CS 524: Introduction to Optimization Fall 2025

Author: Deepa Venkatachalapathi

GPU Cluster Job Scheduling Optimization - Project Spec

(A) What is the issue being addressed?

Modern GPU clusters execute thousands of machine-learning jobs simultaneously. Because each job requests specific GPU and CPU resources, poor scheduling often causes long queue times, uneven utilization, and wasted compute capacity. This project aims to develop optimization model that assigns jobs to available nodes to minimize latency and maximize utilization, while considering communication delays and job priority levels.

(B) Where does the data come from and how will it be obtained?

The two most suitable datasets for this project are the GPU Cluster Trace Dataset released by Alibaba and The MIT SuperCloud Dataset. Both datasets contain detailed job-level attributes such as job duration, number of workers, GPU and CPU requests, priority, and submission timestamps, as well as node-level data on GPU and CPU capacities. Given computational limitations, I plan to use the Alibaba dataset first, as it is smaller and easier to process while still containing all necessary information for modeling. I have disregarded other datasets as they are several gigabytes in size, making them computationally very expensive.

(C) What is the optimization problem underlying this project?

The optimization determines how jobs should be distributed among cluster nodes to achieve three objectives: Minimize execution latency, Maximize GPU utilization and Minimize communication overhead when jobs requiring multiple workers are spread across nodes. The model decides how many workers of each job are placed on each node, whether a job is accepted for scheduling, and which nodes it uses.
Constraints ensure that worker demands are met only if the job is scheduled, GPU and CPU capacities are not exceeded, jobs are executed according to the priority and their age (older jobs have a smaller start timestamp) accordingly. I also need to ensure that jobs run only on compatible GPU models.

(D) What are the deliverables?

Data Preprocessing and analysis to check for any null values and extract all the necessary features. From the obtained result, I will use NetworkX and Plotly to simulate how the jobs are distributed among the nodes and node utilization.
If time permits, I would like to include additional constraints like even load distribution to enable smoother performance. I would also try to run this on the bigger dataset and a Pareto analysis of the utilization and execution latency.

(E) **Other points for me to consider when evaluating?**

My major concern is the computational constraints given the size of the dataset and the constraints.