International Journal of Computer Science and Information Technology for Education
Vol.5, No.1 (2020), pp.31-38
http://dx.doi.org/10.21742/ijcsite.2020.5.1.04

Comparative Study on CPU, GPU and TPU

P Siva Raj[1] and Ch. Sekhar[2]

Dept. of Comp. Science & Engineering, Vignan's IIT (Autonomous), AP, India
[1]shivarajpadala@gmail.com, [2]sekhar1203@gmail.com

Abstract

The recent trend in the hardware technology changed as per need of the processing of huge amount of data receiving in the form of text, image, videos and graphical data. In this paper, we are going to discuss the evolutionary and need for change in hardware such as central processing units, graphical processing units and tensor processing units. Briefly explain the concept of each with architecture. Providing a comparative analysis of each unit.

Keywords: *CPU, GPUs, TPUs, Processing units, Tensor*

1. Introduction

The present trend of usage of the computer system we changed allots to last two decades. Past 10 years tremendous change came to the user perspective. The user used the system not only for the processing of standard data, used for all kind of data accesses of the different forms such as text, image, video, gaming, structured, semi-structured and unstructured data. The usage tendency changed allot of the computer. Initially, we focused more on the processing of flat data by using a CPU based system with single or multi-core processors. Later based the need and processing of graphical data, we started working with GPUs.

Graphical processing unit came into the market in the year 1999 by the manufacturer NVIDIA. Graphical Processing unit simply shortened as GPUs. These have simply settled capacity gadgets, implying with the intention of explicitly progression phases of illustrations pipeline. For example, zenith and pixel shavers, however, they have developed into progressively adaptable coding processors. up to date, Graphical Process Units are completely programmable numerous core chips worked around a variety of comparable processors [1].

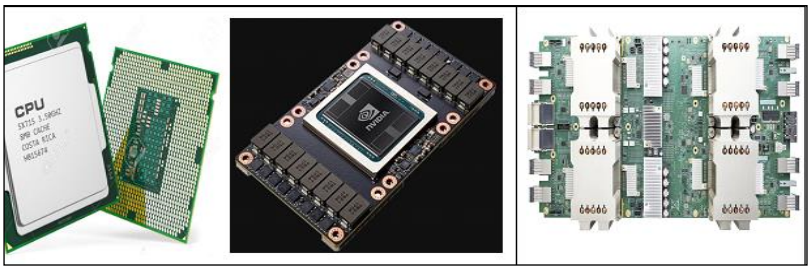


Figure 1. a) Central processing unit b) Graphical processing unit c) Tensor unit

Article history:
Received (February 11, 2020), Review Result (March 15, 2020), Accepted (April 16, 2020)

Print ISSN: 2205-8370, eISSN: 2207-5372 IJCSITE

2. Centralized Processing Unit (CPU)

CPU is the shorten name of Central Processing Unit, is the hardware with a combination of electronics, which functions as brains of the PC that play out the essential number-crunching, sensible, control and information/yield activities indicated by the instructions of a PC program [2]. As far as figuring power, the CPU is an essential component of a PC system. A massive collection of gadgets uses the CPU, with work locale, tablets, PC and workstations, Mobile phones even your smart TV.

The 3 essential elements or components of CPU, the operations of logical operations, arithmetic operations and the centralized unit, using the internal registers up to some size of data based on the bit size of the processor. Control Unit (CU), which removes directions from memory and disentangles and executes them, approaching the ALU when fundamental. Also, Memory administration unit, which is accessible in most top of the line chip to interpret legitimate locations into physical RAM addresses.

Presently, the CPU accompanies single and multi-core variation. Multi-core means having more than one processor units functioning one subsequently to the other implies that the CPU can oversee over one guideline consistently, all the while, radically enhancing execution. [6] The most popular manufacturers are AMD and INTEL work areas, PCs, and servers. At the same time, the enormous models of mobile phone used processor manufacturers are QUALCOMM, APPLE and NVIDIA.

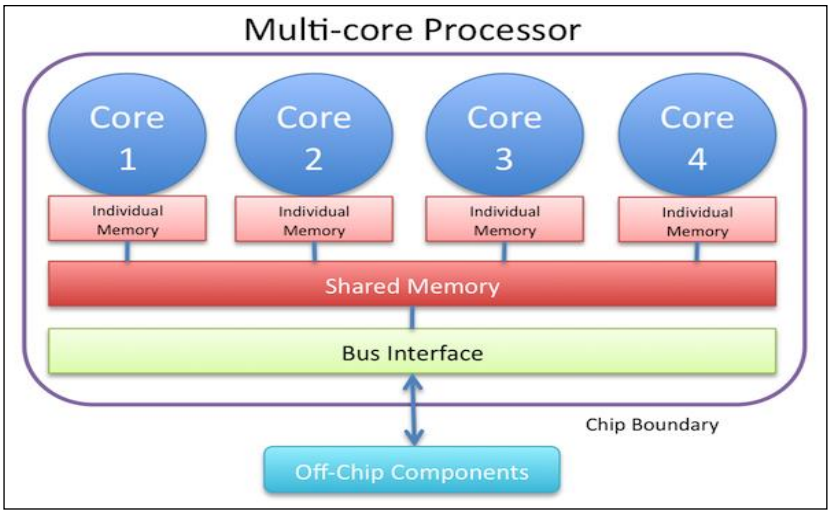


Figure 2. Multi-core processor

The processor can produce utilizing diverse advances - Single-centre CPU and more core processor. As per processors can be isolated into three kinds multiprocessors, multithreaded processors and multi-core processors [3]. There are new patterns in the CPU producing industry which depend on the possibility that while at the same time clock velocities must be expanded as far as possible and there is breaking point to number of electronic parts to be utilized as a part of a center. Numerous different advancements are there to speed things up and open routes for better and all the greater focal handling units [3]. When we can't build the execution of CPU moreover by changing its running recurrence, at that point, an innovation called multi-core engineering makes a difference. In multi-core design, we can put over one centre on a solitary silicon kick the bucket. This new way to deal with improve the speed

accompanied some extra advantages like better execution, better power administration and better cooling as the multicenter processors keep running at a lower speed to disseminate less warmth. It additionally has a few inconveniences like existing projects should be revamped according to new design. On the off chance that we don't compose programs with an exceptional centre for circling parallel centres, we won't get the favourable position of multi-cores.

3. Graphical Processing Unit (GPU)

The need for GPUs advanced to contain extensive quantities of similar threads and numerous cores? The main thrust keeps on being the execution of the ongoing illustrations required to render intricate, high-determination 3D scenes at reasonable edge rates for amusements [4].

Rendering top quality designs scenes is an issue with gigantic characteristic parallelism. An illustrations software engineer composes with one thread program that can provide one pixel, and the GPU runs various cases of this thread in similar—drawing various pixels in parallel. Illustrations programs, written in shading languages, for example, Cg or High- Level Shading Language (HLSL), in this manner scale thread forwardly finished a wide variety of string.

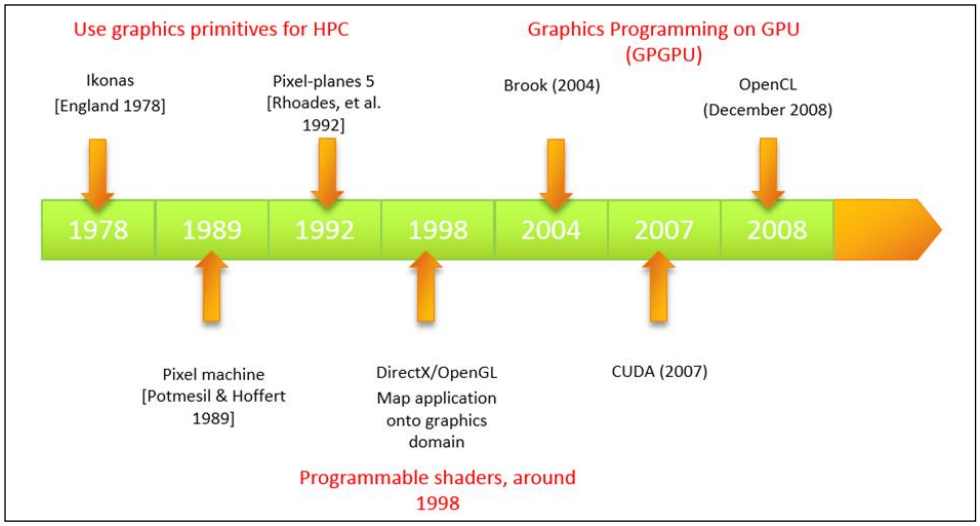


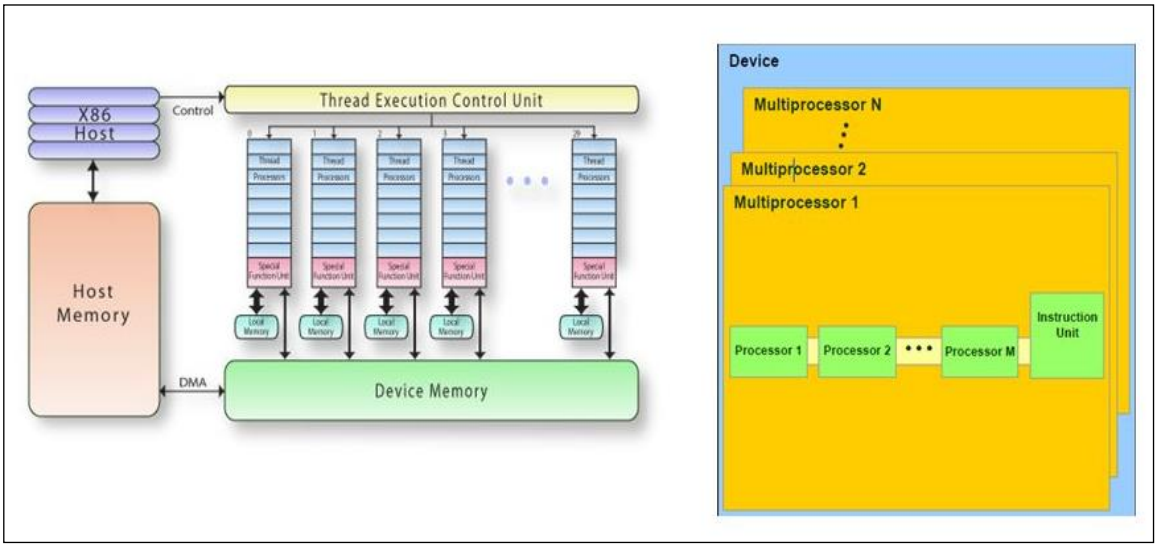
Figure 3. Evolution of the graphics programming on the GPUs: Courtesy [1]

As we come into the period of GPU figuring, requesting applications with considerable parallelism progressively utilize the significantly corresponding figuring abilities of GPUs to accomplish unrivaled execution furthermore, proficiency. Today GPU processing empowers applications that we beforehand attention infeasible because of the lengthy finishing epoch. Faster and quick possible by the GPU's from a configurable illustration workstation to a programmable a parallel central processing unit. The pervasive GPU in each PC, workstation, work area, what's more, the workstation is a many-centre multithreaded multiprocessor that exceeds expectations at the two illustrations and registering applications. The present GPUs utilize several parallel processor centres executing several thousands of parallel strings to fathom extensive issues having generous natural parallelism quickly. They're currently the

most unavoidable hugely parallel preparing stage ever accessible and also the most financially savvy.

During the period of 1990s, there were no GPUs, and Video Graphic Adapter controllers produced 2D designs show for PCs to quicken graphical UIs [4]. During the year 1997's, NVIDIA came with the 3D as single-chip called RIVA 128, quickening agent for recreations and 3Dperception applications, customized with Microsoft Direct3D and OpenGL. Developing to present-day GPUs included including programmability incrementally—from settled work pipelines to microcode processors, configurable processors, programmable processors, what's more, adaptable parallel processors.

The GPU comprises of a variety of Streaming Multiprocessors (SM), every one of which is fit for supporting a great many co-inhabitant simultaneous equipment strings, up to 2048 on current design GPUs. All string administration, including creation, booking, and hindrance synchronization is performed completely in equipment by the SM with basically zero overhead. To productively deal with its extensive string populace, the SM utilizes a SIMT (Single Instruction, Multiple Thread) engineering.



(Image copied from NVIDIA_CUDA_Tuotorial_No_NDA_Aprs08.pdf)

Figure 4. Architecture of GPU

4. Tensor Processing Units (TPU)

A Tensor is an n-dimensional network. This is the fundamental unit of activity in with TensorFlow, the open-source AI system propelled by Google Brain. A Tensor is practically equivalent to a NumPy exhibit and in truth utilizes Numpy. As indicated by their documentation, it is "NumPy is the key bundle for logical figuring with Python. It contains in addition to other things a ground-breaking N-dimensional cluster object Clusters are the key information structures utilized by AI calculations. Increasing and taking cuts from clusters takes a great deal of CPU clock cycles and memory. So Numpy was composed to make composing code to do that simpler. GPUs now make those tasks run quicker than the other models.

Google began searching for a way to support neural networking for the development of their services such as voice recognition Using existing hardware; they would require twice as many datacenters.

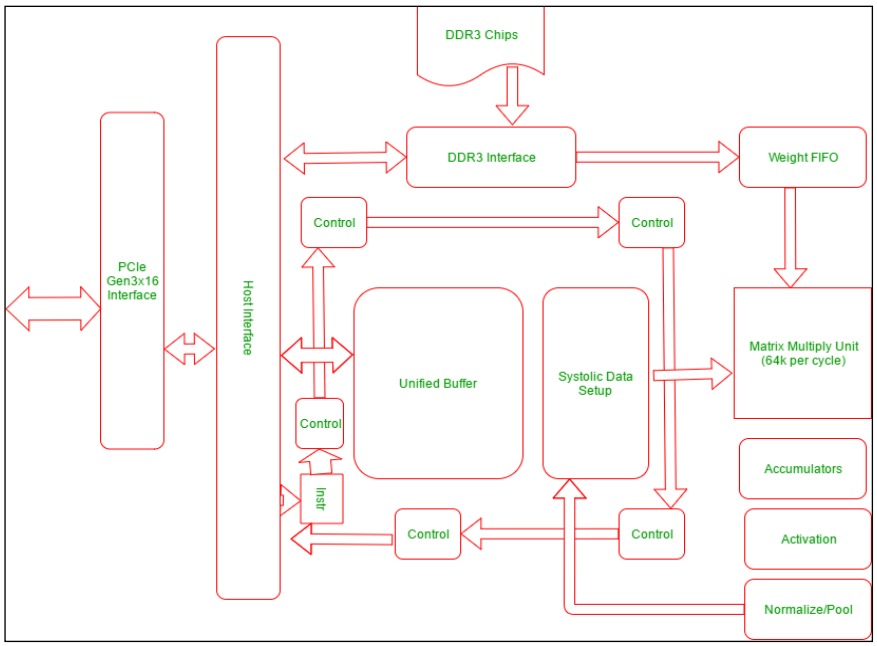


Figure 5. Architecture of tensor

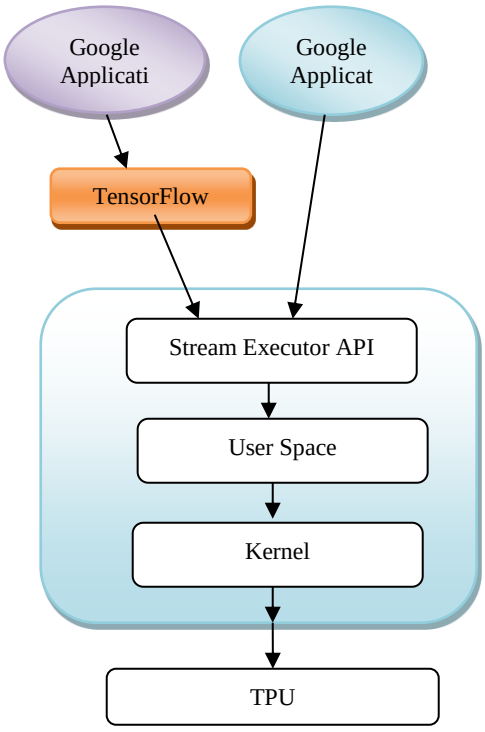


Figure 6. Application stack of tensor

- Development of a new architecture instead Norman Jouppi begins work on a new architecture to support TensorFlow
- FPGA's were not power-efficient enough
- ASIC design was selected for power and performance benefits

- The device would execute CISC instructions on many networks
- The device was made to be programmable, but operate on matrices instead of vector/scalar
- The resulting device was comparable to a GPU or Signal Processor

Based on the above study it the following are comparative analysis given as

Table 1. Comparative analysis

Parameter	CPU	GPU	TPU
Performance	10's operations per Cycle	10-103 operations per Cycle	Up to 128000 operations
Dimension of data	Unit of 1 x 1	Unit of 1 x N	Unit of N x N
Usage	Normal Programming	Graphical Programming	Machine Learning
Manufacturers	Intel, AMD, IBM, Samsung	NVIDIA,AMD	Google
Cost of Machine	10-15 $	150-200 $	350-450$

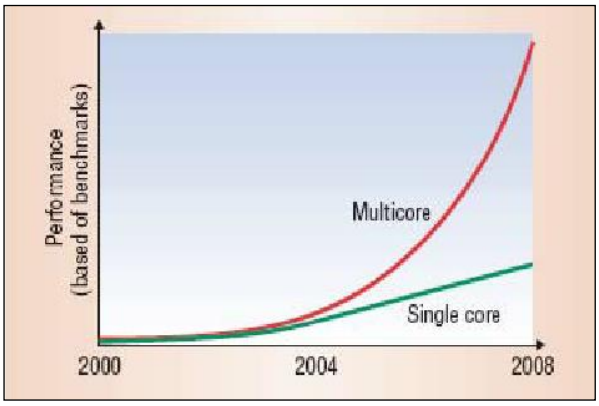


Figure 7. Single processor vs multi-core processor performance

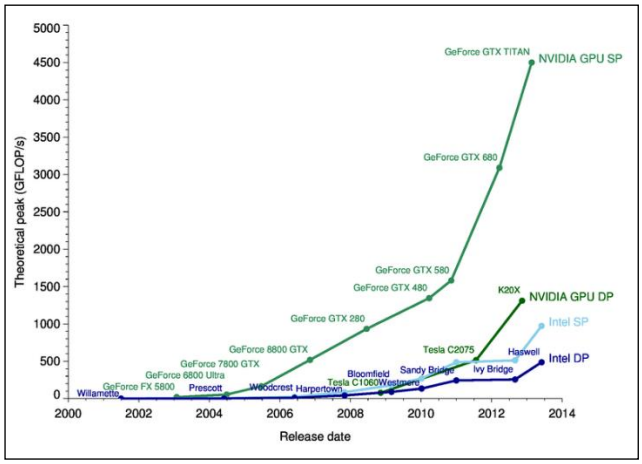


Figure 8. Performance analysis of GPUS vs CPU

5. Conclusion

In this paper, we elaborated the basics architecture views of various processing technologies used in the computer. Such as CPU, GPU and TPU, when we want work parallel task programming we can use Central Processing Units (CPUs), data-parallel over the processors we will choose the Graphics Processing Units (GPUs) [5][7]. Besides their traditional use as graphics coprocessors, the GPUs have been used in recent years for general purpose computations, too. The rapid progress of graphics hardware led to widespread use in both scientific and profitable applications. Numerous papers report high speedups in various domains. This paper presents an effort to bring GPU computing closer to programmers and a wider community of users. The most popular application programming interface, i.e. API, is with the combination of CUDA and NVIDIA to provide general purpose application of GPUs.

References

[1] G. Sukhdev Singh, "Comparison of single-core and multi-core processor", IJARCSSE, vol.6, no.6, pp.423-424, **(2016)**

[2] Hemsoth N., First In-Depth Look at Google's TPU Architecture, https://www.nextplatform.com/2017/04/05/first-depth-look-googles-tpu-architecture, **(2017)**

[3] Geekboots, "CPU vs GPU vs TPU," https://www.geekboots.com/story/cpu-vs-gpu-vs-tpu, **(2020)**

[4] N. John and D. William, "The GPU computing era", Micro, IEEE Computer Society, vol.30, pp.56-69, **(2010)**

[5] M. Misic and D. Durdevic, "Evolution and trends in GPU computing", pp.289-294, **(2012)**

[6] Tatourian A., "NVIDIA GPU Architecture & CUDA Programming Environment," https://tatourian.blog/2013/09/03/nvidia-gpu-archit>ecture-cuda-programming-environment/, **(2013)**

[7] R. Abinash, X. Jingye, and C. Masud., "Multi-core processors: A new way forward and challenges", ICM, pp.454-457, **(2008)**

This page is empty by intention.