DELL AI FACTORY WITH NVIDIA

STEVE MCDOWELL, CHIEF ANALYST
MARCH 20, 2024

CONTEXT

At the NVIDIA GTC conference, Dell Technologies unveiled the Dell AI Factory with Nvidia, a comprehensive set of enterprise AI solutions aimed at simplifying the adoption and integration of AI for businesses.

Dell also announced enhancements to its flagship PowerEdge XE9680 server, including introducing Dell's first liquid-cooled server solution, which allows the server to deliver the full capabilities of NVIDIA's newly announced AI accelerators.

THE DELL AI FACTORY

The Dell AI Factory, in collaboration with NVIDIA, represents a pioneering initiative designed to accelerate the adoption and implementation of artificial intelligence (AI) across enterprises.

This comprehensive, end-to-end solution is tailored to simplify the complex process of deploying AI technologies, addressing several critical challenges that businesses face today.

Key Features and Offerings

- **Integrated AI Solutions:** The Dell AI Factory combines Dell's extensive hardware capabilities, including compute, storage, client devices, and networking, with NVIDIA's advanced AI infrastructure and software suite. This integration ensures a powerful, seamless solution for AI applications.
- **Designed for Enterprises:** Tailored to meet the diverse needs of enterprise environments, the Dell AI Factory supports distributed, democratized AI operations across various locations, from data centers to the edge and cloud. It's specifically engineered to adhere to enterprise data security and governance standards.
- **Support for GenAI Lifecycle:** The solution addresses the entire lifecycle of generative AI (GenAI) operations, including model development, tuning, and

application. It streamlines operations, from data transformation to actionable insights and outcomes.

- **Retrieval-Augmented Generation (RAG):** Among its new capabilities, the Dell AI Factory introduces RAG, a method that integrates an organization's data more effectively, reducing dependence on open-source models and expertise. This feature improves the quality of LLM (large language models) responses and reduces computational overhead.
- **AI-Optimized Infrastructure:** To cater to enterprises' growing needs, the Dell AI Factory leverages servers equipped with powerful NVIDIA GPUs and supports high-speed connectivity and optimized network fabrics. This infrastructure handles multiple AI projects and departments within an organization, ensuring high performance for increasing inferencing demands and model training.
- **Data Visibility and Expertise:** The initiative also focuses on overcoming data silos by enabling better visibility and preparation of data for GenAI projects. Dell Technologies provides professional services and the Dell Data Lake House to help customers identify the right data sets and accelerate their GenAI outcomes.

IMPACT AND ADVANTAGES

The Dell AI Factory with NVIDIA stands as a turnkey solution for businesses seeking to harness the power of AI to drive innovation, enhance productivity, and achieve smarter, higher-value outcomes. Offering a comprehensive suite of tools, services, and infrastructure allows organizations to navigate the complexities of AI implementation more effectively, ensuring a faster return on investment and a scalable path to AI transformation.

UPDATED SERVER & STORAGE PLATFORMS

Building the Dell AI Factory with NVIDIA and supporting NVIDIA's newly announced accelerators and networking components require updating the underlying servers and storage that ultimately comprise the solution.

Dell announced platform updates that provide enterprises with the computational power and storage solutions necessary to drive AI innovation and scale AI operations efficiently.

DELL POWEREDGE XE9680 SERVER ENHANCEMENTS

Dell's flagship PowerEdge XE9680 server has undergone several significant updates to enhance its capabilities for generative AI applications. The focus is on supporting

NVIDIA's advanced and powerful new GPU architectures, expanding memory capacity, and introducing new cooling configurations.

Here's a summary of the critical updates:

- **NVIDIA H200 Tensor Core GPUs:** The PowerEdge XE9680 now supports NVIDIA H200 GPUs, featuring advanced HBM3e memory with 141 GB per GPU. This upgrade facilitates handling more AI model parameters for training and inferencing within an air-cooled 6RU profile, aiming for a lower total cost of ownership.
- **BlueField-3 SuperNICs:** Integration with NVIDIA BlueField-3 SuperNICs enhances network traffic acceleration and offloading, which is crucial for large AI models. This enables more efficient GPU-to-GPU communication and quicker solutions.
- **Air-cooled Configuration with NVIDIA HGX B100 GPUs:** The introduction of NVIDIA Blackwell family HGX B100 GPUs provides the PowerEdge XE9680 with an eight-way GPU configuration that delivers increased processing power and 1.54 TB of HBM3e GPU memory. This setup enhances the server's ability to generate accurate GenAI insights from models with trillions of parameters.
- **Liquid-cooled Configuration with NVIDIA HGX B200 GPUs:** Dell's first liquid-cooled PowerEdge XE9680 configuration uses NVIDIA HGX B200 GPUs to offer up to 25x more AI inferencing performance, significantly reducing TCO and energy utilization compared to previous models.
- **NVIDIA GB200 NVL72 Multi-node Scale-up Server Architecture:** This new architecture, featuring the NVIDIA GB200 Superchip with an ARM CPU and on-chip Blackwell microarchitecture B200 GPUs, promises a dense configuration with 72 NVIDIA NVLink interconnected GPUs, setting a new standard for training and inferencing the largest AI models.

DELL STORAGE SUPERPOD VALIDATION

The Dell PowerScale F710 storage soluitions has been validated with NVIDIA DGX SuperPOD systems, including DGX H100 systems. This validation marks the PowerScale as the first Ethernet-based storage solution compatible with DGX SuperPOD, suitable for data centers already utilizing Ethernet.

The PowerScale architecture fully utilize GPUsd while ensuring high availability, efficiency, and protection against threats like data poisoning and model inversion.

ANALYSIS

As organizations evolve, there's a growing need for robust infrastructure to support expanding AI applications. Dell Technologies is advancing its collaboration with

NVIDIA to meet future demands, enhancing servers with new NVIDIA GPUs and supporting high-performance AI networking fabrics.

NVIDIA enjoys the support of a broad ecosystem, and Dell wasn't the only provider announcing support for its new platforms. Lenovo announced updates to its ThinkSystem AI portfolio to support increased GPU configurations and support for Nvidia's new GB200 Superchip and Quantum-X800 networking products. Likewise, Hewlett Packard Enterprise plans to update its generative AI solution stack, while Supermicro announced similar updates to its AI-focused portfolio. Dell, however, is engaging with NVIDIA more strategically.

The collaboration between Dell and NVIDIA to build the Dell AI Factory is strategic. Each organization has capabilities that complement the other to deliver an infrastructure to simplify AI adoption for businesses. The initiative will lower barriers for enterprises looking to embrace AI, addressing challenges such as data management, security, and the complexities of AI technology.

Dell's PowerEdge enhancements are foundational to this effort, ensuring enterprises have a clear performance path for their AI vision. This is critical for sustaining the increasingly complex projects that underpin enterprises' AI transformation projects.

As organizations expand their AI capabilities, the demand for optimized infrastructure supporting intensive workloads and complex data management requirements will only grow. Dell Technologies' latest moves are a proactive response to this trend, offering enterprises the tools to advance their AI strategies effectively.

Dell's AI Factory with NVIDIA, along with its related product enhancements, is not just about meeting the current demands of the AI and generative AI market; it's about anticipating the industry's future needs.

As enterprises increasingly rely on AI to drive innovation and competitive advantage, the importance of having a robust, scalable, and efficient AI infrastructure cannot be overstated. Dell is sending a clear signal that it intends to be at the center of this evolving ecosystem, enabling enterprises to unlock the full potential of their AI investments.

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.