Next-Generation PE Architecture

Nikhil Bhagdikar

Motivation

- Increase supported application space
- Better integration with AHA flow
- Improve energy and area efficiency

Data Types

- Int4 : ML inferencing
- Int8 : Imaging
- Int16 : ML training/imaging
- B-Floats

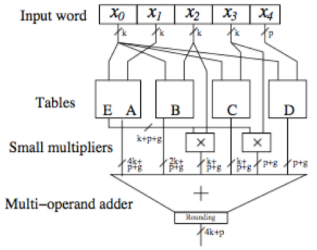
Instructions

- Non linear functions (log, exponentials, trigonometric)
- Packing
- Conversion

Integration

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

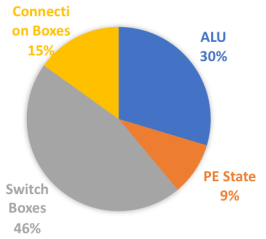
Implementing Non Linear Functions



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