

Update on Ara

09/02/2022

Matteo Perotti

Matheus Cavalcante

Nils Wistoff

Professor Luca Benini

Integrated Systems Laboratory

ETH Zürich

Summary

- **Roi Align - First results**
 - Roi Align
- **T-HEAD C906**
 - Benchmarking
- **fmatmul energy efficiency**
 - Floorplan changes
 - Bug fixes
 - Efficiency
- **Projects about Ara**
 - Add missing instructions for RVV 1.0
 - Verification for Ara

roi_align benchmark

- **Debug** vector code with SPIKE
 - Soon a PR with benchmarks on SPIKE
 - Fully parametrized (scalar implementation from PyTorch)
- **Performance** - whole function, **512 channels** input
 - BxCxHxW memory layout
 - **3.9x speedup** w.r.t. scalar implementation
- **Optimizations**
 - Strided and misaligned memory operations
 - Unoptimized instruction scheduling (all loads - computations - stores)

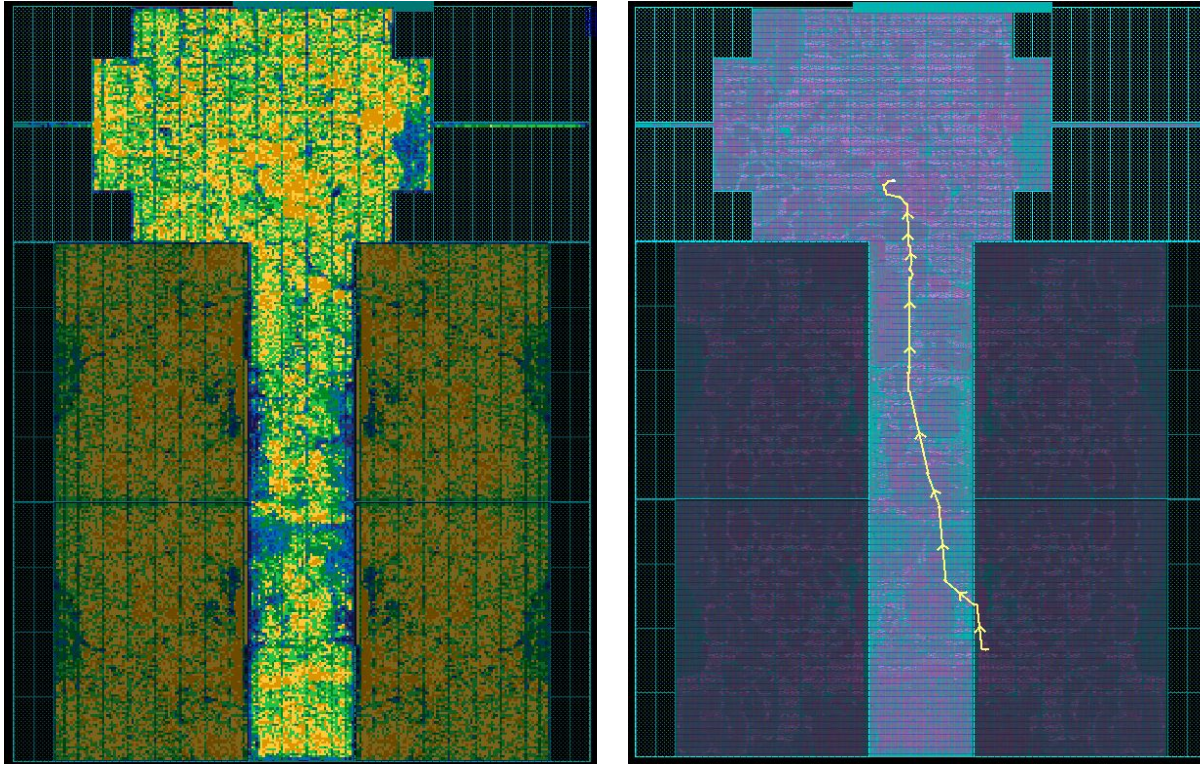
T-HEAD C906

- T-HEAD C906 **open source** industrial vector processor?
- Source code (<https://github.com/T-head-Semi/openc906>) **without V extension** :(
- We ordered the physical **board**, and we are trying it!
- **Benchmark** and **comparison** with Ara


Simulation with SDF

- **Implementation changes**
 - Move and flip D\$ macros
 - Add horizontal channel for buffers and de-congestion
- **Fix input constraints**
 - Input delays w.r.t. an ideal virtual clock
 - Reset
- **fmatmul simulated with SDF back-annotation**

Cell utilization map and critical path of the latest Ara



Efficiency (preliminary)

- **fmatmul 128x128 - 4 Lanes**
- **Power results @1.36 GHz, TT**
 - Old, no SDF: 220 mW
 - New, no SDF: 197 mW  Backend flow improvements!
 - New, with SDF: 212 mW
- **Estimated efficiency**
 - ≈ 37 DP-GFLOPs/W (-2% from AraV1)
 - *Why the efficiency loss?*
 - To Do: power breakdown.

Bug Fixes

- **Mask Unit**

- Ask correct number of elements to the element requesters

- **Multicycled FPU divider/sqrt**

- Old loose multicycle constraints: risk for hold violations in post layout
- Strict multicycle constraints: too much area overhead
- WIP:
 - RTL fix to allow safe loose multicycle constraints
 - Replace the divider with a pipelineable unit (colleague's project)

Projects

- **Ara RVV 1.0 compliance**

- https://iis-projects.ee.ethz.ch/index.php?title=New_RVV_1.0_Vector_Instructions_for_Ara

- **Ara verification**

- <https://github.com/google/riscv-dv>
- <https://github.com/ucb-bar/riscv-torture>

Further

- Develop full net with roi_align
 - Run **full net** to see real performance gains
 - Spot **bottlenecks** for real applications
 - More concrete **memory layout optimization**
- **Optimize roi_align** instruction scheduling / memory accesses
- Vectorize **Embench**