Cerebras Systems: Journey to the Wafer-Scale Engine

Mandy La

Department of Computer Science, University of Chicago

CMSC 22240: Computer Architecture for Scientists

Professor Andrew Chien

March 19, 2020

Abstract

As the once rapid improvement of general purpose computers begins to slow due to the end of Moore's Law and Dennard Scaling, computer scientists must look to more specialized computing devices called accelerators. One frequent use of accelerators is in machine learning. Wafer-scaler processors are an idea that has floated around the technology industry for a while and has been attempted before. Cerebras Systems, a Los Altos, California based startup, has successfully designed and produced a wafer-scale machine learning accelerator. The benefits of condensing all the components of a processor onto a single chip is faster and more efficient performance since components are more densely packed and off-chip wiring introduces longer latency. Cerebras optimizes their Wafer-Scale Engine (WSE) by designing the chip with AI-optimized cores, flexibility in programming, and smaller but faster on-chip memory. Similar to companies that had previously attempted building wafer-scale chips, Cerebras faced numerous of engineering challenges. These challenges included yield, chip packaging, power delivery, and system cooling. The WSE seems to be a promising approach to machine learning accelerators and already has some big customers like the Argonne National Laboratory. However, Cerebras has yet to disclose any benchmark data so it is difficult to compare its performance to more traditional machine learning accelerators.

Introduction

As general purpose computers have reached the end of the monumental age of rapid and consistent improvement driven by Moore's Law and Dennard Scaling, computer scientists must look to more specialized computing devices. These types of devices are called accelerators and allow computers to achieve better performance by exploiting the structure and operations of specific computations.

Graphics processing units or GPUs are one example of an accelerator. GPUs are built with more arithmetic-logic units (ALUs) than central processing units (CPUs) in order to exploit parallelism and achieve higher operation rates. Because of this atypical design, GPUs require programmers to build explicitly parallel programs in order to take full advantage of the accelerator's benefits. In addition, GPUs are only really affective within its area of specialization. When running single thread programs, GPUs can actually perform ten times slower than a CPU thread.

Accelerators are also frequently designed and used for machine learning. Currently, machine learning accelerators consist of many GPUs wired together and attached to an external memory system. Machine learning is considered to be at the forefront of the future of computing. Machine learning and artificial intelligence open up the uses of computers from just computing to inferring. Given large amounts of data, which we have ample access to in the modern day, the computer is able to not only perform operations but use the results of those operations to inform future computations. The uses of this technology are unbounded and have been steadily creeping into our lives for over a decade. Some examples include self-driving cars, personalized ads online, virtual assistants, and facial recognition software.

Because of this great potential in uses, companies are eager to invest more time and money in pursuing machine learning. Thus computer engineers are driven to explore novel ideas that will give their machine learning product a leg up. Some approaches include field programmable gat arrays (FPGAs) that allow the customer to configure the integrated circuit after manufacturing, and application-specific integrated circuits (ASIC) which is an integrated circuit chip customized for a particular use, rather than intended for general-purpose use. As evidenced by these two examples, computer engineers are starting to develop hardware for specific uses instead of advancing the general-purpose computing technology we already have. [1]

Another approach, and the topic of this report, is building wafer-scale processors. The word wafer here refers not a sweet, thin cookie, but a wafer of thin silicon used for the fabrication of

integrated circuits. Typically, one wafer makes many microcircuits that are later separated by wafer dicing and packaged as an integrated circuit. A wafer-scale processor is one that utilizes the entire wafer to make a single large chip. The motivation for this study is to explore this approach; the benefits of condensing the components onto a single chip, and the challenges that engineers faced while making this vision a reality.

Summary and Discussion

The Wafer-Scale Approach to Machine Learning

The obvious benefit of the large size of the chip is the ability to fit an unprecedented amount of technology onto it. The Cerebras CS-1 boasts 1.2 trillion transistor and 400,000 AI optimized cores onto 46,225 square millimeters of silicon. Compare this to NVIDIA's DGX-1, a supercomputer based on an 8 GPU cluster with integrated deep learning software. DGX-1 holds 5,120 Tensor Cores and 40,960 CUDA Cores. NVIDIA's Tesla V100, used within the DGX-1, holds 21 billion transistors and 5,120 CUDA Cores. The Tesla V100 is 815 square mm. Here is a table to comparing specifications of The CS-1 and DGX-1:

	CEREBRAS WSE	**CS-1**	**DGX-1**	**TESLA V100 (SXM2)**
CHIP SIZE	46,225 sq. mm	-	-	815 sq. mm
CHIP COUNT	-	1	8	-
TRANSISTER COUNT	1.2 trillion	1.2 trillion	168 billion	21 billion
CORES	400,000 AI optimized cores	400,000 AI optimized cores	5,120 Tensor / 40,960 CUDA cores	640 Tensor / 5,120 CUDA cores
MEMORY	18 GB on-chip	18 GB	512 GB	32 GB
MEMORY BANDWIDTH	-	9.6 PB/sec	900 GB/sec	-
INTERCONNECT BANDWIDTH	-	100 PB/sec	300 GB/sec	-
TRANSISTOR SIZE	16 nm	-	-	12 nm

[2][3][4]

The new chip size brings a lot of benefits that would be hard to achieve with the traditional technologies. Such a large chip is not comparable to other chips in the market but to entire clusters. As mentioned in the table above, CS-1 contains one WSE while DGX-1 contains a cluster of eight Tesla V100 chips. The on-chip memory puts gigabytes of data within one clock cycle of its cores. The interconnection fabric is fully on-chip and connects all cores allowing it to run orders of magnitude faster than the traditional method of connecting many chips together. Additionally, one large chip allows lower power and space usage compared to multiple small chips.

The large size of the chip allows mapping of the entire neural network onto the chip at once instead of running one layer at a time. Cerebras co-designed its software to be able to apply all the compute power of the chip to a single neural network problem at once. It accepts common machine learning frameworks such as TensorFlow and PyTorch. It then extracts the neural network from the framework and performs placement routing to map the neural network layers to the fabric. In order to do this, it sizes the neural network layers based on compute, memory and bandwidth needs. Larger layers need more resources, while smaller layers require less. Mapping of the entire neural network onto the chip allows for model parallelism and linear performance scaling.

Machine learning requires a very specific type of computation. Cerebra designed CS-1 to tackle some of the key features and problem areas of deep learning. Firstly, machine learning accelerators have to work with an enormous amount of data. Accelerators would have to be working with millions to billions of samples and perform billions to trillions of operations per sample. This quickly becomes a petascale to exascale level of computing. They way NVIDIA handles this abundance of data and computations is buy packing each Tesla V100 with thousands of Compute Unified Device Architecture (CUDA) cores that allows large amounts of data to move through the GPU and efficient parallel computing. CUDA cores, however were designed for graphics processing, not deep learning. To optimize the Tesla V100 for deep learning, NVIDIA included 640 Tensor cores on top of the thousands of CUDA cores. Tensor Cores can accelerate large matrix operations, perform mixed-precision matrix multiply, and accumulate calculations in a single operation.

Cerebras addresses this issue in a similar way. But instead of only optimizing a fraction of cores for deep learning, all 400,000 cores in the CS-1 are AI optimized. Cerebras claims that in order for a core to be optimized for deep learning, it must be flexible because of how fast deep learning technology is progressing and evolving. The core must be able to adjust to these trends and remain useful as the industry moves forward. In addition, the core has to handle tensor operations efficiently and at high performance. These operations form the bulk of the compute of neural networks. In building the WSE, Cerebras included in the core a full array of general instructions with machine learning extensions, flexible general operations for control processing, and optimized tensor operations for data processing. The Cerebras core is a fully programmable core with a set of general instructions used for control processing. These instructions include arithmetic, logical, loads, stores, and branching. On top of these, Cerebras includes tensor operations that give high performance data processing. Notably, instructions can operate on 2D and 3D tensors directly as first class operands in the same way it would on registers.

Another property of deep learning that machine learning accelerators should account for is sparsity. Sparsity is the property that some of the model parameters have a value of zero. When this is the case, multiplications do not have to be performed. This is significant because multiplications comprise most of neural network computations. In addition, models can be represented in sparse matrix formats which are stored and transmitted compactly. [5] Cerebras has native sparse processing in its hardware. It uses data flow scheduling where all computes are triggered by the data. This allows the hardware to filter out all zeros and saves work and thereby saves power and improves performance. Cerebras emphasizes fine-grained execution data paths. This means small cores with independent instructions that allow maximum utilization and efficient processing of dynamic, non-uniform work. NVIDIA's tool for dealing with sparsity is their CUDA Sparse Matrix library (cuSPARSE) provided in their Deep Learning software development kit. cuSPARSE provides GPU-accelerated basic linear algebra subroutines for sparse matrices. Comparing the two approaches, Cerebras's method is certainly more integrated while NVIDA leaves the work on the software developer to utilize the cuSPARSE library. [6]

Not surprisingly, traditional memory architectures are not optimized for deep learning. Neural networks use weights and activations. The weight is simply the importance given to a neuron, and activations are how the network segregates useful and useless information. [7] In the widely used von Neumann architecture, memory is kept separate from the core. This has allowed processor technology and memory technology to develop somewhat independently according to industry needs. However, it means that the bulk of memory is slow and has high access latency. The saving grace for these memories is caches. Traditional memories rely on caches to improve performance by exploiting temporal locality. The issue for deep learning is that the fundamental operation of matrix-vector multiply has low data reuse. This makes caches less effective and potentially even lengthens memory access times as we have to check each filter for data that is unlikely there. The work-around to this issue is to transform the matrix-vector multiplication into a matrix-matrix multiplication which has high data reuse and allow the program to take advantage of the caches. However, this method changes the training dynamics. Cerebras proposes that the better answer to this problem is to house the memory on the chip. WSE has memory uniformly distributed across the cores. Each core is only 10s of microns away from memory. This layout allows the matrix-vector multiply to run at full performance.

Challenges of Wafer-Scale Technology

The idea of building large-scale chips has floated around the tech industry for a while. As with most technological ideas that stray from the norm, there are many difficulties, foreseen and

unforeseen, that may arise. Before delving into the story of Cerebras and the path they traversed to produces the world's largest chip, it is important to remember that wafer-scale chips are not a novel idea, but one that has been attempted long before. Although the wafer-scale chip was never successful until Cerebras's Wafer-Scale Engine, it would be productive to examine previous attempts and why they never made it to market. In 1980, Gene Amdahl, a big name in Silicon Valley at the time, started Trilogy Systems Corporation. At the time Trilogy was the most well-funded startup company in the history of Silicon Valley having raised $230 million from investors. Trilogy's goal was to produce cheaper and more powerful computers than IBM by building a wafer-scale chip that was two and a half inches on each side. This translates to about 4,000 square millimeters which is not as ambitious as Cerebra's 46,225 square millimeters, but still a lot bigger than the average chip at the time which was around 40 square millimeters. [8]

Unfortunately, Trilogy's attempt at the first wafer-scale chip was not successful. A series of unfortunate events befell the company. Amdahl was involved in a car accident, proceeded by a lawsuit regarding the accident. Trilogy's president, Clifford Madden, died of a brain tumor. Trilogy's semiconductor fabrication plant was damaged during construction by a winter storm. Beyond these misfortunes of life, the technologies that Trilogy was developing were riddled with flaws which I discuss later. From there, the company lost its momentum. In 1985, Amdahl decided to stop all Trilogy development and reinvest the remaining $70 million that investors had put into Trilogy. This gave Trilogy the unfortunate label of one of the largest financial failures in Silicon Valley before the dotcom bubble of 2001. The tech industry coined the term "crater" to describe companies that consumed large amounts of venture capital only to later implode, leaving nothing for its investors. [9]

Fast forward four decades, and a new, well-funded startup has brought the idea of wafer-scale chips back to life. Cerebras is a Los Altos, California based startup that has raised more than $200 million in funding from investors to pioneer the Wafer-Scale Engine (WSE) in their deep learning system, the Cerebras CS-1. Cerebras tackles a lot of the same problems that Trilogy faced. These problems included yield, chip packaging, and system cooling. Perhaps Cerebras even drew inspiration from Trilogy in their approaches to some of these challenges. Hopefully, they can learn from Trilogy's shortcomings as well.

The larger area means it is nearly impossible to yield a full wafer of cores with zero defects. To address this unavoidable problem, Trilogy created redundancy on their chips. If one component was improperly manufactured, which happens reliably often, it was simply switched out through on-chip wiring and replaced by a correctly functioning copy. They used "Triple Modular Redundancy", meaning every logic gate was triplicated. Cerebras uses almost an identical approach. They included redundant cores fabric links on the chip to replace the

defective cores and reconnect the fabric to restore the logical 2D mesh. Trilogy, however, would later find that the redundancy schemes they used were not sufficient to produce reasonable manufacturing yields. Cerebras, perhaps having learned from Trilogy's failure, was more thoughtful in designing their redundancy schemes and has not run into this problem.

Another difficulty that arose for Cerebras was the issue of effective power delivery across the large chip. They found that if they delivered power from the edge of the chip, the resistance in the interconnects could be too much and cause all voltage to be lost before it reached the middle of the chip. To solve this issue, they had to deliver power into the chip from above. So they ended up with a power-delivery system that sits about the chip with a watercooled cold plate below. Trilogy did not have this issue, possibly because their chip was not as large as Cerebras's.

Although Trilogy's chip used less power due to the efficiency of being a single chip, the dense packing made heat density a challenge. This required them to develop new cooling technologies such as sealed heat exchangers which use fluids to transfer heat. CS-1 uses up to 20 kilowatts of power to operate. This number is much greater than the maximum power any other single chip. For example the Tesla V100 has a max power consumption of 300 watts. However, it is comparable to an AI cluster like DGX-1 which has a power consumption of 3200 watts of the DGX-1. The heat density on the Cerebras chip is too high for direct air cooling. So naturally, Cerebras would have to find a way to effectively cool its chip. The way Cerebras approaches cooling is by stacking the silicon chip on top of a watercooled cold plate. The water carries heat from the wafer through the cold plate. This way the plate can cool every section of the chip evenly. Cerberas calls this technique utilizing the "Z-dimension". This large cooling system adds a lot of volume to the total CS-1 unit. The majority of the unit is taken up by the cooling mechanism. Trilogy had also decided to vertically stack computer chips. Their reasoning, however, was to allow for extremely dense packing of signal wiring.

A new problem arises with this configuration. When the chip is active, the power-delivery system, chip, and plate stacked on top of each other must warm up to the same temperature. However, as the three components warm, they expand at different rates. The power-delivery system consists of copper which expands rapidly. The silicon of the chip expands minimally. The fiberglass of the plate expands at a more moderate rate. This would be an issue in even typically sized chips. The difference in expanded sizes can be enough to shear away their connection to a printed circuit board. In extreme cases it could produce enough stress to break the chip. In a wafer-scale chip, the issue is even more concerning. Even a small change in size translates to millimeters. Trilogy failed to account for this problem, and their chip interconnect technology had layers that often delaminated. There was no automated way for Trilogy to repair soldering errors. Although Cerebras was prepared to deal with this issue, the solution was not an

easy one to find. The only way to keep the power-delivery posts connected was to find a material that had the right intermediate coefficient of thermal expansion that would sit neatly between the coefficients of the silicon and fiberglass. The engineers at Cerebras was never able to find one so they had to invent one instead. This took them about a year and a half to do.

Analysis and Discussion

Cerebras has been quite a success so far. It received an abundance of media coverage when it came out of "stealth mode" in the summer of 2019, and was presented at Hot Chips, a conference on high-performance microprocessors, in August of 2019. The warm response was likely due to the tech industry's desire for innovative ways to propel computers forward. There is a sort of lack in advancement that consumers are starting to notice now that Moore's Law and Dennard Scaling has ended. Cerebras provides a new and exciting method of advancement.

Cerebras has already had a number of customers purchase the CS-1 which without doubt comes with a very high price tag; so high that they have not even publically disclosed it. Their most notable customer so far is the U.S. Department of Energy's Argonne National Laboratory. In a November 2019 press release, the laboratory speaks very highly of the capabilities of Cerebras's CS-1. They claim "the CS-1 delivers record-breaking performance and scale to AI compute, and its deployment across national laboratories enables the largest supercomputer sites in the world to achieve 100- to 1,000-fold improvement over existing AI accelerators." The CS-1 has not released any benchmark testing data so there is no evidence of its superior performance. Argonne has not only invested a lot of money into the CS-1, but also time. According to Rick Stevens, Argonne Associate Laboratory Director for Computing, Environment and Life Sciences, "We've partnered with Cerebras for more than two years and are extremely pleased to have brought the new AI system to Argonne." The CS-1's first application area is cancer drug response prediction. The Department of Energy is anticipating deploying the CS-1 at the Lawrence Livermore National Laboratory to "accelerate its AI initiatives and further enhance its simulation strengths with the machine learning capabilities of the CS-1." [10]

This sign of interest from the government department and such reputable labs is a promising sign of that the CS-1 can deliver on its claims of superior performance. However, it is difficult to compare the actual performance of the chip because Cerebras has not released benchmark test results. They refuse to even publicly state the clock rate of the processor.

The building of the wafer-scale chip was certainly a feat of ingenuity. The company was able to address all these issues and come out on the other side with a beautiful and enormous chip. However, there are some concerns that Cerebras has yet to even acknowledge. With the grand ambition of squeezing all components of a machine learning accelerate, Cerebras opted to house all the memory on-chip. This makes memory accesses lightning fast, but severely restricts the amount of memory within CS-1. Cerebras advertises 18 gigabytes of on-chip memory and 400,000 cores. Each core is directly connected to its own portion of memory. This means that the amount of memory per core would be:

$$\frac{18\ \text{GB}}{400{,}000\ \text{cores}} = 0.000045\ \text{GB per core} = 45\ \text{KB per core}$$

This is alarmingly small, and comparable to an average L1 cache. For example, Intel's Skylake Processor holds a 64 Kilobyte L1 cache per core; including both data and instruction caches. [11] Upon hearing this comparison to the L1 cache of other processors, one might wonder if Cerebras could use the on-chip memory as a sort of cache and hook the chip up to a large external memory that would serve as the main memory. But, as mentioned previously, deep learning programs typically have a low data reuse rate. Reuse rate is what caches rely on and exploit. They bring frequently used data closer to the processor in order to decrease access latency times to those parts of the memory. To make matters worse, Cerebra engineers realize the only way to input memory through the chip is to have it enter from the edges. This makes connecting the CS-1 to an external memory even more unreasonable.

Given such a small memory size, this then becomes a matter of batch size. The batch size of a machine learning algorithm is a hyperparameter that defines the number of samples to work through before updating the internal model parameters. The size of the memory on the WSE restricts the batch size that can be used. There is extensive debate in the machine learning community on whether smaller or larger batch sizes produce better results. Stochastic gradient descent is an iterative method for optimizing an objective function with suitable smoothness properties. Small batch sizes limit parallelization of stochastic gradient descent in deep learning. Large batch sizes do not have this problem. However, a larger batch size increases computational cost and decreases performance. [12] Typically, engineers use a batch size of 64, 128, or 256. Whether this batch would fit into the 18 gigabytes of memory offered by CS-1 depends on the intrinsic size of the sample data. Thus, in this report, we will not reach a conclusion on whether the on-chip memory of WSE is sufficient, but simply remark on the restriction it places on batch size.

Summary and Learnings

For decades the goal of the computer architects was to continue miniaturizing technology. This was due to the benefits realized by Dennard's scaling which meant smaller was better in every way. That is until engineers began to approach the physical limitations of how small a transistor could be. The approach with wafer-scale technology is to think bigger. The components of the chip stay small, but condensing them onto a large chip allows for higher density and better performance.

Overall, the process of researching and writing this report was very interesting. I learned a lot about very specific aspects of machine learning. I would like to spend more time learning about machine learning so that I can fully grasp some of the concepts required to fully understand the changes Cerebras is bringing to the field. I was able to get an understanding of things like sparsity and weights and activations. However, I wish I could have learned more about matrix-vector to matrix-matrix multiply mini batch.

A limitation, and mild annoyance, to my research process was the amount of information Cerebras has left undisclosed. Cerebras unveiled the CS-1 late last year so there is little data on how it performs in real situations. Andrew Feldman, chief executive of Cerebras Systems, insists that its focus is on real customer trials and reviews. Feldman has no interest in industry benchmarks such as MLPerf, the most widely cited measure of computer chip performance on AI. [13] Perhaps in the next year or so, the performance of Cerebras's CS-1 system will speak for itself and the machine learning industry will know a lot more about weather this is a viable direction for machine learning chips. Until then, we are left with only the biased claims of superior performance made by Cerebras, and hope for a promising future for machine learning.

References

[10] (2019, November 19). Retrieved from https://www.anl.gov/article/argonne-national-laboratory-deploys-cerebras-cs1-the-worlds-fastest-artificial-intelligence-computer

[9] Berg, E. N. (1984, July 8). CAN TROUBLED TRILOGY FULFILL ITS DREAM? *The New York Times*. Retrieved from https://www.nytimes.com/1984/07/08/business/can-troubled-trilogy-fulfill-its-dream.html

[4] Cerebras. (n.d.). *Cs-1 Datasheet*. Retrieved from https://secureservercdn.net/198.12.145.239/a7b.fcb.myftpupload.com/wp-content/uploads/2019/11/CS-1-Datasheet.pdf?time=1574170537

[5] Gale, T., Elsen, E., & Hooker, S. (n.d.). *The State of Sparsity in Deep Neural Networks*. Retrieved from https://arxiv.org/pdf/1902.09574.pdf

[12] Goceri, Evgin & Gooya, Ali. (2018). On The Importance of Batch Size for Deep Learning.

[7] ML Glossary. (n.d.). Retrieved from https://ml-cheatsheet.readthedocs.io/en/latest/nn_concepts.html#weights

[6] NVIDIA. (n.d.). *Nvidia Deep Learning Sdk*. Retrieved from https://docs.nvidia.com/deeplearning/sdk/

[2] NVIDIA. (n.d.). *Nvidia Dgx-1 The Essential Instrument for Ai Research*. Retrieved from https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/dgx-1/dgx-1-rhel-datasheet-nvidia-us-808336-r3-web.pdf

[3] NVIDIA. (n.d.). *Nvidia Tesla V100 Gpu Accelerator*. Retrieved from https://images.nvidia.com/content/technologies/volta/pdf/tesla-volta-v100-datasheet-letter-fnl-web.pdf

[13] Ray, T. (2019, November 19). Cerebras did not spend one minute working on MLPerf, says CEO. Retrieved from https://www.zdnet.com/article/cerebras-did-not-spend-one-minute-working-on-mlperf-says-ceo/

[1] Reuther, A., Michaleas, P., Jones, M., Gadepally, V., Samsi, S., & Jeremy, J. (2019). *Survey and Benchmarking of Machine Learning Accelerators*. MIT Lincoln Laboratory Supercomputing Center. Retrieved from https://arxiv.org/pdf/1908.11348.pdf

[11] Skylake Processors - HECC Knowledge Base. (2019, May 1). Retrieved from https://www.nas.nasa.gov/hecc/support/kb/skylake-processors_550.html

[8] Trilogy Systems. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Trilogy_Systems