LLM-Swap:
Dynamically Scheduling LLMs

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Team Project HWS 2026

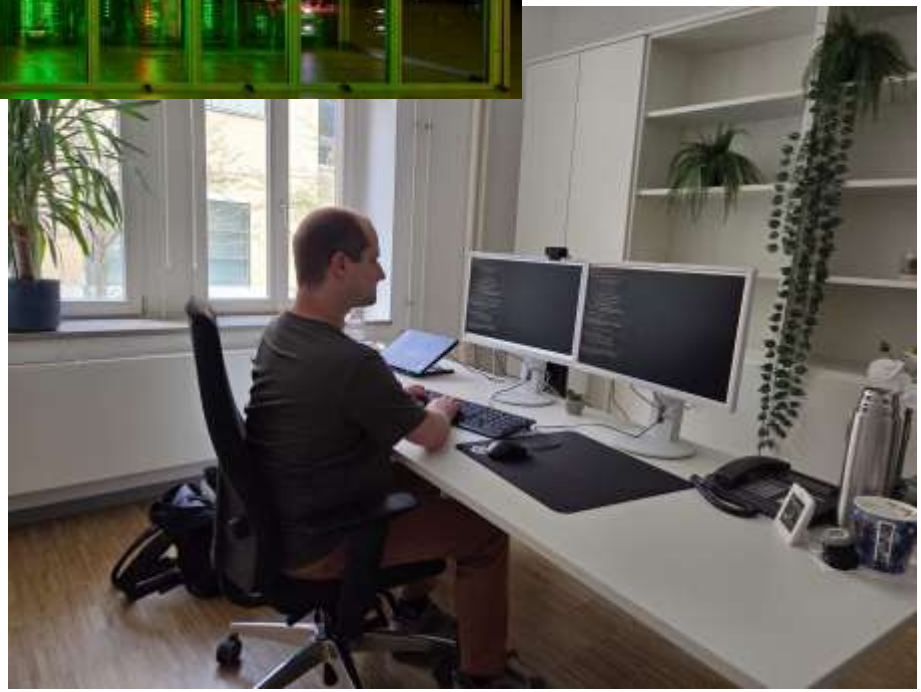
Imagine

- You are doing your master thesis
  - And you probably require LLM access for doing some experiments

Two Options: Option A

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Setup everything yourself (the computer science way) including:
  - Getting access to a cluster like BwUniCluster3.0
  - Typing commands in the terminal
    - Setting up vllm etc
  - But what about non-computer science students (or lazy computer scientists)?



Two Options: Option B

- Just use a service which runs LLMs for you
  - Like

together.ai Fireworks AI Parasail

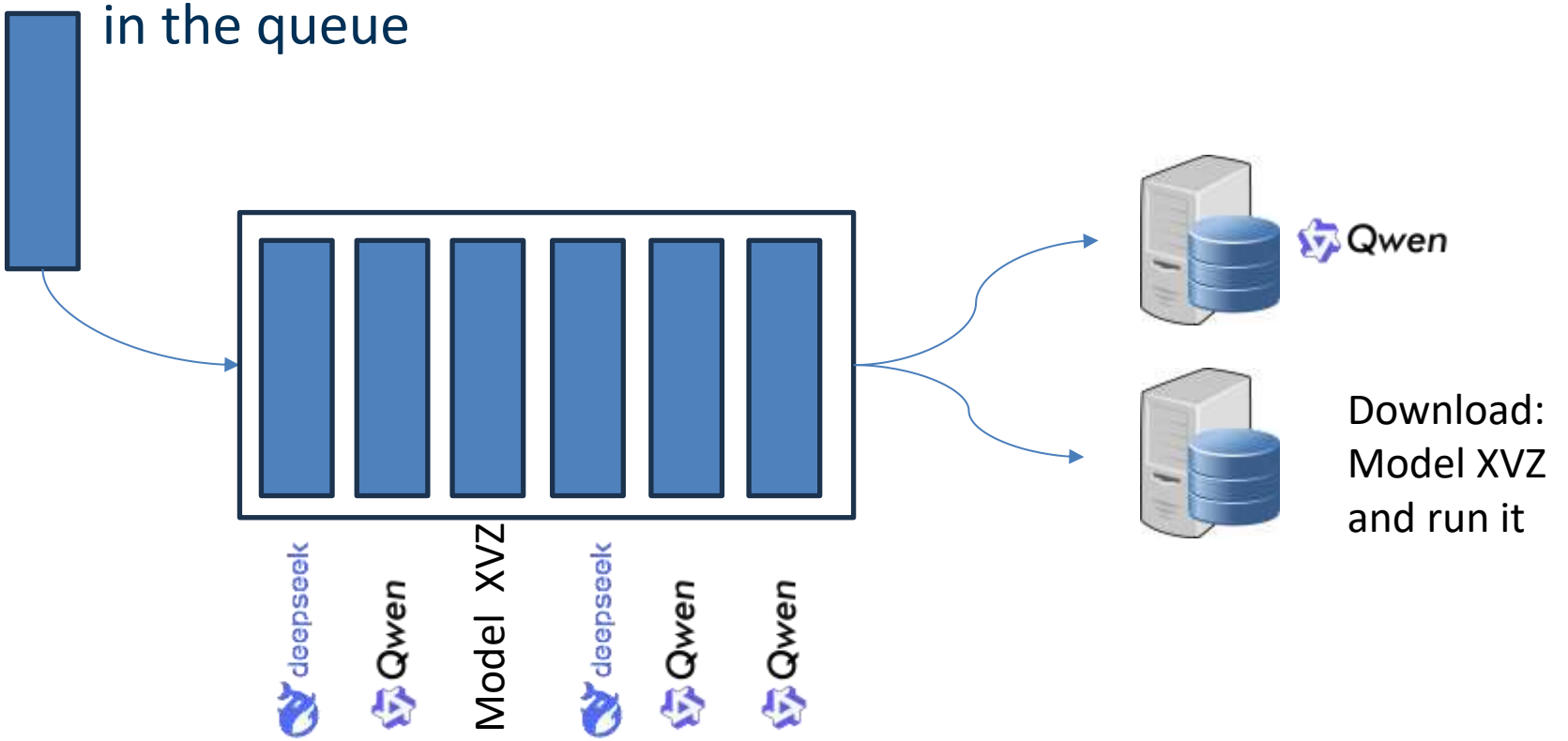
  - Also Uni Mannheim has a service

UNIT UNIVERSITÄTS-IT MANNHEIM maKI

- But those have static hosting of LLMs
  - One LLM per GPU
  - Thus only a limited selection of LLMs available

Task

- Implement a service/scheduler which dynamically load/unload LLMs depending on the requests in the queue



Requirements

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- It should work in a SLURM environment
  - When to allocate resources and for how long?
- Any model from Huggingface should be usable
- If many requests for a specific model are queued maybe host it multiple times to increase throughput?
- How to optimize the model context length?
  - If prompts are smaller (less tokens), model context length can also be smaller which reduces VRAM usage
- How should the scheduling work, when there is a heterogenious cluster with different kinds of GPUs?

Organisational information

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- The project runs for one semester (six months)
- 2-4 students
- There will be regular meetings
  - Some in person, most can be online
- Computing infrastructure is provided by the DWS group
- Required skills of the team
  - Good programming skills in Python

Supervisor

Dr. Sven Hertling
Substitute Professor for Data Science
DWS Group

E-Mail: sven.hertling@uni-mannheim.de

Philipp Hematty
Universität Mannheim – UNIT

E-Mail: philipp.hematty@uni-mannheim.de

Links

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Already existing projects which does parts but not all of our requirements
  - Llama-Swap
    - https://github.com/mostlygeek/llama-swap
  - LiteLLM
    - https://www.litellm.ai/
  - Open On Demand
    - https://www.openondemand.org/
    - https://github.com/OSC/ondemand
- Maki
  - https://maki.uni-mannheim.de

Questions

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics