to change. NAND Research does not warranty or otherwise guarantee the accuracy of the information contained within.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.