Engineering 6 (2020) 6–7

Contents lists available at ScienceDirect

Engineering

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

News & Highlights

The World's Biggest Computer Chip

Marcus Woo

Senior Technology Writer

When it comes to computer chips, size matters. Smaller has always meant better. According to Moore's law, the number of transistors that can squeeze onto a silicon chip doubles every two years, enabling sleeker and more powerful devices.

Now, a California-based company is upending that notion. In August 2019, the Silicon Valley, Los Altos-headquartered Cerebras introduced the largest chip ever produced. The chip, named the Wafer Scale Engine (WSE), is made out of an entire silicon wafer. Roughly the size of an Apple iPad, it measures 46 225 mm^2, an area about 56 times bigger than the next largest chip, NVIDIA'S GV100 Volta graphics processing unit (GPU), which spans 815 mm^2 (Fig. 1). While the NVIDIA GPU contains 21.1 billion transistors, the WSE holds 1.2 trillion [1,2].

"The chip is clearly an integration marvel," said Rakesh Kumar, associate professor of electrical and computer engineering at the University of Illinois at Urbana Champaign. "It is a big deal to have shown it possible to put together a working chip this large."

Such a mammoth chip, according to Cerebras, is needed to meet the growing demands of artificial intelligence (AI). AI algorithms learn to do a task by first training on a huge amount of data. In particular, deep learning algorithms, which use neural networks that roughly mimic how the brain works, require enormous computing power, with training runs that can take hours or even days. According to a recent analysis from OpenAI, a San Francisco-based, AI-focused company backed by Microsoft, the computing power demanded by AI training has, from 2012 to 2018, increased by a factor of 300 000, with a doubling time of 3.5 months. That is 25 000 times faster than Moore's law at its peak [3].

Handling all that computation requires more cores than are on a single standard chip. So multiple chips must work together in concert. But that also means data must be shuttled between chips—a process tens of thousands of times slower than possible within a single chip [1].

Dozens or even hundreds of small chips can be fabricated out of a silicon wafer. But by producing a single integrated chip from one entire wafer, Cerebras has built a processor packed with cores that does not rely on the off-chip communication that can bog down conventional systems. The WSE also holds memory cores close to the compute cores, so the former can constantly feed data to the latter, reducing the idling time of the compute cores. Boasting an architecture designed for machine learning, the chip is optimized for AI training, according to Cerebras. It has 400 000 programmable cores, 18 GB of static random access memory (SRAM), and a memory bandwidth of 9 $PB \cdot s^{-1}$. This is 78 times more cores, 3000 times more on-chip memory, and 10 000 times more memory bandwidth than the next best chip [1,4].

The WSE is a notable achievement, said Mike Demler, a senior analyst with the Linley Group, an analyst firm in Mountain View, California, that focuses on the microprocessor industry. "This has been tried in the past and it never succeeded." In 1980, for example, chip engineer Gene Amdahl founded Trilogy with $230 million USD in funding—the highest at the time—to build a wafer-scale chip [5]. The company, however, did not succeed and folded after only five years [6].

One of the reasons these early attempts failed, Demler said, is that the chip fabrication process produced too many defects. When you fabricate many chips from a wafer, you simply discard the defective chips. That does not work with a chip made from the entire wafer.

While fabrication today is much improved, defects are still inevitable. Cerebras gets around the problem with spare cores and an architecture that bypasses any defects. During manufacturing, any bad cores are identified, and the interconnects are rerouted around the bad cores to the spare ones [7].

In addition to dealing with the defects, building such a large chip poses other challenges, including with cooling and power delivery. Heat causes silicon to expand differently than the

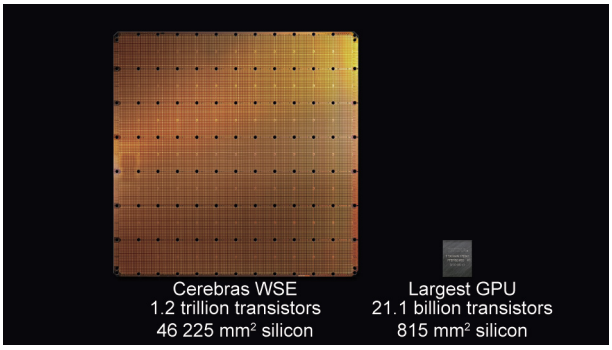

Fig. 1. With the August 2019 introduction of its WSE, the Los Altos, California, company Cerebras claims the distinction of producing the world's largest computer chip, built from a single silicon wafer. The WSE was engineered to accommodate the accelerating demands of artificial intelligence computing. Credit: Cerebras (public domain).

https://doi.org/10.1016/j.eng.2019.11.001

materials in the printed circuit board that connects to the chip. So, according to the company, Cerebras had to create a new proprietary material that can connect the circuit board and the chip while absorbing the thermal stress. A wafer-scale chip also demands 15 kW of power—compared to 250 W for the biggest GPUs—requiring an innovative design to deliver power directly to the middle of the wafer, Kumar said. Bringing power lines through the periphery would be too inefficient and bulky. To cool the wafer uniformly, water flows onto a cold plate attached to the chip [7]. "A lot went into building this thing," Demler said.

The company has not yet publicized the price of their new system, but already has its first customer. In September, Cerebras and the US Department of Energy announced a multi-year partnership to boost deep-learning research at Argonne National Laboratory in Lemont, Illinois, near Chicago, and Lawrence Livermore National Laboratory in California [8]. The WSE is "an ideal instrument to accelerate the US Department of Energy's numerous deep learning experiments," said Rick Stevens, Associate Laboratory Director for Computing, Environment and Life Sciences at Argonne National Laboratory.

Still, there might be limitations with such a large chip, Kumar said. "In general, the larger the chip, the lower the yield," he said. Because the power supply and cooling are specialized, the chip may only be useful for a small number of customers. The fact that the WSE is such an integrated system means you cannot incorporate other kinds of technologies. "This limits the memory capacity of this chip, for example, which in turn can limit the applications it is useful for," Kumar said.

The true test of the WSE will be its real-world performance and how it compares with other systems. "It is an impressive engineering accomplishment," Demler said. "But now they must demonstrate that the software works and provides a real advantage in the end application."

References

[1] Feldman A. Cerebras Wafer Scale Engine: why we need big chips for deep learning [Internet]. Los Altos: Cerebras; 2018 Aug 28 [cited 2019 Oct 15]. Available from: https://www.cerebras.net/cerebras-wafer-scale-engine-why-we-need-big-chips-for-deep-learning/.

[2] NVIDIA Tesla V100 GPU architecture—the world's most advanced data center GPU [Internet]. Santa Clara: NVIDIA; 2017 Aug [cited 2019 Oct 15]. Available from: https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf.

[3] Amodei D, Hernandez D. AI and compute [Internet]. San Francisco: OpenAI; 2018 May 16 [cited 2019 Oct 15]. Available from: https://openai.com/blog/ai-and-compute/.

[4] Cerebras company website [Internet]. Los Altos: Cerebras; [cited 2019 Oct 15]. Available from: https://www.cerebras.net/technology/.

[5] Linder C. This is the world's largest computer chip [Internet]. New York: Popular Mechanics; 2019 Aug 27 [cited 2019 Oct 15]. Available from: https://www.popularmechanics.com/technology/design/a28816626/worlds-largest-computer-chip/.

[6] Metz C. To power AI, start-up creates giant computer chip [Internet]. New York: The New York Times; 2019 Aug 19 [cited 2019 Oct 15]. Available from: https://www.nytimes.com/2019/08/19/technology/artificial-intelligence-chip-cerebras.html.

[7] Schreiber R. Wafer-scale processors: the time has come [Internet]. Los Altos: Cerebras; 2019 Sep 6 [cited 2019 Oct 15]. Available from: https://www.cerebras.net/wafer-scale-processors-the-time-has-come/.

[8] Department of Energy and Cerebras Systems Partner to accelerate science with supercomputer scale artificial intelligence. Business Wire; 2019 Sep 17 [cited 2019 Oct 15]. Available from: https://www.businesswire.com/news/home/20190917005356/en/Department-Energy-Cerebras-Systems-Partner-Accelerate-Science.