

Accelerate Al Innovation

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



Al 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. The 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



GenAl

- 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 humanlike content.
- Models can read, write, code, draw and create, augmenting human creativity and improving productivity across industries to solve the world's toughest problems.



Retrieval-augmented generation (RAG)

- RAG is a technique for enhancing the accuracy and reliability of GenAl 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 there is high demand for immediate, accurate and informed responses.



Agentic Al

- Agentic Al refers to GenAl systems designed to operate autonomously, pursuing goals through multi-step reasoning, planning, and action.
- These agents can make decisions, use tools, adapt to feedback, and coordinate across tasks without constant human prompting.
- Agentic Al unlocks new possibilities in enterprise automation, research, and complex workflows where sustained, context-aware execution is essential.



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 Al

Optimizing organizational use of Al

Every business has multiple goals for leveraging Al. One-off use cases and a one-size-fits-all approach do not meet your 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 Al projects

Most organizations see the potential of AI and GenAI but face challenges with 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 Al, protecting it is critical. You need to avoid solutions that could expose data to external threats, limit the 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 Al 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 blocks 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 ensure that no matter where you need support in your data and Al 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 Al Factory with NVIDIA transforms innovation into value with industry-leading capabilities that simplify development and accelerate Al adoption. This full stack includes GPUs, CPU networking, NVIDIA AI Enterprise software and Dell Professional Services for GenAl, allowing you to embrace GenAl at an enterprisewide scale.



Accelerated insights

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



Simplified operations

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



Trusted Al

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



of organizations say Al/GenAl has become key to their business strategy.1

of organizations have successfully transitioned Al proofs of concept (PoCs) to production.²

79%

of AI use cases in production are running outside public cloud.2

believe data is key, and their GenAl strategy must prioritize its use and protection.2

Learn more

Press releases:

- Dell Technologies Unveils Next Generation Enterprise AI Solutions with NVIDIA
- Dell Technologies Accelerates Enterprise Al Innovation from PC to Data Center with NVIDIA
- Together Al Taps Dell Technologies to Scale Al Acceleration Cloud for the Enterprise

Bloas:

- PowerEdge Server and Networking Announcements at NVIDIA GTC 2025
- · Simplify Enterprise Al Adoption and Scale With Agentic Al
- · Maximize Resources and Drive Al Outcomes at Scale

Dell Technologies survey across 750 business and IT decision-makers across US, UK, DE, FR and JP, all segments, February 2025. ² Dell Chief Strategy Office enterprise AI adoption survey of ~3,800 IT decision-makers (ITDMs) and AI practitioners across US, UK, DE, JP and India, November 2024.



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 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 GenAl 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

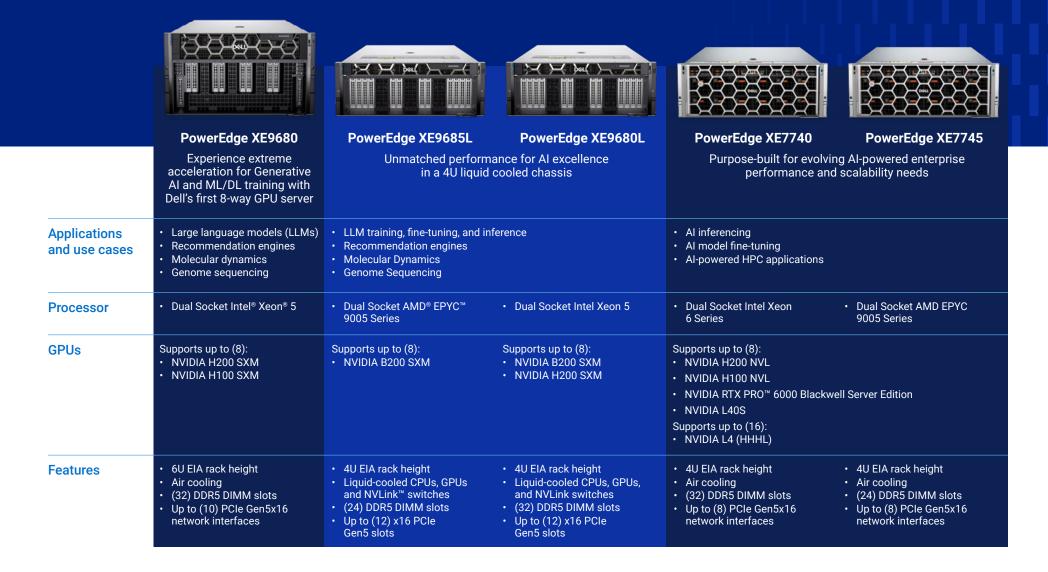
Improve energy efficiency, optimize energy consumption and use recycled materials — validated by recognized eco labels.

PowerEdge XE servers

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

Learn more

Dell.com/PowerEdge



PowerEdge XE servers

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

Learn more
Dell.com/PowerEdge



PowerEdge XE9712

The future of high-performance dense acceleration



PowerEdge XE8712

Extreme GPU density for HPC research and breakthrough discovery



PowerEdge XE9780



PowerEdge XE9785

Applications and use cases

- · Real-time Al inference
- GenAl
- LLMs

- Scientific research
- · Financial market modeling
- Genomic sequencing

- Large model training
- Model fine-tuning
- Large-scale inferencing

CPUs

GPUs

(36) NVIDIA GB200 NVL72 Superchips per IR7000 rack:

- (36) NVIDIA Grace CPUs
- (72) NVIDIA Blackwell B200 or B300 GPUs interconnected by NVLink @ 1.8 TB/s
- Up to 144 GPUs per IR7000 rack:
- (2) NVIDIA Grace CPUs
- (4) NVIDIA Blackwell B200 GPUs interconnected by NVIDIA NVLink
- Dual Socket Intel Xeon 6
- Dual Socket AMD EPYC[™] 9005 series

- nk
 - (8) NVIDIA HGX B200 SXM GPUs or
 - (8) NVIDIA HGX B300 SXM GPUs
- (8) NVIDIA HGX B300 SXM GPUs

Features

- Rack-scale, liquid-cooled IR7000 deployment
- (8) E1.S NVMe drives per tray
- Disaggregated 33kW Powershelves
- 1:1 NIC to GPU ratio for east west GPU traffic
- NVIDIA Bluefield 3 DPUs for storage and in-band traffic
- 10U ORv3 rack height Rack-scale IR7000 blind-mate power and cooling
- Up to (2) E3 NVMe drives
- Liquid cooled CPUs, GPUs, and NVLink Switches
- Disaggregated 33kW Powershelves
- Up to (4) PCle Gen 5x16 network interfaces

- 10U EIA rack height
- Air cooling
- (8) GPU connected ConnectX-8 OSPF network interfaces
- (4) PCIe Gen 5 network interfaces
- (12) PCle Gen5x16 FH network interfaces (B200)
- OCP 3.0 networking slot
- (32) DDR5 DIMM slots
- Up to (12) PCle Gen5x16 network interfaces

- 10U EIA rack height
- Air cooling

Next-generation

- (8) GPU connected ConnectX-8 OSPF network interfaces
- (4) PCIe Gen 5 network interfaces
- (24) DDR5 DIMM slots
- Up to (12) PCle Gen5x16 network interfaces

PowerEdge rack servers

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

Learn more

Dell.com/PowerEdge

	EEEEEET!	E888831		ROBERRA
	PowerEdge R7725	PowerEdge R7715	PowerEdge R670	PowerEdge R770
	Versatility me	eets performance	Elevate your data center efficiencies with optimized power and balanced performance	Drive efficiency in your data center through maximum performance with optimized power
Applications and use cases	 Al and ML HPC Data analytics Virtualization Hyperconverged infrastructure (HCI) 		Hyperscale workloadsCloud-native applicationsScale out databasesVirtualizationAll-flash SDS	Al inferencingVirtualizationCloud-native applicationsHyperscale workloadsScale out databases
Processor	AMD EPYC 9005 series	AMD EPYC 9005 series	• Intel Xeon 6	Intel Xeon 6
GPUs	Supports up to (2): NVIDIA RTX PRO 6000 Blackwell Server Edition Supports up to (2): NVIDIA H100 NVL NVIDIA L40S NVIDIA A16 Supports up to (6): NVIDIA L4 (HHHL)	Supports up to (3): NVIDIA H100 NVL NVIDIA L40S NVIDIA A16 Supports up to (6): NVIDIA L4 (HHHL)	Supports up to (4): • NVIDIA L4 (HHHL)	Supports up to (2): NVIDIA RTX PRO 6000 Blackwell Server Edition NVIDIA H200 NVL NVIDIA H100 NVL NVIDIA A16 Supports up to (4): NVIDIA L4 (HHHL)
Features	 2U rack height Air cooling and direct liquid cooling (DLC) options (24) DDR5 DIMM slots Up to (8) x8 or x16 PCle Gen5 slots 	 2U rack height Air cooling and direct liquid cooling (DLC) options (24) DDR5 DIMM slots Up to (8) x8 or x16 PCle Gen5 slots 	 1U rack height Air cooling and DLC options (32) DDR5 DIMM slots Up to (2) x16 PCle Gen5 slots 	 2U rack height Air cooling and DLC options (32) DDR5 DIMM slots Up to (2) x16 PCle Gen5 slots

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.

NVIDIA HGX H200 SXM The world's leading Al computing platform	NVIDIA HGX B200 SXM NVIDIA HGX B300 SXM Propelling the data center into a new era of accelerated computing		NVIDIA GB200 NVL72 Powering the new era of computing	NVIDIA GB300 NVL72 Built for the age of Al reasoning	NVIDIA GB200 NVL4 Accelerating the rate of scientific discovery
 Al supercomputing Complex simulations Massive datasets	 AI/GenAI Data analytics HPC		Real-time LLM inference	Test-time scaling inference	 HPC AI Modelling and simulation Visualization
 Transformer engine Fourth-generation NVLink NVIDIA Confidential Computing NVIDIA Multi-instance GPU (MIG) DPX instructions 	 (8) NVIDIA Blackwell GPUs Up to 1.4 TB HBM3E 1800GB/s NVLink between GPUs via NVSwitch™ chip 15X faster real-time LLM inference 3X faster training 	 (8) NVIDIA Blackwell Ultra SXM GPUs Fifth-generation NVLink 4x faster training for LLMs Up to 11x higher inference performance 	 Connects (36) Grace CPUs and (72) Blackwell GPUs in an NVIDIA NVLink- connected, liquid- cooled, rack-scale design 30X faster real-time trillion-parameter LLM inference 	 Connects (36) Grace CPUs and (72) Blackwell GPUs in an NVLink-connected, liquid-cooled, rack-scale design 50x higher output for reasoning model inference 10x boost in user responsiveness 	 Connects (4) NVIDIA Blackwell GPUs with (2) Grace CPUs in an NVLink-connected, liquid-cooled, rack-scale design

Source: NVIDIA com, accessed October 2025

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.

RTX PRO 6000 Blackwell server edition Powering the next era of Al	NVIDIA H200 NVL The GPU for GenAl and HPC	NVIDIA L40S Unparalleled AI and graphics performance for the data center	NVIDIA L4 The breakthrough universal accelerator for efficient video, AI and graphics	NVIDIA A16 Unlock an unprecedented virtual desktop infrastructure (VDI) user experience
GenAIRendering and graphicsVisualization and simulationContent and streamingData science	 GenAl LLMs HPC scientific computing	 GenAI LLM inference and training 3D graphics Rendering Video 	VideoAlVisual computingGraphicsVirtualization	• High-density, graphics-rich VDI
 Enhanced Streaming Multiprocessors (SMs) built for neural shaders 5th Gen Tensor Cores support FP4 precision, DLSS 4 Multi Frame Generation 	 1.9X faster Llama2 70B inference 1.6X faster GPT-3 175B inference 110X faster HPC 	 1,466 TFLOPS tensor performance 212 TFLOPS RT core performance 91.6 TFLOPS single-precision performance 	 Hosts up to 1,040 concurrent AV1 video streams at 720p30 120X higher AI video performance than CPUbased solutions 	 More than 2X the encoder throughput Up to 70% better performance using vPC

Source: NVIDIA com, accessed October 2025

NVIDIA technologies are built in

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



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 enterprisegrade 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 4X higher AI training on GPT-3 and 30X faster AI inference performance on the largest models.³

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.⁴

NVIDIA Launchpad

This free, <u>curated lab experience</u> 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.

Customer successes

Worley drives sustainability worldwide through AI innovation

Worley helps customers across energy, resource and chemical industries innovate to become more sustainable and address their most urgent business concerns. To accelerate and support client projects, Worley creates AI solutions built on the Dell AI Factory with NVIDIA.



Faster

time to value for AI initiatives.



Improves

productivity with AI for advanced content generation.



Better

scalability and performance with on-premises Al infrastructure.

"Al is pivotal to Worley achieving our sustainability goals. The Dell Al Factory with NVIDIA plays a critical role in that."

-Summer Husband, Senior Vice President of Data Products and Applied Intelligence, Worley

Northwestern Medicine improves productivity and patient outcomes with GenAl

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.



Blueprint

for GenAl adoption across the healthcare industry

Saves lives

by alerting clinicians to conditions requiring immediate attention

GenAl and Al 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

Learn more

Customer success page: <u>Driving sustainability worldwide through AI innovation</u>

Learn more

Customer success page: GenAl improves productivity and patient outcomes

Why Dell Technologies

Collaborate at worldwide Customer Solution Centers

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

Consume Al-as-a-Service with Dell APEX

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

Speed success with services

<u>Dell Technologies Services</u> 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.

<u>Managed Services</u> and <u>Residency Services</u> can help reduce the cost, complexity and risk of managing IT so you can focus resources on digital innovation and transformation.



Jump-start GenAl objectives

If you're not sure where to begin, the Dell Accelerator Workshop for GenAl is a great way to kick off your journey to developing a winning strategy. This half-day workshop helps you assess your readiness to leverage GenAl across business and IT dimensions.

Dell experts, working with your team, will help you begin to develop a point of view on important GenAl 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 GenAl, synthesized in an executive overview.



Assess current state



Establish a vision



Identify challenges



Develop a roadmap



Define goals

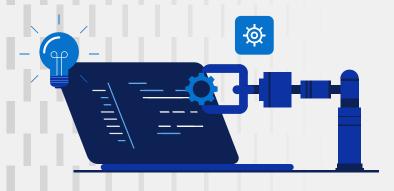


Define expected results

Learn more

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

eBook



Accelerate intelligent outcomes

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

Enabling and accelerating AI workloads for intelligent outcomes

<u>Dell Technologies and NVIDIA</u> 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.





Copyright © 2025 Dell Inc. or its subsidiaries. All Rights Reserved. Dell and other trademarks are trademarks of Dell Inc. or its subsidiaries. NVIDIA®, NVLink™, NVIDIA-Certified Systems™, ConnectX® and BlueField® are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Intel® and Xeon® are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries. AMD® and EPYC™ are trademarks of Advanced Micro Devices, Inc. Other trademarks may be the property of their respective owners. Published in the USA 10/25 eBook

dell-nvidia-ai-EB-103

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