Cray® CS-Storm™ Accelerated GPU Cluster Supercomputer

Purpose-built for the most demanding machine and deep learning workloads, Cray CS-Storm systems provide customers with a powerful, GPU accelerator-optimized solution for moving AI applications from pilot to production.
The Rise of Accelerated GPU Computing

A “perfect storm” is brewing. As data volumes have grown exponentially, new use cases leveraging machine and deep learning are coming to the forefront. Whether it’s scientific research or new devices and services, artificial intelligence (AI) and the application of AI has re-emerged as a key transformation technology. For traditional engineering applications or emerging applications like deep learning neural network training, accelerated computing is rapidly gaining mainstream acceptance in places where high performance computing (HPC) has traditionally been handled by CPUs.

But adding a modern GPU is not as simple as just plugging it in. Factors such as GPU-to-GPU communications, system scaling, file system I/O, power consumption and cooling can rapidly pose daunting challenges for organizations looking to move accelerated computing from pilot use to mainstream production.

The purpose-built Cray CS-Storm supercomputer adds to the industry’s broadest range of integrated systems, ready to tackle problems that benefit from the processing power and scale available from today’s GPU and FPGA accelerators.

Typical applications include deep learning, machine learning, signal processing, reservoir simulation, geospatial intelligence, portfolio and trading algorithm optimization, pattern recognition and in-memory databases. For these classes of applications, the benefits of accelerated computing are immense. A single high-density rack dedicated to GPU computation can deliver up to 658 TF of double-precision performance. For machine learning, where integer operations matter, a single CS-Storm 500GX server node can deliver up to 170 TOPS (tera operations per second). The Cray CS-Storm system is designed specifically for these demanding applications and delivers the compute power to quickly and efficiently convert large streams of raw data into actionable information.

Cray CS-Storm Server Node Configurations

At the heart of the Cray CS-Storm accelerated GPU supercomputer are two unique server nodes, each designed to address a range of machine learning and extreme HPC application requirements.

Cray CS-Storm 500GT System

The Cray CS-Storm 500GT configuration scales up to NVIDIA® Tesla® Pascal P40 or P100 GPUs or Nallatech FPGAs by leveraging the flexibility and economics of PCI Express with both single-root and balanced configuration options.

Cray CS-Storm 500NX System

The Cray CS-Storm 500NX configuration scales up to eight NVIDIA Tesla Pascal P100 SXM2 GPUs using NVIDIA® NVLink™ to reduce latency and increase bandwidth between GPU-to-GPU communications, enabling larger models and faster results for AI and deep learning neural network training.

Customizable Performance

The CS-Storm supercomputer is a highly configurable cluster, able to support options specific to your individual requirements including a range of NVIDIA Tesla accelerators tested to run at full power, multiple network and topology options for InfiniBand™ (including fat tree, single/dual rail) and Intel® Omni-Path, a customizable HPC cluster software stack, and Cray® Sonexion® scale-out parallel file system storage.

CS-Storm Accelerated Cluster Supercomputer

Designed for Speed

Powered by NVIDIA, Intel and Nallatech accelerators and processors, each CS-Storm server node and rack system is integrated by Cray to deliver maximum performance across the broadest range of machine learning and deep learning environments.

Designed for Scale

As machine learning and deep learning become core requirements for business and scientific discovery, the need for consistent and timely processing is driving the use of supercomputer-scale systems, which are designed to handle the largest data sets and perform the most complex calculations. Cray has the depth of experience to guide organizations on the journey from AI trials and pilots to production workloads.

Designed for Simplicity

Designing, implementing and using an AI system doesn’t have to be difficult. You can rely on Cray, the expert in production supercomputing, to simplify your environment. Our systems are built on open, industry-standard technologies. We deliver a complete solution for AI including high-performance storage, system management, developer tools and validated deep learning frameworks.
**Cray Cluster Software Stack**

The Cray CS-Storm accelerated cluster system provides a high-performance software and hardware infrastructure running parallel tasks with maximum efficiency. The system provides optimized I/O connectivity and flexible user login capabilities. It is available with a comprehensive HPC software stack compatible with open-source and commercial compilers, debuggers, schedulers and libraries, such as the optional Cray Programming Environment on Cluster Systems (Cray PE on CS), which includes the Cray Compiler Environment, Cray Scientific and Math Libraries, and performance measurement and analysis tools. Cray LibSci_ACC provides accelerated BLAS and LAPACK routines that generate and execute auto-tuned kernels on GPUs.

**Systems Management**

The CS-Storm system can be easily managed. Bright Cluster Manager™ for HPC provides network, server and cluster capabilities with easy system administration and maintenance for scale-out environments.

Bright Cluster Manager for HPC lets customers deploy complete clusters over bare metal and manage them effectively. It provides single-pane-of-glass management for the hardware, the operating system, the HPC software and users. System administrators can quickly get clusters up and running and keep them running reliably throughout their lifecycle — all with the ease and elegance of a full-featured, enterprise-grade cluster manager.

Additionally, the NVIDIA Data Center GPU Manager (DCGM) can be used in conjunction with the Cray CS-Storm accelerated cluster to manage GPU activity on the CS-Storm system. Designed to work with NVIDIA GPUs, DCGM allows for real-time GPU resource monitoring, system policy implementation and system event diagnostics. It includes the capability to control application power efficiency with dynamic power capping on GPUs, and synchronous clock management to keep GPUs across the system in sync based on target workload and power budgets.

---

**Global Leader in Supercomputing**

As the global leader in supercomputing, Cray's ability to turn pioneering hardware and software technologies into world-renowned supercomputing solutions is the result of decades of dedication to HPC. From technical enterprise- to petaflops-sized solutions, our systems enable tremendous scientific achievement by increasing productivity, reducing risk and decreasing time to solution.

**For More Information**

To learn more about the Cray CS-Storm cluster visit [www.cray.com/cs-storm](http://www.cray.com/cs-storm) or contact your Cray representative.
### Cray® CS-Storm™ Specifications

<table>
<thead>
<tr>
<th></th>
<th>CS-Storm 500GT</th>
<th>CS-Storm 500NX</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Processors</strong></td>
<td>Two Intel® Xeon® Scalable processors</td>
<td>Intel® Xeon® E5-2600 v4 family processors</td>
</tr>
<tr>
<td><strong>Memory Capacity</strong></td>
<td>Up to 2 TB DDR4 (16 x 128 GB DIMMs)</td>
<td>Up to 3 TB DDR4 (24 x 128 GB DIMMs)</td>
</tr>
<tr>
<td><strong>Accelerators</strong></td>
<td>NVIDIA® Tesla® P40 or P100 PCIe GPU accelerators or Nallatech FPGA accelerators. Supports up to 10 400W parts or 8 450W parts.</td>
<td>Up to 8 NVIDIA Tesla P100 SXM2 GPU accelerators; supports up to 300W parts.</td>
</tr>
<tr>
<td><strong>Drive Bays</strong></td>
<td>Multiple local storage configuration options. 12 hot-swappable 2.5&quot; drives (up to 4 x NVMe). Some configurations require an additional add-in storage controller; total number of drives varies by configuration.</td>
<td>16 hot-swappable 2.5&quot; drives (up to 8x NVMe)</td>
</tr>
<tr>
<td><strong>Expansion Slots</strong></td>
<td>12 PCIe 3.0 x 16 slots supporting multiple PCIe topologies and configuration options</td>
<td>4 x16, low-profile (GPU tray), 2 x 8 (motherboard tray) PCIe 3.0</td>
</tr>
<tr>
<td><strong>Power Supply</strong></td>
<td>Four 2200W AC power supplies; N+1 and N+N (limited to specific configurations) redundancy</td>
<td>Four 2200W AC power supplies; 2+2 redundancy; titanium-level efficiency</td>
</tr>
<tr>
<td><strong>Power Input</strong></td>
<td>200-277VAC, 10A max</td>
<td>200-277VAC, 10A max</td>
</tr>
<tr>
<td><strong>Weight</strong></td>
<td>Up to 76 lbs. (without PCIe cards)</td>
<td>Up to 135 lbs.</td>
</tr>
<tr>
<td><strong>Dimensions</strong></td>
<td>5.25&quot; H x 17.6&quot; W x 36.4&quot;D (3U)</td>
<td>7.0&quot;H x 17.6&quot;W x 29&quot;D (4U)</td>
</tr>
<tr>
<td><strong>Temperature</strong></td>
<td>Operating: 10°C–35°C, ASHRAE 2</td>
<td>Operating: 10°C–35°C, ASHRAE 2</td>
</tr>
</tbody>
</table>