CRAY ACCEL AI DEEP LEARNING

Artificial intelligence pioneers are under pressure to succeed. They've committed to AI and deep learning, but the underlying infrastructure — systems used by artificial intelligence teams to complete the entire AI workflow — is complex and evolving. If your toe isn't already in the AI/DL water, it might feel like you're already too late. You're not — because Cray can help. How? By giving you the tools and support to guide you to success, whether you're just starting out or ready for production AI.

Everything You Need to Go From Exploration to Production

Cray Accel AI turns our supercomputing expertise, technologies and best practices into solutions that advance the adoption of deep learning.

These fast-start reference configurations range from a starter system ideal for AI exploration to complete, production-level systems for the entire AI workflow, meeting the needs of data science teams that are under pressure to shorten time to value. We make it easy for your IT team to quickly get started with a single system that makes AI a reality.

Cray has a long track record supporting artificial intelligence and other complex computing challenges. The systems that make AI possible today have evolved over decades in other areas such as medical imaging, cybersecurity, climate modeling and seismic processing — systems driven to success by our supercomputing expertise. We've taken what we've learned in other high-performance computing domains and applied it to AI.

So wherever you are on your AI journey, whether you're just starting out or ready to deploy AI applications in production, we have the high-powered computing systems and industry expertise to make artificial intelligence work for you.

Test, Launch and Grow Your Deep Learning Initiatives

Cray Accel AI reference configurations are for artificial intelligence teams either starting out with AI — exploring the basics of deep learning with GPUs — or developing entire AI workflows where data preparation, model development and model implementation require a mix of CPU and GPU computing capabilities and tools.

They include Cray® CS-Storm™ GPU-accelerated supercomputers featuring NVIDIA® Tesla® V100 "Volta" GPUs and the NVIDIA® NVLink™ GPU-to-GPU interconnect, and Cray® CS500™ CPU systems featuring Intel® Xeon® Scalable processors and Cray® ClusterStor™ HPC storage.

Accel AI Exploration Reference Configuration

If you're at the tool exploration stage — beginning with machine learning and deep learning model exploration — our Accel AI Exploration configuration delivers all the elements small team needs to get started with deep learning trials. You get a single-chassis, fully configured Cray CS-Storm 500NX 8-GPU server with eight NVIDIA Tesla V100 SXM2 GPUs and Cray's Urika-CS AI suite.

Accel AI Prototype Reference Configuration

The Accel AI Prototype reference configuration delivers the technology that small AI teams — usually a handful of data scientists and data engineers — need to bring AI applications from concept to production. In a single rack, you get a complete system for your AI workflow, including Cray CS-Storm 500NX 4-GPU servers and CS500 CPU servers, ClusterStor HPC storage, the Cray® Urika®-CS AI and Analytics suite, and a complete cluster management infrastructure — all selected to meet the needs of data engineers, data scientists and AI engineers tasked with designing and implementing entire AI workflows.

Accel AI Production Reference Configuration

When you're ready for a production-level system, choose the Accel AI Production reference configuration, which is designed for the needs of larger AI teams. The modular Production configuration includes a machine learning partition featuring 32 Cray CS500 nodes that include Intel Xeon Scalable processors to handle tasks associated with the data preparation phase of the AI workflow (and for CPU-based model training when the need arises). The deep learning partition includes four CS-Storm 500NX 8-GPU servers for large-scale distributed deep learning training. Cray ClusterStor HPC storage, featuring flash, supports I/O requirements for data preparation, model development and model implementation tasks. A cluster management and networking infrastructure with high-speed Ethernet® and InfiniBand® completes the configuration.

Complete AI Workflow Environment

The Cray Urika-CS AI and Analytics suite is a comprehensive AI workflow software environment with all the AI frameworks, libraries and tools you need for complex machine learning and deep learning workloads. Featuring the Cray Distributed Training Framework — designed to simplify deep learning model training with TensorFlow — and data preparation tools like Apache Spark™ and Anaconda Python, the Urika-CS suite is a supported environment designed to eliminate the complexity associated with integrating multiple open-source tools on a single system. The Urika-CS AI suite, fully supported by Cray, makes it easy for IT teams to meet the demanding needs of their users.

Where Deep Learning Matters

Cray Accel AI offerings are designed for deep learning applications where models are developed using unstructured data — images, audio, video or text — to make predictions. Across industries, organizations are exploring and implementing AI applications to increase operational efficiency, improve customer experience and develop competitive breakthroughs with new and enhanced business models.

Object Detection in Full Motion Video

Object detection in full motion video is an application of computer vision that uses machines to gain a high-level understanding of digital videos. The goal of object detection is to extract and analyze information from a frame-by-frame video stream analysis. Detecting objects in video is one of the most data- and compute-intensive use cases for deep learning, one in which dense GPU systems and scale allow for timely workflows — completed in hours or days rather than weeks.

Common use cases: full motion video mining; vehicular object detection, identification and avoidance; object detection for surveillance; object tracking; real-time video analytics; content analysis and facial recognition; fault detection and asset performance measurement; enhanced process monitoring and auto-correction; pedestrian, vehicle and object detection and classification; autonomous ride-sharing fleets; and semi-autonomous features such as driver assist.

Image Recognition and Object Detection

Similar to object detection in video, object detection in still images uses elements of computer vision to train machines to extract a high-level understanding from digital images. The goal of image-focused object detection is to automate extraction and classification of useful information from digital images.

Common use cases: static image recognition, classification and tagging; image segmentation; facial recognition; gesture recognition; localization and mapping; medical image analysis; satellite imagery for geoanalytics; object identification in geospatial data; seismic imaging; product image indexing for search optimization; and automated insurance claims processing.

Audio, Text and Natural Language Processing

Natural language processing (NLP) is the ability of a computer to understand human speech as it is spoken. Advances in AI algorithms allow applications to analyze and make use of patterns in data to understand speech and text.

Common use cases: audio mining; text mining and classification; sentiment analysis; emotion AI; speech understanding; speech authentication; text to speech, search; and sensory aid.

High-Dimensional Data Analysis

High-dimensional data can be analyzed using deep learning methods on datasets with large numbers of attributes or features. These models are not easy to work with due to their high dimensionality, which by nature is computationally intensive. With advances in algorithms and compute power, applications for high-dimensional data analysis are used in industries as diverse as life sciences, healthcare, financial services, retail, hospitality, manufacturing and many more.

Common use cases:

- Life sciences – bio-marker discovery and disease research, clustering and phenotype discovery, computational drug discovery, medical diagnosis assistance, enhanced diagnostics, next-generation sequencing, cancer cell morphology, drug discovery, precision medicine
- Healthcare – patient treatment optimization, payer fraud, early identification of potential pandemics
- Financial services – algorithmic trading, risk analysis and mitigation, trading strategies, regulatory compliance, automated trading and stock investment, algorithmic trading, fraud detection, personalized financial planning, insider threat surveillance, anti-money laundering
- Manufacturing – supply chain and production optimization, inventory and delivery management, predictive maintenance

Configuration	Exploration	Prototype	Production
	For tool exploration and early model development	**For small team AI projects**	**A complete system for production-level deep learning**
AI deep learning node(s)	1 x Cray® CS-Storm™ 500NX node with 8 NVIDIA® Tesla® "Volta" V100 SXM2 accelerators with 32 GB GPU memory and 2 Intel® Xeon® Scalable "Skylake" processors	4 x Cray CS-Storm 500NX nodes with 4 NVIDIA Tesla "Volta" V100 SXM2 GPU accelerators with 32 GB GPU memory and 2 Intel Xeon Scalable "Skylake" processors	4 x Cray CS-Storm 500NX nodes with 8 NVIDIA Tesla "Volta" V100 SXM2 GPU accelerators with 32 GB GPU memory and 2 Intel Xeon Scalable "Skylake" processors
AI machine learning node(s)		32 x Cray® CS500™ 2829XT nodes with 18-core Intel Xeon Scalable processors and 192 GB memory	64 x Cray CS500 2829XT nodes with 18-core Intel Xeon Scalable processors and 192 GB memory
Management node(s)		Cray 2828X 2U server with dual Intel Xeon processors	2 x Cray 2828X 2U servers with dual Intel Xeon processors
Storage	4 x 1.92 TB 2.5" SATA SSDs	Cray® ClusterStor™ scalable HPC storage for data preparation and model development, including 35 TB of high-performance L300F scalable flash storage and 640 TB of L300N hybrid (SSD/HDD) storage NetApp™ E-Series E2800 storage for general-purpose use	Cray ClusterStor scalable HPC storage for data preparation and model development, including 70 TB of high-performance L300F scalable flash storage and 640 TB of L300N hybrid (SSD/HDD) storage NetApp™ E-Series E2800 storage for general purpose use
Networking	100 Gb/s InfiniBand® EDR and 10 Gb/s Ethernet®	100 Gb/s InfiniBand EDR and 10 Gb/s Ethernet	100 Gb/s InfiniBand EDR and 10 Gb/s Ethernet
Usability Software	Jupyter Notebook, TensorBoard		
AI Frameworks for Model Development and Training	TensorFlow™, Cray Distributed Training Framework, BigDL		
Data Preparation Tools	Apache Spark™, Anaconda Python		