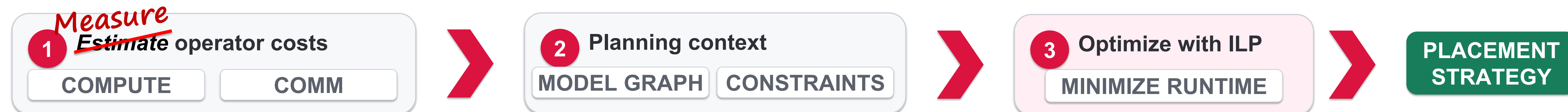


Augmenting AutoParallel with measurement info

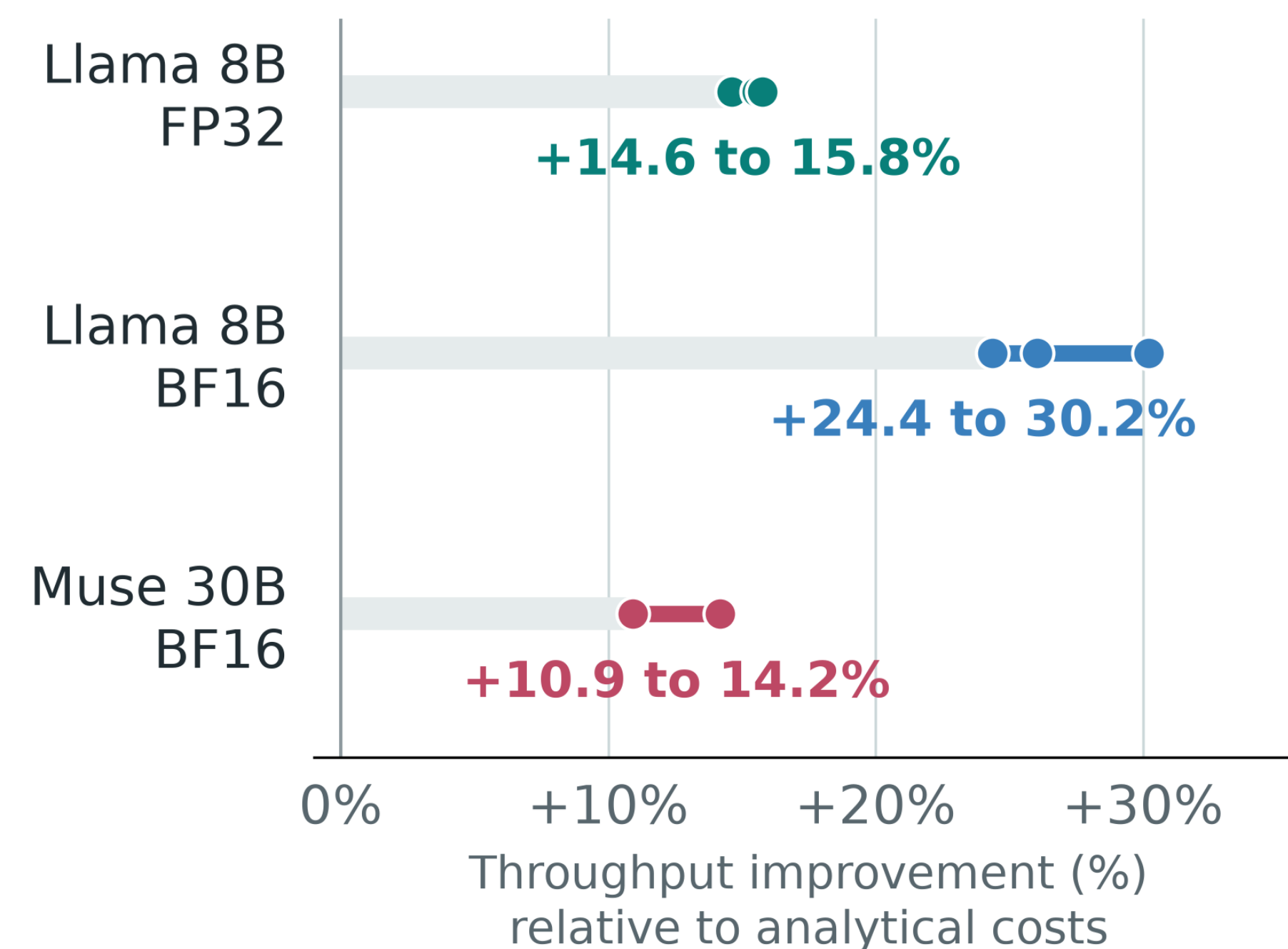
AutoParallel estimates communication and compute costs, maps the parallelization problem to an ILP, then solves the ILP to minimize runtime.

We augment analytical cost estimates with an on-device measurement step to explore actual connectivity and compute capabilities.



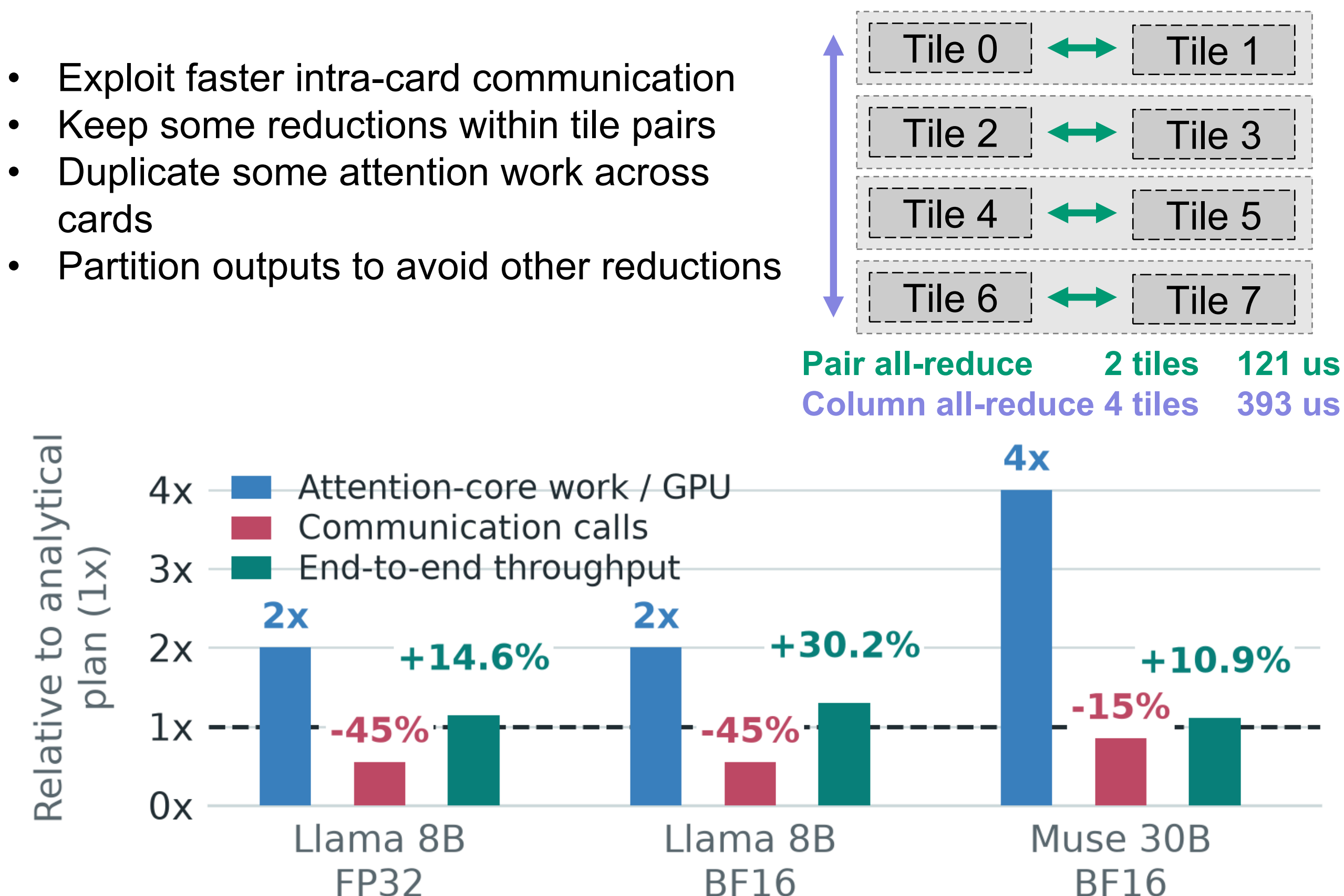
Faster inference with measured costs

Inference on **Intel Max 1550 GPU**, using 8 tiles (4 cards) in a 4x2 configuration.
Global batch 8, sequence length 128



What changes in faster placements?

- Exploit faster intra-card communication
- Keep some reductions within tile pairs
- Duplicate some attention work across cards
- Partition outputs to avoid other reductions

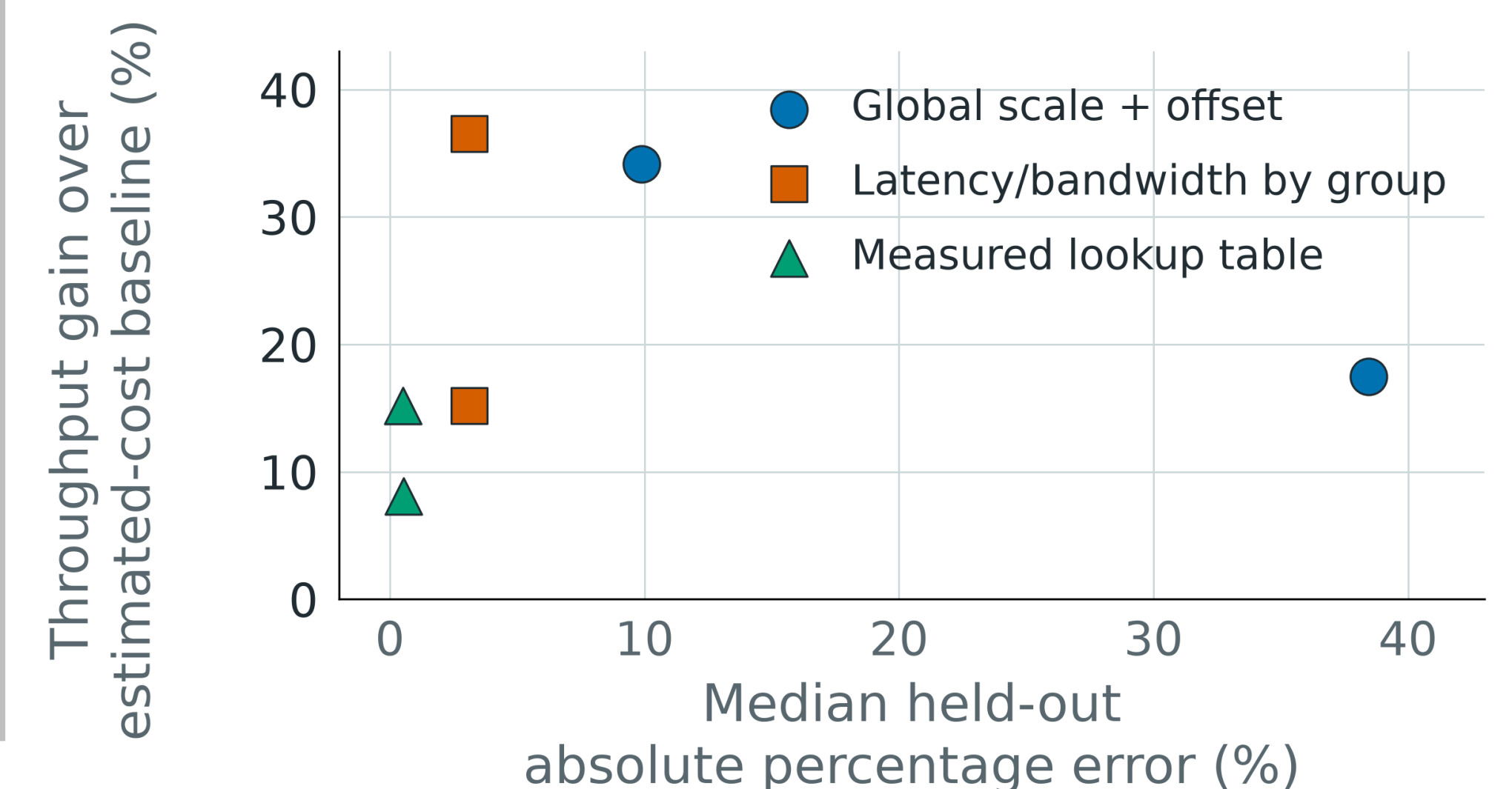


Can simplifying cost estimates still work?

Next, we fit 3 different cost models on a subset of the measured communication data for Llama 3 8B BF16:

1. ● Fit global scale, bias parameters: $t(x) = wx + b$
2. ■ Fit for each collective, mesh dim: $t_g(n) = w_g n + b_g$
3. ▲ A sparser set of measurements in a lookup table

Lower comms prediction error did not reliably produce faster inference.



Measuring costs – what are the cons?

- Significantly higher ILP solve time – up to 50 minutes higher for Muse
- Isolated measurements miss compiled fusion and end-to-end op interactions.
- The most accurate comms lookup table did not produce the fastest plan.
- Selected plans can vary in placement and performance

