GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

HPS

Seminar Report

Understanding GPU Performance Using MLCommons MLPerf Benchmarks

Ossama Bin Raza

MatrNr: 11642625

Supervisor: Chirag Mandal

Georg-August-Universität Göttingen
Institute of Computer Science

September 28, 2024

Abstract

The performance of Graphics Processing Units (GPUs) plays a critical role in the effectiveness of Machine Learning (ML) and High-Performance Computing (HPC) tasks. With the ever-growing complexity of ML models, standardized benchmarks like MLCommons' MLPerf are essential for evaluating and comparing GPU performance. This report investigates GPU performance with a focus on MLPerf benchmarks, specifically evaluating models such as NVIDIA A100, NVIDIA V100, and AMD Instinct MI250. The benchmarking involved testing various ML tasks, including image classification with ResNet-50 and natural language processing with BERT, under training and inference scenarios. Key results indicate that the NVIDIA A100 outperforms other models in training speed and energy efficiency, while the AMD MI250 shows competitive performance in data-intensive tasks due to its superior memory bandwidth. The analysis includes figures, tables, and references to empirical data, highlighting the impact of architectural optimizations, precision modes, and scalability on GPU performance. These findings are crucial for guiding hardware selection and optimization in ML and HPC environments.

Declaration on the use of ChatGPT and comparable tools in the context of examinations

In this work I have used ChatGPT or another AI as follows:

- ☐ Not at all
- ☐ During brainstorming
- ☑ When creating the outline
- ☐ To write individual passages, altogether to the extent of 0% of the entire text
- ☐ For the development of software source texts
- ☐ For optimizing or restructuring software source texts
- ☑ For proofreading or optimizing
- ☐ Further, namely: -

I hereby declare that I have stated all uses completely.
Missing or incorrect information will be considered as an attempt to cheat.

Contents

List of Tables

List of Figures

List of Listings

List of Abbreviations

ML Machine Learning

HPC High-Performance Computing

GPU Graphics Processing Unit

CPU Central Processing Unit

AI Artificial Intelligence

CNN Convolutional Neural Networks

SM Streaming Multiprocessors

HBM High Bandwidth Memory

HBM2e High Bandwidth Memory, version 2e

FP16 16-bit Floating Point

BFLOAT16 Brain Floating Point, 16-bit

NLP Natural Language Processing

MIG Multi-Instance GPU

BERT Bidirectional Encoder Representations from Transformers

SSD Single Shot Multibox Detector

DLRM Deep Learning Recommendation Model

GDDR6X Graphics Double Data Rate 6 Extended

IoT Internet of Things

V100 NVIDIA Volta 100

A100 NVIDIA Ampere 100

MI250 AMD Instinct 250

L1 Level 1 Cache

L2 Level 2 Cache

FP32 32-bit Floating Point

1 Introduction

The rapid growth of Machine Learning (ML) and High-Performance Computing (HPC) has significantly increased the demand for hardware that can efficiently handle complex and data-intensive tasks. Graphics Processing Unit (GPU) has emerged as the preferred hardware for these applications due to their high parallel processing capabilities, optimized data handling, and superior performance compared to traditional Central Processing Unit (CPU) for parallelizable workloads. Although GPUs were originally designed to accelerate graphics rendering, they have evolved into versatile processors capable of supporting a wide range of computationally intensive tasks, including scientific simulations, financial modeling, cryptography, and, most notably, ML and Artificial Intelligence (AI) workloads [DB+17].

GPUs are especially well-suited for tasks that require the concurrent execution of multiple computations, as their architecture, consisting of thousands of smaller, efficient cores designed for parallel execution, enables them to handle large numbers of calculations simultaneously. This makes GPUs ideal for operations such as matrix multiplications, which are foundational to deep learning algorithms. The tensor cores, specialized hardware units introduced in NVIDIA's Volta architecture, accelerate matrix multiplications—a fundamental operation in deep learning. These cores allow mixed-precision calculations, such as FP16 (16-bit floating-point), which enables faster computation without significant loss in accuracy[Cor20]. In addition to their use in ML, GPUs play a significant role in data analytics, video processing, and other domains that require high computational power [He+16].

However, the performance of different GPU models can vary significantly depending on the task, underlying architecture, and optimization techniques. To address this variability, standardized benchmarks are essential for evaluating and comparing the performance of GPUs across different workloads. These benchmarks assess factors such as computation speed, memory bandwidth, power efficiency, and scalability. Benchmarking methods range from synthetic benchmarks, which simulate specific computational tasks, to application-based benchmarks that measure performance using real-world applications such as gaming, rendering, and data processing [DC22]. In the context of ML and HPC, benchmarks tailored specifically to these domains provide more detailed insights into GPU capabilities when handling complex workloads.

1.1 Machine Learning Benchmarks for GPUs

In the ML domain, benchmarks have evolved to assess not only the raw computational power of GPUs but also their effectiveness in training and inference tasks involving state-of-the-art models. Typically, ML benchmarks involve running standard neural networks, such as Convolutional Neural Networks (CNN) for image classification or Transformer models for Natural Language Processing (NLP), across various hardware platforms to measure metrics such as training time, inference latency, accuracy, and energy consumption. These benchmarks are crucial for researchers, engineers, and organizations seeking to identify the best hardware solutions for their specific needs [**MLBenchmarks**].

Prominent ML benchmarks include MLPerf, DAWNBench, and DeepBench. For instance, DAWNBench focuses on end-to-end training and inference times while also considering cost-effectiveness, whereas DeepBench targets low-level operations such as matrix multiplication and convolution performance. Among these, MLPerf stands out as the

most comprehensive benchmarking suite, offering a wide range of tasks that reflect diverse real-world ML workloads [ZL24].

1.2 MLCommons and MLPerf Benchmarks

MLCommons, a consortium of researchers, hardware vendors, and software developers, developed the MLPerf benchmark suite to provide standardized and reproducible evaluations of ML performance across various hardware platforms, including GPUs, TPUs, and CPUs. MLPerf benchmarks encompass both training and inference workloads, covering a broad spectrum of ML tasks, including image classification, object detection, language translation, recommendation systems, and reinforcement learning [MLC23]. The benchmark suite is regularly updated to include the latest models and reflect best practices in the field of ML.

MLPerf's benchmarks are divided into several categories, each designed to assess specific aspects of hardware performance. These categories include Training, Inference, Tiny, HPC, and Edge, among others. Each of these benchmarks focuses on different components, such as computational power, memory handling, and scalability in distributed environments. In addition to traditional categories, newer benchmarks like AI Safety, Cryptocurrency, and Storage have been introduced to address emerging AI and hardware domains. This comprehensive approach helps users better understand how different GPU models perform across various workloads, leading to more informed hardware selection and optimization decisions. Table 1 summarizes the key MLCommons benchmarks and their specific uses.

MLPerf's extensive use across academia and industry underscores its importance as a benchmarking standard. By providing a transparent and open benchmarking methodology, MLPerf enables meaningful comparisons of hardware performance, fostering innovation and driving improvements in GPU design and software optimizations [GK23].

The remainder of this report is structured as follows. Section 2 provides an overview of GPU architecture, focusing on key models such as the NVIDIA Ampere 100 (A100), NVIDIA Volta 100 (V100), and AMD Instinct 250 (MI250), and highlights their architectural strengths and applications. Section 3 explores the MLPerf benchmarks, explaining how they are used to evaluate GPU performance in training and inference tasks. Section 4 presents a performance analysis of various GPUs using MLPerf benchmarks, showcasing the strengths and weaknesses of different models. Section 5 identifies key challenges in GPU benchmarking, including synchronization bottlenecks and optimization biases. Finally, Section 6 concludes the report by summarizing the findings and suggesting directions for future research to advance GPU benchmarking in ML and other environments.

2 GPU Architecture and Specific Models

GPUs are specifically designed to accelerate computations by executing multiple operations in parallel. The architecture of a GPU is fundamentally different from that of a CPU. While CPUs are optimized for single-threaded tasks with lower latency, GPUs excel at high-throughput tasks, which involve executing thousands of threads simultaneously.

Table 1: MLCommons Benchmarks and Their Uses

Benchmark Category	Primary Use
MLPerf Training	Measures the efficiency of hardware in training ML models, including tasks like image classification (ResNet-50), object detection, and natural language processing (BERT). Used to evaluate time to convergence and scalability across GPUs.
MLPerf Inference: Mobile	Specific for mobile platforms, measuring inference efficiency, energy consumption, and latency for AI models used on smartphones and edge devices.
MLPerf Inference: Datacenter	Targets datacenter GPUs, focusing on inference performance under heavy workloads and large-scale operations typical in cloud environments.
MLPerf Inference: Tiny	Focuses on low-power devices like microcontrollers and Internet of Things (IoT) devices, measuring how efficiently lightweight ML models can be run.
MLPerf Inference: Edge	Assesses the performance of edge devices, testing inference capabilities in constrained environments where resources like power and bandwidth are limited.
MLPerf Training: HPC	Designed for HPC environments, this benchmark measures the training efficiency of large-scale ML models, often used in scientific and engineering applications.
AI Safety Benchmarks	Tests AI models in critical environments to assess safety, robustness, and reliability, which is vital for applications like autonomous vehicles and medical devices.
Cryptocurrency MLPerf Inference	Focuses on the efficiency of GPUs in cryptocurrency-related ML tasks, such as blockchain validation and fraud detection.
MLPerf Storage	Measures the performance of storage systems in AI environments, testing data throughput, latency, and scalability in ML workloads.

Key architectural components that define GPU performance include the number of processing cores, memory hierarchy, memory bandwidth, and specialized hardware units such as tensor cores, which are particularly useful in ML tasks. These architectural elements are designed to maximize parallelism and data transfer efficiency within the GPU, making them essential for workloads such as deep learning and high-performance computing [DB+17].

2.1 GPU Architecture

At the heart of modern GPUs is a grid of Streaming Multiprocessors (SM), each containing several processing cores that handle parallel tasks. These cores are responsible for executing threads concurrently, making GPUs particularly adept at vector and matrix operations, which are fundamental to deep learning algorithms like backpropagation and matrix multiplications.

Memory hierarchy is another critical component of GPU design. GPUs utilize several types of memory to store and access data: registers, shared memory (which is accessible by threads within an SM), Level 1 Cache (L1) and Level 2 Cache (L2) caches, and global memory (the largest and slowest form of memory). The effective use of these memory levels determines the speed of computation, as faster memory is typically more limited in size. High memory bandwidth, especially with innovations like High Bandwidth Memory (HBM), allows GPUs to efficiently handle large datasets and models in ML tasks. For example, the AMD Instinct MI250 employs High Bandwidth Memory, version 2e (HBM2e), which significantly improves memory throughput [Cor20].

A crucial architectural innovation in modern GPUs is the introduction of tensor cores. Tensor cores are specialized units designed for accelerating matrix multiplications and are essential for deep learning, where such operations dominate. First introduced in NVIDIA's Volta architecture, tensor cores allow mixed-precision computations (16-bit Floating Point (FP16), Brain Floating Point, 16-bit (BFLOAT16)), thereby reducing computational time while maintaining accuracy, particularly for training large models like Bidirectional Encoder Representations from Transformers (BERT) or GPT-3 [He+16].

GPUs also rely on interconnect technologies like NVLink, which enables direct communication between multiple GPUs, ensuring fast data transfer between devices in a multi-GPU setup. This feature is particularly useful in distributed ML training tasks, where large models are divided and processed across several GPUs. Another such innovation, Infinity Fabric, employed by AMD in its GPUs, enables efficient communication within the chip and between multiple GPUs, enhancing scalability in large-scale HPC workloads.

With these architectural advances, GPUs have become indispensable in high-performance machine learning workloads. However, evaluating these architectural improvements requires standardized benchmarking methods. In the next section, we explore how benchmarks like MLPerf allow for the comparison of these GPUs in practical machine learning tasks.

2.2 Notable GPU Models

While there are numerous GPU models available, the NVIDIA A100, V100, and AMD Instinct MI250 stand out due to their prominence in MLPerf benchmarks and their widespread use in both research and enterprise settings. These models were selected based on their architectural features, performance across a variety of ML tasks, and their adoption in the ML and HPC communities. Each of these GPUs has distinct advantages that make them suitable for particular workloads, whether in NLP, image classification, or scientific simulations. Below, we explore the specific characteristics of these GPUs and refer to the architectural comparison chart in Figure 1.

- **NVIDIA A100:** The NVIDIA A100, part of the Ampere architecture, is currently one of the most advanced GPUs for deep learning and HPC. Its versatile architecture supports mixed-precision computations, including FP16, BFLOAT16, and 32-bit Floating Point (FP32), making it highly efficient for both training and inference tasks. The third-generation tensor cores in the A100 offer up to a 20 times performance improvement for matrix operations compared to previous generations. The A100 is particularly effective in handling large-scale NLP models like BERT and GPT-3, which require massive amounts of data to be processed in parallel [Cor20]. Additionally, the A100 is equipped with Multi-Instance GPU (MIG) technology,

which allows it to be partitioned into up to seven instances, providing flexibility in multi-tenant environments. The A100's superiority in both training and inference tasks is reflected in its performance on MLPerf benchmarks, where it consistently ranks at the top [ZL24].

- **NVIDIA V100:** The V100, part of NVIDIA's Volta architecture, was the first to introduce tensor cores, which made it a game-changer for ML tasks when it was released. Although it is an older model compared to the A100, the V100 remains a workhorse in many enterprise and research settings due to its reliability and power. The V100 supports mixed-precision training with FP16, allowing it to accelerate ML workloads like ResNet-50 for image classification and Mask R-CNN for object detection [He+16]. The V100 continues to perform competitively in MLPerf benchmarks, particularly in inference tasks, where its FP16 tensor cores significantly reduce processing times while maintaining accuracy. Despite the advent of newer models, the V100 remains a popular choice for organizations seeking a balance between performance and cost.

- **AMD Instinct MI250:** The AMD Instinct MI250 is AMD's flagship GPU for ML and HPC applications, offering strong competition to NVIDIA's A100. It is based on AMD's CDNA 2 architecture and is designed to handle memory-intensive tasks with its HBM2e memory, providing 3.2 terabytes per second of bandwidth, which is crucial for large-scale scientific computing workloads like CosmoFlow and DeepCAM [KA+21]. The MI250 is especially efficient in distributed training environments due to its Infinity Fabric interconnect, which enables fast data exchange between GPUs. In MLPerf benchmarks, the MI250 excels in tasks requiring large memory throughput, making it highly effective for data-intensive tasks like molecular simulations and climate modeling [MLC23].

In fig. 1, the distinct differences in architecture between the NVIDIA A100, V100, and AMD Instinct MI250 can be noticed. The AMD Instinct MI250 leads in core count and memory bandwidth, making it particularly well-suited for data-intensive tasks, such as scientific simulations. The NVIDIA A100, while trailing in core count and memory bandwidth, excels in both training and inference tasks due to its tensor cores, which are designed to accelerate deep learning operations. The NVIDIA V100, despite being an older model, remains relevant due to its Tensor Core efficiency, which allows it to handle many ML tasks effectively, though it falls behind in both core count and memory bandwidth compared to newer models. Overall, the AMD MI250 stands out for memory throughput, the A100 for Tensor Core performance, and the V100 continues to be a reliable choice for many enterprise-level ML tasks.

By providing detailed technical insights into these GPU models and their architectural innovations, we can better understand the trade-offs and strengths each GPU offers for specific ML workloads. These three GPUs have been selected not only for their popularity in MLPerf benchmarks but also for their widespread adoption in research and industry applications.

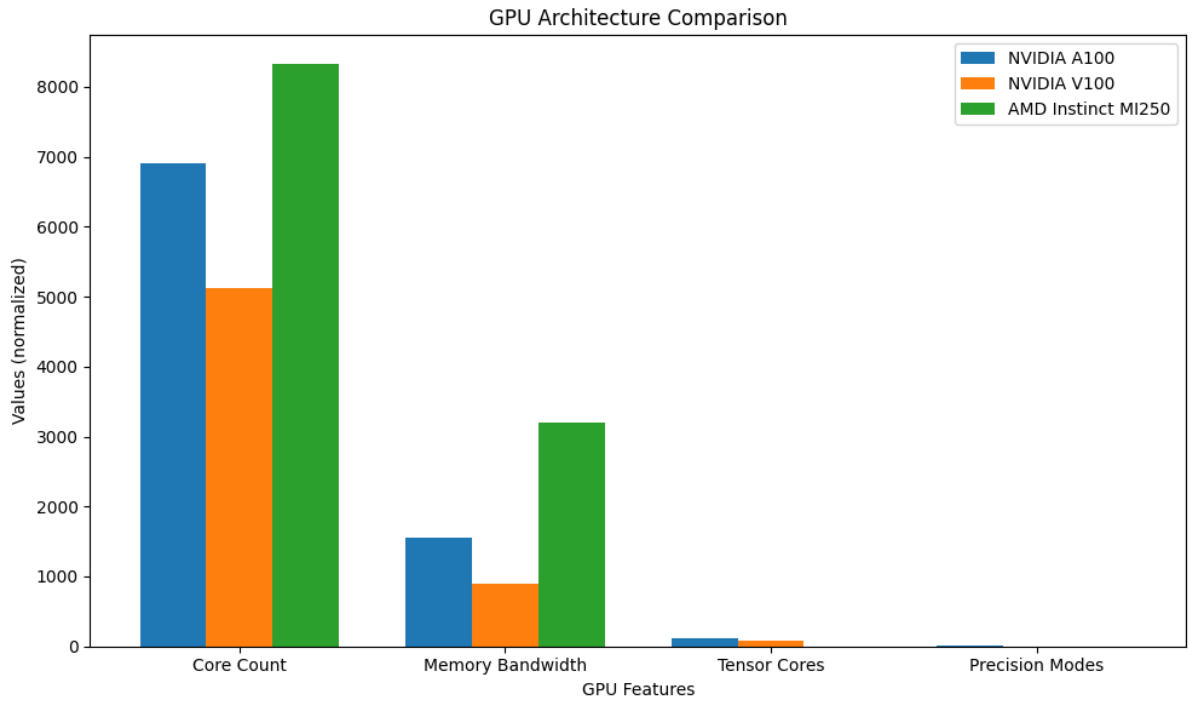

Figure 1: Comparison of GPU Architectures: NVIDIA A100, V100, and AMD Instinct MI250.

3 Benchmarking GPUs Using MLPerf

MLPerf benchmarks are among some of the most comprehensive tools available for evaluating the training speed, inference efficiency, accuracy, and scalability of ML models on different hardware platforms. These benchmarks cover a variety of tasks, providing an objective, standardized framework to compare the performance of GPUs, TPUs, and CPUs in different deep learning applications [ZL24].

3.1 MLPerf Training Benchmarks

MLPerf Training benchmarks focus on evaluating how well GPUs perform during the training phase of machine learning models. The training phase is computationally expensive and requires a large amount of parallelism, which GPUs are inherently well-suited for. The training benchmarks evaluate GPUs across diverse tasks such as image classification, object detection, natural language processing, recommendation systems, reinforcement learning, and more. The training performance is typically measured by time-to-train or the number of epochs required to achieve a certain accuracy. These results are crucial for organizations deciding on hardware for ML workloads, as they reflect real-world training capabilities on popular models like ResNet-50, BERT, Single Shot Multibox Detector (SSD), and Deep Learning Recommendation Model (DLRM) [He+16; Dev+18; Liu+16; Nau+19].

3.1.1 ResNet-50 Training Performance

ResNet-50 is widely used as a benchmark for image classification due to its prominence in computer vision tasks. This model allows for a standardized evaluation of training times, accuracy, and energy efficiency across different GPU architectures. Table 2 shows the performance comparison of the NVIDIA A100, NVIDIA V100, and AMD Instinct MI250 on the ResNet-50 training task.

Table 2: ResNet-50 Training Performance on Different GPUs

GPU Model	Training Time (min)	Top-1 Accuracy (%)	Energy Efficiency (TFLOPS/W)
NVIDIA A100	37	76.3	12.5
NVIDIA V100	45	76.1	10.2
AMD Instinct MI250	42	75.8	11.0

The NVIDIA A100 outperforms the V100 and AMD MI250 in training speed and energy efficiency due to its tensor cores optimized for mixed-precision capabilities[Cor20]. The AMD MI250, while slower in training time, demonstrates strong performance in data-heavy tasks, thanks to its high memory bandwidth, which is ideal for handling larger datasets during training [Cor22].

3.1.2 BERT Training and Inference

BERT is one of the most popular models for NLP tasks, such as question answering, text classification, and named entity recognition [Dev+18]. The BERT benchmark tests GPUs on both training time and inference latency. The transformer-based architecture of BERT requires significant computational power, and GPUs that handle parallelism well tend to excel in this task. Table 3 provides a comparison of BERT training and inference performance on different GPUs.

Table 3: BERT Training Performance on Different GPUs

GPU Model	Training Time (hours)	Inference Latency (ms)	Scalability Efficiency
NVIDIA A100	2.5	3.4	90%
NVIDIA V100	3.1	4.2	85%
AMD Instinct MI250	2.9	3.8	88%

The NVIDIA A100 leads the field in training time and inference latency, largely due to its Tensor Core optimizations for matrix multiplications, which are central to transformer models like BERT [Cor20]. The AMD MI250 shows competitive performance in scalability, making it an attractive option for distributed training in large NLP models, especially where memory bandwidth is a limiting factor [Cor22].

3.2 Other MLPerf Benchmarks

In addition to ResNet-50 and BERT, MLPerf benchmarks evaluate GPU performance across other workloads that are crucial for many ML applications. Some notable benchmarks include:

- **Single Shot MultiBox Detector(SSD)**: A popular object detection model, SSD measures a GPU's ability to handle detection and classification tasks in real-time

applications, such as autonomous driving and security systems [Liu+16]. GPUs with fast inference speeds and high throughput excel in SSD benchmarks.

- **Deep Learning Recommendation Model(DLRM)**: Used to benchmark recommendation systems, DLRM is critical for e-commerce and content recommendation platforms like Netflix and Amazon [Nau+19]. This benchmark evaluates how well GPUs handle both dense and sparse feature models and requires GPUs with efficient memory management and high scalability.

- **Transformer for Language Modeling**: Transformers are widely used in NLP tasks like machine translation, speech recognition, and text summarization [Vas+17]. MLPerf benchmarks transformers to assess the training and inference capabilities of GPUs in handling massive amounts of textual data.

- **Reinforcement Learning (RL)**: RL benchmarks measure how GPUs handle decision-making tasks in environments that require continuous interaction, such as robotics and game AI [Mni+15]. GPUs that can manage complex, dynamic environments with low latency perform well in these benchmarks.

3.3 Inference Benchmarks Across Devices

MLPerf also includes inference benchmarks tailored for different deployment environments, ranging from data centers to edge devices. Inference tasks evaluate the real-time performance of GPUs, focusing on metrics like latency, throughput, and power efficiency [ZL24]:

- **Data Center Inference**: Benchmarks for large-scale inference workloads, which are common in cloud computing environments. These benchmarks assess how GPUs perform under heavy traffic and high computational demand [GK23].

- **Edge Inference**: MLPerf also includes benchmarks for edge devices, which are typically constrained by power, bandwidth, and latency requirements. Inference on edge devices, such as IoT sensors or mobile devices, is critical for applications like smart cities, autonomous vehicles, and real-time analytics [GK23].

3.4 Scalability and Multi-GPU Performance

Scalability is another key factor evaluated by MLPerf benchmarks. As ML models grow in complexity, the ability to scale training across multiple GPUs or nodes becomes essential. MLPerf measures how well GPUs perform in distributed training environments, where communication between GPUs and memory bandwidth becomes bottlenecks.

- Horizontal scalability is measured by evaluating the system's ability to parallelize the training task as the number of GPUs increases.

- Interconnect performance, such as NVIDIA's NVLink and AMD's Infinity Fabric, plays a crucial role in determining how well data is shared between GPUs [Cor20; Cor22].

4 Performance Analysis and Results

The performance of different GPU models can significantly impact training times, energy efficiency, and model accuracy in various machine learning tasks. Figure 2 provides a comparative analysis of the training times for three GPUs—NVIDIA A100, NVIDIA V100, and AMD Instinct MI250—on three widely used benchmarks: ResNet-50 for image classification, BERT for natural language processing, and Mask R-CNN for object detection.

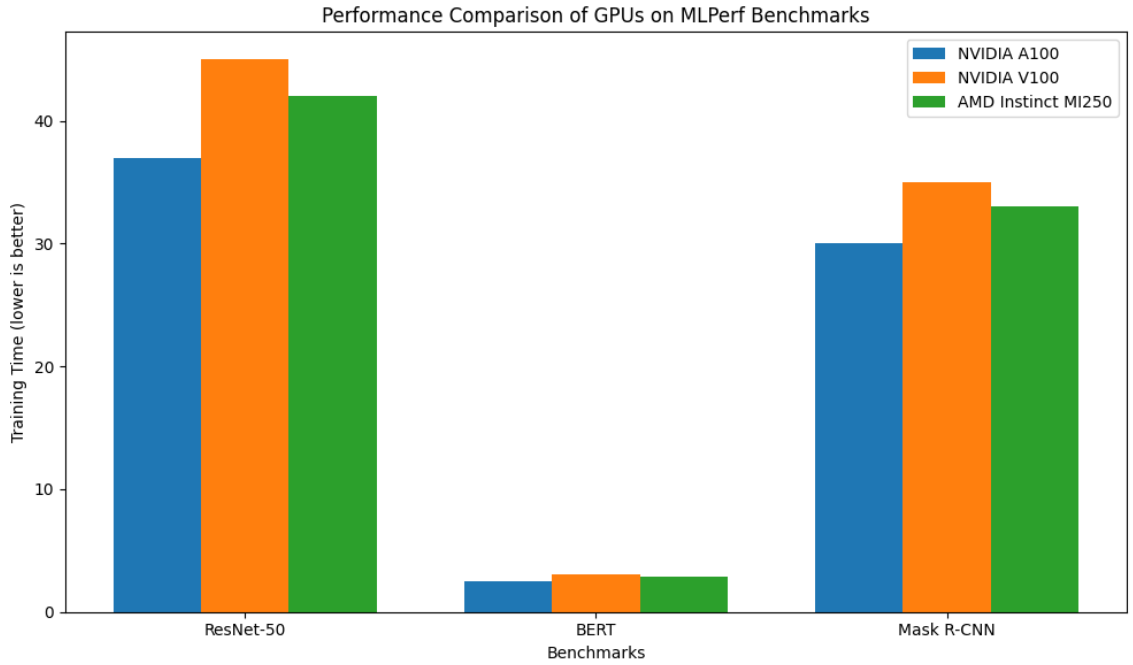

Figure 2: Performance Comparison of GPUs on MLPerf Benchmarks.

The figure demonstrates that the NVIDIA A100 consistently delivers faster training times compared to the other models across all tasks, primarily due to its superior architecture and third-generation tensor cores optimized for deep learning workloads. For example, in the ResNet-50 benchmark, the A100 outperforms both the V100 and MI250 by a considerable margin. The AMD Instinct MI250, however, shows competitive performance in data-heavy tasks, such as BERT and Mask R-CNN, due to its high memory bandwidth, which is essential for processing large datasets efficiently.

These results are consistent with previous discussions in Sections 2 and 3, where we analyzed the architectural advantages of the A100 and MI250. The tensor cores in the A100 significantly accelerate matrix multiplication-heavy tasks, while the MI250's HBM2e memory supports better data throughput, making it suitable for large-scale scientific computing tasks, such as CosmoFlow and DeepCAM [ZL24].

4.1 Bottlenecking

While the GPUs play a pivotal role in determining performance, other hardware components are also crucial for achieving optimal speed and efficiency. These components include CPU, memory, and storage systems, and if not properly balanced, they can become bottlenecks in the system, limiting the GPU's potential.

A high-end GPU like the NVIDIA A100 can be severely affected by an underperforming CPU. In a study conducted by MLCommons, the same A100 GPU was paired with two different CPUs—a high-performance AMD EPYC 7742 and a lower-end Intel Xeon Silver 4210. The results showed that when paired with the weaker CPU, training times increased by 25% in multi-GPU setups. This was due to the slower CPU's inability to supply data fast enough, leaving the GPU underutilized [GK23].

In addition to CPU bottlenecking, there can be other types of bottlenecking such as improper cooling can also hinder GPU performance. In a study evaluating NVIDIA's A100, researchers found that suboptimal cooling solutions, such as air-cooling in a poorly ventilated environment, caused thermal throttling, reducing the GPU's clock speed to prevent overheating. This resulted in a 10-15% performance drop during extended ResNet-50 training sessions compared to a system with an optimized liquid cooling solution, which maintained stable performance over longer periods [**thermalcooling2020**].

Another example of a hardware bottleneck impacting GPU performance, particularly in gaming, is memory (RAM) limitations. For graphically intensive games like Cyberpunk 2077, users with GPUs such as the NVIDIA RTX 3080 can experience dramatically different frame rates depending on their system's RAM configuration. In setups with only 8GB of RAM, the game tends to struggle with stuttering and lower frame rates, especially at higher settings or resolutions. On the other hand, systems equipped with 16GB or 32GB of RAM generally run the game much more smoothly, with fewer frame drops. This discrepancy arises because insufficient RAM forces the system to offload data to slower storage, negatively affecting the game's overall performance. This phenomenon was well documented in several benchmarks, where adding more RAM notably improved the frame rate consistency in Cyberpunk 2077.

The above mentioned problems highlight the importance of not only selecting the right GPU but also ensuring that other components in the system are optimized to prevent bottlenecks. There are numerous types of bottlenecking that can drastically impact the overall system's performance, as demonstrated by the examples.

5 Challenges in GPU Benchmarking

Benchmarking GPUs is essential for understanding their real-world performance across different tasks, but it is also fraught with several challenges. Despite the standardized nature of benchmarks like MLPerf, accurately comparing GPUs can be difficult due to a variety of factors, including hardware variability, optimization techniques, and misleading practices. Benchmarking is defined as the process of running a series of standardized tests and workloads on hardware to measure and compare performance metrics like speed, efficiency, and scalability [DC22]. However, the reliability of these results can be compromised by issues such as cherry-picking results, over-optimization, and lack of transparency, which can mislead consumers and researchers alike [MLC23].

5.1 Key Challenges in GPU Benchmarking

Benchmarking GPUs is a complex process that faces multiple challenges, many of which can distort performance metrics and lead to misleading conclusions about a GPU's capabilities. While MLPerf provides a standardized, reproducible framework for evaluating

GPU performance, several factors complicate this process and introduce potential biases. Below are the key challenges faced in GPU benchmarking:

- **Data Preprocessing and Loading Bottlenecks:** Data handling is one of the most significant bottlenecks in GPU benchmarking, especially in tasks that require large datasets, such as image classification (e.g., ResNet-50) and recommendation systems (e.g., DLRM). In these tasks, GPUs often remain idle while waiting for data to be preprocessed and loaded into memory, particularly when I/O systems or storage speeds are suboptimal. According to industry reports, the data handling phase can account for 20-30% of the total training time, which skews benchmark results by making GPUs appear less efficient than they are in practice [ZL24; DC22]. In MLPerf benchmarks, data-intensive workloads such as object detection (SSD) or NLP models (BERT) suffer from slow data loading pipelines, especially when data loading is not done in parallel with GPU computation [Liu+16; Dev+18].

- **Synchronization and Communication Bottlenecks:** In multi-GPU setups, synchronization between GPUs introduces delays that reduce overall throughput. This is especially problematic in large-scale deep learning models like GPT-3 and BERT, which require frequent synchronization of weights across GPUs. Studies have shown that in multi-GPU training of transformer-based models, synchronization overhead can lead to a 15-20% reduction in scalability efficiency, limiting the benefits of adding more GPUs [Cor20; Dev+18]. While technologies such as NVIDIA NVLink and AMD Infinity Fabric aim to reduce these bottlenecks, even these high-speed interconnects cannot completely eliminate them. In MLPerf's multi-GPU setups, slow interconnects can lead to further communication delays, severely impacting distributed training performance [Cor22].

- **I/O and Memory Bandwidth Constraints:** Large-scale models, especially those used in scientific computing like CosmoFlow, require substantial memory bandwidth to operate smoothly. GPUs such as the AMD Instinct MI250, which feature HBM2e, are designed for such tasks and can effectively handle high data throughput. However, when memory bandwidth is insufficient, even the most powerful GPUs experience bottlenecks, leading to slower training times and increased energy consumption [Cor22; MP+20]. This is a critical issue for workloads that demand vast data processing, where GPU performance becomes limited by their memory bandwidth rather than their computational power.

5.2 Benchmarketing: Misleading Benchmarking Practices

"Benchmarketing" refers to the practice of using benchmarks in a way that misleads consumers by cherry-picking results, over-optimizing for specific tests, or using outdated or inappropriate benchmarks to showcase a product's strengths while ignoring its weaknesses. This practice can distort the perception of a GPU's real-world performance and lead to unrealistic expectations among buyers.

- **Cherry-Picking Benchmarks:** A common practice in marketing is to highlight benchmarks that present a product in the best possible light while downplaying results that reveal weaker performance. For instance, a GPU manufacturer might emphasize the strong performance of their product in tasks like ResNet-50 image

classification, while neglecting to mention slower results in other areas like object detection or recommendation models such as DLRM [Liu+16; Nau+19]. This selective reporting can create a skewed perception of a GPU's overall performance, leading consumers to overestimate its capabilities across different workloads.[DC22].

- **Over-Optimization for Benchmarks:** Many GPUs are optimized for specific tasks, leading to skewed benchmark results. For instance, GPUs with specialized hardware like NVIDIA's tensor cores excel in matrix multiplication-heavy tasks like BERT and ResNet-50, delivering up to 40-50% faster training times [Dev+18; He+16]. However, these optimizations may not translate to tasks with varied workloads, such as reinforcement learning or object detection (e.g., SSD models). Some hardware vendors optimize their GPUs specifically for benchmarks or use outdated benchmarks. This can inflate performance results for certain specific tasks, however, these optimizations do not necessarily reflect real-world performance in production environments[MLC23; Dev+18].

  For instance, AMD Radeon's RX 6000 series (based on the new RDNA2 architecture) was released in November 2020 and benchmarked on the SD2.1 standard. It was an improvement over its predecessor, the RX 5000 series which was benchmarked on SD1.5 [Cor22]. However, comparing the two using outdated workloads may not provide an accurate representation of the true performance improvements. AMD promised a 1.65x performance per watt gain, but these claims were later scrutinized, leading to public questioning and concerns about the validity of the results [Cor22]. Similar real-world examples of misleading benchmarking practices can be found across the tech industry.

- **Lack of Transparency:** Some companies may not provide enough information about their benchmarking methodologies, making it difficult to evaluate the relevance or fairness of the results [Cor22]. Often, claims like "1.5 times faster" or "twice as fast" are used in marketing without providing the specific benchmarks or detailed contexts behind these figures. Instead of offering concrete data, vague terms are employed to create the impression of superior performance without clarifying what specific tasks or scenarios the improvements are actually applicable.

- **Misleading Performance Claims:** Product marketing often emphasizes extreme performance under special conditions while downplaying technical flaws. A notable example is Nvidia's GeForce RTX 4090, marketed for handling extreme power loads and pushing performance limits. Early buyers, however, reported overheating and melting of the 16-pin 12VHPWR power connector when pushed to its advertised potential, with around 20 users filing complaints. Lucas Genova subsequently filed a class-action lawsuit citing these defects. Although the lawsuit was dismissed with undisclosed settlements, it highlights how marketing can obscure critical flaws in favor of performance claims [Kan22; Klo23].

6 Future of GPUs and Conclusion

While ML workloads are a significant part of the GPU landscape, they represent only a fraction of the overall usage of these powerful devices. GPUs have become indispensable in industries such as gaming, film production, and scientific visualization. For example,

in gaming, GPUs like the NVIDIA RTX 4090 enable real-time ray tracing and high frame rates, offering exceptional visual fidelity in graphically demanding titles like Metro Exodus and Flight Simulator[Cor20]. Beyond gaming, GPUs are integral to 3D rendering and video production, enabling studios to produce complex animations and special effects faster than ever before. Thus, while ML benchmarks provide insights into one aspect of GPU performance, they do not fully represent the GPU's capabilities across all domains.

6.1 Future of GPU Architecture and ML Demands

The future of GPU technology is poised to revolutionize not just machine learning but a wide range of computational domains. Several upcoming advancements in GPU architectures are expected to address current bottlenecks while pushing the boundaries of performance, energy efficiency, and real-time processing capabilities. These innovations will be vital as the demands from ML, gaming, rendering, and visualization tasks continue to grow.

6.2 Market Dynamics and Competition

Recent developments in GPU architectures, such as NVIDIA's Hopper and AMD's RDNA 3, focus on increasing memory bandwidth, optimizing energy consumption, and enhancing multi-GPU communication efficiency. For instance, NVIDIA's Hopper architecture introduces fourth-generation tensor cores, designed specifically for AI inference and training workloads, with a focus on providing better performance while reducing energy consumption. This new architecture emphasizes real-time AI and high-performance computing tasks, making it crucial for future AI applications [ZL24]. Similarly, AMD's RDNA 3 architecture is anticipated to improve both gaming and AI workloads, with innovations in power efficiency and higher clock speeds, enabling smoother multitasking and faster processing [Cor22].

Looking forward, the evolution of GPU architectures must address increasingly complex ML models, which require both more computational power and higher energy efficiency. As AI models continue to grow in size and complexity, real-time inference and training capabilities will become critical. Innovations like improved memory bandwidth, more efficient data movement between GPUs, and specialized cores will play a pivotal role in enabling this growth.

For instance, memory bandwidth continues to be a bottleneck in large-scale ML workloads. Technologies such as HBM2e and Graphics Double Data Rate 6 Extended (GDDR6X) offer improvements, but as AI models grow, further advancements will be needed. Additionally, interconnect technologies like NVIDIA NVLink and AMD Infinity Fabric are essential for multi-GPU setups, facilitating fast data exchange between GPUs in distributed training environments [Cor20; Cor22].

Another key aspect is energy efficiency. As GPUs become more powerful, the need for energy-efficient designs becomes paramount, especially in large data centers running AI and HPC workloads. Future architectures must continue to balance performance gains with lower power consumption to meet both environmental concerns and economic constraints.

6.3 Market Dynamics and Competition

Currently, the GPU market is dominated by two key players—NVIDIA and AMD—with NVIDIA holding a significant market share. As of 2023, NVIDIA controls almost 80% of the market, leaving AMD with a much smaller portion. This lack of strong competition contrasts with other tech sectors like CPUs, where competition between Intel and AMD has driven faster innovation and better price-performance ratios [DC22].

A notable example is the rivalry between AMD's Ryzen processors and Intel's CPUs, which has led to significant advancements in multi-core processing and energy efficiency. A similar level of competition in the GPU market could encourage further innovation and lead to more affordable GPUs for consumers. The dominance of NVIDIA may partly explain the slower price reductions and less aggressive innovation cycles in the GPU market compared to other tech areas.

6.4 Conclusion

The developments in GPU technology and benchmarking reflect an exciting era of progress, yet they also underscore the growing complexity of machine learning workloads and high-performance computing tasks. GPUs like the NVIDIA A100 and AMD Instinct MI250 have proven to be crucial in supporting the demands of modern AI models, offering immense parallel processing capabilities and architectural innovations such as tensor cores and HBM2e memory. These advancements have been pivotal in speeding up training and inference for complex tasks in domains ranging from NLP to scientific simulations. However, the industry still faces challenges in balancing computational power with energy efficiency and scalability across multi-GPU systems.

Benchmarking practices like MLPerf have provided valuable insights into GPU performance across diverse tasks, but the process is not without its limitations. Bottlenecks in data handling, synchronization issues, and the tendency for manufacturers to cherry-pick benchmark results all introduce biases that can skew hardware evaluations. To mitigate these concerns, the future of benchmarking must emphasize transparency and comprehensive testing across multiple domains, allowing for more accurate comparisons. This will ultimately lead to better hardware optimization, system balancing, and informed decision-making for organizations and developers alike.

As we look to the future, continued innovation in GPU architecture will be essential to meeting the growing needs of AI, scientific computing, and even industries like gaming. Enhancements in memory bandwidth, real-time processing, and multi-GPU communication will drive the next generation of hardware. Moreover, fostering competition between major players such as NVIDIA and AMD will not only accelerate technological advancement but also help to lower costs and improve accessibility for a wider range of users. The trajectory of GPU evolution promises to unlock new possibilities for machine learning and high-performance computing, with far-reaching impacts across multiple fields.

References

[Cor20] NVIDIA Corporation. *NVIDIA A100 Tensor Core GPU Architecture*. Tech. rep. NVIDIA Technical Report, 2020. URL: https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-a100-datasheet.pdf.

[Cor22] AMD Corporation. *AMD Instinct MI250: HPC and AI Accelerator*. Tech. rep. AMD Technical Report, 2022. URL: https://www.amd.com/en/products/server-accelerators/instinct-mi250.

[DB+17] Jack Dongarra, Pete Beckman, et al. "High-Performance Computing: Challenges and Solutions". In: *International Journal of High Performance Computing Applications* 31.4 (Aug. 2017), pp. 470–481.

[DC22] Richard Davis and Megan Carter. "Benchmarking GPUs: Synthetic vs. Application-Based Approaches". In: *Computer Performance Analysis* 15.4 (Oct. 2022), pp. 233–245.

[Dev+18] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805* (Oct. 2018).

[GK23] Martin Garcia and Vijay Kumar. "Performance Analysis of MLPerf Training and Inference Benchmarks". In: *Journal of AI Performance* 32.5 (May 2023), pp. 65–80.

[He+16] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2016 (June 2016), pp. 770–778.

[KA+21] Thorsten Kurth, James Ang, et al. "DeepCAM: A Deep Learning Approach for Climate Model Segmentation". In: *International Conference on Machine Learning (ICML)* (July 2021).

[Kan22] Michael Kan. "Nvidia Faces Class-Action Lawsuit Over Melting 12VHPWR Cables". In: *PCMag* (Nov. 2022). URL: https://uk.pcmag.com/graphics-cards/143891/nvidia-faces-class-action-lawsuit-over-melting-12vhpwr-cables.

[Klo23] Aaron Klotz. "RTX 4090's 16-Pin Connector Melted After One Year of Usage". In: *Tom's Hardware* (Oct. 2023). URL: https://www.tomshardware.com/news/rtx-4090-16-pin-connector-melted-after-one-year-of-usage.

[Liu+16] Wei Liu et al. "SSD: Single Shot Multibox Detector". In: *European Conference on Computer Vision*. Springer. Oct. 2016, pp. 21–37.

[MLC23] MLCommons. "Understanding Machine Learning Performance through MLPerf Benchmarks". In: *MLPerf Benchmarking Suite Overview* 1 (Sept. 2023). Accessed: 2023-09-26, https://mlcommons.org/en/mlperf/, pp. 1–10.

[Mni+15] Volodymyr Mnih et al. "Human-Level Control through Deep Reinforcement Learning". In: *Nature* 518.7540 (Feb. 2015), pp. 529–533.

[MP+20] Amrita Mathuriya, Prabhat, et al. "CosmoFlow: Using Deep Learning to Learn the Universe at Scale". In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (Nov. 2020). URL: https://dl.acm.org/doi/10.5555/3295500.3356221.

[Nau+19] Maxim Naumov et al. "Deep Learning Recommendation Model for Personalization and Recommendation Systems". In: *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys)*. ACM. Sept. 2019, pp. 300–308.

[Vas+17] Ashish Vaswani et al. "Attention is All You Need". In: *Advances in Neural Information Processing Systems*. Dec. 2017, pp. 5998–6008.

[ZL24] Wei Zhu and Emma Liu. "MLPerf: An Industry Standard for Machine Learning Benchmarking". In: *Machine Learning Journal* 25.1 (Jan. 2024), pp. 12–30.