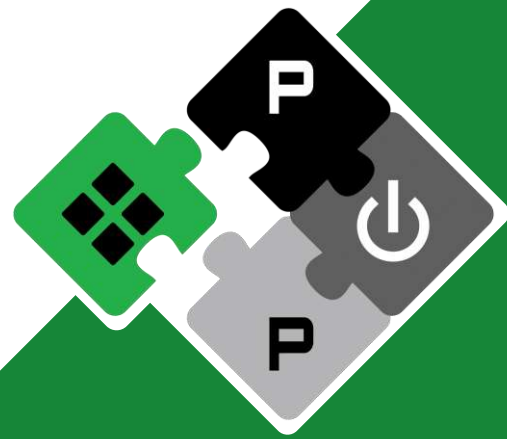
ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

From Cores to Chiplets: PULP's Adventure in Open-Source HPC

Gianna Paulin pauling@iis.ee.ethz.ch
and the **PULP team**

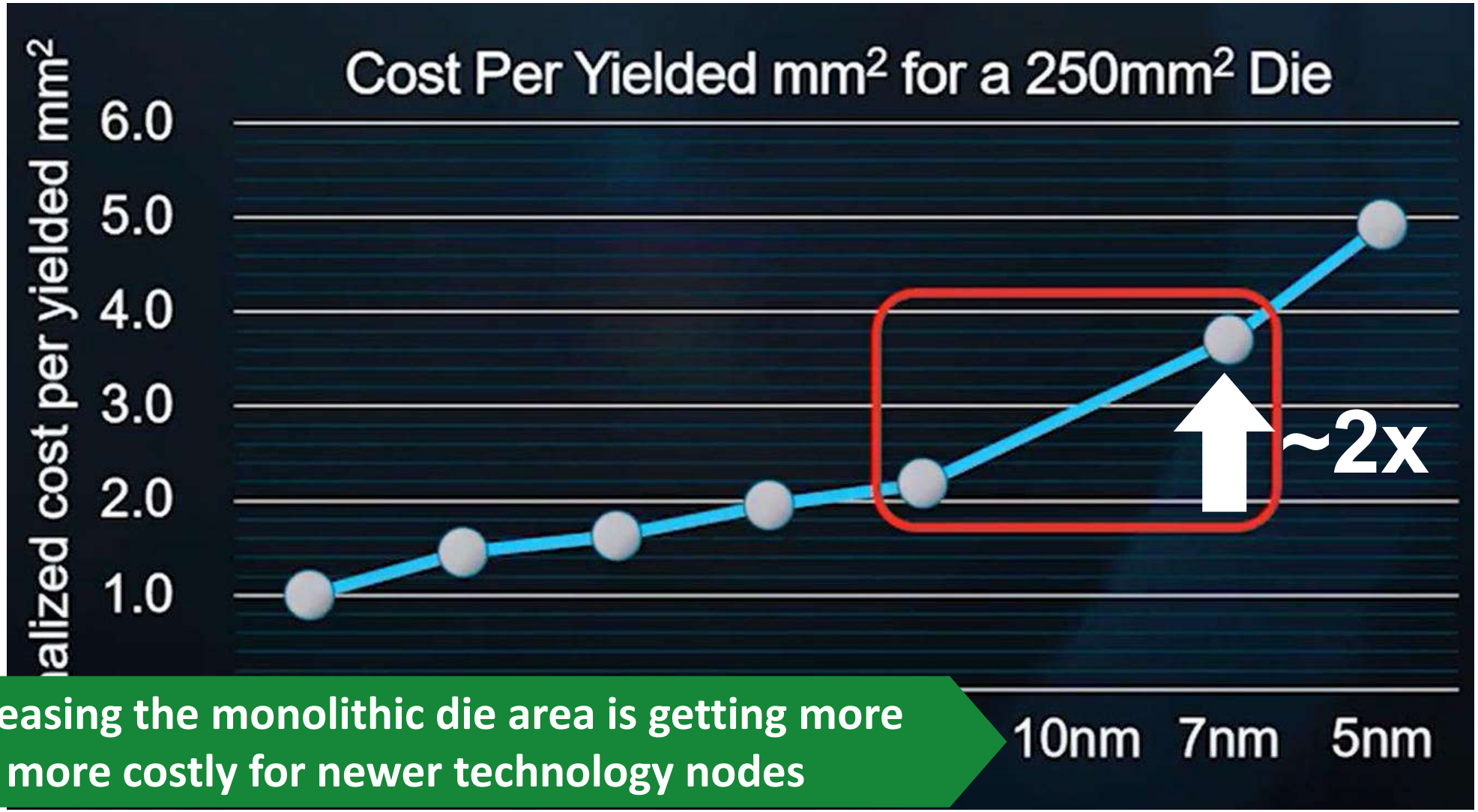


GlobalFoundries™ AVERY design systems
Rambus Micron SYNOPSYS®

PULP Platform
Open Source Hardware, the way it should be!

@pulp_platform
pulp-platform.org
youtube.com/pulp_platform

Cost/Yield increases with more advanced feature nodes



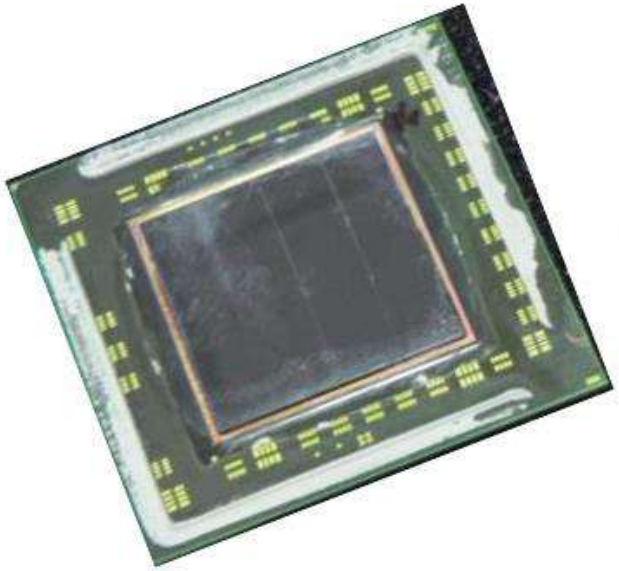
Increasing the monolithic die area is getting more and more costly for newer technology nodes

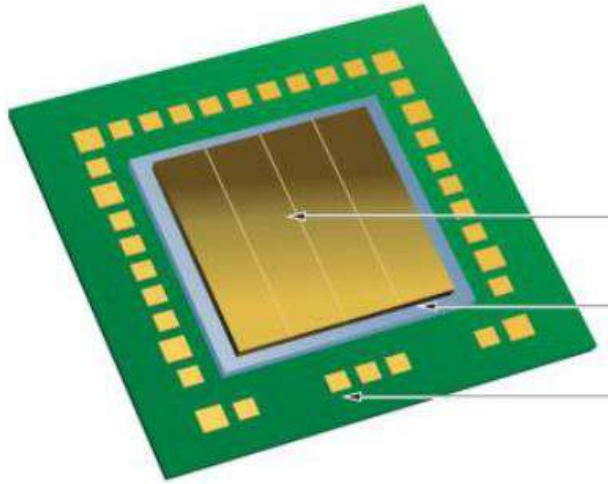
Gabriel Loh, AMD Research, 2021, "An Overview of Chiplet Technology for the AMD EPYC™ and Ryzen™ Processor Families", https://www.youtube.com/watch?v=wqRAG_5KzBE

Xilinx: First 2.5D FPGA Chiplet Design in 2011

eSilicon: Virtex-7-like estimation for 40nm process

- Baseline of **24mm x 24mm monolithic** chiplet**: 25% yield**
- Splitting partitioning into **four 24mm x 6mm** dies: **70% yield**
- Even with the **additional cost of the interposer** the **cost saving >50%**



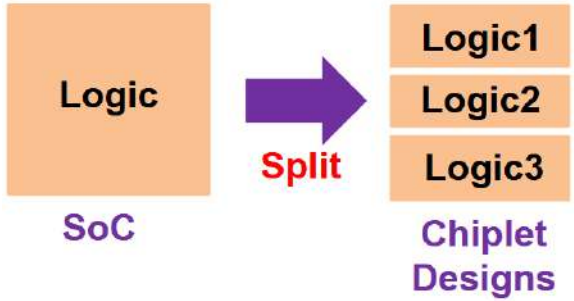


For better manufacturing yield (to save cost), a very large SoC has been split into 4 smaller chips.

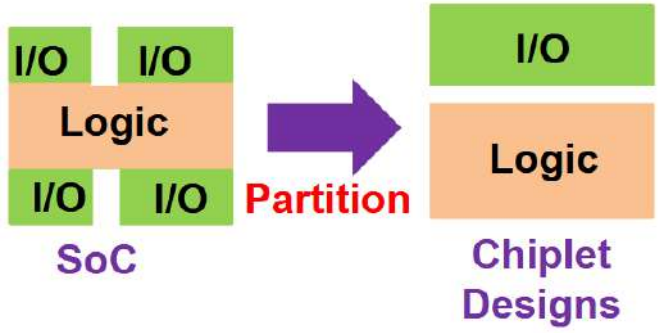
https://semiwiki.com/fpga/xilinx/514-the-worlds-smallest-printed-circuit-boards-interposers/

Chiplet Design and Packaging Technologies

Split Logic



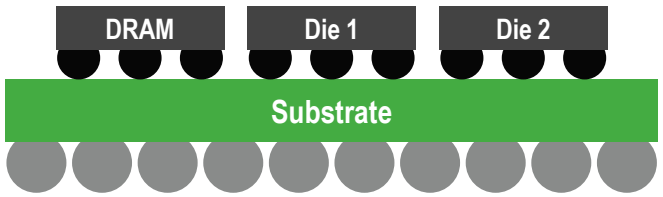
Partition



John H. Lau, Unimicron Technology Corporation, "Chiplet Design and Heterogeneous Integration Packaging", SWISS IEEE/EPS and SSCS Lecture

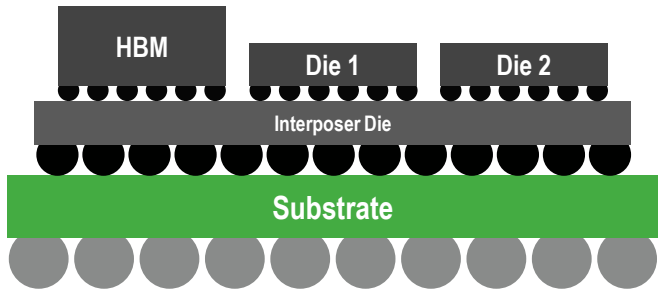
2D / Stnd. Package

Organic substrate



2.5D / Adv. Package

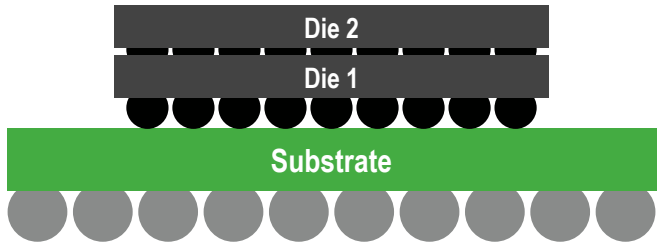
Interposer



Occamy

3D / Die Stacking

Hybrid Bonding



Our latest design Occamy: 0.75 TFLOP/s, 400+ cores

Dual Chiplet System Occamy:

- 216+1 RISC-V Cores per chiplet
- 0.75 TFLOP/s entire system
- GF12LPP
- Area: 73mm^2

2x 16GByte HBM2e DRAMs Micron

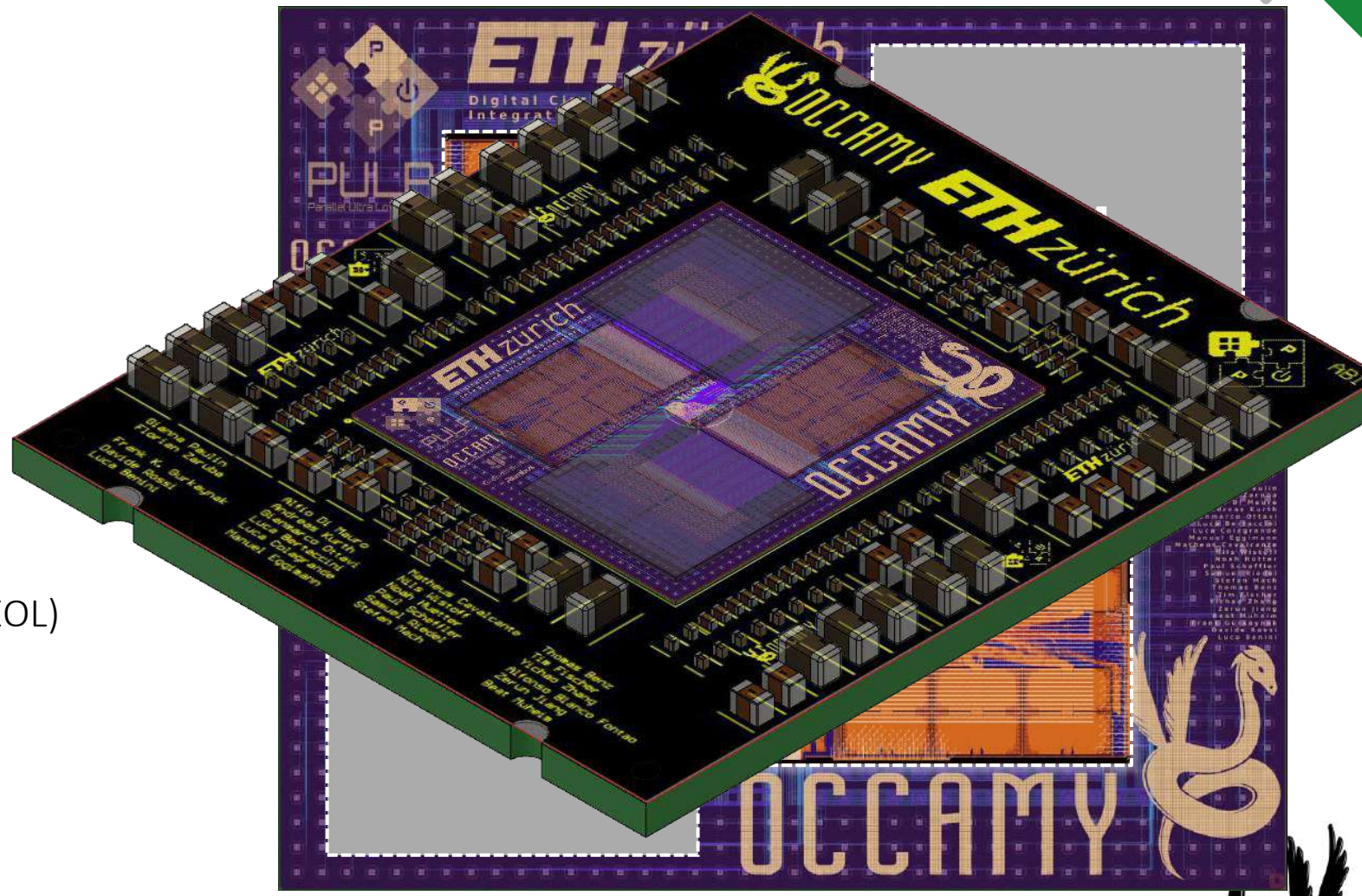
2.5D Integration

Silicon Interposer Hedwig:

- Technology: 65nm, passive (only BEOL)
- Area: 26.3mm x 23.05mm

Carrier PCB:

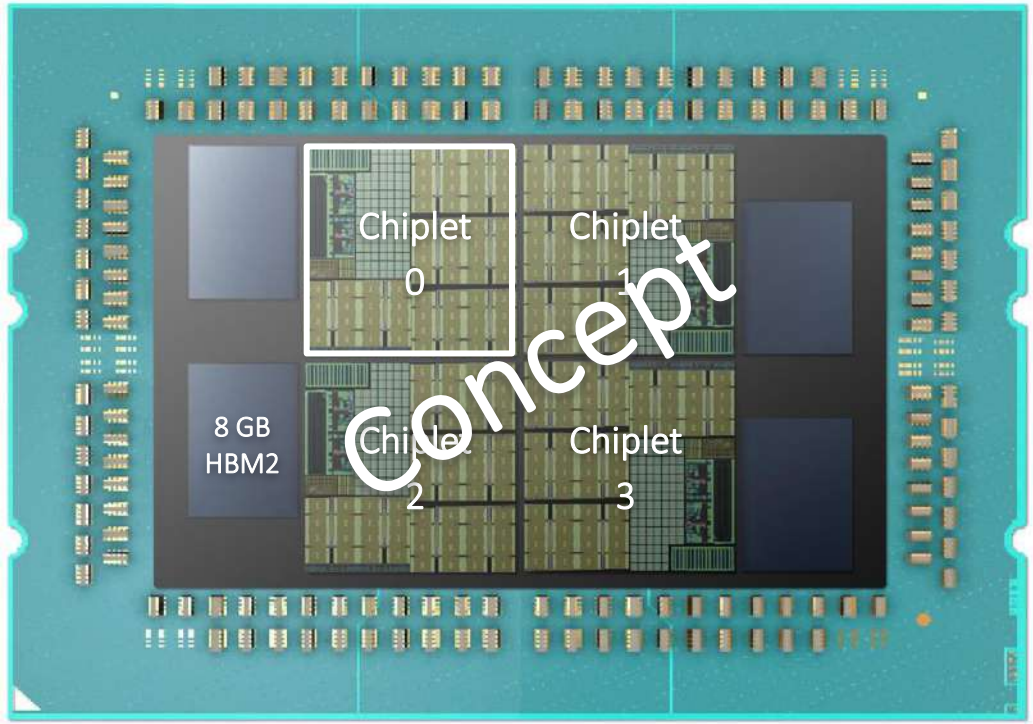
- RO4350B (Low-CTE, high stability)
- 52.5mm x 45mm



How did we get here?

Concept architecture presented at Hotchips 2020 conference [1]

- (Quad-) Chiplet-based architecture
- AI/HPC focused
- Essential components have been manufactured in GF22
- Measured for energy-efficiency
- Extrapolation on larger AI workloads (full training and inference steps)

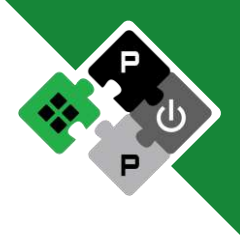


GlobalFoundries™ AVERY design systems Rambus Micron SYNOPSYS®

HOT CHIPS

[1] F. Zaruba et al., "Manticore: A 4096-Core RISC-V Chiplet Architecture for Ultraefficient Floating-Point Computing," in IEEE Micro, vol. 41, no. 2, pp. 36-42, 1 March-April 2021, doi: 10.1109/MM.2020.3045564.

Not All Programs Are Created Equal



- Processors can do two kinds of useful work:

Decide (jump to different program part)

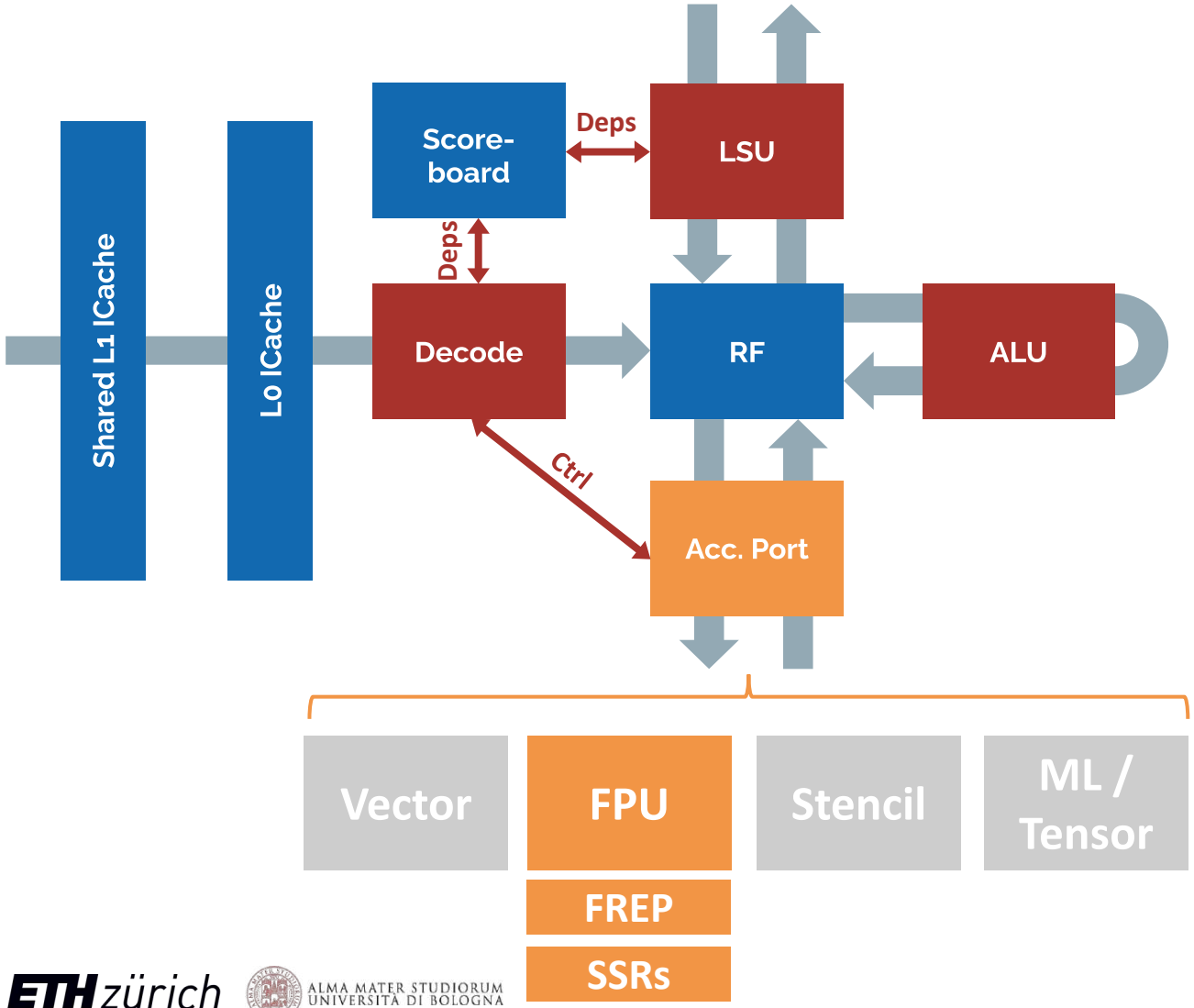
- Modulate flow of instructions
- Smarts:
  - Don't work too much
  - Be clever about the battles you pick (e.g., search in a database)
- Lots of decisions
  Little number crunching

Compute (plough through numbers)

- Modulate flow of data
- Diligence:
  - Don't think too much
  - Just plough through the data (e.g., machine learning)
- Few decisions
  Lots of number crunching

- Many of today's challenges are of the diligence kind:
  - Tons of data, algorithm ploughs through, few decisions done based on the computed values
  - "Data-Oblivious Algorithms" (ML, or better DNNs are so!)
  - Large data footprint + sparsity

Snitch – a Tiny 32b Integer RISC-V Core

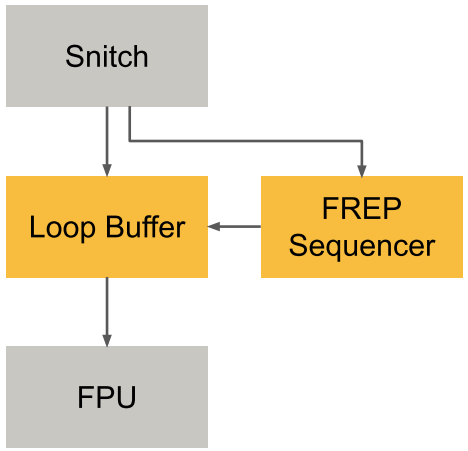


- **Simplest core: around 20KGE**
  - Speed via simplicity (1GHZ+)
  - L0 Icache/buffer for low energy fetch
  - Shared L1 for instruction reuse (SPMD)
- **Extensible → "Accelerator" port**
  - Minimal baseline ISA (RISC-V)
  - Performance through ISA extensions (via accelerator port)
- **Latency-tolerant → Scoreboard**
  - Tracks instruction dependencies
  - Much simpler than OOO support!

Custom ISA extensions

FREP: Remove control flow overhead [2]

- Programmable **micro-loop buffer**
- **Sequencer** steps through the buffer, independently of the FPU
- Integer **core free to operate in parallel**: **Pseudo-dual issue**




```
  mv     r0, zero
loop:
  addi   r0, 1
  fmadd  r2, ssr0, ssr1
  bne    r0, r1, loop
```

↓

```
  frep   r1, 1
loop:
  fmadd  r2, ssr0, ssr1
```

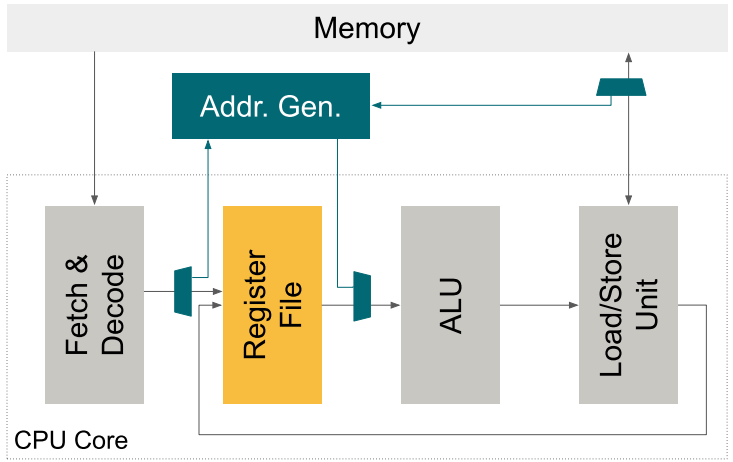
SSRs: Turn Reg R/W into Mem LD/ST [3]

- **Address generation** hardware to Register file
- **SSRs ≠ memory operands**
  - Perfect **prefetching**, latency-tolerant

```
loop:
fld  r0, %[a]
fld  r1, %[b]
fmadd r2, r0, r1
```

↓

```
scfg 0, %[a], ldA
scfg 1, %[b], ldB
loop:
fmadd r2, ssr0, ssr1
```



[2] 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, vol. 70, no. 11, pp. 1845-1860, 1 Nov. 2021, doi: 10.1109/TC.2020.3027900.

[3] F. Schuiki et al., "Stream Semantic Registers: A Lightweight RISC-V ISA Extension Achieving Full Compute Utilization in Single-Issue Cores," in *IEEE Transactions on Computers*, vol. 70, no. 2, pp. 212-227, 1 Feb. 2021, doi: 10.1109/TC.2020.2987314.

Architectural Innovation in Occamy

- **FPU** with **SIMD Mini-float** (ML training, Transformers) and **expanding SDOTP Unit [4]**

Transprecision FPU by Luca B. on Tuesday at 2pm

- **Sparsity support [5]** (Stencils, Sparse Tensors)
  - Extend 2 out of 3 SSRs to **ISSRs** and add index comparison unit
  - Forward result indices to 3rd SSR

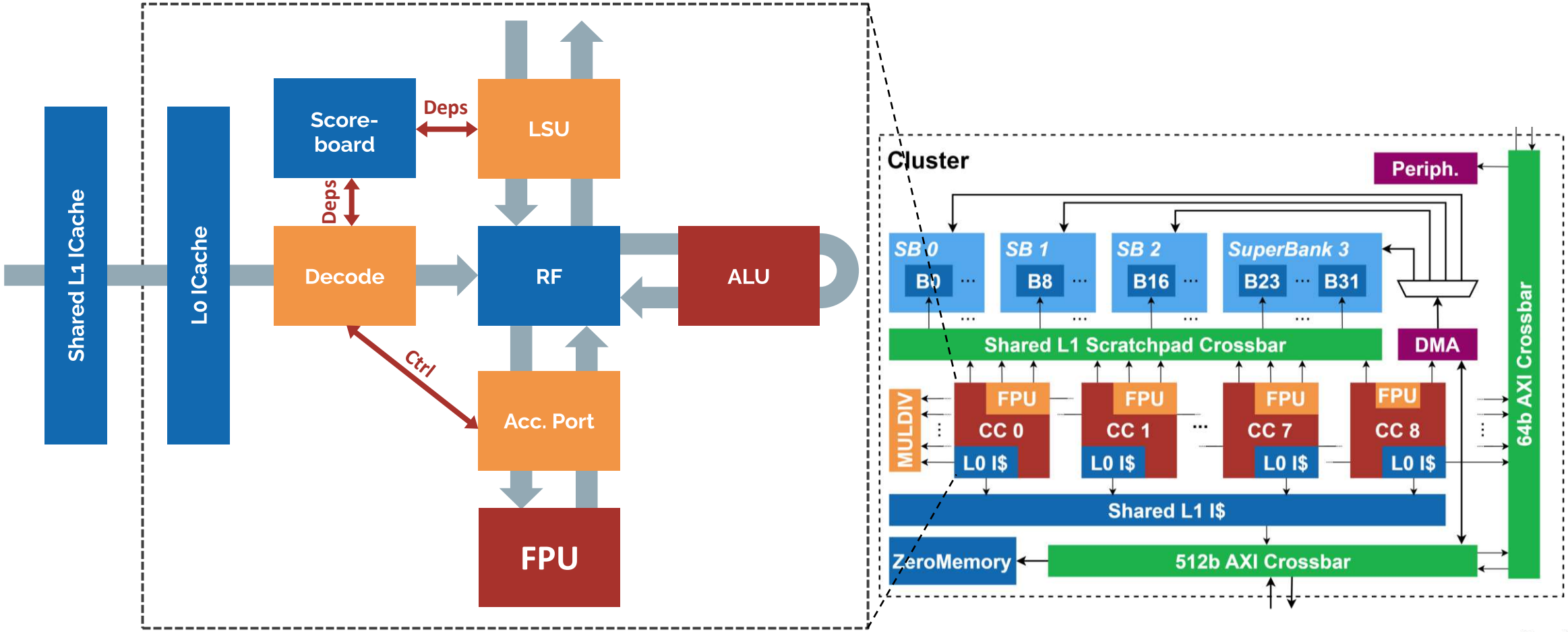
Sparsity Extensions by Paul on Tuesday at 9am

- **Atomics and fast interrupts** (synchro & offload accel.)
- **I-Cache** hierarchy

[4] L. Bertaccini et al., "MiniFloat-NN and ExSdotp: An ISA Extension and a Modular Open Hardware Unit for Low-Precision Training on RISC-V Cores," *2022 IEEE 29th Symposium on Computer Arithmetic (ARITH)*, Lyon, France, 2022, pp. 1-8, doi: 10.1109/ARITH54963.2022.00010.

[5] P. Scheffler, F. Zaruba, F. Schuiki, T. Hoefler and L. Benini, "Sparse Stream Semantic Registers: A Lightweight ISA Extension Accelerating General Sparse Linear Algebra," 2023, arXiv: 2305.05559

Occamy Cluster – 8 MGE



Occamy Cluster – 8 MGE

8 Snitch compute cores

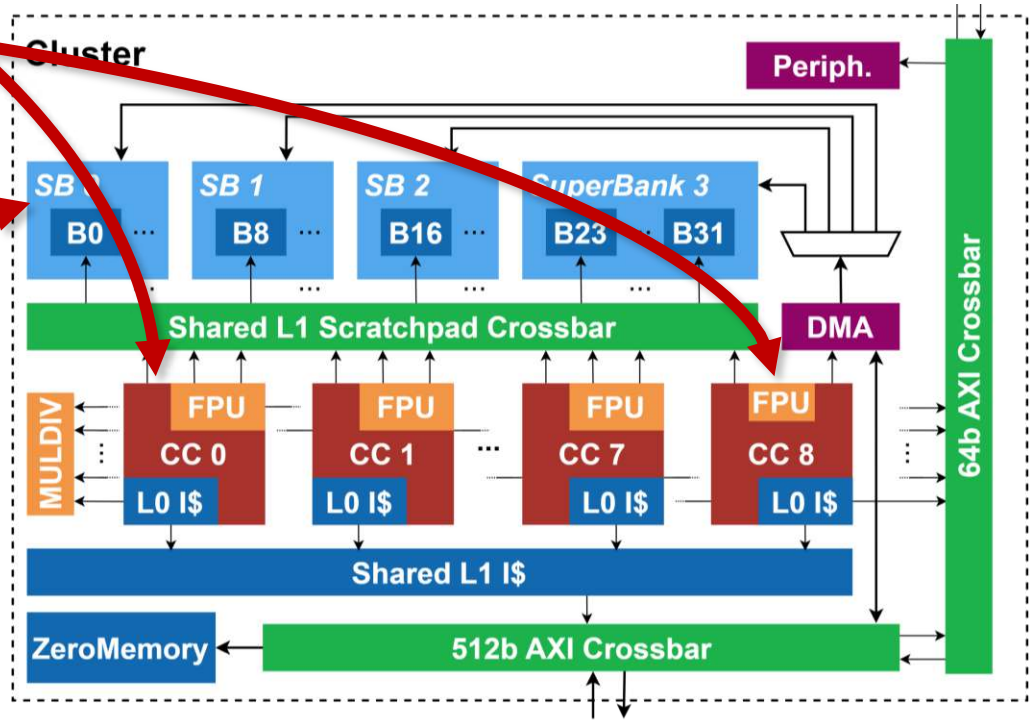
- Single-stage, small Integer control core

9th Core: DMA

- 512 bit data interface
- HW support to autonomously copy 2D shapes
- Higher-dimensionality can be handled by SW

128 kB TCDM

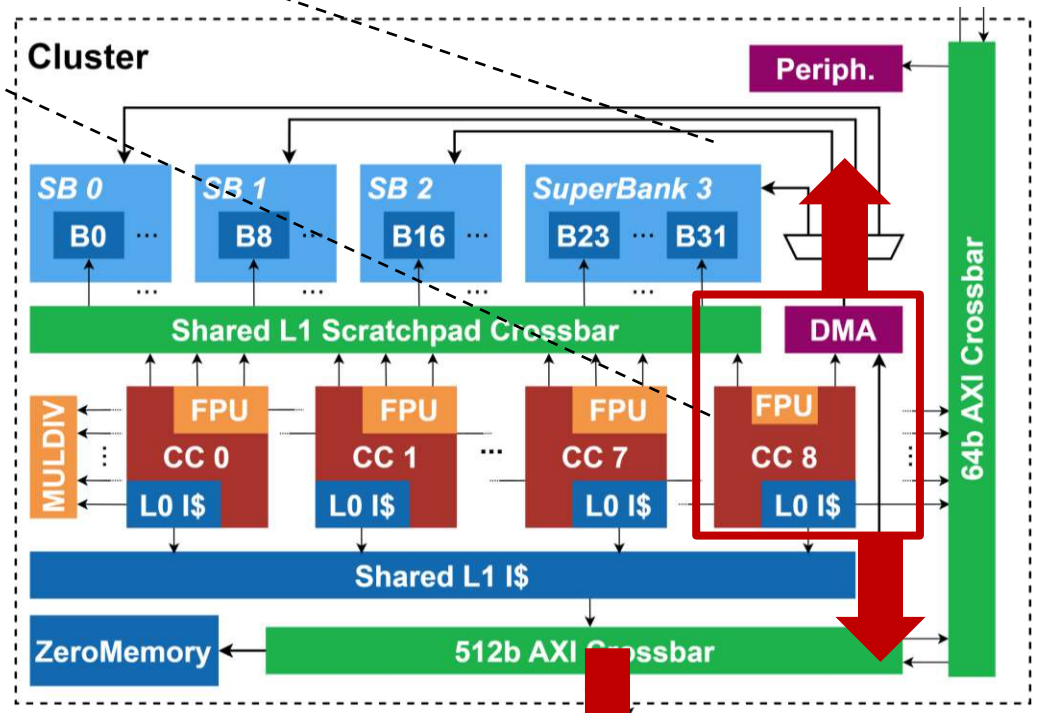
- Scratchpad for predictable memory accesses
- 32 Banks



Efficiently Move Data

- 64-bit AXI DMA
- Operates on **wide 512-bit data-bus**
- **Hardware support** to autonomously copy **2D** shapes
- Higher-dimensionality can be handled by SW

- Intrinsics/library for easy programming
- Exploiting cluster local memory



Data Movers by Thomas on Tuesday at 3.30pm

64GB/s @1GHz per cluster

Occamy Cluster – 8 MGE

8 Snitch compute cores

- Single-stage, small Integer control core

9th Core: DMA

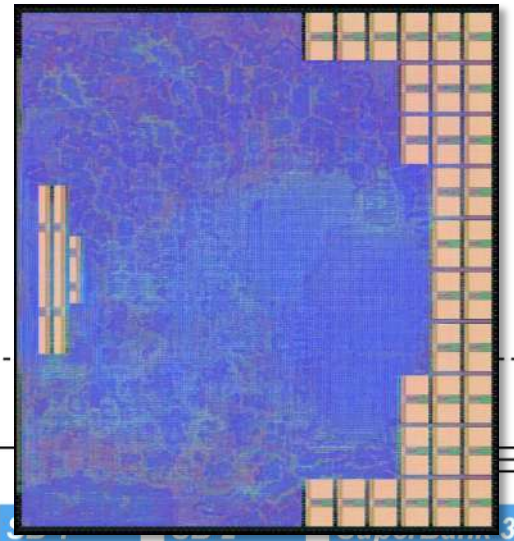
- 512 bit data interface
- HW support to autonomously copy 2D shapes
- Higher-dimensionality can be handled by SW

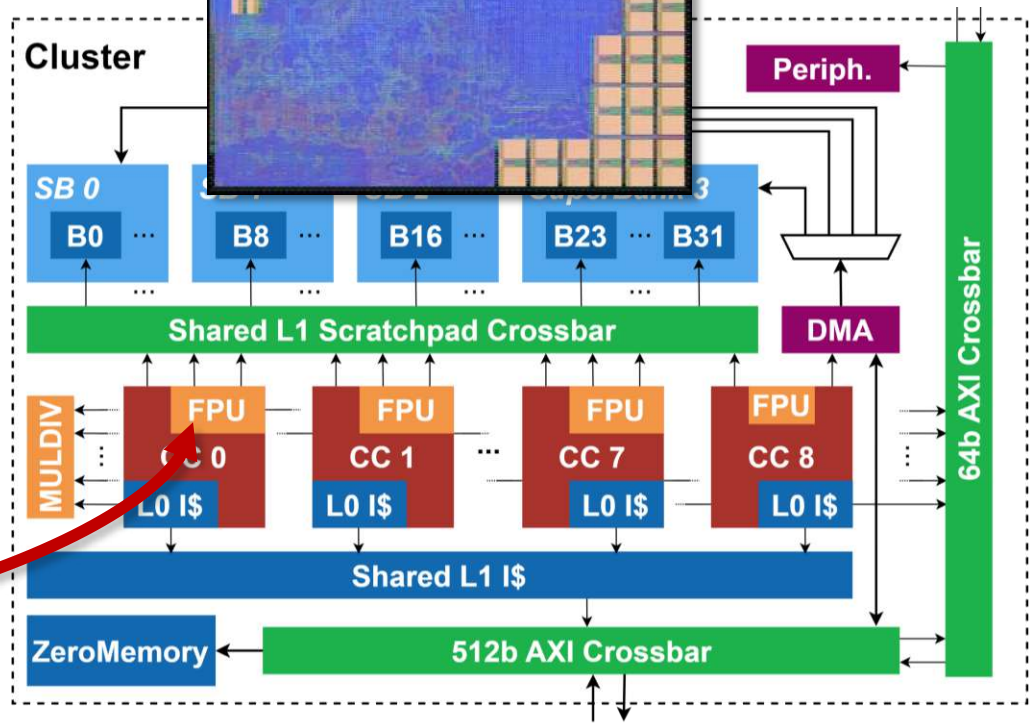
128 kB TCDM

- Scratchpad for predictable memory accesses
- 32 Banks

Custom ISA extensions

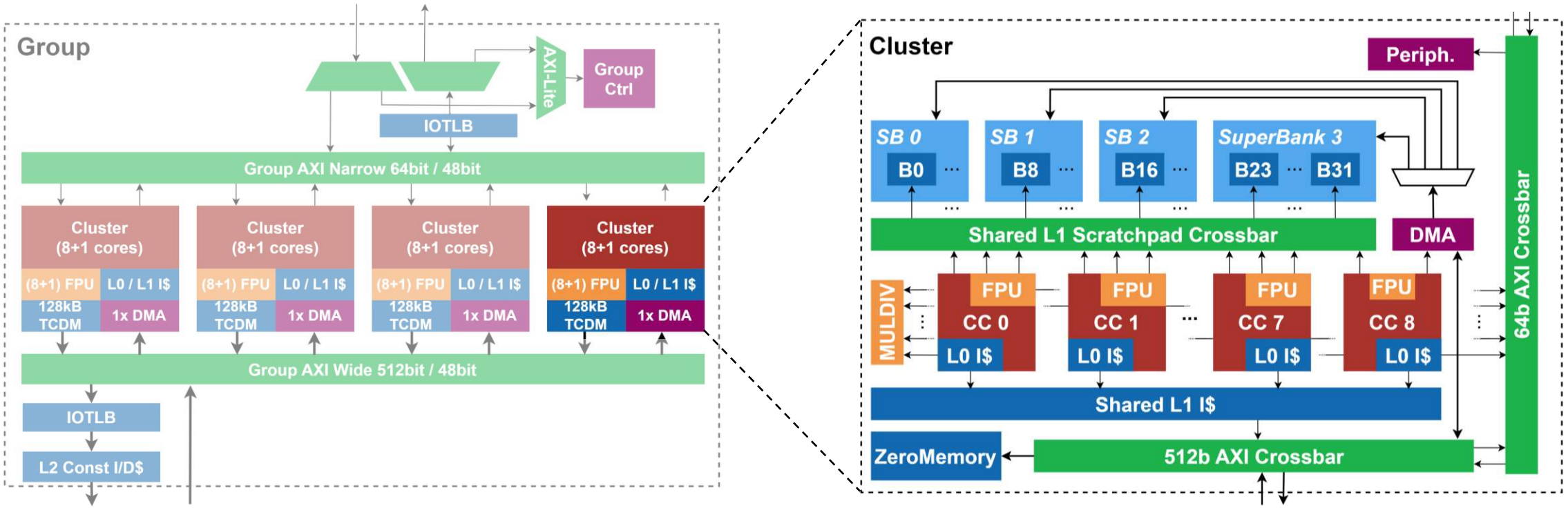
- Xfrep, Xssr
- New: Xissr sparsity support



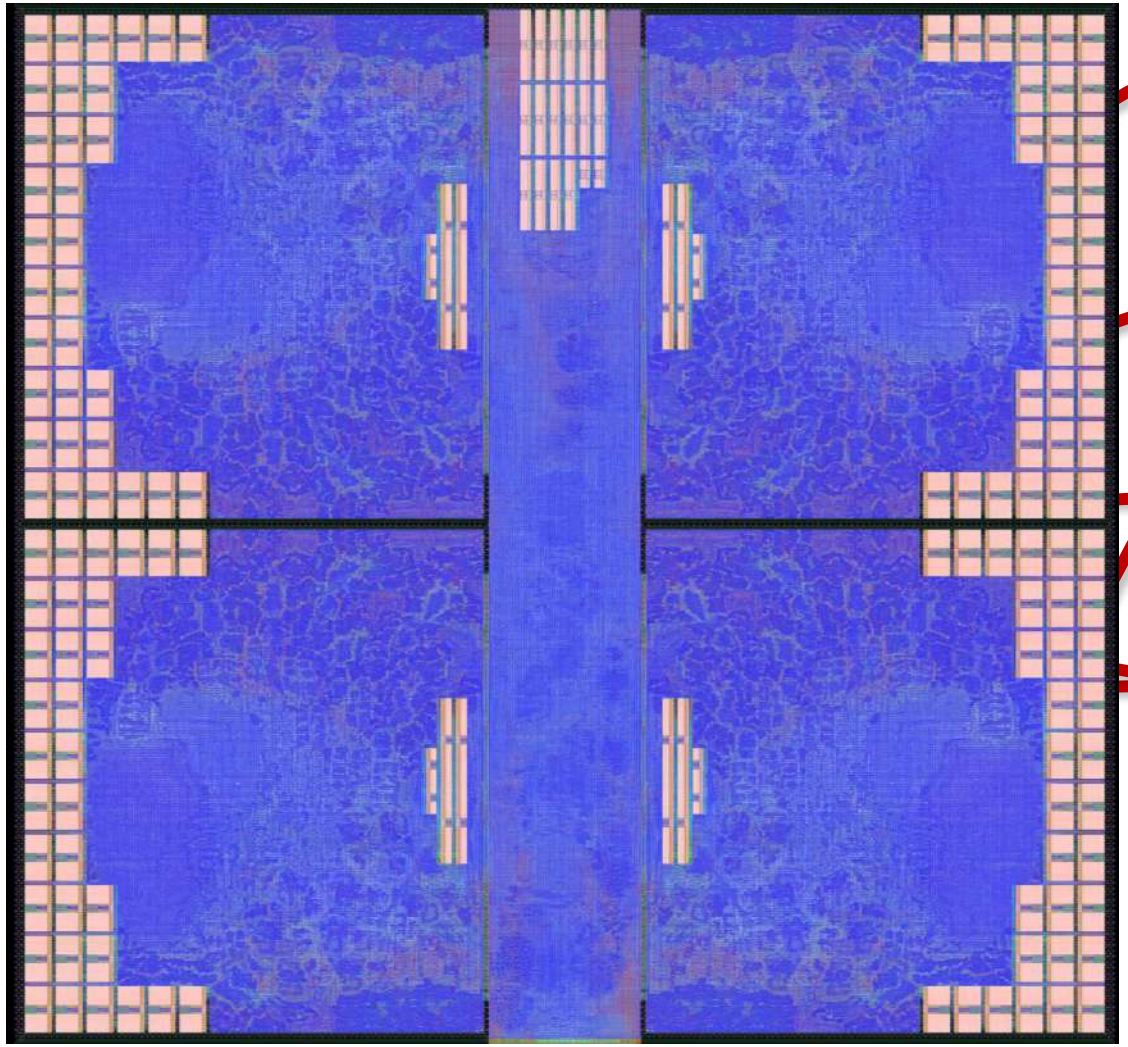


GEMM ≥ 80% FPU util.
Conv2d ≥ 75% PFU util.
Stencils ≤ 60% FPU util.
Sparse Tensors ≤ 50% FPU util.

Occamy Cluster – 8 MGE



Four Snitch Clusters form a Group



4 Clusters per Group

- Single-stage, small Integer control core

2 AXI Busses

- 64-bit narrow interface: config
- 512-bit wide interface: DMA

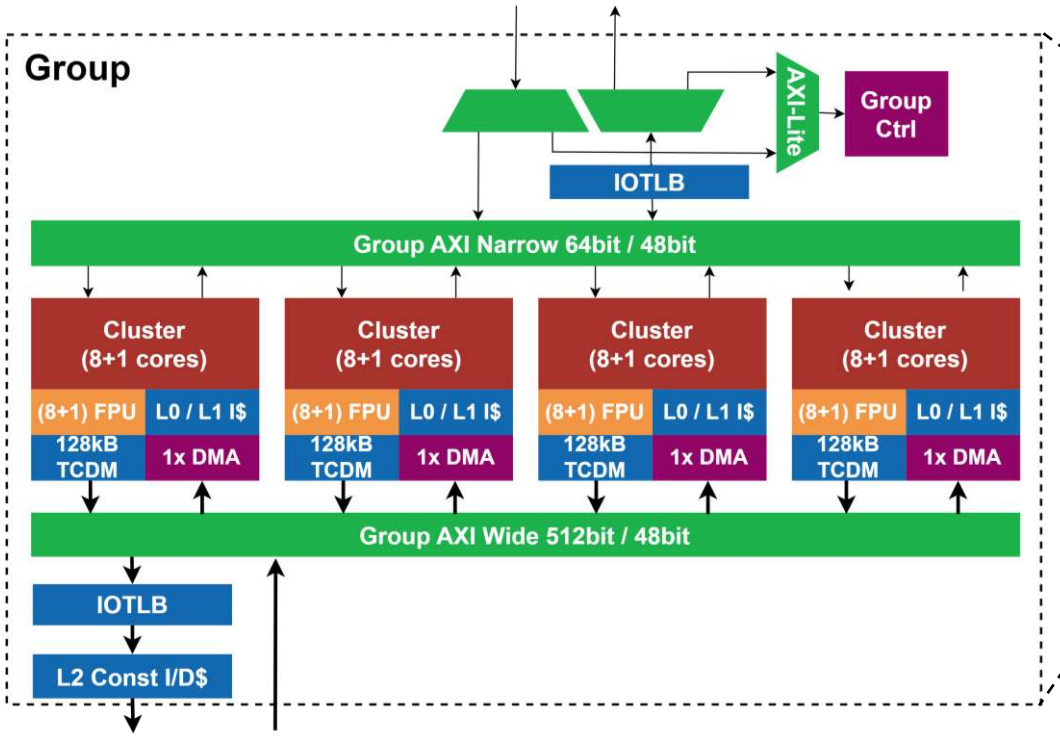
Constant Cache

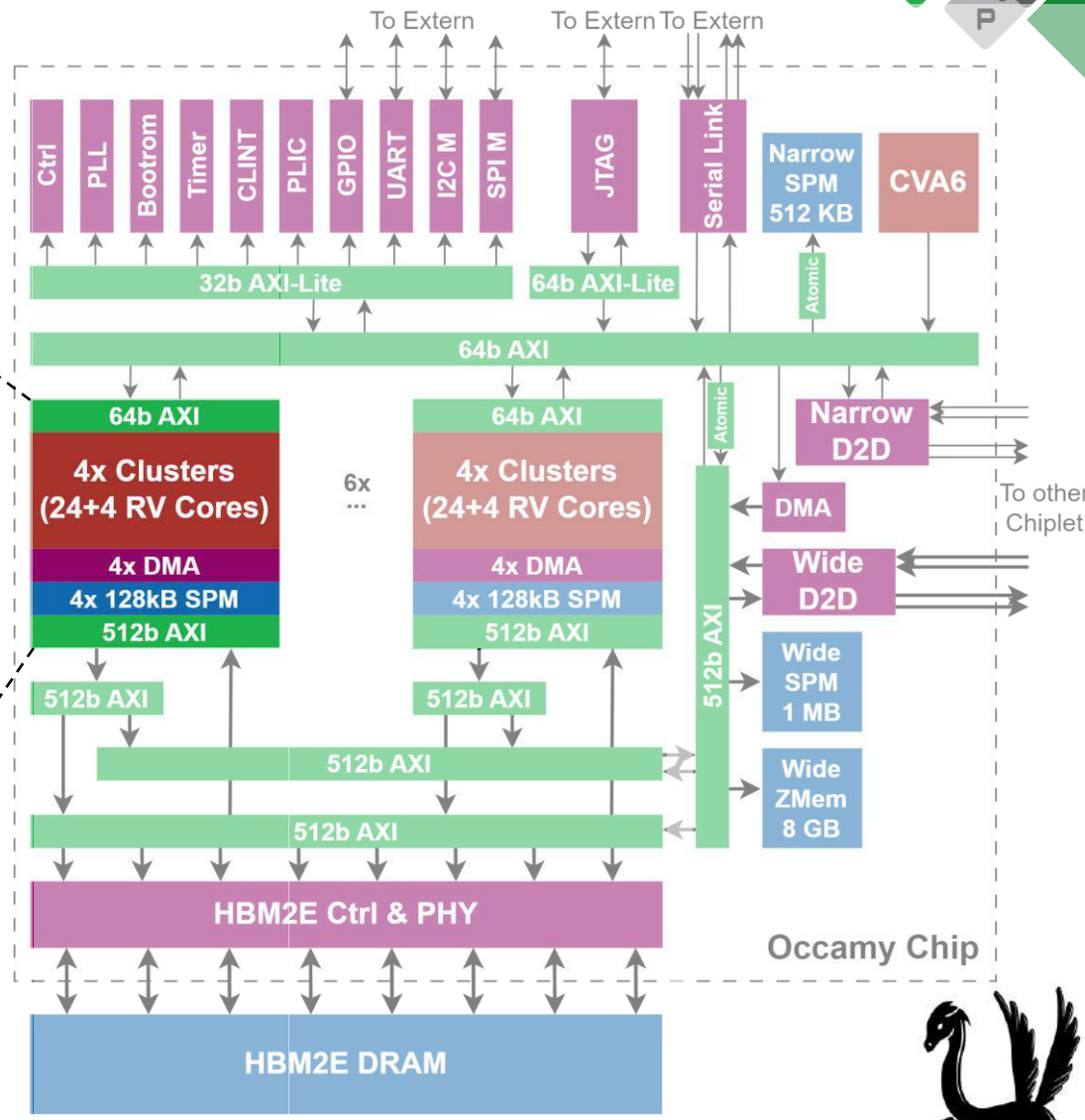
- D/I-Cache hierarchy

Translation Lookaside Buffers (TLBs)

- Virtual Addressing
- 8 PTEs for each (narrow and wide)
- enables:
  - Core access range extension
  - Per-group page remapping
  - Per-group access control

Four Snitch Clusters form a Group





Four Snitch Clusters form a Group

6 Groups: 216 cores/die
- 4 cluster / group:
  - 8 compute cores / cluster
  - 1 DMA core / cluster
- 512bit Constant Cache

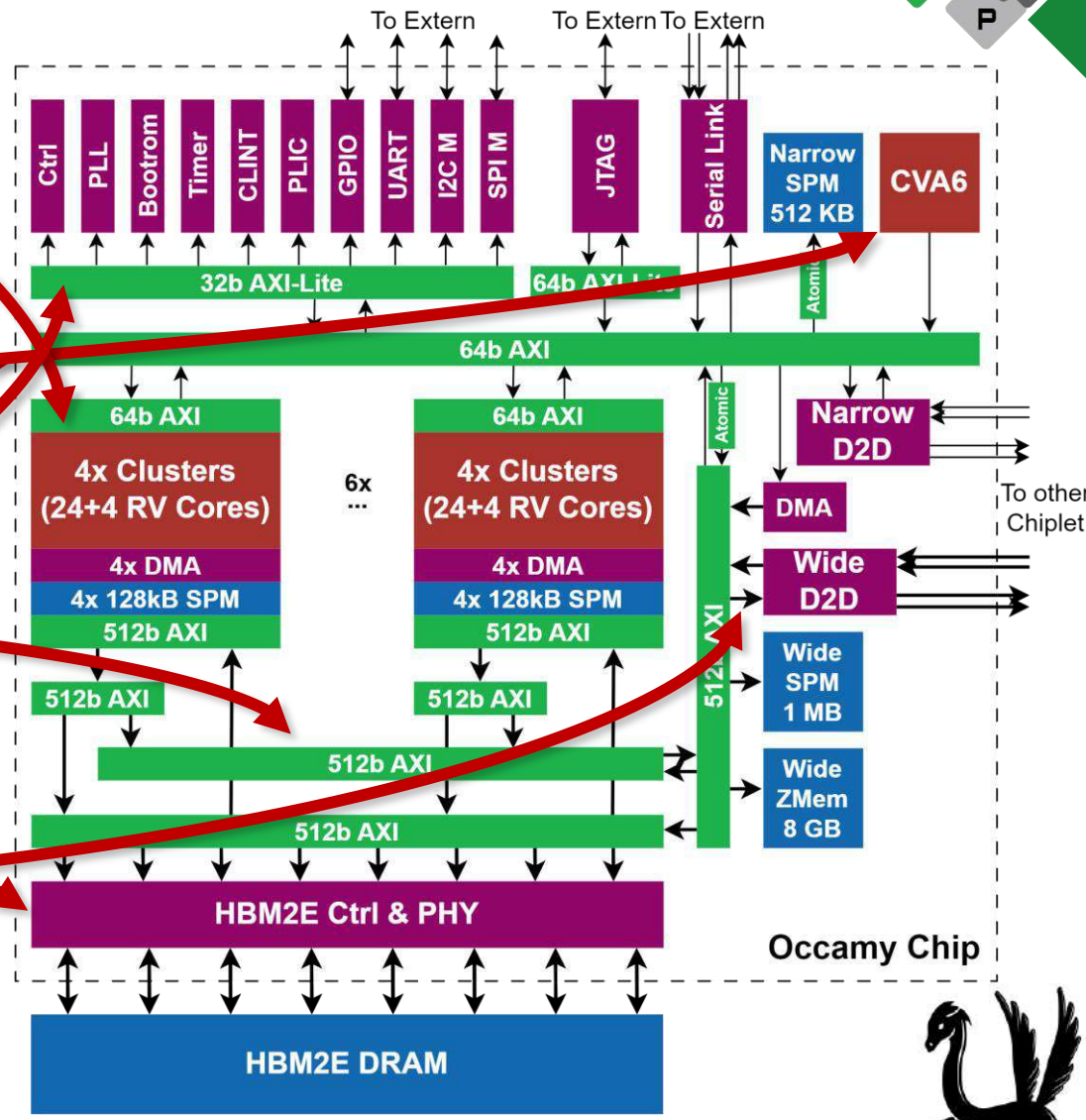
Linux-capable manager core CVA6

2 AXI Busses
- 64-bit narrow interface: config
- 512-bit wide interface: DMA

Peripherals
- SPI, JTAG, I2C, UART, serial link

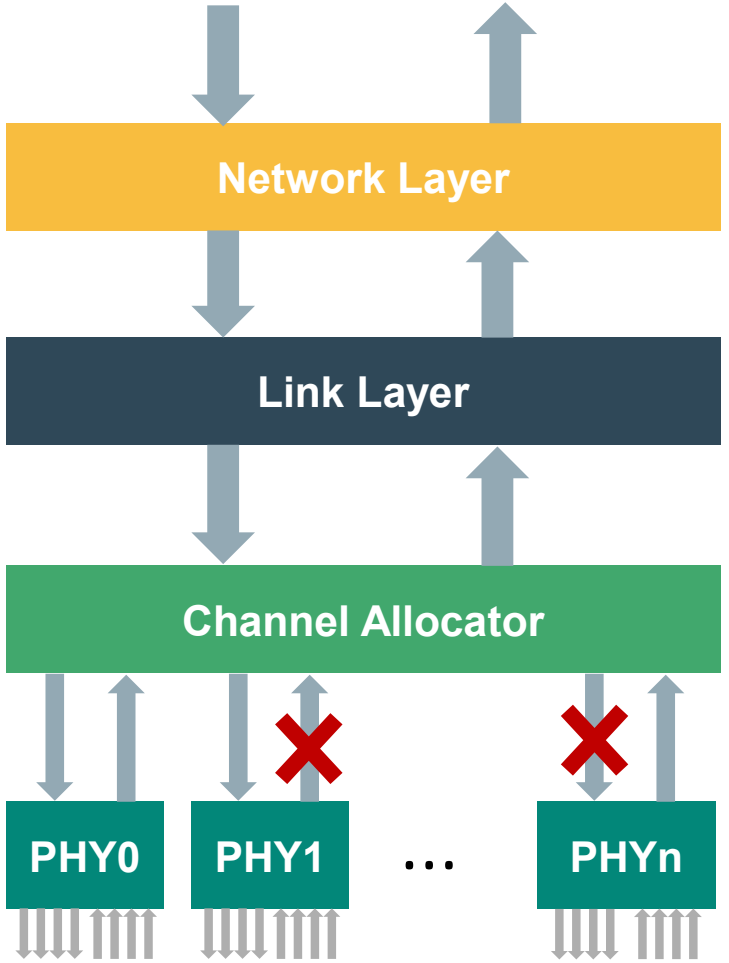
8-channel **HBM2e** (16GB)

Serial DDR D2D link
- Source synchronous
- technology-independent
- Moderate speed: ≤ 125MHz



We designed our own custom Die-to-Die link

- Availability of HBI IP unfortunately not aligned with project timeline
- **Custom solution:**
  - **Source synchronous**
  - Bunch-of-wires (BOW) style
  - **DDR**
  - **technology-independent**
    - Standard digital pads
  - Moderate speed: ≤ 125MHz
  - Front-end + controller usable for faster PHYs



Network Layer:

- Full AXI4 interface
- AXI4 to AXI stream converter

Data Link Layer

- Credit-based flow control
- RX synchronisation

Channel Allocator

- Chops and reshuffles payload
- **Fault tolerance** mechanisms

Physical Layer

- Source-synchronous DDR sampling
- **Scalable** number of channels
  (38 x 8-bit full-duplex DDR channels in Occamy)

https://github.com/pulp-platform/serial_link

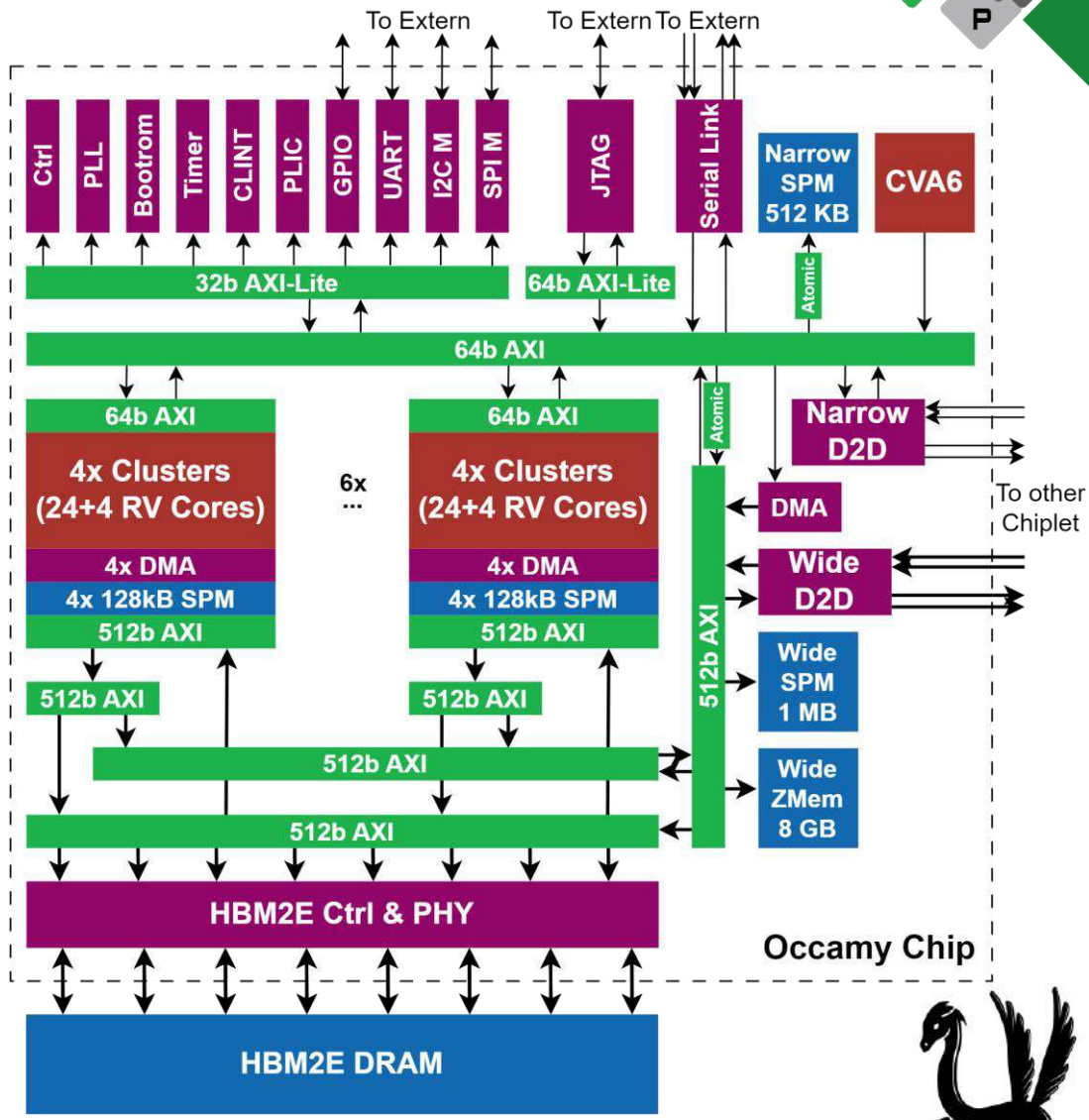
Main Compute architecture is open-source !!!

The main compute architecture is being developed fully open-source !!!

github.com/pulp-platform/snitch

github.com/pulp-platform/serial_link

HBM, DFG, FLL, and any proprietary components are in a separate private repository on our internal Gitlab



Main Compute architecture is open-source !!!

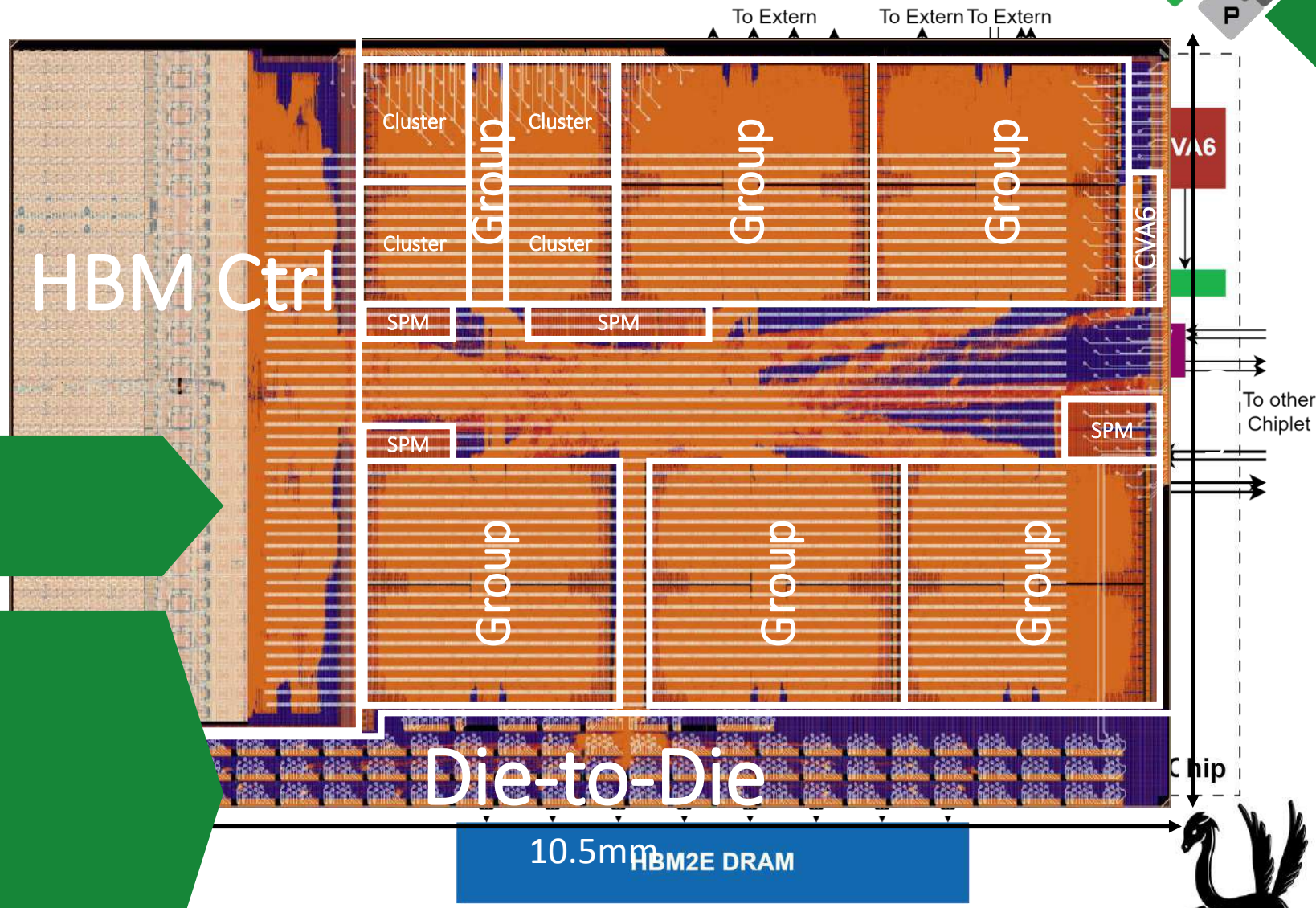
Chiplet Occamy:

- Technology: GF12LP+
- 1 GHz
- Area: 73mm^2
- **>1 billion** transistor

Taped out: 1st of July 2022

Peak Chiplet perf. @1GHz:

FP64: 384 GFLOp/s
FP32: 768 GFLOp/s
FP16: 1.536 TFLOp/s
FP8: 3.072 TFLOp/s



Balancing Bandwidth and Compute

@1GHz

Problem: HBM Accesses are not ideal in terms of

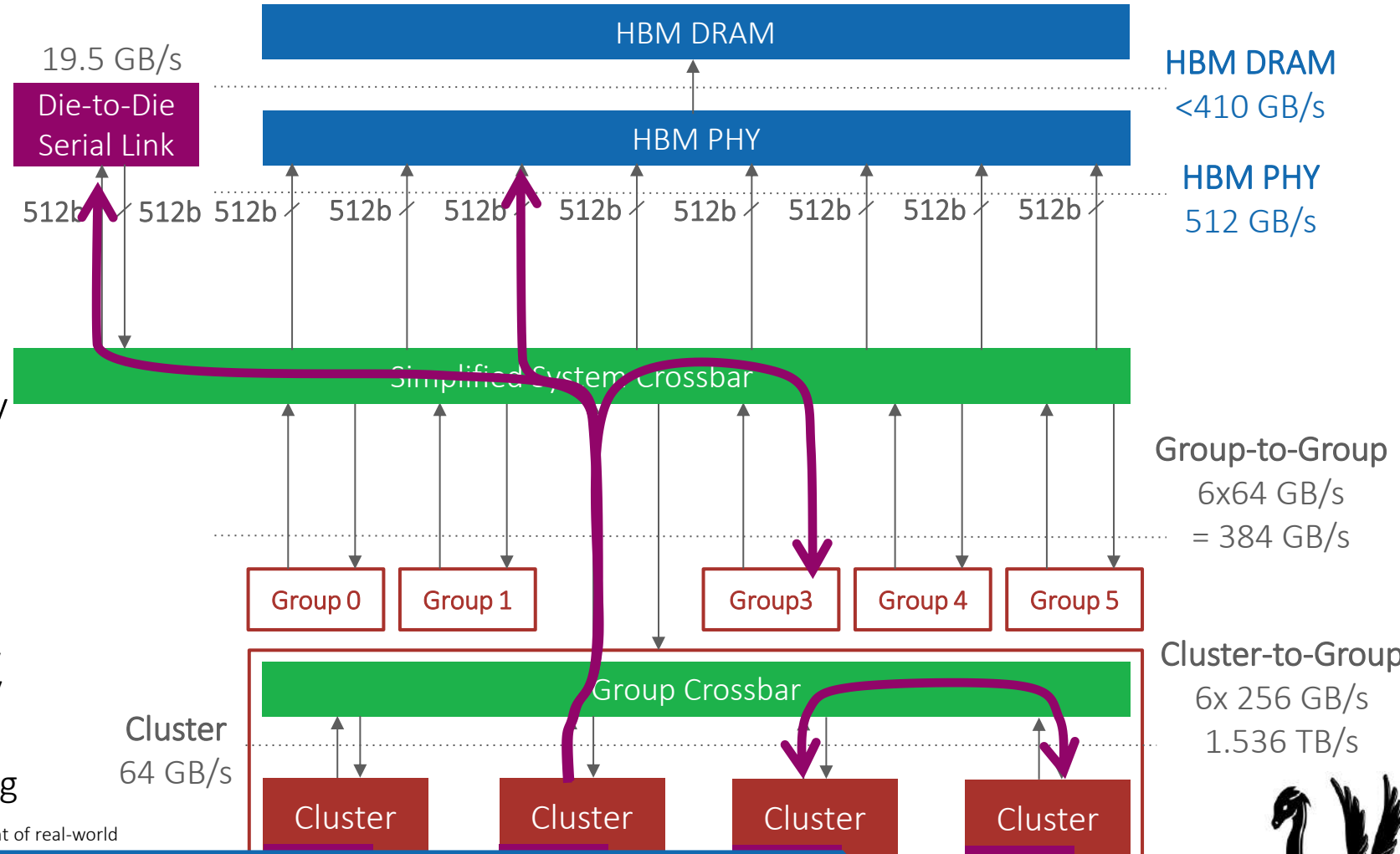
- Access energy
- Congestion
- High latency

Instead reuse data on lower levels of the memory hierarchy

- Between **clusters**
- Across **groups**

Smartly distribute workload

- **Clusters:** *DORY* [5] */ Deeploy* framework for deployment / tiling strategy
- **Chiplets:** E.g. Layer pipelining



[5] Burrello, Alessio, et al. "Dory: Automatic end-to-end deployment of real-world

NAS/Deeploy by Alessio, Mortiz on Tuesday at 11.30am

Programming Model

```
__builtin_ssr_setup_1d(0, repetition, bound, stride, data);
static volatile double d = 42.0;

__builtin_ssr_enable();

__builtin_ssr_push(0, d);
volatile double e;
e = __builtin_ssr_pop(0);
__builtin_ssr_disable();
}
```

Compilers by Giuseppe, Robert on Tuesday at 11am

- Multiple layers of abstraction:
  - …d assembly
  - …sics
    - **MLIR support for Snitch**
    - High-level frameworks:
      DaCE: spcl.inf.ethz.ch/Research/DAPP/
      Dory / Deeploy: deployment and tiling of NNs
- Bare-metal runtime
- Basic OpenMP runtime
- HERO toolchain

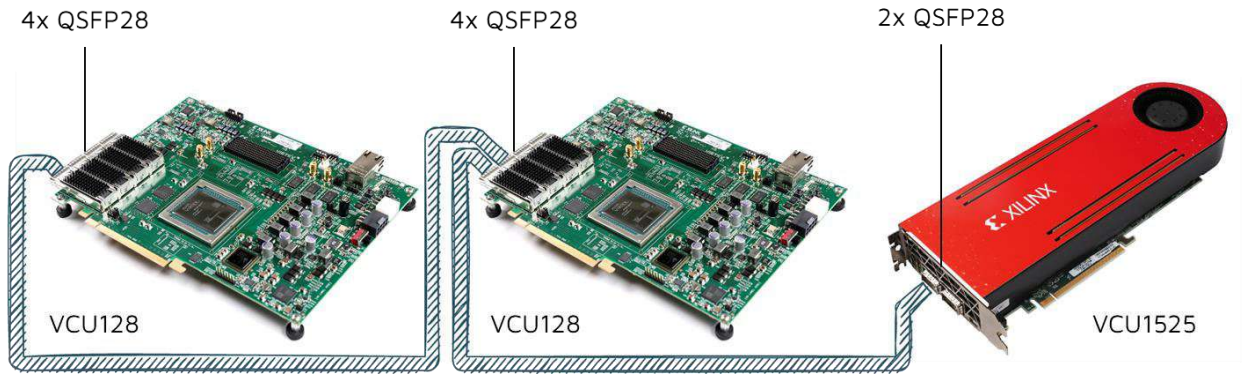
MLIR

OpenMP

Prototyping and Emulation

Occamy mapped onto **2x VCU128 (with HBM) + 1x VCU1525**

- 1x CVA6
- 2-4x Snitch clusters



Attaching a **Xilinx PCIe controller & PHY**

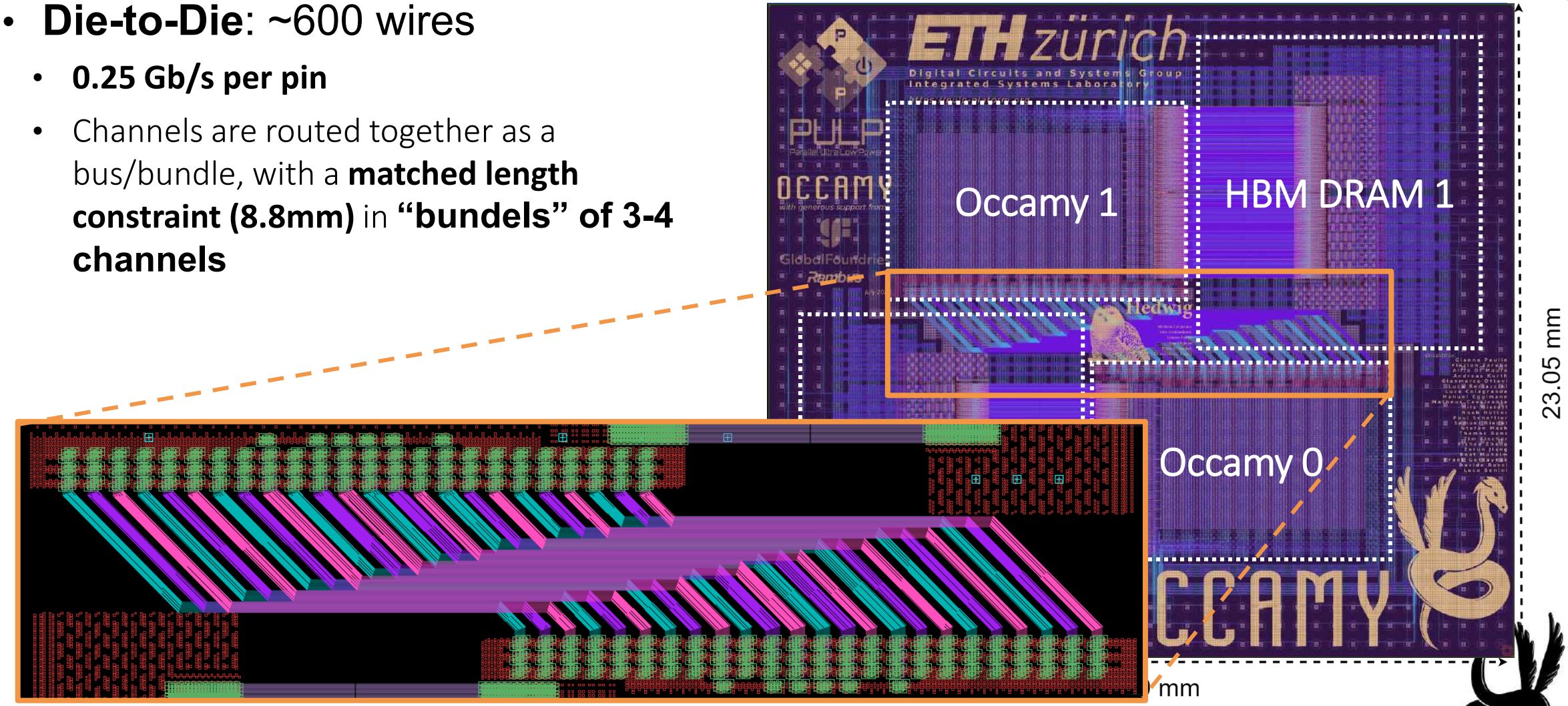
- **x86 Linux host** supported
- **RISC-V Linux host** (**Monte Cimone**) in progress

Supporting **hybrid usage** (High SW stack re-usability)

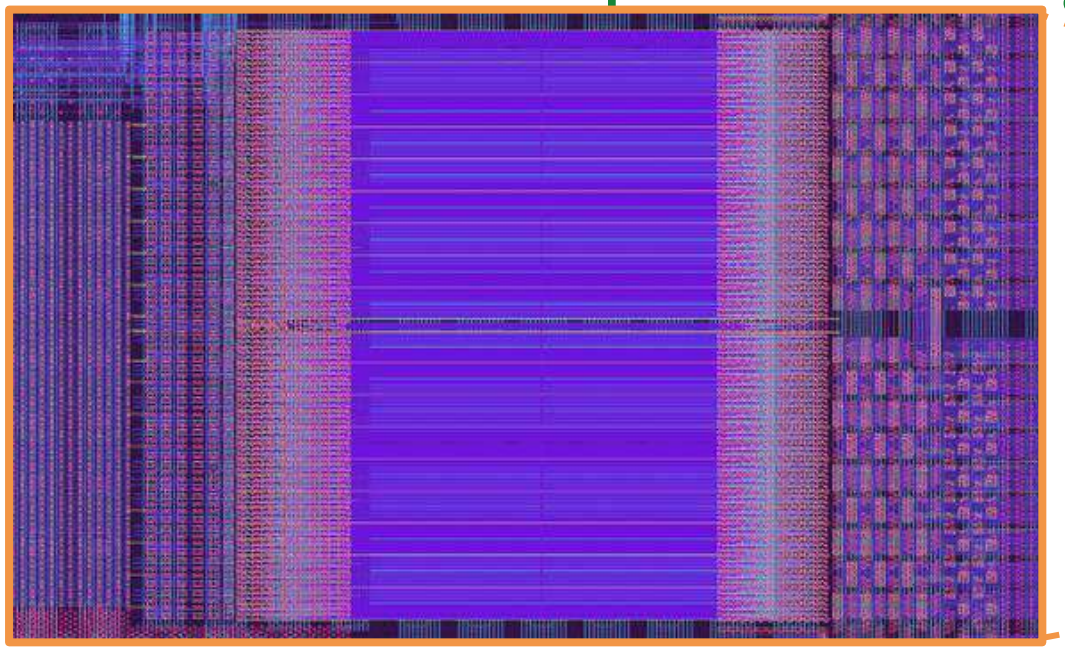
- Boot directly on standalone CVA6
- Do not boot and let the Host control the cluster

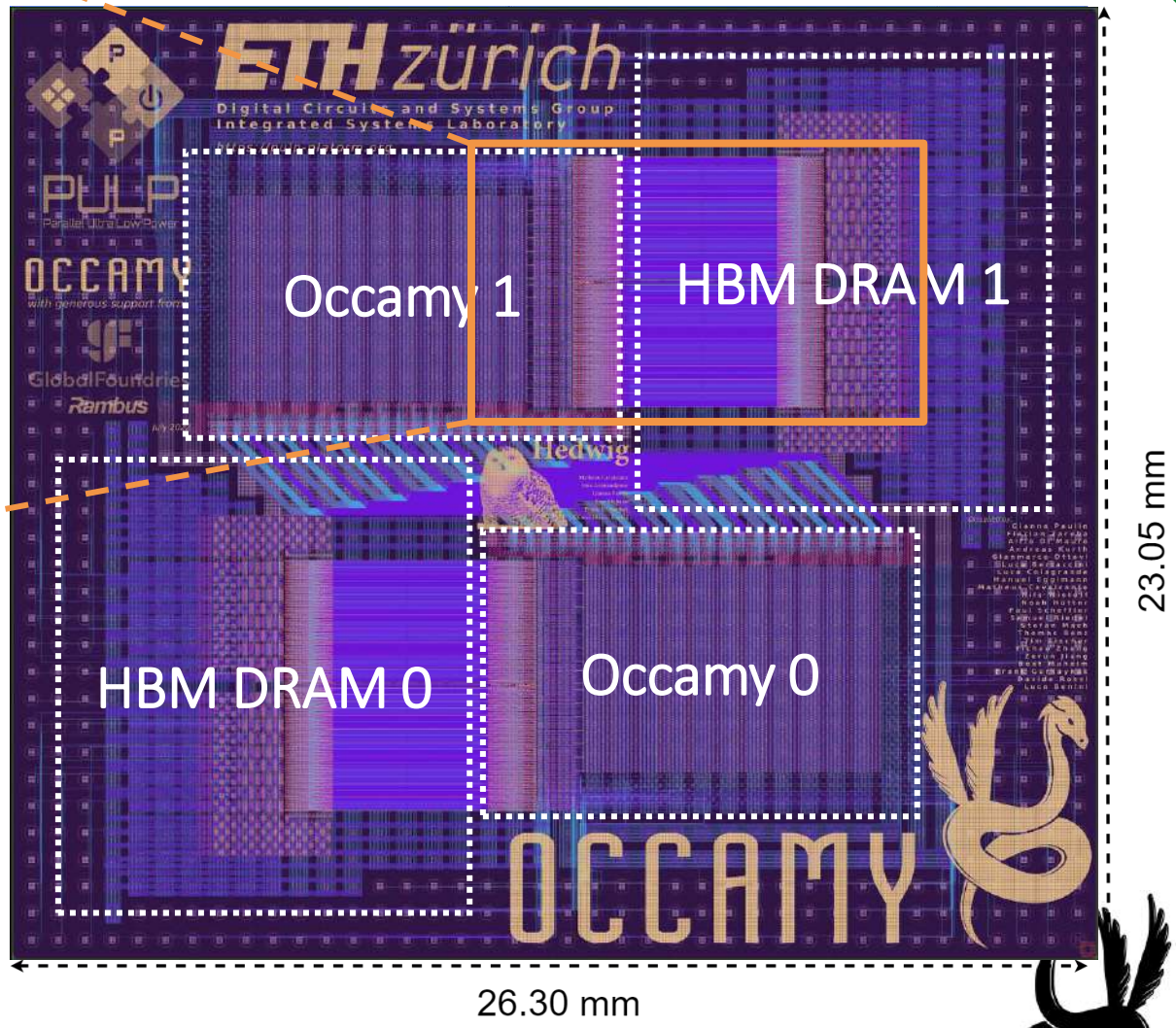
Our Silicon Interposer Hedwig (65nm, passive, GF)

- **Die-to-Die**: ~600 wires
  - **0.25 Gb/s per pin**
  - Channels are routed together as a bus/bundle, with a **matched length constraint (8.8mm)** in **“bundels” of 3-4 channels**



Our Silicon Interposer Hedwig (65nm, passive, GF)



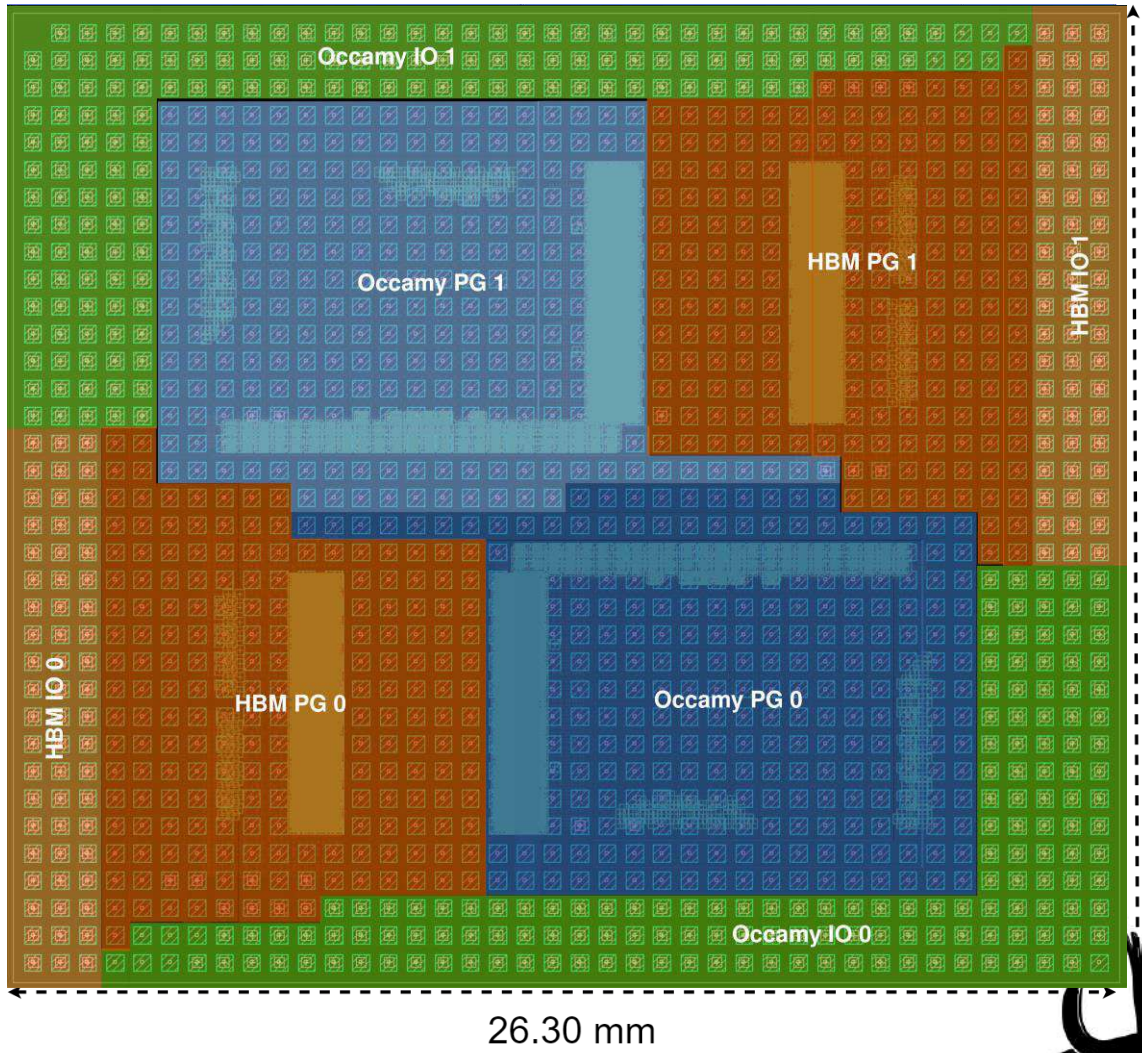


- **HBM**: ~1700 wires
  - **3.2 Gb/s per pin**
  - Nearly 100% track utilization
  - **Ground shielding** planes
  - HBM wire length is **4.9mm**

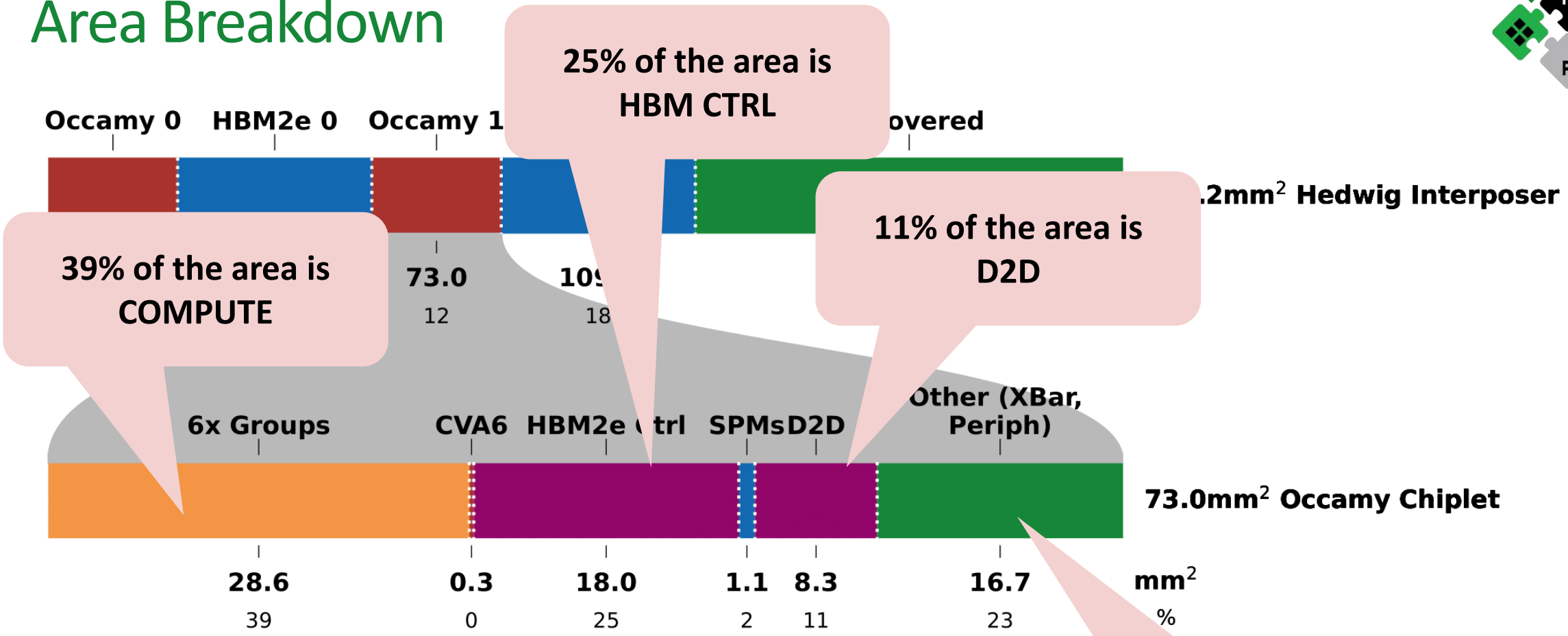
Our Silicon Interposer Hedwig (65nm, passive, GF)

- **Interlocked** die arrangement
  - Prevent bending, increase stability
- **Compact** die arrangement
  - No *dummy dies* or *stitching* needed
- **Fairly low I/O pin count** due to no high-bandwidth periphery
  - Off-package connectivity: ~200 wires
  - Array of **40 x 35 (-1) C4s** (total of 1'399 C4 bumps)
    - Diameter: 400µm, Pitch: 650µm

Taped out: 15th of October 2022



Area Breakdown

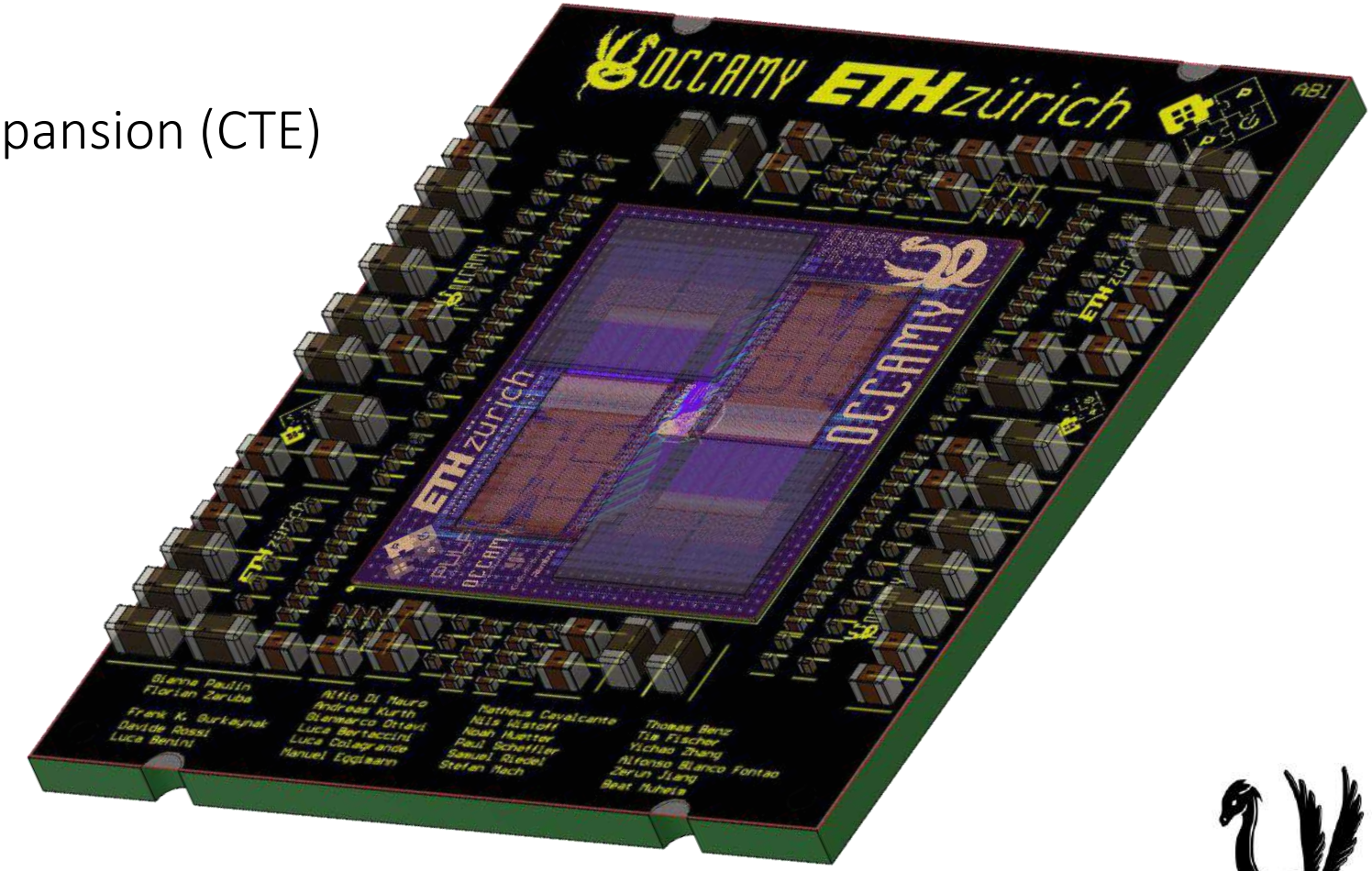


NoC by Tim on Tuesday at 3.30pm

Carrier PCB brings mainly “fan-out” for PCB mounting

Carrier PCB *(52.5 x 45mm)*

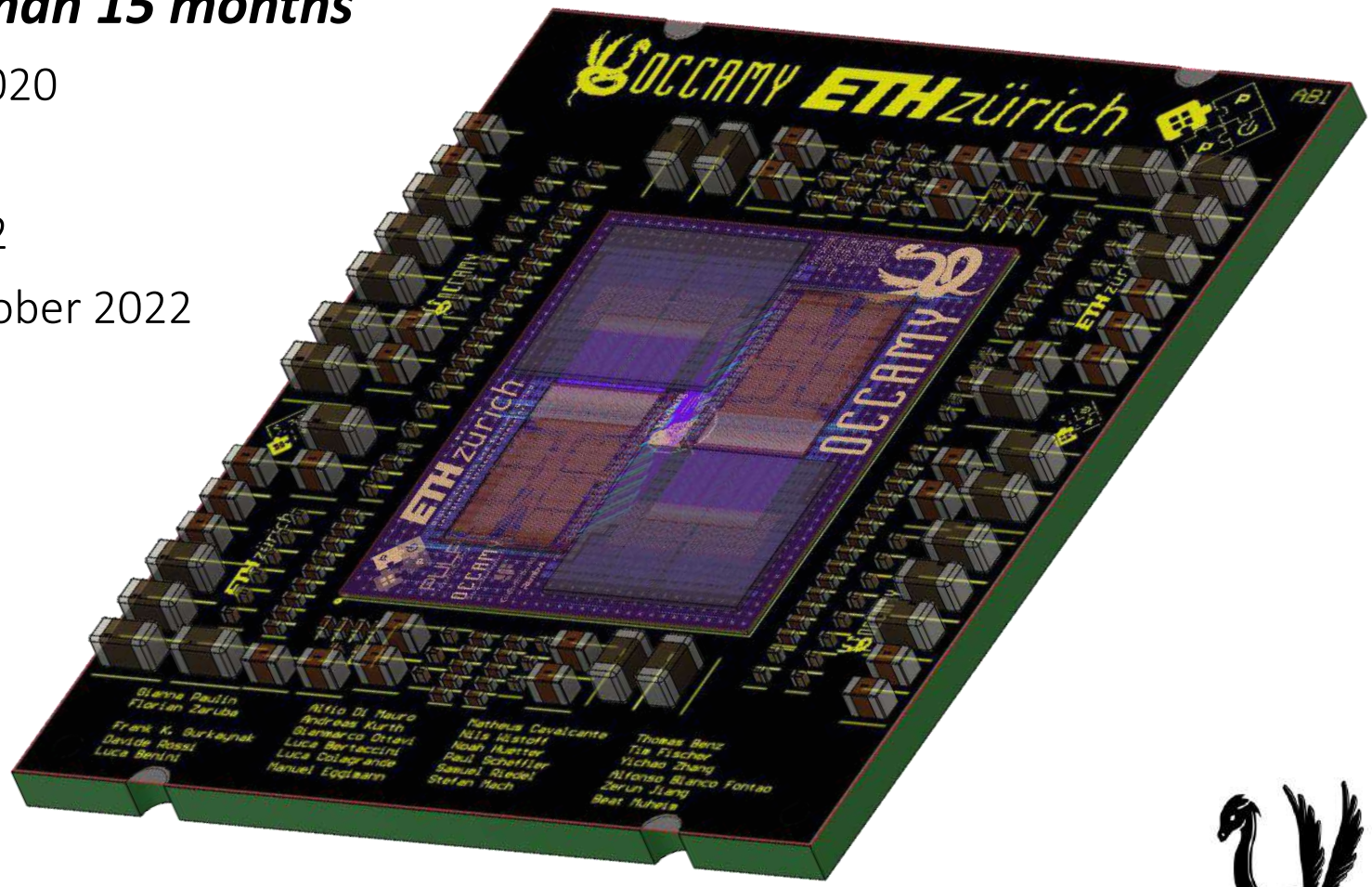
- Material Selection: RO4350B
  - low Coefficient of Thermal Expansion (CTE)
  - High stability
- Decoupling caps
- Custom ZIF socket design



Waiting for the Assembly to complete....

Finished Chiplet Tapeout in less than 15 months

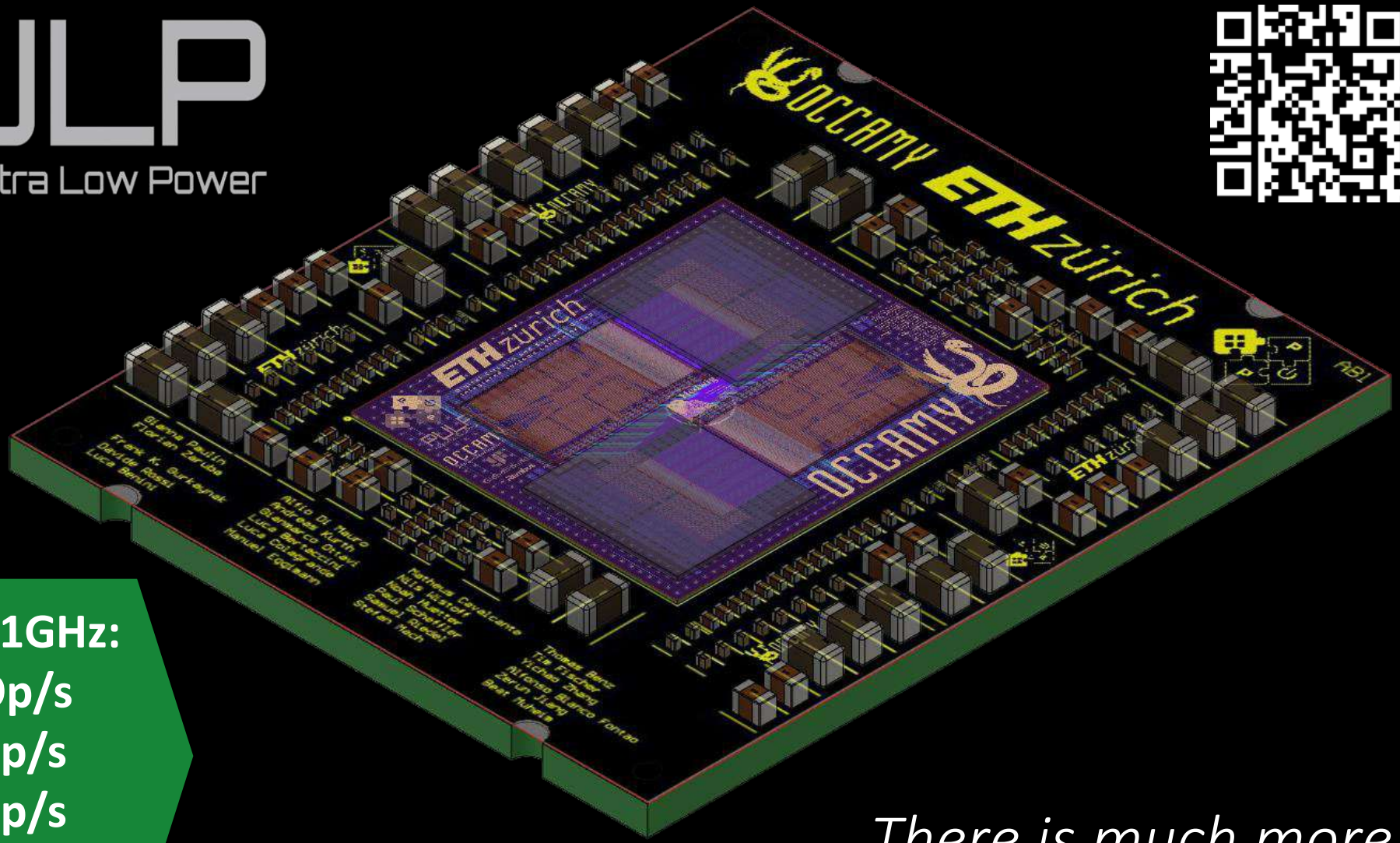
- Initial discussions 20th of October 2020
- Started on 20th of April 2021
- Taped out Chiplet on 1st of July 2022
- Taped out Interposer on 15th of October 2022
- **Currently being assembled**



There is much more to come in Q3-2023 …

PULP

Parallel Ultra Low Power



Peak System perf. @1GHz:

FP64:	**768 GFLOp/s**
FP32:	**1.536 TFLOp/s**
FP16:	**3.072 TFLOp/s**
FP8:	**6.144 TFLOp/s**

There is much more to come in Q3-2023 …

http://pulp-platform.org @pulp_platform