Analyzing ML Training Tasks

Analyzing Time to Success and Failure

Introduction

Given trace data from thousands of training and inference jobs, I conducted an analysis of various variables that may influence failure occurrences. These variables include GPU type, GPU utilization, CPU usage, and GB of main memory. I documented the success and failure rates and mean time to completion (failure or success) associated with variations in these variables.

Overview of the data

As the name indicates each table shows a breakdown of each job, task, and instance. Users submit ML jobs along with application code and specified computational resources. Each job is then translated into tasks that contain one or multiple instances. The trace data for each job is divided into 7 data files. Of those this analysis focuses on 3 tables: pai_job_table, pai_task_table, and pai_instance_table.

There are a total of 1050501 jobs with several tasks per job and any number of instances per job ranging from 2 to 6330.

Data Cleaning and Preprocessing

For each task in DFA I added the runtime of tasks and the task size. I found the runtime of each task by using the earliest start time and the latest end time of all the instances in one task. For terminated cases, this was the same as the start and end times given in the task table.

The task size was the number of instances in each task.

Next, I grouped all the rows based on a selected feature (GPU type, GPU utilization, CPU usage, and GB of main memory) and plotted the average run time for each group to determine the correlation between each feature and the time to completion/failure.

I then divided the test cases into 2: test cases that terminated (meaning success) and the remaining test cases(meaning failure)

Results

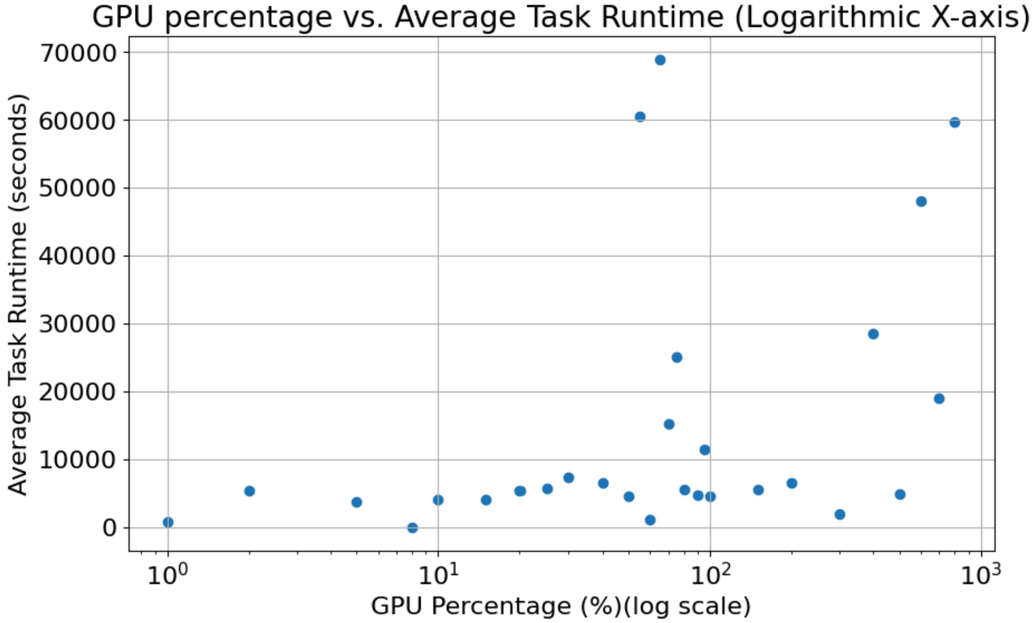
Pearson Correlation is on a scale of -1 to 1 (-1 meaning strong negative correlation and 1 meaning strong positive correlation and zero meaning no or weak correlation)

P-value: Chosen Significance Level (0.05) < 0.05 the null hypothesis can be rejected

Time to Completion for Successful Tasks

GPU Percentage

Pearson Correlation: 0.44
P-value: 2.01e-02



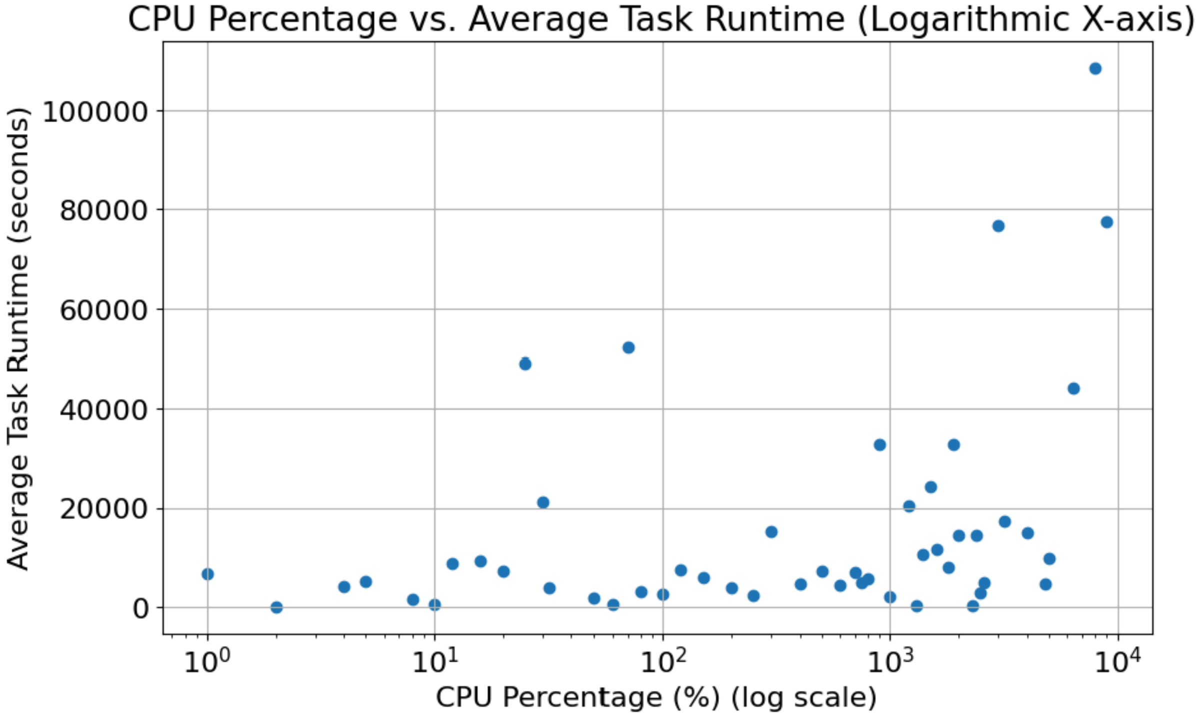
GPU Type

Average time to Completion

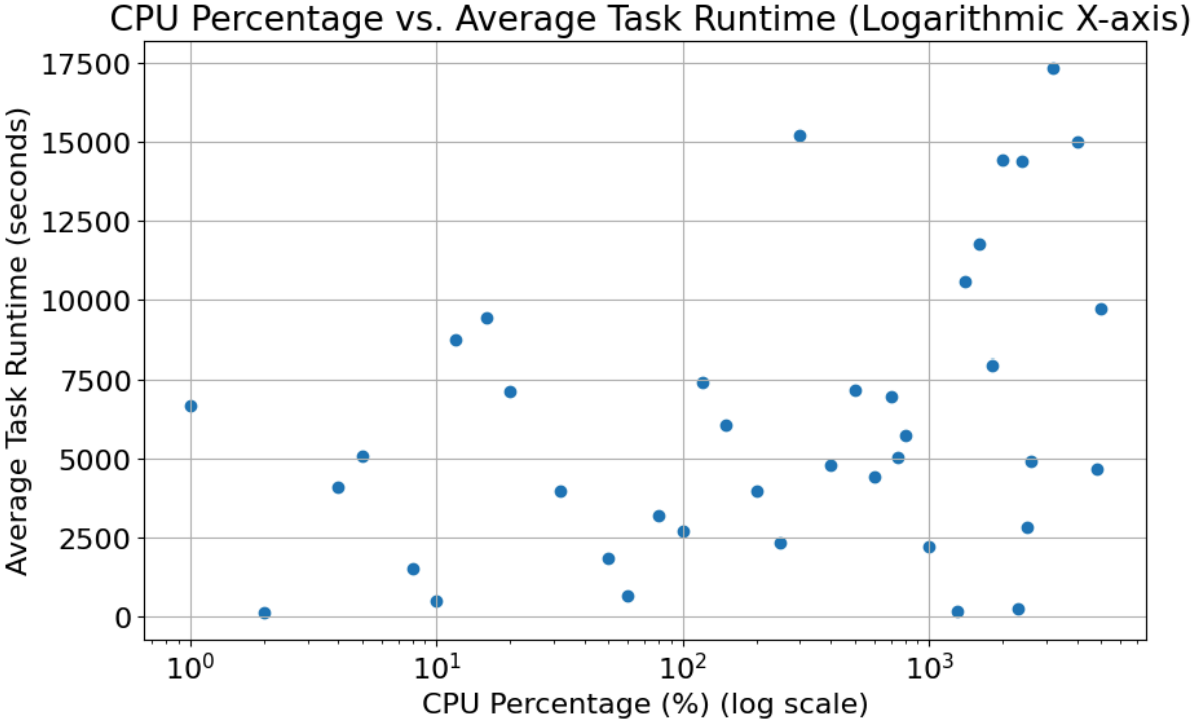
GPU Type	Average task runtime (seconds)
MISC	5197.2
P100	7669.9
T4	2876.3
V100	9707.03
V100M32	11905.27

CPU Percentage

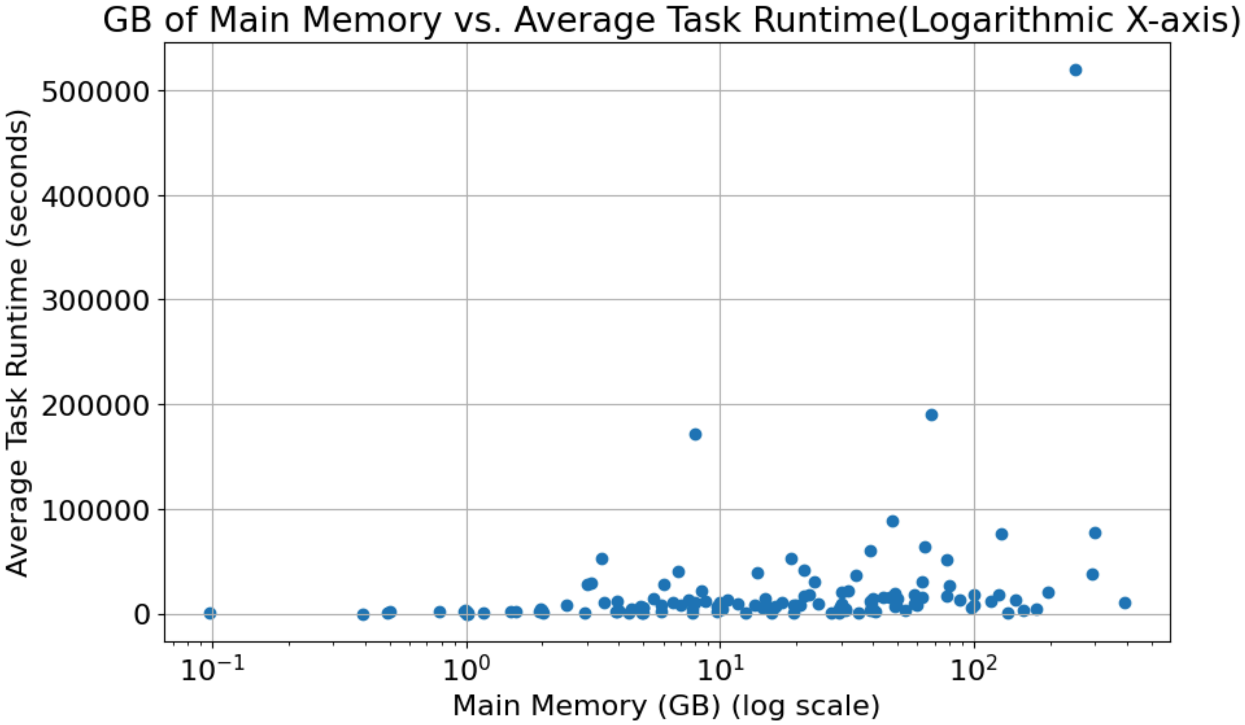
Pearson Correlation: 0.64
P-value: 4.51e-07



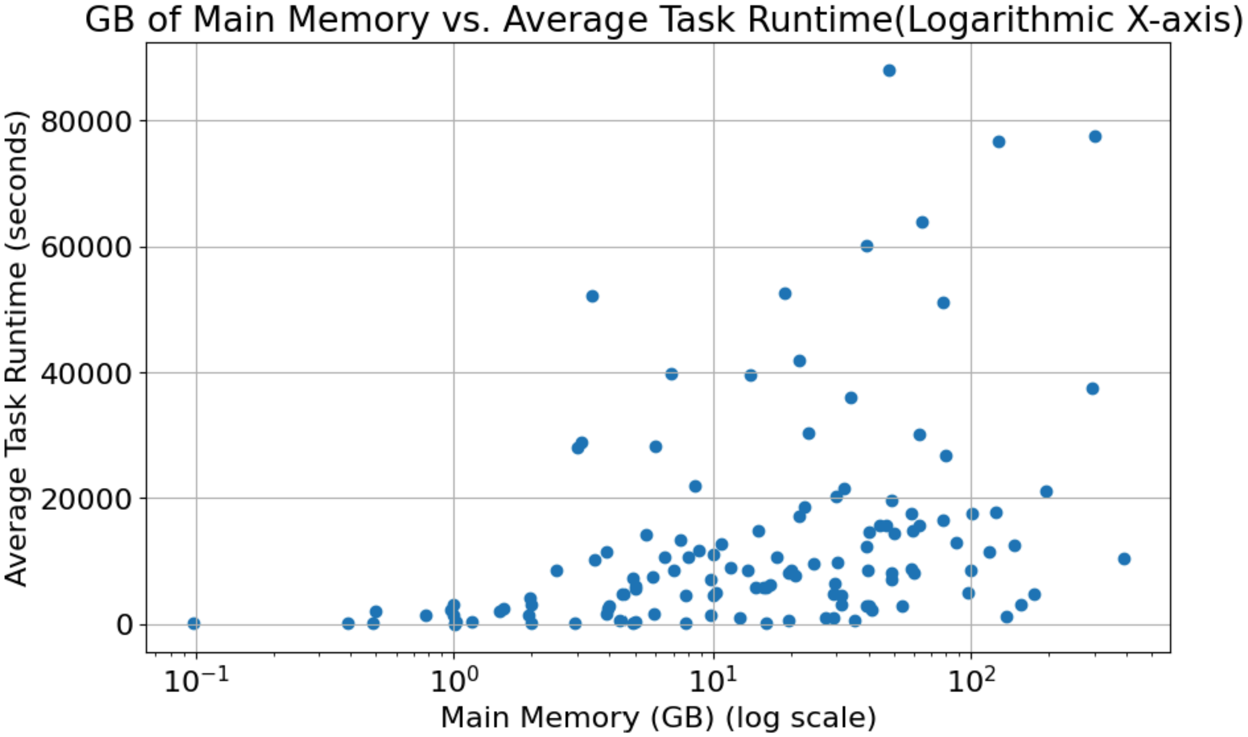
Analyzing only the tasks completed in less than 20000 seconds



GB of Main memory
Pearson Correlation: 0.35
P-value: 3.78e-05

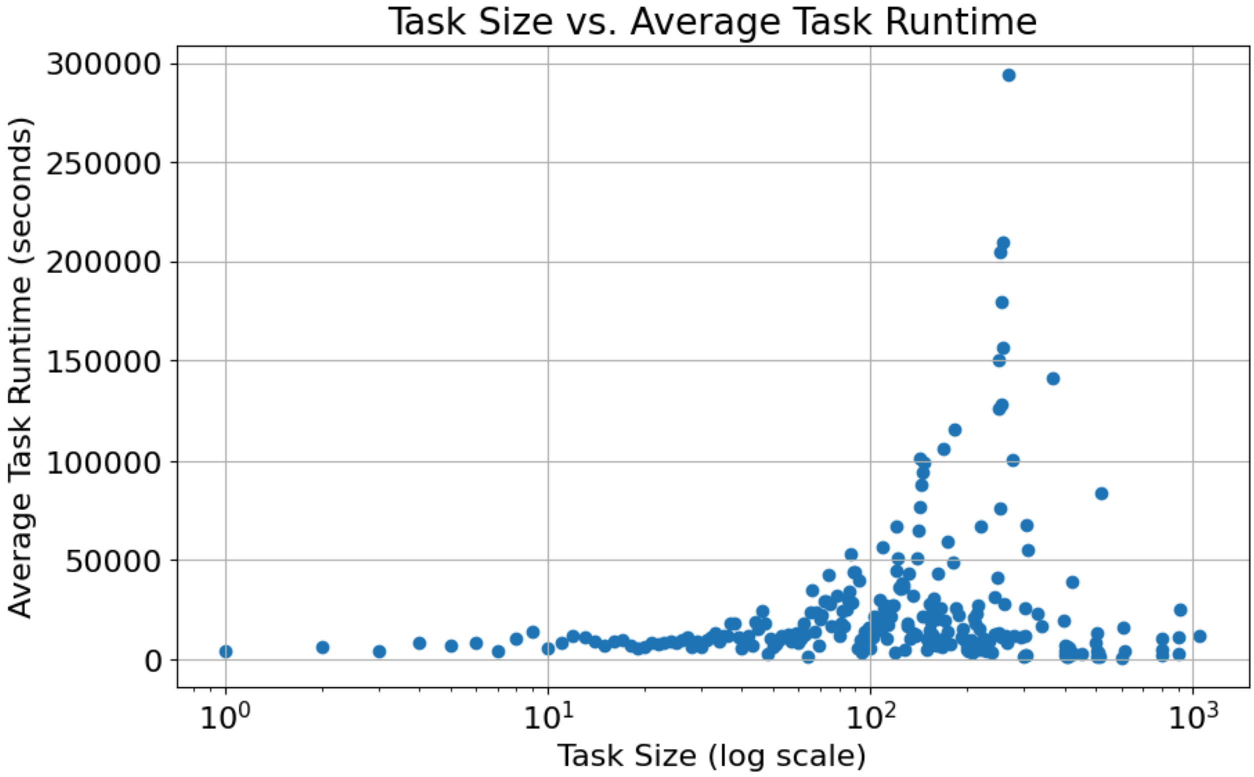


Analyzing only the tasks completed in less than 100,000 seconds



Task Size Comparision

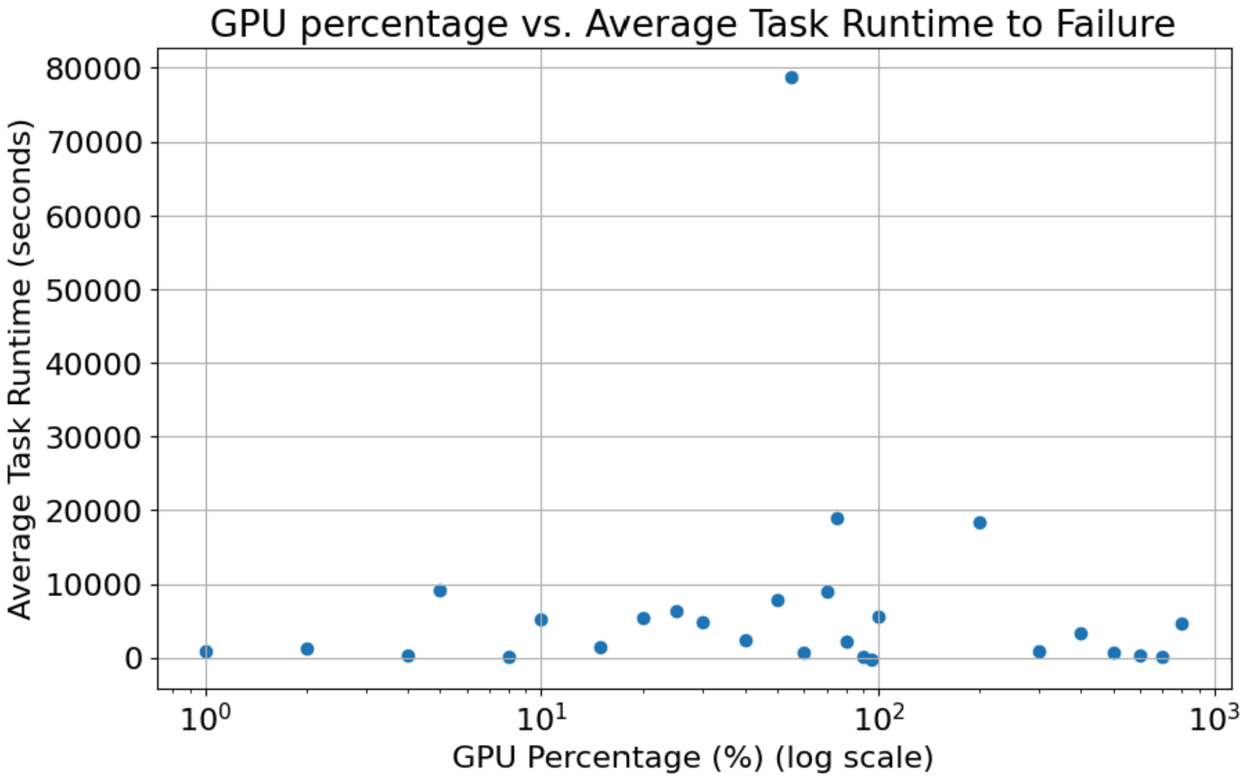
Pearson Correlation: 0.04
P-value: 4.50e-01



Time to Completion for Failure Tasks

GPU Percentage

Pearson Correlation: -0.14
P-value: 4.98e-01



GPU Type

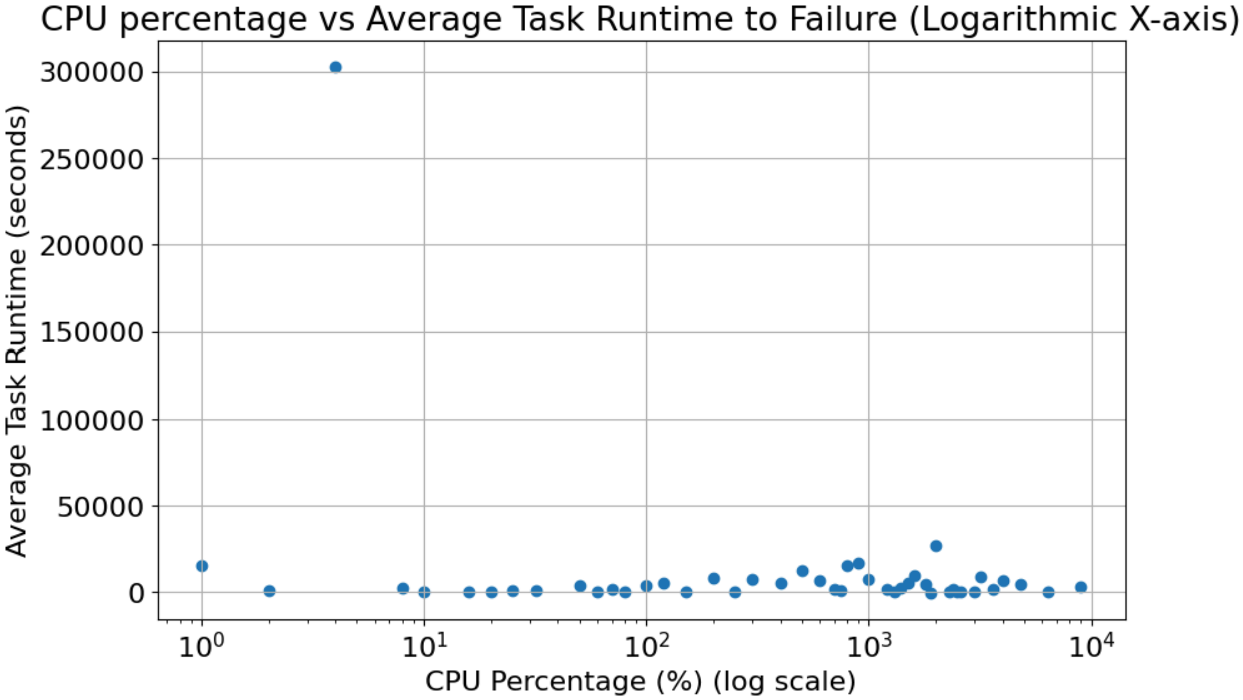
Average time to Failure

GPU Type	Average task runtime (seconds)
MISC	5993.6
P100	7114.5
T4	5064.8
V100	11332.1
V100M32	6889.3

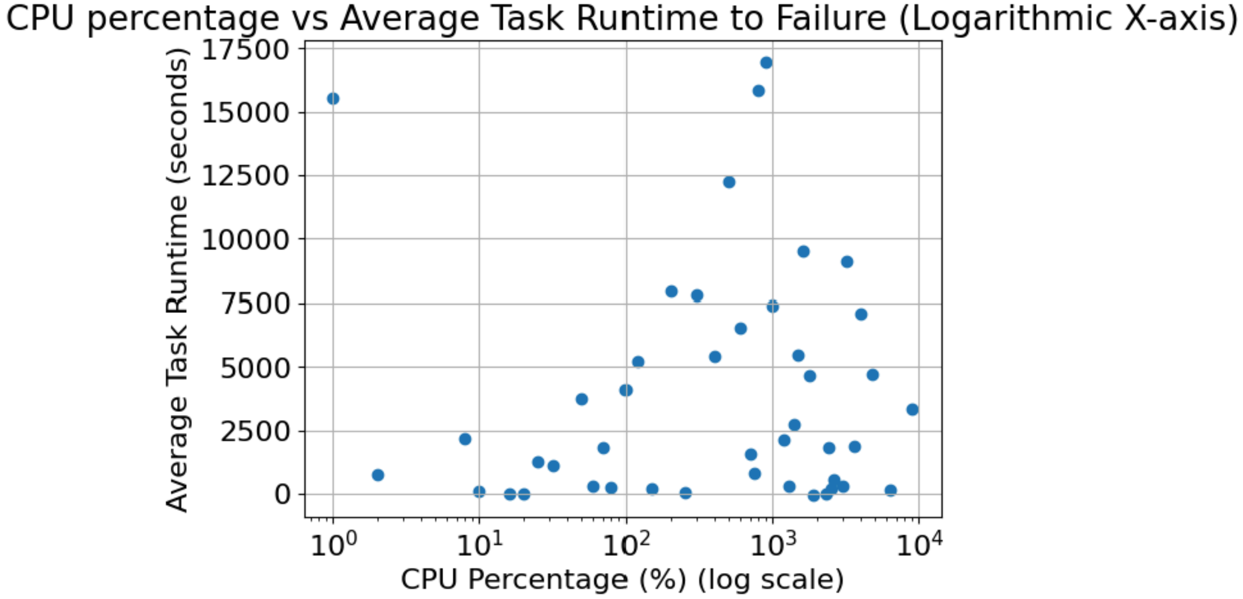
CPU Percentage

Pearson Correlation: -0.11

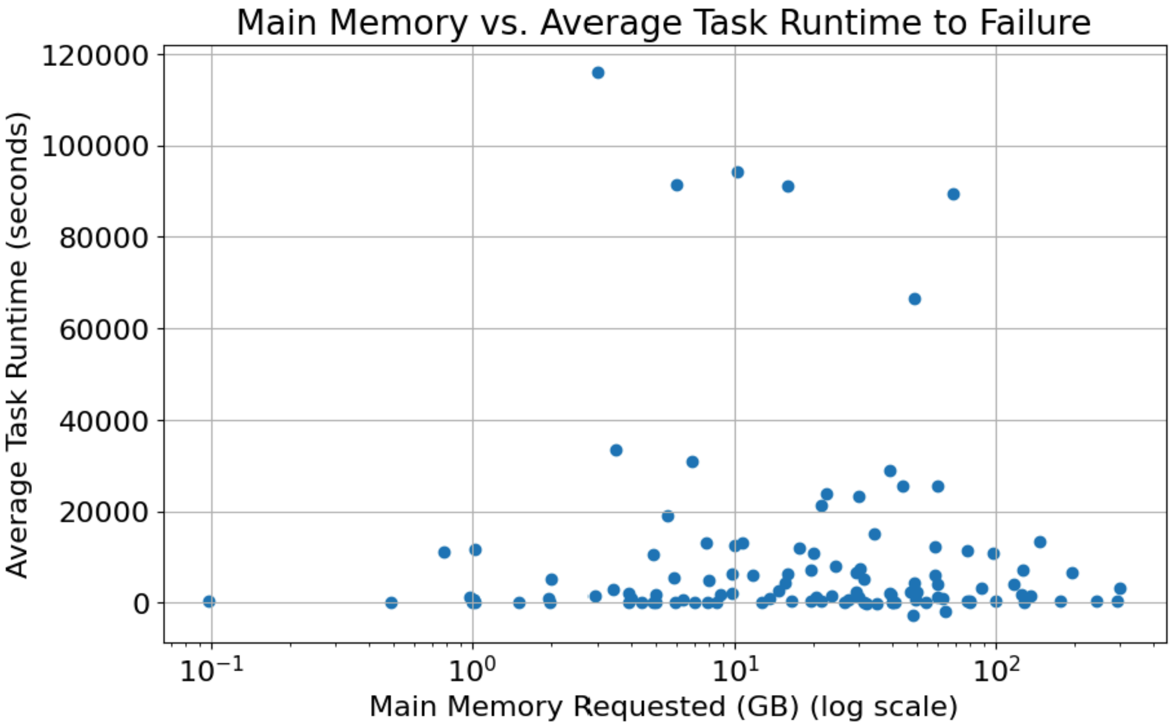
P-value: 4.50e-01



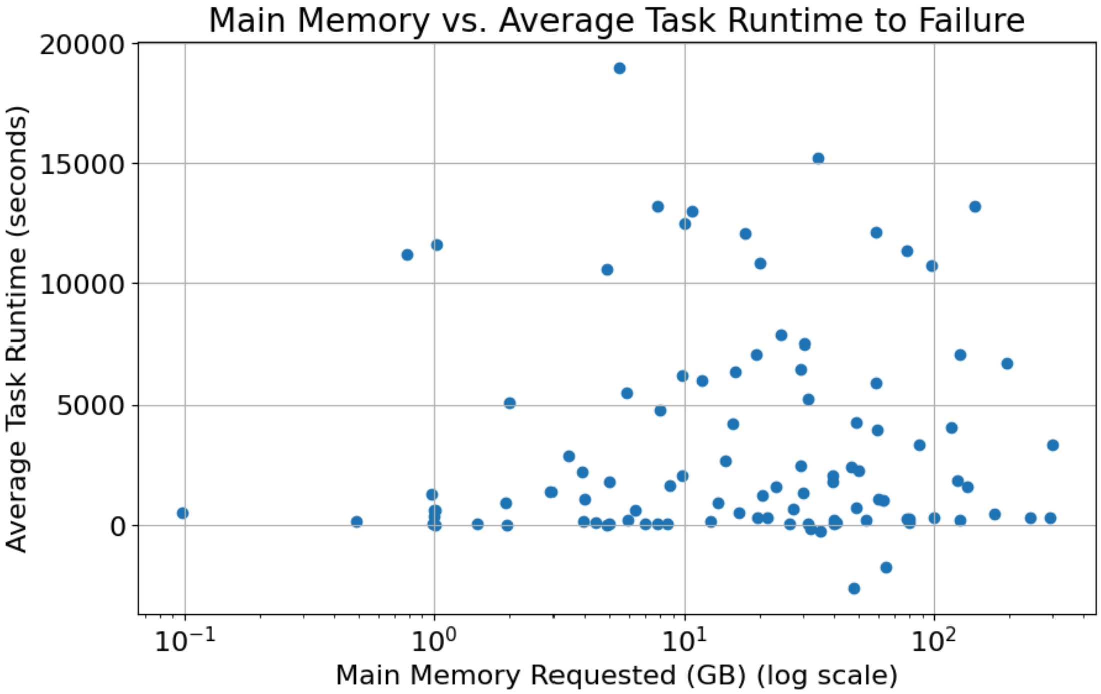
Analyzing only the tasks completed in less than 50,000 seconds



GB of Main memory
Pearson Correlation: -0.09
P-value: 3.19e-0.1

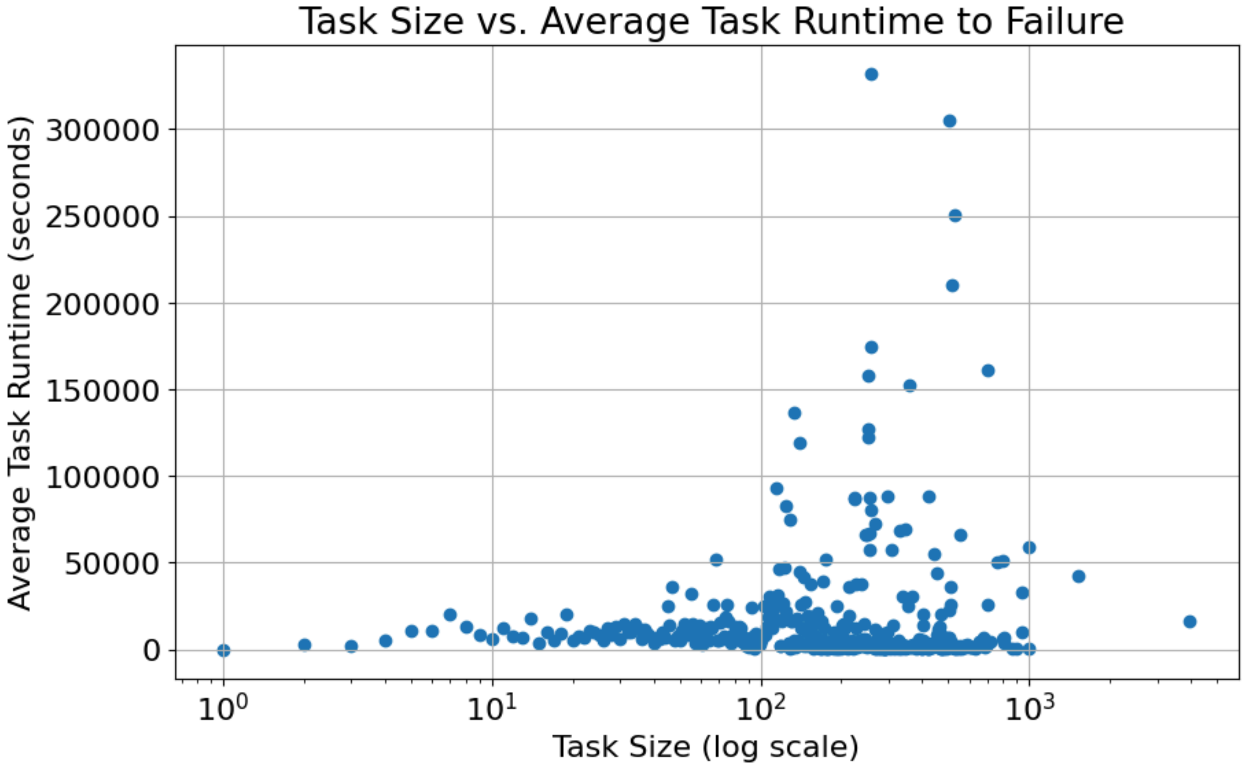


Analyzing only the tasks completed in less than 20,000 seconds



Task Size Comparision

Pearson Correlation: 0.04
P-value: 4.50e-01


Task Size vs. Average Task Runtime to Failure
Average Task Runtime (seconds)
10^0
10^1
10^2
10^3
Task Size (log scale)