**RAJSHAHI UNIVERSITY OF ENGINEERING AND TECHNOLOGY**

|  |  |
| --- | --- |
| RUET 09 | |
| ASSIGNMENT  Course No: CSE 3109  Topic: GPU Acceleration  Date of Submission: 07 September, 2019 | |
| Submitted to:  Biprodip Pal  Assistant Professor,  Department of Computer Science and Engineering  Rajshahi University of Engineering and Technology | **Submitted by:**  Riyad Morshed Shoeb  Roll No: 1603013  Section: A  Department of Computer Science and Engineering  Rajshahi University of Engineering and Technology |

GPU-accelerated computing is the employment of a graphics processing unit (GPU) along with a computer processing unit (CPU) in order to facilitate processing-intensive operations such as deep learning, analytics and engineering applications. It is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device. Developed by NVIDIA in 2007, the GPU provides far superior application performance by removing processing-intensive application sections to GPU. GPU-accelerated computing deployment is growing in popularity due to the large variety of applications in which it could be used, such as artificial intelligence, drones, robots or autonomic cars. GPUs are used in embedded systems, mobile phones, personal computers, workstations, and game consoles. Modern GPUs are very efficient at manipulating computer graphics and image processing. Their highly parallel structure makes them more efficient than general-purpose central processing units (CPUs) for algorithms that process large blocks of data in parallel. In a personal computer, a GPU can be present on a video card or embedded on the motherboard. In certain CPUs, they are embedded on the CPU die.

The GPU helps in providing superior performance for software applications. From the perspective of the user, GPU-accelerated computing makes applications faster. GPU-accelerated computing functions by moving the compute-intensive sections of the applications to the GPU while remaining sections are allowed to execute in the CPU. While the CPU is comprised of cores designed for sequential serial processing, the GPU is designed with a parallel architecture consisting of more efficient yet smaller cores that can easily handle multiple tasks in parallel. As a result, in GPU-accelerated computing, while sequential calculations are performed in the CPU, highly complicated calculations are computed in parallel in the GPU. Another salient feature of GPU-accelerated computing is the support offered to all the parallel programming models, thus helping application designers and developers to provide superior application performance.

Modern GPUs use most of their transistors to do calculations related to 3D computer graphics. In addition to the 3D hardware, today's GPUs include basic 2D acceleration and framebuffer capabilities (usually with a VGA compatibility mode). Newer cards such as AMD/ATI HD5000-HD7000 even lack 2D acceleration; it has to be emulated by 3D hardware. GPUs were initially used to accelerate the memory-intensive work of texture mapping and rendering polygons, later adding units to accelerate geometric calculations such as the rotation and translation of vertices into different coordinate systems. Recent developments in GPUs include support for programmable shaders which can manipulate vertices and textures with many of the same operations supported by CPUs, oversampling and interpolation techniques to reduce aliasing, and very high-precision color spaces. Because most of these computations involve matrix and vector operations, engineers and scientists have increasingly studied the use of GPUs for non-graphical calculations; they are especially suited to other embarrassingly parallel problems.

With the emergence of deep learning, the importance of GPUs has increased. In research done by Indigo, it was found that while training deep learning neural networks, GPUs can be 250 times faster than CPUs. The explosive growth of Deep Learning in recent years has been attributed to the emergence of general purpose GPUs. There has been some level of competition in this area with ASICs, most prominently the Tensor Processing Unit (TPU) made by Google. However, ASICs require changes to existing code and GPUs are still very popular.

GPU-accelerated computing has been extensively used in video editing, medical imaging, fluid simulations, color grading and enterprise applications, and its use is promising in complex fields such as artificial intelligence and deep learning.

Resources:

1. <https://www.techopedia.com/definition/32876/gpu-accelerated-computing>
2. <https://en.wikipedia.org/wiki/Graphics_processing_unit>