Cerebras Systems deploys the ‘world’s fastest AI computer’ at Argonne National Lab

Dean Takahashi@deantak November 19, 2019 5:00 AM

Cerebras Systems is unveiling the CS-1, billed as the fastest artificial intelligence computer in the world and certainly one of the most daring attempts to create a better supercomputer. And it has gained acceptance from the U.S. federal government’s supercomputing program.

The CS-1 has an entire wafer as its brain, rather than a chip. Normally, silicon chips are carved out of processed 12-inch silicon wafers, with many hundreds of chips on a wafer. But Los Altos, California-based Cerebras has designed a computer with lots of little cores, all repeated across an entire wafer. That wafer is sawed into a big rectangle, but it has the equivalent of lots of chips on it. The CS-1 was announced at the Supercomputing 2019 event today.

All told, there are more than 1.2 trillion transistors across all of the cores on one wafer, whereas a typical processor might have 10 billion transistors on one chip. But the CS-1 supercomputer goes even further: It has one of these Cerebras wafers — each called a Wafer Scale Engine — in one system. It’s a behemoth.

And Cerebras has delivered the first CS-1 to the U.S. Department of Energy’s Argonne National Laboratory, one of the world’s biggest supercomputer buyers. It will use the 400,000 cores to handle massive AI computing problems, such as studying cancer drug interactions.

With every component optimized for AI work, the CS-1 delivers more compute performance at less space and less power consumption than any other system, the company said. “This system itself is as tall as 15 racks,” said Andrew Feldman, CEO of Cerebras. “That’s 26 inches tall.”

In August, Cerebras delivered the Wafer Scale Engine (WSE), the only trillion-transistor wafer scale processor in existence. The Cerebras WSE is 56.7 times larger and contains 78 times more compute cores than the largest GPU, setting a new bar for AI processors.

The CS-1 system design and Cerebras software platform combine to extract every ounce of processing power from the 400,000 compute cores and 18 gigabytes of high performance on-chip memory on the WSE.

In AI compute, chip size is profoundly important. Big chips process information more quickly, producing answers in less time. However, exceptional processor performance is necessary but not sufficient. Advanced processors like the WSE must be combined with dedicated hardware systems and software to achieve record-breaking performance. For this reason, every aspect of the Cerebras CS-1 system and the Cerebras software platform was designed for accelerated AI compute.

“This is the largest square that you can cut out of a 300 millimeter wafer,” said Feldman, in an interview with VentureBeat. “Even though we have the largest and fastest chip, we know that an

extraordinary processor is not necessarily sufficient to deliver extraordinary performance. If you want to deliver really fast performance, you need to build a system. And you can't take a Ferrari engine and put it in a Volkswagen to get Ferrari performance. What you do is you move the bottlenecks if you want to get a 1,000 times performance gain."

Cerebras said it is the only company to undertake the ambitious task of building a dedicated system from the ground up. By optimizing every aspect of chip design, system design, and software, the CS-1 delivers unprecedented performance. With the CS-1, AI work that today takes months can now be done in minutes, and work that takes weeks now can be completed in seconds.

Not only does the CS-1 radically reduce training time, but it also sets a new bar for latency in inference. For deep neural networks, single image classification can be accomplished in microseconds, thousands of times faster than alternative solutions.

"We are an AI machine built up of 400,000 dedicated AI processors," Feldman said.

CS-1 system

Above: Cerebras says the CS-1 is the fastest AI computer.

Image Credit: Cerebras

At the Argonne National Laboratory, the CS-1 is being used to accelerate neural networks in cancer studies, to better understand the properties of black holes, and to help understand and treat traumatic brain injuries. The sheer performance of the CS-1 makes it an exceptional solution for the largest and most complex problems in AI.

"The CS-1 is a single system that can deliver more performance than the largest clusters, without the overhead of cluster set up and management," said Kevin Krewell, principal analyst at Tirias Research, in a statement. "By delivering so much compute in a single system, the CS-1 not only can shrink training time but also reduces deployment time. In total, the CS-1 could substantially reduce overall time to answer, which is the key metric for AI research productivity."

Unlike GPU clusters, which can take weeks or months to set up, require extensive modifications to existing models, consume dozens of datacenter racks, and require complicated proprietary InfiniBand to cluster, the CS-1 takes minutes to set up.

Users can simply plug in the standards-based 100Gb Ethernet links to a switch and start training models at record-breaking speed.

Cerebras software platform

Above: Cerebras' chip up close.

Image Credit: Cerebras

The CS-1 is easy to deploy and simple to use. Cerebras' mission is to accelerate not only time-to-train, but also the end-to-end time it takes for researchers to achieve new insights — from model definition to training to debugging to deployment.

The Cerebras software platform is designed to allow machine learning (ML) researchers to leverage CS-1 performance without changing their existing workflows. Users can define their models for the CS-1 using industry-standard ML frameworks such as TensorFlow and PyTorch.

A powerful graph compiler automatically converts these models into optimized executables for the CS-1, and a rich set of tools enables intuitive model debugging and profiling.

"We use open source and make it as easy to program as we can," Feldman said.

The system is neither x86 or Linux based.

"The compute cores are designed for their custom cores by us," Feldman said. "The software stack runs on a host anywhere in your network. And so what happens is, is you take your TensorFlow model and our software arrives in a container. And you point your software at our container, and our container grabs your software, and it compiles it, and it produces a configuration file which it sends to our machines."

The Cerebras software platform is comprised of four primary elements:

- Integration with common ML frameworks like TensorFlow and PyTorch
- Optimized Cerebras Graph Compiler (CGC)
- Flexible library of high-performance kernels and a Kernel API
- Development tools for debug, introspection, and profiling

The Cerebras Graph Compiler

Above: Cerebras has an efficient cooling system.

Image Credit: Cerebras

The Cerebras Graph Compiler (CGC) takes as input a user-specified neural network. For maximum workflow familiarity and flexibility, researchers can use both existing ML frameworks and well-structured graph algorithms written in other general-purpose languages, such as C and Python, to program for the CS-1.

CGC begins the translation of a deep learning network into an optimized executable by extracting a static graph representation from the source language and converting it into the Cerebras Linear Algebra Intermediate Representation (CLAIR). As ML frameworks evolve rapidly to keep up with the needs of the field, this consistent input abstraction allows CGC to quickly support new frameworks and features without changes to the underlying compiler.

Using its knowledge of the unique WSE architecture, CGC then allocates compute and memory resources to each part of the graph and then maps them to the computational array. Finally, a communication path, unique to each network, is configured onto the fabric.

Because of the massive size of the WSE, every layer in the neural network can be placed onto the fabric at once and run simultaneously in parallel. This approach to whole-model acceleration is unique to the WSE — no other device has sufficient on-chip memory to hold all layers at once on a single chip, or the enormous high-bandwidth and low-latency communication advantages that are only possible on silicon.

The final result is a CS-1 executable, customized to the unique needs of each neural network, so that all 400,000 compute cores and 18GB of on-chip SRAM can be used at maximum utilization toward accelerating the deep learning application.

Development tools and APIs

Above: Cerebras' CS-1 system is relatively small.

Image Credit: Cerebras

CGC's integrations with popular ML frameworks mean that popular tools such as TensorBoard are supported out of the box. In addition, Cerebras provides a full-featured set of debugging and profiling tools to make deeper introspection and development easy.

For ML practitioners, Cerebras provides a debugging suite that allows visibility into every step of the compilation and training run.

For advanced developers interested in deeper flexibility and customization, Cerebras provides a Kernel API and a C/C++ compiler based on LLVM that allows users to program custom kernels for CGC. Combined with extensive hardware documentation, example kernels, and best practices for kernel development, Cerebras provides users with the tools they need to create new kernels for unique research needs.

The WSE

Above: The Cerebras wafer.

Image Credit: Cerebras

The Cerebras WSE is the largest chip (if you can call a wafer just one chip) ever made, and the industry's only trillion-transistor processor. It contains more cores, with more local memory and more fabric bandwidth, than any chip in history.

This enables fast, flexible computation at lower latency and with less energy. The WSE is 46,255 millimeters square, which is 56 times larger than the largest GPU. In addition, with 400,000 cores, 18GB of on-chip SRAM, 9.6 petabytes per second (PBps) of memory bandwidth, and 100 petabits per second (Pbps) of interconnect bandwidth, the WSE contains 78 times more compute cores; 3,000 times more high speed, on-chip memory; 10,000 times more memory bandwidth; and 33,000 times more fabric bandwidth than its GPU competitors.

Feldman acknowledged that it's very hard for chip manufacturers like TSMC to build a wafer without any flaws. That's why his team built redundancy into the system. There are as many as 6,000 spare cores on a wafer that holds more than 400,000 cores. If a manufacturing impurity messes up one of the cores, then Cerebras can route around that and replace it with one of the spares, Feldman said.

Feldman believes this system will be serious competition for Nvidia's rival GPUs.

"A Ford F150 is a terrible, terrible vehicle if you want to bring kids to soccer practice," Feldman said. "And so what we've done is we've built a machine that in every aspect is optimized for artificial intelligence work. We do one thing very well, and that's AI work. You can think of the GPU the same way. I mean, the GPU is an extraordinary machine. It makes great graphics. It's got all this AI capability. It's unbelievable what graphics can do — but it was designed for graphics, it wasn't designed for deep learning, and our system in every aspect is tuned and optimized for deep learning."

Working with Argonne

Above: Taking apart a CS-1.

Image Credit: Cerebras

The Argonne lab is a multidisciplinary science and engineering research center. The CS-1 will enable the largest supercomputer sites in the world to achieve 100- to 1,000-fold improvement over existing AI accelerators.

By pairing supercompute power with the CS-1's AI processing capabilities, Argonne can now accelerate research and development of deep learning models to solve science problems not achievable with existing systems.

"We've partnered with Cerebras for more than two years and are extremely pleased to have brought the new AI system to Argonne," said Rick Stevens, Argonne's associate laboratory director for computing, environment, and life sciences, in a statement. "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."

A subset of AI called deep learning allows computer networks to learn from large amounts of unstructured data. However, deep learning models require massive amounts of computing power and are pushing the limits of what current computer systems can handle — until now, with the introduction of Cerebras CS-1.

Argonne deployed the CS-1 to enhance scientific AI models. Its first application area is cancer drug response prediction, a project that is part of a Department of Energy (DoE) and National Cancer Institute collaboration aimed at employing advanced computing and AI to solve grand challenge problems in cancer research. The addition of the Cerebras CS-1 supports efforts to extend Argonne's major initiatives in advanced computing, which will also leverage the AI capabilities of the Aurora exascale system expected in 2021.

Argonne's deployment of the CS-1 is the first part of a multi-laboratory partnership between the DoE and Cerebras Systems. Cerebras has also partnered with DoE's Lawrence Livermore

National Laboratory to accelerate its AI initiatives and further enhance its simulation strengths with the machine learning capabilities of the CS-1.

"At the Department of Energy, we believe public-private partnerships are an essential part of accelerating AI research in the United States," said Dimitri Kusnezov, DoE's deputy undersecretary for Artificial Intelligence & Technology, in a statement. "We look forward to a long and productive partnership with Cerebras that will help define the next generation of AI technologies and transform the landscape of DOE's operations, business and missions."

It's easy to see why Feldman has hired a big staff and raised hundreds of millions of dollars. (He won't say how much.)

"I think we're in for a very exciting career five years ahead of us," said Feldman. "I think small groups of people can change the world. I think that's really the entrepreneur's mantra. And you don't need a giant company. You don't need billions of dollars, just a small group of extraordinary engineers can really change the world. And we believe that every day."