The Cerebras CS-1: Achieving Industry Best Performance Through A Systems Approach

1. A Systems-first Approach to Deep Learning

1a. The Deep Learning Problem

Deep learning has emerged as the most important computational workload of our generation. In the past five years, artificial intelligence (AI) has risen from obscurity to top-of-mind. Its applications are widespread and growing. But, deep learning is profoundly computationally intensive. A recent report by OpenAI showed that, between 2012 and 2018, the compute used to train the largest models increased by 300,000x. In other words, AI compute is doubling every 3.5 months, a growth rate that is 25,000x faster than Moore's law at its peak.

This voracious demand for compute means that AI is not constrained by applications or ideas, but by the availability of compute. Testing a single new hypothesis — training a new model — takes weeks or months and can cost hundreds of thousands of dollars in compute time. This is a significant drag on the pace of innovation. Google, Facebook, and Baidu – among others – have noted that long training time is the key impediment to progress in AI, and that many great ideas are ignored simply because they would take too long to test.

1b. The Need for System-Level Thinking

To solve the problem of slow neural network training, many companies have put forward new chip architectures optimized for AI compute. Because legacy architectures were designed for graphics work, not for AI, there is substantial room to improve on chip architecture. However, to achieve breakthrough AI acceleration, architecture changes to the processor are necessary but not sufficient.

This is because putting a faster chip in a general-purpose server does not vastly accelerate a workload — it simply moves the bottleneck. Delivering performance 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. Delivering compute performance to an application is a system problem.

To accelerate training by a hundred- or thousand-fold requires a fundamental rethinking not merely of the processors, but of all aspects of the system design. This includes 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 and tuned for orders of magnitude performance gain.

Cerebras is the only company in the AI compute space to undertake the ambitious task of building a dedicated system from the ground up. The result is our CS-1. The CS-1 contains the Cerebras Wafer Scale Engine (WSE), the industry's only trillion-transistor processor. The WSE is a single-chip, 400,000 compute core processor that spans an entire 8-inch square silicon wafer. The one-wafer compute solution provides breakthroughs in power efficiency, memory bandwidth and communication bandwidth, removing the key barriers to exceptional AI performance.

The CS-1's system design is every bit as innovative as the WSE. The challenges of powering and cooling the world's largest chip are enormous. So too are the demands on I/O to feed the 400,000 AI optimized

cores. A single system that delivers the compute of hundreds or thousands of graphics processing units requires innovative solutions across all aspects of the system design.

To unleash this performance for users, a powerful, flexible software platform and familiar user workflows are critical. The Cerebras software platform has been tightly co-designed with the WSE. This allows researchers to take full advantage of its computational resources while using industry-standard machine learning (ML) frameworks like TensorFlow and PyTorch. The Cerebras software platform provides a rich tool set for users to introspect and debug, and lower-level kernel APIs for extending the platform. Further, the Cerebras software platform enables clusters of CS-1 devices to behave as if they were a single system — offering solutions to the largest of today's AI challenges, and providing a path to solving tomorrow's.

2. The Cerebras CS-1: The world's only purpose-built Deep Learning computer system

Built from the ground up to accelerate deep learning work, the CS-1 is the world's fastest AI computer. It is comprised of three sets of major technical innovations – the CS-1 system, the Wafer Scale Engine (WSE), and the Cerebras Software Platform.

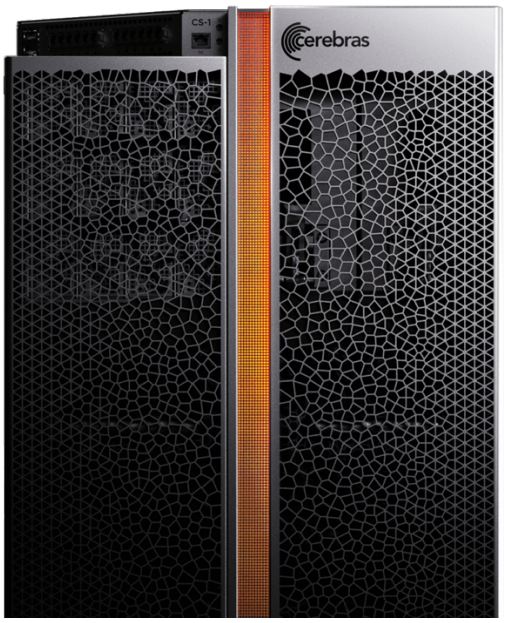

Figure 1: The CS-1, the industry's fastest AI computer

2a. The CS-1 System

The CS-1 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 massive 400,000 AI-optimized compute cores and 18 Gigabytes of high speed, on-wafer memory with 1.2 terabits per second of data.

This combination of enormous Input/Output bandwidth – 12x 100 Gigabit Ethernet lanes – and 18 Gigabytes of on-chip superfast memory, enable the CS-1 to deliver vastly more calculations per unit time

than competitive offerings. And since all the computation and communication remains on-chip – where extraordinarily power efficiency is to be had – the CS-1 communication fabric provides 33,000 times more bandwidth while using less than a tenth of the power and taking up a tenth of the space of alternative solutions.

Powering and cooling the world's largest and fastest processor chip is an exceptionally challenging undertaking. Per unit compute, the WSE uses less than one-tenth the power of graphics processing units. In aggregate, however, since the compute capacity of the WSE is so enormous, it consumes a fair bit of power. The revolutionary power delivery and cooling technology inside the CS-1 keeps the chip running at a temperature well below the operating temperature of traditional processors. In microelectronics, temperature is the enemy of reliability; the lower operating temperature enhances the reliability and performance and extends the life of the CS-1.

Finally, unlike clusters of graphics processing units, which can take weeks or months to set up, require extensive modifications to existing models, occupy dozens of datacenter racks and require complicated and proprietary InfiniBand to cluster, the CS-1 takes minutes to set up. Simply plug in the standards-based 100 Gigabit Ethernet links to a switch and you are ready to start training models at wafer-scale speed.

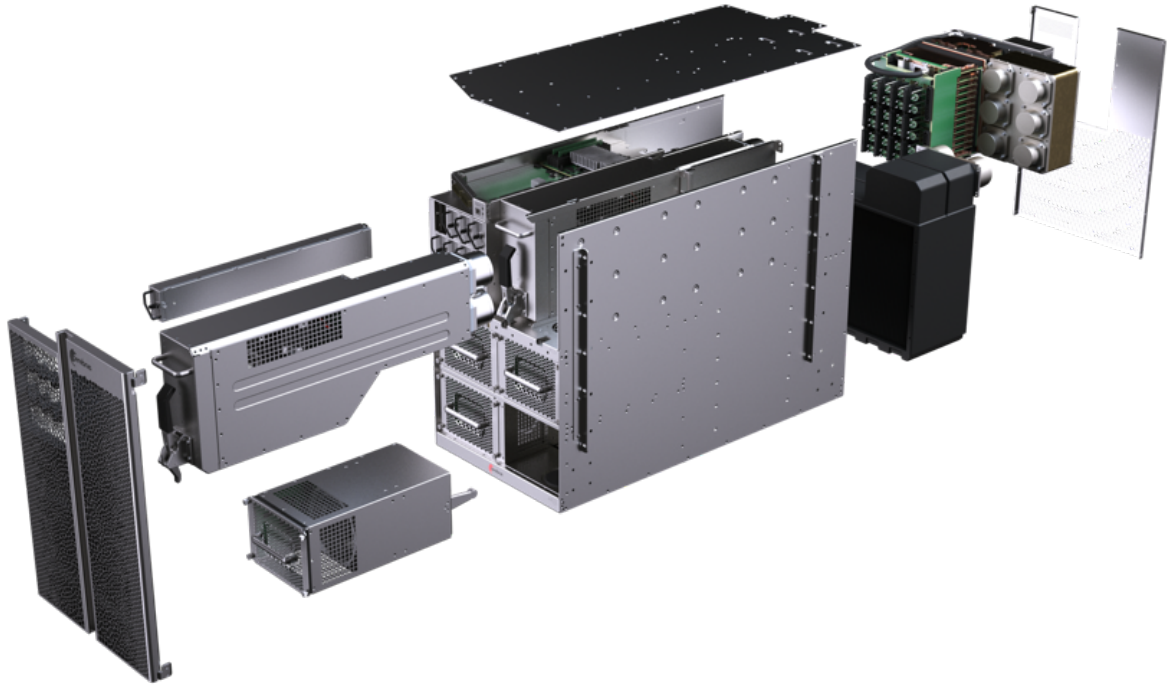

Figure 2: An inside view of the CS-1. Left to right - doors, fans, pumps, power supplies, main chassis, heat exchanger, engine block, back grill.

The CS-1 is an internally water-cooled system. Like a giant gaming PC, the CS-1 uses water to cool the WSE, and then uses air to cool the water. Water circulates through a closed loop internal to the system.

The top right of the system as shown in Figure 3 is for the movement of water. Two hot-swappable pumps move water through a manifold across the back of the WSE, 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-1. 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.

The top left of the system as shown in Figure 3 is for power and signal. Twelve power supplies, in a 6+6 redundant configuration, deliver current. The very top of the system is where 12 x 100 Gigabit Ethernet links connect the CS-1 to datacenter infrastructure.

Figure 3: Front view of the CS-1, 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.

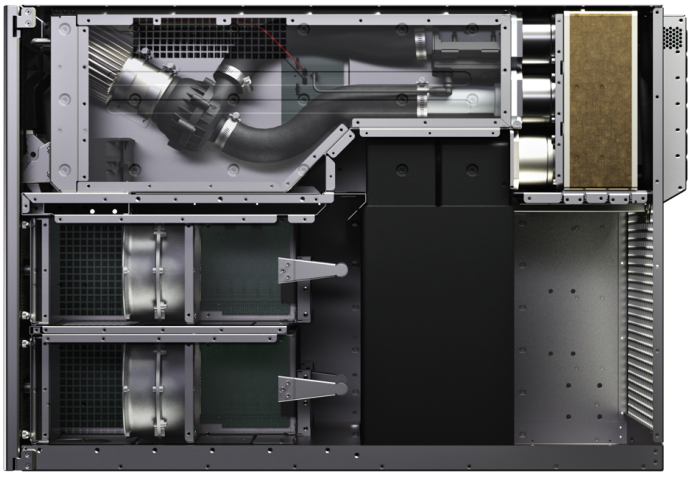

Figure 4: This side view shows the water movement assembly (top), and the air movement infrastructure - fans and a heat exchanger (bottom half)

Feeding the largest processor ever made with power and data is no easy task. The magic occurs in the back of the system, in the engine block. The engine block is an innovation in packaging that solves the challenges of power delivery, cooling, and electrical connectivity to the Wafer Scale Engine.

Here is a quick description of the WSE engine block: the front left side of the engine block contains power pins (see Figure 5). Behind these pins are power step-down modules. The top of the main motherboard can be seen peeking out behind the power delivery module. The brass block, which is the manifold, contains dry quick connectors for the water pumps (the round disks at right in Figure 5). The manifold directs water across the back of the cold plate that is coupled tightly to the wafer, cooling the 1.2 trillion transistors on the WSE.

A key innovation brings power to the wafer through the main board rather than at the edges of the wafer. The high current flux mandates that power as well as I/O arrive at the wafer face; however, the silicon wafer has a different coefficient of thermal expansion (CTE) than the main board. This means that during heating and cooling the main board and the wafer are expanding and shrinking different amounts. A custom connector was developed by Cerebras to maintain electrical connectivity in the face of these stresses.

Overcoming the technical hurdles of power delivery, cooling, packaging, CTE mismatch, with innovative solutions allowed Cerebras to solve the 70-year-old problem of wafer scale compute.

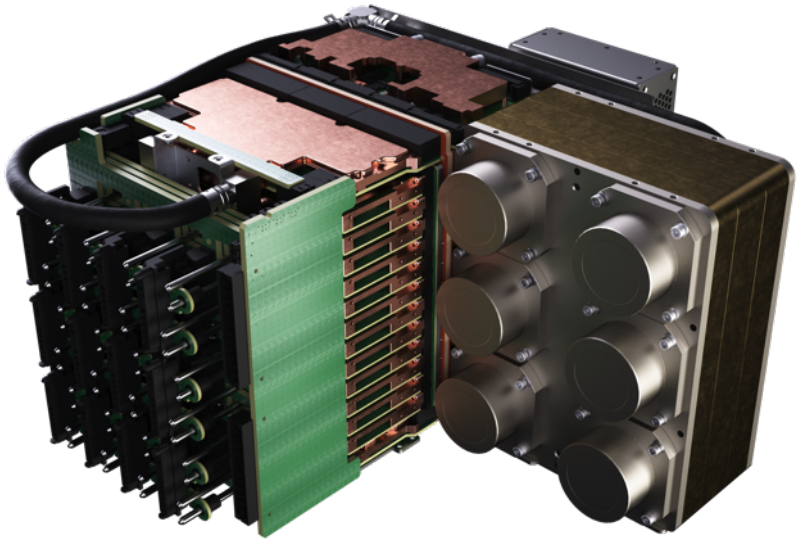

Figure 5: The engine block of the CS-1

2b. The Wafer Scale Engine

The Cerebras Wafer Scale Engine (WSE) is the revolutionary processor at the heart of the CS-1. The WSE is the largest chip ever built. It is the industry's only 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.

The WSE is fabricated on the largest square that can be cut out of the largest circular wafer available today – a 12-inch wafer. The WSE covers 46,255 square millimeters – 56 times larger than the largest graphics processing unit. In addition, with 400,000 cores, 18 Gigabytes of on-chip static random-access memory (SRAM), 9.6 petabytes/sec of memory bandwidth, and 100 petabits/sec of interconnect bandwidth, the WSE contains 78 times more compute cores, 3,000 times more high-speed on-chip memory, and has 10,000 times more memory bandwidth and 33,000 times more fabric bandwidth than its graphics processing competitor. A summary chart of the comparison is below.

	Cerebras WSE	Largest GPU	Cerebras Advantage
Chip size	46,225 mm²	815 mm²	56.7 X
Cores	400,000	5,120	78 X
On chip memory	18 Gigabytes	6 Megabytes	3,000 X
Memory bandwidth	9 Petabytes/S	900 Gigabytes/S	10,000 X
Fabric bandwidth	100 Petabits/S	300 Gigabits/S	33,000 X

Figure 6: This summary table provides an overview of the magnitude of advancement made by the WSE

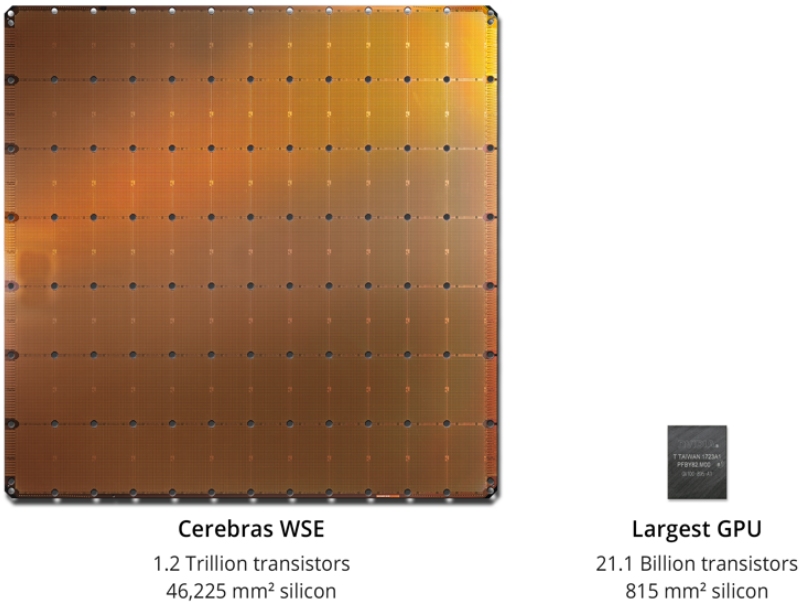

Figure 7: The Cerebras WSE and the largest Graphics Processing Unit in comparison

Computation inside the WSE happens in the 400,000 AI-optimized Sparse Linear Algebra Compute (SLAC) cores. Unlike cores in a graphics processing unit, which rely on Intel CPUs to be programmed, each SLAC core is complete, flexible, programmable, and optimized for the sparse linear algebra that underpins all neural network computation. This programmability ensures cores can run all neural network algorithms in the constantly changing machine learning field, without depending on third-party processors.

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 WSE cores include Cerebras-invented sparsity harvesting technology to accelerate computational performance on sparse workloads (workloads that contain zeros).

This is important because zeros are prevalent in deep learning calculations: often, 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 forward quickly, when most of your steps don't move you toward the finish line. 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.

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 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 Cerebras WSE has 18 Gigabytes of on-chip memory and 9.6 Petabytes/sec of memory bandwidth. As a result, the WSE keeps the entire neural network model – that is all the parameters to be learned – on the same silicon as the compute cores, where they can be accessed at full speed. This is possible because memory on the WSE is widely distributed alongside the computational elements, allowing the system to achieve extremely high memory bandwidth at single-cycle latency, with all model parameters in on-chip memory, all of the time.

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 the traditional communication techniques that are used to aggregate servers, with their graphics processing units, into clusters.

The 400,000 cores on the Cerebras WSE are connected via the Swarm communication fabric in a 2D-mesh with 100 Petabits/sec of bandwidth. Swarm provides a hardware routing engine to each of the compute cores and connects them with short wires optimized for latency and bandwidth. The resulting fabric supports single-word active messages that can be handled by the receiving cores without any software overhead. The fabric provides flexible, all-hardware communication.

Swarm is also fully configurable. Cerebras' software configures all the cores and routers on the WSE to support the precise communication required for training the user-specified model. 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 neural networks are shoehorned.

The Cerebras Wafer Scale Engine includes more cores, with 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.

2c. Cerebras Software Platform

Cerebras's 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-1 performance without changing their existing workflows. Users can define their models 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.

The Cerebras software platform is comprised of four primary elements:

1. Integration with common ML frameworks like TensorFlow and PyTorch
2. The optimized Cerebras Graph Compiler (CGC)
3. A flexible library of high-performance kernels and a kernel-development API
4. Development tools for debug, introspection, and profiling

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 and well-structured graph algorithms written in other general-purpose languages, such as C and Python, to program the CS-1.

To translate a deep learning network into an optimized executable, CGC 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. The result of this matching operation is a kernel graph.

Using its knowledge of the unique WSE architecture, CGC then allocates compute and memory to each kernel in the graph and maps each kernel onto a physical region of the computational array of cores. 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. The computation is parallel at three levels: within the core there is multiple operation per cycle parallelism; across each fabric region, the cores work in parallel on one layer; and all layers run in parallel in separate fabric regions. 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, to prevent bottlenecks from arising when communicating between layers.

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 takes into account both algorithmic efficiency and compute core utilization to generate a result that maximizes locality, minimizes communication requirements, and avoids hotspots and contention.

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

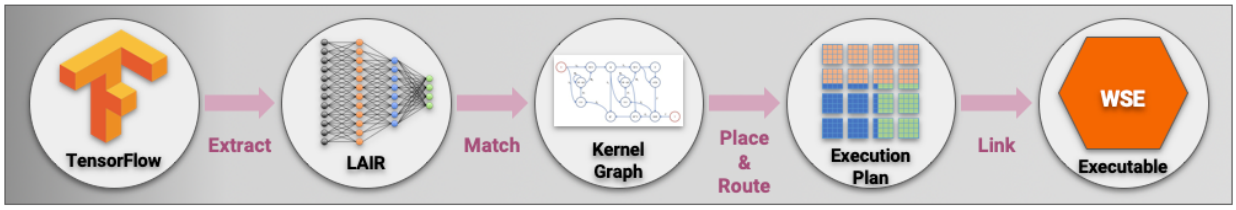

Figure 8: A high-level overview of the compilation process for the WSE

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 fully-featured set of debugging and profiling tools to make deeper introspection and development easy for the unique architecture of the WSE.

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 flexibility and 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 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.

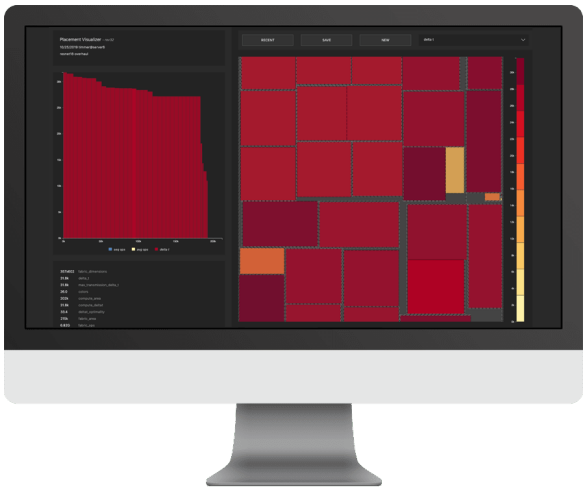

Figure 9: Visualization tools allow researchers to introspect into each step of the CGC compilation process

3. **Cluster**

A cluster of CS-1s 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

To reach new performance records or run extremely large workloads today, researchers must scale-out –in other words use large clusters of graphics processors. However, while it is easy to scale well across a few nodes, it is incredibly difficult and time consuming to scale to hundreds or thousands of nodes.

To build clusters of hundreds or thousands of graphics processing units requires massive investment in systems, ML, and software engineering resources. The systems engineering challenge of managing communication and synchronization overheads is exceptionally hard. ML and software engineering are required to research suitable model architectures and tune hyperparameters to achieve reasonable performance. The need for ever-larger batch sizes to achieve acceptable utilization with data parallel scaling approaches are an ongoing obstacle against network convergence. The resultant model implementation is often brittle, requiring extensive re-tuning and re-engineering to the data or network architecture.

Scaling on the CS-1 cluster

Scaling performance with multiple CS-1s is much easier, with several important advantages beyond existing solutions that use multiple graphics processors.

A single CS-1 delivers orders of magnitude greater deep learning performance than do graphics processors. As such, far fewer CS-1 systems are needed to achieve the same effective compute as large-scale cluster deployments of traditional machines. Scaling to fewer nodes is much simpler and more efficient, due to lower communication and synchronization overheads. This also means that distributed training across CS-1s achieves higher utilization without needing require high batch sizes. The CS-1's custom system design additionally 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-1 provides the compute performance of an entire cluster of graphics processing units, a cluster of CS-1s can replace a datacenter.

Figure 10: Clusters of CS-1s can run in both model parallel and data parallel modes

4. 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-1 unlocks these avenues of research creativity. The WSE 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-1 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-1 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-1, researchers do not need extensive knowledge in parallel programming techniques or experience in configuring complicated multi-node setups. CS-1 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-1'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.

5. 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.

At Cerebras, we think systems first. This thinking is pervasive in our ethos and manifests in our designs. The CS-1 was able to achieve best in industry performance through innovation and technical tradeoffs across software, chip and system hardware. All aspects of the solution work in concert to deliver unprecedented AI performance and ease of use.

The Wafer Scale Engine, the CS-1 System, and the Cerebras software platform - together comprise a complete solution to high performance AI compute. Deploying the solution requires no changes to existing workflows or to datacenter operations. The CS-1 has been deployed in the largest compute environments in the world including the US Department of Energy's supercomputing sites. CS-1s are currently being used to address some of the most pressing challenges of our time including to accelerate AI in cancer research, to better understand and treat traumatic brain injury, and for fundamental science around the characteristics of black holes.

With this breakthrough in performance, the Cerebras CS-1 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-1 reduces the cost of curiosity, accelerating the arrival of the new ideas and techniques that will usher forth tomorrow's AI.