Future Trends of Microprocessor Design

COMP-2660

October 17, 2023

Deni Rakovic

110081508

https://www.ieee.org/content/dam/ieee-org/ieee/web/org/conferences/conference-template-a4.docx

I confirm that I will keep the content of this assignment confidential. I confirm that I have not received any unauthorized assistance in preparing for or writing this assignment. I acknowledge that a mark of 0 may be assigned for copied work.

Future Trends of Microprocessor Design

I. Introduction

The microprocessor, often thought of as the brain of modern electronics has undergone outstanding evolution over the past 3 decades. This rapid pace of development has fueled the digital age we live in today. With applications in a multitude of different domains globally, microprocessors are here to stay with us and will be a part of our lives for decades to come. In 1971 the first microprocessor was introduced to the world, known as the Intel 4004 [1]. Since the release of the Intel 4004, microprocessors have largely followed Moore's law and doubled in performance every two years. Many challenges of microprocessor design lay ahead as the demand for compute power continues to increase. Fundamentally changing the practice of business, personal computing, personal lives in general and reliability of computers. Facing monetary and environmental costs, microprocessor design needs to evolve over time to meet these demands. In this article, we will investigate key trends of microprocessor design that give us a hint of the future.

II. Importance

Microprocessors are vital to the modern economy. Shaping virtually every aspect of our lives, technology is paramount to our modern life and economies functioning. A key component in every smartphone, tablet, computer, and many other devices is a microprocessor. Without advancements in the design of microprocessors many fundamental aspects of our lives today would be largely different or nonexistent. Consider the economic impact without microprocessor companies like Intel and AMD. The e-mail, text, video, or business invoice we communicate with instantaneously would not be possible without microprocessors. All forms of communication infrastructure whether it's an app on your phone, the 5G network itself or global ISP switches and routers require microprocessors. Instead, one would rely on sluggish letter mail. Furthermore, technological advancements such as AI (Artificial Intelligence) face challenges of large quantities of data that require large amounts of compute power. Today, we see GPUs (Graphics processing units) being used to fulfill this compute demand which again would not be possible without microprocessors. Words are not enough to emphasize the importance of microprocessors and their future designs. Instead, I simply ask you to look around you and think of all the different devices that are a part of your life that require microprocessors to function.

III. Miniaturization of transistors

In 1965, Intel co-founder Gordon Moore predicted that the number of transistors on an integrated circuit will double every year in his original paper. In 1975, Moore adjusted his prediction to every two years with minimal rise in cost [4]. Also known as "Moore's Law", Moore established a fundamental motivating objective for Intel and its competitors. In figure 1 we can see that the transistor count within microprocessors has rapidly been expanding over time. For further context, the Intel 4004 released in 2001 held 2,300 transistors, while in 2010 the average Intel processor held 560 million transistors. By the year 2030, Intel aspires to manufacture microprocessors with 1 trillion or more transistors. With the introduction of the Intel 4004 microprocessor in 1971 the transistor size was 10,000 nanometers. For comparison any modern Intel microprocessor is between 45 and 32 nanometers, while a human hair is 100,000 nanometers wide [1].

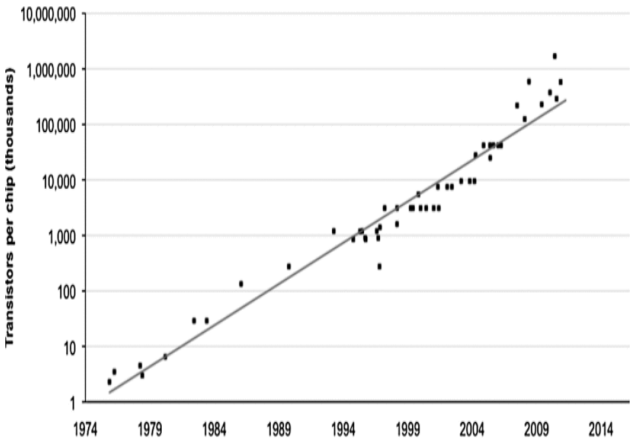

Figure 1 Transistors Per Chip [2]

With more transistors greater computational capabilities are expected and with efficient parallel processing capabilities, allowing greater development of microprocessors for computationally heavy applications like machine learning and artificial intelligence. Increasing the transistor count proves to bring additional challenges to scientists and engineers. Heat dissipation increasingly becomes concerning requiring sophisticated cooling solutions to prevent overheating. Power consumption can increase with more transistors which may not allow more power-hungry processors to be implemented in battery powered devices such as smartphones and tablets. With the miniaturization of transistors also comes improved graphic processing which can enhance real-time rendering in various applications such as Virtual Reality (VR) and Augmented Reality (AR). Perhaps the engineers of the future will be using computer-aided design software in VR, rather than on their monitors. Or perhaps in the future we will be able to render virtual worlds nearly indistinguishable from the real world with advancements in microprocessor design. Coupled with advancements in AI and Machine Learning (ML), the future of microprocessors and continuation of increasing the density of components like transistors is crucial for advancements in these fields to take place due to their computational hunger for performance.

IV. Increasing processor speeds

Over the years, processor speeds have increased tremendously. From 1995 – 2004 processor speeds increased 64% year over year on average. Although, this pace of performance (speed) growth is not expected to last. From 2004 – 2011 performance growth has been about 21% year over year [2]. Many attribute the slow in performance growth to heat dissipation limits which still pose a problem today and

restricts the number of transistors that could be placed within a core. In the mid-2000s, interest in the increase in clock frequency was neglected due to power dissipation barriers. Intel and AMD both aimed for smaller transistor sizes to achieve higher operating frequency, but the chips became too hot and required unreasonable cooling systems [3]. Figure 2 shows the trend of key microprocessor metrics. Even with the given challenges, processor speeds have continued to increase over the long run due to new developments in manufacturing and architecture design. Of particular interest is how the trend of transistors has increased exponentially over time, while clock frequency has stayed relatively flat. Furthermore, single-thread performance has continued to increase, at a slower pace when compared to transistor count. Core count continues to increase largely to compensate for the flat clock frequency trend. Many manufacturers such as Intel have instead placed a higher number of cores within their microprocessors in order to allow more parallel operations to be performed rather than deal with issues of heat dissipation with such dense components in microprocessors. Furthermore, the addition of optimized instruction sets within the processors themselves have added to faster execution of specific tasks. Manufacturers are continuously improving on these instruction set architectures (ISAs) and refine them over time, increasing processing power.

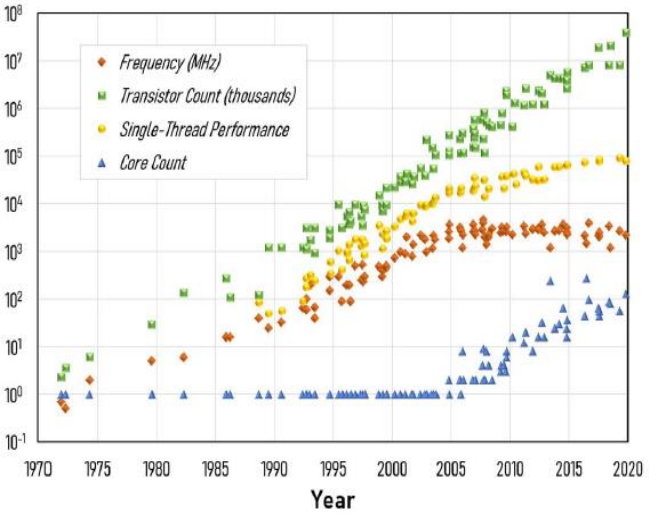

Figure 2 - Trends for key microprocessor metrics [3]

V. INTEGRATION OF AI AND ML IN MICROPROCESSOR ARCHITECTURE

Artificial Intelligence (AI) has recently exploded in interest by professionals around the world. Artificial intelligence is short of a new discovery or thought. British mathematician Alan Turing's work on measuring machine intelligence had a profound impact on field of AI. The Turing test involves a human "judge" who engages in natural conversation with both a human and a machine designed to produce human-like responses. Typically done over text-based communication, if the human "judge" cannot reliably distinguish between the two participants (the machine or the human) they are conversing with, then the machine is said to have passed the Turing Test and demonstrated a form of artificial intelligence. With the recent explosion of Large Language Models (LLMs), there is no doubt that the development and hype around the field of AI is here to stay. Large language models refer to the number of parameters that the model takes as inputs. These parameters can range from millions to billions. Verint a customer engagement company created Davinci, one of the largest and most powerful LLM with 175 billion parameters and trained on 45 Terabytes of text data [8]. Capable of handling most natural language tasks as well as image captioning and visual reasoning, training such large models can be extremely computationally expensive. GPT-3, a popular generative text model was pre-trained using 1,024 GPUs over the course of 34 days, costing $4.6M in compute resources alone [8]. As microprocessors evolve to complete more complex problems in different application domains, it is likely that many microprocessors will cater to the AI field. For example, Convolutional Neural Networks (CCN), a class of neural networks that specializes in processing data was proven useful by the implementation of CCN in ImageNet. ImageNet is a large visual database for use in object recognition software. The most resource hungry calculation in CCN computation is the multiply and accumulate operation between inputs and trained weights [5]. These operations are performed on multiple data points simultaneously to achieve swift and efficient performance. To assist with this processing, many modern processors feature a vector instruction set that allows computational processing of multiple data with a single instruction. Another example is the Intel advanced vector extensions, a 256-bit vector for each core which can support the processing of multiple floating-point operations in a single instruction [5].

VI. ENERGY EFFICENCY

Throughout the history of microprocessors, the push to reduce the size of transistors which in turn reduce the cost per transistor has been largely successful. To continue achieving this goal, we must also reduce the power consumption of transistors and thus microprocessors in general. Field-effect transistors (FET) used to control the flow of a current in a semiconductor have in our recent age experienced diminishing returns in terms of scaling [6]. The need for FETs to meet newer demands for semiconductors and in turn microprocessors has brought many to the attention of carbon nanotubes (CNT) for their energy efficiency. Carbon nanotubes, are nanoscale cylinders generated from a single sheet of carbon atoms with diameters of about 1 -2 nanometers [6]. CNTs offer high electrical conductivity and carrier mobility resulting in less power loss and thus improved energy efficiency. Currently, scientists are struggling with three main challenges of CNTs when applied to modern systems which can consist of billions of FETs. Material defects because of difficulty in precisely controlling CNT diameter results in some percentage of metallic CNTs leading to high leakage current or incorrect logic functionality. Manufacturing defects resulting in high particle contamination rates and lastly, variability of characteristics like uniform threshold voltages. As the fabrication process evolves, the hope is these three issues will be eliminated. However, even with these new issues, over the long run energy efficiency has largely improved over time.

Computations per KWh have been increasing as seen in figure 3. Thankfully, this success
in the long run has positively affected the development of mobile technologies like the smartphone and tablets many of us use today. Without increases in energy efficiency, mobile devices would not be as capable and reliant as we know them since battery technologies have not improved as rapidly as microprocessor and semiconductor technologies. As demands for compute power increase over time, scientists, engineers will have to face the challenges that CNT technology presents or find newer methods to generate newer, more power efficient microprocessors.

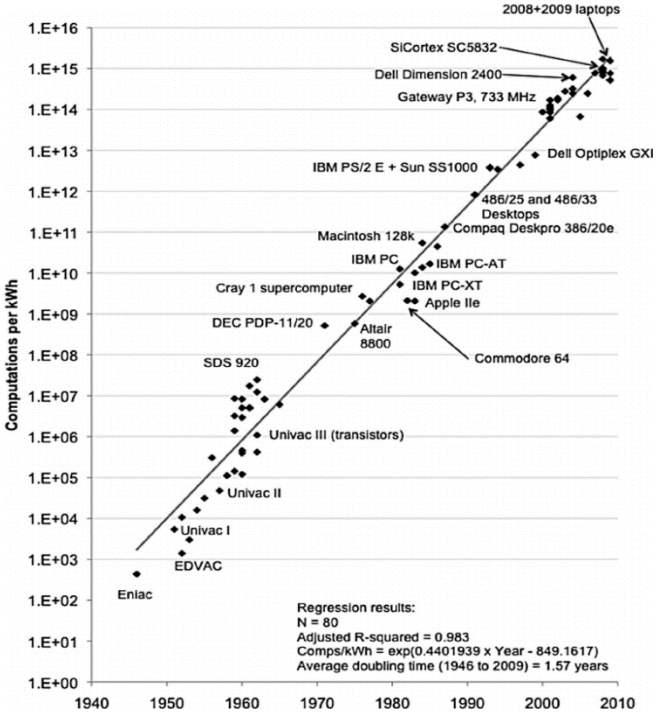

Figure 3 – Computations per KWh [2]

VII. USE OF EMERGING MATERIALS

Modern microprocessors are almost exclusively made with silicon. As advances in microprocessors have created a large density of components such as transistors to meet higher performance demands, challenges like heat dissipation need to be tackled with new methods. With increasing demand for higher integration density and speed as mentioned before, this has put a strain on the limitations of the silicon material and lead to a search for newer emerging materials. Transition metal dichalcogenides, black phosphorus, silicene and other materials known as two-dimensional (2D) materials are being explored as promising candidates for future microprocessors to keep up with demands. 2D materials offer significantly interesting properties such as higher thermal conductivity that graphene offers may prove useful in the search for emerging materials. Recently in 2023, the first functional microchip based on 2D materials has been produced at King Abdullah University of Science and Technology (KAUST) in Saudi Arabia. The microchip utilizes atomically thin layers of graphene which previously have been difficult to use in the fabrication process due to their flimsy and fragile nature [6].

Graphene layer coating has proven promising to avoid limitations of the heat generated by microprocessors. A study published in 2019, tested heat transfer in pure and graphene coated microprocessors. The test was conducted based on CPU utilizations of 0%, 50% and 75% utilization. The graphene was mixed with ethanol and spin-coated on the surface of microprocessors for the study. As graphene has high thermal conductivity, the study found that a maximum of a 5.6 degrees Celsius difference in heat transfer was achieved by introducing the graphene layer on the microprocessor, proving that graphene is a suitable material for heat transfer, resulting in higher amounts of heat dissipation when coupled with the modern-day cooling solutions like heat sinks and fans [6]. Furthermore, using 2D materials for field-effect transistors (FET), a transistor that uses an electric field to control the flow of a current in a semiconductor. 2D materials seem to have a plausible potential for FETs as they require extremely thin channels. Since 2D materials maintain good carrier transport even for atomically thin layers below 1 nano meter, it's possible the future of microprocessors lies with one or many 2D materials that are currently being studied [6]. From the industrial perspective, integrating 2D materials within existing silicon manufacturing processes of semiconductors has been the most important [6]. Furthermore, studies have revealed that 2D materials can also assist with other pieces of technology such as memory, integrated photonics, memory (RAM, SSD, DRAM etc.) and even sensors. If 2D materials are more accepted by the manufacturing industry on a wide scale and integrated into their fabrication processes, we may see 2D materials accepted on more of larger scale, rather than just for FETs used in semiconductors.

VIII. DISCUSSION

Microprocessor design stands at the core of our digital age. Our ambitions as a society largely depend on computing and effects nearly every global industry. As transistors edge closer to the atomic scale, the difficult and number of challenges increase. However, so do the rewards in the context of processor speeds. As this article points out, processor speed is no longer confined to clock frequency yet the efficiency of individual instructions within the ISA of microprocessors. Furthermore, emerging materials provide promise to help us in societies endeavor of advancing microprocessor design. Materials such as graphene and silicene help us advance while we are near the limits of silicone. The properties of these emerging materials may help us redefine the way we think about processing speed and efficiency and possibly assist us with breaking limits we never thought possible. New integrations of AI and ML directly into the architecture of microprocessors may pave the way for new advanced processors, capable of processing hundreds of thousands of TBs of data in short amounts of time. Lastly, we should also address that microprocessor design is incomplete without considering energy efficiency. As devices and microprocessors grow in power so will their demand for power. It should stay a global effort to include power efficiency into microprocessors and continue to advance optimization in this regard. The future of microprocessors is anything but bland, it is an incredible show of ingenuity of the world.

IX. CONCLUSION

As we illustrate the trajectory of microprocessor design in the future, several trends appear to have emerged that shape the future of computing. The relentless miniaturization of transistors continues to be a major objective of the industry, allowing for smaller chip designs and paves the way for unprecedented processor speeds. Integration of AI and Machine Learning (ML) directly into microprocessor architecture provides a paradigm shift. Allowing for more complex and adaptable computational processes to be performed. To add to this, in today's climate the increasing emphasis of lowering environmental and operational costs has led to advancements in energy efficiency. Furthermore, the exploration of promising emerging materials like graphene and other 2D materials offer promising avenues for creating the next generation of highly efficient and high-performance microprocessors. In conclusion, the future of microprocessor design is a promising interwoven construct of speed, intelligence, efficiency and innovation by scientists and engineers around the world.

REFERENCES

[1] Intel Corporation, "The Story of the Intel 4004". [Online]. Available: https://www.intel.com/content/www/us/en/history/museum-story-of-intel-4004.html. [Accessed: 16-Oct-2023].

[2] J. Koomey, S. Berard, "Implications of Historical Trends in Electrical Efficiency of Computing" April 2011.

[3] K. Radhakrishnan, M. Swaminathan, B. Bhattacharyya, "Power Delivery for High-Performance Microprocessors-Challenges, Solutions and Future Trends". April 2021

[4] Intel Corporation, "Moore's Law". [Online] Available: https://www.intel.com/content/www/us/en/newsroom/resources/moores-law.html#gs.72upzz [Accessed: 16-Oct-2023].

[5] F. Khan, M. Pasha, S. Masud, "Advancements in Microprocessor Architecutre for Ubiquitous AI" June 2021.

[6] D. Akinwande, C. Huyghebaert, C. Wang, M. Serna "Graphene and two-dimensional materials for silicon technology". September 2019

[7] T. Thangamuthu, R. Rathanasamy, S. Kulandaivel, G. Palanisamy, "Influence of graphene coating on altering the heat transfer behavior of microprocessors", April 2019 .

[8] D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley "Efficient Large-Scale Language Model Training on GPU clusters Using Megatron-LM", Aug 2021