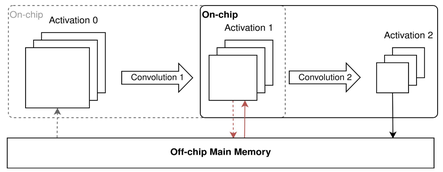
LoopTree: Exploring a More Extensive Fused-layer Dataflow Accelerators Design Space

Michael Gilbert, Nellie Wu, Angshuman Parashar, Vivienne Sze, Joel Emer

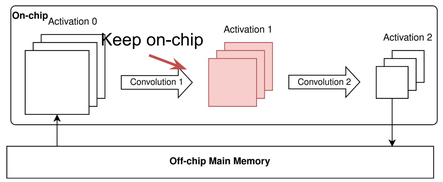
Why Explore Fused-layer Design Space?

(1) Fused-layer Dataflow Reduces Off-chip Transfers



Layer-by-layer processing

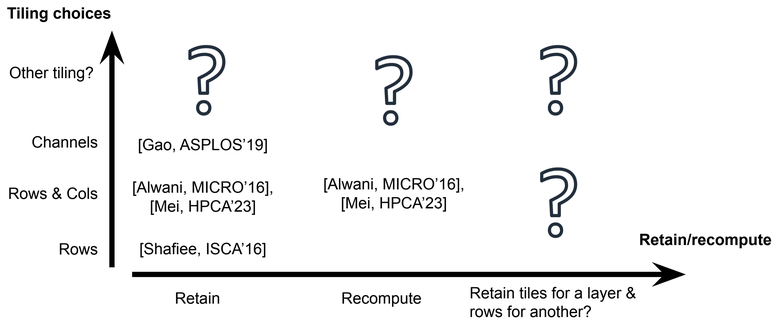
- ✗ Round trip off-chip transfers for every intermediate data



Fused-layer processing

- ✓ No off-chip transfers of intermediate data
- ✗ Requires on-chip buffer for intermediate data

(2) There are Large, Unexplored Portions of the Design Space



Would like to explore this space, but need:

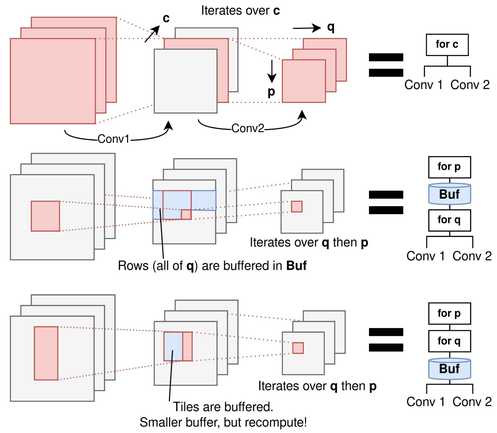
(1) a systematic representation of design space, and

(2) model for an *arbitrary* fused-layer dataflow design?

The LoopTree Framework

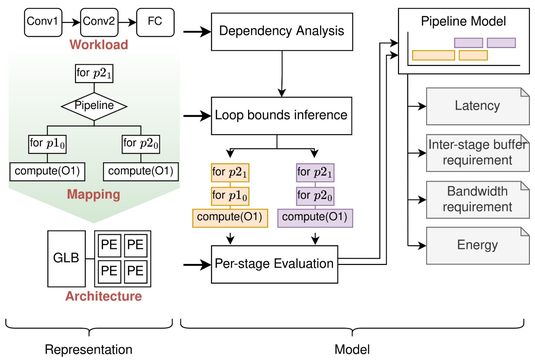
Solving (1): LoopTree Specification

Insight: Dataflow can be described with nested loops and storage levels



Solving (2): LoopTree Analytical Model

Quickly and accurately evaluating the design.



Analysis Steps:

Dependency analysis extracts read and write access patterns.

Loop bounds inference uses dep. relations to infer unspecified loop bounds.

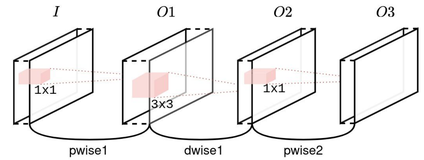
Per-stage evaluation generates metrics of individual stages.

Pipeline model evaluates metrics of entire pipeline.

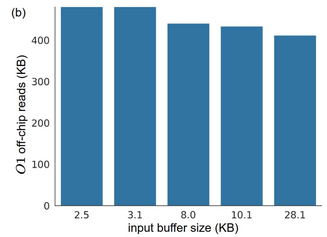
Case Study: Exploration with LoopTree

Workload: MobileNet inverted bottleneck block

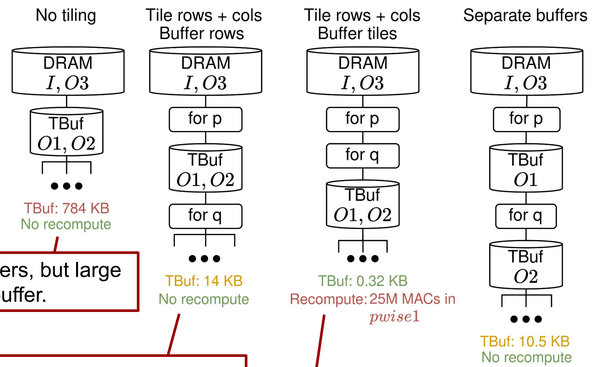
Common in state-of-the-art CNNs.



Challenge: low intra-layer reuse even with large buffers



Exploring fused-layer designs with LoopTree



No off-chip transfers, but large on-chip buffer.

Significantly smaller on-chip buffer.

Even smaller on-chip buffer, but frequent recomputations.
Insight: **all recompute are from O1.**

25% smaller on-chip buffer than state-of-the-art fusion.

2x lower latency
3x lower off-chip transfers
2x lower energy
than optimized layer-by-layer

References

M. Alwani, H. Chen, M. Ferdman, and P. Milder, "Fused-layer CNN Accelerators," in 2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), 2016, pp. 1–12.

M. Gao, X. Yang, J. Pu, M. Horowitz, and C. Kozyrakis, "Tangram: Optimized coarse-grained dataflow for scalable nn accelerators," in Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems, ser. ASPLOS '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 807–820.

L. Song, X. Qian, H. Li, and Y. Chen, "Pipelayer: A pipelined reram-based accelerator for deep learning," in 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA), 2017, pp. 541–552.

A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, "Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," in 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), 2016, pp. 14–26.