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 use 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).

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).

A big thanks to our partners:

**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

- **Die-to-Die 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 (sychro & offload accel.)