Evaluating and Improving Scheduling Times in TVM

R244: Large-Scale Data Processing and Optimisation

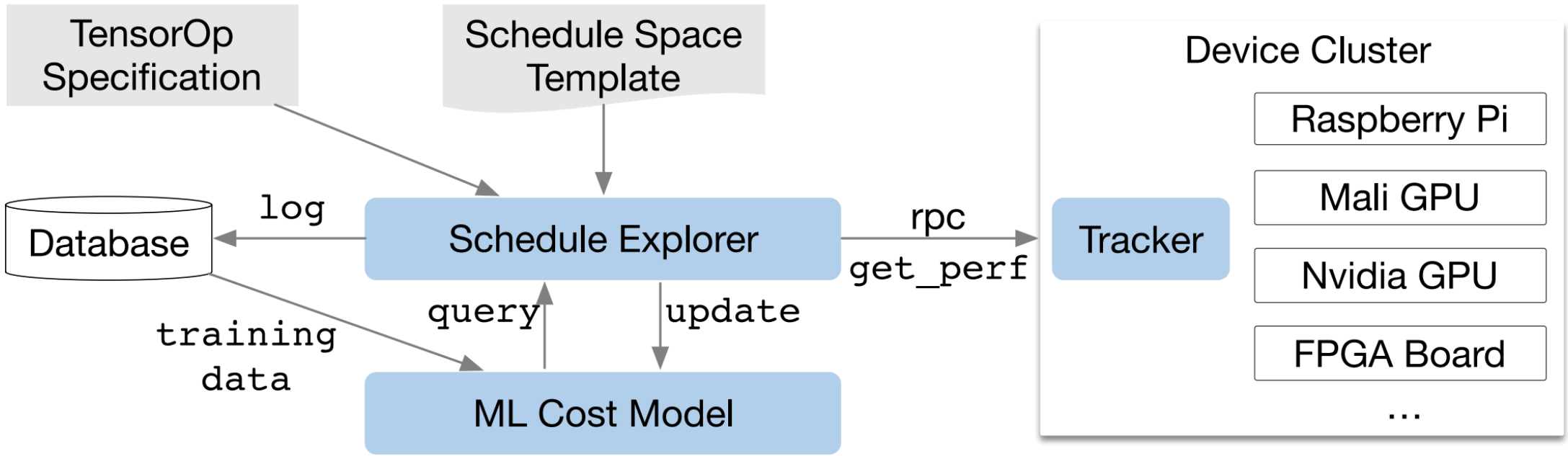
Kian Cross

Pros and Cons

- Comprehensive, well written paper.
- Evaluation shows that TVM performs very well.
- No information on compilation times of models?
- Does not support dynamic input shapes.
- (Only for inference, not training).

Pros and Cons

- Comprehensive, well written paper.
- Evaluation shows that TVM performs very well.
- No information on compilation times of models?
- Does not support dynamic input shapes.
- (Only for inference, not training).


TensorOp Specification
Schedule Space Template
Database
log
Schedule Explorer
rpc
get_perf
Tracker
Device Cluster
Raspberry Pi
Mali GPU
Nvidia GPU
FPGA Board
...
query
update
training data
ML Cost Model

Evaluating and Improving Scheduling Times in TVM

- Obtain baseline results across a variety of models.
- Experiment with transferring the database between different models.
- Identify how far along training can occur before a model can no longer be transferred. What is the optimum point?
- How much can models diverge before the database cannot be transferred?

Why do this?

- Original paper was missing an analysis of scheduling times.
- Time taken to schedule can be a major drawback of TVM.
- Is the current scheduling approach as optimised as possible?

What's next?

Evaluating and Improving Scheduling Times in TVM

Questions