DELL Technologies

Innovate faster with GPU-accelerated AI

Unleash the full potential of artificial intelligence with solutions from Dell Technologies and NVIDIA

Get Started

NVIDIA

Table of Contents

AI is powering an ever-changing world 3

Knock down barriers to entry for AI 4

Go from AI-possible to AI-proven 5

Built to accelerate AI insights 6

- Unleash your AI advantage with Dell PowerEdge servers 6
- Dell PowerEdge XE servers 7
- Dell PowerEdge rack servers 8

Unleash AI with NVIDIA GPUs 9

NVIDIA technologies are built in 11

Recommended configurations 12

Customer successes 13

- Taboola delivers content recommendations on a massive scale 13
- Duos Technologies keeps trains moving at full speed 13
- University of Cambridge accelerates scientific discovery 14
- University of Pisa extends the power of AI 14

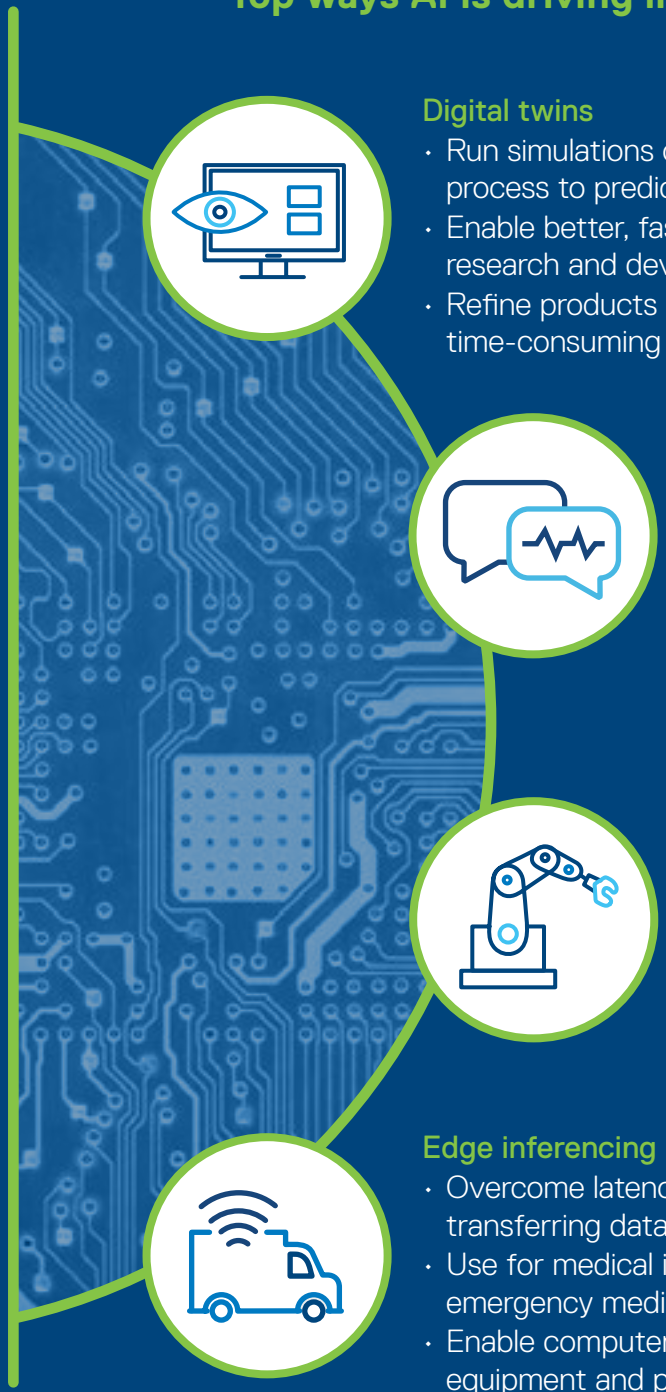
Why Dell Technologies 15

Accelerate intelligent outcomes 16

AI is powering an ever-changing world

In the age of AI, increased access to data and new data management techniques are providing the fuel for AI-driven insights for organizations of all types and sizes. The ubiquity of AI—from advances in processing power to the rise of enterprise multicloud—enables enterprises to benefit from AI on-premises, in private and public clouds and at the edge for a variety of emerging workloads.

Top ways AI is driving innovation



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.
- Refine products before investing in costly and time-consuming physical prototyping.

Generative AI / Natural language processing (NLP)/large language models (LLM)

- Enable machines to understand human language.
- Help humans interact more personally with computerized systems.
- Use for GPT/transformer models, chatbots, digital assistants, sentiment analysis, fraud detection and more.

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.
- Transform design and engineering to power the factory of the future.

Edge inferencing

- Overcome latency and connectivity issues of transferring data from the cloud or core.
- Use for medical imaging analysis to support emergency medical response.
- Enable computer vision to analyze field equipment and power autonomous vehicles.

Knock down barriers to entry for AI

Lack of data science expertise and IT infrastructure skills for AI.
Skills shortages are one of the biggest roadblocks to adopting or expanding AI.

Increasing volume and complexity of data work.
Legacy analytics approaches can't keep up with the volume and complexity of data.

Slow speed to delivery.
Insufficient processing power and lack of skills lead to delays in recognizing value from data.

Taking an AI leadership position

62%

of organizations plan to increase spending on AI including people, processes and technology.[1]

2x

more likely to have AI in production if the right skills are in place.[2]

86%

of organizations identify at least one technology roadblock to AI success.[1]

69.3%

more AI leaders are running AI workloads on servers with an average of 8 GPUs.[3]

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 deep learning (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, density and efficiency required to get started with AI and grow as needed. In addition, PowerEdge servers are available with support for up to 12 NVIDIA® graphics processing units (GPUs) to speed AI workloads — and results.

Partnerships with leading AI software companies 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.

Accelerate time to value with solutions designed for the intelligent business

To speed and simplify your AI journey, Dell offers a portfolio of Validated Designs for AI. These solutions deliver:

AI simplified

Dell Validated Designs for AI are jointly engineered and validated with NVIDIA GPUs, NVIDIA AI Enterprise suite, and other NVIDIA technologies to make it quick and easy to deploy a solution stack optimized to accelerate AI initiatives.

Faster AI insights

NVIDIA GPU–accelerated configurations are delivered with AI tools and frameworks in an optimized infrastructure to enable faster time to production for development and IT teams.

Proven AI expertise

Confidently deploy an engineering-tested AI solution backed by world-class Dell Technologies services and support. Select ProSupport Plus for a single point of contact for software and hardware support.

Day one
readiness to go to work on AI models[4]

10x
faster model generation[4]

60% less
time spent on AI infrastructure management[5]

20% faster
time to value for AI projects using tailor-made systems[5]

50% faster
AI development times[6]

Built to accelerate AI insights

Unleash your AI advantage with Dell PowerEdge servers

Dell helps you put AI to work anywhere in any way to fast-track innovation, powering your AI workloads with accelerated insights, all from the new PowerEdge servers, accelerated by NVIDIA GPUs. Dell PowerEdge advances accelerated compute to drive enhanced AI workload outcomes with greater insights, inferencing and visualization.

Accelerate transformation anywhere with Dell PowerEdge servers

Accelerate AI-driven innovations	Accelerate zero-trust adoption	Accelerate automation
Modernize from edge to cloud	Reinforce your security	Improve operational efficiencies

Sustainability

Dell PowerEdge XE servers

Acceleration-optimized, purpose-built for complex compute, AI/ML/DL and high performance computing (HPC) intensive workloads

	PowerEdge XE9680 Powerful and flexible for no compromise accelerated AI	PowerEdge XE9640* A dense, smart-cooled server to deliver real-time AI insights	PowerEdge XE8640* Superior performance with a GPU-optimized design	PowerEdge XE8545 All-in-one server for AI, machine learning (ML) and HPC
Applications and use cases	• AI/ML/DL training, HPC, CRISP • Generative AI • 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	• AI/ML training and inferencing, small and medium data set language models
Processor	• 2x 4th-generation Intel® Xeon®Scalable processors	• 2x 4th generation Intel Xeon Scalable processors	• 2x 4th-generation Intel Xeon Scalable processors	• 2x 3rd-generation AMD® EPYC™ processors
GPU support	• Up to 8x NVIDIA H100 SXM5 or NVIDIA A100 SXM4 GPUs with full NVLink™ connectivity	• Up to 4 x Intel GPUs	• Up to 4x NVIDIA H100 SXM5 GPUs with full NVLink connectivity	• Up to 4x NVIDIA A100 SXM4 GPUs with NVLink
Features	• 6U rack height • Air-cooled up to 35°C • 32 DDR5 DIMMs • Up to 10 x16 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 4 x PCIe Gen5 slots	• 4U rack height • Air-cooled up to 35°C • 32 DDR4 DIMMs • Up to 4 x16 PCIe Gen4 slots

* Available in 1H2023

Dell PowerEdge rack servers

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

	PowerEdge R760xa* Flagship server for GPU-based workloads	PowerEdge R750xa Purpose-built flexibility	PowerEdge R750/7525/7515 R650/6525/6515 Mainstream performance	PowerEdge XR12 Edge performance
Applications and use cases	• AI/ML/DL training and inferencing, analytics and HPC • Generative AI and dense inferencing • VDI and performance graphics	• AI/ML/DL training and inferencing, analytics and HPC • VDI and performance graphics	• Light-duty AI/ML/DL training and inferencing • VDI, performance graphics • Edge	• Edge AI training and inferencing • Telco • Rendering/modeling
Processor	• 2x 4th-generation Intel Xeon Scalable processors	• 2x 3rd generation Intel Xeon Scalable processors	• Up to 2x 3rd-generation Intel Xeon Scalable or 3rd generation AMD EPYC processors	• 1x 3rd-generation Intel Xeon Scalable processor
GPU support	• Up to 4x double-wide or 12x single-wide NVIDIA PCIe GPUs	• Up to 4x double-wide or 6x 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	• 2U rack height • Air cooled up to 35°C • 32 DDR5 DIMMs • Up to 4x PCIe Gen4 slots	• 1U or 2U rack height • Air-cooled up to 35°C • 32 DDR4 DIMMs • Up to 8 x PCIe 4 Gen4 slots	• 2U rack height • Operational tolerance from –5°C to 55°C • Up to 4x PCIe 4 Gen4 slots

* Available in 1H2023

Achieve near-bare-metal performance

97.5%
of bare metal performance using VMware[7]

66%
increase in performance per watt[8]

66%
increase in High-Performance Linpack (HPL) performance[9]

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.

H100 SXM	H100 PCIe	A100	A30	A10
Highest performance AI, ML training and exascale HPC	Highest performance AI, ML training and exascale HPC	Performance AI, ML training and inference	Mainstream graphics and AI inferencing	Accelerated graphics and video with AI for mainstream enterprise servers
• 3,958 TFLOPS FP8 Tensor Core* • NVLink: 900GB/s PCIe Gen5 • Up to 7 MIGs @ 10GB each	• 3,026 TFLOPS FP8 Tensor Core* • NVLink: 600GB/s PCIe Gen5 • Up to 7 MIGs @ 10GB each • NVIDIA AI Enterprise software included • NVIDIA vGPU • software support	• 312 TFLOPS FP16 Tensor Core* • NVLink Bridge for up to 2 GPUs: 600 GB/s • Up to 7 MIGs @ 10GB each • NVIDIA AI Enterprise software included • NVIDIA vGPU software support	• 165 TFLOPS TF32 Tensor Core* • NVLink Bridge for up to 2 GPUs: 200 GB/s • Up to 4 GPU instances @ 6GB each • NVIDIA AI Enterprise software included • NVIDIA vGPU software support	• 250 TFLOPS FP16* • PCIe Gen4x16 • NVIDIA AI Enterprise software included • NVIDIA vGPU software support

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 AI within reach.

*With structural sparsity enabled.

9X
faster AI training on the largest models[10]

30X
faster AI inference performance on the largest models[11]

3,958
TFLOPS FP8 Tensor Core[12]

L40	A40	A16	L4	A2
Highest performance graphics and rendering	High performance graphics and rendering	Multimedia-rich VDI to enable remote work including CAD/ CAM/CAE	Breakthrough universal accelerator for efficient video, graphics, and AI	Entry-level GPU for AI inferencing at the edge
• 90.5 FP32 TFLOPS (non-Tensor) • 724.1 FP8 Tensor TFLOPS with FP32 accumulate* • NVIDIA AI Enterprise software included • NVIDIA vGPU software support • OVX support for NVIDIA Omniverse	• 299.4 BF16 Tensor TFLOPS with FP32 accumulate* • NVLink 112.5 GB/s (bidirectional) • NVIDIA AI Enterprise software included • NVIDIA vGPU software support	• 4x 35.9 TFLOPS FP16* • PCI Express Gen 4 x16 • NVIDIA AI Enterprise software included • NVIDIA vGPU software support	• 485 TFLOPS FP8* • PCIe Gen4 x16 • NVIDIA AI Enterprise software included • NVIDIA vGPU software support	• 36 TFLOPS FP16 Tensor Core* • PCIe Gen4 x8 • NVIDIA AI Enterprise software included • NVIDIA vGPU software support

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

*With structural sparsity enabled.

NVIDIA technologies are built in

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

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 H100 GPU

The NVIDIA H100 Tensor Core GPU delivers unprecedented performance, scalability and security to every data center. The NVIDIA H100 PCIe GPU includes NVIDIA AI Enterprise software suite for streamlined AI development and deployment. It delivers 9X faster AI training[10] and 30X faster AI inference performance on the largest models.[11]

NVIDIA A100 GPUs

Accelerate AI workloads with up to 20X higher performance over the prior generation. The A100 supports NVLink bridge, the world's first high-speed GPU interconnect offering a significantly faster alternative for multi-GPU systems than traditional PCIe-based solutions.[14]

NVIDIA AI Enterprise on VMware vSphere

NVIDIA AI Enterprise is an end to end, cloud native suite that helps you start your AI journey - without the need for AI expertise - through supported containers, frameworks and workflows. It's certified to run on NVIDIA Certified Systems™ from Dell Technologies, and includes AI development and deployment tools, infrastructure optimization software, and global enterprise support to keep AI projects on track. This enables you to focus on harnessing the business value of AI, not on deploying the infrastructure.

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 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. NVIDIA LaunchPad is proudly built on Dell PowerEdge servers. Learn more at nvidia.com/dell-launchpad.

NVIDIA BlueField data processing units (DPUs)

By offloading, accelerating and isolating a broad range of advanced networking, storage and security services, BlueField DPUs provide a secure and accelerated infrastructure for any workload, in any environment, from cloud to data center to edge.

Recommended configurations

Workload	Use cases	Recommended configurations	
HPC/AI/ML/DL training Generative AI	• Natural language processing (NLP) • Large language models (LLM) • Large recommendation engine training • HPC, modeling and simulation	• PowerEdge XE9680	• H100 SXM GPUs
HPC/AI/database/analytics	• HPC • AI/ML/DL training and inferencing • Medium data set language models • NLP • Image recognition • Modeling and simulation • Molecular dynamics • Genome sequencing	• PowerEdge XE9680 • PowerEdge XE8640	• H100 SXM GPUs • A100 SXM GPUs
Performance graphics/VDI/modeling	• Digital Twins and 3D world/Metaverse • Performance graphics • CAD/CAM/CAE • Virtualization • HPC	• PowerEdge R760xa • PowerEdge R750xa • PowerEdge R750 • PowerEdge R7525	• L40 GPUs • A40 GPUs
Mainstream AI	• HPC • Analytics • GPU database acceleration • AI/ML training and inferencing • Light-duty AI training • A/ML training and inferencing	• PowerEdge R960 • PowerEdge R760xa/R750xa • PowerEdge R760/R750 • PowerEdge R7625/7525 • Other rack servers	• A2, A10, A30 or A100 GPUs • L4 GPUs
VDI and virtualization	• Rich collaboration for power users • VDI for knowledge workers	• PowerEdge R760xa/R750xa • PowerEdge R760/R750 • PowerEdge R7625/R7525 • PowerEdge R960	• A10 or A16 GPUs • L4 GPUs
Mainstream graphics and VDI	• Graphics rendering	• PowerEdge R760/R750 • PowerEdge R7625/R7525 • Other rack servers	• A10 GPUs • L4 GPUs
Inferencing/edge/VDI	• Edge inferencing	• PowerEdge XR12 • PowerEdge R760xa/R750xa • PowerEdge R760/R750 • PowerEdge R7626/R7525 • Other rack servers	• A2 GPUs • L4 GPUs

Customer successes

1 Taboola delivers content recommendations on a massive scale.

Taboola® takes advantage of extraordinary computing power and simplified management to attain the maximum performance, scalability and automation to train and run AI models that provide billions of relevant content recommendations every day.

150,000
AI-driven requests processed per second

6x
improvement in AI-based inferencing

50 milliseconds
to deliver real-time recommendations

"We now get up to six times the performance on our AI-based inferencing...This helps reduce our costs."

— Ariel Pisetzky, VP of IT and Cybersecurity, Taboola

Read the **case study.**

2 Duos Technologies keeps trains moving at full speed.

Duos Technologies® utilizes AI at the edge powered by NVIDIA GPU–accelerated Dell PowerEdge servers to process and analyze data in real time, providing prompt, actionable insights so trains don't have to stop for inspections.

120:1
reduction in inspection time

1.3TB
of data processed and analyzed daily per site

$3,000 USD
savings per instance for server recovery

"We count on PowerEdge servers to process and analyze the images and other data from the cameras and sensors 24x365, using our AI models."

— David Ponevac, Chief Technology Officer, Duos Technologies

Read the **case study.** Watch the **video.**

3 University of Cambridge accelerates scientific discovery

Dell Technologies helps the University of Cambridge build an HPC and data storage system to help solve some of today's most demanding data-driven simulation and AI challenges.

3.8
petaFLOPS

74,000
cores

500
gigabytes per second

"You cannot feed these people enough compute. They will eat whatever you give them. Cambridge's supercomputer provides researchers with the fast and affordable supercomputing power they need for AI work."

—Dr. Paul Calleja, Director of Research Computing Services, University of Cambridge

Read the **case study.**

4 University of Pisa extends the power of AI

Thanks to solutions from Dell Technologies, VMware and NVIDIA, the University of Pisa runs traditional and AI workloads on the same systems, flexing to meet demand while making the most of IT resources.

Zero
silos of AI-specific systems

One
platform for deploying virtual desktops and apps

Multiple
workloads supported on the same infrastructure

"The biggest benefit of virtualized GPUs is flexibility, in the sense that you can design and adapt your enterprise infrastructure to AI workloads."

— Maurizio Davini Chief Technology Officer, University of Pisa

Read the **case study.**

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 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 set up 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.

35K+

services and support members to help create a roadmap to AI success[15]

$0

to collaborate with Dell Technologies AI experts[16]

10

Dell HPC and AI Centers of Excellence worldwide[17]

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, 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 can help you win in the age of AI.

Learn more

Dell.com/PowerEdge

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 NVIDIA AI Enterprise software. With NVIDIA and Dell Technologies, you can take AI where you never thought before.

Published in the USA 02/23 eBook dell-nvidia-ai-EB-101

[1] ESG infographic, Modernize Compute for an AI-driven Future with Dell Servers and NVIDIA, 2022.

[2] Compared to "AI evaluators." Source: IDC Analyst Brief sponsored by Dell Technologies and NVIDIA, Scaling Skills for AI: Lessons from Early Adopters, August 2022.

[3] Compared to "AI evaluators." Source: IDC white paper sponsored by Dell Technologies and NVIDIA, What Businesses with AI in Production Can Teach Those Lagging Behind, August 2022.

[4] With Dell Precision Data Science Workstations. See DSW Ready Day One Guide.

[5] With Dell Validated Designs for AI. Forrester, The Total Economic Impact™ Of Dell Validated Designs For AI, August 2022.

[6] With Dell Precision Data Science Workstations. Dell Technologies case study, AI deep learning extends data science horizons, February 2021.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] For the NVIDIA H100 SXM GPU with structural sparsity enabled. Specifications are one-half lower without sparsity. Source: NVIDIA, NVIDIA H100 Tensor Core GPU, accessed January 2023.

[13] NVIDIA website, Accelerating the Most Important Work of Our Time, accessed June 2022.

[14] NVIDIA website, NVIDIA NVLink, accessed June 2022.

[15] Dell Technologies, Key Facts, 2022.

[16] At Dell Technologies Customer Solutions Centers and HPC & AI Innovation Lab. Speak with your sales representative for more details.

[17] See dell.com for more details.