Next-Generation PE Architecture

Nikhil Bhagdikar

Motivation

- Efficient support for a wider range of applications
- Better integration with AHA flow
- Improve energy and area efficiency

Data Types

- Int4 for machine learning (inferencing)
- Int8 for imaging pipelines
- Int16 for imaging, vision, and machine learning (training)
- B-Floats (16-bit 8+8 floating point)

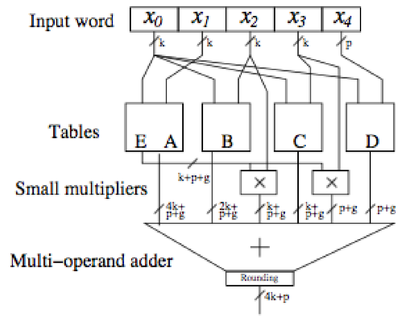
New Instructions

- Non linear functions (log, exponentials, trigonometric)
- Packing (eg: two 8-bit objects to a 16-bit word)
- Conversion (eg: int to float)

Integration

- Create a global spec for the PE
- Improves compiler/mapper to hardware interface
- New instructions are readily absorbed by the flow
- Robust verification

Implementing Non Linear Functions

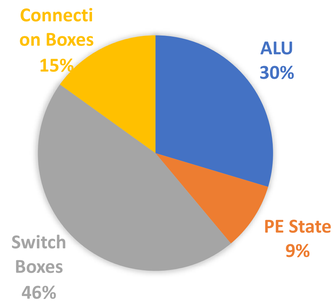


A new scheme for table based evaluation of functions [David et al., 2006]

Current Work: Evaluating efficiency tradeoff

- Specialized units
- Specialized routing in existing units
- Non specialized

Improving Energy/Area Efficiency



PE Area Distribution

**Data from 16nm CGRA chip taped out*

Future Work:

- Heterogeneity
- Improved multipliers and pipelining
- Data/Clock gating

ISTC Agile