APRIL 2020

AI Chips: What They Are and Why They Matter

An AI Chips Reference

AUTHORS
Saif M. Khan
Alexander Mann

Table of Contents

Introduction and Summary

Artificial intelligence will play an important role in national and international security in the years to come. As a result, the U.S. government is considering how to control the diffusion of AI-related information and technologies. Because general-purpose AI software, datasets, and algorithms are not effective targets for controls, the attention naturally falls on the computer hardware necessary to implement modern AI systems. The success of modern AI techniques relies on computation on a scale unimaginable even a few years ago. Training a leading AI algorithm can require a month of computing time and cost $100 million. This enormous computational power is delivered by computer chips that not only pack the maximum number of transistors—basic computational devices that can be switched between on (1) and off (0) states—but also are tailor-made to efficiently perform specific calculations required by AI systems. Such leading-edge, specialized "AI chips" are essential for cost-effectively implementing AI at scale; trying to deliver the same AI application using older AI chips or general-purpose chips can cost tens to thousands of times more. The fact that the complex supply chains needed to produce leading-edge AI chips are concentrated in the United States and a small number of allied democracies provides an opportunity for export control policies.

This report presents the above story in detail. It explains how AI chips work, why they have proliferated, and why they matter. It also shows why leading-edge chips are more cost-effective than older generations, and why chips specialized for AI are more cost-effective than general-purpose chips. As part of this story, the report surveys semiconductor industry and AI chip design

trends shaping the evolution of chips in general and AI chips in particular. It also presents a consolidated discussion of technical and economic trends that result in the critical cost-effectiveness tradeoffs for AI applications.

In this paper, AI refers to cutting-edge computationally-intensive AI systems, such as deep neural networks. DNNs are responsible for most recent AI breakthroughs, like DeepMind's AlphaGo, which beat the world champion Go player. As suggested above, we use "AI chips" to refer to certain types of computer chips that attain high efficiency and speed for AI-specific calculations at the expense of low efficiency and speed for other calculations.*

This paper focuses on AI chips and why they are essential for the development and deployment of AI at scale. It does not focus on details of the supply chain for such AI chips or the best targets within the supply chain for export controls (CSET has published preliminary results on this topic[1]). Forthcoming CSET reports will analyze the semiconductor supply chain, national competitiveness, the prospects of China's semiconductor industry for supply chain localization, and policies the United States and its allies can pursue to maintain their advantages in the production of AI chips, recommending how this advantage can be utilized to ensure beneficial development and adoption of AI technologies.

This report is organized as follows:

Industry Trends Favor AI Chips over General-Purpose Chips

From the 1960s until the 2010s, engineering innovations that shrink transistors doubled the number of transistors on a single computer chip roughly every two years, a phenomenon known as Moore's Law. Computer chips became millions of times faster and more efficient during this period. (Section II.)

* Our definition of "AI chips" includes graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and certain types of application-specific integrated circuits (ASICs) specialized for AI calculations. Our definition also includes a GPU, FPGA, or AI-specific ASIC implemented as a core on system-on-a-chip (SoC). AI algorithms can run on other types of chips, including general-purpose chips like central processing units (CPUs), but we focus on GPUs, FPGAs, and AI-specific ASICs because of their necessity for training and running cutting-edge AI algorithms efficiently and quickly, as described later in the paper.

The transistors used in today's state-of-the-art chips are only a few atoms wide. But creating even smaller transistors makes engineering problems increasingly difficult or even impossible to solve, causing the semiconductor industry's capital expenditures and talent costs to grow at an unsustainable rate. As a result, Moore's Law is slowing—that is, the time it takes to double transistor density is growing longer. The costs of continuing Moore's Law are justified only because it enables continuing chip improvements, such as transistor efficiency, transistor speed, and the ability to include more specialized circuits in the same chip. (Section III and IV.)

The economies of scale historically favoring general-purpose chips like central processing units have been upset by rising demand for specialized applications like AI and the slowing of Moore's Law-driven CPU improvements. Accordingly, specialized AI chips are taking market share from CPUs. (Section V.)

AI Chip Basics

AI chips include graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs) that are specialized for AI. General-purpose chips like central processing units (CPUs) can also be used for some simpler AI tasks, but CPUs are becoming less and less useful as AI advances. (Section V(A).)

Like general-purpose CPUs, AI chips gain speed and efficiency (that is, they are able to complete more computations per unit of energy consumed) by incorporating huge numbers of smaller and smaller transistors, which run faster and consume less energy than larger transistors. But unlike CPUs, AI chips also have other, AI-optimized design features. These features dramatically accelerate the identical, predictable, independent calculations required by AI algorithms. They include executing a large number of calculations in parallel rather than sequentially, as in CPUs; calculating numbers with low precision in a way that successfully implements AI algorithms but reduces the number of transistors needed for the same calculation; speeding up memory access by, for example, storing an entire AI algorithm in a single AI chip; and using programming languages built specifically to efficiently translate AI computer code for execution on an AI chip. (Section V and Appendix B.)

Different types of AI chips are useful for different tasks. GPUs are most often used for initially developing and refining AI algorithms; this process is known as "training." FPGAs are mostly used to apply trained AI algorithms to real-world data inputs; this is often called "inference." ASICs can be designed for either training or inference. (Section V(A).)

Why Cutting-Edge AI Chips are Necessary for AI

Because of their unique features, AI chips are tens or even thousands of times faster and more efficient than CPUs for training and inference of AI algorithms. State-of-the-art AI chips are also dramatically more cost-effective than state-of-the-art CPUs as a result of their greater efficiency for AI algorithms. An AI chip a thousand times as efficient as a CPU provides an improvement equivalent to 26 years of Moore's Law-driven CPU improvements. (Sections V(B) and VI(A) and Appendix C.)

Cutting-edge AI systems require not only AI-specific chips, but *state-of-the-art* AI chips. Older AI chips—with their larger, slower, and more power-hungry transistors—incur huge energy consumption costs that quickly balloon to unaffordable levels. Because of this, using older AI chips today means overall costs and slowdowns at least an order of magnitude greater than for state-of-the-art AI chips. (Section IV(B) and VI(A) and Appendix D.)

These cost and speed dynamics make it virtually impossible to develop and deploy cutting-edge AI algorithms without state-of-the-art AI chips. Even with state-of-the-art AI chips, training an AI algorithm can cost tens of millions of U.S. dollars and take weeks to complete. In fact, at top AI labs, a large portion of total spending is on AI-related computing. With general-purpose chips like CPUs or even older AI chips, this training would take substantially longer to complete and cost orders of magnitude more, making staying at the research and deployment frontier virtually impossible. Similarly, performing inference using less advanced or less specialized chips could involve similar cost overruns and take orders of magnitude longer. (Section VI(B).)

Implications for National AI Competitiveness

State-of-the-art AI chips are necessary for the cost-effective, fast development and deployment of advanced security-relevant AI systems. The United States and its allies have a competitive advantage in several semiconductor industry sectors necessary for the production of these chips. U.S. firms dominate AI

chip design, including electronic design automation (EDA) software used to design chips. Chinese AI chip design firms are far behind and are dependent on U.S. EDA software to design their AI chips. U.S., Taiwanese, and South Korean firms control the large majority of chip fabrication factories ("fabs") operating at a sufficiently advanced level to fabricate state-of-the-art AI chips, though a Chinese firm recently gained a small amount of comparable capacity. Chinese AI chip design firms nevertheless outsource manufacturing to non-Chinese fabs, which have greater capacity and exhibit greater manufacturing quality. U.S., Dutch, and Japanese firms together control the market for semiconductor manufacturing equipment (SME) used by fabs. However, these advantages could disappear, especially with China's concerted efforts to build an advanced chip industry. Given the security importance of state-of-the-art AI chips, the United States and its allies must protect their competitive advantage in the production of these chips. Future CSET reports will analyze policies for the United States and its allies to maintain their competitive advantage and explore points of control for these countries to ensure that the development and adoption of AI technologies increases global stability and is broadly beneficial for all. (Section VII.)

The Laws of Chip Innovation

All computer chips—including general-purpose CPUs and specialized ones like AI chips—benefit from smaller transistors, which run faster and consume less energy than larger transistors. Compared to CPUs, AI chips also gain efficiency and speed for AI applications through AI-optimized designs. However, at least while transistor shrinkage came at a fast rate and produced large speed and efficiency gains through the late 2000s, the value of specialized designs remained low and CPUs were the dominant chip. However, Moore's Law is close to driving transistors to fundamental size limits at atomic scales. For a basic introduction to chips, see Appendix A.

Transistor Shrinkage: Moore's Law

Moore's Law states that the number of transistors in a chip doubles about every two years. Technical innovations that shrink transistors allow increased transistor density. Moore's Law was first observed in the 1960s, and it held until the 2010s, when improvements in transistor density began slowing. Today, leading chips contain billions of transistors, but they have 15 times fewer transistors than they would have if Moore's Law had continued.[2]

Transistor density increases occur in generations, or "nodes." Each node corresponds to the transistor size (expressed in terms of length) that allows a doubling of transistor density relative to the previous node. Fabs began "risk production," i.e. experimental production, of the latest node of 5 nanometers ("nm") in 2019, with mass production expected in 2020.[3] The previous leading nodes were 7 nm and 10 nm.[4]

A companion principle to Moore's Law says that because smaller transistors generally use less power than larger ones, as transistor density increases, power consumption per unit chip area remains constant.[5] However, transistor power reduction rates slowed around 2007.[6]

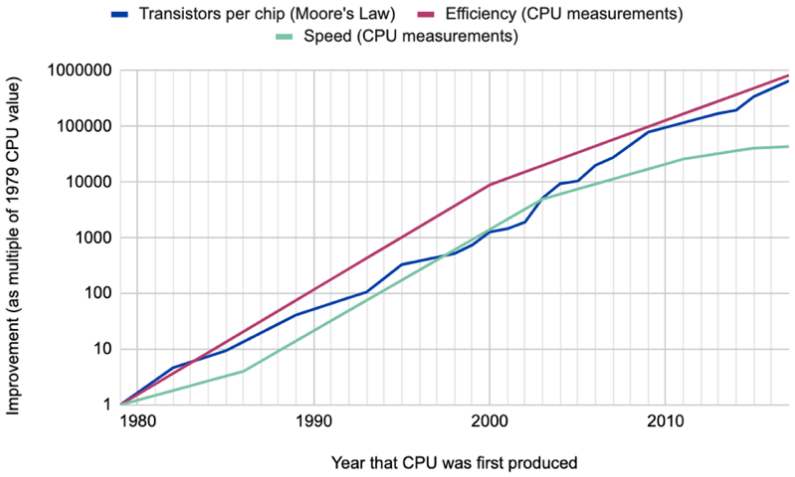
Efficiency and Speed Improvements

CPU speed has improved prodigiously since the 1960s due in large part to Moore's Law. Greater transistor density improved speed primarily via "frequency scaling," i.e. transistors switching between ones and zeros faster to allow more calculations per second by a given execution unit. Because smaller transistors use less power than larger ones, transistor switching speeds could be increased without increasing total power consumption.[7] Figure 1 shows transistor density, speed, and efficiency improvements since 1979.

Between 1978 and 1986, frequency scaling drove 22 percent annual increases in speed. Then, between 1986 and 2003, speed increased by 52 percent annually, due to frequency scaling and design improvements enabling simultaneous calculations to be performed through parallel computing. As frequency scaling slowed, parallelism enabled by multi-core designs powered 23 percent annual speedups between 2003 and 2011. Exploitation of the final remnants of available CPU parallelism brought 12 percent annual gains between 2011 and 2015, after which progress on CPU speed slowed to three percent per year.[8]

Efficiency has also improved dramatically. Because decreased transistor size reduces power use per transistor, overall CPU efficiency during peak chip usage doubled every 1.57 years until 2000.[9] Since then, due to the slowing of transistor power reduction, efficiency has doubled every 2.6 years, equivalent to a 30 percent per year efficiency improvement.[10]

Figure 1: CPU improvement rates normalized relative to 1979[11]



Increasing Transistor Density Unlocks Improved Designs for Efficiency and Speed

As transistors shrink and density increases, new chip designs become possible, further improving efficiency and speed. First, CPUs can include more and different types of execution units optimized for different functions.[12] Second, more on-chip memory can reduce the need for accessing slower off-chip memory. Memory chips such as DRAM chips likewise can pack more memory.[13] Third, CPUs can have more space for architectures that implement parallel rather than serial computation. Relatedly, if increased transistor density enables smaller CPUs, then a single device can house multiple CPUs (also called multiple "cores"), which each run different computations at once.

In the 1990s, design improvement lagged behind transistor density improvement because chip design firms struggled to exploit design possibilities unlocked by rapidly increasing transistor availability.[14] To get around this bottleneck, design firms focused comparatively more on trailing nodes (chips several generations behind the leading-edge), outsourced the brute-force work of creating a large number of chip designs to lower-paid engineers abroad, reused portions ("IP cores") of previous designs, and used EDA software to translate high-level abstract designs—easier for design engineers to work with—into concrete transistor-level designs.[15]

Transistor Design is Reaching Fundamental Size Limits

As transistors have shrunk to sizes only a few atoms thick, they are fast approaching fundamental lower limits on size. Various physics problems at small scales also make further shrinkage more technically challenging. The first significant change arrived in the 2000s when the transistor's insulative layer became so thin that electrical current started leaking across it.[16] Engineers used new, more insulative materials and stopped shrinking the insulative layer even as other components continued to shrink.[17]

More dramatic structural changes followed. From the 1960s to 2011, key transistors were manufactured as thin layers stacked on top of each other.[18] Yet even the more insulative materials could not prevent leakage. Instead, engineers replaced this planar arrangement with a more complex three-dimensional structure. This new structure has been dominant from the 22 nm node—released in 2011—to the current 5 nm node.[19] However, beyond 5 nm, even this structure leaks. A completely new structure has been developed for the future 3 nm node;[20] it includes components measuring only a few atoms in thickness, making further shrinkage beyond 3 nm challenging.[21]

The Slowing of Moore's Law and the Decline of General-Purpose Chips

Today, the trends that sustained CPU progress and primacy over specialized chips are ending. Technical difficulties are increasing the costs of Moore's Law improvements at a faster rate than the growth of the semiconductor market. Ultimately, these economic and technical factors suggest actual transistor densities will fall further behind what Moore's Law predicts and that we may reach the point of no further significant improvements in transistor densities.[22]

The Economies of Scale of General-Purpose Chips

The steady improvement in transistor-switching speeds and transistor power reduction favored CPUs over specialized chips. In the era of general-purpose chip dominance, specialized chips could not generate enough sales volume to recoup steep design costs.[23] Specialized chips earn their task-specific improvements over CPUs from design. But when rapid frequency scaling was still producing large speed and efficiency benefits, the computing premium from specialized chips was quickly erased by next-generation CPUs, whose

costs were spread across millions of chip sales.[24] Today, the slowing of Moore's Law means that CPUs no longer quickly improve. This results in longer useful lifetimes of specialized chips, making them more economical.

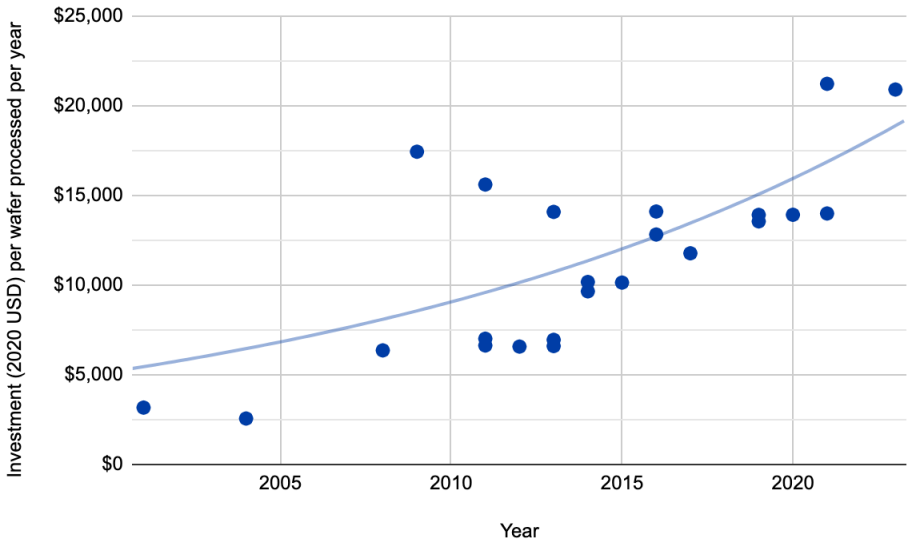
Costs are Increasing Faster than the Semiconductor Market

Increasing technical difficulties at small scales have driven up the costs of high-end semiconductor research and development across the supply chain. Different sectors of the semiconductor industry have localized in different regions based on their comparative advantages.[25]

The highest-value sectors, particularly SME, fabs, and chip design, have seen especially steep rates of cost growth and consolidation.[26] Annual growth rates in the cost of semiconductor fabrication facilities (eleven percent) and design costs per chip (24 percent) are faster than those of the semiconductor market (seven percent).[27] And the approximate number of semiconductor R&D workers has been increasing seven percent per year.

Since the early 2000s, the growth rate of semiconductor fabrication costs, including costs of fabs and SME, has trended at 11 percent per year. Fixed costs increasing faster than variable costs has created higher barriers of entry, squeezing fab profits and shrinking the number of chipmakers operating fabs at the leading nodes.[28] Figure 2 shows increasing construction costs of the largest fabs owned by Taiwan Semiconductor Manufacturing Company (TSMC). Currently, there are only two chipmakers at the 5 nm node: TSMC in Taiwan and Samsung in South Korea. Intel follows at 10 nm with plans to introduce the 7 and 5 nm nodes; GlobalFoundries and Semiconductor Manufacturing International Corporation (SMIC) lag at 14 nm (see Table 1).[29]

Figure 2: TSMC's leading-edge fab costs[30]



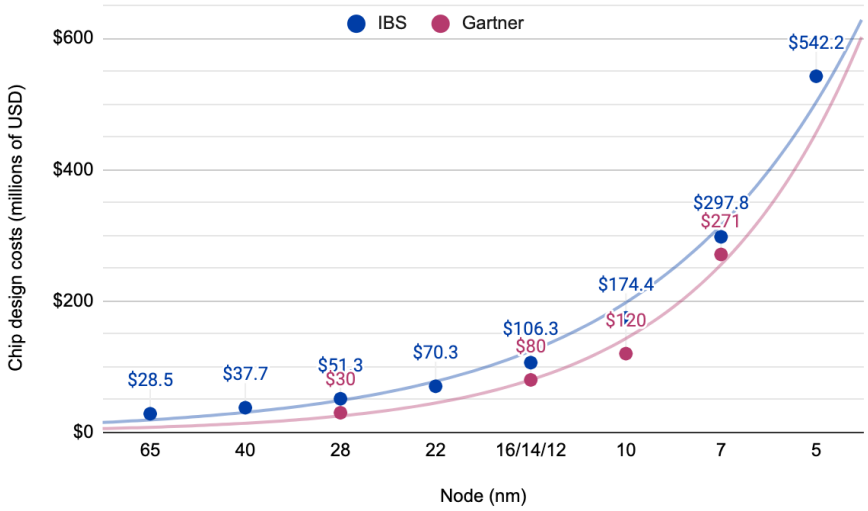
Costs of photolithography tools, the most expensive and complex segment of SME, have risen from $450,000 per unit in 1979 to $123 million in 2019.[31] And only one photolithography company, ASML in the Netherlands, now sells photolithography equipment capable of manufacturing the smallest 5 nm transistors. Nikon in Japan is the only other company making a significant volume of photolithography tools that operate at ≤90 nm (see Table 1). Eventually, increasing research and development costs for photolithography equipment and fabs at the leading node may prevent even a natural monopoly from recouping costs from the slowly growing global semiconductor market.

Table 1: Number of companies at each node

Node (nm)	180	130	90	65	45/ 40	32/ 28	22/ 20	16/ 14	10	7	5
Year mass production	1999	2001	2004	2006	2009	2011	2014	2015	2017	2018	2020
Chipmakers[32]	94	72	48	36	26	20	16	11	5	3	3
Photolithography companies[33]	4	3	2	2	2	2	2	2	2	2	1

Meanwhile, as shown in Figure 3, multiple estimates suggest the cost of chip design has been rising exponentially. When matched with TSMC's node introduction dates, design costs per node according to International Business Strategies (IBS) yields a 24 percent yearly cost increase.[34] Due to their general-purpose usage, CPUs enjoy economies of scale enabling U.S. firms Intel and AMD to maintain a decades-long duopoly in CPU design for servers and personal computers (PCs), such as desktops and laptops.[35]

Figure 3: Chip design costs at each node[36]



As semiconductor complexity increases, demands for high-end talent drive design and fabrication cost overruns. The effective number of researchers, measured by dividing semiconductor R&D spending by wages of high-skilled workers, saw an 18x increase from 1971 to 2015.[37] Put another way, a Moore's Law doubling required eighteen times as much human research effort in 2015 than in 1971, representing a seven percent increase per year.[38]

Overall design and manufacturing cost per transistor may be the best metric to measure whether transistor density improvements remain economical. This cost has historically decreased by around 20-30 percent annually.[39] Some analysts claim that decreases have stopped past the 28 nm node introduced in 2011, while others disagree.[40]

The Semiconductor Industry's Growth Rate is Unlikely to Increase

Unless new chip applications cause growth rates to increase, the semiconductor industry is unlikely to see growth rates sufficient to accommodate the industry's increasing costs. The semiconductor market is already growing at a faster rate than the world economy's three percent rate. Currently, the semiconductor industry produces 0.5 percent of global economic output. Due in part to the trade war between the United States and China, the semiconductor market shrunk in 2019.[41] However, it typically exhibits a year-to-year sawtooth growth trajectory, so a multi-year slowing would better indicate a slowing in long-run growth.[42]

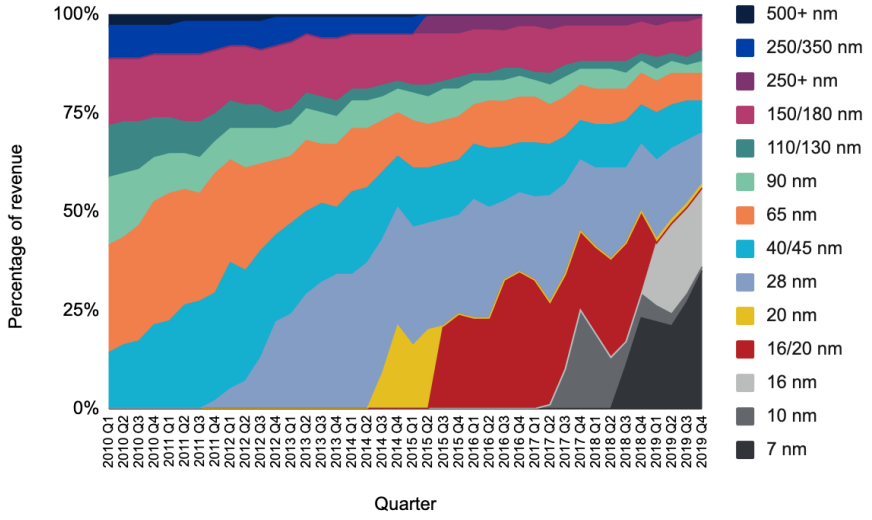
Chip Production at Each Node

Given the technical and economic challenges of chip production, new nodes are being introduced more slowly than in the past. Intel, the standard bearer of Moore's Law, has indeed slowed node introduction. It introduced 32 and 22 nm nodes two years after their predecessors, consistent with Moore's Law, but 14 nm followed three years after 22 nm, and 10 nm four years after 14 nm node chips.[43] Yet the leading foundry services vendor, TSMC, has not slowed node introduction.[44]

Trends in leading node chip sales volumes do not yet suggest a major slowing in the adoption of new nodes. From 2002 to 2016, TSMC's leading node stably represented approximately 20 percent of its revenue.[45] TSMC's 10 nm and 7 nm nodes introduced in 2016 and 2018, respectively, also reached 25 percent and 35 percent respectively, as shown in Figure 4.

TSMC's stable sales rates of new nodes—though slower than in the early 2000s—may mask the fact that the foundry services market as a whole is slowing adoption. TSMC has controlled roughly half of the world's foundry services market share for the last decade.[46] Rising production costs are reducing the number of companies at the leading node. For example, during this time, GlobalFoundries dropped out by failing to progress beyond 14 nm. If this trend is accompanied by less fab capacity at the current leading node than was the case for previously leading nodes, it would indicate that Moore's Law is slowing.[47]

Figure 4: TSMC's rate of introduction and adoption of new nodes has remained stable[48]



Fabs still make chips at the old nodes shown in Figure 4 for several reasons. Fabs incur great costs to build leading fabs or upgrade old ones to manufacture chips at newer nodes, so immediately transitioning world fab capacity to leading nodes is not possible. Instead, fabs continue selling old nodes at lower prices, especially to customers for whom purchase cost is the primary criterion. Many of these customers may be less concerned about efficiency because their applications are not computationally intensive. Similarly, their applications may not require fast speeds or otherwise may complete computations fast enough on old chips. Additionally, some specialized low-volume products like analog chips require trailing nodes to remain cost-effective.[49]

Chip Improvements as Moore's Law Slows

As Moore's Law slows, chips continue to improve in two ways: efficiency and speed improvements of smaller transistors, and efficiency and speed improvements from advanced chip designs exploiting larger numbers of transistors per chip enabled by smaller transistor size. These advanced designs include the ability to pack more specialized cores on a single chip.[50]

Transistor Improvements Continue, but are Slowing

Fortunately, some speed and efficiency improvements are still available, but with considerable technical challenges. Around 2004, when the 65 nm node was reached, transistor density improvements slowed in reducing transistor power usage and increasing transistor switching speed (frequency scaling).[51] Nevertheless, fabs report that transistor-level rather than design-level innovation continues to provide consistent, albeit slowing, improvements from node to node. TSMC and Samsung claim their 5 nm node chips improve upon the transistor speed of their 7 nm node chips respectively by 15 and 10 percent with power usage held constant[52] and reduce power usage by 30 and 20 percent with transistor speed held constant.[53] Figures 5 and 6 show a downward trend in TSMC's claimed node-to-node transistor speed improvements at constant efficiency between 90 nm and 5 nm, but a flat trend in TSMC's claimed transistor power reduction improvements.[54] Samsung trends downward between 14 nm and 5 nm on both metrics, but we lack data at nodes larger than 14 nm.[55] Intel sees slightly dropping transistor speed improvements,[56] but continuing node-to-node transistor power reduction improvements from 65 nm to 10 nm.[57] Intel has not yet introduced its 7 nm node. These improvements in speed and efficiency benefit both general-purpose chips like CPUs and specialized chips like AI chips.[58]

Figure 5: Node-to-node transistor speed improvements

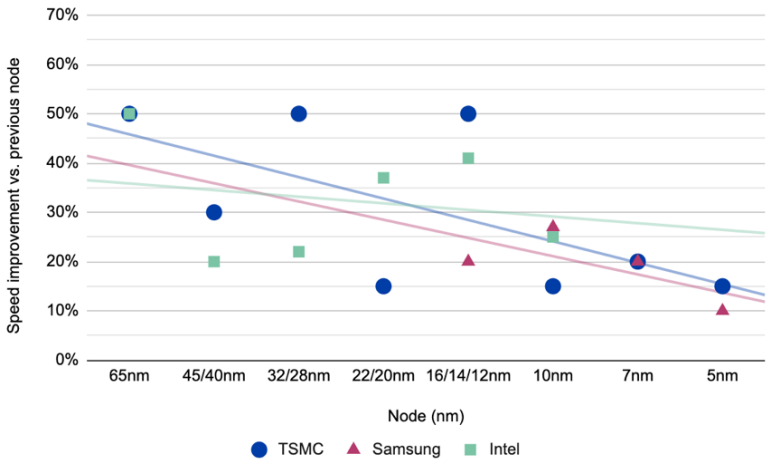
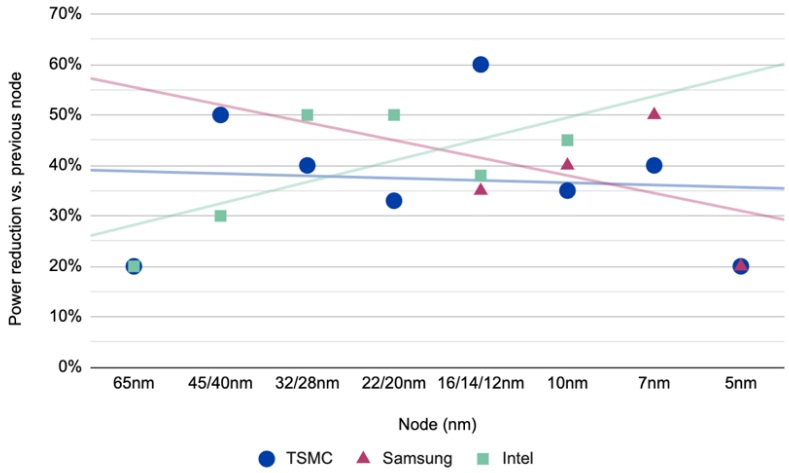
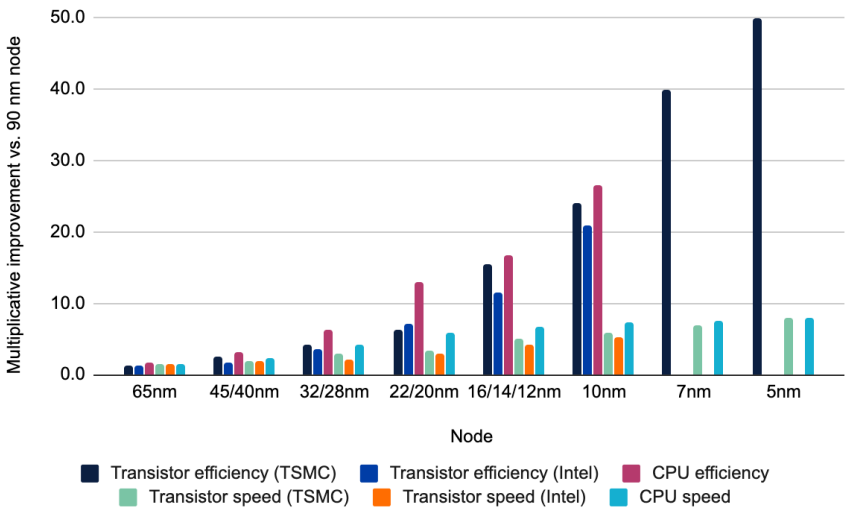


Figure 6: Node-to-node transistor power reduction improvements



Chip design improvements now provide decreasing CPU efficiency and speed improvements. Figure 7 consolidates the speed and efficiency measurements by node, both for CPUs and for transistors. For CPUs, we use data from Figure 1. For transistors, we use data for TSMC's and Intel's nodes from Figures 5 and 6.[59] The sources roughly agree on speed and efficiency improvements. TSMC's and Intel's reported improvements, derived from transistor-level innovation, generally match CPU improvements derived from both transistor-level and design-level innovation. The rough match implies that transistor-level innovation[60] has continued to play a major role in CPU efficiency and speed improvements over the last 15 years,[61] at least for the measured CPU benchmarks.[62] Efficient designs, however, do still play a role.[63]

Figure 7: Measured efficiency and speed improvements against 90 nm node



Improved Transistor Density Enables Specialization

Besides improving transistor function, increasing transistor density enables chips to include more varieties of specialized circuits that perform different types of calculations.[64] A chip can call upon a different specialized circuit depending on which calculation is requested. These circuits can include some optimized for AI algorithms and others specialized for different types of calculations. AI chips, which will be discussed in section V, are chips entirely specialized for AI.

Outside of the use of these specialized circuits, in recent years there has been little left to gain by adding more transistors to general-purpose chips. More transistors could theoretically enable a CPU to include more circuits to perform a larger number of calculations in parallel. However, speedups from parallelism are commonly limited by the percentage of time spent on serial computations, computations performed one after the other because the result of one computation is needed to start another. Parallel computations, conversely, are performed simultaneously. Even when only one percent of an algorithm's calculation time requires serial calculations, 45 percent of processor energy is wasted.[65] Unfortunately, most applications require at least some serial computation, and processor energy waste becomes too high as the serialization percentage increases. As other design improvements have slowed since the mid-2000s, multi-core designs with ever larger numbers of cores have proliferated. But multi-core designs also cannot efficiently

parallelize algorithms requiring a significant percentage of time spent on serial computations.

The AI Chip Zoo

The trend toward chips specialized for AI applications is driven by two factors. First, as discussed in Section IV, the critical improvements in semiconductor capabilities have shifted from manufacturing to design and software.[66] Second, an increasing demand for applications like AI requires highly parallelizable, predictable computations that benefit from specialized chips.[67] Deep neural networks (DNNs)—AI algorithms responsible for most recent AI breakthroughs—fit this bill. DNNs usually implement a type of machine learning called supervised learning, which involves two computing steps: "training" an AI algorithm based on training data (i.e. building the algorithm) and executing the trained AI algorithm (i.e. performing "inference") to classify new data consistent with knowledge acquired from data in the training stage. The training step in particular often requires performing the same computation millions of times. As discussed in Section IV(B), improved transistor density allows more types of specialized circuits on a single chip. AI chips take this to the extreme—the layout of most or all transistors on the chip is optimized for the highly parallelizable, specialized computations required by AI algorithms.

Although analysts disagree widely on the size of the global AI chip market—2018 estimates ranged between $5 and $20 billion—they agree that the market will grow faster than for chips not specialized for AI.[68] Until recently, a small number of firms designing general-purpose chips like CPUs dominated the logic chip design market. They enjoyed economies of scale that enabled them to reinvest into powerful new CPU designs. However, the slowing of Moore's Law is damaging CPU producers' economies of scale; now specialized chips have longer useful lifetime before Moore's Law-driven CPU efficiency and speed gains overcome the benefits of specialized chips. Therefore, the ability of CPU design firms to reinvest in new designs to maintain market dominance is declining. This trend lowers barriers to entry for chip design startups—especially those focused on specialized chips.[69]

AI chips are a common type of specialized chip, and share some features in common. AI chips execute a much larger number of calculations in parallel than CPUs. They also calculate numbers with low precision in a way that successfully implements AI algorithms but reduces the number of transistors

needed for the same calculation. They also speed up memory access by storing an entire AI algorithm in a single AI chip. Finally, AI chips use programming languages specialized to efficiently translate AI computer code to execute on an AI chip. For more detail on these techniques, see Appendix B.

While general-purpose chips include a small number of popular designs, particularly the CPU, AI chips are more diverse. AI chips vary widely in design, the applications they are suited to, efficiency and speed for different AI tasks, generality, and classification accuracy when performing inference. The following subsections categorize AI chips along these axes.

AI Chip Types

AI chips include three classes: graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs).[70]

GPUs were originally designed for image-processing applications that benefited from parallel computation. In 2012, GPUs started seeing increased use for training AI systems and by 2017, were dominant.[71] GPUs are also sometimes used for inference.[72] Yet in spite of allowing a greater degree of parallelism than CPUs, GPUs are still designed for general-purpose computing.[73]

Recently, specialized FPGAs and ASICs have become more prominent for inference, due to improved efficiency compared to GPUs.[74] ASICs are increasingly used for training, as well.[75] FPGAs include logic blocks (i.e. modules that each contain a set of transistors) whose interconnections can be reconfigured by a programmer after fabrication to suit specific algorithms, while ASICs include hardwired circuitry customized to specific algorithms. Leading ASICs typically provide greater efficiency than FPGAs, while FPGAs are more customizable than ASICs and facilitate design optimization as AI algorithms evolve.[76] ASICs, by contrast, grow increasingly obsolete as new AI algorithms are developed.

Different AI chips may be used for training versus inference, given the various demands on chips imposed by each task. First, different forms of data and model parallelism are suitable for training versus inference, as training requires additional computational steps on top of the steps it shares with

inference. Second, while training virtually always benefits from data parallelism, inference often does not. For example, inference may be performed on a single piece of data at a time. However, for some applications, inference may be performed on many pieces of data in parallel, especially when an application requires fast inference of a large number of different pieces of data. Third, depending on the application, the relative importance of efficiency and speed for training and inference can differ. For training, efficiency and speed are both important for AI researchers to cost-effectively and quickly iterate research projects. For inference, high inference speed can be essential, as many AI applications deployed in critical systems (e.g. autonomous vehicles) or with impatient users (e.g. mobile apps classifying images) require fast, real-time data classification. On the other hand, there may be a ceiling in useful inference speed. For example, inference need not be any faster than user reaction time to a mobile app.[77]

Inference chips require fewer research breakthroughs than training chips, as they require optimization for fewer computations than training chips. And ASICs require fewer research breakthroughs than GPUs and FPGAs; because ASICs are narrowly optimized for specific algorithms, design engineers consider far fewer variables. To design a circuit meant for only one calculation, an engineer can simply translate the calculation into a circuit optimized for that calculation. But to design a circuit meant for many types of calculations, the engineer must predict which circuit will perform well on a wide variety of tasks, many of which are unknown in advance.

An AI chip's commercialization has depended on its degrees of general-purpose capability. GPUs have long been widely commercialized, as have FPGAs to a lesser degree.[78] Meanwhile, ASICs are more difficult to commercialize given high design costs and specialization-driven low volume. However, a specialized chip is relatively more economical in an era of slow general-purpose chip improvement rates, as it has a longer useful lifetime before next-generation CPUs attain the same speedup or efficiency. In the current era of slow CPU improvements, if an AI chip exhibits a 10-100x speedup, then a sales volume of only 15,000-83,000 should be sufficient to make the AI chip economical.[79] The projected market size increase for AI chips could create the economies of scale necessary to make ever narrower-capability AI ASICs profitable.

AI chips come in different grades, from more to less powerful. At the high-end, server grade AI chips are commonly used in data centers for high-end

applications and are, after packaging, larger than other AI chips. At the medium-end are PC grade AI chips commonly used by consumers. At the low-end, mobile AI chips are typically used for inference and integrated into a system-on-a-chip that also includes a CPU. A mobile system-on-a-chip needs to be miniaturized to fit into mobile devices. At each of these grades, AI chip market share increases have come at the expense of non-AI chips.[80]

Supercomputers have limited but increasing relevance for AI. Most commonly, server grade chips are distributed in data centers and can be executed sequentially or in parallel in a setup called "grid computing." A supercomputer takes server grade chips, physically co-locates and links them together, and adds expensive cooling equipment to prevent overheating. This setup improves speed but dramatically reduces efficiency,[81] an acceptable tradeoff for many applications requiring fast analysis. Few current AI applications justify the additional cost of higher speed, but training or inference for large AI algorithms is sometimes so slow that supercomputers are employed as a last resort.[82] Accordingly, although CPUs have traditionally been the supercomputing chip of choice,[83] AI chips are now taking an increasing share.[84] In 2018, GPUs were responsible for the majority of added worldwide supercomputer computational capacity.[85]

AI Chip Benchmarks

There is no common scheme in the industry for benchmarking CPUs versus AI chips, as comparative chip speed and efficiency depends on the specific benchmark.[86] However, for any given node, AI chips typically provide a 10-1,000x improvement in efficiency and speed relative to CPUs, with GPUs and FPGAs on the lower end and ASICs higher.[87] An AI chip 1,000x as efficient as a CPU for a given node provides an improvement equivalent to 26 years of CPU improvements. Table 2 shows our estimates for efficiency and speed gains for GPUs, FPGAs, and ASICs relative to CPUs (normalized at 1x) for DNN training and inference at a given node. No data is available for FPGA training efficiency and speed, as FPGAs are rarely used for training. These estimates are informed by benchmarking studies, which are summarized in Appendix B. Table 2 also lists the generality and inference accuracy of these chips.

Table 2: Comparing state-of-the-art AI chips to state-of-the-art CPUs

	Training		Inference		Generality[88]	Inference accuracy[89]
	Efficiency	Speed	Efficiency	Speed		
CPU	1x baseline				Very High	~98-99.7%
GPU	~10-100x	~10-1,000x	~1-10x	~1-100x	High	~98-99.7%
FPGA	-	-	~10-100x	~10-100x	Medium	~95-99%
ASIC	~100-1,000x	~10-1,000x	~100-1,000x	~10-1,000x	Low	~90-98%

The Value of State-of-the-Art AI Chips

Leading node AI chips are increasingly necessary for cost-effective, fast training and inference of AI algorithms. This is because they exhibit efficiency and speed gains relative to state-of-the-art CPUs (Table 2 and Appendix C) and trailing node AI chips (Figure 7). And, as discussed in subsection A, efficiency translates into overall cost-effectiveness in chip costs—which are the sum of chip production costs (i.e. design, fabrication, assembly, test, and packaging costs). Finally, as discussed in subsection B, cost and speed bottleneck training and inference of many compute-intensive AI algorithms, necessitating the most advanced AI chips for AI developers and users to remain competitive in AI R&D and deployment.

The Efficiency of State-of-the-Art AI Chips Translates into Cost-Effectiveness

Efficiency translates into overall cost-effectiveness. For trailing nodes, chip operating costs—due to energy consumption costs—dominate chip production costs and quickly balloon to unmanageable levels. Even for leading nodes, operating costs are similar to production costs, implying the need to continue optimizing for efficiency.

Table 3 presents the results of a CSET model of chip production and operating costs for nodes between 90 and 5 nm with the same number of transistors as a generic server-grade 5 nm chip modeled according to the specifications similar to those of the Nvidia P100 GPU. This means that an above-5 nm chip would require a larger surface area. For above-5 nm

nodes, the model could equivalently be interpreted as accounting for production of multiple chips that together have the transistor count of one 5 nm chip. The model takes the perspective of a fabless design firm that, in 2020, designs the chip, buys foundry services from TSMC, then runs the chip in its own server. This mirrors the approach of companies like Google, which designs its TPU in-house, outsources fabrication to TSMC, then runs its TPUs in Google servers for its own AI applications or cloud-computing services to external customers.

The costs break down as follows. The foundry sale price paid by the fabless firm includes capital consumed (i.e. costs of building a fab and purchasing SME), materials, labor, foundry R&D, and profit margin. The fabless firm additionally incurs chip design cost. After fabrication, an outsourced semiconductor and test firm assembles, tests, and packages (ATP) the chip. The sum of foundry sale price, chip design cost, and ATP cost equals the total production cost per chip. The fabless firm also incurs energy cost when operating the chip. We estimate energy cost based on an electricity cost of $0.07625 per kilowatt-hour. See Appendix D for explanations of how each line-item in Table 3 is calculated. We make two findings.

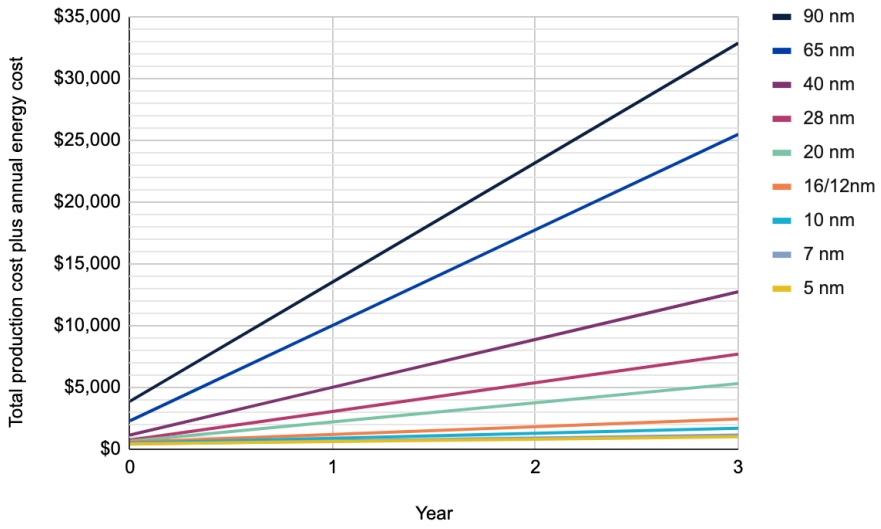
Table 3: Chip costs at different nodes with 5 nm-equivalent transistor count

Node (nm)	90	65	40	28	20	16/12	10	7	5
Year of mass production	2004	2006	2009	2011	2014	2015	2017	2018	2020
Foundry sale price to fabless firm per chip (i.e. costs + markup)	$2,433	$1,428	$713	$453	$399	$331	$274	$233	$238
Fabless firm's design cost per chip given chip volume of 5 million[90]	$630	$392	$200	$135	$119	$136	$121	$110	$108
Assembly, test, and packaging cost per chip	$815	$478	$239	$152	$134	$111	$92	$78	$80
Total production cost per chip	*$3,877*	*$2,298*	*$1,152*	*$740*	*$652*	*$577*	*$487*	*$421*	*$426*
Annual energy cost to operate chip	$9,667	$7,733	$3,867	$2,320	$1,554	$622	$404	$242	$194

First, in less than two years, the cost to operate a leading node AI chip (7 or 5 nm) exceeds the cost of producing said chip, while the cumulative electricity cost of operating a trailing node AI chip (90 or 65 nm) is three to four times the cost of producing that chip.[91] Figure 8 presents total chip costs for

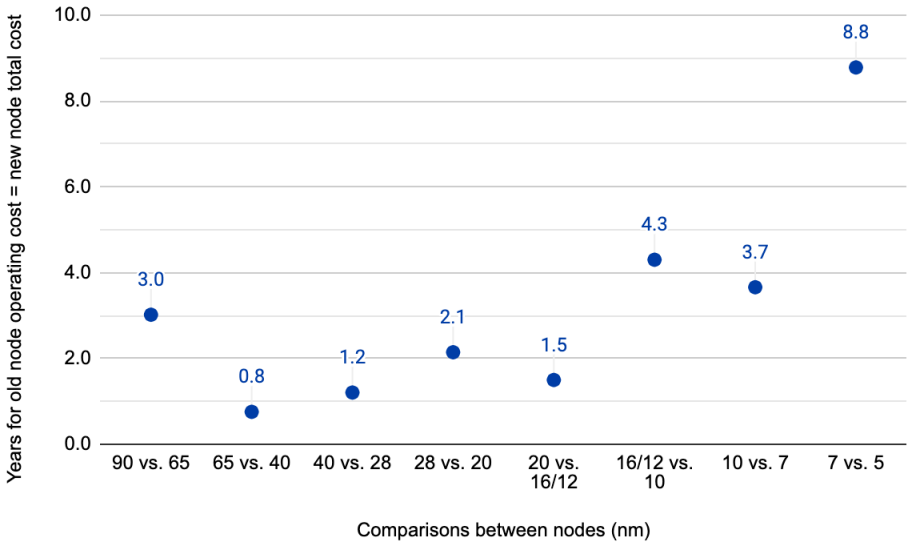
continuous use up to three years: total production cost per chip is added in year zero, with annual energy cost of using the chip added in each subsequent year. These results suggest that leading node AI chips are 33 times more cost-effective than trailing node AI chips when counting production and operating costs. Likewise, because leading node AI chips exhibit one to three orders of magnitude greater efficiency than leading node CPUs (Table 2 and Appendix C), we expect leading node AI chips are also one to three orders of magnitude more cost-effective than leading node CPUs when counting production and operating costs.

Figure 8: Cost of AI chips over time for different nodes



Second, it takes 8.8 years for the cost of producing and operating a 5 nm chip to equal the cost of operating a 7 nm chip.[92] Below 8.8 years, the 7 nm chip is cheaper, and above, the 5 nm chip cheaper. Therefore, users have an incentive to replace existing 7 nm node chips (assuming they do not break down) only when expecting to use 5 nm node chips for 8.8 years. Figure 9 shows node-to-node comparisons between 90 nm and 5 nm. We find that the timeframe where these costs become equal has increased, with a dramatic rise at the 7 versus 5 nm comparison.[93] Firms typically replace server-grade chips after about three years of operation, which is consistent with recent timeframes for introduction of new nodes—that is, firms relying on leading node chips purchase newly introduced node chips as soon as they are available. However, if firms begin purchasing 5 nm node chips, they may expect to use these chips for much longer.[94] This would constitute a market prediction that Moore's Law is slowing, and that the 3 nm node may not be introduced for a long time.[95]

Figure 9: Node transition economics



Compute-Intensive AI Algorithms are Bottlenecked by Chip Costs and Speed

AI firms' time and money spent on AI-related computing have become a bottleneck on AI progress. Given leading node AI chips are vastly more cost-effective and faster (Table 4 and Figure 7) than trailing node AI chips or leading node CPUs, these AI labs therefore need leading node AI chips to continue AI progress.

First, training costs of AI lab DeepMind's leading AI experiments, such as AlphaGo, AlphaGo Zero, AlphaZero, and AlphaStar, have been estimated at $5 to $100 million each.[96] One cost model suggests AlphaGo Zero's training cost was $35 million.[97,98] AI lab OpenAI reports that of their $28 million total 2017 costs, $8 million went to cloud computing.[99] Multiplying these computing costs by thirty for trailing node AI chips, or even more for leading node CPUs, would make such experiments economically prohibitive. And computing costs for some AI companies have increased so quickly that a cost ceiling may soon be reached, necessitating the most efficient AI chips.[100,101]

Second, leading AI experiments can take days or even a month for training,[102] while deployed critical AI systems routinely require fast or real-time inference. Increasing these times by using trailing node AI chips or leading node CPUs would make the required iteration speed for AI R&D and inference speed of deployed critical AI systems unacceptably slow. A

company with slower chips could attempt to pay the enormous energy costs to increase speed by using large numbers of slower chips in parallel. But this gambit would fail for two reasons. For one, as discussed in Section A of Appendix A, leading experiments require AI researchers to tune AI algorithms to support more data and model parallelism. AI researchers can do this to a limited degree, but may face difficulty if attempting to use a dramatically greater number of AI chips in parallel than currently used by leading AI experiments. For another, even if algorithmically possible, such parallelism requires complementary software and networking technology to enable it.[103] Scaling up hundreds or thousands of GPUs in parallel is extremely difficult.[104] Scaling up an even larger number of trailing node GPUs would likely be beyond current capabilities. The new Cerebras Wafer Scale Engine chip presents an intriguing potential workaround to networking technology. It is the first wafer-scale chip, having a much larger surface area than any other AI chip, meaning a large degree of parallelism can be accomplished on a single chip, reducing the need for advanced networking technology between multiple chips.[105]

A caveat to this analysis is that some recent AI breakthroughs have not required a significant amount of computing power.[106] Furthermore, there is ongoing research in developing AI algorithms requiring minimal training (e.g. "few shot" learning techniques).[107] For these AI algorithms, multiplying a small cost or speed by a large number may still yield a small cost or speed.

U.S. and Chinese AI Chips and Implications for National Competitiveness

Cost-effectiveness and speed of leading node AI chips matter from a policy perspective. U.S. companies dominate AI chip design, with Chinese companies far behind in AI chip designs, reliant on U.S. EDA software to design AI chips, and needing U.S. and allied SME and fabs to fabricate AI chips based on these designs. The value of state-of-the-art AI chips, combined with the concentration of their supply chains in the United States and allied countries, presents a point of leverage for the United States and its allies to ensure beneficial development and adoption of AI technologies.[108]

U.S. companies Nvidia and AMD have a duopoly over the world GPU design market, while China's top GPU company, Jingjia Microelectronics, fields dramatically slower GPUs.[109] Likewise, U.S. companies Xilinx and Intel dominate the global FPGA market, while China's leading FPGA companies

Efinix, Gowin Semiconductor, and Shenzhen Pango Microsystem have only developed trailing node FPGAs thus far.[110]

The AI ASIC market, especially for inference, is more distributed with lower barriers to entry, as ASICs and inference chips are easier to design (see Section VI(A)). Unlike GPUs and FPGAs, companies active in AI such as Google, Tesla, and Amazon have begun designing AI ASICs specialized for their own AI applications. Google's TPU is a leading commercial AI ASIC.[111] Intel is also developing powerful commercial AI ASICs,[112] and claims even greater improvements for research ASICs in the range of 10,000x and 1,000x for efficiency and speed respectively.[113] Competitive Chinese companies in the AI ASIC space include Baidu, Alibaba, Tencent, HiSilicon (owned by Huawei), Cambricon Technologies, Intellifusion, and Horizon Robotics. Chinese researchers have also produced high-end research ASICs.[114] However, they are largely limited to inference, although Huawei recently announced the development of an AI training ASIC.[115]

Table 4 lists world-leading server grade U.S. AI chip designs alongside leading Chinese counterparts.[116,117] The data tells two stories.

Table 4: Leading U.S. and Chinese AI chips

Type	Firm HQ	Design firm	AI chip	Node (nm)	Fab
GPU	United States	AMD[118]	Radeon Instinct	7	TSMC
		Nvidia[119]	Tesla V100	12	TSMC
	China	Jingjia Micro[120]	JM7200	28	Unknown
FPGA	United States	Intel[121]	Agilex	10	Intel
		Xilinx[122]	Virtex	16	TSMC
	China	Efinix[123]	Trion	40	SMIC
		Gowin Semiconductor[124]	LittleBee	55	TSMC
		Shenzhen Pango[125]	Titan	40	Unknown
ASIC	United States	Cerebras[126]	Wafer Scale Engine	16	TSMC
		Google[127]	TPU v3	16/12 (est.)	TSMC
		Intel[128]	Habana	16	TSMC

		Tesla[129]	FSD computer	10	Samsung
	China	Cambricon[130]	MLU100	7	TSMC
		Huawei[131]	Ascend 910	7	TSMC
		Horizon Robotics[132]	Journey 2	28	TSMC
		Intellifusion[133]	NNP200	22	Unknown

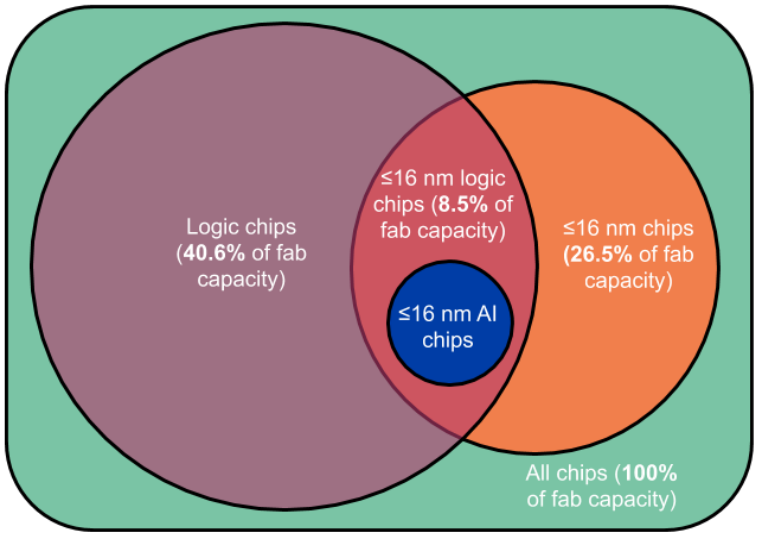
First, Table 4 shows that U.S. AI chip design firms fab exclusively at TSMC, Samsung, or Intel, with chips either at the leading commercial node (7 nm) or close behind. U.S. GPUs use more leading nodes than U.S. FPGAs and ASICs—possibly due to their generality and therefore higher sales volumes that recoup higher leading node design costs.[134]

Experts disagree on the need for leading nodes for AI chips. An executive of the EDA company Cadence Design Systems said, "everybody who wants to do AI needs the performance, power and form factor of 7nm and below."[135] Meanwhile, a semiconductor researcher at Hong Kong Applied Science and Technology Institute was more skeptical: "For AI chips ... manufacturing costs will be much lower if you use 28nm technology and not 10 or 14nm tech ... you need to spend a lot of effort from scratch [to design at leading nodes]—mathematical models, the physical layers, the computational language, all these need investment."[136]

The data in Table 4 settles this question: near-leading-edge nodes (i.e. ≤16 nm) are used for all of the leading U.S. AI chips we investigated. This data is consistent with the CSET chip economics model discussed in Section VI(A). Specifically, the model's results in Figure 8 show an especially high cost-effectiveness for chips at ≤16 nm, with ≥20 nm having much higher costs.

Few fabs are capable of manufacturing near-state-of-the-art AI chips, as shown in Figure 10. Only approximately 8.5% of global fab capacity could be used to fabricate near-state-of-the-art AI chips, and only a subset is currently used for it. The actual percentage used to fabricate near-state-of-the-art AI chips is difficult to calculate and varies year-to-year.

Figure 10: Near-state-of-the-art AI chips comprise a small percentage of all chips[137]



Red intersection: chip fab capacity capable of making near-state-of-the-art AI chips
Blue circle: chip fab capacity actually used to make near-state-of-the-art AI chips

Second, Table 4 shows that Chinese AI chip design firms use trailing nodes for GPUs and FPGAs, and a mix of leading nodes and trailing nodes for ASICs. Even though China has some local fabrication capacity at a number of these trailing nodes, China's AI chip design firms still mostly outsource fabrication of trailing node chips to the Taiwanese fab TSMC. This likely reflects TSMC's more reliable fabrication processes than those of Chinese domestic fabs like SMIC. SMIC has capacity as advanced as 14 nm, but only at a low volume.[138] Some of these chip design firms do use SMIC, but SMIC relies on SME imports from the United States, the Netherlands, and Japan. This is because China's SME industry includes only a small number of companies that are not at the state-of-the-art.[139] Chinese AI chip design firms also rely on U.S. EDA software to design their AI chips. Therefore, China remains dependent on the United States and its allies for AI chip production capabilities.

China has achieved the most design success in AI inference ASICs, as its large and well-educated population of engineers is well-suited to the labor-intensive work of designing a chip that performs extremely well on a specific

task.[140] However, given China's relatively young AI chip design industry, Chinese companies have yet to acquire the implicit know-how needed to navigate the large optimization space and higher complexity of mastering GPUs and FPGAs.

Chinese companies also heavily incorporate Western IP cores into their designs. For example, Huawei licenses British chip design firm ARM's instruction set architecture and IP cores.[141] Chinese FPGA makers also license Intel and Xilinx FPGA IP cores.[142] Licenses for IP cores become exponentially more expensive at leading nodes.[143]

China's lack of development in key sectors of AI chip supply chains—including AI chip designs, EDA software, SME, and fabs—means the United States and its allies maintain a competitive advantage in the production of leading-edge AI chips. As discussed in Section VII, leading-edge AI chips have critical strategic value for the development and deployment of advanced security-relevant AI systems. Therefore, it is vital to U.S., allied, and global security to maintain this advantage.

Future CSET reports will more deeply analyze AI chip industry competitiveness of the United States and China, China's semiconductor industry and its plans for chip independence and supply chain localization, and recommend policies the United States and its allies should pursue to maintain their advantages in the production of AI chips.

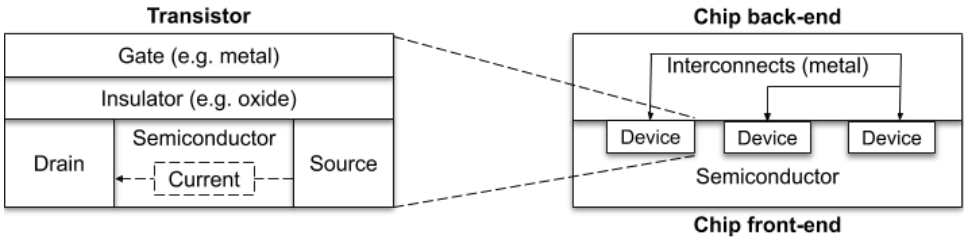
Appendix A: Basics of Semiconductors and Chips

A **semiconductor** is a material with an electrical conductivity between that of a conductor, which allows the flow of electrical current, and an insulator, which does not. A semiconductor can switch between being conductive and insulative in different circumstances. Silicon is the most commonly used semiconductor. Semiconductors are used in a wide array of devices, such as transistors, resistors, capacitors, and diodes, each of which perform distinct functions. These devices can be manufactured separately as "discrete" devices or multiple devices can be combined into an integrated circuit, also called a "chip."

Transistors are especially important devices for computing, as they can be switched between on and off states representing 1 and 0. The metal-oxide-semiconductor field-effect transistor (MOSFET) has been the dominant transistor type since the 1960s. The name is explanatory: a MOSFET includes an insulator (e.g. an oxide) between a gate (e.g. a conductive metal) and a semiconductor channel (e.g. silicon[144]) that connects a source and a drain (see Figure 11). When a voltage (i.e. an electric field) is applied to the gate, the channel is put in an "on" state so that current flows between the source and the drain. When voltage is not applied, the channel is put in an "off" state such that current does not flow between the source and the drain.

The structure of a chip includes a "front-end" and "back-end." The front-end has silicon layers embedded with electrical devices such as transistors. The back-end sits on top of the front-end and consists of layers formed of insulators through which conductive metal wires called interconnects connect the electrical devices of the front-end (see cross-sectional side view in Figure 11).[145] Different combinations of transistors and other electrical devices, wired in particular ways, create various types of "logic gates," which perform basic logical operations. Seven basic logic gates serve as building blocks to create larger "execution units," which implement any desired computation.[146] "Chip design" refers to the layout and structure of these electrical devices and their interconnections.

Figure 11: Transistor and chip structure



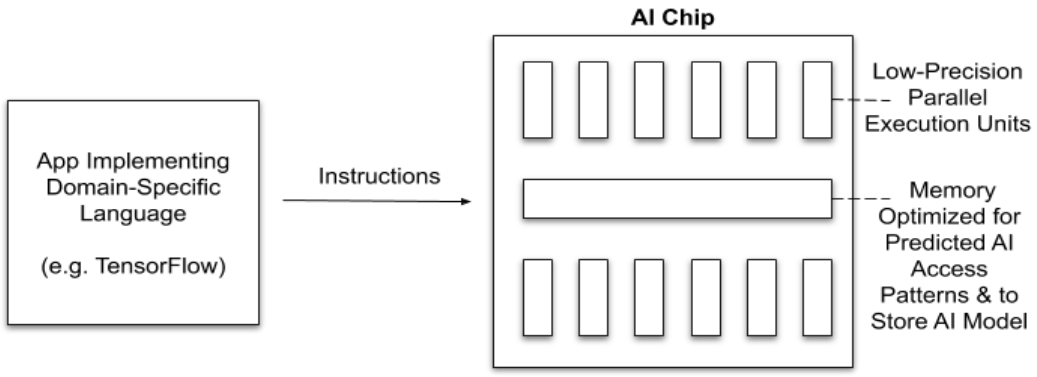
Chips today perform virtually all computing and include many types. First, **logic chips** perform calculations on digital data (0s and 1s) to produce an output. Examples include CPUs, which are general-purpose processors suitable for a wide variety of computing tasks but not specialized for any given tasks, and specialized chips like graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs). GPUs, FPGAs, and ASICs are specialized for improved efficiency and speed for specific applications—such as AI—at the expense of worse-than-CPU efficiency and speed on other applications.

In contrast to logic chips, **memory chips** store the digital data on which logic devices perform calculations. Examples include "dynamic random-access memory" (DRAM), NAND flash memory, and solid-state hard drives. **Analog chips** convert between analog (continuous) data and digital (0s and 1s) data. **Mixed-signal chips** include both digital and analog functions. A **system-on-a-chip (SoC)** is a single chip that includes all necessary computer functions, including logic functions and memory.[147]

Appendix B: How AI Chips Work

AI chips implement specific techniques to increase efficiency and speed relative to CPUs. See Figure 12 for a top-down view of a generic AI chip and a pictorial representation of these techniques, which are described in detail in the following subsections.

Figure 12: Generic AI Chip



Parallel Computing

The most important improvement an AI chip provides over traditional CPUs is parallel computing. AI chips can run a much larger number of simultaneous computations than a CPU can.

Computations for DNNs are especially parallelizable because they are identical and not dependent on the results of other computations. DNN training and inference require a large number of independent, identical matrix multiplication operations, which in turn requires performing many multiplications that are then summed—so called "multiply-and-accumulate" operations.[148,149]

AI chip designs typically include large numbers of "multiply-accumulate circuits" (MACs) in a single chip to efficiently perform matrix multiplication operations within a massively parallel architecture.[150] Performing calculations in parallel also enables the AI chip to complete calculations faster than in sequence. Multiple AI chips connected in a parallel architecture can further increase the degree of parallelism.[151] While advanced CPUs have some degree of parallel architectures, AI chips achieve significantly greater parallelism.[152]

Parallel processing operations use several techniques. **Data parallelism**, the most common form of parallelism, splits the input dataset into different "batches," such that computations are performed on each batch in parallel. These batches can be split across different execution units of an AI chip or across different AI chips connected in parallel. Data parallelism works for any type of neural network. Across a wide variety of neural networks, data parallelism using hundreds to thousands of batches during training achieves the same model accuracy without increasing the total number of required computations. However, greater numbers of batches start requiring more compute to achieve the same model accuracy. Beyond a certain number of batches—for some DNNs, over a million—increasing data parallelism requires more compute without any decrease in time spent training the model, thereby imposing a limit on useful data parallelism.[153]

Model parallelism splits the model into multiple parts on which computations are performed in parallel on different execution units of an AI chip or across different AI chips connected in parallel.[154] For example, a single DNN layer includes many neurons, and one partition may include a subset of those neurons and another includes a different subset of the same neurons. An alternative technique performs calculations on different neural network layers in parallel.[155]

Given the limits on parallelism, scaling up the amount of compute through more AI chips in parallel is not on its own a viable strategy for further AI progress.[156] Instead, research is necessary to produce AI algorithms allowing greater degrees of data and model parallelism, including research to combine techniques to multiply the degree of parallelism.[157]

Low-Precision Computing

Low-precision computing—which sacrifices numerical accuracy for speed and efficiency—is especially suitable for AI algorithms.[158] An x-bit processor contains execution units each built to manipulate data that is represented by x bits. A transistor stores a bit, which can take a value of 1 or 0; therefore, x bit values allow 2^x different combinations. Table 5 shows common values of x for processor data types.

Table 5: Data types

Data types	64-bit	32-bit	16-bit	8-bit
Possible values	18 quintillion (1.8×10^{19})	4.3 billion (4.3×10^{9})	65,536 (6.5×10^{4})	256 (2.5×10^{2})

Higher-bit data types can represent a wider range of numbers (e.g. a larger set of integers) or higher precision numbers within a limited range (e.g. high precision decimal numbers between 0 and 1). Fortunately, with many AI algorithms, training or inference perform as well, or nearly as well, if some calculations are performed with 8-bit or 16-bit data representing a limited or low-precision range of numbers.[159] Even analog computation can suffice for some AI algorithms.[160] These techniques work for the following reasons. First, trained DNNs are often impervious to noise, such that rounding off numbers in inference calculations does not affect results. Second, certain numerical parameters in DNNs are known in advance to have values falling within only a small numerical range—precisely the type of data that can be stored with a low number of bits.[161]

Lower-bit data calculations can be performed with execution units containing fewer transistors. This produces two benefits. First, chips can include more parallel execution units if each execution unit requires fewer transistors. Second, lower-bit calculations are more efficient and require fewer operations. An 8-bit execution unit uses 6x less circuit area and 6x less energy than a 16-bit execution unit.[162]

Memory Optimization

If an AI algorithm's memory access patterns are predictable, AI chips can optimize memory amounts, locations, and types for those predictable uses.[163] For example, some AI chips include sufficient memory to store an entire AI algorithm on-chip.[164] Intra-chip memory access provides major efficiency and speed improvements compared to communication with off-chip memory. Model parallelism becomes an especially useful tool when a model becomes too large to store on a single AI chip; by splitting a model, different portions can be trained on different AI chips connected in parallel.[165]

By contrast, most CPUs have a "Von Neumann" design, which includes a single central bus—a communication system that shares data between the CPU and a separate memory chip storing program code and data. Given the bus' limited bandwidth, the CPU must separately access the code and data sequentially and experiences a "Von Neumann bottleneck," whereby memory-access latency prevents CPUs from achieving speeds enabled by high transistor-switching speeds.[166] The Von Neumann design is useful for general-purpose computing. AI chips, on the other hand, do not require a Von Neumann design or exhibit the Von Neumann bottleneck.

Domain-Specific Languages

Domain-specific languages (DSLs) provide efficiency gains for specialized applications run on specialized chips.[167]

Programmers use computer languages to write computer code (i.e. instructions to a computer) in a human-understandable way. A computer program called a compiler (or an interpreter) then translates this code into a form directly readable and executable by a processor. Different computer languages operate at various levels of abstraction. For example, a high-level programming language like Python is simplified for human-accessibility, but Python code when executed, is often relatively slow due to complexities of converting high-level instructions for humans into machine code optimized for a specific processor. By contrast, programming languages like C operating at a lower-level of abstraction require more complex code (and effort by programmers), but their code often execute more efficiently because it is easier to convert into machine code optimized for a specific processor.[168] However, both examples are general-purpose programming languages whose code can implement a wide variety of computations, but is not specialized to translate efficiently into machine code for specific computations.

By contrast, DSLs are specialized to efficiently program for and execute on specialized chips. A notable example is Google's TensorFlow, which is DSL whose code runs with higher efficiency on AI chips than any general-purpose language would.[169] Sometimes, the advantages of DSLs can be delivered by specialized code libraries like PyTorch: these code libraries package knowledge of specialized AI-processors in functions that can be called by general-purpose languages (such as Python in this case).[170]

Appendix C: AI Chip Benchmarking Studies

Many researchers have attempted to benchmark DNN efficiency and speed of AI chips against CPUs and each other, with varying results depending on variables including chip type, whether the computation is training or inference, and DNN type (i.e. the benchmark). DNN types include fully connected neural networks (FCNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM), residual networks (ResNets), and others. Table 6 presents results for a sampling of key recent studies on various comparisons between server grade and PC grade chips.[171] Notably, even some CPUs are being designed with improved AI capabilities (e.g. 200x speed increases), which may reduce the difference between CPU and AI chip results.[172] Finally, all of the chips listed below are U.S. chips, except for the U.K. Graphcore chip and the Chinese Cambricon chip. Little rigorous benchmarking data exists for Chinese AI chips.

Table 6: AI Chip Efficiency and Speed Benchmarking Studies for DNNs

Author and year	Chip comparison	Computation type	DNN types	Efficiency	Speed
Harvard-1 (2019)[173]	Nvidia Tesla V100 **GPU** vs. Intel Skylake **CPU**	Training	FCNN	-	1-100x
	Google TPU v2/v3 **ASIC** vs. Nvidia Tesla V100 **GPU**		CNN, RNN, FCNN	-	0.2-10x
MLPerf (2019)[174]	Google TPU v3 **ASIC** vs. Nvidia Tesla V100 **GPU**	Training	ResNet, SSD, R-CNN, NMT, Transformer, MiniGo	-	0.8-1.2x
Graphcore (2019)[175]	Graphcore IPU **ASIC** vs. **GPU**	Training	Transformer, MLP, Autoencoder, MCMC	-	1-26x
		Inference	Transformer, ResNext	-	3-43x
Google (2017)[176]	Nvidia K80 **GPU** vs. Intel Haswell **CPU**	Inference	Weighted average of MLP, CNN, RNN	3x	2x

	Improved Google TPU v1 **ASIC** vs. Intel Haswell **CPU**			196x	50x
Stanford (2017)[177]	Nvidia Tesla K80 or P100 **GPU** vs. 16 Intel Broadwell **vCPUs**	Training	ResNet	-	2-12x
		Inference		-	5-3x
Hong Kong Baptist (2017)[178]	Nvidia GTX 1080 **GPU** vs. Intel Xeon **CPU**	Training	FCNN, CNN, RNN, ResNet	-	7-572x
Harvard-2 (2016)[179]	Nvidia GeForce GTX 960 **GPU** vs. Intel Skylake **CPU**	Training	CNN, RNN, FCNN, MemNet	-	3-1,700x
		Inference		-	2-500x
Bosch (2016)[180]	Nvidia GTX Titan X **GPU** vs. Intel Xeon **CPU**	Training	CNN, RNN, FCNN	-	7-29x
		Inference		-	9-30x
Stanford / Nvidia (2016)[181]	Nvidia GeForce Titan X **GPU** vs. Intel Core i7 **CPU**	Inference	Geometric mean of CNN, RNN, LTSM	4-7x	15-16x
	EIE **ASIC** vs. Nvidia GeForce Titan X **GPU**			1,052x	4x
Rice (2016)[182]	Nvidia Jetson TK1 **GPU** vs. Nvidia Jetson TK1 **CPU**	Inference	CNN	4x	16x
Texas State (2016)[183]	Nvidia GeForce Titan X **GPU** vs. Intel Xeon **CPU**	Training	CNN	12x	19x
UCSB / CAS / Cambricon (2016)[184]	Cambricon-ACC **ASIC** vs. Nvidia K40M **GPU**	Mean of training and inference	Geometric mean of MLP, CNN, RNN, LTSM, Autoencoder, BM, RBM, SOM, HNN	131x	3x
Michigan-1 (2015)[185]	Nvidia GTX 770 **GPU** vs. Intel Haswell **CPU**	Inference	DNN	7-25x	5-9x

	Xilinx Virtex-6 **FPGA** vs. Intel Haswell **CPU**			20-70x	10-18x
Michigan-2 (2015)[186]	Nvidia K40 **GPU** vs. Intel Xeon **CPU**	Inference	CNN, DNN	-	40-180x
Peking / UCLA (2015)[187]	Xilinx Virtex-7 **FPGA** vs. Intel Xeon **CPU**	Inference	CNN	25x	5x
Microsoft (2015)[188]	Nvidia GeForce Titan X **GPU** vs. Intel Xeon **CPU**	Inference	CNN	77x	78x
	Intel Arria 10 GX1150 **FPGA** vs. Intel Xeon **CPU**			102x	16x
ETH Zurich / Bologna (2015)[189]	Nvidia GTX 780 **GPU** vs. Intel Xeon **CPU**	Inference	CNN	-	23x
NYU / Yale (2011)[190]	Nvidia GTX 480 **GPU** vs. Intel DuoCore **CPU**	Inference	CNN	34x	267x
	Xilinx Virtex-6 **FPGA** vs. Intel DuoCore **CPU**			368x	134x

Appendix D: Chip Economics Model

This appendix explains the assumptions and calculations underlying the line-item values for the CSET chip economics model presented in Table 3. The model takes the perspective of a fabless firm designing a chip, purchasing foundry services to fabricate the chip, and operating the chip, all in 2020 when TSMC expects to mass produce 5 nm node chips.

Chip Transistor Density, Design Costs, and Energy Costs

Chip transistor density. Our model uses, as a baseline, a hypothetical 5 nm GPU with the specifications of Nvidia's Tesla P100 GPU, which OpenAI used in 2018 to train the breakthrough AI algorithm OpenAI Five.[191] The P100 GPU is fabricated at TSMC at the 16 nm node and contains 15.3 billion transistors in a chip (die) area of 610 mm^2, translating to a transistor density of 25 MTr/mm^2.[192] A 300 mm diameter silicon wafer produces 71.4 610 mm^2 GPUs on average.[193] Our hypothetical 5 nm GPU has a chip area of 610 mm^2 and given its greater transistor density than the P100 GPU, 90.7 billion transistors.[194] Table 7 presents estimated TSMC transistor densities for nodes between 90 and 5 nm. For nodes in the 90 to 7 nm range, our model uses a hypothetical GPU with identical specifications, including transistor count, as the hypothetical 5 nm GPU, except with a transistor density associated with the hypothetical node. Therefore, GPUs with nodes larger than 5 nm will respectively have an area greater than 610 mm^2, resulting in differing numbers of GPUs fabricated per wafer as shown in Table 7. However, the model could equivalently be interpreted as accounting for one chip at the 5 nm node, but at any given larger node, multiple chips totaling the same transistor count as one 5 nm chip.

Table 7: TSMC transistor density[195]

Node (nm)	90	65	40	28	20	16/12	10	7	5
Density (MTr/mm^2)	1.6	3.3	7.7	15.3	22.1	28.9	52.5	96.3	171.3
Average chips per wafer	0.7	1.4	3.2	6.4	9.2	12.0	21.9	40.1	71.4

Design costs per chip. For chip design costs for nodes between 5 to 65 nm, we use the IBS estimates presented in Table 1 for 5 to 65 nm. For the 90 nm

node, we extrapolate the cost based on the IBS data.[196] We assume production of 5 million units.[197] For the 5 nm node, we obtain a design cost per chip of $108. For larger nodes, the chips in our model require a larger chip area (or equivalently, more chips), therefore for larger nodes the per chip cost is determined by dividing by a smaller number of units.[198] In practice, the design cost per chip could vary widely due to varying production volume for different AI chips or depending on whether a fabless firm reuses old chip designs or IP cores.[199]

Annual energy cost per chip. The Nvidia Tesla P100 GPU runs at 9.526 teraflops for 32-bit floating point calculations with a thermal design power (TDP) of 250 watts.[200] When a typical high-end GPU is idle, it uses 31 percent of TDP,[201] while peak utilization uses 100 percent of TDP. We adopt OpenAI's assumption that a typical GPU exhibits a utilization rate of 33 percent during training.[202] For simplicity, we assume a linear relationship between utilization and power consumption,[203] yielding an estimate that the Nvidia Tesla P100 GPU uses 54 percent of TDP during training.[204] We then use an estimated electricity cost of $0.07625 per kilowatt-hour to determine chip annual energy usage.[205] We then increase the energy costs by 11 percent to account for cooling and other costs based on Google's report that its data centers have an average power usage effectiveness (PUE) of 1.11.[206] We also increase energy costs to account for a power supply efficiency of 95 percent. For nodes other than 16 nm, we adjust electricity cost according to TSMC's node-to-node comparative power consumption data presented in Figure 6.[207]

Foundry, Assembly, Test, and Packaging Costs

We first use TSMC's historical financial data to estimate foundry sale price per chip for each node. Initially, we note foundry revenue equals capital assets consumed (i.e. depreciated) plus other costs plus operating profit. Table 8 breaks down the unweighted yearly average of the percentage contributions of these components for the period from 2004 to 2018. Table 8 also lists the unweighted yearly average of TSMC's capital depreciation rate for this period. For the remainder of the calculations, we use these values.[208]

Table 8: Costs used in the model (taken from TSMC's financials)[209]

Financial line-item	Average from 2004 to 2018
Capital consumed (i.e. depreciated)	24.93%

Other costs	39.16%
Operating profit	35.91%
Revenue	*100%*
Capital depreciation rate	25.29%

We first calculate capital consumed per wafer for each node based on TSMC's capital investments, annual wafer capacity of its foundries, and the capital depreciation rate as follows. Then, we will infer other costs and markup per chip using Table 8.

To obtain capital consumed per wafer, we first calculate capital investment per wafer processed per year. TSMC currently operates three GigaFabs (Fabs 12, 14, and 15) with a fourth (Fab 18) scheduled to come online in 2020 with expansion thereafter.[210] These four fabs include a total of 23 fab locations each with a known initial capital investment in 2020 USD—representing investments in facilities, clean rooms, and purchase of SME—and annual 300 mm wafer processing capacity. Dividing these two values produces the capital investment per wafer processed per year for each fab location. Figure 13 plots these 23 values according to the year in which each fab location began processing wafers.[211] When fit to an exponential trendline, capital investment per wafer processed per year shows an 8.3 percent increase per year, with a value of $4,649 in 2004 and $16,746 in 2020.

Figure 13: Capital investment per 300 mm wafer processed per year[212]

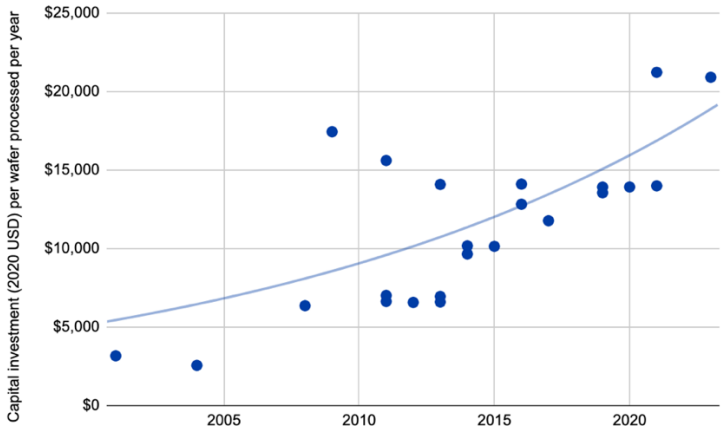


Table 9 on line 2 lists the trendline-fitted capital investment per wafer processed per year for each node based on the year and quarter of introduction of that node listed in line 1.[213] Based on the yearly depreciation rate of 25.29 percent from Table 8, line 3 lists net capital depreciation rate for each year's capital investment per wafer processed per year from the perspective of the year 2020. Typical capital depreciation schedules reach a maximum. Here, we assume a maximum of 65 percent.[214] Line 4 lists undepreciated capital remaining at the start of 2020, which we obtain by depreciating the capital investment per wafer processed per year using the net capital depreciation rate. Line 5 lists how much of any given year's undepreciated capital the processing of one wafer would consume in 2020. This value is obtained by multiplying any given year's undepreciated capital by the capital depreciation rate of 25.29 percent.[215] Line 6 lists other costs and markup per chip for each node, which we obtain by multiplying capital consumed per chip by the ratio of other costs and operating profit as a percentage of revenue (75.07 percent) and capital consumed as a percentage of revenue (24.93 percent), as obtained from Table 8. To avoid complexity, for each node we assume a flat ratio of capital consumed to other costs and markup.[216] Line 7 lists the foundry sale price per wafer, which is the sum of capital consumed per wafer (line 5) and other costs and markup per wafer (line 6).[217] In line 8, we convert the per wafer value to a per chip value by dividing by the number of chips per wafer of a given year's node listed in Table 7.[218] Foundry sale price per chip values in line 8 are not integer fractions of foundry sale price per wafer values in line 7, as for each node the average number of chips per wafer is not an integer.[219]

Table 9: Calculation of foundry sale price per chip in 2020 by node

Line	Node (nm)	90	65	40	28	20	16/12	10	7	5
1	Mass production year and quarter[220]	2004 Q4	2006 Q4	2009 Q1	2011 Q4	2014 Q3	2015 Q3	2017 Q2	2018 Q3	2020 Q1
2	Capital investment per wafer processed per year	$4,649	$5,456	$6,404	$8,144	$10,356	$11,220	$13,169	$14,267	$16,746
3	Net capital depreciation at start of 2020 (25.29% / year)	65%	65%	65%	65%	65%	65%	55.1%	35.4%	0.0%
4	Undepreciated capital per wafer processed per year (remaining value at start of 2020)	$1,627	$1,910	$2,241	$2,850	$3,625	$3,927	$5,907	$9,213	$16,746
5	Capital consumed per wafer processed in 2020	$411	$483	$567	$721	$917	$993	$1,494	$2,330	$4,235

6	Other costs and markup per wafer	$1,293	$1,454	$1,707	$2,171	$2,760	$2,990	$4,498	$7,016	$12,753
7	Foundry sale price per wafer	$1,650	$1,937	$2,274	$2,891	$3,677	$3,984	$5,992	$9,346	$16,988
8	Foundry sale price per chip	$2,433	$1,428	$713	$453	$399	$331	$274	$233	$238

Finally, we calculate assembly, test, and packaging (ATP) costs per chip. Under the fabless-foundry model, fabless firms design chips, and purchase foundry services from foundries and assembly, test, and packaging (ATP) services from outsourced semiconductor assembly and test (OSAT) firms. We can derive OSAT costs based on the ratio of the fab market to the ATP market. Total 2018 OSAT revenue was $30 billion.[221] Because OSAT revenues are about 36.8% of ATP revenues, the total ATP market was $81.5 billion.[222] Total 2018 foundry revenue was $62.9 billion.[223] Total 2018 IDM revenue was $312.8 billion and total 2018 fabless revenue was $108.9 billion for a ratio of 2.9.[224] We multiply total 2018 foundry revenue by this ratio to obtain an estimated $180.6 billion in fab revenue attributable to IDMs. Adding this value to 2018 foundry revenue gives us a total semiconductor fab revenue of $243.5 billion. Finally, dividing the ATP market of $81.5 billion by the fab market of $243.5 billion equals 33.48% percent. We calculate OSAT costs for each node by multiplying the foundry sale price by this percentage.[225]

Authors

Saif M. Khan is a Research Fellow at Georgetown's Center for Security and Emerging Technology (CSET). His work focuses on AI policy, semiconductor supply chains, China's semiconductor industry, and U.S. trade policy. Alexander Mann is a Research Collaborator with Georgetown's Center for Security and Emerging Technology (CSET) and a Research Associate at the University of Maryland.

Acknowledgments

For helpful discussions, comments, and input, great thanks go to Jeff Alstott, Zachary Arnold, Carrick Flynn, Michael Fritze, Lorand Laskai, Igor Mikolic-Torreira, Ilya Rahkovsky, Jacob Strieb, Helen Toner, Alexandra Vreeman, and Lynne Weil. The authors are solely responsible for all mistakes.

Document Identifier: doi: 10.51593/20190014

Endnotes

[1] Saif M. Khan, "Maintaining the AI Chip Competitive Advantage of the United States and its Allies" (Washington, DC: Center for Security and Emerging Technology, December 2019), https://cset.georgetown.edu/wp-content/uploads/CSET-Maintaining-the-AI-Chip-Competitive-Advantage-of-the-United-States-and-its-Allies-20191206.pdf.

[2] John L. Hennessy and David A. Patterson, "A New Golden Age for Computer Architecture," *Communications of the ACM* 62, no. 2 (February 2019), Vol. 62 No. 2: 52, https://dl.acm.org/doi/10.1145/3282307.

[3] Samuel K. Moore, "Another Step Toward the End of Moore's Law," *IEEE Spectrum*, May 31, 2019, https://spectrum.ieee.org/semiconductors/devices/another-step-toward-the-end-of-moores-law.

[4] A new node name typically always represents 70% of the transistor length of a previous node. And 70% length translates into a transistor doubling per unit area. Historically, node names referred to actual lengths of transistors, but in recent years transistors have been changing shapes as they shrink. Therefore, each new node name is meant to represent a doubling in transistor density rather than any transistor feature size.

[5] This principle is called Dennard Scaling. Relatedly, both current and voltage scale linearly down with transistor length.

[6] Hennessy et al., "New Golden Age," 52.

[7] Chris A. Mack, "Lecture 2: Moore's Law," CHE323/CHE384: Chemical Processes for Micro- and Nanofabrication, University of Texas, Austin, TX, 2013, http://www.lithoguru.com/scientist/CHE323/Lecture2.pdf.

[8] The speed measurements are based on a benchmark test called SPECint, which involves calculations of integers. The 1978 to 1986 period used a chip design architecture called "complex instruction set computer" (CISC) and the 1986 to 2003 period used a simpler and more efficient architecture called "reduced instruction set computer" (RISC). In the latter period, the most important design improvement was a form of parallel computing called instruction-level parallelism (ILP) which allowed multiple instructions to be executed by a single CPU chip simultaneously. Hennessy et al., "New Golden Age," 54. Other types of parallel computing include bit-level parallelism and task parallelism.

[9] This trend is called Koomey's Law. Jonathan G. Koomey, Stephen Beard, Marla Sanchez, and Henry Wong, "Implications of Historical Trends in the Electrical Efficiency of Computing," *IEEE Annals in the History of Computing* 33, no. 3 (March 29, 2010), https://ieeexplore.ieee.org/document/5440129.

[10] This trend is based on measurements of AMD CPUs. An analog relating to typical use rather than peak use still shows doublings every 1.5 years driven by improved designs and power management. Jonathan Koomey and Sam Naffziger, "Energy Efficiency of Computing: What's Next?", *Electronic Design*, November 28, 2016,

https://www.electronicdesign.com/microprocessors/energy-efficiency-computing-what-s-next. These efficiency improvements have not typically come at the expense of speed. That is, new chips continually incorporate both the efficiency and speed improvements shown in Figure 1. Koomey et al., "Electrical Efficiency of Computing," 50.

[11] For transistor data, see Max Roser and Hannah Ritchie, "Technological Progress," *Our World in Data*, 2019, https://ourworldindata.org/technological-progress. For efficiency data, see Koomey et al., "Energy Efficiency of Computing." For speed data, see Hennessy et al., "New Golden Age," 54.

[12] Ed Sperling, "Why Scaling Must Continue," *Semiconductor Engineering*, August 15, 2019, https://semiengineering.com/why-scaling-must-continue/.

[13] An example of on-chip memory is static random-access memory (SRAM).

[14] Clair Brown and Greg Linden, *Chips and Change: How Crisis Reshapes the Semiconductor Industry* (Cambridge, MA: MIT Press, August 19, 2011), 67.

[15] Ibid, 68–69, 72–73. This functionality of EDA software is called electronic system-level (ESL) design and verification.

[16] Specifically, the insulative layer is the thinnest layer of the MOSFET. Because it became too thin, electrical current leaked across the insulative layer between the metal gate and the semiconductor channel of the MOSFET.

[17] Mark T. Bohr, Robert S. Chau, Tahir Ghani, and Kaizad Mistry, "The High-k Solution," *IEEE Spectrum*, October 1, 2007, https://spectrum.ieee.org/semiconductors/design/the-highk-solution.

[18] Specifically, a planar field-effect transistor (FET), which is a type of MOSFET, includes metal, insulator, and semiconductor regions that are each flat layers stacked on each other, as shown in Figure 11.

[19] This structure is called the FinFET, which is another type of MOSFET. The FinFET reduces the contact region between the insulator and the semiconductor, and the metal gate surrounds the insulator on three sides, resulting in less current leakage into the semiconductor. Peide Ye, Thomas Ernst, and Mukesh V. Khare, "The Nanosheet Transistor Is the Next (and Maybe Last) Step in Moore's Law," *IEEE Spectrum*, July 30, 2019, https://spectrum.ieee.org/semiconductors/devices/the-nanosheet-transistor-is-the-next-and-maybe-last-step-in-moores-law. Besides FinFET, other new MOSFET structures with some production included the ultrathin body silicon-on-insulator (UTB SOI) and fully depleted silicon-on-insulator (FD SOI), and the Tri-Gate. Khaled Ahmed and Klaus Schuegraf, "Transistor Wars," *IEEE Spectrum*, October 28, 2011, https://spectrum.ieee.org/semiconductors/devices/transistor-wars; Handel Jones, "FD SOI Benefits Rise at 14nm," *EE Times*, June 13, 2016, https://web.archive.org/web/20190828144514/https://www.eetimes.com/author.asp?section_id=36&doc_id=1329887.

[20] This new MOSFET structure is called the nanosheet transistor and may be the final structure for all future MOSFET-based nodes. Ye et al., "Nanosheet Transistor."

[21] Various other options to squeeze out the final remnants of MOSFET-based improvements include yet another MOSFET structure called the gate-all-around MOSFET, 3D stacking of MOSFETs, near- and in-memory placement of MOSFETs, chiplets, optical interconnects, and advanced packaging techniques. Mark Lapedus, "What's the Right Path for Scaling?", *Semiconductor Engineering*, January 2, 2019, https://semiengineering.com/whats-the-right-path-for-scaling/; Joel Hruska, "Chiplets Are Both Solution to and Symptom of a Larger Problem," *ExtremeTech*, April 30, 2019, https://www.extremetech.com/computing/290450-chiplets-are-both-solution-and-symptom-to-a-larger-problem; Joel Hruska, "Samsung Unveils 3nm Gate-All-Around Design Tools," *ExtremeTech*, May 16, 2019, https://www.extremetech.com/computing/291507-samsung-unveils-3nm-gate-all-around-design-tools. Eventually, only a new computing paradigm that uses a new logic switch—different than the MOSFET, which has been used for the last 60 years—will enable further progress. New logic switches could be implemented using carbon nanotubes, memristors, silicon photonics, and qubits. For a more comprehensive list of new computing paradigms, see President's Council of Advisors on Science and Technology, *Ensuring Long-Term U.S. Leadership in Semiconductors* (Washington, DC: Executive Office of the President, January 2017), 28, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_ensuring_long-term_us_leadership_in_semiconductors.pdf.

[22] See also Hassan N. Khan, David A. Hounshell, and Erica R. H. Fuchs, "Science and research policy at the end of Moore's law," *Nature Electronics* 1, no. 1 (January 8, 2018): 14–18, https://www.nature.com/articles/s41928-017-0005-9.

[23] In 2007, a $20 million design needed $400 million in sales to generate a normal level of profitability for a chip design firm. Brown et al., *Chips and Change*, 64.

[24] Neil Thompson and Svenja Spanuth, "The Decline of Computers As a General Purpose Technology: Why Deep Learning and the End of Moore's Law are Fragmenting Computing," December 12, 2018, 12 and 24–32, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3287769.

[25] The supply chain necessary for making advanced chips includes: basic research; the production of electronic design automation (EDA) software used to design chips; chip design; the production of semiconductor manufacturing equipment (SME); the procurement and processing of materials such as silicon; the manufacture of chips in fabs based on chip designs; assembly, test, and packaging of manufactured chips using SME; and distribution and end-use of chips. These functions are usually separated in different companies, except integrated device manufacturers (IDMs), most notably Intel and Samsung, often perform chip design, fabrication, assembly, packaging, and testing. Otherwise, chip design occurs in "fabless" firms who then outsource designs to fabs called "foundries" that provide contract manufacturing (also called "foundry services"). "OSAT" firms then perform assembly, test, and packaging.

[26] Thompson et al., "The Decline of Computers As a General Purpose Technology," 32–35; John VerWey, "The Health and Competitiveness of the U.S. Semiconductor Manufacturing

Equipment Industry," *Office of Industries and Office of Economics of the U.S. International Trade Commission*, 5–6, July 2019, https://www.usitc.gov/publications/332/working_papers/id_058_the_health_and_competitiveness_of_the_sme_industry_final_070219checked.pdf.

[27] Each of these values are in nominal U.S. dollars, i.e. not adjusted for inflation. For the semiconductor market growth rate, see Semiconductor Industry Association, "2019 Factbook," May 20, 2019, 2, https://www.semiconductors.org/resources/2019-sia-factbook/.

[28] Thompson et al., "The Decline of Computers As a General Purpose Technology," 32–36.

[29] Ibid, 34; Moore, "End of Moore's Law." GlobalFoundries in 2018 announced it would not progress beyond the 14 nm node.

[30] This data includes 23 fab locations for TSMC's four GigaFabs (Fabs 12, 14, 15, and 18). SEMI, *World Fab Forecast*, May 2019 edition.

[31] The 1979 cost is for a g-line photolithography tool. Chris A. Mack, "Lecture 2: Semiconductor Economics," CHE323/CHE384: Chemical Processes for Micro- and Nanofabrication, University of Texas, Austin, TX, 2013, http://www.lithoguru.com/scientist/CHE323/Lecture3.pdf. The 2019 cost is based on ASML's average sale price—reported in its 2019 Q4 financial statements—of eight extreme ultraviolet photolithography tools sold in that quarter. In the 1980s and 1990s, photolithography tool prices increased 17% per year. Chris A. Mack, "Milestones in Optical Lithography Tool Suppliers," University of Texas, Austin, TX, 2005, 25, http://www.lithoguru.com/scientist/litho_history/milestones_tools.pdf; see also "McKinsey on Semiconductors" (McKinsey & Company, Autumn 2011), 10, https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/semiconductors/pdfs/mosc_1_revised.ashx. For info on rising photomask costs, one subsegment of photolithography, see Moein Khazraee, Lu Zhang, Luis Vega, and Michael Bedford Taylor, "Moonwalk: NRE Optimization in ASIC Clouds or, accelerators will use old silicon," *Proc. 22nd Int'l Conf. Architectural Support for Programming Languages and Operating Systems* (2017), 514, https://homes.cs.washington.edu/~vegaluis/pdf/asplos17_khazraee_moonwalk.pdf; "Semiconductor Wafer Mask Costs," anysilicon, September 15, 2016, https://anysilicon.com/semiconductor-wafer-mask-costs/.

[32] CSET analysis of data from SEMI, *World Fab Forecast*, May 2019 Edition. The data represents the number of companies currently operating at the listed node or smaller.

[33] CSET research. For each node, the data represents the number of companies operating when that node was introduced. However, it remains the case today that only one company has reached 5 nm and only two companies sell a significant volume of photolithography equipment for ≤90 nm. For photolithography technology evolution, see
SET research. For each node, the data represents the number of companies operating when that node was introduced. However, it remains the case today that only one company has reached 5 nm and only two companies sell a significant volume of photolithography equipment for ≤90 nm. For photolithography technology evolution, see Robert Castellano,

"Canon's Nanoimprint Lithography: A Chink In ASML Holding's Armor", Seeking Alpha, March 19, 2019, https://seekingalpha.com/article/4249762-canons-nanoimprint-lithography-chink-asml-holdings-armor. For history of photolithography companies, see Atsuhiko Kato, "Chronology of Lithography Milestones," May 2007, http://www.lithoguru.com/scientist/litho_history/Kato_Litho_History.pdf and Mack, "Optical Lithography Tool Suppliers."

[34] Chip design costs per transistor appear flat, as between TSMC's 28 and 7 nm nodes, transistor density increased by 6.3x while chip design costs increased by 5.8x.

[35] See, e.g., "AMD vs Intel Market Share," PassMark Software, January 4, 2020, https://www.cpubenchmark.net/market_share.html. By contrast, the mobile system-on-a-chip market, which includes CPUs, is controlled by Qualcomm (United States), Apple (United States), Samsung (South Korea), MediaTek (Taiwan), and HiSilicon/Huawei (China). "MediaTek lost glory as Qualcomm gained smartphone processor market share," *telecomlead*, May 29, 2018, https://www.telecomlead.com/telecom-statistics/mediatek-lost-glory-as-qualcomm-gained-smartphone-processor-market-share-84373. Additionally, many of these firms license British chip design firm ARM's instruction set architecture and IP cores. Meanwhile, China's CPU design industry is not competitive. In 2016, a Chinese supercomputer company Sugon Information Industry Co. licensed AMD's x86 CPU designs, giving Sugon a boost in developing its own x86 CPUs. Kate O'Keefe and Brian Spegele, "How a Big U.S. Chip Maker Gave China the 'Keys to the Kingdom,'" *Wall Street Journal*, June 27, 2019, https://www.wsj.com/articles/u-s-tried-to-stop-china-acquiring-world-class-chips-china-got-them-anyway-11561646798. Separately, China has been developing a homegrown CPU called Loongson for the last decade. Gong Zhe, "Backgrounder: What is 'Loongson' computer chip?", *CGTN*, July 18, 2019, https://news.cgtn.com/news/2019-07-18/Backgrounder-What-is-Loongson-computer-chip--Inz7pUO4O4/index.html. The latest generation of Loongson CPUs have reached the 12 nm node and have improved in performance enough to convince Chinese PC-maker Lenovo and Chinese supercomputer manufacturer Sugon to use Loongson CPUs. Chen Qingqing, "High-tech de-Americanization accelerated," *Global Times*, December 25, 2019, https://www.globaltimes.cn/content/1174804.shtml.

[36] These data are not adjusted for inflation. For IBS data, see "FinFET and FD SOI: Market and Cost Analysis" (International Business Strategies, September 18, 2018), 8, http://soiconsortium.eu/wp-content/uploads/2018/08/MS-FDSOI9.1818-cr.pdf; for Gartner data, see Mark Lapedus, "Foundry Challenges in 2018," *Semiconductor Engineering*, December 17, 2017, https://semiengineering.com/foundry-challenges-in-2018/; "AI chips may be only the vassal of FPGA," South China Venture Capital, Accessed January 4, 2020, http://www.scvc.cn/index.php?m=Article&a=show&id=405&l=en.

[37] Nicholas Bloom, Charles I. Jones, John Van Reenen, and Michael Webb, "Are Ideas Getting Harder to Find?", 2, 18–19, https://web.stanford.edu/~chadj/IdeaPF.pdf. The actual number of workers is likely to be higher or lower depending on prevailing wages.

[38] Ibid. The semiconductor innovation rate measured by Moore's Law is faster than innovation rates in the economy as a whole, causing researcher productivity to fall more rapidly than in other sectors of the economy, which are also seeing declines in researcher

productivity. Ibid, 37–44, 49. The semiconductor innovation rate has been fast because demand for computing power is high given its general-purpose economic value. Ibid, 44.

[39] Kenneth Flamm, "Measuring Moore's Law: Evidence from Price, Cost, and Quality Indexes," April 2018, 7 (Figure 3), 8 (Table 1), 21 (Table 7), 23 (Table 8 and Figure 8), https://www.nber.org/papers/w24553.

[40] Compare Intel's optimistic estimates in Ibid with Global Foundries' and Handel Jones' pessimistic estimates in Ibid, 13 (Figures 5–6). See also Jones, "FD SOI Benefits Rise at 14nm"; Khan et al., "Science and research policy at the end of Moore's law."

[41] "WSTS Semiconductor Market Forecast Fall 2019" (World Semiconductor Trade Statistics, December 3, 2019), 2, https://www.wsts.org/esraCMS/extension/media/f/WST/4298/WSTS_nr-2019_11.pdf.

[42] Statista, Semiconductor sales revenue worldwide from 1987 to 2020, June 2019, https://www.statista.com/statistics/266973/global-semiconductor-sales-since-1988/.

[43] Kaizad Mistry, "10 nm Technology Leadership," Intel Technology and Manufacturing Day, 2017, https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/03/Kaizad-Mistry-2017-Manufacturing.pdf.

[44] "Logic Technology," TSMC, accessed January 4, 2020, https://www.tsmc.com/english/dedicatedFoundry/technology/logic.htm.

[45] Thompson et al., "The Decline of Computers As a General Purpose Technology," 41. In the early 2000s, approximately 80 percent of TSMC's sales came from the three leading nodes, but by around 2009, this percentage shrank and stabilized at approximately 55 percent. Ibid.

[46] Kim Eun-jin, "Samsung Electronics Narrows Gap with First-Ranked TSMC in Foundry Business," *Business Korea*, March 29, 2019, http://www.businesskorea.co.kr/news/articleView.html?idxno=30407.

[47] Thompson et al., "The Decline of Computers As a General Purpose Technology," 32–35, 40.

[48] Data from TSMC's financial statements. "Financial Results," TSMC, accessed January 16, 2020, https://www.tsmc.com/english/investorRelations/quarterly_results.htm.

[49] Brown et al., *Chips and Change*, 42.

[50] These factors explain why new nodes continue to be introduced at all. Other factors that are difficult to measure and validate include hopes that unexpected innovations make new nodes more profitable than they currently appear and marketing benefits for being seen as a technology leader for any company at a leading node.

[51] Flamm, "Measuring Moore's Law," 9 (Figure 4), 11. First, transistors at ≤65 nm can stop working if power is reduced too much, as electrical current does not properly flow across the semiconductor channel from the source to the drain and instead leaks or is disrupted by ambient heat. Ahmed, "Transistor Wars"; Mack, "Moore's Law"; Hennessy et al., "New Golden Age," 52. Second, as transistors shrink, the lengths of interconnects between transistors stay the same, such that when transistor delays become sufficiently low, unchanging interconnect lengths become a bottleneck preventing transistor speed improvements. In fact, resistance-capacitance (RC) delays increase at small scales. Chris A. Mack, "Lecture 2: Device Interconnect, part 2," CHE323/CHE384: Chemical Processes for Micro- and Nanofabrication, University of Texas, Austin, TX, 2013, http://lithoguru.com/scientist/CHE323/Lecture29.pdf; Mark Lapedus, "Big Trouble at 3nm," *Semiconductor Engineering*, June 21, 2018, https://semiengineering.com/big-trouble-at-3nm/.

[52] Moore, "End of Moore's Law"; Anton Shilov, "TSMC Announces Performance-Enhancing 7nm and 5nm Process Technologies," *AnandTech*, July 30, 2019, https://www.anandtech.com/show/14687/tsmc-announces-performanceenhanced-7nm-5nm-process-technologies. The most recent speed improvements are attributable to faster transistor shapes, more advanced chip packaging to fight against current leakage and interference from ambient heat, and replacement of copper with cobalt for interconnects which reduces the interconnect delay. Sundeep Bajikar, "Technology Trends and Innovations in 2019," *Applied Materials*, January 23, 2019, http://blog.appliedmaterials.com/technology-trends-innovations-2019.

[53] Moore, "End of Moore's Law"; Shilov, "TSMC Announces Performance-Enhancing 7nm and 5nm Process Technologies."

[54] For TSMC data, see TSMC, "Logic Technology."

[55] For 5 nm, see "Samsung Successfully Completes 5nm EUV Development to Allow Greater Area Scaling and Ultra-low Power Benefits," *Samsung Newsroom*, April 16, 2019, https://news.samsung.com/global/samsung-successfully-completes-5nm-euv-development-to-allow-greater-area-scaling-and-ultra-low-power-benefits. For 7 nm, see "Samsung Electronics Starts Production of EUV-based 7nm LPP Process,' *Samsung Newsroom*, October 18, 2018, https://news.samsung.com/global/samsung-electronics-starts-production-of-euv-based-7nm-lpp-process. For 10 nm, see "Samsung Starts Industry's First Mass Production of System-on-Chip with 10-Nanometer FinFET Technology," *Samsung Newsroom*, October 17, 2016, https://news.samsung.com/global/samsung-starts-industrys-first-mass-production-of-system-on-chip-with-10-nanometer-finfet-technology. For 14 nm, see "Samsung Announces Mass Production of Industry's First 14nm FinFET Mobile Application Processor," *Samsung Newsroom*, February 16, 2015, https://news.samsung.com/global/samsung-announces-mass-production-of-industrys-first-14nm-finfet-mobile-application-processor.

[56] Some studies show that frequency scaling has ended for Intel CPUs. Joel Hruska, "The death of CPU scaling: From one core to many — and why we're still stuck," *ExtremeTech*, February 1, 2012, https://www.extremetech.com/computing/116561-the-death-of-cpu-scaling-from-one-core-to-many-and-why-were-still-stuck. These results can be reconciled

with Intel's reported speed improvements: given tradeoffs between further improvements in speed and efficiency, Intel may have chosen efficiency.

[57] For 10 nm, see Mistry, "10 nm Technology Leadership," 31. For 14 nm, see Mark Bohr, "14 nm Process Technology: Opening New Horizons," Intel Developer Forum, 2014, 39, https://www.intel.com/content/dam/www/public/us/en/documents/technology-briefs/bohr-14nm-idf-2014-brief.pdf at 39; William Holt, "Investor Meeting 2014," Intel, 2014, 4, https://www.sec.gov/Archives/edgar/data/50863/000005086314000078/exhibit99_2.pdf. For 22 nm, see Mark Bohr and Kaizad Mistry, "Intel' Revolutionary 22 nm Transistor Technology," Intel, May 2011, 20, 22, https://www.intel.com/content/dam/www/public/us/en/documents/presentation/revolutionary-22nm-transistor-technology-presentation.pdf. For 32 nm, see Mark Bohr, "Moore's Law Leadership," Intel Technology and Manufacturing Day, 2017, 33 https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/03/Mark-Bohr-2017-Moores-Law.pdf (power reduction estimated based on chart); see also Mark Bohr, "Intel 32nm Technology," Intel, February 10, 2009, http://download.intel.com/pressroom/kits/32nm/westmere/Mark_Bohr_32nm.pdf. For 45 nm, see "Introducing the 45nm Next-Generation Intel® Core™ Microarchitecture" (Intel, 2007), 4, https://www.intel.com/content/dam/doc/white-paper/45nm-next-generation-core-microarchitecture-white-paper.pdf. For 65 nm, see David Lammers, "IBM-AMD, Intel Describe 65-nm Transistors," *CRN*, December 6, 2005, https://www.crn.com/news/components-peripherals/174901579/ibm-amd-intel-describe-65-nm-transistors.htm; "Intel Updates 65nm Plans," *Electronics Weekly*, August 31, 2004, https://www.electronicsweekly.com/news/research-news/device-rd/intel-updates-65nm-plans-2004-08/.

[58] For some chip designs, interconnect delays can nevertheless bottleneck speeds for the chip as a whole. As a result, transistor speed gains are not always realized as chip speed gains.

[59] CPU efficiency data is based on measurements of AMD CPUs, as reported in Figure 1. The last empirical study of AMD CPU efficiency was published around the time of the 10 nm node's introduction. Therefore, we do not include values for 7 and 5 nm. Additionally, for each node, we use CPU efficiency and speed values according to the time that TSMC introduced the node, as reported in Table 9. Therefore, in Figure 7, TSMC's data is a better comparison than Intel's data to the CPU efficiency and speed data, as Intel's node introduction timeline differs greatly from TSMC's, especially in recent years as Intel's node introduction has slowed.

[60] See, e.g., "Improving high-performance transistor technology," Research Impact, University College London, December 12, 2014, https://www.ucl.ac.uk/impact/case-studies/2014/dec/improving-high-performance-transistor-technology (describing gate dielectrics that reduce energy loss).

[61] Data on cost per computation per second for different GPUs over time are consistent with the hypothesis that relatively more GPU cost improvement has resulted from transistor-level innovation than design-level innovation. "Recent trend in the cost of computing," *AI Impacts*, November 11, 2017, https://aiimpacts.org/recent-trend-in-the-cost-of-computing/.

[62] The CPU speed here is based on a benchmark involving integer calculations. Other benchmarks may involve calculations that exhibit improved speed and efficiency as a result of design innovations across the 65 nm to 5 nm nodes.

[63] See, e.g., Adam Hadhazy, "New microchip demonstrates efficiency and scalable design," *Princeton University*, August 23, 2016, https://www.princeton.edu/news/2016/08/23/new-microchip-demonstrates-efficiency-and-scalable-design (describing energy-reducing chip designs); and Koomey et al., "Energy Efficiency of Computing" (describing multi-core designs as reducing energy).

[64] Sperling, "Why Scaling Must Continue." Potentially slowing improvements in transistor density have also incentivized specialization as discussed in Section V. Some go further and argue that specialization is purely a response to this slowing, and is not itself a factor driving transistor density increases.

[65] This observation is called Amdahl's Law.

[66] Brown et al., *Chips and Change*, 157–158.

[67] Michael Feldman, "The Era of General Purpose Computers is Ending," *The Next Platform*, February 5, 2019, https://www.nextplatform.com/2019/02/05/the-era-of-general-purpose-computers-is-ending/. Some companies integrate CPUs with specialized processors on the same chip. Id. For a more detailed model on cost-benefit ratio of specialization, see Thompson et al., "The Decline of Computers As a General Purpose Technology," 26–32.

[68] See Gaurav Batra, Zach Jacobson, Siddarth Madhav, Andrea Queirolo, and Nick Santhanam, "Artificial-intelligence hardware: New opportunities for semiconductor companies" (McKinsey and Company, January 2019), Exhibits 3 and 5, https://www.mckinsey.com/industries/semiconductors/our-insights/artificial-intelligence-hardware-new-opportunities-for-semiconductor-companies (projecting AI hardware market growth from $17B in 2017 representing 7% of total semiconductor market to $65B in 2025 representing 19% of total semiconductor market, including strong growth for all sectors including server, edge, training, and inference); Joy Dantong Ma, "Chip on the Shoulder: How China Aims to Compete in Semiconductors," *Macro Polo*, September 10, 2019, Figure 2, https://macropolo.org/china-chips-semiconductors-artificial-intelligence/ (projecting AI hardware market growth from $5.66B in 2018 to $83.25B in 2027); "Hitting the accelerator: the next generation of machine-learning chips" (Deloitte, 2018), Figure 13, https://www2.deloitte.com/content/dam/Deloitte/global/Images/infographics/technologymediatelecommunications/gx-deloitte-tmt-2018-nextgen-machine-learning-report.pdf (projecting 4x increase in minimum AI chip sales from 2016 to 2018); "Artificial Intelligence Market Forecasts" (Tractica, 2019), https://www.tractica.com/research/artificial-intelligence-market-forecasts/ (projecting AI hardware market growth from $19.63B in 2018 to $234.6B in 2027); "AI and Semiconductors" (J.P. Morgan, February 7, 2018), 12, (projecting AI hardware market growth from $6B in 2018 representing 1% of total semiconductor market to $33B in 2022 representing 6% of total semiconductor market). A future CSET report will analyze the market for AI chips in more detail.

[69] Cade Metz, "Big Bets on A.I. Open a New Frontier for Chip Start-Ups, Too," *New York Times*, January 14, 2018, https://www.nytimes.com/2018/01/14/technology/artificial-intelligence-chip-start-ups.html.

[70] ASICs specific to AI uses go by many names, such as tensor processing units (TPUs), neural processing units (NPUs), and intelligence processing units (IPUs).

[71] Batra et al., "Artificial-intelligence hardware," Exhibit 6; Dario Amodei and Danny Hernandez, "AI and Compute," *OpenAI*, May 16, 2018, https://openai.com/blog/ai-and-compute/ (see "Eras" section); Ben-Nun, "Demystifying Parallel and Distributed Deep Learning," 1:7 (Figure 3a).

[72] Batra et al., "Artificial-intelligence hardware," Exhibit 6.

[73] Kaz Sato, "What makes TPUs fine-tuned for deep learning?", *Google Cloud Blog*, August 30, 2018, https://cloud.google.com/blog/products/ai-machine-learning/what-makes-tpus-fine-tuned-for-deep-learning.

[74] However, both ASICs and FPGAs typically sacrifice inference accuracy compared to CPUs and GPUs. Brandon Reagen, Paul Whatmough, Robert Adolf, Saketh Rama, Hyunkwang Lee, Sae Kyu Lee, José Miguel Hernández-Lobato, Gu-Yeon Wei, and David Brooks, "Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators," *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, August 25, 2016, 1 (Figure 1), https://ieeexplore.ieee.org/document/7551399. For example, ASICs and FPGAs may implement lower precision computing than CPUs and GPUs, resulting in more opportunities for inference errors.

[75] Batra et al., "Artificial-intelligence hardware," Exhibit 6; David Patterson, "Domain-Specific Architectures for Deep Neural Networks," Google, April 2019, 53, http://www-inst.eecs.berkeley.edu/~cs152/sp19/lectures/L20-DSA.pdf (the TPU v1 is used only for inference, but the TPU v2/v3 is used for both training and inference); Amodei et al., "AI and Compute" (see "Eras" section.)

[76] Tim Hwang, "Computational Power and the Social Impact of Artificial Intelligence," January 18, 2019, 10, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3147971.

[77] Norman P. Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit," April 16, 2017, 2, 8, 12, https://arxiv.org/pdf/1704.04760.pdf.

[78] GPUs may also be more popular than other AI chips due to lock-in of customer use of the GPU-maker Nvidia's supporting software CUDA, which translates AI code written by programmers to different types of machine code that can run on a variety of GPUs. Jeffrey Ding, "ChinAI # 71: What I Got Wrong re: Entity List & Chinese AI Startups," *ChinAI Newsletter*, October 21, 2019, https://chinai.substack.com/p/chinai-71-what-i-got-wrong-re-entity.

[79] Thompson et al., "The Decline of Computers As a General Purpose Technology," 31.

[80] Ibid, 22.

[81] "Price-performance trend in top supercomputers," *AI Impacts*, November 8, 2017, https://aiimpacts.org/price-performance-trend-in-top-supercomputers/.

[82] A few supercomputers are optimized for AI, such as a new MIT Lincoln Lab supercomputer. Kylie Foy, "Lincoln Laboratory's new artificial intelligence supercomputer is the most powerful at a university," *MIT News*, September 27, 2019, http://news.mit.edu/2019/lincoln-laboratory-ai-supercomputer-tx-gaia-0927. Google found that it takes weeks to months for training production runs on its TPU v1, and proposes that DNN supercomputers using its TPU v2/v3 are suitable for large training runs. Patterson, "Domain-Specific Architectures," 40–41.

[83] S. Haug, M. Hostettler, F. G. Sciacca, and M. Weber, "The ATLAS ARC backend to HPC," *Journal of Physics: Conference Series* 664, No. 062057, 2 (Figure 1), https://iopscience.iop.org/article/10.1088/1742-6596/664/6/062057.

[84] Thompson et al., "The Decline of Computers As a General Purpose Technology," 22–23.

[85] Michael Feldman, "New GPU-Accelerated Supercomputers Change the Balance of Power on the TOP500," *Top 500*, June 26, 2018, https://www.top500.org/news/new-gpu-accelerated-supercomputers-change-the-balance-of-power-on-the-top500/. Nvidia GPUs alone power 136 of the TOP500 supercomputers. Nvidia, "Annual Report Pursuant to Section 13 or 15(d) of the Securities Exchange Act of 1934," February 20, 2020, 4, https://s22.q4cdn.com/364334381/files/doc_financials/2020/q4/174b2f06-ba5b-4e0a-b288-e33c46e9a0a4.pdf.

[86] Hwang, "Computational Power," 14.

[87] The Defense Advanced Research Projects Agency gave a similar estimate specifically for the AI chip efficiency premium over CPUs. Defense Advanced Research Projects Agency, *A DARPA Approach to Trusted Microelectronics* (Arlington, VA: Department of Defense), 2 (Figure 1), https://www.darpa.mil/attachments/Background_FINAL3.pdf.

[88] Hwang, "Computational Power," 11–12.

[89] Reagen, "Minerva," 1 (Figure 1).

[90] As discussed in Appendix D, older nodes in our model require larger chips to achieve the same transistor count as a 5 nm node chip. Alternatively, if chips at each node are assumed to be the same size, then the equivalent interpretation is that more chips are produced at older nodes to achieve the same transistor count as the 5 nm node. Given the first interpretation—equivalent-transistor-count chips rather than equivalent-size chips—for nodes larger than 5 nm, volume is smaller than 5 million, as that lower volume would be equivalent to a volume of 5 million chips at the 5 nm node. Additionally, depending on the scenario being modeled, we could have instead assumed that chip design costs for trailing nodes are zero, as the fabless firm could use its old design.

[91] For trailing node chips, operating costs in 2020 dominate production costs in 2020 because the capital used to produce trailing node in 2020 has depreciated. Therefore, capital consumed when producing a trailing chip in 2020 is dramatically lower than it would

have been when that trailing node chip was first introduced for mass production. If our model does not include depreciation, then for all nodes, chip production and operating costs differ by less than a factor of three. See Appendix D for more details on the methodology for capital depreciation.

[92] We observe a similar pattern between other successive nodes (e.g. between 7 nm and 10 nm and between 10 nm and 16/12 nm).

[93] For all comparisons except the 7 versus 5 nm comparison, we use a modified version of the model described in detail in Appendix D. The modified version assumes that for each comparison, the current year is the newer node's introduction year. This causes the depreciation rates to differ for each comparison. For example, with the 10 versus 7 nm comparison, net depreciation would be 0% in 2018 Q3 (the 7 nm introduction time) and 30.5% in 2017 Q2 (the 10 nm introduction time).

[94] Microsoft replaces its server FPGAs after three years. Dan Fay and Derek Chiou, "The Catapult Project - An FPGA view of the Data Center," *Microsoft Research*, 2018, 2, http://www.prime-project.org/wp-content/uploads/sites/206/2018/02/Talk-7-Dan-Fay-The-Catapult-Project-%E2%80%93-An-FPGA-view-of-the-Data-Center.pdf. In a similar vein, Intel reports that servers over 4 years in age provide only 4% of performance capability but are responsible for 65% of energy consumption. "Updating IT Infrastructure," *Intel IT Center*, December 2013, 5, https://www.intel.com/content/dam/www/public/us/en/documents/guides/server-refresh-planning-guide.pdf. Companies with high variance in demand may keep older servers in their racks for use only during demand surges. Amy Nordrum, "How Facebook Keeps Messenger From Crashing on New Year's Eve," *IEEE Spectrum*, December 28, 2018, https://spectrum.ieee.org/tech-talk/computing/software/how-facebooks-software-engineers-prepare-messenger-for-new-years-eve.

[95] The node transition economics calculations in Figure 9 rely on an idealized model involving the manufacture of only a single type of chip—a generic GPU. In reality, many different types of chips are manufactured and may start using newly introduced nodes at varying times after their introduction. Immediately after a node is introduced, new mobile chips (such as a mobile system-on-a-chip) often use that node. By comparison, a node may be one or two years old before desktop- and server-grade chips including GPUs use that node. Given these real-world complications, the time-of-use values in Figure 9 only approximate their real-world counterparts. However, our finding of an increasing expected time-of-use for chips for newer nodes should be robust to changing assumptions on the types of manufactured chips and their differing lags before using newly introduced nodes.

[96] Although these numbers are based on the market value of cloud compute, that likely overstates the cost of DeepMind's access to its sister company Google's TPUs. Jeffrey Shek, "Takeaways from OpenAI Five (2019)," *Towards Data Science*, April 23, 2019, https://towardsdatascience.com/takeaways-from-openai-five-2019-f90a612fe5d.

[97] Dan Huang, "How much did AlphaGo Zero cost?", 2019, https://www.yuzeh.com/data/agz-cost.html.

[98] These computing costs contributed to DeepMind's $572 billion in losses in 2018 and over $1 billion in losses between 2016 and 2018. Gary Marcus, "DeepMind's Losses and the Future of Artificial Intelligence," *Wired*, August 14, 2019, https://www.wired.com/story/deepminds-losses-future-artificial-intelligence/.

[99] OpenAI, Form 990 for fiscal year ending Dec. 2017, 11, https://projects.propublica.org/nonprofits/organizations/810861541/201920719349300822/IRS990.

[100] Amodei et al., "AI and Compute"; Ross Gruetzemacher, "2018 Trends in DeepMind Operating Costs (updated)," August 8, 2019, http://www.rossgritz.com/uncategorized/updated-deepmind-operating-costs/.

[101] A single training run is more compute-intensive than inference of the same AI algorithm. For example, DeepMind's AlphaZero experiment was trained in two steps, the first requiring 5,000 TPUv1s in parallel and the second requiring 64 TPUv2s in parallel. By contrast, inference required only 4 TPUs in parallel. However, inference is performed many times on an AI algorithm while training is performed only once. Therefore, as many as five times as many AI chips may be allocated to inference as to training. Gaurav Batra, Zach Jacobson, Siddarth Madhav, Andrea Queirolo, and Nick Santhanam, "Artificial-intelligence hardware: New opportunities for semiconductor companies" (McKinsey & Company, January 2019), Ex. 5, https://www.mckinsey.com/industries/semiconductors/our-insights/artificial-intelligence-hardware-new-opportunities-for-semiconductor-companies. Therefore, to increase training capacity, companies can reallocate chips from inference to training. Amodei et al., "AI and Compute." However, as leading node AI chips are much more than five times as cost-effective as CPUs or trailing node AI chips, users without access to leading node AI chips will still face prohibitive costs. Moreover, swiftly switching from inference to training is not viable for many AI chips that are suitable for one but not the other. Google's TPUs are an exception, but Google does not sell its high-end TPUs directly; instead, it sells cloud-compute access to its TPUs.

[102] François Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," April 4, 2017, 1, https://arxiv.org/abs/1610.02357.

[103] Chen Huiling, "Will 'Open-Source Traps' Like TensorFlow Strangle China's AI Companies?", *DeepTech official WeChat microblog*, June 10, 2019, https://mp.weixin.qq.com/s?__biz=MzA3NTIyODUzNA==&mid=2649570328&idx=1&sn=cdf2a32921732f18812b4aa5d5b91971&chksm=876a2801b01da117e433d7914e46a6aa83b8e573481053667afee66559e0a5725f6876c0c0d8&scene=21.

[104] Nvidia acquired the networking company Mellanox to develop and use its networking technology to enable more GPU parallelism for AI. Tiffany Trader, "Why Nvidia Bought Mellanox: 'Future Datacenters Will Be...Like High Performance Computers,'" *HPC Wire*, March 14, 2019, https://www.hpcwire.com/2019/03/14/why-nvidia-bought-mellanox-future-datacenters-will-belike-high-performance-computers/.

[105] Tom Simonite, "To Power AI, This Startup Built a Really, *Really* Big Chip," *Wired*, August 19, 2019, https://www.wired.com/story/power-ai-startup-built-really-big-chip. Cerebras is now selling a computing system that includes the Wafer Scale Engine chip. Danny

Crichton, "The Cerebras CS-1 computes deep learning AI problems by being bigger, bigger, and bigger than any other chip," *TechCrunch*, November 19, 2019, https://techcrunch.com/2019/11/19/the-cerebras-cs-1-computes-deep-learning-ai-problems-by-being-bigger-bigger-and-bigger-than-any-other-chip/.

[106] Amodei et al., "AI and Compute" (see "Appendix: Recent novel results that used modest amounts of compute.")

[107] Yaqing Wang, Quanming Yao, James Kwok, and Lionel M. Ni, "Generalizing from a Few Examples: A Survey on Few-Shot Learning," May 13, 2019, https://arxiv.org/abs/1904.05046.

[108] Multilateral export controls on SME are an especially promising tool for the United States and its allies to maintain the exclusive ability to manufacture leading-edge AI chips. Saif M. Khan, "Maintaining the AI Chip Competitive Advantage."; National Security Commission on Artificial Intelligence, *Interim Report* (Washington DC: November 2019), 41–42, https://drive.google.com/file/d/153OrxnuGEjsUvlxWsFYauslwNeCEkvUb/view.

[109] Jingjia has released two GPUs, which have not met expectations. "Jingjia Microelectronics: China's leader in "independent and controllable" microelectronics—military-civil fusion opens up new growth space," *Finance World*, June 13, 2019, http://finance.sina.com.cn/stock/relnews/hk/2019-06-13/doc-ihvhiqay5448234.shtml. The latest model JM7200 is 25x slower compared to a leading U.S. GPU. Jia Xiao, "Jingjia Microelectronics' GPU JM7200 is still far behind global leaders," *PConline*, September 5, 2018, https://diy.pconline.com.cn/1168/11684164.html.

[110] He Huifeng, "'Made in China 2025': the Guangzhou start-up aiming big in semiconductors," *South China Morning Post*, September 24, 2018, https://www.scmp.com/business/companies/article/2165558/made-china-2025-guangzhou-start-aiming-big-semiconductors; Paul Triolo and Jimmy Goodrich, "From Riding a Wave to Full Steam Ahead," *New America*, February 28, 2018, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/riding-wave-full-steam-ahead/.

[111] Although Google prefers calling its TPU a domain-specific architecture (DSA) rather than an ASIC. David Patterson, "Domain-Specific Architectures," 39.

[112] Natasha Mascarenhas, "Intel's Latest Swing At AI Is A $2 Billion Deal From Israel," *Crunchbase News*, December 17, 2019, https://news.crunchbase.com/news/intels-latest-swing-at-ai-is-a-2-billion-deal-from-israel/.

[113] A leading research ASIC is Intel's neuromorphic research chip Loihi, which achieves a 10,000x improvement in efficiency and 1,000x improvement in speed relative to CPUs for specialized applications. Samuel K. Moore, "Intel's Neuromorphic System Hits 8 Million Neurons, 100 Million Coming by 2020," *IEEE Spectrum*, July 15, 2019, https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/intels-neuromorphic-system-hits-8-million-neurons-100-million-coming-by-2020.

[114] For example, the DianNao family of ASICs perform well against benchmarks. See, e.g., Zidong Du, Robert Fasthuber, Tianshi Chen, Paolo Ienne, Ling Li, Tao Luo, Xiaobing Feng, Yunji Chen, and Olivier Temam, "ShiDianNao: Shifting Vision Processing Closer to the Sensor," *ISCA '15: Proceedings of the 42nd Annual International Symposium on Computer Architecture*, 102, June 2015, https://dl.acm.org/doi/10.1145/2749469.2750389; Daofu Liu, Tianshi Chen, Shaoli Liu, Jinhong Zhou, Shengyuan Zhou, Olivier Teman, Xiaobing Feng, Xuehai Zhou, Yunji Chen, "PuDianNao: A Polyvalent Machine Learning Accelerator," *ASPLOS '15: Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems*, March 2015, https://dl.acm.org/citation.cfm?id=2694358; Yunji Chen, Tao Luo, Shaoli Liu, Shijin Zhang, Liqiang He, Jia Wang, Ling Li, Tianshi Chen, Zhiwei Xu, Ninghui Sun, and Olivier Temam, "DaDianNao: A Machine-Learning Supercomputer," *MICRO-47: Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture*, December 2014, https://dl.acm.org/doi/10.1109/MICRO.2014.58.

[115] Eileen Yu, "Huawei unleashes AI chip, touting more compute power than competitors," *ZDNet*, August 23, 2019, https://www.zdnet.com/article/huawei-unleashes-ai-chip-touting-more-compute-power-than-competitors/.

[116] A leading U.S. example of a mobile system-on-a-chip that includes an AI chip is Apple's 7 nm TSMC-fabricated A13 Bionic processor and a leading Chinese example is Huawei's 7 nm TSMC-fabricated Kirin 980 processor. Mark Gurman and Debby Wu, "Apple Partner Starts Building Chips for the Next Generation of iPhones," *Bloomberg*, May 10, 2019, https://www.bloomberg.com/news/articles/2019-05-10/apple-partner-tsmc-starts-building-chips-for-next-gen-iphones; "The Most Powerful and Intelligent Ever: The World's First 7nm Mobile Chipset," Huawei, accessed January 4, 2020, https://consumer.huawei.com/en/campaign/kirin980/. A mobile system-on-a-chip is typically fabricated at the leading node for two reasons. First, it is mass produced which allows fixed costs to be recouped. Second, mobile devices often demand miniaturization.

[117] We thank Lorand Laskai for his insights and data collection on the Chinese AI chip industry.

[118] "AMD Radeon Instinct™ MI50 Accelerator (16GB)," AMD, accessed January 4, 2020, https://www.amd.com/en/products/professional-graphics/instinct-mi50; Joel Hruska, "Report: TSMC 7nm Utilization Improves on Orders From AMD, HiSilicon," *ExtremeTech*, April 4, 2019, https://www.extremetech.com/computing/288917-report-tsmc-7nm-utilization-improves-on-orders-from-amd-hisilicon.

[119] "Nvidia Tesla V100 GPU Architecture," (Nvidia, August 2017), 2, https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf.

[120] Joel Hruska, "Chinese Vendor Designs PCIe 4.0 GPU, Targets GTX 1080 Performance," *ExtremeTech*, August 13, 2019, https://www.extremetech.com/computing/297099-chinese-vendor-designs-pcie-4-0-gpu-targets-gtx-1080-performance. As Jingjia's GPUs are used by the Chinese military, Jingjia may be using China's domestic fab SMIC, which has 28 nm fabs. Jingjia's first-generation GPU, the JM5400, was fabricated at 65 nm. Their upcoming GPUs remain at 28 nm. Mark Tyson, "Jingjia Micro developing GTX1080

perform-alike GPU," *Hexus*, August 26, 2019, https://hexus.net/tech/news/graphics/134084-jingjia-micro-developing-gtx1080-perform-alike-gpu/.

[121] "Intel® Agilex™ FPGAs and SoCs," Intel, accessed January 4, 2020, https://www.intel.com/content/www/us/en/products/programmable/fpga/agilex.html.

[122] "Delivering a Generation Ahead at 20nm and 16nm," Xilinx, accessed January 4, 2020, https://www.xilinx.com/about/generation-ahead-16nm.html.

[123] "Trion® FPGAs," Efinix, accessed January 4, 2020, https://www.efinixinc.com/products-trion.html.

[124] He Huifeng, "'Made in China 2025': the Guangzhou start-up aiming big in semiconductors," *South China Morning Post*, September 24, 2018, https://www.scmp.com/business/companies/article/2165558/made-china-2025-guangzhou-start-aiming-big-semiconductors; "LittleBee ®," Gowin Semiconductor, accessed January 4, 2020, https://www.gowinsemi.com/en/product/detail/2/.

[125] David Manners, "China FPGA strategy takes shape," *Electronics Weekly*, October 30, 2017, https://www.electronicsweekly.com/news/business/china-fpga-strategy-takes-shape-2017-10/.

[126] Dean Takahashi, "Cerebras Systems unveils a record 1.2 trillion transistor chip for AI," *VentureBeat*, August 19, 2019, https://venturebeat.com/2019/08/19/cerebras-systems-unveils-a-record-1-2-trillion-transistor-chip-for-ai/.

[127] Joel Hruska, "Google Announces 8x Faster TPU 3.0 For AI, Machine Learning," *ExtremeTech*, May 9, 2018, https://www.extremetech.com/extreme/269008-google-announces-8x-faster-tpu-3-0-for-ai-machine-learning. Google's TPUv1 used the 28 nm node, but Google has not disclosed the nodes for its more advanced TPU v2/v3. Analysts believe Google used TSMC's 16/12 nm node for the TPU v3. Paul Teich, "Tearing Apart Google's TPU 3.0 AI Coprocessor," *Next Platform*, May 10, 2018, https://www.nextplatform.com/2018/05/10/tearing-apart-googles-tpu-3-0-ai-coprocessor/.

[128] Ian Cutress, "Hot Chips 31 Live Blogs: Habana's Approach to AI Scaling," *AnandTech*, August 19, 2019, https://www.anandtech.com/show/14760/hot-chips-31-live-blogs-habanas-approach-to-ai-scaling; Mascarenhas, "Intel's Latest Swing."

[129] Stephen Shankland, "Meet Tesla's self-driving car computer and its two AI brains," *CNET*, August 20, 2019, https://www.cnet.com/news/meet-tesla-self-driving-car-computer-and-its-two-ai-brains/.

[130] Ian Cutress, "Cambricon, Makers of Huawei's Kirin NPU IP, Build A Big AI Chip and PCIe Card," *AnandTech*, May 26, 2018, https://www.anandtech.com/show/12815/cambricon-makers-of-huaweis-kirin-npu-ip-build-a-big-ai-chip-and-pcie-card.

[131] Timothy Prickett Morgan, "Huawei Jumps into the ARM Server Fray," *Next Platform*, January 8, 2019, https://www.nextplatform.com/2019/01/08/huawei-jumps-into-the-arm-server-chip-fray/.

[132] Linda Trego, "Horizon Robotics releases 2nd generation AI processor," *Autonomous Vehicle Technology*, September 25, 2019, https://www.autonomousvehicletech.com/articles/2014-horizon-robotics-releases-2nd-generation-ai-processor.

[133] "intellifusion has won the second list of Internet Weekly chips," intellifusion, August 30, 2018, http://www.intellif.com/news_description?_l=en&article_id=255. Analysts believe TSMC fabricated the 22 nm chip. Paul Triolo and Graham Webster, "China's Efforts to Build the Semiconductors at AI's Core," *New America*, December 7, 2018, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/chinas-efforts-to-build-the-semiconductors-at-ais-core/.

[134] For more discussion on why AI chips are less likely to use the leading node than CPUs, see Khazree, "Moonwalk."

[135] Anna Stefora Mutschler, "Frenzy At 10/7nm," *Semiconductor Engineering*, September 14, 2017, https://semiengineering.com/the-rush-to-107nm/. A "form factor" refers to the size and shape of a device.

[136] Zen Soo, Sarah Dai, Meng Jing, "Lagging in semiconductors, China sees a chance to overtake the US with AI chips as 5G ushers in new era," *South China Morning Post*, September 18, 2019, https://www.scmp.com/tech/enterprises/article/3027775/lagging-semiconductors-china-sees-chance-overtake-us-ai-chips-5g.

[137] SEMI, *World Fab Forecast*, May 2019 edition. "Logic chips" broadly includes chips that make calculations such as AI chips and CPUs. Only 8.5 percent of global chip fab capacity is configured to fabricate logic chips at ≤16 nm—this value represents the intersection of global chip fab capacity suited to logic chips (40.6 percent) and suited to ≤16 nm chips (26.5 percent).

[138] Arjun Kharpal, "China's biggest chipmaker is still years behind its global rivals," *CNBC*, August 5, 2019, https://www.cnbc.com/2019/08/06/smic-chinas-biggest-chipmaker-is-still-years-behind-its-rivals.html.

[139] International Trade Administration, *2016 Top Markets Report Semiconductors and Semiconductor Manufacturing Equipment* (U.S. Department of Commerce, 2016), 1, https://www.trade.gov/topmarkets/pdf/Semiconductors_Executive_Summary.pdf. A Chinese company called SMEE is purportedly developing 90 nm photolithography tools, although there is little evidence that SMEE's tools are commercially viable at scale or technically reliable. "Current Status of the Integrated Circuit Industry in China," *J. Microelectron. Manuf.* 2, No. 19020105, March 29, 2019, http://www.jommpublish.org/p/26/. China also has little other SME capacity besides photolithography. Two other significant Chinese SME companies are AMEC and Naura.

[140] A good proxy for China's ability to compete in ASIC design is China's dominance in bitcoin mining hardware. The Chinese company Bitmain designs ASICs specialized for bitcoin mining, and has achieved a 75% global market share in the bitcoin mining market. David Floyd, "Bitmain By the Numbers: An Inside Look at a Bitcoin Mining Empire," *CoinDesk*, September 28, 2018, https://www.coindesk.com/bitmain-by-the-numbers-an-inside-look-at-a-bitcoin-mining-empire.

[141] However, many Chinese chip design firms are attempting to reduce their reliance on foreign proprietary IP by instead using the open-source instruction set architecture RISC-V. Shuhei Yamada, "Cutting off Arm: China leads exodus from top chip architect," *Nikkei Asian Review*, October 29, 2019, https://asia.nikkei.com/Business/China-tech/Cutting-off-Arm-China-leads-exodus-from-top-chip-architect.

[142] "AI Policy and China: Realities of State-Led Development" (Stanford-New America DigiChina Project, October 29, 2019), 4, https://newamerica.org/documents/4353/DigiChina-AI-report-20191029.pdf.

[143] Khazree, "Moonwalk," 6 (Figure 3.)

[144] To enhance its semiconductive properties, silicon is typically mixed with added impurities called "dopants." Common dopants include boron, phosphorus, arsenic, and gallium.

[145] Interconnects were historically made of aluminum but now more commonly made of copper or cobalt.

[146] These seven gates are: AND, OR, XOR, NOT, NAND, NOR and XNOR.

[147] "Beyond Borders: The Global Semiconductor Value Chain" (Semiconductor Industry Association, May 2016), 41, https://www.semiconductors.org/wp-content/uploads/2018/06/SIA-Beyond-Borders-Report-FINAL-June-7.pdf.

[148] Inference involves a forward pass through the DNN to classify unlabeled data to yield a label. Training involves additional computations. First, the classification of training data from a forward pass is compared with an existing correct label to determine the degree of classification error. Second, a "gradient descent" computation backward propagates the error through the DNN to update the DNN's parameters to better match and learn from the training data. Michael Andersch, "Inference: The Next Step in GPU-Accelerated Deep Learning," *NVIDIA Developer Blog*, November 11, 2015, https://devblogs.nvidia.com/inference-next-step-gpu-accelerated-deep-learning/,

[149] Other calculations important for DNNs include "vector operations, application of convolutional kernels, and other dense linear algebra calculations." Jeffrey Dean, "The Deep Learning Revolution and Its Implications for Computer Architecture and Chip Design," November 13, 2019, 6, https://arxiv.org/abs/1911.05289. Essentially all of these operations can be implemented as a series of multiply-accumulate operations.

[150] Tiernan Ray, "AI is changing the entire nature of compute," *ZDNet*, June 30, 2019, https://www.zdnet.com/article/ai-is-changing-the-entire-nature-of-compute/; Deloitte, "Hitting the accelerator," 22 (note 12). This massively parallel architecture is an example of

"single instruction multiple data" (SIMD), i.e. identical operations performed on different data, unlike "multiple instruction multiple data" (MIMD), i.e. different operations performed on different data. Hennessy et al., "New Golden Age," 52.

[151] Leading node AI chips, due to extreme transistor densities achieved, may produce too much heat if all execution units are run simultaneously. This is due to increasing power consumption per unit chip area for recent nodes. Ibid, 52. Consequently, some execution units of a chip must be turned off while others operate to prevent overheating. This limits parallelism. The turned-off execution units are called "dark silicon." Nikos Hardavellas, Michael Ferdman, Babak Falsafi, and Anastasia Ailamaki, "Toward Dark Silicon in Servers," *IEEE Computer Society*, July/August 2011, https://infoscience.epfl.ch/record/168285/files/darksilicon_ieeemicro11.pdf.

[152] In one example, CPUs achieve a 212x speed increase for matrix multiplication by using four types of parallelism in combination. Patterson et al., *Computer Organization*, 562. Consistent with this speed increase, in 2018, Intel said they had modified their CPUs (e.g. Xeon) over the previous several years to improve AI algorithm training performance by 200x. Stephen Nellis, "Intel sold $1 billion of artificial intelligence chips in 2017, *Reuters*, August 8, 2018, https://www.reuters.com/article/us-intel-tech/intel-sold-1-billion-of-artificial-intelligence-chips-in-2017-idUSKBN1KT2GK. By comparison, GPUs can perform thousands of parallel computations. Patterson et al., *Computer Organization*, 524-525.

[153] Chris Shallue and George Dahl, "Measuring the Limits of Data Parallel Training for Neural Networks", *Google AI Blog*, March 19, 2019, https://ai.googleblog.com/2019/03/measuring-limits-of-data-parallel.html citing Christopher J. Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E. Dahl, "Measuring the Effects of Data Parallelism on Neural Network Training," November 18, 2018, https://arxiv.org/abs/1811.03600; Sam McCandlish, Jared Kaplan, and Dario Amodei, "How AI Training Scales," *OpenAI*, December 14, 2018, https://openai.com/blog/science-of-ai/ citing Sam McCandlish, Jared Kaplan, and Dario Amodei, "An Empirical Model of Large-Batch Training," December 14, 2018, https://arxiv.org/pdf/1812.06162.pdf; Tal Ben-Nun and Torsten Hoefler, "Demystifying Parallel and Distributed Deep Learning: An In-Depth Concurrency Analysis," September 15, 2018, https://arxiv.org/abs/1802.09941.

[154] Yanping Huang, "Introducing GPipe, an Open Source Library for Efficiently Training Large-scale Neural Network Models," *Google AI Blog*, March 4, 2019, https://ai.googleblog.com/2019/03/introducing-gpipe-open-source-library.html; Ben-Nun, "Demystifying Parallel and Distributed Deep Learning."

[155] Ben-Nun et al., "Demystifying Parallel and Distributed Deep Learning."

[156] AI algorithms are typically trained until they reach "convergence," that is, they reach a low-as-possible inference error rate given the data they were trained on. Larger AI algorithms with more parameters typically require more computing power to train.

[157] McCandlish et al., "How AI Training Scales"; see also Rangan Majumder and Junhua Wang, "ZeRO & DeepSpeed: New system optimizations enable training models with over 100 billion parameters," *Microsoft Research Blog*, February 13, 2020,

https://www.microsoft.com/en-us/research/blog/zero-deepspeed-new-system-optimizations-enable-training-models-with-over-100-billion-parameters/.

[158] For inference using FPGAs and ASICs, see Hwang, "Computational Power," 12–13. For training using the TPU v2, see Jeff Dean, Google Brain, "Machine Learning for Systems and Systems for Machine Learning," NIPS 2017, 7, 12, 27, http://learningsys.org/nips17/assets/slides/dean-nips17.pdf.

[159] Pete Singer, "AI Chips: Challenges and Opportunities," *Semiconductor Manufacturing & Design*, September 12, 2018, https://web.archive.org/web/20181231034201/http://semimd.com/blog/2018/09/12/ai-chips-challenges-and-opportunities/; Jouppi et al., "Tensor Processing Unit," 1; Andres Rodriguez, Eden Segal, Etay Meiri, Evarist Fomenko, Young Jim Kim, Haihao Shen, and Barukh Ziv, "Lower Numerical Precision Deep Learning Inference and Training" (Intel, January 2018), https://www.intel.ai/nervana/wp-content/uploads/sites/53/2018/05/Lower-Numerical-Precision-Deep-Learning-Inference-Training.pdf.

[160] See, e.g. Robert LiKamWa, Yunhui Hou, Julian Gao, Mia Polansky, Lin Zhong, "RedEye: Analog ConvNet Image Sensor Architecture for Continuous Mobile Vision," *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture*, 2016, https://www.recg.org/publications/likamwa2016isca.pdf; Ali Shafiee, Anirban Nag, Naveen Muralimanohar, Rajeev Balasubramonian, John Paul Strachan, Miao Hu, R. Stanley Williams, Vivek Srikumar, "ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars," October 5, 2016, https://www.cs.utah.edu/~rajeev/pubs/isca16.pdf.

[161] Manas Sahni, "Making Neural Nets Work With Low Precision," December 7, 2018, https://sahnimanas.github.io/post/quantization-in-tflite/.

[162] Jouppi et al., "Tensor Processing Unit," 1.

[163] Hennessy et al., "New Golden Age," 56-57.

[164] Ray, "AI is changing the entire nature of compute," In one example, AI chips can also include specialized memory optimized for AI that improves bandwidth by 4.5x to allow storage of large amounts of data needed for AI, albeit at a 3x cost. Batra et al., "Artificial-intelligence hardware." For an overview of the optimization of different compute elements for AI, see Ibid, Exhibits 1–2.

[165] "MegatronLM: Training Billion+ Parameter Language Models Using GPU Model Parallelism," *Nvidia Applied Deep Learning Research*, August 13, 2019, https://nv-adlr.github.io/MegatronLM; Yu Emma Wang, Gu-Yeon Wei, and David Brooks, "Benchmarking TPU, GPU, and CPU Platforms for Deep Learning," October 22, 2019, 9, https://arxiv.org/abs/1907.10701.

[166] Joel Hruska, "Amazon adds Nvidia GPU firepower to its compute cloud," *ExtremeTech*, October 3, 2016, https://www.extremetech.com/computing/236676-amazon-adds-nvidia-gpu-firepower-to-its-compute-cloud.

[167] Specifically, TensorFlow's code runs efficiently on Google's tensor processing units (TPUs). Hennessy et al., "New Golden Age," 57.

[168] Matrix multiplication operations run 47x faster when programmed in C versus Python. Ibid, 56.

[169] Ibid, 57.

[170] David A. Patterson and John L. Hennessy, *Computer Organization and Design: The Hardware/Software Interface* (Burlington, MA: Morgan Kaufmann 5th Edition, 2013), 525.

[171] We exclude results for any mobile system-on-a-chip, which often includes a CPU in combination with an AI chip, because they are less powerful than server and PC chips and therefore less relevant for high-end AI applications. Additionally, other researchers have comprehensively benchmarked these chips. Andrey Ignatov, Radu Timofte, William Chou, Ke Wang, Max Wu, Tim Hartley, and Luc Van Gool, "AI Benchmark: Running Deep Neural Networks on Android Smartphones," October 15, 2018, 11 (Table 2), https://arxiv.org/abs/1810.01109.

[172] Patterson et al., *Computer Organization*, 562; Nellis, "Intel sold $1 billion of artificial intelligence chips in 2017.

[173] The GPU vs. CPU results reach 100x as DNN models get large and therefore benefit from parallelism, while the TPU vs. GPU results range between 1-10x depending on the types of DNNs and the model size. Yu Emma Wang, Gu-Yeon Wei, and David Brooks, "Benchmarking TPU, GPU, and CPU Platforms for Deep Learning," October 22, 2019, 8–10 (Figures 8–10).

[174] "MLPerf Training v0.6 Results," MLPerf, July 10, 2019, https://mlperf.org/training-results-0-6. The values in the table represent comparisons for the same number of TPUs and GPUs.

[175] Graphcore say which GPU was used—the Nvidia Tesla V100—only for its inference test on a transformer model. Dave Lacey, "New Graphcore IPU Benchmarks," *Graphcore*, November 18, 2019, https://www.graphcore.ai/posts/new-graphcore-ipu-benchmarks.

[176] "Improved" means a version of the first generation TPU with improved memory. The values represent a weighted average across types of DNNs, where the weights represent industry-wide usage of each type. Jouppi et al., "Tensor Processing Unit," 8–9.

[177] Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia, "DAWNBench: An End-to-End Deep Learning Benchmark and Competition," *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017, 5 (Figure 5), https://cs.stanford.edu/~matei/papers/2017/nips_sysml_dawnbench.pdf.

[178] The presented results compare optimal threading for the server class Intel Xeon CPUs, and includes tests only where the Nvidia GTX 1080 outperformed other GPUs, which includes all tests but one. Shaohuai Shi, Qiang Wang, Pengfei Xu, and Xiaowen Chu, "Benchmarking

State-of-the-Art Deep Learning Software Tools," February 17, 2017, 6 (Table 7), https://arxiv.org/abs/1608.07249.

[179] Robert Adolf, Saketh Rama, Brandon Reagen, Gu-Yeon Wei, and David Brooks, "Fathom: Reference Workloads for Modern Deep Learning Methods," August 23, 2016, 7 (Figure 5), https://arxiv.org/abs/1608.06581. Google claims its inference results differ from Harvard's because Harvard's "CPU and GPU are not server-class, the CPU has only four cores, the applications do not use the CPU's AVX instructions, and there is no response-time cutoff." Jouppi et al., "Tensor Processing Unit," 14.

[180] For the training benchmark, only gradient computation is done, not the entire training process including updating weights. The ranges represent the best-performing frameworks for both the GPU and CPU, and best-performing multithreading for the CPU. Soheil Bahrampour, Naveen Ramakrishnan, Lukas Schott, and Mohak Shah, "Comparative Study of Caffe, Neon, Theano, and Torch for Deep Learning," *ICLR 2016*, 2016, 5 (Table 3), 6 (Table 4), 8 (Table 6), https://openreview.net/pdf?id=q7kEN7WoXU8LEkD3t7BQ.

[181] Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A. Horowitz, and William J. Dally, "EIE: Efficient Inference Engine on Compressed Deep Neural Network," May 3, 2016, 6 (Figures 6–7), https://arxiv.org/abs/1602.01528.

[182] LiKamWa et al., "RedEye," 263.

[183] The efficiency results for maximum batching, and the speed results compare the fastest respective frameworks for the GPU and CPU. Da Li, Xinbo Chen, Michela Becchi, Ziliang Zong, "Evaluating the Energy Efficiency of Deep Convolutional Neural Networks on CPUs and GPUs," *2016 IEEE International Conferences on Big Data and Cloud Computing, Social Computing and Networking, Sustainable Computing and Communications*, October 31, 2016, 3 (Figure 2), 5 (Figure 4), 6 (Figure 6), https://ieeexplore.ieee.org/document/7723730.

[184] Shaoli Liu, Zidong Du, Jinhua Tao, Dong Han, Tao Luo, Yuan Xie, Yunji Chen, and Tianshi Chen, "Cambricon: an instruction set architecture for neural networks," *ISCA '16: Proceedings of the 43rd International Symposium on Computer Architecture*, June 2016, 400, 401, 403 (Figures 12–13) https://dl.acm.org/citation.cfm?id=3001179.

[185] Some results are based on a combined use of a DNN and Gaussian mixture model. Johann Hauswald, Michael A. Laurenzano, Yunqi Zhang, Cheng Li, Austin Rovinski, Arjun Khurana, Ronald G. Dreslinski, Trevor Mudge, Vinicius Petrucci, Lingjia Tang, and Jason Mars, "Sirius: An Open End-to-End Voice and Vision Personal Assistant and Its Implications for Future Warehouse Scale Computers," *ASPLOS '15: Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems*, March 2015, 231 (Table 5), 232 (Figure 15), https://dl.acm.org/doi/10.1145/2694344.2694347.

[186] We report results where the calculations utilize batching, not results without batching. Johann Hauswald, Yiping Kang, Michael A. Laurenzano, Quan Chen, Cheng Li, Trevor Mudge, Ronald G. Dreslinski, Jason Mars, and Lingjia Tang, "DjiNN and Tonic: DNN as a Service and Its Implications for Future Warehouse Scale Computers," *ISCA '15*, June 2015,

5 (Table 1), 8 (Figure 10), http://web.eecs.umich.edu/~jahausw/publications/hauswald15djinn.pdf.

[187] Chen Zhang, Peng Li, Guangyu Sun, Yijin Guan, Bingjun Xiao, and Jason Cong, "Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks," *FPGA '15: Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, February 2015, 9 (Tables 7–8). https://dl.acm.org/doi/10.1145/2684746.2689060.

[188] Kalin Ovtcharov, Olatunji Ruwase, Joo-Young Kim, Jeremy Fowers, Karin Strauss, and Eric S. Chung, "Toward accelerating deep learning at scale using specialized hardware in the datacenter," *2015 IEEE Hot Chips 27 Symposium*, July 7, 2016, 35, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7477459.

[189] Lukas Cavigelli, Michele Magno, and Luca Benini, "Accelerating real-time embedded scene labeling with convolutional networks," *DAC '15: Proceedings of the 52nd Annual Design Automation Conference*, June 2015, 4 (Table 2), https://dl.acm.org/citation.cfm?id=2744788.

[190] Clement Farabet, Berin Martini, Benoit Corda, Polina Akselrod, Eugenio Culurciello, and Yann LeCun, "NeuFlow: A Runtime Reconfigurable Dataflow Processor for Vision," *CVPR 2011 Workshops*, August 12, 2011, 115 (Table 5), https://ieeexplore.ieee.org/document/5981829.

[191] Jonathan Raiman, Przemysław Dębiak, Brooke Chan, Jie Tang, Michael Petrov, Christy Dennison, David Farhi, Susan Zhang, Filip Wolski, Szymon Sidor, Jakub Pachocki, Henrique Pondé, Greg Brockman, "OpenAI Five," *OpenAI*, June 25, 2018, https://openai.com/blog/openai-five/.

[192] "NVIDIA Tesla P100 PCIe 16 GB," TechPowerUp, accessed January 4, 2020, https://www.techpowerup.com/gpu-specs/tesla-p100-pcie-16-gb.c2888.

[193] We estimated chips per wafer by assuming that chips are square-shaped, separated from other chips by scribe lines with widths of 100 microns, and do not extend less than 3 millimeters from the wafer's edge. We performed the calculations based on the calculator provided in "Die Per Wafer Formula and (free) Calculator," *anysilicon*, April 9, 2013, https://anysilicon.com/die-per-wafer-formula-free-calculators/. We then multiplied the result by 85 percent to account for an 85 percent yield. "Yield" refers to the percentage of chips successfully fabricated without errors. Josef Biba, "Yield Calculation," Advanced MOSFETs and Novel Devices, University of Munich, Munich, Germany, 6, https://dokumente.unibw.de/pub/bscw.cgi/d10465215/%C3%9Cbung-1.pdf.

[194] We calculate the 5 nm GPU's transistor density as: (Nvidia Tesla P100 16 nm GPU transistor density x TSMC 5 nm node transistor density) / TSMC 16/12 nm node transistor density = (25 MTr/mm^2 x 171.3 MTr/mm^2) / 28.9 MTr/mm^2 = 148.2 MTr/mm^2.

[195] For 5, 7, 10, and 16/12 nm densities, see David Schor, "TSMC Starts 5-Nanometer Risk Production," *WikiChip Fuse*, April 6, 2019, https://fuse.wikichip.org/news/2207/tsmc-starts-5-nanometer-risk-production/. For 20 nm density, we interpolate between 16/12 nm

and 28 nm. For 28 nm density, we use data on Intel's 22 nm node, which Intel claims has similar transistor density as competitor 28 nm nodes. See Mistry, "10 nm Technology Leadership," 8, 25. We infer 40, 65, and 90 nm densities based on TSMC's claimed node-to-node improvements. "TSMC's 28nm To Be a Full Node Process," *Design & Reuse*, September 29, 2008, https://www.design-reuse.com/news/19173/tsmc-28nm.html; "40nm Technology," TSMC, accessed January 4, 2020, https://www.tsmc.com/english/dedicatedFoundry/technology/40nm.htm; "TSMC Unveils Nexsys 65nm Process Technology Plans," *Phys.org*, May 3, 2005, https://phys.org/news/2005-05-tsmc-unveils-nexsys-65nm-technology.html.

[196] The IBS data do not appear to adjust for inflation. For our calculations, we convert all values to equivalent 2020 USD values.

[197] In 2019 Q3, a total of 89 million GPUs were sold. This includes discrete (i.e. standalone) GPUs plus GPUs included on a system-on-a-chip. Robert Dow, "Global GPU shipments up in Q2'19 reports Jon Peddie Research," *Jon Peddie Research*, August 27, 2019, https://www.jonpeddie.com/press-releases/global-gpu-shipments-up-in-q219-reports-jon-peddie-research/. In all of 2017, 92.8 million discrete GPUs were sold. Statista, Global shipments of discrete graphics processing units from 2015 to 2018, August 2018, https://www.statista.com/statistics/865846/worldwide-discrete-gpus-shipment/.

[198] For example, the 7 nm node requires chips with 1.8x the chip area as 5 nm node chips. This means the 7 nm design cost per chip is determined by 5,000,000 but then multiplying by 1.8.

[199] Our model imagines a 5 nm-equivalent transistor count for each node. Naively, this may suggest that design costs should be similar across nodes. However, we assume that chips at nodes larger than 5 nm use designs appropriate to their nodes, and if necessary, use duplicate logic blocks to achieve the 5 nm-equivalent transistor count.

[200] TechPowerUp, "NVIDIA Tesla P100 PCIe 16 GB."

[201] We obtain the 31% value by averaging power consumption at idle as a percentage of TDP for seven high end GPUs. Marco Chiapetta, "NVIDIA TITAN V Review: Volta Compute, Mining, And Gaming Performance Explored," *Hot Hardware*, December 15, 2017, 6, https://hothardware.com/reviews/nvidia-titan-v-volta-gv100-gpu-review?page=6.

[202] Amodei et al., "AI and Compute."

[203] We use the formula: average power consumption = power consumption at idle + (utilization * (1 - power consumption at idle)). We neglect several complicating factors. First, GPUs exhibit a non-linear relationship between utilization and power consumption, for example Nvidia Volta V100's "most efficient range on the curve might be 50-60% of TDP, where the GPU still attains 75-85% of maximum performance." Nvidia, "Nvidia Tesla V100 GPU Architecture," 11. Second, Nvidia does not publish figures on power consumption at idle for its GPUs, and different AI experiments will have different utilization rates and therefore different ratios. Third, actual power consumption can exceed TDP due to cooling requirements and inefficiency in converting AC to low-voltage regulated DC power.

[204] We assume that the GPU is hosted in the cloud and run continuously. One way that cloud providers run chips continuously is by offering preemptable services in which chips experiencing down-time for one customer are rented to other customers. See, e.g. "Preemptible Virtual Machines," Google Cloud, accessed January 4, 2020, https://cloud.google.com/preemptible-vms/.

[205] In October 2019, U.S. industrial users paid an average of $0.0685 per kilowatt-hour with a regional range between $0.0550 and $0.2306 per kilowatt-hour. "Electric Power Monthly," U.S. Energy Information Administration, December 23, 2019, https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=epmt_5_6_a. Chinese industrial users pay an average of $0.084 per kilowatt-hour. Zoey Ye Zhang, "China Electricity Prices for Industrial Consumers," *China Briefing*, April 23, 2019, https://www.china-briefing.com/news/china-electricity-prices-industrial-consumers/. We use the average of these values: $0.07625 per kilowatt-hour.

[206] "Efficiency," Google Data Centers, accessed January 4, 2020, https://www.google.com/about/datacenters/efficiency/.

[207] For some leading node chip designs, energy costs of communicating data are larger than energy costs of operating transistors and do not show significant node-to-node improvements. Our model does not take into account this complication. However, ongoing research and development is focusing on new methods reduce the energy costs of data communication at leading nodes.

[208] These values approximately match the 2018 values. Therefore, the 2004 to 2018 averages are still valid today.

[209] Data obtained from TSMC, "Financial Results." The capital depreciation rate is calculated by dividing TSMC's reported depreciated assets for a given year by its reported net capital assets for that year.

[210] TSMC claims its GigaFabs offer the lowest operating costs and highest flexibility of all fab sizes. "GIGAFAB® Facilities," TSMC, accessed January 4, 2020, https://www.tsmc.com/english/dedicatedFoundry/manufacturing/gigafab.htm. Larger foundries achieve the greatest economies of scale, but not all firms have sufficient demand to support the largest foundries.

[211] Fabs are frequently expanded and updated over time. However, when estimating capital investment per wafer processed per year, we use original cost and fabrication capacity figures.

[212] Capital investment and annual wafer capacity are from SEMI, *World Fab Forecast*, May 2019 edition.

[213] The model equates year with the most advanced node in production in that given year, as fabs are largely constructed at the leading node. For example, we assume that a fab constructed in 2011 will predominantly be equipped for the 28 nm node.

[214] In some depreciation schedules, SME fully depreciates in about four years, consistent with our model.

[215] For nodes at 16/12 nm and larger, capital is already fully depreciated at 65%, at which point we define "consumption" as follows. We assume that our use of the 25.29% depreciation rate to calculate capital "consumption" approximately accounts for the yearly likelihood that old capital breaks down or is retired.

[216] Other costs include materials, labor, and R&D. Materials and labor costs are tricky to allocate. Presumably, these costs are more closely correlated with the number of wafers processed than the node of the wafers processed. On the other hand, new nodes typically require more intensive engineering efforts to bring yields up to production levels while more advanced nodes can require more exotic materials. In practice, these costs likely increase sharply as a new node comes online, then eventually decrease to be more closely correlated to quantity of wafers processed than the costs of capital assets used for processing. R&D is likewise tricky to allocate. R&D is primarily motivated by the desire to stay at the leading node, for which TSMC has a near monopoly for foundry services. Samsung is the only other fab operating at 5 nm, but its business operates mostly under an integrated device manufacturer (IDM) model, therefore its foundry services capacity is small compared to TSMC's. Additionally, TSMC could continue fabrication of chips at older nodes even if R&D ceased. Finally, markup may also vary between nodes. TSMC may place a greater markup on leading node chips on which they have a near monopoly, although declining benefits from leading nodes means that customers could choose older nodes if TSMC introduces too high of a markup.

[217] Our calculated foundry sale prices per wafer for each node are close to 2018 estimates of foundry revenue per logic wafer by IC Insights: $1,800 for 90 nm, $2,110 for 65 nm, $2,655 for 45/40 nm, $3,010 for 28 nm, and $6,050 for ≤20 nm. "Advanced Technology Key to Strong Foundry Revenue per Wafer," *IC Insights*, October 12, 2018, http://www.icinsights.com/news/bulletins/advanced-technology-key-to-strong-foundry-revenue-per-wafer/.

[218] The Semiconductor Industry Association (SIA) reports an annual industry cost of $0.57 per chip. Semiconductor Industry Association, "2019 SIA Databook," 2019, iv, https://www.semiconductors.org/wp-content/uploads/2019/07/2019-SIA-Databook-For-Purchase-TOC.pdf. This calculation includes sales of all chips, discretes, sensors, and optoelectronics, many of which are low cost. However, GPUs are among the most expensive chips in production, which is why our per-chip GPU costs differ from SIA's industry-wide calculations.

[219] The modeled 90 nm chip technically cannot be manufactured on a 300 mm wafer, as Table 7 says only 0.7 90 nm node chips fit on a 300 mm wafer. But in reality, no companies are attempting to manufacture wafer scale 90 nm node chips. Instead, the model should be interpreted as quantifying the cost, at each node, of an equivalent number of transistors as a 5 nm chip with an area of 610 mm^2. This means the model could equivalently be interpreted as accounting for multiple 90 nm node chips totaling the transistor count of one 5 nm chip.

[220] Up to the 7 nm node, the quarter is based on when TSMC first reported at least 1% of its revenue from that node. TSMC, "Financial Results." TSMC is planning mass production of 5

nm node chips in the first half of 2020. Ian Cuttress, "Early TSMC 5nm Test Chip Yields 80%, HVM Coming in H1 2020," *AnandTech*, December 11, 2019, https://www.anandtech.com/show/15219/early-tsmc-5nm-test-chip-yields-80-hvm-coming-in-h1-2020.

[221] OECD, *Measuring distortions in international markets: The semiconductor value chain* (Paris, France: OECD Trade Policy Papers, No. 234, OECD Publishing, December 12, 2019), 22, http://dx.doi.org/10.1787/8fe4491d-en.

[222] "The Strength of the OSAT Companies," *New Venture Research*, December 7, 2013, https://newventureresearch.com/the-strength-of-the-osat-companies/.

[223] Ibid, 21.

[224] "McLean Report," IC Insights, accessed February 25, 2020, http://www.icinsights.com/services/mcclean-report/report-contents/.

[225] We lack access to data on assembly, test, and packaging costs by node. Therefore, we use the same industry-wide percentage for each node.