Server accelerators

Turbo-charge your applications
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.

7 in 10 execs ranked AI in the top 3 most significant technologies over the next 5-10 years

up to 100x CPUs in a single GPU

>600 applications accelerated and all deep learning frameworks

---

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.

The PowerEdge C4140 server with SXM2 modules can leverage NVLINK™ high-speed interconnect between GPUs.

FPGAs
Field programmable gate arrays (FPGAs) are accelerators with specific technical characteristics for executing certain types of algorithms up to 1,000X faster than traditional software solutions. FPGAs can be configured to precisely match the requirements of specific tasks or applications.

Dell offers a range of FPGAs as PCIe cards that fit into server PCIe slots.

IPUs
Graphcore’s intelligence processing unit (IPU) emphasizes graph computing with massively parallel, low-precision floating-point computing. It has more than 1,000 processors which communicate with each other to share the complex workload required for machine learning.

Dell offers the Graphcore IPU in PCIe slots inside the DSS 8440 server.

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.

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.

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 P40, M10 and V100/S GPUs; and, Xilinx® Alveo™ U200 and Intel® Programmable Acceleration Cards (PAC) D5005 FPGAs with Stratix® 10 SX. Graphcore Intelligence Processing Units (IPUs) are also double-wide. Single-wide (SW) accelerators take up one PCIe slot and include: NVIDIA P4 and 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 T4 GPU has 2560 CUDA cores, 16GB memory, and uses up to 70 watts. The NVIDIA Tesla® V100 has 5120 CUDA cores, up to 32GB memory, and uses up to 300 watts. Intel and Xilinx FPGA specifications include embedded and off-chip memory capacity ranging from 35 to 64MB using from 10 to 225 watts. Graphcore® IPU specifications include 1216 IPU-Tiles™ with independent IPU-Core™ and tightly coupled in-processor-Memory™.

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.

The Poplar® software stack is world’s very first graph tool chain specifically designed for machine intelligence with IPUs. It integrates with development tools and frameworks, includes libraries and building blocks.
NVIDIA Tensor Core GPUs

NVIDIA Tensor 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).

- T4 Tensor Core GPU
- V100/S Tensor Core GPU
- NVLink™ Fabric interconnect
- GPU CLOUD™ containers
- Software application catalog and developer resources

NVIDIA GPUs are available in Dell EMC PowerEdge servers including VRTX, T440, T640, R640, R6515, R740, R740xd, R7525, R7425, R7515, R6525, C6525, R840, R940xa, C4140, DSS 8440 and in Dell EMC Ready Solutions for HPC and AI.

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.

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.

Graphcore IPUs

Intelligence Processing Units (IPUs) are designed for arithmetic efficiency on small batch sizes for both training and inference. The result is faster model convergence in training, models that generalize better. IPUs also have the ability to parallelize over many more IPU processors to reduce training time for a given batch size while delivering high throughput at lower latency for inference.

- Graphcore® IPU
- Poplar® software stack

Available in DSS 8440 servers.
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.

<table>
<thead>
<tr>
<th>NVIDIA GPUs</th>
<th>INTEL FPGAs</th>
<th>XILINX FPGA</th>
<th>GRAPHCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>V100S</td>
<td>V100 32GB</td>
<td>V100 16GB</td>
<td>M10</td>
</tr>
<tr>
<td>DW PCIe</td>
<td>SW PCIe</td>
<td>DW PCIe</td>
<td>SW PCIe</td>
</tr>
<tr>
<td>C4140</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>C6525</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R940XA</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>R840</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>R740/XD</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>R7515</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R7425</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>R6525</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R6525</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R640</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T640</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>R440</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VRTX</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DW = DOUBLE WIDE. SW = SINGLE WIDE. DSS8440 AND C4140 ARE SET CONFIGS. R7515 WITH V100S AND XILINX U200 FQ2FY21.

Accelerated Dell Technologies Solutions

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

Dell EMC Ready Solutions

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.

4 "NVIDIA Tesla V100 Tensor Core GPU," May 2019.
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.

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.

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 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.

Contact us
To learn more, visit delltechnologies.com/accelerators, contact your local representative or authorized reseller.