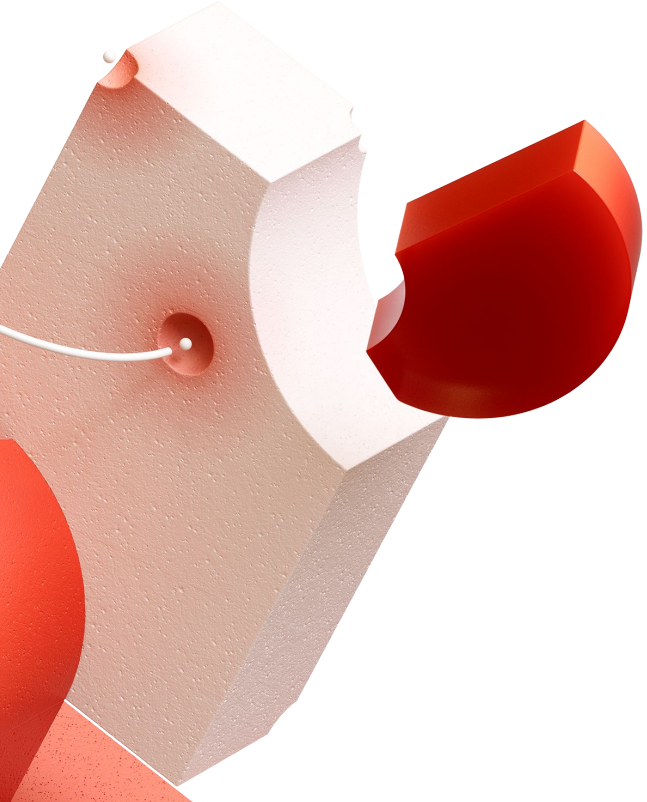
Google DeepMind

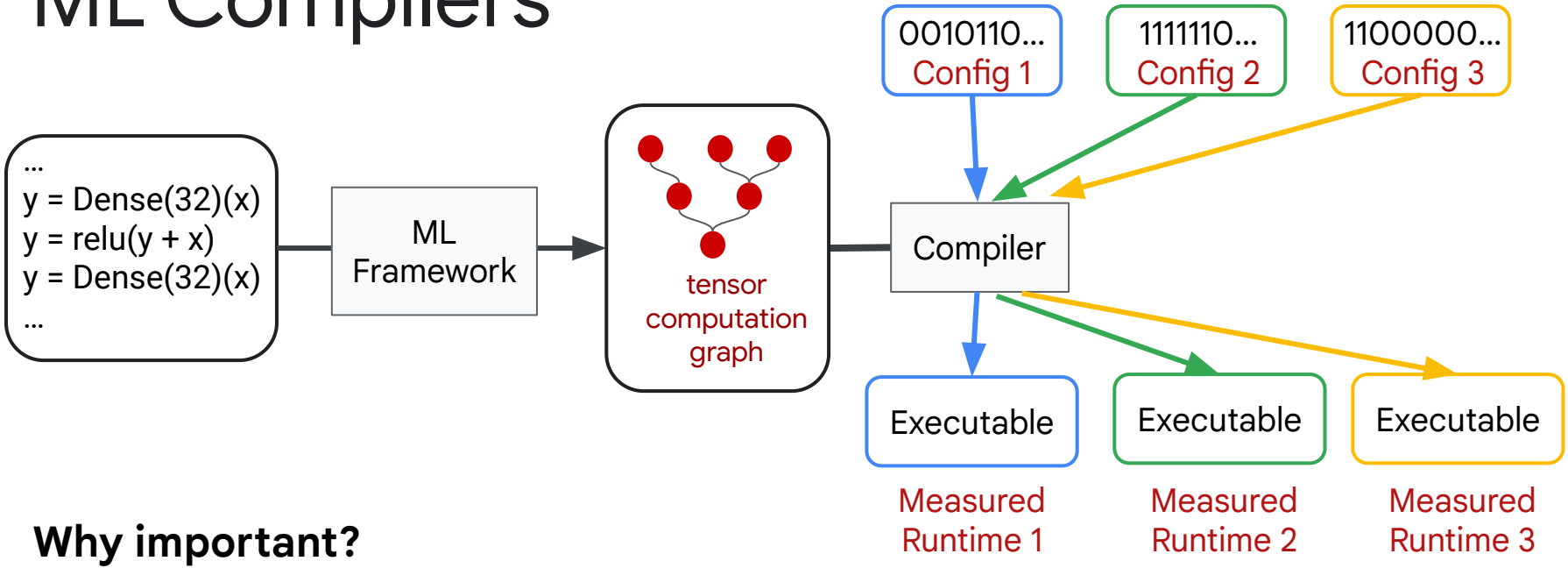


TpuGraphs: A Performance Prediction Dataset on Large Tensor Computational Graphs

Phitchaya Mangpo Phothilimthana*,
Sami Abu-El-Haija*, Kaidi Cao◆, Bahare Fatemi*,
Mike Burrows*, Charith Mendis◇, Bryan Perozzi*

*Google, ◆Stanford, ◇UIUC

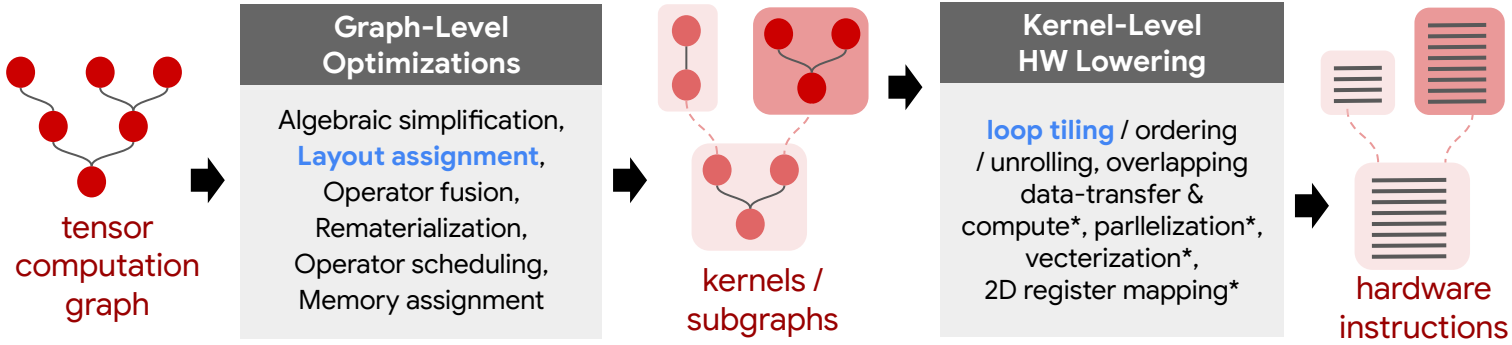
ML Compilers



Why important?
Make ML models faster!

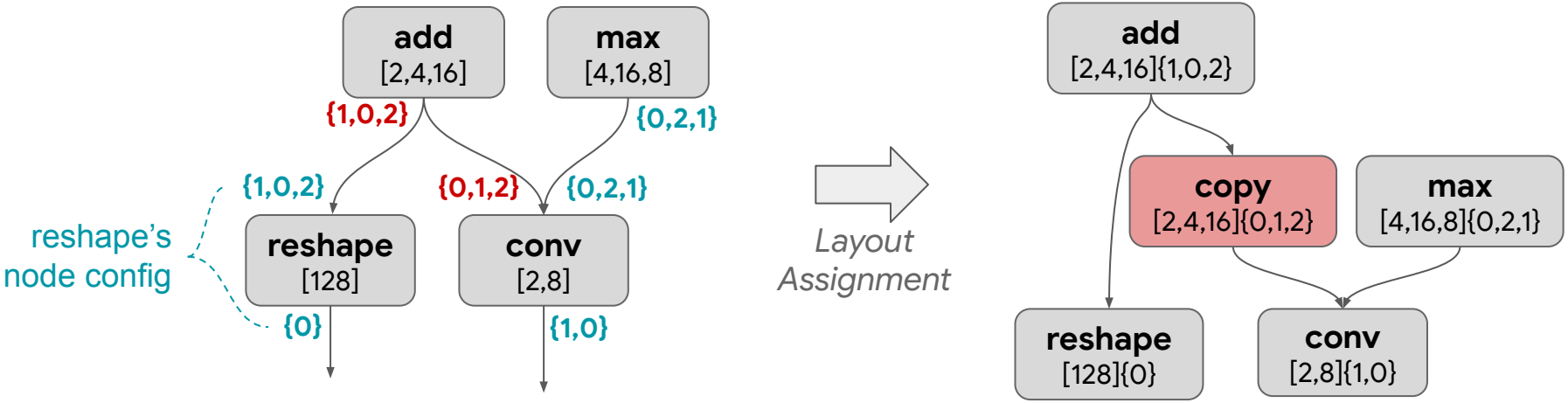
- At Google, we achieve **5-25% speedup** on important production models by searching config space.
- ***Learned cost model*** reduces the search time.

Target Optimizations

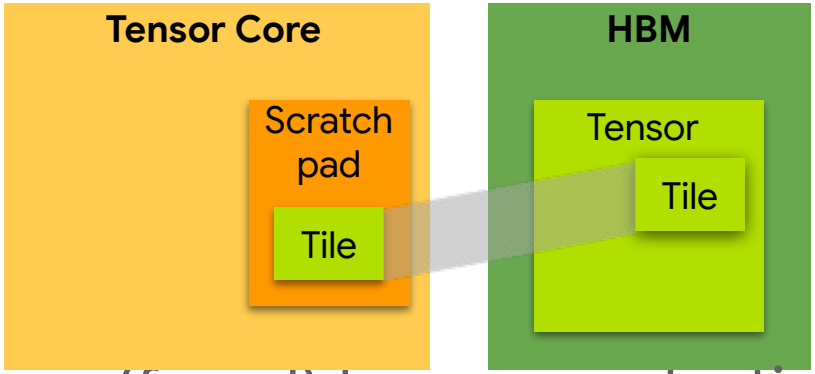


Layout Assignment

Example:



Tiling Optimization: Tile Size Selection



- TPUs process one (fused) tensor op at a time
  - Fetch tiles of input tensors, copy output tiles into scratchpad
  - Supports arbitrary intermediates in scratchpad

TpuGraphs Dataset Collections

{opt}:{src}:{space}

layout
tile

xla: **diverse** types of **ML models** (e.g. vision, NLP, speech, audio, and recommendation)

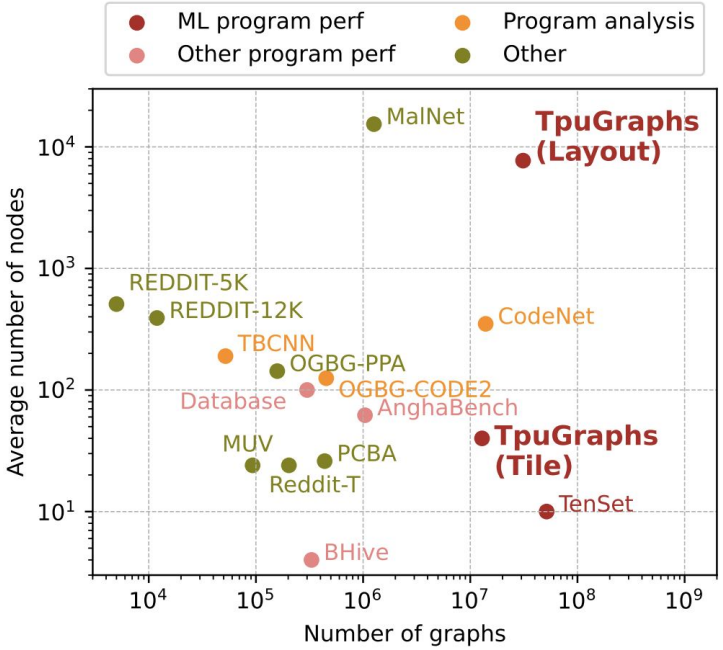
nlp: variety of **transformer** models

default: configs generated from **genetic algorithm**

random: configs generated from **random search**

TpuGraphs Statistic

Collection {opt}:{src}:{space}	Avg # of Nodes	# of Graphs + Configs
Layout:XLA:Default	14,105 (372 - 43,614)	771,496
Layout:XLA:Random		908,561
Layout:NLP:Default	5,659 (876-21,919)	13,285,415
Layout:NLP:Random		16,125,781
Tile:XLA	40	12,870,077



Graph property prediction datasets

Evaluation Metrics

Top-K Error: slow down compared to optimal

$$\frac{\text{The best runtime of the top-k predictions}}{\text{The best runtime of all configurations}} - 1 = \frac{\min_{i \in K} y_i}{\min_{i \in A} y_i} - 1$$

Ranking Correlation: ability to guide the search

```
Kendall-Tau(model rank, gound-truth rank)
```

Best GNN Baselines

Collection	Kendall τ		Top-K_1 E %		Top-K_2 E %		Top-K_3 E %	
	Val	Test	Val	Test	Val	Test	Val	Test
Layout:XLA:Random	0.19	0.34	19.8	10.9	12.3	5.7	9.7	1.6
Layout:XLA:Default	0.12	0.21	3.8	14.1	1.9	0.6	0.3	0.2
Layout:NLP:Random	0.58	0.53	2.1	4.6	2.0	1.0	0.6	0.09
Layout:NLP:Default	0.30	0.28	4.0	4.0	3.7	3.1	3.5	0.13
Tile:XLA	–	–	10.5	10.8	3.9	3.4	2.7	2.1

$$(K_1, K_2, K_3) = \begin{cases} (1, 10, 100) & \text{for layout} \\ (1, 5, 10) & \text{for tile} \end{cases}$$

TpuGraphs

Dataset: github.com/google-research-datasets/tpu_graphs

Competition: kaggle.com/competitions/predict-ai-model-runtime

Come learn more about the conclusion of the competition and the winning strategies at **ML for Systems Workshop @ NeurIPS**