WORKING PAPER

The Decline of Computers as a General Purpose Technology:

Why Deep Learning and the End of Moore's Law are Fragmenting Computing

Neil C. Thompson*

Laboratory for Innovation Science at Harvard &

MIT Computer Science and Artificial Intelligence Lab

Svenja Spanuth*

MIT Sloan School of Management &

RWTH Aachen University

November 2018

Abstract

It is a triumph of technology and of economics that our computer chips are so universal. Countless applications are only possible because of the staggering variety of calculations that modern chips can compute. But, this was not always the case. Computers used to be specialized, doing only narrow sets of calculations. Their rise as a 'general purpose technology (GPT)' only happened because of the technical breakthroughs by computer scientists like von Neumann and Turing, and the mutually-reinforcing economic cycle of general purpose technologies, where product improvement and market growth fuel each other.

This paper argues that technological and economic forces are now pushing computing in the opposite direction, making computer processors less general purpose and more

*N. Thompson and S. Spanuth contributed equally to this work.

specialized. This process has already begun, driven by the slow-down in Moore’s Law and the algorithmic success of Deep Learning. This trend towards 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 fades.

The rise of general purpose computer chips has been remarkable. So, too, could be their fall. This paper outlines the forces already starting to fragment this general purpose technology.

1 Introduction

The history of computing has been a triumphant one. Since their first invention in the mid-20th century, computers have become pervasive, shaping the work and personal lives of much of humanity. Perhaps in no other technology have there been such large year-on-year improvements over so many decades. The economic implications of this rise are substantial, for example it is estimated that a third of all productivity increases in the United States since 1974 came from information technology (Byrne, Oliner and Sichel, 2013), making it one of the largest contributors to national prosperity.

The rise of computers is due to technical successes and to the economics forces that financed them. The economics of this process were first identified by Bresnahan and Trajtenberg (1992) as the virtuous cycle of a general purpose technology (GPT). This cycle begins with expensive computers that only benefit a few high-value applications (military, space, etc.). But, as computer chip manufacturers invest in innovation, they produce ever-better performance at lower cost, which causes more and more industries to adopt computers. Increased demand then finances further improvements and the cycle continues. For computer chips, this GPT cycle has held for decades and the resultant improvements (often described as Moore's Law[1]) have been transformative.

But, just as Bresnahan and Trajtenberg (1992) predicted, GPTs at the end of their lifecycle can run into challenges. As progress slows, the possibility arises for other technologies to displace the GPT in particular niches. We are observing just such a transition today as some applications move to specialized computer processors - which can do fewer things than traditional 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. This paper outlines why this transition is happening. It shows how we are moving from the traditional model of computer hardware that is universal, providing broad-based benefits to many and improving rapidly, to a model where different applications use different computer hardware and where the benefits are uneven. In the long term, this fragmentation of computing could also slow the overall pace of computer improvement, jeopardizing this important source of economic prosperity.

With this background, we can now be more precise about what we mean about "The Decline of Computers as a General Purpose Technology." We do *not* mean that computers, as a whole, will 'forget' how to do some calculations. We *do* mean that we are moving away from an era when almost everyone was using a similar computing platform, and thus where improvements in that platform were widely felt, to an era where different users are on different computing platforms and many improvements are only narrowly felt. This fragmentation will mean that parts of computing will progress at different rates. This will be okay for applications that get to be in the 'fast lane,' where improvements continue to be rapid, but very bad for applications that no longer get positive spill-overs from these leading domains and are thus consigned to a 'slow lane' of computing improvements.

Specialized processors and the significant speedup they often offer is not a recent invention. 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 new specialized processor chips (Lapedus, 2017b), and universal processors dominated the market until at least the mid-2000s.

Today, this trend has started to reverse itself because advances in universal processors have slowed considerably. Whereas chip performance-per-dollar improved 48% per year from 2000-2004, it has been less than 10% since 2008 (BLS, 2018). This slowing of improvement in the general purpose chips makes specialized processors more attractive because the one-time jump in performance they get from being more efficient provides an advantage for longer.

Not only is performance improvement slowing for universal processor users, but universal processor producers face rapidly escalating costs. Semiconductor manufacturing has always been capital intensive industry, but it is becoming ever-more so. 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)(Dent, 2018). 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. In 2014, for the first time since the economy-wide adoption of computers in the 1990s, semiconductor industry giant Intel's fixed costs surpassed their variable costs.

The worsening economics of chip manufacturing poses an important threat to the advancement of universal processor performance[2] because the reinforcing economic cycle of general purpose technologies also works in the reverse direction: 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 further slows performance improvement, and so on. This process is further exacerbated if consumers, facing slowing universal chips improvement, switch to specialized chips and thus further diminish market demand for the universal chips. Using a theoretical model and empirical evidence we show that this is indeed going on, pushing more and more applications to specialize and thus steadily draining the market that fuels improvements in universal chips. Put another way, as the improvements in the general purpose technology slow, movement to fragmented, niche technologies accelerate.

The performance advantage that comes from moving to specialized processors can be substantial and lead to significant breakthroughs. For example, Deep Learning is a machine learning algorithm that can be run on specialized chips, and where the benefits of doing so has been transformative for tasks such as image recognition (Russakovsky et al., 2015). Today, Deep Learning makes fewer errors than humans when categorizing images. Before moving to specialized processors, Deep learning was not even competitive with other image recognition algorithms whose error rates were roughly five times as high as humans'.

Another major advantage of 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 the power bill for datacenters.

Based on these advantages, transitioning to specialized processors, and thereby displacing the universal processors, seems a logical choice. But for some applications there will be technical or economic reasons that preclude the move to specialized processors. These applications will get left behind. Worse, the technology they get left behind on, the universal processors, 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.

Our argument proceeds as follows: Section 2 reviews the historical triumph of universal computers as a general purpose technology. Section 3 documents the rise of specialized chips and how that trend has been accelerated by Deep Learning. Section 4 explains how the general purpose technology cycle, which has underpinned computing improvement for so many decades, is now reversing itself and fragmenting computing. Section 5 outlines the consequences of this change and Section 6 concludes.

2 The Rise of Universal Computers

2.1 From Specialized to Universal Computers

In 1969, the Japanese company Busicom decided to re-design one of their products, a calculator, and so they went to the newly-founded semiconductor chip manufacturer, Intel. In a seminal decision, Intel chose not to build Busicom a specialized chip that could only be used for that calculator, but instead to build a universal processor that could be used for any type of computing, and which could be programmed to do the functions of a calculator (Malone, 1995). Understanding why this choice was so important requires some historical perspective on early electronics and early universal computers.

Early electronics were not universal computers, but dedicated pieces of equipment, such as radios or televisions. Inside, their specialized electronics were designed to do one task, and only one task. This is the type of chip that Busicom had in mind when they came to Intel. It would be specialized, performing exactly the functions that the calculator needed to do, and nothing else. This approach has advantages: the design complexity is manageable and the processor is highly efficient, working fast and using little power. But this approach also has a key drawback: it lacks flexibility. Functionality can't be added after the chip is created.

In contrast to specialized processors, ‘universal’ processors are ones that can perform many different calculations. The goal of creating a ‘universal’ computer has a long history. As early as 1837, Charles Babbage attempted to build a mechanical version, an “analytical engine”, to solve “numerical computations of the most varied kind” (Davis, 2012). But Babbage never succeeded, and it would be another century until Alan Turing proved that computers could be universal.

Early electronic computers[3], 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. This requirement to re-implement hardware designs for new functionality continued even after the switch-over from vacuum tubes to semiconductors (Noyce & Hoff, 1981). The key to resolving this problem was invented by John von Neumann in 1945. Based on work by Eckert and Mauchly, he developed a computer architecture that could store instructions and thus could 'reprogram' the hardware for each step of a calculation. This made it possible to execute algorithms in software on a universal computer, rather than on specialized hardware. This von Neumann architecture has been so successful that it continues to be the basis of virtually all universal computers today.

The pioneering work on universal computers and the progress in semiconductor electronics came together in the 4004 microprocessor that Intel built for Busicom in 1971. It was the first commercial chip to house a universal processor[4], and thus the first one that could be adapted to many applications by just changing the software. Thus began a cycle that has revolutionized computing and led us to the pervasive use of universal computers that we enjoy today.

2.2 The Virtuous Economic Cycle of a General Purpose Technology

Many technologies, when they are introduced into the market, experience a virtuous reinforcing cycle that helps them develop. Early adopters buy the product, which finances investment to make the product better.[5] 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 markets become saturated.

For general purpose technologies, where there is enormous potential for market expansion because they can be useful to so many consumers across so many industries, this virtuous economic cycle can continue for long periods of time so long as technical constraints allow it. For universal processors, the quintessential example of a general purpose technology (Bresnahan & Trajtenberg, 1992), this is exactly what has happened: the virtuous general purpose technology cycle (shown in Figure 1) has lasted for decades as more and more applications throughout the economy became digitized.

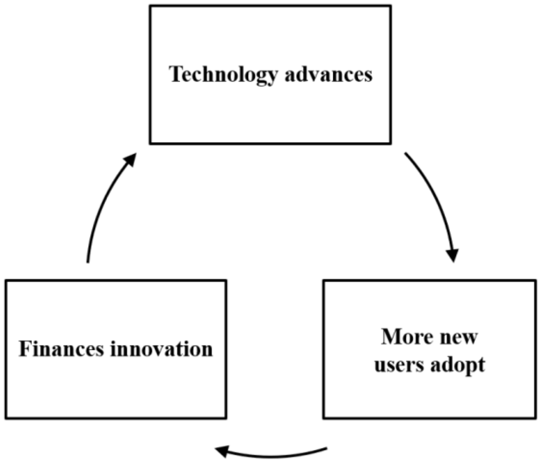


Figure 1: The virtuous cycle of computers as a general purpose technology

The extent to which the virtuous general purpose technology cycle has shaped computing is hard to overstate. From its nascent state with the Intel 4004 processor, there has been enormous market expansion. From 2000 to 2010, the number of personal computers (PCs) sold grew an average of 9%[6] per year (Wong et al., 2017), and 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[7]. This has paid 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 (an important technical trend that underlies much of the improvement) phrases this as hardware performance doubling every two years at constant cost[8].

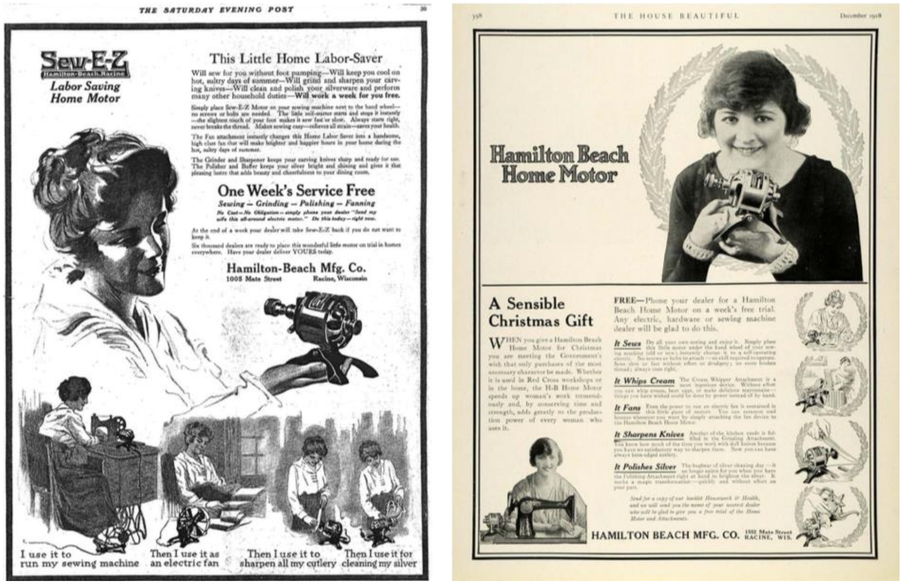
Not surprisingly, the effect of computing on the economy has been substantial. Jorgenson and Stiroh (2000) conclude that computer hardware, software and communication equipment contributed 0.74 percentage points annually towards U.S. economic growth in the late 1990s. Byrne, Oliner and Sichel (2013) 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.

2.3 When a General Purpose Technology Became Less General

Based on the compelling economics of general purpose technologies, it might be easy to conclude that once processors became universal, they would never return to being specialized. But this paper argues the opposite. This argument unfolds over the coming sections, but here we present an example of another general purpose technology that fragmented into smaller, less-general pieces and which illustrates the forces that will be relevant for our discussion.

At the beginning of the 20th century, electric motors (another general purpose technology) became small enough to automate kitchen appliances, but were still expensive. For example, a 1917 sewing machine with an integrated electrical motor sold for $35 (Western Electric, 1917), representing 28% of an average household monthly income (Chao & Utgoff, 2006). Rather than requiring customers to buy specialized motors for each appliance, the Hamilton-Beach Company invented a universal home motor ($11.50) which could attach to, and power, a variety of existing manual appliances (original advertisements shown in Figure 2).


THE SATURDAY EVENING POST 30

Sew-E-Z
Hamilton-Beach, Racine
Labor Saving
Home Motor

This Little Home Labor-Saver

Will sew for you without foot pumping—Will keep you cool on hot, sultry days of summer—Will grind and sharpen your carving knives—Will clean and polish your silverware and perform many other household duties—**Will work a week for you free.**

Simply place Sew-E-Z Motor on your sewing machine next to the hand wheel—no screws or bolts are needed. The little self-starter starts and stops it instantly—the slightest touch of your foot makes it sew fast or slow. Always starts right, never breaks the thread. Makes sewing easy—relieves all strain—saves your health.

The Fan attachment instantly changes this Home Labor Saver into a handsome, high class fan that will make brighter and happier hours in your home during the hot, sultry days of summer.

The Grinder and Sharpener keeps your carving knives sharp and ready for use. The Polisher and Buffer keeps your silver bright and shining and gives it that pleasing lustre that adds beauty and cheerfulness to your dining room.

One Week's Service Free
Sewing — Grinding — Polishing — Fanning
No Cost—No Obligation—simply phone your dealer "Send my wife this all-around electric motor." Do this today—right now.

At the end of a week your dealer will take Sew-E-Z back if you do not want to keep it.

Six thousand dealers are ready to place this wonderful little motor on trial in homes everywhere. Have your dealer deliver YOURS today.

Hamilton-Beach Mfg. Co.
1008 Main Street Racine, Wisconsin

I use it to run my sewing machine

Then I use it as an electric fan

Then I use it to sharpen all my cutlery

Then I use it for cleaning my silver

358 THE HOUSE BEAUTIFUL December 1918

Hamilton Beach
Home Motor

A Sensible Christmas Gift

WHEN you give a Hamilton Beach Home Motor for Christmas you are meeting the Government's wish that only purchases of the most necessary character be made. Whether it is used in Red Cross workshops or in the home, the H-B Home Motor speeds up woman's work tremendously and, by conserving time and strength, adds greatly to the production power of every woman who uses it.

FREE—Phone your dealer for a Hamilton Beach Home Motor on a week's free trial. Any electric, hardware or sewing machine dealer will be glad to do this.

It Sews Do all your own sewing and enjoy it. Simply place this little motor under the hand wheel of your sewing machine (old or new), instantly change it to a self-operating electric. No screws or bolts to attach—no skill required to operate. Sews slow or fast without effort or drudgery; no more broken thread; always runs right.

It Whips Cream The Cream Whipper Attachment is a most ingenious device. Without effort you can whip cream, beat eggs, or make delicious mayonnaise—things you have wished could be done by power instead of by hand.

It Fans Even the power to run an electric fan is contained in this little piece of motors. You can summon cool breezes whenever you want by simply attaching the fan device to the Hamilton Beach Home Motor.

It Sharpens Knives Another of the kitchen needs is fulfilled in the Grinding Attachment. You know how much of the time you work with dull knives because you have no satisfactory way to sharpen them. Now you can have always keen-edged cutlery.

It Polishes Silver The bugbear of silver cleaning day—it no longer exists for you when you have the Polishing Attachment right at hand to brighten the silver. It works a magic transformation—quickly and without effort on your part.

Send for a copy of our booklet Housework & Health, and we will send you the name of your nearest dealer who will be glad to give you a free trial of the Home Motor and Attachments.

HAMILTON BEACH MFG. CO. 1552 Main Street RACINE, WIS.

Figure 2: Home Motor advertisements[9]

One can imagine a world where universal home motors became the standard. Fueled by increasing sales, Hamilton-Beach could have invested in improvements to make the motor cheaper and better, which would further expand the market, financing more improvements and so on. Had the economics and technical improvements worked out, the motor's performance and costs would become so good that universal home motors would become the standard embedded into most home devices.

But appliance electric motors didn't stay universal; they specialized. One hundred years later, instead of a single type of motor that can power anything, we have a panoply: from tiny motors in hand-held fans that can run on AAA batteries to blender motors that can crush ice because they are 350 times as powerful (see Figure 3).[10]

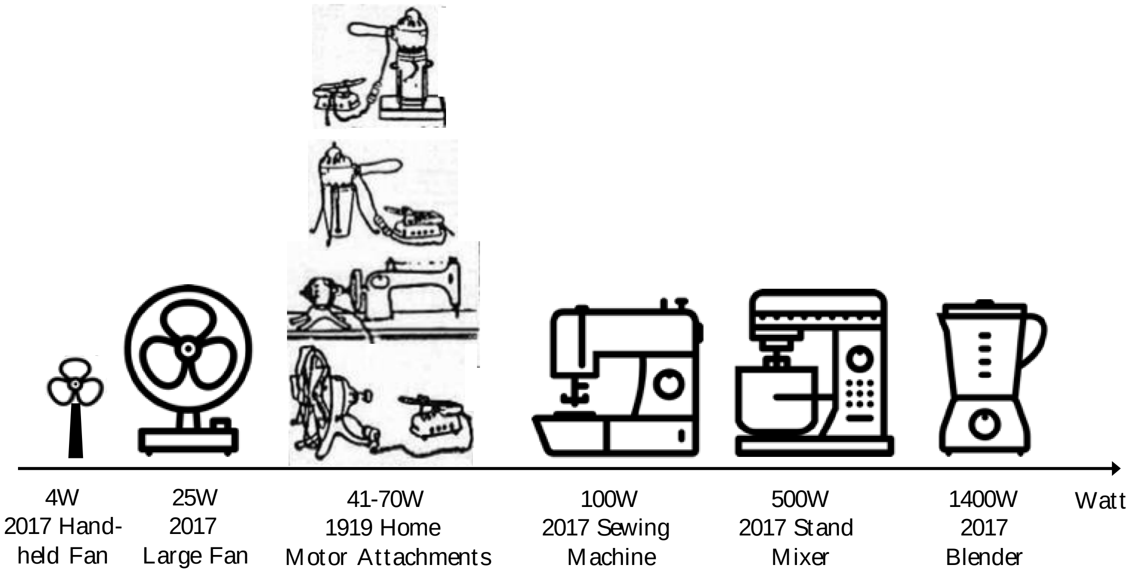


Figure 3: Power rating of different household appliances

So, what differentiates the history of computers and electric motors? Why did computers become more universal whereas electric motors became specialized? This work argues that the main difference between them is the rate of performance-per-dollar improvement over time.

Had the performance-per-dollar of universal home motors improved rapidly, a single design might have had the power needed for heavy-duty applications like a blender, while costing no more than a current motor for a small fan. With these type of improvements, universal home motors might have persisted. Instead, the costs of powerful motors remained high and the power of cheap motors stayed low, limiting adoption. And thus, home motors became specialized rather than universal.

In computing, in stark contrast to home motors, the performance-per-dollar of universal processors has improved exponentially. This meant that even for those wanting high-end performance, the benefits of moving to a specialized processor could be quickly eclipsed by subsequent versions of the universal processor - particularly if the specialized processor was also costlier. We will formalize this intuition with a model in succeeding sections, but first we provide more detail on specialized processors and the recent developments that are favoring their adoption.

3 Deep Learning and the Rise of Specialized Processors

After a long quiescence, there has recently been remarkable growth in the use of specialized processors for cell phones, internet-of-things (IoT), and Deep Learning. Thus far in this paper, we have asserted that this might be because of performance advantages that arise from using a specialized processor. In this section we justify this assertion, explaining specialized processors' advantages and why these were insufficient to advantage them historically but may be sufficient to advantage them now.

3.1 The Advantages of Specialized Processors

All else equal, if one had to choose between two processors, one universal and one specialized, it is hard to imagine not choosing the processor that performs a greater variety of tasks (i.e. the universal one). But, of course, all else is not equal. Achieving the breadth of a universal processors requires making compromises that specialized processors can avoid. To understand how this advantages specialized processors, it is useful to understand the broad technical challenges to running a processor efficiently. To demystify some of the electrical engineering jargon around processors, we explain the challenges in processor operation via two analogies to car manufacturing: supply-chain management and production scale-up.

Supply-chain management is critical for keeping a manufacturing plant utilized because some raw material inputs, for example steel from China, have long lead times. Without managing these properly, a plant could go idle while it waits for inputs to arrive. The analogous process for computer processors is sourcing the data for a calculation. When coordinated well, the data needed as an input is foreseen and stored near the processor (in a 'cache') so that it is available when needed for the calculation. However, if there is insufficient storage nearby, or an unexpected calculation is performed, a processor may go idle as the data is sourced from far away (the hard drive). The downtime associated with this can be astounding. In the time it takes to get new data from the hard drive, a processor can perform millions of calculations! This highlights how important data management is and why specialized processors, by designing the data flow (e.g. memory bandwidth) specifically for one problem, can yield large speedups (Hennessy & Patterson, 2017).

A second challenge in car manufacturing is scaling-up to produce more. This could involve modifying an existing line to make it run more quickly or adding production lines to be run in parallel. Similar options are available for processors. Processors can be run more quickly if the heat produced (usually the limiting factor in processor speed) can be dissipated effectively. Specialized processors can be designed to waste much less power, reducing heat production. Specialized chips can also be designed so that they can do many calculations in parallel, for example through the addition of additional processors (or 'cores'). This can speed up a calculation massively if the larger problem can be transformed into many smaller, independent ones capable of running in parallel.

With this model of processor operation in mind, it is easier to see why making a processor more universal can hinder performance for specialized tasks. For example, consider the trade-offs involved in allocating chip 'real-estate' (which is valuable).[11] One use for the space would be the addition of extra processor cores so that more calculations can be done in parallel. Another use is to add cache so that more data can be stored closer to the processor. Which choice is better depends on the calculation being done. Because universal processors must be able to do many different calculations well, design choices are made to benefit a broad set of calculations, even if that means that they are not optimal for any particular one. In practice, this has meant that universal processors are designed with lots of cache (to avoid those expensive delays sourcing data from the hard drive)[12] and have some, but not that much, parallelism.[13]

Because of these limitations, there are certain types of problems where a specialized processor is much better than a universal processor. These can be divided into cases where (i) calculations can be done with much greater amounts of parallelism, (ii) the computations to be done are very stable and arrive at regular intervals (called ‘regularity’), (iii) few memory accesses are needed (called ‘locality’), and (iv) calculations can be done with fewer significant digits of precision[14] (Hennessy & Patterson, 2017). In each of these cases, specialized processors perform better because different trade-offs can be made to tailor the hardware to the calculation. Broadly speaking, the more this changes the design of the chip, the larger the gains from switching to a specialized processor. The two main ways that these gains manifest are better performance and better energy efficiency.

3.1.1 *Performance*

The extent to which specialization leads to changes processor design can be seen in the comparison of a typical central processing unit (CPU – the dominant universal processor) and a typical graphics processing unit (GPU - the most-common type of specialized processor).

Table 1: Technical specifications of a CPU compared to a GPU[15]

Processor	Model	Calculations in parallel16	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 cycles17
GPU	NVIDA P100	3,584	1.1 GHz	732 GB/s	80 clock cycles

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. Testing validates this, showing that for workloads with high parallelism GPUs are cheaper per calculation (Brodtkorb et al., 2013)

The memory systems are also designed differently, with GPUs having almost 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.

Table 2 shows a compilation, put together by NVIDIA, the leading manufacturer of GPUs, of how these characteristics translate into performance gains for various applications that are well-suited to GPUs[18]. Notice particularly, how well Deep Learning (in the final row) performs, as this is a theme we'll return to.

Table 2: Speedups of various applications through GPU implementation[19] (NVIDIA Corporation, 2017a)

Application	Field	Description	Speedup
GTC-P	Physics	A development code for optimization of plasma physics	5x
RTM	Oil and Gas	Reverse time migration (RTM) modeling is a critical component in the seismic processing workflow of oil and gas exploration	5x
LAMMPS	Molecular Dynamics	Classical molecular dynamics package	6x
MILC	Physics	Lattice Quantum Chromodynamics (LQCD) codes simulate how elemental particles are formed and bound by the 'strong force' to create larger particles like protons and neutrons	6x
VASP	Quantum Chemistry	Package performing ab-initio quantum-mechanical molecular dynamics (MD) simulations	10x
HOOMD-Blue	Molecular Dynamics	Particle dynamics package is written from the ground up for GPUs	14x
Specfem 3D	Oil and Gas	Simulates Seismic wave propagation	18x
LSMS	Quantum Chemistry	Materials code for investigating the effects of temperature on magnetism	25x
Amber	Molecular Dynamics	Suite of programs to simulate molecular dynamics on biomolecule	34x
AlexNet with Caffe	Deep Learning	Winning network of ImageNet competition 2012 combined with a popular, GPU-accelerated Deep Learning framework developed at UC Berkeley	>35x

3.1.2 Energy Efficiency

Another important benefit of specialized processors 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 / data centers, supercomputing). For example, our cell phones typically have a host of specialized processors to handle common computations (e.g. phone connectivity, security, navigation, graphics operations).

Energy efficiency is also a major cost driver for large data centers. In 2014, datacenters made up 1.8% of the US overall energy consumption (Shehabi et al., 2016) and NVIDIA estimates that using GPUs lowers energy cost by more than an order of magnitude for certain applications (NVIDIA, 2011).

Thus, both by providing significant speedups to amenable calculations and lower energy usage, specialized chips can provide large, tangible economic benefits.

3.2 The Disadvantages of Specialized Processors

The freedom that specialized chips have *in the design phase*, allows them to greatly speed up certain calculation. But once a particular design is manufactured, this freedom disappears. The NVIDIA GPU in Table 1 can be used by many more applications than most specialized processors, since the task that GPUs accelerate is high-dimensional vector multiplication[20], which has many applications in science and engineering. Nevertheless, *the vast majority of software cannot be run on a GPU*.

The customization of hardware in specialized chips can also be a problem because it makes programming[21] them idiosyncratic and difficult. For example, IBM, Sony and Toshiba collaborated to develop the Cell specialized processor, which in 2006 founds its first major application in the PlayStation 3 game console. Only three years later, the line was discontinued, in part because of difficulty in programming it (Stokes, 2009). A similar fate awaited the company ClearSpeed Technology, which was founded to build specialized chips with lots of parallelism, but which downsized after three years because of low sales (Wilson, 2008). Because many tasks on mobile phones are run on specialized processors, Apple has hundreds of programmers who work to ensure the compatibility of Apps across iPhone generations.

The extent to which programming difficulty dissuades potential adopters can be seen in the usage of GPUs. Because they are used to render images on computer screens, virtually every computer has a GPU. Nevertheless, the difficultly in programming them has meant that only the most sophisticated, performance-hungry programmers used them for non-graphics purposes. This was such an important barrier to their adoption that special programming languages, e.g. CUDA, have been created to simplify GPU programming.[22] Figure 4 shows the increase in number of applications that have developed over time in response, including examples from manufacturing, finance and weather modeling industries.

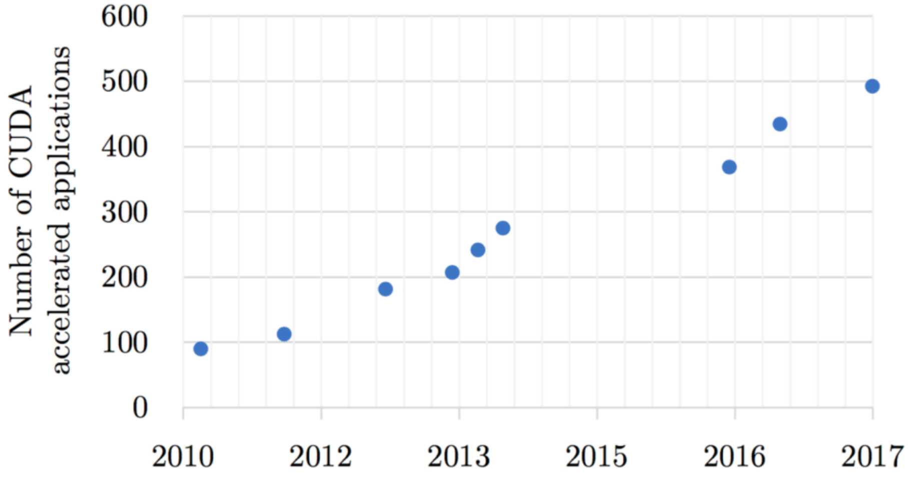


Figure 4: Number of applications implemented on a NVIDIA GPU using CUDA[23]

Thus, even though GPUs had many advantages, e.g. being well-known and ubiquitously installed in computers to do graphics (Steinkraus et al., 2005), ease-of-programming was still a major barrier. For other specialized processors that were not as broadly available as GPUs, this effect would be even stronger.

The disadvantages that we have articulated so far all manifest after a specialized processor chip is developed. But perhaps the biggest drawback of specialized processors are their fixed costs. For universal processors, the fixed costs (also called non-recurring engineering costs (NRE)) are distributed over a large number of processors. In contrast, the market size for specialized processors is very limited. Overall, the cost to manufacture a specialized processor using leading-edge technology is about $80 million[24] (as of 2018). For workloads that can be significantly accelerated by specialized processors, using older generations of technology still provides a performance benefit over universal processors. This can bring the cost down to about $30 million (Lapedus, 2017b). Section 4.1.1 goes into greater detail about how high investment costs influence the trade-off between buying a universal or a specialized processor.

Despite all the advantages of specialized chips articulated in Section 3.1, the disadvantages were large 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 or cryptocurrency mining. Recent years have overturned this trend because of the emergence of one particularly transformative technique, Deep Learning.

3.3 How Specialized Hardware is driving Deep Learning

In 2012, the machine learning community was rocked by the victory of a program, AlexNet, in an important image recognition contest - the ImageNet Large Scale Visual Recognition Challenge ("ImageNet"). In this contest, competing programs analyze visual images and attempt to correctly label them (e.g. 'starfish', 'cowboy hat', 'cauliflower'). Traditionally, the winning programs included approaches such as Support Vector Machines (SVM) or Fisher Vectors. In 2011, the best of these models made classification errors on 25.8% of the contest images[25].

AlexNet was remarkable because it shattered the performance record, achieving an error rate of only 16.4% (Russakovsky et al., 2015) and also because it won with a different algorithm: Deep Learning. Deep Learning is the modern incarnation of the neural networks that originated in the late 1950s (Rosenblatt, 1958), but has more layers of neurons and is thus deeper (hence the name). The neural networks heyday was from the 1980s to mid-1990s, but then interest faded because even the calculations for a modest-sized network were beyond the capabilities of computers of the time (Hof, 2013). AlexNet changed this because it was implemented on a specialized processor (Krizhevsky et al., 2012)[26]. By leveraging the excellent match between GPU's hardware parallelism and big memory and Deep Learning's algorithmic parallelism, a vastly larger network could be trained in the same amount of time, which led to AlexNet's win.

Further optimizations have taken the progress of AlexNet even further, and today 35 times as many images can be trained per second with a GPU than with a universal processor implementation (NVIDIA Corporation, 2016)[27]. The effect has been transformative. Since AlexNet, every winning entry in the ImageNet contest has used Deep Learning, and every one of those has been implemented on GPUs (Sze et al., 2017).

In the years since Deep Learning on specialized chips triumphed in computer vision, it 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). Again, the increase in available computing performance was highlighted as a key factor for the transition of natural language processing theory into real-world applications (Hirschberg & Manning, 2015).

Deep Learning using specialized chips gained rapid and extensive adoption by industry. For example, 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) are all based on such systems. Facebook uses Deep Learning to help with picture tagging, to filter for hate-speech and to customize advertisements to the users (Marr, 2016).

The power of specialized hardware is demonstrated with remarkable clarity in the rise of Deep Learning. 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.

3.4 How Deep Learning Is Driving Hardware Specialization

Deep Learning's success created an enormous appetite from industry for specialized processors. Bill Dally, Chief Scientist with GPU-producer NVIDIA, called it “one of the big killer applications in the datacenter today” (Falsafi et al., 2017). NVIDIA datacenter revenue has more than quadrupled from 2015 to 2017 – and about 50% of those sales can be attributed to Deep Learning (Tanner, 2016).

In 2013, Google looked at anticipated demand and predicted that if people used voice search for three minutes a day, Google's Deep Learning needs would require a doubling of their datacenter capacity (Metz, 2017). Instead of building more universal processor-equipped datacenters or even using GPUs, Google decided to develop an even-more specialized processor, called the Tensor Processing Unit (TPU). Such custom processors are known as Application-Specific Integrated Circuits (ASICs). And where GPUs could accelerate a variety of substantial number of tasks, Google's Tensor Processing Unit (TPU) could only handle neural networks.

Creating a customized processor was very costly for Google, with experts estimating the fixed cost as tens of millions of dollars. And yet, the benefits were also great – they claim that their performance gain was equivalent to seven years of Moore's law (Jouppi et al., 2017) – and that the avoided infrastructure costs made it worth it. So much so that, in 2017, Google released a second-generation TPU that is eight times as fast as leading-edge GPUs (as measured by how long it takes to train a large-scale translation model).[28]

Google is not the only one investing heavily in Deep Learning hardware, there are also efforts at Facebook collaborating with Intel on their Neural Network Processor, Lake Crest. While the majority of the computationally expensive tasks are performed at datacenters, cell phone chip manufacturers have started to dedicate part of the highly constrained space on mobile chips to Deep Learning. Amongst the pioneers are Apple's iPhone X and Huawei's Mate 10 (Vincent, 2017).

Start-ups are also getting into the action. Estimates are that at least 12 companies have been formed to develop chips for artificial intelligence ("12 AI Hardware Startups", 2017). In 2016, venture capitalists invested about $113 million in these type of companies, three times as much as the previous year (Martin Giles, 2017).

3.5 The State of Specialized Processors Today

Computing platforms are changing. The industry is moving away from PCs towards mobile devices and datacenters. In 2017, five times as many smartphones were sold as PCs[29] (Gartner, 2017)(Wong et al., 2017), and a major semiconductor manufacturer expects high-performance computing to be their fastest growing segment (Feldmann, 2018a). There is also a broad consensus that the internet of things (IoT) will be a key source of future growth. In all of these computing platforms, PCs, mobile, datacenters and IoT, specialized processors are dominant or growing.

Mobile chips are so-called ‘systems on a chip’ that integrate multiple processors in one package. This allows for a combination of a universal processors and specialized processors. Because mobile phones are constrained by battery life, much of a smartphone chip has been specialized to provide better energy efficiency (Shao et al., 2014). This has progressed to the point where, today, universal processors cover less than 20% of the area of an iPhone chip, while the rest of the processing space is allocated to specialized circuits, such as a graphics processor, motion coprocessor and, for the latest generation, a neural engine. A simple regression analysis that we conducted about the trends across the past eight iPhone generations suggests that the percentage of the chip that is universal is holding constant or falling[30].

IoT devices are probably even more sensitive to power concerns than mobile phones, and much of the end-user hardware, such as the sensors and RFID-tags for data collection, is based on specialized hardware (Cavin et al., 2012) (Eastwood, 2017).

As we already reported in section 3.4, there is a significant rise in the use of specialized processors by Google and other data centers / cloud providers. This has also happened in supercomputing which shares many of the same trends as data centers but has better public data reporting, and thus is easier to analyze.

Until 2010, only a handful of the world’s 500 best supercomputers used specialized processors (Top 500, 2018). Figure 5 shows the percentage of supercomputers added to the Top 500 List (the list of the world's 500 most powerful computers) each year that were equipped with specialized processors. Regression results affirm what is also clear in the figure, there is a highly statistically significant increase of 3 percentage points per year in the share using specialized processors[31]. In line with this analysis, in 2018, for the first time the performance added by specialized processors was greater than the performance added by universal processors (Feldmann, 2018b).

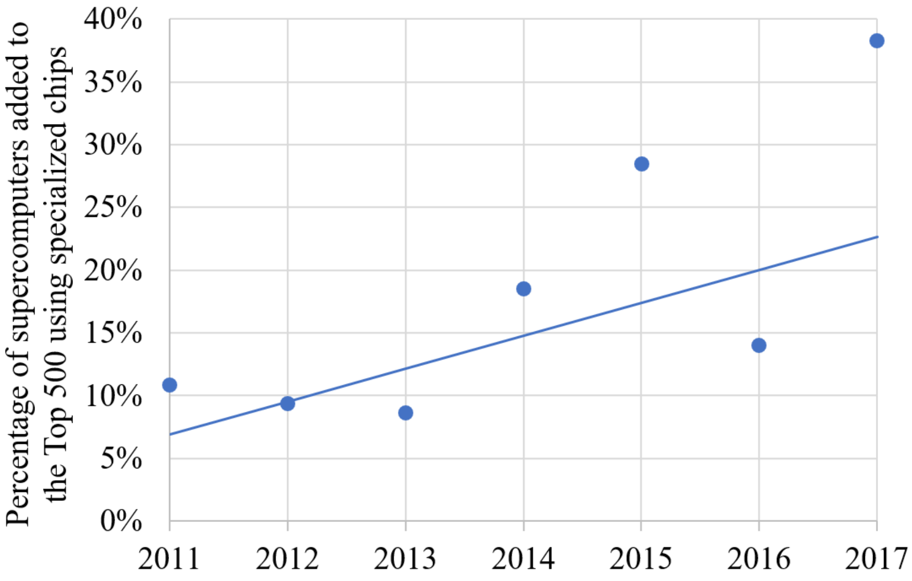


Figure 5: Share of supercomputers added to the Top 500 list using specialized chips

We also performed a regression-analysis of how increasing use of specialized chips in supercomputing is changing energy efficiency. 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[32].

Of all of the computing platforms, PCs remain the most universal, probably because they need to perform a greater variety of tasks and power concerns are less critical. Because PCs represent the most conservative case, they are the base case that we consider throughout this paper. But, if anything, other platforms are even farther along in the move to specialization.

With all of the major computing platforms dominated by specialized processors or moving towards them, it becomes important to ask what effect this will have on the economics of producing universal processors. We argue in the next section that it will undermine it, fundamentally changing the direction of computer hardware advancement.

4 The Fragmentation of a General Purpose Technology

The virtuous cycle that underpins general purpose technologies comes from a mutually reinforcing set of technical and economic forces. Unfortunately, this mutual reinforcement also applies in the reverse direction: 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, as we will show, it has the potential to fragment the general purpose technology, leaving a set of loosely-related technologies advancing at different rates.

The fragmenting cycle has three parts:

- Fewer new users adopt
- Financing innovation is harder
- Technology advances slow

Figure 6 shows how these parts relate and how they change the virtuous general purpose technology cycle identified by Bresnahan and Trajtenberg (1992), shown in (a), to a fragmenting cycle, shown in (b):

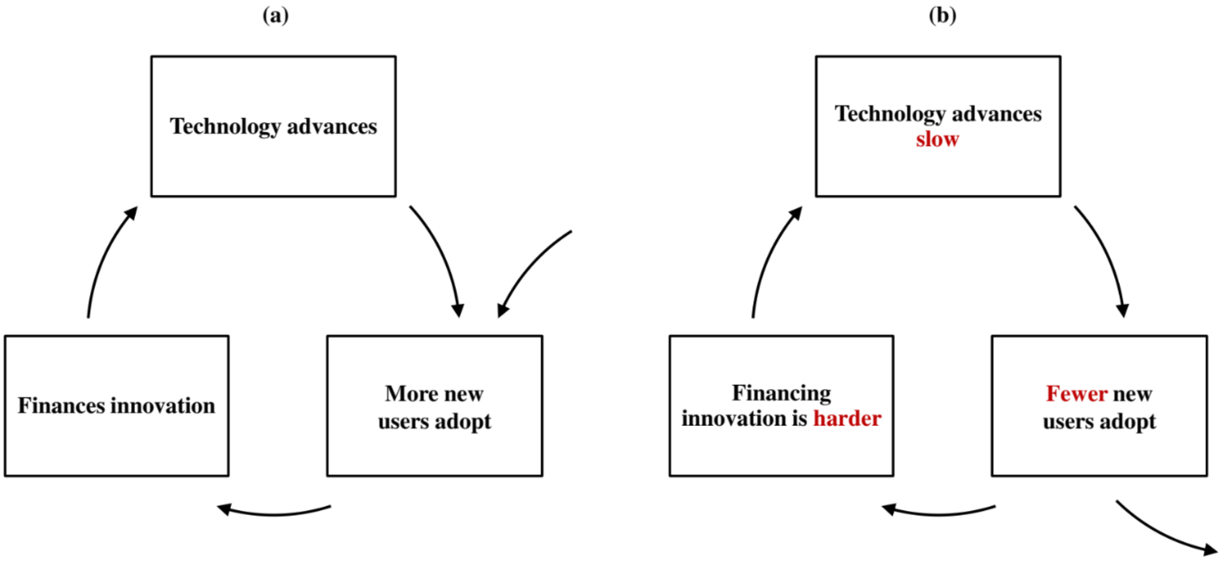


Figure 6: The historical virtuous cycle of universal processers (a) is turning into a fragmentation cycle (b)

In addition to these parts of the fragmentation cycle, external forces also play a role, either augmenting or weakening parts of the cycle. As we describe below, it is changes in these external forces that have pushed computing from a virtuous cycle to a fragmenting one.

4.1 Fewer New Users Adopt

The reduction in the number of new adopters for universal processors comes from two sources: slowing demand from those using the universal processors and movement of users away from universal processors to specialized ones.

Slowing demand from those using the universal technology is not surprising because, as we will show in section 4.3.2, there has been an enormous fall-off in performance improvements for universal processors. Under such conditions, we would expect customers to replace their computing devices less often. In 2016, Intel CEO Krzanich has confirmed exactly this, saying that the replacement rate has risen from every 4 years to every 5-6 years (Krzanich, 2016). Sometimes, customers even skip multiple generations of processor improvement before it is worth updating (Patton, 2017). This same trend manifests itself in smartphone sales. In 2014, U.S. consumers were upgrading every 23 months. In 2018, they were waiting an average of 31 months (Martin & Fitzgerald, 2018).

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. As already shown in Section 3, specialized processors are gaining market share, revealing that customers find them increasingly attractive. Deep Learning has been a noteworthy contributor to this because it has been so successful across so many domains. To explore why there has been a movement of users out of universal processors to specialized ones, we model the choice that consumers make between specialized and universal processors.

4.1.1 Modelling of processor choice - Intuition

A consumer choosing between a universal processor and a specialized one[33] must decide which will provide the best performance at the lowest cost. Figure 7 shows three scenarios of how processor performance could evolve, each showing a variation of two key parameters: the size of the initial jump in performance that comes from specialization, and the rate of improvement of the universal processor (which, over time, erodes or eclipses the gains from that performance jump). In this figure, we assume that T is chosen so that 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 a superior performance profile for one also implies that it has a superior performance-per-dollar profile. This is also why we depict the specialized processor as having constant performance over this period.

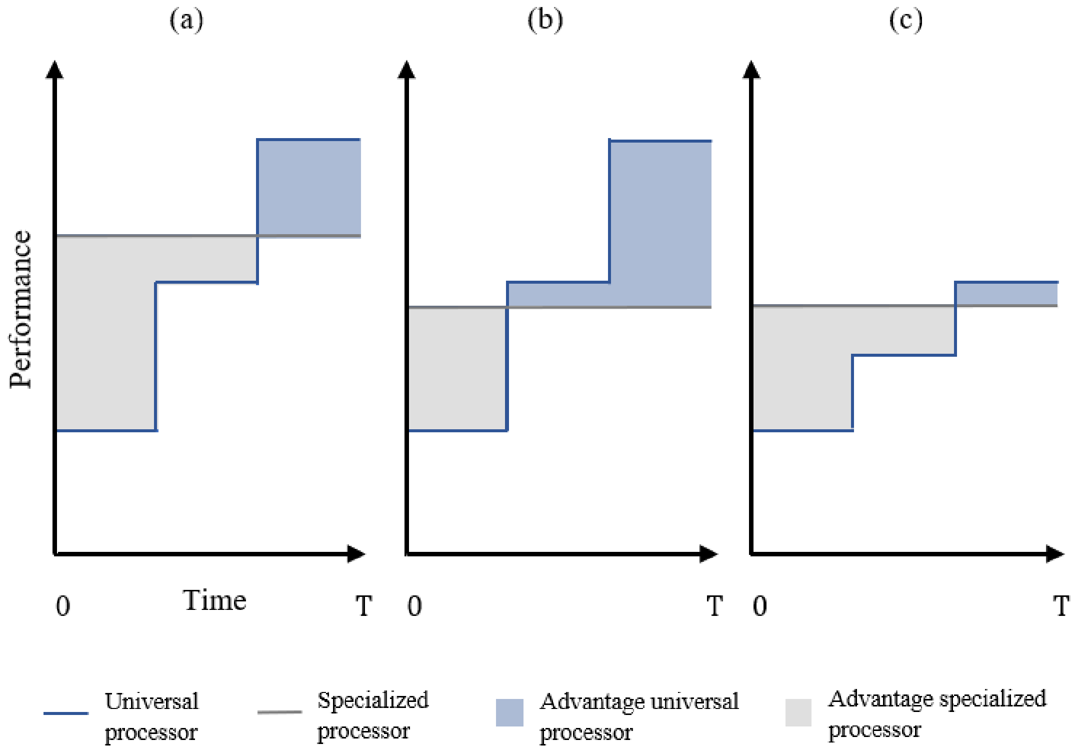


Figure 7: Optimal processor choice depends on the performance speed-up that the specialized processor provides, as well as the rate of improvement of the universal technology. (a) Big performance jump from specialization, (b) Small performance jump from specialization, (c) Slower improvement in the universal processor

In **Figure 7**, a specialized processor is more attractive than 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).

4.1.2 Modelling Processor Choice – Formal Model

Consider two types of processors, universal and specialized, that deliver performance[34] P_u, and P_s at prices[35] π_u and π_s. The performance of the specialized processor is higher initially, but so is its price. This means that, for the price of one specialized processor, a user can instead buy a sequence of improving universal processors (with the first period's performance being denoted as P_{u,t_0}, and each universal processor being used for m years). Which choice is better depends on: the relative performance $\left(\frac{P_s}{P_u}\right)$, the annual performance improvement rate of the universal processor[36], λ, (which we typically represent via $r = \ln(1 + \lambda)$) and how long the specialized processor needs to be used in order to be cost-equivalent to the sequence of universal processors (T).

If we model chip updating as continuous[37] (to avoid unnecessary and uninsightful complications from discretization), then specialized chips are preferred when the time integral of performance from the specialized chip is greater than that of the universal chip, i.e. when

$$\int_0^T P_s\, dt \geq \int_0^T P_{u,t_0} e^{rt}\, dt \tag{1}$$

Solving for the integral and re-arranging yields:

$$P_s\, T \geq P_{u,t_0} \left(\left[\frac{e^{rt}}{r}\right]_{t=T} - \left[\frac{e^{rt}}{r}\right]_{t=0} \right) \tag{2}$$

$$\frac{P_s}{P_{u,t_0}} \geq \frac{1}{T} \left(\frac{e^{rT}}{r} - \frac{1}{r} \right) \tag{3}$$

$$\frac{P_s}{P_{u,t_0}} \geq \frac{e^{rT} - 1}{Tr} \tag{4}$$

We’re interested in the cutoff point between when specialized and universal processors are preferred, which is when equation (4) holds at equality (also substituting to express the formula in terms of the annual improvement rate of the general purpose processor, λ):

$$\frac{P_s}{P_{u,t_0}} = \frac{e^{\ln(1+\lambda)T} - 1}{T\ln(1+\lambda)} \tag{5}$$

Where T is the number of years that the specialized chip needs to be used to have the same cost as the sequence of universal chips. Thus T is equal to the number of universal chips that can be purchased for the same cost, $\frac{\pi_s}{\pi_u}$, multiplied by how many years each universal chip is used, m. If we assume equal markups, price can be re-written as $\gamma(VC_c + FC_c)$, where γ is the markup, VC_c is the variable costs and FC_c is the unit share of fixed costs. We further assume, based on interviews we conducted with hardware experts, that the variable cost, κ, of specialized and universal chips are the same[38]. Analysis of Intel's cost-base suggests that the variable and fixed cost contributions for universal chips are roughly similar (discussed further in Section 4.2.2), so we assume this as well. Finally, for the fixed cost contribution for specialized chips we substitute in the total fixed cost, TFC_s, divided by the number of specialized chips produced (N_s). Together these assumptions yield:

$$T = m * \frac{\pi_s}{\pi_u} = m * \frac{\gamma(VC_s + FC_s)}{\gamma(VC_u + FC_u)} = m * \frac{\kappa + \frac{TFC_s}{N_s}}{\kappa + \kappa} = m * \frac{\kappa N_s + TFC_s}{2\kappa N_s} = \frac{m}{2} * (1 + \frac{TFC_s}{\kappa N_s}) \tag{6}$$

Thus, in Equations (5) and (6) we have derived a simple function for the cutoff between chip types that only requires information on the performance gain from switching to a specialized chip $\left(\frac{P_s}{P_{u,t_0}}\right)$, the number of specialized chips that will be ordered (N_s), and the annual improvement rate of the universal processors (λ). This equation does not have a simple analytical solution, but we will estimate it numerically in the following section.

4.1.3 Implications for the fragmenting of the general purpose technology

In the analysis above, it was clear that the higher fixed costs of specialized chips requires that that they be amortized over a longer period than the typical ownership for a universal chip. During this period, the extent to which the specialized chip is a better choice for consumers depends on whether the performance of universal chips overtakes them and therefore depends on the rate of progress for universal chips, λ. The Bureau of Labor Statistics (BLS) estimates that λ=48% from 2000-2004[39]. This is the case when universal processor improvement was at its highest. In recent years this annual improvement rate has slowed considerably to 29% in 2004-08, and to 8% in 2008-13 (BLS, 2018).

κ is approximately \$50[40], TFC_s is approximately \$30 million (Lapedus, 2017b), and m is approximately 4 (Krzanich, 2016). Substituting these into Equation (5) allows us to plot the volume of chips used and speedup from specialization (Figure 8).

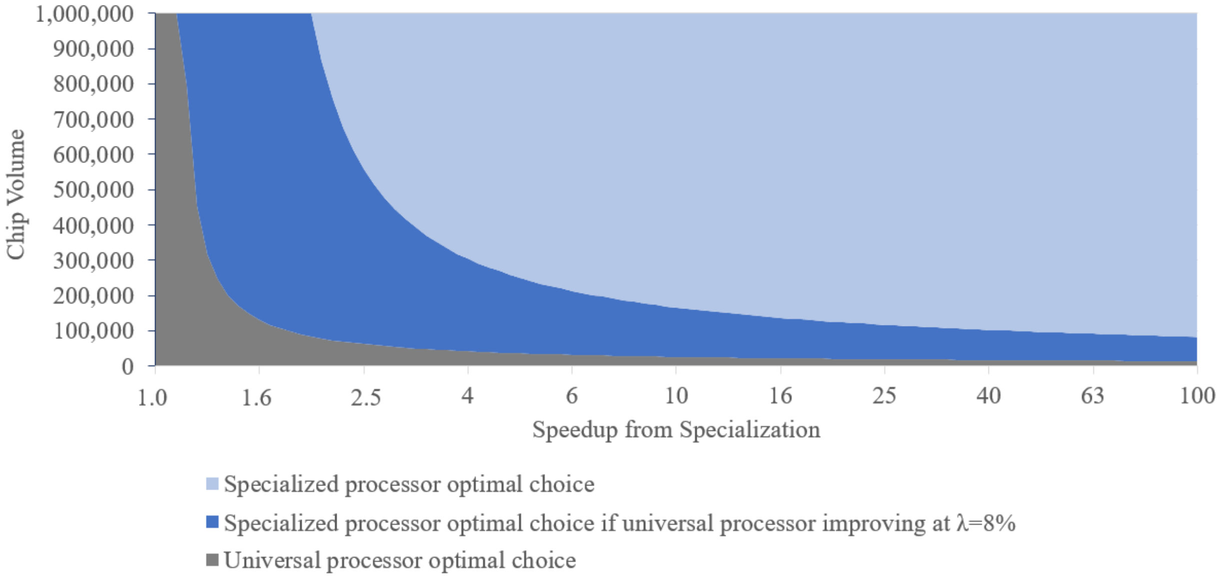


Figure 8: Optimal processor choice depends on performance ratios and volume

As we can see, speedup and volume both strongly impact whether specialized processor chips are attractive. At the peak of Moore's Law, when λ=48%, if specialized chips were 100x faster than universal ones, i.e. $\frac{P_s}{P_u} =$

100 (a huge difference), then at least ~83,000 had to be produced for the investment in a specialized processor to pay off. While that volume of chips is realizable for some applications, only a few would be able to achieve such large speedups from specialization. A speedup factor of 10 is more realistic, as was observed for a range of applications that benefited from Google's TPU (Sato et al., 2017)[41]. For a speedup of 10x, at least ~167,000 chips are needed to make specialization attractive. At the other extreme, if the performance benefit were only 2x, it would take ~1,000,000 chips to make specialization attractive. These results make it clear that during the heyday of Moore's Law, when universal processors were improving rapidly, market demand would have to have been hundreds-of-thousands of chips for specialized chips to be attractive.

As we saw in Figure 7, specialized chips become more attractive if universal chips improve more slowly. Thus, we repeat our processor choice calculations, but using $\lambda = 8\%$, the improvement rate from 2008-2013. For applications with 100x speed-up, the number of processors needed falls from 83,000 to 15,000, for those with 10x speed-up it drops from 167,000 to 27,000, and for those with 2x speed-up it drops from 1,000,000 to 81,000. Thus, for many (but not all) applications it will now be economically viable to get specialized processors – at least in terms of hardware. Another way of seeing this is to consider that during the 2000-2004 period, an application with a market size of ~83,000 processors would have required that specialization provide a 100x speed-up to be worthwhile. In 2008-2013 such a processor would only need a 2x speedup.

Thus far, this analysis has ignored the cost of re-developing code to run on a specialized processor. If one assumes a re-development cost of $11 per line of code[42], then a 100,000 line program requires an extra $1 million in fixed costs to specialize, whereas a larger program like MS Office would cost about $500 million to re-write its about 45 million lines of code (Dvorak, 2013), assuming such an adaptation was even technically possible.[43] Figure 9 shows the effects that such additional costs would have on the trade-off for choosing a specialized processor:[44]

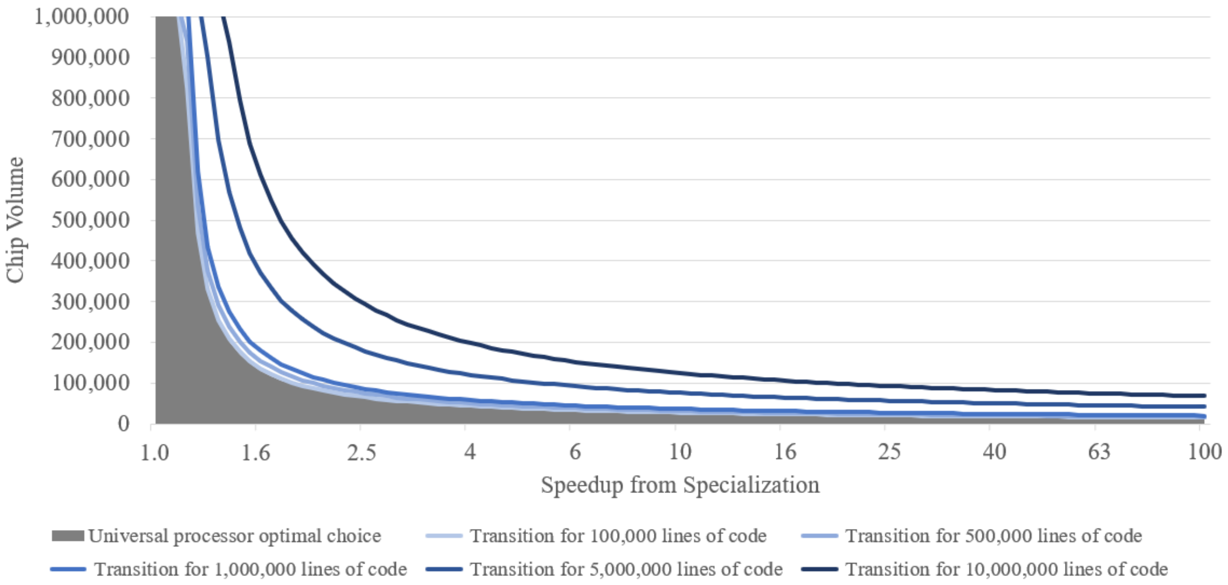


Figure 9: Added friction due to code re-development cost

Notice that the cutoffs rise substantially. Importantly, the cost of code re-development is a friction – it inhibits movement in either direction. Thus, code re-development also makes it makes it more likely for those who switch to specialized processors to stick with them.

The move to specialized processors undermines the general purpose technology 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.[45,46]

4.2 Financing Innovation is Harder

4.2.1 External forces

The fixed cost of chip manufacturing is enormous. The Semiconductor Industry Association (2017) estimates that the cost to build and equip a fabrication facility ('fab') for the next-generation technology is roughly $7 billion. By next-generation here, we mean the next miniaturization of chip components (often referred to as the next process 'node'). It is this miniaturization that has underpinned much of the hardware improvement since the 1960s (Leiserson et al., forthcoming).

The costs invested in chip manufacturing facilities must be justified by the revenues that they produce. Perhaps as much as 30%[47] of the industry's $343 billion annual revenue (2016) comes from cutting-edge chips. But while revenues are substantial, costs are growing. In the past twenty-five years, the investment to build leading-edge fab (as shown in Figure 10) rose 11% per year! Including process-development cost increases into this estimate further accelerates cost increases to 13% per year (as measured for 2001 to 2014 by Santhanam et al., 2015)[48].

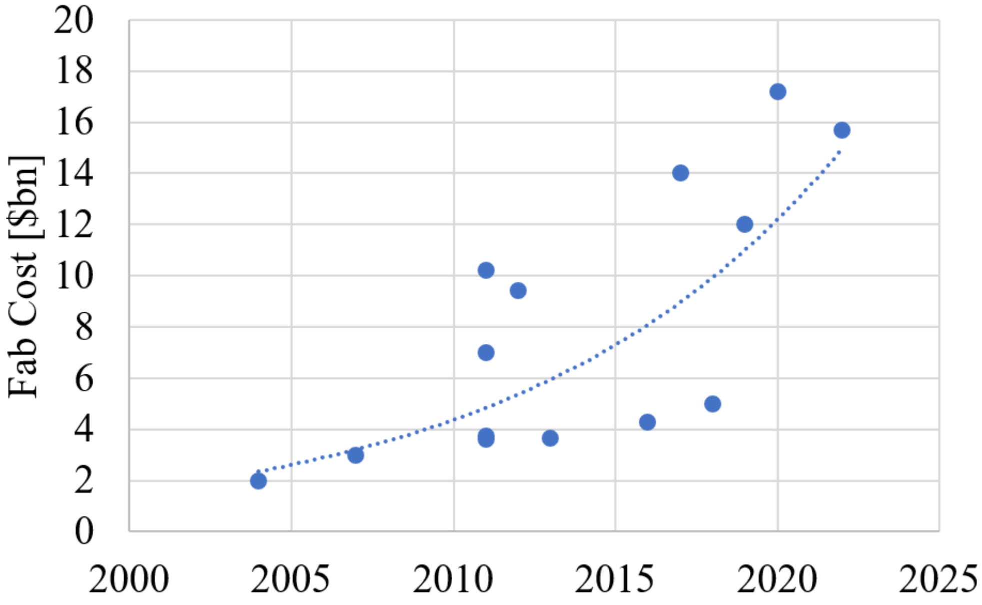


Figure 10: Leading-edge fab costs over time

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[49] (SIA, 2017)), which allowed semiconductor manufacturers to amortize fixed costs across greater volumes.[50] 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 11 shows, there has indeed been enormous consolidation in the industry, with fewer and fewer companies 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 [51]). And GlobalFoundries recently announced that they would not pursue development of the next node (Dent, 2018).

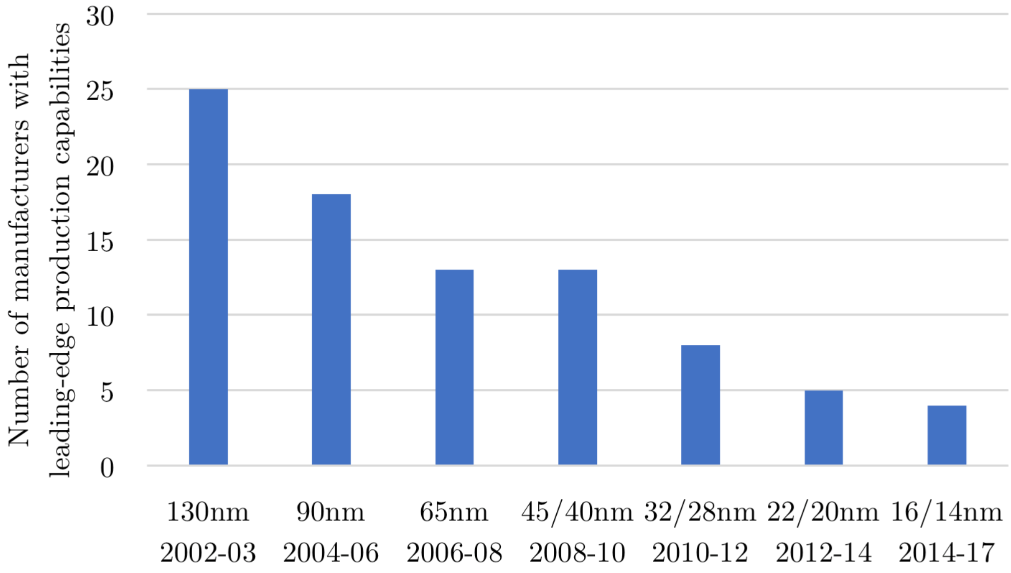


Figure 11: Number of manufacturers with leading-edge production capabilities by node size and year (Smith, 2017)

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% ($\frac{100\%}{26}$) in 2002/2003 to 25% ($\frac{100\%}{4}$) in 2014-2016. Expressed as a compound annual growth rate, this would be 14% - more than enough for an average remaining manufacturer to overcome rising costs. That is, market share growth coming from consolidation was sufficient to offset the worsening economics. Of course, in practice, the market was far from evenly divided because of Intel's dominant share. As a result, Intel should have been less able than smaller players to control fixed cost growth this way (although they would have been starting from a lower base anyway, and so would have remained highly competitive).

However, consolidation can only proceed for so long. And if we project forward current trends, then 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 revenues (see endnote for details[52]). 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.

4.2.2 Current state of semiconductor manufacturing economics

As rising fixed costs worsen the overall economics of semiconductor manufacturing, it is worth considering the effect this is having on Intel specifically, since traditionally the majority of Intel's revenue comes from selling leading-edge processors. In 2017, Intel spent $25 billion (almost 40% of their net revenue) on new fabrication facilities (as measured by property, plant and equipment expenditures) and on R&D (Intel Corporation, 2017).

Figure 12 illustrates how Intel's fixed costs have risen in comparison with their variable costs (as measured by cost of goods sold). We use variable costs, rather than revenues, for our denominator because it allows us to evaluate our concern about rising fixed costs without worrying about changes in pricing behavior (for example those that could arise from the market concentration mentioned above).[53]

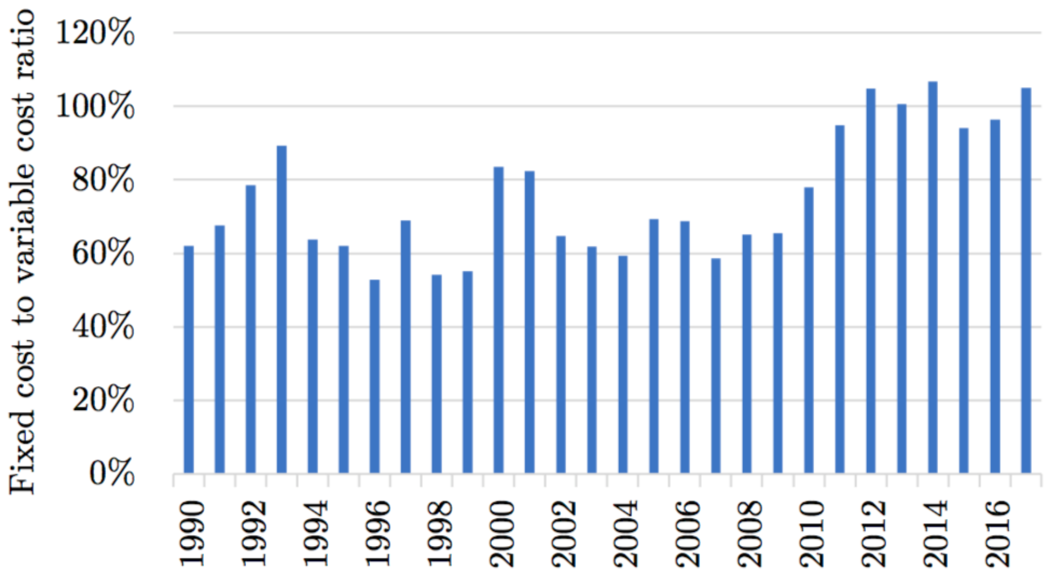


Figure 12: Intel's fixed cost to variable cost ratio over time

Over the past decade, Intel's fixed costs have risen from 60% of their variable costs, to over 100%. 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.

4.2.3 Implications for the benefits of advancing chip technology

Faced with rising fixed costs and slowing market growth, chip manufacturers face tough choices – all of which hurt the performance-per-dollar provided by universal chips.

One option for dealing with rising costs would be to raise prices. Intel announced raising average selling prices as one key component to maintain high profit margins (Flamm, 2018). A second option for dealing with rising fixed costs would be to amortize them over more chips by delaying the issuance of the next generation of processors with smaller node sizes. Intel seems to be pursuing this option as well. In 2016, Intel replaced their cycle of providing a smaller node sized processor every two years with a three-year cycle and has already announced that they will continue pursuing this slower timeline for the successive generation (7nm).

Consistent with this approach for dealing with rising fixed costs, when Intel switched to this new model they extended the depreciable life of their equipment from four to five years (Wong et al., 2016).

In Figure 13, we consider how Intel's fixed costs might have risen had they kept to a higher pace for introducing smaller node-sized processors. The red bars represent a rough calculation of the additional cost that Intel would have incurred had they introduced processors at a higher pace. For example, if Intel spent $20 billion on R&D and CAPEX per year when they were on a 3-year cycle, then to calculate their spend on a 1.5-year cycle we would multiply it by $\frac{3}{1.5}$, i.e. 2x, to arrive at $40 billion per year[54]. We use 1.5 years as the baseline because it was the fastest introduction cycle that Intel had over this period (in 1993 and 2000), and thus makes year-to-year comparisons easiest. This normalization shows an even more dramatic rise in fixed costs if this cycle-adjustment is added back.

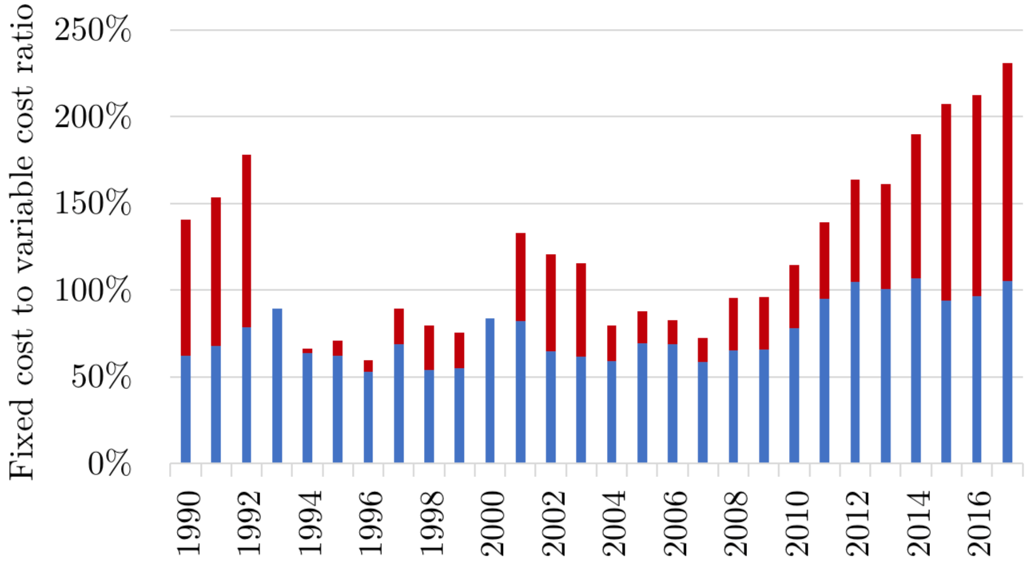


Figure 13: Effect of Intel prolonging the time until the introduction of the next generation of processors on the fixed cost to variable cost ratio

In theory, Intel's adoption of a slower introduction cycle for smaller-node sized processors does not need to decrease λ, the rate of universal processor improvement, if the subsequent steps they take are larger and thus provide equivalent benefit. This is precisely what Intel and others are claiming, but this not supported by the data, as we show below. But, even if it were true, longer gaps between the introduction of new smaller-node size processors might lead consumers to replace their computers less often, undermining the key amortization goal.

4.3 Technology Advancements Slow

To measure the rate of improvement and competitiveness of processors, there are two key sets of metrics that are important. One is overall performance, the other is performance-per-dollar. In this section, we show that improvement in both has slowed dramatically.

4.3.1 External forces

As already hinted at in the last section, an enormously important source of performance improvement has been the miniaturization of transistors. Indeed, this forms the basis for Moore's law, probably the most famous description of the improvements in computers (see Thompson 2017 for a more complete discussion).

While miniaturization continues today, the benefits that it provides have fallen enormously since 2004/2005 because of technical challenges. Manufacturers are hitting the physical limits of what existing materials and designs can do (Shalf & Leland, 2015). And these limits take ever more effort to overcome (Bloom et al., 2017). This external force of slowing improvements of universal processors performance is perhaps the most important external cause of the switch from a general purpose technology cycle to a fragmenting one.

The scale of this change can be seen, for example, in one widely accepted benchmark suite: SPECint, which is designed to test performance of computer systems when handling compute intensive workloads. Hennessy and Patterson (2017) show that, by this measure, over 1985-2005 universal computer performance improved 52% per year. Put another way, the compounding effect of this improvement means that every 5 to 6 years, universal processors improved 10x. Around 2004/2005 this changed dramatically. Prior to that time, processors had gained enormously from a long-term trend called Dennard scaling, which meant that as transistors were miniaturized, computers could be run more quickly.[55] Running processors faster provided widespread gains, and no longer being able to run them faster caused widespread diminishment of those gains (Thompson, 2017). After 2005, the rate of performance improvement dropped to 22% per year, roughly doubling the time needed for a 10x improvement.[56]

A second slow-down to the improvement of universal processors is expected around 2020-2025, as Moore's Law comes to an end (Leiserson et al., forthcoming). The magnitude of this effect is also likely to be large.

4.3.2 Current state of performance-per-dollar of universal processors

The slowdown in chip improvement also happened in performance-per-dollar. The U.S. Bureau of Labor Statistics (BLS) publishes a producer-price index that explicitly tries to account for both price and quality changes. As can be seen in Figure 14, they find that improvements in performance-per-dollar have dropped from 48% annually in 2000-2004 to 8% annually in 2008-2013.[57]

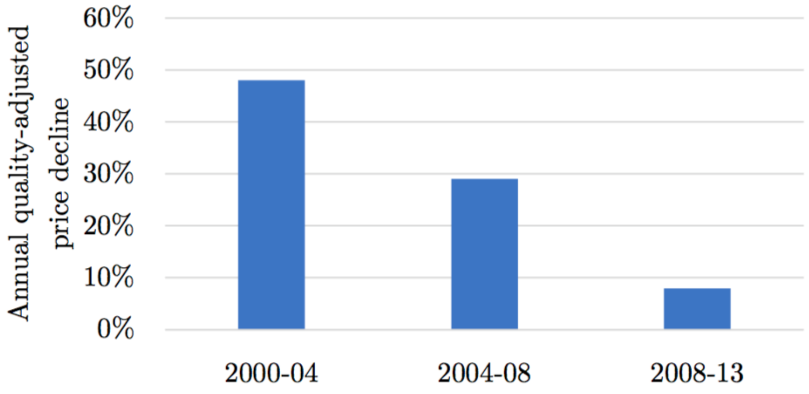


Figure 14: Quality-adjusted annual price declines of microprocessors (BLS, 2018)

With such a remarkable slow-down in performance and performance-per-dollar, it is not surprising that we see the drop in adoption discussed in section 4.

4.3.3 Implications for the attractiveness of specialized processors

The improvement of universal computers has already slowed considerably from one tectonic shift in computing (when the benefits of node miniaturization started to diminish at the end of Dennard Scaling) and is likely to do so again as another approaches (the end of Moore's Law). As the model in Section 4.1.2 shows, these declines (denoted as λ) make specialized processors more attractive, because they decrease the opportunity cost of universal processors.

One might imagine that deteriorating benefits from new processors might hurt specialized processors, not just universal ones. However, for specialized processors it is often not cost-optimal to use the latest manufacturing technology (Khazraee et al., 2017), and thus they are much less affected.

An important reason why specialized processors use old technologies is cost. Samsung and IC Knowledge Market Research believe that (in contrast to historical trend) the cost per transistor is now *increasing* (Low, 2016).[58] Qualcomm agrees, arguing that we are now paying a premium for diminishing performance improvements, whereas the cost-minimizing size might be 2011's 28nm technology (Arabi, 2014).

This increase in sales for older-generation chips can been seen to some extend in the percentage of TSMC's revenue attributed to different generations, as shown in Figure 15. But this likely considerably understates the extent of this change, since (as discussed in section 4.2.1) TSMC remains one of the three manufacturers that continue to make leading-edge chips. By definition, all manufacturers but those three are now working only on older-generation chips. In fact, market demand for older generations has surged so much that there is a shortage of manufacturing equipment for the old technologies (Lapedus, 2017a).

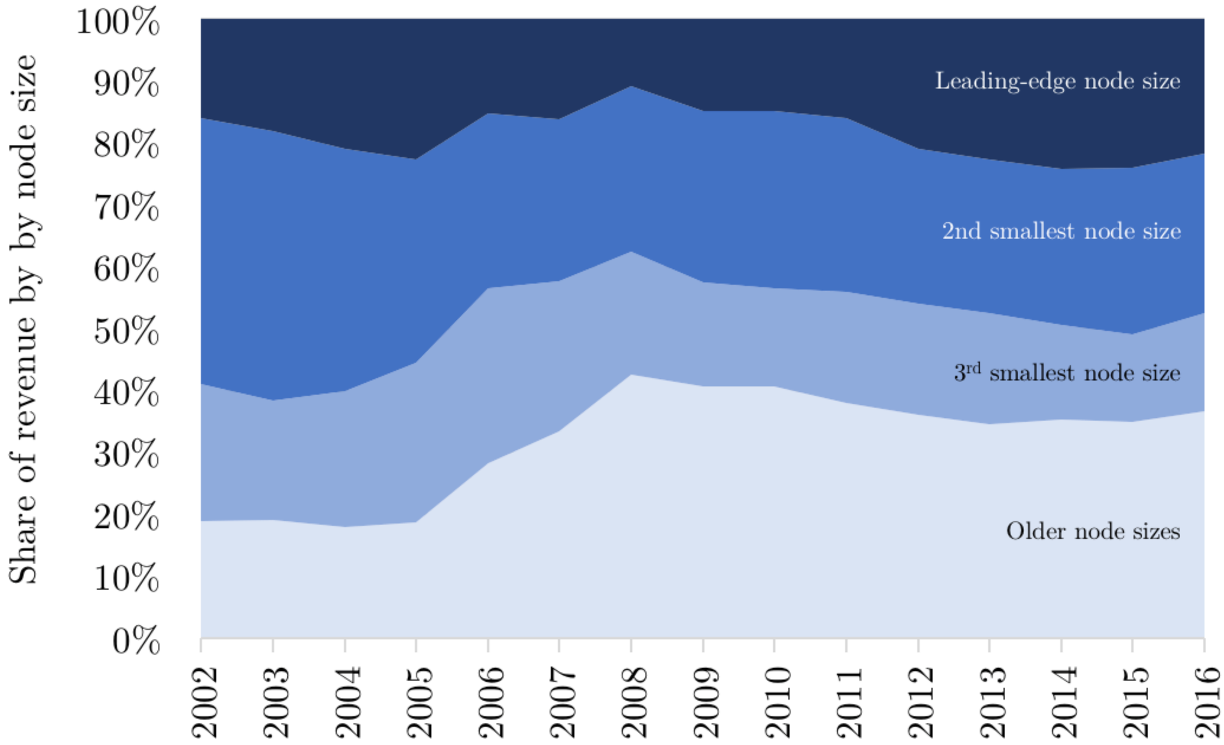


Figure 15: Share of revenue attributed to node size over time. Leading-edge node size starts at 350nm in 2002 and concludes at 16/20nm in 2016. Percentage is a three-year moving average around the respective year to minimize effects of the introduction phase of a smaller node size

Industry experts implicitly confirmed this shift 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 their 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 this paper is arguing is now happening.

4.4 The Cycle

We have argued that there exists a mutually-reinforcing fragmenting cycle, wherein universal processors have fewer new adopters, which worsen the financing of innovation, slowing improvement in universal processors, leading to less new adoption, and so on. This cycle leads to an iterative slowing of universal processor improvement and ever-greater attractiveness of specialized chips. Figure 16 illustrates the induced deteriorating spiral.

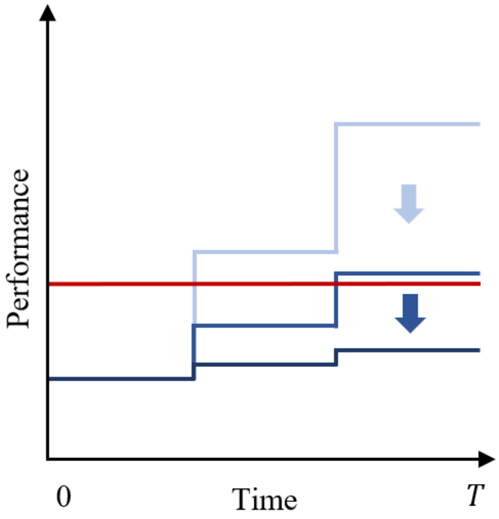


Figure 16: Self-reinforcing cycle of users favoring specialized processors. The more users switch to specialized processors, the less are left to finance advances in universal processors

We are already observing that computation is starting to fragment. If our argument is correct, we would expect a future where different applications become siloed, being run on different processors with different software and algorithms. In this world, some applications would get massive investment and become orders of magnitude more efficient than they could be on universal processors (as with Deep Learning today). In contrast, other applications are likely to fare much worse, garnering little investment in specialized processors. And, since these areas no longer benefit from fast-improving universal processors, their performance could stagnate. The next section explores these issues.

5 The Fragmentation of Computing

5.1 Universal processors become a mature technology

Universal processors have been the superior computing platform for decades, but recent slow-downs of improvements in performance-per-dollar may be a sign that technology is maturing. The final performance improvements that can come from miniaturization will be at a high price premium, and are only likely to be paid for by important commercial applications such as data centers (Johnson et al., 2017). 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. But many of the interviewees that we contacted for this study doubt whether it will be economically worthwhile miniaturizing past 7nm.

If the existing technology is now too mature to advance at historically high rates, might another technological improvement restore the pace of universal processor improvements? Certainly, there is a lot of discussion of exotic technologies: quantum computing, carbon nanotubes, optical computing. Unfortunately, it seems unlikely that these will produce major gains in the near term.

For example, quantum computers are often mentioned as the next universal computing platform. In reality, the technology is far away from achieving this goal. Google claims that their superconducting qubits system can outperform classical computers (a.k.a. "quantum supremacy") (Neill et al., 2017). But this only holds true for a very narrow set of tasks answering specific questions in physics, chemistry, and cryptography. 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 (Prickett Morgan, 2017).

Other technologies might hold broader promise, but these would most likely need strong public financial support to become production-ready. Slowly, governments are acting. For example, in 2017 DARPA announced the 'Electronics Resurgence Initiative' which provides $216 million funding for microelectronic technology research (DARPA, 2017). Highlighting the severe lack of research ideas to move forward, they increased the funding to $1.5 billion in mid-2018. While laudable, that entire initiative's budget is less than 10% of Intel's R&D budget, and thus it represents only a small fraction of what we should be investing. At the same time, China announced at $47 billion fund to support the local semiconductor industry to catch up to the leading-edge technology, and potentially, move beyond. (Kubota, 2018)

5.2 More and more users switch to specialized processors

Based on our analysis in Section 4.1.3 it is clear that specialized processors will be adopted by those (i) that will get a large speedup from moving to specialized chips, and/or (ii) where there will be enough demand for the volume of specialized chips needed to justify the fixed costs.

In this context, it is perhaps not surprising that Google was one of the first of the recent wave of adopters of specialized chips with their Tensor Processing Unit (TPU) for Deep Learning. Deep Learning gets large benefits from specialization because of the high parallelism in the algorithm, and Google's internal usage represents a sufficiently large market to justify the investment in time and money. And this was substantial. It took Google 15 months to develop the first TPU version from scratch (Jouppi et al., 2017) and is estimated to have cost them tens of millions of dollars. For similar reasons, other large companies are also moving towards specialized chips, for example Microsoft (Putnam et al., 2016), Baidu (Hemsoth, 2017), and Alibaba (Pham, 2018).

But coordination of those already using an algorithm is only part of the demand that will exist for specialized chips. Once these chips exist and can provide large benefits, application designers that currently use other algorithms will try to adapt their algorithms to use the new hardware (recall, this is how Deep Learning gained prominence). And, since continued moves to specialized chips will themselves generate further movement to specialization (via the fragmenting cycle), we expect an increasing share of the algorithms that are technically amenable to be re-written to run on specialized chips.

We show this schematically in Figure 17, where each rectangle represents an application, and its size represents the number of users. Filled rectangles represent specialized processor adoption. Applications adopt if they have enough market size and speed-up. Thus, chronologically from left to right we first see GPUs used for graphics. Over time, they were adopted by other applications, likely because they share an underlying algorithm, e.g. now also being used for Deep Learning. Later, Google develops their TPU which takes over part of the GPU market. Other specialized processors, such as Microsoft FPGAs for networking and search engine requests, emerged as well.

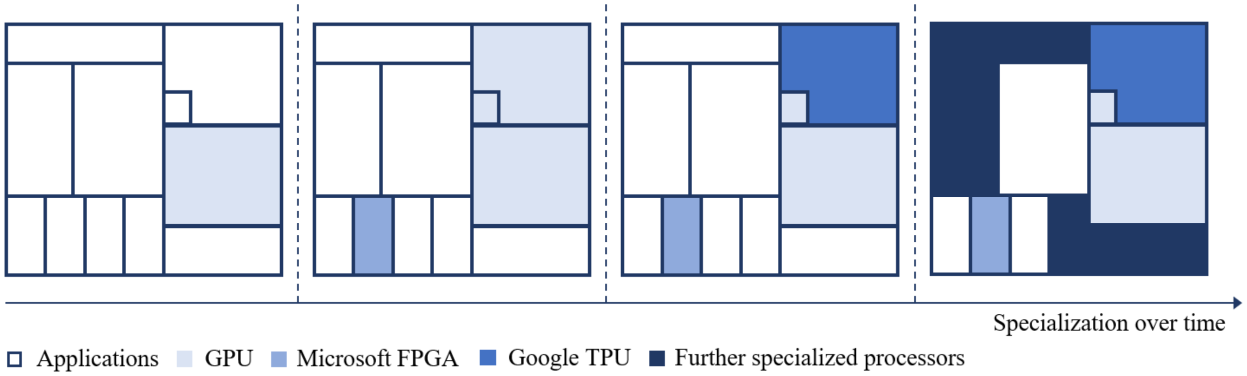


Figure 17: Schematic representation of adoption of specialized processors

5.3 Who gets left-behind

The fragmenting cycle that we have identified in this paper is pushing many applications towards specialized processors. For these applications, specialized processors can provide significant performance gains. But what about other applications? Will there be ones that don't transition to specialized chips? What performance improvements will they see?

Applications that do not move to specialized chips will likely do so because they (i) will get little speed-up from them, (ii) do not comprise a sufficient market to justify the upfront fixed costs, or (iii) cannot coordinate their demand.

These applications will get 'left-behind'. Not only will they not get their own, improved processors. But the improvement from universal processors, that they rely on, will diminish. And so, they will end up in the slow lane of computing while those that move to specialized chips get to be in the fast lane.

5.3.1 Left-behind: Applications whose current algorithms are ill-suited to specialization

In Section 3.1 we described the characteristics of a calculation that allows for specialized processors to accelerate it. Absent these characteristics, there are only minimal performance gains, if any, to be had from specialization. For some applications, for example databases, this will be a challenge. Over the past decades, it has been clear that a specialized chip for databases could be very useful. Databases are widely used and important for many businesses, and thus there would easily be enough demand from applications like credit card transactions to justify the fixed costs of a specialized chip. And yet, as one expert we interviewed told us, there are no widely used specialized chips for transactional databases. The reason is that the calculations needed for databases are poorly-suited to being on a specialized chip. In particular, they are hard to predict (and thus to build a pipeline around) because transactions are generated seemly-randomly by consumers. It is also dangerous to calculate these transactions in parallel. For example, if two separate $100 bank withdrawal transactions are made on an account with a balance of $125 the right conclusion is NOT that both should be approved (which is what would happen if they were done strictly in parallel).

5.3.2 Left-behind: Applications with insufficient demand

Another important category of those that will not get specialized chips are those where even if the application did switch, there would be insufficient demand to justify the upfront investment. As Goodfellow et al. (2016) have written, and we have now derived, demand on the order of a thousand processors will most likely not be sufficient to finance a specialized processor.

This can matter a lot for those doing intensive computing on a small scale, for example research scientists. One particular example that was emphasized to us in our interviews is the small community of climate modelers. Although there is a theoretical potential to solve more precise climate models on specialized hardware (Wehner et al., 2008), most calculations are still performed on commercial off-the-shelf technology.

Another important reason why an application might have insufficient demand is a lack of stability in the calculations that underpin an application. This is because demand for a specialized chip can be aggregated over many users, but also over time. But if the calculations needed for an application change frequently, then any particular specialized processor design will become obsolete quickly, hindering the ability to spread the fixed costs over time.

5.3.3 Left-behind: Fragmented application users that fail to coordinate

The problem of insufficient demand can also come from cases where, in theory, there is enough demand, but potential users cannot coordinate to aggregate it. This is likely to be of biggest concern to small users, where aggregating across thousands could introduce substantial coordination costs.

We anticipate that cloud computing companies will be important in mitigating this effect, by providing a platform that allows small users to aggregate. Already, we see Google providing the TPU on its cloud (Google Cloud, 2018), and Amazon Web Services (and others) providing GPUs (AWS, 2017).

Less tangible, but perhaps more important in this category of failed coordination, are potential future users. Being willing to invest in a new processor requires the ability to foresee what you would do with it and provide some estimate of the gains that it will provide. Even in interviews that we've done with sophisticated, computing-centric companies, they report enormous uncertainty about whether redesign projects will even work, never mind any precise indication of the likely gains. Instead, we see lots of after-the-fact exploration, where users test out *already-existing* hardware and check to see if adapting to it will provide benefits. This is what happened with Deep Learning using GPUs, and (in a different field) with researchers using Microsoft's Kinect (Teodoridis, 2014). This difficulty in forecasting demand will make it hard to assess market demand and thus will, in our judgment, be a major barrier to the development of specialized chips.

The forecasting difficulty highlighted above arises because users may know that additional performance will be useful but cannot reliably forecast how much performance improvement they will get. A second forecasting difficulty arises because users may not even know that additional computing performance would be useful. For example, many people argue that users don't need additional computing speed. But these same users have unfailingly taken advantages of orders-of-magnitude improvements in processors over the past decades. We would posit that this is because it is hard to see the causal chain that leads from speedups to useful functionality. For example, a performance speedup could lead a designer to realize that a voice recognition algorithm could be adapted to work better. This, in turn, could lead a third-party software developer, say Microsoft, to incorporate it into their program and allow users to react verbally to behavior in Excel. That functionality could be very useful, but it could be very hard for an end user, or even an application developer, to predict from just hearing about a potential processor speed-up. In the words of Steve Jobs, “A lot of times, people don't know what they want until you show it to them” (Mui, 2011).

The important consequence for this lack of visibility of how future performance gains translate into useful functionality is that it makes it much harder to coordinate those that would benefit but don't realize it.

5.4 Welfare Implications

There are both short-term and long-term welfare implications of the fragmentation of computing. In both timeframes it is useful to consider three groups: Always Specializers, Induced Specializers, and the Left-Behinds.

Always Specializers are those users whose applications benefit from hardware specialization and who have sufficient scale that, even if universal computers had continued to improve rapidly, would still have specialized. Users of Deep Learning and Bitcoin miners probably fall into this category. In the short run, these users should get welfare gains because they get access to cost effective performance improvements.

Induced Specializers are users who switched to specialized chips but who would have stayed with universal processors, had they continued to improve, i.e. those induced to switch. Axiomatically, these users are worse off in the short run because their choice of what to do has been constrained and it is forcing them to choose something they would otherwise have thought of as inferior.

The Left-Behinds, which we might also call Always Universalists, are those that won't get specialized chips because of the algorithm they are using, the volume of chips that they would demand, or the difficulty in forecasting or coordinating their demand. In the short run, these users will be significantly hurt by the slowing of progress in universal processors and the ever-deteriorating general purpose technology cycle.

If we consider these three groups together, the short run welfare effect is ambiguous. The overall welfare gains from the Always Specializers may be sufficient to off-set the losses from the Induced Specializers and the Left-Behinds, particularly because some applications like Deep Learning that have broad reach fall into this category. However, because (by definition) Always Specializers would have used specialized chips anyway, their switch is not a result of the GPT slow-down. Taking them out of the calculation produces the expected unambiguous result that a slowdown in the universal technology is detrimental to welfare in the short run.

In the longer-term, all groups may be hurt by fragmentation because it delays the introduction of improved chips (via the GPT fragmenting cycle). Perhaps more importantly, it could also delay the introduction of whatever successor technology to silicon microprocessor could propel computing forward. To understand this, it is worth considering the series of technologies that have underpinned computing: from electromechanical in the early 1900s to relays in the 1930s, to vacuum tubes in the 40s and 50s, transistors in the 60s, and integrated circuits since. In such transitions, a new technology is introduced and the market must adopt it. If such a successor technology had been introduced before computing began to fragment, there would have been strong pressure to make it backwards compatible. And thus the transition would have simply involved each user replacing their hardware in a completely analogous way to how they have adopted new chip designs over the past decades, and their software would just continue to run.[59] This promise of an easy transition to the new technology would provide strong R&D incentives for firms looking for the successor technology to integrated circuits because it would mean that the technology's adoption would likely be widespread and rapid.

Consider instead the outcome if users have already transitioned to specialized chips. In this case, they will have software designed specifically for use on those specialized chips. Moving back to universal processors will require re-writing their software, requiring significant investment. To justify these extra costs, the gains from moving to the universal processor will have to be larger and thus adoption will be slower. And, of course, slower adoption means diminished R&D incentives for firms producing the successor technology, and thus that next breakthrough technology could arrive later.

Thus, while the fragmentation of computing may be good or bad for welfare in the short run, in the long-run any lengthening in the time before we get a successor universal processor technology will diminish overall computing improvement, hampering applications that could have benefited from it and ultimately diminishing economic growth. In such circumstances, welfare would be notably worse in the long term.

6 Conclusion

Traditionally, the economics of computing were driven by the general purpose technology (GPT) model (Bresnahan & Trajtenberg, 1992)(Geiger, 2017), where universal processors (e.g. CPUs) grew ever-better and market growth fueled 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 paper provides evidence that this general purpose technology cycle is being replaced by a fragmenting cycle where computing separates into specialized domains that are largely distinct and provide few benefits to each other.

We show evidence that an application switching to a specialized chip can provide large gains for the applications that use it (e.g. Deep Learning or cryptocurrency mining). We argue, however, that it also worsens the economics of chip manufacturing, particularly for universal processors, leading them to improve at a slower rate. But, if universal processors improve less quickly, they become less attractive and more users switch to specialized chips. Hence the move to specialized chips perpetuates itself, fragmenting the general purpose model and splitting off more and more applications.

Not all applications will benefit from specialized chips. Some applications are too small to be worth the high investment cost needed for such chips. Others do not fulfill the technical requirements needed to for specialization to be valuable. These 'left-behind' applications will not only miss the benefits of specialization, but will also be left with slower performance improvements from universal processors.

The migration of computing from a general purpose technology to a fragmented one will fundamentally alter it. Some applications will move to a fast lane, gaining big benefits, while others will only improve slowly. Because of the pervasive and growing importance of computing in our society, we expect this to have important impacts on computation-driven innovations in the future. In particular, we expect the gains from computing improvement to be become much more unequal, to the detriment of many.

7 Bibliography

12 AI Hardware Startups Builing New AI Chips. (2017). Retrieved January 18, 2018, from https://www.nanalyze.com/2017/05/12-ai-hardware-startups-new-ai-chips/

Arabi, K. (2014) More Problems Ahead / Interviewer: Sperling, E. Retrieved September 10, 2017 from http://semiengineering.com/more-problems-ahead/

AWS. (2017). Elastic GPUs. Retrieved March 23, 2018, from https://aws.amazon.com/de/ec2/elastic-gpus/

Bloom, N., Jones, C., Van Reenen, J., & Webb, M. (2017). Are Ideas Getting Harder to Find?, 1-55.

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

Bohr, M. (2007). A 30 Year Retrospective on Dennard's MOSFET Scaling Paper. *IEEE Solid-State Circuits Newsletter*, 12(1), 11–13.

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

Brodtkorb, A. R., Hagen, T. R., & Sætra, M. L. (2013). Graphics processing unit (GPU) programming strategies and trends in GPU computing. *Journal of Parallel and Distributed Computing*, 73(1), 4–13.

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

Byrne, D. M., Oliner, S. D., & Sichel, D. E. (2017). How Fast are Semiconductor Prices Falling? Review of Income and Wealth, (March), 1-58.

Cavin, R. K., Lugli, P., & Zhirnov, V. V. (2012). Science and Engineering Beyond Moore's Law. Proceedings of the IEEE, 100 (Special Centennial Issue), 1720–1749.

Chao, E. L., & Utgoff, K. P. (2006). 100 Years of US Consumer Spending. Data for the Nation, New York City, and Boston (Report No. 991). Retrieved from BLS https://www.bls.gov/opub/uscs/report991.pdf

DARPA. (2017). DARPA Rolls Out Electronics Resurgence Initiative. Retrieved December 20, 2017, from https://www.darpa.mil/news-events/2017-09-13

Davis, M. (2012). The Universal Computer (Turing Cen). Boca Raton, FL: CRC Press Taylor & Francis Group.

Dennard, R. H., Gaensslen, F. H., Yl, H., Rideout, V. L., Bassous, E., & Leblanc, A. R. (1974). Design of lon-Implanted MOSFET ' s with Very Small Physical Dimensions. IEEE Journal of Solid-State Circuits, SC-9(5), 256–268.

Dent, S. (2018). Major AMD chip supplier will no longer make next-gen chips. Retrieved October 14, 2018 from https://www.engadget.com/2018/08/28/global-foundries-stops-7-nanometer-chip-production/

Dvorak, J. (2013). Microsoft Office's Spaghetti Code Mess. Retrieved August 7, 2018, from https://uk.pcmag.com/opinion/15915/microsoft-officesspaghetti-code-mess

Eastwood, G. (2017). How chip design is evolving in response to IoT development. Retrieved March 24, 2018, from https://www.networkworld.com/article/3227786/internet-of-things/how-chip-design-is-evolving-in-response-to-iot-development.html

Esmaeilzadeh, H., Blem, E., St. Amant, R., Sankaralingam, K., & Burger, D. (2012). Dark Silicon and the End of Multicore Scaling. IEEE Micro, 32(3), 122–134.

Falsafi, B., Dally, B., Singh, D., Chiou, D., Yi, J. J., & Sendag, R. (2017). FPGAs versus GPUs in Data centers. IEEE Micro, 37(1), 60–72.

Feldmann, M. (2018a). TSMC Expects HPC to be Fastest Growing Segment. Retrieved March 22, 2018, from https://www.top500.org/news/tsmc-expects-fastest-growth-in-hpc-silicon/

Feldmann, M. (2018b). New GPU-Accelerated Supercomputers Change the Balance of Power on the Top500. Retrieved August 6, 2018, from https://www.top500.org/news/new-gpu-accelerated-supercomputers-change-the-balance-of-power-on-the-top500/

Flamm, K. (2017). Has Moore's Law been repealed? – An economist's perspective. Computing in Society and Engineering, 19(2), 29-40.

Flamm, K. (2018). Measuring Moore's Law: Evidence from price, cost, and quality indexes. NBER Working Paper Series, 1-46.

Gartner. (2017). Gartner Says Worldwide Sales of Smartphones Grew 7 Percent in the Fourth Quarter of 2016. Retrieved March 22, 2018, from https://www.gartner.com/newsroom/id/3609817

Gartner. (2018). Gartner Says Worldwide Semiconductor Revenue Grew 22.2 Percent in 2017; Samsung Takes Over No. 1 Position. Retrieved March 24, 2018, from https://www.gartner.com/newsroom/id/3842666

Geiger, D. (2017) Record Spending for Fab Equipment Expected in 2017 and 2018. Retrieved November 24, 2017, from http://www.semi.org/en/record-spending-fab-equipment-expected-2017-and-2018

Giles, Mike (2017). Lecture 2: different memory and variable types. Retrieved March 19, 2018, from http://people.math.umass.edu/~johnston/CUDA_WG_2012/memory_hierarchy_lec2-2x2.pdf

Giles, Martin (2017). The Race to Power AI's Silicon Brains. Retrieved January 12, 2018, from https://www.technologyreview.com/s/609471/the-raceto-power-ais-silicon-brains/

Glassdoor (2018). Computer Programmer Salaries. Retrieved November 5, 2018, from https://www.glassdoor.com/Salaries/computer-programmer-salary-SRCH_KO0,19.htm

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

Google Cloud. (2018). TPU Release Notes. Retrieved March 23, 2018, from https://cloud.google.com/tpu/docs/release-notes

Harrington, M. (2017). Celebrating 100 Years: The History of the Home Motor. Retrieved March 6, 2018, from https://everydaygoodthinking.com/2017/09/18/celebrating-100-years-home-motor/

Hemsoth, N. (2017). An Early Look at Baidu's Custom AI and Analytics Processor. Retrieved March 24, 2018, from https://www.nextplatform.com/2017/08/22/first-look-baidus-custom-ai-analytics-processor/

Hennessy, J., & Patterson, D. (2017). Domain-Specific Architectures. In Computer Architecutre: 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.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261–266.

Hof, R. D. (2013). Deep Learning. Retrieved January 26, 2018, from https://www.technologyreview.com/s/513696/deep-learning/

IEA. (2009). Gadgets and Gigawatts. International Energy Agency. Retrieved August 6, 2018, from http://www.iea.org/publications/freepublications/publication/gigawatts2009.pdf

Intel Corporation. (2017). Annual Report. Retrieved February 10, 2018 from https://www.intc.com/investor-relations/financials-and-filings/annual-reports-and-proxy/default.aspx

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

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.

Johnson, B., Tuan, S., Brady, W., Jim, W., & Jim, B. (2016). Predicts 2017: Semiconductor Technology in 2026. Gartner. Retrieved from https://www.gartner.com/doc/3528117/predicts--semiconductor-technology-

Jones, S. (2016). Presentation at Semicon West: Technology and Cost Trends at Advanced Nodes [PDF slides].

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.

Jouppi, N. P., Borchers, A., Boyle, R., Cantin, P., Chao, C., Clark, C., Coriell, J., Daley, M., Dau, M., Dean, J., Gelb, B., Young, C., Ghaemmaghami, T. V., Gottipati, R., Gulland, W., Hagmann, R., Ho, C. R., Hogberg, D., Hu, J., Hundt, R., Hurt, D., Ibarz, J., Patil, N., Jaffey, A., Jaworski, A., Kaplan, A., Khaitan, H., Killebrew, D., Koch, A., Kumar, N., Lacy, S., Laudon, J., Law, J., Patterson, D., Le, D., Leary, C., Liu, Z., Lucke, K., Lundin, A., MacKean, G., Maggiore, A., Mahony, M., Miller, K., Nagarajan, R., Agrawal, G., Narayanaswami, R., Ni, R., Nix, K., Norrie, T., Omernick, M., Penukonda, N., Phelps, A., Ross, J., Ross, M., Salek, A., Bajwa, R., Samadiani, E., Severn, C., Sizikov, G., Snelham, M., Souter, J., Steinberg, D., Swing, A., Tan, M., Thorson, G., Tian, B., Bates, S., Toma, H., Tuttle, E., Vasudevan, V., Walter, R., Wang, W., Wilcox, E., Yoon, D. H., Bhatia, S., & Boden, N. (2017). In-Datacenter Performance Analysis of a Tensor Processing Unit. Proceedings of the 44th Annual International Symposium on Computer Architecture - ISCA ’17, 1–12.

Kelly, J. (2016). Bitcoin “miners” face fight for survival as new supply halves. Retrieved January 26, 2018, from https://www.reuters.com/article/us-markets-bitcoin-mining/bitcoin-miners-face-fight-for-survival-as-new-supply-halves-idUSKCN0ZO2CW

Khazraee, M., Zhang, L., Vega, L., & Taylor, M. B. (2017). Moonwalk : NRE Optimization in ASIC Clouds or, accelerators will use old silicon. ACM, ASPLOS ’17, 1–16.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. 26th Annual Conference on Neural Information Processing Systems 2012, NIPS 2012, 2, 1097–1105.

Krzanich, B. (2016). Presentation at Sanford C Berstein Strategic Decisions Conference [online transcript]. Retrieved September 21, 2017 from https://seekingalpha.com/article/3979164-intel-corporations-intc-ceo-brian-krzanich-presents-sanford-c-bernstein-strategic-decisions

Kubota, Y. (2018). China Plans $47 Billion Fund to Boost Its Semiconductor Industry. Retrieved August 7, 2018, from https://www.wsj.com/articles/china-plans-47-billion-fund-to-boost-its-semiconductor-industry-1525434907

Lapedus, M. (2017a). 200mm Crisis? Retrieved March 16, 2018, from https://semiengineering.com/200mm-crisis/

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

Lee, V. W., Hammarlund, P., Singhal, R., Dubey, P., Kim, C., Chhugani, J., Deisher, M., Kim, D., Nguyen, A. D., Satish, N., Smelyanskiy, M., & Chennupaty, S. (2010). Debunking the 100X GPU vs. CPU myth. ACM SIGARCH Computer Architecture News, 38(3), 451-460.

Leiserson, C. E., Thompson, N., Emer, J., Kuszmaul, B. C., Lampson, B. W., Sanchez, D., & Schardl, T. B. (forthcoming). There's Plenty of Room at the Top - What will drive growth in computer performance after Moore's Law ends?, 1–13.

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/

Low, K. (2016). Presentation at Semicon West: Is it the end of Moore's Law or Just the Start of a New Journey? . Retrieved from http://www.monolithic3d.com/blog/archives/08-2016

Mahal, D. (2014). The Programmer Productivity Paradox. Retrieved August 7, 2018 from https://dzone.com/articles/programmer-productivity

Malone, M. (1995). The Microprocessor: A Biography. Santa Clara, California: Allan M. Wylde.

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

Martin, T. W., & Fitzgerald, D. (2018). Your Love of Your Old Smartphone Is a Problem for Apple and Samsung. Retrieved March 16, 2018, from https://www.wsj.com/articles/your-love-of-your-old-smartphone-is-a-problem-for-apple-and-samsung-1519822801

McCarren, D. & Govett, M. (2018). Future HPC Needs for Earth System Prediction Models. [PDF slides]

Metz, C. (2017). Google's TPU Chip Helped It Avoid Building Dozens of New Data Centers. Retrieved December 5, 2017, from https://www.wired.com/2017/04/building-ai-chip-saved-google-building-dozen-new-data-centers/

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

Mui, C. (2011). Five Dangerous Lessons to Learn From Steve Jobs. Retrieved March 23, 2018, from https://www.forbes.com/sites/chunkamui/2011/10/17/five-dangerous-lessons-to-learn-from-steve-jobs/#dddf9dc3a95c

Neill, C., Roushan, P., Kechedzhi, K., Boixo, S., Isakov, S. V., Smelyanskiy, V., Barends, R., Burkett, B., Chen, Y., Chen, Z., Chiaro, B., Dunsworth, A., Fowler, A., Foxen, B., Graff, R., Jeffrey, E., Kelly, J., Lucero, E., Megrant, A., Mutus, J., Neeley, M., Quintana, C., Sank, D., Vainsencher, A., Wenner, J., White, T. C., Neven, H., & Martinis, J. M. (2017). A blueprint for demonstrating quantum supremacy with superconducting qubits. Retrieved from https://arxiv.org/pdf/1710.09659.pdf

Noyce, R. N., & Hoff, M. E. (1981). A History of Microprocessor Development at Intel. IEEE Micro, 1(1), 8-21.

NVIDIA Corporation. (2016). Tesla P100 Datasheet. Retrieved from https://images.nvidia.com/content/tesla/pdf/nvidia-tesla-p100-datasheet.pdf

NVIDIA Corporation. (2017a). Tesla P100 Performance Guide - HPC and Deep Learning Applications. Retrieved from http://images.nvidia.com/content/pdf/v100-application-performance-guide.pdf

NVIDIA Corporation. (2017b). Tesla V100 Datasheet. Retrieved from https://images.nvidia.com/content/technologies/volta/pdf/tesla-volta-v100-datasheet-letter-fnl-web.pdf

Patterson, D., & Hennessy, J. (2014). In Praise of Computer Organization and Design : The Hardware / Software Interface (5th edition). Waltham, MA: Morgan Kaufmann

Patton, G. (2017). Forging Intelligent Systems in the Digital Era [Slides]. Retrieved from https://www-mtl.mit.edu/mtlseminar/garypatton.html#simple3

Period Paper. (2018). 1918 Ad Hamilton Beach Home Motor Sewing Sharpening. Retrieved March 26, 2018, from https://www.periodpaper.com/products/1918-ad-hamilton-beach-home-motor-sewing-sharpening-original-advertising-096931-thb1-048

Pham, S. (2018). Who needs the US? Alibaba will make its own computer chips. Retrieved October 14, 2018 from https://edition.cnn.com/2018/10/01/tech/alibaba-chip-company/index.html

Prickett Morgan, T. (2017). Intel Takes First Steps To Universal Quantum Computing. Retrieved March 3, 2018, from https://www.nextplatform.com/2017/10/11/intel-takes-first-steps-universal-quantum-computing/

Putnam, A., Gray, J., Haselman, M., Hauck, S., Heil, S., Hormati, A., Kim, J.-Y., Lanka, S., Larus, J., Peterson, E., Pope, S., Caulfield, A. M., Smith, A., Thong, J., Xiao, P. Y., Burger, D., Chung, E. S., Chiou, D., Constantinides, K., Demme, J., Esmaeilzadeh, H., Fowers, J., & Gopal, G. P. (2016). A reconfigurable fabric for accelerating large-scale datacenter services. Communications of the ACM, 59(11), 114–122.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408.

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.

Santhanam, N., Wiseman, B., Campbell, H., Gold, A., & Javetski, B. (2015). McKinsey on Semiconductors. McKinsey. Retrieved from: https://www.mckinsey.com/~/media/McKinsey/Industries/Semiconductors/Our%20Insights/McKinsey%20on%20Semiconductors%20Issue%205%20-%20Winter%202015/McKinsey%20on%20Semiconductors%20Winter%202015.ashx

Sato, K., Young, C., & Patterson, D. (2017). An in-depth look at Google’s first Tensor Processing Unit (TPU). Retrieved March 24, 2018, from https://cloud.google.com/blog/big-data/2017/05/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

Shalf, J. M., & Leland, R. (2015). Computing beyond Moore’s Law. Computer, 48(12), 14–23.

Shao, Y.S., Reagen, B., Wei, G-Y. & Brooks, D. (2014) Alladin: A Pre-RTL, Power-Performance Accelerator Simulator Enabling Large Design Space Exploration of Customized Architechtures. *International Symposium on Computer Architecture (ISCA)*.

Shehabi, A., Smith, S. J., Sartor, D. A., Brown, R. E., Herrlin, M., Koomey, J. G., Masanet, E. R., Horner, N., Azevedo, I. L., & Lintner, W. (2016). United States Data Center Energy Usage Report. Ernest Orlando Lawrence Berkeley National Laboratory. Retrieved from http://eta-publications.lbl.gov/sites/default/files/lbnl-1005775_v2.pdf

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

Skype. (2014). Skype Translator – How it Works | Skype Blogs. Retrieved January 30, 2018, from https://blogs.skype.com/news/2014/12/15/skype-translator-how-it-works/

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

Sperling, Ed. (2014). More Problems Ahead. Retrieved March 10, 2018, from http://semiengineering.com/more-problems-ahead/

Steinkraus, D., Buck, I., & Simard, P. Y. (2005). Using GPUs for machine learning algorithms. Eighth International Conference on Document Analysis and Recognition (ICDAR'05), Vol. 2, 1115–1120.

Stokes, J. (2009). End of the line for IBM's Cell. Retrieved December 13, 2017, from https://arstechnica.com/gadgets/2009/11/end-of-the-line-for-ibms-cell/

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.

Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. (2017). Efficient Processing of Deep Neural Networks: A Tutorial and Survey, 1–31.

Tanner, P. (2016). Why Did Nvidia's Data Center Revenue Double in Fiscal 2Q17? Retrieved March 15, 2018, from https://marketrealist.com/2016/08/nvidias-data-center-revenue-double-fiscal-2q17

Teodoridis, F. (2014). Generalists, Specialists, and the Direction of Inventive Activity. DRUID Society Conference 2014, CBS, Copenhagen.

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

Thompson, N. (2017). The Economic Impact of Mooore's Law: Evidence from When it Faltered. SSRN Electronic Journal, 1-58.

Top 500. (2018). Development over Time. Retrieved March 19, 2018, from https://www.top500.org/statistics/overtime/

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/

Vincent, J. (2017). A brief guide to mobile AI chips. Retrieved March 15, 2018, from https://www.theverge.com/2017/10/19/16502538/mobile-ai-chips-apple-google-huawei-qualcomm

Wehner, M., Oliker, L., & Shalf, J. (2008). Towards Ultra-High Resolution Models of Climate and Weather. The International Journal of High Performance Computing Applications, 22(2), 149–165.

Western Electric. (1917). Western Electric Sweing Machine Advertisement. Retrieved March 22, 2018, from https://witness2fashion.files.wordpress.com/2015/08/1917-mar-p-47-western-electric-portable-sewing-machine-ad.jpg

Wilson, R. (2008). ClearSpeed cuts costs despite sales boost. Retrieved December 14, 2017, from https://www.electronicsweekly.com/news/business/finance/clearspeed-cuts-costs-despite-sales-boost-2008-09/

Wong, D., Kan, K., Chanda, A., & Zhang, J. (2016). Intel Corp.: Primer 2016 - Provider of Key Technology for the Future. Wells Fargo.

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/

[1] In reality, Moore's Law has only been one part of this – although a crucial one. See (Leiserson et al., forthcoming) for a more complete discussion.

[2] Over the longer-term this may also become important for specialized chips, as will become clear later in the paper.

[3] In this work, the term "computer" describes both, universal and specialized devices.

[4] Chips, or processors, will be used largely interchangeably for this work. In practice, one chip often houses multiple processors. A universal chip houses only universal processors, while a specialized chip might host a combination of universal and specialized processors. Processors are the technology within each computer which responds to and executes the basic instructions. Since chip/processor improvement is what makes computers better, for the remainder of this paper, those will be the unit of analysis.

[5] This financing could be direct (using retained earnings for investment) or indirect (ability to raise external financing because of expectations of future earnings performance). For the purpose of this paper, this distinction is not important.

[6] Estimated by combined microprocessor sales of desktop and laptop computers. For this period, laptops were the major driver of growth (CAGR 26%).

[7] Calculated as 2008-2017 R&D and additions to PPE spending

[8] This is not exactly accurate but suffices for the purpose of this paper. See (Moore, 1995) for a more-technical discussion.

[9] Advertisements retrieved from (Harrington, 2017) and (Period Paper, 2018)

[10] Interestingly, today, we see some cross-usage of the most powerful household motors, in the form of various attachments to food processors (mixing, blending, grinding).

[11] The space is valuable because the smaller the processor, the more can be made on a single semiconductor wafer and thus the lower the manufacturing cost. (Flamm, 2017)

[12] Technical note: This is because the cache will speculatively bring additional data from the hard drive ahead of time, in the hope that it will be useful later. The larger the cache, the less often the processor searches for the data in cache, doesn't find it (a cache 'miss'), and needs to go to a higher (i.e. slower) level of memory to find it. (Patterson & Hennessy, 2014)

[13] Technical note: This parallelism exists at multiple levels: vector units, pipelining, multicore, etc.

[14] In our analogy: Parallelism is when parts of the car can be assembled independently. For example, the seats and the chassis can be produced at the same time, in different locations, and combined later. Regularity occurs with repetitive tasks, for example as done by robots working on an assembly line. The robot fastens the screws in the same way, in the same amount of time, for every object passing by. Locality suggests that multiple process steps can be done one after another on the same piece. For example, in rapid succession robots could punch holes in steel, fasten screws in those holes, and apply a covering. There is significant locality if it is efficient to do operations at the same stop of the conveyer belt. Less precision might manifest in cutting the fabric for car seats, which can safely be 'off' by a few millimeters (whereas mechanical gear tolerances might be much smaller). This lower need for precision could mean, for example, that less sophisticated manufacturing equipment is needed to achieve equivalent performance.

[15] Data from Intel and NVIDIA data sheets, 'Access to L1 Cache' from (Mike Giles, 2017)

[16] Approximated by number of threads for CPU and number of CUDA cores for GPU.

[17] This is not the specific value for the Intel Xeon E5-2690v4 but approximated with data from similar CPUs

[18] Research Scientists from Intel (the largest producer of CPUs) contest the magnitude of these performance benefits, highlighting how these comparisons often do not sufficiently optimize the software to get the most out of the CPU hardware. (Lee et al., 2010) However, optimizing software to this extent is rare for applications, and even with it, they still find a significant (2.5x) throughput advantage of the GPU.

[19] Comparing performance equivalence of applications on server accelerated with 2x NVIDIA P100 (12GB) GPUs to CPU Dual Xeon E5-2690 v4 server

[20] The latest generation GPUs now also allow for matrix multiplication

[21] Programming refers to running any problem on a specialized processor. Due to different interfaces than the universal processors, most commercially available software will not run on a specialized processor, but needs to be developed explicitly.

[22] NVIDIA created the CUDA programming language and the Khronos group created the open-source language OpenCL.

[23] Data from NVIDIA's publications "GPU accelerated applications", 2012-2017

[24] This true for the 16/14nm node size. Lithography mask cost are by far the biggest cost component of the NRE (Khazraee et al., 2017). Further costs include labor and design tools, as well as IP Licensing cost.

[25] Correct classification in the contest means that one of the *five* labels assigned the highest probability by the algorithm is correct.

[26] Additionally, AlexNet implemented the algorithm more efficiently, which also played an important role

[27] The latest NVIDIA GPU (Volta) also has a 30x higher throughput on inference (NVIDIA Corporation, 2017b)

[28] While the first generation could only execute neural networks, the second generation also speeds-up training. Many of Google's speedup numbers for the TPU are contested by NVIDIA, who argue that the comparison was against a previous generation of a GPU, rather than the cutting-edge version.

[29] Approximated by microprocessor sales

[30] The regression estimates for $share\ universal = \beta_0 + \beta_1(iphone\ generation)$ were $\beta_0 = 0.167\ (^{**})$ and $\beta_1 = -0.005\ (-)$.

[31] The regression estimates for $share\ of\ new\ entries\ with\ specialized\ chips = \beta_0 + \beta_1(2011 - year)$ were $\beta_0 = 0.069\ (^{***})$ and $\beta_1 = 0.026\ (^{***})$. Interestingly, some of these supercomputers are equipped with specialized chips from Intel. Traditionally known for producing universal processors, Intel now offers an increasing number of specialized processors. For their datacenter segment, they project specialized hardware to grow at a 30% CAGR and become a major contributor to its datacenter sales (Flamm, 2017).

[32] 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. For each supercomputer, they report energy-efficiency (in MFlops/Watt) and whether they use specialized processors. For the change in power usage over time we considered only computers that were newly added to the list in that year. We estimate regression coefficients for $(energy\ efficiency) = \beta_0 + \beta_1 * \left(specialized\ processor_{0/1}\right) + \beta_2 * (year - 2011) + \beta_3 * (year - 2011) * \left(specialized\ processor_{0/1}\right)$, and find at $\beta_0 = 347.27$ (***), $\beta_1 = -129.23$ (-), $\beta_2 = 197.18$ (***), $\beta_3 = 736.50$(***).

[33] Theoretically, every application can be either run on specialized or universal hardware. In practice, some compute-intensive problems would have such long runtimes on universal hardware that they would either be implemented on specialized hardware, or not at all.

[34] 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 (e.g. energy efficiency)

[35] For the sake of simplicity, we have for now abstracted away from the distinction between fixed and variable costs. Equation (6) goes into more detail about the cost components, but it is important to understand that this price is implicitly the *average* price / cost and thus includes a fixed cost share.

[36] We model this as a yearly improvement rate (rather than per technology node) to abstract away from how often new technology nodes are introduced, and how big their performance gains are.

[37] 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.

[38] In practice they are probably similar, but not identical. The variable cost depends on the space each chip takes up on the silicon wafer and the batch size. Specialized chips are generally produced in smaller volumes, which lowers their yield and throughput, but also tend to be smaller.

[39] In our model, we assume that prices are constant. To accommodate this, we use the performance-per-dollar increase rate, rather than a pure performance measure.

[40] Flamm (2017) estimates that Intel sold ~400M processors in 2015 and that Intel's Cost of Goods Sold was ~$20 billion, yielding a per unit variable cost of $50 per processor

[41] The speedup for some Deep Learning applications was significantly higher, but since the greater variety of accelerated applications certainly helped to amortize the cost, an average value of 10 seems reasonable.

[42] Assuming 5 lines of code per hour (which is already the upper limit, especially for complex programs (Mahal, 2014)), an average annual salary of $75,000 (Glassdoor, 2018) (plus 30% in taxes and benefits), while working 1780 hours per year (US employee average per OECD).

[43] In fact, this would probably be even more expensive, as specialized processors often rely on lower-level programming languages which is a skill are less widely amongst programmers than the current C++/C#.

[44] In this analysis we assume that all the redevelopment costs must be amortized over the life of the specialized processor.

[45] 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 of these decrease demand for cutting-edge chips, further undermining the economics of producing new, cutting-edge chip manufacturing plants.

[46] These transition dynamics also occurred in the past, when supercomputer users slowly made their way from specialized chips to massive numbers of universal processors.

[47] $23 billion of Foundry revenue can be attributed to leading-edge nodes (Smith, 2017). This accounts for TSMC and GlobalFoundries. Assuming the majority (90%) of Intel's ($54 billion) and Samsung's ($40 billion) total revenues (Gartner, 2018) derives from leading-edge nodes, yields an upper bound of $108 billion/$343 billion≈30%.

[48] Technical note: This is driven overwhelmingly by lithography costs – the cost of etching the design onto the chip.

[49] 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 further accentuate our overall point.

[50] The 5% increase is across all process nodes, so it is possible that there would be a different growth rate for the most advanced nodes. But, as we show later, such rates of growth would be, if anything, *lower* than 5%.

[51] The past decade saw the emergence of pure-play foundries. Foundries own fabs and produce other company's processor designs. Until 2009, GlobalFoundries was the manufacturing department of Advanced Micro Devices (AMD). By vertically disintegrating, the cost of the fab can be amortized over larger volumes.

[52] Calculations based on values / rates derived earlier this section: assuming 30% of market sales going to leading edge; 13% cost increase, 2026: 0% market growth / 2032: 5% market growth. New facilities are assumed to be needed every two years, and fixed costs spread out over that period. We (conservatively) assume that all market demand can be met with a single facility. If more than that is needed, the date moves earlier.

[53] An alternative explanation for the increasing ratio could be that Intel is decreasing variable costs. We see no evidence of this. Intel's gross margin has remained fairly stable since 2010. In contrast, their EBIT margin which includes fixed costs, has (as expected) been decreasing.

[54] This number is NOT precise and is not intended to be. To calculate this precisely would require detailed numbers on internal cost allocation at Intel, how they accrue, and for which chips. Intel does not share such data. Rather, this is intended to convey an order-of-magnitude comparison about how these costs have changed over time.

[55] To be more technically precise, Dennard Scaling meant that power density of transistors decreased as they were miniaturized. This allowed them to be switched faster. The operating speed of a processor is largely constrained by the ability to dissipate heat as power density rises, but due to Dennard Scaling, the cooling problem per unit area remained largely unchanged (Dennard et al., 1974), so many more transistors could run in parallel. Since around 2005, significant leakage currents and voltage limits led to an increase in energy density, limiting further speedups from faster transistor switching and creating areas of dark silicon to ensure operation without overheating (Bohr, 2007) (Esmaeilzadeh et al., 2012)

[56] Importantly, these losses were felt unevenly. Some applications experienced the drop fully, whereas others did not (Thompson (2017) discusses these differences and their reliance on the parallelism of software in more detail).

[57] For a discussion about these numbers and the debates in calculating them, see (Byrne, Oliner, & Sichel, 2017)

[58] There is a debate about this. Taiwan Semiconductor Manufacturing (TSCM) claims that the cost per transistor is still decreasing at historical rates (Jones, 2016). Intel and GlobalFoundries agree, but admit to shrinking transistors by more than 50% each generation to make it happen (Bohr, 2017) (Patton, 2017).

[59] In practice this is not strictly true, both Microsoft and Intel do significant work to ensure backwards compatibility because users find remarkably unexpected (and off-standard) ways of using computers. Nevertheless, this work is only the tiniest fraction of what would be involved if a non-backwards compatible processor were introduced.