Dell AI Factory

Dell Generative AI Solution with AMD

Accelerate innovation, reduce costs and protect data with a scalable and modular architecture for complex GenAI.

75% more cost-effective inferencing vs. public cloud IaaS[1]

86% reduction in time to value[2]

Key use cases demand power, flexibility and scale

- Assistants, chatbots and content creation
- Accelerator as a service
- Multimodal retrieval augmented generation (RAG)

Simplified

Streamline GenAI deployments with proven validated solutions, backed by more than 340K engineering hours.

Optimize performance
High-performance accelerator, open-architecture and AI-optimized fabrics

AI anywhere
On data everywhere with multicloud storage flexibility

Turnkey multi-node
Proven full-stack AI foundations delivering faster outcomes

Tailored

AMD ROCm™ open-source software and open ecosystems boost AI development and operations.

Innovate faster
Use open-source software and ecosystems to develop unique applications.

Accelerate development
Leverage industry-standard frameworks with flexible technology stacks.

Activate your data
Efficiently run multiple AI use cases simultaneously.

Trusted

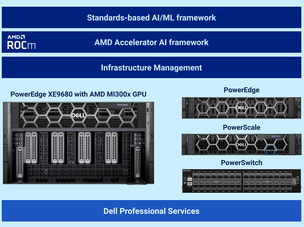
82% of ITDMs prefer an on-premises or hybrid model.[3]
Your data determines your outcomes. Protect it.

Start fast
On-premises foundations with root of trust security and full control

Streamline connectivity
Secure, feature-rich fabrics with scalability and optimized traffic flows

Automate provisioning
Open-source foundation for deploying and managing high-performance clusters

Dell GenAI Solutions with AMD



Inference

Run a 70B parameter model on a single AMD Instinct™ MI300X Accelerator.[4]

Customize

Deploy and fine-tune eight concurrent 70B models on a single Dell PowerEdge XE9680.[4]

Augment

Incorporate your data into the generative process.

Achieve competitive differentiation with a proven, open solution delivering secure, on-premises AI applications at scale.

Learn More

[1] Enterprise Strategy Group, Maximizing AI ROI: Inferencing On-premises With Dell Technologies Can Be 75% More Cost-effective Than Public Cloud, April 2024.
[2] Estimate based on Dell analysis in May 2024 comparing time to set up a 2-node Kubernetes cluster for a general-purpose LLM using automated scripts vs deploying a common design manually. Setup time includes base installation only. Actual setup time will vary depending on solution configuration.
[3] Dell Technologies, Generative AI Pulse Survey, August and September 2023.
[4] Dell Technologies blog, Silicon Diversity: Deploy GenAI on the PowerEdge XE9680 with AMD Instinct MI300X Accelerators, May 2024.