Cerebras Systems: Achieving Industry Best AI Performance Through A Systems Approach

A Systems Approach to Deep Learning

The Deep Learning Problem

Deep learning has emerged as one of the most important computational workloads of our generation. Its applications are widespread and growing. But deep learning is profoundly computationally intensive. Between 2015 and 2020, the compute used to train the largest models increased by 300,000x. In other words, AI compute demand is doubling every 3.5 months.

Because of this voracious demand, AI is constrained by the availability of compute; not by applications or ideas. Testing a single new hypothesis — training a new model — can take weeks or months and can cost hundreds of thousands of dollars in compute time. This is a significant drag on the pace of innovation, and many important ideas are ignored simply because they take too long to test.

The Need for System-Level Thinking

Delivering performance for deep learning is an end-to-end problem. One cannot put a Ferrari engine in a Volkswagen and expect Ferrari performance. To achieve Ferrari performance, every aspect of the car must be tuned and co-designed with the engine. So too it is with compute.

Putting a faster chip in a general-purpose server cannot vastly accelerate a workload on its own — it simply moves the bottleneck. A PCIe form factor limits the chip's size; the type of connections it can accept; even the communication standard it must conform to. A standard server enclosure defines what cooling capabilities are available and the ultimate power envelope. Inside this tight bounding box, innovation is heavily constrained.

To accelerate training by a hundred- or thousand-fold requires a fundamental rethinking of more than just the processor. It requires reinvention of all aspects of the system design, including the system architecture, the design of the core, the memory architecture, the communication fabric, the chip I/Os, the power and cooling infrastructure, the system I/Os, the compiler, the software toolchain — to name just a few of the elements that need to be optimized to achieve orders of magnitude performance gain.

Cerebras is the only company to undertake the ambitious task of designing a system from the ground up to accelerate AI applications. It has allowed us to ask the fundamental question — what is the ideal processor for deep learning? — and make tradeoffs across every domain to realize it. We have created the world's largest, most powerful deep learning-optimized chip and built the entire system around its needs.

The result was the Cerebras CS-1 — the world's fastest AI accelerator — and then two years later, the CS-2, which more than doubled the best-in-industry performance of the CS-1.

The Cerebras CS-2:
The world's only purpose-built Deep Learning solution

The CS-2 is a system solution that consists of innovations across three dimensions: a) the second generation Cerebras Wafer Scale Engine (WSE-2) — the industry's largest and only multi-trillion-transistor processor, b) the Cerebras System and c) the Cerebras software platform.

Cluster-scale deep learning compute in a single system

15 RU tall
Fits in a standard datacenter rack

1.2 Terabits/sec
System IO over 12x standard 100 GbE

23 kW
Maximum power draw

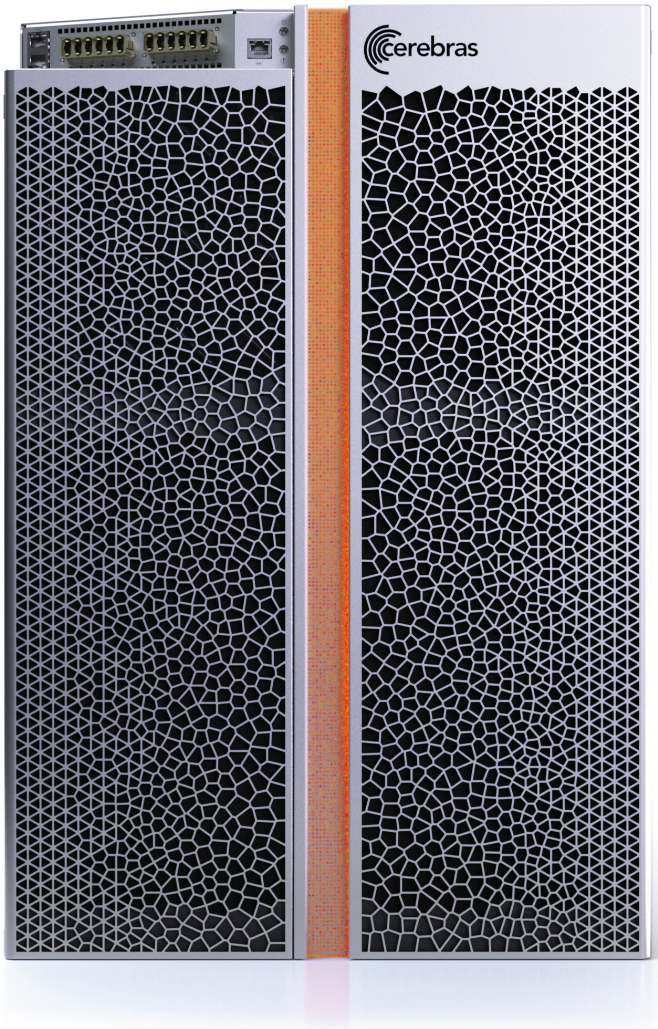


Figure 1. The CS-2, the industry's fastest AI computer

The Second Generation Wafer Scale Engine 2

The Cerebras second generation Wafer Scale Engine (WSE-2) is the processor at the heart of the CS-2. The WSE-2 is the largest chip ever built. It is the industry's only multi-trillion transistor processor, and contains more cores, more local memory, and more fabric bandwidth than any chip in history. This enables fast, flexible computation at lower latency and with less energy.

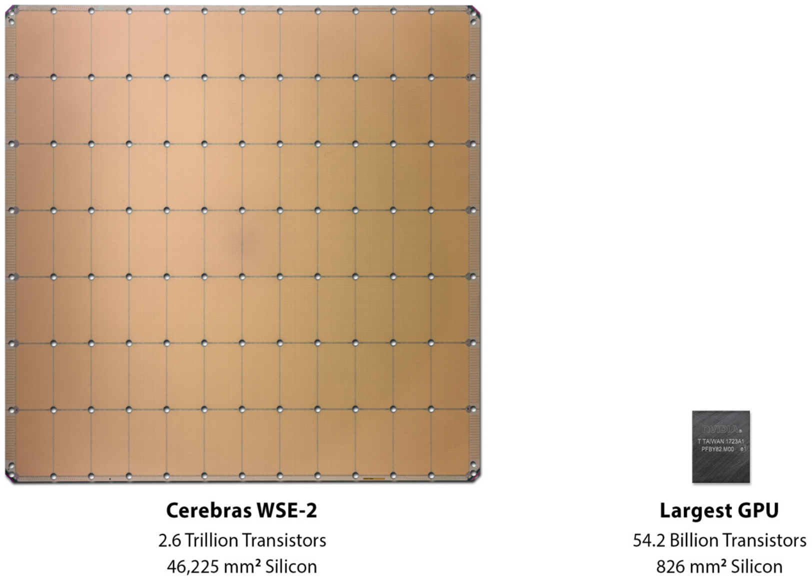


Figure 2. The Cerebras WSE-2 and the largest Graphics Processing Unit in comparison

Processors are typically fabricated by printing many copies of the same chip onto a wafer. The wafer is cut into individual chips, which are packaged and sold as individual processors. People spend an enormous amount of time, money, and effort to re-connect these chips — via InfiniBand and other interconnect technologies — back into clusters. It's like purposefully breaking Humpty Dumpty, only to try to piece him back together again. At Cerebras, we keep Humpty Dumpty whole. We cut the largest square from a single wafer of silicon, and this is our processor — the WSE-2.

The WSE-2 covers 46,255 square millimeters — 56 times larger than the largest graphics processing unit. With 850,000 cores, 40 Gigabytes of on-chip SRAM, 20 petabytes/sec of memory bandwidth, and 220 petabits/sec of interconnect bandwidth, the WSE-2 contains 123 times more compute cores, 1,000 times more high-speed on-chip memory, 12,862 times more memory bandwidth and 45,833 times more fabric bandwidth than its graphics processing competitor. In effect, it provides the compute capacity of an entire cluster in a single chip, without the cost, complexity, and bandwidth bottlenecks involved with lashing together hundreds of smaller devices. A summary table of the comparison is below.

	Cerebras WSE-2	A100	Cerebras Advantage
Chip size	46,225 mm^2	826 mm^2	**56 X**
Cores	850,000	6,912 + 432	**123 X**
On chip memory	40 Gigabytes	40 Megabytes	**1,000 X**
Memory bandwidth	20 Petabytes/sec	1,555 Gigabytes/sec	**12,862 X**
Fabric bandwidth	220 Petabits/sec	600 Gigabytes/sec	**45,833 X**

Table 1. Overview of the magnitude of advancement made by the Cerebras WSE-2.

Computation inside the WSE-2 happens in 850,000 AI-optimized Sparse Linear Algebra Compute (SLAC) cores. The SLAC cores are designed for the sparse linear algebra primitives that underpin all neural network computation. This programmability ensures cores can run all neural network algorithms in the constantly changing machine learning field.

Because the SLAC cores are optimized for neural network compute primitives, they achieve industry-best utilization — often double or triple that of a graphics processing unit. In addition, the cores include proprietary sparsity-harvesting technology which can accelerate computational performance on sparse workloads (workloads that contain zeros). Often, in calculations for deep learning, the majority of the elements in the vectors and matrices that are to be multiplied together are zero. And yet multiplying by zero is a waste of silicon, power, and time. No new information is made.

Because graphics processing units and tensor processing units are dense execution engines — engines designed to never encounter a zero — they multiply every element even when it is zero. When 50 to 98 percent of the data are zeros, as is often the case in deep learning, most of the multiplications are wasted. Imagine trying to run somewhere quickly, when most of your steps don't move you forward at all. The Cerebras SLAC cores never multiply by zero. All zero data is filtered out and skipped in the hardware. Instead, useful work is done in its place.

> *"If you look back over the last 30 years or so, there's a handful of inflection points in technology where somebody has a new idea, there's a new product, and that product sets the standard for how the future evolves, and I think the CS-2 is in that category."*
>
> ***Rick Stevens***
> *Argonne National Laboratory*

Memory is a key component of every computer architecture. Memory closer to compute translates to faster calculation, lower latency, and better power efficiency for data movement. High performance deep learning requires massive compute with frequent access to data. This requires especially close proximity between the compute cores and memory. This is a big problem for graphics processing units where the vast majority of the memory is slow and far away — off-chip.

The WSE-2 has 40 Gigabytes of on-chip memory, all uniformly distributed alongside the cores, and 20 Petabytes/sec of memory bandwidth. As a result, the WSE-2 can keep the entire neural network model in on-chip memory, all of the time, on the same piece of silicon as the compute cores. This means that all model parameters can be accessed at extremely high bandwidth at single-cycle latency.

The Cerebras Swarm communication fabric creates a massive on-chip network that delivers breakthrough bandwidth and low latency, at a fraction of the power draw of traditional communication techniques that are used to aggregate servers of graphics processing units into large clusters.

Swarm connects all 850,000 cores on the Cerebras WSE-2 in a 2D-mesh with 220 Petabits/sec of bandwidth. Swarm provides a hardware routing engine to each of the cores and connects them with short wires optimized for bandwidth and low-latency. The resulting fabric supports single-word active messages that can be received by the cores without any software overhead, providing flexible, all-hardware communication.

Swarm is also fully configurable. Cerebras' software configures all the cores and routers on the WSE to support a unique, optimized communication pattern for each particular neural network. This is different from the approach taken by central processing units and graphics processing units that have one hard-coded, on-chip communication path into which all workloads are shoehorned.

The Cerebras WSE-2 includes more cores, more local memory, and more core-to-core communication than any chip in history. This enables fast, flexible computation, at lower latency and with less energy — unlocking unprecedented deep learning performance on a single chip.

The CS-2 System

A revolutionary chip of unparalleled size and power requires a revolutionary system to drive it.

The CS-2 is 26-inches (15 rack units) tall and fits in one-third of a standard datacenter rack. It houses the Cerebras Wafer Scale Engine and feeds the 850,000 AI-optimized compute cores and 40 Gigabytes of high-speed, on-wafer memory with 1.2 terabits per second of data.

This combination of enormous input/output bandwidth — 12 x 100 Gigabit Ethernet lanes – and 40 Gigabytes of fast, on-chip memory enables the CS-2 to deliver more calculations per unit time than any machine ever built. And since all the computation and communication remains on-chip, where extraordinary power efficiency is to be had, the CS-2 consumes only a small fraction of the power and space of alternative solutions.

Powering and cooling such a large chip requires fundamental technology innovations. The custom power delivery and cooling technology inside the CS-2 keeps the chip running at full performance at a temperature well below the operating temperature of traditional processors. In microelectronics, high temperatures are the enemy of reliability; the lower operating temperature of the WSE-2 enhances both its reliability and performance.

Finally, unlike clusters of graphics processing units — which can take weeks or months to set up, require extensive hyperparameter tuning to converge on, occupy dozens of datacenter racks and require proprietary InfiniBand to cluster — the CS-2 takes minutes to set up. Simply plug in the standards-based 100 Gigabit Ethernet links to a switch, and you are ready to train models at wafer-scale speed.

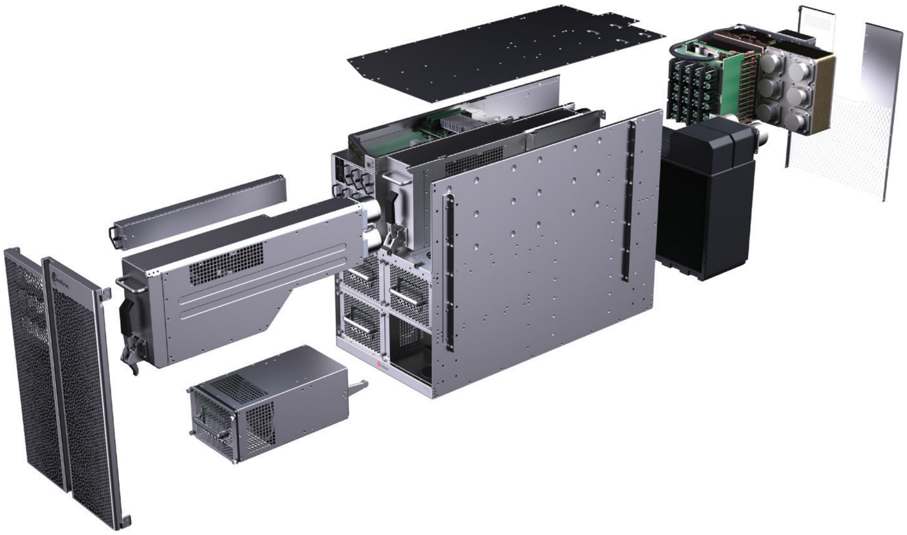


Figure 3. An inside view of the CS-2. Left to right - doors, fans, pumps, power supplies, main chassis, heat exchanger, engine block, back grill.

To supply power and signal to the world's largest chip, the CS-2 uses tailored configurations of standard parts and interfaces. In the top left of the system, as shown in Figure 4, twelve standard power supplies (PSUs), in a 9+3 redundant configuration, deliver up to 23,000 watts to the WSE-2. 12 x 100 Gigabit Ethernet links sit above the PSUs to bring in up to 1.2 Terabits per second of input data bandwidth from the surrounding datacenter infrastructure.

The rest of the system is dedicated to removing heat from the WSE-2. To provide the cooling horsepower the WSE-2 needs while keeping datacenter integration simple, the CS-2 is internally water-cooled. Water circulates through a closed loop, fully self-contained within the system. Like a giant gaming PC, the CS-2 uses water to cool the WSE-2, and then air to cool the water.

The top right of the system is for the movement of water. Two hot-swappable pumps move water through a manifold across the back of the WSE-2, cooling the wafer and warming the water. Warm water is then pumped into a heat exchanger. This heat exchanger presents a large surface area for the cold air blown in by the four hot-swappable fans at the bottom of the CS-2. These fans move air from the cold aisle, cool the warm water via the heat exchanger, and exhaust the warm air into the warm aisle.

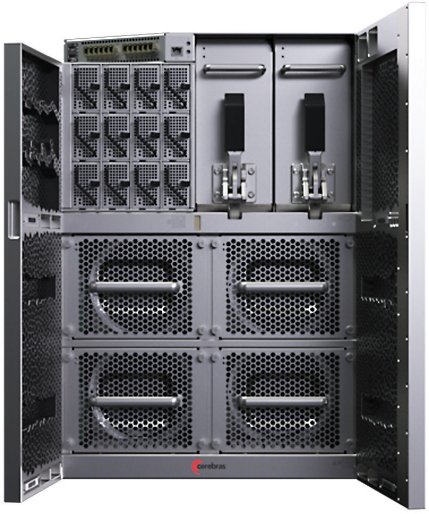


Figure 3: Front view of the CS-2, with doors open. Fans in the bottom half move air; pumps in the top right move water, power supplies and I/O in the top left provide power and data.

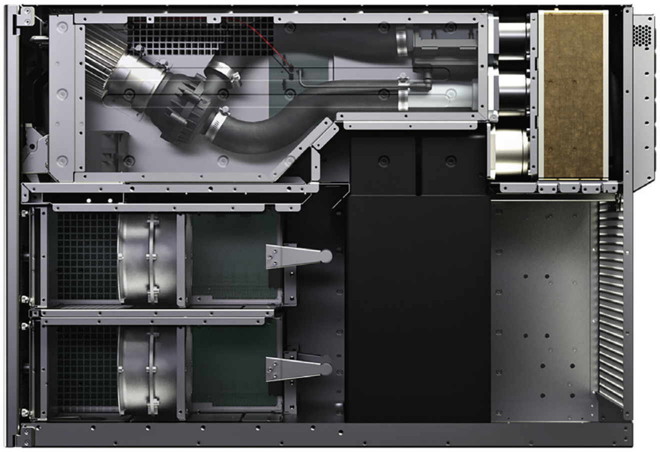


Figure 4: This side view shows the water movement assembly (top), and the air movement infrastructure — fans and a heat exchanger (bottom half).

Once power and data are brought into the system, the challenge remains to deliver them to the chip. The computational magic occurs in the back of the system, in the engine block, where Cerebras' innovations in chip packaging, power delivery, data streaming, cooling, and electrical connectivity are married to the Wafer Scale Engine.

In the engine block, thousands of watts delivered through the power pins are stepped down to the sub-volt thresholds used by the WSE-2. Because the current density required is so high, the traditional method of distributing power through the edges of the board results in too much dissipation at the chip's center. The custom packaging solution of the engine block instead brings power and data through the main board, perpendicularly to the wafer face, rather than at the edges. A novel, flexible connector between the silicon wafer and main board maintains electrical connectivity to the chip as the pieces expand and contract at different rates when heated and cooled.

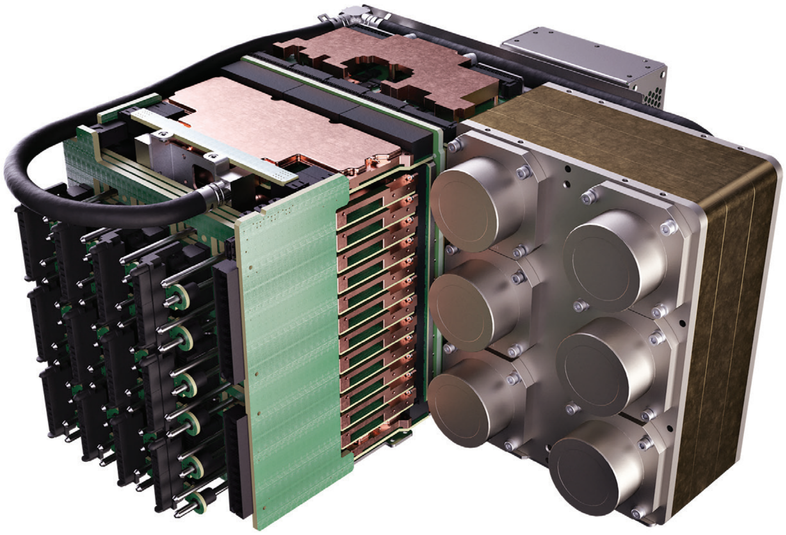


Figure 5: The engine block of the CS-2.

The innovations required to enable the Wafer Scale Engine are only possible because of the flexibility afforded by a full system solution. Every component of the CS-2 — from power and data delivery to cooling to software has been codesigned and optimized to take full advantage of this massive, deep learning chip.

Cerebras Software Platform

Cerebras' mission is to accelerate not only time-to-train, but the end-to-end time it takes for researchers to achieve new insights — from model definition, to training, to debugging and deployment.

The Cerebras software platform allows machine learning (ML) researchers to leverage CS-2 performance without changing their existing workflows. Users can define their models using popular ML frameworks such as TensorFlow and PyTorch. A flexible graph compiler automatically converts these models into optimized executables for the CS-2, and a rich set of tools enables intuitive debugging and profiling.

The Cerebras software platform is comprised of four primary elements:

1. The optimized Cerebras Graph Compiler (CGC)
2. A flexible library of high-performance kernels and a kernel-development API
3. Development tools for debug, introspection, and profiling
4. Clustering software

The Cerebras Graph Compiler

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 — such as TensorFlow and PyTorch and well-structured graph algorithms written in other general-purpose languages, such as C and Python, to program the CS-2.

To translate a deep learning network into an optimized executable, GCC extracts a static graph representation of the problem from the source language and converts 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.

Once the CLAIR graph has been extracted, CGC performs a matching and covering operation that matches subgraphs to kernels from the Cerebras kernel library. These kernels are optimized to provide high-performance compute at extremely low latency on the fabric of the WSE-2. The result of this matching operation is a kernel graph. CGC then allocates compute and memory to each kernel in the graph and maps every kernel onto a physical region of the computational array of cores. Finally, a communication path, unique to each network, is configured onto the fabric.

During this compilation process, kernel placement is formulated as a multi-constraint problem on 1) memory capacity and bandwidth, 2) computation requirements, and 3) communication costs. The placement engine then takes into account both algorithmic efficiency and compute core utilization to generate a result that maximizes locality, minimizes routing distances, and avoids hotspots. The final result is a CS-2 executable, customized to the unique needs of each neural network, so that all 850,000 SLAC cores and 40 Gigabytes of on-chip SRAM can be used at maximum utilization towards accelerating the deep learning application.

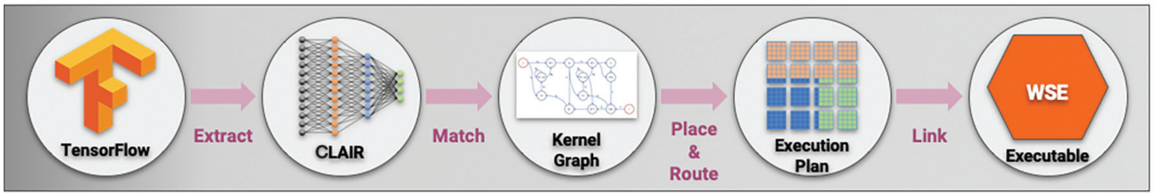


Figure 6: A high-level overview of the compilation process for the WSE-2.

Because of the massive size of the WSE-2, every layer in the neural network can be placed onto the fabric at once and run simultaneously. The computation is parallel at three levels: within the core, there is multiple operation per-cycle parallelism; across each fabric region, the cores can work in parallel on one layer; and all layers can run in parallel on separate fabric regions. This approach to whole-model acceleration is unique to the WSE-2 — 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, to prevent bottlenecks from arising when communicating between layers.

On other devices, hardware characteristics will typically constrain applicable distribution modes. In contrast, CS-2's memory and bandwidth advantages combined with the uniformity of the cores mean that CGC can support any hybrid execution mode combining data-parallel, layer-parallel, and layer-pipelining techniques. It can run in a traditional, layer-sequential mode to support exceptionally large networks. It can combine model and layer-parallelism, with the entire model spread across the cores of the WSE-2, and each layer running pipeline-parallel. It can map multiple copies of a layer-parallelized model to the fabric at once, and train them all in a data-parallel fashion. In summary, CGC can distribute a network across the WSE-2 with complete flexibility, for maximum utilization, automatically — choosing the optimal distribution strategy to suit the computational and memory requirements of a given model.

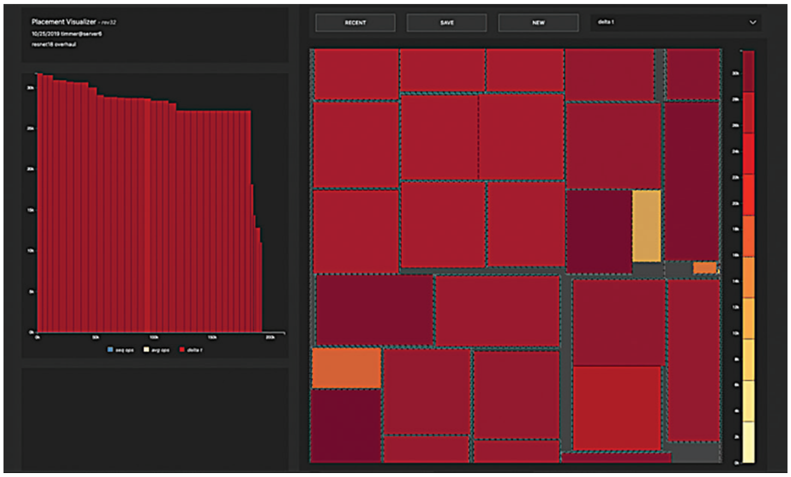
Development Tools and APIs

CGC's integrations with popular ML frameworks means that industry-standard tools such as TensorBoard are supported out of the box. In addition, Cerebras provides a comprehensive set of debugging and profiling tools to make introspection and development easy for the WSE-2.

For ML practitioners, Cerebras provides a debugging suite that allows visibility into every step of the compilation and training run. This enables visual introspection into details like:

- Validity of the compilation on the fabric
- Latency evaluations across a single kernel vs. through the entire program
- Hardware utilization on a per-kernel basis to help identify bottlenecks

For advanced developers interested in deeper customization, the Cerebras software platform includes a kernel API and C/C++ compiler based on the LLVM toolchain that allows users to program custom kernels for CGC. Combined with extensive documentation, example kernels, and best practices for kernel development, Cerebras enables users to write new kernels for their own unique research needs.



Placement Visualizer - rev32
10/25/2019 timmer@server6
resnet18 overhaul

Figure 7: Visualization tools allow researchers to introspect into each step of the CGC compilation process.

Clustering

A cluster of CS-2s enables performance scaling beyond what is possible with a much larger cluster of many small processors — at greater power and space efficiency with simpler deployment.

Scaling on many small devices today

Today, to reach new performance records or run extremely large workloads, researchers must scale-out — in other words, use large clusters of many small graphics processors. However, while it is easy to scale well across a few nodes, it is very difficult and time-consuming to scale to tens or hundreds of nodes.

To build clusters of hundreds or thousands of graphics processors requires massive investment in systems, ML, and software engineering resources. The systems engineering challenge of managing communication and synchronization overheads is exceptionally hard. The need for ever-larger batch sizes to achieve acceptable utilization using data parallel scaling is an ongoing obstacle. Researchers need to extensively retune hyperparameters and even develop new optimizers to achieve network convergence. The resultant model implementation is often brittle and must be adapted again to each new hardware setup.

> *"Integrating Cerebras technology into the Lawrence Livermore National Laboratory supercompute infrastructure enabled us to build a truly unique compute pipeline with massive computation, storage, and thanks to the Wafer Scale Engine, dedicated AI processing."*
>
> ***Bronis de Supinski***
> *Lawrence Livermore National Laboratory*

Figure 8: Clusters of CS-2s can run in both model parallel and data parallel modes.

Scaling on the CS-2 cluster

Scaling performance with multiple CS-2s is much simpler, with several important advantages beyond existing solutions that use multiple graphics processors.

A single CS-2 delivers orders of magnitude greater deep learning performance than do graphics processors. As such, far fewer CS-2 systems are needed to achieve the same effective compute as large-scale cluster deployments of traditional machines. Scaling to fewer nodes is simpler and more efficient, due to lower communication and synchronization overheads. This also means that distributed training across CS-2s achieves higher utilization without needing high batch sizes. The CS-2's custom system design allows it to sustain enormous I/O bandwidth at the system edge — 1.2Tb/s. In a cluster implementation, this translates to much larger communication bandwidth between systems to alleviate communication bottlenecks, larger than is provided by any other deep learning system today.

If a single CS-2 provides the compute performance of an entire cluster of graphics processing units, a cluster of CS-2s can replace a datacenter.

What This Means for Deep Learning Researchers

Today, deep learning researchers are constrained by hardware. It is common to choose model topologies and hyperparameters because they will speed up training, and not because they will necessarily result in the best model. There are many esoteric "do's" and "don'ts" that come into play when optimizing for graphics processing units. For example, needing to choose layer and batch sizes so all tensor dimensions are divisible by 8, to prevent significant performance degradations.

Neural architecture searches show that models built of irregular, heterogeneous blocks can often fit the data better than regular ones, given the same parameter budget. Adding sparsity to input data through sampling, to activations and to the weights of a model, has also been proposed to reduce algorithmic complexity of both training and inference jobs. Graph Convolutional Networks, with less regular and less dense structures, are promising areas of exploration. But all of these new ideas are difficult to test and leverage, in large part because they are difficult to run quickly on existing hardware.

The CS-2 unlocks these avenues of research creativity. The WSE-2 has been architected from the ground up for the neural network workload. Because each core is individually programmable, researchers have wide flexibility to explore different tensor shapes and sizes, and network and layer types. Support for sparsity has also been built directly into the hardware so that zeroes are never multiplied, and sparsity directly translates into acceleration. The CS-2 gives researchers the freedom to push the frontiers of deep learning and experiment with strange tensor shapes, irregular network structures, very sparse networks, and much more, without the performance penalties levied by existing devices.

Because the CS-2 can deliver cluster-scale compute in a single system, it also makes these fast training speeds more accessible to a much wider audience of DL researchers. With existing hardware, researchers must rely on multi-GPU and multi-node training, which require delicate system and software configuration, careful synchronization, and extensive model tuning. With CS-2, researchers do not need extensive knowledge of parallel programming techniques or experience in configuring complicated multi-node setups. CS-2 abstracts away the complexities of highly parallel execution, allowing researchers to focus on deep learning rather than on systems engineering problems.

In summary, entire classes of models and novel learning algorithms that cannot be effectively run on graphics processing units are unlocked by the CS-2's unique architecture. And with cluster-scale resources on a single chip, researchers are no longer constrained by the costs and neural network architecture paradigms imposed upon them by the graphics processing approach.

Such a multi-generational leap is only made possible by a full, end-to-end systems-driven approach to solving the problem of compute for deep learning.

Conclusion

Cerebras Systems is a team of pioneering computer architects, computer scientists, deep learning researchers, and engineers of all types who love doing fearless engineering. We have come together to build a new class of computer to accelerate artificial intelligence work.

"Its nice to have a system that trains so quickly that when you get the answer, you can still remember the question you had asked."

Rick Stevens
Argonne National Laboratory

At Cerebras, we think systems-first. This thinking is pervasive in our ethos and manifests in our designs. The CS-2 is able to achieve best-in-industry performance through innovation and technical tradeoffs across software, chip design and system hardware. All aspects of the solution work in concert to deliver unprecedented AI performance and ease-of-use.

The second generation Wafer Scale Engine, the CS-2 system, and the Cerebras software platform together comprise a complete solution to high-performance deep learning compute. Deploying the solution requires no changes to existing workflows or to datacenter operations. Cerebras solutions have been deployed in some of the largest compute environments in the world, including Argonne National Laboratory, Lawrence Livermore National Laboratory, Pittsburgh Supercomputing Center, the European Parallel Computing Centre, as well as pharmaceutical companies like GalaxoSmithKline, heavy industry, military and intelligence customers alike. Cerebras solutions are currently being used to address some of the most difficult challenges of our time — from accelerating AI in cancer research, to better understanding and treating traumatic brain injury, to furthering discovery in fundamental science around the characteristics of black holes.

With this breakthrough in performance, the Cerebras CS-2 eliminates the primary impediment to the advancement of artificial intelligence, reducing the time it takes to train models from months to minutes and from weeks to seconds, allowing researchers to be vastly more productive. In so doing the CS-2 reduces the cost of curiosity, accelerating the arrival of the new ideas and techniques that will usher forth tomorrow's AI.