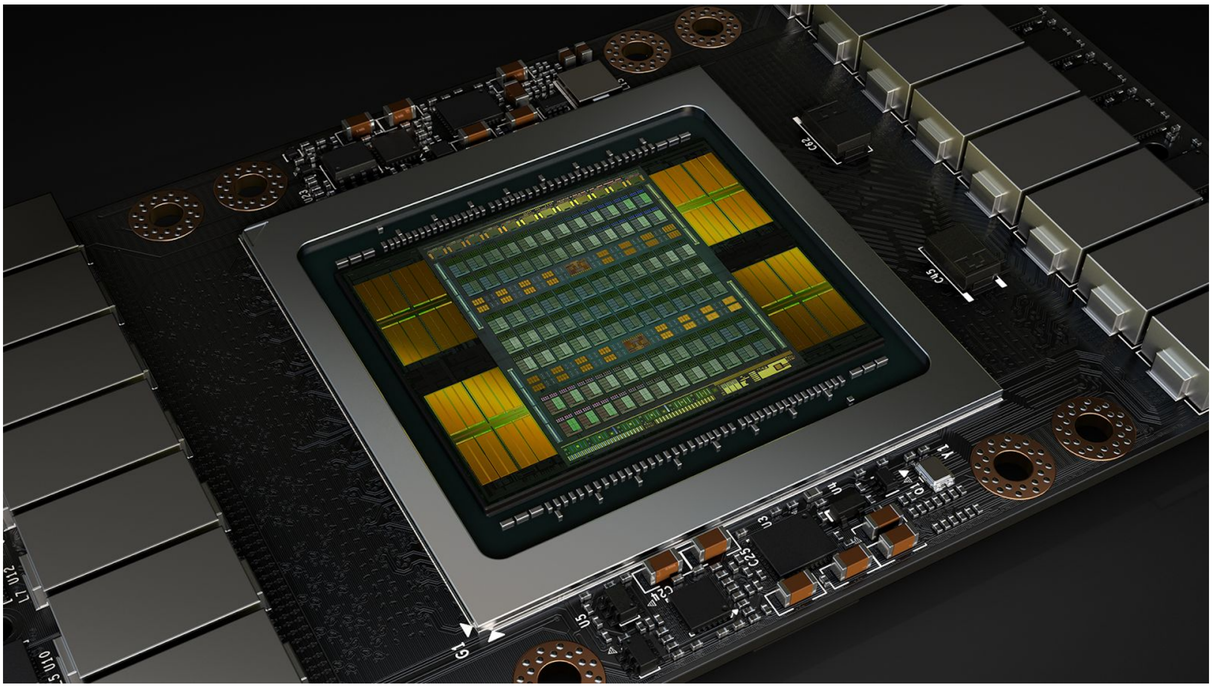
eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Advanced Techniques in Artificial Intelligence

The development in AI thanks to advancements in hardware



Iker García

Eritz Yerga

igarcia945@ikasle.ehu.eus
eyerga001@ikasle.ehu.eus

Index:

Abstract

In this report we will cover how the advancements made in hardware have permitted the advancements in Artificial Intelligence, Machine Learning, Neural Networks, Deep Learning…
We will cover the history of AI, the history of hardware advances related to this field and the type of hardware used as of today in AI.

Introduction

When the field of AI began the future advancements on it were not fully realised at all times, cycles of high expectation and disappointment were often created when researchers came up with different ideas for AI.
Herbert Simon predicted in 1957 that in ten years a computer would be a champion in chess or that it would be possible to demonstrate an important mathematical theory with a machine. Those predictions, however, were at some degree wrong, as that kind of development wouldn't happen in ten years but forty. That excess of confidence in Simon's prediction happened because the first AI systems had a really good performance in simple problems and that led to high expectations. [0]

The reality was problems started to emerge when more different or more complex problems were faced. The knowledge these systems had about the subject of study wasn't enough to face more than really simple problems and in reality the amount of knowledge needed to solve those more complex problems was a lot higher.

Most of the problems that were trying to be solved using AI were also unfeasible to be solved the way they were trying to. The focus on that moment was to try to solve the problem with different experimentation, trying diverse steps until the solution was reached. This worked for problems that dealt with a small amount of objects, at that time hardware was also expensive and not very powerful which caused complex problems to be impossible to solve in reasonable time.

The impossibility to handle combinatory explosion was one of the main critique AI received at that time, this caused that a lot of pioneering universities in this field stopped receiving funds for research on AI. This all led to the decrease in interest and expectations in AI until time passed on. [0]

AI First Failures

First years of the Artificial Intelligence development were full of success. In the 50s computers were still considered instruments for arithmetic calculations and nothing more. Scientific community believed that a computer could never perform task as humans do. Artificial Intelligence researchers prove them wrong, in 1959 Herbert A. Simon, J.C. Shaw, and Allen Newell developed the General Problem Solver or G.P.S. was a computer program that imitated the human protocols for problem resolving. Herbert Gelernter also in 1959 developed a geometry problem demonstrator. In 1952 Arthur Samuel wrote a serie of programs that learned to play checkers game with an amateur level, this also proved wrong the idea that a computer could only do what they are said to do. It didn't take long until the program learned to play better than his creator, when the program was presented in the television in 1956 made a big impression.

February 24, 1956. Arthur Samuel demonstration of his program in an IBM 701 computer

This early success generated great optimism about the Artificial Intelligence development. For example, in 1957 Herbert Simon commented:
“ I’m not trying to impress you or leave you amazed, but the easiest way I have to sum it up is to say that at the moment machines capable of thinking, learning and creating exist in the world. Furthermore, their ability to do that will quickly increase until (in a near future) the magnitude of problems they will be able to solve will be on par with human mind itself on the same tasks. ” [0]
But between 1966 and 1973 their beliefs crashed with the reality. Their algorithms worked for problems that dealt with a small amount of objects, but when they tried to use that methods with bigger problems they failed strongly. For example the government of US founded a project to develop a program for translating Russian scientific documents. The program translated “the spirit is strong but the flesh is weak” as “the vodka is good but the flesh is rotten”, the problem was that for a good translation you need to know the context of the sentence, but that translator was based in syntax transformations and word placement with the help of a digital dictionary. The report of the project concluded that they didn't succeed in getting any translation of scientific texts and they didn't expect to advance in the short term. [0]

The development of the Artificial Intelligence has been cycling like this since the early beginning. A small amount of successes fire up expectancies but afterwards the reality had to be faced.

However, that cycle has recently been broken.

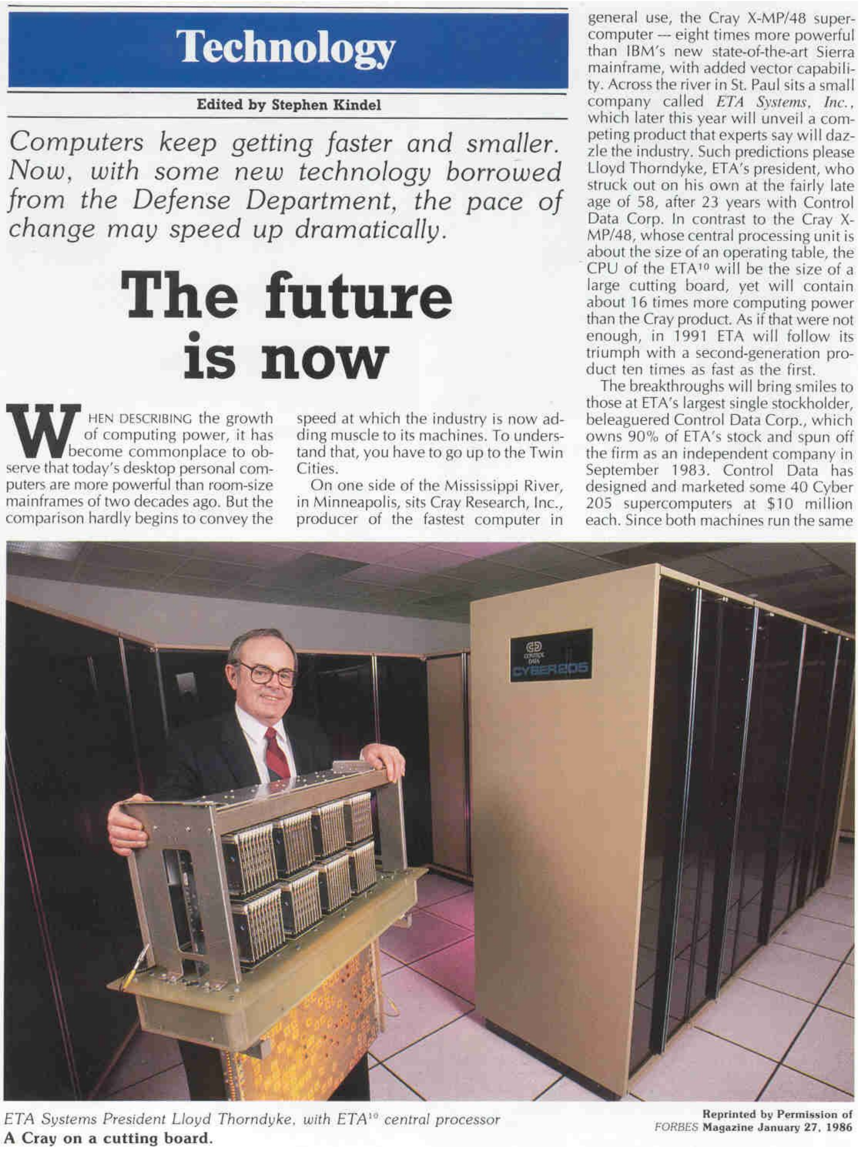
AI is skyrocketing. Why now?

AI development as of today is having a bigger impact as never before because of a lot of advances in various fields of technology. Most influential advances are cheaper computing power, easier access to diverse digitalized contents and more sophisticated algorithms and methodologies.

Not only the hardware is more capable of doing more computations per second today [*], but also modern manufacturing processes make this hardware accessible to almost anyone who wants to buy it due to costs being greatly cut down. The difference in processing power today compared to not so much time ago is breathtaking. For example: The MOS 6502 CPU found inside the Apple II in 1977 was capable of 500000 flops and that computer was US$1298 in its most cheapest variants [1], today you can use a AMD Ryzen 7 1800X CPU capable of 84600000000 flops (169 200 times more floating point operations per seconds) for just US$400 (US$800 if you want a whole operable system) [2].

But what really impacts are the advancements in GPUs. GPUs were designed to display graphical interfaces on monitors. The development of the GPUs was greatly influenced since 1999 by videogames, which requires a lot of computing power to simulate 3D environments, and made them have a key special design that is very different compared to CPUs. Contrary to CPUs which are designed to perform consecutive instructions and are composed by a small amount of cores, GPUs make use of a lot (thousands) of cores designed to perform parallel calculations. This gives them a high computational power. For instance, the Nvidia GTX 1080 Ti is capable of 11000000000000 flops (approximately 130 times more than the AMD Ryzen 7 1800X, one of the most powerful CPUs as of today) [3]. This attribute makes them the best option for diverse computational tasks, including AI.

One of the best examples that illustrate the jump in processing power since the beginning of AI to today is the difference between supercomputers of that era vs what we have today. The ETA10 made in early 1987 had 10 GFLOPS of computing power and needed liquid nitrogen to cool its CPU, the high cost of the computer coupled with the high electricity consumption and the difficulty of cooling it makes it only accessible to few universities and institutions such as Florida State University or Johnson Space Center [4] [5]. Today a US$110 Nvidia GTX 1050 has 1.8 TFLOPS of computing power and It has a power consumption lower than 75 watts, this means that it doesn't even need a fan for cooling it, can be cooled passively [6] [3]. This makes this GPU accessible to almost everyone, high computational power is no longer reserved for few universities and research centers, everyone can have access to it.



Technology

Edited by Stephen Kindel

Computers keep getting faster and smaller. Now, with some new technology borrowed from the Defense Department, the pace of change may speed up dramatically.

The future is now

When describing the growth of computing power, it has become commonplace to observe that today's desktop personal computers are more powerful than room-size mainframes of two decades ago. But the comparison hardly begins to convey the speed at which the industry is now adding muscle to its machines. To understand that, you have to go up to the Twin Cities.

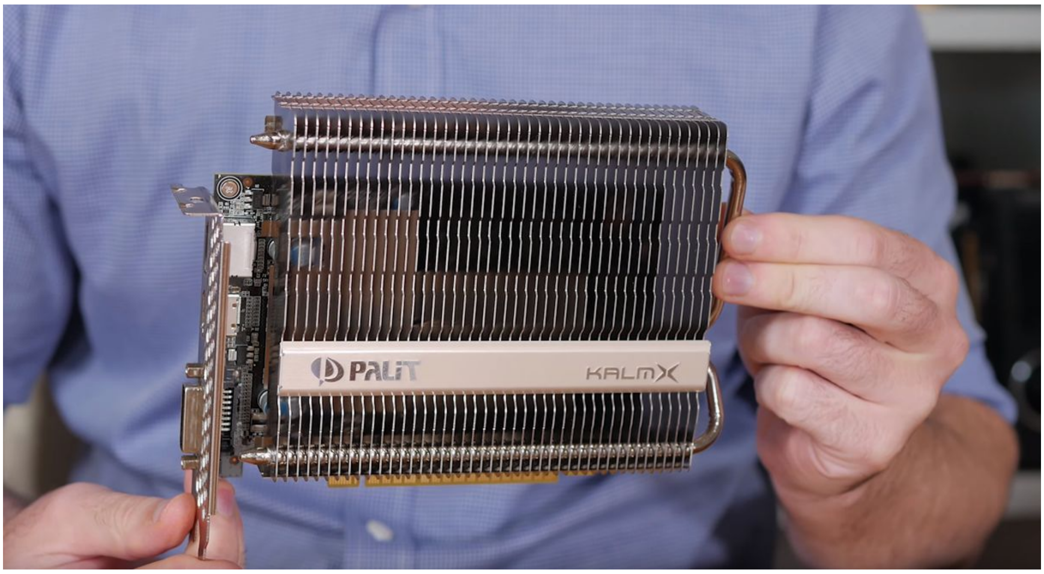
On one side of the Mississippi River, in Minneapolis, sits Cray Research, Inc., producer of the fastest computer in general use, the Cray X-MP/48 supercomputer — eight times more powerful than IBM's new state-of-the-art Sierra mainframe, with added vector capability. Across the river in St. Paul sits a small company called *ETA Systems, Inc.*, which later this year will unveil a competing product that experts say will dazzle the industry. Such predictions please Lloyd Thorndyke, ETA's president, who struck out on his own at the fairly late age of 58, after 23 years with Control Data Corp. In contrast to the Cray X-MP/48, whose central processing unit is about the size of an operating table, the CPU of the ETA^{10} will be the size of a large cutting board, yet will contain about 16 times more computing power than the Cray product. As if that were not enough, in 1991 ETA will follow its triumph with a second-generation product ten times as fast as the first.

The breakthroughs will bring smiles to those at ETA's largest single stockholder, beleaguered Control Data Corp., which owns 90% of ETA's stock and spun off the firm as an independent company in September 1983. Control Data has designed and marketed some 40 Cyber 205 supercomputers at $10 million each. Since both machines run the same

ETA Systems President Lloyd Thorndyke, with ETA^{10} central processor
A Cray on a cutting board.

Reprinted by Permission of FORBES Magazine January 27, 1986

Forbes article about the ETA10, Juanary 27, 1986



[7] NVIDIA GTX 1050 with passive cooler. Hardware unboxed. February 23, 2017

The boom of internet and the tendency to digitize all systems has made it easy to have really big databases and datasets to train systems in machine learning, deep learning and AI.

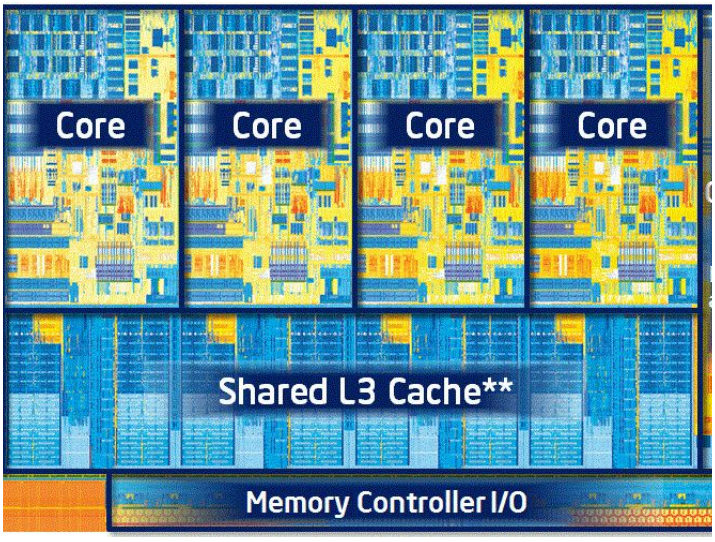
Finally, the development of machine learning and deep learning has gained a great importance in current businesses, becoming a hot topic in computer science. This programming paradigm and methodologies has changed the way professional environments focus on AI, and its utilities make really big corporations spend a lot on research as they try to exploit the potentials these algorithms give them.

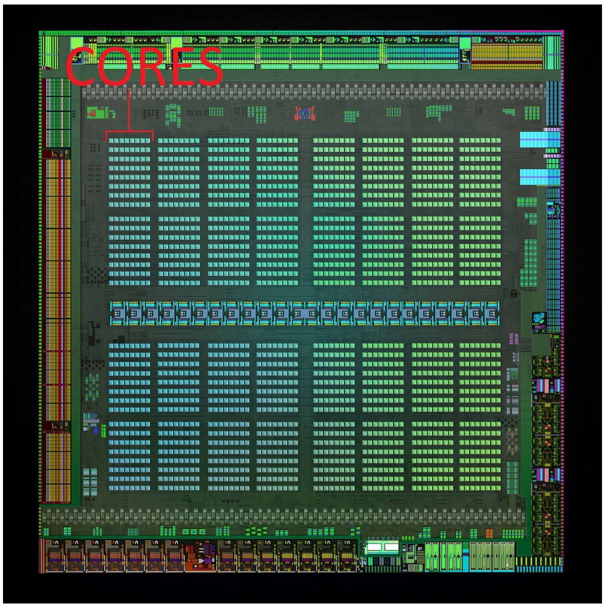
Why GPUs?

For years CPUs were the kings. They were used for everything, including computation and of course Artificial Intelligence. The advance of the technology was governed by the advance in CPUs. But recently, that has changed, now the focus is on GPUs.

In the beginning GPUs were not designed with artificial intelligence in mind, the first GPUs were chips designed to show interfaces, freeing up the main processor from this task. More advanced chips were used to create some of the first consoles, such as Atari 800, a computer designed for video games. With this machines video games started to become more and more popular. Video Games soon encountered the limitations of CPUs, CPUs were designed with serial computing in mind, CPU cores are intended to be as powerful as possible and to operate at the highest possible frequencies, the ultimate goal of CPUs is to achieve the highest number of instructions per second or I.P.C. But if you want to simulate a 3D environment you will need to perform millions of operations, vertex transformations, lightning calculations… It doesn’t matter how high the IPC of you CPU is, it isn’t enough [8].

That’s why Nvidia developed the GeForce 256, considered the “first GPU”, a term Nvidia defined at the time as “a single-chip processor with integrated transform, lighting, triangle setup/clipping, and rendering engines that is capable of processing a minimum of 10 million polygons per second." [9]. Since the GeForce 256 GPUs has followed the opposite path of CPUs, while modern CPUs consists of a maximum of 32 cores, GPUs have up to 5376 cores. The cores inside a GPU are smaller than the cores inside a CPU, this make them less powerful, also more cores implies more heat and power consumption, so they need to work at lower frequencies, modern intel processors can work up to 4,7Ghz while the fastest Nvidia GPU works at 1,7Ghz.

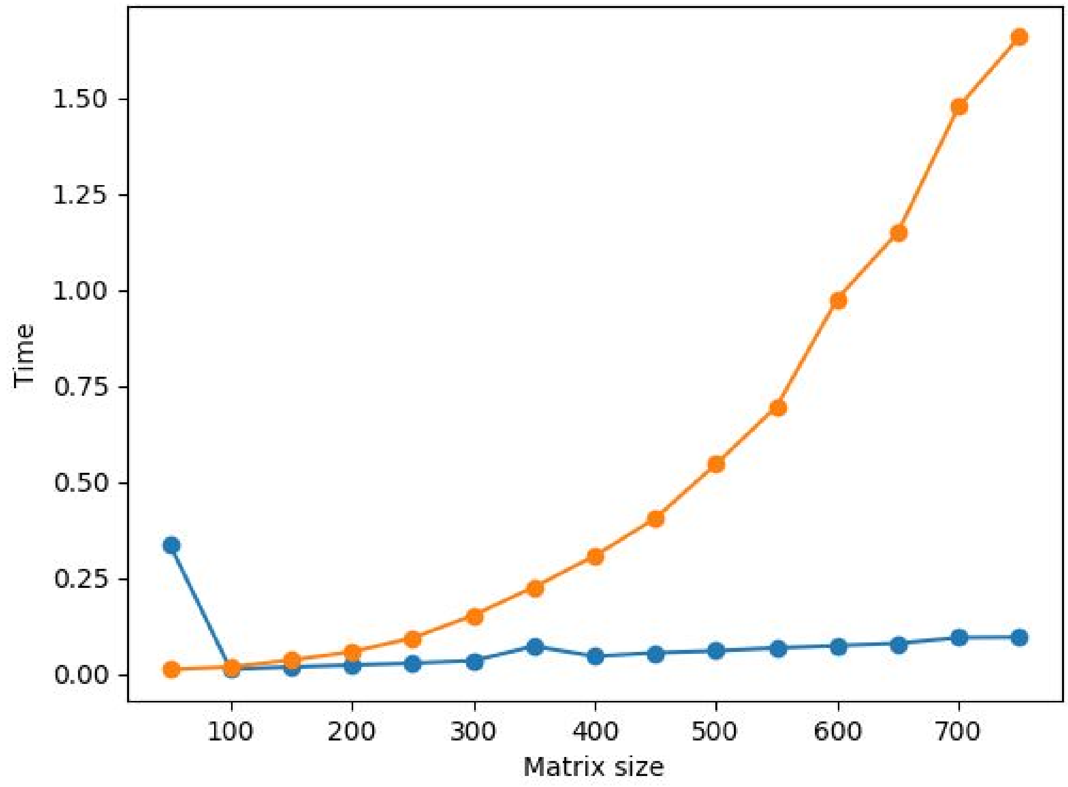




Intel i7. (4 Cores)

Nvidia GTX 1080 (2560 Cores)

But even if the cores inside a CPU are faster than the cores inside a GPU, when you have tasks that can run in parallel, the high amount of cores allow GPUs to be much faster than CPUs [10] [11].
For testing this affirmation we used a simple benchmark, a program written in python that uses TensorFlow for perform the multiplications of 2 matrix [12]. Matrix multiplication is one of the most easily parallelizable task, because each position of the result matrix does not depend of the result of other positions. We tested the Intel Core i5 4690K a 4 core CPU at 4GHz against the GTX 970 a 1664 core GPU at 1,350GHz. The result clearly shows that witch big matrices, the GPU is much faster than the CPU. The GTX 970 can perform much more operations like these per second than the i5 4690K.



🟧 Intel Core i5 4690K

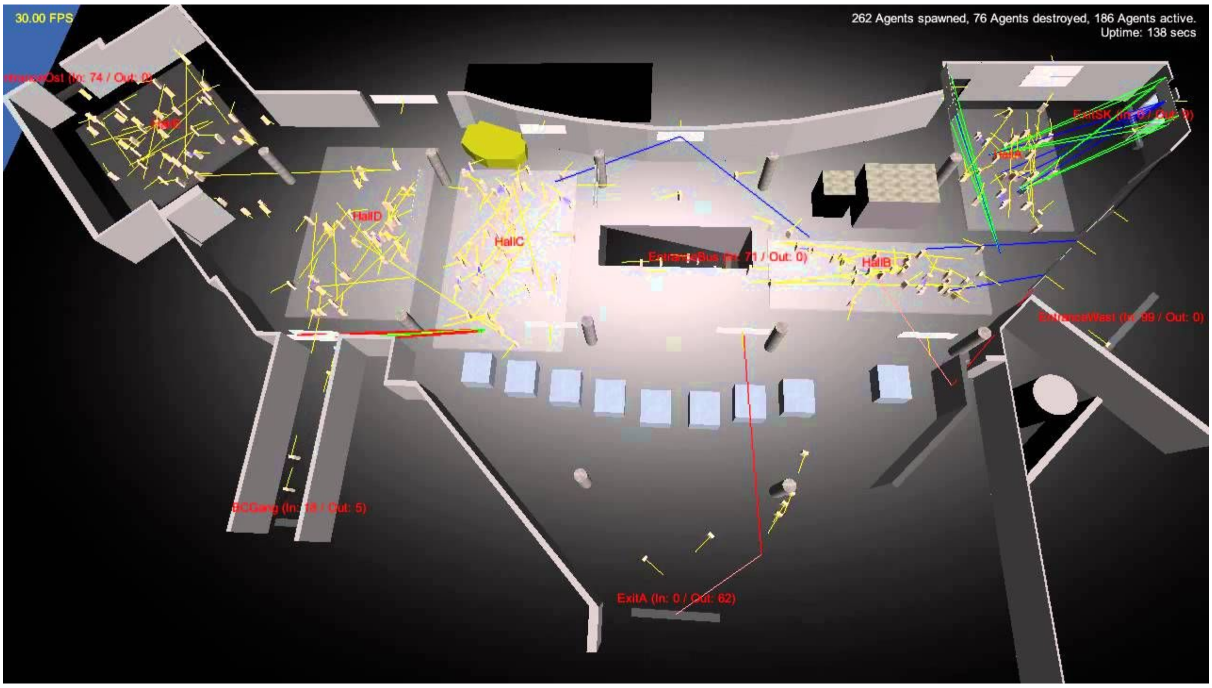
🟦 Nvidia GTX 970

CPU vs GPU on matrix multiplication operations.

We have proved that when parallel computing is possible, GPU are much faster than CPUs, now were are going to focus on artificial intelligence, and why a chips initially designed for video games now drive the world of the Artificial intelligence.

Multi agent AI systems are often used to do rich complex simulations, these systems instead of having a unique AI agent that performs the simulation integrate a lot of agents, one for each being involved in the simulation. An agent is defined as “Autonomous agents are computational systems that inhabit some complex dynamic environment, sense and act autonomously in this environment, and by doing so realize a set of goals or tasks for which they are deigned”. [13]

Some examples of simulations that are made with these systems are: city traffic simulations, airport passenger simulations, consumer good distribution simulations, electricity transportation simulations, business process simulations, … [14]



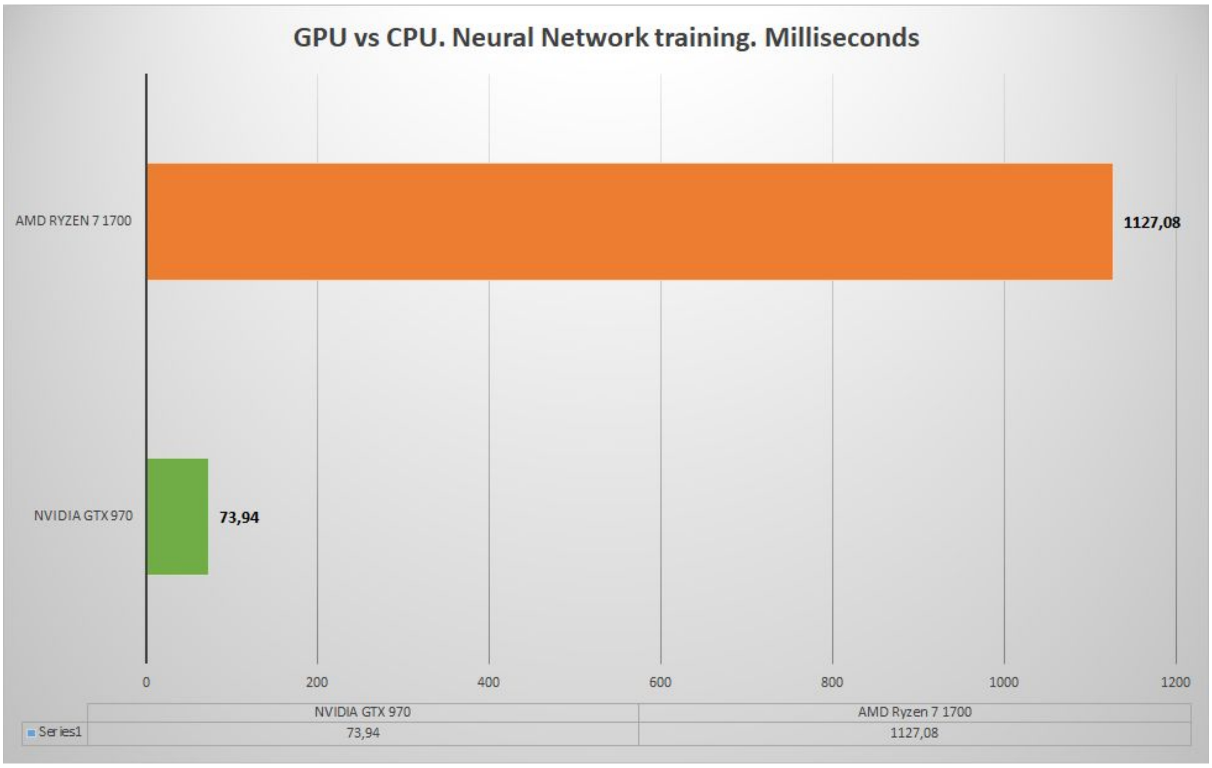
[15] Example of airport simulation using multi agent systems.

It is often more valuable to be able to run the most agents as possible all together, since most CPUs don't have as much threads needed to run complex multi agent simulations it is often prefered to use the GPUs despite their lower single-thread performance because with their higher clock count it'd be able to run a really high number of agents in real time. Counting the fact that most modern computers are compatible with having more than one GPU, it is possible to run a lot more of agents on real time. It doesn't matter how fast a CPU can process each agent, a GPU processing thousands of agents at the same time will be faster.

Today "Deep" is the buzzword of the Artificial intelligence, deep learning and neural networks has become recently one of the hot topics in AI. Deep learning involves a huge amount of matrix multiplications, and as we seen before, GPUs are much faster than CPUs for this task. But the advantages of GPUs doesn't end here, CPUs are designed for general purpose and cache memories and the RAM memories that it uses are optimized for low latencies, while GPUs are optimized for high bandwidth. Usually modern RAM memories can achieve a bandwidth of 50GB/s while GPUs VRAM memories can achieve up to 900GB/s. When you need to access a single or very few values from memory a CPU will be able to access them faster, but when you are working with neural networks and huge amount of data like big matrixes, a GPU would be able to access faster to all the information. [16] This is due to the CPUs not being able to fetch big amounts of "packages" of information, unlike GPUs. The CPUs can get the information faster but it cannot drive a lot of information at the same time, meanwhile the GPUs are slower to get the information but have a lot of lanes to drive higher amount of information at the same time.This not made GPU well suited only for neural networks, it also makes them really fast for any task that involves high bandwidth usage, such as data mining.

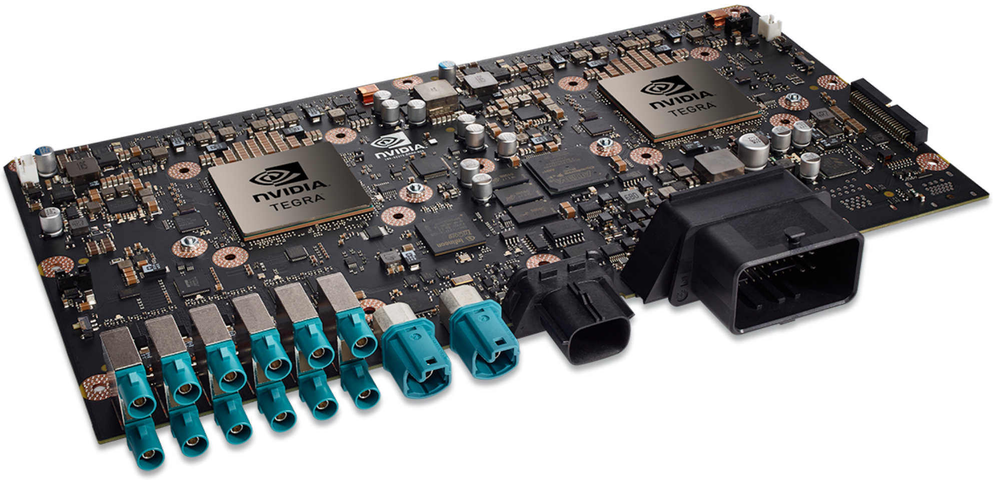
We also tested this. We used one of the sample benchmark from TensorFlow, the MNIST convolutional model [17]. This benchmark trains a convolutional neural network with the MNIST dataset. We tested the Ryzen 7 1700 a 8 cores / 16 threads CPU at 3,2Ghz from AMD, and the NVIDIA GTX 970 a 1664 core GPU at 1,350GHz. The Ryzen 7 1700 is considered one of the faster consumer CPU, was released in 2017, and it costs around US$320, While GTX 970 is a midrange GPU released in 2014 (3 years before the Ryzen 7) at a price of US$329.

As we can see in the results, the GPU is 15.24 times faster than the CPU, this means that we can train more complex neural networks with a GPU or use this advantage to solve a large number of problems in which the same amount of time.



As we have seen, evolution of artificial intelligence has been influenced by hardware. In the past only few people that worked in universities and research centers could have access to computers capable of execute artificial intelligence algorithms in a reasonable time, but that computers were very limited, and could only run simple algorithms. Modern hardware, especially GPUs allows to execute much more complex algorithms and deal with a much larger amount of information. They have also allowed everyone to have access to high computational power.

But this progress has also allowed many things impossible before, for example, today is possible to equip a car with a GPU powerful enough for autonomous driving. This completely changes the world of the Artificial Intelligence because now they are a lot of real applications for AI that can generate a lot of money, thus being reasons for companies to invest money in the development of AI. This is the main reason why Artificial intelligence is skyrocketing.

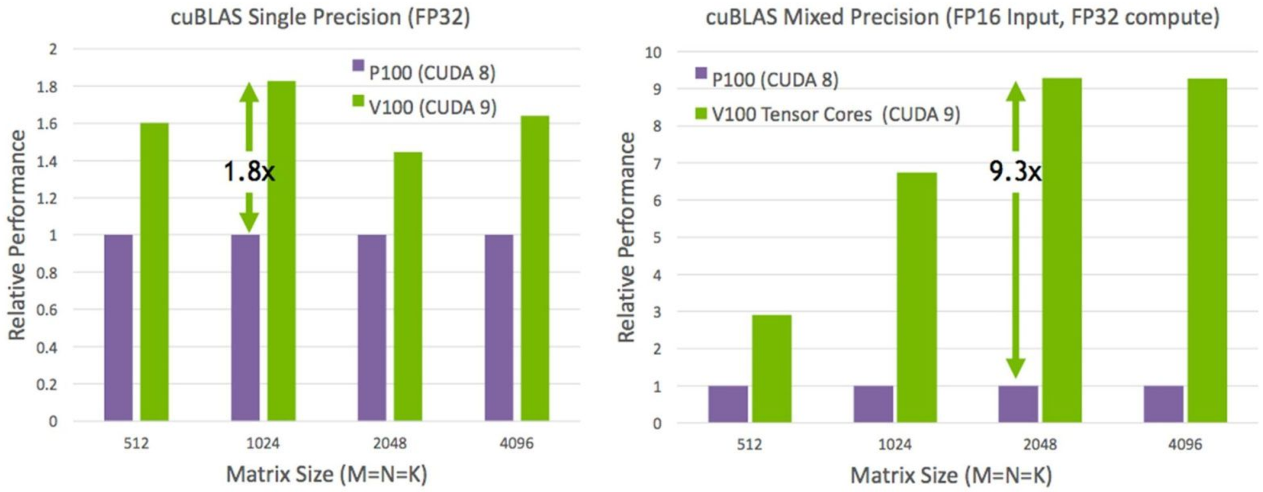


NVIDIA parker, a computer for autonomous cars with a 256 core GPU

A vision of the future

Companies are investing important amount of money in Artificial Intelligence right now. It is a world that is expected to report great benefits, that's why Google, Amazon, Facebook… are investing on it, and AI has become one of their priorities.
Investing in Artificial intelligence means that this companies are building huge servers, or that they are buying hardware to mount it in autonomous driving. That's why for the biggest hardware companies Artificial Intelligence is also one of their priorities. There is a huge competition between Nvidia, AMD, intel, IBM and other smaller companies to develop the best GPUs and CPUs for Artificial Intelligence.

Nvidia has recently announced their new GPU architecture, Volta, that introduced dedicated hardware for Artificial intelligence, the Tensor Cores. The first GPU based in Volta architecture is the GV100, and thanks to Tensor Cores this GPU can perform an incredible amount of calculations per second in 16 bit precision (used in neural networks). A GV100 is as fast as 10 GP100 GPUs, It's predecessor that was released a year ago.



Matrix multiplication comparison between Nvidia P100 and NVIDIA V100

Two super computer are being created around this GPU which is the most advanced GPUs for Artificial Intelligence in the market.
The Nvidia DGX-1 is equipped with 8 Tesla V100 and is capable of 1 Petaflop, this computer uses a new connection between the GPUs NVlink that offers a higher bandwidth that the traditional PCIe 3.0 [18].

Nvidia DGX-1

IBM released “Minsky” last year, and become a very popular super computer, and now their are developing its successor that will compete which the Nvidia DGX-1. It will also use Tesla V100 GPUs.

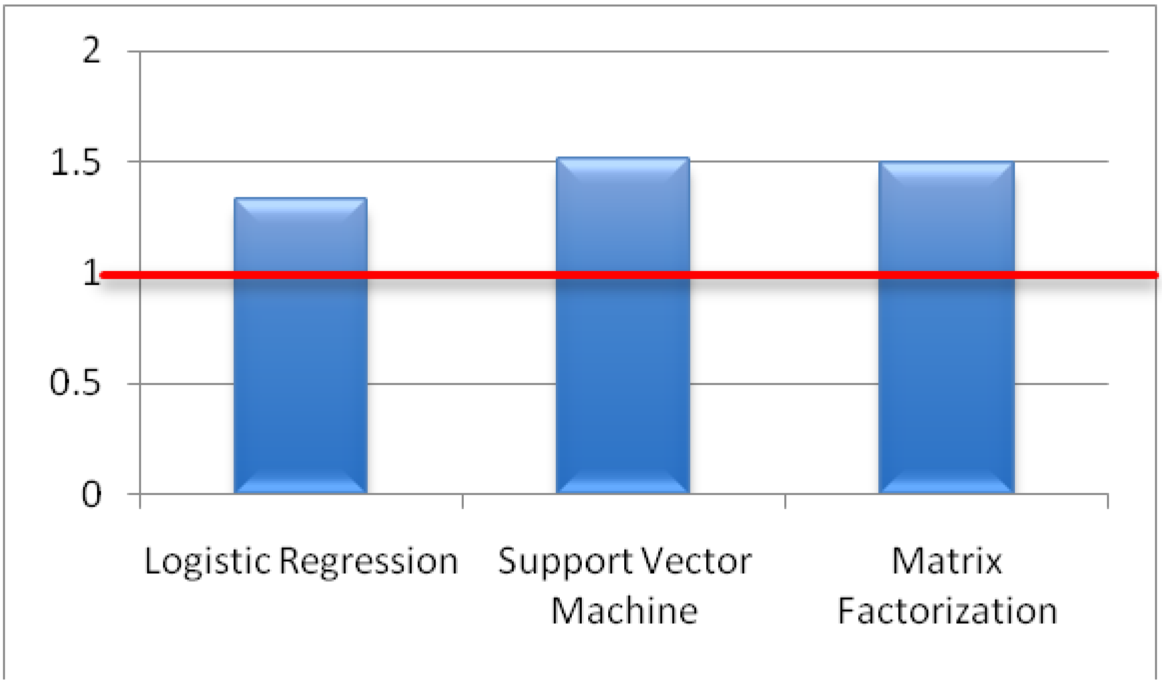
IBM Minsky

The main difference between this supercomputers is the architecture of the CPUs. DGX-1 uses 2 Intel Xeon ES-2698 v4, a 20 core CPU at 2,2Ghz, based x86 architecture. While the IBM computer will use CPUs based in Power9 architecture.

The x86 architecture, designed by Intel, is a family of backward compatible instruction set architectures based on the Intel 8086 CPU and its Intel 8088 variant. The whole architecture is designed with backwards compatibility in mind, adding new instruction sets through extensions. So, it has compatibility with 16-bit, 32-bit and 64-bit instructions. It is widely used in personal computers, laptops and even most of servers as of today, most powerful x86 CPUs a total of 64 threads [19].

The POWER architecture on the other hand is a reduced instruction set computing (RISC) instruction set architecture designed mainly by IBM and with its designed available for license under the OpenPOWER Foundation. Each new iteration of the architecture introduces a new version of the instruction set and it isn't necessarily backwards compatible. They are designed to maintain the best efficiency possible on the workloads they are doing. The current version of the architecture is POWER8, POWER9 being currently in development. Most powerful POWER8 CPUs a total of 96 threads, higher than x86 CPUs [20].

Both architectures support DDR4 RAM with ECC and PCIExpress, so the main differences are usually on the CPUs itself. x86 CPUs have better single-threaded performance but when it comes to multi-threaded workloads POWER CPUs with their higher thread count get an edge and with their better efficiency focus are getting a really important position on workloads like these.



Power Systems S822L [24 cores 192 threads] (Power8) vs Intel Xeon E5-2690 V3 [24 cores 48 threads] (x86). 256 GB RAM.

References and useful links

[0] Artificial Intelligence: A Modern Approach. By Stuart Russell and Peter Norvig.
http://aima.cs.berkeley.edu/

[*] Instruction per second calculation:
https://en.wikipedia.org/wiki/Instructions_per_second

[1] Apple II:
https://en.wikipedia.org/wiki/Apple_II_series

[2] AMD Ryzen 7 1800x:
http://www.amd.com/en/products/cpu/amd-ryzen-7-1800x

[3] GeForce 10 Series spec sheet:
https://en.wikipedia.org/wiki/GeForce_10_series#GeForce_10_.2810xx.29_series

[4] The ETA10 Supercomputer System. Charles D. Swanson
https://link.springer.com/chapter/10.1007/978-3-642-83221-5_3

[5] ETA10 Vector supercomputer:
https://en.wikipedia.org/wiki/ETA10

[6] Palit passively cooled Nvidia GTX 1050. Anandtech.
https://www.anandtech.com/show/11106/palit-announces-kalmx-passively-cooled-geforce-gtx-1050-ti-graphics-card

[7] NVIDIA GTX 1050 with passive cooler. Hardware unboxed.
https://www.youtube.com/watch?v=rYtG3InRHI0

[8] Graphics processing unit.
https://en.wikipedia.org/wiki/Graphics_processing_unit

[9] Nvidia GeForce 256.
https://en.wikipedia.org/wiki/GeForce_256

[10] GPU Accelerated computing. Nvidia.
http://www.nvidia.com/object/what-is-gpu-computing.html

[11] General-purpose computing on graphics processing units.
https://en.wikipedia.org/wiki/General-purpose_computing_on_graphics_processing_units

[12] Introduction to TensorFlow - CPU vs GPU. By Erik Hallström:
https://medium.com/@erikhallstrm/hello-world-tensorflow-649b15aed18c

[13] P. Maess, “Artificial life meets entertainment: Life like autonomous agents”, Communications of the ACM, vol. 38, no. 11, pp. 108-114, 1995.
http://web.media.mit.edu/~pattie/CACM-95/alife-cacm95.html

[14] Applications of Multi-Agent Systems. Mihaela Oprea. University of Ploiesti, Department of Informatics, Bd. Bucuresti Nr. 39, Ploiesti, Romania.
https://link.springer.com/content/pdf/10.1007/1-4020-8159-6_9.pdf

[15] Multi-Agent System with Unity3D for airport passenger simulation video by Christian Becker-Asano. Programmed by Felix Ruzzoli.
https://www.youtube.com/watch?v=uVpN136q7N8

[16] Why are GPUs well-suited to deep learning. Multiple Sources:
https://www.quora.com/Why-are-GPUs-well-suited-to-deep-learning

[17] GitHub TensorFlow models repository. Convolutional MNIST model benchmark:
https://github.com/tensorflow/models/blob/master/tutorials/image/mnist/convolutional.py

[18] Nvidia DX-1 Spec Sheet
https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/dgx-1/dgx-1-ai-supercomputer-datasheet-v4.pdf

[19] x86 CPU Architecture
https://en.wikipedia.org/wiki/X86

[20] POWER CPU Architecture
https://en.wikipedia.org/wiki/Power_Architecture

Other used and useful references:

The AI Resurgence: Why now? Badak Hoojat:
https://www.wired.com/insights/2015/03/ai-resurgence-now/

How Nvidia went from powering Video Games to Revolutionizing Artificial Intelligence. Forbes, Aaron Tilley:
https://www.forbes.com/sites/aarontilley/2016/11/30/nvidia-deep-learning-ai-intel/#4d0b7d247ff1

Nvidia reinvents the GPU for artificial Intelligence. Forbes. Jim McGregor:
https://www.forbes.com/sites/tiriasresearch/2016/04/05/nvidia-reinvents-the-gpu-for-artificial-intelligence-ai/#7eccb9472a3b

Why graphic card are hacking the future. Fred Benenson
https://medium.com/@fredbenenson/why-graphics-cards-are-hacking-the-future-390262edc247

Computer development vs Car development. Jason Torchinsky:
https://jalopnik.com/heres-what-cars-would-be-like-if-they-advanced-at-the-p-1791938679

Reasons why artificial intelligence is skyrocketing Huy Nguyen Trieu
http://www.disruptivefinance.co.uk/2015/03/18/6-reasons-why-artificial-intelligence-is-skyrocketing-and-will-grow-even-faster/