Accelerate AI Innovation

Unleash the full potential of artificial intelligence with Dell AI Factory with NVIDIA

Get Started

AI is powering an ever-changing world

Welcome to the artificial intelligence (AI) era. Whether you're already innovating with AI and generative AI (GenAI) models or looking for the best path forward for new use cases, cutting-edge technology and expertise are mission-critical. Dell AI Factory with NVIDIA® can help you seize the power of AI and GenAI to empower intelligent applications and experiences across your organization.

Top ways AI is driving innovation

GenAI

- Enable machines to identify patterns and structures within existing data inputs including text, image, audio, video and code.
- Quickly and automatically generate new and original content such as text, images, sounds, animation and 3D models.
- Streamline workflows for creatives, engineers, researchers, scientists and more.

Large language models (LLMs)

- LLMs are deep learning (DL) algorithms that can recognize, summarize, translate, predict and generate content using very large data sets.
- Training on data sets with hundreds of billions of parameters has unlocked the ability for AI to generate human-like content.
- Models can read, write, code, draw and create, augmenting human creativity and improving productivity across industries to solve the world's toughest problems.

Natural language processing (NLP)

- NLP enables AI to derive meaning from human language — written or spoken — by processing and analyzing text or voice data in order to understand, interpret, categorize and/or derive insights from the content.
- NLP includes natural language generation (NLG), which is the ability to create human language text. It also includes natural language understanding (NLU), which takes text as input, understands context and intent and generates an intelligent response.

Retrieval augmented generation (RAG)

- RAG is a technique for enhancing the accuracy and reliability of GenAI models using facts from external sources.
- Chatbots use RAG to deliver responses that are more relevant to the context of the user's query and enriched with the most current information available without the need for retraining the underlying LLM.
- RAG profoundly impacts user engagement, particularly in customer service, education and entertainment, where the demand for immediate, accurate and informed responses is paramount.

Learn more

Dell AI solutions webpage: Dell.com/AI | Code generation: Codeium Enterprise | RAG: Dell Scalable Architecture for RAG with NVIDIA Microservices

Digital twins

- Run simulations on a virtual object, system or process to predict real-world behavior.
- Enable better, faster and more cost-effective research and development (R&D) cycles.
- Bridge the digital and real world to optimize design, improve performance and provide real-time predictive maintenance.

Computer-aided design, manufacturing and engineering (CAD/CAM/CAE)

- Gain insights for radical new methods of product design and production.
- Speed time to market with more innovative and higher-quality products.
- Refine products before investing in costly and time-consuming physical prototyping.

Knock down barriers to entry for AI

Optimizing organizational use of AI
Every business has multiple goals for leveraging AI. A one-off use case and one-size-fits-all approach does not meet their needs. The solution is a strategy that addresses all your use cases and delivers comprehensive platforms and turnkey solutions that accelerate them to production.

Increasing volume and complexity of AI projects
Most organizations see and have captured the opportunity of AI and GenAI but face challenges understanding the technologies involved, building a scalable strategy and powering more of the business with AI for competitive advantage.

Data protection is more critical than ever
With data determining the outcomes for AI, protecting it is critical. You need to take steps to avoid solutions that expose data to external threats or that could limit data's value or result in theft of intellectual property.

Go from AI possible to AI proven

Dell Technologies is prepared to meet you wherever you are on your AI journey. Whether you're just getting started with AI or are ready to deploy a DL cluster, Dell Technologies has a complete portfolio of solutions that can help you recognize and take advantage of untapped market opportunities.

Dell PowerEdge servers are the foundational building block for AI solutions, providing the performance, GPU density and efficiency required to get started with AI and grow as needed. In addition, NVIDIA-Certified PowerEdge servers are available with NVIDIA accelerated compute to speed AI workloads — and results.

Dell Technologies works with NVIDIA and other leading AI software companies to help ensure that no matter where you need support in your data and AI portfolio, we have the right solution to meet you there. You can take advantage of an integrated ecosystem of technology innovations from the workstation to the data center, edge and cloud, enabling a holistic approach to AI that leads to success.

Bring AI to your data: Dell AI Factory with NVIDIA

The Dell AI Factory with NVIDIA transforms innovation into value with industry-leading capabilities that simplify development and accelerate AI adoption. This is a full stack that includes GPUs, CPUs networking, NVIDIA AI Enterprise software and Dell Professional Services for Generative AI, allowing you to embrace GenAI at an enterprise-wide scale.

Accelerated insights

Innovative compute performance across the AI lifecycle delivers AI, HPC, and modeling and simulation operations at the speed of business.

Simplified operations

Boost AI infrastructure automation to effectively control and manage AI and HPC infrastructure and workloads, anywhere.

Trusted AI

Reduce risks and accelerate your AI lifecycle with trustworthy, high-quality solutions and infrastructure.

Up to **86%**

reduced time to value compared to doing it yourself[1]

#1

Broadest GenAI solutions portfolio from desktop to data center to cloud, all in one place[2]

82%

of IT decision-makers prefer an on-premises or hybrid model for GenAI.[3]

65%

of IT leaders that have moved beyond pilot stages expect near-immediate value.[3]

Learn more

Press releases:

- Dell Technologies Expands Dell AI Factory with NVIDIA to Turbocharge AI Adoption
- Dell Technologies Fast-Tracks AI-Driven Innovation with the Dell AI Factory

Articles:

- Forbes: Dell Launches AI Factory To Accelerate Enterprise AI Integration
- Forbes: NVIDIA and Dell Build an AI Factory Together
- Tech Target: Dell AI Factory curates AI tech for customers

Blogs:

- Transform Innovation into Value: The Dell AI Factory with NVIDIA
- Simplifying AI in the Enterprise: The Dell AI Factory with NVIDIA
- How Dell Makes the AI Factory Real

Videos:

- Bloomberg Television: NVIDIA, Dell Are Building Their Own AI 'Factories'

[1] 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.

[2] Based on Dell analysis, August 2023. Dell Technologies offers solutions engineered to support AI workloads from workstations PCs (mobile and fixed) to servers for HPC, data storage, cloud-native software-defined infrastructure, networking switches, data protection, HCI and services.

[3] Dell Technologies, Generative AI Pulse Survey, September 2023.

Built to accelerate AI insights

Unleash your AI advantage with Dell PowerEdge servers
NVIDIA-Certified Dell PowerEdge servers are acceleration-optimized and purpose-built for AI, GenAI and high performance computing (HPC). With superior acceleration and diverse GPU options, these powerful platforms are optimized to turn ideas into action faster.

Accelerate transformation anywhere with PowerEdge servers

Accelerate innovations

Deliver greater insights with GenAI and accelerate AI/ML/DL operations at the speed of business.

Security from concept to retirement

Harness cryptographic verification, system lockdown and safeguards, anchored by silicon root of trust.

Intuitive systems management

Facilitate effortless discovery, deployment, monitoring, securing and updating of PowerEdge servers.

Sustainability

Improve energy efficiency, optimize energy consumption and use recycled materials — validated by recognized eco labels.

PowerEdge XE servers

Acceleration optimized, purpose built for complex compute, AI/ML/DL and HPC-intensive workloads

	 PowerEdge XE9680 Powerful and flexible for no-compromise accelerated AI	 **PowerEdge XE9640** A dense, direct liquid cooled (DLC) server to deliver real-time AI insights	 **PowerEdge XE8640** Superior performance with a GPU-optimized design
Applications and use cases	• AI/ML/DL training, HPC, CRISP • Healthcare, cloud service providers (CSPs), finance, academia	• AI/ML/DL training, HPC modeling and simulation	• Medium data set language models, NLP, modeling and simulation • AI/ML/DL training and inferencing, image recognition
CPU	• 2x 4th and 5th Generation Intel® Xeon® Scalable processors	• 2x 4th and 5th Generation Intel Xeon Scalable processors	• 2x 4th and 5th Generation Intel Xeon Scalable processors
GPU support	• Up to 8x NVIDIA H100 or H200 SXM5 GPUs with full NVLink™ connectivity	• Up to 4x NVIDIA H100 SXM5 GPUs with full NVLink connectivity	• Up to 4x NVIDIA H100 SXM5 GPUs with full NVLink connectivity
Features	• 6U rack height • Air cooled up to 35°C • 32 DDR5 DIMMs • Up to 10x 16 PCIe Gen5 slots	• 2U rack height • Liquid-cooled CPU and GPU operation • 32 DDR5 DIMMs • Up to 2 x PCIe Gen5 slots	• 4U rack height • Air cooled up to 35°C • 32 DDR5 DIMMs • Up to 4x PCIe Gen5 slots

PowerEdge rack servers

Flexible, mainstream computing foundations for a wide range of applications, use cases and workloads

Achieve near-bare-metal performance

97.5%
of bare-metal performance using VMware®[4]

66%
increase in performance per watt[5]

67%
increase in high-performance LINPACK (HPL) performance[6]

	PowerEdge R760xa Flagship server for GPU-based workloads	PowerEdge R750/R650 Mainstream performance	PowerEdge XR12 Edge performance
Applications and use cases	• AI/ML/DL training and inferencing, analytics and HPC • Virtual desktop infrastructure (VDI) and performance graphics	• Light duty AI/ML/DL training and inferencing • VDI, performance graphics • Edge	• Edge AI training and inferencing • Telco • Rendering/modeling
CPU	• 2x 4th and 5th Generation Intel Xeon Scalable processors	• Up to 2x 3rd Generation Intel Xeon Scalable processors	• 1x 3rd Generation Intel Xeon Scalable processor
GPU support	• Up to 4x double-wide or 8x single-wide NVIDIA PCIe GPUs	• Up to 3x double-wide or 6x single-wide NVIDIA PCIe GPUs	• Up to 2x double- or single-wide NVIDIA PCIe GPUs
Features	• 2U rack height • Air cooled up to 35°C • 32 DDR5 DIMMs • Up to 4x PCIe Gen5 slots	• 1U or 2U rack height • Air cooled up to 35°C • 32 DDR4 DIMMs • Up to 8x PCIe Gen4 slots	• 2U rack height • Operational tolerance from -5°C to 55°C • Up to 4x PCIe Gen4 slots

Learn more

Dell PowerEdge servers webpage: Dell.com/PowerEdge

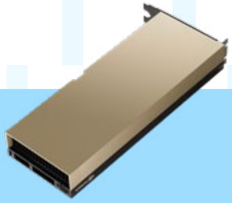
[4] In performance testing, configurations using Dell Technologies and VMware achieved up to 97.5% of bare-metal performance on the same server. Source: Principled Technologies report: Achieve near-bare-metal inference throughput for image classification workloads with the Dell PowerEdge R7525 server using virtual GPUs, July 2022.

[5] 66% increase in performance/watt on the Dell PowerEdge R750xa with the NVIDIA H100s configuration vs the A100 configuration. Source: Dell Technologies tech note, PowerEdge R750xa and NVIDIA H100 PCIe GPU: 66% Increase in HPC Performance per Watt, 2022.

[6] The PowerEdge R750xa with NVIDIA H100s configuration achieved a 67% increase in HPL benchmark performance compared to the NVIDIA A100 configuration. Source: Dell Technologies tech note, PowerEdge R750xa and NVIDIA H100 PCIe GPU: 66% Increase in HPC Performance per Watt, 2022.

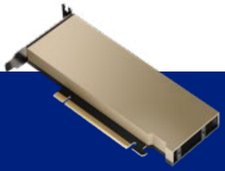
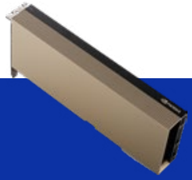
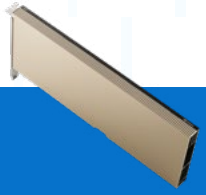
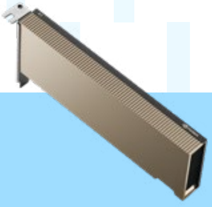
Unleash AI with NVIDIA GPUs

Dell Technologies works closely with NVIDIA, the only vendor offering a complete portfolio with Hopper and Ampere GPUs from entry level to mainstream to the highest performance. Each provides the versatility to accelerate the widest range of AI applications, whether at the edge, in the cloud or on-premises.

				
H200 SXM The world's most powerful GPU for supercharging AI and HPC workloads	**H100 SXM** Extraordinary performance, scalability and security for every data center	**H100 NVL** Highest performance AI, ML training and exascale HPC	**L40S** Unparalleled AI and graphics performance for the data center	**L40** High-performance graphics and rendering
AI and HPC	AI and HPC	AI and HPC	AI and HPC and performance graphics	Performance graphics and VDI
• 3,958 TFLOPS FP8 Tensor Core* • NVLink: 900GB/s PCIe Gen5 • Up to 7 MIGs @ 16.5GB each • NVIDIA vGPU software support	• 3,958 TFLOPS FP16 Tensor Core* • NVLink: 600GB/s PCIe Gen5 • Up to 14 MIGs @ 12GB each	• 3,026 TFLOPS FP8 Tensor Core* • NVLink: 600GB/s PCIe Gen5 • Up to 7 MIGs @ 10GB each • NVIDIA AI Enterprise software included	• 1,466 TFLOPS Tensor performance** • 212 TFLOPS RT core performance • NVIDIA vGPU software support • OVX support for NVIDIA Omniverse™	• 90.5 FP32 TFLOPS (non-Tensor) • 724.1 FP8 Tensor TFLOPS with FP32 accumulate* • NVIDIA vGPU software support • OVX support for NVIDIA Omniverse

* With structural sparsity enabled

** Peak rates are based on GPU boost clock.

 **L4** Breakthrough universal accelerator for efficient video, graphics and AI	 **A16** Multimedia-rich VDI to enable remote work including CAD/CAM/CAE	 **A10** Accelerated graphics and video with AI for mainstream enterprise servers	 **A2** Entry-level GPU for AI inferencing at the edge
AI inferencing, edge and VDI	VDI	Mainstream graphics and VDI	AI inferencing, edge and VDI
• 485 TFLOPS FP8* • PCIe Gen4 x16 • NVIDIA vGPU software support	• 4x 35.9 TFLOPS FP16* • PCI Express Gen4 x16 • NVIDIA vGPU software support	• 250 TFLOPS FP16* • PCIe Gen4x16 • NVIDIA vGPU software support	• 36 TFLOPS FP16 Tensor Core* • PCIe Gen4 x8 • NVIDIA vGPU software support

* With structural sparsity enabled

An order-of-magnitude leap: NVIDIA H100 Tensor Core GPU

Deploying H100 GPUs at data center scale delivers outstanding performance and brings the next generation of exascale HPC and trillion-parameter language models within reach.

Up to **4X**
higher AI training on GPT-3[7]

30X
faster AI inference performance on the largest LLMs[7]

Up to **7X**
higher performance for HPC[7]

For details on which PowerEdge servers support which NVIDIA GPUs, see the GPU matrix.

Learn more
Dell Technologies accelerators web page: Dell.com/GPU

[7] NVIDIA.com, NVIDIA H100 Tensor Core GPU, accessed July 2024.

NVIDIA technologies are built in

The PowerEdge servers at the heart of your solution come with integrated NVIDIA technologies that help speed AI workloads — and results.

NVIDIA AI Enterprise: The operating system for enterprise AI

NVIDIA AI Enterprise is a cloud-native platform that makes it easy to create and deploy optimized AI solutions including RAG, computer vision, speech AI and more. Deploy anywhere — cloud, data center, edge and workstations. Assembling, optimizing and securing production deployments is no longer complex or time-consuming. Included in NVIDIA AI Enterprise is NVIDIA NIM, a set of easy-to-use microservices designed for secure, reliable deployment of high-performance AI model inferencing.

NVIDIA-Certified Systems

As NVIDIA-Certified Systems®, Dell VxRail HCI and Dell PowerEdge bring together NVIDIA GPUs, NVIDIA ConnectX® smart network interface cards (SmartNICs), and NVIDIA BlueField® DPUs in optimized configurations. These are validated for performance, manageability, security and scalability and are backed by enterprise-grade support from NVIDIA and Dell Technologies.

NVIDIA H100 GPU

The NVIDIA H100 Tensor Core GPU delivers unprecedented performance, scalability and security to every data center and includes NVIDIA AI Enterprise software suite for streamlined AI development and deployment. It delivers 9X faster AI training[8] and 30X faster AI inference performance on the largest models.[9]

NVIDIA Virtual GPUs (vGPUs)

NVIDIA vGPU software enables sharing GPU resources across multiple VMs to make them accessible to any device, anywhere.

NVIDIA Multi-Instance GPUs (MIGs)

NVIDIA MIGs expand the performance and value of GPUs by partitioning them into as many as seven instances to support every workload and extend accelerated resources to more users.

NVIDIA BlueField-3 Data Processing Units (DPUs)

The NVIDIA BlueField-3 DPU is a 400Gb/s infrastructure computing platform for data center infrastructure workloads. By offloading, accelerating, and isolating networking, storage, and security services, BlueField-3 DPUs enhance performance, optimize efficiency, and bolster security within AI data centers.

NVIDIA Spectrum-X

The NVIDIA Spectrum-X networking platform improves the performance and efficiency of Ethernet-based AI clouds and enterprise deployments. It achieves 1.6X better networking performance for AI, along with consistent, predictable performance in multi-tenant environments.[10]

NVIDIA Launchpad

This free, curated lab experience enables you to get immediate, short-term access to the hardware and software stacks you need to experience end- to-end solution workflows for AI, data science, 3D-design collaboration and simulation and more.

[8] H100 features fourth-generation Tensor Cores and the Transformer Engine with FP8 precision that provides up to 9X faster training over the prior generation for mixture-of-experts (MoE) models. Source: NVIDIA, NVIDIA H100 Tensor Core GPU, accessed January 2023.

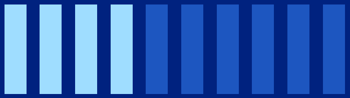
[9] Compared to the previous generation. Source: NVIDIA, NVIDIA H100 Tensor Core GPU, accessed January 2023.

[10] NVIDIA.com, NVIDIA Spectrum-X Networking Platform, accessed July 2024.

Customer successes

Northwestern Medicine improves productivity and patient outcomes with GenAI

Northwestern Medicine wanted to advance the healthcare ecosystem to improve patient outcomes and accelerate healthcare delivery. To realize the promise of GenAI, it followed a unified approach that would allow caregivers to act more quickly to save lives and be more effective in helping patients.



40% improvement in radiology performance

Blueprint
for GenAI adoption across the healthcare industry

Saves lives
by alerting clinicians to conditions requiring immediate attention

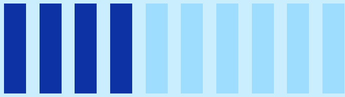
> “GenAI and AI offer a tremendous opportunity to help us take better care of our patients and give time back to care providers.
>
> — Dr. Mozziyar Etemadi, Clinical Director of Advanced Technologies at Northwestern Medicine

The City of Amarillo delivers accessibility with GenAI

In order to bridge the language gap for residents of Amarillo, TX, Dell Professional Services consulted Assistant City Manager and CIO Rich Gagnon and his team on the creation of a GenAI digital assistant with the ability to communicate in multiple languages.

Emma
The GenAI digital assistant lives on the city’s website.

62 languages
and dialects are now available for accessing services.



24%
of the population can now access services in their own language.

> “We're not afraid of the future. We're embracing this wholeheartedly.
>
> — Rich Gagnon, Assistant City Manager and Chief Information Officer, City of Amarillo

Learn more
Customer success page: GenAI improves productivity and patient outcomes

Learn more
Customer success page: Delivering accessibility through GenAI

Why Dell Technologies

Collaborate at worldwide Customer Solution Centers

Collaborate with Dell Technologies engineering teams at one of our worldwide Customer Solution Centers, tap into the resources of one of our HPC & AI Centers of Excellence or test and tune real-world systems at the HPC & AI Innovation Lab.

Consume AI-as-a-Service with Dell APEX

With simple and consistent cloud experiences delivered as-a-Service (aaS), Dell APEX for Generative AI can help you get the AI-optimized solutions you need to fast-track intelligent outcomes everywhere. Dell APEX can deliver a cloud operating model for AI on-premises, off-premises and at the edge so you can create measurable value from data at any scale.

Speed success with services

Dell Technologies Services include consulting, deployment, support and education to help drive the rapid adoption and optimization of AI environments from initial setup and upskilling of resources through to ongoing support. Managed Services and Residency Services can help reduce the cost, complexity and risk of managing IT so you can focus resources on digital innovation and transformation.

Jump-start GenAI objectives

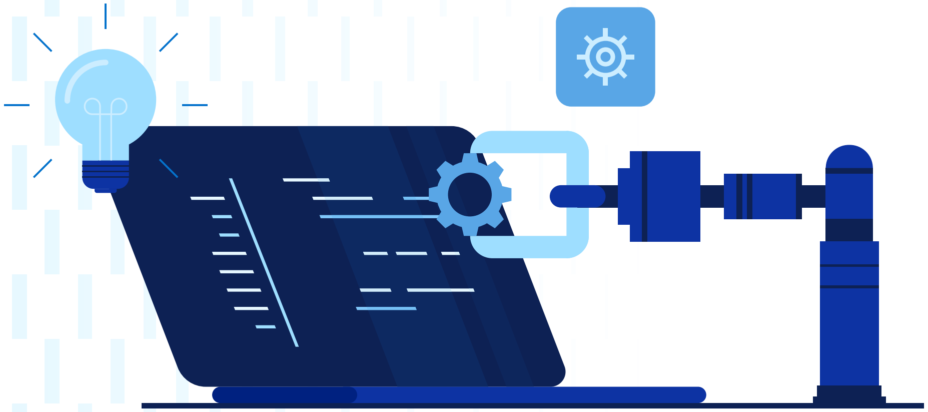
If you're not sure where to begin, you can leverage the Dell Accelerator Workshop for Generative AI to start your journey to developing a winning strategy. This half-day workshop is a great place to start, helping you address your readiness in leveraging GenAI across business and IT dimensions.

Dell experts, working together with your team, will help you begin to develop a point of view on important GenAI questions and create a vision for your future state. Utilizing our "AS-IS"/"TO-BE" methodology, we will conduct interviews and review the existing environment to identify challenges and opportunities and drive consensus for GenAI, synthesized in an executive overview.

- Assess current state
- Establish a vision
- Identify challenges
- Develop a roadmap
- Define goals
- Define expected results

Learn more

Read the brochure: Accelerate the power of AI for your data



Accelerate intelligent outcomes

Dell Technologies helps organizations of all types and sizes illuminate opportunity and reveal the full potential of their data. With 35+ data science teams driving 450+ AI projects and 1,800+ team members dedicated to extracting insights from data,[11] Dell Technologies brings proven AI expertise to improve IT efficiencies and mitigate risk to deliver better customer insights and experiences. And we do this in a consistent way across hybrid clouds, on-premises, off-premises and at the edge.

Dell Technologies and NVIDIA can help you win in the age of AI.

Dell Technologies and NVIDIA

Enabling and accelerating AI workloads

Dell Technologies and NVIDIA work together to deliver engineering-validated hardware and software to accelerate AI, ML and DL workloads. Dell Technologies also invests heavily in servers and solutions that incorporate leading-edge NVIDIA GPUs, SmartNICs with DPUs and AI Enterprise software. With NVIDIA and Dell Technologies, you can take AI where you never thought possible.

DELL Technologies AI Factory WITH NVIDIA

Learn More

The Dell AI Factory with NVIDIA

[11] Dell Technologies internal data sources.

Dell Technologies believes the information in this document is accurate as of its publication date. The information is subject to change without notice.