GPU resources management for model serving as a service and serverless clouds

Aiza Maksutova

Higher School of Economics, Moscow, Russia

Subject area description

The subject area of GPU resources management for model serving focuses on the management and allocation of computing resources provided by GPUs to support the deployment and operation of machine learning models in production environments. This area encompasses the optimization of GPU usage and capacity to support the real-time delivery of predictions or insights generated by machine learning models, including tasks such as scheduling, load balancing, and resource allocation. It also includes strategies for scaling GPU resources to meet demand, monitoring and troubleshooting performance, and ensuring the reliability and availability of GPU infrastructure for model serving.

Relevance of the field

GPU resource management is becoming increasingly relevant in serverless computing, which is a cloud computing model that enables users to run code without the need for server management. In serverless computing, functions are executed in response to events, and resources are provisioned on-demand, based on the workload requirements. GPUs are well-suited for running compute-intensive workloads in serverless environments, such as machine learning models and data processing tasks. However, managing GPU resources in serverless environments is challenging due to the dynamic nature of workloads and the need for efficient resource utilization. To address this challenge, researchers and practitioners are developing novel approaches for GPU resource management in serverless computing, such as dynamic resource allocation, task scheduling, and workload prediction. These approaches aim to optimize the performance and cost-effectiveness of serverless applications that use GPUs, and have important implications for improving the overall efficiency and scalability of cloud computing.

Research Objectives

- **Objective 1:** Analyze the impact of PCIe contention on GPU resource management and identify the key factors affecting system performance.
- **Objective 2:** Evaluate different mitigation techniques for PCIe contention, such as prioritization, traffic shaping, and QoS, and determine their effectiveness in optimizing system performance
- **Objective 3:** Measure the latency of a DNN model under varying traffic loads and identify the performance bottlenecks in the system.
- **Objective 4:** Investigate the relationship between traffic load and system performance in real-time applications such as self-driving cars and image processing.
- **Objective 5:** Propose and test new strategies for GPU resource management that can improve the efficiency, scalability, and reliability of the system in different use cases.

Proposed methods

To understand the effect of PCIe contention and explore the bottlenecks we implement the below architecture on a NVIDIA Tesla GPU P100 with a maximum bandwidth of 15.75 GB/s.

In our study, we applied a load by transferring data in blocks of size (1024 * 1024) * pressure level * sizeof(float) bytes. To create contention, we utilized a function that spawns multiple threads equal to the pressure level parameter, which is externally set and increases depending on how much load we need to apply. We then called this function 300 times in the main function, which allowed us to perform detailed calculations of the system performance under varying load conditions.

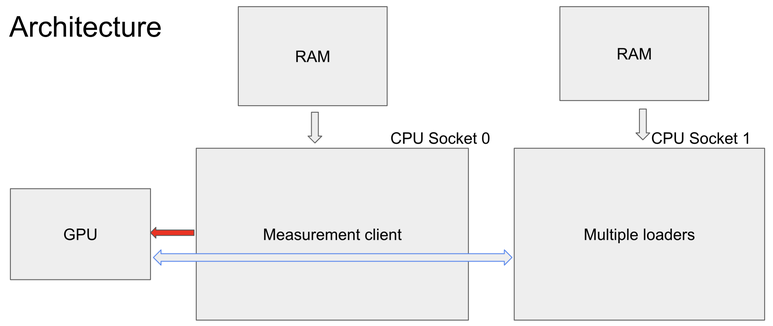

Figure 1. Architecture which allows us to stress the PCIe and measure model inference under different loads

Results and Discussion

In the first experiments we measured the inference of a Bert-Base-Uncased model with 110M parameters. As we see, the results can be interpreted as intuitive - the more load we apply, the more time it takes on average to perform inference. Despite that, the novel insight would be that the inference time increases exponentially depending on the percentage of PCIe bandwidth utilization.

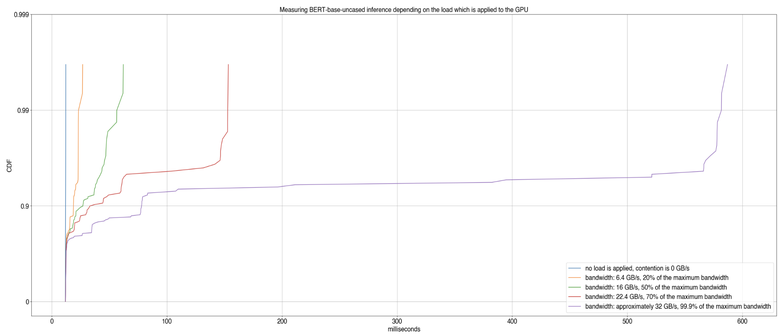

Figure 2. The proportion of values that are less than or equal to each corresponding time measurement of BERT inference depending on the load which is applied to the PCIe

Future Perspectives

This research has still a lot to discover in terms of improving efficiency of model serving

- **Perspective 1** Discover the impact of PCIe contention on various DNN models
- **Perspective 2** Evaluate different techniques on how to mitigate PCIe contention depending on different models
- **Perspective 3** Discover the impact of GPU contention on heterogeneous and homogeneous models
- **Perspective 4** Establish the main bottlenecks of GPU and PCIe contention and suggest an architecture for their mitigation

What is already known about this subject?

- **Resource management techniques** Workload scheduling, resource allocation and task offloading
- **Optimization techniques** Manipulations with cold start, unified frameworks for parameter-efficien transformers serving
- **Storage optimizations** Elastic serverless cloud storages, Transparent Auto-Scaling Cache for Serverless Applications

What does this study add?

- **Insight 1** Correlation between the PCIe bandwidth and Model inference
- **Insight 2** Solutions to the bottlenecks created by PCIe and GPU contention
- **Insight 3** Discovery of the possibility of deploying heterogeneous DNN models on the same GPU

Relevant works

Figure 3. QR code which you can use to get access to a list of all the relevant works on the field