
DELL AI FACTORY

STEVE MCDOWELL, CHIEF ANALYST
MAY 20, 2024

CONTEXT

At NVIDIA GTC earlier this year, Dell announced a [collaboration with NVIDIA](#) to deliver the Dell AI Factory with NVIDIA that was heavily based on NVIDIA technology. At this year's Dell Tech World, Dell went further, introducing its own Dell AI Factory, while also updating the Dell AI Factory with NVIDIA.

First things first, what is an AI factory? An AI factory, as described by NVIDIA and Dell Technologies, is a modern approach that bundles the elements required to quickly enable enterprises to painlessly deploy AI applications and innovations.

Unlike traditional IT infrastructures and operating models, which may not suffice for AI's intensive demands, an AI factory is tailored specifically for this purpose. It functions somewhat analogously to a physical factory that drove the Industrial Revolution, but instead of manufacturing physical goods, an AI factory produces actionable intelligence, fresh content, and new insights.

Setting the metaphor aside, an AI Factory is an intelligent bundling of the components required to deploy AI solutions. This bundling includes traditional infrastructure like servers, networking, and storage, all pre-qualified to run AI workloads at scale, and combines that with the right set of software and services.

Let's take a look at what Dell announced at Dell Tech World.

DELL AI FACTORY

Dell introduced its version of an AI Factory, the cleverly named Dell AI Factory. This includes all the elements required to quickly deploy AI at scale within an enterprise. As part of the announcement, Dell introduced new servers, storage, networking, and services to its portfolio, all pre-qualified to support demanding AI workloads.



Key Components of the Dell AI Factory include:

1. **Infrastructure Solutions:** The foundational elements of the AI Factory include advanced servers, storage solutions, data protection, and networking that are specifically designed to meet the demanding requirements of AI workloads.
2. **AI PCs and Copilot+:** New AI PCs equipped with Copilot+ and powered by Snapdragon X Elite and Snapdragon X Plus processors are designed to enhance productivity and creativity by allowing users to focus on strategic tasks.
3. **Edge to Cloud Services:** Dell provides solutions that stretch from the edge, where data is generated and often processed, to centralized cloud services, ensuring a cohesive and scalable AI deployment strategy. This includes orchestration platforms like Dell NativeEdge for automating the deployment of AI applications at the edge.
4. **Partnerships and Ecosystem:** Dell collaborates with leading technology partners like NVIDIA and Hugging Face and integrates with major platforms like Microsoft Azure to offer optimized and pre-tested AI solutions. These partnerships enhance the capabilities of the AI Factory by providing specialized software and hardware ready for enterprise deployment.
5. **Professional Services:** Dell has expanded its AI Professional Services portfolio to include tools and frameworks necessary for building AI platforms and deploying AI copilots. Services range from implementation support for AI deployments to strategic consulting to align AI technologies with business objectives.

NEW: DELL POWEREDGE XE9680L

The new Dell PowerEdge XE9680L server is a significant update to Dell's server lineup, specifically designed to support high-performance AI workloads with NVIDIA GPU acceleration.

Dell PowerEdge XE9680L
Extreme AI performance and scalability for NVIDIA HGX B200

- 8-GPUs in ultra dense 4U form factor
- Industry-leading network fabric throughput with 12 PCIe slots
- Direct to chip Liquid Cooling

Densest rack-scale architecture in the industry

Highest throughput in the industry with full support for 400G Ethernet and InfiniBand

Turn-key **liquid coolant distribution** solutions ecosystem

Factory integration of pre-validated networking, power distribution and cooling

- Up to **72** GPUs per rack
- 2x** I/O and throughput²
- 2.5x** energy efficiency³

Here are the key features of the new server:

- GPU Support:** The PowerEdge XE9680L supports up to eight NVIDIA Blackwell Tensor Core GPUs.
- Form Factor:** Despite its powerful capabilities, the server comes in a compact 4U form factor. This relatively small size for its capacity allows for efficient use of data center space while still providing significant computational power.
- High-Density GPU Deployment:** The server offers the industry's highest possible rack-scale density for NVIDIA GPUs within an x86 standard rack environment. This design increases GPU density per node by 33%, allowing more processing power in a smaller physical footprint.
- Enhanced Cooling System:** Utilizing Direct Liquid Cooling (DLC) technology, the PowerEdge XE9680L achieves greater efficiency with an enhanced cooling capacity.
- Increased Expansion Capacity:** Compared to previous models, the server features 20% more PCIe Gen. 5 slots and doubles the North/South network expansion capacity.
- Serviceability and Deployment:** Designed for easy serviceability, the PowerEdge XE9680L is tailored for fast onsite installation and fully configured with advanced factory integration.

DELL POWERSCALE & PROJECT LIGHTNING

The Dell PowerScale F910 and Project Lightning update Dell's storage portfolio, designed to enhance data handling capabilities for AI and other demanding applications.

Dell PowerScale: F910 and Project Lightning
For AI factories of today and tomorrow

PowerScale F910
The latest addition to our purpose-built all-flash lineup, optimized for AI

PowerScale: Project Lightning
Future-proof your storage with the addition of a true parallel file system

Extreme performance unlocks GPU potential for large-scale training

World's first ethernet storage for NVIDIA DGX SuperPOD

Available on-prem or in the cloud

Secure AI data against data poisoning and model inversion

Faster time to AI insights with up to **127%** improved performance¹

Up to **20X** Greater performance²

up to **6x** Greater performance vs. Azure NetApp Files³

Dell Technologies **APEX**

The Dell PowerScale F910 is an all-flash storage system engineered to cater to the intensive demands of AI workloads. Here are the technical specifics:

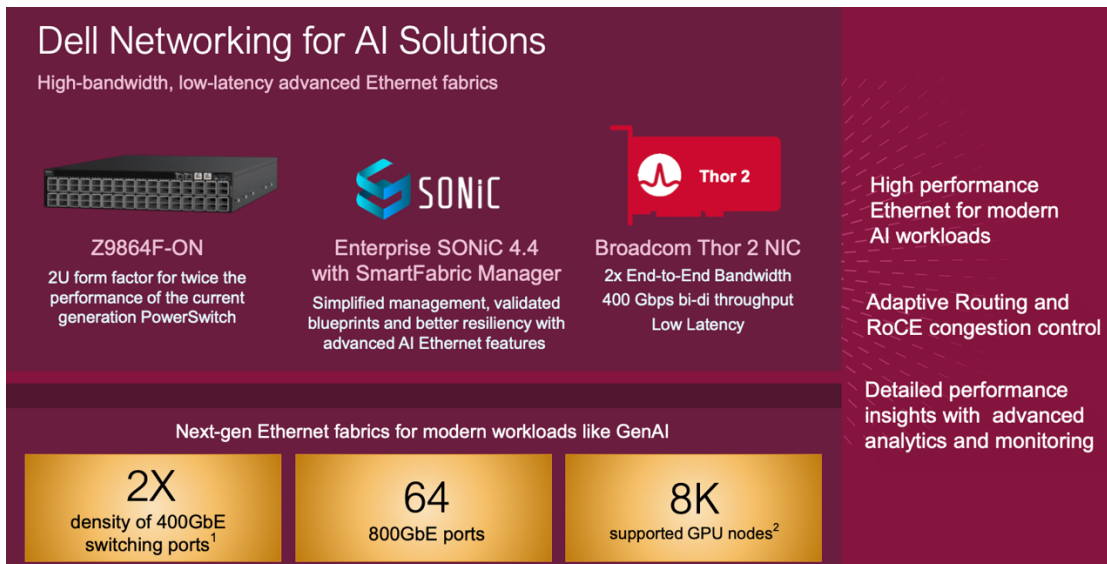
1. **High-Density Configuration:** The PowerScale F910 is designed in a compact yet highly dense 2U rack configuration, making it suitable for data centers where space efficiency is crucial.
2. **Storage Capacity:** Each F910 node has 24 NVMe SSDs, allowing organizations to scale raw storage capacity from 92 TB to 737 TB per node. This scalability can extend up to a staggering 186 PB of raw capacity per cluster, supporting extensive data growth and storage needs.
3. **Efficiency Features:** The system includes built-in in-line compression and deduplication technologies, which maximize storage efficiency by reducing data redundancy and conserving space.
4. **Cluster Configuration:** The PowerScale F910 can scale in clusters with a minimum of three nodes and a maximum of 252 nodes.
5. **Targeted Workloads:** The F910 is ideally suited for demanding sectors such as media and entertainment, high-frequency trading, and healthcare. It is particularly well-suited for supporting various phases of the AI lifecycle, especially in Generative AI applications. Its high performance and vast capacity make it an optimal choice for environments that require rapid data processing and large-scale data management.

Project Lightning is Dell's new Parallel File System for PowerScale, focusing on optimizing file storage for parallel processing:


1. **Parallel File System Architecture:** Project Lightning utilizes a parallel file system that enables data to be processed simultaneously across multiple nodes. This architecture is crucial for reducing the time required for data-intensive operations, such as training complex AI models or processing large datasets.
2. **AI Workflow Optimization:** The software is specifically tuned to accelerate AI training workflows, often requiring large-scale data ingestion and rapid iteration across datasets. Project Lightning enhances these processes by improving file read/write speeds and reducing input/output (I/O) overhead.
3. **Seamless Integration with Existing Infrastructure:** As an integrated software update to the PowerScale system, Project Lightning enhances existing hardware without necessitating significant infrastructural changes.
4. **Management and Maintenance:** Despite its advanced capabilities, Project Lightning is designed to integrate smoothly with the existing PowerScale management interfaces, ensuring that system administrators can manage their storage resources without a steep learning curve.

DELL NETWORKING UPDATES


Dell has updated its networking portfolio for the Dell AI Factory, with the updates designed to provide the performance and scalability required to support demanding AI applications and data-intensive environments.




Dell Networking for AI Solutions
High-bandwidth, low-latency advanced Ethernet fabrics



Z9864F-ON
2U form factor for twice the performance of the current generation PowerSwitch



Enterprise SONiC 4.4 with SmartFabric Manager
Simplified management, validated blueprints and better resiliency with advanced AI Ethernet features



Broadcom Thor 2 NIC
2x End-to-End Bandwidth
400 Gbps bi-di throughput
Low Latency

High performance Ethernet for modern AI workloads

Adaptive Routing and RoCE congestion control

Detailed performance insights with advanced analytics and monitoring

Next-gen Ethernet fabrics for modern workloads like GenAI

2X density of 400GbE switching ports ¹	64 800GbE ports	8K supported GPU nodes ²
---	---------------------------	---

Here's an overview of the new networking products:

1. **Dell PowerSwitch Z9864F-ON:** The Dell PowerSwitch Z9864F-ON is a high-performance switch powered by the Broadcom Tomahawk® 5 chipset. The chipset provides enhanced processing capabilities.
2. **Integration with Broadcom 400G PCIe Gen 5.0 Ethernet Adapters:** integration of Broadcom 400G PCIe Gen 5.0 Ethernet adapters with Dell PowerEdge servers and the PowerSwitch Z9864F-ON switch **Enterprise SONiC**

NEW: DELL NATIVEEDGE

Dell NativeEdge is a new edge orchestration platform introduced by Dell Technologies, designed to significantly enhance the deployment and management of AI applications at the edge.

Here are some key features and benefits of Dell NativeEdge:

1. **Automation of AI Software Delivery:** Dell NativeEdge automates the delivery and deployment of NVIDIA AI Enterprise software, which includes a suite of tools and frameworks designed to facilitate AI development and deployment across various environments.
2. **Support for NVIDIA Technologies:** The platform integrates seamlessly with NVIDIA's cutting-edge AI and machine learning technologies. This includes support for NVIDIA Metropolis for video analytics, NVIDIA Riva for speech and translation capabilities, and NVIDIA NIM for inference microservices.
3. **Deployment Blueprints:** Dell NativeEdge comes with deployment blueprints, pre-configured templates or guidelines that help simplify setting up and deploying AI applications at the edge. These blueprints are tailored to specific use cases and industries, making it easier for organizations to implement AI solutions that fit their unique requirements.
4. **Edge Data Processing:** By bringing AI processing closer to where data is generated, Dell NativeEdge helps businesses reduce latency, enhance data security, and improve the efficiency of their operations.
5. **Simplification of AI Deployment:** The platform aims to simplify the complex process of deploying AI applications at the edge by providing a centralized system that orchestrates necessary software and hardware components.
6. **Enhanced IT Operations:** Dell NativeEdge provides tools and frameworks for developers and IT operators that streamline the management of edge devices and applications. This includes features for monitoring, updating, and maintaining AI systems, ensuring they operate smoothly and efficiently.

SERVICES

Dell broadened its AI Professional Services to help organizations effectively implement AI solutions and achieve better business outcomes:

1. **Implementation Services for Microsoft Copilot Solutions:** These services are designed to assist organizations in transforming their operational models by integrating Microsoft Copilot experiences across various platforms, including GitHub, Security, Windows, and Sales. Dell provides expert guidance on adoption, aiming to enhance productivity and innovation within enterprises.
1. **Accelerator Services for Dell Enterprise Hub on Hugging Face:** Dell offers specialized support for organizations looking to quickly prototype AI using the Dell Portal. These services include strategic advice on tool selection, model selection, and alignment with specific use cases, helping to reduce time to value for AI projects.

PARTNERSHIPS & ECOSYSTEM

Dell made several strategic moves to enhance its ecosystem and partnerships, focusing on creating more integrated and efficient AI solutions.

Here are the key updates in Dell's partnerships and ecosystem:

Collaboration with Hugging Face

Dell Technologies is the first infrastructure provider to partner with Hugging Face, a leading platform for AI builders. This partnership focuses on:

1. **Optimized On-Premises Deployment of Generative AI Models:** Dell and Hugging Face have collaborated to optimize the deployment of large language models (LLMs) on-premises using Dell infrastructure. This allows organizations to train and deploy customized AI models securely within their data centers.
2. **Dell Enterprise Hub on Hugging Face:** This initiative provides organizations with direct access to Hugging Face's platform via the Dell Enterprise Hub. It enables secure and easy training and deployment of AI models, facilitating rapid prototyping and implementation of AI applications like chatbots and customer support tools.

Collaboration with Meta

Dell continues its partnership with Meta (formerly Facebook) with a focus on:

1. **Simplifying Deployment of Meta Llama 3 Models:** Dell and Meta are working together to streamline the deployment of Meta's Llama 3 models on Dell infrastructure. This partnership includes sharing test results, performance data,

and deployment recipes to ensure organizations can leverage Meta's AI capabilities efficiently.

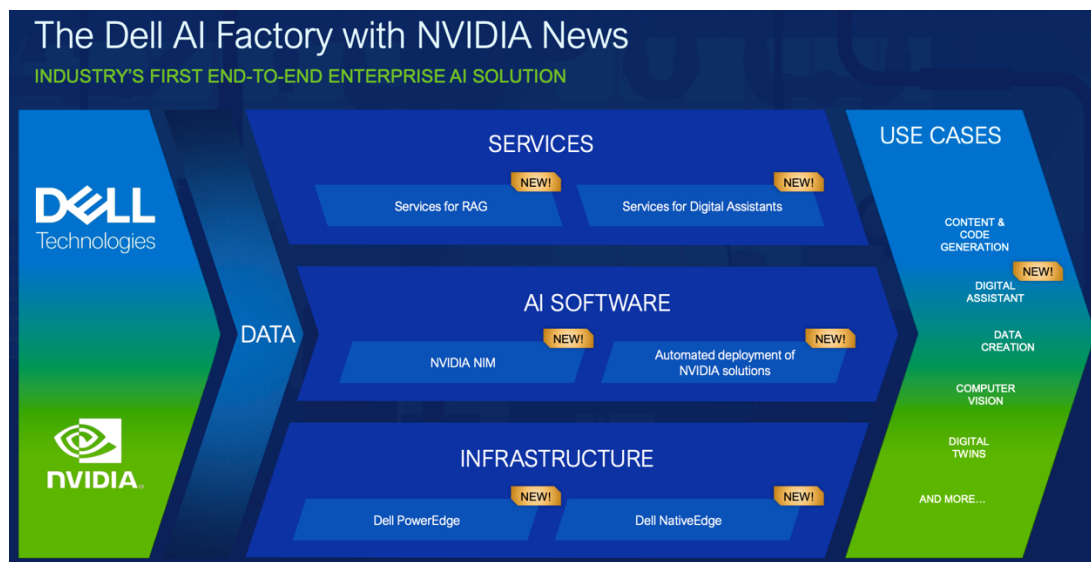
Dell AI Solution for Microsoft Azure AI Services

Dell's ongoing collaboration with Microsoft has led to the creation of specialized solutions that enhance the deployment of AI services:

1. **Microsoft Azure AI Services on Dell APEX Cloud Platform:** This solution speeds up the deployment of various Azure AI services, such as speech transcription and translation capabilities. It leverages Dell's APEX Cloud Platform to offer a flexible, scalable infrastructure for AI applications.

UPDATED: DELL AI FACTORY WITH NVIDIA

Dell updated its Dell AI Factory with NVIDIA, introducing several new and innovative elements designed to accelerate AI deployment and enhance performance across various applications.



Here's a detailed look at what's new in this collaboration:

1. **Dell PowerEdge XE9680L Server:** This new server model supports eight NVIDIA Blackwell GPUs and is designed for high performance in a compact 4U form factor. It achieves a higher density of GPU deployment within a standard x86 rack, which is crucial for environments demanding extensive computational power and efficiency.
2. **Direct Liquid Cooling:** The Dell PowerEdge XE9680L incorporates direct liquid cooling (DLC) technology, enhancing overall efficiency by providing superior cooling capacity for both CPUs and GPUs. This technology is essential for

maintaining performance stability and longevity in high-load AI computational tasks.

3. **Edge AI Orchestration with Dell NativeEdge:** Dell NativeEdge is a pioneering edge orchestration platform that automates the delivery of NVIDIA AI Enterprise software. It facilitates the easier and faster deployment of AI applications at the edge, enabling businesses in sectors like manufacturing and retail to process data locally and make decisions in real time.
4. **Full Stack Deployment Automation:** Dell and NVIDIA have engineered full stack deployment automation to reduce the time to value by up to 86% compared to manual setups. This automation spans the installation and configuration of both hardware and software, streamlining the setup process for AI environments.
5. **Generative AI Solution for Digital Assistants:** This solution is specifically designed to expedite the deployment of digital assistants, leveraging Dell and NVIDIA's technologies to offer personalized user experiences. Implementation services are also available to assist organizations in designing and scaling these solutions effectively.
6. **NVIDIA NIM Microservices:** These microservices are now part of the Dell AI Factory offerings, providing enterprise developers with production-ready, optimized inference engines. This supports a variety of popular AI models, facilitating faster deployment and scaling of AI applications.
7. **Dell Accelerator Services for RAG:** Available on Precision AI Workstations, these services help shorten the AI development cycle. They provide tailored support for deploying large language models using Retrieval-Augmented Generation (RAG), enhancing the performance of AI applications.

These advancements highlight Dell and NVIDIA's commitment to providing cutting-edge solutions that simplify AI adoption and maximize its potential across various industry sectors. This collaboration ensures that organizations can access the necessary tools, infrastructure, and support to leverage AI effectively and innovate within their operations.

ANALYSIS

Enterprise IT shops struggle with how to best approach deploying AI. The technology remains fast-moving, while the infrastructure required to support GPU-accelerated AI technologies, like generative AI, doesn't always look like what most IT practitioners are used to working with.

Businesses need to move fast to take advantage of AI's powerful potential. Solutions like the new Dell AI Factory enable IT, providing an integrated and scalable platform that simplifies the traditionally complex process of AI deployment. Dell expands

beyond its comfort zone of servers, storage, and networking, nurturing an expanding ecosystem of AI partners and further simplifying the solution for its enterprise customers.

The only questionable part of the announcement is Dell's Project Lightning. Providing high-performance, scalable storage with a unified namespace is critical for large AI training clusters. This is a market in which companies like WEKA excel, while others leverage existing parallel file systems like the open-source Lustre parallel file system and IBM's Spectrum Scale. VAST Data also has a significant footprint here. These well-proven solutions would all work just fine atop Dell PowerScale storage.

Tying the unproven Project Lightning to Dell PowerScale limits its reach. This isn't a technology that can be leveraged for AI training the in cloud, for example, so organizations with a heterogeneous training environment won't find use for the technology. Organizations looking to do small-scale training, however, may find Project Lightning perfectly acceptable. This is an area where the industry will look for further proof-points from Dell.

Dell's primary competitor in delivering AI infrastructure to the enterprise isn't other traditional OEMs like Lenovo and Hewlett Packard Enterprise; the fight is between on-prem deployments and the cloud. The early days of generative AI were a cloud-first experience precisely because of the cost and complexity of building AI training clusters. That's changing.

As the industry moves towards the next wave of generative AI, where businesses of all sizes look to the technology to provide new levels of digital transformation, the conversation changes. It's here where Dell is in its comfort zone, providing high-performance on-prem infrastructure backed by the service and support that's kept Dell the number one provider of servers and storage for the past decade.

Dell Technologies' AI Factory is a big step in democratizing AI technology. By providing a robust, scalable, and integrated suite of AI solutions, Dell empowers organizations to harness the full potential of AI to drive transformation and gain a competitive edge in the digital era. As AI evolves, comprehensive solutions like the Dell AI Factory will be crucial in enabling businesses to adapt and thrive in an increasingly AI-driven world.



RESEARCH BRIEF

© Copyright 2024 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.