RECENT TRENDS IN THE COST OF COMPUTING. IS INNOVATION COMING TO AN END?

Cătălin ȚURCAN

Technical University of Moldova

***Abstract**: Since the creation of modern CPUs, computing has seen a tremendous growth in terms of performance, while the price decreased each generation. This growth has seen diminishing returns in the recent years. It is only logical to assume that this technology is converging to its maximal potential. As you will see, this is not an assumption, it is a fact. The main question is how much time we have left and what solutions do we have in mind that will allow us to pursue a performance increase. Are we close to disruptive innovation?*

***Keywords**: computing, performance, price, chiplet, ASIC*

Introduction

The first microprocessor was invented in 1971. Its importance for the modern society is comparable to the discovery of fire. The microprocessor uses transistors as its main components to control the flow of electrons within the processor. Since there are two types of computing, parallel and sequential, there are 2 ways of increasing performance. You can either increase the amount of computing cycles per second or you can increase the amount of cores. Both of these methods do have a physical limit, therefore it is only logical to think that innovation will come one day to an end, and we are closer than you may think. For what reason are present day PCs such powerful when compared to the older ones? One clarification identifies with the colossal number of advances which have occurred at an architectural level over the past decades.

1. Moore's Law

As a rule of the thumb, the quantity of transistors that can be put onto an incorporated circuit duplicates every 18 months. This is known as the "Moore Law", even though it is more an observation than an actual law. Either way, this has been indeed the case for the last decades and this led to an exponential growth over the years. In the past decades we jumped from thousands of transistors to billions of them in a processor with relatively same die size. As we can see in the graph below, the amount of transistors increased tremendously over the past 40 years. In order to see how far we got, let's compare the original intel 8086 to the 2018's i7 8086K. This process happened at an exponential growth, but as you can imagine this growth is not sustainable, simply because there are physical limits that block us from doing so. It is important to note that this observation, denotes the fact that the number of transistors increase because of the density and not by sheer numbers. Increasing the size of the die will naturally increase the number of transistors, but that is cheating and it is not a solution simply because it will increase the temperature, power consumption and cost. These factors vastly outweigh the increased performance.

In the following paragraph we will compare the original intel 8086 to the one that was released in 2018, in terms of transistor count.

An intriguing thing to note is that Moore's Law turned out at an indistinguishable time from the Intel 8086 was created. The 8086 included around 29,000 transistors, each roughly 3.2 μm. The chip had a die size of 33 mm². The limited edition CPU released in 2018 highlights around 3 billion transistors. On the off chance that you figure it out the quantity of transistors has expanded by multiple times and remembering the die size is 149 mm², the transistor size has diminished by multiple times. These are exceptionally huge numbers and are no little accomplishments, without a doubt. Shrinking the die size is currently the most preferred method when planning to increase the speed of the CPU without having temperature constraints. The thing is that there is a certain point, in which shrinking the transistor will have a negative effect. At less than 5nm the electron is able to jump over the logic gates in the CPU causes by an effect called "quantum tunneling", according to Mohsen Razavy in his book "Quantum Theory Of Tunneling (2nd Edition)", "Quantum tunneling or tunneling is the quantum mechanical phenomenon where a subatomic particle passes through a potential barrier that it cannot surmount under the provision of classical mechanics." At such small sizes, even a hydrogen atom is considered big. In 2018, at the September event, Apple included in its Iphone XS series the world's first commercial 7nm processor. That is a truly big achievement, but it only proves that the end of innovation is near. In a matter of years, 5nm or 4nm transistor sizes will be achieved. Extreme

cooling (LN2, liquid Helium) will still be used to achieve record-high speeds, but this buff is temporary and has no practical use.

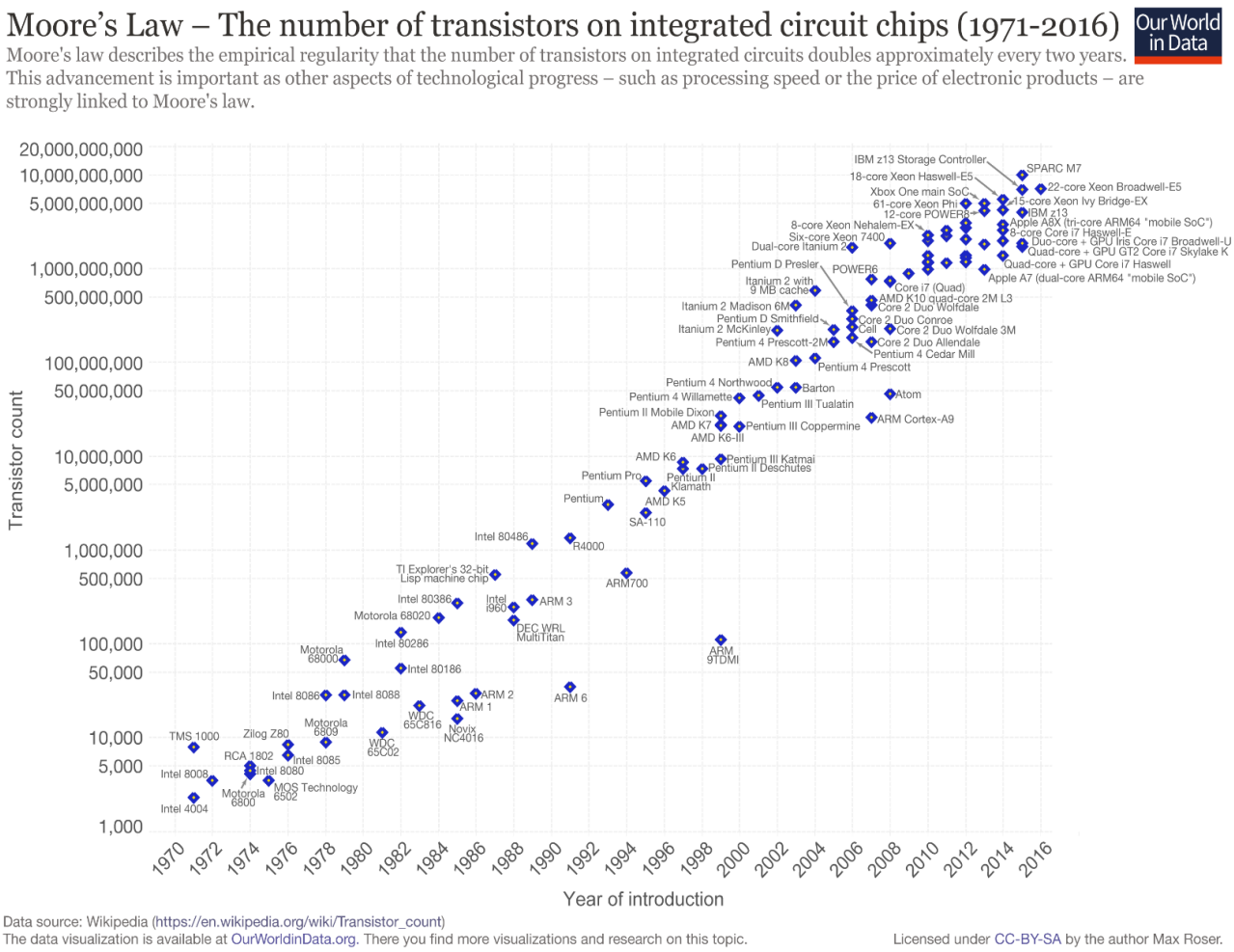

Fig 1. Number of transistors over the years.

2. Parallel computing

The story repeats itself this time, because adding more cores will always yield diminishing returns. Today, most programming is done sequentially and the code must be specifically designed to have parts that can be computed in parallel. Parallel computing will always be bottlenecked by the sequential part of the algorithm. In order to understand better, let's look at the newest products from Nvidia, specifically because GPUs are the best general-computing devices designed for parallelization. The "RTX 2080" has 48 tensor cores, specifically designed for ray-tracing, one of the most parallelized algorithms known today, in which the sequential code has a proportion of only 5%. The performance increase can be calculated Amdahl's law: "In computer architecture, Amdahl's law is a formula which gives the theoretical speedup in latency of the execution of a task at fixed workload that can be expected of a system whose resources are improved. It is named after computer scientist Gene Amdahl, and was presented at the AFIPS Spring Joint Computer Conference in 1967. This law is used to predict the amount of performance return proportional to adding more threads to a workload. (s is the speedup of the part of the task that benefits from improved system resources; p is the proportion of execution time that the part benefiting from improved resources originally occupied.)

$$S_{\text{latency}}(s) = \frac{1}{(1-p) + \frac{p}{s}}$$

Let's suppose the next gen RTX cards will have 64 cores, that will increase the performance only by 12%. Increasing it to 128, resulting in a 27% increase of performance. Let's suppose the card will have 4096 cores, returning only 34% more performance. In fact, above 128 cores, there will be no financial reason to

increase the cardinality of cores, because the price of implementing this amount of cores will vastly outweigh the performance return. At about 4096 cores, the performance return will plateau.

3. Is the end of innovation in computing coming to an end?

It is easy to determine the fact that we are at a major road-block when trying to increase the computing power while decreasing the price. Methods that worked for the past 40 years seem to fail. Is there any solution? First we must understand that our main focus is increasing the power in terms of general-purpose computing. Enterprise computing has less problems because of the appearance of ASICs (application-specific integrated circuit), a microchip designed for a special application, such as a particular kind of transmission protocol or a hand-held computer. You might contrast it with general integrated circuits, such as the microprocessor and the random access memory chips in your PC.). While being much less versatile and do only a certain type of operations (which it was specifically designed for), it will do those operations at less power, less cost and much faster than any general-purpose computer. Amazon is the world's biggest online retailer and it is no wonder they need a lot of computing power for their own services. After ditching Intel's processors, this is what AWS (Amazon web service) had to say: "We run our own custom-made hardware, made to our specifications, and we have our own protocol-development team. It was cost that caused us to head down our own path, and though there's a big cost (improvement) . . . the biggest gain is in reliability." This custom-made gear "has one requirement, from us, and we show judgment and keep it simple. As fun as it would be to have a lot of tricky features, we just don't do it, because we want it to be reliable." While ASICs will prove an efficient solution for some time, the ordinary citizen won't truly benefit from them. Currently, to keep pace with Moore's Law, Chipmakers Turn to 'Chiplets'. In this type of architecture, it is much easier to add features, since most designs feature 3D stacks. One of the biggest bottlenecks in a modern computer is the slow memory. Every time the CPU runs out of cache, it looks in the system memory. Even though the system RAM (random access memory) is quicker than any hard-drive or SSD on the market, it is about 10 times slower than the cache inside the processor. Chiplets mitigate this flaw by being able to add huge amounts of cache, resulting in a decrease in the number of times the CPU runs out of memory during an operation. The latest, greatest, and smallest transistors are also the trickiest and most expensive to design and manufacture. In processors made up of chiplets, that cutting-edge technology can be reserved for the pieces of a design where the investment will most pay off. Other chiplets can be made using more reliable, established, and cheaper techniques. Smaller pieces of silicon are also inherently less prone to manufacturing defects. Basically the processor is not build in a single process, but in multiple components, then all the pieces are assembled together resulting in cheaper production and lower-fail rate. The only disadvantage is the increase amount of dissipated head when operating, but given enough time, researchers will find a way, like they always do.

Conclusions

- General computing's market has seen no innovation for years.
- Older solutions that resulted in better performance are failing.
- Non-expensive solutions will result in small incremental improvements.
- Diminishing returns in computing are already present.
- There is no disruptive innovation at the moment.
- Chiplets may prove to be the path, CPU architecture will take in the nearest future.

References

1. Razavy Moshen *Quantum Theory Of Tunneling (2nd Edition)*, 2003
2. Apek Mulay *Sustaining Moore's Law: Uncertainty Leading to a Certainty of IoT Revolution*, 1965
3. Brock David *Understanding Moore's Law*, 2006
4. Fayez Gebali *Algorithms and Parallel Computing*, 2011
5. https://www.sciencedirect.com/topics/computer-science/amdahls-law
6. https://www.investopedia.com/terms/m/mooreslaw.asp
7. https://www.wired.com/story/keep-pace-moores-law-chipmakers-turn-chiplets/