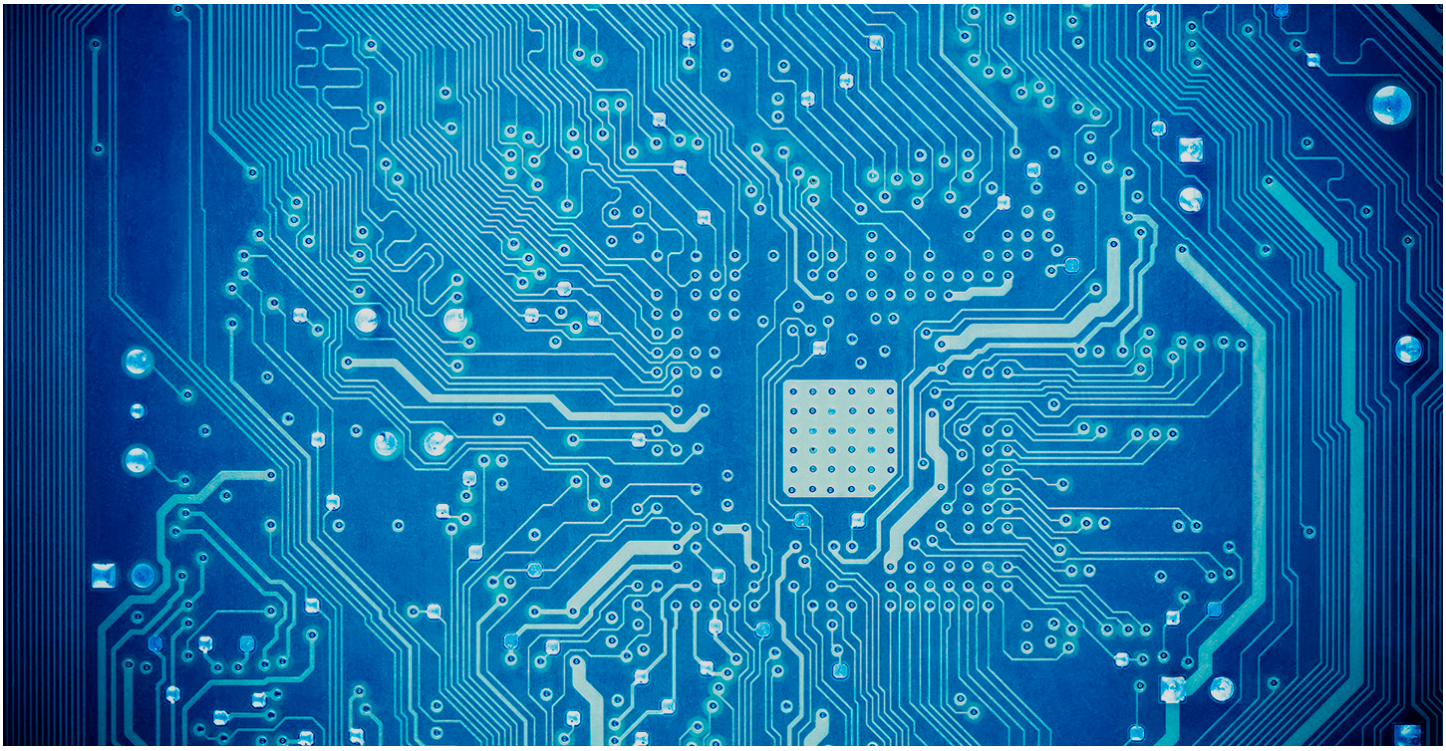
DELL Technologies

Acceleration-Optimized servers and accelerator portfolio

Redefine data visualization and insights with AI



64%

of executives believe AI is considerably important for their future[1]

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. 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 operationswith 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.

[1] ESI Thoughtlab, Driving ROI through AI, sponsored by Dataiku, Sept 2020

Emerging and traditional use cases for AI

- **Large language models** – Accelerators are powering the AI performance for language processing technologies which can enable more intelligent systems with a richer understanding of language than ever before. These tools can now read, summarize and translate texts and predict future words in a sentence letting them generate sentences similar to how humans talk and write.
- **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.
- **Predictive analytics** – AI, enabled by accelerators, can supercharge analytics, enabling dynamic correlation and delivering predictive outcomes with staggering speed, accuracy and scale.
- **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.

22

#1 performance positions with MLPerf(TM) Inference 1.1 suite, September 2021

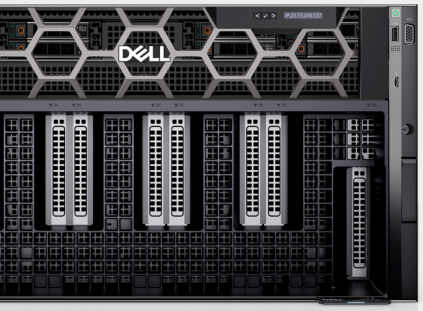
The PowerEdge Adaptive Compute approach enables servers engineered to optimize the latest technology advances for predictable profitable outcomes. Here are a few of the improvements in the PowerEdge portfolio:

- **Focus on Acceleration** – Support for the most complete portfolio of GPUs, delivering maximum performance for HPC Modeling & Simulation, 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

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 NVIDIA H100 or A100 Tensor core GPUs with NVLink
- Air-cooled operation

Ideal workloads: Large language Models, Natural Language processing, Large Recommendation engine training, Modeling & Simulation, molecular dynamics and genomic sequencing

Applicable GPUs:

NVIDIA H100 SXM or A100 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 enables highest performance AI operations
- 4x Intel Data Center Max GPUs
- Liquid-cooled CPU and GPU operation

Ideal workloads: Natural Language processing, Large Recommendation engine training, Modeling & Simulation, Artificial Intelligence, ML/DL Training for object recognition

Applicable GPUs: Intel Data Center Max 1550 OAM GPUs

Purpose-built performance

XE8640* helps businesses unlock insights with purpose-built performance in a highly dense 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

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

* available 1H, 2023

PowerEdge servers for accelerated workloads

Purpose-built server for highly intensive GPU workloads

R750xa is optimized to tackle GPU workloads and deliver outstanding performance for demanding and emerging applications.

- Maximize performance
- Front-to-back air-cooled design
- Supports all GPU cards

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

Applicable GPUs:

NVIDIA H100, A100*, A40*, A30*, A16, A2, A10

AMD MI100, MI210

Cutting edge AI, ML and HPC processing

XE8545 delivers optimized CPU and GPU performance for AI and ML training and inferencing by pairing the maximum core count AMD EPYC™ processors, highest performing Nvidia A100 GPUs, and NVLINK to maximize the time to value.

- Supercharge AI/ML and HPC performance
- Interconnected 4-way NVLINK architecture
- GPU Virtualization

Ideal workloads: AI & ML training and inferencing, HPC, GPU virtualization

Applicable GPUs: NVIDIA A100 SXM* (40GB and 80GB)

Provide extreme acceleration

R940xa is optimized to tackle workloads that are compute-intensive, combining up to 4 CPUs, up to 112 cores, with four GPUs in a powerful 1:1 ratio to drive artificial intelligence, machine learning and deep learning workloads.

- Accelerate applications
- Scale dynamically
- Streamline IT operations

Ideal workloads: GPU database acceleration, data analytics, artificial intelligence, machine learning

Applicable GPUs: NVIDIA A100 (80GB)

* NVIDIA-Certified PowerEdge server configurations. Read more here: https://www.nvidia.com/en-us/data-center/data-center-gpus/qualified-system-catalog/

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 4th Generation Intel® Xeon® Scalable processors	8 x 2.5”	4 x 700W or 4 x 500W SXM	Family page Family Video
PowerEdge XE9640	AI ML DL Training, HPC, Modeling & Simulation, Healthcare, Life Sciences, Finance	32 (4TB)	Two 4th Generation Intel® Xeon® Scalable processors	4 x 2.5”	4 x 600W OAM	Family page Family Video
PowerEdge XE8640	AI ML DL Training, HPC, Oil & Gas, Healthcare, Life Sciences, Finance	32 (4TB)	Two 4th Generation Intel® Xeon® Scalable processors	8 x 2.5”	4 x 700W SXM	Family page Family Video
PowerEdge R750xa	AI-ML/DL training and inferencing, HPC, render farms and virtualization	32 (4TB)	Two 3rd Generation Intel® Xeon® Scalable processors	12 x 3.5” or 24 x 2.5” or 16 x 2.5”	4 x 300W DW or 6 x 75W SW	Shop PowerEdge 750xa Spec Sheet Video
PowerEdge XE8545	AI ML Training and inferencing	32 (4TB)	Two 3rd Generation AMD EPYC™ processors	10 x 2.5”	4 x 500W or 4 x 400W SXM	Spec Sheet Video
PowerEdge R940xa	Data analytics, database acceleration and ML	48 (15.36TB)*	4 x 2nd Generation Intel® Xeon® Scalable processors	32 x 2.5”	4 x DW GPUs or 8 x SW GPUs or FPGAs	Shop PowerEdge 940xa Spec Sheet

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, and as SXM2 modules mounted to 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.

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: NVIDIA H100 A100, A30 and A40 GPUs; and, AMD MI210, Single-wide (SW) accelerators, including the NVIDIA A2, take up one PCIe slot. Dell PowerEdge engineering qualifies accelerators, including the NVIDIA A2, 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 CUDA cores, amount of memory, and power and cooling requirements. For example, the NVIDIA Hopper® H100 has up to 80GB memory, and uses up to 700 watts.

Software

Compute Unified Device Architecture (CUDA©) gives direct access to the GPU virtual instruction set and parallel computational elements, for the execution of compute kernels.

Via hardware description language (HDL), FPGAs can be configured to match the requirements of specific tasks or applications, in essence mimicking application-specific integrated circuits (ASICs). Both Intel and Xilinx have FPGA acceleration software stacks and development tools available for download.

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 VMware vSphere 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 for VMware

Model	Workloads	Mem-ory	Graphic Bus/ System interface	Slot width	Max Power Consumption	Server support
H100	HPC/AI/ML/DL Training	80 GB	PCIe Gen4x16/ NVLink bridge	Double-wide	350W	R750xa, R750, R7525
H100	HPC/AI/ML/DL Training	80 GB	NVLink bridge	N/A	700W	XE9680 (8xH100), XE8640 (4xH100)
A100	HPC/AI/Database Analytics	80 GB	PCIe Gen4x16/ NVLink bridge	Double-wide	300W (80GB)	R750xa, R750, R7525, XR12, R940xa, R740/XD, DSS8440
A100	HPC/AI/Database Analytics	40 / 80 GB	NVLink bridge	N/A	500W (80GB) 400W (40GB)	XE8545
A40	Performance graphics/VDI	48 GB	PCIe Gen4x16/ NVLink bridge	Double-wide	300W	R750xa, R750, R7525, XR12, DSS8440, R740, R740xd, T550
A30	Mainstream AI	24 GB	PCIe Gen4x16/ NVLink bridge	Double-wide	165W	R750xa, R750, R7525, R7515, R740, R740xd, XR12, XE2420, T550
A16	VDI, Virtualization	32 GB	PCIe Gen4x16	Double-wide	250W	R750xa, R750, R7525, R7515, R740, R740xd
A10	Mainstream graphics/VDI	24 GB	PCIe Gen4x16	Single-wide	150W	R750xa, R750, R7525, R740, R740xd, XE2420
A2	Inferencing/Edge/ VDI	16 GB	PCIe Gen4x8	Single-wide	60W	R750xa, R750, R7525, R7515, R650, C6520, R6525, R6515, C6525, XR12, XR11, R740, R740xd, R640, T550
T4	Inferencing/Edge/ VDI	16 GB	PCIe Gen3x16	Single-wide	70W	R750xa, R750, R7525, R7515, R650, C6520, R6525, R6515, C6525, XR12, XR11, DSS8440, R740, R740xd, R640, XR2, XE2420, XE7100

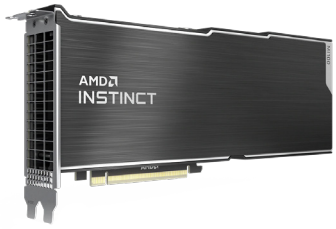
NVIDIA-Certified Dell Systems brings together NVIDIA GPUs and NVIDIA networking in servers and hyperconverged infrastructure from Dell Technologies in optimized configurations.

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.



AMD GPUs

Built on CDNA architecture, AMD MI210 delivers industry best single-precision (FP32) performance.

The AMD Instinct GPU family accelerates HPC, AI workloads, and reduces the overall cost of ownership.

Now available on Dell PowerEdge R750xa and PowerEdge R7525 servers.

- Explore MI210 Accelerator
- Read MI100 Brochure
- Read the AMD MI100 Whitepaper
- Learn how the ROCm™ open software platform enables HPC GPU computing

Model	Workloads	Memory	Graphic Bus/ System interface	Slot width	Max Power Consumption	Server support
MI100	HPC/Machine learning training	32 GB	PCIe Gen4x16	Double-wide	300W	R750xa, R7525
MI210	HPC/Machine learning training	64 GB	PCIe Gen4x16	Double-wide	300W	R750xa, R7525, R7515

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.

Accelerated Dell Technologies Solutions

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

Dell Validated Designs

Achieve more, deliver quick results and maximize efficiency.

Dell Validated Designs are purpose-designed with IT's transformation journey in mind to run intelligent applications and processes in the digital business.

Along with Dell PowerEdge servers, Dell Technologies partners and collaborates with industry leaders including Intel, Microsoft, NVIDIA, and others to optimize IT for your critical business workloads together with emerging technologies such as AI, machine learning, and blockchain.

- Validated Design for AI, including Deep Learning with NVIDIA and Cloudera
- Ready Solutions for Data Analytics
- Ready Solutions for HPC
- Ready Architectures for VDI

Validated Designs for HPC make adopting advanced computing faster and simpler. Dell delivers a choice of flexible and scalable high performance computing solutions, with servers, networking, storage, solutions and services optimized together to address use cases in a variety of industries.

Dell Ready Solutions for AI include everything you need to accelerate your AI initiatives. Making AI simpler, these integrated systems are ideal for machine and deep learning so you can get faster, deeper insights into your customers and your business.

Solutions available with Dell Technologies partners

Amulet Hotkey® virtual desktop solutions combine enterprise-class servers with virtual GPU accelerators to deliver high-density, data center–optimized solutions to simplify the transition to Windows® 10. In addition, virtual GPUs help address the growing demand for graphics-accelerated virtualization of everyday programs like Windows 10, Microsoft® Office 365®, YouTube® and more for an exceptional virtual desktop experience. Read about Amulet Hotkey customer successes.

Kinetica® is an insight engine that includes a GPU-accelerated database, visual discovery and machine learning capabilities, and accelerated parallel computing. Running on Dell EMC PowerEdge servers with NVIDIA GPUs, Kinetica helps organizations meet the challenges that come with huge quantities of complex, unpredictable data. Read the article: Explaining GPUs to Your CEO: The Power of Productization.

Tracewell Systems® deliver powerful, off-the-shelf computing technology for businesses, government agencies and OEMs in places where environmental factors create unique computing challenges, such as in the air, at sea or on the ground, in fixed and mobile installations, or in situations where integration with specialty hardware or software is required. Get data sheets, videos and resources.

[5] "Intel Arria 10 FPGAs Features," May 2019.

[6] "Alveo U200 Data Center Accelerator Card," May 2019.

Dell Technologies Acceleration Software partners

VMware® BitFusion® software disaggregates GPUs, FPGAs and/or ASICs and dynamically attaches them anywhere in the data center.

NVIDIA GRID™ Virtual Apps improve virtual desktops and accelerate server applications, with proven performance built on NVIDIA® GPUs.

AMD ROCm™ delivers an open-source exascale-class platform for accelerated computing in HPC and cluster deployments.

Kinetica® software dramatically speeds up traditional online analytics processing (OLAP) workloads using GPUs for parallel computing.

SQream Technologies® GPU-accelerated data warehouse is capable of scaling from terabytes to petabytes, adapting to any scale and workload.

FASTDATA.io PlasmaENGINE® GPU-native software enables real-time processing of infinite data in motion, over multiple nodes, with multiple GPUs.

RAPIDS is a suite of data science libraries built on NVIDIA CUDA-X for executing end-to-end data science training pipelines in NVIDIA GPUs.

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 education.emc.com. Learn how to solve problems with deep learning at the Deep Learning Institute by Dell Technologies.

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.

Discover more about PowerEdge servers

Learn more

Consult the Dell accelerators site for accelerated servers and GPUs

Technical documentation

See performance results, reference architectures and blogs from HPC engineering at hpcatdell.com

Virtual Rack

See servers and solutions in the virtual rack esgvr.dell.com

Join the Dell Technologies HPC Community

A worldwide technical forum that fosters the exchange of ideas dellhpc.org