Adam Deryło
MISMaP

Beyond Moore's Law: The Future of Computing

Moore's Law has served as the guiding principle of the technology industry for over half a century, facilitating exponential increases in computing power and performance. However, as we near the physical limits of miniaturization, some experts predict that Moore's Law may soon reach its end. While this may represent a significant turning point for the industry, the future of computing is not necessarily dire. In fact, parallelization and the abundance of renewable energy sources present opportunities for a bright future. Supercomputers capable of addressing the most complex problems and devices powered by the sun are just a few examples of the limitless potential for progress. As we bid farewell to Moore's Law, it is important that we consider the exciting possibilities of a new era of computing.

However, let's not get ahead of ourselves. Where did Moore's Law come from, and why might it come to a sudden end? Moore's Law was originally proposed by Gordon Moore, co-founder of Intel, in a 1965 paper[1] in which he predicted that the number of transistors that could be placed on a microchip would double approximately every two years. This prediction has held true for many years, leading to exponential increases in computing power and decreases in cost. However, as transistors have become smaller and more densely packed, the challenges of miniaturization have become increasingly difficult to overcome. At the smallest scales, transistors are prone to a variety of problems that can limit their performance. For example, as transistors become smaller, they also become more vulnerable to fluctuations in temperature and voltage, which can affect their behavior. Additionally, as the size of transistors approaches the atomic scale, it becomes increasingly difficult to control their behavior, which could further limit the potential for improvement. Furthermore, smaller transistors require more power to operate, which can lead to increased heat generation and potentially cause damage to the chip[2]. Finally, as transistors are packed more densely, there is a greater risk of electrical interference (leakage) between them, which can also affect their performance[3]. All of these problems could contribute to the end of Moore's Law. In January 2019, NVIDIA CEO Jensen Huang declared that Moore's Law was dead. But what does this mean for the future of computing?

To understand the implications of the end of Moore's Law, it's important to consider the nature of the prediction itself. Moore's Law is a techno-economic forecast, which means it is influenced by both technological developments and economic incentives. While the demand for faster computing is unlikely to dissipate, the end of Moore's Law does signify the end of one of the most feasible paths for achieving this goal. In other words, we have reached the limits of what can be accomplished through miniaturization and transistor density, and we must seek new methods for continuing to improve computing performance. Nonetheless, the global semiconductor market, valued at approximately 555.9 billion dollars in 2021[4], is still

[1] https://hasler.ece.gatech.edu/Published_papers/Technology_overview/gordon_moore_1965_article.pdf
[2] https://iopscience.iop.org/article/10.1088/0034-4885/68/12/R01
[3] https://en.wikipedia.org/wiki/Leakage_(electronics)
[4] https://www.statista.com/statistics/266973/global-semiconductor-sales-since-1988/

projected to double by the end of the decade[5], making it a trillion dollar industry despite any concerns about miniaturization and transistor density.

In the realm of semiconductor design, there are three main avenues to explore. One option is to optimize architecture and packaging, as demonstrated by the contrasting approaches of chiplet and monolithic die manufacturing[6]. Second, there are new materials that could potentially challenge the paradigm of silicone semiconductors. And last but not least, there are completely new models of computation such as quantum computing. Each one of these options has a slightly different timeline for possible impact on the industry. Advancements in more efficient architecture, be it through better packaging or hardware specialization, can be expected to hit the market continuously with each development cycle. This means that we can expect to see regular improvements in this area, albeit at a potentially slower pace than what has been seen under Moore's Law. Conversely, devising solutions that utilize novel materials or rely on an advance in basic device physics is a much more complex and research-intensive process. Estimates suggest that, based on the history of the silicon fin field-effect transistor (FinFET), it would take a decade to three decades of development for any new material to potentially disrupt the semiconductor industry. This implies that while new materials offer considerable potential, their impact is likely to be longer term. Alternative models of computation, such as quantum computing, are generally viewed as supplementary approaches that may or may not eventually be adopted in the market. There is currently no concrete timeline for their adoption, and some experts have cautioned against the possibility of a "quantum winter". It is important to approach these endeavors with a long-term perspective and manage expectations accordingly[7].

It is important to consider the end users of the computing power facilitated by Moore's Law. A review of semiconductor spending reveals that data centers constitute a significant portion of the market, accounting for nearly half of all spending[8]. This raises the question of how data centers and high-performance computing (HPC) in particular can continue to improve their performance, as they have been the primary drivers of rapid advancements in transistor miniaturization.

In addition to optimizing conventional CPU architecture, it is likely that we will see increased reliance on specialized hardware such as graphics processing units (GPUs) and field-programmable gate arrays (FPGAs). At present, the majority of the top 10 fastest supercomputers utilize GPU acceleration provided by AMD Instinct MI250X or Nvidia's Ampere A100. The "Frontier" supercomputer, which utilizes GPU acceleration, has even recently achieved exascale performance![9] The use of specialized hardware, such as GPUs and FPGAs, allows for greater parallelism and efficient processing of certain types of tasks, making them well-suited for use in high-performance computing applications. It is probable that we will witness continued adoption of these types of hardware in the development of future

[5] https://www.mckinsey.com/industries/semiconductors/our-insights/the-semiconductor-decade-a-trillion-dollar-industry
[6] https://www.verilogpro.com/how-chiplets-assemble-into-the-most-advanced-socs/
[7] https://royalsocietypublishing.org/doi/10.1098/rsta.2019.0061
[8] https://www.mckinsey.com/industries/semiconductors/our-insights/the-semiconductor-decade-a-trillion-dollar-industry
[9] https://www.top500.org/lists/top500/2022/11/highs/

supercomputers, as they offer a promising means of achieving ever-higher levels of performance.

Furthermore, there is already a rising trend developing problem specific hardware such as TPUs. Google's Tensor Processing Units (TPUs) are custom application-specific integrated circuits (ASICs) that are specifically designed to accelerate machine learning workloads. They are well-suited for these types of workloads because they are optimized for the matrix operations that are commonly used in machine learning, and they also have a high memory bandwidth, which is important for efficiently accessing the large amounts of data that are often used in machine learning applications. One key advantage of TPUs is their use of mixed precision arithmetic, which involves using both 16-bit and 32-bit floating point numbers in computation. This allows TPUs to perform many machine learning operations with increased numerical accuracy while using fewer resources (e.g., less energy and memory) compared to using only 32-bit floating point numbers[10].

Another alternative catalyst that could substitute for the miniaturization of transistors as a leading driving force for improved performance is the increasing abundance of cheap renewable energy[11]. As the availability of renewable energy sources continues to grow, it becomes increasingly feasible to power large-scale computing systems with sustainable energy sources. This could potentially alleviate some of the constraints on performance improvement imposed by the physical limits of transistor miniaturization. Instead of solely relying on the shrinking of transistors to increase performance, the use of renewable energy could enable the use of more powerful and energy-intensive hardware, such as specialized accelerators and multi-core processors. This could result in significant improvements in the performance of computing systems across a range of applications, including machine learning, scientific simulation, and data analytics.

In light of these considerations, the potential end of Moore's Law could have the unforeseen consequence of financially incentivizing data centers and other high-performance computing (HPC) users to subsidize the construction of renewable energy sources in order to utilize their excess energy for computing purposes. There are several factors that could motivate such a decision, including the desire to reduce energy costs, the potential for financial incentives related to the use of renewable energy, and the desire to decrease the environmental impact of computing. One potential model for this type of arrangement is for data centers and other HPC users to enter into long-term power purchase agreements (PPAs) with renewable energy providers. Under a PPA, the data center or HPC user agrees to purchase a predetermined amount of renewable energy over a set period of time, often at a fixed price[12]. This can provide a stable source of funding for the construction of new renewable energy facilities and enable the data center or HPC user to access renewable energy at a predictable cost. Another potential approach is for data centers and HPC users to invest directly in the construction of renewable energy facilities, either by building their own or by partnering with other organizations to fund new projects. In either case, the excess energy generated by these

[10] https://cloud.google.com/blog/products/ai-machine-learning/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

[11] https://www.irena.org/publications/2022/Jul/Renewable-Power-Generation-Costs-in-2021

[12] https://www.projectfinance.law/publications/2020/october/powering-data-centers/

facilities could then be used to power the data center or HPC system, providing a reliable and sustainable source of energy for computing purposes.

In conclusion, it is clear that the world of computing will continue to evolve, even as the physical limits of transistor miniaturization may be reached. The transition to renewable energy sources, the ongoing efforts to improve parallel processing, and the significant advances in machine learning are all likely to play a major role in driving scientific breakthroughs and shaping the future of computing. These trends hold the promise of enabling new technologies and applications that could have a profound impact on the way we live and work. Although the future is always uncertain, these prospects offer a glimpse of the exciting possibilities that lie ahead, even in a post-Moore's Law world.