Case Study

High Performance Computing
3rd and 2nd Generation Intel® Xeon® Scalable Processors
Habana Labs GAUDI® and GOYA AI Processors

Voyager AI Supercomputer—Habana Labs' First On-Prem Training Deployment—Enables New Explorations

The AI-focused system at San Diego Supercomputer Center will allow scientists to develop new approaches for accelerated training and inferencing

Solution Summary:

- 42 Supermicro X12 Habana GAUDI® AI Training System nodes with 3rd Generation Intel® Xeon® Scalable processors
- Total of 336 Habana GAUDI HL-205 AI training processors with ten integrated 100 GbE RoCE ports per chip
- Two Supermicro SuperServer GOYA inference nodes with 2nd Gen Intel Xeon Scalable processors
- Total of 16 Habana GOYA HL-100 PCIe AI inference processors

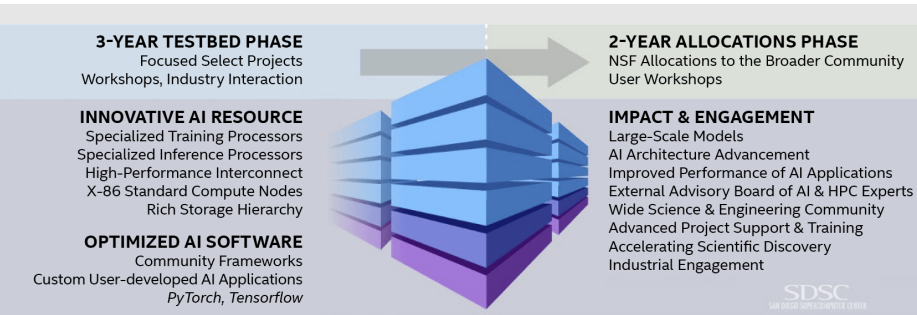
Executive Summary

Scientists around the world are inundated with petabytes and exabytes of data, which have the potential to unlock insights and discoveries that lead to breakthroughs in health, physics, and many other sciences. To sift through and make sense of these mountains of data, researchers are turning to Artificial Intelligence (AI), deep learning (DL), and machine learning (ML). An AI-focused supercomputer called Voyager at the San Diego Supercomputer Center (SDSC) will help scientists around the world discover and develop new methods for applying AI to their fields. Funded by the National Science Foundation (NSF) as an experimental system, Voyager will allow scientists to try new approaches and design, build, and optimize algorithms for accelerated machine learning model training and inference.

Challenge

The availability of more data across all scientific domains, driven by new acquisition sources, higher resolution models, and federated repositories, is resulting in massive and increasingly complex datasets. To sort and synthesize this data, scientists are applying AI, especially DL techniques and algorithms, using new purpose-built computational methods to extract insights.But there is yet a lot to learn about how to best analyze these massive and complex datasets.

GPUs have traditionally been the go-to architecture for large-scale deep learning training workloads, but AI is in a phase where emerging technologies are offering



Key aspects of the experimental Voyager supercomputer funded by the National Science Foundation. Image by Ben Tolo, SDSC/UC San Diego.

alternatives to existing AI methods. As new solutions appear from companies like Habana Labs (an Intel company), researchers need access to these technologies at-scale in order to explore new approaches to address the most pressing research challenges. Scientists need to understand how to train and deploy their models on these new machines and, importantly, be able to share their experiences with the research community.

Voyager at the San Diego Supercomputer Center, located at the University of California San Diego (UC San Diego), offers a platform for AI acceleration research and development.

"We talked to several scientists about what they needed for their research," said Amit Majumdar, Director of the Data Enabled Scientific Computing (DESC) Division at SDSC. "AI is becoming an important component of their research. When the National Science Foundation requested proposals for unique experimental supercomputers, we began architecting Voyager and sought an NSF grant. Voyager is one of the first AI-focused experimental systems to join the NSF ecosystem."

Solution

Voyager is NSF's first AI-focused supercomputer. Over the years, AI workloads have been adapted and designed for GPUs. Unlike these general purpose systems, Habana processors were purpose-designed expressly to drive improved AI compute efficiencies. As a result, Habana accelerators enable dramatic performance improvements for training and inferencing.

Voyager comprises 42 training nodes of Supermicro X12 GAUDI® Training Servers powered by dual-socket 3rd Gen Intel® Xeon® Scalable processors. Each training node contains eight GAUDI HL-205 training processor cards. Two Supermicro SuperServer nodes are deployed for inferencing, each with two 2nd Gen Intel Xeon Scalable processors and eight GOYA HL-100 inferencing PCIe cards. The combination of Intel Xeon Scalable processors and Habana training and inferencing processors provides a unique and powerful architecture for accelerated training and inference. Using Voyager, scientists can explore how to best take advantage of these computational devices for particular types of workloads.

Funded by NSF's experimental high-performance computing program, Voyager will enable access to a small group of scientists for the first three years of its 5-year operations. Several disciplines will be represented, including high-energy physics, biology, genetics, materials science, atmospheric and astronomic sciences, plus others.

"AI is becoming a discipline itself," added Majumdar. "Unlike general purpose computing, tools and technologies focused on deep learning are different. This is hardware specially built for AI—Gaudi for training and Goya for inference. We need this hardware to experiment, test, and learn in order to advance AI approaches. We will be building and testing algorithms, optimizing them, and contributing to the community what we learn from Voyager and its technologies."

After three years, the new system will be open to the entire NSF scientific community for accelerated AI projects via an open allocation process.

GAUDI and GOYA—accelerating AI

Gaudi is an AI training processor designed from the ground up for performance and scalability. Gaudi accelerates key AI training workloads, including image classification, object detection, natural language processing, text to speech, sentiment analysis, recommender systems, and others.

With eight fully programmable Tensor processor cores (TPC), Gaudi presents a high-performance, cost-efficient option for deep learning training.

AI training often consumes a very large number of nodes for extended periods. At such a scale, Gaudi accelerators provide AI compute and operational efficiencies for training DL models whether on-prem or in the cloud. Amazon Web Services added them to its EC2 instances to offer lower cost alternatives to their GPU instances. Amazon Web Services expects their Gaudi processor-enabled instances to deliver up to 40 percent better price- performance over their current generation of GPU offerings.[1]

The growing size and complexity of data sets, neural networks and AI workload—combined with the increased demand for AI accuracy—is resulting in training systems requiring massive amounts of compute capability. To meet this need, systems must provide high-performance scaling with efficient communications across nodes.

> *"We're working with scientists who will run both training and inference. Some will migrate their workloads built on other technologies to Voyager; others will develop their models directly on Voyager. They will also need to transfer their models from training to inference, so it's good to have both in one system."*
>
> –Amit Majumdar, Director of the Data Enabled Scientific Computing (DESC) Division at SDSC.

To address this demand for a high-performance, flexible and scalable AI system, every Gaudi processor integrates ten 100 gigabit RDMA over Converged Ethernet (RoCE) ports. In Voyager, each of the eight Gaudi cards contained within the servers dedicates seven 100 Gb ports to connect in an all-to-all non-blocking configuration to the other cards. The other three 100 Gigabit ports in each Gaudi are dedicated to scale out, giving each Voyager node 24 100 Gigabit ports, for highly flexible, massive scalability.

Built on Habana's TPC architecture with eight programmable cores in each inference card, Goya accelerates AI inference workloads, irrespective of the architecture on which they were trained. It natively supports many mixed-precision data types, including FP32, INT32/16/8, and UINT32/16/8.

"We're working with scientists who will run both training and inference," added Majumdar. "Some will migrate their workloads built on other technologies to Voyager; others will develop their models directly on Voyager. They will also need to transfer their models from training to inference, so it's good to have both in one system."

The AI training and inference processors run with 3rd Gen Intel Xeon Scalable host processors and 2nd Gen Intel Xeon Scalable host processors, respectively. The next-gen datacenter CPUs are designed around a balanced

architecture, with built-in acceleration and advanced security capabilities. They are optimized for many workload types and performance levels, all with the consistent, open Intel architecture scientists have trusted for decades. 3rd Gen Intel Xeon Scalable processors are built for cloud, enterprise, HPC, network, security, and IoT workloads. They provide eight to 40 powerful cores and a wide range of frequency, feature, and power levels. And they are the only data center CPU with built-in AI and HPC acceleration. Plus, they offer end-to-end data science tools, and an ecosystem of smart solutions.

Model Migration Made Easy

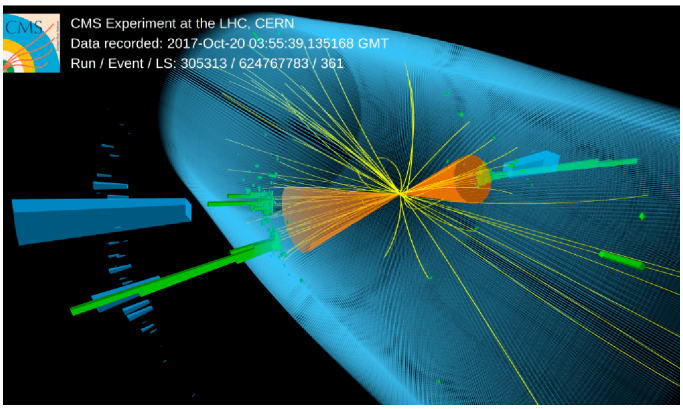
In addition to Gaudi and Goya, Habana's software suite is designed for ease of development—to build new models or simplify the migration from existing GPU-models to Habana. Habana's SynapseAI software suite for training and inference includes a graph compiler and runtime communication libraries, TPC kernel library, firmware, drivers, and data center deployment and management tools. The software suite is integrated with TensorFlow and PyTorch Frameworks, and is performance-optimized for Gaudi training and Goya inference training.

Results

For Voyager's first three years, a small group of scientists will work with the system, in close collaboration with SDSC and Habana application experts. Some will migrate existing workloads and others will develop new ones specifically on Voyager. After three years of research, experimentation, development, and sharing their findings with their communities, Voyager will be available for general scientific research. Two of the scientists using Voyager following its deployment are Javier Duarte and Mai Nguyen, both from UC San Diego.

Crashing Protons

Javier Duarte is an Assistant Professor of Physics. He is also a member of the Compact Muon Solenoid (CMS) Collaboration, a project that involves over 3,000 physicists and engineers around the world. The CMS experiment is run on the Large Hadron Collider (LHC) at CERN in Switzerland. Protons are accelerated in the LHC to nearly the speed of light and smashed into each other in the middle of the CMS detector, which measures the outgoing particles. CMS researchers want to explore what the collisions produce to help them better understand the laws of nature.



A Higgs boson decays into a collimated bottom quark-antiquark pair, which are reconstructed as a single large-radius jet with two-prong substructure, represented by the orange cone on the left part of the display. The Higgs boson is recoiling against a jet, represented by the orange cone on the right side of the display. The electromagnetic and hadronic contributions are represented by green and blue boxes, respectively. © 2020-2021 CERN

"There are many aspects of the universe we don't yet understand and have not been accounted for in our theories," Duarte explained. "Will these collisions produce new particles we don't know about? Will they help us understand what dark matter is? How will these subatomic particles interact with each other at such high energies?"

The CMS experiment collects about 50 petabytes a year from collisions. That's more data than a human can parse for meaningful investigations. Most collisions are background with only very rare events constituting important data. Duarte's job involves using machine learning algorithms to better detect patterns in the data compared to what a person could do to 'disentangle' these very rare signals from the background.

"We currently discard about 99.98 percent of the data we generate, simply because we don't have the computational capacity to analyze it all," he added. "Collisions happen at about 40 MHz, but we can only save data, after some upgrades, at a rate of about 7.5 kHz. That's a huge reduction in data. So, we look for certain patterns that trigger data collection. And my work involves improving our triggering algorithms so we can better select our data, with the goal of doing it in real-time."

Duarte also uses graph neural networks to help reconstruct the events from the data that was captured. According to Duarte, the datasets for the graphs are massive and require highly scalable machine learning, difficult to do efficiently with GPUs and their limited memory capacity. He expects that Voyager's large memory capacity per node and memory access pattern will allow his work to scale more efficiently on Voyager to better train these kinds of networks.

Fighting Fire with Better Knowledge

Mai Nguyen is the Lead for Data Analytics at SDSC
She applies deep learning techniques to a range of interdisciplinary problems, including image analysis, disaster management, and natural language processing, among others.

One of her projects for Voyager will be running deep learning algorithms on satellite images, and possibly images from aircraft, to determine land covers across different areas.

"The goal is to understand the land cover composition of an area in the context of wildland fire management," stated Nguyen. "Fire behavior depends on a lot of environmental conditions, and the fuel available to it is important. Grass creates one kind of fire, while dead wood presents another, potentially more intense fire."

"There is a lot of interest in using AI for wildfire management," she added. "UC San Diego and the University of California system as a whole are involved in this line of research. At SDSC, the WIFIRE Lab, led by Ilkay Altintas, Chief Data Science Officer at SDSC, offers a knowledge cyberinfrastructure for monitoring and responding to hazards such as wildfires. Our work is to combine AI

Combining AI techniques with fire science models and expertise will help scientists better understand how to mitigate wildfires.

techniques with fire science models and fire science expertise, and then put together a platform that integrates all these different technologies for researchers to go to. They can study and simulate fire behavior under certain conditions to better understand how to mitigate wildfires."

For her satellite image analysis work, Nguyen uses a processing pipeline she and her WIFIRE colleagues have applied to several different applications. These projects include understanding the composition of a city, how quickly a refugee camp can build up, and detecting where schools are in rural parts of Africa.

She has developed her algorithms on the TensorFlow framework for running on GPU-based systems. Voyager offers her an alternative architecture to test and develop her work with an easy tr;ansition to the new accelerators.

"For the most part, it is straightforward to adapt my deep learning code to run on the Habana system," she concluded. "Easy migration is important for adoption and ease of use for researchers like me."

Solution Summary

SDSC was awarded the NSF grant to build a unique, AI-focused supercomputer. Voyager is built with Supermicro servers using 2nd and 3rd Gen Intel Xeon processors and Habana Gaudi training and Goya inference AI processors. The new system will allow researchers to migrate their existing machine learning and inference projects to Voyager, plus design, develop, and optimize new algorithms. Voyager is expected to enter production operation in early 2022, where it will operate for five years on behalf of the national research community.

Where to Get More Information

Learn more about Habana Labs and their Gaudi and Goya accelerators.

Find out more about SDSC and Voyager.

Explore the capabilities of the 3rd Generation Intel Xeon Scalable processors with integrated Intel Deep Learning Boost capabilities for accelerated AI inferencing.

Solution Ingredients

- 42 Supermicro X12 Habana GAUDI AI Training System nodes with 3rd Gen Intel Xeon processors
- Total of 336 Habana GAUDI HL-250 accelerators with ten integrated 100 GbE RoCE ports/chip
- Two Supermicro SuperServer 4029GP-T inference nodes with 2nd Gen Intel Xeon processors
- Total of 16 Habana GOYA HL-100 PCIe inference accelerators.

intel

1 https://aws.amazon.com/ec2/instance-types/habana-gaudi/

1021/RJMJ/PA/PDF ♻ Please Recycle 348530-001US