Argonne National Laboratory Brings Cerebras Computational Tools to the Forefront of Artificial Intelligence Research

Partnership Produces Award-Winning Work, Spanning Cancer Treatment, CoVID-19 Drug Discovery, and Advances in Physics

Vishal Subbiah, Cerebras Systems

Abstract

At Argonne National Laboratory (ANL), researchers work to gain a deeper understanding of our planet, our climate, and the cosmos. When they ran into major challenges associated with scaling large AI models across a cluster of GPUs, ANL contacted Cerebras Systems to explore how we could accelerate their important work – ultimately reducing training time from weeks to hours. Our multi-year partnership with ANL has supported a wide range of successful research projects, as well as multiple Cerebras system purchases and upgrades.

"Cerebras' inventions, which will provide a 100 times increase in parameter capacity, may have the potential to transform the industry. For the first time we will be able to explore brain-sized models, opening up vast new avenues of research and insight."

- Rick Stevens, Associate Laboratory Director of Computing, Environment and Life Sciences, Argonne National Laboratory[1]

Contents

Background

Argonne National Laboratory (ANL) is a U.S. Department of Energy-funded facility. It is a multidisciplinary science and engineering research center, where scientists and engineers work together to answer the biggest questions facing humanity, from how to obtain affordable clean energy, to protecting ourselves and our environment. At ANL, Cerebras is working with research staff in the Computing, Environment, and Life Sciences (CELS) directorate to accelerate groundbreaking scientific and technical innovation in biology and life sciences.

Researchers at ANL are working on a wide range of cutting-edge research, from work with the U.S. National Cancer Institute (NCI) and the National Institutes of Health (NIH) to develop treatments for cancer, to Convolutional variational autoencoders (CVAE) for CoVID-19 research, as well as topics in physics, including time series data signatures of black hole collisions in gravitational wave data and the processing of particle accelerator data.

Relationship Overview

Cerebras announced its partnership with the Department of Energy (DoE), including a partnership with ANL, in September of 2019.[2] The original Cerebras CS-1 system was launched publicly in November of 2019 and was the world's most powerful solution for AI computing, with the largest chip ever made. When the CS-1 was deployed at ANL, Rick Stevens, Argonne Associate Laboratory Director for Computing, Environment and Life Sciences said, "By deploying the CS-1, we have dramatically shrunk training time across neural networks, allowing our researchers to be vastly more productive to make strong advances across deep learning research in cancer, traumatic brain injury and many other areas important to society today and in the years to come."[3]

ANL first employed the CS-1 to develop treatments for cancer by working on the "drug response problem," which is the complex challenge of modeling the response of a disease to candidate therapeutic compounds to accelerate development of treatments. The research aimed to use AI-powered predictive models for drug response to optimize pre-clinical drug screening and to drive precision medicine-based treatments for cancer patients.

When CoVID-19 reached pandemic status in March 2020, ANL quickly pivoted their research to help in the race to find life-saving treatments and therapeutics.[4] In this research, ANL focused on determining which existing drugs are effective in the fight against coronavirus – that is, which molecules from already-approved drugs can best bind to the docking sites of the virus' proteins to inhibit them from attaching to human host cells. Instead of employing brute force computation, which takes substantially more time and requires significantly

more computing power, ANL used the full force of its supercomputing infrastructure and state-of-the-art AI technology, including the Cerebras CS-1 system. This work resulted in a study on CoVID-19 reproduction that earned the multi-institutional team a finalist nomination for the Association of Computing Machinery (ACM) Gordon Bell Special Prize for High Performance Computing-Based COVID-19 Research. The team presented at the SC21 conference in November, 2021.[5] Another multi-institutional team went one better in 2022 with another CoVID-19 study, that this time actually won the Gordon Bell Special Prize at the SC22 conference.

LEARN MORE

2020 Gordon Bell Prize Work:

- Paper
- HPCwire article

2021 Gordon Bell Prize Finalist:

- Paper
- HPCwire article

2022 Gordon Bell Prize Winner:

- Blog
- HPCwire article

ALCF Fireside Chat (YouTube)

AI for Science Report (DoE Town Hall)

Other teams at ANL have leveraged the Cerebras CS-1 and CS-2 to work on problems in astrophysics and fundamental physics. This work includes building and running an AI model used to classify time series data signatures of black hole collisions in gravitational wave data. Traditional computer systems could run this model, but performance was slower than real-time, meaning that analysis could be delayed, or valuable data could be left on the table. The Cerebras systems enable researchers to run the model in real time, enabling faster and more complete data analysis, accelerating time to insight. Additionally, the Cerebras team worked closely with an ANL team and collaborators at the DoE-funded Stanford Linear Accelerator Center (SLAC) using deep learning to model and process particle accelerator data for fundamental physics to support simultaneous training of multiple small model replicas on the wafer. This work achieved extraordinary performance gains over existing systems.

In April of 2021, Cerebras Systems announced the CS-2 system, based on second-generation Wafer-Scale Engine, the WSE-2, which more than doubled the CS-1's performance. The CS-2 boasts 850,000 AI-optimized processor cores, 40 GB of high-performance on-wafer memory, 20 Petabytes of memory bandwidth, and 220 Petabits of core-to-core communication bandwidth. Thanks to the Wafer-Scale Engine, the CS-2 helps machine learning researchers reduce training times and reach solutions faster.

ANL's Leadership Computing Facility (ALCF) upgraded from a single CS-1 system to two CS-2 systems in 2021. One of these systems is part of the ALCF AI Testbed,[6] which enables researchers the ability to explore next-generation machine learning applications and workloads with the goal of advancing the use of AI for science.[7]

Finally, a notable highlight of the Cerebras partnership with ANL is the Exascale effort that Cerebras was selected to be a part of in March of 2020.[8] Here, multiple national labs joined the effort to build a nationwide AI program, modeled loosely on the U.S. Exascale Initiative, and Cerebras was named as a key player.

Before Cerebras: Clusters of GPUs as a Limiting Factor

For ANL to perform their ground-breaking research, they are using machine learning (ML) and large computational simulations. ANL's IT environment has thousands of graphics processing units (GPUs), as well as several of the largest supercomputers in the world. Why were these machines not sufficient to do the AI computational work required?

There are several major challenges associated with scaling large AI models across a cluster of GPUs. First, performance – as measured by wall clock training time to accuracy – does not scale linearly as more GPUs are added to a cluster. The performance of a GPU cluster is sublinear, while power draw and cost are super-linear. The reason for this is that as more GPUs are added to a cluster, the demand for cross-GPU communication explodes. This in turn necessitates layers of switches and routers, all of which add cost, power draw, and latency. The resultant communication "tax" eats into the compute resources available for useful work.

Second, clusters of GPUs are difficult to program. Researchers often spend weeks or even months refining their model, tuning hyperparameters such as batch size and learning rate, and orchestrating device-to-device communication patterns. Moreover, much of this ML engineering work is model and cluster-specific, so when a researcher wants to modify, change, or scale their model – an integral part of the research and development process in deep learning – they must start over. In a sense this can be thought of as a recurring ML engineering tax levied on research progress for large GPU cluster users.

Third, as clusters grow, researchers' hyperparameter choices are constrained. They need to use a larger and larger batch size, for example, to enable reasonable utilization on the cluster to make up for significant device-to-device communication overheads. This limits the ML methods available and restricts innovation.

All three of these limitations hampered ANL researchers' work using existing cluster computing infrastructure. In doing studies on clusters of GPUs, ANL researchers faced not only long training times for their models, but also long set-up times and a limited set of hyperparameter choices to get their models training on a cluster.

Together, these factors created a situation in which the total time to solution precluded research progress. Each time ANL researchers wanted to run a new experiment, they were forced to start from scratch – requiring engineering and tuning their cluster code before they could even execute a complete training run.

The status quo left many research questions unanswered as new ideas were simply too time-consuming to set up and test.

ANL Chooses Cerebras

To see why ANL partnered with Cerebras to address these computational bottlenecks, let's examine their work in three key areas: Cancer treatment, CoVID-19 drug discovery, and physics research.

Cancer Treatment

Training Drug Response Models (Uno)
Researchers at ANL are working with the U.S. National Cancer Institute (NCI) and the National Institutes of Health (NIH) to develop treatments for cancer as part of the CANDLE project. They are using ML and large computational simulations to do this. One of the largest projects is aimed at addressing the "drug response problem." The drug response problem is the complex challenge of modeling the response of a disease to candidate therapeutic compounds to accelerate development of treatments. ANL aims to develop AI-powered predictive models for drug response that can be

> **"We've been working on a series of models that run on the CS-1 predicting tumor response to drugs and these models are achieving speed ups of many hundreds of times on the CS-1 compared to our GPU baselines. We're quite happy with these. These are models that are running on large scale machines across the different architectures."**
>
> - Rick Stevens, Associate Laboratory Director of Computing, Environment and Life Sciences, Argonne National Laboratory[13]

used to optimize pre-clinical drug screening and to drive precision medicine-based treatments for cancer patients.

One of the first models deployed to the CS-1 was a multi-tower, fully connected, supervised learning model used to predict cancer cell response to candidate therapeutic compounds. According to ANL, "The scale in this problem derives from the number of relevant parameters to describe:

1. The properties of a drug or compound (O(106)),
2. The number of measurements of important tumor molecular characteristics (O(107)), and
3. The number of drug/tumor screening results (O(107))."

This model has tremendous scientific value and therapeutic potential but was challenging to experiment with and deploy using the existing GPU cluster infrastructure.

ANL put the Cerebras CS-1 to work on this important model immediately. Deployment was complete, and the systems were up and running within hours, rather than the weeks or months typically required to deploy a traditional cluster of GPUs.

> **"Cerebras allowed us to reduce the experiment turnaround time on our cancer prediction models by 300X, ultimately enabling us to explore questions that previously would have taken years, in mere months."**
>
> - Rick Stevens, Associate Laboratory Director of Computing, Environment and Life Sciences, Argonne National Laboratory

Because the CS-1 is programmed with standard TensorFlow, ANL researchers were able to begin training their exact existing models without modification, and immediately saw significant reduction in wall clock training time to desired accuracy.

The CS-1 trains ANL's drug response deep learning models in hours, rather than the days or weeks that had been the case with their legacy GPU cluster. When ANL researchers saw the first results from CS-1, they immediately nominated dozens of other model training experiments and new model architecture ideas, which had been impractical to test on clusters of GPUs, to run on the CS-1.

Put simply, the combination of extraordinary performance and ease of use of the CS-1 allowed the ANL research team to test more ideas per unit time and arrive at new solutions faster than ever before. Speaking about the massive Wafer-Scale Engine processor (WSE) at the heart of the CS-1 system, Rick Stevens, Associate Lab Director at ANL, told leading trade publication HPCwire that his teams were able to get their AI model running "many hundreds of times faster on the CS-1 than it runs on a conventional GPU."

Next Generation Drug Response Models (RUno)

The next generation model for the CANDLE project, called RUno, was the first model to have a use case of Online Normalization for training neural networks. Online Normalization is a new technique, proposed by machine learning scientists at Cerebras, for normalizing the hidden activations of a neural network. Like Batch Normalization, it normalizes the sample dimension. While Online Normalization does not use batches, it is as accurate as Batch Normalization.

LEARN MORE

- How AI is used in cancer treatment research
- CANDLE Research Program
- MIT Tech Article
- Online Normalization (NeurIPS, 2019)

CoVID-19 Drug Discovery

Predicting Docking Scores

Inventing a new drug, obtaining regulatory approvals, and reaching commercialization generally takes decades. A faster alternative to finding a therapeutic for CoVID-19 is to reuse drugs that are already approved for use in humans. The problem then becomes determining which existing drug will be effective in the fight against coronavirus – that is, which molecules from already-approved drugs can best bind to the docking sites of the virus' proteins to inhibit them from attaching to human host cells.

"This program is really aimed at generating new molecules that could be used as drugs. We've been applying this to our work in CoVID-19. It's also related to our cancer work in other areas where we're trying to rapidly search through large spaces of drug molecules to find [candidate] molecules that meet certain criteria. Again, we're getting great speedups there."

- Rick Stevens, Associate Laboratory Director of Computing, Environment and Life Sciences, Argonne National Laboratory

The first step in rendering the virus inactive is to identify which of the candidate molecules from existing approved drugs can bind to the pockets of the COVID-19 virus' proteins and block its docking sites. Historically, the way to tackle this challenge was through brute-force computation. Each molecule would be computed a "docking score" based on its characteristics or descriptors, which ends up being a very compute-intensive task given there are billions of molecules, and each of the virus' proteins have dozens of docking sites.

ANL took a different approach, leveraging the power of AI and ML to develop a model which much more efficiently processes ANL's massive datasets to determine which molecules have the best docking scores. This ML model allows ANL to quickly predict which drug molecules are the best candidates for the next stage of testing. Instead of employing brute force computation, which would take substantially more time and require significantly more computing power, ANL used the full force of its supercomputing infrastructure and state-of-the-art AI technology, including the Cerebras CS-1 system. ANL is using the CS-1 to train the models which predict docking scores. The speed in turnaround is allowing the team to churn through the massive datasets much more quickly, enabling faster experimentation at a time when the problem is urgent, and the potential exploration space is enormous.

"This iterative workflow of supporting streaming AI and MD techniques on emerging hardware platforms will pave the way for advancing our knowledge of how proteins function."

- Arvind Ramanathan, Argonne computational biologist[14]

Training the CVAE Model for Molecular Analysis

ANL and Cerebras also collaborated on training a second model which uses an image-based representation rather than a numerical representation of the molecular characteristics, and has, in testing, yielded excellent model results. By using image-based representations of the drug molecules and virus proteins, ANL can train the CS-1 to learn using the same "language" that chemists use to communicate about molecules – diagrams that detail molecular structure and shape. Once the CS-1 is running the ML models, the next step is to run inference over several billion samples to pick the ones most likely to bind to the virus' pocket and prevent it from binding to other host cells. From there, ANL is building in-silico models to understand the interactions of the selected drug molecules further. Finally, the most promising drug candidates are passed to the wet-lab for verification. This enables drug development facilities to test only the most viable drug molecules, saving valuable time and resources. This research resulted in a study that was nominated for a Gordon Bell Special Prize and presented at the SC21 supercomputing conference in 2021.

The process starts with three-dimensional images of the virus captured using cryo-electron microscopy. This technique can achieve near-atomic resolution, but the images are still not good enough, or dynamic enough, to show us how the mechanism really works. To fill in the missing data, the research team layered on top two completely different, but complementary techniques, working at different scales. First, we can treat biomolecules the way we treat any materials problem. We can use a type of the finite element analysis tools we routinely use to design continuum-scale objects such as engine parts. And second, we can simulate molecules atom-by-atom like a much more sophisticated version of the ball-and-stick models we all remember from chemistry class.

Putting all this together is a mammoth task. Innovation was needed at, as it were, every scale, from a novel workflow architecture that allows widely-distribute computing resources to mesh seamlessly and automatically, to improving the computational efficiency of the individual models.

That last part – improving computational efficiency – is where Cerebras comes in. In the past, these simulations took so long to create that it was only possible to study a few tens of nanoseconds of motion at one time. However, to reach a broader understanding, they needed to study microseconds – a 50x longer period of time. The team realized that the machine learning steps were the bottleneck to achieving the 50x speedup needed when integrating simulations with AI.

Each simulation experiment ties up a supercomputer with thousands of processing nodes for a long time. To avoid wasted time, it's vitally important to "steer"

LEARN MORE

2020 Gordon Bell Prize Work:

- Stream-AI-MD: streaming AI-driven adaptive molecular simulations for heterogeneous computing platforms (PASC '21 paper)
- HPCwire article

2021 Gordon Bell Prize Finalist:

- Intelligent Resolution: Integrating Cryo-EM with AI-driven Multi-resolution Simulations to Observe the SARS-CoV-2 Replication-Transcription Machinery in Action (Sage paper, 2021)
- Press Release (ANL)
- HPCwire article

2022 Gordon Bell Prizewinner:

- Genomics in Unparalleled Resolution: Cerebras Wafer-Scale Cluster Trains Large Language Models on the Full COVID Genome Sequence (blog)
- GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics (bioRxiv paper)
- HPCwire article

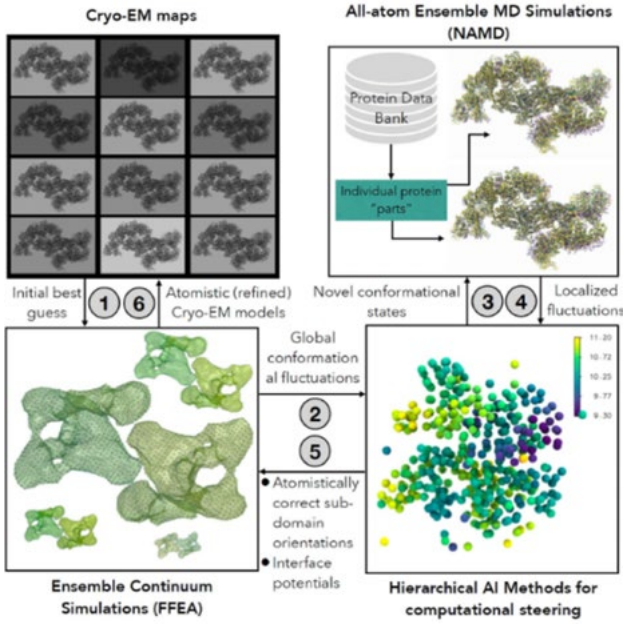


Diagram from the paper showing how the components of the study fit together. Our work is in the computational steering part.

these experiments, recognizing and halting simulations that are going down dead ends, and encouraging, so to speak, those that may prove fruitful. This is easier said than done. It's very difficult to specify the characteristics of a "bad" simulation beforehand. But it's easy after the fact to recognize that a bad thing happened. This is a classic ML opportunity: you know what the answer looks like, but you don't know how to define rules to describe it.

We address this with a machine learning model called a "convolutional variational autoencoder", or CVAE. Oversimplifying, a CVAE takes a complex "high-dimensional" input and transforms or "encodes" it into a smaller form. You can think of this as a kind of figure of merit. We train the model by letting it observe snapshots of the simulations. We then run the reverse transformation – or decode it. If the decoded version is a good match for the original, we know the CVAE is working. That trained model can then be used during "real" experiments by another algorithm that does the actual steering. However, as the paper points out: "CVAE is quadratic in time and space complexity and can be prohibitive to train."

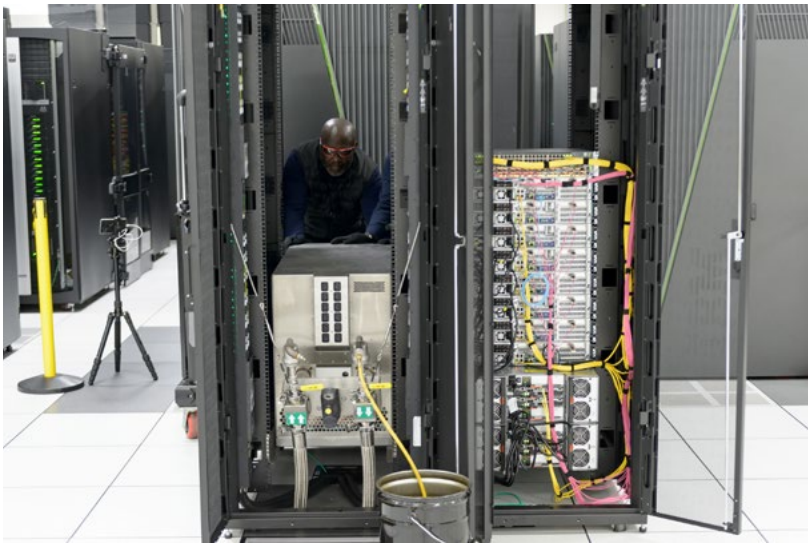
Cerebras comes into the picture here because this bit of the problem was being explored at Oak Ridge National Laboratory on the Summit supercomputer and on the Argonne AI-Testbed at ANL, which just happens to feature a Cerebras accelerator. The ANL researchers compared training their CVAE model on 256 nodes of Summit, for a total of 1,536 GPUs, and on a single Cerebras CS-2 system.

And how did Cerebras do? In terms of pure performance, rather well. Quoting the paper again: "the CS-2 delivers out-of-the-box performance of 24,000 samples/s, or about the equivalent of 110-120 GPUs."

> **"Because a single CS-2 here delivers the performance of over 100 GPUs, it is a practical alternative for organizations interested in this workflow who do not have extremely large GPU clusters."**
>
> - Anda Trifan et al, ANL[15]

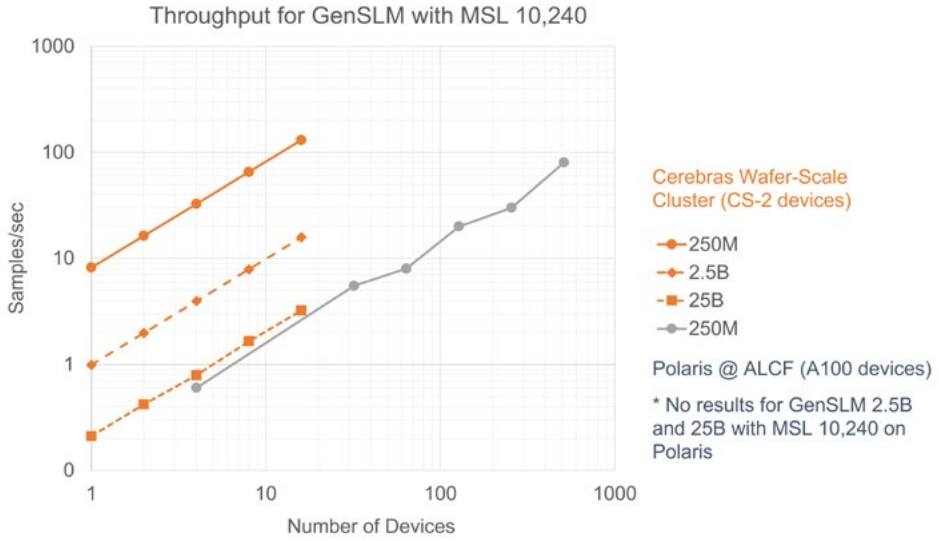
As impressive as this number is, perhaps even more impressive is the "out of the box" comment. Distributing a promising algorithm across a large cluster of compute nodes is difficult and time-consuming even for experts in the field. The CS-2 system, by contrast, is intentionally architected as a single, ultra-powerful node with cluster-scale performance. Our software makes it easy to get a neural network running by changing just a couple of lines of code.

Finally, it's important to bear in mind that while this study has direct benefits in the treatment of COVID-19, the new tools and workflow may ultimately prove much more significant. This methodology can be applied to any kind of molecular machinery, paving the way for more rapid and better understanding of molecular interactions across a wide range of use cases, including treatment discovery for a range of diseases.

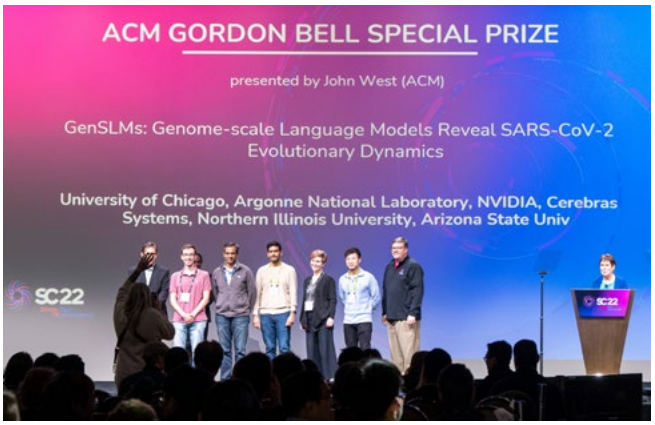


Training Large Language Models on the Full COVID Genome Sequence
For this work, ANL wanted to train full-genome models enabled by the unique long sequence length capabilities of the CS-2. The challenge of training these complex models was addressed using weight streaming to accelerate training using a Cerebras Wafer-Scale Cluster.

In genomics, a physical characteristic (or phenotype) can be coded for by genes that are far apart in the genetic sequence. The SARS-CoV-2 virus genome, which causes COVID-19, can have a sequence of 30,000 nucleotides or 10,000 codons. LLMs can make sense of genomic sequences, but their self-attention mechanism, which can relate different positions of a sequence, is severely limited by conventional computer systems. Fortunately, the Cerebras CS-2 has the capacity to train state-of-the-art LLMs with input sequences longer than the SARS-CoV-2 genome. That allows researchers to study genomics from a whole new perspective.



Cerebras engineers worked with the lab of Dr. Arvind Ramanathan at Argonne National Laboratory (ANL) to develop genome-scale language models (GenSLM) that reveal evolutionary dynamics of SARS-CoV-2. For the first time, we trained LLMs using between 123 million and 25 billion parameters with the full SARS-CoV-2 genome of sequence length 10,240, on a single CS-2 system, and to speed up the training we used as many as 16 CS-2s in a Cerebras Wafer-Scale Cluster. This work resulted in a paper, "GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics", which was presented at SC22.[9]



This joint work with Argonne National Laboratory (ANL) and NVIDIA was honored with the 2022 Gordon Bell Special Prize for HPC-Based COVID-19 Research.[10]

Physics Research

Researchers at ANL, in collaboration with other organizations, have used the CS-1 and CS-2 to accelerate advances in astrophysics and fundamental physics. This work includes projects on modeling black-hole collisions, a collaboration with the DoE-funded Stanford Linear Accelerator Center (SLAC) to process particle accelerator data, and the enhancement of X-Ray data analysis with AI.

Building and Running the Gravewave Model

A research group at ANL initially used Cerebras' CS-1 to build and run an AI model used to classify time series data signatures of black hole collisions in gravitational wave data as part of ANL's core fundamental astrophysics enterprise. Traditional computer systems could run this model, but performance was slower than real-time, meaning that analysis could be delayed or valuable data could be left on the table. Here, the CS-1 enabled researchers to run the model in real time, enabling faster and more complete data analysis, accelerating time to insight. "This is a project that's working on gravity wave detection [and] is building models that are suited for AI processing, the data coming off of things like LIGO to look for subtle signals of gravity waves. The project has been ongoing for about a year and is making good progress and achieving good speed ups on the CS-1," said Stevens in October of 2020.[11]

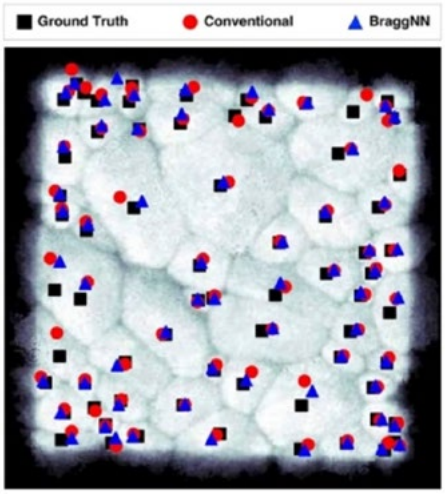
Comparing Training Times for Particle Accelerator Data (BraggNN and AutoEncoder)

Next, the Cerebras team worked closely with a different ANL team and collaborators at SLAC using deep learning to model and process particle accelerator data for fundamental physics. Here, the AI model used was relatively small, and the Cerebras team developed a new software mechanism to support simultaneous training of multiple small model replicas on the wafer to achieve extraordinary performance gains over existing systems. On CS-1 this resulted in a 290x training time acceleration beyond GPU; on CS-2 this resulted in a 1040x training time acceleration beyond GPU. The two main models used in this work are the BraggNN and AutoEncoder model. The ANL and SLAC team measured performance of these models on various hardware types:

Table 1: Time breakdown of the workflow steps/actions when using either a remote Cerebras DCAI system, a remote 8-GPU server or a remote SambaNova (only used 1 out of 8 RDUs per node) versus using one local GPU. The purpose of listing the performance of different systems is not to compare them, as this is not a systematical benchmarking study. The purpose is rather to demonstrate the feasibility of using powerful yet remote DCAI systems for AI at edge applications.

Mode \ Time	Neural Network	Data Transfer (s)	Model Training (s)	Model Transfer (s)	End-to-End (s)
Local (one GPU)	BraggNN [2]	N/A	1102	N/A	1102
Remote (Cerebras, Entire Wafer)		7	19	5	31
Remote(SambaNova, 1-RDU)		7	139	5	151
Local (one GPU)	CookieNetAE	N/A	517	N/A	517
Remote (Cerebras, Entire Wafer)		5	6	4	15
Remote (multi-GPU server)		5	88	4	97

Enhancing X-ray Data Analysis (BraggNN)
Finally, a different research group at ANL also worked on enhancing X-ray data analysis with artificial intelligence.[12] This research reused the BraggNN model. To quote their paper, "Training takes about 1 hr using one NVIDIA V100 GPU, and less than 1 min using the Cerebras artificial intelligence system."



> **"BraggNN can be trained in as little as 19 seconds when using the Cerebras system."**
>
> - Zhengchun Liu et al, ANL[16]

Powering the Future

Cerebras Systems builds compute solutions for the hardest AI problems. Our flagship product, the CS-2, is not only the fastest AI computer in existence, but is also easy to set up, quick to configure, and enables blisteringly fast training of models. ANL researchers have translated this exceptional compute performance into meaningful research in the development of treatments for cancer, advances in Convolutional variational autoencoders (CVAE) for CoVID-19 research, as well as applications in physics, including time series data signatures of black hole collisions in gravitational wave data and the processing of particle accelerator data. Our systems allow ANL to do in days and weeks what in the past took months or even years.

The original Cerebras CS-1 system was launched in November of 2019 and was the world's most powerful solution for AI computing with the largest chip ever made. In April of 2021, Cerebras Systems announced the CS-2 system, based on our second-generation Wafer-Scale Engine, the WSE-2, which more than doubled the CS-1's performance. The CS-2 brings 850,000 AI-optimized processor cores, 40 GB of high-performance on-wafer memory, 20 Petabytes of memory band-width, and 220 Petabits of core-to-core communication bandwidth. Thanks to the Wafer-Scale Engine, the CS-2 helps machine learning researchers reduce training times and reach solutions faster.

The CS-1 and CS-2 are currently deployed at national laboratories and super compute sites throughout the U.S. Europe, and Japan. These systems also provide best-in-market AI performance for enterprise customers across numerous market segments including pharmaceuticals, heavy manufacturing, and web services. In addition, our systems are deployed in the U.S. military and intelligence communities.

To explore how Cerebras systems can accelerate your research or to see a demo, please contact us at www.cerebras.net/get-demo.

References

1 Sam Cox, "New chip technology could enable 'brain-scale' AI", Silicon Republic, 2021 https://www.siliconrepublic.com/machines/cerebras-systems-brain-scale-ai

2 Business Wire, "Department of Energy and Cerebras Systems Partner to Accelerate Science with Supercompute Scale Artificial Intelligence", 09/17/2019, https://www.businesswire.com/news/home/20190917005356/en/Department-of-Energy-and-Cerebras-Systems-Partner-to-Accelerate-Science-with-Supercompute-Scale-Artificial-Intelligence

3 ANL, "Argonne National Laboratory Deploys Cerebras CS-1, the World's Fastest Artificial Intelligence Computer", 11/19/2019, https://www.anl.gov/article/argonne-national-laboratory-deploys-cerebras-cs1-the-worlds-fastest-artificial-intelligence-computer

4 Cerebras Systems, "Argonne National Laboratory and Cerebras Systems Leverage the World's Fastest AI Supercomputer to Advance COVID-19 Research", 05/13/2020, https://www.cerebras.net/blog/argonne-national-laboratory-and-cerebras-systems-leverage-the-worlds-fastest-ai-supercomputer-to-advance-covid-19-research/

5 Matt Lakin, "Waltzing the virus: Study on COVID-19 reproduction earns Argonne researchers Gordon Bell Special Prize nomination", 11/16/2021, https://www.anl.gov/article/waltzing-the-virus-study-on-covid19-reproduction-earns-argonne-researchers-gordon-bell-special-prize

6 Jim Collins, "ALCF AI Testbed's Cerebras and SambaNova systems now available to research community", 05/02/2022, https://www.alcf.anl.gov/news/alcf-ai-testbeds-cerebras-and-sambanova-systems-now-available-research-community

7 Business Wire, "Leading Supercomputer Sites Choose Cerebras for AI Acceleration", 05/31/2022, https://www.businesswire.com/news/home/20220531005775/en/Leading-Supercomputer-Sites-Choose-Cerebras-for-AI-Acceleration

8 John Russell, "Conversation: ANL's Rick Stevens on DoE's AI for Science Project", 03/23/2020, https://www.hpcwire.com/2020/03/23/conversation-anls-rick-stevens-on-does-ai-for-science-project/

9 Maxim Zvyagin, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, et al, "GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics", SC22, 2022, https://www.biorxiv.org/content/10.1101/2022.10.10.511571v2

10 ACM Press Release "ACM Gordon Bell Special Prize for HPC-Based COVID-19 Research Awarded to Team for Modelling How Pandemic-Causing Viruses, Especially SARS-CoV-2, are Identified and Classified", 2022, https://www.acm.org/media-center/2022/november/gordon-bell-special-prize-covid-research-2022

11 John Russell, "LLNL, ANL and GSK Provide Early Glimpse into Cerebras AI System Performance", 10/13/2020, https://www.hpcwire.com/2020/10/13/llnl-anl-and-gsk-provide-early-glimpse-into-cerebras-ai-system-performance/

12 Jared Sagoff, "Hitting a new peak: Scientists enhance X-ray data analysis with artificial intelligence", 05/11/2022, https://www.alcf.anl.gov/news/hitting-new-peak-scientists-enhance-x-ray-data-analysis-artificial-intelligence

13 Cerebras, Customer Spotlight – Argonne National Laboratory, https://www.cerebras.net/cerebras-customer-spotlight-overview/spotlight-argonne-national-laboratory/

14 Alexander Brace, Michael Salim, Vishal Subbiah et al, "Stream-AI-MD: streaming AI-driven adaptive molecular simulations for heterogeneous computing platforms" PASC '21, https://dl.acm.org/doi/abs/10.1145/3468267.3470578

15 "Intelligent Resolution: Integrating Cryo-EM with AI-driven Multi-resolution Simulations to Observe the SARS-CoV-2 Replication-Transcription Machinery in Action", https://journals.sagepub.com/doi/full/10.1177/10943420221113513

16 Zhengchun Liu et al, "Bridging Data Center AI Systems with Edge Computing for Actionable Information Retrieval", v3, 2022, https://arxiv.org/abs/2105.13967