IEEE
micro
The magazine for chip and silicon systems designers

VOLUME 41, NUMBER 6 NOVEMBER/DECEMBER 2021

Microprocessor at 50.

The Path to Successful Wafer-Scale Integration: The Cerebras Story

IEEE
75 YEARS
IEEE COMPUTER SOCIETY
www.computer.org/micro

The Path to Successful Wafer-Scale Integration: The Cerebras Story

Gary Lauterbach, *Cerebras Systems, Sunnyvale, CA, 94085, USA*

SCALING: TECHNOLOGY, CLOCK, ARCHITECTURE, AND DIE SIZE

There has been an impressive increase in single-chip processing power since the Intel 4004 was launched in 1971. This is usually attributed to Moore's law, but there are additional factors to consider. In understanding the components of prior improvements, we can gain insight into the potential for future improvements and potential limits to scaling.

We begin by considering how several of the most significant design and performance metrics have changed over the last 50 years. In Table 1, we list fabrication technology node, clock frequency, die size, transistor count, and operations per second performance using the Intel 4004 and the Nvidia A100 as representative end points.

The growth in operations-per-second and transistor count is particularly impressive. If the manufacturing technology geometries of these two designs were measured using the same features, we would expect the transistor count ratio to be very near the square of the technology ratio times the die size ratio. In fact, it is quite close despite the fact that the feature measurements differ and the transistor usage (SRAM versus logic) is not equivalent. The simple scaling results in a surprisingly close transistor ratio of 141 million rather than 24 million, surprisingly close. Even more impressive is the performance ratio: 13 orders of magnitude increase in the last years. Truly spectacular!

Of additional interest is how these components contribute to the performance of the micro-processor. The "architecture factor" component of performance is derived by dividing operations per second by the clock frequency times the transistor count: the amount of op/sec each transistor is producing per cycle. The relative contributions to performance over the last five decades can now be ranked as shown in Figure 1.

The Intel 4004 performed add operations serially on 4-bit binary coded decimal (BCD) nibble search using ten instructions and requiring eight clock cycles per instruction resulting in 640 clock cycles for a 32-bit BCD addition. Unsurprisingly, the first improvements in microprocessor architecture provided the greatest benefit: parallel adders and pipelined instruction execution. These two features alone could have made the 4004 640× faster, based on the number of cycles required for an addition to be completed, albeit with a significant increase in the number of transistors. While this may make it seem like the 365× architecture factor is entirely composed of parallel adder and pipelining benefits this is far from the case. The dramatic increase in clock rate demanded architectural improvements, such as cache memory systems, to enable performance to increase with clock rate without encountering the "von Neumann bottleneck."

HISTORY OF MICROPROCESSOR DIE SIZE

Die size improvements have not kept pace with other advancements. In the remainder of this article, we explore why this is the case, and how the industry and we at Cerebras Systems have worked to break historical die size barriers.

Widening the adder datapath and reducing overhead in instruction execution were the first architectural steps taken in the Intel 8008 and 8080 chips. The 8080 was fabricated using 6-μm technology, rather than the 10-μm technology used to make the 4004, yet an increase from 12- to 20-mm^2 die area was required to implement the 8080 beyond that provided by Moore's law geometry reduction. This trend in die size increases, which provided more silicon area for the required transistor count beyond that delivered by the technology shrink, continued until recently when die sizes have reached an asymptote. Figure 2 shows the trend and some of the notable microprocessors on the die size increase curve.

The maximum reticle size has not increased significantly since 2018, for the reasons discussed below. The cadence of geometry shrinks is also slowing down. The compound result of these two effects is a dramatic slowing of the increase in the number of transistors available to implement a processor. Fortuitously, many traditional

0272-1732 © 2021 IEEE
Digital Object Identifier 10.1109/MM.2021.3112025
Date of current version 19 November 2021.

TABLE 1. Microprocessor scaling components.

	Manufacturing Technology node (μm)	Clock frequency	Op/Sec[1]	Die size (mm)^2	Transistor count	Architecture factor	Launch Date
Intel 4004	10	740 kHz	1176	12	2250	7×10^{-7}	1971
Nvidia A100	0.007	1.4 GHz	19.5×10^{15}	826	54 billion	2.6×10^{-4}	2021
Ratio	1429	1892	1.66×10^{13}	68.9	24 million	365	

[1]For comparison, Op is defined as a 32-bit BCD addition for the 4004 and a 32-bit integer add for the A100.

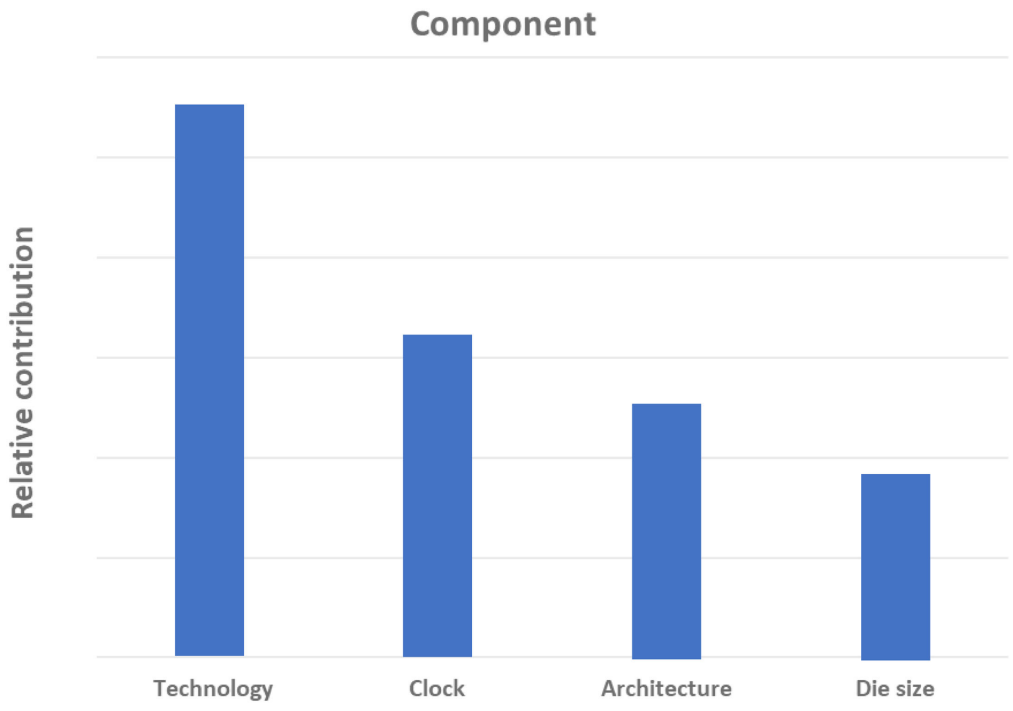

FIGURE 1. Performance contributions of the last 50 years.

workloads such as web browsers, word processing, spreadsheets, presentation drafting, etc. are not performance-bound by processing speed. This has led to a decrease in die size for many large production volume microprocessors; some are even as small as 1/10 the reticle limit for an 8-core chip.[2] At the same time, there are emerging workloads that have a massive and rapidly growing demand for more processing performance. Perhaps the most important emerging workload is neural network training, which is central to deep learning and artificial intelligence. These workloads can also take advantage of abundant parallelism. This combination drives the demand for more transistors, whether on a single piece of silicon or many.[3]

LARGE DIE AND RENT'S RULE

Why drive to one larger die rather than multiple die? In the 1960s, E. F. Rent discovered a remarkable correlation between the number of pins at the boundary of integrated circuit designs and the number of internal components such as logic gates. On a log–log plot, the relation of the number of logic gates to boundary pins lie on a straight line implying a power-law relation. More recently this has been extended to bandwidth sections of parallel programs.[4] The implications of this for large die are that the off-die bandwidth increases proportionally with the log of the die area.

Since processing power grows linearly with die area, while off-die bandwidth growth is a log function of area, the ratio of compute to off-die bandwidth increases with die area. This strongly encourages keeping as much compute as possible on-die to decrease the impact of relatively slow off-die communication, for workloads that require large amounts of compute.

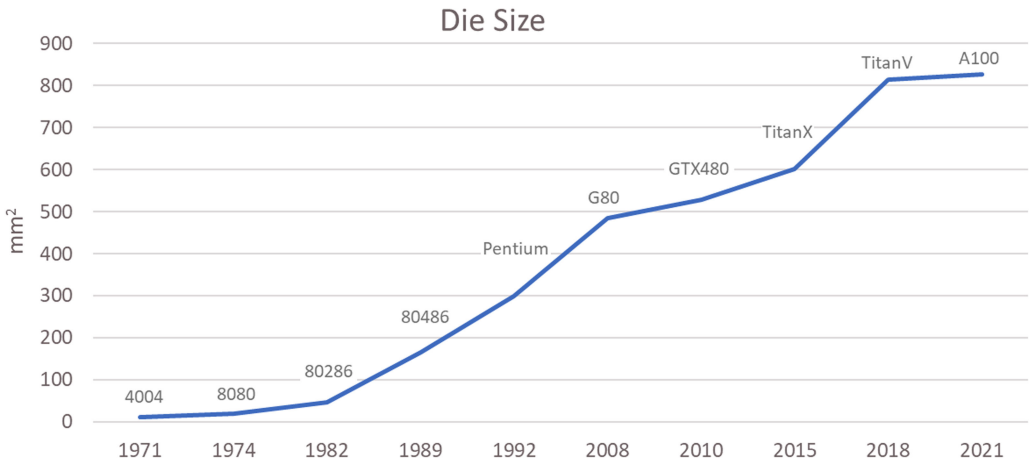

FIGURE 2. Maximum die size versus time.

Clearly, one way to make the most powerful processor possible is to use the entire wafer to make a single chip. This is called wafer-scale integration (WSI). The name extends the term very large scale integration that was the state of the art when WSI was being developed. So why is not wafer-scale integration commonly used?

LIMITS TO DIE SIZE

As shown in Figure 2, increasing reticle size (i.e., the largest area of the wafer that can be patterned using a lithography stepper system) becomes more difficult technology dimensions shrink. Maintaining lithography focus over ever larger areas as exposure wavelengths decrease is increasingly difficult. This is a primary contributor to the asymptotic growth in die size that is observed. Die size growth rate is not expected to increase in the near future and may actually decrease with high-NA EUV lithography because of the extremely short, 13.5-nm wavelength employed by these systems.[5]

An equally challenging limiter maximum die size is the yield of good die per wafer as die size increases. While there are several commonly used models of yield using different distributions of defects across a wafer—for example, Murphy's model, Poisson's model, the binomial model, Moore's model, and Seeds' model—common to all is a general trend of exponentially decreasing yield with increasing die size. Although extending the die size to a whole wafer is highly desirable to maximize compute and bandwidth, traditional yield models essentially guarantee zero yield since a defect-free wafer is an extremely rare event. Yielding a wafer-scale design is a fundamental obstacle.

EARLY EFFORT AT WAFER-SCALE INTEGRATION

In 1980, Trilogy Systems attempted to build an IBM-compatible mainframe using wafer-scale integration, producing a single-chip 2.5 inches (in) on a side. Their motivation was to decrease the performance and power costs of communicating between the large number of chips that were used to build mainframe computers. At that time, the reticle limit constrained chip size to a maximum of 0.25 in per side. This led to mainframes containing thousands of chips. Trilogy encountered several obstacles that eventually led to the failure of this early effort at wafer-scale integration. Notably

1) Exceeding the reticle limit forced the resulting design to use transistor geometries that were larger than the minimum available at that time due to photolithography tolerances over the larger area. This sacrificed transistor density and thus performance in order to produce the large die.
2) Triple modular redundancy was used to address the yield problem of a large silicon area. Each logic gate and flip-flop were triplicated employing a binary two-out-of-three voting at each flip-flop. Employing this triple modular redundancy at the lowest level came at a huge cost by effectively reducing transistor density by 3× compared to a single redundancy design.
3) The resulting 2.5-in die needed 1200 pins for off-chip interconnect, a herculean effort at that time that required yet more innovation in packaging. In addition, cooling the 2.5-in wafer-scale package required new types of heat exchangers to be designed.

TABLE 2. Reticle-to-reticle communication.

	Reticle size mm^2	Inter-reticle BW TB/s	Areal BW GB/s/ mm^2	Signaling energy pJ/bit	Inter-reticle signaling power W
Nvidia A100	826	0.6	0.762	10[1]	60
Cerebras WSE-2	525	3.2	6.1	0.15	4.8
Ratio		5.33	8	66.6	12.5

[1]Estimated based on PCIe 5.0 SERDES design in 7-nm technology.

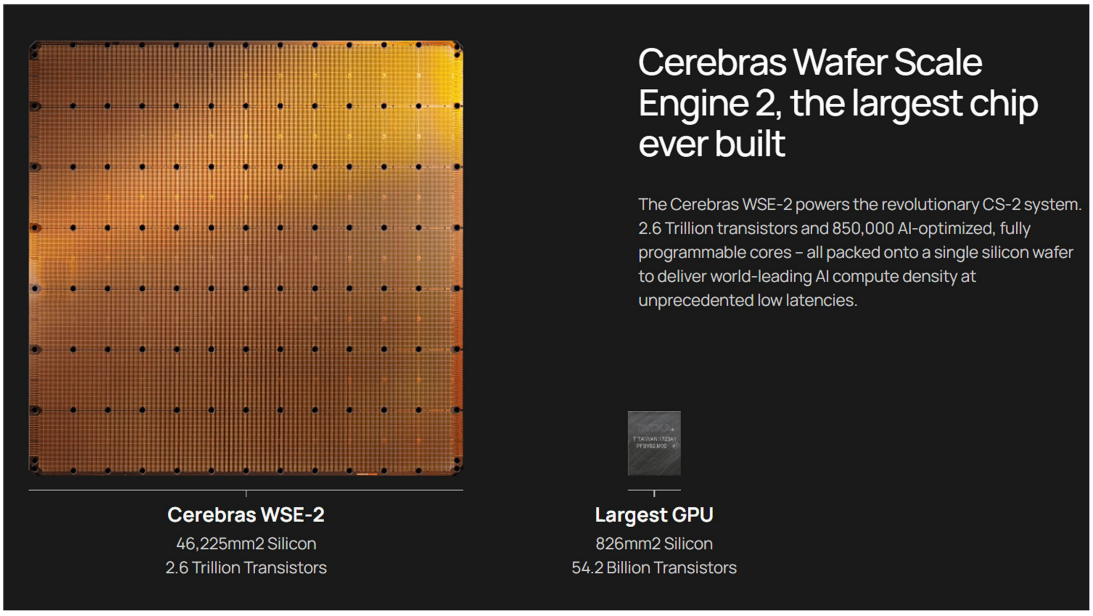

FIGURE 3. Cerebras' current, WSE-2.

Ultimately, Trilogy's effort failed and wafer-scale integration technology was abandoned by the industry. This early failure was so visible that for decades afterward, the Silicon Valley investment community refused to consider any startup efforts along this path.

ADVANTAGE OF LARGER SILICON

The compelling advantage to build a larger die always has been to reduce the time and energy required to communicate between functional units. Communication costs impact power and performance, with performance having two components: latency and bandwidth. Table 2 and Figure 3 compare two state-of-the-art designs. The Nvidia A100 is a conventional design, where each wafer is cut up to make a few hundred separate devices. Communication between A100 chips could be described as "off-silicon reticle-to-reticle communication." In contrast, the second-generation Cerebras Wafer-Scale Engine (WSE-2) is a single chip that uses the largest square area that fits on a 300-mm wafer. The WSE-2 uses on-silicon reticle-to-reticle communication. Keeping the communication on silicon results in much higher performance.

The A100 uses a high-speed serial interface mechanism called SERDES for inter-reticle (package-to-package) communication using traces on a printed circuit board and is fabricated at the 7-nm technology node. The WSE-2 is also fabricated at 7 nm but uses on-silicon wires for inter-reticle communication instead. The benefit of keeping reticle-to-reticle communication on silicon is apparent: the WSE-2 achieves nearly an order of magnitude greater bandwidth per square millimeter while using less than 1/12th of the power. This difference of nearly 100× in performance per watt is rarely seen between two contemporary technologies. For the SERDES-based technology to achieve comparable bandwidth to the WSE-2, it would need to consume 480 W just for the inter-reticle signaling, calculated from the PCIe Gen 4 specification. That is more than the power budget for the entire chip.

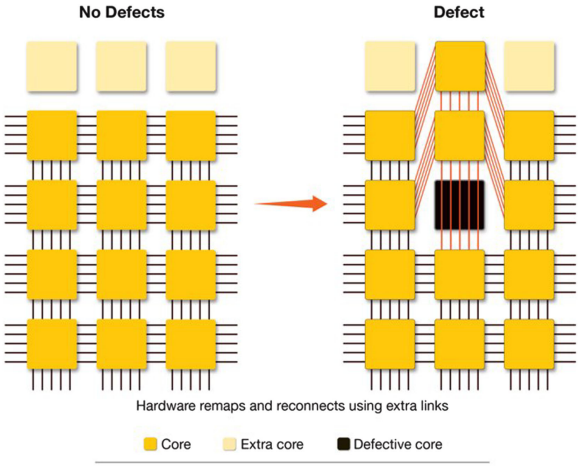

FIGURE 4. Routing around a defective PE.

OVERCOMING THE OBSTACLES TO WSI

In 2015, Cerebras Systems embarked on an effort to increase die size in five years matching that of the whole industry in the previous 50 years. The Cerebras WSE, first introduced in 2019, successfully overcame both the lithography and yield limits and for the first time delivered a computing system using WSI to dozens of commercial customers.

The first obstacle to overcome was wafer-scale interconnect without sacrificing the performance scaling benefits of the technology geometry as Trilogy did. This turned out to be fairly straightforward: using a standard step and repeat reticle of 525 mm^2, the full wafer was exposed as it would be in a standard process. An offset reticle of wiring between the 525-mm^2 reticles was then used to "stitch" together the standard reticles. The offset exposures do not fabricate any active circuitry, they are strictly reserved for upper level interconnect metal. This process of "field stitching" had been published but not previously applied to a commercial WSI product.[6]

The yield challenge was overcome by implementing redundancy at a much higher level than in the prior Trilogy design. We use a homogenous array of processing elements (PE) rather than redundant gates. The WSE-2 wafer comprises over 850,000 individual PEs with a small fraction, approximately 1%, held in reserve in order to "repair" defective PEs. This is a significant improvement beyond the approximately 66% redundancy overhead required by the earlier Trilogy design. Coupled with a uniform fine-grained repair mechanism (see Figure 4), this level of redundancy is sufficient to deliver very high effective wafer yield with a low cost to "repair" expected manufacturing defects.

A successful WSI design also requires redundancy in the infrastructure required to support the PE array: power, clocking, testing, initialization, and external communication. The Cerebras WSE provisions power, clocking, testing, and initialization separately to each reticle basis. Any defects in these global resources can be worked around by disabling that reticle. The amount of silicon area used for the global infrastructure is small so the probability of needing to disable a reticle is correspondingly low. For external IO, a simple overprovi-

FIGURE 5. WSE-2 package.

sioning of about 10% is used and any defective IO path can be "muxed" over to an alternate IO path.

As with Trilogy, Cerebras encountered more obstacles to be overcome, notably packaging, powering, and cooling the wafer. As shown in Figure 5, the WSE-2 package is unlike a normal chip package residing on a printed circuit board. The packaging challenges were manifold. Supplying the peak 20,000 A into the wafer at less than 1 V while maintaining good regulation may have been the hardest challenge. Cerebras' innovative solution employs more than 300 individual voltage regulation modules (VRMs) distributed over the wafer surface that drive current into the wafer perpendicular to its surface. Using multiple VRMs per reticle ensures redundancy in the power distribution and gives individual control of each reticle's power domain.

Extracting more than 15 kW of heat from the wafer was another challenge. This was solved by having the wafer "float" on a large copper heat exchanger. Water is pumped through micro-fins on the backside of the heat exchanger to remove heat from the powered wafer; the wafer is allowed to expand and contract while remaining in contact with the polished front side of the exchanger. The ability to float the wafer to maintain thermal connection to the heat exchanger despite the different coefficients of thermal expansion of copper and silicon is crucial.

The final challenge was to hold the assembly together while maintaining electrical contacts on the front side of the wafer for power and IO. For this, the Cerebras team "sandwiched" the wafer in an assembly comprising a thick PCB, a flexible membrane, the WSE and its heat exchanger. An array of clamping fasteners distribute the packaging force evenly across the assembly while still allowing the wafer to expand and contract.

While the journey to the first commercial WSI product has not been an easy one, the result is immensely satisfying. Our solution is delivering transformative performance and value to customers around the world.[7] The Cerebras Systems team is proud to have built a piece of microprocessor history.

REFERENCES

1. J. W. Backus, "Can programming be liberated from the von Neumann style? A functional style and its algebra of programs," *Commun. ACM*, vol. 21, no. 8, pp. 613–641, Aug. 1978. [Online]. Available: https://doi.org/10.1145/359576.359579
2. I. Cutress, "AMD ryzen mobile 4000: Measuring renoir's die size," *AnandTech*, Jan. 2020. [Online]. Available: https://www.anandtech.com/show/15381/amd-ryzen-mobile-4000-measuring-renoirs-die-size
3. F. Lardinois, "Google's newest cloud TPU pods feature over 1,000 TPUs," *TechCrunch*, May 2019. [Online]. Available: https://techcrunch.com/2019/05/07/googles-newest-cloud-tpu-pods-feature-over-1000-tpus
4. W. Heirman, J. Dambre, D. Stroobandt, and J. Campenhout, "Rent's rule and parallel programs: Characterizing network traffic behavior," in *Proc. Int. Workshop Syst. Level Interconnect Prediction*, 2008, pp. 87–94.
5. M. Lapedus, "Multi-Patterning EUV vs. High-NA EUV," Semiconductor Engineering, Dec. 2019. [Online]. Available: https://semiengineering.com/multi-patterning-euv-vs-high-na-euv/
6. W. Flack and G. E. Flores, "Lithographic manufacturing techniques for wafer scale integration," in *Proc. Int. Conf. Wafer Scale Integration*, 1992, pp. 4–13.
7. Cerebras Press Release, "Cerebras systems smashes the 2.5 trillion transistor mark with new second generation wafer scale engine," *Business Wire*, Apr. 2021. [Online]. Available: https://www.businesswire.com/news/home/20210420005955/en/Cerebras-Systems-Smashes-the-2.5-Trillion-Transistor-Mark-with-New-Second-Generation-Wafer-Scale-Engine

GARY LAUTERBACH is a Co-Founder and the Chief Technology Officer of Cerebras Systems. He is widely recognized as one of the industry's leading computer architects. Prior to Cerebras, he was Co-Founder and CTO of SeaMicro, where his inventions helped pioneer the microserver category. Following SeaMicro's acquisition by AMD, he was a Corporate Fellow and CTO for the server and server-CPU business units. Earlier in his career, he was a Distinguished Engineer with Sun Microsystems, where he was the Chief Microprocessor Architect for the UltraSPARC III and UltraSPARC IV microprocessors. He holds more than 50 patents. Contact him at gary@cerebras.net.