JCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.7, Special Issue.3, Feb. 2019 E-ISSN: 2347-2693

A Comparative Study of GPU Computing Techniques: A Review

[1*]Durgesh Kumar Keshar, [2]Sanjay Kumar, [3]V.K. Patle

[1,2,3] School of Study in Computer science & IT Pt. Ravishankar Shukla University, Raipur (Chhattisgarh) 492010 India

*Corresponding Author:durgeshkeshar.prsu@gmail.com

Available online at: www.ijcseonline.org

***Abstract* —** Nowadays, time is very important in computational field. Today every field in computer science has a huge amount of data, and we need to process data to get valuable information out of it. To reduce the processing time and using of maximum capacity of processor, we divide a large computation problem in to small chunks that is processed by individual processor. Recent microprocessors, becomes possible to utilize the parallelism using multi-cores which support improved SIMD instructions. In this paper we present the GPU based of Parallel Processing architecture, working and its applications for performing fast execution of a task.

***Keywords*—** Parallel computing technique. GPU Architecture, Memory architecture.

I. INTRODUCTION

A huge computational task may take very long time to find accurate result when it processed on Single Processor. Every processor has its own physical architectural limit and maximum processing speed. To overcome this condition multiple processors are introduced in which each processor co-ordinate to other that leads to Parallel Computing. In parallel computing architecture huge computational problem can be divided into multiple problems that can be handled by different cores independently and result comes after recombination of all solutions of problem. GPU and CPU have heterogeneous architecture; it has unique opportunity for energy conservation [1]. GPU performs over large number of cores which gives better results compare to CPU. A graphics processing unit or GPU is a specialized microprocessor that discharges and accelerates 3D or 2D graphics. It is used in embedded systems, mobile phones, personal computers, workstations, and game consoles. GPU is widely used for cost effective and high performance computing. Currently GPU and CPU have also competence to solve complex problem with high performance. GPU needs specific software and libraries like MPI and OpenMP that execute any parallel application [2]. Currently GPU processes image like highly parallel multithreads, with marvelous computation speed and very eminent memory bandwidth.

II. ARCHITECTURE OF CPU AND GPU COMPUTING

A GPU (Graphics Processing Unit) acts as a co-processor to accelerate CPU for computing. The GPU accelerates application running on the CPU (Central Processing Unit) by loading some complex and time consuming problems on GPU, rest of the application as is on CPU.

CPU core support fewer threads and does not require additional data communication; on the other hand GPU needs huge number of current threads.
Mostly, CPU has up to 12 cores whereas GPU have up to 512 cores in a single chip [3].

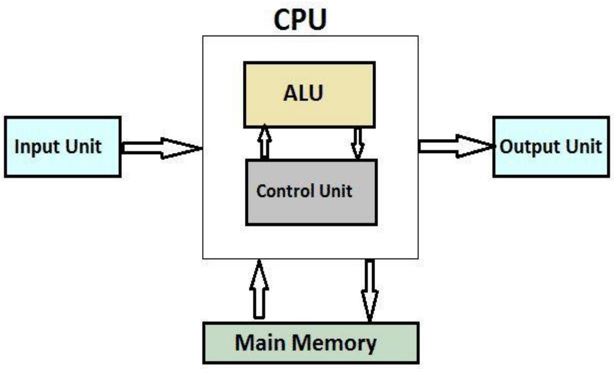

Figure.1 CPU Architecture [20]

Control unit fetches the instruction/data from memory decode the instruction and then sequentially executed the programming task, whereas parallel computing also have same design but little bit changes in its processing units.

A GPU is a heterogeneous chip multiprocessor.

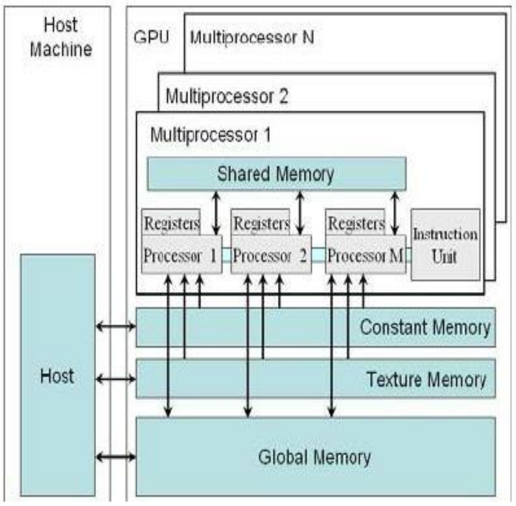

Figure. 2 GPU Architecture [21]

NVIDIA introduces first GPU **(GeForce256 announce by Nvidia on 31 August 1999 and release on October 11,1999)** [5] that operates on scientific application which lead to GPGPU (General Purpose Graphics Processing Unit) computing. NVidia realizes the complexity of GPGPU so that it adds high level language like C and C++.

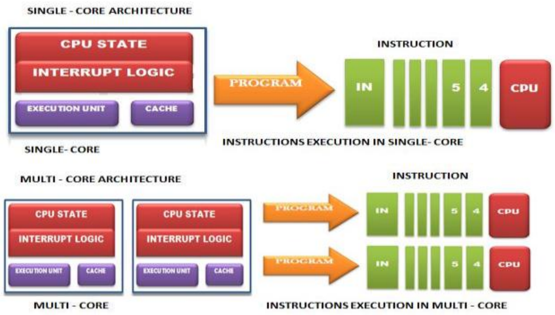

Figure.3 Single & Multicore Core Architecture

III. FLYNN'S CLASSICAL

To Understand Parallel computers one of the most widely classification is Flynn's Classification introduced in 1966 [8]. Parallel computer also have different memory architecture such as UMA (Uniform Memory Access), NUMA (Non Uniform Memory Access), and Hybrid distributed shared memory.

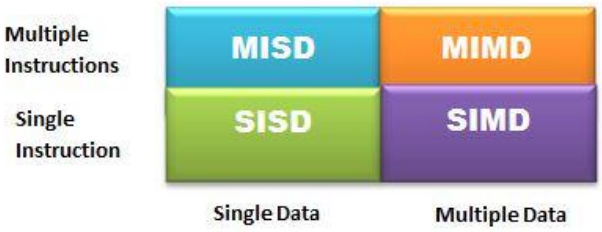

Figure. 4 Flynns Classification

A. SISD (Single Instruction Single Data)
A serial (Non Parallel) computer that have only one instruction stream is being acted on by the CPU during any one clock cycle and only one data stream is being used as input during any one clock cycle, It act as deterministic execution.
Exa.: UNIVAC,IBM360,Cray etc.

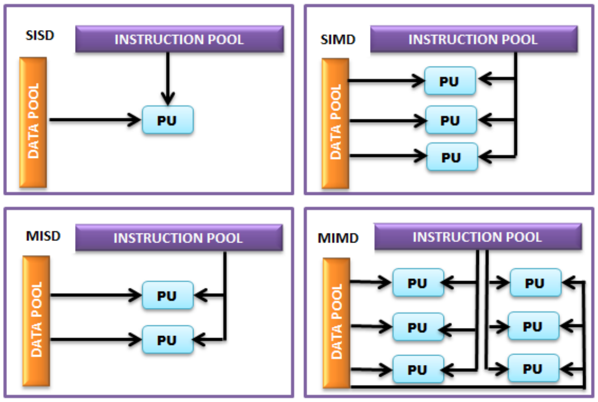

Fig. 5 SISD, SIMD, MISD, MIMD [22]

B. SIMD (Single Instruction Multiple Data)

A type of parallel computer which has all processing units execute the same instruction at any given clock cycle and each processing unit can be operated on a different data element.
Exa.:- Cray X-MP, CrayY-MP, ILLIAC IV etc.

C. MISD (Multiple Instruction Single Data)

A type of parallel computer which has all processing units operates on the data independently via separate instruction stream and single data steam is fed into multiple processing units.

D. MIMD (Multiple Instruction Multiple Data)

A type of parallel computer which have every processor may be executing a different instruction stream and every processor may be working with different data stream.
Exa.:- IBM POWER5, Intel IA32, IBM BG/L etc.

IV. LITERATURE REVIEW

Suichi Asano,T Sutomu and Yoshiki Yamaguchi [4]. proposed a performance of image processing, it also becomes possible to utilize the parallelism using multi-cores which supports improved SIMD instructions, though programmers have to use them explicitly to achieve high performance. They use Nvidia GTX280 because it supports programming environment called CUDA (Compute Unified Device Architecture) [5]. GPU can show its best performance. T. Brandes, A. Arnold, T. Soddemann, and D. Reith, discuss how well the Gram-Schmidt method performs on different hardware architectures, including both state-of-the-art GPUs and CPUs. NVIDIA GeForce GTX580, GPU is about 50% faster than a corresponding Intel X5650 Westmere hexacore CPU [6]. Ben Cope studied a 2D convolution algorithm implementation on data streaming and pipeline, GPU architectures Throughput exceeds that of the CPU for all filter sizes [7]. Gurindar singh and Aman kaurra try to explain how a big task can be divided in to small number of sub task which can be handled by different processors. They also try to explain considerable different forms of parallel computing like instruction, bit, data and task level parallelism. In GPU computing, a job is broken down in multiple similar subtasks that can be processed separately and form a results after the completion of the job processing [8]. Bhavna Samel, Shubhrata and Prof. A.M. Ingole try to explain performance of GPU that's optimized for which are highly effectively than a CPU that's optimized [9]. One of the members of IEEE Craig A. Lee try to explains GPU performance and architecture on remote sensing areas [10]. Kai Ma explains GPU-CPU adopted in high performance computing, because of their competence of providing high computational throughput.

For GPU, CPU heterogeneous architectures GPU dynamically split. Current research goes on performance of the GPU-CPU, while the energy expertise of such systems receives much less attention. GreenGPU randomly control the frequencies of GPU cores, based on their utilizations, for maximized energy savings with only marginal performance degradation [1]. Christian Martin explained importance of architectural alternatives they clear up the reasons of the Post- Dennard scaling regime and discussed the consequences for future multicore designs [11]. Many time for Speed-up measures a GPU works faster than a CPU one when both programs solve the same problem. Speed-up is denoted by Sp which is the ratio of Ts and Tp and can be represented as

$$Sp = \frac{Ts}{Tp}$$

Hence Sp measures the benefit of GPU over CPU [12]. So According to the Amdahl's law, Due to the communication cost, the speed-up is always less than equal to the number of processors used in parallel computer [13].

V. WHY PARALLEL COMPUTING AND IT'S APPLICATION

GPU also important for making super computer like **Tianha - 2 (China),** It has 3120000 cores, **Titan (USA),** It uses a total of 561,000 Opteron 6274 16-core 2.2GHz processors along with NVIDIA GPUs, **Sequoia (USA),** IBM machine with 1.6 million 16- core processors. It has a performing speed of 17.2 petaflops [14]. Parallel computing provides concurrency execution of task, solve huge and complex problem, the real world is monolithic, save time and cost and also better performance of software and hardware. Parallel computing plays a vital role on different fields like science, engineering, industrial & commercial, space, aeronautics and medical area.

Medical: For the diagnostic of cancer patient, Using image processing it is easy to describe specific description of tumor [15, 16].

Science & Engineering: NVidia CUDA is an extension to C language offer programming efficiency to General purpose GPU (GPGPU). It offers high performance computing platform in remote sensing areas [17, 18, 19].

GPU Industry: Many applications have been developed to use GPUs for supercomputing in various fields Like Scientific Computing: Molecular Dynamics, Genome Sequencing, Mechanical Simulation, Quantum Electrodynamics, Image Processing,
Registration, interpolation, feature detection, recognition, filtering Data Analysis Databases, sorting and searching, data mining.

VI. CONCLUSION

In this paper, we studied detailed overview of architecture, memory structure and application of GPU based parallel computing. Using GPU computing our system performance is increases. We must use parallelism for the increased performance required to deliver more value to users. A GPU that's optimized for throughput delivers parallel performance much more powerfully than a CPU. We tried to explain importance and difference between CPU and GPU. In this era it is important to work on GPU computing so that we can explore the power of Digital World and Artificial

Intelligence system (AI). In future researchers may try to focus on energy consumption with performance.

REFERENCES

[1] Kai Ma†, Xue Li‡, Wei Chen†, Chi Zhang‡, and Xiaorui Wang†, *"GreenGPU: A Holistic Approach to Energy Efficiency in GPU-CPU Heterogeneous Architectures,"* 2012 41st International Conference on Parallel Processing.

[2] Rafiqul Zaman Khan, Md Firoj Ali, *"Current Trends in Parallel Computing, International Journal of Computer Applications"* (0975 – 8887) Volume 59– No.2, December 2012.

[3] Bhavna Samel, Shubhrata and Prof. A.M. Ingole, *"GPU Computing and Its Applications, International Research Journal of Engineering and Technology (IRJET),"* Volume: 03 Issue: 04 | Apr-2016.

[4] Shuichi Asano, Tsutomu Maruyama and Yoshiki Yamaguchi, *"Systems and Information Engineering,"* University of Tsukuba 1-1-1 Ten-ou-dai Tsukuba Ibaraki 305-8573 JAPAN, 978-1-4244-3892-1/09/$25.00 ©2009 IEEE

[5] http://www.nvidia.com/object/cuda home.html.

[6] T. Brandes1,a, A. Arnold2, T. Soddemann1, and D. Reith1, *"The European Physical Journal Special Topics*," Eur. Phys. J. Special Topics 210, 73–88 (2012)c_EDP Sciences, Springer-Verlag 2012

[7] DOI: 10.1140/epjst/e2012-01638-7

[8] Ben Cope,Department of Electrical & Electronic Engineering, Imperial College London benjamin.cope@imperial.ac.uk

[9] Gurindar Singh and Aman Kaura, *"Recent Trends in Parallel Computing,"* GIAN JYOTI E-JOURNAL, Volume 6, Issue 1 (Jan-Apr 2016).

[10] Bhavna Samel, Shubhrata and Prof. A.M. Ingole, *"GPU Computing and Its Applications, International Research Journal of Engineering and Technology (IRJET),"* Volume: 03 Issue: 04 | Apr-2016.

[11] Craig A. Lee, Member, IEEE, Samuel D. Gasster, Senior Member, IEEE, Antonio Plaza, Senior Member, IEEE, Chein-I Chang, Fellow, IEEE, and Bormin Huang, *"Recent developments in High Performance Computing for Remote Sensing: A Review,"* IEEE Journal of selected topics in applied earth observations and remote sensing," VOL. 4, NO. 3, SEPTEMBER 2011.

[12] Christian Märtin, Augsburg University of Applied Sciences Augsburg, Germany, embedded world 2014, exhibition and conference

[13] Rafiqul Zaman Khan, Md Firoj Ali, *"Current Trends in Parallel Computing, International Journal of Computer Applications,"* (0975 – 8887) Volume 59– No.2, December 2012.

[14] Amdahl G. M. *"Validity of the Single-processor Approach to Achieving Large Scale Computing Capabilities*". In AFIPS Conference Proc., Atlantic City, New Jersey, pp.483-485, 1967.

[15] https://fossbytes.com/top-5-supercomputers-of-the-world/ till july 2018.

[16] George Teodoro, Rafael Sachetto, Olcay Sertel,Metin N. Gurcan, Wagner Meira Jr., Umit Catalyurek, Renato Ferreira1, *"Coordinating the Use of GPU and CPU for Improving Performance of Compute Intensive Applications,"* 978-1-4244-5012-1/09/$25.00 ©2009 IEEE.

[17] L. Teot, R. Khayat, S.Qualman, G.Reaman, and D.Parham, *"The Problems and Promise of Central Pathology Review: Development of a Standardized Procedure for the Children's Oncology Group,"* Pediatric and Developmental Pathology, vol. 10, no. 3, pp. 199–207, 2007.

[18] Craig A. Lee, Member, IEEE, Samuel D. Gasster, Senior Member, IEEE, Antonio Plaza, Senior Member, IEEE, Chein-I Chang, Fellow, IEEE, and Bormin Huang, *"Recent Developments in High Performance Computing for Remote Sensing: A Review,"* IEEE Journal of selected topics in applied earth observations and remote sensing," VOL. 4, NO. 3, SEPTEMBER 2011.

[19] J. Nickolls and W. J. Dally, *"The GPU computing era,"* IEEE Micro, vol. 30, pp. 56–69, 2010.

[20] T. Balz and U. Stilla,*"Hybrid GPU-based single- and double-bounce SAR simulation,"* IEEE Trans. Geosci. Remote Sens.," vol. 47,no. 10, pp. 3519–3529, 2009.

[21] http://codesandtutorials.com/hardware/computerfundamentals/cpu-block_diagram-working.php as on July 7, 2018.

[22] https://www.researchgate.net/figure/GPU-Schematic-Architecture_fig1_220489693 as on July 7, 2018

[23] https://computing.llnl.gov/tutorials/parallel_comp/#Flynn as on July 7, 2018.