Accelerated servers and accelerator portfolio

Redefine data visualization and insights
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), field programmable gate arrays (FPGAs) and Intelligence Processing Units (IPUs) 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 EMC PowerEdge servers, so you can turbo-charge your applications.

Use cases for accelerators

- **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.

64% of executives believe AI is considerably important for their future.¹

---

¹ ESI Thoughtlab, Driving ROI through AI, sponsored by Dataiku, Sept 2020
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 use 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.

Adaptive Compute – 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, AI-ML/DL training and inferencing, DB analytics and VDI 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
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 A100*, A40*, A30*, A16
AMD MI100

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 (40GB)

Efficient Machine Learning platform built for density
The Dell DSS 8440 allows businesses to scale-up significantly to support demanding AI workloads, with the ability to leverage up to 10 GPU accelerators for machine learning, or up to 16 GPU accelerators for inferencing.
- Scale-up GPU density as needed
- Provision GPU virtualization to workloads and users
- Support multi-tenant environments

Ideal workloads: Artificial intelligence, machine learning training and inferencing, Deep Learning

Applicable GPUs: NVIDIA A100* (40GB), A40*
Accelerated GPU servers, at-a-glance

<table>
<thead>
<tr>
<th>Model</th>
<th>Workloads</th>
<th>Memory</th>
<th>Processor</th>
<th>Storage</th>
<th>Accelerators</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>PowerEdge R750xa</td>
<td>AI-ML/DL training and inferencing, HPC, render farms and virtualization</td>
<td>32 (8TB)</td>
<td>Two 3rd Generation Intel® Xeon® Scalable processors</td>
<td>12 x 3.5&quot; or 24 x 2.5&quot; or 16 x 2.5&quot;</td>
<td>4 x 300W DW or 6 x 75W SW</td>
<td>Shop PowerEdge 750xa Spec Sheet Video</td>
</tr>
<tr>
<td>PowerEdge XE8545</td>
<td>AI ML Training and inferencing</td>
<td>32 (8TB)</td>
<td>Two 3rd Generation AMD EPYC™ processors</td>
<td>10 x 2.5&quot;</td>
<td>4 x 500W DW or 4 x 400W DW</td>
<td>Spec Sheet Video</td>
</tr>
<tr>
<td>PowerEdge R940xa</td>
<td>Data analytics, database acceleration and ML</td>
<td>48 (15.36TB)*</td>
<td>4 x 2nd Generation Intel® Xeon® Scalable processors</td>
<td>32 x 2.5&quot;</td>
<td>4 x DW GPUs or 8 x SW GPUs or FPGAs</td>
<td>Shop PowerEdge 940xa Spec Sheet</td>
</tr>
<tr>
<td>Dell DSS8440</td>
<td>AI ML Training and High performance compute</td>
<td>24</td>
<td>2 x 2nd Generation Intel® Xeon® Scalable processors</td>
<td>10 x 2.5&quot;</td>
<td>Up to 10 x DW GPUs or 16 x SW GPUs</td>
<td>Spec Sheet</td>
</tr>
</tbody>
</table>

GPUs, FPGAs and IPUs for Dell EMC PowerEdge servers

Turbo-charge your applications with performance accelerators available in select Dell EMC 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 A100, A30 and A40 GPUs; and, AMD MI100, Xilinx Alveo™ U200 and Intel® Programmable Acceleration Cards (PAC) D5005 FPGAs with Stratix® 10 SX. Single-wide (SW) accelerators take up one PCIe slot and include: NVIDIA T4 GPUs, and Intel PAC Arria® 10 GX. Dell EMC 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 CUDA cores, amount of memory, and power and cooling requirements. For example, the NVIDIA Ampere® A100 has up to 80GB memory, and uses up to 500 watts. Intel and Xilinx FPGA specifications include embedded and off-chip memory capacity ranging from 35 to 64MB using from 10 to 225 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 Ampere and Tensor Core GPUs

NVIDIA 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.

- Ampere core GPU
- NVLink™ Fabric interconnect
- GPU CLOUD™ containers
- Software application catalog and developer resources
- NVIDIA AI Enterprise for VMware

<table>
<thead>
<tr>
<th>Model</th>
<th>Workloads</th>
<th>Memory</th>
<th>Graphic Bus/ System interface</th>
<th>Slot width</th>
<th>Max Power Consumption</th>
<th>Server support</th>
</tr>
</thead>
<tbody>
<tr>
<td>A100</td>
<td>HPC/AI/Database Analytics</td>
<td>40 / 80 GB</td>
<td>PCIe Gen4x16/ NVLink bridge</td>
<td>Double-wide</td>
<td>300W (80GB) 250W (40GB)</td>
<td>40GB: R750xa, R750, R7525, XR12, DSS8440, R940xa, R740, R740xd 80GB: R750xa, R750, R7525</td>
</tr>
<tr>
<td>A100</td>
<td>HPC/AI/Database Analytics</td>
<td>40 / 80 GB</td>
<td>NVLink bridge</td>
<td>N/A</td>
<td>500W (80GB) 400W (40GB)</td>
<td>XE8545</td>
</tr>
<tr>
<td>A30</td>
<td>Mainstream AI</td>
<td>24 GB</td>
<td>PCIe Gen4x16/ NVLink bridge</td>
<td>Double-wide</td>
<td>165W</td>
<td>R750xa, R750, R7525, R740, R740xd</td>
</tr>
<tr>
<td>A40</td>
<td>Performance graphics/VDI</td>
<td>48 GB</td>
<td>PCIe Gen4x16/ NVLink bridge</td>
<td>Double-wide</td>
<td>300W</td>
<td>R750xa, R750, R7525, XR12, DSS8440, R740, R740xd</td>
</tr>
<tr>
<td>V100S</td>
<td>HPC/AI/Database Analytics</td>
<td>32 GB</td>
<td>PCIe Gen3x16</td>
<td>Double-wide</td>
<td>250W</td>
<td>R7525, R7515, C4140, DSS8440, R940xa, R840, R740, R740xd, T640, XE2420, XE7100</td>
</tr>
<tr>
<td>T4</td>
<td>Inferencing/ Edge/VDI</td>
<td>16 GB</td>
<td>PCIe Gen3x16</td>
<td>Single-wide</td>
<td>70W</td>
<td>R750xa, R750, R7525, R7515, R650, C6520, R6525, R6515, C6525, XR12, XR11, DSS8440, R740, R740xd, R640, XR2, XE2420, XE7100</td>
</tr>
</tbody>
</table>

NVIDIA-Certified Dell EMC 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 MI100 delivers industry best single-precision (FP32) performance. MI100 accelerates HPC, AI workloads, and reduces the overall cost of ownership. Now available on Dell EMC PowerEdge R750xa and PowerEdge R7525 servers.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Workloads</th>
<th>Memory</th>
<th>Graphic Bus/System interface</th>
<th>Slot width</th>
<th>Max Power Consumption</th>
<th>Server support</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI100</td>
<td>HPC/Machine learning training</td>
<td>32 GB</td>
<td>PCIe Gen4x16</td>
<td>Double-wide</td>
<td>300W</td>
<td>R750xa, R7525</td>
</tr>
</tbody>
</table>

### Intel FPGAs

Intel FPGAs can be dynamically reprogrammed with a data path that exactly matches your workloads, such as data analytics, image inference, encryption and compression.

- Intel FPGA Programmable Acceleration Card (PAC) D5005 with Intel Stratix 10 SX
- Intel PAC with Intel Arria 10 GX FPGA
- Software acceleration stack
- Intel FPGA Acceleration Hub

Available in Dell EMC PowerEdge servers including the R640, R740/xd, R840, R940xa and in HPC and AI solutions. Download the configuration guide.

**1.5 TFLOPs**

(tera floating-point operations per second) and power efficiency up to 40 GFLOPs/watt

### Xilinx FPGAs

Built on the Xilinx 16nm UltraScale™ architecture, Xilinx Alveo accelerator cards are adaptable to changing acceleration requirements and algorithm standards, capable of accelerating any workload without changing hardware, and reduce total cost of ownership.

- Alveo U200
- Software ecosystem
- Developer tools

Available in Dell EMC PowerEdge servers including the R7425, R7515, R740/xd, R7525, R840, R940xa and in HPC and AI solutions.

**up to 90x**

higher performance than CPUs for machine learning inference, video transcoding, and database search and analytics

---

© 2021 Dell Inc. or its subsidiaries.
Dell EMC 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 EMC 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

Dell EMC Ready Solutions for HPC make adopting advanced computing faster and simpler. Dell EMC 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 EMC 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.

The BittWare Stratix® 10-based FPGA Accelerated Compute Node allows you to run the most demanding data center workloads using a high-density rackmount server optimized for FPGA accelerators — up to four Intel Stratix 10 FPGAs per 1U. Systems can be purchased directly from BittWare or Dell Technologies. Watch the video.

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.

© 2021 Dell Inc. or its subsidiaries.
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.

Kinecta® 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 EMC 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

© 2021 Dell Inc. or its subsidiaries. All Rights Reserved. Dell, EMC and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be trademarks of their respective owners.