Chipping away at Moore's law

MODERN CPUs ARE JUST CHIPLETS CONNECTED TOGETHER.

JESSIE FRAZELLE

CPUs are made up of billions of teeny, tiny transistors. Transistors are electrical gates that can be switched on and off individually. Since each transistor can be in two distinct states (on or off), it can store two different numbers: zero and one. With billions of transistors, a chip can store billions of zeros and ones, and almost as many ordinary numbers and characters. The smaller the transistor, the less power is required for a chip to function.

You might have heard of AMD's 7nm chips or Intel's 10nm chips. A nanometer (nm) is a useful metric for judging how powerful a CPU is since it denotes the measurement of transistor size (or linewidth).

Smaller transistors can do more calculations without overheating, which makes them more power efficient. It also allows for smaller die sizes, which reduce costs and can increase density, allowing more cores per chip. The silicon wafers that chips are made of vary in purity, and none are perfect, which means every chip has a chance of having imperfections that differ in effect. Manufacturers can limit the effect of imperfections by using *chiplets*.

CHIPLETS

Traditionally, chip manufacturers connect two monolithic CPUs together in an MCM (multichip module). An MCM is a package with pins where multiple ICs (integrated circuits,

Chiplets are the individual ICs that make up an MCM.

or chips), semiconductor dies, and/or other components are integrated. This is usually done on a unifying substrate, so when the MCM is in use it can be treated as if it were one large chip. An MCM is sometimes referred to as a *hybrid IC*.

Chiplets are the individual ICs that make up an MCM. They provide a way to minimize the challenges of building with cutting-edge transistor technology. Spawning out of DARPA, the idea of chiplets has existed since the 1980s.[2]

The smallest transistors are also the most expensive and difficult with which to design and manufacture. In processors made up of chiplets, the most cutting-edge technology can be reserved for the pieces of a design where the investment will pay off the most. Then other chiplets manufactured using more reliable and economical techniques can be combined with the latest-technology chiplets into the same package. Intel does this with its Foveros Project, which combines, by way of stacking, 10nm chiplets for power-efficient activities and 14nm chiplets that serve higher power functions.[13] The section on 7nm chips later in this article will provide more examples of this, although you can undoubtedly anticipate that 7nm chiplets are combined with larger transistor chiplets into one package.

Smaller pieces of silicon are also inherently less prone to manufacturing defects and have a reduced risk of getting destroyed by dust particles during semiconductor fabrication. Using chiplets over monolithic die architecture guarantees no single core on a package will act as a single point of failure for the MCM since there are multiple individual cores. Manufacturers can also use a process known as *binning* to select the best cores to be paired with

each other.

Both AMD and Intel are breaking down monolithic processors into chiplets that are connected on a multichip module. AMD was the first in the mass-market CPU industry to move to chiplets when it announced its third-generation Ryzen CPU.[7] This improves performance and reduces costs since splitting a large monolithic IC into smaller chiplets allows for more transistors split across multiple chiplets, more ICs per silicon wafer, and better yields since smaller dies have less risk of defects or manufacturing errors.

In his 1965 paper, Gordon Moore wrote, "It may prove to be more economical to build large systems out of smaller functions, which are separately packaged and interconnected."[10] This is very true for chiplets. AMD gave a talk at the International Electron Devices Meeting in 2017[14] on how it split up its 32-core EPYC server-class chip onto four 8-core chiplets.[19] Doing so resulted in a 10 percent overhead for the added I/O to the four chiplets. Since it could fit 9 percent more of the smaller dies onto a wafer than larger dies, however, it was able to reduce the overall cost because of the improvements in die yield.

CHIPLET INTERCONNECTS

There are a few different ways manufacturers are interconnecting chiplets to form MCMs, most of which are proprietary to each vendor. Think of die-to-die interconnect (or chiplet interconnect) as the interface connecting the various chiplets. A chiplet is not just a piece of hardware; it is also a controller and a PHY (physical layer of software). The die-to-die interconnect is the

controller and the PHY for a chiplet. When designing die-to-die interconnects, power is the main constraint. Low power is table stakes because any additional power is considered overhead.

AMD calls its die-to-die interconnect Infinity Fabric, or IF (https://en.wikichip.org/wiki/amd/infinity_fabric). IF is not just a die-to-die interconnect but a processor-to-processor interconnect. It is made up of two parts: Infinity SDF (Scalable Data Fabric) and SCF (Scalable Control Fabric). SDF is the primary way data flows throughout the system. This can connect memory controllers, PCIe devices, USB hubs, and other peripherals. This interface is PHY level. The SCF, on the other hand, handles communication for controls such as power management, thermal controls, security, and tests.

Because most die-to-die interconnects are proprietary, the benefits of building with chiplets as if they are Lego blocks has not entirely come to fruition. Adding functionality such as an FPGA (field-programmable gate array) or an accelerator from a third party requires adopting the proprietary die-to-die interconnect.

As a response and solution to this problem, Intel open sourced its AIB (Advanced Interface Bus), which is a die-to-die interconnect standard that enables system design in a modular way with libraries containing chiplet IP (intellectual property) blocks.[8] (The specification is on GitHub.[1]) It is PHY level like AMD's SCF, but it will take a lot more than just Intel working on this problem for it to be solved for the wider industry. This effort is ongoing as a part of the Open Compute Project's subgroup ODSA (Open Domain-Specific Architecture).[11]

There hasn't been a transistor shrink from Intel since 2014.

7NM CHIPS

Moore's law observed that the number of transistors on a chip doubles every year, while the costs are halved. While this theory has held for a long time, it has been slowing down lately. In the late 1990s and early 2000s, transistors shrunk in size by half every two years, leading to massive improvements on a regular schedule. Making transistors smaller and smaller has gotten more complicated, however, and there hasn't been a transistor shrink from Intel since 2014 when it promised 10nm;[3,18] 7nm is possible but has not yet been done by Intel. The slowdown of transistor size in semiconductor devices led to the development of multichip modules and other innovations, since semiconductor designers are always looking for new ways to provide increased compute capabilities.

TSMC (Taiwan Semiconductor Manufacturing Company) began production of 256-Mbit SRAM (static random-access memory) chips using a 7nm process in 2017.[16] Later in 2018, Samsung and TSMC began mass production of 7nm devices.[15] The Apple A12 Bionic, a 7nm chip that went mainstream, was released at Apple's September 2018 event.[6] Technically, Huawei had announced its own 7nm processor, the Kirin 980, on August 31, 2018,[12] before the A12 Bionic, but the A12 was first to market and was released to consumers before the Kirin 980. Both the Kirin 980 and Apple A12 Bionic are manufactured by TSMC.

AMD released its Rome family of processors for servers and data centers, which are based on TSMC's 7nm node and feature up to 64 cores and 128 threads. AMD also released its Matisse family of consumer desktop processors with 16 cores and 32 threads. Technically, Intel does have one 7nm

part obtained with its acquisition of Barefoot Networks: the Tofino 2 chip, an Ethernet-switch ASIC, based on TSMC's 7nm node.

Shrinking transistor size isn't just about performance, though; it has huge implications for mobile, data center, and laptop chips. Compared with 14nm, 7nm can achieve 25 percent more performance under the same power, or the same performance for half the power. This means longer battery life for laptops and phones, and more power-efficient data centers, with the same performance. For smaller devices, since twice as much performance can effectively fit into the limited power target, this translates into much more powerful chips. Apple did this with its A12 Bionic chip.

You might have noticed a trend in that all these 7nm chips rely on TSMC for manufacturing. Companies using these chips are focusing the 7nm technology in small special-purpose chips in order to increase the yield and effectiveness since it is in high demand. For example, AMD's second-generation EPYC is built as nine die packages with eight 7nm CCD (complex core die) chiplets, each with up to eight cores, surrounding a 14nm I/O die.

The question is, if 7nm parts are in such high demand, why is TSMC the only foundry manufacturing them?

SEMICONDUCTOR FOUNDRIES

TSMC is the world's largest semiconductor foundry. Most of its production is in Taiwan, with only one plant in Shanghai, China. (It is interesting to note where each foundry has facilities since at the time of this writing the CoVID -19 coronavirus and U.S./China tariffs are

relevant, and some might be curious about their effect on the semiconductor industry.) Aside from TSMC, other common foundries include Globalfoundries and UMC (United Microelectronics Corporation). Globalfoundries, which has plants in Singapore, Germany, and the U.S., had set out to manufacture 7nm chips and stopped because of the expense.[9] After Globalfoundries publicly stated it would not manufacture 7nm chips, UMC followed and said it would not manufacture anything under 14nm because of the expense.[17] UMC has plants in Taiwan, Singapore, and the two Chinese cities of Suzhou and Xiamen.

What makes 7nm manufacturing so capital intensive? Foundries need to increase their capital expenditures to deal with the technical difficulties of decreasing the size of transistors, among which lithography remains the biggest hurdle. Lithography, also known as photolithography, is the process used in microfabrication to pattern transistor circuits onto silicon. Shining light on certain regions of silicon can create a specific pattern. Equipment for lithography is very expensive, preventing foundries from investing further in research and development to shrink transistor size.

Lithography technology has been improving with immersion lithography and multipatterning but with significant increases in the number of masks and processes. When additional exposures are needed to create complicated patterns on silicon, the costs of masks increase immensely. According to eBeam Initiative's survey,[5] the average number of masks per mask set has reached 76 for 7–10nm process node, and the number reaches more than 100 for manufacturers. Because of the

increase in mask cost, 7nm manufacturing processes have been outside the economical scope for most small and medium-sized design houses.

While phones, servers, graphics, and data centers all benefit from enhanced computing performance and power efficiency, the cost to manufacture bleeding-edge chips is increasing significantly. Therefore, not all foundries can handle the economics of 7nm chip manufacturing. The companies that have built products with 7nm chips include AMD, Apple, Samsung, Huawei, NVIDIA, and Barefoot Networks. What all these companies have in common is the motivation to be on the bleeding edge of technology in order to remain or become a market leader. They all benefit from economies of scale for their large production needs, so they can share the costs of masks, design, and manufacturing. Foundries without such high demand from a mass market cannot afford to invest in advanced technology since the risk might outweigh the reward.

Related articles

Getting Gigascale Chips
Challenges and Opportunities in Continuing Moore's Law
Shekhar Borkar
https://queue.acm.org/detail.cfm?id=957757

CPU DB: Recording Microprocessor History
With this open database, you can mine microprocessor trends over the past 40 years.
Andrew Danowitz, et al.
https://queue.acm.org/detail.cfm?id=2181798

Reconfigurable Future
The ability to produce cheaper, more compact chips is a double-edged sword.
Mark Horowitz
https://queue.acm.org/detail.cfm?id=1388771

FUTURE

What does all this mean for the future? It's likely that semiconductor designers will continue to innovate without needing to shrink transistor

size, as happened with multichip modules. Hopefully, the increase in demand for 7nm chips will give other foundries the economic incentive to enter the market, so the industry is not tied to TSMC alone. This is easier said than done, of course. I also hope by the end of reading this, you have learned that modern-day CPUs are actually just a few chiplets connected together into one package. The ability to construct packages of chiplets gives a Lego-like building power to people for designing their own MCMs and innovating faster.

References

1. Advanced Interface Bus (AIB) die-to-die hardware open source; https://github.com/chipsalliance/aib-phy-hardware.
2. Coldewey, D. 2017. DARPA project aims to make modular computers out of chiplets. *Techcrunch*; https://techcrunch.com/2017/08/26/darpa-project-aims-to-make-modular-computers-out-of-chiplets/.
3. Cuttress, I. 2019. Intel's 10nm Cannon Lake and Core i3-8121U deep dive review. *Anandtech*; https://www.anandtech.com/show/13405/intel-10nm-cannon-lake-and-core-i3-8121u-deep-dive-review.
4. eBeam Initiative. 2019. Mask makers' survey; https://www.ebeam.org/docs/2019-mask-maker-survey_en.pdf.
5. eBeam Initiative. 2020. Survey: 2019 eBeam Initiative mask makers' survey results. *Semiconductor Engineering*; https://semiengineering.com/survey-2019-ebeam-initiative-mask-makers-survey-results/.
6. Frumusanu, A. 2018. The iPhone XS and XS Max review: unveiling the silicon secrets. *Anandtech*; https://www.

anandtech.com/show/13392/the-iphone-xs-xs-max-review-unveiling-the-silicon-secrets/2.

7. Hruska, J. 2019. Chiplets are both solution to and symptom of a larger problem. *Extremetech*; https://www.extremetech.com/computing/290450-chiplets-are-both-solution-and-symptom-to-a-larger-problem.
8. Leibson, S. 2019. Intel releases royalty-free high-performance AIB interconnect standard to spur industry's chiplet adoption and grow the ecosystem. Intel Programmable Logic; https://blogs.intel.com/psg/intel-releases-royalty-free-high-performance-aib-interconnect-standard-to-spur-industrys-chiplet-adoption-and-grow-the-ecosystem/.
9. 15. McGregor, J. 2019. Globalfoundries' change in strategy pays off. *Forbes*; https://www.forbes.com/sites/tiriasresearch/2019/09/17/globalfoundries-change-in-strategy-pays-off/#40990f3c1d37.
10. Moore, G. 1965. Cramming more components onto integrated circuits. *Electronics* 38(8); https://newsroom.intel.com/wp-content/uploads/sites/11/2018/05/moores-law-electronics.pdf
11. Open Compute Project. 2020. Open Domain-Specific Architecture subgroup; https://www.opencompute.org/wiki/Server/ODSA.
12. Savov, V. 2018. Huawei promises its 7nm Kirin 980 processor will destroy the Snapdragon 845. *The Verge*; https://www.theverge.com/2018/8/31/17803682/huawei-kirin-980-processor-soc-qualcomm-snapdragon-845-ifa-2018.
13. Savov, V. 2018. Intel unveils Foveros 3D chip stacking and new 10nm 'chiplets.' *The Verge*; https://www.

theverge.com/2018/12/12/18137401/intel-foveros-3d-chip-stacking-10nm-roadmap-future.

14. Schor, D. 2017. IEDM 2017: AMD's grand vision for the future of HPC. *WikiChip Fuse*; https://fuse.wikichip.org/news/523/iedm-2017-amds-grand-vision-for-the-future-of-hpc/.
15. Tallis, B., Shilov, A. 2018. Samsung starts mass production of chips using its 7nm EUV process tech. *Anandtech*; https://www.anandtech.com/show/13496/samsung-starts-mass-production-of-chips-using-its-7nm-euv-process-tech.
16. TSMC. 7nm technology; https://www.tsmc.com/english/dedicatedFoundry/technology/7nm.htm.
17. Wang, L. 2018. UMC not to rejoin race to develop 7nm technology. *Taipei Times*; http://www.taipeitimes.com/News/biz/archives/2018/09/04/2003699736.
18. Williams, C. 2015. Intel TOCK BLOCK: 10nm Cannonlake delayed to 2017, bonus 14nm Kaby Lake to '16. *The Register*; https://www.theregister.co.uk/2015/07/16/intel_10nm_14nm_plans/.
19. Yeric, G. 2018. Three dimensions in 3DIC, Part 1. *Arm Research*; https://community.arm.com/developer/research/b/articles/posts/three-dimensions-in-3dic-part-1.

Jessie Frazelle *is the co-founder and Chief Product Officer of the Oxide Computer Company. Before that, she worked on various parts of Linux including containers as well as the Go programming language.*

CONTENTS