

News & Highlights

世界上最大的计算机芯片

Marcus Woo

Senior Technology Writer

对于计算机芯片来说，尺寸至关重要。芯片尺寸越小越好。根据Moore定律，集成在硅片上的晶体管数量每两年翻一番，从而使器件更轻薄以及功能更强大。

目前，加利福尼亚州的一家公司正在颠覆这一观念。2019年8月，总部位于美国硅谷洛斯阿图斯的Cerebras公司推出了有史以来世界上尺寸最大的芯片。这个芯片被命名为Wafer Scale Engine (WSE)，它是由整个硅片制成。它的面积达到了46 255 mm²，相当于一个Apple iPad的大小。尺寸第二大的芯片是由NVIDIA公司研制的GV100 Volta图形处理器 (GPU)，其面积为815 mm²(图1)。WSE的面积比GPU的面积大56倍以上。NVIDIA的GPU包含了211亿个晶体管，而WSE包含了1.2万亿个晶体管[1,2]。

伊利诺伊大学厄巴纳-香槟分校电子和计算机工程学院的副教授Rakesh Kumar表示，该芯片显然是一个集

成奇迹。能把如此大的一块芯片组装在一起是一件了不起的事情。

根据Cerebras公司的说法，研制这种芯片是为了适应人工智能 (AI) 日益增长的需求。AI算法首先通过处理大量数据来学习执行一项任务。尤其像深度学习算法，该算法利用神经网络来模拟大脑的工作方式，它需要巨大的计算能力，因此训练可能要花费数小时甚至数天。根据总部位于美国旧金山、专注于AI研发的OpenAI公司 (该公司已得到了微软公司的支持) 最近的一项分析，2012—2018年，AI训练所需的计算能力增长了30万倍，倍增时间延长至3个半月。这比Moore定律的最高值快了25 000倍[3]。

处理所有的这些计算需要比单个标准芯片更多的内核。因此，多个芯片必须协同工作。但是，这也意味着数据必须能在多个芯片之间进行传输。然而，数据在这个传输过程中所花费的时间要比在单个芯片中传输所花费的时间长1万倍以上[1]。

一个硅片可以被制成数十个甚至数百个小芯片。但是，Cerebras公司通过将完整的晶片制成一个集成芯片，设计了一个内核处理器，该处理器不依赖于芯片外通信，因为芯片外通信通常会干扰常规系统。WSE使存储内核更靠近计算内核，因此前者可以不断向后者提供数据，从而减少了计算内核的空转时间。据Cerebras公司介绍，该芯片拥有专为机器学习设计的架构，它优化了芯片的AI训练。该芯片具有40万个可编程内核、18 GB静态随机存取存储器 (SRAM) 和每秒9 PB的存储带宽。与GPU相比，该芯片内核数量增加

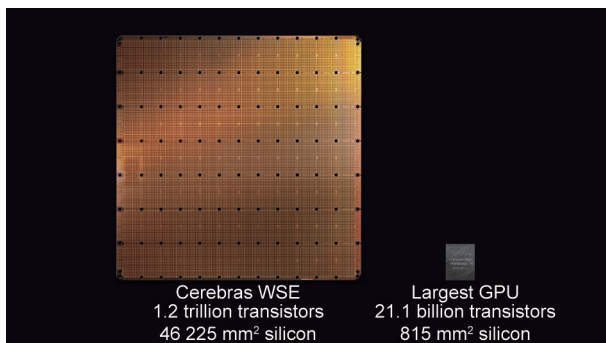


图1. 随着2019年8月WSE的推出，总部位于美国加利福尼亚州洛斯阿图斯的Cerebras公司宣布生产出世界上最大的计算机芯片，且该计算机芯片由单晶硅制成。WSE旨在适应人工智能计算速度不断增长的需求。图片来源：Cerebras (公有领域)。

了78倍、片上内存增加了3000倍以及内存带宽增加了1万倍[1,4]。

位于美国加利福尼亚州山景城的Linley Group是一家专注于微处理器产业的高级分析公司。Linley Group的一位资深分析师Mike Demler表示，WSE是一项卓越的成就。过去已有类似的尝试，但从未成功过。例如，在1980年，芯片工程师Gene Amdahl用2.3亿美元（当时是最高的投资）创立了Trilogy公司，用于制造晶圆级芯片[5]。然而，该公司没有成功，5年后就倒闭了[6]。

Demler表示，这些早期尝试失败的原因之一是芯片在制造过程中会产生很多的缺陷。当你把一个晶片制成多个芯片时，你只需丢弃有缺陷的芯片。但是这对由整个晶片制成的单个芯片并不起作用。

尽管现在的制造技术有了很大的改进，但缺陷仍然不可避免。Cerebras公司利用备用内核和完美的架构解决了这个问题。在制造过程中，任何有缺陷的芯片都会被识别出来，并且互连线会将有缺陷的内核路由到备用内核[7]。

除了需要解决这些缺陷，制造如此大的芯片还需要克服一些技术障碍，如散热和功率传输。热量导致硅的膨胀和与芯片连接的印刷电路板中材料的膨胀是不同的。因此，Cerebras公司必须设计出一种新的材料，这种材料在吸收热应力的同时还可以将电路板和芯片连接起来。Kumar说，单晶片级芯片需要15 kW的功率，而最大的GPU才需要250 W。因此，我们需要创新设计，使得功率能被直接传递到晶片中间。然而，将电源线直接穿过晶片外围的做法是低效且笨拙的。为了均匀地冷却晶片，水流会经过附着在芯片上的冷却板[7]。

Demler表示，制造这个单芯片需要做出很大的努力。

该公司尚未公布新系统的价格，但该系统已经有了它的第一位客户。2019年9月，Cerebras公司与美国能源部（US Department of Energy）宣布了一项长期合作计划，以促进美国伊利诺伊州芝加哥附近的莱蒙特市阿贡国家实验室（Argonne National Laboratory）和加

利福尼亚州的劳伦斯·利弗莫尔国家实验室（Lawrence Livermore National Laboratory）之间的深度学习研究[8]。阿贡实验室的计算、环境与生命科学实验室副主任Rick Stevens表示，WSE是促进美国能源部众多深度学习实验的理想工具。

据Kumar表示，如此大的芯片可能仍然存在局限性。一般来说，芯片尺寸越大，产出率就越低。因为电源和冷却系统都是专用的，所以该芯片可能仅适用于少量客户。并且，WSE是一个集成系统，它不能再与其他类型的技术合并。这限制了该芯片的存储容量，进而限制了其应用程序的适用性。

对WSE来说，真正考验它的是其在实际应用中心能以及与其他系统比较时的表现。Demler表示，WSE是一项杰出的工程成就，但是我们现在必须要证明WSE的有效性和它在最终应用程序中所显现出的真正优势。

References

- [1] Feldman A. Cerebras Wafer Scale Engine: why we need big chips for deep learning [Internet]. Los Altos, CA: 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, CA: 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, CA: 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, CA: 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>.