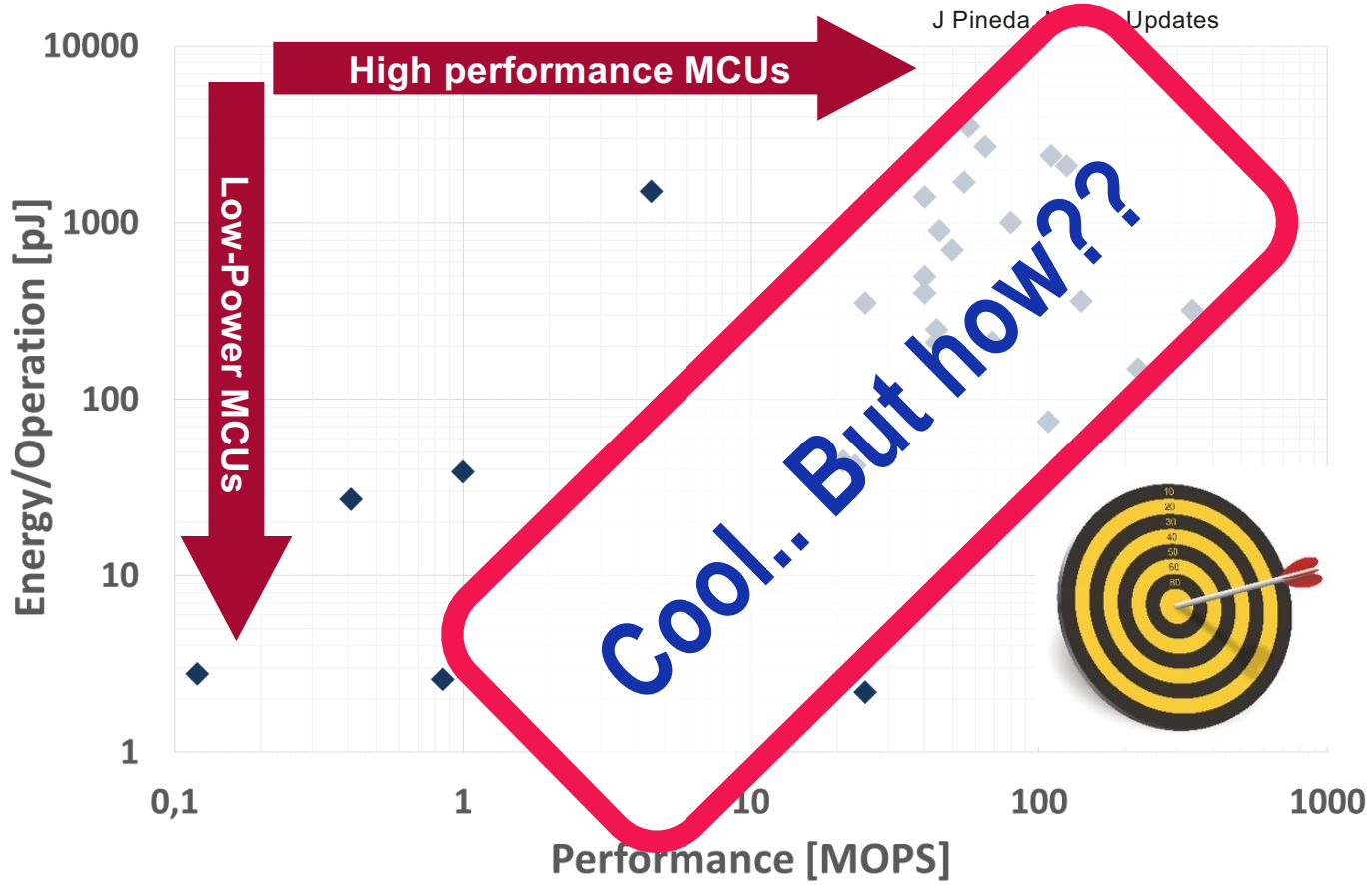
Multicore systems for HPC and edge AI acceleration

Yvan Tortorella

yvan.tortorella@chips.it

EFCL Winter School 2026

Energy efficiency @ GOPS is THE Challenge

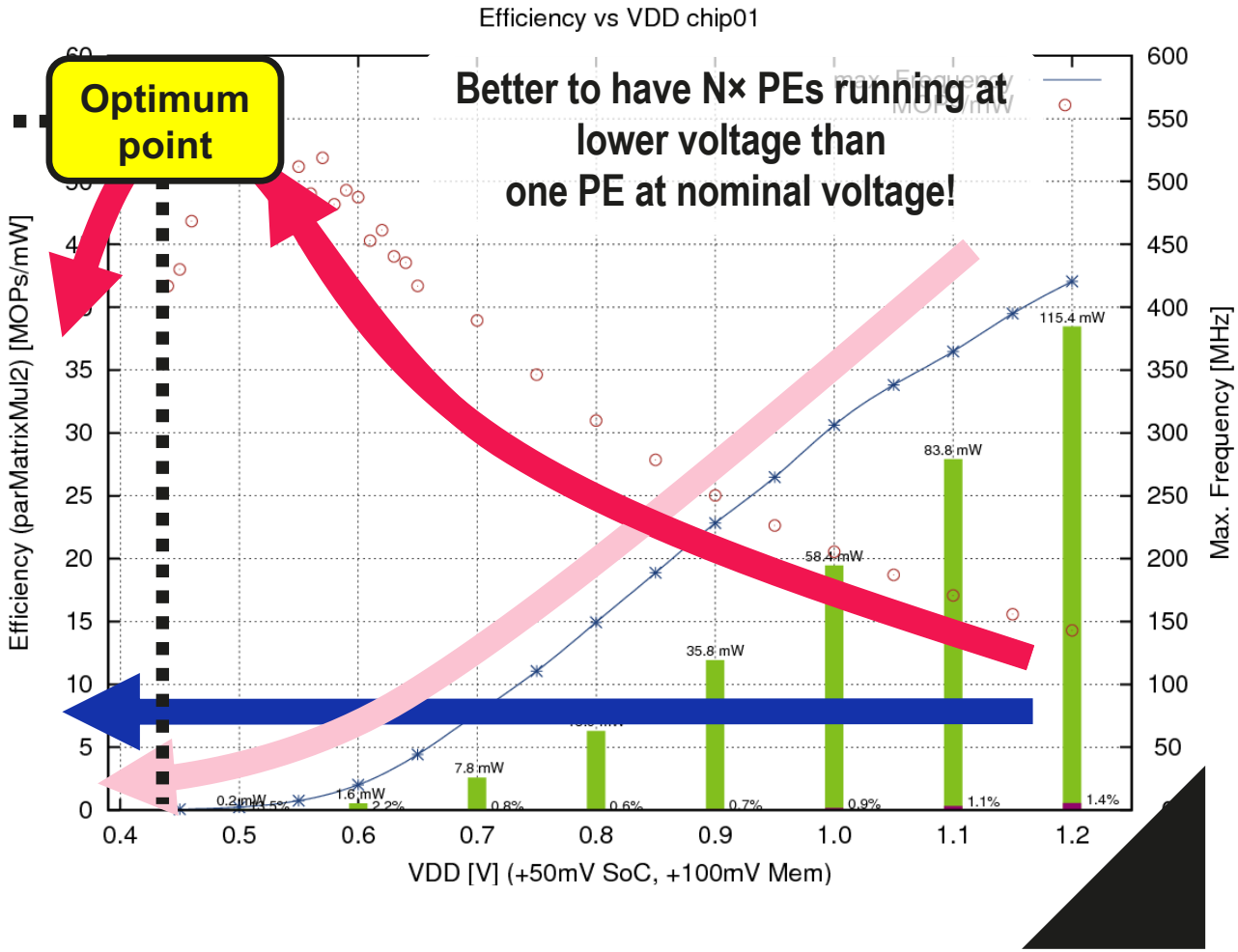


ML & Parallel, Near-threshold: a Marriage Made in Heaven

Chips-IT

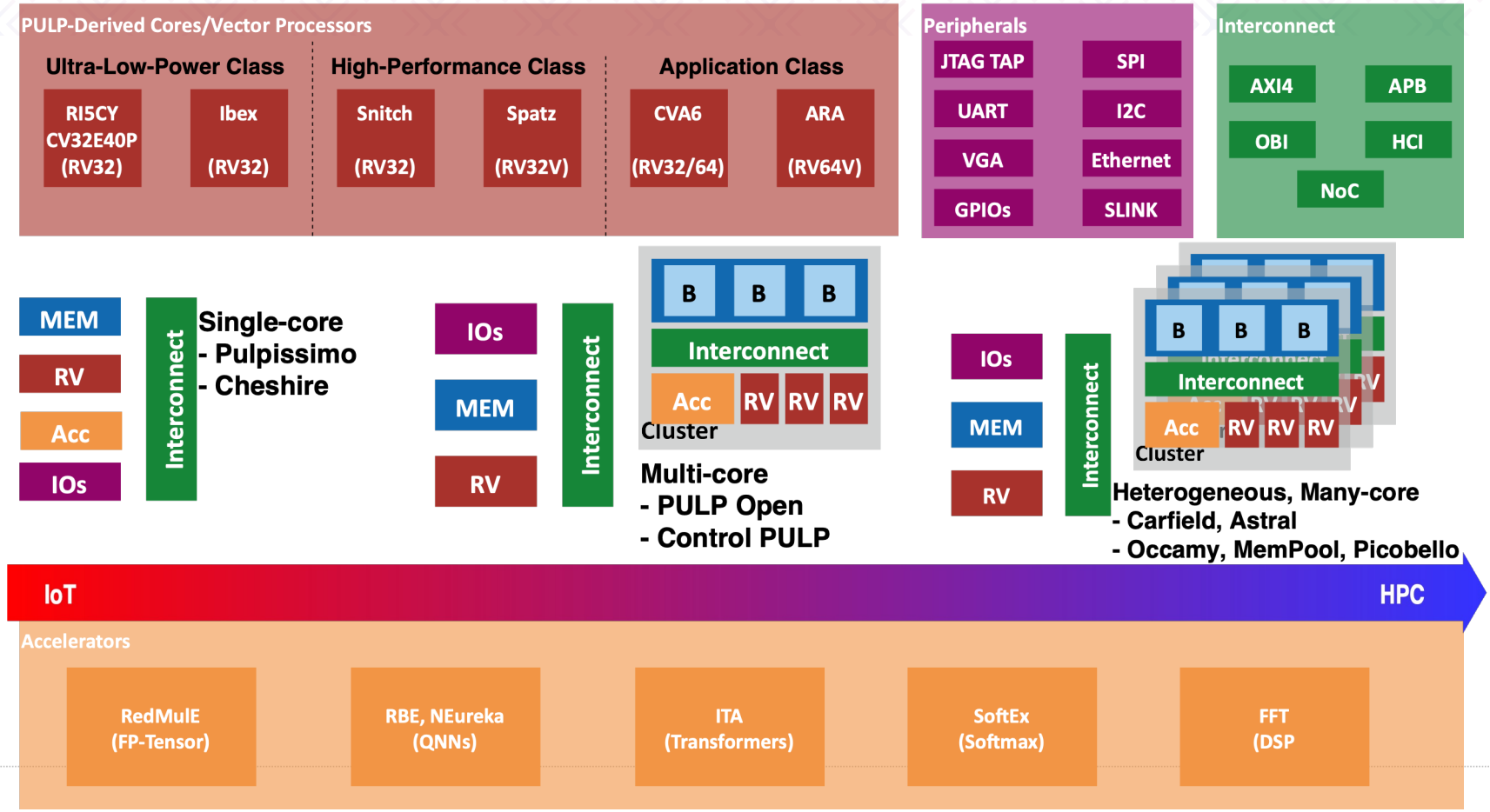
- As VDD decreases, operating speed decreases
- However efficiency increases→ more work done per Joule
- Until leakage effects start to dominate
- Put more units in parallel to get performance up and keep them busy with a parallel workload

ML is massively parallel and scales well



PULP Ecosystem

Chips-IT



Not All Programs Are Created Equal

Chips-IT

- Processors can do two kinds of useful work:

Decide (jump through program parts)

- 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 just ploughs through, few decisions done based on the computed values
  - **"Data-Oblivious Algorithms" (ML, or better DNNs are so!)**

Not All Programs Are Created Equal

Chips-IT

- Processors can do two kinds of useful work:

Decide (jump through program parts)

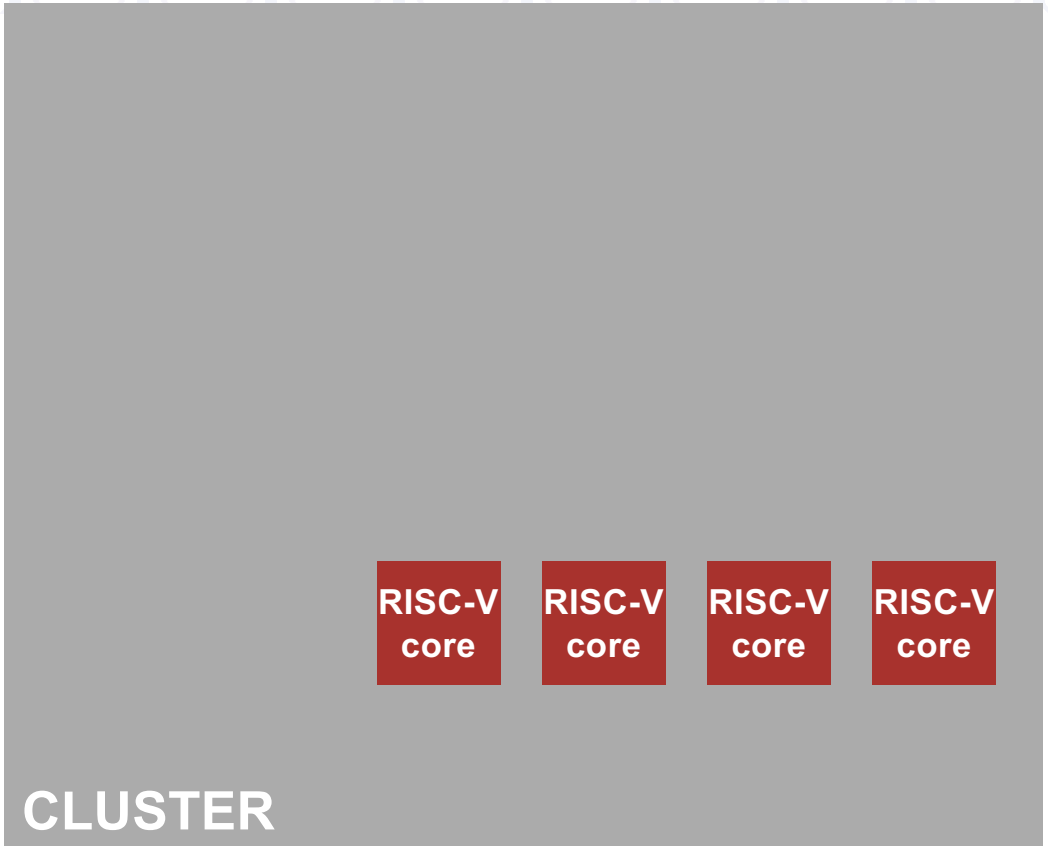
- Modulate flow of **instructions**
- **Smarts**:
  - Don't work too much
  - Be clever about the battles you pick (e.g., search in a database)
  - Lots of decisions
    Little number crunching

Compute (plough through numbers)

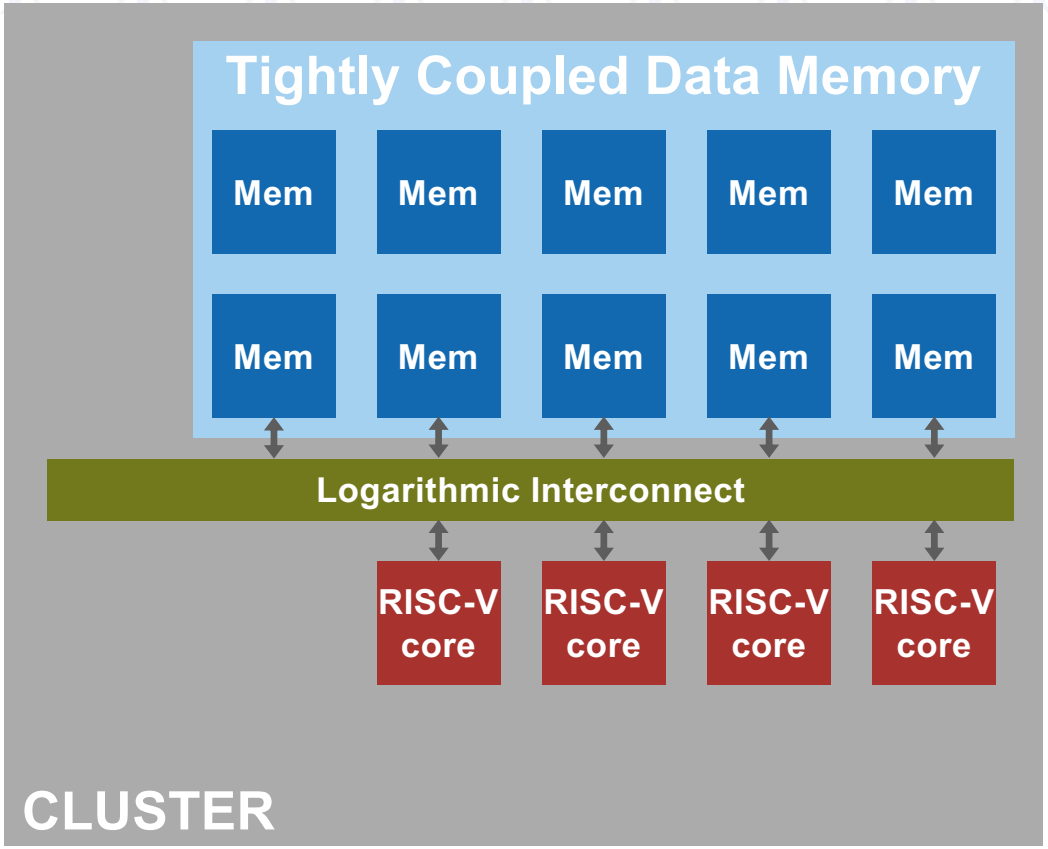
- Modulate flow of **data**
- **Diligence**:
  - Don't think too much
  - Just plough through the data (e.g., machine learning)
  - Few decisions
    Lots of number crunching

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

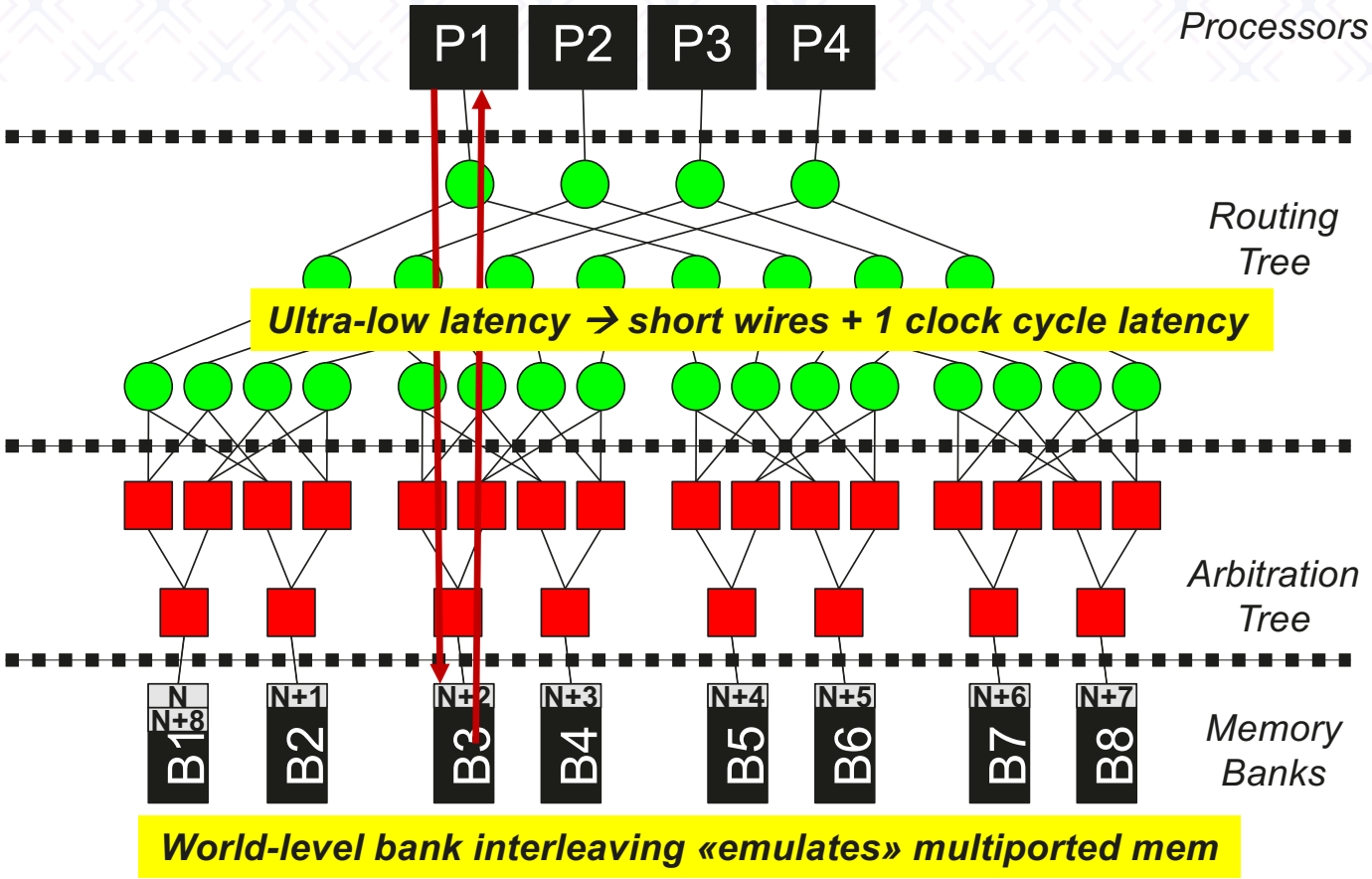
Multiple RI5CY Cores (1-16)



Low-Latency Shared TCDM

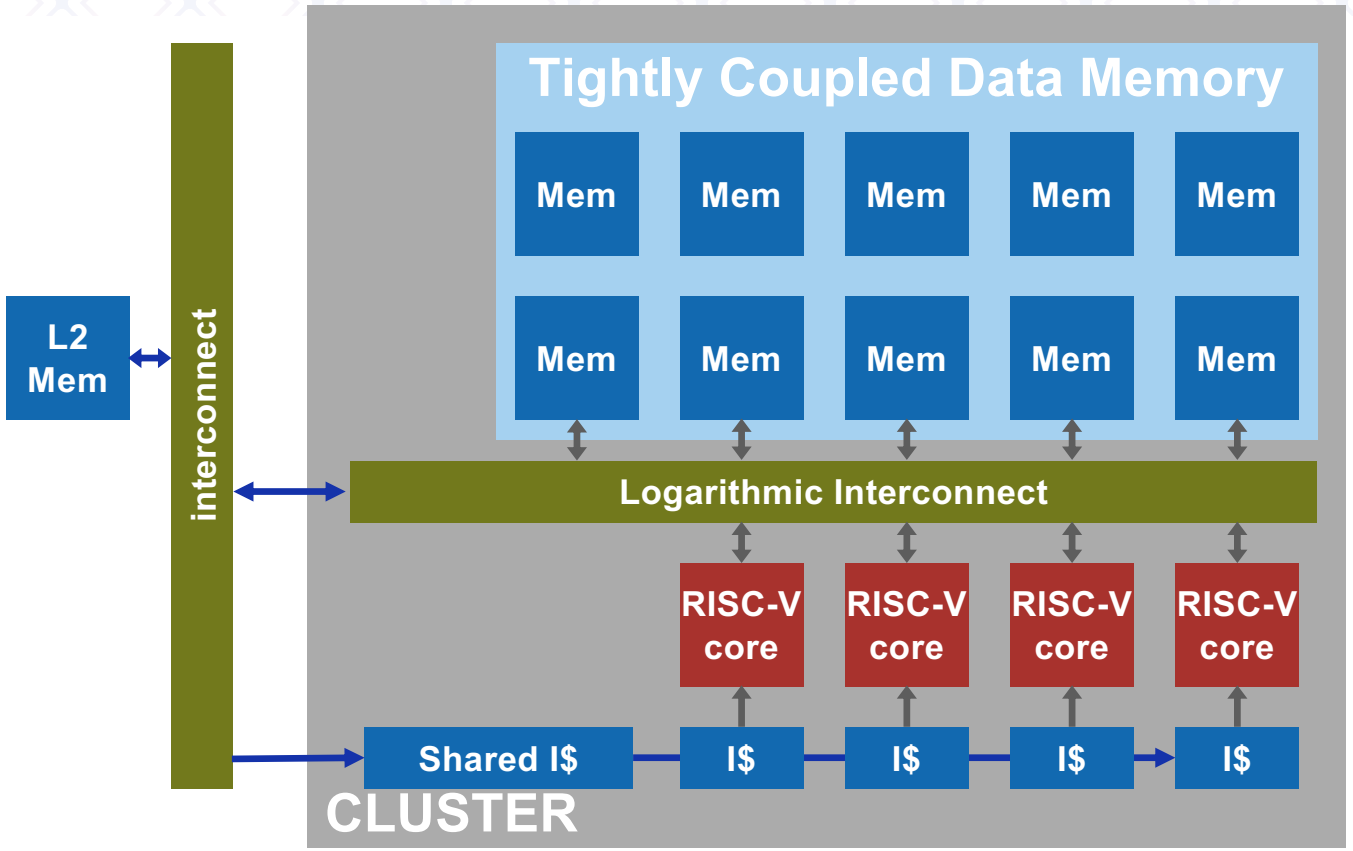


High speed single clock logarithmic interconnect

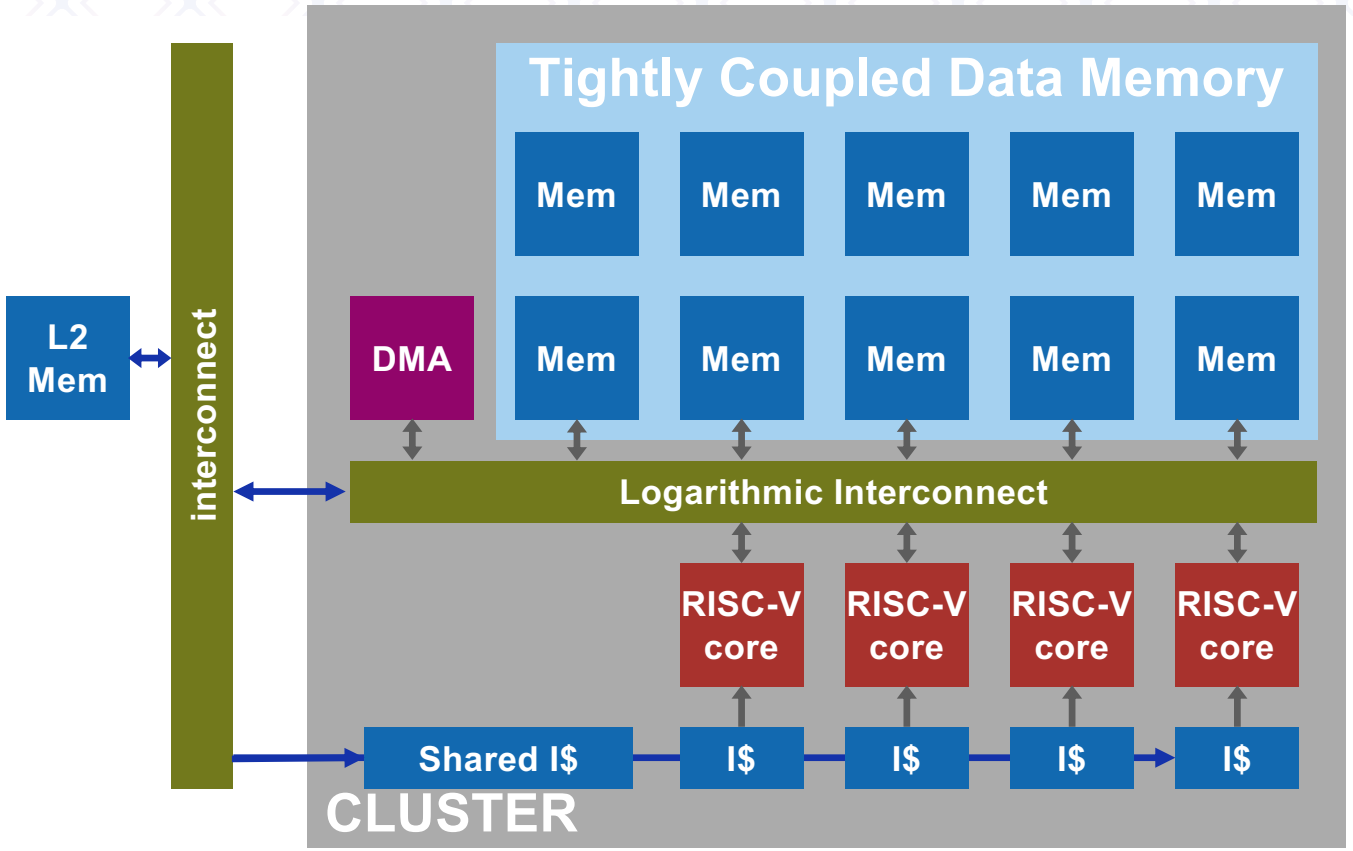


A. Rahimi, I. Loi, M. R. Kakoee and L. Benini, "A fully-synthesizable single-cycle interconnection network for Shared-L1 processor clusters," *2011 Design, Automation & Test in Europe*, 2011, pp. 1-6.

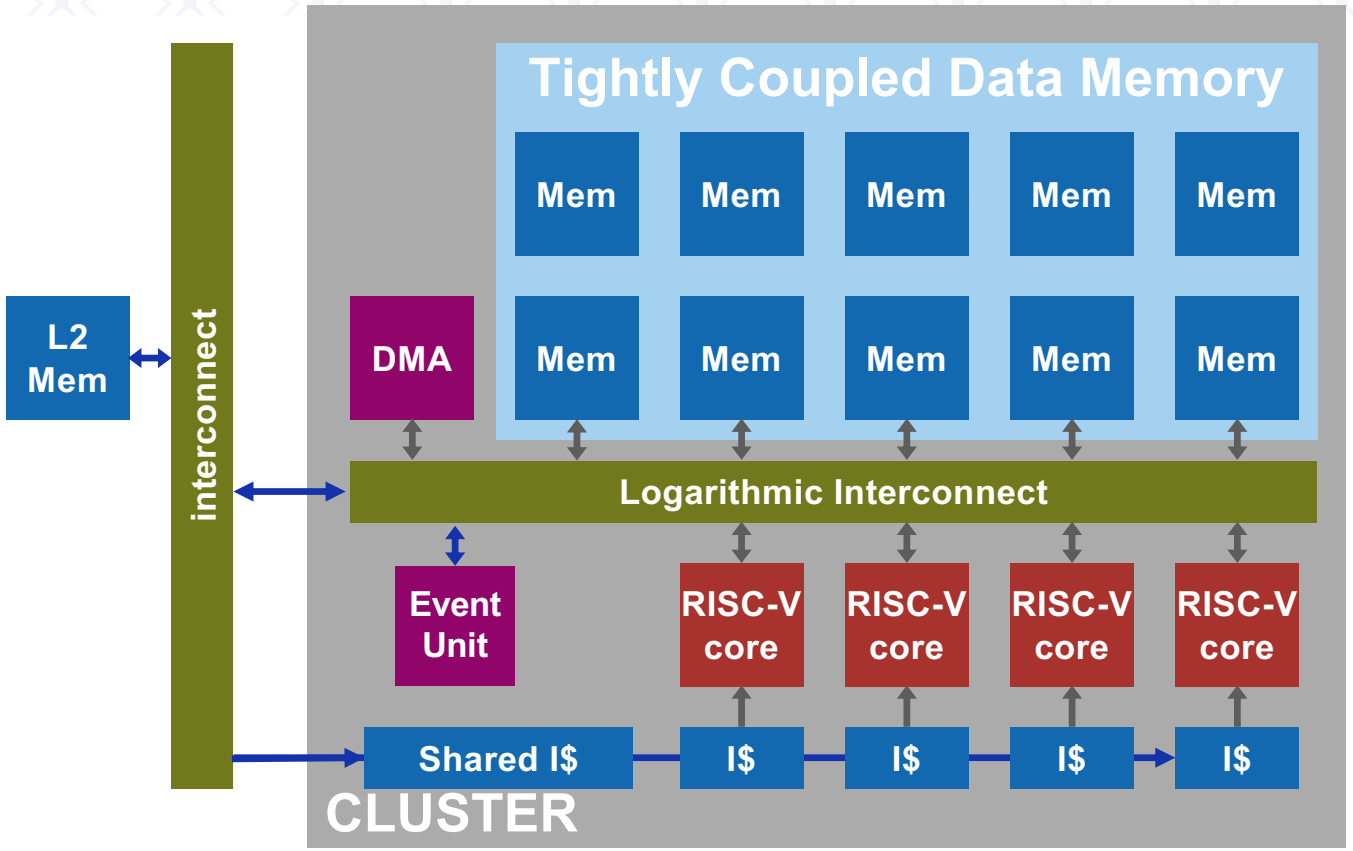
Shared instruction cache with private “loop buffer”



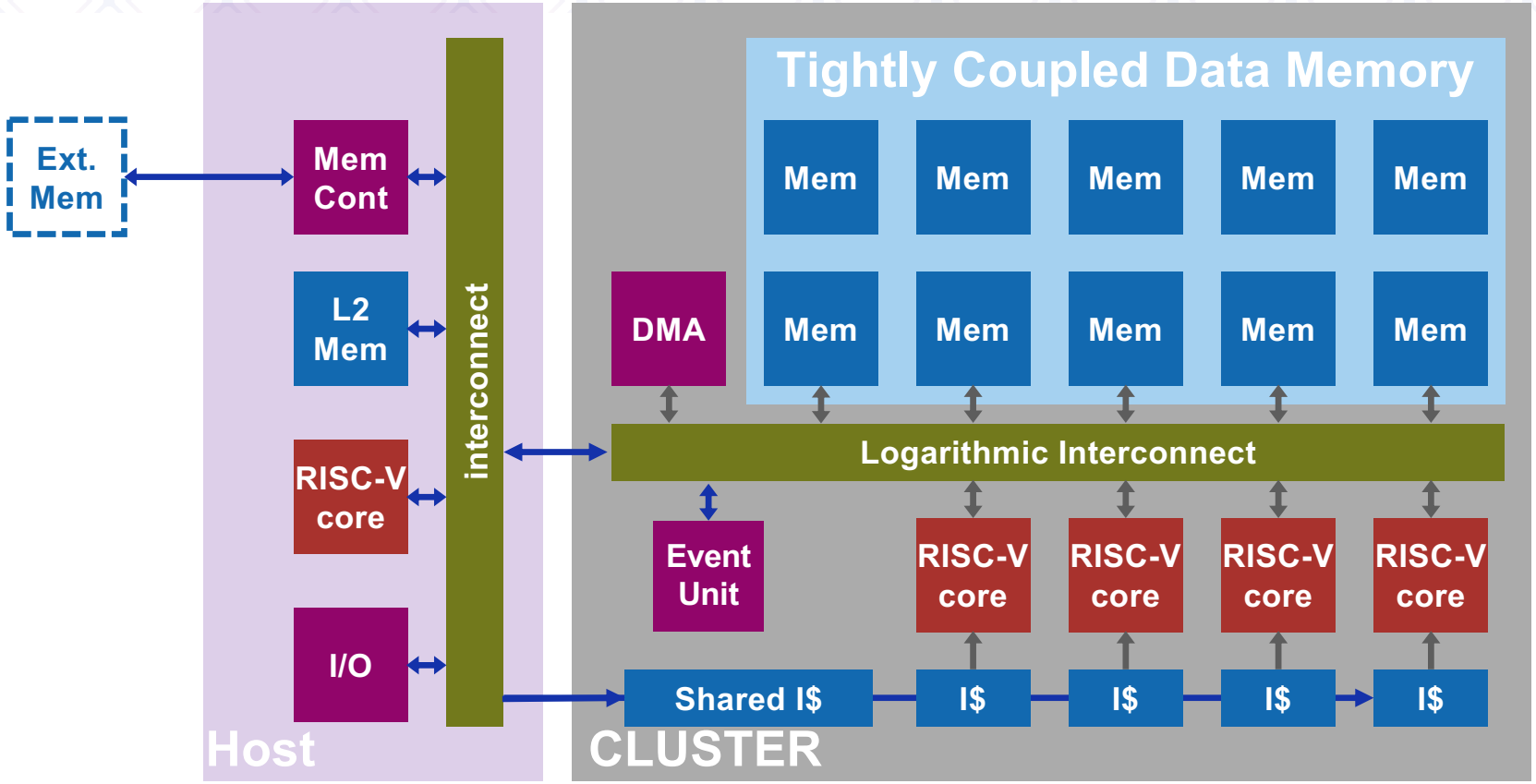
DMA for data transfers from/to L2



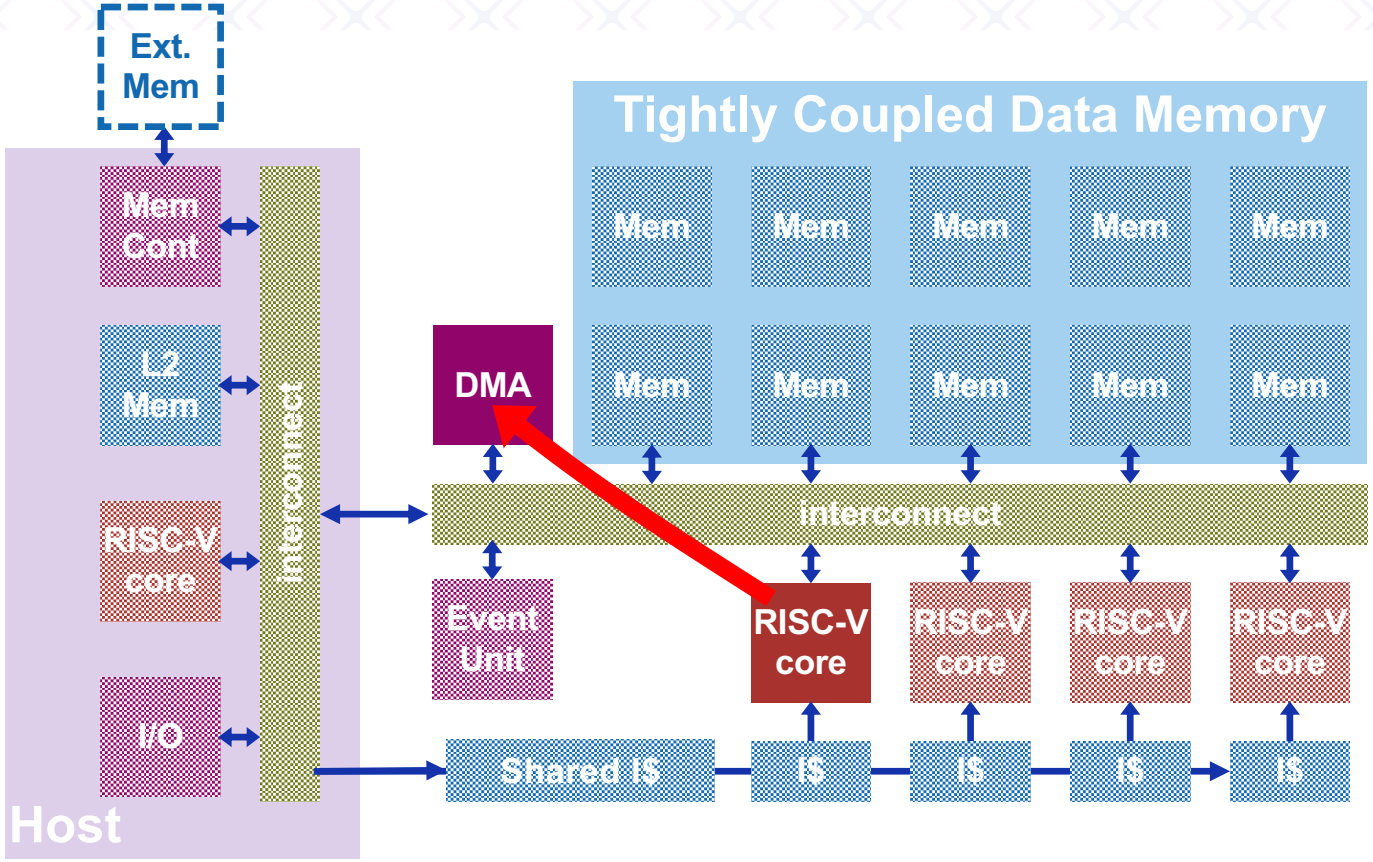
Event Unit for synchronization on events



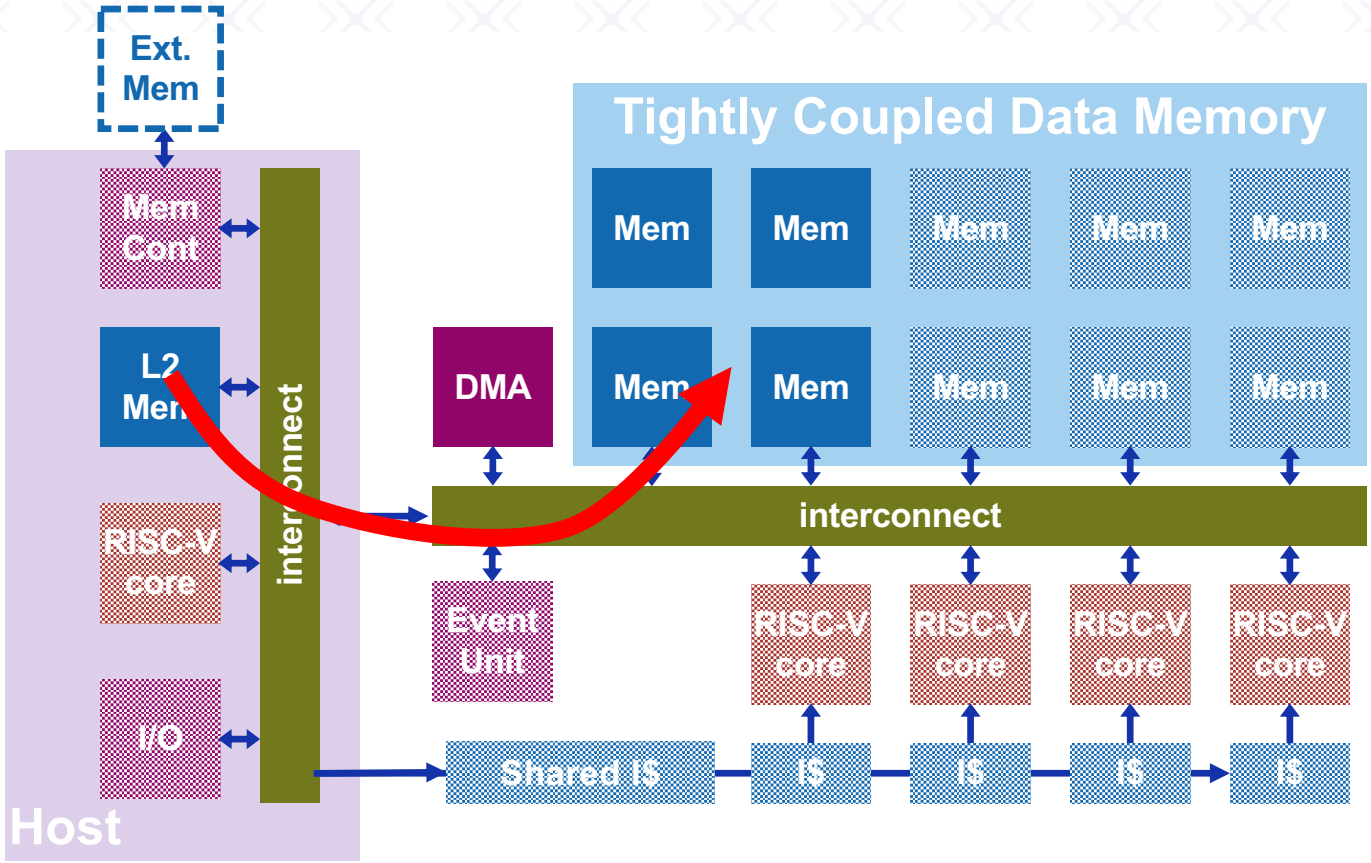
An additional Host controller is used for IO



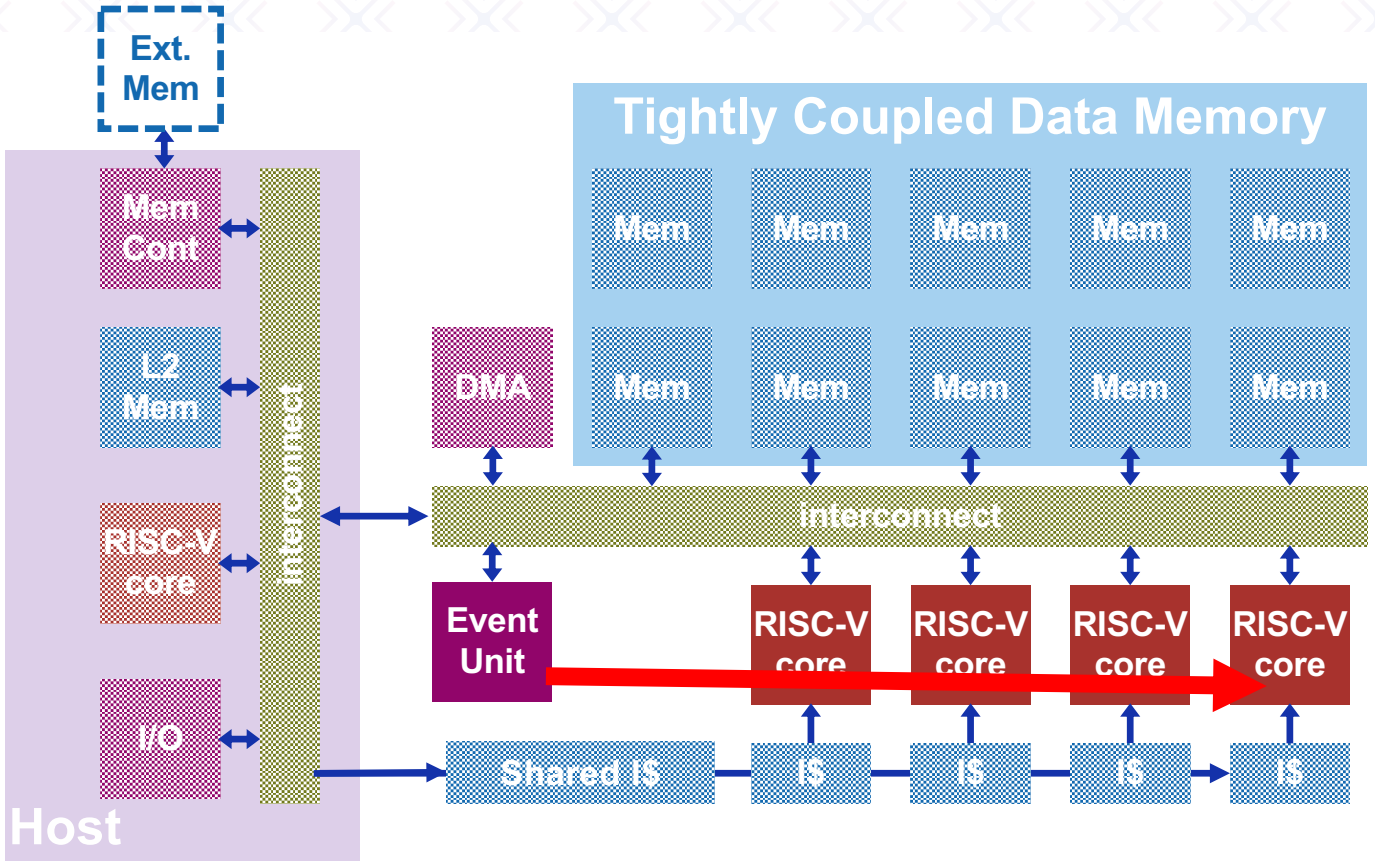
How do we work: Initiate a DMA transfer



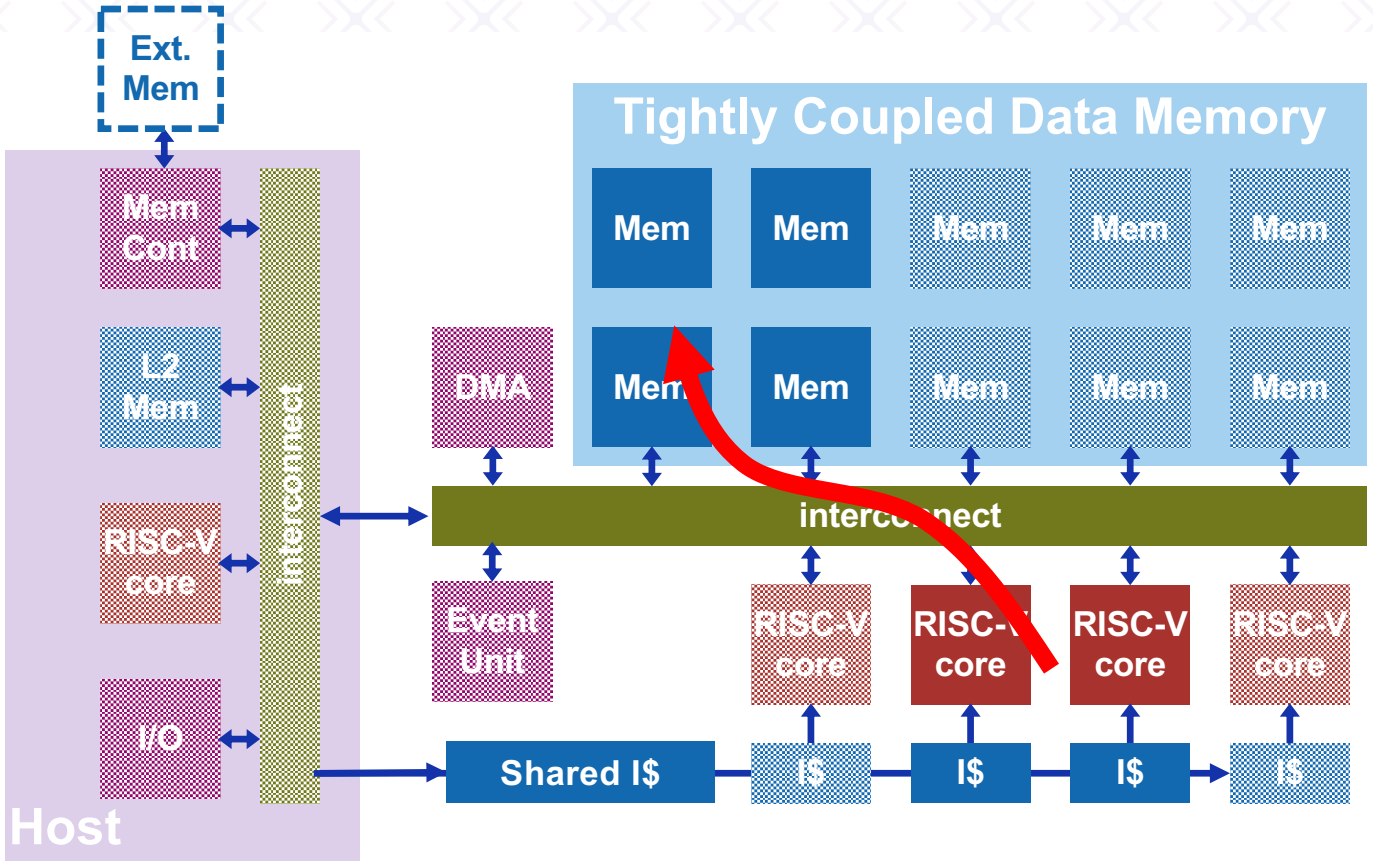
Data copied from L2 into TCDM



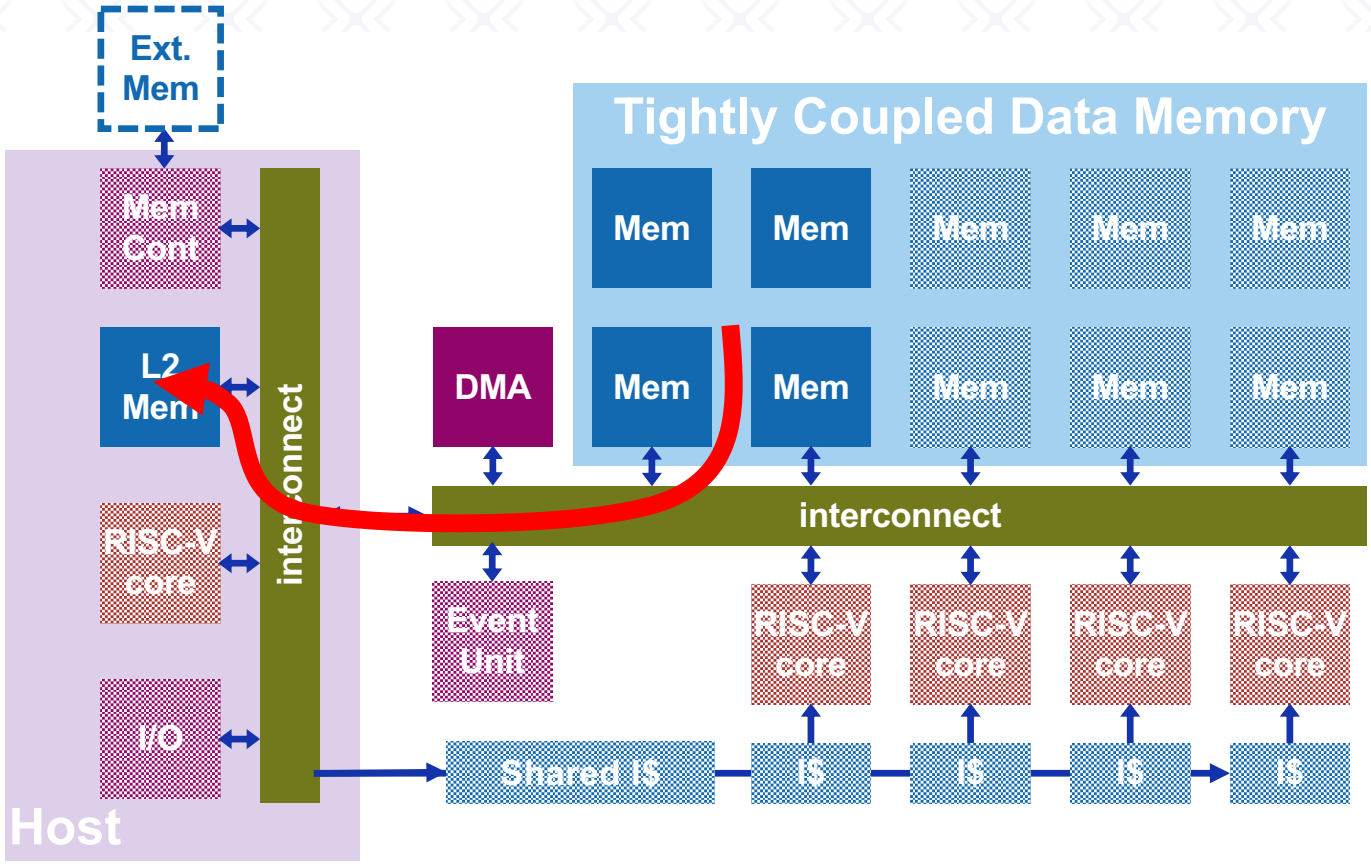
Once data is transferred, event unit notifies cores



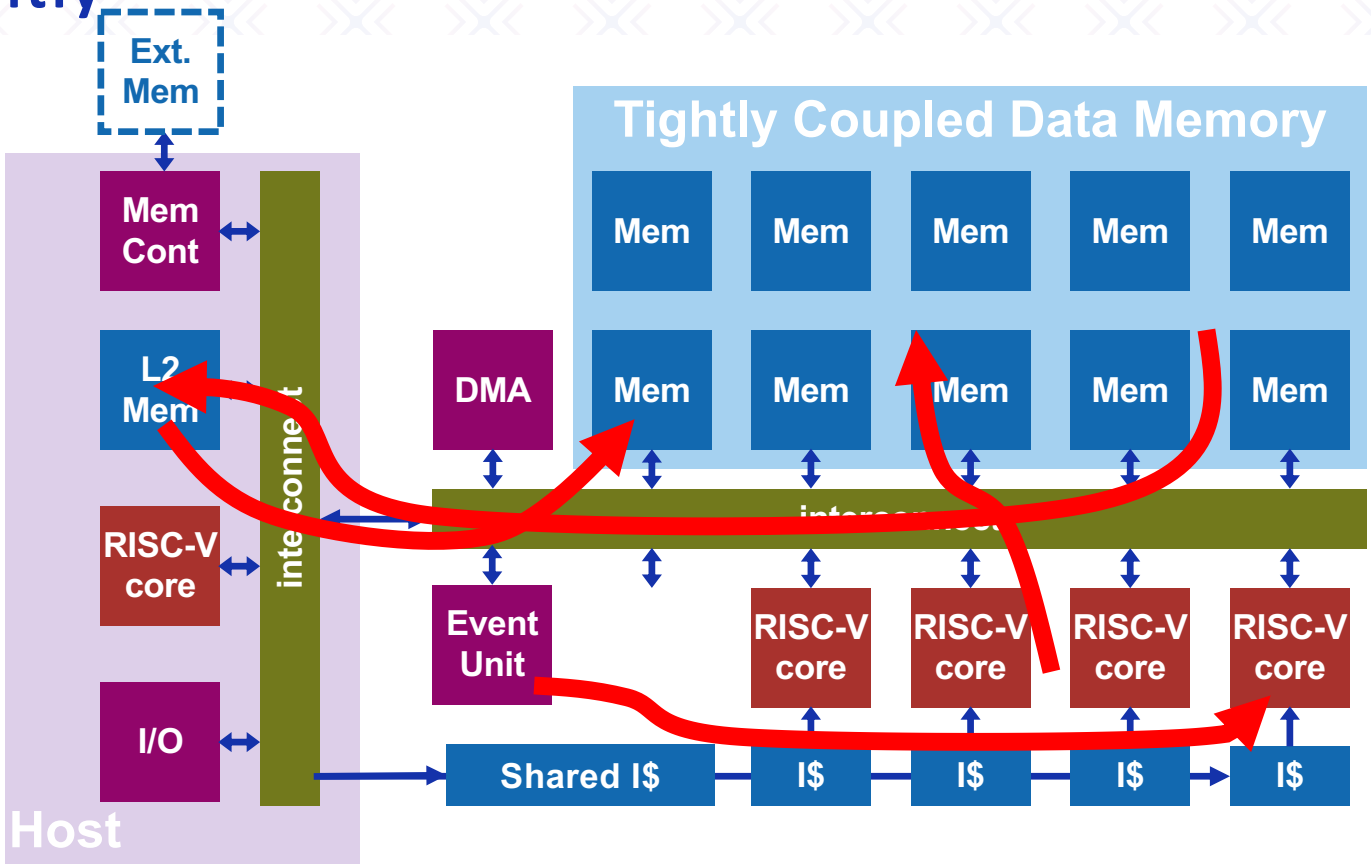
Cores can work on the data transferred



Once our work is done, DMA copies data back



During normal operation all of these occur concurrently



Not All Programs Are Created Equal

Chips-IT

- Processors can do two kinds of useful work:

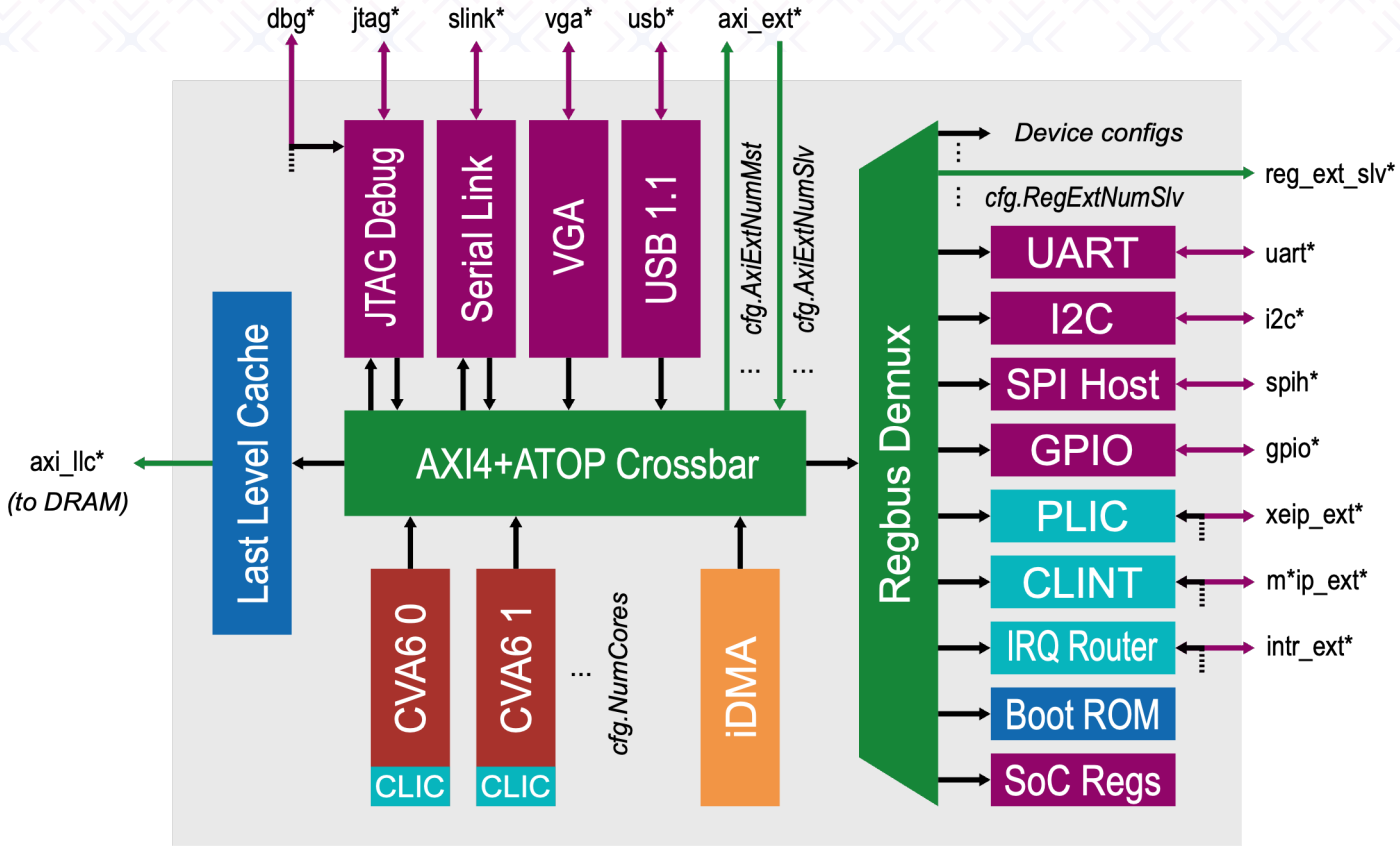
Decide (jump through program parts)

- 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

Cheshire: Not only extreme edge IoT

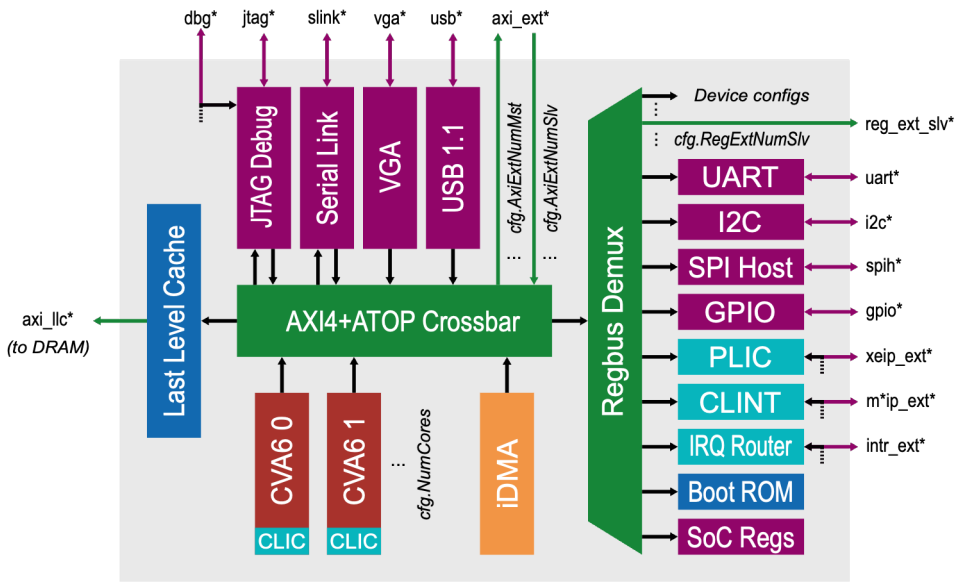


Cheshire: Not only extreme edge IoT

Chips-IT

Modular Platform

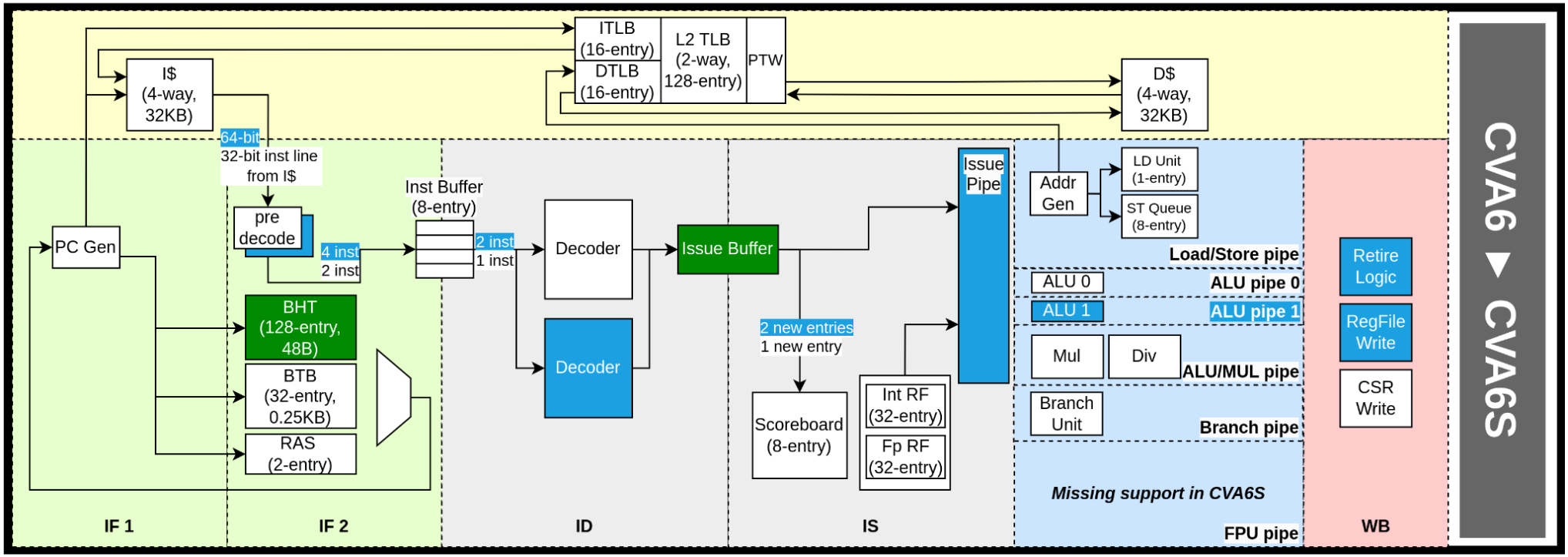
- Based on 64-bit Application-class processor
- Linux-capable and configurable for a wide range of applciations
- Standard interrupt, debug, and memory interfaces
- Access to external DRAMs (HyperRAMs)



CVA6(S+): Application-class 64-bit RISC-V core

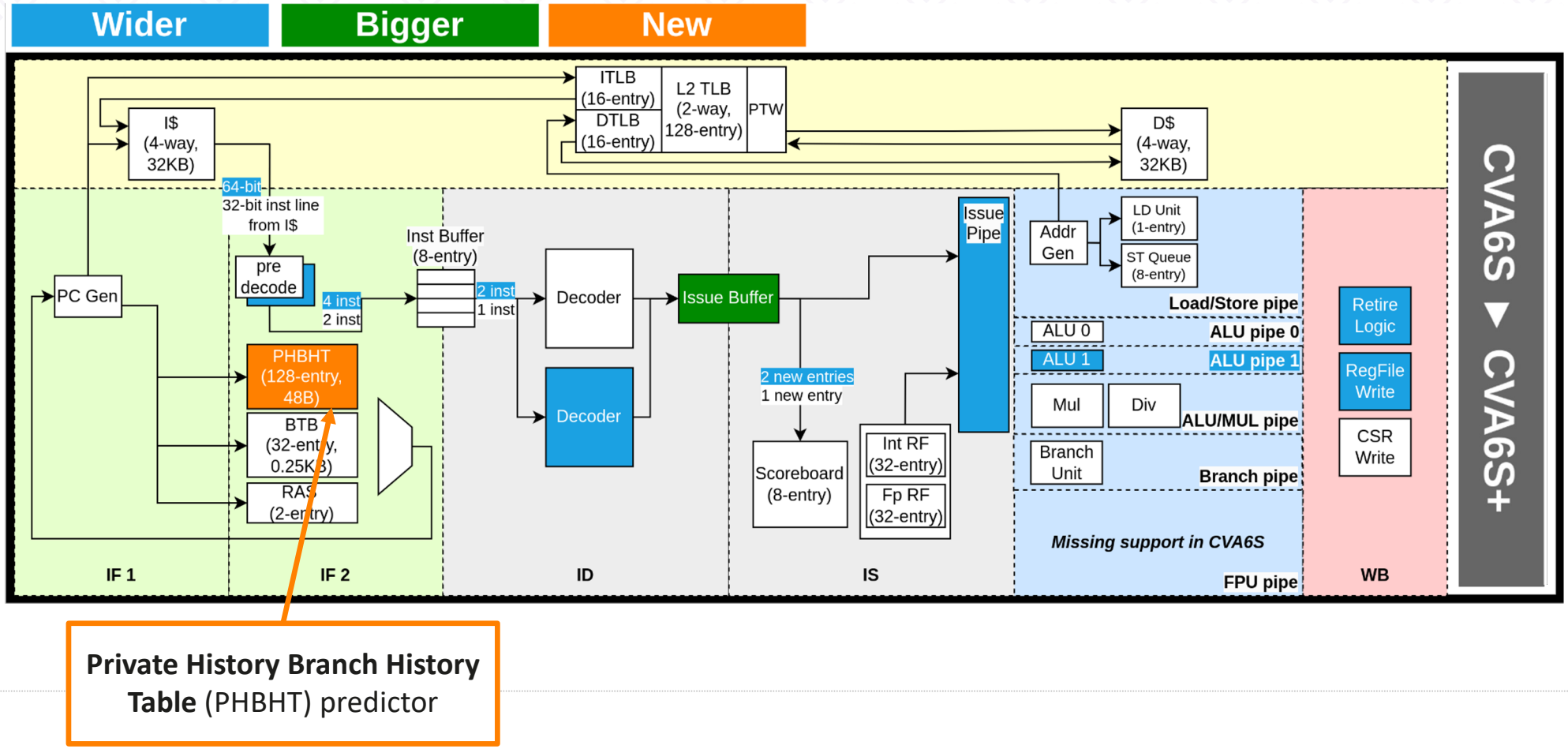
Chips-IT

Wider **Bigger**



CVA6(S+): Application-class 64-bit RISC-V core

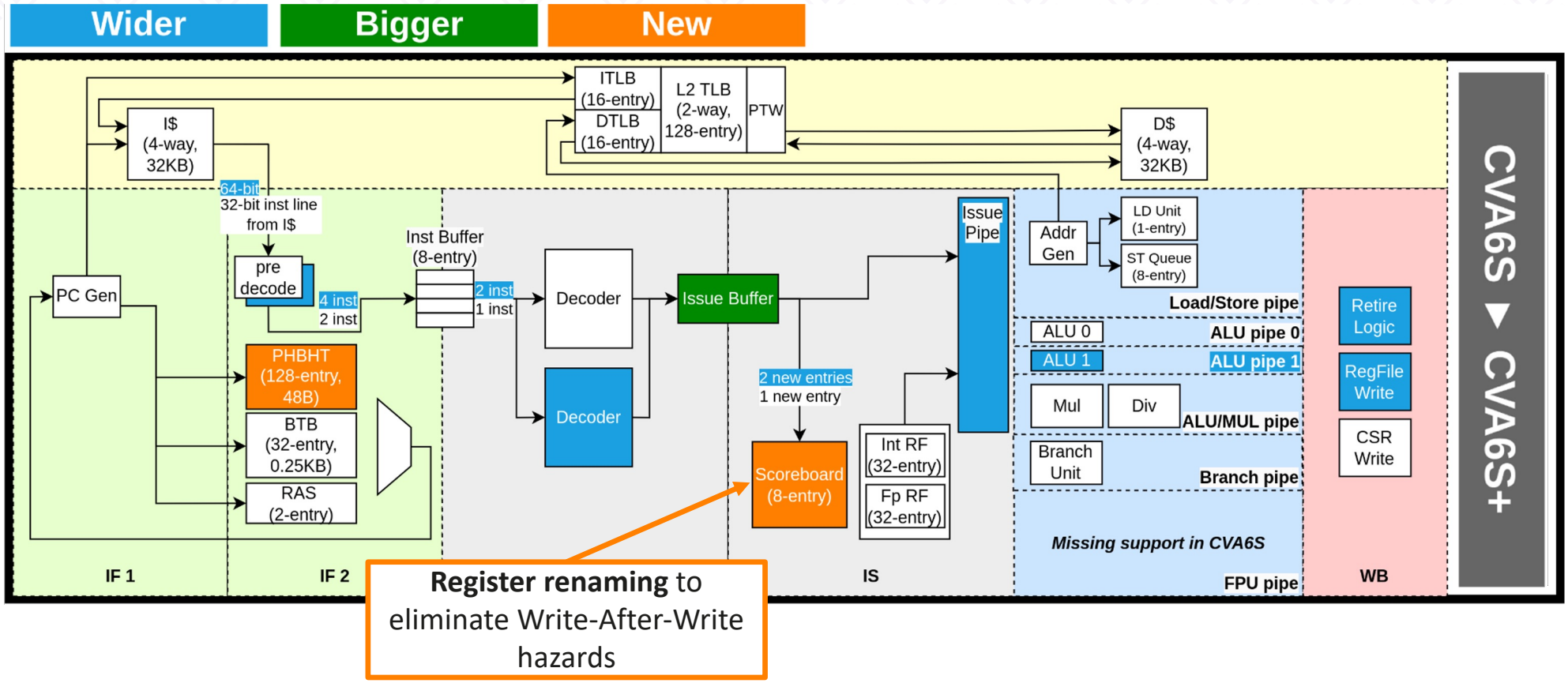
Chips-IT



Private History Branch History Table (PHBHT) predictor

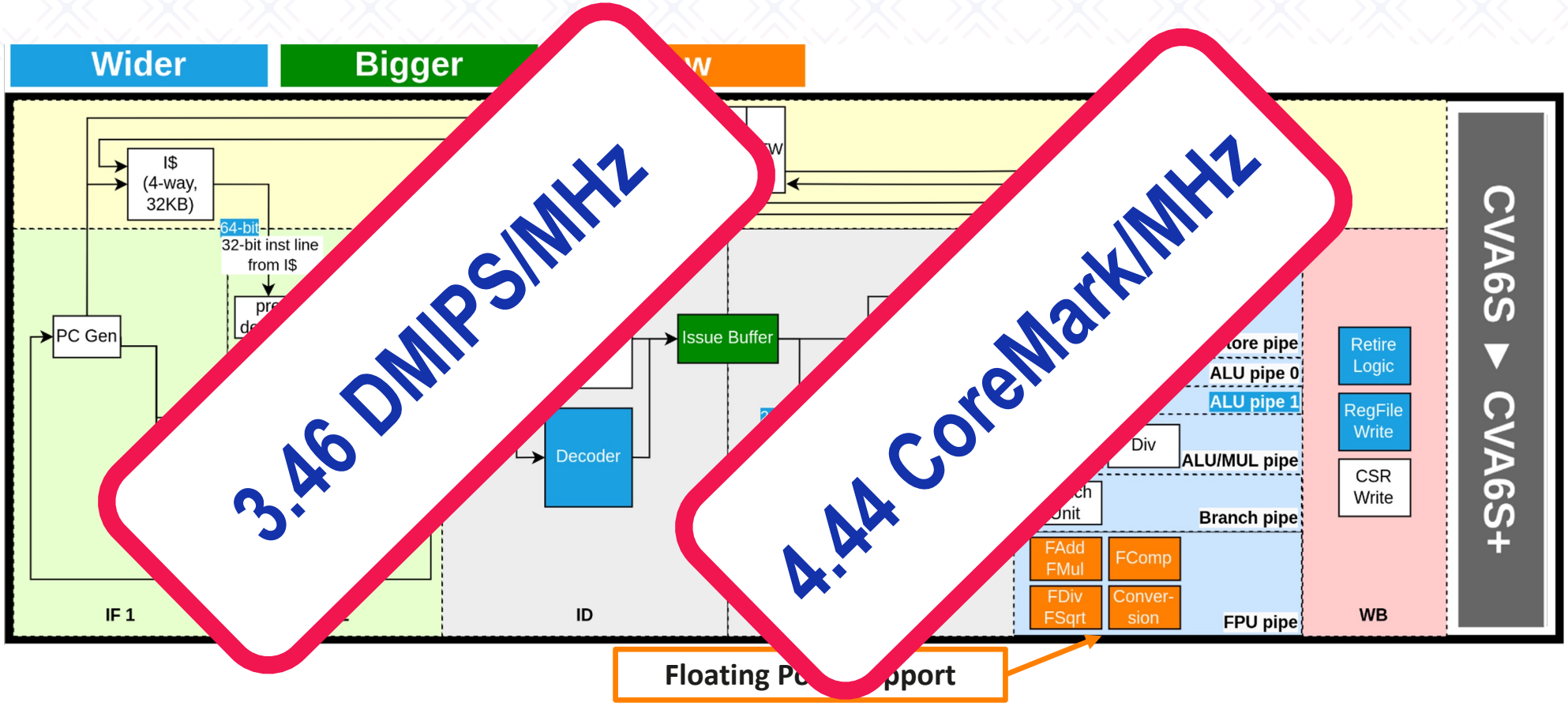
CVA6(S+): Application-class 64-bit RISC-V core

Chips-IT



CVA6(S+): Application-class 64-bit RISC-V core

Chips-IT



Heterogeneous systems

Chips-IT

- Processors can do two kinds of useful work:

Decide (jump through program parts)

- 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

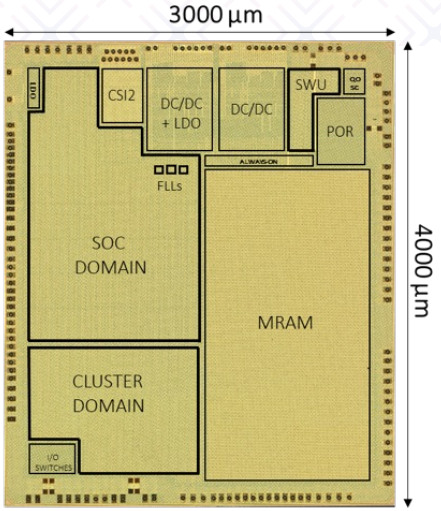
+

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

Vega: IoT Heterogeneous SoC (ISSCC 2021)

- **Transprecision Floating-Point RISC-V Extensions (16/32-bit float)**
- 4 MB non-volatile MRAM (Weight Storage)
- Programmable Cognitive Wake-up Unit based on HD Computing

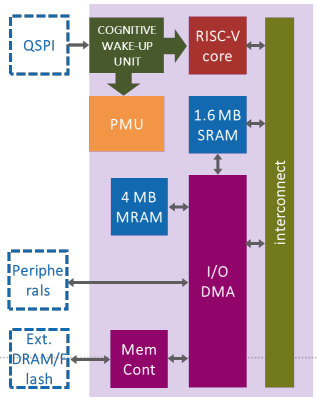


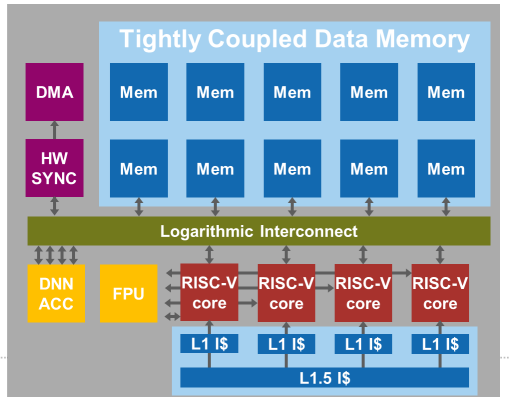
ISSCC 2021 / SESSION 4 / PROCESSORS / 4.4

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]

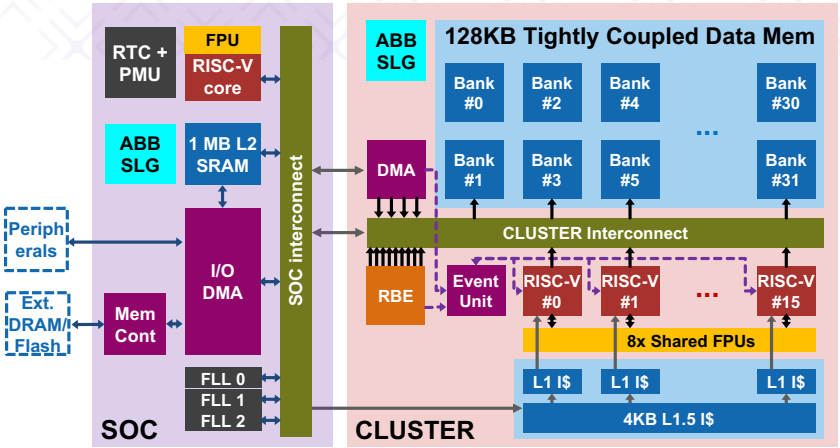
[1]University of Bologna, Bologna, Italy
[2]ETH Zurich, Zurich, Switzerland
[3]Greenwaves Technologies, Grenoble, France





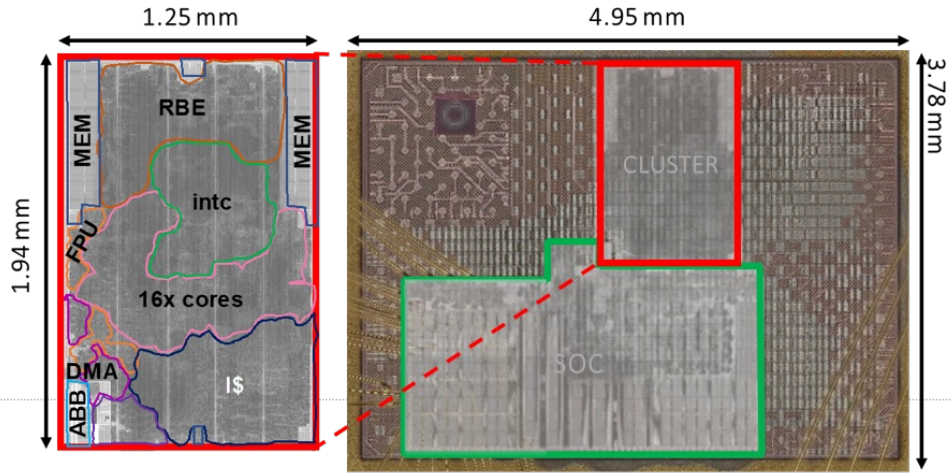
MARSELLUS: AI-IoT Heterogeneous SoC (ISSCC 2023)

- XpulpNN + M&L RISC-V Extensions
- 2-8b Reconfigurable Binary Engine for 3x3, 1x1 DNN kernels
- Adaptive Body Biasing with on-the-fly control

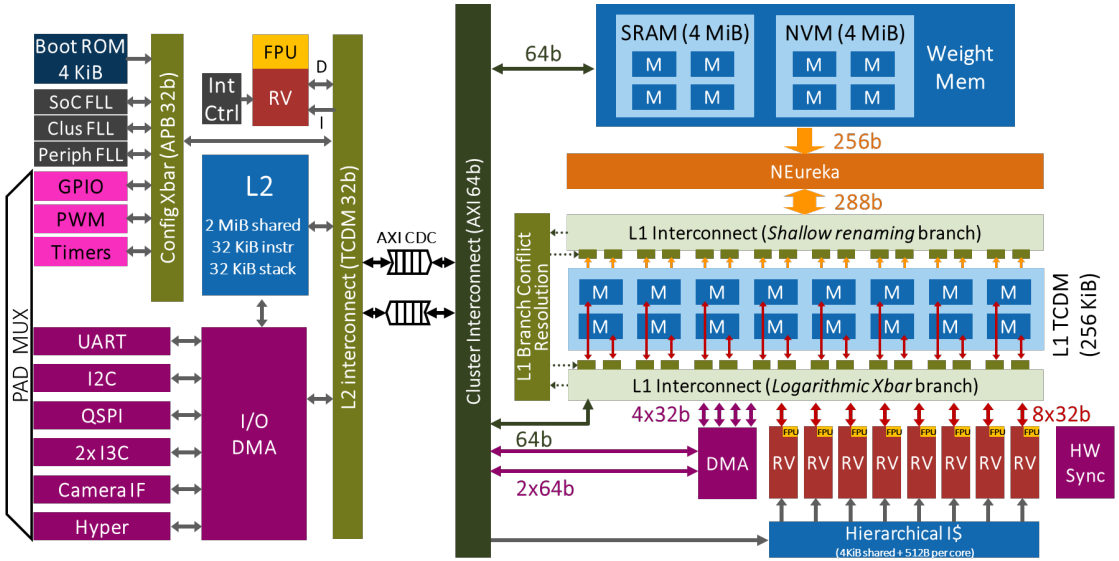


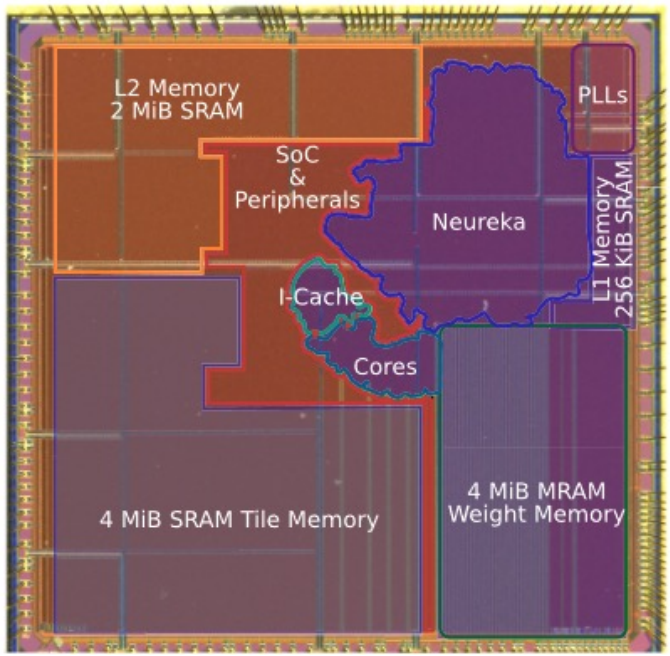
22.1 A 12.4TOPS/W @ 136GOPS AI-IoT System-on-Chip with 16 RISC-V, 2-to-8b Precision-Scalable DNN Acceleration and 30%-Boost Adaptive Body Biasing

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]



Siracusa: PULP-based XR glasses SoC



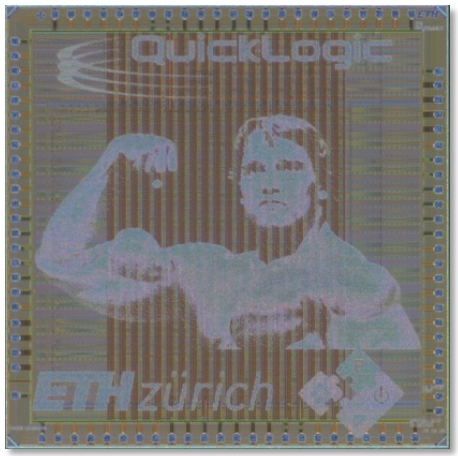


RISC-V + Archi Extensions for AR/VR Applications

A. S. Prasad et al., "Siracusa: A 16 nm Heterogenous RISC-V SoC for Extended Reality With At-MRAM Neural Engine," in IEEE Journal of Solid-State Circuits, vol. 59, no. 7, pp. 2055-2069, July 2024, doi: 10.1109/JSSC.2024.3385987.

The open model led to successful industry collaborations

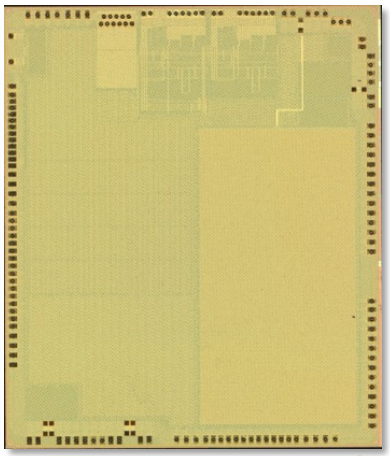
Chips-IT



QuickLogic®

(GF22)

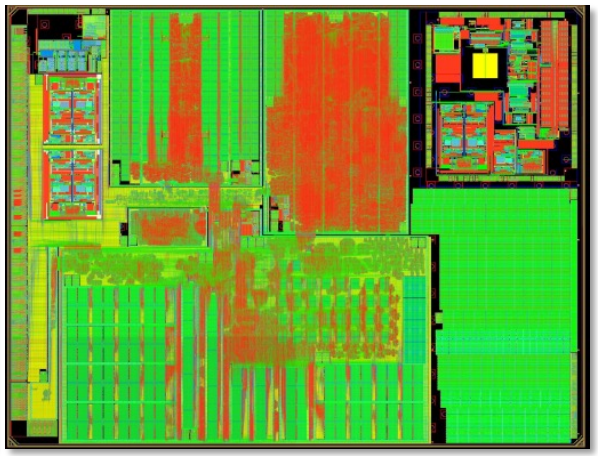
eFPGA with RISC-V core



GREENWAVES TECHNOLOGIES

(GF22)

IoT Processor with ML acceleration



DOLPHIN INTEGRATION
The enabler of low-power Systems-on-Chip

(GF22)

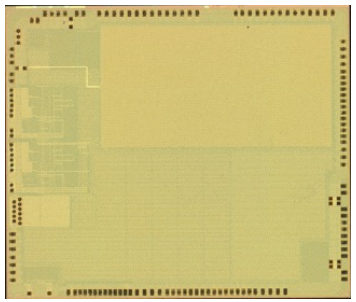
IoT Processor with low power modes and event based computing

Meta numem

(TSMC16)

XR Processor with NVM technology

And Products…



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

GAP9 SoC ***(commercial)***
65% of GAP9 is based on PULP open-source IPs

DNN Accelerator

Leading Edge Products…

Chips-IT

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

Submitter	Board Name	SoC Name	Processor(s) & Number	Accelerator(s) & Number	Software	Notes	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)	
							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	20[illegible]236.3	99.8	5[illegible]30.3	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]50.5	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]01.9	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	10[illegible]381.6	75.1	3[illegible]471.7	7.57	323.0
STMicroelectronics	NUCLEO-U575ZI-Q	STM32U575ZIT6Q	Arm® Cortex®-M33		X-CUBE-AI v7.3.0	160MHz, 1.8Vbat		133.4	33[illegible]4.5	139.7	[illegible]342.0	44.2	[illegible]38.5	4.84	119.1
Syntiant	NDP9120-EVL	NDP120	M0 + HiFi	Syntiant Core 2 (98MH	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 (30MH	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

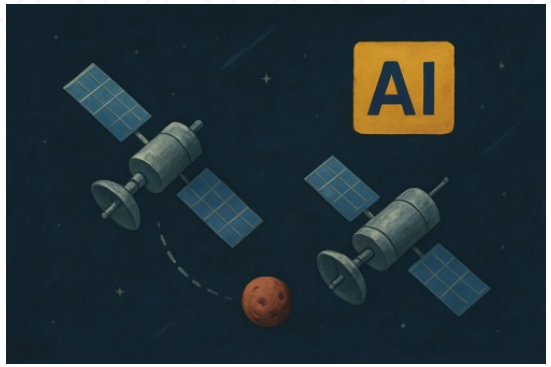
1.7x 3.7x 1.7x

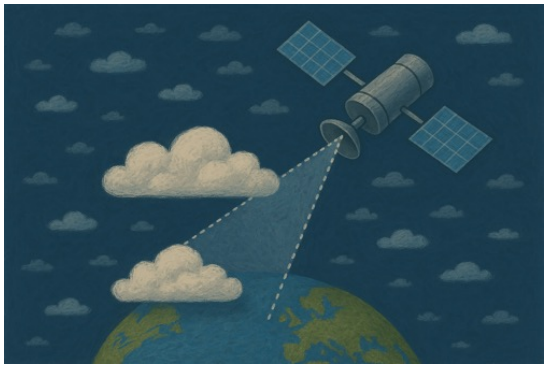
RISC-V (PULP) Currently Dominates TinyML benchmarks

AI in space cyber-physical systems (satellites)

Autonomous operations [1]

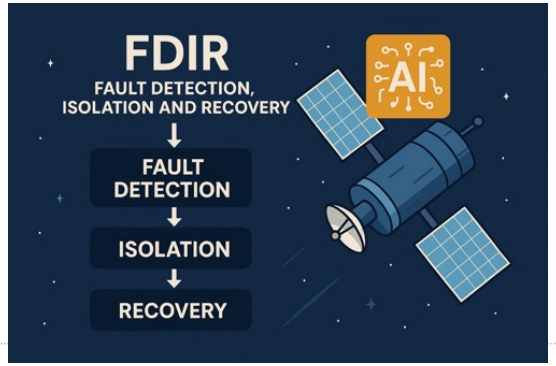
- Orbital collision hazard avoidance
- Asset reconfiguration
- Dynamic mission planning and reconfiguration [2]





Object detection and **scene understanding**

- **Phi-Sat-1**: AI-based image selection for cloud filtering [3]
- **MANTIS AI**: Cubesat for AI cloud detection and filtering [4]

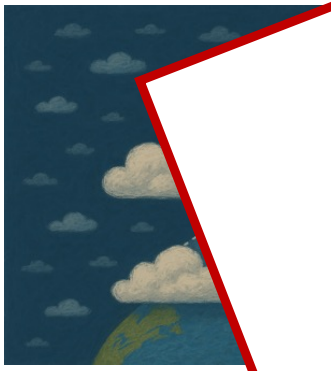


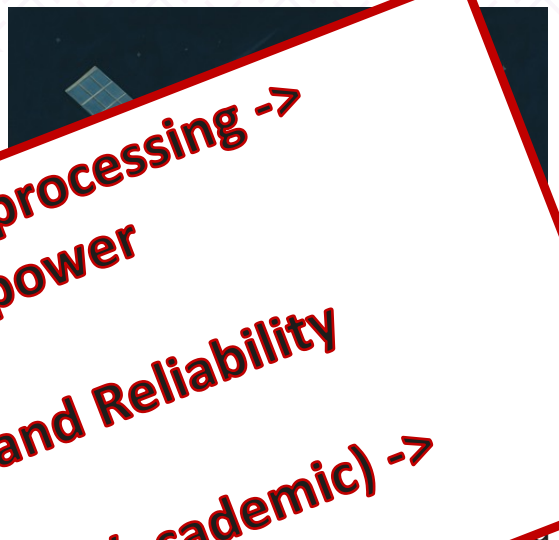
Fault detection, Isolation, and Recovery (**FDIR**) [5]

AI in space cyber-physical systems (satellites)

Autonomous operations [1]

- Orbital collision hazard avoidance
- Asset reconfiguration
- Dynamic mission planning and reconfiguration [2]





Real-time edge data processing -> Computing power

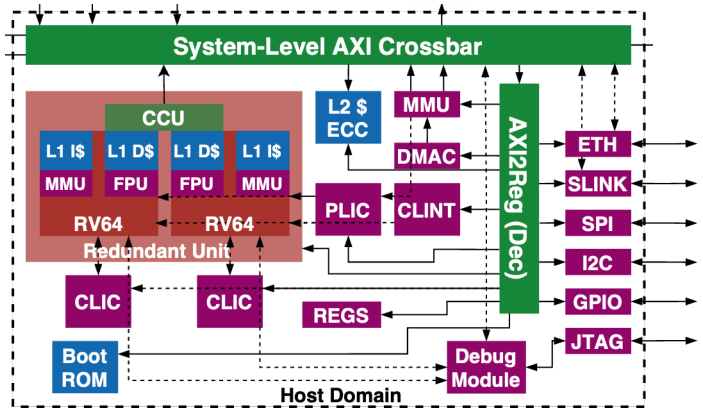
Fault Tolerance and Reliability

Research (industrial/academic) -> Open platforms

Fault detection, Isola

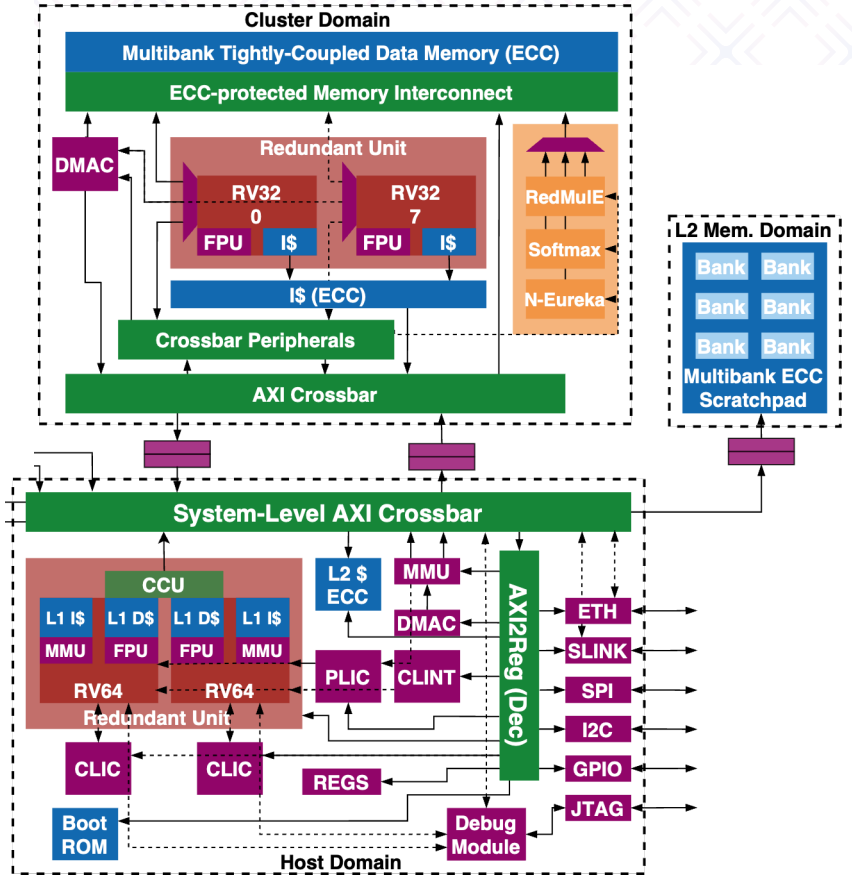
[5]

Astral Soc – unleash AI acceleration in space



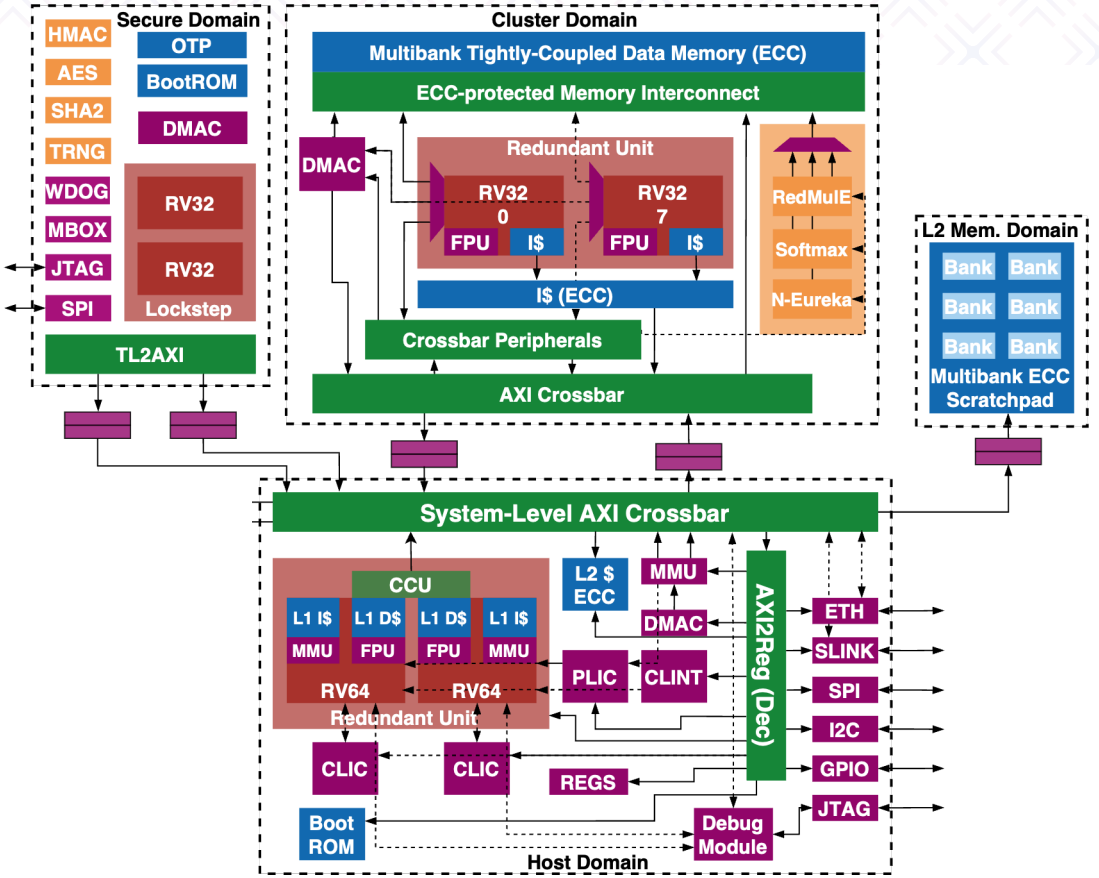
Astral Soc – unleash AI acceleration in space

Chips-IT



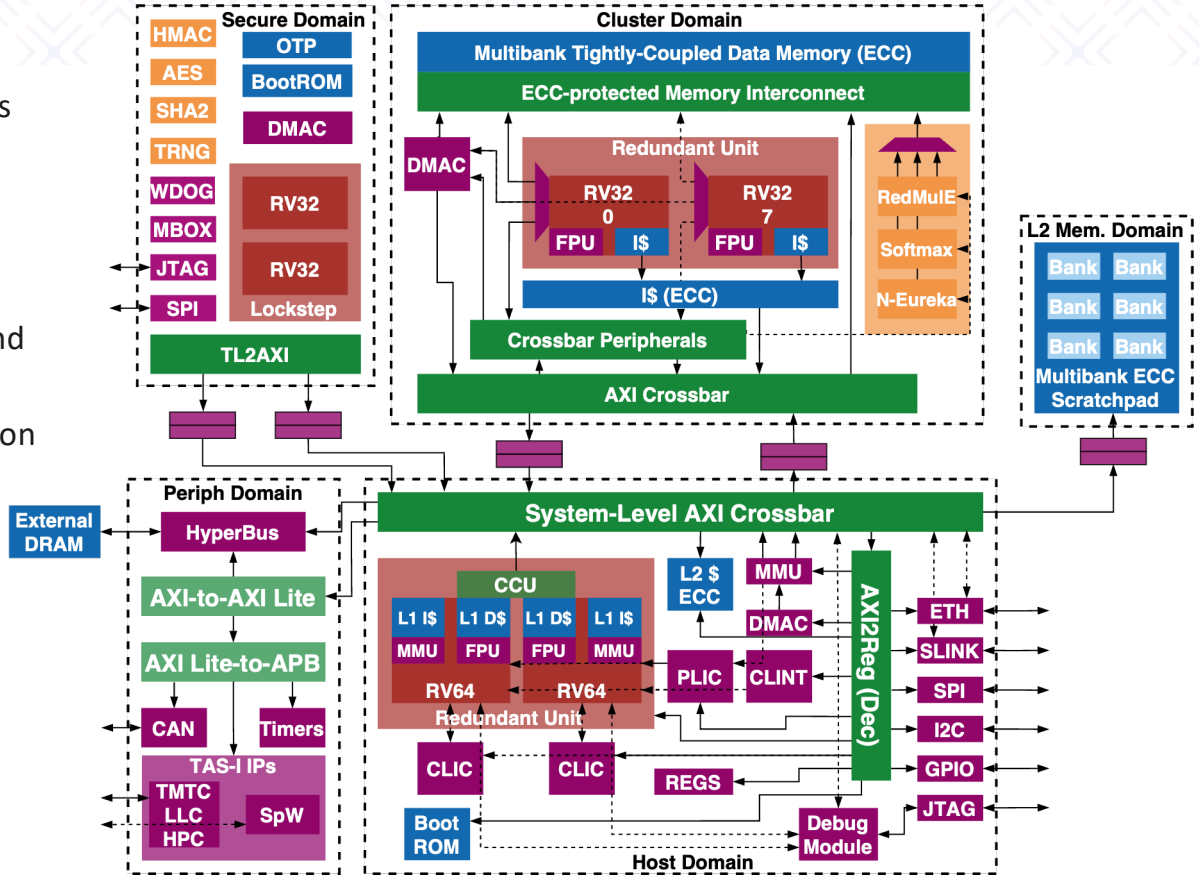
Astral Soc – unleash AI acceleration in space

Chips-IT

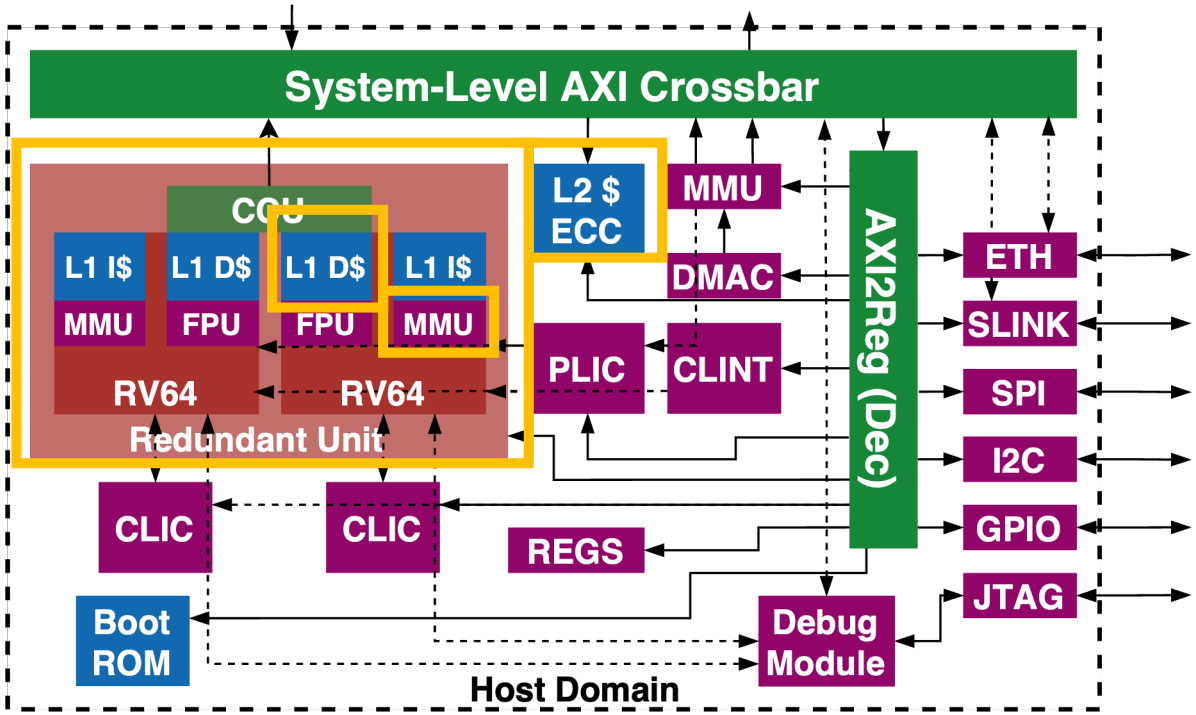


Astral Soc – unleash AI acceleration in space

- First-of-a-kind mixed-criticality heterogeneous platform
- Fault tolerance – architectural extensions for redundancy, error correction, and real-time recovery
- Fully open-source – all IPs are open-source and released under Apache-like license
- Highly parametric – straight-forward integration of additional IPs
- Ease of programmability – boots full-fledge Linux OS
- Performance&Security – powerful multicore accelerator + Root of Trust
- Development - Plug&Play flow for AMD Xilinx Ultrascale+ VCU118

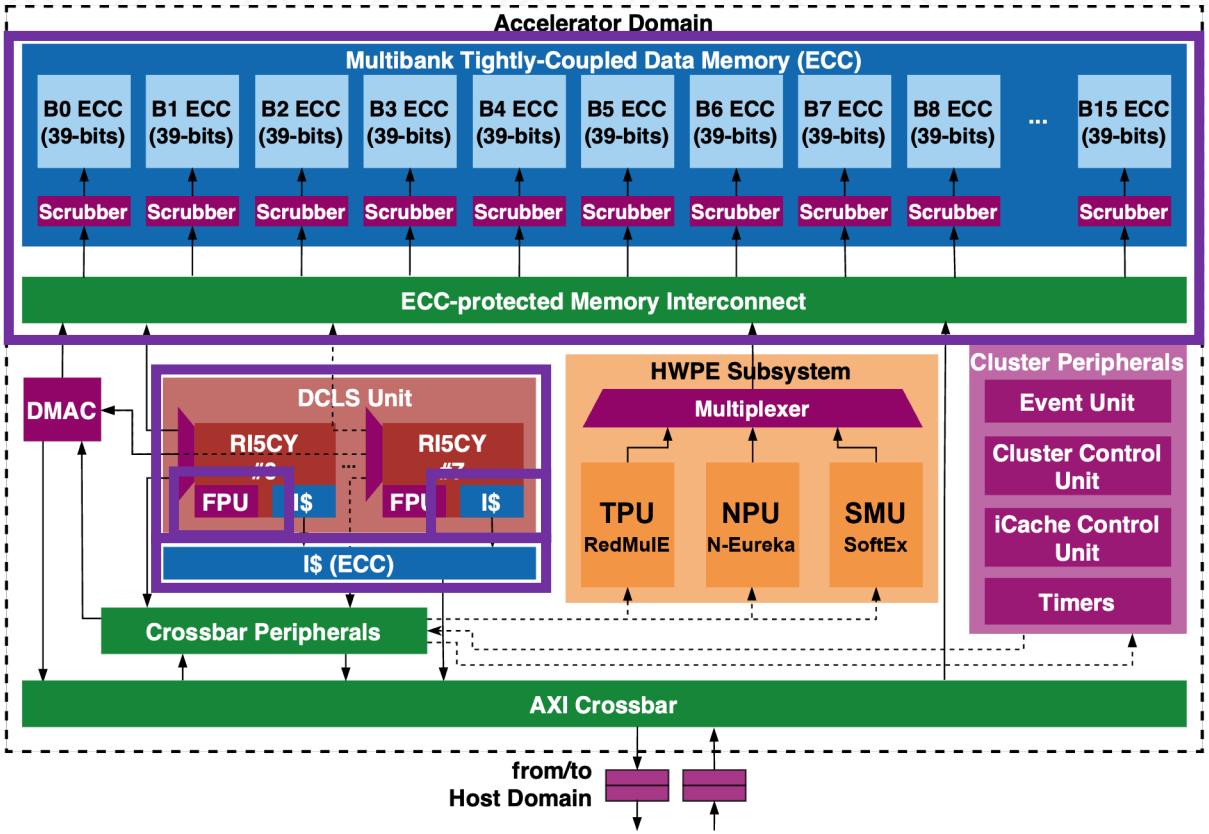


Astral soc – host domain architecture



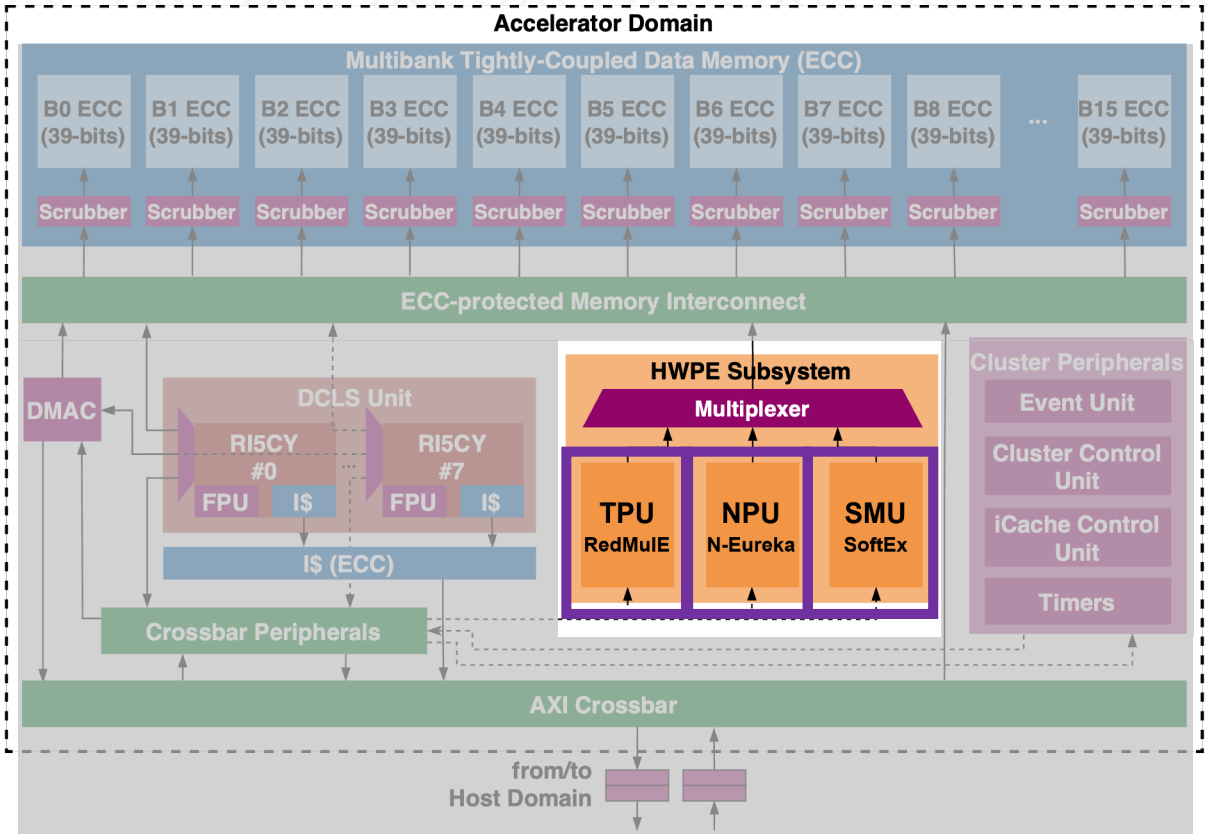
- **Linux-capable** multicore (1 - 4) cache-coherent cluster of **64-bit CVA6** cores based on the **ACE** protocol
- Runtime-programmable **redundant grouping** of the CVA6 cores to balance **reliability** and **performance**
  - **Less then 100 clock cycles fault recovery** with hardware extension
- **SEC-DED** extension of the CVA6 **TLB**s and **Branch Predictor**
- **ECC** extension of the **L1 data cache** of the cores for SECDED
- **ECC** extension of the **L2 cache** for single error automatic correction, and cache re-fetch/IRQ notification in case of multiple errors on clean/dirty lines

Astral soc – accelerator domain architecture



- Hybrid Modular Redundant cluster for runtime reconfiguration (**TMR/DMR/Independent**)
  - **Less then 30 clock cycles fault recovery** with hardware extension
- Extension of the cores' **FPU**s with **temporal repetition** with detection + correction of internal faults
- **L0 + L1 instruction cache** protected with **parity** for lines replacements in case of errors
- **Memory subsystem** protected with **SECDED** from the cores and accelerators branches down to the memory banks (storing ECCs) + memory **scrubbers**

Astral soc – accelerator domain architecture

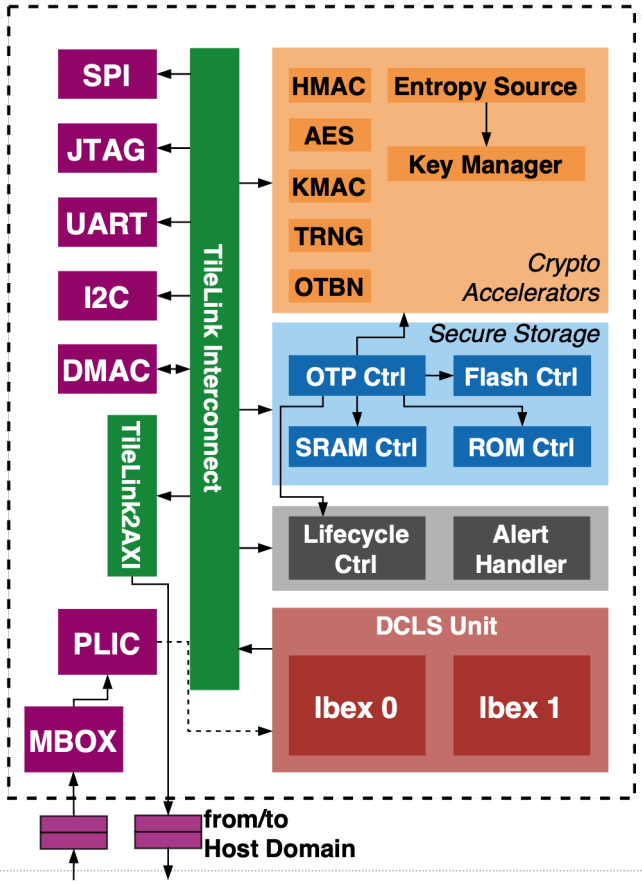


Coupling **performance** with **reliability**

Wide **accelerator subsystem** for on-chip inference and training acceleration

- **FP16 TensorCore** accelerator extended for fault-tolerance for on-chip matrix-multiplication boost
- Transprecision **Integer Neural Engine** for onchip CNN inference acceleration
- **FP16 SoftMax** accelerator for transformers acceleration

Astral soc - Secure domain (Opentitan integration)



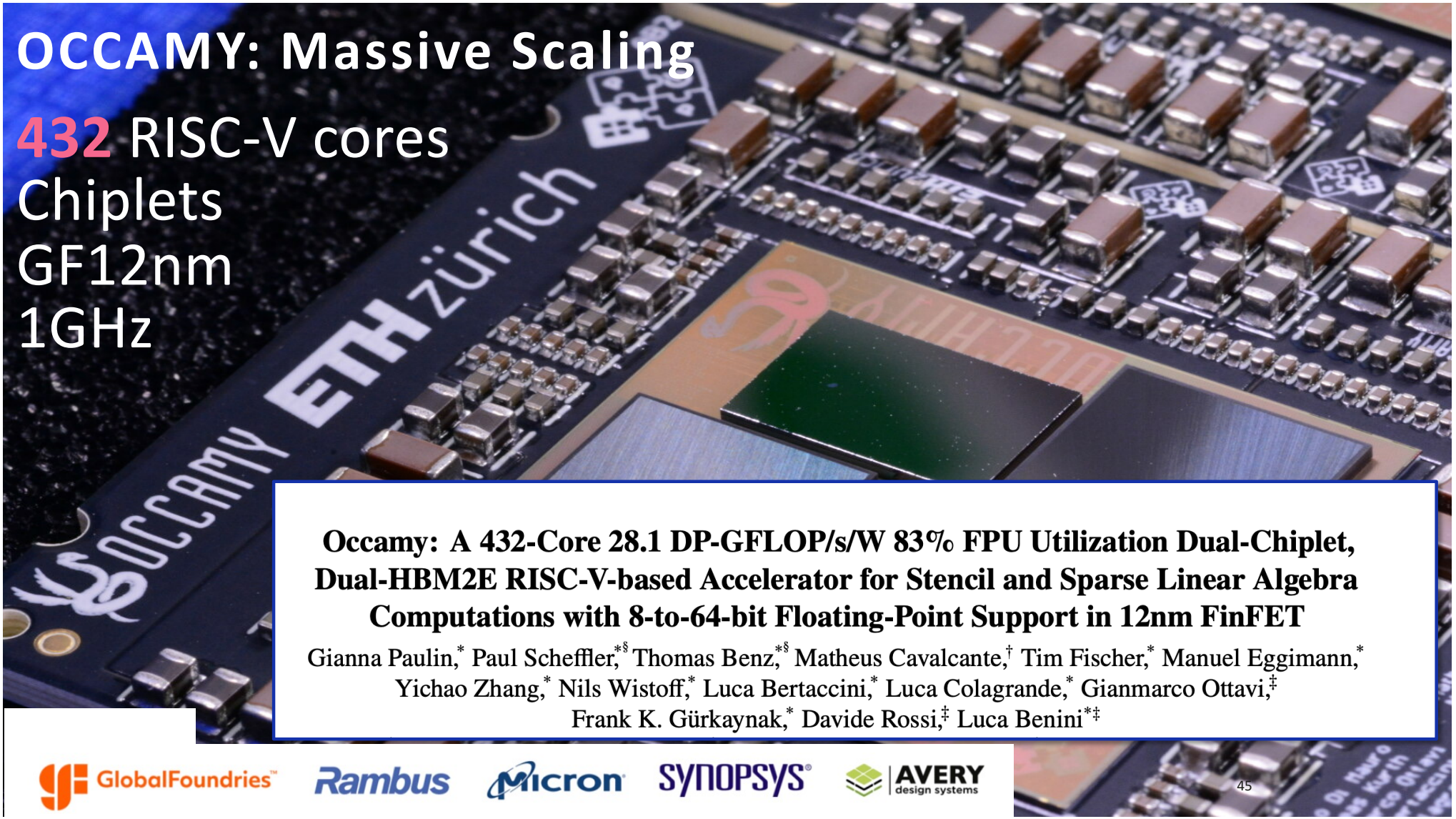
- **OpenTitan**: open-source Root of Trust from lowRISC
- Provides **secure-boot** services verifying the security of the Host domain CVA6 code
- Internal **DMAC**
  - Speed up data transfer during secure boot stage
  - Efficiently serving internal encryption/decryption accelerators
- Refer to OpenTitan docs for more details (https://opentitan.org)

Astral made its way into silicon

- Taped out in september 2024
- Global Foundries GF12LP+ 12 nm
- Received in spring 2025
- Currently under submission (results hidden prior official publication)

OCCAMY: Massive Scaling

432 RISC-V cores
Chiplets
GF12nm
1GHz



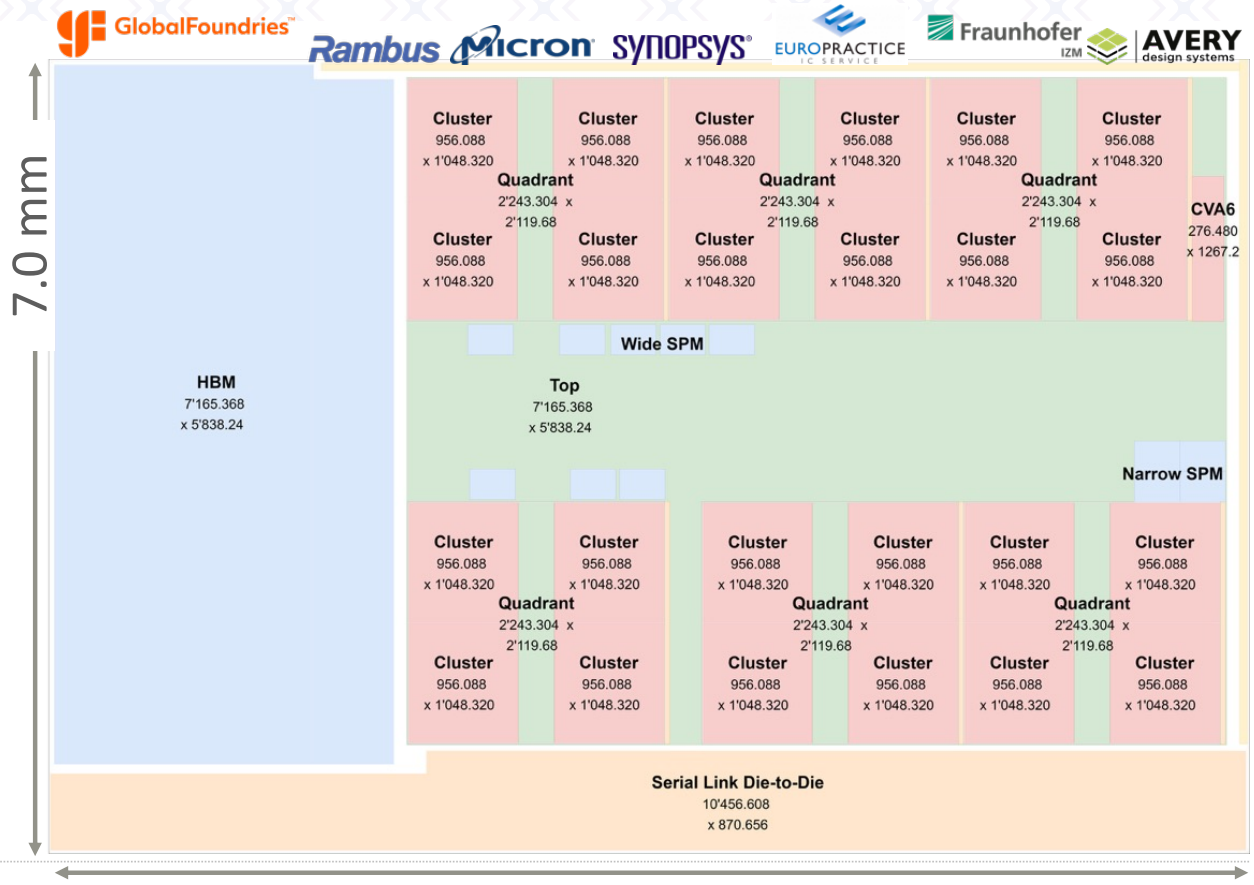
Occamy: A 432-Core 28.1 DP-GFLOP/s/W 83% FPU Utilization Dual-Chiplet, Dual-HBM2E RISC-V-based Accelerator for Stencil and Sparse Linear Algebra Computations with 8-to-64-bit Floating-Point Support in 12nm FinFET

Gianna Paulin,[*] Paul Scheffler,[*§] Thomas Benz,[*§] Matheus Cavalcante,[†] Tim Fischer,[*] Manuel Eggimann,[*]
Yichao Zhang,[*] Nils Wistoff,[*] Luca Bertaccini,[*] Luca Colagrande,[*] Gianmarco Ottavi,[‡]
Frank K. Gürkaynak,[*] Davide Rossi,[‡] Luca Benini[*‡]

GlobalFoundries™ Rambus Micron SYNOPSYS® AVERY design systems

Occamy: Next-Generation HPC SoC

2.5D 2xOccamy+2HBM2 on interposer



Occamy – Massive Scaling

Dual Chiplet System Occamy:

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

Interposer Hedwig:

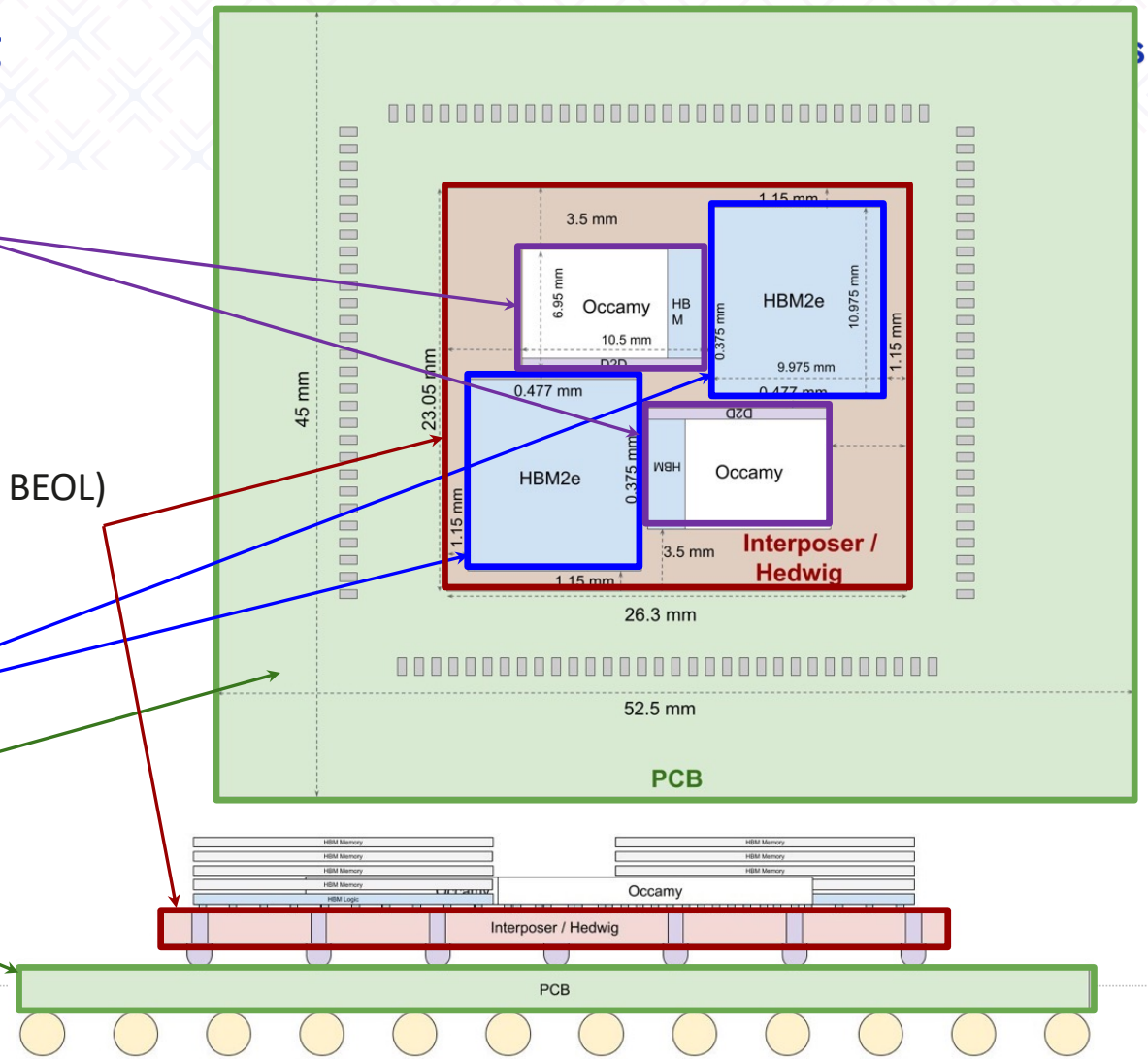
- Technology: 65nm, passive (only BEOL)
- Area: 26.3mm x 23.05mm

HBM2e:

- 16GB HBM2e (Micron)

Fan-out PCB:

- RO4350B
- 52.5mm x 45mm



Thank you

Chips-IT