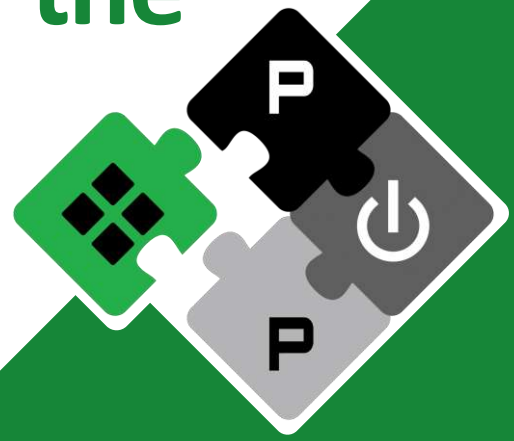
ETH zürich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Ten years of PULP: The Evolution of the Species from IoT to HPC



Davide Rossi davide.rossi@unibo.it

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

@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

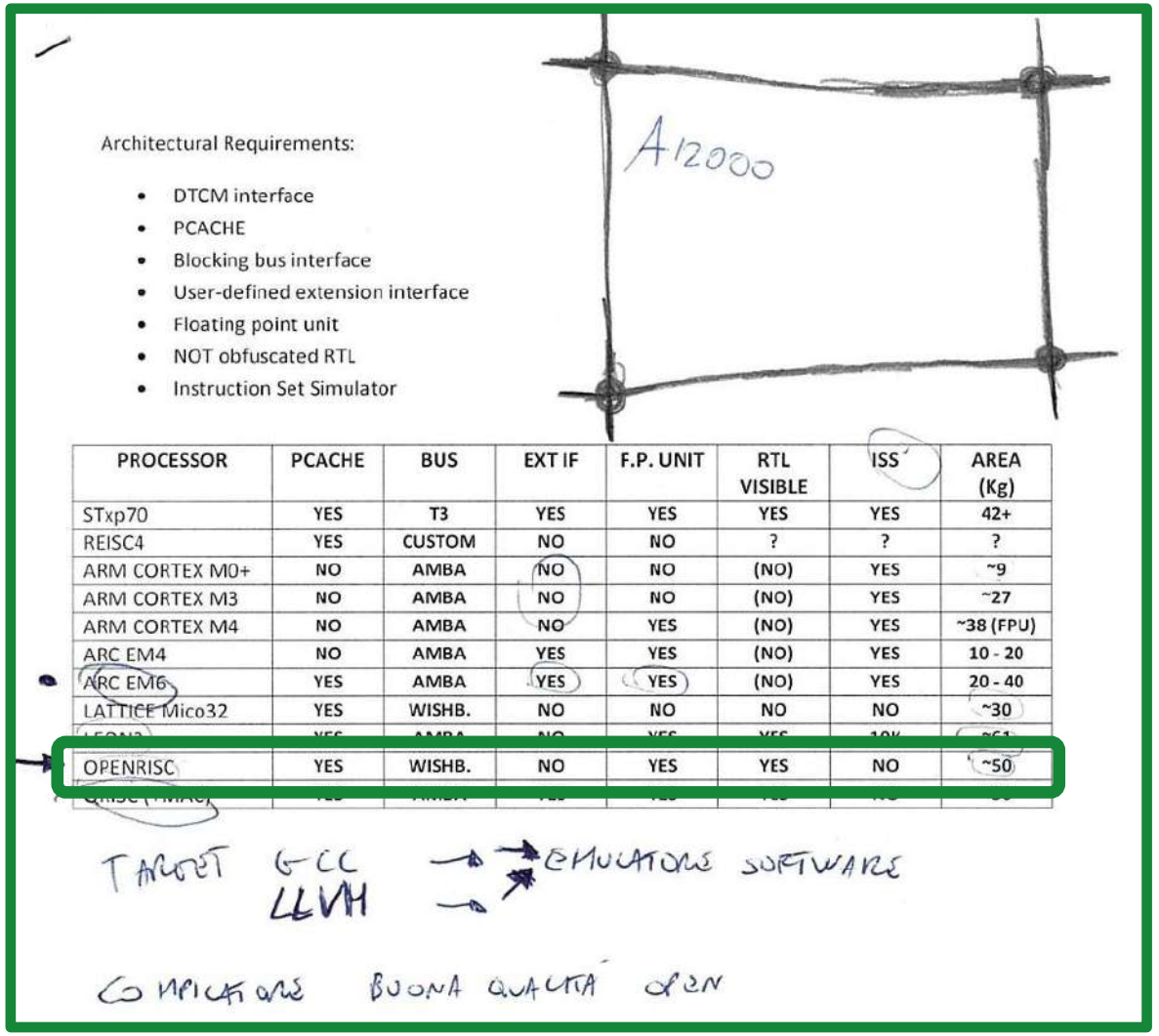
PULP: The Origins

- **Project started in 2013**
- **A collaboration between University of Bologna and ETH Zürich**
  - We wanted to start with a clean slate, no need to remain compatible to legacy systems, no dependencies with any commercial IP
- **Original Academic/Research goal**
  - Push energy efficiency of IoT computing systems as much as possible
  - Create a compute platform and an ecosystem used for research on Computer Architectures by our team as well as other groups in the World
- **Original Approach**
  - Exploit Open Source Ips as much as possible
  - “We’ll never do a processor”
  - “We’ll never do a compiler”
  - Team of 5-6 people at the beginning

Back to 2013 Open (And Close) Source Processors Landscape

- **Reasonably Good Quality IPs**
- **Area around 50 kGates**
- **Open Source**
- **Compiler Support**
- **Community support**

- **We decided to go with OR1200**



Architectural Requirements:

- DTCM interface
- PCACHE
- Blocking bus interface
- User-defined extension interface
- Floating point unit
- NOT obfuscated RTL
- Instruction Set Simulator

A12000

PROCESSOR	PCACHE	BUS	EXT IF	F.P. UNIT	RTL VISIBLE	ISS	AREA (Kg)
STxp70	YES	T3	YES	YES	YES	YES	42+
REISC4	YES	CUSTOM	NO	NO	?	?	?
ARM CORTEX M0+	NO	AMBA	NO	NO	(NO)	YES	~9
ARM CORTEX M3	NO	AMBA	NO	NO	(NO)	YES	~27
ARM CORTEX M4	NO	AMBA	NO	YES	(NO)	YES	~38 (FPU)
ARC EM4	NO	AMBA	YES	YES	(NO)	YES	10 - 20
ARC EM6	YES	AMBA	YES	YES	(NO)	YES	20 - 40
LATTICE Mico32	YES	WISHB.	NO	NO	NO	NO	~30
OPENRISC	YES	WISHB.	NO	YES	YES	NO	~50

TARGET GCC LLVM → EMULATORS SOFTWARE

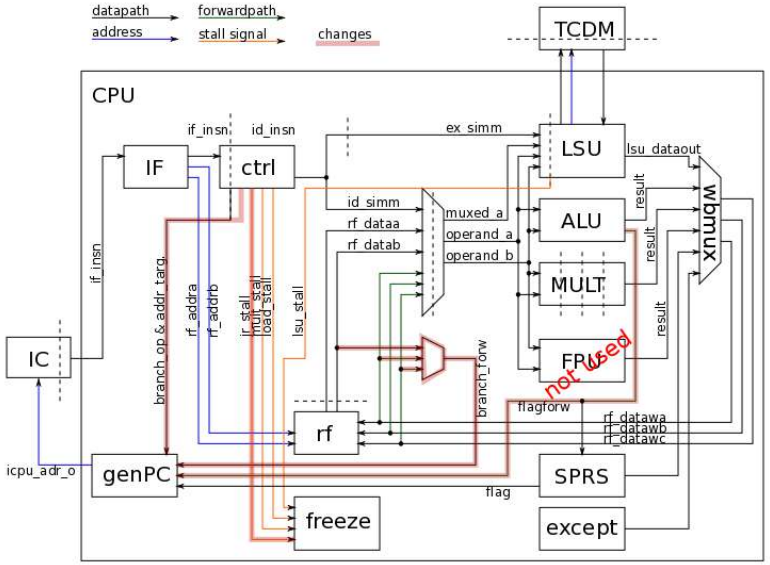
COMPLICATIONS BUONA QUALITÀ OPEN

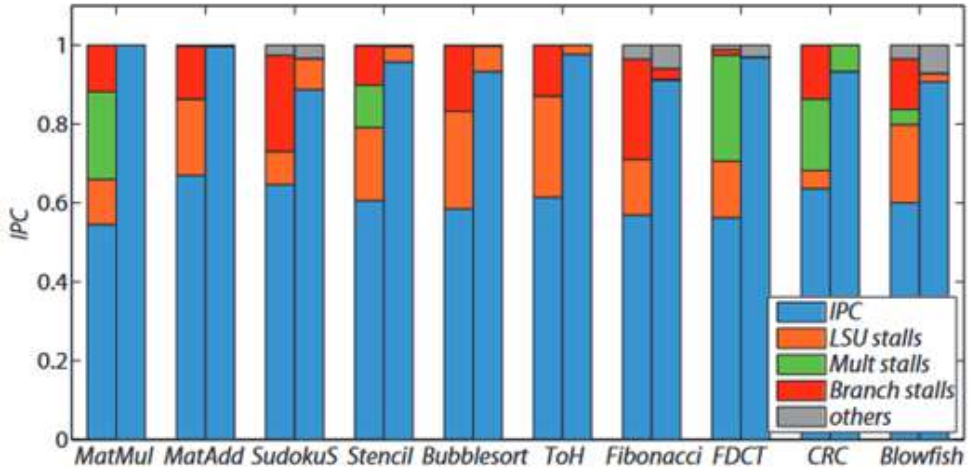
First Presence of PULP @ ORCONF in 2013

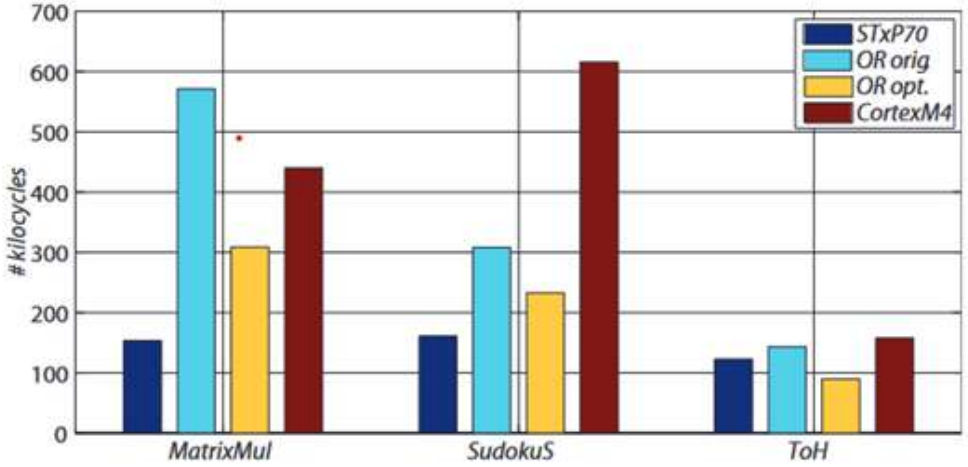
ORCONF 2013 5-6/10/2013

OR1200 Microarchitectural Analysis (And Optimizations)

- Blocking LSU (2 cycles)
- Blocking Multiplier (3 cycles)
- I$ →TCDM Combinational Path
- No Support for compressed instructions

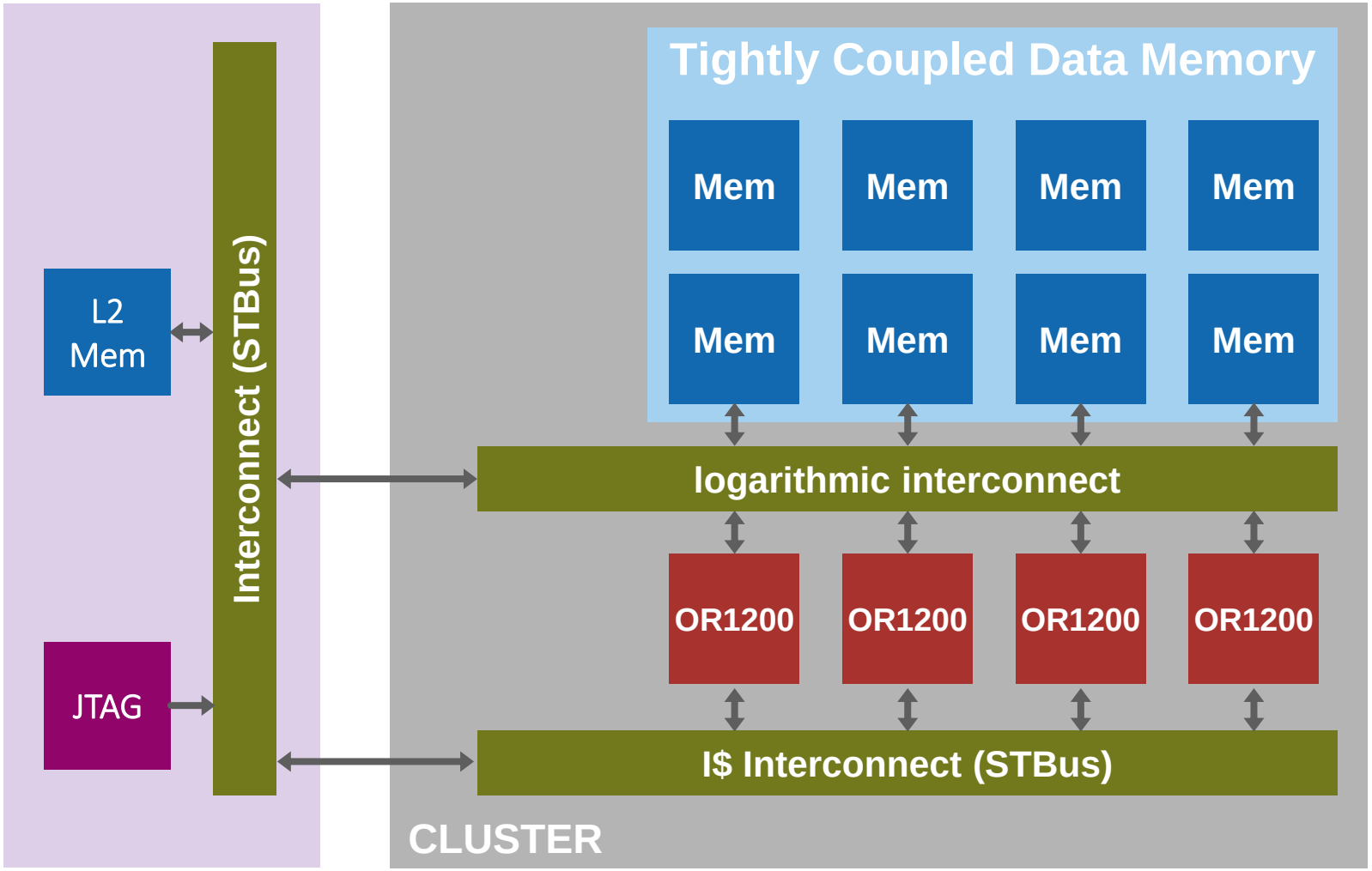






The first “PULP cluster”: PULPv1 - TO Dec 2013

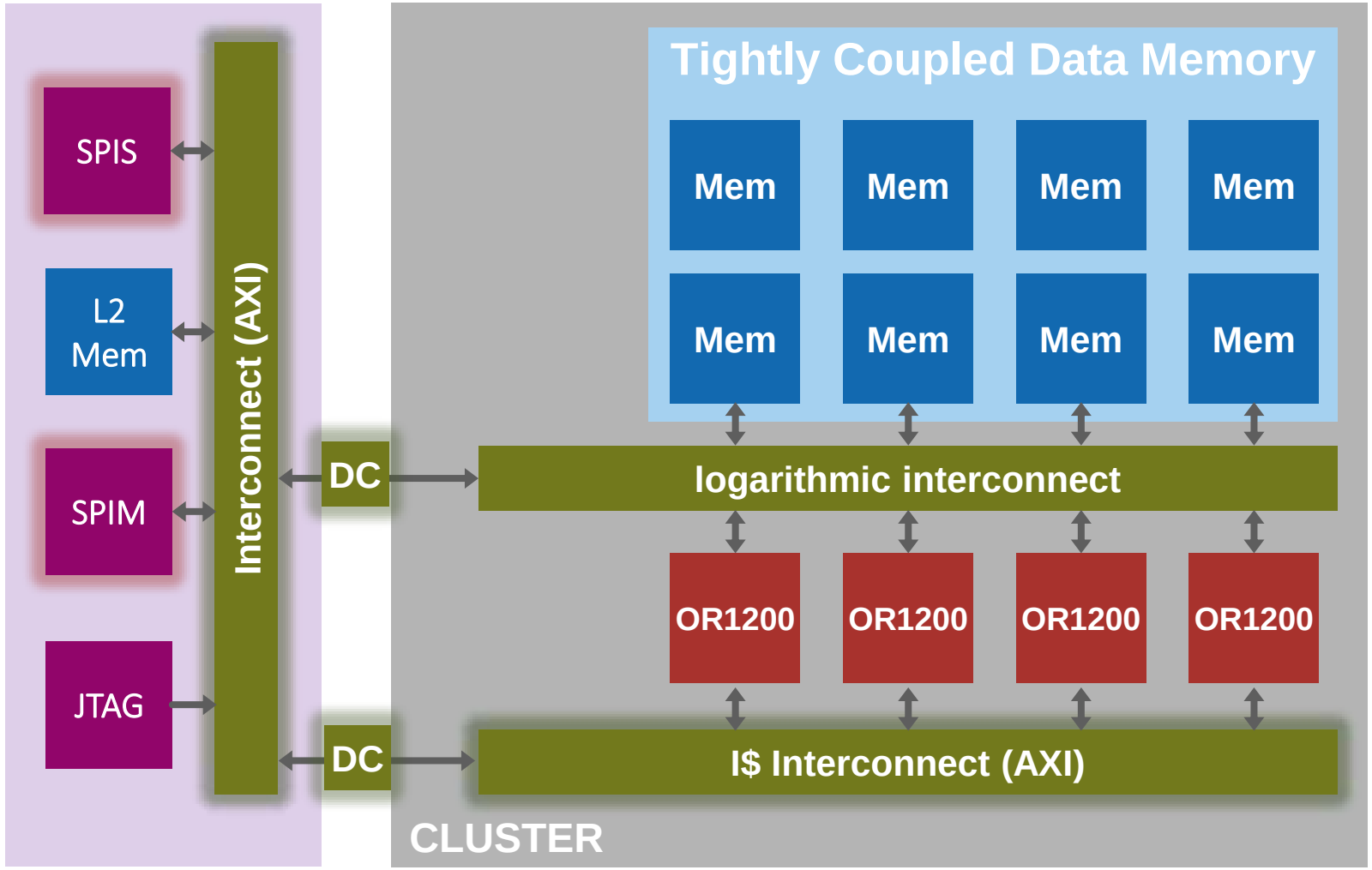
- 4-cores + I$ cluster
- Logarithmic interco.
- I$ + L2 Interco.
- L2 memory
- JTAG



....not really “Open Source”...☺

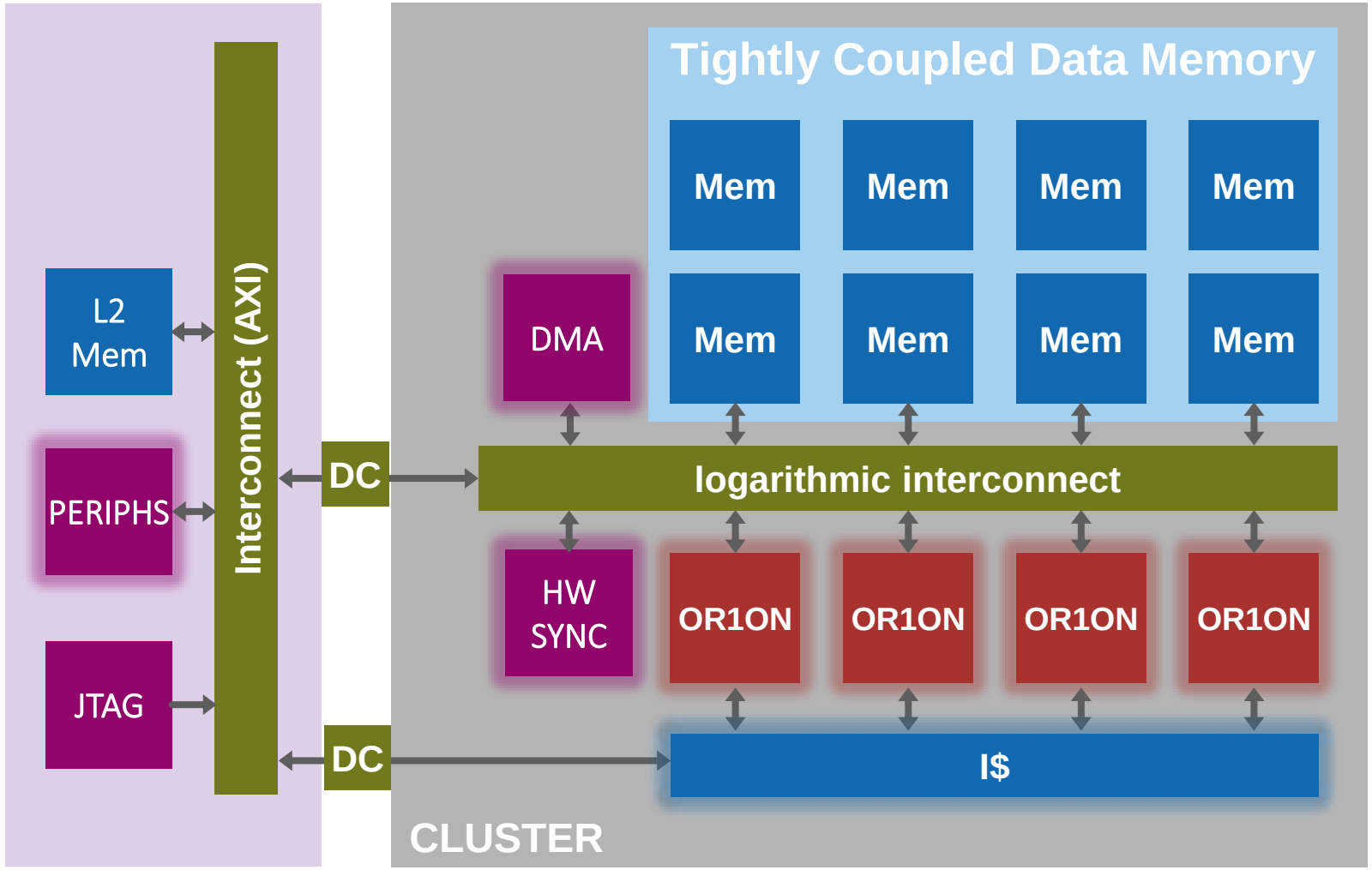
The Evolution of the Species: PULPv2 - TO Dec 2014

- PULPv1 +
- New (very rudimental) but Open Source AXI interconnect
- Some new peripherals (SPIM, SPIS)
- Instruction Cache
- DVFS-ready



The Evolution of the Species: PULPv3 - TO Dec 2015

- PULPv2 +
- OR1ON cores with DSP Extensions (brand new core)
- Hardware Synchronizer
- Instruction Cache
- Standard Peripherals



PULP 1,2,3 at Hot Chips 27, 2015 + Other Open Source Contribs.

MIAOW: An Open Source GPGPU

www.miaowgpu.org

Vinay Gangadhar, Raghu Balasubramanian, Mario Drumond, Ziliang Guo, Jai Menon, Cherin Joseph, Robin Prakash, Sharath Prasad, Pradip Vallathol, **Karu Sankaralingam**

Vertical Research Group
University of Wisconsin - Madison

IAOW

ASPIRE BWRC RISC-V

Raven: A 28nm RISC-V Vector Processor with Integrated Switched-Capacitor DC-DC Converters and Adaptive Clocking

Yunsup Lee, Brian Zimmer, Andrew Waterman, Alberto Puggelli, Jaehwa Kwak, Ruzica Jevtic, Ben Keller, Stevo Bailey, Milovan Blagojevic, Pi-Feng Chiu, Henry Cook, Rimas Avizienis, Brian Richards, Elad Alon, Borivoje Nikolic, Krste Asanovic

University of California, Berkeley

ETH zürich

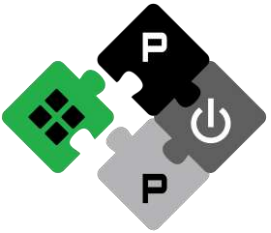
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Wind of Change….First RISC-V Workshop, January 2015

- We started to push the accelerator…
  - Switch to RISC-V (RTL, compilers, etc…)
  - size of the group increased (20-30 people)

January 14-15, 2015
Marriott Hotel, Monterey, CA

PULP - Open HW: RISC-V Cores,... and more



RISC-V Cores

RI5CY	Ibex	Snitch	Ariane + Ara
32b	32b	32b	64b

Peripherals

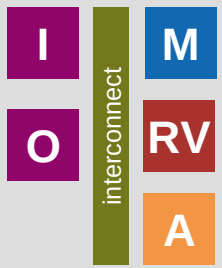
JTAG	SPI
UART	I2S
DMA	GPIO

Interconnect

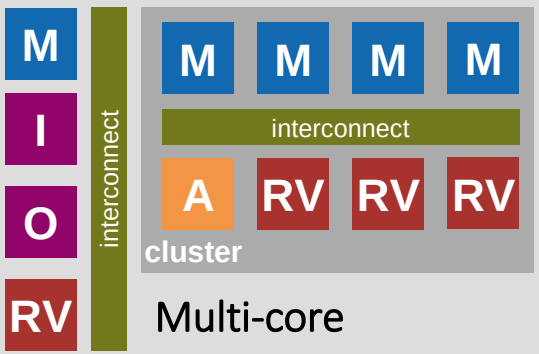
- Logarithmic interconnect
- APB – Peripheral Bus
- AXI4 – Interconnect

Platforms

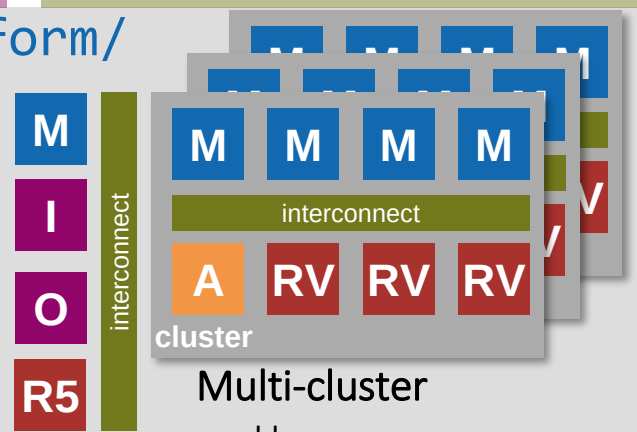
https://github.com/pulp-platform/



Single Core
- PULPino
- PULPissimo



Multi-core
- Control-PULP
- Kraken



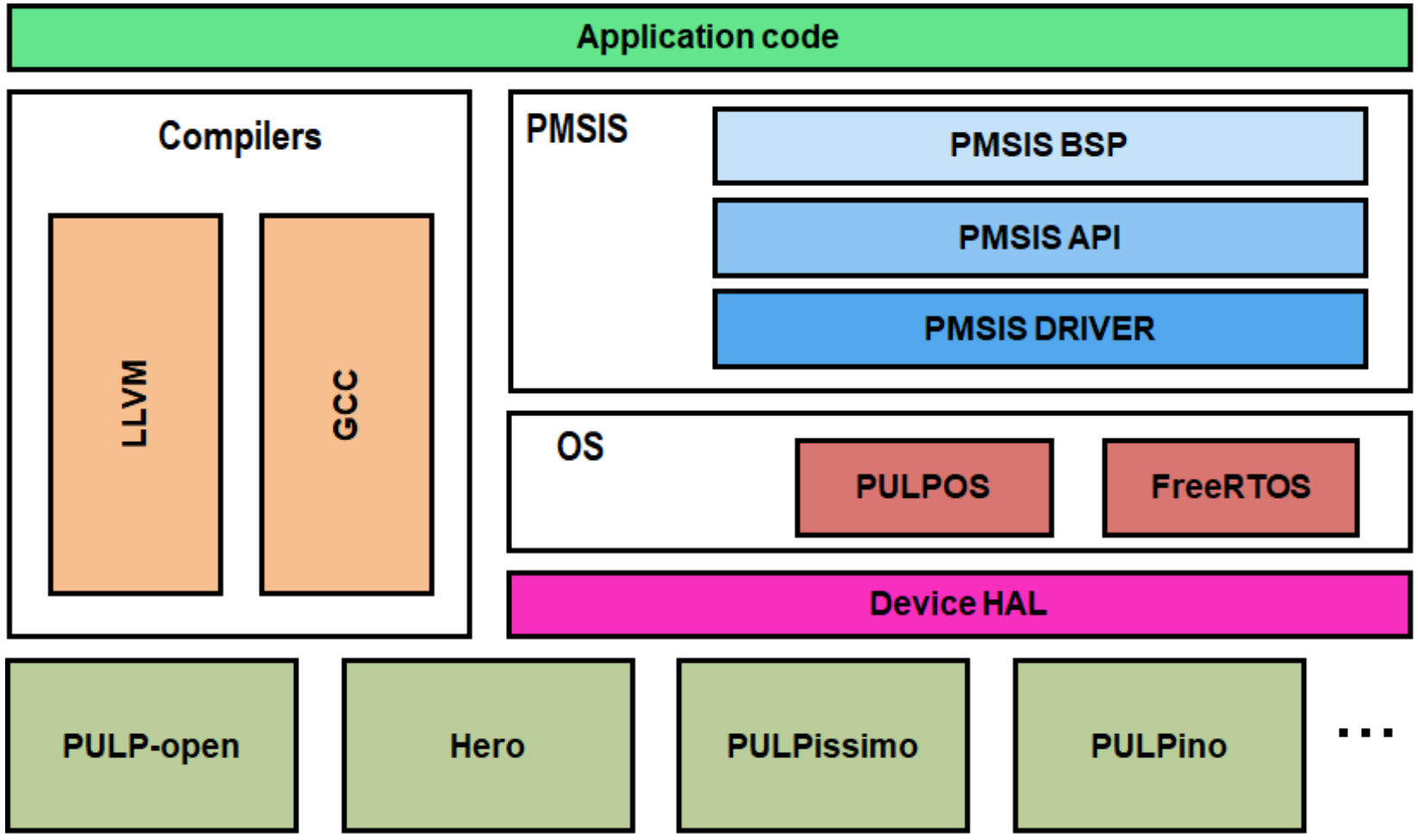
Multi-cluster
- Hero
- Occamy

IOT → HPC

Accelerators

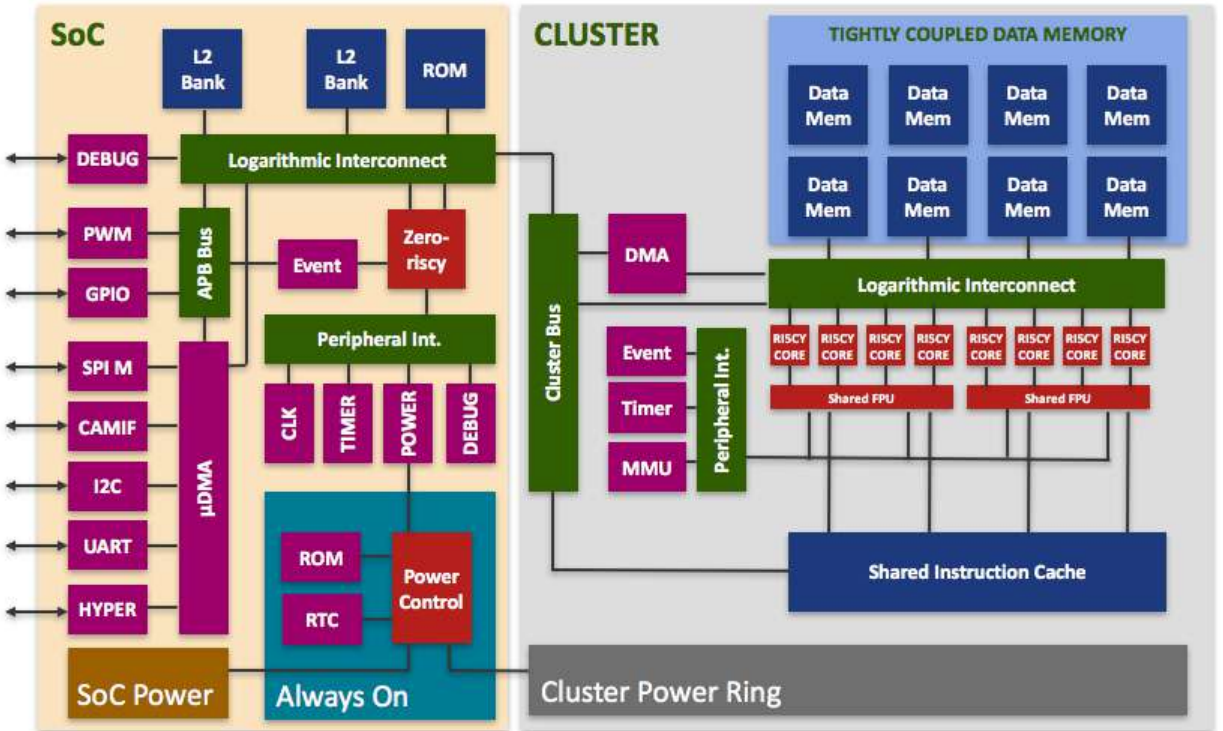
HWCE,1,2,3 (convolution)	Neurostream (ML train)	FPNEW (64/32/16/8)	ARA (DFP vector)

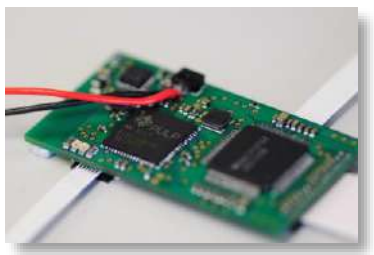
PULP Software Stack



Mr. Wolf SoC: First Board-Ready Chip (Dec 2018)

- First SoC with Fabric controller: 32-bit RISC-V processor (Zero-RISKY)
- Autonomous IO DMA Subsystem
- Rich set of peripherals
- Parallel Programmable Accelerator:
  - First pulp chip with RI5CY
  - First Floating-Point Capable PULP chip





GREENWAVES TECHNOLOGIES

bitcraze

ETH zürich

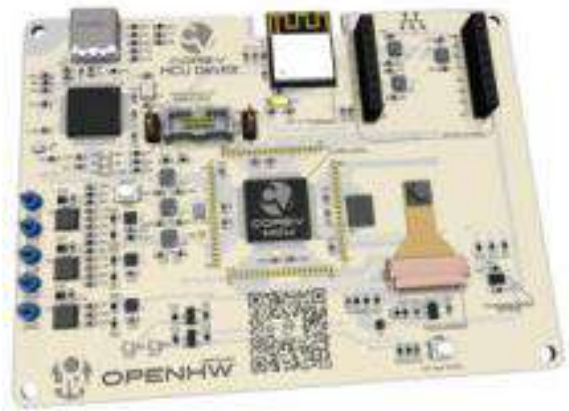
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

From Academia to Industry

- **OpenHW Group** is a not-for-profit, global organization (EU,NA,Asia) where HW and SW designers collaborate in the development of open-source cores, related IP, tools and SW such as the **Core-V** family
- **OpenHW Group** provides an infrastructure for hosting high quality open-source HW in line with industry best practices.

PULP
Parallel Ultra Low Power

RI5CY, ARIANE RISC-V cores, FPU, AXI4 Components.



OPENHW GROUP
PROVEN PROCESSOR IP

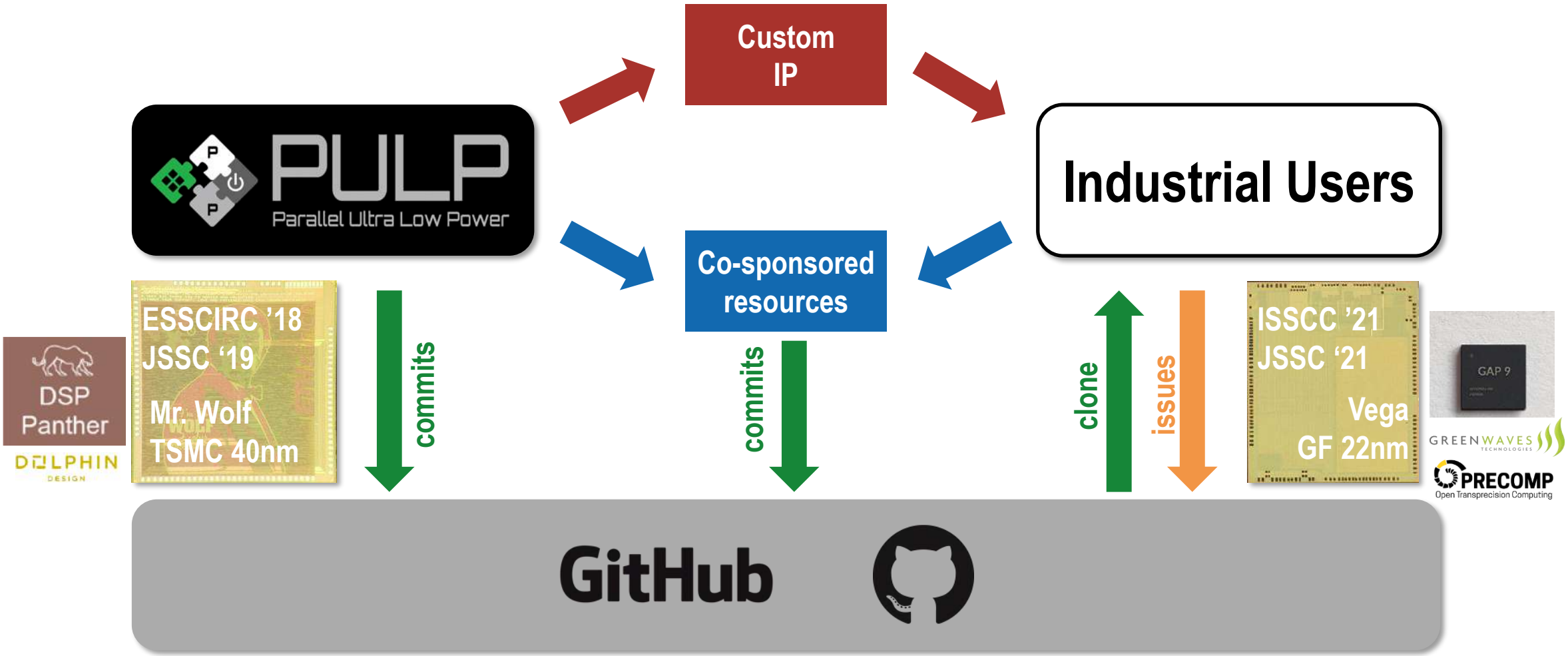
ETH zürich
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

OPENHW GROUP
PROVEN PROCESSOR IP

Members

10X ENGINEERS, aLean-Tec, Alibaba Group, aws, AMIQ EDA, ASHLING, Barcelona Supercomputing Center, beagleboard.org, 北京开源芯片研究院, bluespec, cea, CIRCUIT SUTRA, CMC MICROSYSTEMS, FRONTGRADE, Codasip, CRYPTO QUANTIQUE, csem, DATUM, DOLPHIN, ECSPEC, em microelectronic, EMBECOSM, Empaiot., ESPRESSIF, ETH zürich, FUTUREWEI, GF, GREENWAVES, GroupGets, HARVEY MUDD COLLEGE, HUAWEI, 中国科学院计算技术研究所, imperas, intel, INTRINSIX, ISCAS 中国科学院软件研究所, Jade, KALRAY, lpf Low Power Futures, McMaster University, mu ELECTRONICS, NXP, OKLAHOMA STATE UNIVERSITY, Plan, Politecnico di Torino, POLYTECHNIQUE MONTRÉAL, QuickLogic, RIOS, Rumble Development, SIEMENS, Silicon Assurance, SILICON LABS, SFU SIMON FRASER UNIVERSITY, SCIENCES SORBONNE UNIVERSITÉ, Synthara, TECHNION Israel Institute of Technology, THALES, Science and Technology Facilities Council, Universidade do Minho, ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA, UC SANTA BARBARA, University of Reading, UNIVERSITY OF SASKATCHEWAN, UNIVERSITY OF Southampton, whalemicro, XMOS, ZERODAY LABS

Industrial Paths for PULP IPs & Mutual Benefits



IoT & TinyML Applications and Chips

Vega: IoT Heterogeneous SoC

GREENWAVES TECHNOLOGIES

- Transprecision Floating-Point Units (16/32-bit float)[1]
- 4 MB non-volatile MRAM (Weight Storage)[2]
- Programmable Cognitive Wake-up Unit based on HD Computing

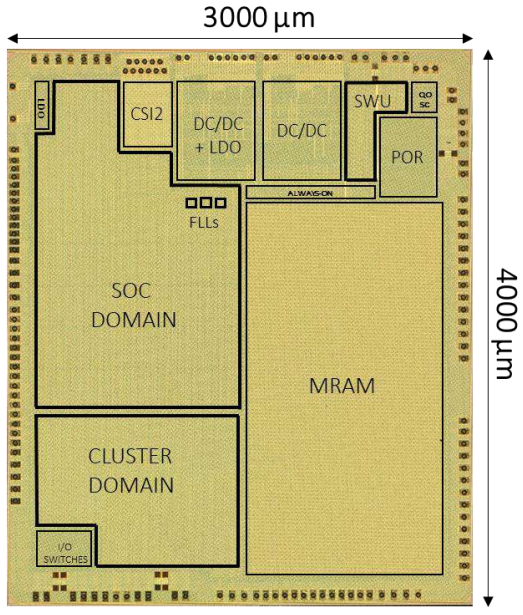
Presented at ISSCC 2021

ISSCC 2021 / SESSION 4 / PROCESSORS / 4.4

4.4 A 1.3TOPS/W @ 32GOPS Fully Integrated 10-Core SoC for IoT End-Nodes with 1.7µW Cognitive Wake-Up From MRAM-Based State-Retentive Sleep Mode

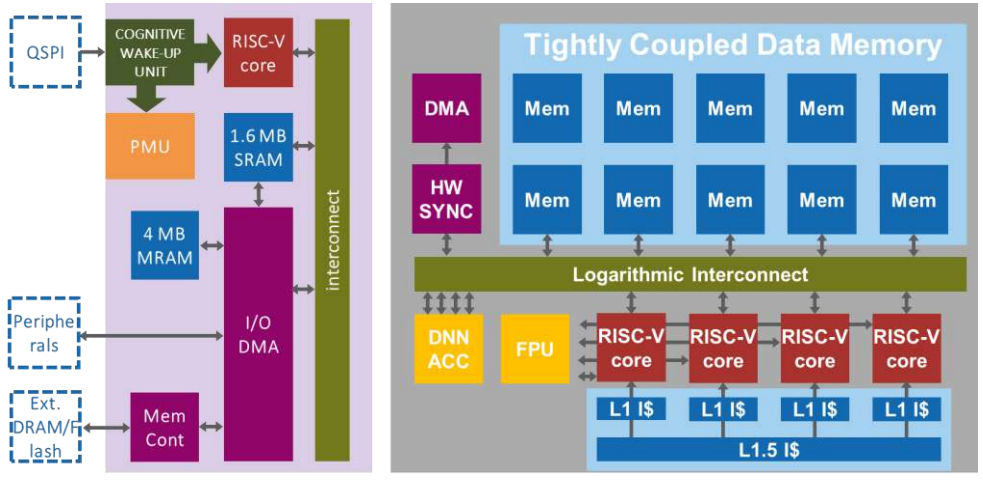
Davide Rossi[1], Francesco Conti[1], Manuel Eggiman[2], Stefan Mach[2], Alfio Di Mauro[2], Marco Guermandi[1,3], Giuseppe Tagliavini[1], Antonio Pullini[2,3], Igor Loi[3], Jie Chen[1,3], Eric Flamand[2,3], Luca Benini[1,2]

[1]University of Bologna, Bologna, Italy
[2]ETH Zurich, Zurich, Switzerland
[3]Greenwaves Technologies, Grenoble, France



Prototype implemented in GF 22FDX

flip-well LVT & SLVT cells, 12mm2 for Full SoC



ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

State-of-the-Art Comparison

	SleepRunner [2]	Mr.Wolf [3]	SamurAI [4]	*VEGA*
Embedded NVM	-	-	SoA Int & FP Perf & Efficiency; Boost of 4.3x & 2.8x compared to Wolf	4 MB MRAM
Wake-up Sources	WiC	GPIO, RT		PIO, RTC, Cognitive
Best Int Perf. Best.Int Eff. @ Perf.	31 MOPS (32b) 97 MOPS/mW (32b) @ 18.6 MOPS (32b)	12.1 GOPS 190 GOPS/W @ 3.8 GOPS	1.5 GOPS 230 GOPS/W @110 MOPS	15.6 GOPS 614 GOPS/W @7.6 GOPS
Best FP Perf. Best FP Eff. @Perf	-	1 GFLOPS 18 GFLOPS/W @350 MFLOPS	-	4 GFLOPS 158 GFLOPS/W @ 2 GFLOPS
Best ML Perf. Best ML Eff. @Perf	-	-	36 GOPS 1.3 TOPS/W @ 2.8 GOPS	32.2 GOPS 1.3 TOPS/W @15.6 GOPS

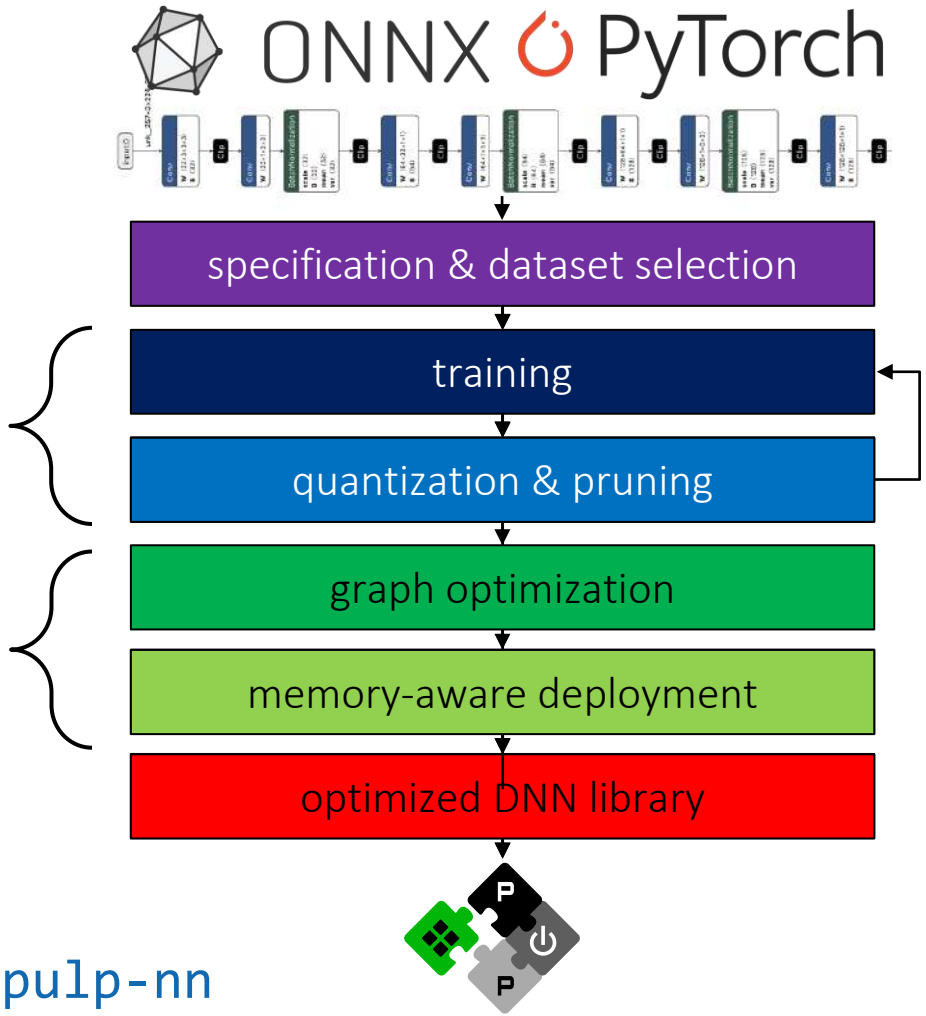
Open Software Stack for VEGA: *QuantLab, Nemo, Dory*

QuantLab
Quantization Laboratory

NEMO
NEural **M**inimization for pyt**O**rch

DORY
Deployment **O**riented to memo**RY**

PULP-NN
PULP **N**eural **N**etwork backend



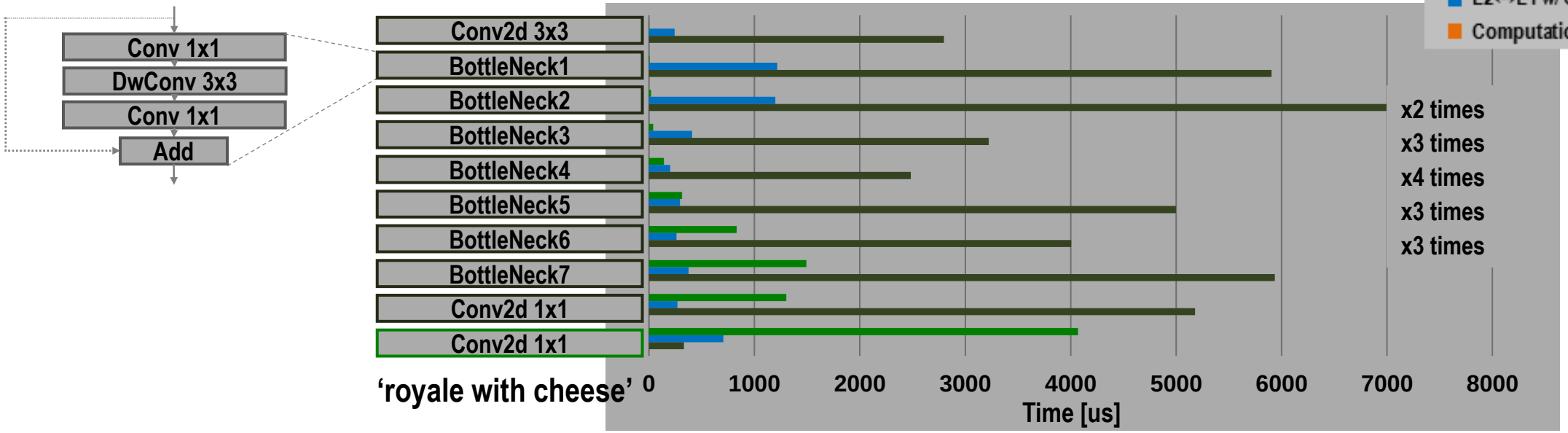
`github.com/pulp-platform/nemo,/dory,/pulp-nn`

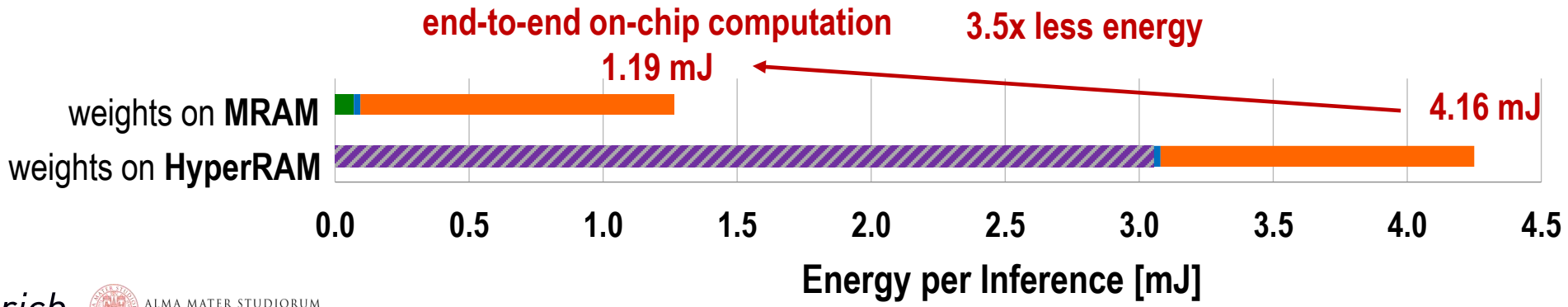
ETH zürich
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

A. Burrello, A. Garofalo, N. Bruschi, G. Tagliavini, D. Rossi and F. Conti, "DORY: Automatic End-to-End Deployment of Real-World DNNs on Low-Cost IoT MCUs," in *IEEE Transactions on Coputers*, vol. 70, no. 8, pp. 1253-1268, 1 Aug. 2021.

First Fully Integrated MobileNetV2 on an IoT SoC

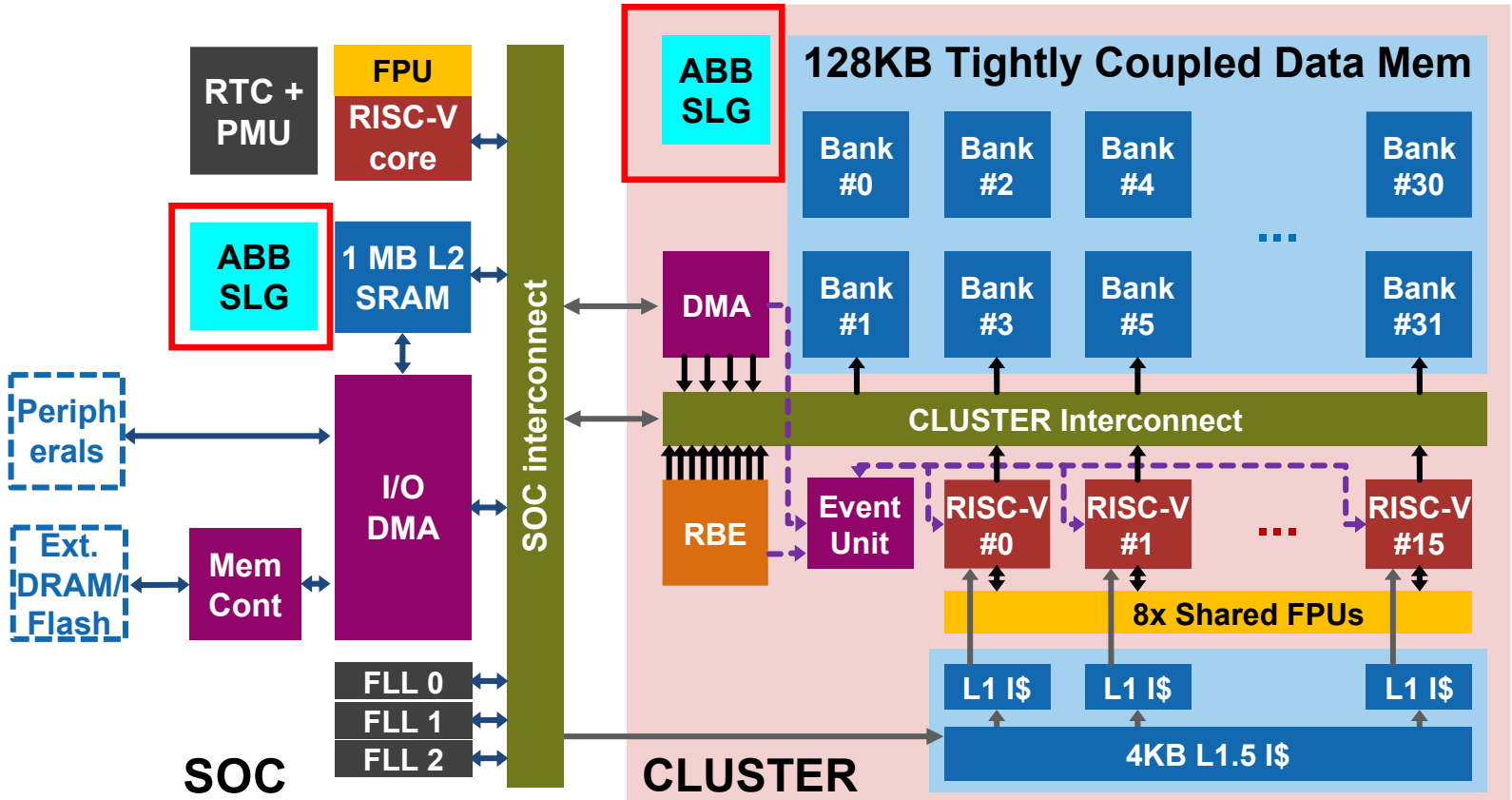
@ Vdd_SOC=0.8V, f_SOC=250 MHz, f_CL=250 MHz





MARSELLUS: AI-IoT Heterogeneous SoC

- Machine Learning ISA Extensions
- 2-8b Reconfigurable Binary Engine for 3x3, 1x1 DNN kernels
- Adaptive Body Biasing with on-the-fly control



22.1 A 12.4TOPS/W @ 136GOPS AI-IoT System-on-Chip with 16 RISC-V, 2-to-8b Precision-Scalable DNN Acceleration and 30%-Boost Adaptive Body Biasing

Francesco Conti[1], Davide Rossi[1], Gianna Paulin[2], Angelo Garofalo[1], Alfio Di Mauro[2], Georg Ruetishauer[2], Gianmarco Ottavi[1], Manuel Eggimann[2], Hayate Okuhara[1], Vincent Huard[3], Olivier Montfort[3], Lionel Jure[3], Nils Exibard[3], Pascal Gouedo[3], Mathieu Louvat[3], Emmanuel Botte[3], Luca Benini[1,2]

Presented at ISSCC 2023

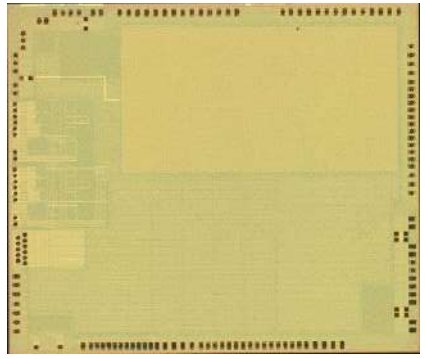
State-of-the-Art Comparison (HW Accelerated)

	VEGA [1] ISSCC'21	SAMURAI [2] VLSIC'20	DIANA [3] ISSCC'22	*MARSELLUS (this work)*
Technology	22nm FDX	28nm FD-SOI	22nm FDX + AIMC	22nm FDX
Die Area	10 mm^2	4.5 mm^2	10.24 mm^2	18.7 mm^2 (cluster 1.9 mm^2)
Cores	10x RV32IMCFXpulp + HWCE	1x RV32IMCFXpulp + Digital Accel.	1x RV32CIMFXpulp + Digital Accel. + AIMC	1x RV32IMCFXpulp + 16x RV32IMCFXpulpnn + RBE
Max Frequency	450 MHz	350 MHz	320 MHz	420 MHz
Power range	1.7 uW - 49.4 mW	6.4 uW - 96 mW	10-129 mW (digital)	12.8 mW - 123 mW
Best SW (INT) Perf	15.6 GOPS (8 RISC-V)	1.5 GOPS (1 RISC-V)	-	**90 GOPS (16 RISC-V M&L 2x2b, 0.8V+ABB)**
Best SW (INT) Eff	614 GOPS/W @ 7.6 GOPS (8 RISC-V)	230 GOPS/W @ 110 MOPS (1 RISC-V)	-	**1.66 TOPS/W @ 19 GOPS (16 RISC-V M&L 2x2b)**
Best HW-Accel Perf	32.2 GOPS (HWCE)	36 GOPS (Dig	180 GOPS (Digital), AIMC)	**637 GOPS (RBE 2x2b, 0.8V+ABB)**
Best HW-Accel Eff	1.3 TOPS/W @ 15.6 GOPS (HWCE)	1.3 @ 2 (Digital Accel.)	/W 600 TOPS/W (AIMC)	**12.4 TOPS/W @ 136 GOPS (RBE 2x2b)**

DNN Acceleration
Perf & Efficiency boost
>10x w.r.t. VEGA

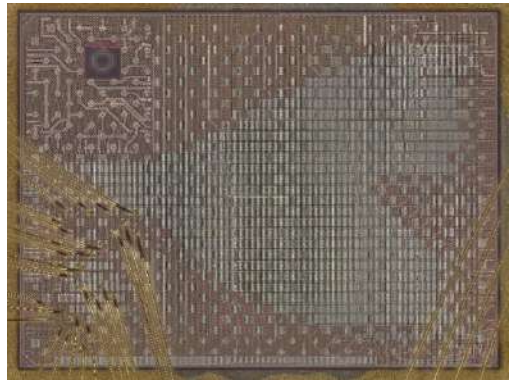
More on Francesco Conti's talk

IoT SoC Playground: Success Stories



Vega 22nm, *ISSCC'21*
(UNIBO + GreenWaves)

GREENWAVES TECHNOLOGIES



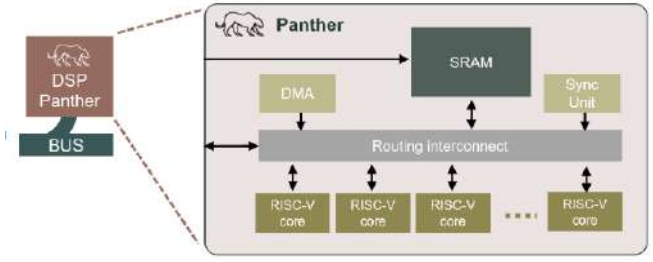
Marsellus 22nm, *ISSCC'23*
(UNIBO + Dolphin Design)

DOLPHIN DESIGN

Fixed weight precision to 8/16-bit

GAP9 SoC *(commercial)*

65% of GAP9 is based on PULP open-source IPs



Panther DSP IP *(commercial)*

90% HW and 20% SW of Phanter based on PULP open-source IPs

DOLPHIN DESIGN

RISC-V (PULP) Currently Dominates TinyML benchmarks

November 09, 2022 – Inference: Tiny v1.0 mlcommons.org

							Benchmark Results							
							Task	Visual Wake Words		Image Classification		Keyword Spotting		Anomaly Deteciton	
							Data	Visual Wake Words Dataset		CIFAR-10		Google Speech Commands		ToyADMOS (ToyCar)	
							Model	MobileNetV1 (0.25x)		ResNet-V1		DSCNN		FC AutoEncoder	
							Accuracy	80% (top 1)		85% (top 1)		90% (top 1)		0.85 (AUC)	
Submitter	Board Name	SoC Name	Processor(s) & Number	Accelerator(s) & Number	Software	Notes	Units	Latency in ms	Energy in uJ	Latency in ms	Energy in uJ	Latency in ms	Energy in uJ	Latency in ms	Energy in uJ
Greenwaves Technologies	GAP9 EVK	GAP9	RISC-V Core (1+9)	NE16 (1)	GreenWaves GAPFlow	GAP9 (370MHZ, 0.8Vcore)		1.13	58.4	0.62	40.4	0.48	26.7	0.18	7.29
Greenwaves Technologies	GAP9 EVK	GAP9	RISC-V Core (1+9)	NE16 (1)	GreenWaves GAPFlow	GAP9 (240MHZ, 0.65Vcore)		1.73	40.8	0.95	27.7	0.73	18.6	0.27	5.25
OctoML	NRF5340DK	nRF5340	Arm® Cortex®-M33		microTVM using CMSIS-NN backend	128MHz		232.0		316.1		76.1		6.27	
OctoML	NUCLEO-L4R5ZI	STM32L4R5ZIT6U	Arm® Cortex®-M4		microTVM using CMSIS-NN backend	120MHz, 1.8Vbat		301.2	155[illegible]1.4	389.5	2[illegible]236.3	99.8	5[illegible]230.3	8.60	443.2
OctoML	NUCLEO-L4R5ZI	STM32L4R5ZIT6U	Arm® Cortex®-M4		microTVM using native codegen	120MHz, 1.8Vbat		336.5	171[illegible]1.6	389.2	2[illegible]342.3	144.0	7[illegible]950.5	11.7	633.7
Plumerai	B_U585I_IOT02A	STM32U585	Arm® Cortex®-M33		Plumerai Inference Engine 2022.09	160MHz		107.0		107.1		35.4		4.90	
Plumerai	CY8CPROTO-062-4343w	PSoC 62 MCU	Arm® Cortex®-M4		Plumerai Inference Engine 2022.09	150MHz		192.5		[illegible]		[illegible]		6.70	
Plumerai	DISCO-F746NG	STM32F746	Arm® Cortex®-M7		Plumerai Inference Engine 2022.09	216MHz		57.0		64.8		19.1		2.30	
Plumerai	NUCLEO-L4R5ZI	STM32L4R5ZIT6U	Arm® Cortex®-M4		Plumerai Inference Engine 2022.09	120MHz		208.6		173.2		71.7		5.60	
Silicon Labs	xG24-DK2601B	EFR32MG24	Arm® Cortex®-M33	Silicon Labs MVP(1)	TensorFlowLite for Microcontrollers, CMSIS-NN, Silicon Labs Gecko SDK			111.6	11[illegible]9.2	120.9	[illegible]234.7	36.3	[illegible]401.9	5.43	47.3
STMicroelectronics	NUCLEO-H7A3ZI-Q	STM32H7A3ZIT6Q	Arm® Cortex®-M7		X-CUBE-AI v7.3.0	280MHz, 3.3Vbat		50.7	79[illegible]8.5	54.3	[illegible]707.3	16.8	2[illegible]721.8	1.82	266.5
STMicroelectronics	NUCLEO-L4R5ZI	STM32L4R5ZIT6U	Arm® Cortex®-M4		X-CUBE-AI v7.3.0	120MHz, 1.8Vbat		230.5	100[illegible]6.6	226.9	1[illegible]681.6	75.1	3[illegible]971.7	7.57	323.0
STMicroelectronics	NUCLEO-U575ZI-Q	STM32U575ZIT6Q	Arm® Cortex®-M33		X-CUBE-AI v7.3.0	160MHz, 1.8Vbat		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4.84	119.1
Syntiant	NDP9120-EVL	NDP120	M0 + HiFi	Syntiant Core 2 (98MHz	Syntiant TDK	Syntiant Core 2 (98MHz, 1		4.10	97.2	5.12	139.4	1.48	43.8		
Syntiant	NDP9120-EVL	NDP120	M0 + HiFi	Syntiant Core 2 (30MHz	Syntiant TDK	Syntiant Core 2 (30MHz, 0		12.7	71.7	16.0	101.8	4.37	31.5		
Qualcomm Innovation Center	Next Generation Snapdragon Mobile Platform HDK	Next Generation Snapdragon Mobile Platform	Qualcomm Kryo CPU(1)	Qualcomm Sensing Hub(1)	Qualcomm AI Stack									0.098	

ENERGY GAP

1.7x 3.7x 1.7x

Higher-Performance PULP?

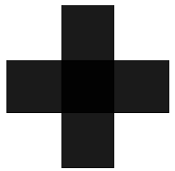
- So far, focus on **low-power** applications: near-sensor processing, nano-UAVs, etc…
- Two niches left out: **Linux-capable low-power MPUs**…
  - can exploit clusters for performance boost
- … and what about **high-performance computing**?
  - energy efficiency of capital importance (**power = $$$** spent for **cooling**, **energy bill**)

- **Hardware accelerators** provide a key technology for **HPC**
  - compute-dominated workloads
  - highly parallel workloads
  - efficiency in Joules/op and power envelope in kW are important metrics
  - flexibility is also of primary importance

Heterogeneous Systems-on-Chip (HeSoCs)

HOST

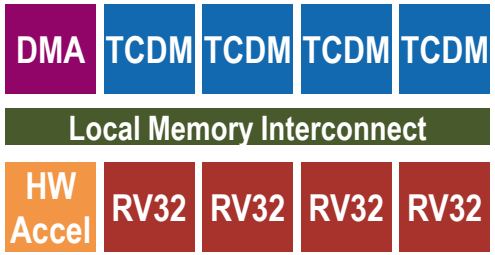
- General-purpose
- Linux-capable
- Versatility
- Programmability



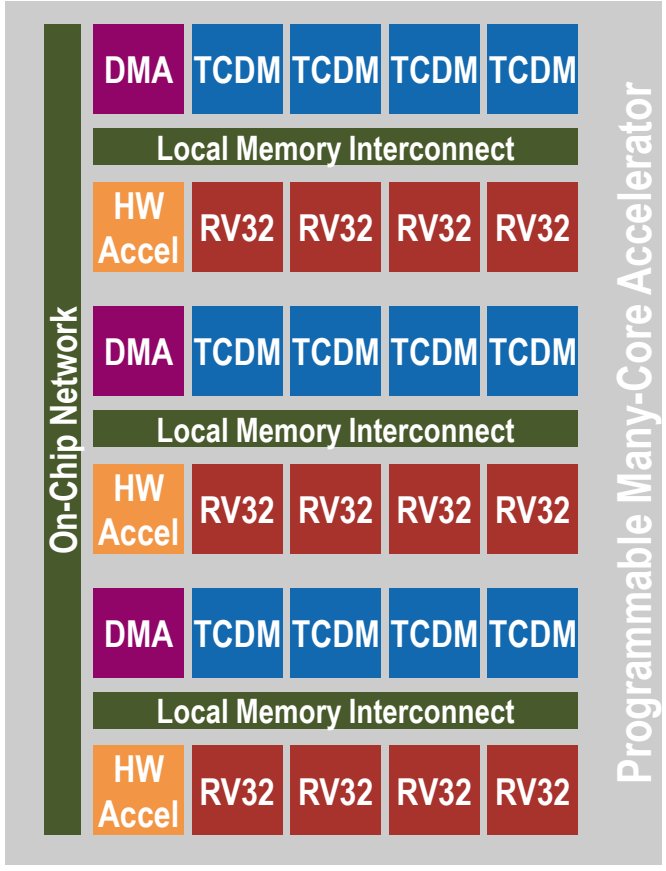
PMCA

- Parallel Manycore Accelerator (PMCA)
- Domain-specialized
- Energy-efficient

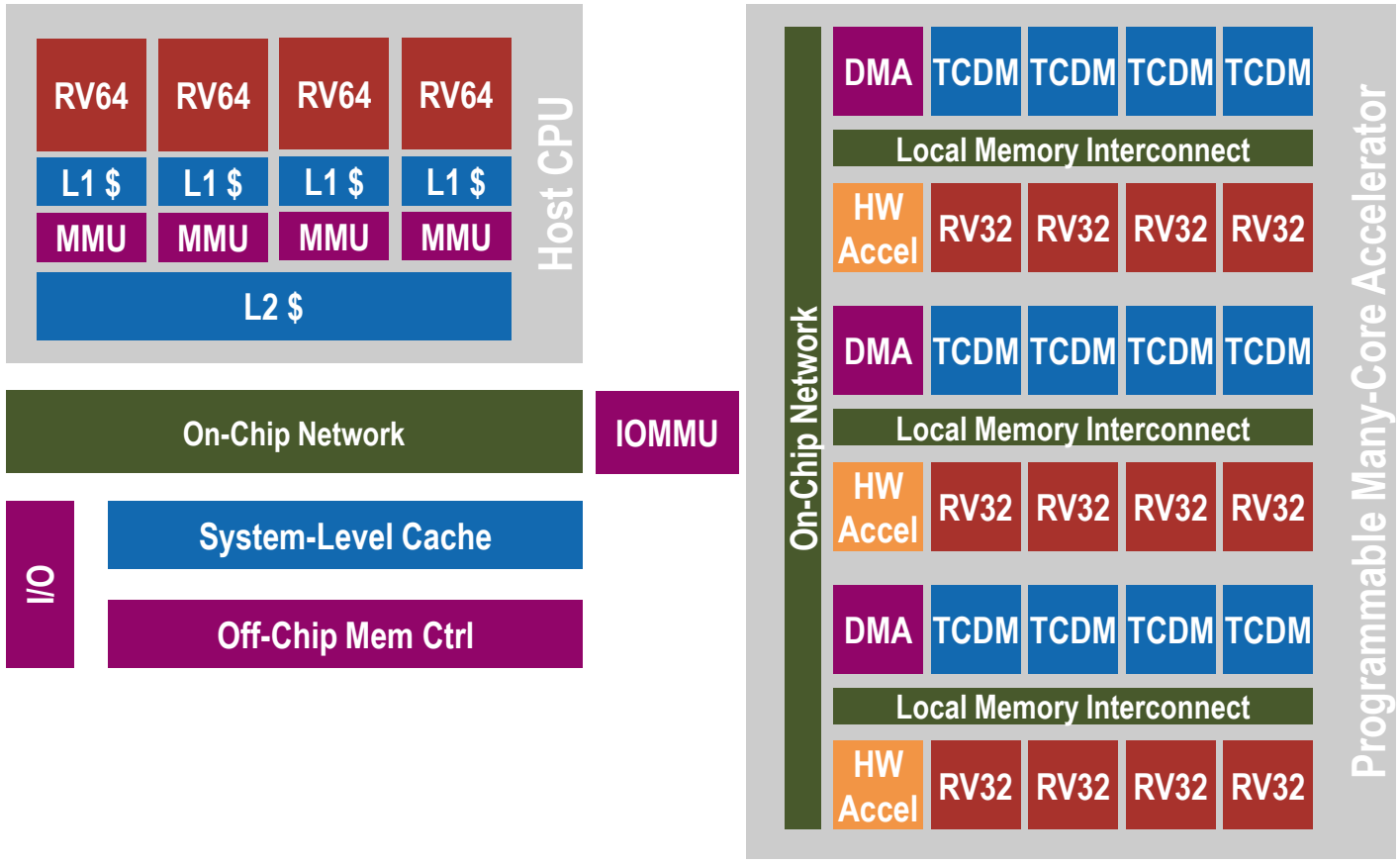
HERO: Hardware Architecture



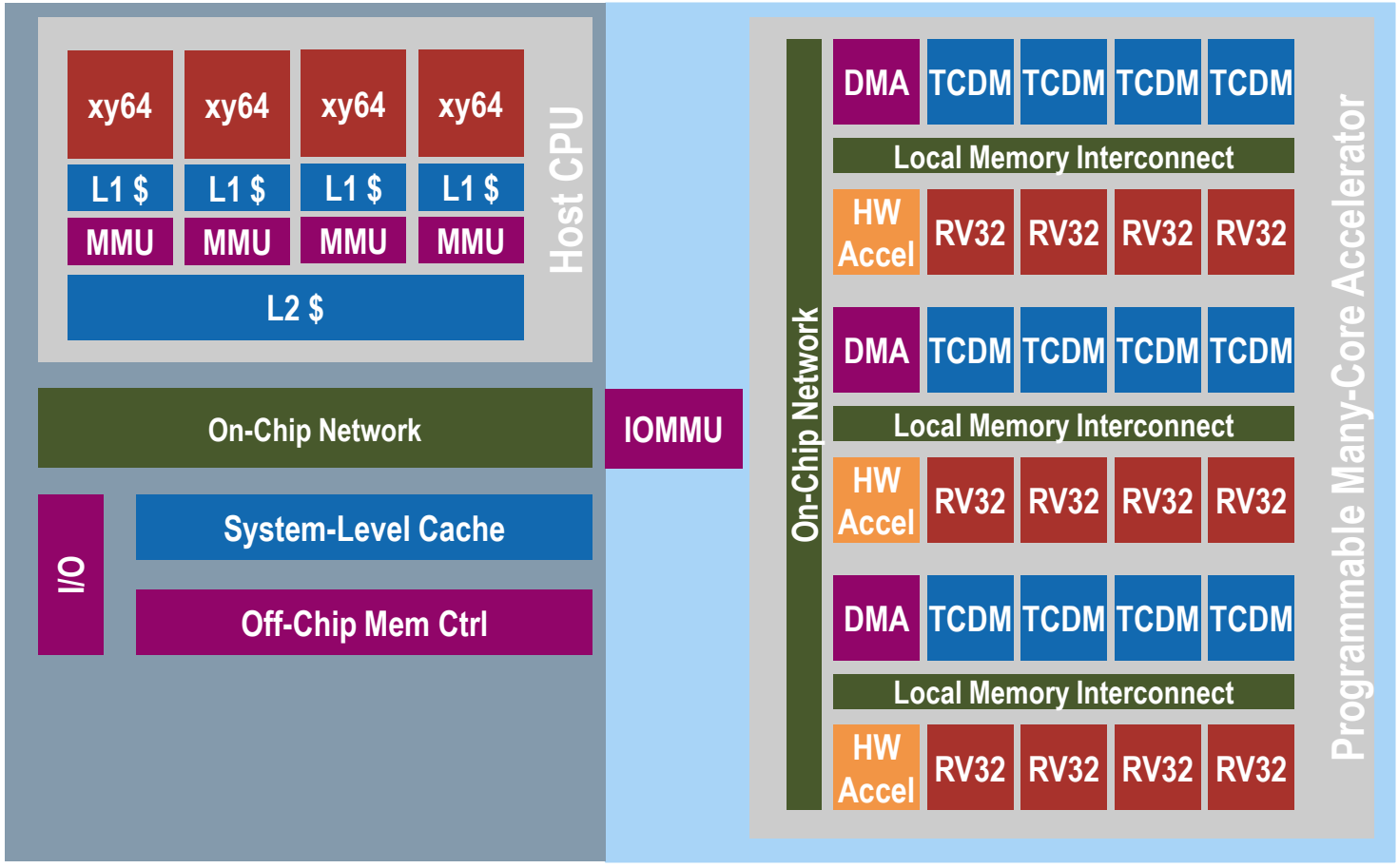
HERO: Hardware Architecture



HERO: Hardware Architecture



HERO: Hardware Architecture

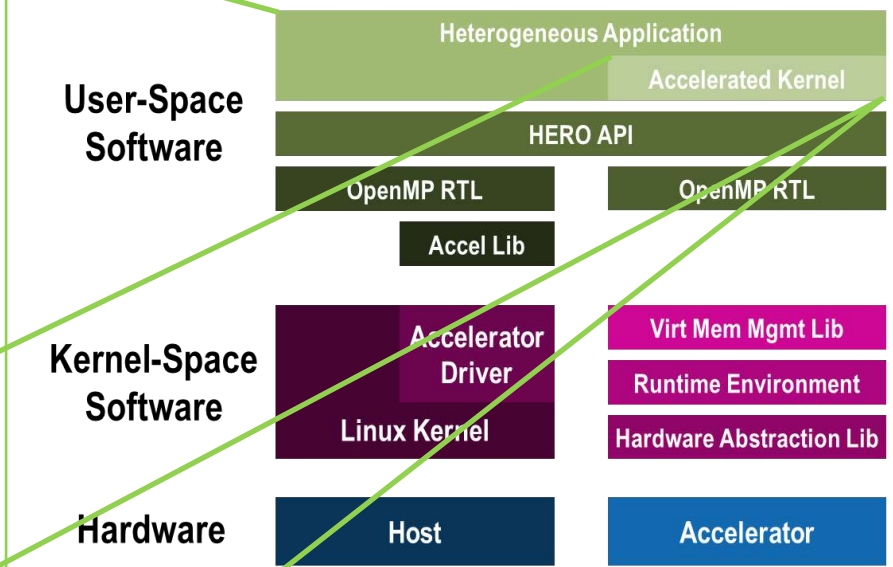


L. Valente et. al., "HULK-V: a Heterogeneous Ultra-low-power Linux capable RISC-V SoC", DATE 2023

HERO: Software Architecture

Principle: single-source heterogeneous programming. Offload with OpenMP 4.5 target semantics.
Example:

```
int main ()
{
  vertex vertices[N];
  load(&vertices, N);
  #pragma omp target map(tofrom: vertices)
  {
    #pragma omp parallel
    for  (i = 0; i < N; ++i)
      process(vertices[i]);
  }
}
```



HERO: Heterogeneous Compilation

Single-source, single-binary heterogeneous compilation

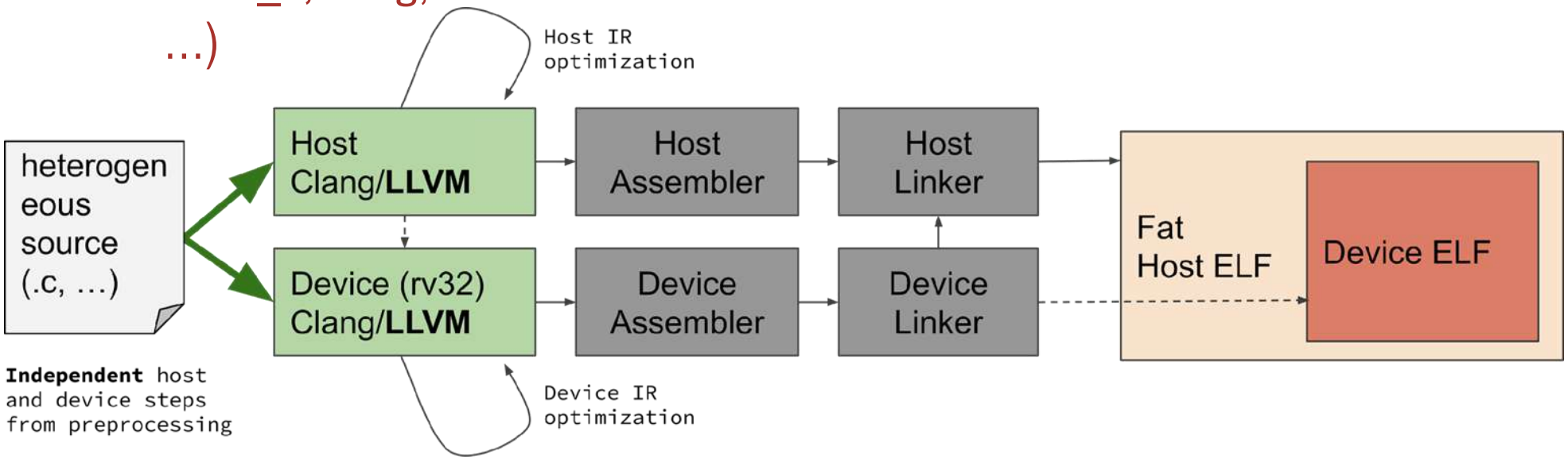
Single Source, Common API, Specialized Code

Separate data models (width of pointers, `size_t`, `long`, …)

Separate ISAs

Separate Libraries

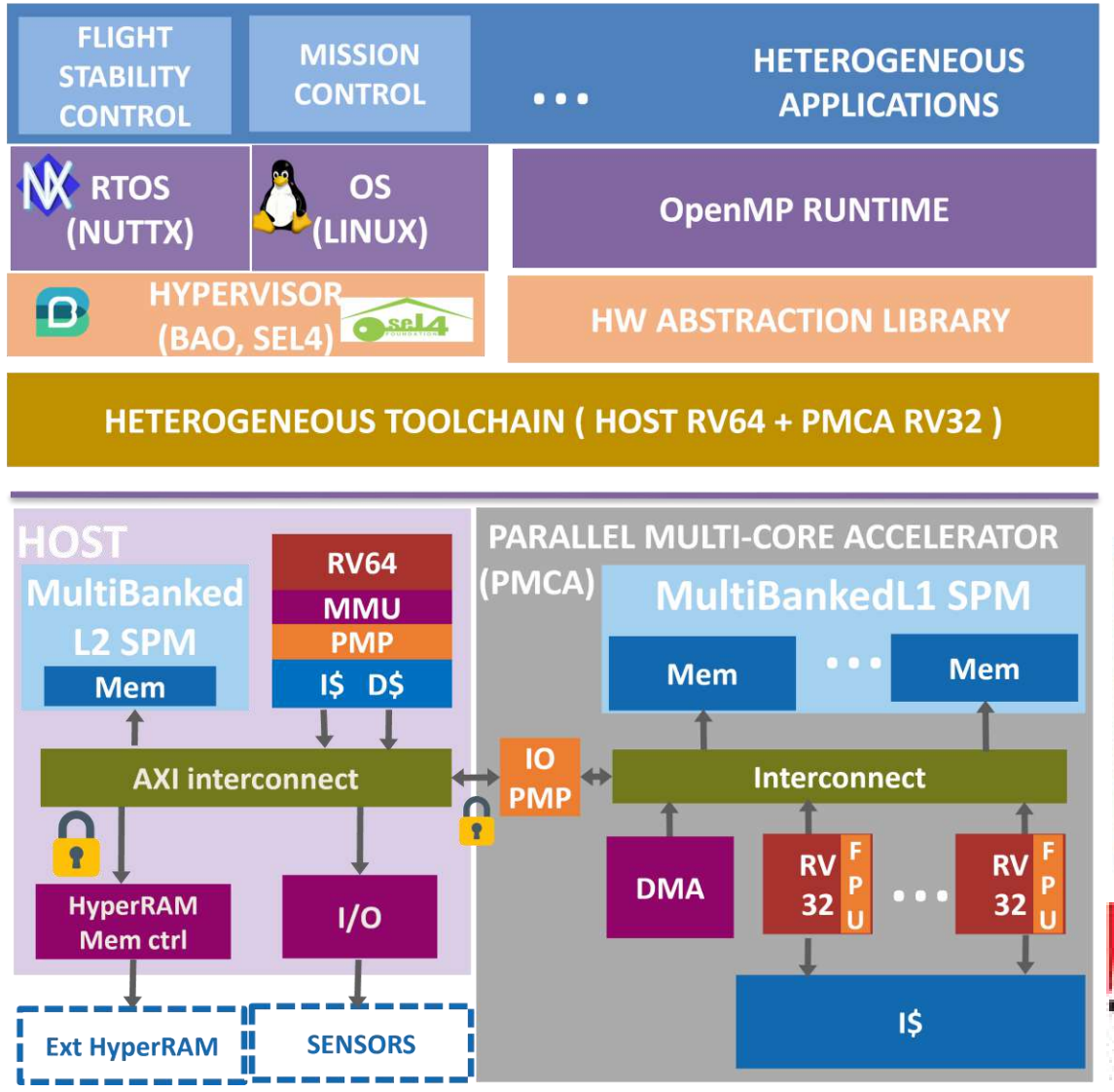
Single Binary, Nested Executables

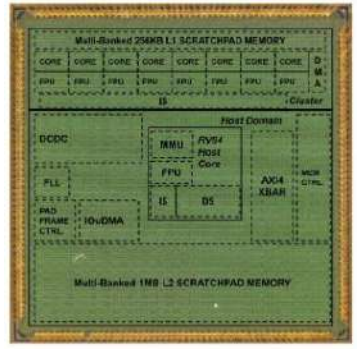


OpenMP offloading requires a **host compiler** plus **one target compiler for each PMCA ISA** in the system.

Shaheen: Heterogeneous SoC for Nano-UAVs

- **Secure Heterogeneous Application Processor**
- **Host Subsystem**
  - CVA6
  - H-Extensions →BAO, Sel4
  - Timing channel attack protection
- **PULP Cluster with Mixed-Precision Extension**
- **HyperRAM Memory Controller**
  - Up to 512 Mbit
  - Up to 1.6 Gbit/s





ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

PULP in HPC

Occamy – Massive Scaling

Dual Chiplet System Occamy:

- Technology: GF12LP+
- Area: 73mm^2

Interposer Hedwig:

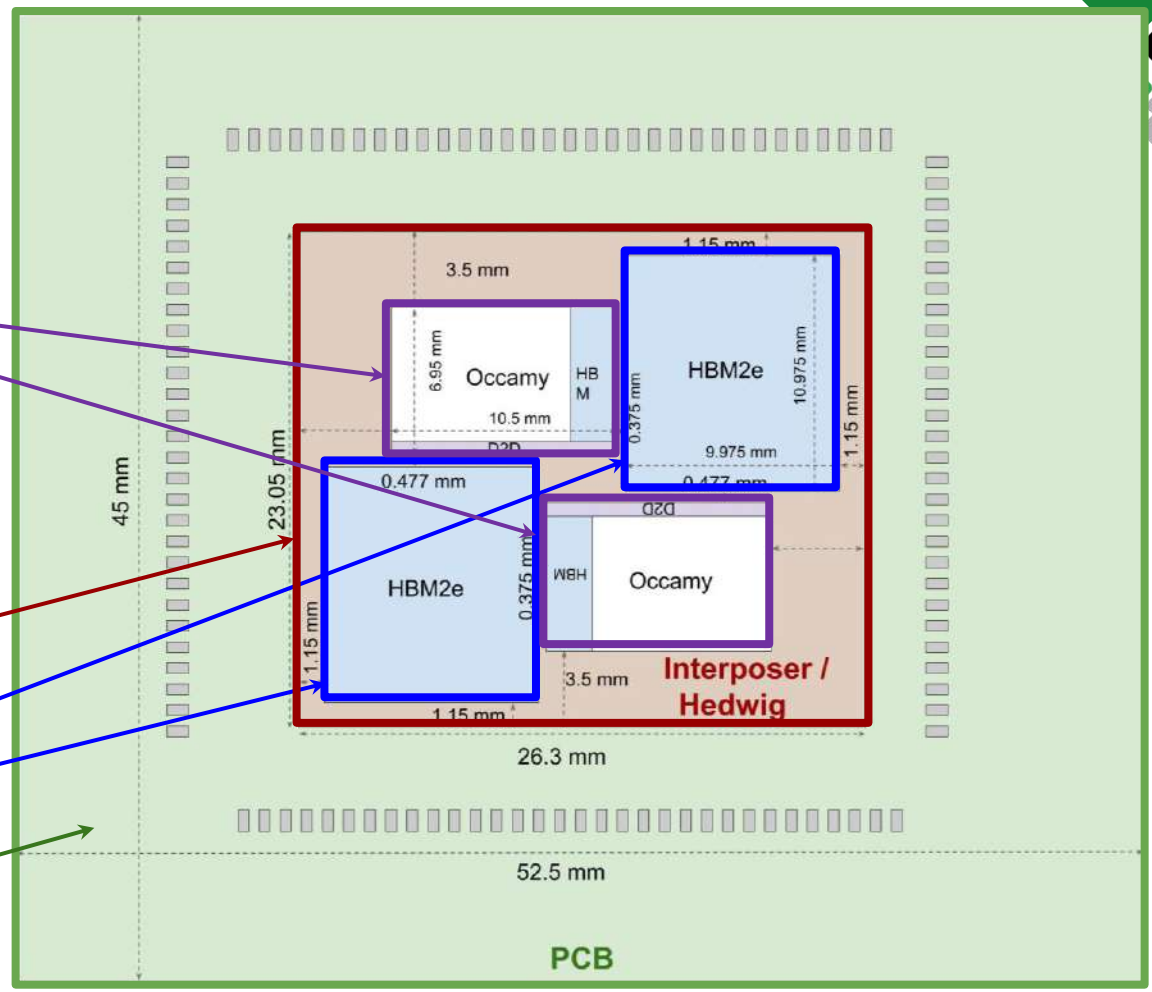
- Technology: 65nm, passive (only BEOL)
- Area: 26.3mm x 23.05mm

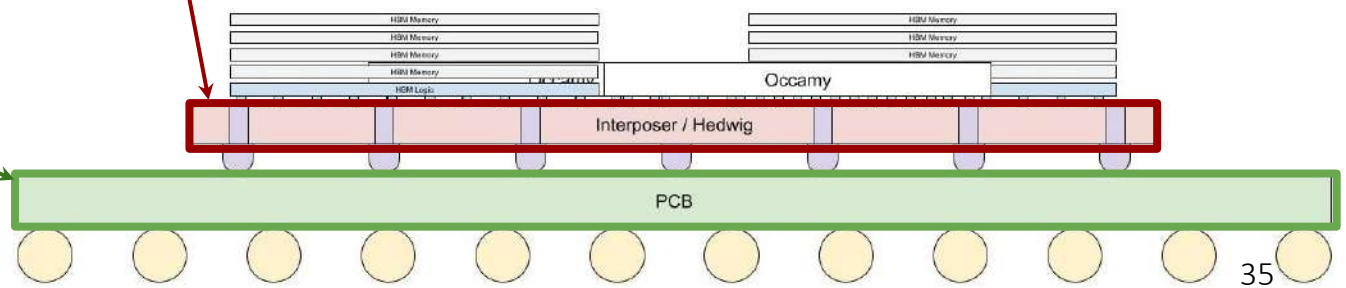
HBM2e:

- 16GB HBM2e (Micron)

Fan-out PCB:

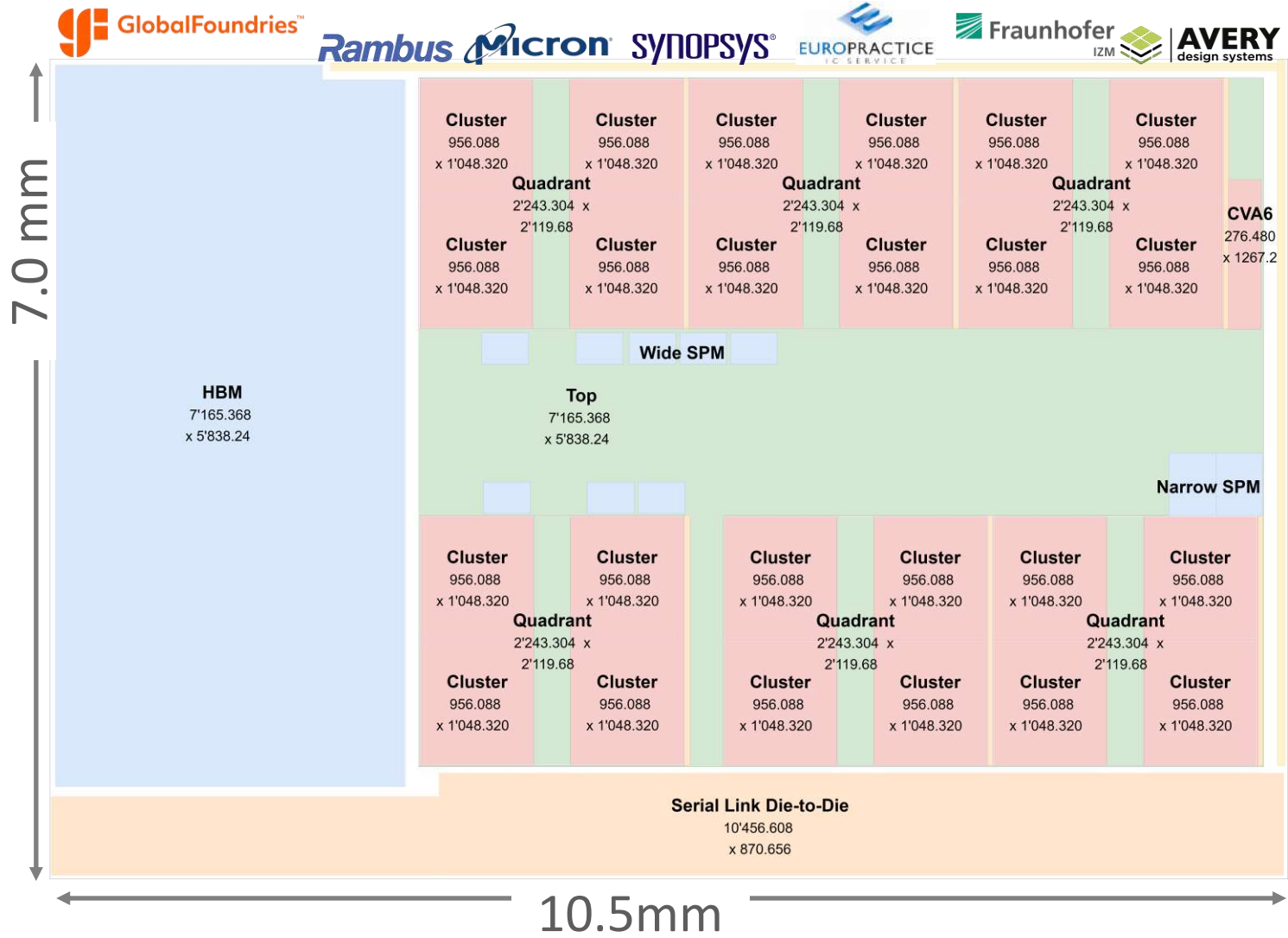
- RO4350B
- 52.5mm x 45mm





Occamy: Next-Generation HPC SoC

2.5D 2xOccamy+2HBM2 on interposer



RV-Based Snitch Cluster

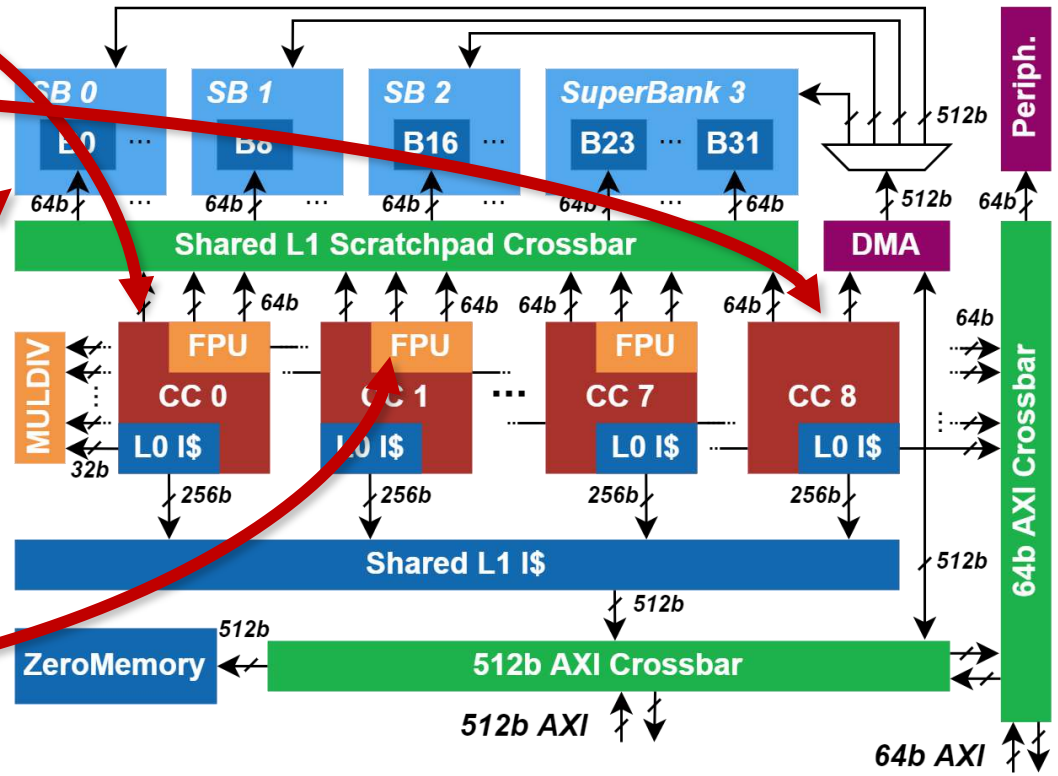
8 Snitch compute cores
- Single-stage, small Integer control core

9th Core: DMA
- 512 bit data interface
- Efficient data movement

128 kB TCDM
- Scratchpad for predictable memory accesses
- 32 Banks

Custom ISA extensions
- Xfrep, Xssr
- New: Xissr sparsity support

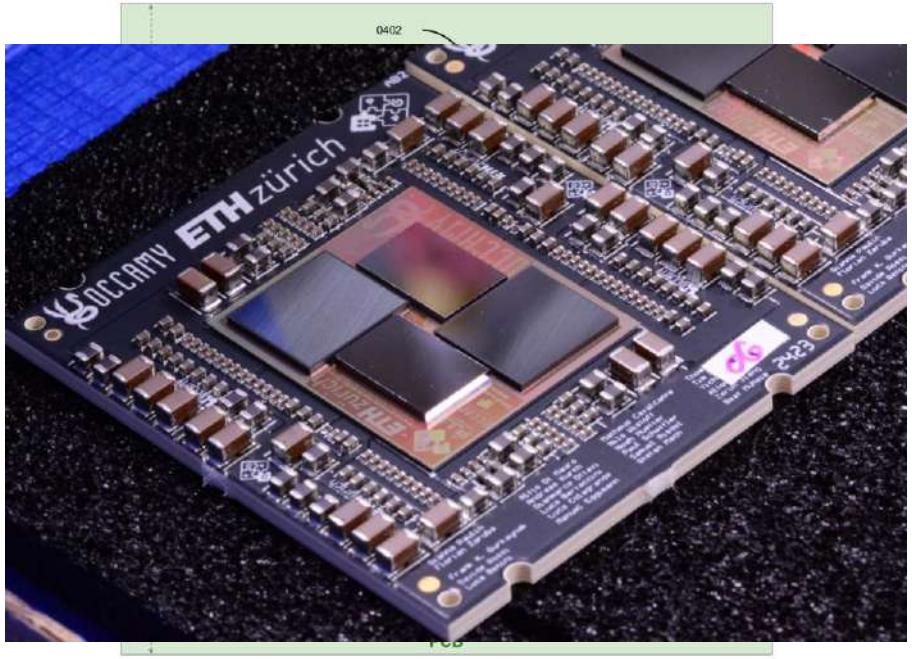
1 FPU per Snitch core
- Decoupled and heavily pipelined
- Multi-format FPU (+SIMD)
- New: Minifloat support + SDOTP

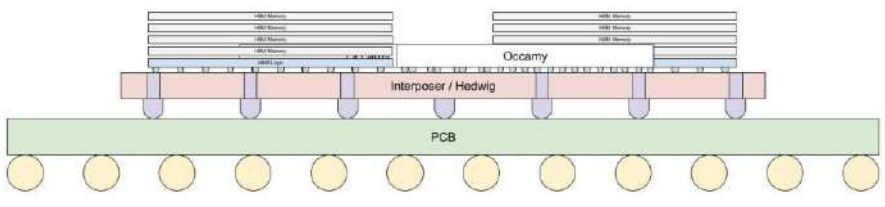


64GB/s in each direction

8GB/s in each direction

Occamy is On The Tester Now



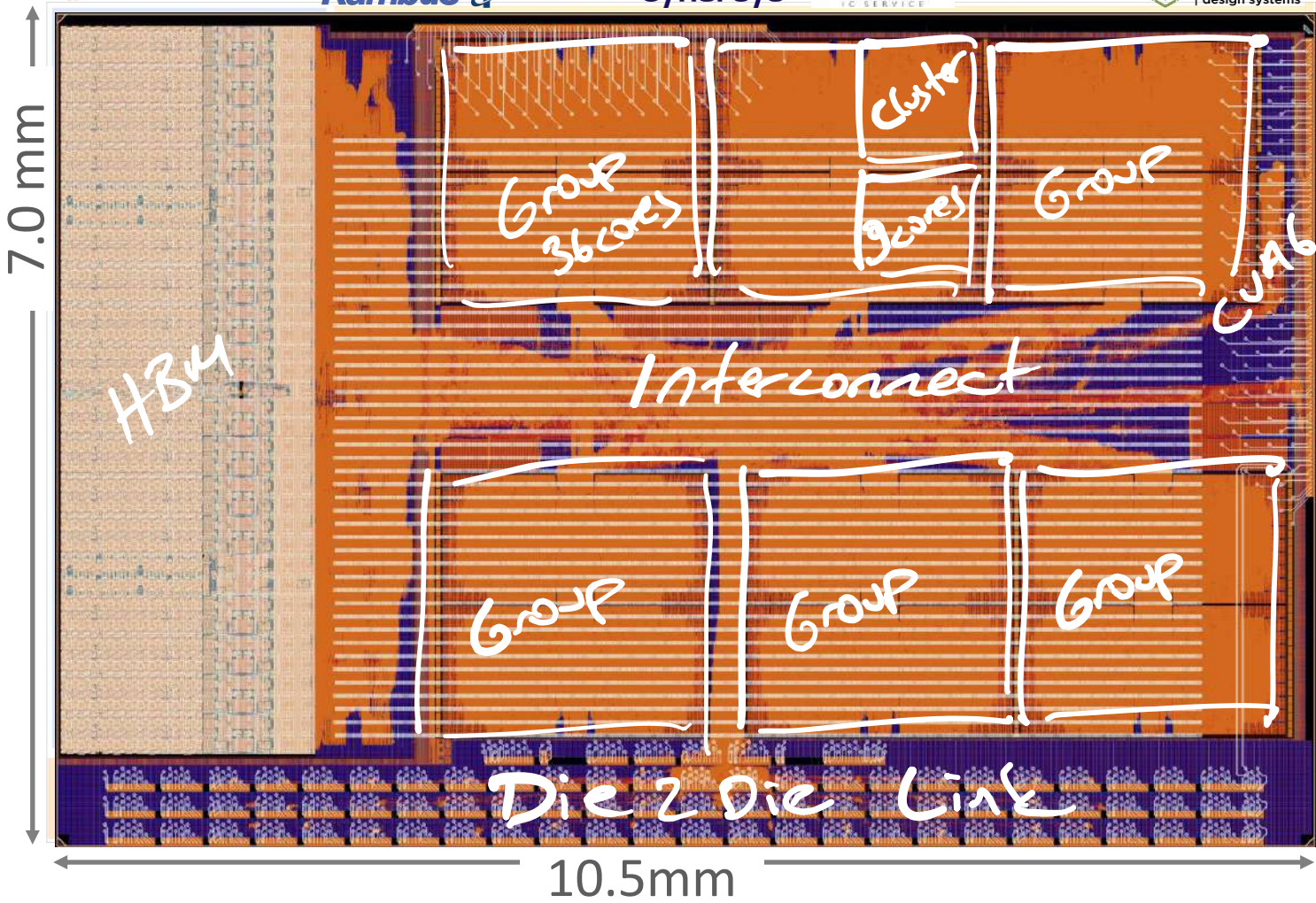


2.5D 2xOccamy+2HBM2 on interposer

GF12 @1GHz: FP64 384GFLOP/s, FP32,16,8

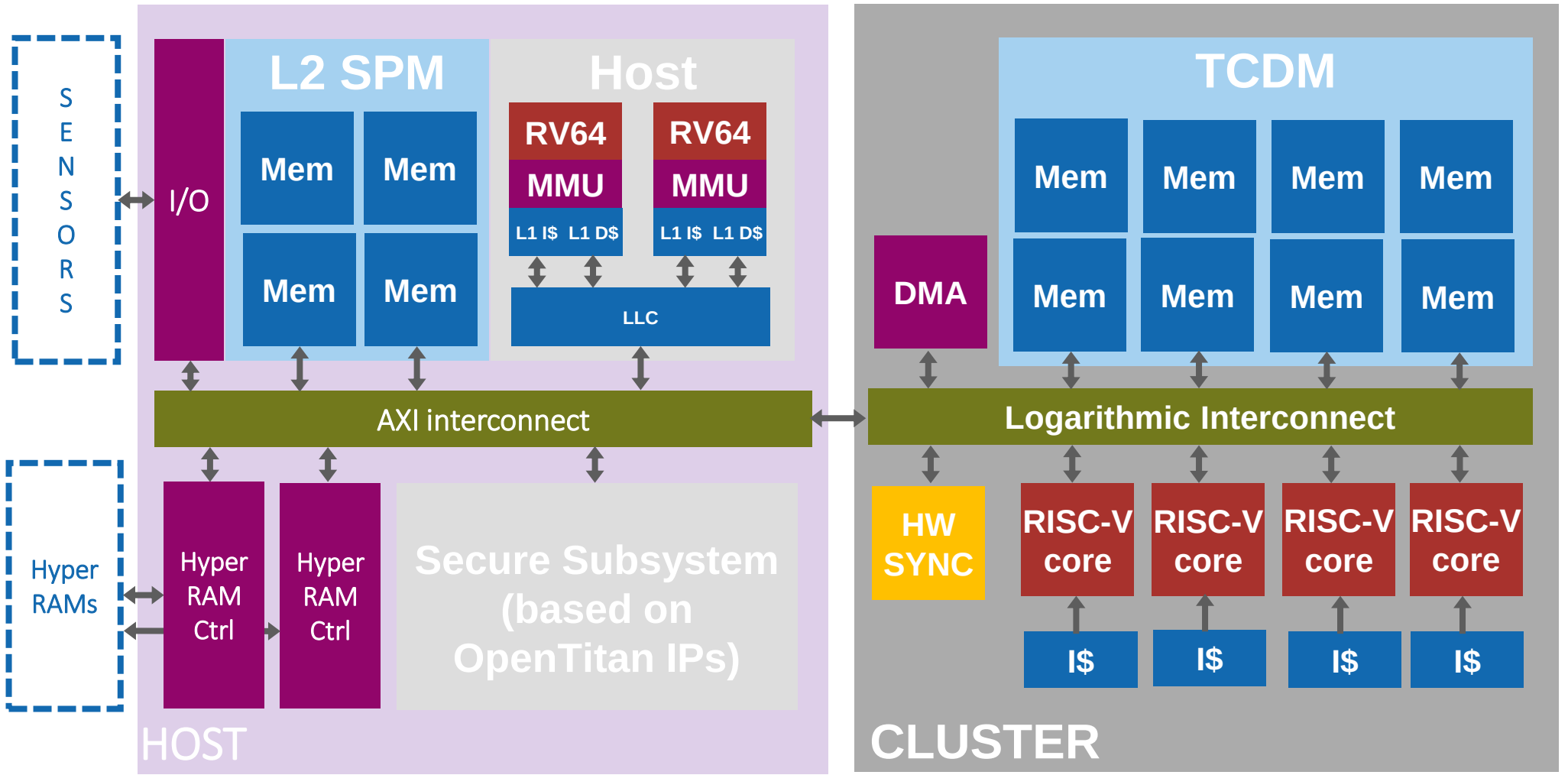
ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

GlobalFoundries Rambus Micron SYNOPSYS EUROPRACTICE IC SERVICE Fraunhofer IZM AVERY design systems



What's Next?

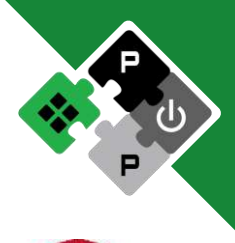
Al Saqr: Open-Research Platform for Secure UAVs



ETH zürich · ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

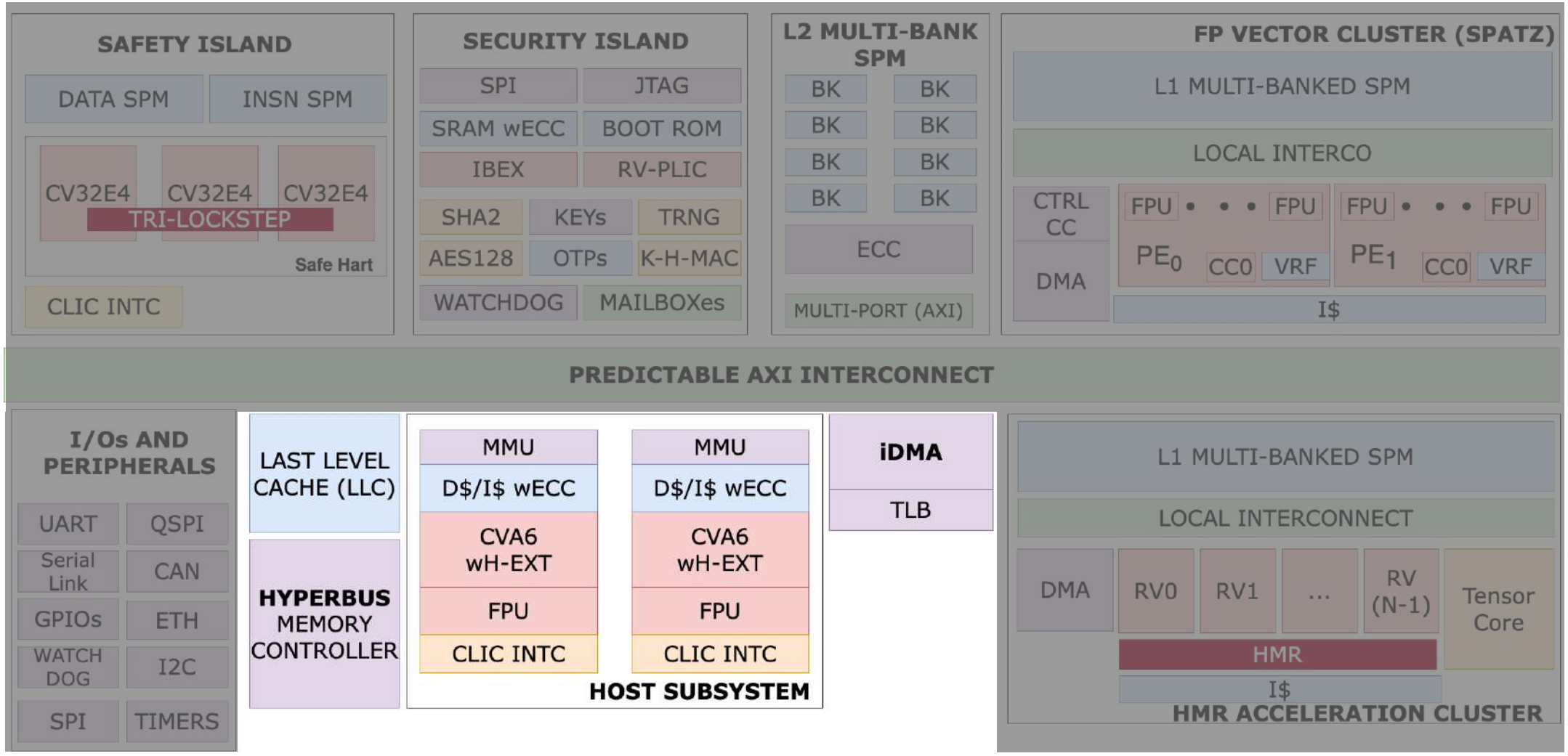
TAPE OUT: Dec 2023, GF22nm Technology

Carfield: Open-Research Platform for Automotive Systems

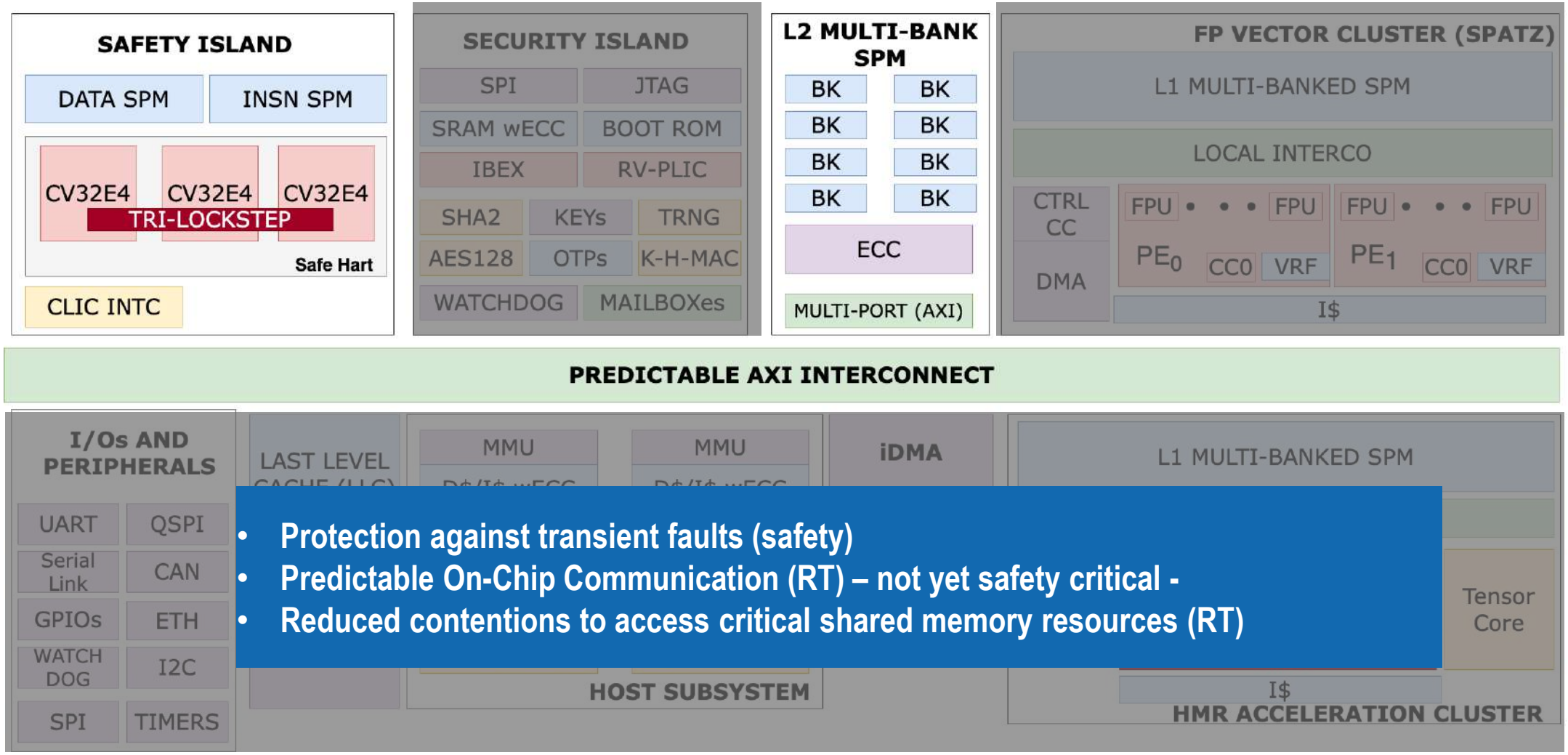


- Develop a precompetitive RISC-V Automotive SoC
  - Based on **fully open-source** HW and SW IPs
  - Scalable **and** configurable **architectural template** based on our PULP architectural ball-park
    - To adapt to different computing requirements
  - **Full SW stack** to address requirements of RISC-V based automotive applications
- Collaborative research roadmap for automotive-driven computing architectures
  - Functional safety, Hardware/Software Acceleration, Time-Predictability, Fast-Interrupts, HW-based Virtualization
- Benchmark RISC-V ecosystem and architectural solutions for automotive
  - Contributing to European interest to build an automotive reference platform around RISC-V
- Close the gap between RISC-V and ARM-powered solutions for automotive SoCs
- Fully aligned with EU Initiative (DG-connect) on RV for automotive (HW and SDV)!

Host-Domain for Low-Criticality Linux-Based Applications

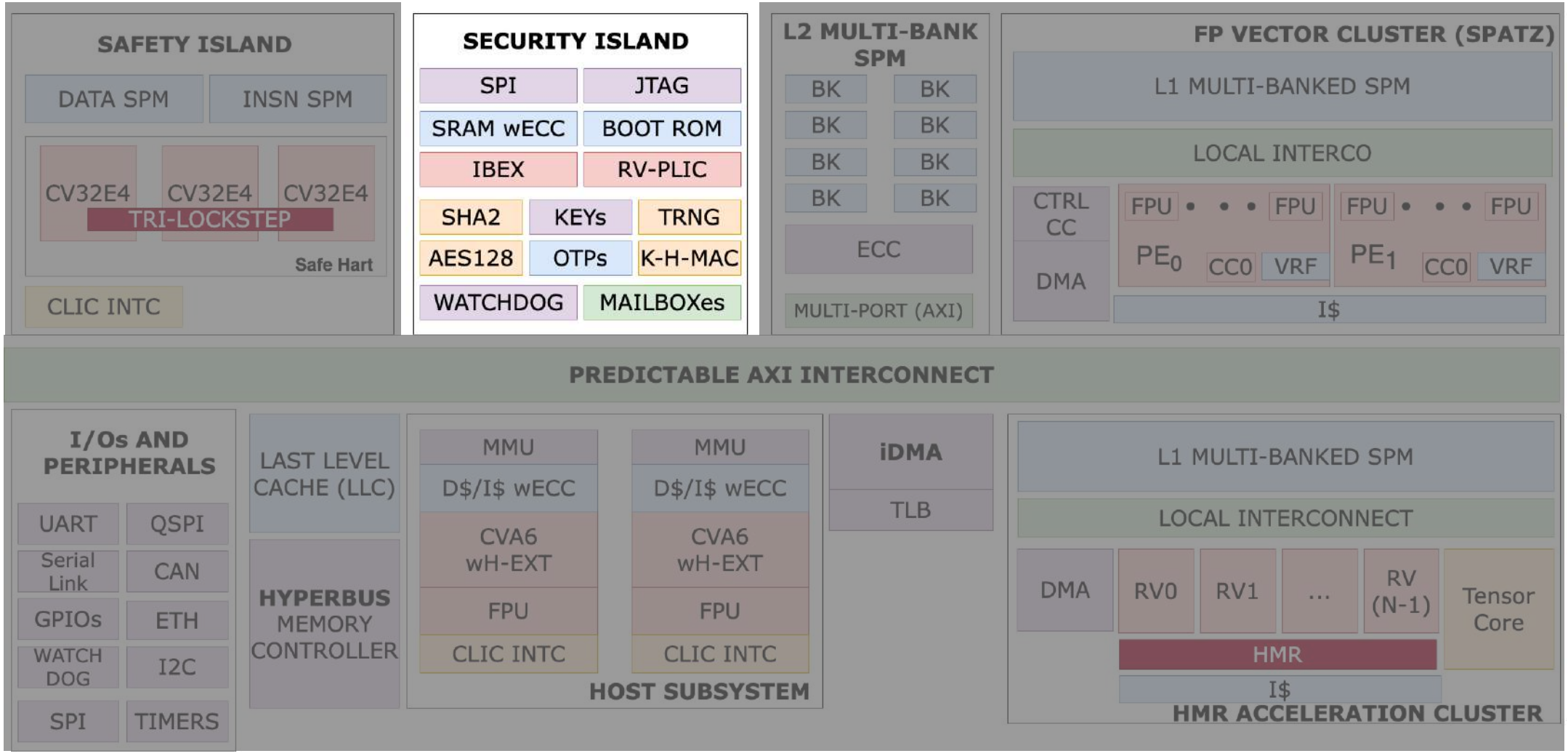


How Do We Handle Safety-Critical and Real-Time Tasks?

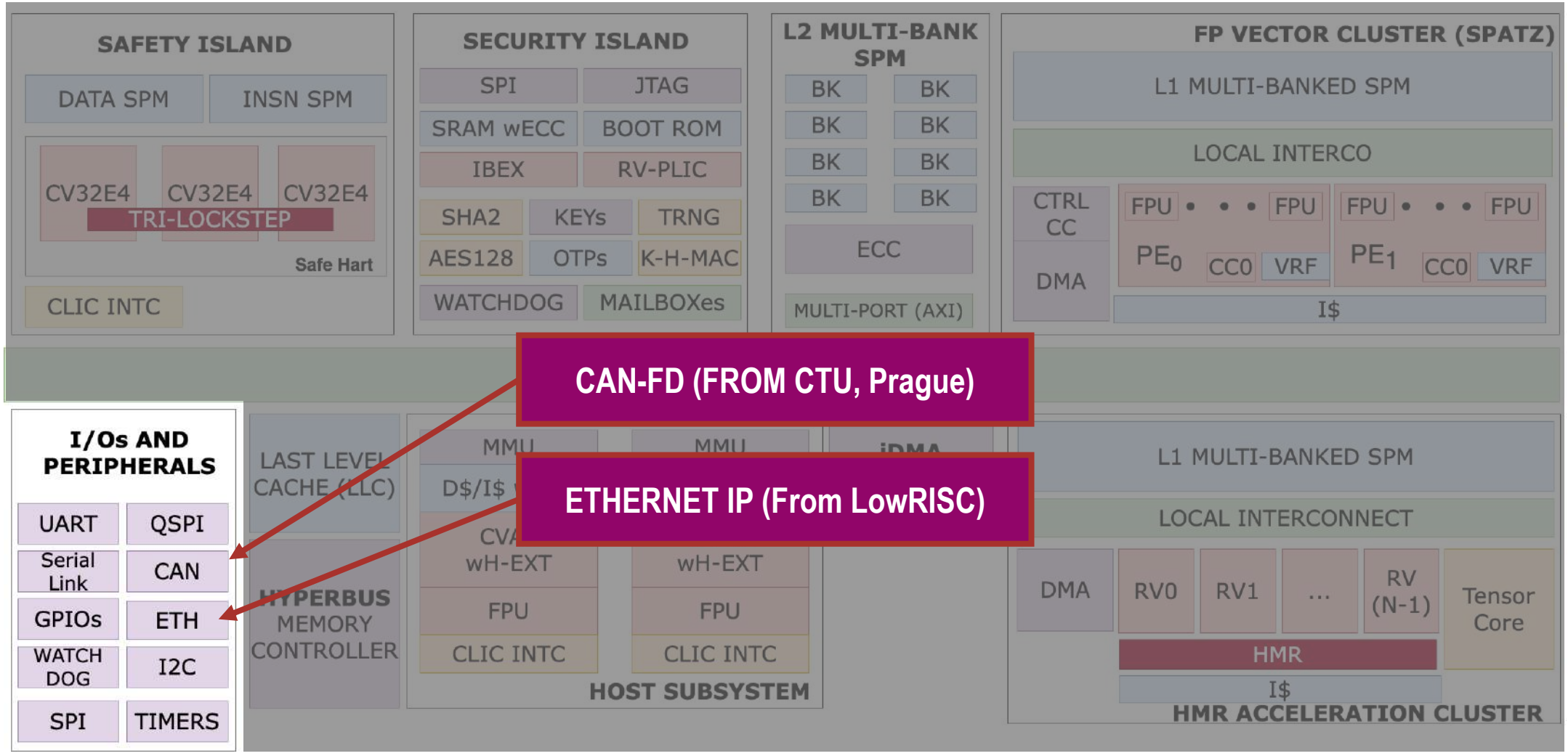


- Protection against transient faults (safety)
- Predictable On-Chip Communication (RT) – not yet safety critical -
- Reduced contentions to access critical shared memory resources (RT)

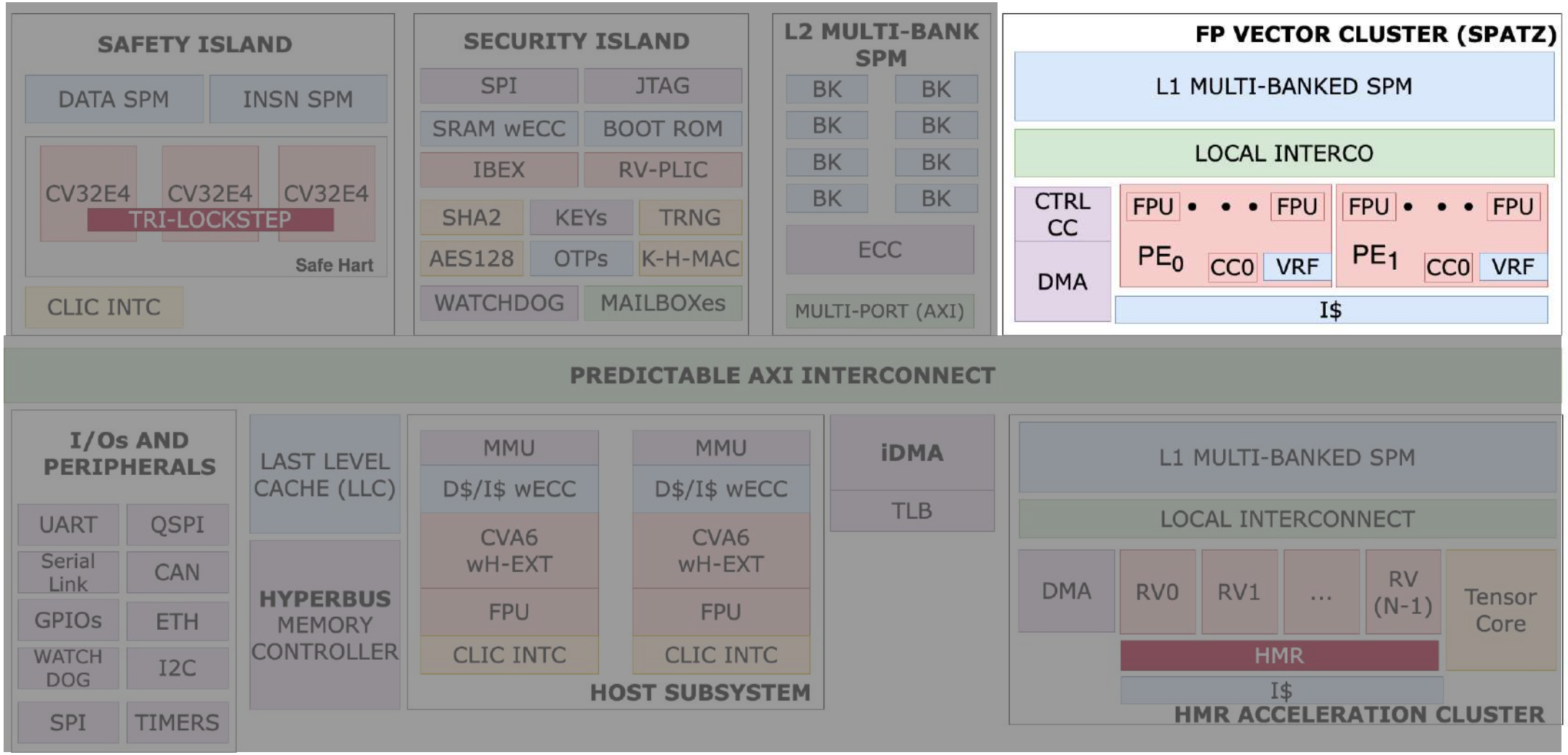
What About Security and Data Encryption/Decryption?



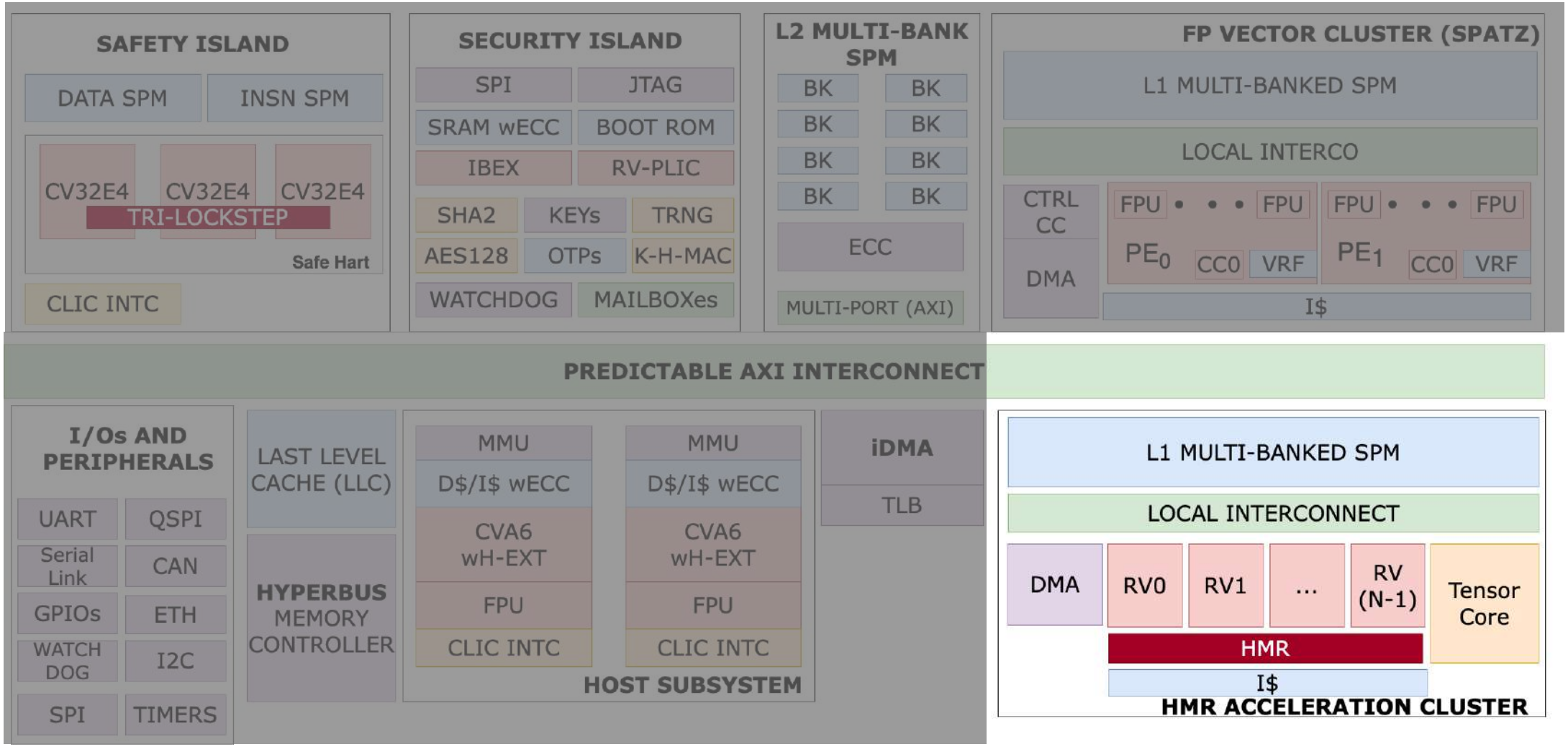
The I/O Communication



The Spatz Acceleration Cluster



The Inference Acceleration Cluster



PULP
Parallel Ultra Low Power

Questions?

http://pulp-platform.org

@pulp_platform