The Past, Present, and Future of CPUs and Moore's Law

Honors University Physics II Project
By Joey Castrodale
H2

4/25/2011

Living in today's world, it is hard to imagine a society without technology. Cell phones and personal computers have become essential to the daily workings of businesses and industries. Behind these marvels of technology is what is known as the central processing unit. The central processing unit (CPU) is the component in a piece of technology that carries out the functions of a program for a device. The CPU consists of semiconductor devices called transistors which move the data across it. The more transistors on a CPU, the faster the data can be processed. In the advancement of CPUs, the question has always been how to fit more transistors onto a given space, as that would increase the speed. In the scheme of history, CPU technology is relatively young, having started around the early 1960's. However, CPUs have come a long way in a short time in terms of speed and performance. Gordon Moore, a co-founder of Intel, noticed a trend in this rapid expansion of CPU advancement. What is now known as Moore's Law states that the number of transistors on a chip will double every two years. This prediction was made in 1965 when Moore predicted this trend would continue for ten years. Surprisingly, it has still held true today. While it seems this law will never come to a halt, some scientists, including Moore himself, believe the law will eventually come to a stop. This paper will examine the beginnings of CPUs, what they are capable of now, future uses of the CPU, and also will examine Moore's Law in detail.

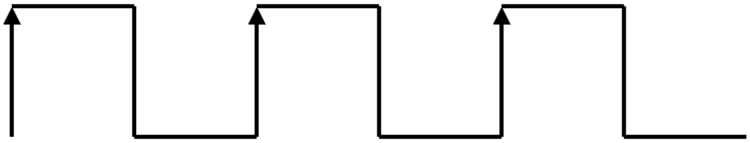
Before the idea of the CPU came into existence, the usage of semiconductors came first. In the late nineteenth and early twentieth centuries, now-historic scientists such as Lord Rutherford, J. J. Thomson, and Neils Bohr contributed greatly to the understanding of sub-atomic particles. Their studies of electrons laid the foundation for the study of semiconductors. A semiconductor is an element that is not a true conductor, nor is it an insulator; it has qualities of both. In the 1930s, scientists at the Bell Telephone Laboratories researched how to alter the electrical properties of semiconductors. Through a process called doping, they discovered that conductive paths could be formed along the crystals of semiconducting material when impurities were infused with them (Simon 1986, 3-5). “Their research culminated, in 1948, in the creation of a complete transistor within a piece of semiconductor material.” This transistor served as a prototype, and “…by 1954…Texas Instruments announced the first commercial silicon transistor.” (Simon 1986, 5). After the transistor was developed, exactly how to implement it became the next question.

The solution came in the form of the integrated circuit. The integrated circuit (IC) was the perfect complement to the newly developed transistor. Integrated circuits are basically pieces of silicon that host a number of connected transistors. Two scientists, Robert Noyce and Jack Kilby, both patented their own ICs that were commercially produced starting in 1961. As ICs became more common, the different classifications of ICs began to emerge. These classifications were identified by the number of transistors that could be placed on a piece of silicon. In this time period, there was Small Scale Integration (up to 64 transistors), Medium Scale Integration (up to 1,024 transistors), and Large Scale Integration and Very Large Scale integration (upwards of tens of thousands of transistors) (Simon 1986, 6). The way these digital circuits operated was “…on the basis of the presence or absence of an electrical voltage or current.” (Simon 1986, 7). In other words, the data can only be one thing or the other; on or off; up or down. This system is called binary logic. It is usually represented by 1s and 0s, which is referred to as binary code. When used with mathematical functions such as AND, OR, and NOT, most operations can be reduced to binary code (Simon 1986, 7).

In the early 1970s, Intel Corporation released the first microprocessors as they came to be known; the 4004 and the 8008. “The 4004 [was] for a calculator and the 8008 for a video display terminal." (Simon 1986, 8). Intel advertised these processors not as a replacement for minicomputers, but rather as components to be used to improve already-existing devices. Over the next few years, other companies began producing microprocessors (another name for modern-day CPUs), and products such as sewing machines and electronic games began to implement them (Simon 1986, 8-9). As the popularity of CPUs grew, the demand for faster CPUs grew as well. The main way to increase the speed of a CPU is to increase the number of transistors on that CPU. As a result, over the past fifty years, engineers have had to integrate more transistors onto a chip. Significant advancement has taken place since the beginning: from Small Scale Integration containing only sixty-four transistors, CPUs can now have well over one billion transistors (Murph 2010). In fact, according to Intel's website, the CPU inside the PC being used to write this paper, Intel's i7 720QM, has 774 million transistors!

Regardless of how many transistors a CPU has, the basic concept of how they work is still the same. All CPUs process data to perform a function. Whether it is a sewing machine or a computer, the CPU is the component that processes the data to carry out the function. It is important to note that the CPU is only one *component* of a device. The CPU does not actually perform a physical task; rather, it processes data and sends it to the other components that perform an action. In a personal computer (PC) for example, the CPU is attached to the motherboard along with other components, such as the Random Access Memory (RAM) and the graphics processing unit (GPU). However, the CPU is a crucial component, as it interprets the commands from a program. Running a video game on a PC is a good example. The user runs the application, which is stored on a hard drive. Then, the data is transferred through the RAM to the CPU. The CPU then processes the data and, depending on the program, sends it to the GPU to display an image, continues loading, or executes a different aspect of the program (Torres 2005, 1).

As mentioned earlier, the greater the number of transistors on a CPU, the greater the speed. The speed at which a CPU runs is called a clock rate, measured in Hertz (Hz). The clock rate is a measure of clock cycles per second. A clock cycle starts when a square wave (pictured below) changes from 0 to 1 (represented by the arrow).



(**Figure 1** – **Source:** http://www.hardwaresecrets.com/fullimage.php?image=1987)

For example, a clock of 50Hz means there are fifty clock cycles in one second (Torres 2005, 2). In a computer, the CPU must read data from the RAM. This causes a major slow down, as the clock speed of the typical RAM stick is much slower than that of a CPU. Processor developers invented a new method called clock multiplying to clarify this difference. The clock at which the CPU reads data from the RAM is called the *internal* clock and the clock at which the CPU processes data is called the *external* clock. For example, when one purchases a computer, it may have a "2.8 GHz processor." This means the internal clock of the processor is 2.8 GHz. However, the computer's RAM runs at 400 MHz. This means the processor's external clock is 400 MHz. That means the processor has to reduce its speed a factor of seven when it reads from the RAM, which is a significant slowdown.

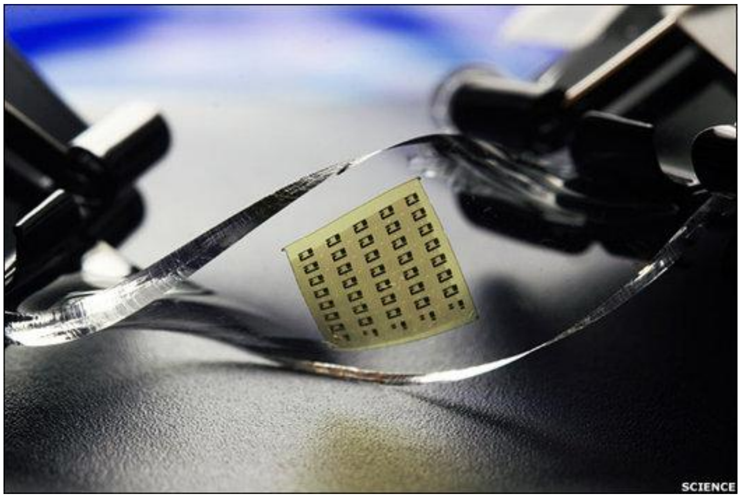
To compensate for this, CPU manufacturers have developed methods such as integrating memory cache inside the CPU, as well as increasing the amount of data transferred per clock cycle (Torres 2005, 3). Increasing the amount of data transferred per clock cycle makes sense for increasing external clock speed, but integrating memory cache inside the CPU might not be as clear at first hearing. How it works is actually rather simple. With cache memory installed into the processor, the CPU gathers data from the RAM like normal. Then, it begins processing that data. While that bit of data is being processed, the cache memory finds the next bit of data that is to be processed, so the CPU does not have to spend time finding more data itself (Torres 2005, 5).

The latest trend in CPU architecture is multi-core processors. A multicore processor is exactly what it sounds like: a processor that has multiple cores on the same chip. If a processor is a “dual-core” processor, it functions as two processors, while being one unit. This is made possible by parallel processing. With parallel processing, a CPU is able to perform multiple tasks at one time, or perform one task with the speed of two processors (TechTarget, 2004). It is important to note that a dual-core processor does not always function at the speed of two processors. That depends on the type of software. The software has to be specifically designed to utilize multi-core technology. Most personal computers on sale today come with multi-core processors, usually a dual-core processor. However, quad-core processors have broken into the mainstream lately with the release of Intel’s i7 processors.

In laboratories across the world, faster processors are needed to run supercomputers. These computers are not used for playing games or loading operating systems like personal computers. A supercomputer’s purpose is to crunch numbers. While one function of a supercomputer may be to visually display something such as a black hole on a screen, the main work of the supercomputer is running the calculations so the screen can accurately simulate said black hole. When speaking of the calculation speed of a CPU, scientists measure in terms of FLOPS. This is an acronym that stands for floating point operations per second. A floating point operation is a calculation involving very large or very small numbers with a decimal “floating” around in the calculation. Since they are very large or very small, the numbers used in a FLOP are expressed in scientific notation. Also, since supercomputers usually contain many CPU and GPU cores, the floating point operations per second are very high, which requires the SI prefixes (such as mega, giga, tera, etc…). So 1 TFLOPS means one TeraFLOPS (1×10^{12} FLOPS).

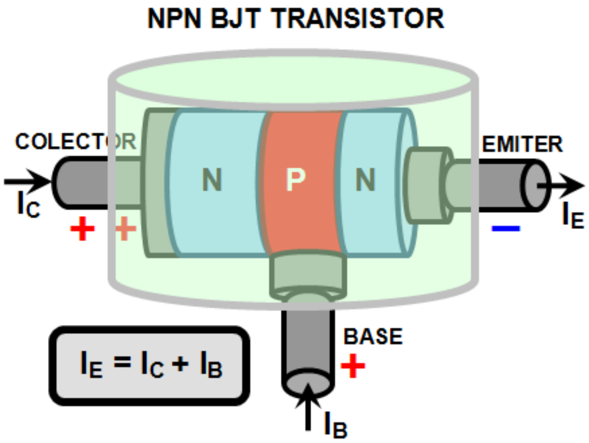
Every device that performs calculations can be measured in FLOPS. For example, a simple calculator runs 10 FLOPS (Wikipedia/FLOPS). An iPhone 4 runs 33.35 MFLOPS (TheBestHandPhone.com, 2011). According to Microsoft, an Xbox 360 gaming console can produce 1 TFLOPS, though many find this to be an erroneous overestimation (DeBoer, 2009). The University of Arkansas' own supercomputer, the Star of Arkansas, has a theoretical peak performance of 13.36 TFLOPS (engr.uark.edu, 2008). Currently, the world's fastest supercomputer is the Chinese-made Tianhe-1A. Inside this 155,000 kilogram behemoth are 14,000 Intel CPUs working with 7,000 Nvidia GPUs to produce a total of 2.5 PFLOPS. To explain just how extremely fast that is, 2.5 PFLOPS means the computer can carry out 2,500 *trillion* calculations per second. Now that is fast! Currently, this astounding machine is being put to use by the local weather service and the National Offshore Oil Corporation (BBC.com, 2010). One device that is highly debated in terms of how many FLOPS it can perform is the human brain. Some argue that the human brain runs 10 PFLOPS (Orca, 2010). The argument is based on the fact that the brain not only has to run mathematical calculations, but also process sensual things such as smell and sight. Others contend that the human brain is very slow in terms of FLOPS due to the average person's to do calculations by hand. Further criticism states that the human brain cannot be measured in FLOPS at all since it is a biological system as opposed to the being a computer system.

With all this discussion on computer speed and performance, what is next for the world of CPUs? Of course, engineers will still be piling more transistors onto CPUs, but also, they are seeking to make the chips smaller. Smaller chip size is important because it means less power consumption. For Intel, that is a main goal of theirs. When Brian Krzanich, a senior vice president and general manager for manufacturing and supply chain at Intel, was asked in an interview in October 2010 about advancing to fifteen nanometer-sized chips, he stated, "It's absolutely in the labs." (Crothers). Krzanich went on to say that "It's [designing CPUs] no longer how just fast [*sic*] you can do the quarter mile. Now, by integrating memory controllers and graphics you're going to see more and more memory on the parts…It's like going from a drag racer to a Porsche, not only does it go fast but it can turn quickly and stick to the roads. … [But] we're still adding cores." (Crothers). One interesting route for the future of CPU technology is changing how silicon is used in semiconductors, such as creating flexible CPUs (Hsu, 2010).



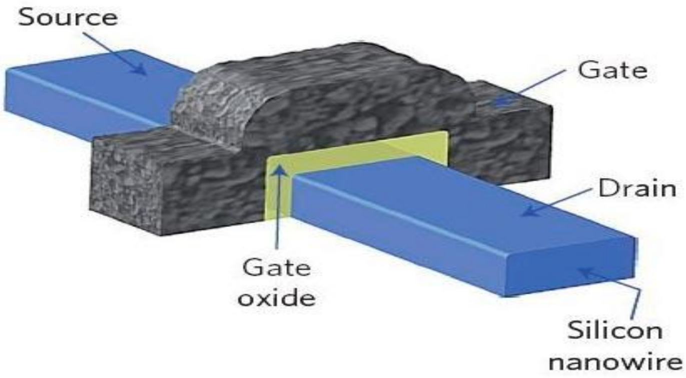
(**Figure 2** – Flexible silicon. **Source:** http://www.popsci.com/technology/article/2010-03/researchers-look-beyond-silicon-toward-computings-future)

Another great improvement that is being looked into by semiconductor companies is the junctionless transistor. Semiconductors in today's CPUs, still using the aforementioned doping method, have junctions. In a junction (pictured on page 8), a current enters the collector, while another current is pumped through the base. The more current pumped through the base means more current will be amplified out of the emitter.



(**Figure 3** – a basic diagram of a transistor.
Source: http://www.corollarytheorems.com/Design/ed_images/transistor_g_1.png)

The problem with this type of transistor is that current easily leaks. However, with the junctionless transistor developed at the Tyndall National Institute, a small silicon wire only a few atoms in diameter can carry a current, but can be controlled by "A component nicknamed the 'wedding ring' [that] regulates the flow of current by electrically 'squeezing' the wire to stop the electron flow…" (Dillow, 2010). The junctionless transistor is cheaper and leaks much less current than a regular transistor. This is a relatively new technology that has been developed, so it is still being researched by top CPU companies.



(**Figure 4** – a diagram of a junctionless transistor.
Source: http://physicsworld.com/cws/article/news/41881)

When discussing the future of CPUs, Moore’s Law has to come up. This law was popularized by Gordon Moore’s paper “Cramming More Components Onto Integrated Circuits.” In the paper, he plotted the $\log_2$ of the number of components per integrated circuits versus the time in years, and his graph showed a trend of that number doubling every two years (Moore 1965, 84).

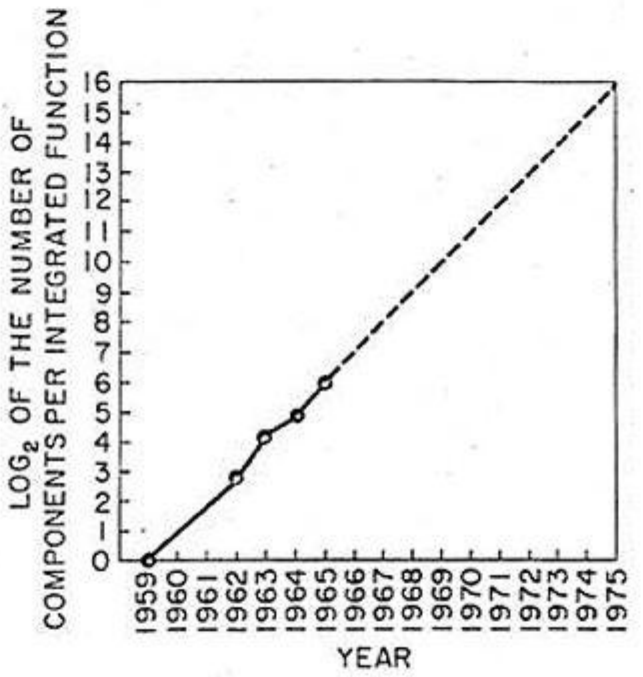
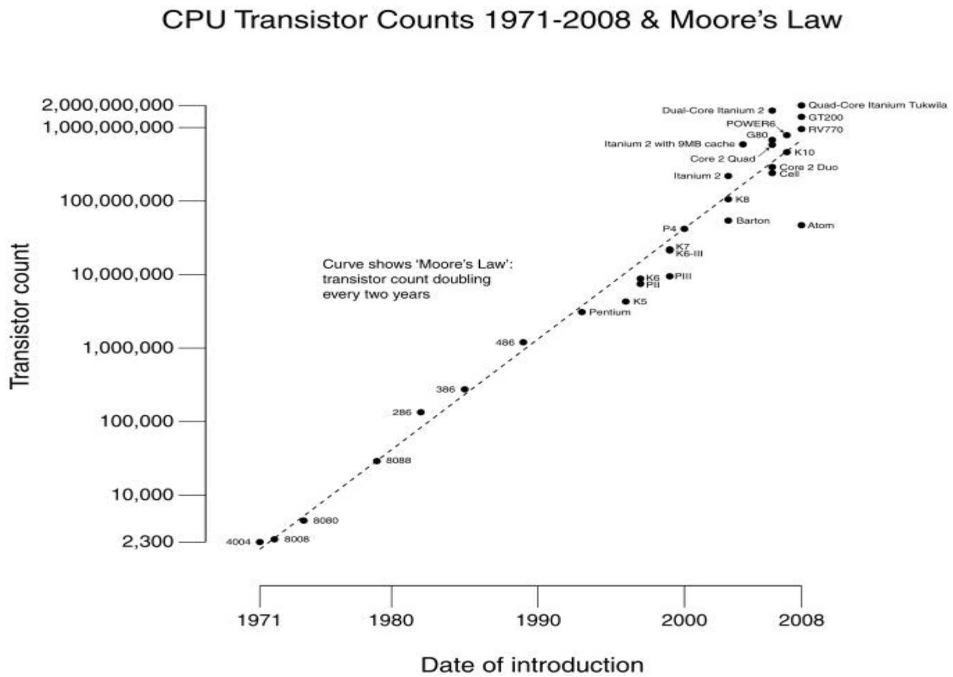


Fig. 2 Number of components per Integrated function for minimum cost per component extrapolated vs time.

(**Figure 5:** Moore’s original plot.
Source: http://www.intel.com/pressroom/kits/events/moores_law_40th/Images_Assets/graph.jpg)

While the law is commonly understood as stating that the number of transistors doubles every eighteen months, Moore himself set the record straight in an interview with Intel. “I think it was Dave House, who used to work here at Intel, did that, [*sic*] he decided that the complexity was doubling every two years and the transistors were getting faster, that computer performance was going to double every 18 months…but that’s what got on Intel’s Website…and everything else. I never said 18 months that’s [*sic*] the way it often gets quoted.” (Intel Corporation, 2005).

Originally, he predicted the trend would continue for the next ten years. What makes the law famous is that the number of transistors is still doubling forty-five years later!



(**Figure 6** – An updated version of Moore's original plot. The points labeled are processors. **Source:** http://3.bp.blogspot.com/-y16B1TXpuwQ/TWVpi4asjLI/AAAAAAAAAFQ/DKTYftwuuqo/s1600/683px-Transistor_Count_and_Moore%2527s_Law_-_2008.jpg)

If Moore's Law stays true for the next half decade, some scientists believe it could have serious implications. One such scientist is Rich Terrell, from the Jet Propulsion Laboratory in California. On the Science Channel's mini-series *Through the Wormhole with Morgan Freeman*, Terrell discusses the possibility of simulating the entire world and all the living beings in it. He says, "Right now, the fastest computers on the planet are now comparable or exceeding the computational ability of the human brain, as we estimate it. That's about one million-billion operations per second. ...In the next year, that'll double. In the next decade, that'll increase by a factor of five hundred. So a decade from now, our supercomputers will be about five hundred times faster than the human brain." Getting a bit more philosophical, Terrell speculates what will happen in fifty years. He imagines a laptop computer in fifty years placed inside of a box with a human brain. If he began to ask questions to the box and he was not able to tell which device was answering, one would have to agree the objects are qualitatively the same. “And if I believe that the human is conscious and self-aware, I must also believe that the machine has the same qualities,” he says (*Through the Wormhole*, 2010). This speaks volumes for Moore’s Law and just how important it is for the development of future CPUs. It will allow scientists and engineers to set goals and hypothesize about the future.

Realistically though, it does not seem that Moore’s Law will be able to continue indefinitely. Strangely enough, Gordon Moore himself believes it will come to an end. “It can’t continue forever,” he says. “In terms of [transistor] size you can see that we’re approaching the size of atoms which is a fundamental barrier, but it’ll be two or three generations before we get that far. …We have another 10 to 20 years before we reach a fundamental limit.” (Dubash, 2005). How exactly engineers plan to create transistors beyond the atomic level is yet to be seen.

The microprocessor industry is a very fast-growing one. While the technology is only fifty years old, how far it has advanced is astounding. As the number of transistors in CPUs continues to double with Moore’s Law, the speed will continue to increase. Also, with companies like Intel decreasing the size of chips, efficiency will increase, causing price per unit to decrease. Efficiency will also increase with new technologies such as the junctionless transistor. If the history of the CPU is any predictor, these new advancements are only the beginning in terms of what will be cutting-edge in the industry in only a few short years.

Bibliography

BBC News. “China claims supercomputer crown.” *BBC.com*, Last modified October 28, 2010. http://www.bbc.co.uk/news/technology-11644252.

Crothers, Brooke, “A peek into the future of Intel processors,” *Cnet.com*, October 19, 2010, accessed March 24, 2011, http://news.cnet.com/8301-13924_3-20020078-64.html.

DeBoer, Clint, “Playstation 3 vs. Xbox 360,” *Audioholics.com*, last modified July 23, 2009. http://www.audioholics.com/buying-guides/how-to-shop/playstation-3-vs-xbox-360.

Dubash, Manek, “Moore’s Law is dead, says Gordon Moore,” *TechWorld.com*, April 13, 2005. accessed March 24, 2011, http://news.techworld.com/operating-systems/3477/moores-law-is-dead-says-gordon-moore/.

Engr.uark.edu, “University of Arkansas Installing Supercomputer; ‘Star of Arkansas’ to Be State’s Fastest,” accessed March 24, 2011. http://www.engr.uark.edu/home/2378.php.

Hsu, Jeremy, “Researches Look Beyond Silicon Toward Computing’s Future,” March 26, 2010, accessed March 24, 2011. http://www.popsci.com/technology/article/2010-03/researchers-look-beyond-silicon-toward-computings-future.

Intel Corporation, *Excerpts from* A Conversation with Gordon Moore: Moore’s Law, 2005.

Moore, G.E., “Cramming More Components Onto Integrated Circuits,” *Proceedings of the IEE* 86 (1965): 84, accessed April 10, 2011, doi: 10.1109/JPROC.1998.658762.

Murph, Darren, “IBM claims world’s fastest processor with 5.2 GHz z196,” *Engadget.com*, September 6, 2010, accessed March 22, 2011. http://www.engadget.com/2010/09/06/ibm-claims-worlds-fastest-processor-with-5-2ghz-z196/.

Orca, Surfdaddy, “The Ultimate Connection Machine,” *Hplusmagazine.com*, March 5, 2010, accessed March 24, 2011. http://hplusmagazine.com/2010/03/05/ultimate-connection-machine/.

Simmons Jr. J., John, *From Sands to Circuits*. Cambridge, Massachusetts: Harvard University Office for Information Technology, 1986.

TechTarget.com, “parallel processing,” accessed March 24, 2010. http://searchdatacenter.techtarget.com/definition/parallel-processing.

Through the Wormhole with Morgan Freeman. Episode no. 1, “Is There a Creator?” First broadcast June 9 2010 by The Science Channel. Created by Geoffrey Sharp.

Torres, Gabriel. “How a CPU Works.” *Hardwaresecrets.com*, last modified September 26, 2005. http://www.hardwaresecrets.com/article/How-a-CPU-Works/209/1.

Wikipedia, “FLOPS,” accessed March 23, 2011. http://en.wikipedia.org/wiki/FLOPS.