

Acceleration-Optimized servers and accelerator portfolio

Redefine data visualization and insights with AI



Accelerate insight and innovation

For the digital enterprise, success hinges on leveraging big, fast data. But as data sets grow, traditional data centers are starting to hit performance and scale limitations — especially when it comes to ingesting and querying real-time data sources.

While some have long taken advantage of accelerators for speeding visualization, modeling and simulation, today, more mainstream applications than ever before can leverage accelerators to boost insight and innovation with generative AI models customized to deliver precise results with enterprise business data. Accelerators such as graphics processing units (GPUs), complement and accelerate CPUs, using parallel processing to crunch large volumes of data faster. Accelerated data centers can also deliver better economics, providing breakthrough performance with fewer servers, resulting in faster insights and lower costs.

Organizations in multiple industries are adopting server accelerators to outpace the competition — honing product and service offerings with data-gleaned insights, enhancing productivity with better application performance, optimizing operations with fast and powerful analytics, and shortening time to market by doing it all faster than ever before.

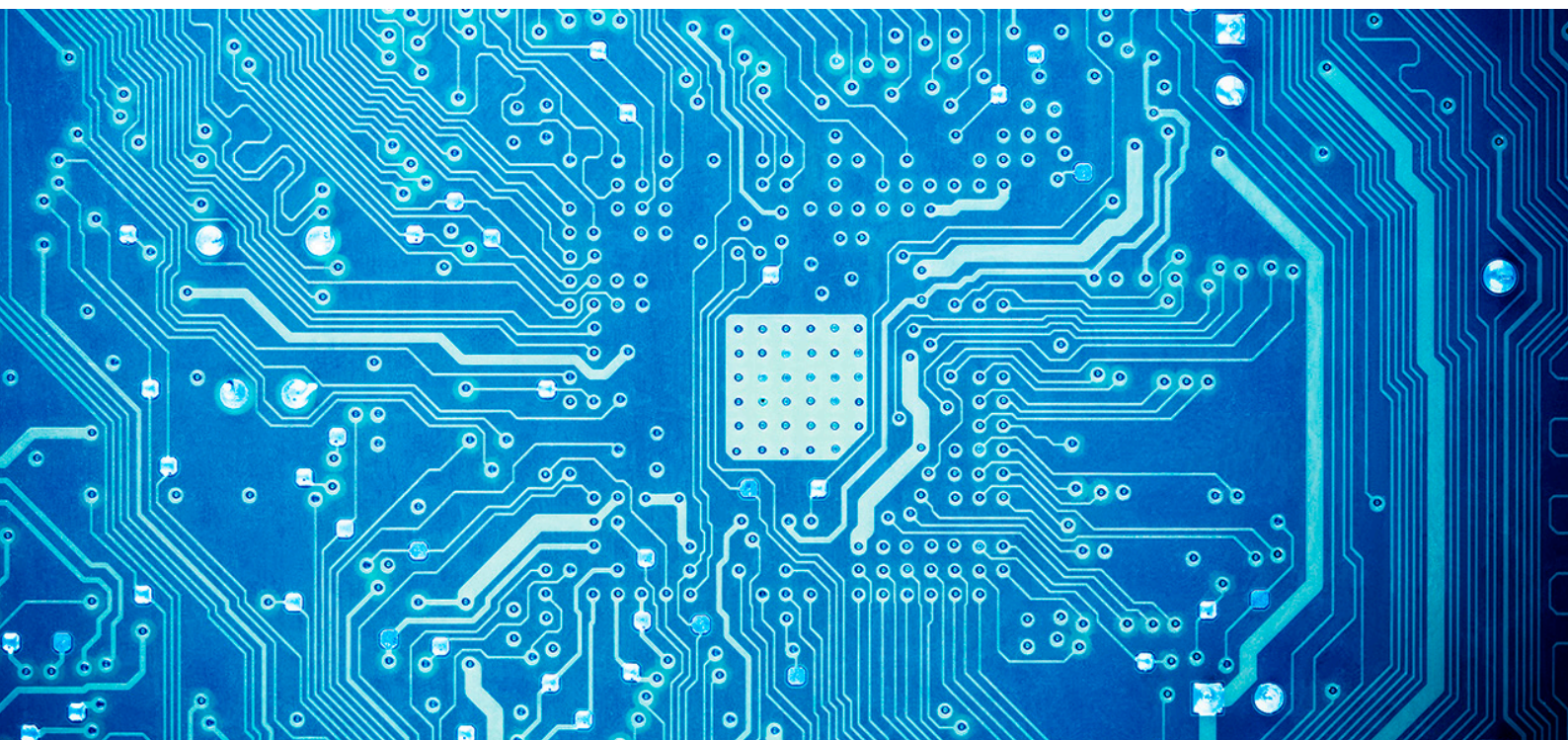
Dell Technologies offers a choice of server accelerators in Dell PowerEdge servers, so you can turbo-charge your applications.

over

77%

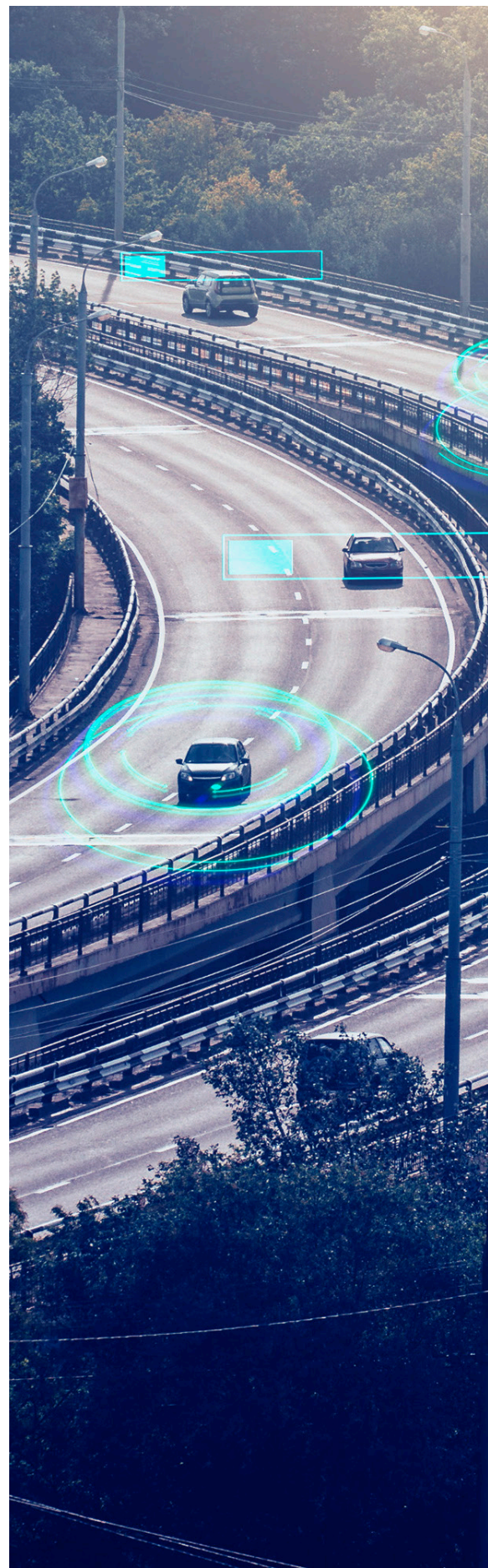
of organizations are either exploring potential use cases or investing significantly in GenAI technologies¹

¹ IDC Future Enterprise Resiliency and Spending Survey, Wave 6, July 2023



Emerging and traditional use cases for AI

- **Generative AI and large language models** – Accelerators are powering generative AI transformers and language processing technologies which can enable more intelligent systems with a richer understanding of language than ever before. These tools can now combine natural language processing, computer vision, and audio analysis to accept complex queries and deliver multimodal results.
- **Large-scale recommendation engines** – Accelerators excel in powering deep learning models to continuously improve advertising and search recommendations, both on relevance and timeliness, from advertisers to reach their audience and affect ad ranking models, for example.
- **Natural Language Processing (NLP)** – accelerators help boost, via machine learning, the programming of systems to process and analyze language data from spoken to written. A model can then accurately extract information and insights as well as learn new natural language tasks including language modeling, parsing, summarizing and other syntactic/semantic analysis methods, across global languages.
- **Digital twins** – these are virtual representations of objects, systems, processes, updated from real-time data and using simulation, machine learning and reasoning to drive decision-making. Digital twins are synchronized to real-world systems and data to help organization simulate, optimize products, people, equipment, and processes in real-time before ever going to production.
- **Machine and deep learning** – Accelerators have taken AI from theory to mainstream by enabling the parallel processing power required to speed both training and inferencing workloads.
- **Accelerated databases** – Accelerators can help speed aggregations, sorts and grouping operations to solve complex analytics operations that overload traditional databases.
- **Streaming data** – The Internet of Things (IoT) has created a firehose of data. Accelerators enable simultaneous ingestion, exploration and visualization of streaming data for real-time analysis.
- **Visualization** – Accelerators enhance performance for 3D visualization applications such as computer-aided design, enabling software to draw models in real time as the user moves them.
- **Modeling and simulation** – Accelerators can provide modeling and simulation for early evaluation, fast testing of design modifications enabling more iterations.
- **Financial modeling** – Accelerated HPC and artificial intelligence (AI) solutions are revolutionizing analytics tools, enabling the industry to leverage massive data sets to better understand risk and return.
- **Seismic processing** – Oil & Gas companies are finding new and better ways to extract information from massive seismic data stores, leveraging accelerators to speed time to results and shave costs.
- **Signal processing** – Accelerators enable providers to model and analyze signal data streams coming in from computers, radios, videos and cell phones in real-time.



Leveraging Innovation and accelerated architectures

As the prior uses cases suggest, the continued adoption of AI, ML, HPC workloads and VDI is adding complexity to data center and business operations, as workforce grows globally and remotely, as well as demanding use cases becoming more mainstream. For example, Artificial Intelligence has generated a wide range of new and hyper-tailored solutions for customers. Companies now leverage AI to automate many business processes, shifting human resources from one business unit to other areas for value creation.

Choosing GPUs and other accelerated architectures and products is a key decision IT teams have in their hands. And once that decision is made, for the appropriate workloads, then infrastructure strategy and product choices are addressed.

Accelerated Insights – the leading edge of innovation from PowerEdge Servers

To design an infrastructure to deliver the capabilities which can make organizations successful with AI and other demanding workloads, requires a modern architecture approach where one of the biggest innovations is improved performance with the addition of dense acceleration, at scale. Improved performance is not only about implementing complete solution and infrastructure strategy, but also starts with innovations in the building blocks to also help provide other benefits, including improved costs, security, and thermal/power design.

There are a number of innovations within the PowerEdge server family which enable drastic performance improvements. From architectures specifically designed to support acceleration to thermally optimized designs, today's workloads demand higher quality components and subsystems to flawlessly drive workload operations.

The PowerEdge Adaptive Compute approach enables servers engineered to optimize the latest technology advances for predictable profitable outcomes. Here are a few of the improvements in the PowerEdge portfolio:

- **Focus on Acceleration** – Support for the most complete portfolio of GPUs, delivering maximum performance for HPC modeling & simulation, generative AI/ML/DL training and inferencing, analytics and rich-collaboration application suites and workloads
- **Thoughtful Thermal Design** – New thermal solutions and designs to address dense heat-producing components, and in some cases, front-to-back air-cooled designs
- **Dell Multi Vector Cooling** – Streamlined, advanced thermal design for airflow pathways within the server
- **Dell Direct Liquid Cooling** – Extending liquid cooling support across more PowerEdge servers and their CPUs for exceptional heat removal capability

Dell PowerEdge XE9680
delivers the industry's best
AI performance²

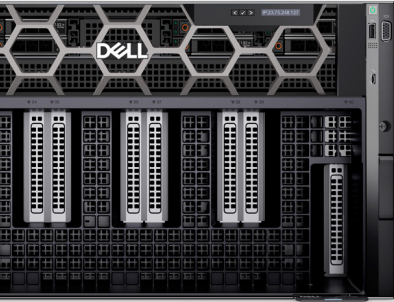
² Based on Dell analysis of publicly available performance results and specifications of comparable OEM Servers as of 17 May 2023.

Accelerated AI Insights

Engineered to optimize the latest technology advances for predictable profitable outcomes



PowerEdge servers for accelerated workloads



No-compromise accelerated AI

XE9680 is designed to drive business insights in the most demanding Deep Learning and modeling applications, from large natural language processing models and recommendation engines to complex research and academia problems.

- Highest performance for HPC and Enterprise
- 8x AMD Instinct MI300X GPUs with Infinity Fabric, 8x Intel Gaudi 3 GPUs with RoCE interconnectivity, 8x NVIDIA H100 or 8x NVIDIA H200 Tensor core GPUs with NVLink
- Air-cooled operation

Ideal workloads: Generative AI, Large Language Models, Natural Language Processing, large recommendation engine training, molecular dynamics, genomic sequencing modeling and simulation.

Applicable GPUs:

Either AMD MI300X OAM, Intel Gaudi 3 OAM or, NVIDIA H100 SXM, or NVIDIA H200 SXM



Dense acceleration

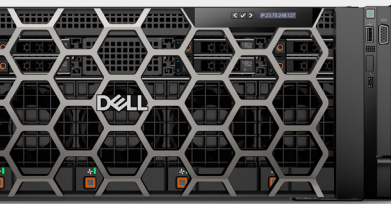
XE9640 boosts insights from your growing data sets with AI acceleration technology designed for optimal performance, fastest time-to-value, in a liquid-cooled environment.

- Mainstream 2U form factor enables highest GPU density per rack AI operations
- 4x NVIDIA H100 Tensor core GPUs with NVLink
- Liquid-cooled CPU and GPU operation

Ideal workloads: Natural Language Processing, large recommendation engine training, modeling & simulation, Artificial Intelligence and ML/DL training for object recognition

Applicable GPUs: NVIDIA H100 SXM

PowerEdge servers for accelerated workloads



Purpose-built performance

XE8640 helps businesses unlock insights with purpose-built performance in a dense air-cooled server for AI, removing traditional computational boundaries of real-time insights.

- Optimized balance of performance for diverse applications
- 4x NVIDIA H100 Tensor core GPUs with NVLink
- Air-cooled operation with liquid-assisted CPU/GPU cooling radiator

Ideal workloads: Medium data set language Models, Natural Language Processing, modeling & simulation, Artificial Intelligence, ML/DL training and inferencing, image recognition

Applicable GPUs: NVIDIA H100 SXM



Purpose-built scale up server for GPU applications

R760xa maximizes results from AI to Modeling & Simulation applications with maximum flexibility and the latest 4th or 5th Generation Intel® Xeon® Scalable Processors.

R760xa is optimized to tackle GPU workloads and deliver outstanding performance for demanding and emerging applications.

- Maximize performance with Dell's broadest selection of PCIe GPU configurations
- Front-to-back air-cooled design, optimized for PCIe GPUs
- R760xa supports up to 12 Single-wide GPUs or 4 Double-wide GPUs, up to 350W
Supports all GPU cards

Ideal workloads: AI & ML training and inferencing, data analytics, HPC, VDI & Performance graphics

Applicable PCIe GPUs:

AMD MI210

Intel Flex 140

NVIDIA H100 NVL, L40S, L40, L4, A16, A2

Accelerated GPU servers, at-a-glance

Model	Workloads	Memory	Processor	Storage	Accelerators	Details
PowerEdge XE9680	AI ML DL Training, HPC, CRISP, Healthcare, CSP/HPCaaS, Finance, Academia	32 (4TB)	Two 4 th or 5 th Generation Intel® Xeon® Scalable processors	8x 2.5" or 16x E3.S	8x 700W SXM or 8x 750-850W OAM	Family page Product Video Specification Sheet Technical Guide
PowerEdge XE9640	AI ML DL Training, HPC, Modeling & Simulation, Healthcare, Life Sciences, Finance	32 (4TB)	Two 4 th or 5 th Generation Intel® Xeon® Scalable processors	4x 2.5"	4x 700W SXM or 4x 600W OAM	Family page Product Video Specification Sheet Technical Guide
PowerEdge XE8640	AI ML DL Training, HPC, Oil & Gas, Healthcare, Life Sciences, Finance	32 (4TB)	Two 4 th or 5 th Generation Intel® Xeon® Scalable processors	8x 2.5"	4x 700W SXM	Family page Product Video Specification Sheet Technical Guide
PowerEdge R760xa	AI-ML/DL training and inferencing, HPC, render farms and virtualization	32 (4TB)	Two 4 th or 5 th Generation Intel® Xeon® Scalable processors	8x 2.5" or 6x 2.5" NVMe 6x E3.S	4x 350W DW or 12x 75W SW	Family Page Product Video Specification Sheet Technical Guide

PCIe GPUs, DPUs for Dell PowerEdge servers

Turbo-charge your applications with performance accelerators available in select Dell PowerEdge tower and rack servers. The number and type of accelerators that fit in PowerEdge servers is based on the physical dimensions of the PCIe cards.

Double-wide (DW) accelerators take up two slots and include: AMD MI210, Intel Flex 140, NVIDIA H100 NVL, L40S, L40, and A16 GPUs and, Single-wide (SW) accelerators, including the NVIDIA L4 and A2, take up one PCIe slot. Dell PowerEdge engineering qualifies accelerators with servers based on demand. Dell Technologies also works with a wide range of partners to create and sell specific combinations for particular vertical market applications.

GPUs vary in number of cores, amount of memory, and power and cooling requirements. For example, the NVIDIA H100 NVL has up to 94GB memory, and uses up to 400 watts.

GPUs

Graphics processing units (GPUs) are co-processors designed to accelerate compute performance. A GPU typically has thousands of cores designed for efficient execution of mathematical functions. Portions of a workload are offloaded from the CPU to the GPU, while the remainder of the code runs on the CPU, improving overall application performance.

Dell offers a range of GPUs as PCIe cards that fit into server PCIe slots. The Dell PowerEdge XE product family offers support for 4x or 8x GPU assemblies on Open Compute Project Accelerator Module (OAM) or NVIDIA SXM modules mounted on the server motherboard.

DPUs

A Data Processing Unit (DPU) combines computing, networking, and programmability to offload CPUs and deliver software-defined, hardware-accelerated solutions for the most demanding workloads.

Parallel processing

Parallel processing is a method of simultaneously breaking up and running program tasks on multiple microprocessors, reducing processing time.

Optimize the code

To take full advantage of server accelerators, optimize the software code. For many applications, four lines of code can provide a boost.

NVIDIA Hopper and Ampere and Tensor Core GPUs



NVIDIA Hopper and Ampere Core GPUs deliver the horsepower needed to run deep learning training, high performance data analytics, visualization and other workloads faster than ever before. Plus, NVIDIA GPUs deliver high performance and user density for virtual desktop infrastructure (VDI). Deliver mainstream AI on with NVIDIA AI Enterprise.

- [Hopper core GPU](#)
- [Ampere core GPU](#)
- [NVLink™ Fabric interconnect](#)
- [GPU CLOUD™](#) containers
- [Software application catalog](#) and [developer resources](#)
- [NVIDIA AI Enterprise](#)

Model	Workloads	Memory	Graphic Bus/ System interface	Slot width	Max Power Consumption	Server support
H200	AI/HPC	141 GB HBM3e	SXM5/NVLink	N/A	700W	XE9680 (8x H200)
H100	HPC/AI/ML/DL Training	80 GB HBM3	SXM5/NVLink	N/A	700W	XE9680 (8x H100), XE8640/XE9680 (4x H100)
H100 NVL	AI/HPC	94 GB HBM3	PCIe Gen 5x16	Double-wide	350-400W	R660, R760, R760xa
L40S	AI/Performance Graphics/VDI	94 GB HBM3	PCIe Gen 4x16	Double-wide	700W	R760, R760xa, R7625, R7615
L40	Performance graphics/VDI	48 GB GDDR6	PCIe Gen4 x16	Double-wide	300W	R750, R7525
A16	VDI	64 GB GDDR6	PCIe Gen4 x16	Double-wide	250W	R750, R7525, R7515
A10	mainstream graphics/VDI	24 GB GDDR6	PCIe Gen4 x16	Single-wide	150W	R750, R7525
L4	Inferencing/Edge/VDI	24 GB GDDR6	PCIe Gen4 x16	Single-wide	72W	R750, R7525, R650
A2	Inferencing/Edge/VDI	16 GB GDDR6	PCIe Gen4 x8	Single-wide	60W	R750, R7525, R7515, R650, R6525, R6515, C6525

³ w/Nvlink bridge is supported on; w/Nvlink bridge is supported on and; w/NVLink bridge is supported on and; MI210 w/Infinity Fabric Link bridge is supported on H100 and w/Nvlink bridge will be supported on Max1100 w/XeLink bridge is supported on R760XA

DW - Double Wide, SW - Single Wide, FH- Full Height, FL - Full Length, HH - Half Height, HL - Half Length

NVIDIA-Certified Dell Systems brings together NVIDIA GPUs and NVIDIA networking in servers and hyperconverged infrastructure from Dell Technologies in optimized configurations.

These systems are validated for performance, manageability, security, and scalability and are backed by enterprise-grade support from NVIDIA and Dell Technologies.

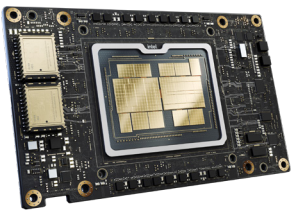


- Deliver infrastructure to drive a diverse range of accelerated workloads for the enterprise
- Excellent performance
- Reduce time to deployment
- Secured, no-compromise operations and workflows
- Designed for single to multi-node configs, optimal Scale-out and clusters

Learn more about Dell PowerEdge servers with NVIDIA-Certified solutions [here](#).

Consult our [matrix of supported PowerEdge servers and partner accelerators](#) to deliver the optimal configuration for your applications and workloads.

Intel GPUs



The Intel® Data Center GPU Max Series is designed to take on the most challenging high-performance computing (HPC) and AI workloads.

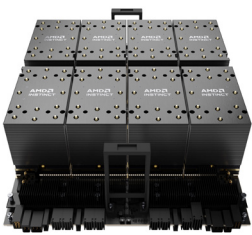
Available on Dell XE9640 servers.

Unleash the Power of Intel Data Center GPU Max Series through software: For the data center GPU, Intel oneAPI and AI tools help you realize maximum performance from the innovative hardware's advanced capabilities like Intel® Xe Matrix Extensions (Intel® XM), vector engine, Intel® Xe Link, data type flexibility, and more.

- [Intel® Data Center GPU Flex Series Overview](#)

Model	Workloads	Memory	Graphic Bus/ System interface	Slot width	Max Power Consumption	Server support
Intel Gaudi 3	AI inferencing and training	128 GB HBM3	OAM with RoCE interconnectivity	N/A	850W	XE9680 8X Gaudi 3
Flex 140	Inferencing/Edge	12 GB GDDR6	PCIe Gen4 x8	SW	300W	R760xa, R760, R660

AMD GPUs



Built on CDNA architecture, AMD MI300X delivers top-tier High Bandwidth Memory support with 192 GB of HBM3 per accelerator with over 1.5 TB per Dell PowerEdge XE9680 server.

- [Learn more about Dell and AMD solutions for AI](#)
- [Watch Dell PowerEdge XE9680 with AMD Instinct MI300X product video](#)
- [Read top 5 reasons to choose Dell PowerEdge XE9680 with AMD Instinct MI300X Infographic](#)
- [Read Dell PowerEdge XE9680 with AMD Instinct MI300X and porting LLMs to AMD ROCm Infographic](#)
- [Read Dell PowerEdge XE9680 with AMD Instinct MI300X and ROCm eBook](#)
- [Learn how multimodal RAG-based AI accelerates Healthcare](#)

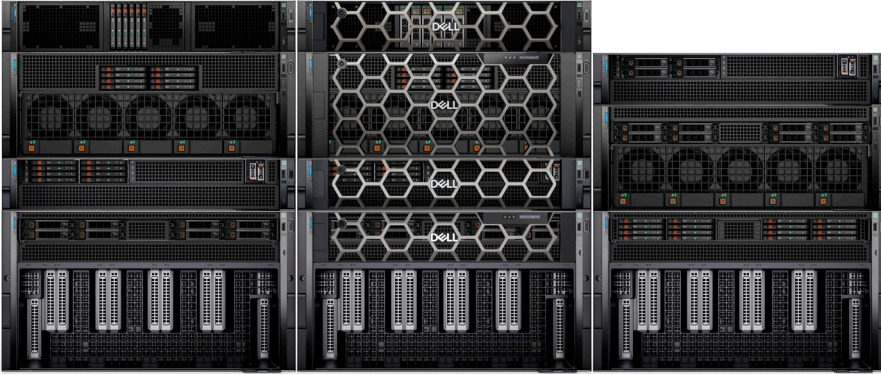
Model	Workloads	Memory	Graphic Bus/ System interface	Slot width	Max Power Consumption	Server support
MI300X	AI / HPC	192 GB HBM3	OAM with AMD Infinity Fabric Links	N/A	750W	XE9680 8x MI300X
MI210	AI / HPC	64 GB HBM2e	PCIe Gen4x16/ Infinity Fabric Link bridge8	D/W	300W	R7625, R7615, R760xa

⁴ w/Nvlink bridge is supported on; w/Nvlink bridge is supported on and; w/NvLink bridge is supported on and; MI210 w/ Infinity Fabric Link bridge is supported on; H100 and w/Nvlink bridge will be supported on Max1100 w/XeLink bridge is supported on R760XA

DW - Double Wide, SW - Single Wide, FH- Full Height, FL - Full Length, HH - Half Height, HL - Half Length

Dell AI Solutions

Save time with Dell Technologies and partner solutions with accelerators inside.



Dell PowerEdge Server – Accelerator Combinations

The number and type of accelerators that fit in [PowerEdge servers](#) is based on the number and type of PCIe slots in the server chassis and the accelerator form factor (FF), or the physical dimensions of the PCIe cards.

Solution	Description	Resources
AI MLOps with cnvrg.io	Standardize machine learning pipelines with cnvrg.io to minimize friction for data science and engineering teams from research to production.	<ul style="list-style-type: none"> • Solution Brief • Validated Design
Automatic Machine Learning	Automate algorithm selection, feature generation, hyperparameter tuning, and model assessment to ease and speed time to AI.	<ul style="list-style-type: none"> • Solution Brief • Validated Design
Conversational AI	Deliver extraordinary, effective, and efficient AI-enabled customer and employee experiences on voice and digital channels (including chatbots and virtual assistants)	<ul style="list-style-type: none"> • Solution Brief • Validated Design
Generative AI with NVIDIA for Inferencing	Quickly get up and running with a pre-trained model and start producing outputs and value with a joint architecture from Dell Technologies and NVIDIA.	<ul style="list-style-type: none"> • Solution Brief • Validated Design
Generative AI with NVIDIA for Model Customization and Tuning	Learn how to re-train an existing GenAI model for your own use cases, with examples of standard customization techniques such as transfer learning and prompt tuning.	<ul style="list-style-type: none"> • Solution Brief • Validated Design
Red Hat OpenShift AI on APEX Cloud Platform for Red Hat OpenShift	Implement a digital assistant by leveraging a Large Language Model (LLM) and the Retrieval Augmented Generation (RAG) framework.	<ul style="list-style-type: none"> • Solution Brief • Validated Design

Rest easier from day one with our comprehensive services

Utilize Dell Technologies Services to maximize the life and value of your PowerEdge Servers on a global scale, across 170 locations and benefit from the deep expertise of our 60K+ employees and partners.

- [ProDeploy Factory Configuration](#) – Factory-based services deliver PowerEdge servers configured to your specifications, ready to install
- [ProDeploy Rack Integration](#) – Receive PowerEdge fully configured and racked direct from our facility with optional onsite final configuration
- [ProDeploy or ProDeploy Plus](#) – ProDeploy experts are here to help, with 24/7 field-based deployment services, from planning through implementation and beyond. Choose from guided remote to fully onsite hardware and software implementation
- [Data Migration Services](#) – Efficiently move data from where it is to where it will drive innovation
- [Dell Professional Service](#) – Leverage predictive issue detection and proactively improve the performance of your critical systems, while taking advantage of an assigned Service Account Manager

Availability and terms of services vary by region. For more information and details on our entire range of offerings, please contact your Dell Technologies representative or visit us online at Dell.com/services.

Become a Dell Technologies Partner

When you join the Dell Technologies Partner Program, you are joining a partner ecosystem that together is making digital, IT, workforce, and security transformation real to organizations across the globe - every single day. Underpinning the industry's most robust portfolio from the edge to the core to the cloud is the Dell Technologies Partner Program, designed to be Simple. Predictable. Profitable.

Resources

Ready your data center to handle any workload with PowerEdge Servers PowerEdge tower servers are designed to grow with your organization, at your pace. PowerEdge rack servers combine a highly scalable architecture and optimum balance of compute and memory to maximize performance across the widest range of applications. Shop Dell PowerEdge servers at dell.com/poweredge.

Server advanced engineering provides guidance at [Support for Servers Solution Resources](#). White papers are also available at delltechnologies.com/accelerators > [resources](#) > [white papers](#). For reference architectures, visit delltechnologies.com/referencearchitectures.

See performance results

Get benchmarking data by workload, reference architectures and blogs from HPC/AI engineering at hpcatdell.com and download from [GitHub](#).

Access Education Services

Get the skills, training and certifications you need at learning.dell.com

Community resources

Join the Dell Technologies HPC/AI Community at dellhpc.org. Connect with the AI Builders Community at builders.intel.com/ai.

Visit a Dell Technologies Customer Solution Center

Experience our solutions and products with a customized engagement designed to help you address your business challenges or innovate for success. Work with our subject matter experts in our dedicated labs – stacked with the latest and greatest products and solution showcases. Remote connectivity enables you to include global team members, or work with us from your own location. Learn more at delltechnologies.com/csc.



[Learn more](#) about
Dell solutions



[Contact](#) a Dell
Technologies Expert



[View more](#) resources



[Join the conversation](#) with
[#PowerEdge](#)