Integrated Hardware Architecture and Device Placement Search

Irene Wang, Jakub Tarnawski, Amar Phanishayee, Divya Mahajan

Georgia Tech · Microsoft · ICML International Conference On Machine Learning

Problem: Fast and efficient training of large deep learning models depends on a multiple important factors including device placement strategies and capabilities and architectural details of hardware accelerators.

Our Approach: We devise Phaze, a novel framework for co-optimizing hardware architecture, device placement strategy, and per-chip operator scheduling.

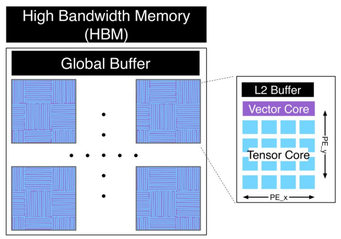
Findings: Our study demonstrated the benefits of co-optimization instead of examining each problem in isolation. Phaze-searched architectures delivers higher throughput compared to state-of-the-art accelerator architectures and other hardware-search framework baselines.

Paper Link

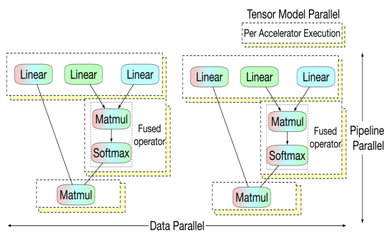
How to Train DNN Efficiently?

Training DNN models require **2 simultaneous design choices** to be made to balance resource utilization and memory footprint

1. Hardware Architecture

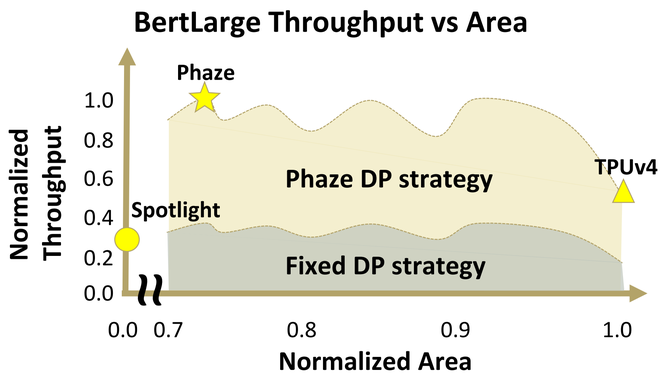


2. Device Placement Strategy



Need for Co-optimization

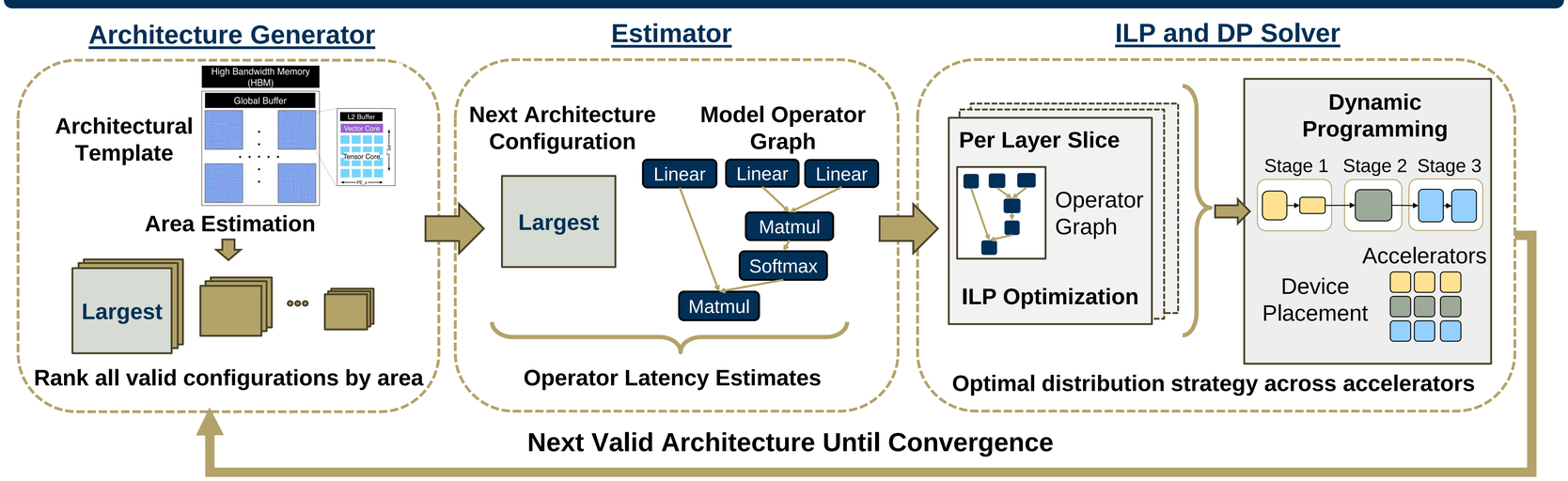
Hardware architecture and device placement strategy should be co-optimized instead of explored in isolation



Fixed device placement in architecture search may lead to hardware under-utilization

Fixed hardware architecture in device placement search limit the search space of memory footprint and networking overhead

The Phaze Framework



Phaze Architecture Characteristics

- Over 91% area utilization of area constraint
- High number of vector cores for parallelization of operations
- Large tensor cores for better reuse
- ≤ 64 GB HBM: Larger memory ≠ Higher throughput

Experiments – Throughput Improvement

Phaze's common architecture across all models, on average, delivers **2.9×** higher throughput compared to TPUv4 with expert device placement strategy