Efficient Cluster Scheduling for Fine-tuning LLMs Using Historical Configuration Data

Matthijs Jansen[1], Sacheendra Talluri[1], Animesh Trivedi[2], Vassilis Vassiliadis[3], Michael Johnston[3], Alexandru Iosup[1], Srikumar Venugopal[3]

VU VRIJE UNIVERSITEIT AMSTERDAM

[1]VU Amsterdam
[2]IBM Research Zurich
[3]IBM Research Dublin

IBM

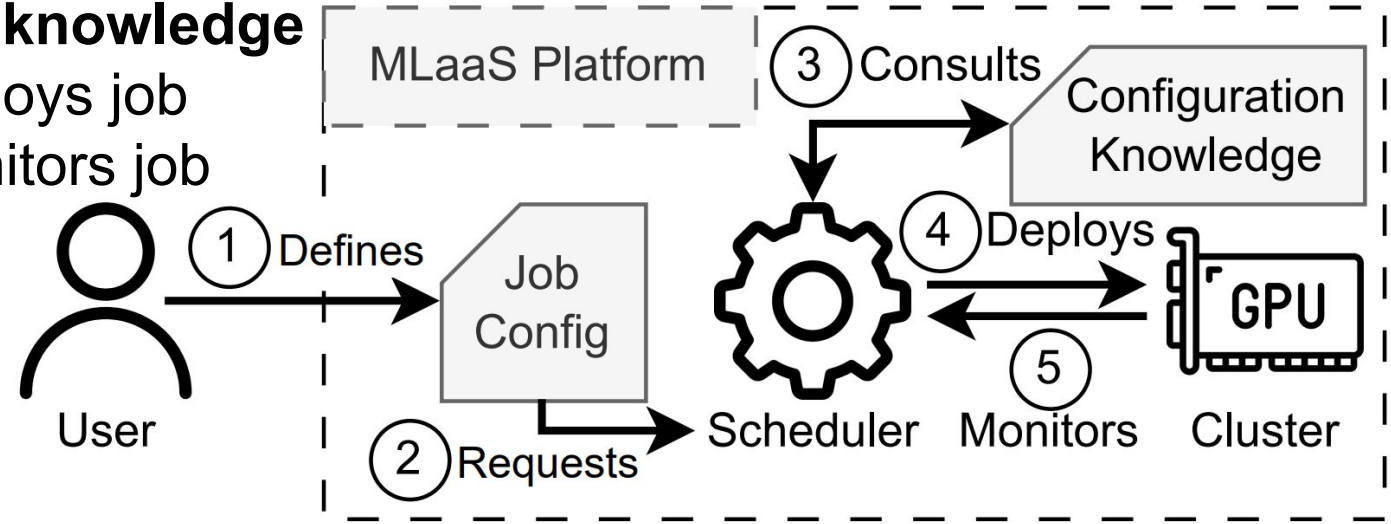
m.s.jansen@vu.nl

https://atlarge-research.com/mjansen

Machine-Learning-as-a-Service

1. User defines job
2. Submits job to batch scheduler
3. Scheduler fills in missing configuration params using **configuration knowledge**
4. Scheduler deploys job
5. Scheduler monitors job

Vision: User defines few required parameters, others are optional

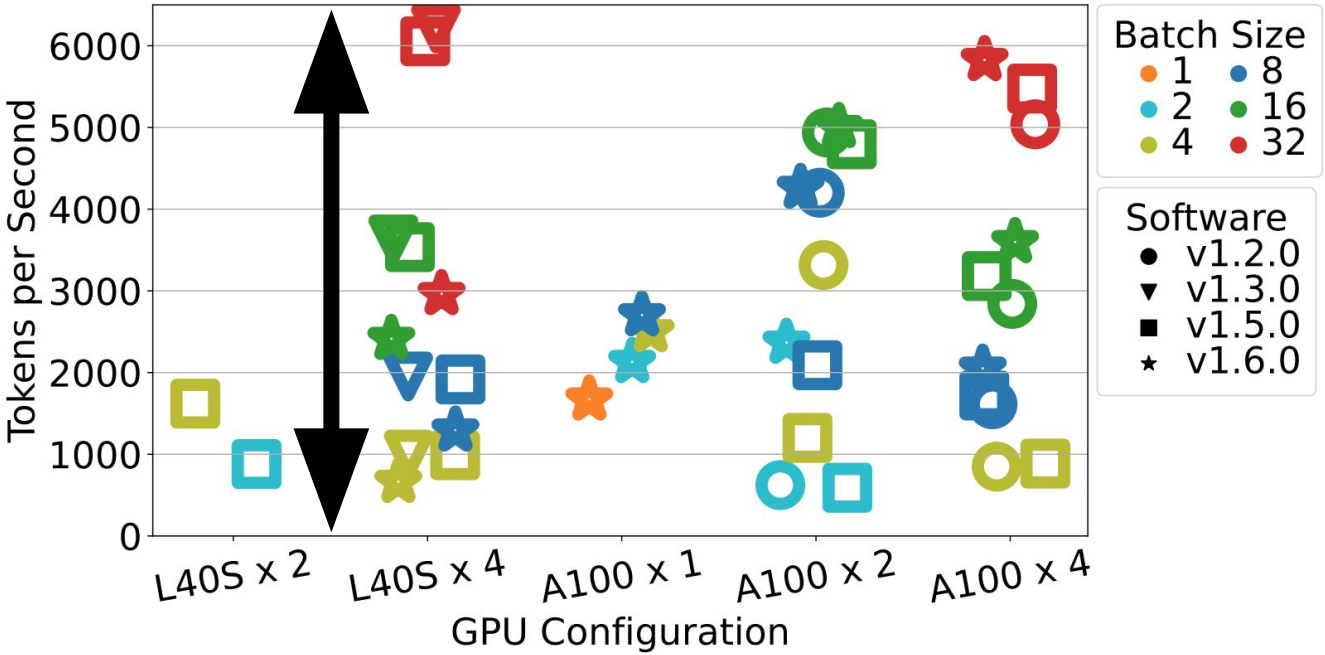


Fine-tuning Configurations Matter

Up to 10.53x performance difference with only 4 parameters

In MLaaS:

- >>10.000 config combinations
- SotP: Default values cannot capture complexity
- SotA: Ignores config knowledge availability

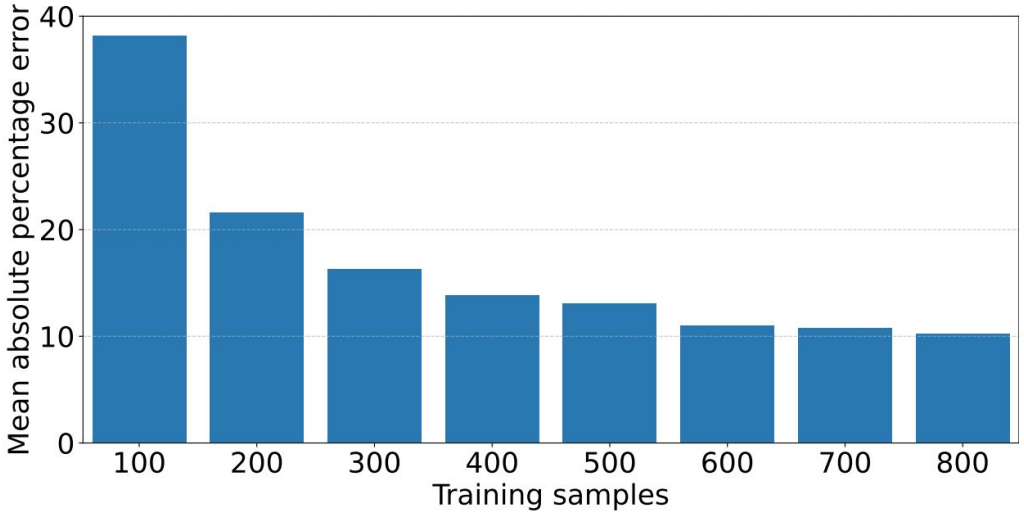


How to Create Configuration Knowledge

- Consider 7 parameters, with ~72.000 combinations
- Collected execution data of 5.000 configurations
- 220 days of execution

Predict, multivariate regression
- Validity: 4.2% error
- Performance: 9.4% error
- Small impact: Configurations differ by up to 4000% (40x)

Data impact on regression accuracy
- Diminishing returns, especially from 600 samples with 11% error



Impact of Configuration Knowledge

CoTune:

- Expose config knowledge sources for schedulers to integrate and evaluate

Evaluation:

- FIFO: 78% reduced JCT with predict performance
- Sia: 70% reduced JCT with predicted knowledge