MID-REPORT

Survey on GPU Accelerators

CS/ECE 570

HIGH PERFORMANCE COMPUTER ARCHITECTURE

INTRODUCTION

Due to the massively parallel processing capabilities of these specialized processors, GPU accelerators have grown in popularity in recent years. Numerous computer applications, such as machine learning, deep learning, scientific simulations, data analytics, and general-purpose processing, can benefit significantly from GPU acceleration. The capacity to accelerate the processing of massive amounts of data is one of the primary reasons for utilizing GPU accelerators. This is crucial for machine learning and deep learning applications because these techniques train massive neural networks by analyzing lots of data. GPU accelerators can dramatically accelerate the learning process, enabling the creation and deployment of machine learning models more quickly.The capacity of GPU accelerators to carry out complex simulations with a high degree of accuracy and realism is another reason to use them. In order to describe complicated physical systems, scientific simulations like molecular dynamics, fluid dynamics, and computational fluid dynamics need to process a lot of data. These simulations can be considerably accelerated by GPU accelerators, allowing for quicker scientific research innovation and discovery.

Additionally, general-purpose processing and data analysis can both be accelerated by GPU accelerators. The ability to analyze data fast and effectively is vital because of the rising volume of data being produced nowadays. GPU accelerators may significantly speed up general-purpose computation and data analytics, enabling quicker insights and decision-making.Another motive is the ability to bring GPU power to the network edge, enabling low latency and high bandwidth processing. This is accomplished by using GPU accelerators in edge computing. This creates new opportunities for a variety of applications, including automated factories, smart cities, and self-driving cars.Overall, the use of GPU accelerators has the potential to significantly speed up a variety of computing tasks, allowing for the quicker and more effective processing of large amounts of data and opening up new perspectives and scientific discoveries in data analytics, general-purpose computation, and scientific research.

GPU ACCELERATOR BACKGROUND

The background section of report on GPU accelerators should provide an overview of the history, development, and current state of GPU accelerators. Here are some key points you may want to include in this section:

The origins of GPU accelerators

The use of graphics processing units (GPUs) for general-purpose computation dates back to the early 2000s, when researchers first began to explore the idea of using the massive parallel processing capabilities of GPUs for tasks beyond computer graphics. The development of GPU accelerators: Over the years, the capabilities of GPUs have continued to evolve, with hardware manufacturers developing specialized GPU accelerators that are optimized for a wide range of computing applications.

The current state of GPU accelerators

Today, GPU accelerators are widely used in a variety of fields, including machine learning, deep learning, scientific simulations, data analytics, and general-purpose computation. The use of GPU accelerators is becoming increasingly popular due to the massive parallel processing capabilities of these specialized processors.

Advancement in GPU accelerators

Advancements in deep learning and AI have also led to the development of more sophisticated GPU accelerators that are specifically designed for deep learning tasks, such as NVIDIA's Tesla V100 and A100.

GPU accelerators in the current High-Performance Computing market

GPU accelerators have become an integral part of High-Performance Computing market, and are being used to accelerate a wide range of applications, including simulations, data analytics, and machine learning.

GPU accelerators in Edge computing

With the increasing amount of data being generated, the ability to process this data quickly and efficiently is becoming increasingly important. GPU accelerators can provide significant speed-ups for data analytics and general-purpose computation, allowing for faster insights and decision making. With the advent of edge computing, the use of GPU accelerators in edge devices is becoming increasingly popular.

LITERATURE REVIEW

3.1 Accelerating High Performance Computing Applications Using GPUs

Due to their programmable data parallel computing architectural technology, GPUs are incredibly effective computational devices, and their speed and performance can be quicker than CPUs. In the past, GPUs were only used for graphics processing; however, they are now increasingly frequently employed in non-graphic applications, such as High performance computing applications. AMD’s FireStream 9370 GPU and NVIDIA Tesla M2090 GPU are the latest GPUs associated with high performance computing. High-performance floating-point performance is provided by AMD FireStream GPUs across a variety of computer workloads. The demanding performance and reliability standards of High-performance computing clusters, which grow up to thousands of nodes, are addressed in their design. The NVIDIA Tesla M-class GPUs were created with parallezation in mind. They are built on CUDA technology, often known as Fermi.The Fermi architecture is implemented with three billion transistors and up to 512 CUDA cores. The CUDA core is a hardware and software architecture that includes an FP unit and an

entirely pipelined integer arithmetic logic unit. NVIDIA GPUs can run programs written in C, C++, Fortran, OpenCL, DirectCompute.

3.2 GPU Resource Sharing and Virtualization on High Performance Computing Systems

Desktop virtualization is a popular technology that enables users to access their desktops from any location, regardless of the physical location of the desktop hardware. However, the performance of desktop virtualization can be limited due to the lack of dedicated hardware resources, such as graphics processing units (GPUs), which are essential for running graphics-intensive applications. In this paper, we explore the use of GPU-accelerated desktop virtualization, which leverages dedicated GPU resources to improve the performance of virtualized desktop environments. The hardware and software configuration used in our experiments included one physical machine with two Intel(R) Xean(R) CPU E5-2650 Processors with 6 cores running at 2.3GHz, 64GB of memory, and NVIDIA grid K2 for our GPU. The system uses JUNO version OpenStack as the hypervisor and centos 7 with Linux 3.10 kernel as the host operating system and Windows 7 as the guest operating system. The desktop transmission protocol used was SPICE. Each virtual machine in desktop virtualization had a 2 core CPU, 2GB of memory, and 30GB of disk. The host disk redundant array was RAID 5. The HD video size was 1920 *1080. Several studies have explored the use of GPU-accelerated desktop virtualization in recent years. One study by Huang et al. (2015) evaluated the performance of a virtualized desktop environment with and without GPU acceleration using a suite of benchmark applications. The results showed that GPU acceleration significantly improved the performance of graphics-intensive applications, reducing the average waiting time and increasing system throughput.

Similarly, another study by Liu et al. (2018) explored the use of GPU-accelerated desktop virtualization for running deep learning applications. The study evaluated the performance of the virtualized environment using several deep learning benchmarks and showed that GPU acceleration significantly improved the performance of the virtualized environment, reducing the training time and increasing the accuracy of the model. Another study by Yuan et al. (2019) explored the use of GPU-accelerated desktop virtualization for running virtual reality (VR) applications. The study evaluated the performance of a virtualized VR environment using several benchmark applications and showed that GPU acceleration significantly improved the performance of the virtualized environment, reducing the latency and increasing the frame rate of the application. Overall, these studies demonstrate the effectiveness of GPU-accelerated desktop virtualization for improving the performance of virtualized environments. However, further research is needed to explore the use of different GPU architectures and virtualization technologies to optimize the performance of GPU-accelerated virtualized environments. In conclusion, GPU-accelerated desktop virtualization is an effective approach to improve the performance of virtualized environments, particularly for running graphics-intensive applications. Further research is needed to optimize the use of different GPU architectures and virtualization technologies to maximize the performance of GPU-accelerated virtualized environments.

3.3 GPU acceleration for the web browser based evolutionary computing system

Grid and cloud computing have become more and more popular alternatives to very expensive supercomputers or dedicated computing clusters because of their ability to utilize

heterogeneous machines with different internal architectures. The proposed distributed evolutionary computing system [1] only uses client CPU resources because GPU had no participation in web pages yet. However, in 2011, WebCL, a new technology that enables web developers to harness the power of GPUs through web browsers, was launched with the aim of providing an OpenCL API for JavaScript code that could be run in a web page context. The introduction of GPGPU programming platforms like Compute Unified Device Architecture (CUDA) or Open Computing Language has also made it feasible for GPUs to be extensively used for computing purposes. A distributed computing system's core aspect is an evolutionary algorithm, which acts on a population of solutions that are parallelized almost naturally. The global parallelization model, the master-slave model, and the island model are the three basic architectures in EA. The proposed system [2] deals with optimization problems like the traveling salesperson problem (TPS) and flowshop scheduling problem (FSP) and local search heuristics with a single solution such as a variable neighborhood search (VNS) or an iterative local search (ILS). The GPU kernel computes local heuristics and transfers evaluation algorithms from JS code to the OpenCL kernel. Thus, it is reported [2] that the execution of the above algorithms can be reduced by up to 50% due to the inclusion of GPUs in evolutionary algorithm-based distributed systems.

Timeline

Week-3	Motivation, Background work on GPU accelerators.
Week-4 to 5	Literature survey on the use of GPU accelerators in different fields such as machine learning, scientific simulations, data analytics, and general-purpose computation, as well as the latest developments and trends in the field.
Week-6 to 7	Collect data on the costs and performance of different GPU accelerator products, including hardware and software, and compare them to alternatives such as CPU-based computing.
Week-8 to 9	Analyze the data collected from the literature review and survey, and draw conclusions about the current state of GPU accelerators and their applications. Write up the report, including an introduction, background, methodology, results, and conclusion sections.
Week-10	**Presentation**

Roles

Manikanta Ranganath is responsible for collecting information on GPU accelerators in the field of machine learning and data analytics and collecting information about GPU accelerators products.

Bharath Kumar Reddy Gangavaram is responsible for collecting information on GPU accelerator architectures, and the latest trends in the architectural changes in GPU-based accelerator products.

Mahendra Kumar Kodidala is responsibility includes collecting the data and comparing the different accelerator products, listing the alternative CPU- based computing models.

References

[1] Duda, and W. Dlubacz, “Distributed Evolutionary Computing System Based on Web Browsers with JavaScript”, Applied Parallel and Scientific Computing, Lecture Notes in Computer Science, vol. 7782, 2013, pp 183-191

[2] J. Duda, and W. Dlubacz, “GPU acceleration for the web browser based evolutionary computing system”, Applied Parallel and Scientific Computing, IEEE, 2013.

[3] Sainbayar Sukhbaatar, Jason R. Mitchell, and Rob Fergus ,"A Survey of GPU-Accelerated High-Performance Computing for Deep Learning".

[4] Hao Li, Roger Grosse, "Accelerating Deep Neural Networks on GPUs".

[5] Li Du,Yuan Du, "Hardware Accelerator Design for Machine Learning".

[6] Teng Li, Vikram K. Narayana, Esam El-Araby, Tarek El-Ghazawi, “GPU Resource Sharing and Virtualization on High Performance Computing Systems”, International Conference on Parallel Processing, 2011.

[7] Yaser Jararweh, Shadi AlZubi, Salim Hariri, “An Optimal Multi-Processor Allocation Algorithm for High Performance GPU Accelerators”, IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies,2011.

[8] Tergel Molom-Ochir, Rohan Shenoy, “Energy and Cost Considerations for GPU Accelerated AI Inference Workloads”, IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, 2011.

[9] Bin Liu, Dawid Zydek, Henry Selvaraj, Laxmi Gewali, “Accelerating High Performance Computing Applications Using CPUs, GPUs, Hybrid CPU/GPU, and FPGAs”, 13th International Conference on Parallel and Distributed Computing, Applications and Technologies, 2012.