Cerebras Systems Enables Brain-scale AI

WAFER-SCALE ENGINE INVENTOR LAYS THE FOUNDATIONS FOR RUNNING MASSIVE AIS

INTRODUCTION

This research paper explores Cerebras System's approach to create a brain-scale AI and the new technologies that could enable that feat. But first, let's put this discussion into the proper context. Just how big is a 120 trillion-parameter model?

Many technology prognosticators have predicted it would take at least until 2045 for the industry to create Artificial Intelligence (AI) technology that could rival the human brain, as measured by the number of synapses in the human brain or parameters in an AI. Note that we are not talking about Terminator-style AI, or "General AI" here. We are talking about an AI that can process a single though complex task, such as natural language processing. The average human cerebral cortex has about 80-100 billion neurons and 120 trillion synapses. For the sake of argument, let's assume the parameters in an AI model roughly equates to synapses. The largest AI ever trained is the GPT3 natural language model from OpenAI.org, at 175 billion parameters, or roughly 1/1000th the size of a brain. So, 120 trillion is enormous, approximately 1000 times larger than today's state-of-the-art.

It is important to note that AI model size expansion will certainly not stop at 100 trillion parameters, if and only if hardware is invented that can train such massive models at a reasonable cost. It is estimated that Open.AI spent some $12M to train GPT-3. And development of GPT-4 is already well underway. Bringing the cost down will depend on more efficient processing and through techniques such as sparsity harvesting and dynamic network pruning. And the research into expanding the scale of networks is producing tantalizing results, as seen already with GPT-3, so we are just beginning to comprehend the scope of what is possible.

HOW DOES CEREBRAS CREATE BRAIN-SCALE AI?

Andrew Feldman co-founded Cerebras Systems in 2016 with a vision to create the industry's largest and fastest AI technology. The company introduced the second-generation Wafer-Scale Engine (WSE-2) in 2021, which is 56 times larger than the largest GPU, has 123 times more compute cores, and 1000 times more high-performance on-chip memory. Packaged in a CS-2 accelerator, which takes up about 1/3 of the data center rack, this platform boasts 850,000 AI-optimized cores. That certainly sounds big. But, if you want to go huge and train multi-trillion parameter models, you have three more fundamental problems to solve in addition to having a high-performance computational platform:

1. You need to have a lot more memory to hold the model's weights.
2. You need to scale efficiently to many compute nodes to reduce training time.
3. You need to be smart about executing the training; attacking this problem with brute force simply won't work.

The latest technology announcements from Cerebras made at the annual Hot Chips 33 conference in August 2021 lay the foundations to solve these challenges. If Cerebras can deliver on the 120 trillion-parameter promise, the world will enter a new era. Today, it can take a team of scientists months to train GPT-3 on a massive cluster of state-of-the-art GPUs. Imagine doing so in just a day

or being able to train a trillion-parameter model in just a few days. Now imagine running models that are three orders of magnitude more complex. Doing so will transform medicine, chemistry, astronomy, physics, material science, and perhaps human life and society.

This all begs a big question. Do we seriously need trillion-parameter models? And if so, when? Over the last two years, we have seen model size increase by 1000-fold, requiring 1000 times more computation power. If we keep up that pace, and there is no reason to believe we will not, we will need multi-trillion parameters in less than two years.

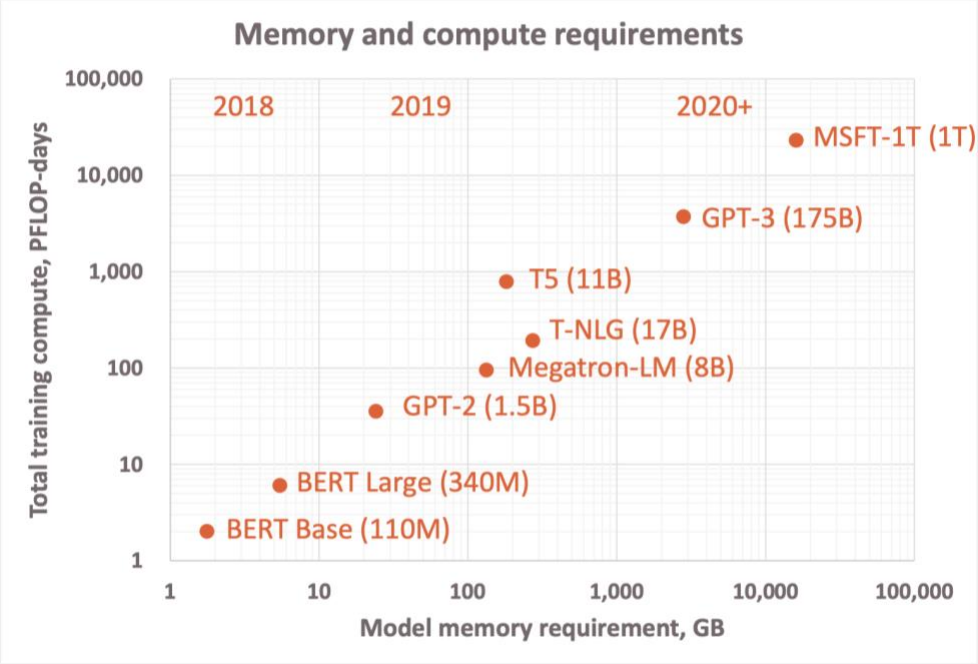

Figure 1: AI Model sizes have been doubling every 3.5 months, according to Open.AI. Source: Cerebras Systems

MEMORY

Ok, so how much more memory could one need for a trillion-parameter scale model and how could a CS-2 access it efficiently? Models like GPT-3, with "only" 175 billion weights require over 2 TB of memory, but a multi-trillion parameter model requires petabytes of memory to hold the weights. So Cerebras created the MemoryX technology with up to 2.4 PB of DRAM and Flash memory. Think of it as a high-speed data store that also performs data transformation operations to feed the CS-2 more efficiently, reducing processing demands on the server. As we will cover later, the MemoryX technology also facilitates weight sparsity – a key technique to reduce the computational intensity required.

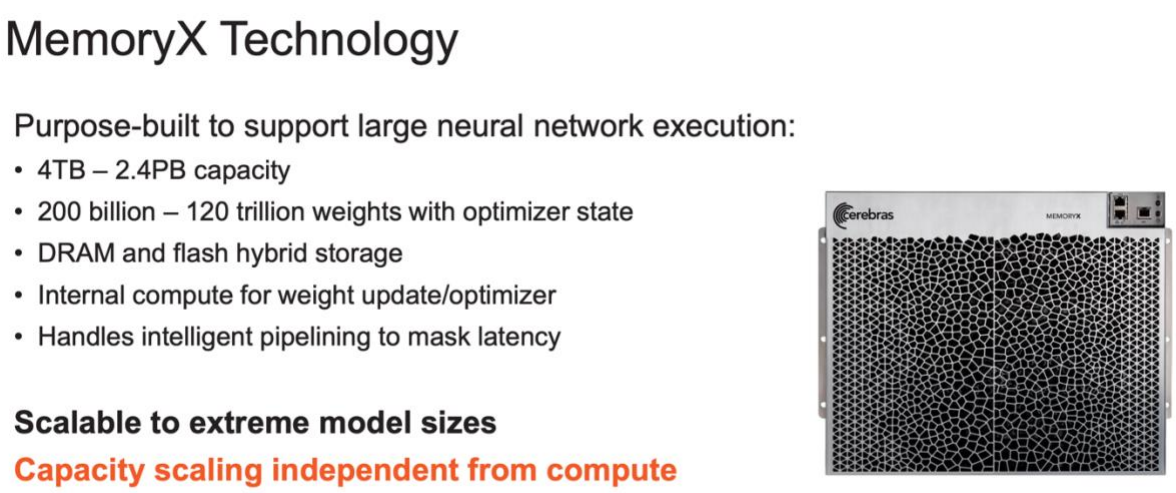

Figure 2: The MemoryX platform streams weights to a CS-2 system or an entire cluster of CS-2 systems. Source: Cerebras Systems

SCALABILITY

Cerebras solved the memory limitations for large models with MemoryX and the scalability problem with a new fabric interconnect called SwarmX. Since each CS-2 system can hold any layer of a multi-trillion parameter model, there is no need for model parallelism. With this approach, all compute parallelism is data parallelism, with each CS-2 in a network containing the entire layer, not the entire model. The execution model remains the same regardless of the size of the cluster, greatly simplifying the programming experience, broadcasting only weights and reducing gradients across the fabric.

The SwarmX fabric uses a binary tree topology. The fabric uses 12 x 100 Gb optical Ethernet connections per CS-2 system for its Layer 1 protocol. So, there is plenty of available bandwidth at 1.2 Tb per server. Cerebras plans the SwarmX fabric to scale up to 192 CS-2 systems.

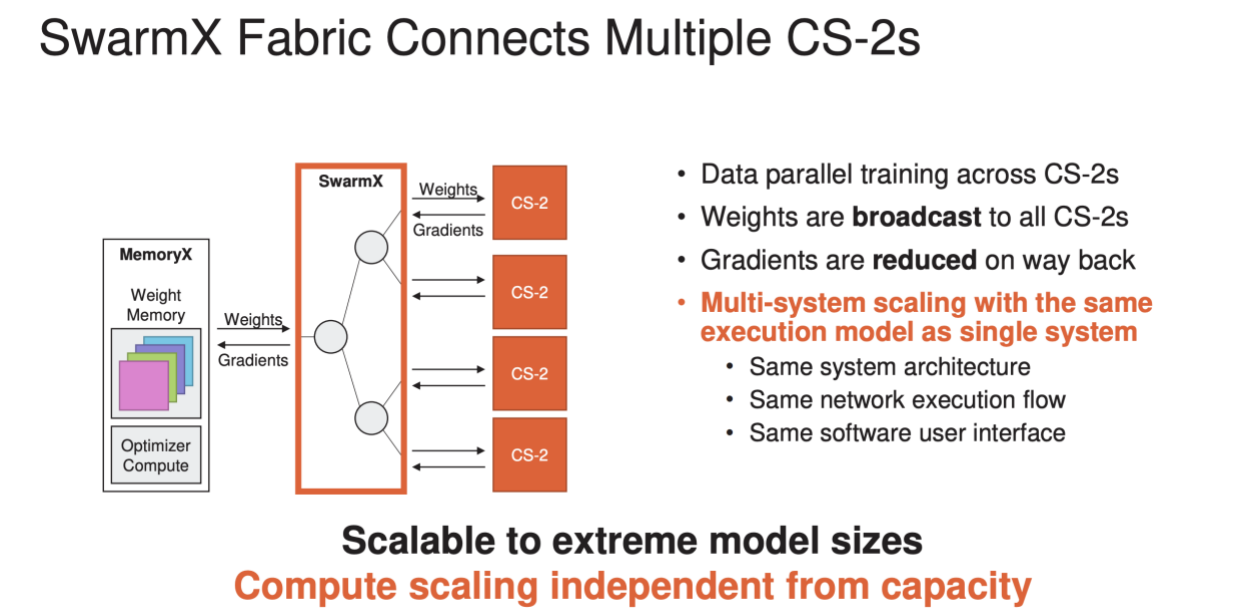

Figure 3: The SwarmX fabric enables scaling to 192 CS-2 systems. Source: Cerebras Systems

Cerebras also shared projected performance data to demonstrate the linear scalability of the CS-2 + SwarmX + MemoryX architecture. While this is only a projection at this point, the hardware will likely scale as advertised. We look forward to actual measurements that corroborate the simulations.

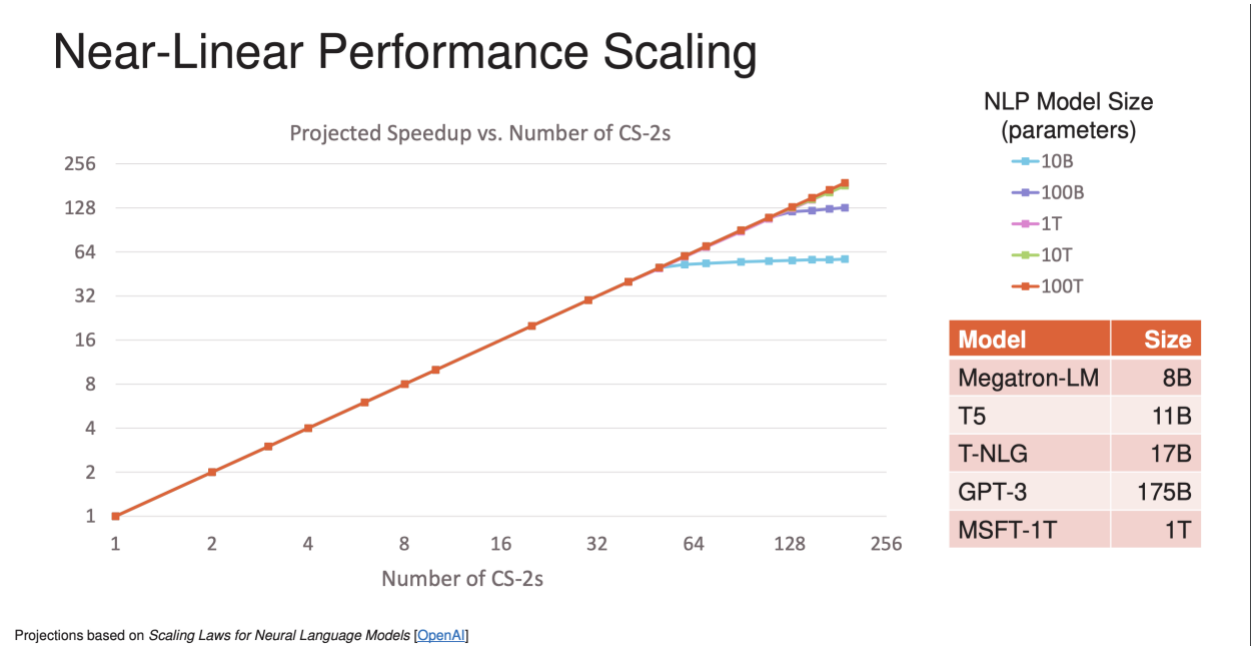

Figure 4: Projected scalability of the Cerebras data center predicts near-linear scalability up to at least 256 CS-2 systems. Source: Cerebras

WORKING SMARTER, NOT JUST HARDER

Cerebras, like most other AI organizations, accepts that brute-force computational density alone cannot solve large AI problems. Doing so runs afoul of Amdahl's law of diminishing returns of high levels of scalability. The team at Cerebras has two tricks up its sleeve in this regard. First, Cerebras has rethought the compute/memory model to streamline execution. Second, Cerebras enables sparsity to reduce the amount of computation that is required. If a number is zero, you do not need to perform a multiplication to produce another zero. Let's look at these two optimizations built on MemoryX and the Cerebras dataflow cores themselves, respectively.

THE WEIGHT STREAMING EXECUTION MODEL

In this execution model, the weights are stored on the MemoryX, and the activations that are applied to the weights for the current mini-batch reside on the CS-2 system's on-die memory. Weights stream into the CS-2, where the activations for the current mini-batch of data are computed. When the CS-2 system completes a forward and backward pass, it transfers the weight gradients back to the MemoryX, which updates the weights; the CS-2 system is not burdened by this compute load. As we saw above, using this memory / compute disaggregation, a cluster of 192 CS-2 systems could train a multi-trillion parameter model with near-linear performance scaling, according to Cerebras. While the company has not yet published benchmark results and prediction accuracy, the company is running these models today in-house.

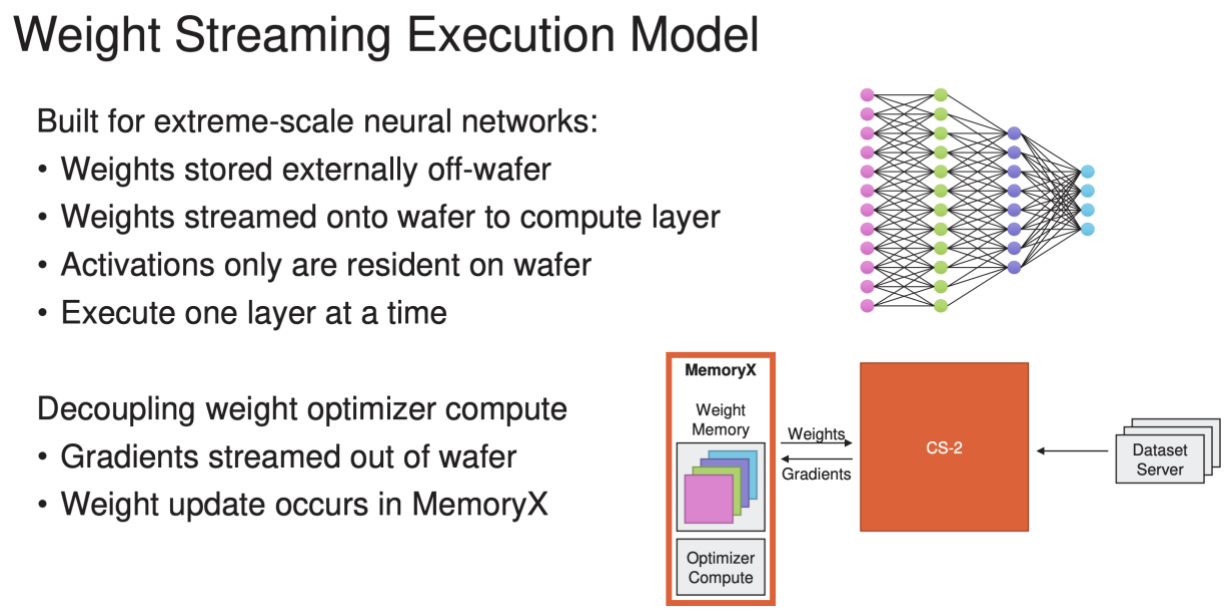

Figure 5: The Cerebras Weight Streaming Execution Model enables the CS-2 to train huge models. Source: Cerebras Systems

SPARSITY

It takes a tremendous amount of investment, innovation, and energy to create a 10-fold performance improvement in silicon alone. Relying on sparsity and harvesting sparsity is another potential path to

acceleration, but it has proven only modestly successful in practice. This is partly due to higher training costs and complexity, requiring those seeking a sparse model to train a dense network before pruning, plus performing additional fine-tuning to preserve model prediction accuracy. And sparsity has barely emerged out of research to date, limited by today's available hardware.

However, Cerebras designed the cores themselves for sparse linear algebra. Each core triggers a multiply only if both factors are non-zero. This allows Cerebras to harvest sparsity efficiently. If an ML model can be trained in a sparse model, it can be trained in less time. The advantage is that Cerebras can handle non-structured sparsity, which has been shown to be a better way to reduce model size while preserving accuracy. When the sparsity processing complements this ability to handle zeros at the core in the MemoryX, Cerebras can take all sources of sparsity, whether the zeros stem from ReLu, rounding, activations, or weights. The system can also handle dynamic sparsity, where a parameter that has no effect through dozens of epochs can be pruned out. To our knowledge, Cerebras is the only hardware that can deal with all these sources and forms of sparsity induction. Existing hardware to date has never been able to accelerate unstructured or dynamic sparsity.

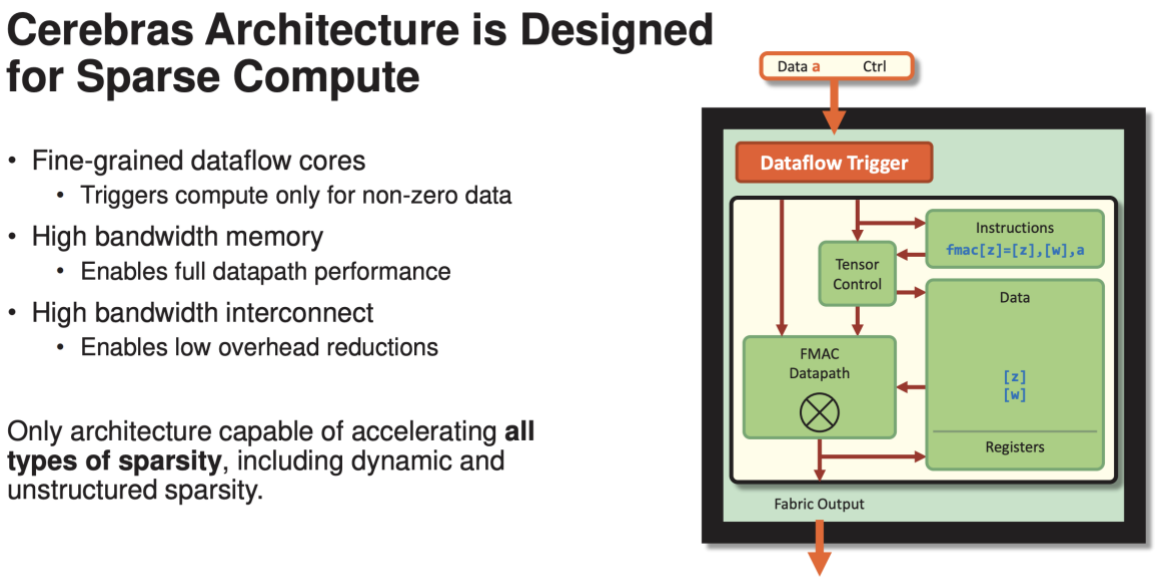

Figure 6: Cerebras designed sparsity awareness into each core with a trigger that skips the MAC data path in case a zero is detected. Source: Cerebras Systems

In addition to the native fine-grained sparsity triggers in the cores, the MemoryX system induces sparsity in the weights, and performs the weight updates when it receives the new gradients from the CS-2 servers via SwarmX. Sparsity is induced in the weight matrix within the MemoryX unit using any ML sparsity algorithm, and then the sparse weights are streamed to all CS-2s. This means that the SwarmX interconnect broadcasts weights in sparse form already. When the CS-2 receives them, it performs a sparse computation. On the delta pass, the CS-2 produces gradients, also sparse – computing gradients only for the non-zero weights. And those sparse gradients are streamed back out, reduced through the SwarmX interconnect, and back to the MemoryX unit. Finally, the weight

update is performed, all sparse as well, of course! All of this happens natively, with no change to the weight streaming model. In fact, it's exactly the same Weight Streaming flow Cerebras uses for dense matrices.

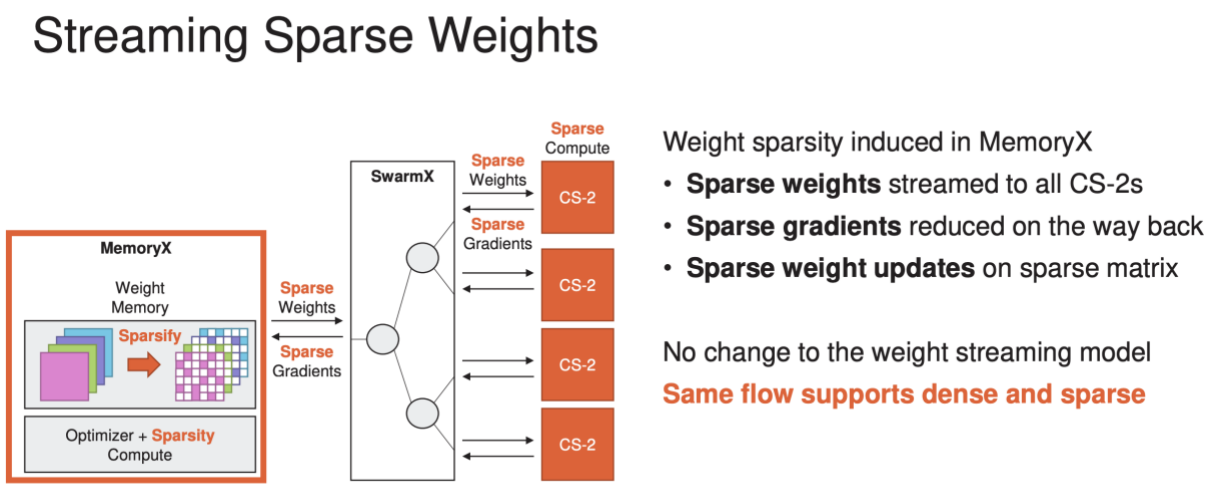

Figure 7: This image shows how a four CS-2 system works, with MemoryX providing sparse weights through SwarmX to the WSE servers, which return the gradients to the memory server to update the weights. Source: Cerebras Systems

DO WE REALLY NEED SUCH MASSIVE AI'S? DO WE WANT THEM?

Yes, we need them to solve the most difficult challenges facing science today, from cancer research to molecular dynamics and material science. However, let's be clear that we are not implying the ability of a brain-scale AI built on Cerebras or any other hardware technology to be used as a general artificial intelligence, which is defined as the hypothetical ability of an intelligent agent or computer system to learn or understand any intellectual task that a human being can. Weak or narrow AI as it exists today is not intended to have human-like cognitive abilities and personality. Narrow AI is limited to study or accomplish specific pre-learned problem-solving or reasoning tasks. However, transitioning from brain-scale computing to General AI may be shifting from a compute-bound challenge to a software and ethical one. In this arena, Google Deep Mind and Open.ai are two organizations that continue to conduct research.

CONCLUSIONS

Cerebras Systems has again demonstrated the company's unique approach to AI systems and software, developing a platform that does not resemble anything else in the competitive landscape, and that can handle AI that is beyond the reach of would-be competitors. Cerebras provides a comprehensive acceleration system, whose design, memory, and I/O are up to the task of the very largest AI models in development. Taking a different path than all other AI startups and major players alike should position Cerebras uniquely well as AI continues to grow. It is one thing to say

"my chip is faster than yours". It is quite another to be able to say "I can solve problems no one else can."

IMPORTANT INFORMATION ABOUT THIS PAPER

AUTHOR: Karl Freund, Founder and Principal Analyst at Cambrian-AI Research

INQUIRIES:

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in context, displaying the author's name, author's title, and "Cambrian-AI Research." Non-press and non-analysts must receive prior written permission from Cambrian-AI Research for any citations.

LICENSING

DISCLOSURES

This document was developed with Cerebras funding and support. Although the document may utilize publicly available material from various vendors, including Cerebras, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Cambrian-AI Research provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.