Soft Tiles: Capturing Physical Implementation Flexibility for Tightly-Coupled Parallel Processing Clusters

Gianna Paulin[1], Matheus Cavalcante[1], Paul Scheffler[1], Luca Bertaccini[1], Yichao Zhang[1], Frank Gürkaynak[1], Luca Benini[1,2]
[1]ETH Zurich; [2]University of Bologna

Modern high-performance computing architectures (Multicore, GPU, Manycore) are based on **tightly-coupled clusters of processing elements** which are **physically** implemented as **rectangular tiles.**

Goal: achieve a high utilization for the top-level die floorplan.:

- **size** and **aspect ratio** strongly impact the achievable **QoR**
- as **flexible** as possible to achieve a high utilization for the top-level die floorplan.

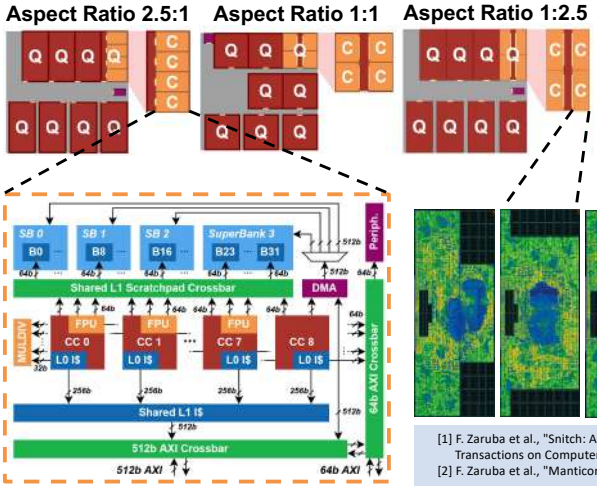
We focus on an **open-source, high-performance cluster tile** with **8x compute** (+1x control) **RISC-V cores** connected to a shared L1 SPM through a low-latency interconnect [1].

Similar to the state-of-the-art architectures, the cluster tile is then **replicated** to build a **scaled-up high performance acceleration system** [2].

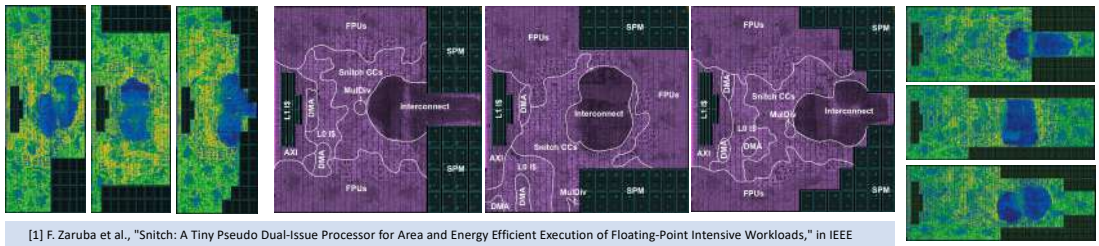
We **explore** the **QoR** of the **physical implementation** of this cluster as a **soft tile** based on a **flexible range of aspect ratio** and **memory macro placement** styles for a **fixed area** of **0.9 $\mathbf{mm^2}$**.

We used *Synopsys Fusion Compiler 2020.09* to synthesize, place, and route the cluster in *Globalfoundries'* ***12 nm advanced FinFET*** technology node at 1 GHz worst-case conditions (SS, 0.72V, 125 °C).

Aspect Ratio 2.5:1 **Aspect Ratio 1:1** **Aspect Ratio 1:2.5**

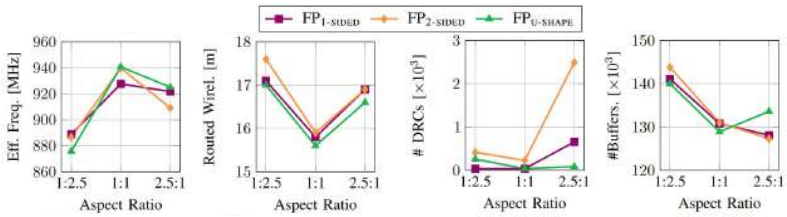


Aspect Ratio	1:2.5			1:1			2.5:1		
Floorplan	$FP_{1\text{-SIDED}}$	$FP_{2\text{-SIDED}}$	$FP_{U\text{-SHAPE}}$	$FP_{1\text{-SIDED}}$	$FP_{2\text{-SIDED}}$	$FP_{U\text{-SHAPE}}$	$FP_{1\text{-SIDED}}$	$FP_{2\text{-SIDED}}$	$FP_{U\text{-SHAPE}}$
Eff. Freq. [MHz]	888.8	886.5	875.7	927.6	939.8	940.7	921.7	909.1	925.1
TNS [ns]	-33.8	-48.2	-103.3	-25.5	-30.2	-24.7	-37.5	-40.2	-78.2
#Violating Paths	5352	5787	6819	4890	5372	4459	6163	5871	8271
RtWL [m]	17.1	17.6	17.0	15.8	15.9	15.6	16.9	16.9	16.6
#DRCs	36	417	259	38	227	38	654	2943	86
#Buffers	141.1 E3	143.8 E3	140.0 E3	130.8 E3	131.0 E3	128.9 E3	138.1 E3	137.3 E3	133.6 E3
Cell Density	59.5 %	60.7 %	59.7 %	57.3 %	57.9 %	57.4 %	58.7 %	58.9 %	58.5 %



[1] F. Zaruba et al., "Snitch: A Tiny Pseudo Dual-Issue Processor for Area and Energy Efficient Execution of Floating-Point Intensive Workloads," in IEEE Transactions on Computers
[2] F. Zaruba et al., "Manticore: A 4096-Core RISC-V Chiplet Architecture for Ultraefficient Floating-Point Computing," in IEEE Micro

A big thanks to our partners:
GLOBALFOUNDRIES
SYNOPSYS
Rambus



RISC-V

Occamy: a 432-core RISC-V Based 2.5D Chiplet System with > 1 Billion Transistors per Chiplet

Peak performance per chiplet:

- **384 Gflop/s DP @1GHz**
- **768 Gflop/s SP @1GHz**

HBM DRAM Bandwidth:

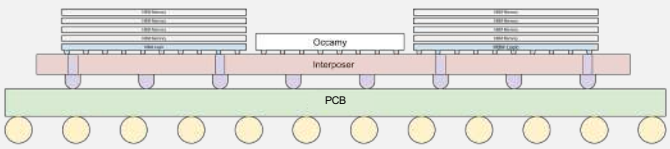
- **358 GB/s @1GHz**

Die-to-Die Bandwidth:

- **70 Gb/s @ 125MHz pad speed**
- **2Gb/s @ 125MHz pad speed**

Off-system Bandwidth:

- **2 Gb/s @ 125MHz pad speed**



Key features:

- **FPU** with **Mini-float** (ML training, Transformers):
  - **FP8** (1, 5, 2)
  - **FP8ALT** (1, 4, 3)
  - FP16 (1, 5, 10)
  - FP16ALT (1, 8, 7)
  - **Expanding SDOTP** Unit
- **Sparsity** support (Stencils, Sparse Tensors)
- **Atomics** and fast **interrupts** (synchro & offload accel.)

ETH zürich

github.com/pulp-platform/snitch

pauling@iis.ee.ethz.ch

DESIGN AUTOMATION CONFERENCE