Update on Ara

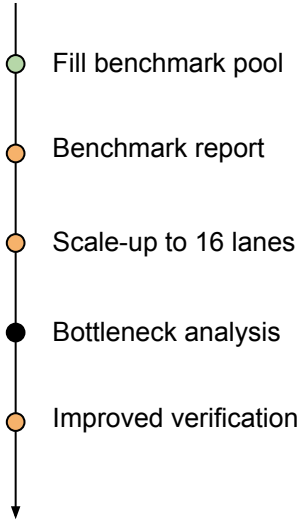
26/10/2022

Matteo Perotti
Matheus Cavalcante
Nils Wistoff

Professor Luca Benini
Integrated Systems Laboratory
ETH Zürich

Summary

- **Software**
  - New benchmarks upstream
- **Hardware (RTL + Backend)**
  - Fixed-Point support
  - Scale to 16 lanes
  - Merge Fixed-Point support

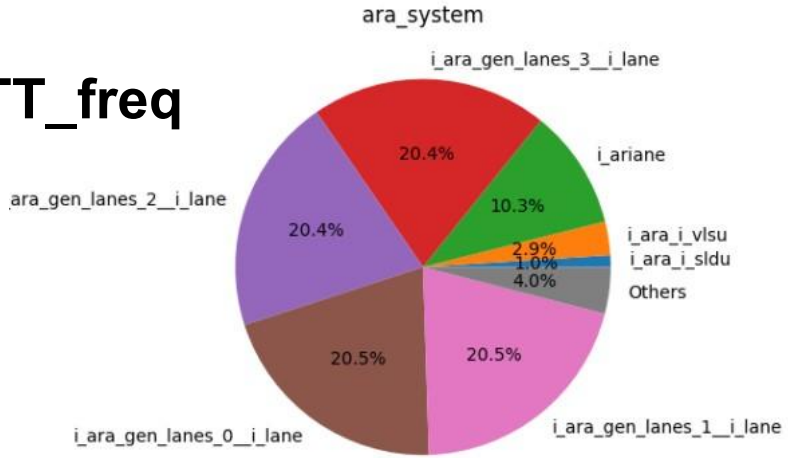


Power

- Set up generic flow to get **power analysis**

- **`fmatmul` 128x128, 4-lanes**
  - Updated system with FP-reductions
  - **~same power than before**
  - **From 280 mW to 290 mW @TT_freq**

- **~36 DP-GFLOPs/W**



HW - Scale up to 8 lanes

- **First** run:
  - **~950 MHz***
  - * without aggressive DRC/hold fixing
- **Second** run:
  - **Vertical channels among the $ banks**
  - **More space to solve DRC/hold**
- **Next** run:
  - **Need for larger channels!**

HW - Scale up to 16 lanes

- **First** run:
  - >10 days run
  - May require RTL modifications
  - May require different die shape

- **Next** run:
  - Pipelined Slide Unit?
  - Lanes around the system?

PR

- **Merged:**
  - Jacobi2d, Dropout, DWT, FFT
  - Ideal-dispatcher
  - Misc (EEW fix, VRF size, SPIKE patch)

- **Ongoing:**
  - softmax, [f]dotp, pathfinder, roi-align
  - Fixed-Point support

- **Next:**
  - AWB, spmv, lavaMD

Further

- **Software**
  - spmv, AWB, lavaMD
- **Hardware (RTL + Backend)**
  - Merge fixed point support
  - Try different die shapes
  - Close timing with 16 lanes