

Assignment 4:

MadsServe: Own the Serving Engine

Overview

- Setup the basic SGLang serving engine
- Run agent with the self-hosted serving engine
- Observe the bottleneck
- Optimize for the bottleneck you have observed
- Submit and iterate

Basic SGLang Setup

- [First time setup in the ATG cluster](#)
- Allocate a RTX Pro 4500 Node
- Follow the instructions on the Assignment 4 page to set up a SGLang instance and connect to your agent

Measure the Performance

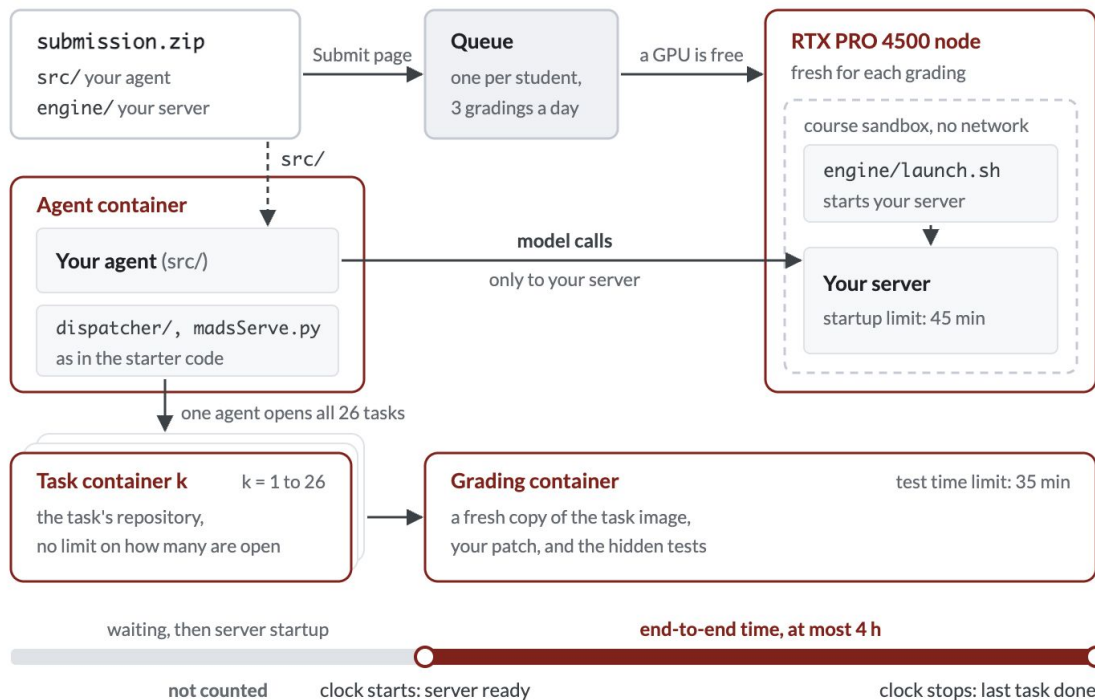
- Run your agent with all 25 tasks distributed to you
 - Log your agent and serving engine
 - /metrics endpoint in SGLang is useful
 - The task set is different from the task set on the leaderboard
- Dig deep into where the time goes
- Also run the SGLang bench to see how different the real serving is different from benchmarks
 - Does the benchmark show the same trend or different trend due to some reason?

Make Serving Efficient

- Think about the possible solutions to the problem you found in the measurement
- Try the solutions
 - One solution at a time
- Allowed changes:
 - Serving engine changes: configuration and code
 - The agent: How it use the server
- Model change is not allowed in the assignment (not the focus of this assignment)

Leaderboard

- We will run the serving engine in a sandbox
- Grading: end-to-end time, and solved tasks
- The others are similar to Assignment 3
- The serving engine loading time is not counted



Writeup & Video

- Show your understanding
- Required content in Writeup
 - The bottleneck you found in the measurement
 - The findings compared to the SGLang benchmark
 - Changes made with detailed observation and reasoning
 - Lessons learned
- Video
 - 3-minute video with your face and real human interaction

Notices

- Please be considerate. Use at most 1 GPU at a time!
 - We have limited GPU resources (80 RTX 4500 reserved)
- The leaderboard will stop receiving submissions right after the deadline
 - Submitted requests will be graded