THE DECLINE OF COMPUTERS AS A GENERAL PURPOSE TECHNOLOGY

By Neil C. Thompson and Svenja Spanuth

It is easy to forget that our computers weren't always highly-flexible systems capable of the staggering variety of computational wizardry that we take for granted today. They used to be special-purpose machines built to calculate things like ballistic trajectories or breaking codes. The rise of computers as a general purpose technology (GPT) only happened because of concurrent technical and economic breakthroughs where product improvement and market growth fueled one another.

Our research finds that technological and economic forces are now pushing computing in the opposite direction, making computer processors less general-purpose and more specialized. This process has already begun, driven by a slowing of Moore's Law and the success of algorithms like deep learning. So, what are the repercussions? The trend toward specialization threatens to fragment computing into "fast lane" applications that get powerful customized chips, and "slow lane" applications that get stuck using general-purpose chips whose progress is fading.

The rise of general-purpose computer chips had a profound impact on society; their decline could too. Our work outlines the forces already starting to fragment this GPT and what may lie ahead.

IN THIS BRIEF

- Technological and economic forces are making computer processors less general-purpose and more specialized. This process has already begun, driven by a slowing of Moore's Law and the success of algorithms like deep learning.
- Specialization threatens to fragment computing into "fast lane" applications that get powerful customized chips, and "slow lane" applications that get stuck using general-purpose chips whose progress is fading.
- The virtuous, general-purpose technology (GPT) cycle that has driven computing for decades is ending and is being replaced by a fragmented cycle where computing separates into specialized domains that are largely distinct and provide few benefits to each other.
- In the long term, this fragmentation could slow the overall pace of computer improvement, jeopardizing an important source of economic prosperity.

BACKGROUND

The technical and financial successes of computing are well-recognized. Bresnahan and Trajtenberg (1992) first noted the virtuous cycle of GPT, which began with expensive computers that only benefited a few, high-value applications (military, space, etc.). As computer chip manufacturers invested in innovation, however, they produced ever-better performance at lower cost, which caused more industries to adopt computers. Increased demand financed further improvements, and the virtuous cycle flourished. For computer chips, this GPT cycle has continued for decades and the resultant improvements (often, colloquially described as Moore's Law) have been transformative.

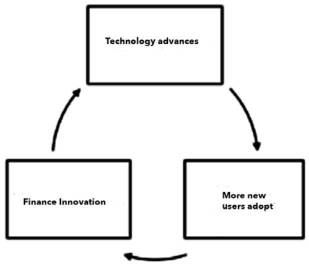


Figure 1: The virtuous cycle of computers as a general purpose technology

The extent to which the virtuous GPT cycle has shaped computing is hard to overstate. Since the early days of the Intel 4004 processor, there has been enormous market expansion. For example, from 2000 to 2010, the sale of personal computers (PCs) grew an average of 9% per year (Wong et al., 2017). There are now more than 2 billion PCs in use worldwide (Worldometers, 2018). This market growth has fueled ever-greater investments to improve chips.

Over the last decade, Intel spent $183 billion on R&D and new fabrication facilities[1] with enormous dividends: By one estimate, processor performance has improved about 400,000 times since 1971 (The Future of Computing, 2016). Indeed, one popular description of Moore's Law phrases this growth as hardware performance doubling every two years at constant cost.

Not surprisingly, the effect of computing on the economy has been substantial, too. Byrne, Oliner, and Sichel (2013) estimate that since 1974 information technology has been responsible for more than a third of the annual labor productivity growth in the U.S. non-farm sector.

1. Calculated as 2008-2017 R&D and additions to PPE spending.

THE DECLINE OF COMPUTERS AS A GENERAL PURPOSE TECHNOLOGY

By Neil C. Thompson and Svenja Spanuth

THE PULL OF SPECIALIZATION

Based on the compelling economics of GPTs, it might be easy to conclude that once processors became universal, they would never return to being specialized. But our research shows how the opposite can also occur. We show the forces pulling computer chips from a GPT into smaller, less-general pieces.

As Bresnahan and Trajtenberg predicted, at the end of their lifecycle GPTs can run into challenges. As progress slows, other technologies can displace the GPT in particular niches. We are observing just such a transition today as some applications move to specialized computer processors that perform fewer functions, but perform them better. Many high-profile applications are already following this trend, including Deep Learning and Bitcoin mining.

Our research illustrates how we are moving from the traditional, universal model of computer hardware--providing broad-based benefits to many--to a model where different applications use different computer hardware with uneven benefits. In the long term, this fragmentation of computing could also slow the overall pace of computer improvement, jeopardizing an important source of economic prosperity.

Specifically, we are moving away from an era when almost everyone was using a similar computing platform and improvements in that platform were widely felt. Instead, we are heading to an era where different users are on different computing platforms and many improvements are only narrowly felt. As a result, some applications will get to be in the "fast lane," where improvements continue to be rapid. Other applications will no longer get positive spill-overs from these leading domains and will be consigned to a "slow lane" of computing improvements.

Specialized processors have existed for some time. For example, in the early years of computing many supercomputers used specialized hardware such as Cray's architecture. But the attractiveness of this option diminished because universal processor performance improved exponentially. As a result, it became unattractive to invest millions of dollars to develop specialized, proprietary processor chips (Lapedus, 2017b), and universal processors dominated the market until at least the mid-2000s.

Today, advances in universal processors have slowed considerably. Whereas, chip performance-per-dollar improved 48% per year from 2000-2004, improvement has been less than 10% since 2008 (BLS, 2018). This anemic progress in general purpose chips makes specialized processors more attractive becaus their performance jump provides a long-term boost.

SEMICONDUCTOR TRENDS

Not only is universal processor performance improvement slowing, producers face rapidly escalating costs. Semiconductor manufacturing has always been a capital-intensive industry, but the costs are accelerating. Of the 25 chip manufacturers that made cutting-edge chips at the beginning of the millennium, all but three have ceased making the necessary investments to stay at the cutting-edge (Smith, 2017). This isn't surprising. It currently costs a staggering $7 billion to build a manufacturing plant (Semiconductor Industry Association, 2017) and a roughly equivalent amount to design and operationalize the production of a new generation of chips – and both of these are still increasing.

The worsening economics of chip manufacturing poses an important threat to the advancement of universal processor performance because the economic cycle of GPTs also works in reverse: if higher costs and technical challenges slow performance improvement, then market growth will slow, which makes financing the next round of improvements less attractive, which slows performance improvement, and so on.

Using a theoretical model and empirical evidence we show that this reversal is indeed under way, encouraging specialized applications, and steadily draining the market that fuels improvements in universal chips. Put another way, as the improvements in GPT slow, movement to fragmented, niche technologies accelerate.

DEEP LEARNING REAPS ADVANTAGES

The performance advantage from moving to specialized processors can be substantial. For instance, deep learning, a machine learning algorithm that can run on specialized chips for tasks such as image recognition (Russakovsky et al., 2015), is one beneficiary of the trend. Before specialized processors took hold, deep learning was not even competitive with other image recognition algorithms. In some cases today, deep learning makes fewer errors than humans when categorizing images.

Deep learning has also proven to be the superior algorithm for many natural language processing tasks, including machine reading, speech recognition (Hinton et al., 2012) and machine translation (Sutskever et al., 2014) (Jean et al., 2015). Familiar systems include the voice systems for Google Home (Marr, 2017), Apple's Siri (Levy, 2016), Amazon's Alexa (Strom, 2015), and machine translation systems such as Skype Translator (Skype, 2014) and Google translate (Turovsky, 2016). Facebook uses

THE DECLINE OF COMPUTERS AS A GENERAL PURPOSE TECHNOLOGY

By Neil C. Thompson and Svenja Spanuth

deep learning to help with picture tagging, to filter for hate-speech, and to customize advertisements to the users (Marr, 2016).
Absent faster hardware, deep learning would still be in the doldrums of its neural network ancestors (Goodfellow et al., 2016). Instead, it has proven to be transformative, infusing itself into numerous applications that we use every day.

Another major advantage of using specialized processors is energy-efficiency. This not only allows much higher performance of smartphones or Internet-of-Things (IoT) devices without immediately draining the battery, but also reduces datacenter energy costs. We find that over time, supercomputers with specialized processors are improving the number of calculations that they can perform per watt almost five times as fast as those that only use universal processors, and that this result is highly statistically significant[2].

MARKET FRAGMENTATION

Based on these advantages, transitioning to specialized processors, and thereby displacing universal processors, seems like a logical choice; but there will be tradeoffs. For some applications, technical or economic reasons will preclude the move to specialized processors, and they will get left behind. Worse, the universal processors on which they are built will be improving more slowly. So, as the virtuous cycle of universal chips is replaced by a fragmenting one, access to ever-better computers will no longer be guaranteed for all users. Instead of computing improvements being "a tide that raises all boats," they will become uneven, ranging from highly accelerated to stagnating. Key among the left-behind applications will be those whose current algorithms are ill-suited to specialization, and those with insufficient demand or fragmented application users.

Moreover, if improvements slow in one part of the cycle, so will improvements in other parts of the cycle. We call this latter cycle a fragmenting cycle because it has the potential to fragment the GPT, leaving a set of loosely related technologies advancing at different rates.
The fragmenting cycle has three parts:
• Fewer new users adopt the technology
• Financing innovation is harder
• Technology advances slow

The move to specialized processors, therefore, undermines the GPT cycle in two ways: It diminishes the number of new users adopting universal processors, and it anchors many of the switchers there so that even if processor performance were to speed up again, it would require more time and greater improvement to move those users back. [3]

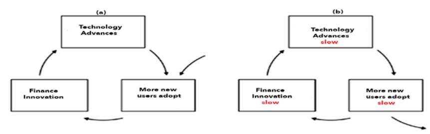


Figure 2: The historical virtuous cycle of universal processers (a), is turning into a fragmentation cycle (b).

Nevertheless, we also expect consolidation only to proceed for so long. If we project current trends forward, by 2026 to 2032 (depending on market growth rates) leading-edge semiconductor manufacturing will 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 revenue. 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 ten years (!) manufacturers will be forced to dramatically slow the release of new technology nodes and find other ways to control cost--which will further slow progress on universal processors.

Industry experts confirmed this shift toward specialized processors in the final report of the International Technology Roadmap for Semiconductors (ITRS), the group which coordinated the technology improvements needed to keep Moore's Law going. In its final report in 2015, ITRS acknowledges that "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" (ITRS, 2015). This is precisely the fragmentation of the general technology that our research uncovered.

WHO WILL BE THE 'WINNERS'?

The switch to specialized processors may not be better for everyone, but it will be better for some. For problems where the is enough market demand for specialized chips and where the technical details are amenable, specialization can provide big

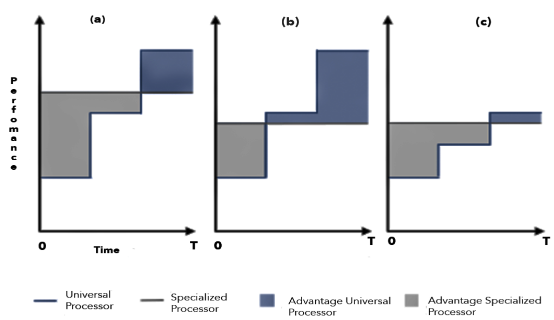
2. We based our regression on data from the Top 500 list. This list is released twice a year and ranks the world's 500 best supercomputers.

3. There is a subtle, but important third effect. Specialized chips are likely to have longer replacement cycles (because of the high fixed costs) and use older process technology. Both effects decrease demand for cutting-edge chips, further undermining the economics of producing new, cutting-edge chip manufacturing plants. These transition dynamics also occurred in the past, when supercomputer users slowly made their way from specialized chips to massive numbers of universal processors.

benefits. We already see this, for example, in Google's usage of Tensor Processing Units (TPUs), which are designed to do their type of deep learning very efficiently.

The types of computations that work well for specialization are those where:

- Calculations can be done with much greater amounts of parallelism,
- The computations to be done are very stable and arrive at regular intervals (called regularity),
- Few memory accesses are needed (called locality),
- Calculations can be done with fewer significant digits of precision (Hennessy & Patterson, 2017).



For calculations with these properties, specialized processors perform better because different design choices can be made than were done with universal chips. Broadly speaking, the more this changes the design of the chip, the larger the gains from switching to a specialized processor. As noted, the two main ways that these gains manifest are better performance and better energy efficiency.

Figure 3 shows how the performance gains from specialized processors can be eroded quickly (or not) depending on the pace of improvement of the universal processors. In the figure, a specialized processor is more attractive tha[4]n a universal processor when the grey shaded region is larger than the blue shaded region. Thus, a specialized processor is more attractive if it provides a larger initial gain in performance, as in panel (a), or if the gains that it provides take longer to erode because the universal processor is improving more slowly, as in panel (c). In contrast, universal processors are more attractive when their rate of improvement quickly eclipses any performance jump from specialization, as in panel (b).

On the other hand, if universal processors improve less quickly, they become less attractive and more users will want to switch to specialized chips. In this way, the move to specialized chips perpetuates itself, fragmenting the general purpose model and splitting off more and more applications.

CONCLUSION

We conclude that the virtuous GPT cycle that has driven computing for decades is ending. This paper provides evidence that the GPT cycle is being replaced by a fragmented cycle where computing separates into specialized domains that are largely distinct and provide few benefits to each other. This trend will have important implications for individual users and for the economy more broadly.

For users who can profitably switch to specialized chips, there are likely to be significant gains, as we've seen with deep learning and cryptocurrency. For those who can't switch, the picture will be bleaker as universal chip progress slows and with it, much of their computing performance improvements. On a larger scale, we argue that the switch to specialization will worsen the economics of chip manufacturing, leading to slower improvements. Therefore, the move to specialized chips perpetuates itself, fragmenting the general -purpose model and splitting off more and more applications.

BIBLIOGRAPHY

Bresnahan, T. F., & Trajtenberg, M. (1992). General Purpose Technologies: "Engines of growth." NBER Working Paper Series, 1–43.

BLS. (2018). PPI industry data for Semiconductors and related device manufacturing - Microprocessors. Retrieved March 24, 2018, from https://beta.bls.gov/dataViewer/view/timeseries/PCU33441333441312

Byrne, D. M., Oliner, S. D., & Sichel, D. E. (2013). Is the Information Technology Revolution Over?, International Productivity Monitor, (25), 20-36.

4. The duration of time in the plot is assumed to equate the cost of the two types of processors: either a sequence of improving universal processors or a single specialized processor (that is more expensive because of higher fixed costs).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. Cambridge, MA: MIT Press.

Hennessy, J., & Patterson, D. (2017). Domain-Specific Architectures. In Computer Architecture: A Quantitative Approach (6th Edition, pp. 432–498). Waltham, MA: Morgan Kaufmann Publishers.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine, 29(6), 82–97.

ITRS. (2015). Executive Report. International Technology Roadmap for Secmiconductors 2.0. Retrieved January 17, 2018 from https://www.semiconductors.org/clientuploads/Research_Technology/ITRS/2015/0_2015%20ITRS%202.0%20Executive%20Report%20(1).pdf

Jorgenson, D. W., & Stiroh, K. J. (2000). Raising the Speed Limit: U.S. Economic Growth in the Information Age. Brookings Papers on Economic Activity, 2000(1), 125–210.

Jean, S., Firat, O., Cho, K., Memisevic, R., & Bengio, Y. (2015). Montreal Neural Machine Translation Systems for WMT15. Proceedings of the Tenth Workshop on Statistical Machine Translation, (September), 134–140.

Lapedus, M. (2017b). Foundry Challenges in 2018. Retrieved March 19, 2018, from https://semiengineering.com/foundry-challenges-in-2018/

Levy, S. (2016). An Exclusive Look at How AI and Machine Learning Work at Apple. Retrieved January 30, 2018, from https://www.wired.com/2016/08/an-exclusive-look-at-how-ai-and-machine-learning-work-at-apple/

Marr, B. (2016). 4 Mind-Blowing Ways Facebook Uses Artificial Intelligence. Retrieved January 30, 2018, from https://www.forbes.com/sites/bernardmarr/2016/12/29/4-amazing-ways-facebook-uses-deep-learning-to-learn-everything-about-you/#7432dab2ccbf

Marr, B. (2017). The Amazing Ways Google Uses Deep Learning AI. Retrieved January 30, 2018, from https://www.forbes.com/sites/bernardmarr/2017/08/08/the-amazing-ways-how-google-uses-deep-learning-ai/#28cd5b823204

Moore, G. E. (1995). Lithography and the future of Moore's law. In R. D. Allen (Ed.), Proceedings of SPIE, Vol. 2437, 2–17.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211–252.

SIA. (2017). SIA 2017 Factbook. Semiconductor Industry Association. Retrieved from http://go.semiconductors.org/2017-sia-factbook-0-0-0

Smith, S. J. (2017). Intel Technology and Manufacturing Day – Strategy Overview [PDF slides]. Retrieved from https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/09/stacy-smith-on-milestones-in-intels-process-technology-roadmap.pdf

Strom, N. (2015). Scalable distributed DNN training using commodity GPU cloud computing. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2015–January, 1488–1492.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. NIPS, 9.

The future of computing - After Moore's law. (2016). Retrieved January 25, 2018, from https://www.economist.com/news/leaders/21694528-era-predictable-improvement-computer-hardware-ending-what-comes-next-future

Turovsky, B. (2016). Found in translation: More accurate, fluent sentences in Google Translate. Retrieved January 30, 2018, from https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/

Wong, D., Kan, K., Chanda, A., & Zhang, J. (2017). Processor Review Q3 2017 Semiconductors. Wells Fargo

Worldometers. (2018). Computers sold in the world this year. Retrieved March 2, 2018, from http://www.worldometers.info/computers/

REPORT

A full version of the Working Paper can be found here: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3287769

THE DECLINE OF COMPUTERS AS A GENERAL PURPOSE TECHNOLOGY

By Neil C. Thompson and Svenja Spanuth

ABOUT THE AUTHORS

Neil C. Thompson is a Research Scientist at MIT's Computer Science and Artificial Intelligence Lab and a Visiting Professor at the Lab for Innovation Science at Harvard. He is also an Associate Member of the Broad Institute, and was previously an Assistant Professor of Innovation and Strategy at the MIT Sloan School of Management, where he co-directed the Experimental Innovation Lab (X-Lab).

Svenja Spanuth is a recent M.Sc. graduate from RWTH Aachen University, Germany. She obtained a dual degree in electrical engineering and business administration. During her graduate studies she spent a year at MIT Sloan School of Management combining her two fields of study. Her research centers on the economic impact of developments in the semiconductor industry.