Hasso Plattner Institut
Digital Engineering · Universität Potsdam

GPU Resource Sharing in Interactive User Sessions

Felix Grzelka
HPI

Exploratory Programming

Exploratory Programming

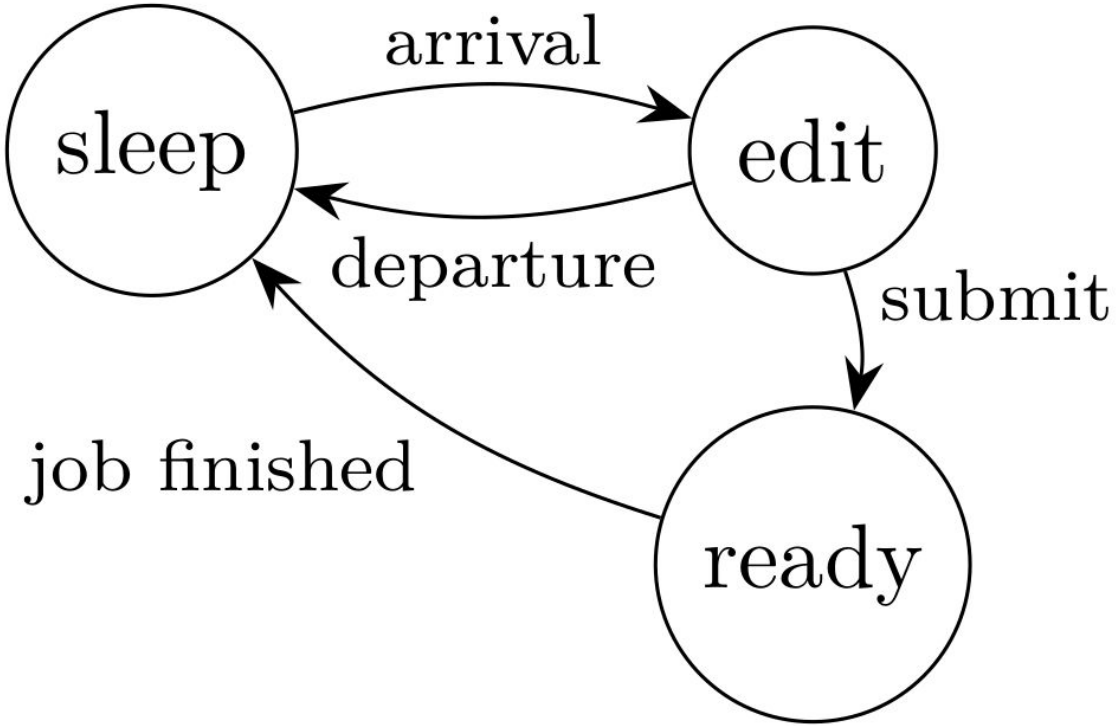
Außenwirkung — Phase 1: Selbstverständnis und Kernkompetenzen

⚠ **Wie uns Kollegen und Studierende teilweise auffassen**

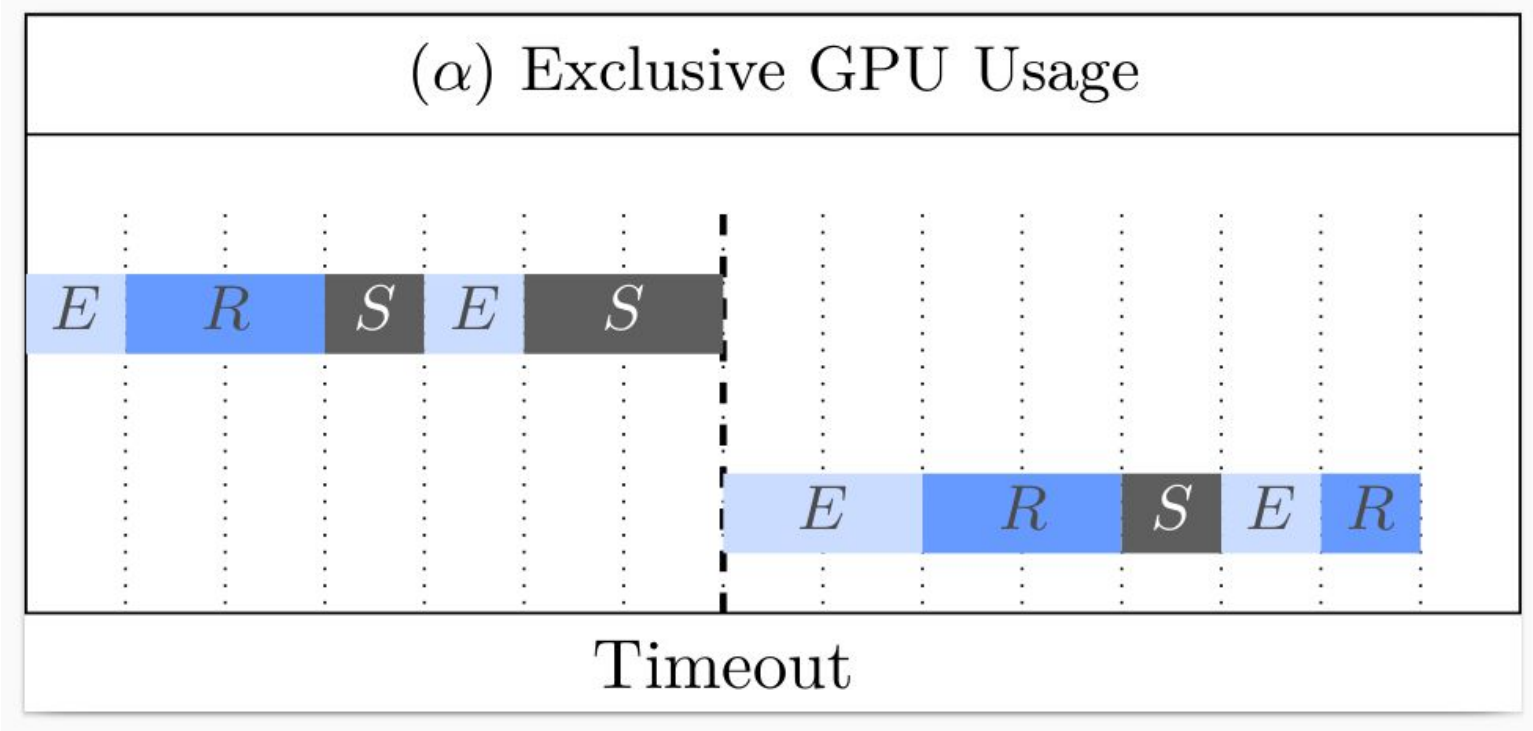
- Kannst Du nicht mal *Moderne Betriebssysteme* machen?
- Wir brauchen mehr Python. Ein *Linux-Führerschein* würde ansonsten reichen!
- Wir brauchen (auf unserer heterogenen Plattform) kein Betriebssystem und entwickeln lieber direkt auf der Hardware!
- Ressourcenverwaltung spielt in meiner Anwendungsdomäne keine Rolle, ich belege Betriebsmittel exklusiv.

Exclusive Resource Usage

Modelling Users

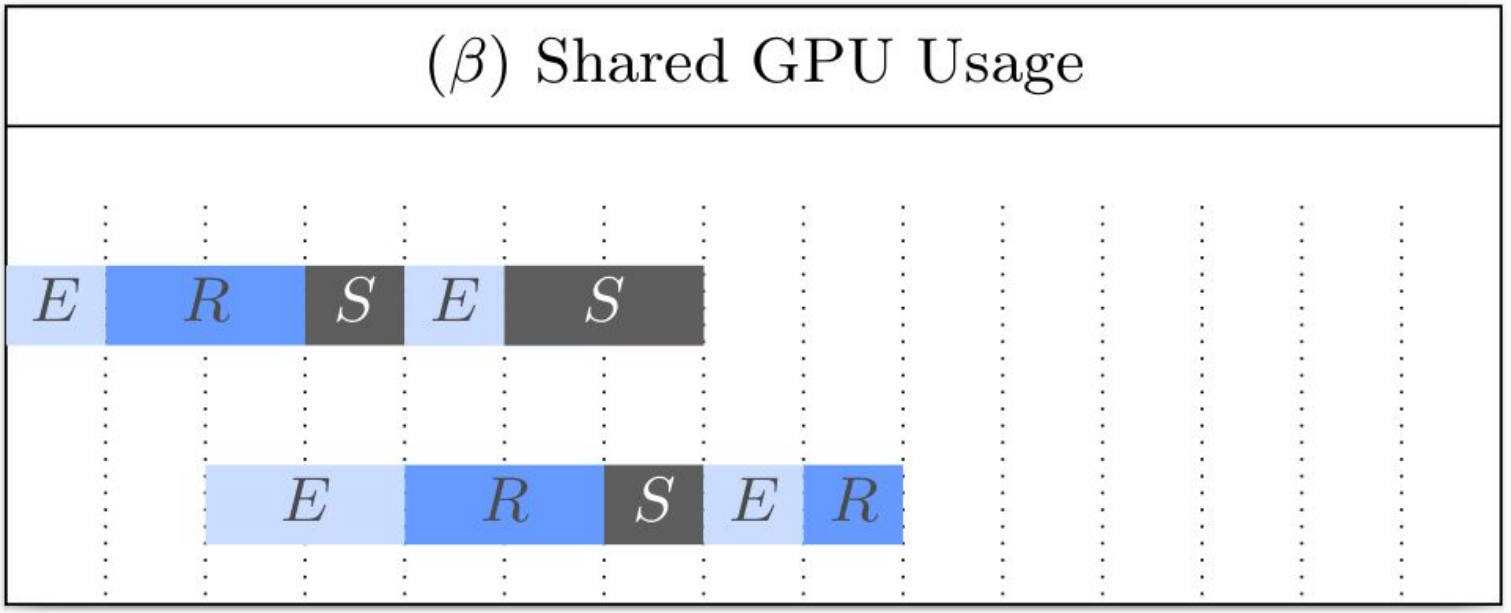


Modelling Users



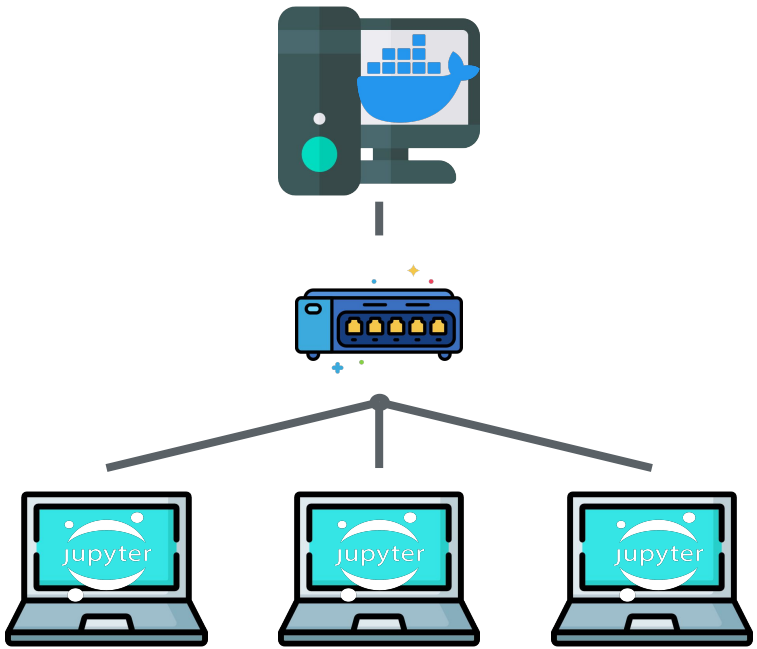
E: Edit
R: Ready
S: Sleep

Modelling Users



E: Edit
R: Ready
S: Sleep

The Classroom Scenario



GPU Server with
#GPUs < N

N students work on
notebooks without GPUs

Icons by Flaticon.com

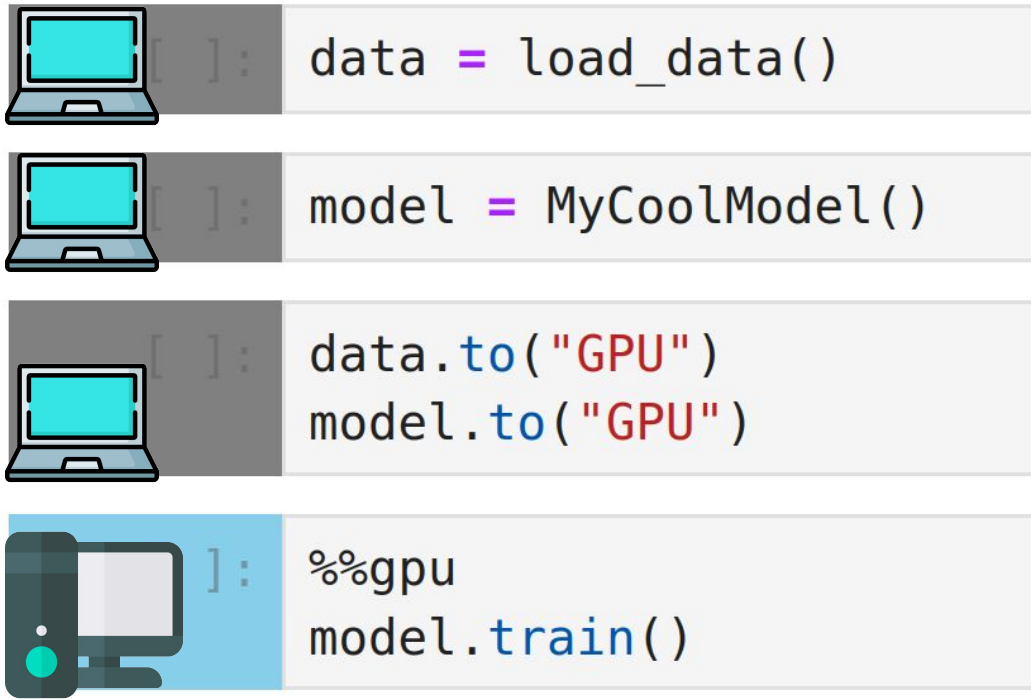
Example: Jupyter Notebooks

```
[ ]: data = load_data()

[ ]: model = MyCoolModel()

[ ]: data.to("GPU")
     model.to("GPU")

[ ]: model.train()
```

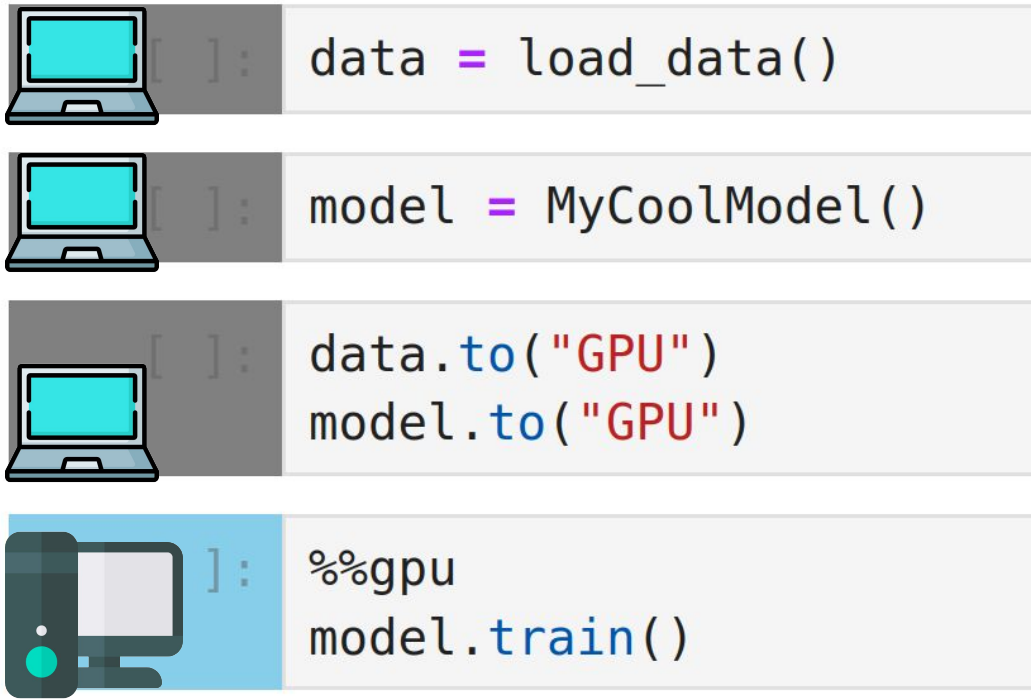
What about Google Colab (The Cloud)?

image source: https://blog.netdata.cloud/assets/images/colab-a699aa465989ccdc63feb80cdd107224.png

Prototype

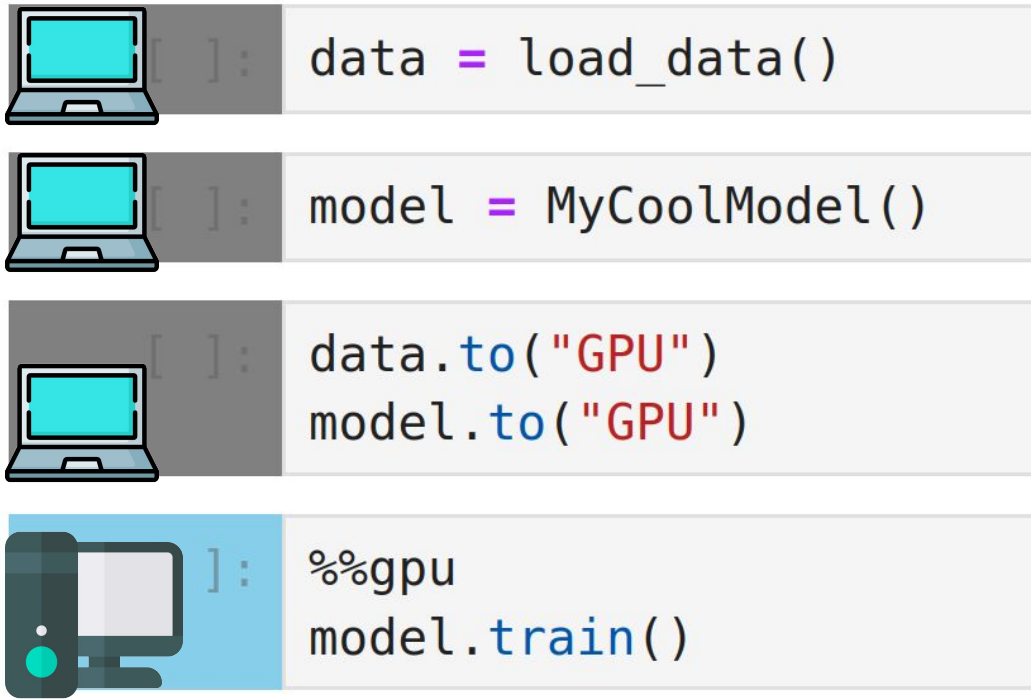

[ ]:
data = load_data()
[ ]:
model = MyCoolModel()
[ ]:
data.to("GPU")
model.to("GPU")
]:
%%gpu
model.train()

Prototype



serialize and transfer data, model and state

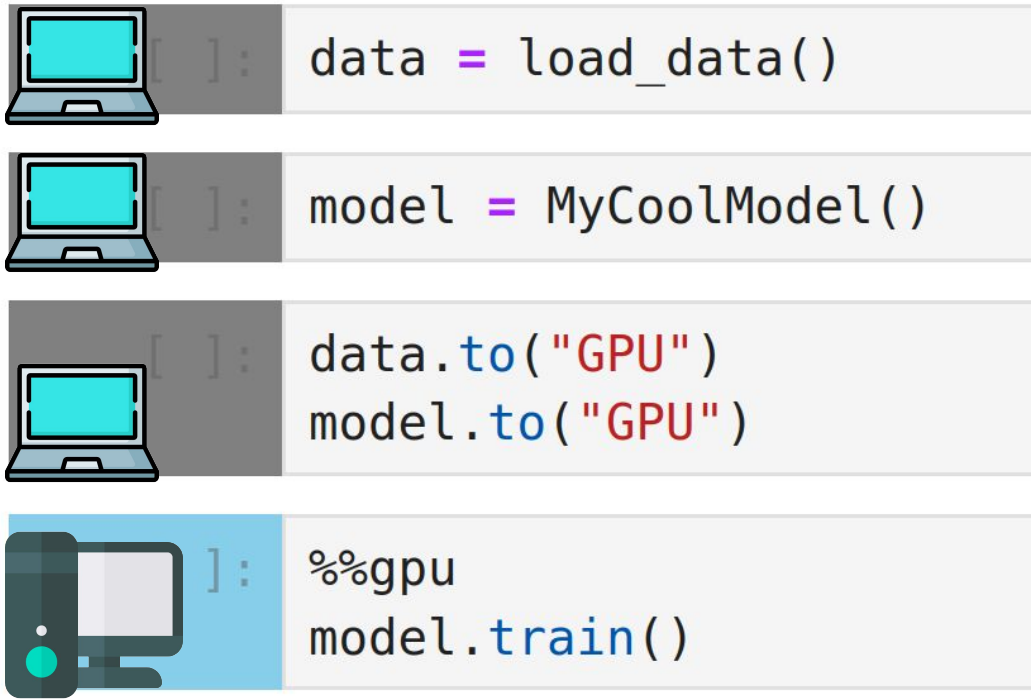
Prototype



serialize and transfer data, model and state

mock GPU availability on local node

Prototype

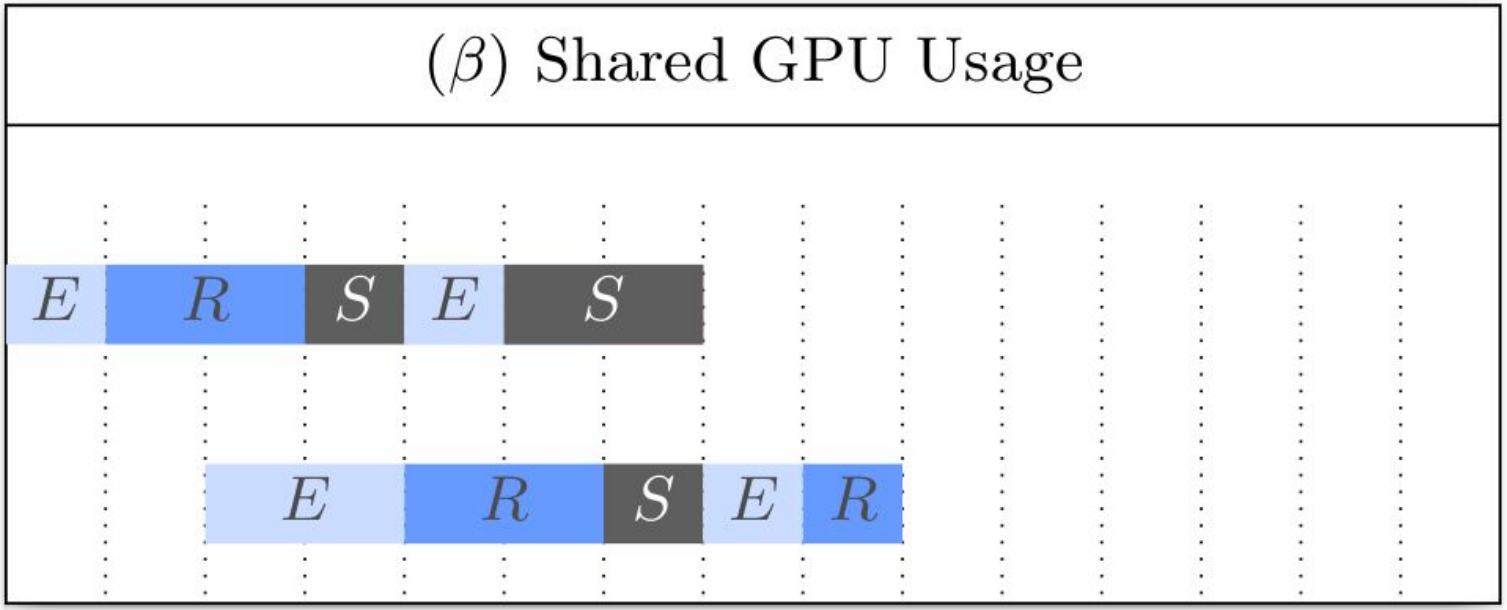


serialize and transfer data, model and state

mock GPU availability on local node

execute annotated code on remote GPU node

Modelling Users



E: Edit
R: Ready
S: Sleep

Research Questions

RQ1: How much additional latency is caused?

Research Questions

RQ1: How much additional latency is caused?

RQ2: How many users can realistically be served by one GPU?

Research Questions

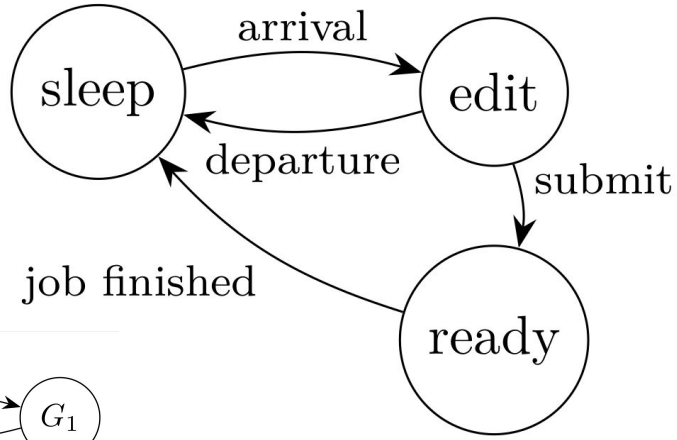
RQ1: How much additional latency is caused?

RQ2: How many users can realistically be served by one GPU?

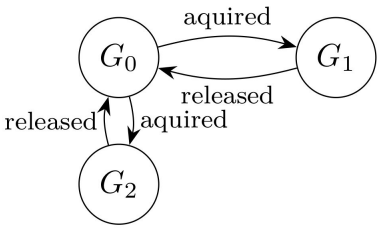
RQ3: How long does each user have to wait (on average/99 percentile)?

Modelling as CTMC

N x



1 x GPU



(N -1) x queue slots

Example with 2 Users

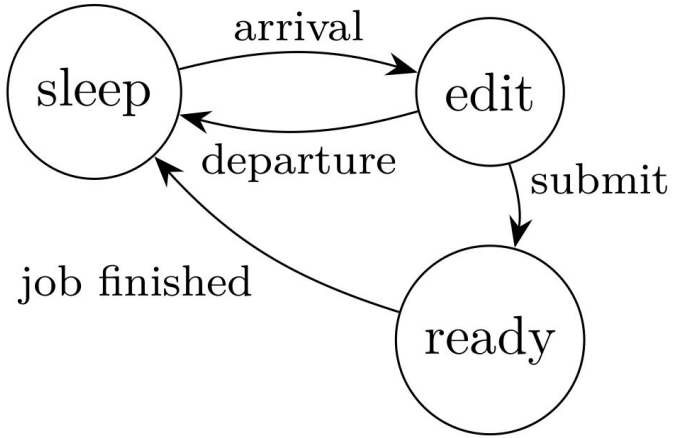
ssQ0G0, eeQ0G0,
ssQ0G1, eeQ0G1,
ssQ0G2, eeQ0G2,
seQ0G0, erQ2G1,
seQ0G1, erQ0G2,
seQ0G2, rsQ0G1,
srQ2G1, rsQ1G2,
srQ0G2, reQ0G1,
esQ0G0, reQ1G2,
esQ0G1, rrQ2G1,
esQ0G2, rrQ1G2

22 States

State Transitions

Model Parameters:

- arrival rate
- departure rate
- submission rate
- execution rate



Exploratory Programming

Außenwirkung

Phase 1: Selbstverständnis und Kernkompetenzen

Wie uns Kollegen und Studierende teilweise auffassen

- Kannst Du nicht mal *Moderne Betriebssysteme* machen?
- Wir brauchen mehr Python. Ein *Linux-Führerschein* würde ansonsten reichen!
- Wir brauchen (auf unserer heterogenen Plattform) kein Betriebssystem und entwickeln lieber direkt auf der Hardware!
- Ressourcenverwaltung spielt in meiner Anwendungsdomäne keine Rolle, ich belege Betriebsmittel exklusiv.

Exclusive Resource Usage

Questions?