Dell Technologies Fuels Enterprise AI Innovation with Infrastructure, Solutions and Services

May 19, 2025

- Advancements to the Dell AI Factory — from industry-first AI PCs to edge and data center enhancements — simplify and speed AI deployments for organizations of any size
- Powerful AI infrastructure and solutions, backed by a broad partner ecosystem and global services, empower organizations to embrace applications from building foundational models to running agentic AI

LAS VEGAS--(BUSINESS WIRE)--May 19, 2025-- DELL TECHNOLOGIES WORLD-- Dell Technologies (NYSE: DELL), the world's No. 1 provider of AI infrastructure,[1] announces Dell AI Factory advancements, including powerful and energy-efficient AI infrastructure, integrated partner ecosystem solutions and professional services to drive simpler and faster AI deployments.

Why it matters

AI is now essential for businesses, with 75% of organizations saying AI is key to their strategy[2] and 65% successfully moving AI projects into production.[3] However, challenges like data quality, security concerns and high costs can slow progress.

The Dell AI Factory approach can be up to 62% more cost effective for inferencing LLMs on-premises than the public cloud[4] and helps organizations securely and easily deploy enterprise AI workloads at any scale. Dell offers the industry's most comprehensive AI portfolio designed for deployments across client devices, data centers, edge locations and clouds.[5] More than 3,000 global customers across industries are accelerating their AI initiatives with the Dell AI Factory.[6]

Dell infrastructure advancements help organizations deploy and manage AI at any scale

Dell introduces end-to-end AI infrastructure to support everything from edge inferencing on an AI PC to managing massive enterprise AI workloads in the data center.

Dell Pro Max AI PC delivers industry's first enterprise-grade discrete NPU in a mobile form factor [7]

The **Dell Pro Max Plus laptop** with Qualcomm® AI 100 PC Inference Card is the world's first mobile workstation with an enterprise-grade discrete NPU.[8] It offers fast and secure on-device inferencing at the edge for large AI models typically run in the cloud, such as today's 109-billion-parameter model.

The Qualcomm AI 100 PC Inference Card features 32 AI-cores and 64 GB memory, providing power to meet the needs of AI engineers and data scientists deploying large models for edge inferencing.

Dell redefines AI cooling with innovations that reduce cooling energy costs by up to 60%[9]

The industry-first **Dell PowerCool Enclosed Rear Door Heat Exchanger (eRDHx)** is a Dell-engineered alternative to standard rear door heat exchangers. Designed to capture 100% of IT heat generated with its self-contained airflow system, the eRDHx can reduce cooling energy costs by up to 60%[10] compared to currently available solutions.

With Dell's factory integrated IR7000 racks equipped with future-ready eRDHx technology, organizations can:

- Significantly cut costs and eliminate reliance on expensive chillers given the eRDHx operates with water temperatures warmer than traditional solutions (between 32 and 36 degrees Celsius).
- Maximize data center capacity by deploying up to 16% more racks[11] of dense compute, without increasing power consumption.
- Enable air cooling capacity up to 80 kW per rack for dense AI and HPC deployments.[12]
- Minimize risk with advanced leak detection, real-time thermal monitoring and unified management of all rack-level components with the **Dell Integrated Rack Controller**.

Dell PowerEdge servers with AMD GPUs maximize performance and efficiency

Dell PowerEdge XE9785 and XE9785L servers will support AMD Instinct™ MI350 series GPUs, which offer 288 GB of HBM3E memory per GPU and deliver up to 35 times greater[13] inferencing performance.[14] Available in liquid-cooled and air-cooled configurations, the servers will reduce facility cooling energy costs.

Dell advancements power efficient and secure AI deployments and workflows

Because AI is only as powerful as the data that fuels it, organizations need a platform designed for performance and scalability. The **Dell AI Data Platform** updates improve access to high quality structured, semi-structured and unstructured data across the AI lifecycle.

- **Dell Project Lightning** is the world's fastest parallel file system per new testing, delivering up to two times greater throughput than competing parallel file systems.[15] Project Lightning will accelerate training time for large-scale and

complex AI workflows.
- **Dell Data Lakehouse** enhancements simplify AI workflows and accelerate use cases — such as recommendation engines, semantic search and customer intent detection — by creating and querying AI-ready datasets.

"We're excited to work with Dell to support our cutting-edge AI initiatives, and we expect Project Lightning to be a critical storage technology for our AI innovations," said **Dr. Paul Calleja, director, Cambridge Open Zettascale Lab and Research Computing Services, University of Cambridge**.

With additional portfolio advancements, organizations can:

- Lower power consumption, reduce latency and boost cost savings for high performance computing (HPC) and AI fabrics with **Dell Linear Pluggable Optics**.
- Increase trust in the security of their AI infrastructure and solutions with **Dell AI Security and Resilience Services**, which provide full stack protection across AI infrastructure, data, applications and models.

Dell expands AI partner ecosystem with customizable AI solutions and applications

Dell is collaborating with AI ecosystem players to deliver tailored solutions that simply and quickly integrate into organizations' existing IT environments. Organizations can:

- Enable intelligent, autonomous workflows with a first-of-its-kind on-premises deployment of **Cohere North**, which integrates various data sources while ensuring control over operations.
- Securely run scalable AI agents and enterprise search on-premises with **Glean**. Dell and Glean's collaboration will deliver the first on-premises deployment architecture for Glean's Work AI platform. [16]
- Innovate where the data is with **Google Gemini** and Google Distributed Cloud on-premises available on Dell PowerEdge XE9680 and XE9780 servers.
- Prototype and build agent-based enterprise AI applications with Dell AI Solutions with Llama, using **Meta's** latest Llama Stack distribution and Llama 4 models.
- Build and deploy secure, customizable AI applications and knowledge management workflows with solutions jointly engineered by Dell and **Mistral AI**.

The Dell AI Factory also expands to include:

- Advancements to the **Dell AI Platform with AMD** add 200G of storage networking and an upgraded AMD ROCm open software stack for organizations to simplify workflows, support LLMs and efficiently manage complex workloads. Dell and AMD are collaborating to provide Day 0 support and performance optimized containers for AI models such as Llama 4.
- The new **Dell AI Platform with Intel** helps enterprises deploy a full stack of high performance, scalable AI infrastructure with Intel® Gaudi® 3 AI accelerators.

Dell also announced advancements to the **Dell AI Factory with NVIDIA** and updates to **Dell NativeEdge** to support AI deployments and inferencing at the edge.

Perspectives

"It has been a non-stop year of innovating for enterprises, and we're not slowing down. We have introduced more than 200 updates to the Dell AI Factory since last year," said **Jeff Clarke, chief operating officer, Dell Technologies**. "Our latest AI advancements — from groundbreaking AI PCs to cutting-edge data center solutions — are designed to help organizations of every size to seamlessly adopt AI, drive faster insights, improve efficiency and accelerate their results."

"We leverage the Dell AI Factory for our oceanic research at Oregon State University to revolutionize and address some of the planet's most critical challenges," said **Christopher M. Sullivan, director of Research and Academic Computing for the College of Earth, Ocean and Atmospheric Sciences, Oregon State University**. "Through advanced AI solutions, we're accelerating insights that empower global decision-makers to tackle climate change, safeguard marine ecosystems and drive meaningful progress for humanity."

Additional Resources:

- Blog: Continuing to Power the Future of AI with Dell's Direct Liquid Cooling and Computing Innovations
- Blog: Redefining AI Connectivity with Dell's Optimized Infrastructure
- Blog: Driving Innovation in AI: Continuous Updates to the AI Platform with AMD
- Blog: Smart, Simple, Secure Enterprise AI with Dell and Cohere
- Blog: Gemini on Google Distributed Cloud for On-premises Computing with Dell Customers
- Blog: Dell and Glean On-Premises Solution Redefines Enterprise AI Search
- Blog: Now shipping - Dell AI Platform with Intel
- Blog: Bringing Mistral AI's Platform On-Premises with Dell AI Factory
- Blog: Build Agent-based Applications Faster with Dell AI Solutions
- Blog: IT leader's guide to build trust in AI solutions
- Connect with Dell on X and LinkedIn

About Dell Technologies

Dell Technologies (NYSE: DELL) helps organizations and individuals build their digital future and transform how they work, live and play. The company provides customers with the industry's broadest and most innovative technology and services portfolio for the AI era.

[1] IDC Semiannual Artificial Intelligence Infrastructure Tracker, 2024H1 (Feb 2025).

[2] Source: Dell Technologies survey across 750 business and IT decision makers across US, UK, DE, FR and JP, all segments, Feb 2025.

[3] Dell Technologies Chief Strategy Office enterprise AI adoption survey (US findings), November 2024 (N of 1,661 including 1,302 ITDMs and 359 AI practitioners).

[4] Based on Enterprise Strategy Group White Paper commissioned by Dell, "Understanding the total cost of inferencing large language models," April 2025. Analyzed models show a 70B parameter LLM leveraging RAG for an organization of 10k users being up to 52% more cost effective and for 50k users being up to 62% more cost effective over 4 years. Actual results may vary.

[5] Based on Dell analysis, July 2024. Dell offers solutions with NVIDIA hardware and software engineered to support AI workloads from PCs with AI-powered features and workstations to Servers for High-performance Computing, Data Storage, Cloud Native Software-Defined Infrastructure, Networking Switches, Data Protection, HCI and Services.

[6] Based on April 2025 Dell analysis of customer order data.

[7-8] Based an internal analysis of workstation providers, no one has an "enterprise-grade" discrete NPU in market. May 1, 2025.

[9-12] Based on Dell analysis in April 2025. Assumes 36°C facility water supply and ASHRAE A3 inlet server air compared to 20°C facility water supply and ASHRAE A3 inlet server air. Actual savings will vary.

[13] Based on internal analysis of PowerEdge XE9785 and XE9785L featuring MI350 series GPUs vs. previous generation XE9680 with MI300X, April 2025.

[14] AMD Accelerates Pace of Data Center AI Innovation and Leadership with Expanded AMD Instinct GPU Roadmap.

[15] Based on Dell preliminary testing comparing random and sequential throughput per rack unit, May 2025. Actual performance may vary.

[16] Based on Dell internal analysis, May 2025.

View source version on businesswire.com: https://www.businesswire.com/news/home/20250519816407/en/

Dell Technologies Media Relations: Media.Relations@Dell.com

Source: Dell Technologies