Beyond Data and Model Parallelism for Deep Neural Networks

ZHIHAO JIA, MATEI ZAHARIA, ALEX AIKEN

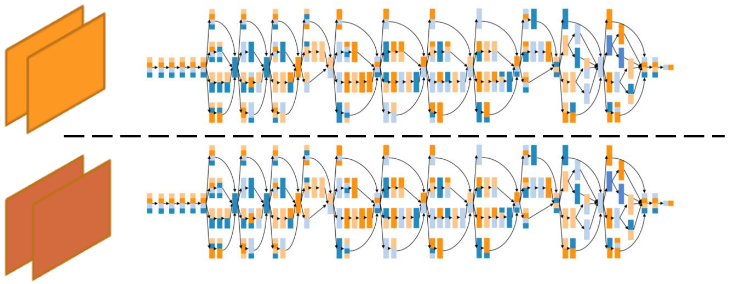
SYSML 2019

PRESENTED BY JULIUS LISCHEID

Existing Parallelisation Approaches (1/2)

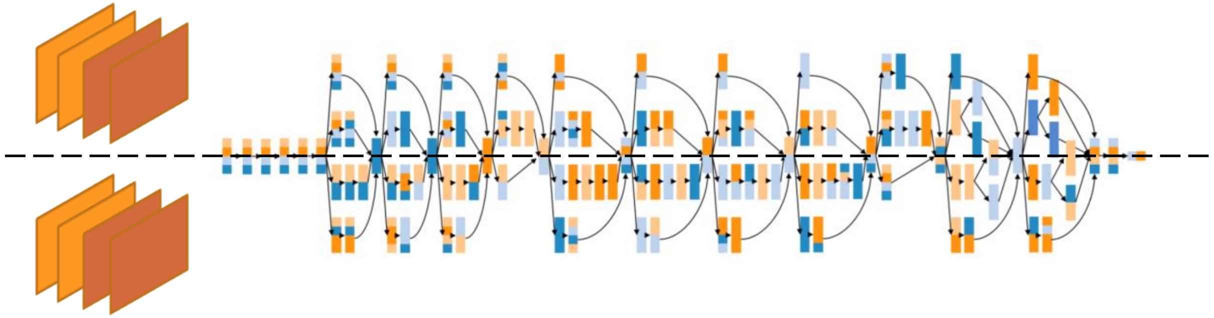
DATA PARALLELISM

- Replica of neural network on each device
- Each device processes subset of training data
- After each iteration, parameters are synchronised
- Works well for compute-heavy operations with few parameters (e.g. convolutions)



MODEL PARALLELISM

- Disjoint subsets of neural network assigned to devices
- No parameter synchronisation, but requires data transfers between operations



Existing Parallelisation Approaches (2/2)

EXPERT-DESIGNED STRATEGIES

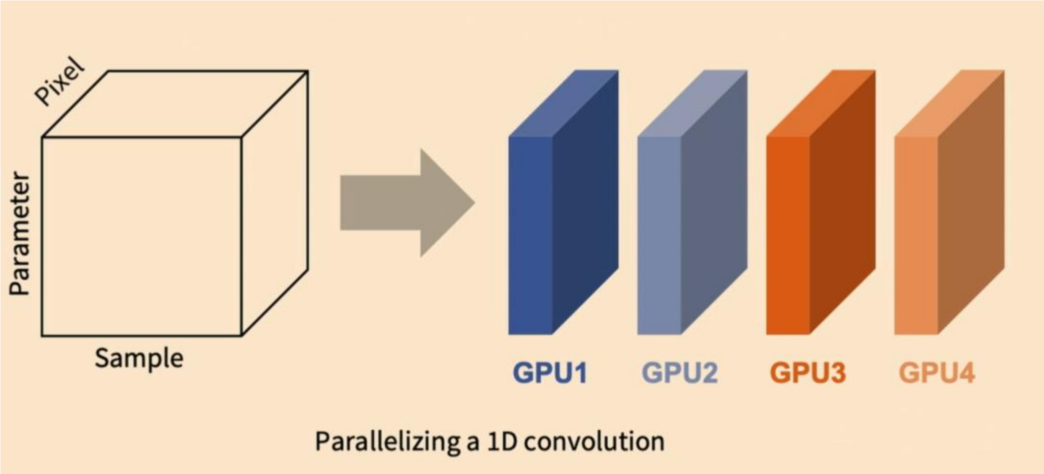
- A. Krizhevsky. One weird trick for parallelizing convolutional neural networks. CoRR 2014.
  - Data parallelism for convolutional layers, model parallelism for fully-connected layers
- Y. Wu et al. Google's neural machine translation system: bridging the gap between human and machine translation. CoRR 2016.
  - Data parallelism for compute nodes, model parallelism for intra-node computation

AUTOMATED FRAMEWORKS

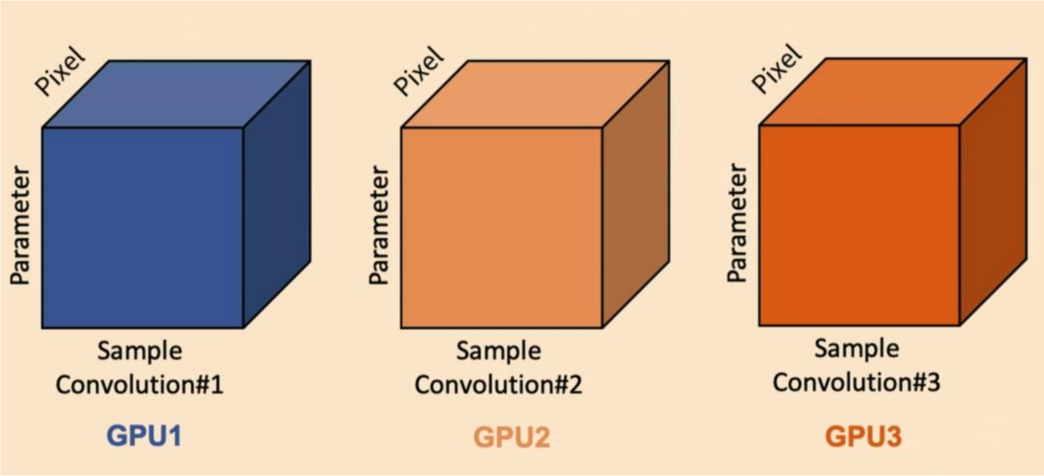
- A. Mirhoseini et al. Device Placement Optimization with Reinforcement Learning. ICML 2017.
  - Reinforment learning for model parallelism
- Z. Jia et al. Exploring hidden dimensions in parallelizing convolutional neural networks. CoRR 2018.
  - Dynamic Programming for parallelisation of DNNs with linear computation graphs
- D. Narayanan et al. PipeDream: generalized pipeline parallelism for DNN training. SOSP 2019.
- …

The SOAP Search Space

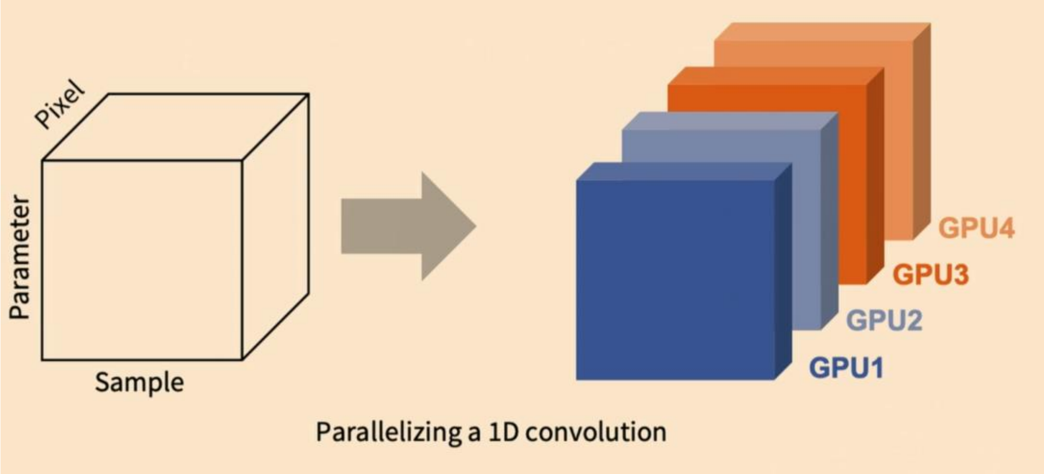
Samples (data parallelism)



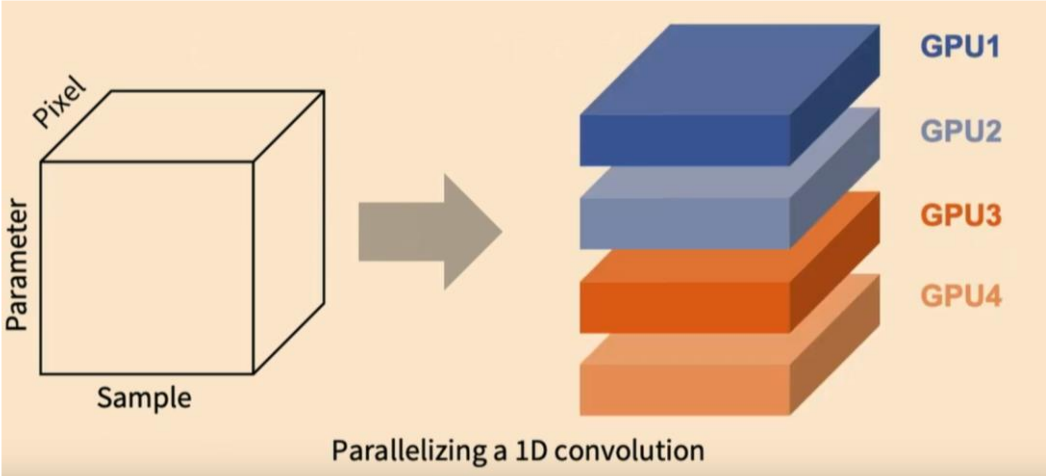
Operators (model parallelism)



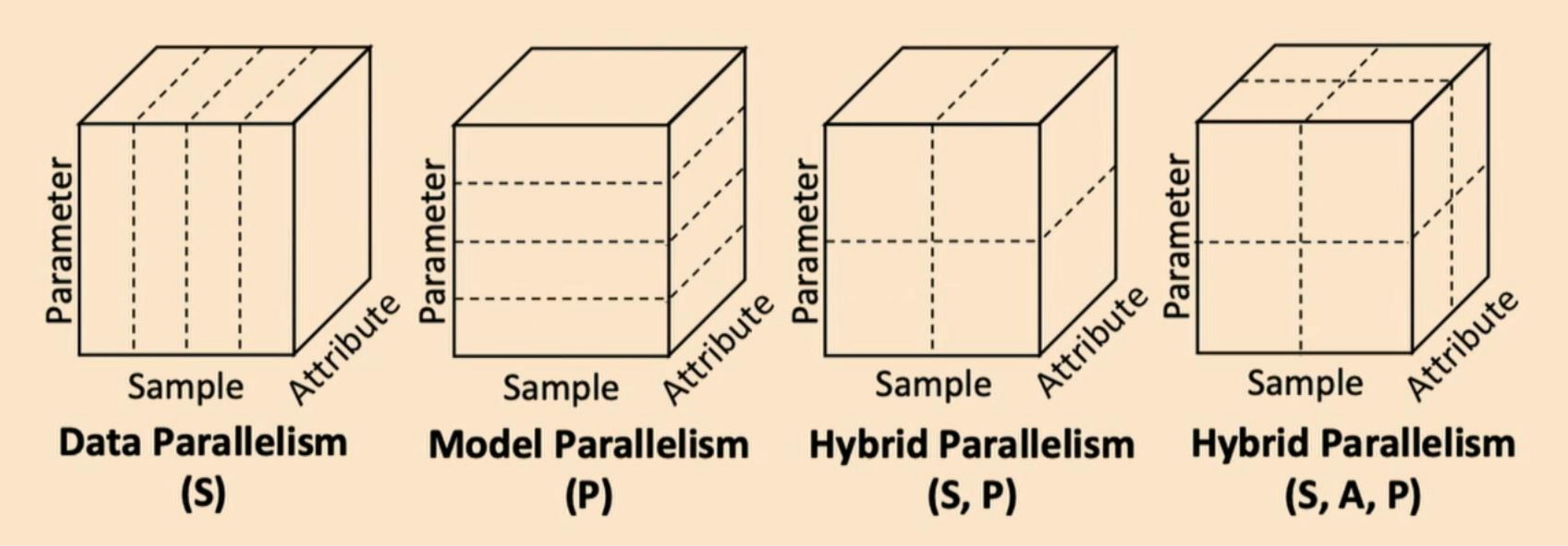
Attributes (e.g. pixels)



Parameters (≈model parallelism)

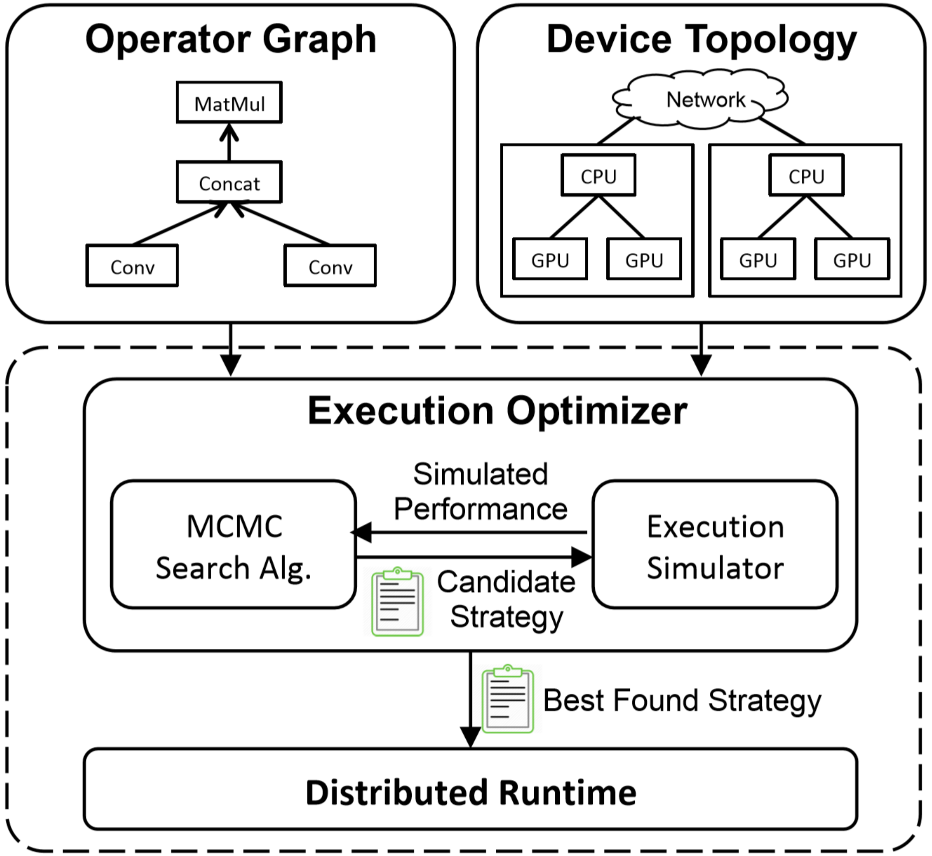


Hybrid Parallelism in SOAP



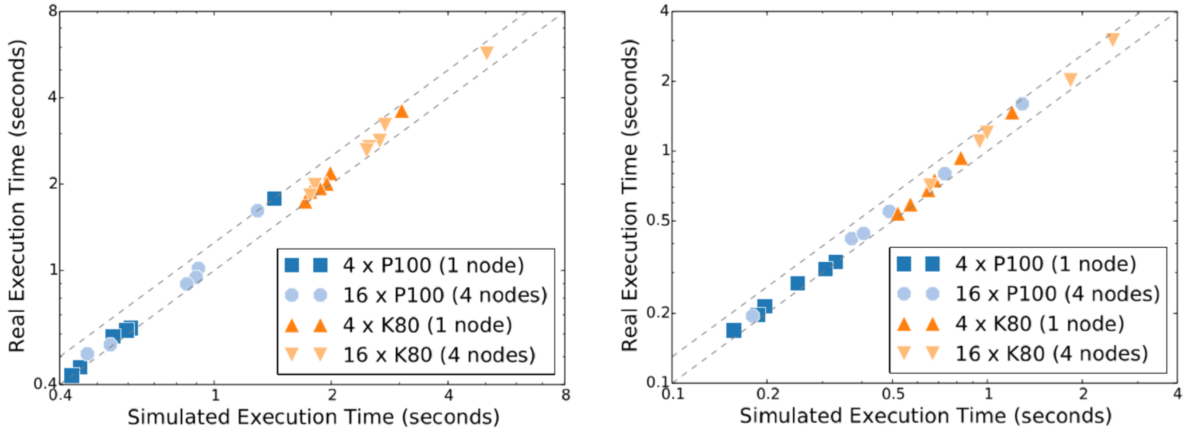
Example parallelization strategies for 1D convolution

FlexFlow



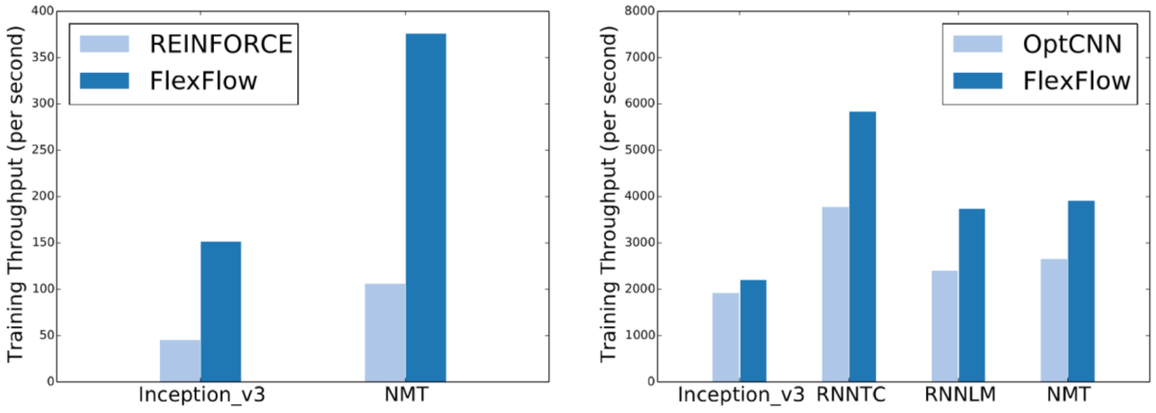
- Trying out strategies on hardware is expensive due to long iteration times
- Execution Optimizer uses simulator instead
  - Measures operator runtime on hardware
  - Estimates runtime of parallelisation strategies
  - Delta simulation algorithm uses incremental updates for acceleration
- Execution optimizer explores search space with Markov Chain Monte Carlo algorithm

Evaluation (1/2)



(a) Inception-v3 (b) NMT

Figure 11: Comparison between the simulated and actual execution time for different DNNs and device topologies.



(a) REINFORCE (b) OptCNN

Figure 10: Comparison among the parallelization strategies found by different automated frameworks.

Evaluation (2/2)

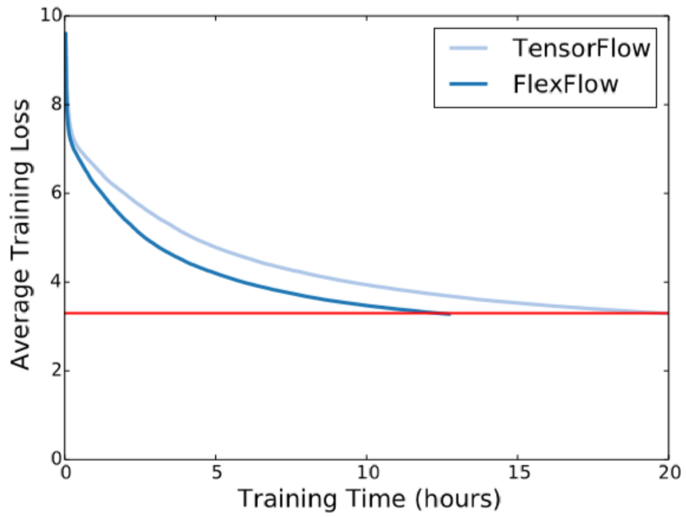
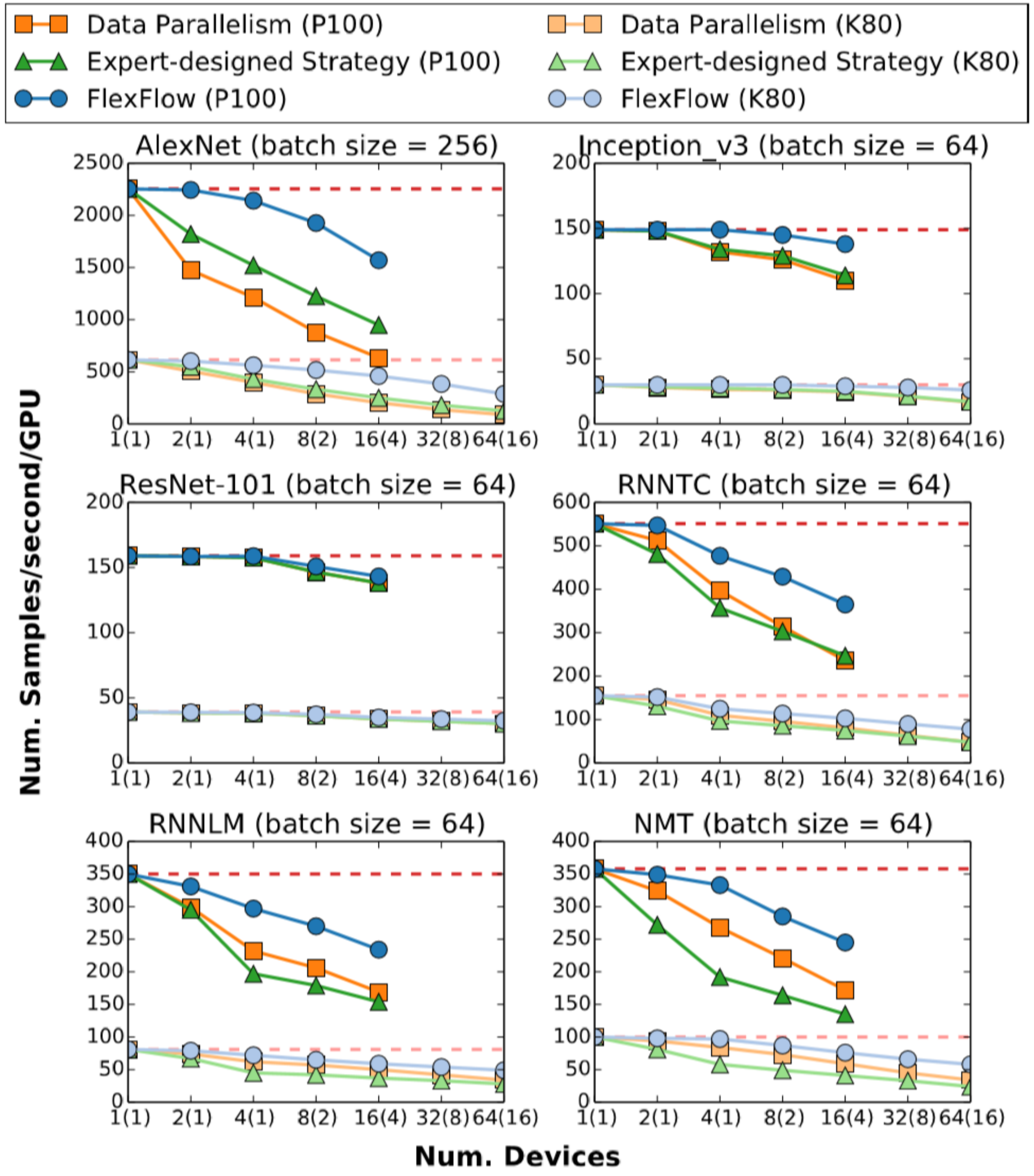


Figure 9: Training curves of Inception-v3 in different systems. The model is trained on 16 P100 GPUs (4 nodes).



Review (1/2)

STRENGTHS/AGREEMENTS

- Expands search space for parallelisation strategies
- Proposes a way to efficiently explore that search space
- Leads to an actual speed-up

WEAKNESSES/DISAGREEMENTS

- Unclear how much SOAP and execution optimiser contribute to training acceleration
- Usefulness of Attribute dimension is questionable
- More end-to-end performance benchmarks would have been useful

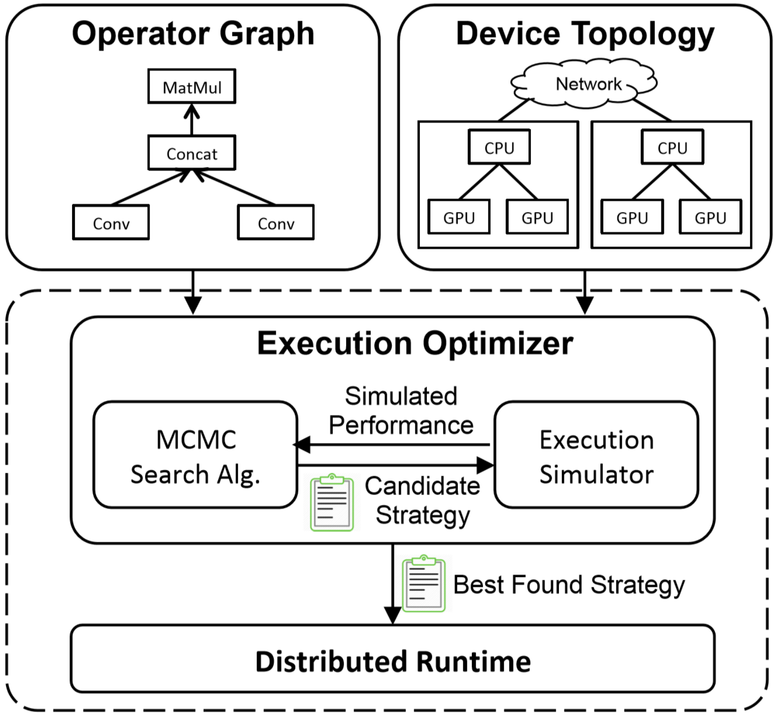
Review (2/2)

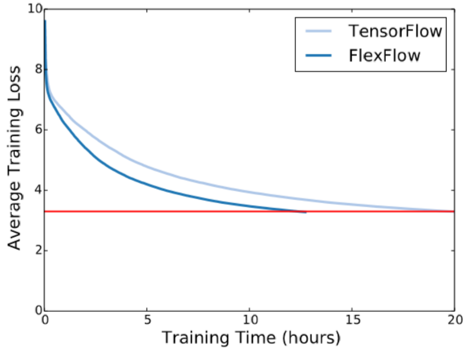
KEY TAKEAWAYS

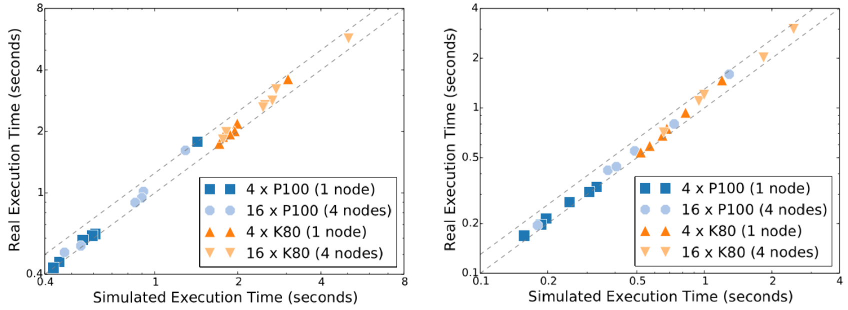
- Training performance of parallelisation strategies can be efficiently and accurately predicted
- The resulting speed-up allows for the exploration of a wider search space

POTENTIAL IMPACT

- Usage of other search algorithms to explore parallelisation search space in simulation
- Combination of parallelisation search space with computation graph substitutions (compare Tim’s presentation next week)

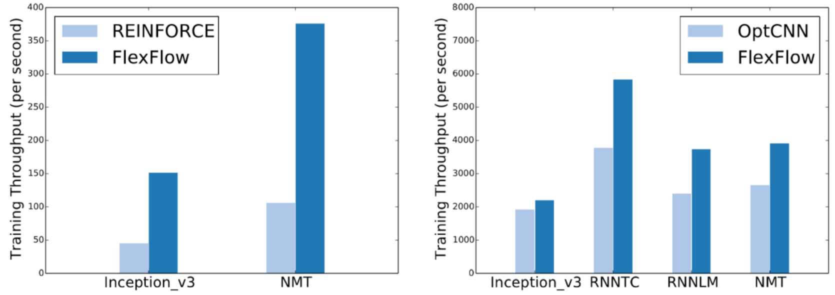






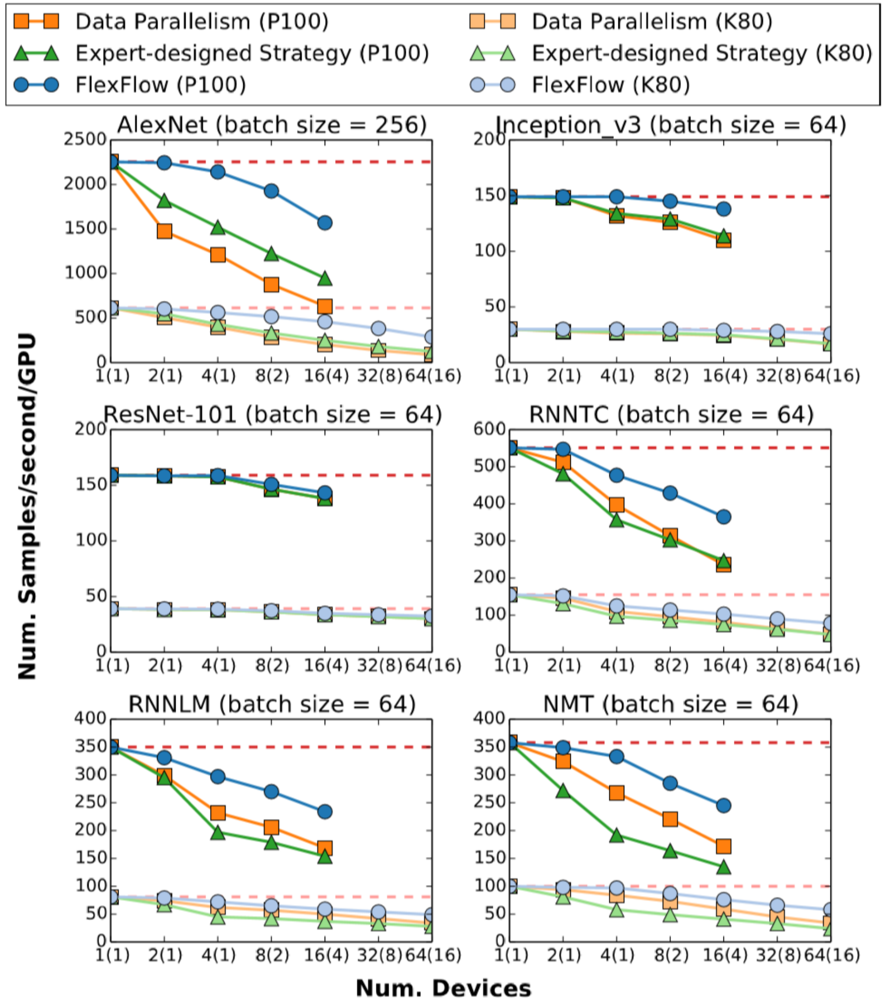
(a) Inception-v3 (b) NMT

Figure 11: Comparison between the simulated and actual execution time for different DNNs and device topologies.

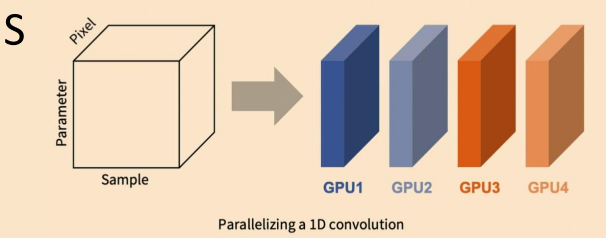


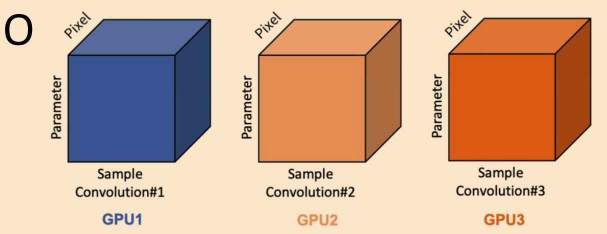
(a) REINFORCE (b) OptCNN

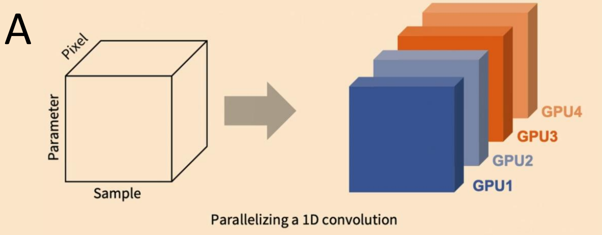
Figure 10: Comparison among the parallelization strategies found by different automated frameworks.



Questions?







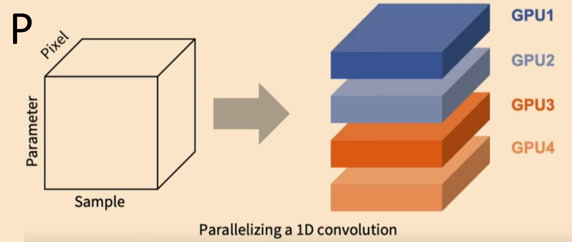


Image Citations

Images with beige background retrieved from Jia Zhihao's SysML 19 talk: https://www.youtube.com/watch?v=81l6kkV-OkE

All other images extracted from Z. Jia, M. Zaharia, and A. Aiken: Beyond Data and Model Parallelism for Deep Neural Networks, SYSML, 2019.