Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

News & Highlights

Can Giant Microchips Become a Big Deal?

Mitch Leslie

Senior Technology Writer

For decades, manufacturers have boasted about how small they can make microchip components. Transistors have shrunk by about 1000-fold over the last 50 years, for example [1]. But Cerebras Systems, Inc. of Sunnyvale, CA, USA, takes pride in how big its chips are. Produced from a single silicon wafer, its Wafer-Scale Engine (WSE)-2 chips measure 46 225 mm^2, 56 times the size of a standard Nvidia microprocessor (Fig. 1) [2].

Companies have tried to make giant chips before—and failed dismally [3]. Cerebras chose this unconventional approach to meet the exploding demand for computing power, driven largely by artificial intelligence (AI) [4]. The company claims that the WSE-2—on which memory and computing cores, or processing units, are near each other to improve speed and efficiency—is hundreds of times faster than the standard graphics processing units (GPUs) that power most computers running AI models [5]. Unlike its unsuccessful predecessors, Cerebras is making sales. In 2019, the United States Department of Energy's Argonne National Laboratory in Lemont, IL, USA, became the first customer for the company's Cerebras system (CS)-1 AI accelerator, which is built around a single, equally large forerunner of the WSE-2 [6]. The lab has since added the CS-2, a newer accelerator that relies on the WSE-2 [7]. In 2023, Cerebras connected 64 of its CS-2 units to create—allegedly in ten days—an AI supercomputer for the Abu Dhabi-based tech company G42 of the United Arab Emirates (Fig. 2) [4,8]. The company has announced plans to build eight more such supercomputers for G42 [4].

Cerebras is the first company to offer a product with wafer-scale integration (WSI), in which complete circuits reside on a single silicon wafer [9]. However, researchers are working on other designs that can take advantage of WSI's benefits [10,11]. "You want to consolidate as much computing as possible on one chip," said Rakesh Kumar, professor of electrical and computer engineering at the University of Illinois Urbana-Champaign, IL, USA, who described one of these designs with his colleagues in a 2021 paper [10].

It is not clear whether other oversized chips will go into production or whether this format can grab a bigger share of the semiconductor market. The chips and their accompanying components are expensive and face stiff competition from more conventional approaches for ramping up AI performance [12]. "Cerebras is getting some traction," said Naveen Verma, professor of electrical and computer engineering at Princeton University (Princeton, NJ, USA). But in the long term, "it is hard to see the computation space going in that direction," said Verma, who co-founded a company, EnCharge AI, based in Santa Clara, CA, USA, that will soon start selling a standard-sized microchip for accelerating AI.

Most computer chips are manufactured in the same way. Multiple integrated circuits are printed onto a single wafer of silicon that is typically 300 mm in diameter [13]. The individual chips are then cut out and go through further processing before they are ready for use. Although conventional chips vary in size depending on their functions, they are all much smaller than the WSE-2. Nvidia's A100, an AI workhorse and the largest GPU on the market, spans 826 mm^2 [2].

Whether chips are running a cell phone or an AI supercomputer, they must communicate with each other. Processing chips typically have little storage and need to summon data from separate memory chips, for example. Signals between chips may need to travel distances of several centimeters. "Communication is very expensive in terms of energy, reduced bandwidth, and increased latency," said Kumar. These limitations hamper the performance of AI systems, which need to analyze enormous amounts of data rapidly as they train for and then perform particular applications [14]. Kumar said that uniting processing and memory on a single chip can reduce the energy use per bit, shrink latency, and boost bandwidth—and thereby accelerate AI results.

That is the rationale for the approach taken by Cerebras. The company's manufacturing process starts out roughly the same as for conventional chips, with several microprocessors etched onto a silicon wafer. But instead of cutting up the wafer to free individual chips, the company links each of these units, known as tiles, with a proprietary conducting material called interconnect [15]. The company says the benefits of locating processing and memory units on the same chip are substantial, claiming that the bandwidth of its interconnect is 45 000 times higher than the bandwidth of the wires that link conventional GPUs [16]. Cerebras released its first giant chip, the WSE-1, in 2019 [17] and premiered the WSE-2, which contains more than twice as many computing cores, in 2021 [2,3].

The company needed some unorthodox engineering to supply power to its behemoth chips and keep them cool. Each WSE-2 requires a large amount of electricity, about 20 kW. Standard chips plug into a printed circuit board (PCB) with a power supply

https://doi.org/10.1016/j.eng.2024.04.005

Please cite this article as: M. Leslie, Can Giant Microchips Become a Big Deal?, Engineering, https://doi.org/10.1016/j.eng.2024.04.005

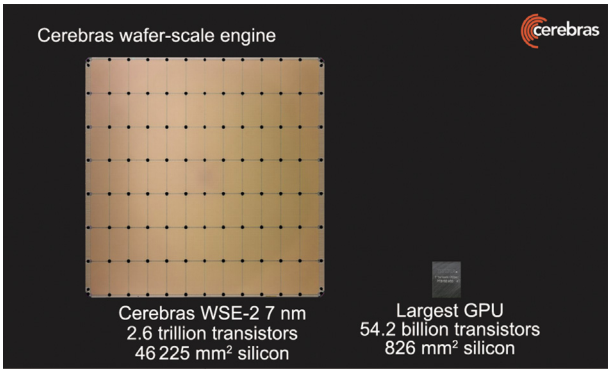

Fig. 1. Cerebras' enormous WSE-2 chip (left) dwarfs the largest graphics processing unit (GPU) on the market. The WSE-2 features 850 000 cores and 40 GB of memory. According to the WSE-2 product specifications, its conducting fabric can carry data at 220 Pbit per second. Credit: Cerebras (public domain).

Fig. 2. In 2023 Cerebras built an AI supercomputer for the company G42 by linking 64 of its CS-2 AI accelerator units, each built around a single WSE-2 chip. According to the company, its 54 million cores enable the supercomputer to reach a computing speed of 4×10^{18} floating operations per second. Credit: Cerebras (public domain).

connected to the edge. Electricity for the chips moves across the PCB, but it would have to travel so far in a massive chip that some tiles would receive less power—or none at all—and some might overheat [15,18,19]. In Cerebras' chips, each tile receives electricity through the PCB, evening out the power distribution [15,18]. Temperature control for such a large chip is also a challenge, but Cerebras uses a modular water-cooling system. Each tile has its own circulation loop that brings in cool water and removes hot water that has absorbed the tile's heat [15,18].

The challenges of building large chips with WSI have defeated several companies over the last 50 years. Texas Instruments, based in Dallas, TX, USA, failed at the task in the 1960s [3]. In the 1980s, Trilogy Systems collected more than 200 million USD from investors but never delivered a working chip, becoming a cautionary tale in Silicon Valley [3].

What doomed Trilogy Systems was that it could not solve the so-called yield problem, said Puneet Gupta, professor of electrical and computer engineering at the University of California, Los Angeles, USA. A certain number of microchips have manufacturing defects—although companies will not disclose how many of their chips are duds. Manufacturers can throw out or recycle the flawed conventional chips, which are relatively cheap to replace. Like its smaller counterparts, an oversized chip with one or more defective processors may not work. But because such chips cost much more to manufacture than standard chips, tossing one out results in a bigger financial hit. For various reasons, Trilogy Systems could not get its yield high enough [3].

Cerebras says it has found a solution. After the chips are etched, they go through testing that identifies faulty cores. The manufacturer then reconfigures the interconnect to route communication around the bad core or cores, allowing the chip to function [15]. To make this step easier, the company includes spare cores on its chips, a strategy it claims can theoretically result in a 100% yield [18].

Although Cerebras has not disclosed its prices, industry observers have estimated the each of its CS-2 AI accelerator units could cost as much as 3 million USD [20]. The giant chips are expensive to make, and thus are only being used by purchasers "willing to pay a premium," said Kumar. Another downside, said Gupta, is that the processors and memory on the chips are homogeneous because of how they are manufactured. The chips cannot incorporate the same diversity of functions, such as different types of memory, as can conventional circuits. "That is very limiting," said Gupta.

Gupta, Kumar, and their colleagues have developed an alternative approach that involves chiplets, or small, specialized chips [10,21]. Their design would include 2048 memory and compute/processing chiplets plugged into a single silicon wafer with an embedded copper interconnect for communication [10]. The completed circuit would measure 15 100 mm^2, about one-third the size of Cerebras' chip. The researchers have built a small prototype with eight chiplets and continue to work on their design, but there are no plans to put it into production, said Kumar. "It is a proof of principle," said Gupta, to show that WSI is possible with a different approach than the one taken by Cerebras.

Cerebras may not face immediate competition from chiplets, but a slew of companies is developing competing solutions to speed AI. Verma and colleagues, for example, have devised a standard-sized chip that reduces the delays and energy losses that result from storing memory separately. They used a strategy known as in-memory computation, in which data is housed in random access memory (RAM) on processing chips that need it [22]. Verma said EnCharge AI, the company he co-founded, has produced working chips and will soon begin marketing its products.

The long-term prospects for supersized chips are uncertain. So far, Cerebras is the only company selling them, and whether other companies will introduce their own versions remains unclear. However, the motivation for producing chips with WSI remains in place, said Gupta. "There is a huge demand for large amounts of compute connected to large amounts of memory with a high-performance interconnect."

References

[1] Moore SK. The node is nonsense. IEEE Spectrum 2020;57(8):24–30.

[2] Moore SK. Cerebras' new monster AI chip adds 1.4 trillion transistors [Internet]. New York City: IEEE Spectrum; 2021 Apr 20 [cited 2024 Jan 31]. Available from: https://spectrum.ieee.org/cerebras-giant-ai-chip-now-has-a-trillions-more-transistors.

[3] Linder C. This is the world's largest computer chip [Internet]. New York City: Popular Mechanics; 2019 Aug 27 [cited 2024 Jan 31]. Available from: https://www.popularmechanics.com/technology/design/a28816626/worlds-largest-computer-chip/.

[4] Lu Y. Supersizing computers and chips. The New York Times 2023 Jul 21;Sect B:1.

[5] High-performance computing [Internet]. Sunnyvale: Cerebras Systems, Inc.; [cited 2024 Jan 31]. Available from: https://www.cerebras.net/applications/high-performance-computing/.

[6] Moore SK. Cerebras unveils first installation of its AI supercomputer at Argonne National Labs [Internet]. New York City: IEEE Spectrum; 2019 Nov 19 [cited 2024 Jan 31]. Available from: https://spectrum.ieee.org/cerebras-unveils-ai-supercomputer-argonne-national-lab-first-installation.

[7] ALCF Cerebras CS-2 system available to users [Internet]. Argonne: Argonne National Laboratory; 2023 Apr 28 [cited 2024 Feb 12]. Available from: https://www.alcf.anl.gov/support-center/facility-updates/alcf-cerebras-cs-2-system-available-users-0.

[8] Wang J. Introducing Condor Galaxy 1: a 4 exaFLOPS supercomputer for generative AI [Internet]. Sunnyvale: Cerebras Systems, Inc.; 2023 Jul 20 [cited 2024 Feb 12]. Available from: https://www.cerebras.net/blog/introducing-condor-galaxy-1-a-4-exaflop-supercomputer-for-generative-ai/.

[9] Wafer scale integration [Internet]. New York City: PCMag; [cited 2024 Jan 31]. Available from: https://www.pcmag.com/encyclopedia/term/wafer-scale-integration.

[10] Pal S, Liu J, Alam I, Cebry N, Suhail H, Bu S, et al. Designing a 2048-chiplet, 14336-core waferscale processor. In: Proceedings of 2021 58th ACM/IEEE

Design Automation Conference; 2021 Dec 5–9; San Francisco, CA, USA. Piscataway: IEEE; 2021. p. 1183–8.

[11] Zhang T. 'Big Chip': China is building a wafer-sized processor to beat US sanctions on supercomputers and AI [Internet]. Hong Kong: South China Morning Post; 2024 Jan 16 [cited 2024 Jan 31]. Available from: https://www.scmp.com/news/china/science/article/3248176/big-chip-china-building-wafer-sized-processor-beat-us-sanctions-supercomputers-and-ai.

[12] Castelvecchi D. 'Mind-blowing' IBM chip speeds up AI. Nature 2023;623(7985):17.

[13] Li A, Timings J. 6 crucial steps in semiconductor manufacturing [Internet]. Veldhoven: ASML; 2023 Oct 4 [cited 2024 Jan 31]. Available from: https://www.asml.com/en/news/stories/2021/semiconductor-manufacturing-process-steps.

[14] Hutson M. The world's largest computer chip [Internet]. New York City: The New Yorker; 2021 Aug 20 [cited 2024 Jan 31]. Available from: https://www.newyorker.com/tech/annals-of-technology/the-worlds-largest-computer-chip.

[15] Ugnius. Wafer-scale processors: the time has come [Internet]. Sunnyvale: Cerebras Systems, Inc.; 2019 Sep 6 [cited 2024 Jan 31]. Available from: https://www.cerebras.net/blog/wafer-scale-processors-the-time-has-come/.

[16] Wafer-scale engine: the largest chip ever built [Internet]. Sunnyvale: Cerebras Systems, Inc.; 2021.

[17] Woo M. The world's biggest computer chip. Engineering 2020;6(1):6–7.

[18] McGregor J. AI startup Cerebras develops the most powerful processor in the world [Internet]. New York City: Forbes; 2019 Aug 20 [cited 2024 Jan 31]. Available from: https://www.forbes.com/sites/tiriasresearch/2019/08/20/ai-start-up-cerebras-develops-the-most-powerful-processor-in-the-world/.

[19] Moore SK. Huge chip smashes deep learning's speed barrier. IEEE Spectrum 2020;57(1):24–7.

[20] Cutress I. Cerebras unveils wafer scale engine two (WSE2): 2.6 trillion transistors, 100% yield [Internet]. Los Angeles: AnandTech; 2021 Apr 20 [cited 2024 Feb 14]. Available from: https://www.anandtech.com/show/16626/cerebras-unveils-wafer-scale-engine-two-wse2-26-trillion-transistors-100-yield.

[21] Orcutt M. Chiplets: 10 breakthrough technologies 2024 [Internet]. Cambridge: MIT Technology Review; 2024 Jan 8 [cited 2024 Jan 31]. Available from: https://www.technologyreview.com/2024/01/08/1085120/chiplets-moores-law-avanced-micro-devices-intel-chips-breakthrough-technologies/.

[22] Wiggers K. EnCharge raises $22.6M to commercialize its AI-accelerating chips [Internet]. San Francisco: TechCrunch; 2023 Dec 5 [cited 2024 Jan 31]. Available from: https://techcrunch.com/2023/12/05/encharge-raises-22-6m-to-commercialize-its-ai-accerating-chips/.