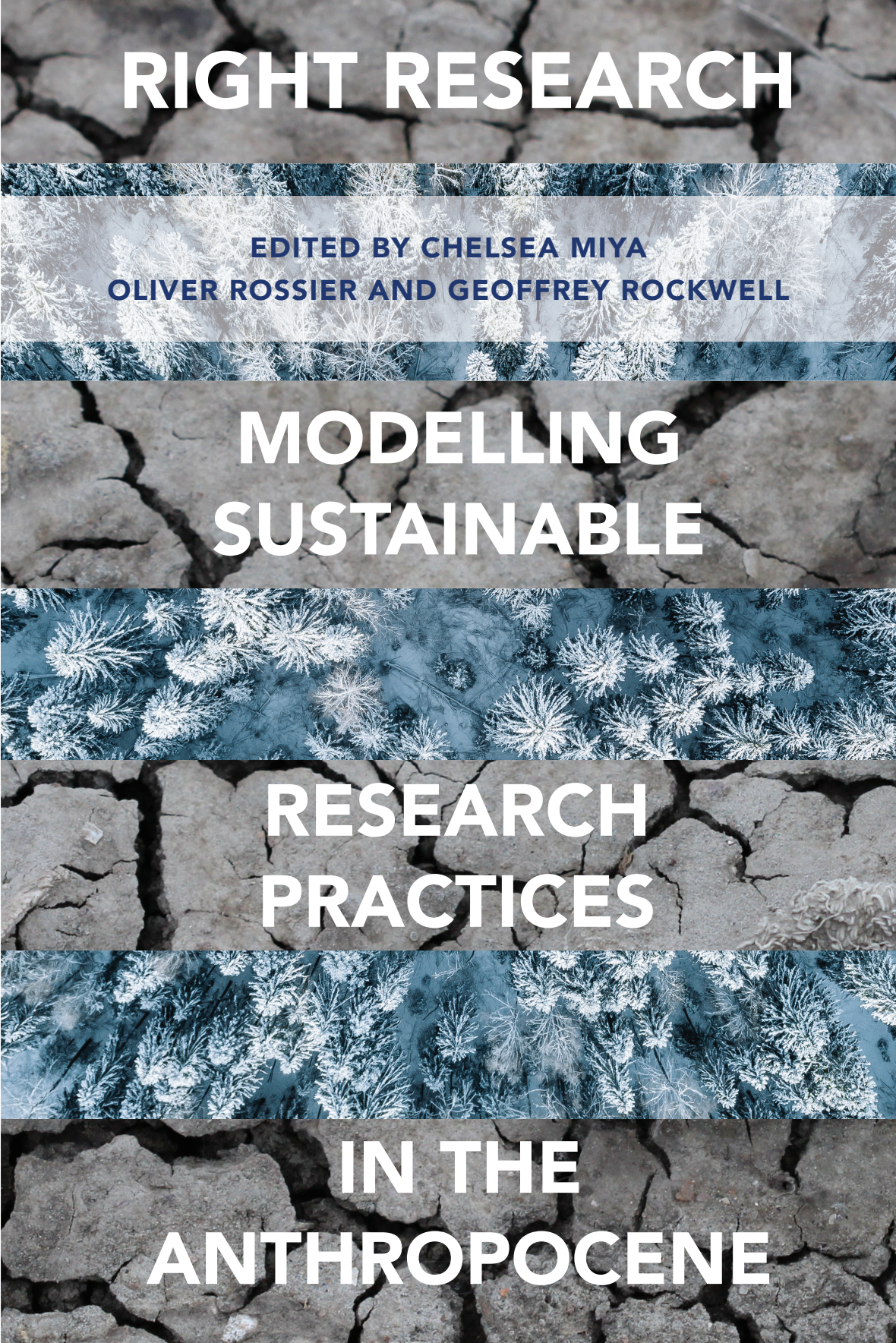
RIGHT RESEARCH
EDITED BY CHELSEA MIYA
OLIVER ROSSIER AND GEOFFREY ROCKWELL
MODELLING
SUSTAINABLE
RESEARCH
PRACTICES
IN THE
ANTHROPOCENE

https://www.openbookpublishers.com

Chelsea Miya, Oliver Rossier and Geoffrey Rockwell (eds), *Right Research: Modelling Sustainable Research Practices in the Anthropocene*. Cambridge, UK: Open Book Publishers, 2021. https://doi.org/10.11647/OBP.0213

ISBN Paperback: 9781783749614
ISBN Hardback: 9781783749621
ISBN Digital (PDF): 9781783749638
ISBN Digital ebook (epub): 9781783749645
ISBN Digital ebook (mobi): 9781783749652
ISBN Digital (XML): 9781783749669
DOI: 10.11647/OBP.0213

Cover image by Leanne Olson, The Clay at Ryley, CC-BY-NC-ND.
Cover design by Emilie St-Hilaire.

5. Impact of the Digital Revolution on Worldwide Energy Consumption

Doug Barlage and Gem Shoute

We Tweet, Facebook, Netflix and YouTube in the palm of our hand. We are aware of the amount of energy that it takes from how many times that we need to recharge our devices. However, this is just the tip of the iceberg. For every joule of energy we expend locally, many more joules are spent in the backbone of the Internet. While our appetite for data has largely been insatiable over the last thirty years, the energy required to sustain this has been held in check by Moore's Law's driving creed that density of function in a computer chip increases by two every two years, and energy/function decreases by a similar amount. With that said, this driving relationship between power consumption and computing density is slowing due to a multitude of physical constraints when the density of transistor packing approaches the limits. In the following chapter, the authors examine these relationships and outline some of the challenges that the world is facing as we continue to meet and exceed the expectations of our data-driven world with a finite growth in worldwide power generation capacity.

Introduction

Increases in computational demand has led to rising demands on the power grid for energy efficiency. When we use our cell phone, our

 https://doi.org/10.11647/OBP.0213.05

computer or one of the many things that interact with the larger cloud of data, we forget that we are exercising a certain amount of energy to do so. Unlike a light bulb where the energy spent is all at the point of use, when we watch a viral cat video on our phone, we spend about 10–100 times more energy supporting the data path to get it to you. That energy is spent away from you, but nonetheless it is required so that you can download information and fully enjoy the fruits of the information age. For that matter, the conference on sustainability that inspired the initial presentation documented in this paper depends on the World Wide Web. Streaming information across the planet forces energy to be expended. Whether it is watching Netflix or researching remote documents to write a paper, the computations required are not an amount that can be brushed aside as trivial. While this energy to communicate across the Web is considerably smaller than the amount of energy that it would take to bring everyone physically to the same room at the same time, it is not zero and in our growing, increasingly interconnected world, it can no longer be considered an afterthought of the total energy that we expend in our daily lives. Computational energy has become something that must be considered when planning on future energy demands. Our increasing demand for instantaneous information consumes energy that is definitely not zero and is actually the fastest growing area of energy consumption around the world. In this chapter, we introduce the reader to the basic principles of energy consumption by digital computation and the limitations in reduction of this energy using the current technology available. The realities of Moore's Law reaching a plateau are discussed. We follow this with a calculation of the demand that is occurring on a rapid level as the information revolution continues to unfold before us, and we identify the limiting economic principles that drive this energy consumption phenomena. Lastly, we propose some approaches to finding a solution to this dilemma and look at the role that industry and national governments are playing and should play as this increasingly demanding economic sector continues to require an ever-growing quantity of energy.

Energy Consumption of Computation

Computation machines preceded the discovery of the transistor by John Bardeen and Walter Brattain by several thousand years.[1] The abacus was one of the first computational machines in existence. The abacus served the ancients through the accountants and traders that used it. The energy consumption of these abacuses should be thought of as two distinct entities: the actual energy used to drive the abacus and the energy used to sustain the driver of the abacus. This could be considered the energy consumed by the system. In the days of the ancients, this would be considered the amount of food required to feed the accountants who drove the abacus. We refer to this as E_{System}. Today this energy is the amount of energy that our computational machines require. Furthermore, it is possible to directly analyze the amount of energy required to do the computation. The energy required to do a computation could be determined by the amount of energy required to move the beads. This energy is of the non-recoverable variety and could be easily calculated if the amount of friction was known. In a very general way, the energy could be calculated as follows:

$$E_{Computation} = \int_{X_{state1}}^{X_{state2}} F(x)dx$$

Where the force, $F(x)$, is the force required to move the beads and X_{State1} is the starting position and X_{State1} is the ending position. A first-year physics student is taught to make this calculation. To try and make this a little more tangible, we can look back to the Babbage machines of the 1800s. Modern computational apparatus and architectures are at least an inspirational descendent of these devices. It is possible to see what's going on the inside with your eyes rather than relying on electron microscopes and what really amounts to applied imagination when you look at the working mechanisms of a modern system. Every piston that moved to perform a calculation expended an amount of energy that is easily observed. This energy is of the kind that follows thermodynamic principles of non-recovery and introduces some

1 John Bardeen and Walter Hauser Brattain, 'The transistor, a semi-conductor triode', *Physical Review*, 74.2 (1948), 230–231, https://doi.org/10.1103/physrev.74.230.

entropy. These systems had hundreds of tiny pistons that executed algorithms at several calculations per second. To perform a binary calculation many pistons would have to change positions many times to perform one operation. It might be said that our mechanical computation forerunners had a superior way to do calculations. It was able to accomplish more output with less internal operations. In modern terms, this is referred to as a floating point operation or FLOP. Our modern systems require that many individual beads must change position to perform one calculation.

In principal, the driving notion behind reducing energy in computation from an engineering standpoint remains essentially the same today. But instead of beads we move electrons from one energetic state to another. In a very tangible, though not computationally sufficient way, reductions in energy on the per bit level comes from moving states closer together with lower amounts of energy between those states. Using these ideas, we can calculate the efficiency of the system and put this directly in the context of power and bits per second. While the modern computational system is a little more complicated than an abacus, the principle is the same. A user tells the computational system what it wants and the computational system does what it takes to provide the user what it desires. Every time the computational system moves a bead on the abacus, the computational system takes energy. A modern system such as your cell phone has well over a billion such beads that calculate at well over a billion calculations per second. When the calculation system does its work, it takes energy. Some of the energy is used to move the beads on the abacus, and other energy is required just to maintain the computer. You have to supply both to make the whole system work and keep it healthy and so this has to be included when you consider energy consumption. Just as if we had a million accountants from ancient times to do our bidding, we have to feed those accountants to keep them well and we have to feed them the energy it takes to do our bidding. Likewise, we feed our computers to allow them to do work.

It was not until after World War II that vacuum tubes powered the first electronic computers. The first system, ENIAC, was constructed at the University of Pennsylvania, consumed about 150kW and fit in a fair-sized room. For reference, 150 kW is pretty close to the power that

a modern electric vehicle expends when it is accelerating to cruising speed. This is something that is completely manageable by today's standards, but in 1946, when the only real loads on the grid were light bulbs, heaters and refrigerators, this was a significant amount of energy. The first computer on the campus at the University of Pennsylvania took about as much energy as a 100-person dormitory to operate. While this is not a small amount of energy, when you consider a campus of about 25000 students, the energy consumption due to this room-sized computer is barely noticeable. For an individual operation it took around 100 J on this system. It is hard to make a comparison to what this energy actually means. There are 11 calories in a peanut, 4,184 J in a calorie, therefore a perfectly converted peanut gives you about 460 operations. For every joule of energy spent, ENIAC took 5 J of energy just to stay on. So it might be said that ENIAC may have been ever so slightly more efficient than the people operating it. But ENIAC was faster than the people operating it. It could execute around 300 multiplication operations per second and so it was very useful.

Solid state electronics came into full force with the introduction of a practical integrated circuit process by Robert Noyce in 1958.[2] This led to a realization that was documented by Gordon Moore in 1965. Moore's Law famously states that the complexity of integrated circuits increases by a factor of two every two years. The semiconduction industry was advanced by the invention of the Metal Oxide Semiconductor (MOS) Transistor in 1963[3] and the road to perfection of the initial process for metal oxide transistors was initiated by Andrew Grove in 1964. Since then, the creation of a well-established set of business and engineering rules has continued to push the industry forward.[4] In 1970, Intel, which was founded by Noyce, Moore and Grove, reported a revenue of just over $4.2 million versus expenses of around $5.6 million.[5] The minimum printed transistor dimension was over ten microns (ten one millionths

2 Robert N. Noyce, 'Semiconductor device-and-lead structure', U.S. Patent No. 2,981,877 (April 25, 1961).

3 Kahng Dawon, 'Electric field controlled semiconductor device', U.S. Patent No. 3,102,230 (August 27, 1963).

4 Tim Jackson, *Inside Intel: Andrew Grove and the Rise of the World's Most Powerful Chip Company* (New York: Viking Penguin, 1997).

5 Intel, *1970 Intel Financial Statement* (Arthur Young & Company, 1971), https://www.intel.com/content/dam/www/public/us/en/documents/corporate-information/history-1970-financial-statement.pdf.

of a meter). In 2018, Intel's revenue was approximately $70 billion with a minimum printed dimension approaching ten nanometers (ten one billionths of a meter).[6] In 1974, another set of rules was established for the governing and scaling of the MOS transistor by Robert Dennard.[7] When the complementary process was introduced in production around 1980, the minimum printed dimension of the transistor was reduced by a factor of 40% every two years (area reduction by a factor of 2). This became a well-documented and executable realization of Moore's Law. These governing economic and engineering principals, which allowed for continued functionality enhancements from device scaling held until at least 2015. What was also true during this time was that the energy per computation reduced at the same rate as minimum dimension. This has a direct analogy to the abacus example that was discussed previously. Smaller devices consume less energy. This minimum amount of energy per 'bead' movement can be found simply by recognizing that there are about thirty electrons moving through one volt of electric potential in the modern device. The minimum energy for this single bit change is then around $5*10^{-18}$ joules, or 5 aJ. When the bit changes it must remove thirty electrons in addition to adding thirty holes (electron voids which behave as particles in semiconductors) or vice versa. The total energy under this scenario is 10 aJ. This is a good estimation of the capabilities of CMOS technology for a single bit. The present complementary MOS (CMOS) technology has not improved upon this minimum energy significantly since about 2011. In other words, the leading-edge energy efficiency has not really improved considerably during that time. In truth, by lowering the voltage swing and by reducing the total number of carriers per bit, there is still room to improve and billions of dollars are spent every year by multiple governments and private companies to improve that figure. However, these efforts have not yielded a significant amount of difference in the minimum amount of energy required to create and destroy a single bit that can meet the stringent requirements necessary to begin plans for

6 NASDAQ, *INTC Company Financials* (2019), https://www.nasdaq.com/market-activity/stocks/intc/financials

7 Robert H. Dennard et al., 'Design of ion-implanted MOSFET's with very small physical dimensions', *IEEE Journal of Solid-State Circuits*, 9.5 (1974), 256–268. DOI: 10.1109/JSSC.1974.1050511.

production. When it is discussed that Moore's Law may be coming to an end, it is this fact that dominates the conversation.[8]

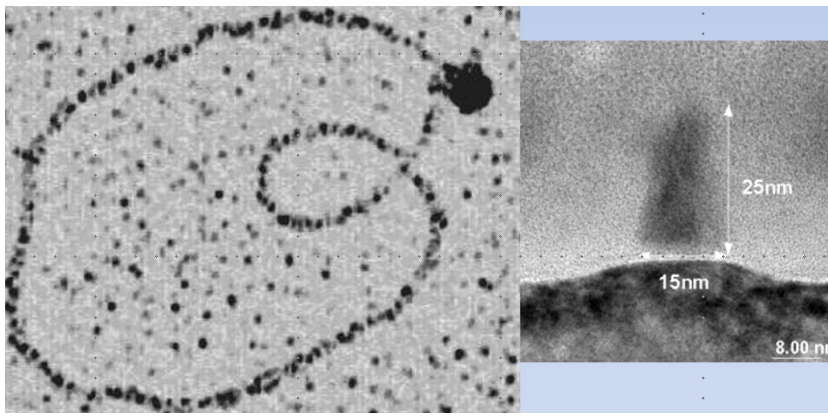

Fig. 1 From the author's (Barlage) notebook, c. 2000. A stock photo of a strand of DNA, adjusted to be at the same scale as a prototypical 15nm gate length MOS transistor produced by his group at Intel in 2000. This prototype transistor from dimensions are actually slightly larger than production transistors in 2019. From an atomistic view point, there is little room for further scaling.

What does the size 10 nm mean? Around the year 2000, prototypes of this size of device were made. To explain what this meant, the prototype device was compared to a strand of DNA. In Figure 1, a transmission electron microscope of a DNA strand is compared to the same scale as a prototype 10 nm CMOS device. The DNA strand is on the same scale. In fact, within the prototype device, there is an oxide film with a critical dimension less than 1 nm in thickness. This is truly approaching the limits that scaling can achieve as 1 nm corresponds to a single molecular thickness of the material silicon dioxide, a key building component in this device. This is informative because typical production devices today are roughly at the same scale. There is simply not much room to become smaller. Our expected increase in efficiency may truly be at an end, or at least within sight of the end.

8 Jonathan Koomey and Samuel Naffziger, 'Moore's Law might be slowing down but not energy efficiency', *IEEE spectrum* 52.4 (2015), p. 35.

While the CMOS technology is physically limited to a minimum amount of energy as described above, a far more fundamental limit—Landauer's Limit—is determined by the second law of thermodynamics as it pertains to information. The minimum energy required to erase a classical bit is given by $k_B T \cdot ln(2) \approx 0.003$aJ at room temperature. There is room to move towards this fundamental limit by at least a factor of one thousand, even more if the system is cooled, if an ideal switch can be found. But at present there is no device that has a lower energy per switch than the CMOS device. An examination of available competing technologies that rely on charge transport shows that at best, the minimum energy of bit flipping could be reduced to 1 aJ (reducing the number of electrons/holes to ten to make bit and reducing the supply voltage to 0.3 V, regardless of how the transistor is made, is about the best that can be achieved) and this too would be limited by the statistical nature of quantum mechanics at room temperature. So, the Landauer Limit remains fairly far away from being functionally realized.

In its most idealized realization of computational efficiency, we see the supercomputer. The supercomputer is actually extremely efficient and makes use of the approximately 10 MW of power that it consumes for computation more efficiently than many other computing devices. An examination of the current champion of supercomputers,[9] Summit, located at the US Department of Energy Oak Ridge Laboratory (ORNL) near Knoxville, Tennessee, finds that it can execute 200 Thousand Tera FLOPs or 0.2 ExaFLOPS every second.[10] A quick calculation that looks at the power consumption and the rate of calculation:

$$\frac{Energy}{Flop} = \frac{Power\left(\frac{Joules}{second}\right)}{Computation\ Rate\left(\frac{FLOPs}{second}\right)} = \frac{10MW}{0.2 \cdot 10^{18}\left(\frac{FLOP}{second}\right)} = \frac{50pJ}{FLOP}$$

indicates that every FLOP consumes 50 pJ or $50*10^{-12}$ J. This is 10,000 times more than the minimum energy per bit required in state art microprocessors and about 10,000,000 times more than the minimum

9 'US Debuts world's fastest supercomputer' (n.a.), *BBC News* (June 11, 2019), https://www.bbc.com/news/technology-44439515.

10 Jonathan Hines, 'Stepping up to summit', *Computing in Science & Engineering*, 20.2 (2018), 78–82. DOI: 10.1109/MCSE.2018.021651341.

energy per bit defined by the Landauer Limit. In truth, each operation requires many individual bit transfers to perform one calculation. A further investigation shows that this disparity can be explained by the weakest link in energy efficiency. This is when the processing unit needs to access memory. This is the cause for hope in terms of engineering. While the reduction in energy and size in producing devices to access information is reaching an end there are simply a lot of inefficiencies to reduce to make the overall system more efficient.

10 MW, the power required to drive the world's more powerful supercomputer, is a lot—nearly enough power to meet the energy demand for 10,000 homes. While large, this amount of power available is not nearly as large as the amount of power required to supply a typical server farm that is currently being operated by companies such as Google or Facebook. Just 200 km or so east of the Summit supercomputer is a Facebook server farm. The power consumption of this operation is not published. The majority of the people that operate this server farm are in locations that are far away from this modest looking Forest City, North Carolina information hub. The physical size of this facility is about ten times the size of the Summit facility. The density of microprocessors inside the facility is about the same. If an extrapolation is made, you could note that the Facebook facility consumes about ten times the amount of power as the Summit supercomputer. That would imply that the energy consumed by this server farm would be equivalent to 100 MW. This is enough power to supply 100,000 homes. Both Google and Facebook, have committed to moving to renewable energy to help manage the high energy bills that they face from their data centers. Putting it into perspective, it takes over 70,000,000,000 kWh per year to run all data centers in the U.S.,[11] and the average household takes about 10,000 kWh per year.[12] It should be noted that the decisions are predominately driven by profit motive and not an altruistic desire to save the earth. Renewable energy has become more cost effective than non-renewable sources and these

11 Arman Shehabi et al., *United States Data Center Energy Usage Report* (Berkeley: Lawrence Berkeley National Laboratory, 2016), https://www.osti.gov/servlets/purl/1372902/.

12 US Energy Information Administration (EIA), *How Much Electricity Does an American Home Use?* (Washington: 2019), https://www.eia.gov/tools/faqs/faq.php?id=97&t=3.

technology companies are taking advantage of this. In April of 2018, Google announced that it had achieved 100% use of renewable energy in its operations.[13]

The Rising Demand of Computation

The total consumption of energy use for computation is on the rise and this rate of increase shows no signs of slowing down. Depending on the source, the percentage of energy consumption dedicated to computation has risen from under 2% in 2000 to over 6% in 2012 and the figure continues to rise.[14] While the exact figure is currently unavailable, recent trends would suggest that today at least 10% of all electrical power produced in the world is dedicated to computation. It is estimated by 2025, this figure could rise to as much as 20% of total worldwide power generation.[15] It should be noted that this figure includes the energy that is used for displays and that in 2000, that made up a large percentage of the resources required for computation. Today, that human interface is a mere fraction of the total energy expenditure required in calculation, most of the energy that is consumed as you are watching the most recent viral cat video is consumed in a server farm far away from you, not from the screen directly in front of you. To put this in perspective, in the year 2000, worldwide energy consumption for computation was still a minor consideration for overall energy consumption in the world. Refrigerators were a bigger concern than computers in 2000.

2000 is a good year to choose as the time when we transitioned from using discrete computers as our primary resource for computation to when we use networked resources to create our primary value. This dramatically pressured the demand for computation. Metcalfe's Law states the more similar devices that are connected on a network, the greater the value of the collective network by a factor of N^2 (it has been recently suggested to modify this expression to N*ln(N) instead of

13 Urs Hozle, '100% renewable is just the beginning', *Google* (5 April, 2018), https://sustainability.google/projects/announcement-100/.

14 United States, US Energy Information Administration (EIA), *2012 Energy Consumption Survey* (Washington: Government Publications, 2017), https://www.eia.gov/consumption/commercial/reports/2012/energyusage/.

15 Shehabi, et al. (2016).

N^2).[16] When this happened value increased dramatically for companies that made use of networked machines to add value. Among the best example to illustrate the increasing value of networked computers is the Amazon market cap stock evaluation. Amazon is the leader in server farms for hire. Since 2005, Amazon, a company that depends dramatically on the number of interconnected machines, has increased its value from $14 billion to nearly a trillion dollars. In that same time period, Intel the primary maker of computation machines rose from a value of $100 billion to $200 billion. More machines, N (Intel's value driver), were attached to the Internet and the value of a server farm was recognized by businesses, universities and governments worldwide and increased as N^2 (Amazon's value driver).

The energy consumption required for computation has increased at a steady but large rate but less than the amount of operations that are being performed. Furthermore—on demand video and social networking became increasingly dominant and the energy consumption of computation started to become a more relevant number. These two applications drove the increasing demand for bits as the energy cost per bit became increasingly lower. In the year 2000, the power supplies driving the server farms were barely 50% efficient and today that number stands closer to 90% and yet the total power consumption continues to grow. Computation, and its supporting infrastructure continued to be more efficient and this increase in efficiency was outpaced by the demand for more bits. An irony of technology development, known as Jevon's Paradox, says that as you make more efficient use of a resource, society as a whole will not use less of that resource, it will actually use more.[17] When things get more efficient more people want those things. We lived in a world where advanced computation was found only in the domains of businesses, governments and the scientists who needed advanced computation. Jevon's Paradox of consumption started to prevail as computation became more efficient. Computation spread to the masses and the energy consumption rose dramatically. The number of operations

16 Carl Shapiro and Hal R. Varian, *Information Rules: A Strategic Guide to the Network Economy* (Boston: Harvard Business Press, 1998).

17 William Stanley Jevons, *The Coal Question* (New York; The Macmillan Company, 1906).

used has been increasing at an extremely high rate while the energy per operation decreases at an almost equally fast rate through 2015. The combination of these factors is what drives the demand side of computation. It was the continued drive of reduced transistor size that kept the energy demand in check for much of this time, however, as discussed previously, the gains that come from transistor scaling are no longer available with CMOS scaling alone. It can be expected that in the near future, the energy expended per bit at the leading technology will be restricted to no less than 10 aJ. Working towards the 1 aJ limit and the Landauer Limit will be an unlikely path toward greater energy efficiency, despite the fact that this was the driver from more than fifty years.

This insatiable demand for bits of information is captured in Figure 2. This approach to the representation of this type of data was originally presented in an advocacy report from the Semiconductor Research Corporation and the Semiconductor Industry Association[18] and is updated here with a further breakdown of the entire world energy supply. Here, we can see the rising demand for information that is literally catching up with the total capacity for world energy production. By estimating the total number of bits that are used and monitoring its rise we can start to grasp the potential problems at hand. In 2010, it was estimated that for every bit of information that was processed it took 10 fJ to process. This is in fact about 2000 times the minimum amount of energy it should take today. This extra amount of energy comes from the fact that when a bit is processed it must also access memory and move this bit from one part of the computer system to another. The number of bits processed in the world increased by a factor of approximately 1000 from the year 2000 to the year 2010. From 2010 to 2015, this trend continued and there is little likelihood that this trend will slow down. Fortunately, as the world's demand for data increased, the efficiency of the way in which data was processed also increased. The minimum energy per bit reduced by a factor of about 100 during that time. Thus, the total energy still increased. The world went from having almost no noticeable energy consumption due to

18 Semiconductor Industry Association, *Rebooting the IT Revolution: A Call to Action* (2015), https://www.semiconductors.org/resources/rebooting-the-it-revolution-a-call-to-action-2.

computation, <2% of all energy produced on the grid being consumed in computation, to nearly 5% in 2010. The only reason that the impact on energy consumption has not been higher has been the steady gains throughout the electronic ecosystem that Moore's Law has given us for over fifty years.

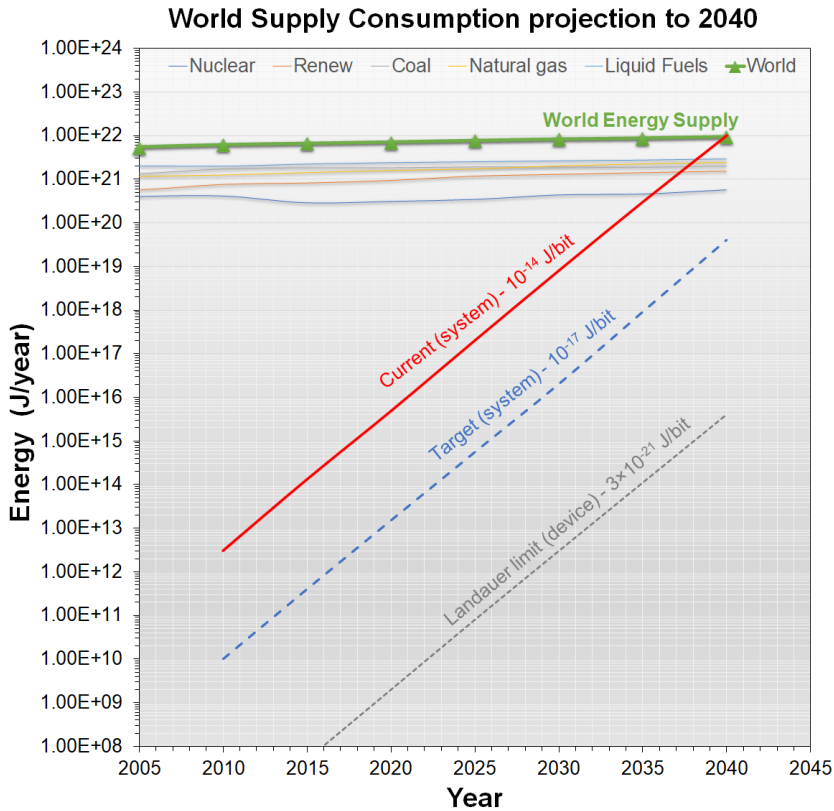

Fig. 2 Computational energy used with respect to total number of bits that are being demanded and the increasing amount of worldwide energy supply. This data is adapted from a joint report from Semiconductor International Association and the Semiconductor Research Corporation (2015).

Table 1 Comparison of fundamental units of energy consumption for computation.

Energy Consumption	Type of Operation	Source
50 pJ	Floating point operation (FLOP) requiring many bits to achieve an operation	Top Super Computers in 2018, Sumitt Statistics[19]
10 fJ	Average energy required to create or destroy one bit when considering memory access	Considerations from 2011 estimate that include memory access[20]
10 aJ	Energy required to create or destroy one bit of information in the microprocessor	Dimensions and operating conditions of reported devices of major semiconducting manufactures (14 nm node of Intel, TSMC and Samsung)[21]
0.003 aJ	Thermodynamic limit of the creation of a single bit of information	Landauer Limit[22]

A comparison of the required energy consumption indicates paths to strategies to reduce the total system energy in Table 1. The differences in the total energy consumed yields pathways that can be followed in meaningful engineering strategies to meet the expected 1000X/decade increase in demand for computation. Reducing the number of bits per operation is the most notable path to meaningful gains in energy

19 Chibuzor Aguwa, 'Top 5 fastest supercomputers in the world 2018', *Blogginfotech* (July 10, 2018), https://blogginfotech.com/top-5-fastest-supercomputers-in-the-world-2018/.

20 Semiconductor Industry Association, *2011 International Technology Roadmap for Semiconductors* (*ITRS*) (2015), https://www.semiconductors.org/resources/2011-international-technology-roadmap-for-semiconductors-itrs/.

21 Jin Cai (organizer), 'Scaling survival guide in the more than Moore Era', International Electron Devices Meeting (IEDM) 2018, San Francisco, 2 December 2018, short course.

22 Rolf Landauer, 'Irreversibility and heat generation in the computing process', *IBM Journal of Research and Development*, 5.3 (1961), 183–191, https://doi.org/10.1147/rd.53.0183.

efficiency. In fact, great strides have been made in this area. The most obvious example of this can be found within the supercomputers built at ORNL. Reducing the complexity of operation, i.e. reducing the number of bits required to perform an operation, allowed the energy per bit to be reduced by a factor of three from Summit, commissioned in 2018 to Titan, which was first turned on in 2012. The more impressive gains have come from the server world, where introductions to approximate computing has led Google to be able to perform searches with 200 times less energy with the same technology than they were able to use just a few years ago. To achieve these ends, Google uses what is commonly referred to as accelerators to perform specific computations and to streamline the computational efficiency per operation. They changed the amount of energy by changing the floating point energy computational energy required. They moved from 50 pJ to 400 fJ (125 times less) without changing the underlying technology. Reducing access to memory is also a means of improving computational efficiency. If the memory does not have to be accessed, each act of bit destruction could be reduced to the technology minimum of 10 aJ. This is a factor of 1000 to be gained by incorporating memory directly into the central processing unit. This gain could be achieved without further altering the core transistor technology. Changing the core technology to an ultimate limit that still uses principles of well-established semiconductor physics, could allow the number of electrons and holes to be reduced by a factor of 10 and the voltage that the system is operated to be reduced by a factor of 2. The maximum possible gain that could be obtained would be a factor of 20. This pales in comparison to the amount that stands to be gained from more efficient architectures and designs. The last is that driving towards the Landauer Limit and the gains that can be found from improving the underlying semiconductor technology in some capacity. Can there be a single electron device that can operate without error and be manufactured and considered reliable? That answer is not certain and will certainly be pursued in the coming years by the organizations that have the most to gain from the increased gain in efficiency.

The architecture gains that were alluded to in the last paragraph form the foundation for the applications that will drive the new sources of demand in the coming ten to twenty years. What was loosely recognized as the smarter way to do computing, and is being

implemented, is the foundation for what could loosely be referred to as neuromorphic computing. This type of computing is, in some sense, the hardware realization of artificial intelligence. We have become familiar with artificial intelligence in our daily lives through many applications that exist just below our awareness. These applications have been almost universally software driven. This next step, incorporating these algorithms directly into the hardware has already begun to yield fruits in efficiency and stands ready to be the next driving force in computer evolution. This is where Jevon's Paradox will again take hold as the next wave of computation will drive artificial applications such as self-driving vehicles. While watching the latest viral cat video took a measurable amount of energy, being driven to work by an algorithm operating on a remote computer connected by the 5G network will take 100–1000 times more energy. Yet that algorithm when run today will be executed in a way that is 1000 times more efficient than if it were to be executed with technology available in the year 2000. Autonomous vehicles are just one of the many artificial intelligence applications that are becoming more prevalent and constitute just one more increase to the energy demands of computation.

Pursuing neuromorphic computing is a realizable approach to computing improvements, just like CMOS scaling was the most realizable approach to gains in efficiency from 1980 through 2015. An important fact that should be noted when considering advanced CMOS processes: in 2000, there were nearly thirty companies in the world that could produce the most advanced silicon devices. Today there are only three. In 2000, the typical minimum dimension in production at the advanced node was 180 nm and today it is 10 nm. The capital investment for every 40% shrink has increased even faster than 40% and slowly eliminated the number of companies able to keep up at the leading edge. In terms of energy consumption per bit, that represents a nearly 400 times improvement from the year 2000 to today. Of these three companies, one is what is referred to as a foundry service (Taiwan Semiconductor Manufacturing Company, TSMC), one offers both internal and external services (Samsung) and the last (Intel) only manufactures products that are issued under its brand. Foundry services enable companies to produce chips with their own designs with the most advanced silicon. Recently former participants in this economic sector, IBM and Global

Foundries have bowed out of the race at the edge as neither could compete due to financial considerations.

Today, Apple and Huawei do not produce silicon microprocessors; but they do build the most advanced cell phone microprocessors in the world in the same multi-billion-dollar facility in Taiwan.[23] The most advanced products and computationally energy efficient devices in the world are built side-by-side, using the exact same technology. What does this mean for experimental research in neuromorphic computing at the leading edge? Both advanced microprocessors and neuromorphic computing require an extraordinary amount of capital; this largely prohibits universities and other smaller facilities from being able to do experimental research in this area. What can be said about computing is that, ultimately, the most energy efficient endeavors have prevailed in the marketplace and slowly eliminated competition. Neuromorphic computing, and its related approaches, shows a high potential for energy gains and for increasing its chances of a successful and significant market penetration. These gains, however, only occur when operating within the smallest node. Only the largest players who have access to the smallest nodes will be able to compete in the field. The best ideas that enable energy savings will largely not make it out of the laboratory without access to opportunities for experimenting within these advanced nodes.

Navigating the Future of Computing

Unlike the last fifty years, where the gains came largely from material, device design and technology implementation, the gains of tomorrow will come from improved architectures. This approach is commonly referred to as accelerators or heterogeneous computing. Accelerators are application specific integrated circuits (ASICs) that offer advantages in speed or energy efficiency for very specific tasks. The number of people that can be trained in these new architectures will be limited by the companies that control the leading edge of technology. In the past, access

23 Malcolm Owen, 'Apple, Huawei both using 7nm TSMC processors, beating out Qualcomm and Samsung', *Apple Insider* (October 2, 2018), https://appleinsider.com/articles/18/10/02/apple-huawei-both-using-7nm-tsmc-processors-beating-out-qualcomm-and-intel.

to advanced device manufacturing could be realized with relatively small investments. Today, at least in the near future, that capability is lost. While companies such as Apple, Amazon or Google can afford to hire either Samsung or TSMC to build speciality chips, the rest of the world cannot. This limits the open expression of ideas that has largely defined the semiconductor and digital applications world for most of its existence, and has driven the innovation cycle. The expense of access to a resource which has largely become a commodity, as necessary to modern life as petroleum, will be held in the hands of fewer and fewer organizations as time progresses. Recognition at the government level is what is required to ensure this openness, and to ensure that this ever-increasing commodity is not unnaturally limited.

National strategies to maintain and create access in this area are being pursued by both China and Saudi Arabia. Both nations have committed an extraordinary amount of resources to enhance the sovereignty of the data that flows through both nations. India has also announced plans to build its own national foundry to rival that of TSMC. These are examples of governments that recognize that computation is fundamental to their growth and survival as a nation. China has committed to invest $150 billion USD through an investment fund (Tsinghua Unigroup) that is matched by another $300 billion in corporate investment to develop competitive manufacturing at the most advanced nodes of semiconductor manufacturing. In Saudi Arabia, there is an investment fund, nearly the size of China's investment fund, to establish semiconductor manufacturing in the region. The United Arab Emirates privately owns Global Foundries, the third largest semiconductor foundry in the world. The United States has more recently recognized the attention that is required by initiating legislative discussion of this with the Semiconductor Foundry Act of 2020.[24]

Our future world of computing will look a lot like the world that we live in today, however the biggest change will be what drives that world behind the scenes. There will be an increasing number of specialty chips

24 Shannon Davis, 'American Foundries Act would provide needed investments in U.S. semiconductor manufacturing, Research', *Semiconductor Digest* (June 26, 2020), https://www.semiconductor-digest.com/2020/06/26/american-foundries-act-would-provide-needed-investments-in-u-s-semiconductor-manufacturing-research/.

to perform specific tasks within the computing environment. We already see the first steps in phones that offer facial recognition, delivered by an additional chip that is designed to specifically address facial recognition algorithm—not general computing. The energy that would be required in a general-purpose chip would be too high to maintain, so an accelerator designed to do nothing but provide facial recognition enables energy efficiency for this task. This is a good example of the many accelerators that are likely to come. The energy efficiency that is gained is obvious. Future tasks will see a master controller central processor with many accelerators providing functions of which we have not yet fully conceived. The nations, companies and individuals that can offer advancements in these areas will lead the semiconductor industry and subsequently the information technology of the future.

Acknowledgement

The authors acknowledge support of the work through a grant from the University of California, Lawrence Berkeley National Laboratories and the U.S. Department of Energy. For directed nanoscale research in Computational Electronics. Subcontract No. 7334354.

Bibliography

Aguwa, Chibuzor, 'Top 5 fastest supercomputers in the world 2018', *Blogginfotech* (July 10, 2018), https://blogginfotech.com/top-5-fastest-supercomputers-in-the-world-2018/

Bardeen, John, and Walter Hauser Brattain, 'The transistor, a semi-conductor triode', *Physical Review*, 74.2 (1948), 230–231, https://doi.org/10.1103/physrev.74.230

Cai, Jin, 'Scaling survival guide in the more than Moore Era', International Electron Devices Meeting (IEDM) 2018, San Francisco, December 2, 2018, short course.

Dawon, Khang, 'Electric field controlled semiconductor device', U.S. Patent No. 3,102,230 (August 27, 1963).

Davis, Shannon, 'American Foundries Act would provide needed investments in U.S. semiconductor manufacturing, research', *Semiconductor Digest* (June 26, 2020), https://www.semiconductor-digest.com/2020/06/26/

american-foundries-act-would-provide-needed-investments-in-u-s-semiconductor-manufacturing-research/

Dennard, Robert H., et al., 'Design of ion-implanted MOSFET's with very small physical dimensions', *IEEE Journal of Solid-State Circuits*, 9.5 (1974), 256–268.

Hines, Jonathan, 'Stepping up to summit', *Computing in Science & Engineering*, 20.2 (2018), 78–82.

Hozle, Urs, '100% renewable is just the beginning', *Google* (April 5, 2018), https://sustainability.google/projects/announcement-100/

Intel, *1970 Intel Financial Statement* (Arthur Young & Company, 1971), https://www.intel.com/content/dam/www/public/us/en/documents/corporate-information/history-1970-financial-statement.pdf

Jackson, Tim, *Inside Intel: Andrew Grove and the Rise of the World's Most Powerful Chip Company* (New York: Viking Penguin, 1997).

Jevons, William Stanley, *The Coal Question* (New York: The Macmillan Company, 1906).

Koomey, Jonathan, and Samuel Naffziger, 'Moore's Law might be slowing down but not energy efficiency', *IEEE spectrum*, 52.4 (2015).

Landauer, Rolf, 'Irreversibility and heat generation in the computing process', *IBM Journal of Research and Development*, 5.3 (1961), 183–191, https://doi.org/10.1147/rd.53.0183

NASDAQ, *INTC Company Financials* (2019), https://www.nasdaq.com/market-activity/stocks/intc/financials

Noyce, Robert N., 'Semiconductor device-and-lead structure', U.S. Patent No. 2,981,877 (April 25, 1961).

Owen, Malcolm, 'Apple, Huawei both using 7nm TSMC processors, beating out Qualcomm and Samsung', *Apple Insider* (October 2, 2018), https://appleinsider.com/articles/18/10/02/apple-huawei-both-using-7nm-tsmc-processors-beating-out-qualcomm-and-intel

Semiconductor Industry Association, *Rebooting the IT Revolution: A Call to Action* (2015), https://www.semiconductors.org/resources/rebooting-the-it-revolution-a-call-to-action-2

Semiconductor Industry Association, *2011 International Technology Roadmap for Semiconductors* (*ITRS*) (2015), https://www.semiconductors.org/resources/2011-international-technology-roadmap-for-semiconductors-itrs/

Shapiro, Carl, and Hal R. Varian, *Information Rules: A Strategic Guide to the Network Economy* (Boston: Harvard Business Press, 1998).

Shehabi, Arman, et al., *United States Data Center Energy Usage Report* (Berkeley: Lawrence Berkeley National Laboratory, 2016), https://www.osti.gov/servlets/purl/1372902/

'US Debuts world's fastest supercomputer' (n.a.), *BBC News* (June 11, 2019), https://www.bbc.com/news/technology-44439515

United States, US Energy Information Administration (EIA), *2012 Energy Consumption Survey* (Washington: Government Publications, 2017), https://www.eia.gov/consumption/commercial/reports/2012/energyusage/

US Energy Information Administration (EIA), *How Much Electricity Does an American Home Use?* (Washington: EIA, 2019), https://www.eia.gov/tools/faqs/faq.php?id=97&t=3