cerebras

CASE STUDY

Lawrence Livermore National Laboratory

Customer	Industry	Application
Lawrence Livermore National Laboratory	Scientific Computing	High-throughput DNN inference for nuclear fusion simulation experiments

The Opportunity

Lawrence Livermore National Laboratory in Livermore, California, is a federal research facility primarily funded by the US Department of Energy's National Nuclear Security Administration (NNSA). LLNL's mission is to strengthen the United States' security by developing and applying world-class science, technology and engineering.

LLNL is home to the National Ignition Facility (NIF) which conducts nuclear fusion research using the world's most powerful laser. This is big science. However, their inertial confinement experiments are very expensive and time consuming. In order to do more science, more efficiently, LLNL runs simulated experiments using a multi-physics software package called HYDRA on the Lassen supercomputer. Real-world data from NIF is used to validate and fine-tune the HYDRA models. This allows the models to more accurately predict the outcome of real-world experiments which, in a virtuous circle, can guide the design of real-world experiments.

The part of HYDRA that models atomic kinetics and radiation is called CRETIN. This module predicts how an atom will behave under the conditions pertaining at that point in the simulation. CRETIN can represent tens of percent of the total compute load for HYDRA[1]. LNLL's researchers have shown that they can greatly reduce computational intensity by replacing CRETIN with a deep neural network model (DNN), dubbed CRETIN-surrogate, which was trained using data from CRETIN. This inference is performed in every time step of the simulation.

This work is part of a broader initiative at LLNL that aims to blend traditional high-performance computing (HPC) simulation

> Pairing the AI power of the CS-1 with the precision simulation of Lassen creates a CogSim computer that kicks open new doors for inertial confinement fusion (ICF) experiments at the National Ignition Facility.
>
> **Brian Spears**
> **Principal Investigator, LLNL**

The National Ignition Facility (Photo courtesy of Lawrence Livermore National Laboratory)

Installing the Cerebras system (Photo courtesy of Lawrence Livermore National Laboratory)

Further reading

Keynote from 2021 AI Systems Summit by LLNL CTO Bronis de Supinski: "Heterogeneous System Architectures Effective Use of Diverse Components" https://vimeo.com/531218771/b7d9b279a2

More information on Lassen: https://hpc.llnl.gov/hardware/platforms/lassen

and modelling workloads with artificial intelligence to create "cognitive simulation", or CogSim, for short.

The Challenge

Lassen is an extremely powerful supercomputer. It has 792 IBM Power9 compute nodes, each equipped with 4 NVIDIA® Volta graphics processing units (GPUs) for a grand total of 3,168 GPUs, spread across 44 racks.

So why not just use that horde of GPUs to run the DNN? The problem is one of load-balancing. The goal was to improve overall system performance by offloading the computationally-intensive inference to a separate device.

A significant trend in HPC over the last few years is to incorporate machine learning (ML) elements to guide and improve traditional simulation and modelling codes which are parallelized to run across hundreds or thousands of identical compute nodes. There is a growing realization that the traditional scale-out approach of thousands of identical compute nodes loaded up with GPUs or other accelerators ("node-level heterogeneity") has limitations. In addition to the difficulty of sharding large models across many nodes, there's the practical issue of overprovisioning.

In systems designed for flexibility and multi-tenancy (also a trend away from the old HPC world where systems were purpose-built to run one workload), you don't want stranded resources that are inaccessible to the rest of the system or sit idle much of the time.

The solution is to add accelerators as complete, independent compute nodes. Bronis de Supinski, CTO of LLNL's Livermore Computing Center, describes this architecture as "system-level heterogeneity". LLNL has built novel system software to steer workloads, automatically, towards the optimum node type. LLNL is a leader in this field. They have implemented workload steering in a way that is transparent to the researcher, allowing them to focus on their work, rather than thinking about how to tune their codes to take into account available system resources.

The challenge is to leverage ML compute nodes so that the training and inference can fit "in the loop" with time-step HPC codes without slowing the overall system down. This drives the need to achieve terabit-scale bandwidth to the ML compute nodes, and also

to have very powerful special-purpose processors to complete ML tasks in the time available.

The Cerebras Solution

LLNL chose the Cerebras CS-1 system to perform their CRETIN-surrogate inference. A CS-1 system was integrated with the Lassen supercomputer at LLNL using spare InfiniBand ports. de Supinski described the installation as "one of the smoothest sitings ever".

Installation took less than 20 hours from crate to "hello world". Before the system itself arrived, Cerebras technicians installed a "cooling shell", along with the CS-1 system's mechanical support rails and hardware. The cooling shell is an empty CS-1 chassis with a normal external cooling loop interface that we use to verify that the facility's liquid cooling system is functioning correctly. The mechanical installation of the CS-1 system itself and support cluster took about ten hours. Software bring-up was completed in one hour.

In parallel, Cerebras' machine learning software engineers worked alongside their LLNL colleagues to write a C++ API which is used to allow HYDRA code to call the CRETIN-surrogate model. The model uses an autoencoder to compress the input data into lower dimensional representations which are then processed by a predictive model built with a novel deep neural network algorithm called DJINN[2]. DJINN automatically chooses an appropriate neural network architecture for the given data without requiring the user to manually tune the settings.

The Cerebras System with part of the Lassen supercomputer (Photo courtesty of Lawrence Livermore National Laboratory)

The Impact

Early results show that the combination of the Lassen system with the Cerebras accelerator works brilliantly. Plugging the CS-1 system into Lassen's InfiniBand network enables 1.2 terabits-per-second bandwidth to the CS-1 system. The CS-1 system has 19GB of super-fast SRAM memory tightly coupled to 400,000 AI compute cores, so it was possible to run many instances of the relatively compact DNN model in parallel. This combination of bandwidth and horsepower allows HYDRA to perform inference on 18 million samples every second. This is well within the ingest capacity of the CS-1 system for this application. To put that in perspective, that figure is 37x higher than one of Lassen's GPUs can achieve, or 16x faster than an entire compute node equipped with four GPUs.

According to de Supinski, these initial results yield a 5x performance improvement per transistor over GPUs, a metric that he uses as a proxy for cost. Thus, the Cerebras system is allowing LLNL to run experiments that were previously computationally intractable, with simple integration, at much lower cost.

Future Work

The present work to improve the "in the loop" performance of the multi-physics simulation is just the first step in the CogSim program (figure 1).

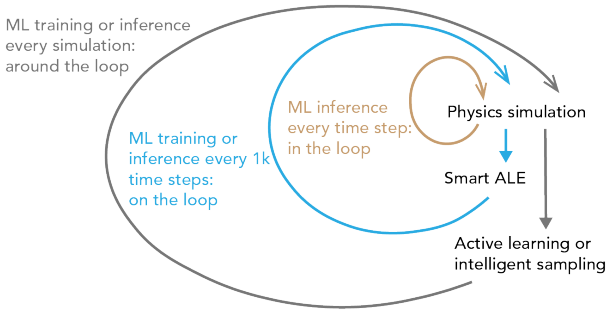


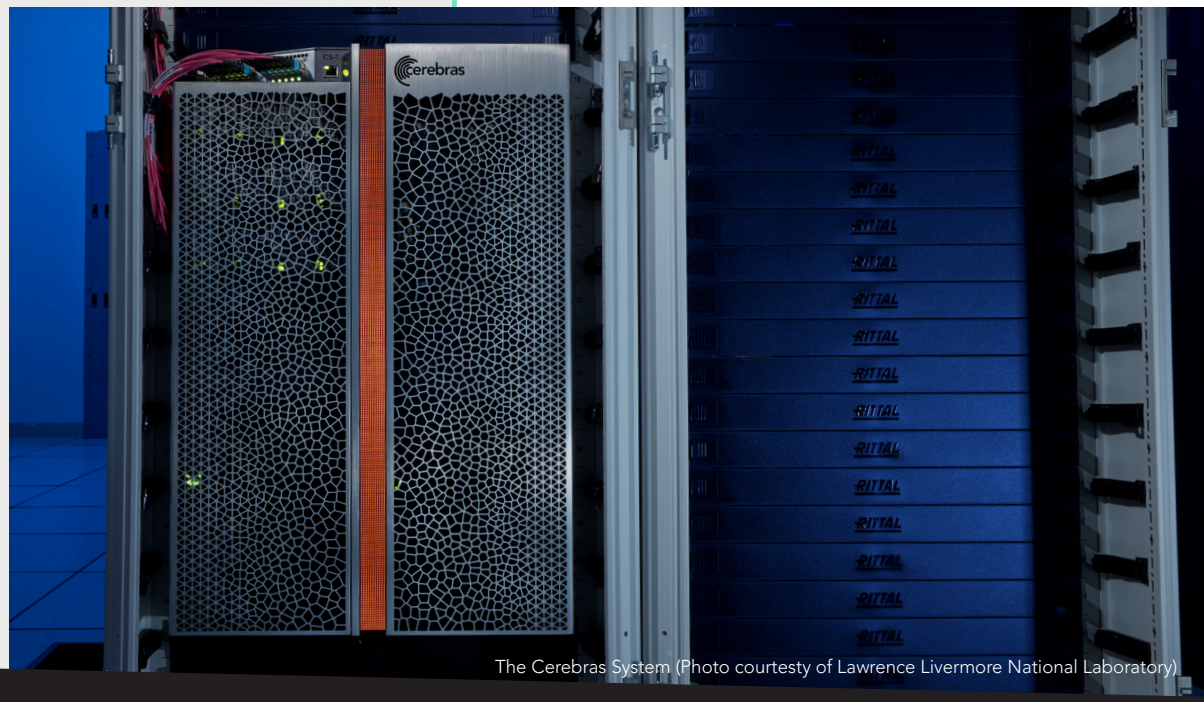
Figure 1. Wrapping simulations in multiple layers of machine learning (Graphic reproduced from LLNL[1])

Key Results

18 million DNN inferences per second

37x performance of Lassen GPU

20 hours from crate to "hello world!"



The Cerebras System (Photo courtesy of Lawrence Livermore National Laboratory)

To learn more or to see a demo, please contact us at cerebras.net/get-demo.

The next layer, called "on the loop" is intended to steer the simulation and provide insight into the simulation while it is still running. This allows researchers to monitor and potentially halt the run if the simulation is not working well. The results from each run become part of the training set for the model, allowing continuous training of that model. We expect the Cerebras system to be able to perform that retraining without significantly impacting overall execution time.

The final layer, "around the loop" uses the characteristics and results of entire simulation runs as a training set to create an "active learning" model which could be used to optimize future runs by picking the parameters and initial boundary conditions for the next experiment.

Conclusion

Cerebras Systems builds compute solutions for the hardest AI problems. Our systems are not only the fastest AI accelerators in existence, but are also easy to install, quick to configure, and enables blisteringly fast model training. Researchers at LLNL have integrated this exceptional compute performance into their Lassen supercomputer to enable meaningful advances in nuclear fusion simulations.

As Brian Spears, Principal Investigator at LLNL put it: "Pairing the AI power of the CS-1 with the precision simulation of Lassen creates a CogSim computer that kicks open new doors for inertial confinement fusion (ICF) experiments at the National Ignition Facility. Now, we can combine billions of simulated images with NIF's amazing X-ray and neutron camera output to build improved predictions of future fusion designs."

Endnotes

1 LLNL News article, "Machine Learning Speeds Up and Enhances Physics Calculations" https://lasers.llnl.gov/news/machine-learning-speeds-up-enhances-physics-calculations

2 Humbird, Peterson and McClarren, "Deep neural network initialization with decision trees", 2017, https://arxiv.org/abs/1707.00784v3

Cerebras Systems is revolutionizing compute for Deep Learning with the CS-2 system, powered by the Wafer Scale Engine. The Wafer Scale Engine delivers more compute, more memory, and more communication bandwidth to enable artificial intelligence research at previously-impossible speeds and scale. Pioneering computer architects, computer scientists, and deep learning researchers have come together to build a new class of computer system that accelerates AI by orders of magnitude beyond the current state of the art.