2nd National Conference on Engineering Applications of Emerging Technology in association with International Journal of Scientific Research in Science, Engineering and Technology | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
[https://ijsrset.com | doi : https://doi.org/10.32628/IJSRSET]

A Review on GPU Architectures and Programming

Usha Kiran K, Supritha S, Vijayalaxmi BV, Yogitha P, Jyothi R
Department of Computer Science and Engineering, Global Academy of Technology, Bangalore, Karnataka, India

ABSTRACT

Graphic Programming Unit (GPU) is a parallel processor designed with high computational ability. The extensive use of GPU was in the field of gaming and rendering of 30 graphics. With the fast-growing technology GPUs are employed in many other fields like life sciences block chains and crypto currency mining and many more. In this review paper we cover a brief evolution,

Keywords- Include at least 5 keywords or phrases

I. INTRODUCTION

Graphic processing unit (GPU) are identified for the parallel computation and for rendering 3D graphics. GPUs are evolved from fixed pipelining to programmable parallel processor over a period of time. Graphic processing was accelerated by implementing parallel instruction handling by vector architectures, and instruction level parallelism (ILP) [3]. GPUs were started with the design goals of: to cover the latency of memory bandwidth so that wide range of data could be fetched in a high frequency, to support parallel graphics shader programming models, to simplify the parallel programming models [3]. The modern GPUs allows the programmers to exploit more parallelism with the help of programming languages as OpenCL and CUDA [4].

Central processing unit (CPU) are task parallel and latency oriented whereas Graphic programming unit (GPU) are data parallel and throughput oriented [5]. The major advantage of GPUs are they have hundreds of computational cores where the processing of a particular instruction is at a faster rate, GPUs have higher memory bandwidth compared to CPU.

Apart from processing graphics GPUs are used in: Data science, analytics, bioinformatics, life sciences, machine learning, imaging and computer visions and in media and entertainment and more [2].

II. THE PATH OF ADVANCEMENT

The first generation of GPU was before 256, which was introduced in 1999 which contained fixed function vertex pixel shaders programmed with open GL and the Microsoft DX? API and later in 2001 Ge Force 3 was introduced which was the first programmable vertex processor which were executing vertex shaders [3].

In 2002 the Radeon 9700 was introduced which was a programmable 24-bit floating-point fixed fragment processed which was programmed with DX9 and OpenGL. Later in 2005 the X Box 360

introduced an GPU which allowed vertices and pixel to be executed on the same processor [2].

In 2012 NVIDIA launched Kepler technology which used a dynamic parallelism. In 2016 bought a Pascal which was advanced with Al Deep Learning and other computer intensiv technology [2].

III. INFRASTRUCTURE

The typical GPU consists of several computing cores. The functional unit in GPU contain Streaming Multiprocessors (SM) and each SM in turn contain streaming/scalar processors (SP) cores and special function units (SFUs), a multithreaded instruction fetches and issue unit (Ml Issue), a 16-byte read or write shared memory and a instruction cache and read only cache [3].

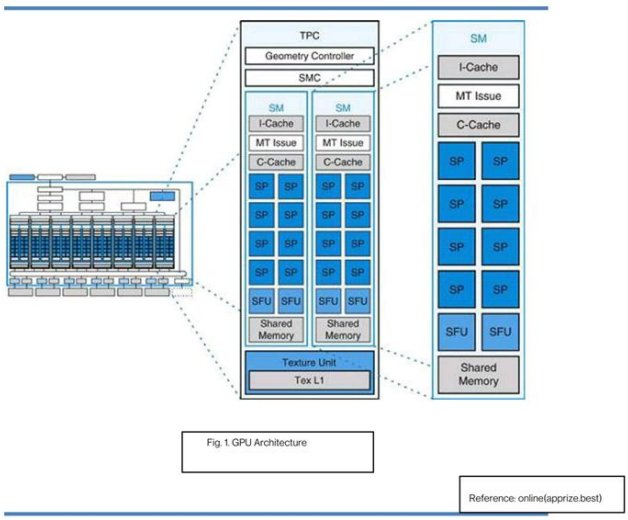
Scalar multiply add unit (MAD) is present in each SP cores. SFU units are used as transcendental functions by SM. SFUs also contain four floating point multipliers [3].

In GPUs the threads are executed in the group of 32 to 96. Group of 32 threads are called warps. The warp scheduler manages and schedules the group of 32 threads. The SM warp scheduler selects the warps to be executed where the threads are mapped to SP cores [3].

SM has large vector register files. Registers are divided across SPs.

A. Memory Distribution

The memory hierarchy of GPU is unique and implemented as on chip memory and off chip memory. The local memory for each thread for local access of a single thread and a Global memory where all threads have a access to This segment is implemented on DRAM [3].



Shared memory gives access to data shared by threads in the SM which facilitates reuse of on chip data and helps in reducing off chip traffic.

Each SM has 64 KB of on chip memory which can be configured as 48 KB of shared memory with 16 KB of L1 cache or 48 KB of L1 cache and 16 KB of shared memory

Memory are private to a particular SM. Where L2 cache are unified for all SMs [3].

B. Programming Models

OpenGL is a type of Application Programming Interface (API) used for accessing GPU hardware. It helps in implementation of video games image processing and in altering 3D graphics and images OpenGL is a hardware independent interface [2].

Direct3D is another type of low-level application programming interface from Microsoft Windows Proprietary graphics API which helps in revealing advanced graphical features [2].

Computer unified device architecture (CUDA). The arrival of CUDA reassured the application developers to have clear idea of complex graphic pipelines and also has made GPUs useful for general applications including graphics. CUDA has made the developers to easily utilize the GPI as an accelerator. CUDA programming has significantly

erased the kernel functions which is an application hotspot that runs parallel on GPU. The research studies use the GPU for computation and memory intensive applications [5] [6].

IV. CHALLENGES

All the parallel processing chips including GPU have some challenges regarding the infrastructure or regarding the implementation of the hardware. Some of those challenges are:

A. Power and Energy

The consumption of the power and energy has been increasing with the time. Meanwhile the requirement to place a greater number of devices on chip is increasing.

Since we place a greater number of devices on chip, we have to afford the power and energy so that the power is sufficient to all the devices to work normally. To save the amount of power consumed the number of instructions executed and the data movement overhead has to be reduced.

Most of the energy and power is spent in the overheads of data movement. Therefore, the future architecture should plan accordingly the proper usage of power. So that there would be promising architecture delivering most of the energies to useful works [4].

B. Memory usage

The allocation of memory plays a very important role in increasing the through put of the processor. Usage of memory efficiently is the most important tasks in the computing units. Even though GPUs have higher bandwidth than the CPU the on-chip memory is not promising when compared to off chip memory.

Since accessing the on-chip memory requires less time and power compared to accessing the off-chip memory. The future architectures should tend to increase the on-chip memory to increase the through put as well as to save the energy of a processor [4].

C. Programming

Since the cost of power and memory usage are increasing the programming should be simple enough so that more number of data fetch shouldn't be taking place the programmer should be aware of the memory management and on what kind of processors their tasks are running on and which processor would be more suitable for their tasks to execute. [4]

V. IMPLEMENTATIONS

GPUs are not only used just 3D graphics and in videogaming they have uses in the other industries as well:

A. In DNA Sequencing

Researches and Clinical trials require the entire 3 billion pairs of genomes present in our chromosomes. Followed by sequencing the computational analysis has to be performed to detect variants in the genome extracted.

GPU is used in the secondary analysis of the genomic data. The software developed by Para bricks team which runs on GPU analyses the genetic data at faster rate and identifies the mutations and variants if present. The need of entire genome analysis is increasing rapidly. Since GPU works as parallel processor the Speed Up helps researchers in the drug discovery and disease treatment.

When there is a huge amount of data to be sequenced which takes CPU years together GPU does the sequencing in the minimal time. And this is the game changer in the field of medicine which helps the doctors to quickly analyse the problem and to find a cure for it [8].

B. In Crypto Currency Mining

Earlier crypto currency m1n1ng was done using Central processing unit. But due to the using Central processing unit. But due to the less processing speed and maximum power consumption led to inefficient mining process. After the emergence of GPUs, the GPU mining became more popular with higher benefits. Compared to CPU the power potential of GPU has become more beneficial for the blockchain mining because of its high speed and efficiency. For example, A standard GPU like Radeon HD 5970 clocked processing speeds of executing 3200 32-bit instructions per clock cycle which is almost 8000 times more than the speed of a CPU which executed only 4 32 bit instructions per clock cycle. As mining process requires higher efficiency in executing similar kinds of repetitive computations. GPU has become more suitable for crypto currency mining [9].

VI. ACKNOWLEDGMENT

In this review paper we focused on the overview of the GPU and its architecture and its use in different fields. This paper delivers some basic information of evolution and trends of GPU, the basic architecture of GPU and programming languages of GPU, it also gives a brief of the challenges faced and the future adaptation's to be made to make GPU more efficient and it delivers some knowledge of emerging applications of GPU.

This paper gives some basic information about GPUs for those who are completely new to the word GPU.

VII. REFERENCES

[1]. Erik Lindholm, John Nickolls, Stuart Oberman, John Montrym, and NVIDIA, "NVIDIA TESLA: A UNIFIED GRAPHICS AND NVIDIA, "NVIDIA TESLA: A UNIFIED GRAPHICS AND COMPUTING ARCHITECTURES ", IEEE computer society 0272-1732/08 Map-Apr 2008.

[2]. Shilan Ahmed Mohammed, Rezgar Hasan Saeed, Jihan Abdulazeez Ahmed, Shilan Bashar Muhammad, Zainab Salih Ageed, Zryan Najat Rashid, "GPU Concepts and Graph Application Challenges: A Review," International Journal of Multidisciplinary Research and Publications ISSN(Online): 2581-6187.

[3]. SWAYAM, NPTEL Online course, "GPU Architectures and Programming," Prof. Soumyajit Dey.Stephen W. Keckler,William J. Dally, Brucek Khailany, Michael Garland, David Glasco, NVIDIA."GPUs AND THE FUTURE OF PARALLEL COMPUT ING." Article in IEEE Micro. November 2011 DOI: 10.1109/MM.2011.89.

[4]. Marko J. Misic, Dorde M. Durdevic, and Milo V.Tomasevic, "Evolution and Trends in GPU Computing." Conference Paper. January 2012.

[5]. Fumihiko INO."The Past, Present and future of GPU-Accelerated Grid Computing". Conference paper. December 2013.

[6]. "What's in Your Genome? Startup Speeds DNA Analysis with GPUs" [Online]. Available:https://blogs.nvidia.com/blog/2018/09/05/parabricks-genomic- analysis/

[7]. "GPU Usage in Cryptocurrency Mining." [Online]. Available: https://www.investopedia.com/tech/gpu - cryptocurrency-mining/