Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/49/20241253

Enhancing GPU performance and energy efficiency: Innovative strategies for sustainable computing

He Shen

Department of Engineering, University College London, London, WC2N 5BY, British

zceehs3@ucl.ac.uk

Abstract. With the rapid advancement of science and information technology, Graphics Processing Units (GPUs) have become indispensable tools in contemporary scientific research and industrial production. This paper delves into optimizing GPU performance and energy efficiency, with a specific focus on five aspects: core count and layout, memory hierarchy and cache design, clock speed, specialized hardware units for specific tasks, and interconnections among GPU components. Increasing core count generally enhances GPU performance but also elevates energy consumption. Optimizing core count and layout strikes a balance between performance and energy use. Memory hierarchy and cache design are crucial for handling the inherent parallelism of GPU architecture; optimizing these aspects boosts GPU efficiency and sustainability. Clock speed is a vital performance indicator, with the right speed achieving an optimal balance between performance and energy use. Tailoring hardware units for specific tasks enhances computational efficiency and lowers energy consumption. Optimizing interconnections between GPU components enhances data transfer efficiency, yielding higher performance. Through comprehensive research and optimization in these five areas, this paper introduces innovative strategies and techniques to boost GPU performance and reduce energy consumption, laying the foundation for sustainable computing development.

Keywords: GPU, low consumption, higher efficiency

1. Introduction

In recent years, with the rapid development of science and information technology, the high-performance computer has become an important tool in modern scientific research and industrial production. The relentless growth in computational demands, coupled with a universal emphasis on energy efficiency, has thrust the design and optimization of GPU architectures into the limelight. Graphics Processing Units (GPUs) have transcended their initial role in rendering graphics, becoming indispensable accelerators for a myriad of applications, from deep learning to simulations. However, the sheer power and computational capabilities they bring also come with heightened energy costs. Balancing raw performance with energy consumption has become a paramount challenge. As such, optimizing GPU architecture and design for both these facets isn't just an engineering necessity but also a cornerstone for sustainable computing [1].

This article will thoroughly examine the intricate balance between performance and energy efficiency within the realm of Graphics Processing Units (GPUs). It will delve deep into the strategies, techniques, and fundamental principles that underpin the continuous evolution of GPU designs. By dissecting the

DOI: 10.54254/2755-2721/49/20241253

intricacies of this crucial interplay, the article aims to shed light on the innovative approaches and cutting-edge developments driving the optimization of GPU performance while concurrently addressing energy consumption concerns.

2. A Comprehensive Overview of GPU and its optimization

2.1. Definition of GPU

GPU were originally designed to accelerate graphics rendering, a type of task that requires a lot of parallel computing, so GPU have relatively simple cores. They are specifically designed to perform similar repetitive tasks, so more cores can be integrated on the same chip. As the number of cores increases, it is inevitable that their power consumption will increase, as each core consumes power while running. At the same time, more cores mean more parallel processing power, which often leads to higher performance, especially in highly parallel tasks [2]. While increasing the number of cores in a GPU generally improves performance, this increase is not always linear. The structure of a GPU is shown in figure 1.

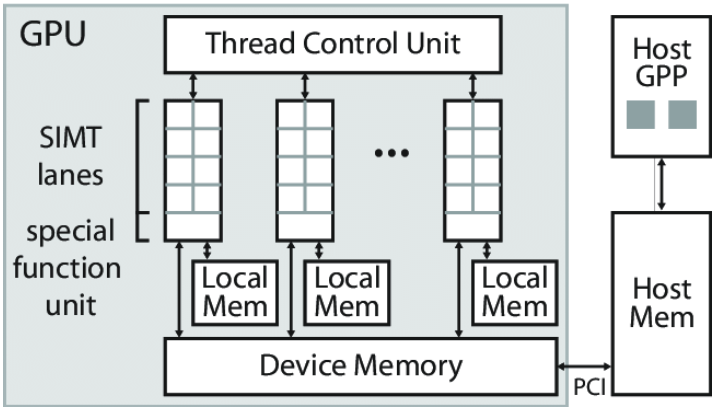

Figure 1. The structure of a GPU [2]

2.2. Optimizing GPU Performance and Efficiency

Graphics Processing Units have a unique memory hierarchy and cache design, which is instrumental in meeting the demands of massive parallelism inherent in their architecture. In terms of performance, different memory ranges in the GPU hierarchy have different read speeds from registers to global memory, generally speaking, the read speed of registers is faster, and the read speed of global memory is slower. Performance may be affected. At the same time, in terms of cache hierarchy, the hierarchy of different designs may have a greater impact on the mind. In terms of energy consumption. Accessing different memory types may consume different amounts of energy. Registers and shared memory typically consume less energy than global memory. Caches, on the other hand, can speed up memory access, but they also consume power [3].

Clock speed is a key indicator of GPU performance because it indicates how many operations the GPU can perform per second. In the condition of all else environment being equal, higher clock speeds mean higher performance. As the clock speed increases, so does the power consumption of the GPU. This is not just a linear relationship; When the GPU is operating at higher clock speeds, its voltage may also need to be increased to maintain stable operation. Energy consumption is proportional to the square of the voltage, so even a slight increase in voltage will significantly increase energy consumption.

Over time, in order to meet the increasing demands for graphics and computing, GPU have undergone advancements by incorporating multiple dedicated hardware units designed for specific tasks. These specialized units enhance efficiency, reduce power consumption, and in certain cases offer superior performance compared to general-purpose computing units. Each dedicated unit serves a distinct purpose and typically exhibits higher efficiency and lower energy consumption when executing

DOI: 10.54254/2755-2721/49/20241253

specific tasks. By leveraging these dedicated hardware components, modern GPU are capable of delivering exceptional performance across various application scenarios [4].

When designing a GPU, it is critical to ensure efficient communication between individual components, as communication bottlenecks can limit overall performance. To meet this demand for high-speed, high-bandwidth communication, GPU designers use specialized interconnection technologies.

3. Research on optimizing GPU performance and energy efficiency

For the area about design the GPU. The research on GPU load balancing scheduling model in distributed heterogeneous computing framework [5]. This article focuses on how to utilize and integrate the computing power of GPU in the field of high-performance computing and deep learning more effectively. This paper pays special attention to how to achieve efficient GPU management and scheduling in CPU-GPU heterogeneous cluster environment. While Spark has added support for GPU acceleration, Flink does not yet support GPU task scheduling. In order to solve the problem of workload imbalance between different hardware and GPU models in heterogeneous cluster, this paper proposes a multi-GPU load balancing scheduling model named MLSM. The MLSM model adopts fine-grained task mapping mechanism, unified device resource management mechanism, feedback based asynchronous flow adjustment strategy, and resource-aware GPU task scheduling strategy to achieve workload balancing among heterogeneous Gpus. The experimental results show that MLSM model based on Spark framework has excellent performance and can effectively integrate GPU into distributed computing framework.

Special cache designs can also make power consumption even lower, based on research data. To solve the problem of power consumption, a data-dependent GPU power management method (DDPM) is proposed, which reduces the power consumption of GPU system by optimizing thread allocation and cache displacement strategy. The experimental results show that DDPM improves the hit rate of L1 cache by 7% and reduces the DRAM data transmission by 8.1% compared with the shared awareness data management method. The energy efficiency of MC-aware-ORI, MC-aware-LoSe and MC-aware-SiOb methods increased by 2.7%,2.65% and 8.67% respectively [6].

Low power clock tree: clock gating technology is adopted [7]. By analyzing the performance and structure of the clock tree of the GPU module, this paper proposes a new optimization method, that is, to further reduce the dynamic power consumption of the GPU module by clustering the triggers in the clock network on the premise of ensuring the performance of the GPU. The method of trigger cluster is based on minimum spanning tree algorithm, and threshold distance and threshold capacitance are used to constrain the scale of trigger cluster. Compared with the GPU module with common low power optimization methods, the clock tree dynamic power consumption of the GPU module for clustering is reduced by 5%, and the switching power consumption is reduced by 8%. The trigger clustering algorithm adopted in this study is based on the principle of nearby grouping on the premise of not changing the original physical location of the trigger. The algorithm adopts TCL language and can be integrated into the EDA tool, which has strong practicability. This innovative approach realizes the optimization of dynamic power consumption of GPU modules by clustering triggers, and provides a new possibility and direction for low power design in integrated circuit design.

ChatGPT runs on a massive computing cluster managed by OpenAI, which consists of multiple high-performance Gpus. These Gpus are specifically optimized for running deep learning models and are capable of quickly processing a large number of computational tasks in parallel. The GPU models used by OpenAI may change over time to accommodate evolving computing needs and technological advances. ChatGPT is based on the Large Language model (LLM) launched, this technology attracted attention, and then Microsoft, Google and other technology companies in the field of AI competition. LLM computing power demand is huge, in response to this NVIDIA launched a new GPU - NVIDIA H100 NVL. The H100 is based on Nvidia's Hopper architecture and uses a Transformer engine. It has 94GB of memory and is equipped with a PCIE H100 GPU with dual GPU NVLINK, which can handle GPT-3 with 175 billion parameters. Compared to the HGX A100 for ChatGPT, a standard server with

DOI: 10.54254/2755-2721/49/20241253

four pairs of H100 and dual NVLINK can process up to 10 times faster. Huang says it can reduce the processing cost of large language models by an order of magnitude [8].

A large number of component interconnections can optimize GPU performance and power consumption in equal proportion, especially in some specific scenarios. Article: "Design and Implementation of GPU Resource Management Components in Star Ring Container Platform", introduces a scheme called Krux, which allows multiple workloads to share physical Gpus by integrating GPU sharing function in Kubernetes to improve GPU utilization and reduce task average latency [9]. Krux implements a GPU resource management component including resource discovery module, scheduling plug-in module, resource limit module and container runtime module based on Kubernetes extension mechanism. This solution allows for more flexible and efficient use and sharing of GPU resources in the Kubernetes platform, effectively improving GPU utilization, reducing the average task latency, and helping enterprises reduce the cost of hardware resources. This GPU resource management component has been integrated into Starring's container platform version 3.0, and has been widely used in the big data and artificial intelligence platform supported by the platform, showing the value and potential of practical applications.

4. Challenges in GPU Architecture and Design

GPU architecture and design face many challenges in terms of performance and power consumption optimization. With the slowing down of Moore's Law, it is increasingly difficult to reduce the size of transistors, while increasing power consumption and heat dissipation issues. Power wall, i.e. increasing frequency and number of cores can lead to exponential growth in power consumption, placing higher demands on cooling and power management; Storage bottleneck, memory bandwidth may limit GPU performance; Coordination issues in heterogeneous integration, how to efficiently distribute tasks among CPU, GPU, and other hardware accelerators for optimal performance and energy efficiency; As well as adaptability to new technologies and materials, applications such as 3D stacking, 2D materials, and quantum dot technology may face technical and cost challenges in practical designs [10].

5. Conclusion

On the whole, as computational demand continues to grow, GPU have become an indispensable part of many domains, including deep learning, graphics rendering, and high-performance computing. However, with the increasing energy consumption, the optimization of performance and energy efficiency has become a focal point of current research. Multi-level power optimization is essential to address this issue. Merely increasing the frequency can lead to exponential growth in power consumption; hence, GPU architectures might lean towards more layered designs, where different layers cater to varied application scenarios and power requirements. Meanwhile, GPU integrated with central processing units have been around for a while, but the integration of more function-specific hardware, like AI accelerators, dedicated graphics rendering units. With GPU allows for more flexible resource allocation and power management based on workload requirements. Notably, advancements like 3D stacking technologies, as well as novel semiconductor materials and techniques such as 2D materials and quantum dot technologies, might usher in higher performance and reduced power consumption for GPU. Additionally, utilizing AI technologies to dynamically adjust GPU operating parameters and the co-design of software and hardware are believed to be pivotal for future enhancements in GPU performance and efficiency. From materials to architecture, and the integration of software and hardware, every aspect holds vast potential and research possibilities.

In conclusion, the evolution of GPU designs is guided by the intertwining principles of performance enhancement and energy conservation. The relentless pursuit of computational power, coupled with a global emphasis on energy efficiency, mandates continual advancements in GPU architecture and design. By delving into the intricacies of GPU components and their impact on performance and energy consumption, we gain insights that are instrumental in fostering the development of more robust and sustainable computing solutions, ultimately contributing to the advancement of science, technology, and industrial production.

DOI: 10.54254/2755-2721/49/20241253

References

[1] Mittal S, Vetter J S. A survey of methods for analyzing and improving GPU energy efficiency. ACM Computing Surveys (CSUR), 2014, 47(2): 1-23.

[2] Anzt H, Tomov S, Dongarra J. On the performance and energy efficiency of sparse linear algebra on GPUs. The International Journal of High Performance Computing Applications, 2017, 31(5): 375-390.

[3] Rodríguez-Borbón J M, Kalantar A, Yamijala S S, et al. Field programmable gate arrays for enhancing the speed and energy efficiency of quantum dynamics simulations. Journal of chemical theory and computation, 2020, 16(4): 2085-2098.

[4] Ali G, Side M, Bhalachandra S, et al. Performance-Aware Energy-Efficient GPU Frequency Selection using DNN-based Models. Proceedings of the 52nd International Conference on Parallel Processing. 2023: 433-442.

[5] Du Li-Fan. Research on GPU Load balancing scheduling model in distributed heterogeneous computing Framework. Hunan University, 2021.

[6] Wei Xiong, Wang Qiuxian, Hu Qian, Yan Kun, Xu Pingping. Research on GPU Power Management Based on Data Dependence. Computer and Networks, 2019, 47(15): 66-72.

[7] Du Wenjing. Physical Design of low power consumption of GPU module based on TSMC6nm process. Xi 'an University of Technology, 2023.

[8] Wen Qiao. 10 times faster GPU for ChatGPT What else to watch at Nvidia GTC Conference?. National Business Daily,2023-03-28(005).

[9] Gong Chenmiao. Design and Implementation of GPU resource management component in Star Ring container Platform. Nanjing University,2021.

[10] Jiao Q, Lu M, Huynh H P, et al. Improving GPGPU energy-efficiency through concurrent kernel execution and DVFS. 2015 IEEE/ACM International Symposium on Code Generation and Optimization (CGO). IEEE, 2015: 1-11.