DOI:10.1145/3430936

Technological and economic forces are now pushing computing away from being general purpose and toward specialization.

BY NEIL C. THOMPSON AND SVENJA SPANUTH

The Decline of Computers as a General Purpose Technology

PERHAPS IN NO other technology has there been so many decades of large year-over-year improvements as in computing. It is estimated that a third of all productivity increases in the U.S. since 1974 have come from information technology,^{a,4} making it one of the largest contributors to national prosperity.

The rise of computers is due to technical successes, but also to the economics forces that financed them. Bresnahan and Trajtenberg³ coined the term *general purpose technology* (GPT) for products, like computers, that have broad technical applicability *and* where product improvement and market growth

could fuel each other for many decades. But, they also predicted that GPTs could run into challenges at the end of their life cycle: as progress slows, other technologies can displace the GPT in particular niches and undermine this economically reinforcing cycle. We are observing such a transition today as improvements in central processing units (CPUs) slow, and so applications move to *specialized processors*, for example, graphics processing units (GPUs), which can do fewer things than traditional universal processors, but perform those functions better. Many high profile applications are already following this trend, including deep learning (a form of machine learning) and Bitcoin mining.

With this background, we can now be more precise about our thesis: “The Decline of Computers as a General Purpose Technology.” We do *not* mean that computers, taken together, will lose technical abilities and thus ‘forget’ how to do some calculations. We *do* mean that the economic cycle that has led to the usage of a common computing platform, underpinned by rapidly improving universal processors, is giving way to a fragmentary cycle, where economics push users toward divergent computing platforms driven by special purpose processors.

» key insights

- **Moore's Law was driven by technical achievements and a “general purpose technology” (GPT) economic cycle where market growth and investments in technical progress reinforced each other. These created strong economic incentives for users to standardize to fast-improving CPUs, rather than designing their own specialized processors.**
- **Today, the GPT cycle is unwinding, resulting in less market growth and slower technical progress.**
- **As CPU improvement slows, economic incentives will push users toward specialized processors, which threatens to fragment computing. In such a computing landscape, some users will be in the ‘fast lane,’ benefiting from customized hardware, and others will be left in the ‘slow lane,’ stuck on CPUs whose progress fades.**

a Their analysis excludes the farming sector.

Figure 1. The historical virtuous cycle of universal processors (a) is turning into a fragmentation cycle (b).

This fragmentation means that parts of computing will progress at different rates. This will be fine for applications that move in the ‘fast lane,’ where improvements continue to be rapid, but bad for applications that no longer get to benefit from field-leaders pushing computing forward, and are thus consigned to a ‘slow lane’ of computing improvements. This transition may also slow the overall pace of computer improvement, jeopardizing this important source of economic prosperity.

Universal and Specialized Computing

Early days—from specialized to universal. Early electronics were not universal computers that could perform many different calculations, but dedicated pieces of equipment, such as radios or televisions, designed to do one task, and only one task. This specialized approach has advantages: design complexity is manageable and the processor is efficient, working faster and using less power. But specialized processors are also ‘narrower,’ in that they can be used by fewer applications.

Early electronic computers,^b even

those designed to be ‘universal,’ were in practice tailored for specific algorithms and were difficult to adapt for others. For example, although the 1946 ENIAC was a theoretically universal computer, it was primarily used to compute artillery range tables. If even a slightly different calculation was needed, the computer would have to be manually re-wired to implement a new hardware design. The key to resolving this problem was a new computer architecture that could store instructions.¹⁰ This architecture made the computer more flexible, making it possible to execute many different algorithms on universal hardware, rather than on specialized hardware. This ‘von Neumann architecture’ has been so successful that it continues to be the basis of virtually all universal processors today.

The ascent of universal processors. Many technologies, when they are introduced into the market, experience a virtuous reinforcing cycle that helps them develop (Figure 1a). Early adopters buy the product, which finances investment to make the product better. As the product improves, more consumers buy it, which finances the next round of progress, and so on. For many products, this cycle winds down in the short-to-medium term as

product improvement becomes too difficult or market growth stagnates.

GPTs are defined by the ability to continue benefiting from this virtuous economic cycle as they grow—as universal processors have for decades. The market has grown from a few high-value applications in the military, space, and so on, to more than two billion PCs in use worldwide.³⁸ This market growth has fueled ever-greater investments to improve processors. For example, Intel has spent \$183 billion on R&D and new fabrication facilities over the last decade.^c This has paid enormous dividends: by one estimate processor performance has improved about 400,000x since 1971.⁸

The alternative: Specialized processors. A universal processor must be able to do many different calculations well. This leads to design compromises that make many calculations fast, but none optimal. The performance penalty from this compromise is high for applications well suited to specialization, that is those where:

- ▶ substantial numbers of calculations can be parallelized
- ▶ the computations to be done are stable and arrive at regular intervals (‘regularity’)
- ▶ relatively few memory accesses are needed for a given amount of computation (‘locality’)
- ▶ calculations can be done with fewer significant digits of precision.¹⁵

In each of these cases, specialized processors (for example, Application-specific Integrated Circuits (ASICs)) or specialized parts of heterogeneous chips (for example, I.P. blocks) can

^b In this article, the term “computer” describes both, devices with solely a universal processor and those that also contain specialized functionality.

Technical specifications of a CPU compared to a GPU.

Processor	Model	Calculations in parallel ⁱ	Speed	Memory Bandwidth	Access to Level 1 Cache
CPU	Intel Xeon E5-2690v4	28	2.6–3.5 GHz	76.8 GB/s	5–12 clock cycles ⁱ
GPU	NVIDIA P100	3,584	1.1 GHz	732 GB/s	80 clock cycles

ⁱ Data from Intel and NVIDIA data sheets, ‘Access to L1 Cache’ from Giles^{50ppx}

ⁱⁱ Approximated by number of threads for CPU and number of CUDA cores for GPU.

^c Calculated as 2008–2017 R&D and additions to PPE spending.

perform better because custom hardware can be tailored to the calculation.²⁴

The extent to which specialization leads to changes in processor design can be seen in the comparison of a typical CPU—the dominant universal processor—and a typical GPU—the most-common type of specialized processor (see the accompanying table).

The GPU runs slower, at about a third of the CPU's frequency, but in each clock cycle it can perform ~100x more calculations in parallel than the CPU. This makes it much quicker than a CPU for tasks with lots of parallelism, but slower for those with little parallelism.^d

GPUs often have 5x–10x more memory bandwidth (determining how much data can be moved at once), but with much longer lags in accessing that data (at least 6x as many clock cycles from the closest memory). This makes GPUs better at predictable calculations (where the data needed from memory can be anticipated and brought to the processor at the right time) and worse at unpredictable ones.

For applications that are well-matched to specialized hardware (and where programming models, for example CUDA, are available to harness that hardware), the gains in performance can be substantial. For example, in 2017, NVIDIA, the leading manufacturer of GPUs, estimated that Deep Learning (AlexNet with Caffe) got a speed-up of 35x+ from being run on a GPU instead of a CPU.²⁷ Today, this speed-up is even greater.²⁶

Another important benefit of specialized processors^e is that they use less power to do the same calculation. This is particularly valuable for applications limited by battery life (cell phones, Internet-of-things devices), and those that do computation at enormous scales (cloud computing/datacenters, supercomputing).

As of 2019, 9 out of the top 10 most power efficient supercomputers were using NVIDIA GPUs.³⁷

Specialized processors also have important drawbacks: they can only run a limited range of programs, are hard to program, and often require a universal processor running an operating system to control (one or more of) them. Designing and creating specialized hardware can also be expensive. For universal processors, their fixed costs (also called non-recurring engineering costs (NRE)) are distributed over a large number of chips. In contrast, specialized processors often have much smaller markets, and thus higher per-chip fixed costs. To make this more concrete, the overall cost to manufacture a chip with specialized processors using leading-edge technology is about \$80 million^f (as of 2018). Using an older generation of technology can bring this cost down to about \$30 million.²³

Despite the advantages that specialized processors have, their disadvantages were important enough that there was little adoption (except for GPUs) in the past decades. The adoption that did happen was in areas where the performance improvement was inordinately valuable, including military applications, gaming and cryptocurrency mining. But this is starting to change.

The state of specialized processors today. All the major computing platforms, PCs, mobile, Internet-of-things

(IoT), and cloud/supercomputing, are becoming more specialized. Of these, PCs remain the most universal. In contrast, energy efficiency is more important in mobile and IoT because of battery life, and thus, much of the circuitry on a smartphone chip,³⁴ and sensors, such as RFID-tags, use specialized processors.^{5,7}

Cloud/supercomputing has also become more specialized. For example, 2018 was the first time that new additions to the biggest 500 supercomputers derived more performance from specialized processors than from universal processors.¹¹

Industry experts at the International Technology Roadmap for Semiconductors (ITRS), the group which coordinated the technology improvements needed to keep Moore's Law going, implicitly endorsed this shift toward specialization in their final report. They acknowledged the traditional one-solution-fits-all approach of shrinking transistors should no longer determine design requirements and instead these should be tailored to specific applications.¹⁶

The next section explores the effect that the movement of all of the major computing platforms toward specialized processors will have on the economics of producing universal processors.

The Fragmentation of a General Purpose Technology

The virtuous cycle that underpins GPTs comes from a mutually reinforcing set of technical and economic forces. Unfortunately, this mutual reinforcement also applies in the reverse direc-

^f This true for the 16/14nm node size. Lithography cost are by far the biggest cost component of the manufacturing NRE.²¹ Other costs include labor and design tools, as well as IP licensing.

Figure 2. Rate of improvement in microprocessors, as measured by (a) Annual performance improvement on the SPECint benchmark,^{7appx} and (b) Annual quality-adjusted price decline.^{1appx}

^d Of course, many tasks will have multiple parts, some parallelizable and some not. In which case, speed-ups will be constrained by Amdahl's Law.

^e For brevity we will use the term "specialized processors" throughout this article to refer both to stand-alone processors as well as specialized functionality on heterogeneous chips (for example, I.P. blocks)

tion: if improvements slow in one part of the cycle, so will improvements in other parts of the cycle. We call this counterpoint a ‘fragmenting cycle’ because it has the potential to fragment computing into a set of loosely-related siloes that advance at different rates.

As Figure 1(b) shows, the fragmenting cycle has three parts:

- ▶ Technology advances slow
- ▶ Fewer new users adopt
- ▶ Financing innovation is more difficult

The intuition behind this cycle is straightforward: if technology advances slow, then fewer new users adopt. But, without the market growth provided by those users, the rising costs needed to improve the technology can become prohibitive, slowing advances. And thus each part of this synergistic reaction further reinforces the fragmentation.

Here, we describe the state of each of these three parts of the cycle for computing and show that fragmentation has already begun.

Technology advancements slow. To measure the rate of improvement of processors we consider two key metrics: *performance*^g and *performance-per-dollar*. Historically, both of these metrics improved rapidly, largely because miniaturizing transistors led to greater density of transistors per chip (Moore’s Law) and to faster transistor switching speeds (via Dennard Scaling).²⁴ Unfortunately, Dennard Scaling ended in 2004/2005 because of technical challenges and Moore’s Law is coming to an end as manufacturers hit the physical limits of what existing materials and designs can do,³³ and these limits take ever more effort to overcome.² The loss of the benefits of miniaturization can be seen vividly in the slowdown of improvements to performance and performance-per-dollar.

Figure 2(a), based Hennessy and Patterson’s characterization of progress in SPECInt, as well as Figure 2(b) based on the U.S. Bureau of Labor Statistics’ producer-price index, show how dramatic the slowdown in performance improve-

GPTs are defined by the ability to continue benefiting from this virtuous economic cycle as they grow—as universal processors have for decades.

ment in universal computers has been. To put these rates into perspective, if performance per dollar improves at 48% per year, then in 10 years it improves 50x. In contrast, if it only improves at 8% per year, then in 10 years it is only 2x better.

Fewer new users adopt. As the pace of improvement in universal processors slows, fewer programs with new functionality will be created, and thus customers will have less incentive to replace their computing devices. Intel CEO Krzanich confirmed this in 2016, saying that the replacement rate of PCs had risen from every four years to every 5–6 years.²² Sometimes, customers even skip multiple generations of processor improvement before it is worth updating.²⁸ This is also true on other platforms, for example U.S. smartphones were upgraded on average every 23 months in 2014, but by 2018 this had lengthened to 31 months.²⁵

The movement of users from universal to specialized processors is central to our argument about the fragmentation of computing, and hence we discuss it in detail. Consider a user that could use either a universal processor or a specialized one, but who wants the one that will provide the best performance at the lowest cost.^h Figures 3(a) and 3(b) present the intuition for our analysis. Each panel shows the performance over time of universal and specialized processors, but with different rates at which the universal processor improves. In all cases, we assume that the time, T , is chosen so the higher price of a specialized processor is exactly balanced out by the costs of a series of (improving) universal processors. This means that both curves are cost equivalent, and thus superior performance also implies superior performance-per-dollar. This is also why we depict the specialized processor as having constant performance over this period. (At the point where the specialized processor would be upgraded, it too would get the benefit of whatever process improvement had benefited the universal processor and the user would again repeat this same decision process.)

A specialized processor is more at-

^g While we have in mind a measure of performance based on computational power/speed, this model is actually more general and could refer to other characteristics (for example, energy efficiency).

^h Computing at larger scales (including the massive parallelism of current deep learning models) are scaled-up versions of this same problem, and the logic of our analysis (and thus our results) also carry over to them.

tractive if it provides a larger initial gain in performance. But, it also becomes more attractive if universal processor improvements go from a rapid rate, as in panel (a), to a slower one, as in panel (b). We model this formally by considering which of two time paths provides more benefit. That is, a specialized processor is more attractive if

$$\int_0^T P_s dt \geq \int_0^T P_{u,t_0} e^{rt} dt$$

Where universal and specialized processors deliver performanceⁱ P_u , and P_s , over time T , while the universal processor improves at r .^j We present our full derivation of this model in the online appendix (<https://doi.org/10.1145/3430936>). That derivation allows us to numerically estimate the volume needed for the advantages of specialization to outweigh the higher costs (shown in Figure 3(c) for a slow-down from 48% to 8% in the per-year improvement rate of CPUs).

Not surprisingly, specialized processors are more attractive when they provide larger speedups or when their costs can be amortized over larger volumes. These cutoffs for when specialization becomes attractive change, however, based on the pace of improvement of the universal processors. Importantly, this effect does not arise because we are assuming different rates of progress between specialized and universal processors overall—all processors are assumed to be able to use whatever is the cutting-edge fabrication technology of the moment. Instead, it arises because the higher per-unit NRE of specialized processors must be amortized and how well this compares to upgrading universal processors over that period.

A numerical example makes clear the

importance of this change. At the peak of Moore's Law, when improvements were 48% per year, even if specialized processors were 100x faster than universal ones, that is, $\frac{P_s}{P_u} = 100$ (a huge difference), then ~83,000 would need to be built for the investment to pay off. At the other extreme, if the performance benefit were only 2x, ~1,000,000 would need to be built to make specialization attractive. These results make it clear why, during the heyday of Moore's Law, it was so difficult for specialized processors to break into the market.

However, if we repeat our processor choice calculations using an improvement rate of 8%, the rate from 2008–2013, these results change markedly: for applications with 100x speed-up, the number of processors needed falls from 83,000 to 15,000, and for those with 2x speed-up it drops from 1,000,000 to 81,000. Thus, after universal processor progress slows, many more applications became viable for

specialization.^k

Financing innovation is harder. In 2017, the Semiconductor Industry Association estimated that the cost to build and equip a fabrication facility ('fab') for the next-generation of chips was roughly \$7 billion.³⁵ By "next-generation," we mean the next miniaturization of chip components (or process 'node').

The costs invested in chip manufacturing facilities must be justified by the revenues that they produce. Perhaps as much as 30%^l of the industry's \$343 billion annual revenue (2016) comes from cutting-edge chips. So

k In the online appendix (<https://doi.org/10.1145/3430936>) we also consider how these values change with code development costs.

l \$23 billion of Foundry revenue (TSMC and GlobalFoundries) can be attributed to leading-edge nodes.³⁶ Assuming the majority (90%) of Intel's (\$54 billion) and Samsung's (\$40 billion) total revenues¹² derives from leading-edge nodes, yields an upper bound of \$108 billion/\$343 billion ≈ 30%.

Figure 3. Optimal processor choice depends on the performance speed-up that the specialized processor provides, as well as the rate of improvement of the universal technology.

i Here we assume that the cost of the CPU running the OS and controlling the specialized processor(s) does not materially affect this calculation. Relaxing that assumption would not change our model but would require incorporating of these costs into the specialized processor parameter estimates.

j In practice, manufacturers do not update continuously, but in large steps when they release new designs. Users, however, may experience these jumps more continuously since they tend to constantly refresh some fraction of their computers. The continuous form is also more mathematically tractable.

revenues are substantial, but costs are growing. In the past 25 years, the investment to build leading-edge fab (as shown in Figure 4a) rose 11% per year (!), driven overwhelmingly by lithography costs. Including process-development costs into this estimate further accelerates cost increases to 13% per year (as measured for 2001 to 2014 by Santhanam et al.³²). This is well known by chipmakers who quip about Moore's "second law": the cost of a chip fab doubles every four years.⁹ Historically, the implications of such a rapid increase in fixed cost on unit costs was only partially offset by strong overall semiconductor market growth (CAGR of 5% from 1996–

2016^{m,35}), which allowed semiconductor manufacturers to amortize fixed costs across greater volumes. The remainder of the large gap between fixed costs rising 13% annually and the market growing 5% annually, would be expected to lead to less-competitive players leaving the market and remaining players amortizing their fixed costs over a larger number of chips.

As Figure 4(b) shows, there has indeed been enormous consolidation in the industry, with fewer and fewer com-

panies producing leading-edge chips. From 2002/2003 to 2014/2015/2016, the number of semiconductor manufacturers with a leading-edge fab has fallen from 25 to just 4: Intel, Taiwan Semiconductor Manufacturing Company (TSMC), Samsung and GlobalFoundries). And GlobalFoundries recently announced that they would not pursue development of the next node.⁶

We find it very plausible this consolidation is caused by the worsening economics of rapidly rising fixed costs and only moderate market size growth. The extent to which market consolidation improves these economics can be seen through some back-of-the-envelope calculations. If the market were evenly partitioned amongst different companies, it would imply a growth in average market share from $4\% = \frac{100\%}{25}$ in 2002/2003 to $25\% = \frac{100\%}{4}$ in 2014/2015/2016. Expressed as a compound annual growth rate, this would be 14%. This means that producers could offset the worsening economics of fab construction through market growth and taking the market share of those exiting ($13\% < 5\% + 14\%$).

In practice, the market was not evenly divided, Intel had dominant share. As a result, Intel would not have been able to offset fixed cost growth this way.ⁿ And indeed, over the past decade, the ratio of Intel's fixed costs to its variable costs has risen from 60% to over 100%.^o This is particularly striking because in recent years Intel has slowed the pace of their release of new node sizes, which would be expected to *decrease* the pace at which they would need to make fixed costs investments.

The ability for market consolidation to offset fixed cost increases can only proceed for so long. If we project forward current trends, then by 2026 to 2032 (depending on market growth rates) leading-edge semiconductor manufacturing will only be able to support a single monopolist manufacturer, and yearly fixed costs to build a single new facility for each node size will be equal to yearly industry

m We implicitly assume that this is also the rate of growth for the leading-edge nodes. In practice it may be somewhat lower, which would only accentuate our point.

n Despite their rate of change being less favorable, Intel's large market share meant that they started from a lower base, so they remained highly competitive.
o Calculated from Intel financial statements, with fixed costs as R&D + Property, Plant and Equipment, and variable costs as cost of goods sold.

Figure 4. Deteriorating economics of chip manufacturing.

revenues (see endnote for details^p). We make this point not to argue that in late 2020s this will be the reality, but precisely to argue that current trends *cannot* continue and that within only about 10 years(!) manufacturers will be forced to dramatically slow down the release of new technology nodes and find other ways to control costs, both of which will further slow progress on universal processors.

The fragmentation cycle. With each of the three parts of the fragmentation cycle already reinforcing each other, we expect to see more and more users facing meager improvements to universal processors and thus becoming interested in switching to specialized ones. For those with sufficient demand and computations well-suited to specialization (for example, deep learning), this will mean orders of magnitude improvement. For others, specialization will not be an option and they will remain on universal processors improving ever-more slowly.

Implications

Who will specialize. As shown in Figure 3(c), specialized processors will be adopted by those that get a large speedup from switching, and where enough processors would be demanded to justify fixed costs. Based on these criteria, it is perhaps not surprising that big tech companies have been amongst the first to invest in specialized processors, for example, Google,¹⁹ Microsoft,³¹ Baidu,¹⁴ and Alibaba.²⁹ Unlike the specialization with GPUs, which still benefited a broad range of applications, or those in cryptographic circuits, which are valuable to most users, we expect narrower specialization going forward because only small numbers of processors will be needed to make the economics attractive.

We also expect significant usage from those who were not the original designer of the specialized processor, but who re-design their algorithm to take advantage of new hardware, as deep learning users did with GPUs.

^p Assumes new facilities are needed every two years; 30% of market sales go to leading edge chips; and 13% annual increase in fixed costs. 2026: 0% market growth / 2032: 5% market growth. We (conservatively) assume all market demand can be met with a single facility. If more than that is needed, the date moves earlier.

It is expected the final benefits from miniaturization will come at a price premium, and are only likely to be paid for by important commercial applications.

Who gets left behind. Applications that do not move to specialized processors will likely fail to do so because they:

- ▶ Get little performance benefit,
- ▶ Are not a sufficiently large market to justify the upfront fixed costs, or
- ▶ Cannot coordinate their demand.

Earlier, we described four characteristics that make calculations amenable to speed-up using specialized processors. Absent these characteristics, there are only minimal performance gains, if any, to be had from specialization. An important example of this is databases. As one expert we interviewed told us: over the past decades, it has been clear that a specialized processor for databases could be very useful, but the calculations needed for databases are poorly-suited to being on a specialized processor.

The second group that will not get specialized processors are those where there is insufficient demand to justify the upfront fixed costs. As we derived with our model, a market of many thousands of processors are needed to justify specialization. This could impact those doing intensive computing on a small scale (for example, research scientists doing rare calculations) or those whose calculations change rapidly over time and thus whose demand disappears quickly.

A third group that is likely to get left behind are those where no individual user represents sufficient demand, and where coordination is difficult. For example, even if thousands of small users would collectively have enough demand, getting them to collectively contribute to producing a specialized processor would be prohibitively difficult. Cloud computing companies can play an important role in mitigating this effect by financing the creation of specialized processors and then renting these out.^q

Will technological improvement bail us out? To return us to a convergent cycle, where users switch back to universal processors, would require rapid improvement in performance and/or performance-per-dollar. But technological trends point in the opposite direction. For example on performance, it is

^q Already, Google provides TPUs on its cloud,¹³ and Amazon Web Services (and others) provide GPUs.¹

expected that the final benefits from miniaturization will come at a price premium, and are only likely to be paid for by important commercial applications. There is even a question whether all of the remaining, technically-feasible, miniaturization will be done. Gartner predicts that more will be done, with 5nm node sizes being produced at scale by 2026,¹⁸ and TSMC recently announced plans for a \$19.5B 3nm plant for 2022.¹⁷ But many of the interviewees that we contacted for this study doubt were skeptical about whether it would be worthwhile miniaturizing for much longer.

Might another technological improvement restore the pace of universal processor improvements? Certainly, there is a lot of discussion of such technologies: quantum computing, carbon nanotubes, optical computing. Unfortunately, experts expect that it will be at least another decade before industry could engineer a quantum computer that is broader and thus could potentially substitute for classical universal computers.³⁰ Other technologies that might hold broader promise will likely still need significantly more funding to develop and come to market.²⁰

Conclusion

Traditionally, the economics of computing were driven by the general purpose technology model where universal processors grew ever-better and market growth fuels rising investments to refine and improve them. For decades, this virtuous GPT cycle made computing one of the most important drivers of economic growth.

This article provides evidence that this GPT cycle is being replaced by a fragmenting cycle where these forces work to slow computing and divide users. We show each of the three parts of the fragmentation cycle are already underway: there has been a dramatic and ever-growing slowdown in the improvement rate of universal processors; the economic trade-off between buying universal and specialized processors has shifted dramatically toward specialized processors; and the rising fixed costs of building ever-better processors can no longer be covered by market growth rates.

Collectively, these findings make it clear that the economics of processors has changed dramatically, pushing computing into specialized domains

that are largely distinct and will provide fewer benefits to each other. Moreover, because this cycle is self-reinforcing, it will perpetuate itself, further fragmenting general purpose computing. As a result, more applications will split off and the rate of improvement of universal processors will further slow.

Our article thus highlights a crucial shift in the direction that economics is pushing computing, and poses a challenge to those who want to resist the fragmentation of computing. **C**

References

1. Amazon Web Services: Elastic GPUs, 2017; <https://aws.amazon.com/de/ec2/elastic-gpus/>
2. Bloom, N., Jones, C., Van Reenen, J. and Webb, M. Are Ideas Getting Harder to Find? Cambridge, MA, 2017; <https://doi.org/10.3386/w23782>
3. Bresnahan, T.F. and Trajtenberg, M. General purpose technologies 'Engines of growth'? *J. Econom.* 65, 1 (Jan. 1995), 83–108; [https://doi.org/10.1016/0304-4076\(94\)01598-T](https://doi.org/10.1016/0304-4076(94)01598-T)
4. Byrne, D.M., Oliner, S.D. and Sichel, D.E. Is the information technology revolution Over? *SSRN Electron. J.* (2013), 20–36; <https://doi.org/10.2139/ssrn.2303780>
5. Cavin, R.K., Lugli, P. and Zhirnov, V. V. Science and engineering beyond Moore's Law. In *Proceedings of the IEEE 100, Special Centennial Issue* (May 2012), 1720–1749; <https://doi.org/10.1109/JPROC.2012.2190155>
6. Dent, S. Major AMD chip supplier will no longer make next-gen chips, 2018; <https://www.engadget.com/2018/08/28/global-foundries-stops-7-nanometer-chip-production/>
7. Eastwood, G. How chip design is evolving in response to IoT development. *Network World* (2017); <https://www.networkworld.com/article/3227786/internet-of-things/how-chip-design-is-evolving-in-response-to-iiot-development.html>
8. *Economist*. The future of computing—After Moore's Law (2016); <https://www.economist.com/news/leaders/21694528-era-predictable-improvement-computer-hardware-ending-what-comes-next-future>
9. *Economist*. The chips are down: The semiconductor industry and the power of globalization (2018); <https://www.economist.com/briefing/2018/12/01/the-semiconductor-industry-and-the-power-of-globalisation>
10. ENIAC Report. Moore School of Electrical Engineering, 1946.
11. Feldmann, M. New GPU-accelerated supercomputers change the balance of power on the TOP500, 2018; <https://www.top500.org/news/new-gpu-accelerated-supercomputers-change-the-balance-of-power-on-the-top500/>
12. Gartner Group. Gartner Says Worldwide Semiconductor Revenue Grew 22.2 Percent in 2017. Samsung Takes Over No. 1 Position. Gartner, 2018; <https://www.gartner.com/newsroom/id/3842666>
13. Google Cloud. Google: Release Notes, 2018; <https://cloud.google.com/tpu/docs/release-notes>
14. Hemsoth, N. An Early Look at Baidu's Custom AI and Analytics Processor. The Next Platform; <https://www.nextplatform.com/2017/08/22/first-look-baidus-custom-ai-analytics-processor/>
15. Hennessy, J. and Patterson, D. *Computer Architecture: A Quantitative Approach* (6th ed.). Morgan Kaufmann Publishers, Cambridge, MA, 2019.
16. International Technology Roadmap for Semiconductors 2.0: Executive Report. International technology roadmap for semiconductors, 79, 2015; http://www.semiconductors.org/main/2015_international_technology_roadmap_for_semiconductors_itrs/
17. Jao, N. Taiwanese chip maker TSMC to build the world's first 3nm chip factory. *Technode*, 2018; <https://technode.com/2018/12/20/taiwanese-chip-maker-tsmc-to-build-the-worlds-first-3nm-chip-factory/>
18. Johnson, B., Tuan, S., Brady, W., Jim, W. and Jim, B. *Gartner Predicts 2017: Semiconductor Technology in 2026*.
19. Jouppi, N.P. et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the*

- 44th Annual Int. Symp. Comput. Archit. (2017), 1–12; <https://doi.org/10.1145/3079856.3080246>
20. Khan, H.N., Hounshell, D.A. and Fuchs, E.R.H. Science and research policy at the end of Moore's Law. *Nat. Electron.* 1, 1 (2018), 14–21; <https://doi.org/10.1038/s41928-017-0005-9>
21. Khazraee, M., Zhang, L., Vega, L. and Taylor, M.B. Moonwalk? NRE optimization in ASIC clouds or accelerators will use old silicon. In *Proceedings of ACM ASPLOS 2017*, 1–16; <https://doi.org/http://dx.doi.org/10.1145/3037697.3037749>
22. Krzanich, B. Intel Corporation Presentation at Sanford C Bernstein Strategic Decisions Conference, 2016.
23. Lapedus, M. Foundry Challenges in 2018. *Semiconductor Engineering*, 2017; <https://semiengineering.com/foundry-challenges-in-2018/>
24. Leiserson, C.E., Thompson, N., Emer, J., Kuszmaul, B.C., Lamson, B.W., Sanchez, D. and Schardl, T.B. There's plenty of room at the top: What will drive growth in computer performance after Moore's Law ends? *Science* (2020).
25. Martin, T.W. and Fitzgerald, D. Your love of your old smartphone is a problem for Apple and Samsung. *WSJ* (2018); <https://www.wsj.com/articles/your-love-of-your-old-smartphone-is-a-problem-for-apple-and-samsung-1519822801>
26. Mims, C. Huang's Law is the new Moore's Law, and explains why Nvidia wants arm. *WSJ* (2020); <https://www.wsj.com/articles/huangs-law-is-the-new-moores-law-and-explains-why-nvidia-wants-arm-11600488001>
27. NVIDIA Corporation. Tesla P100 Performance Guide. HPC and Deep Learning Applications, 2017.
28. Patton, G. Forging Intelligent Systems in the Digital Era. MTL Seminar, 2017; <https://www-mtl.mit.edu/mtlseminar/garypatton.html#simple3>
29. Pham, S. 2018. Who needs the US? Alibaba will make its own computer chips. *CNN Business*, 2018; <https://edition.cnn.com/2018/10/01/tech/alibaba-chip-company/index.html>
30. Prickett Morgan, T. Intel Takes First Steps To Universal Quantum Computing. Next Platform, 2017; <https://www.nextplatform.com/2017/10/11/intel-takes-first-steps-universal-quantum-computing/>
31. Putnam, A. et al. A reconfigurable fabric for accelerating large-scale datacenter services. *Commun. ACM* 59, 11 (Oct. 2016), 114–122; <https://doi.org/10.1145/2996868>
32. Santhanam, N., Wiseman, B., Campbell, H., Gold, A. and Javetski, B. McKinsey on Semiconductors, 2015.
33. Shalf, J.M. and Leland, R. Computing beyond Moore's Law. *Computer* 48, 12 (Dec. 2015), 14–23; <https://doi.org/10.1109/MC.2015.374>
34. Shao, Y.S., Reagen, B., Wei, G.Y., and Brooks, D. Aladdin: A pre-RTL, power-performance accelerator simulator enabling large design space exploration of customized architectures. In *Proceedings of the Int. Symp. Comput. Archit.* (2014), 97–108; DOI:<https://doi.org/10.1109/ISCA.2014.6853196>
35. Semiconductor Industry Association: 2017 Factbook; <http://go.semiconductors.org/2017-sia-factbook-0-0-0>
36. Smith, S.J. Intel: Strategy Overview, 2017; <https://doi.org/10.1016/B978-0-240-52168-8.10001-X>
37. Top500. The Green500 List (June 2019); <https://www.top500.org/green500/lists/2019/06/>
38. Worldometers. Computers sold in the world this year, 2018; <http://www.worldometers.info/computers/>

Neil C. Thompson (neil_t@mit.edu) is an innovation scholar in the Computer Science and Artificial Intelligence Lab and the Initiative on the Digital Economy at the Massachusetts Institute of Technology, Cambridge, MA, USA.

Svenja Spanuth (sspanuth@ethz.ch) is a Ph.D. candidate in the Department of Management, Technology, and Economics at ETH Zurich, Switzerland.

The authors contributed equally to the work.

Copyright held by authors.

Watch the authors discuss this work in the exclusive *Communications* video. <https://cacm.acm.org/videos/the-decline-of-computers>