

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

Working with RISC-V

Part 5 of 5 : PULP based chips

Davide Rossi Luca Benini

<davide.rossi@unibo.it> <luca.benini@unibo.it>















ETH zürich

- Part 1 Introduction to RISC-V ISA
- Part 2 Advanced RISC-V Architectures
- Part 3 PULP concepts
- Part 4 PULP Extensions and Accelerators
- Part 5 PULP based chips
 - From concept to reality
 - Single core microcontrollers PULPino to PULPissimo
 - Many core systems OpenPULP
 - Advanced systems with accelerators
 - Lessons learned, the good, the bad and the ugly.



We will discuss chips we have made with PULP

• Why make chips at all?

- MPW: Only limited samples
- Use cases

Single core PULP chips

- PULPino (Imperio)
- PULPissimo (Arnold)

Many core PULP chips

- Cluster only (Honey Bunny, Dustin)
- PULPopen (Mr. Wolf, Vega)

Advanced PULP chips

- Kosmodrom: 2x 64b Ariane cores
 + ML accelerators
- Making use of technology: Body biasing

Lessons learned

- There are many pitfalls
- We had great success, but..
- Sometimes you have embarrasing failures. Part of the process



Multi Project Wafer, chips for prototyping

Cost sharing method for ICs EUROPRACTICE

- Multiple ICs are manufactured together. They share the mask costs
 - 1.5M cost / 10 projects = 150k per project
 - But you only get 1 / 10 of the area
- Dedicated MPW services available
 - Europractice-IC for SMEs and academia
- You only get few chips
 - Usually 50 to 200
 - Per chip costs very high (few kUSD)
- All our chips through MPWs

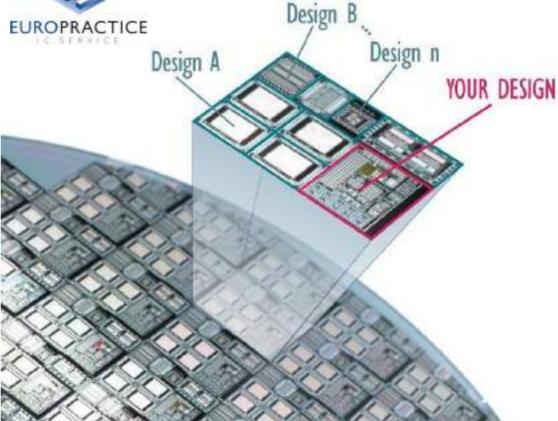


Image taken from https://europractice-ic.com/mpw-prototyping/general/mpw-minisic/



Our ASICs have different use cases

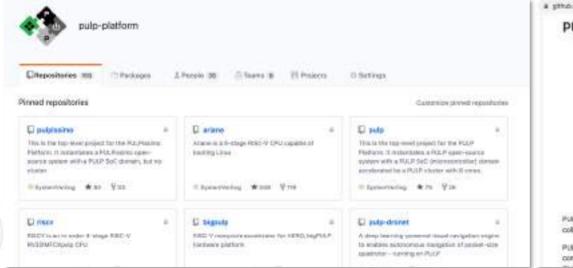
- Chips characterized on an IC tester (Poseidon 22nm)
- Research demonstrators (Nano drone with Mr. Wolf/GAP8)
- Industrial uses of our cores/peripherals (open-isa.org Vega)

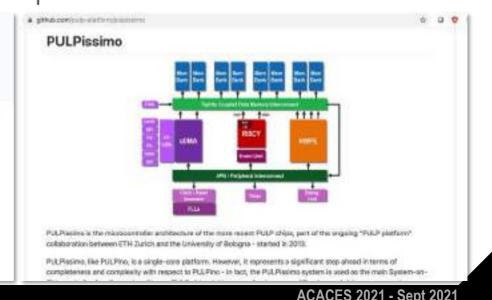


H zürich

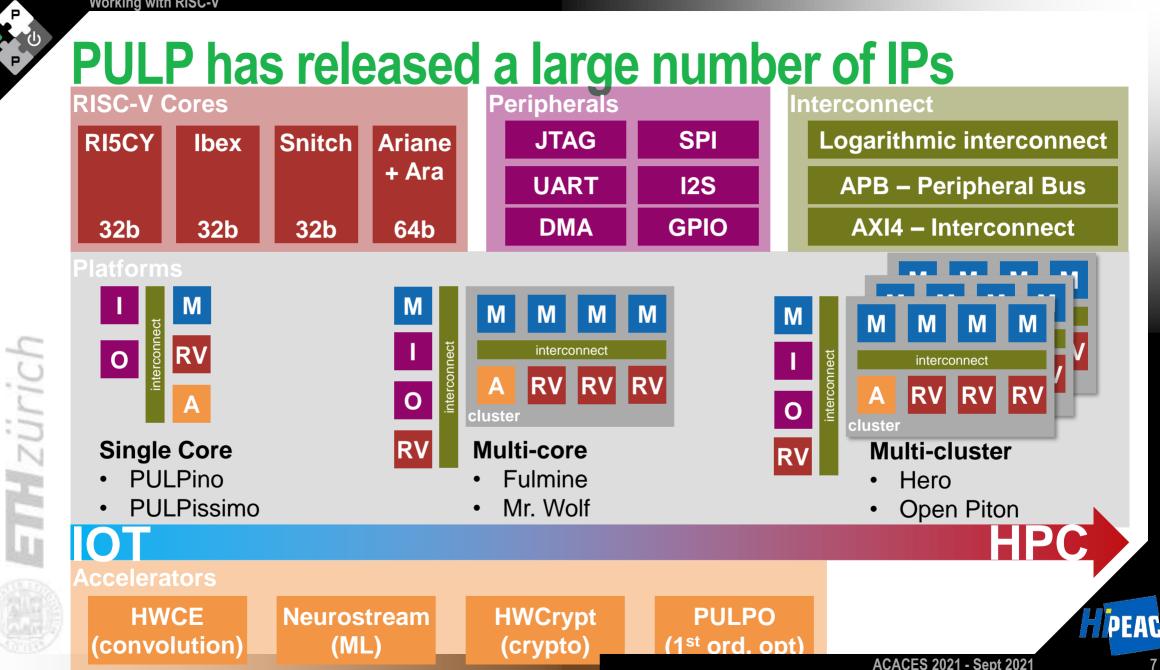
Most of what we show is openly available

- All our development is on GitHub
 - HDL source code, testbenches, software development kit, virtual platform
 https://github.com/pulp-platform
- PULP is released under the permissive Solderpad license
 - Allows anyone to use, change, and make products without restrictions.



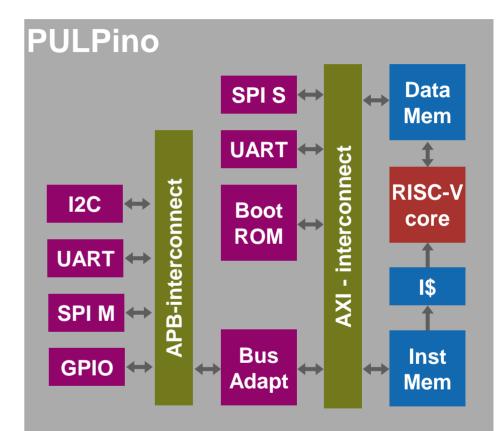


6



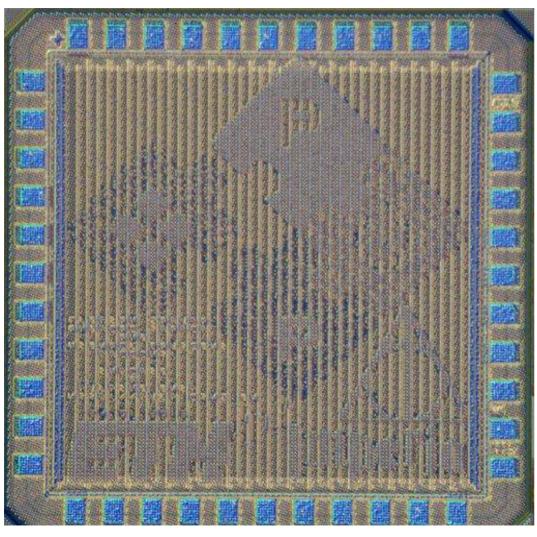
PULPino: Our first open source release

- Simple design
 - Meant as a quick release
- Separate data and inst. mem
 - Makes it easy in HW
 - Not meant as a Harvard arch.
- Can use all our 32bit cores
 - RI5CY, Zero/Micro-Riscy (Ibex)
- Peripherals from other projects
 - Any AXI and APB peripherals could be used



zürich

Imperio – 65nm RISC-V core



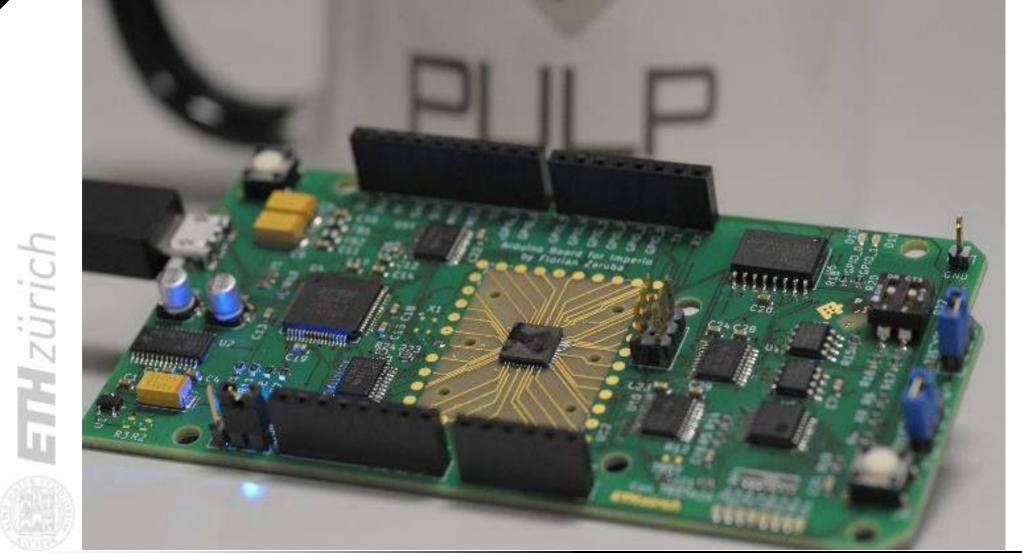
Chip implemented in 65nm

- Using RI5CY (RV32IMC) core
- 64 kBytes of memory
- Basic peripherals (GPIO, SPI, I2C)
- Working debug interface

Functional up to 500 MHz

- Main challenge was to find fast memory cuts to work at that speed.
- Memory made of multiple smaller cuts to maximize the operating speed.

Working chip on an Arduino compatible board



PEAC

10

Arnold (2018) – Fastest collaboration

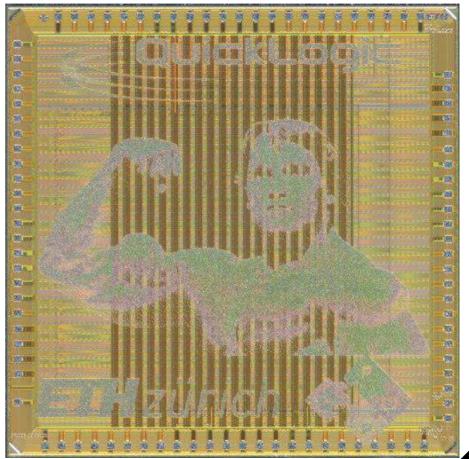
GF22nm

zürich

- RISC-V microcontroller with eFPGA
- Based around PULPissimo

Collaboration with Quicklogic

- Met at GTC 2017 by coincidence
- In one year chip was taped out
- Only possible because of open source nature
- Quicklogic is going open source
 - They announced June 2020 the Quicklogic
 Open Reconfigurable Computing
 https://www.quicklogic.com/QORC/

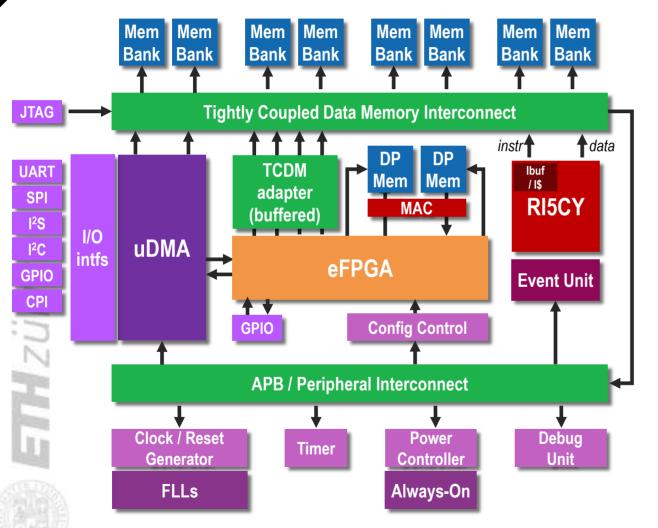


ACACES 2021 - Sept 2021

P. D. Schiavone et al., "Arnold: An eFPGA-Augmented RISC-V SoC for Flexible and Low-Power IoT End Nodes," TVLSI, vol. 29, no. 4, pp. 677-690, April 2021.



PULPissimo: very good platform for extensions

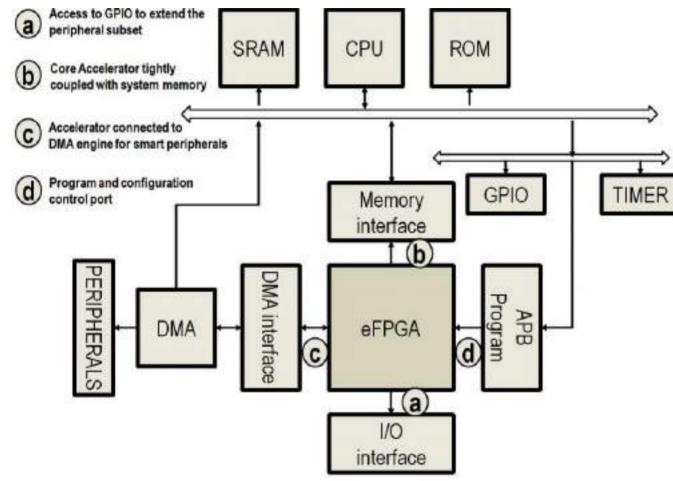


- eFPGA added as accel.
 - Easy plug and play
 - Configuration over APB
 - Additional ALU and memory
 - Uses the same memory
- Multiple operation modes
 - Configurable peripheral
 - Accelerator for core
 - Accelerator for independent I/O



ETH zürich

Experimental platform with many configurations

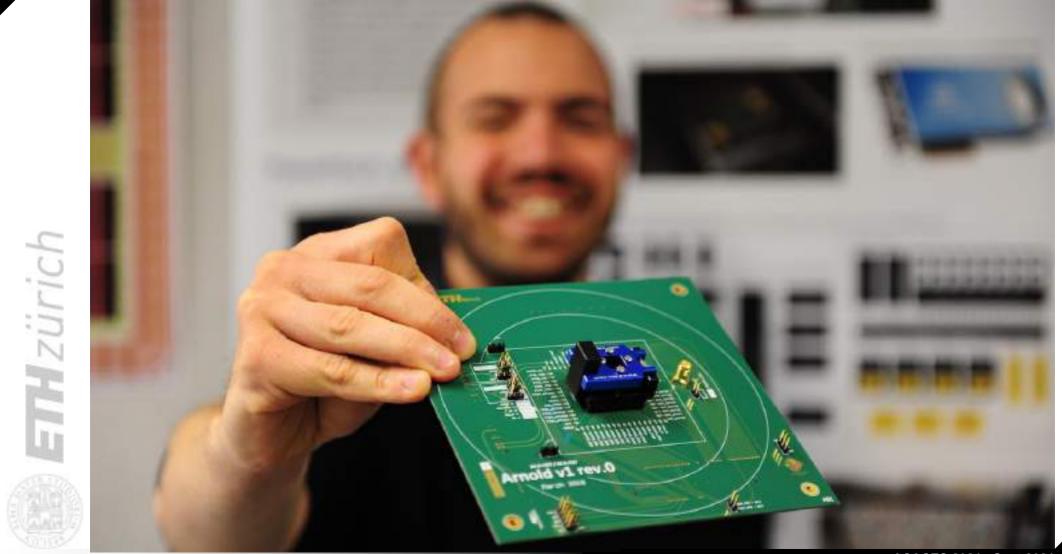


I/O subsystem accel

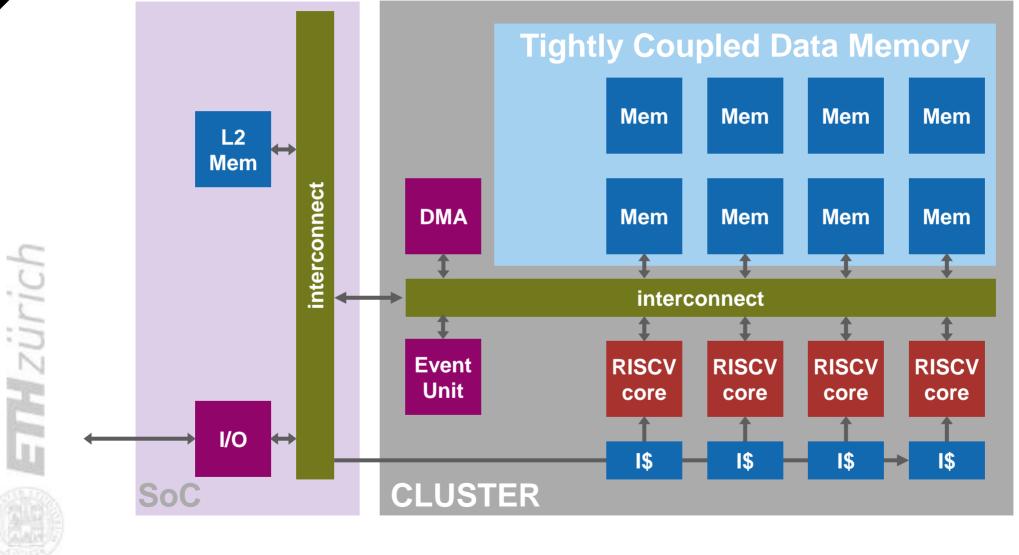
• 6.0mW, 2.5x

- Custom I/O interface
 - BNN interface 12.5mW
 2.2x
- CPU accelerator
 - CRC 7.5mW 42x
- Many more ideas
 - Dynamic reconfiguration

Arnold test board with D. Schiavone



Full Multi-Cluster SoCs

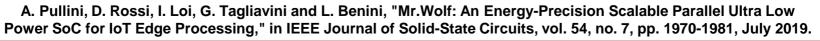


ACACES 2021 - Sept 2021

Mr. Wolf (TSMC 40): 8+1 core IoT Processor

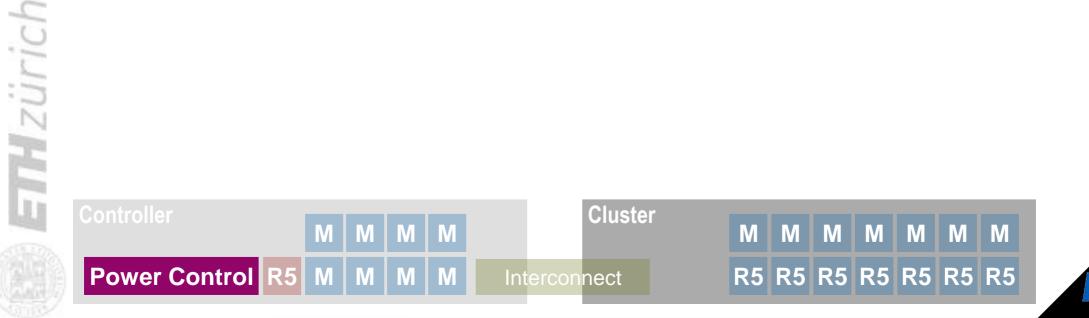
One cluster with

- 8 RISC-V cores
- 2x shared FPU units
- 64 kByte of TCDM
- One controller with
 - 512 kByte L2 RAM
 - Peripherals
- On chip voltage regulators
 - By Dolphin Integration



On-chip regulators allow different power modes

Power Mode	VDD	Frequency Range	Power
Deep Sleep	0.8 V	n.A.	72 µW



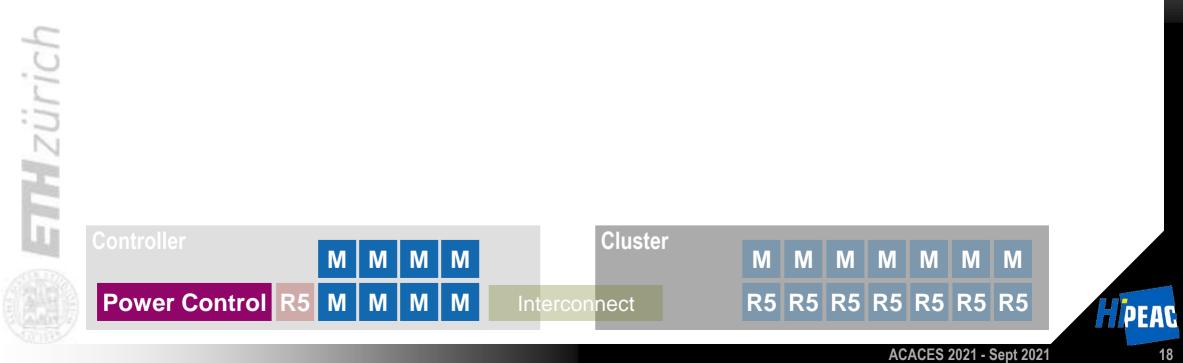


ACACES 2021 - Sept 2021

17



Power Mode	VDD	Frequency Range	Power
Deep Sleep	0.8 V	n.A.	72 µW
State Retentive Deep Sleep	0.8 V	n.A.	77 – 108 µW



SoC is awake but is clock gated

Power Mode	VDD	Frequency Range	Power
Deep Sleep	0.8 V	n.A.	72 µW
State Retentive Deep Sleep	0.8 V	n.A.	77 – 108 µW
SoC Idle	0.8 – 1.1V	SoC clock gated	0.55 – 1.96 mW
Controller M M M	M	uster M M M M	M M M
Power Control R5 M M M	M Interconneo	ct R5 R5 R5 R5	5 R5 R5 R5

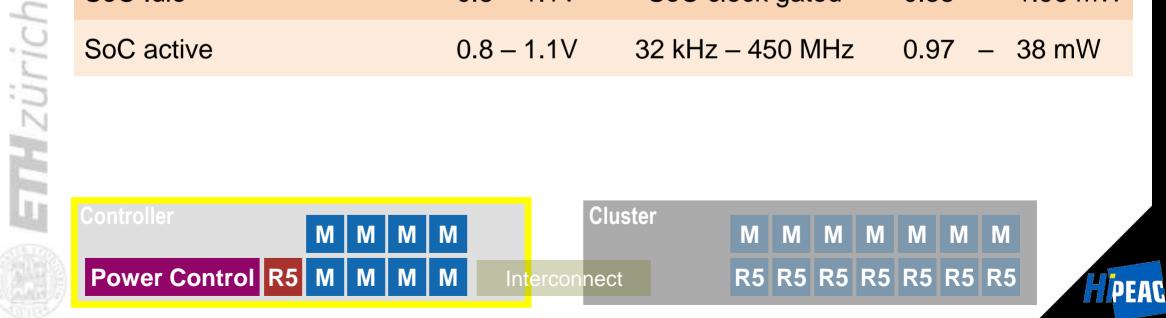
ETH zürich

ACACES 2021 - Sept 2021

19



Power Mode	VDD	Frequency Range	Power
Deep Sleep	0.8 V	n.A.	72 µW
State Retentive Deep Sleep	0.8 V	n.A.	77 – 108 µW
SoC Idle	0.8 – 1.1V	SoC clock gated	0.55 – 1.96 mW
SoC active	0.8 – 1.1V	32 kHz – 450 MHz	0.97 – 38 mW



ACACES 2021 - Sept 2021

ETH zürich

Cluster is active, but clock gated

Power Mode	VDD	Frequency Range	Power
Deep Sleep	0.8 V	n.A.	72 µW
State Retentive Deep Sleep	0.8 V	n.A.	77 – 108 µW
SoC Idle	0.8 – 1.1V	SoC clock gated	0.55 – 1.96 mW
SoC active	0.8 – 1.1V	32 kHz – 450 MHz	0.97 – 38 mW
Cluster Idle	0.8 – 1.1V	Cluster clock gated	1.2 – 4.6 mW



ACACES 2021 - Sept 2021



Power Mode	VDD	Frequency Range	Power
Deep Sleep	0.8 V	n.A.	72 µW
State Retentive Deep Sleep	0.8 V	n.A.	77 – 108 µW
SoC Idle	0.8 – 1.1V	SoC clock gated	0.55 – 1.96 mW
SoC active	0.8 – 1.1V	32 kHz – 450 MHz	0.97 – 38 mW
Cluster Idle	0.8 – 1.1V	Cluster clock gated	1.2 – 4.6 mW
Cluster Active	0.8 – 1.1V	32 kHz – 350 MHz	1.6 – 153 mW
Controller M M M	M	luster M M M M	MMM
Power Control R5 M M M	M Interconne	ct R5 R5 R5 R5	R5 R5 R5

ACACES 2021 - Sept 2021

AC

22

Our OpenPULP release is Mr. Wolf

- With Mr. Wolf, most of what we have is open sourced
 - This is a **complex IoT processor**, not like the much simpler PULPino
 - 8 + 1 cores, FPUs, shared accelerators, multiple power down modes.
- Still many parts can still not be open source
 - Technology specific information, P&R scripts
 - Memory macros, selected cuts, their performance
 - I/O cells

Working with RISC-

- FLL, analog macros, I/O cells, memory cuts (affects performance), P&R scripts
- Interesting industry collaboration
 - Greenwaves, BitCraze, Dolphin

Mr. Wolf has been used in multiple systems

- Designed as an application processor
 - We still build boards with it
 - Despite only 200 manufactured
- Widespread industrial use:
 - Dolphin IP was validated on this chip
 - Greenwaves GAP8 is based on the open source release OpenPULP
 - BitCraze AI Deck is related



ETH zürich

24

Complete Application: DroNET on NanoDrone



Only onboard computation for *contronomous* flight + obstacle avoidance no human operator, no ad-hoc external signals, and no remote base-station!

H zürich



R

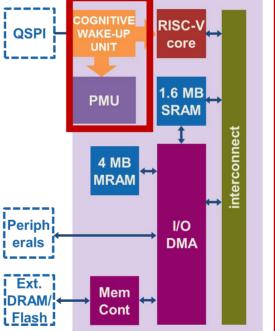
VEGA: Extreme Edge IoT Processor

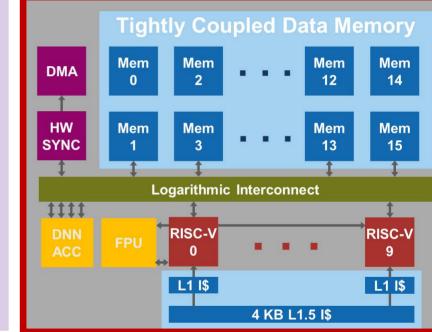
Fully programmable RISC-V based cluster targeting highly dynamic Near-Sensor Analytic Applications (NSAA)

26

1.7 μW cognitive unit for autonomous wake-up from retentive sleep mode

ETH zürich





GREENWA



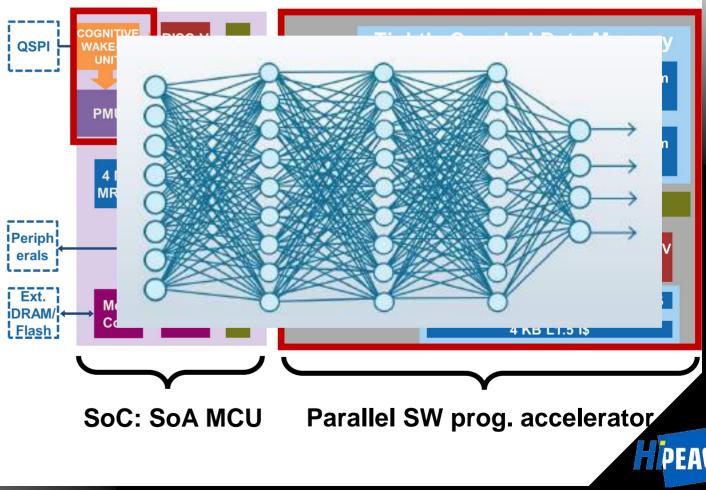
DEAC

VEGA: Extreme Edge IoT Processor

Fully programmable RISC-V based cluster targeting highly dynamic Near-Sensor Analytic Applications (NSAA)

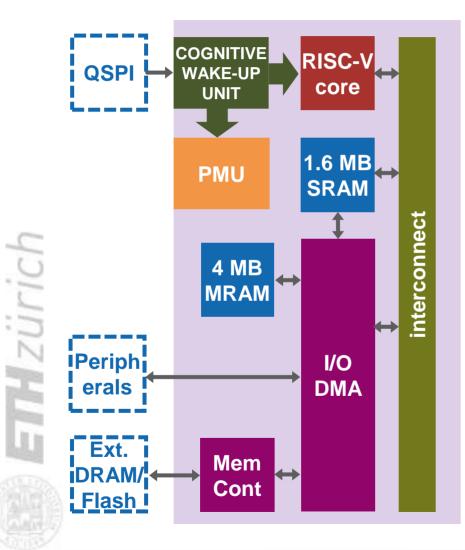
27

- 1.7 μW cognitive unit for autonomous wake-up from retentive sleep mode
- Fully integrated execution of real-life DNN from 4 MB of nonvolatile MRAM (first time for an IoT end-node)



SoC Overview

28



- 32-bit RISC-V core (Fabric Controller)
- 1.6 MB L2 SRAM

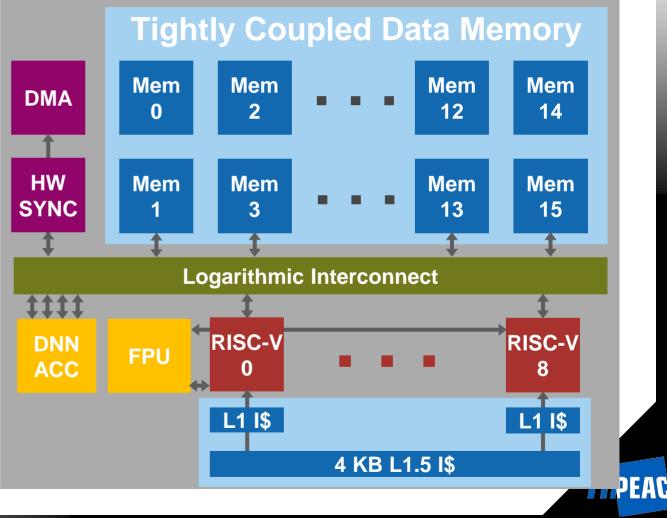
- 4 MB non-volatile MRAM
- Standard set of peripherals (SPI, I2C, UART, CSI2...)
- Off-chip memory (*HyperRAM™ DRAM / Flash)
- Autonomous I/O DMA
- Cognitive smart wake-up
- 3 Frequency-Locked Loops (FLL)
 - 2 Voltage regulators (HP/LP) + 1 LDO (COTS) + PMU

*https://www.cypress.com/products/hyperram-memory

Software-Programmable Accelerator

- 9 RISC-V DSP cores
- 128KB 16-Banks TCDM (scratchpad, no cache)
- Single-cycle latency, word-level interleaved Interconnect
- DMA for explicit memory mgmt.
- I\$: 9x 0.5kB L1 I\$ + 4KB L1.5 I\$
- Hardware Synchronizer (HW SYNC)
- Shared SIMD Floating-Point Unit (FPU)

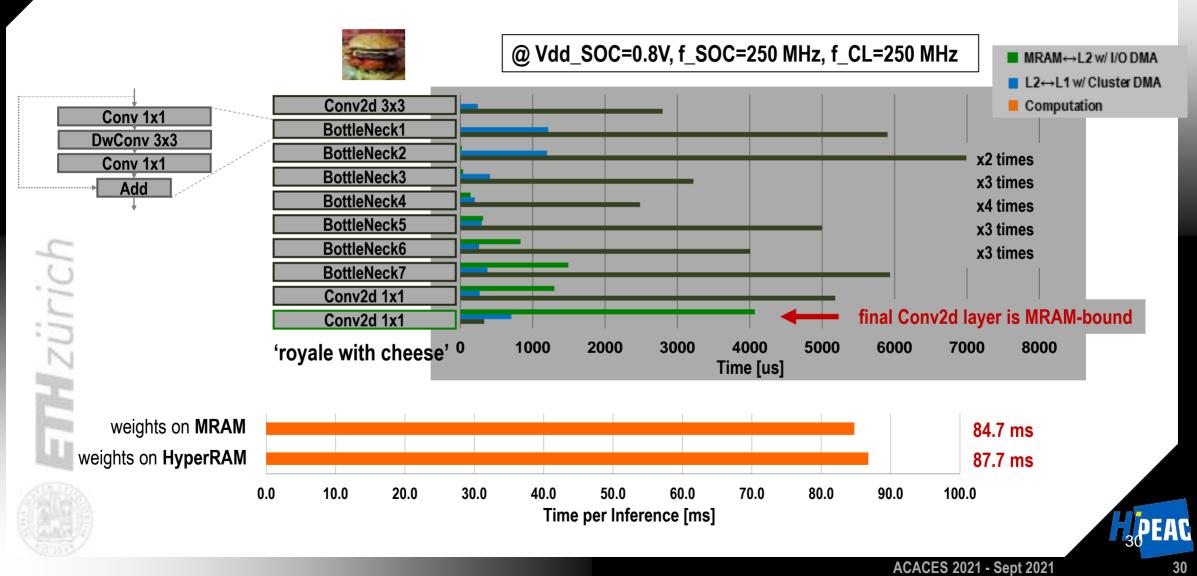
DNN Accelerator



ACACES 2021 - Sept 2021

Digital computing platforms for near-sensor processing at the extreme edge of the IoT

Full DNN Performance (MobileNetV2)

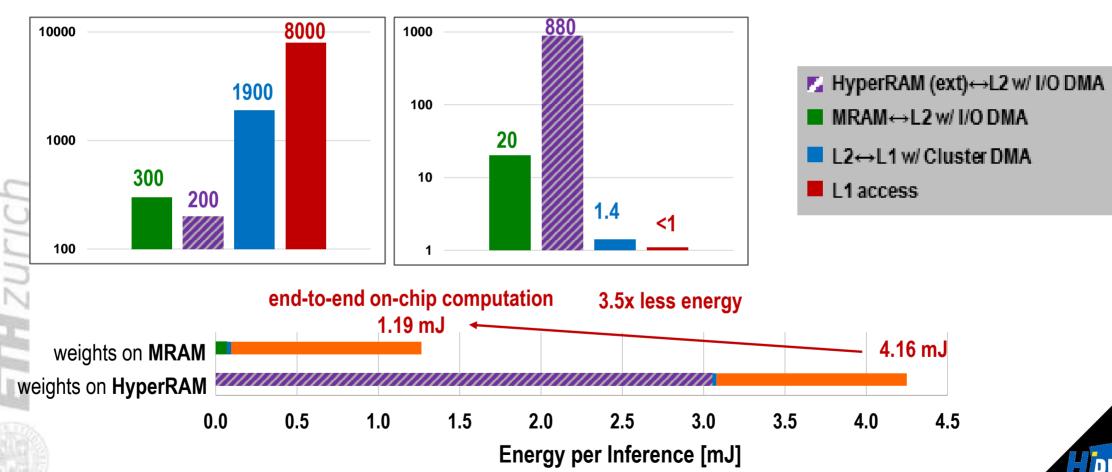


Digital computing platforms for near-sensor processing at the extreme edge of the IoT



Full DNN Energy (MobileNetV2)





Energy per byte [pJ/B]

World-Record Efficiency Among IoT Processors

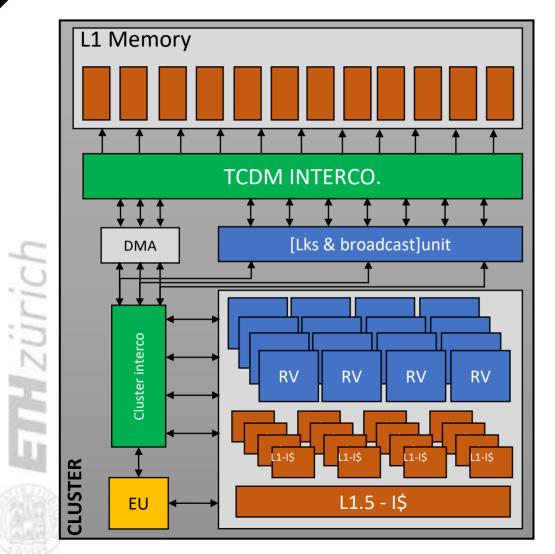
	SleepRunner [2]	Mr.Wolf [3]	SamurA' [4]	work)
Embedded NVM	-	-	-	MB MRAM
Wake-up Sources	WiC	GPIO, RTC	MuR, RTC, InC GPIO	GPIO, RTC ognitive
Best Int Perf.	31 MOPS (32b)	12.1 G	1.5 G02S	1 J GOPS
Best.Int Eff.	97 MOPS/mW (32b)	190 JPS/W	230 GC S/W	14 GOPS/W
@ Perf.	@ 18.6 MOPS (32b)	J.8 GOPS	MOPS	@7.6 GOPS
Best FP Perf.		1 GFLOPS		2 GFLOPS
Best FP Eff.	- (18 GFLOPS	s - /	79 GFLOPS/W
@Perf		2350 MFL 25		@ 1 GFLOPS
Best ML Perf.		8,	36 JPS	32.2 GOPS
Best ML Eff.	-		1 rOPS/W	1.3 TOPS/W
@Perf			3, 2.8 GOPS	@15.6 GOPS

3.2x Efficiency
(a) 2x Performance
4.3x Efficiency
(a) 2.8x Performance
Similar Efficiency
(a) 5.5x Performance

D. Rossi *et al.*, "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," *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, 2021, pp. 60-62



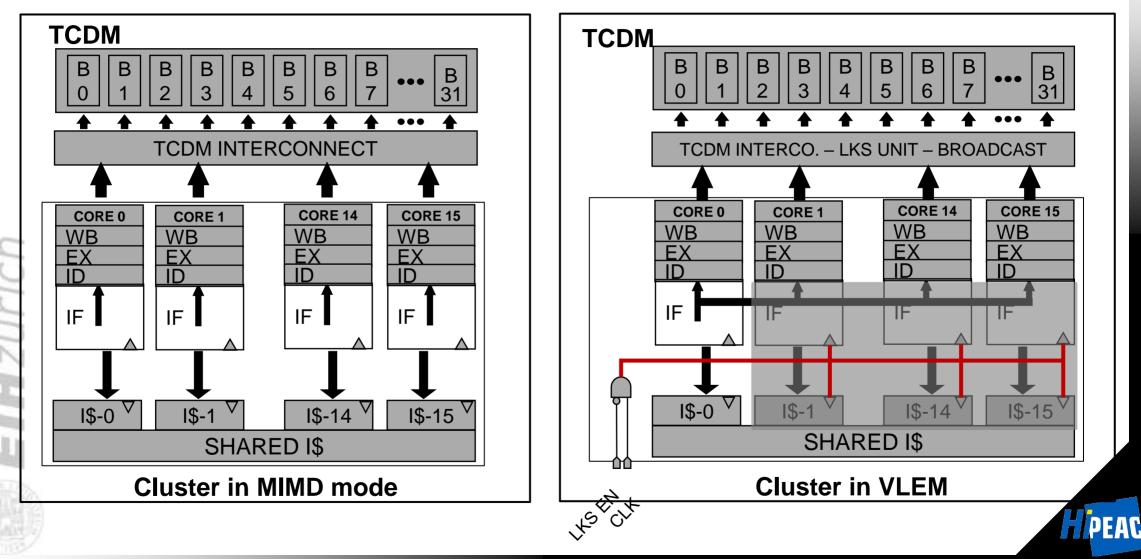
DUSTIN: Mixed-Precision Cluster



□ Accelerator Cluster

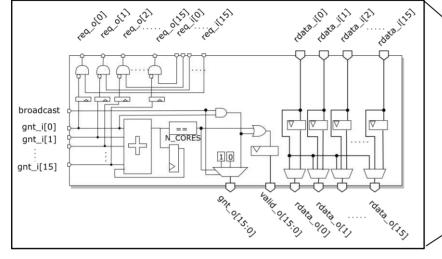
- 16 RI5CY (*) cores augmented with 2b-to-32b SIMD instructions;
- Software Configurable Vector Lockstep Execution Mode (VLEM);
- Single-cycle latency TCDM interco.
 leveraging a req/gnt protocol, word-level interleaved scheme.
- 128 kB of Shared Tightly-Coupled L1 Data Memory;
- Hierarchical Instruction Cache;
- □ High performance DMA (L2 <-> L1);
- Event Unit supporting efficient synchronization among the cores;

Vector Lockstep Exec. Mode (VLEM)

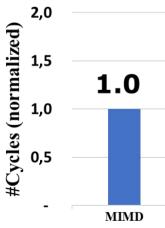


ACACES 2021 - Sept 2021

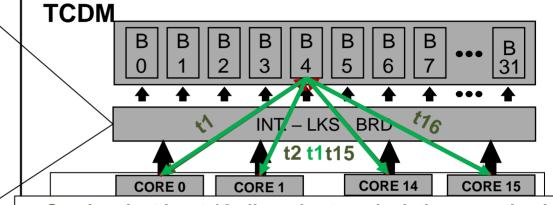
VLEM: Broadcast Unit



MatMul exec. kernels



Hzürich



Overhead: at least 16 clk cycles to unlock the execution in case of concurrent accesses;

Solution:

HE FOIT

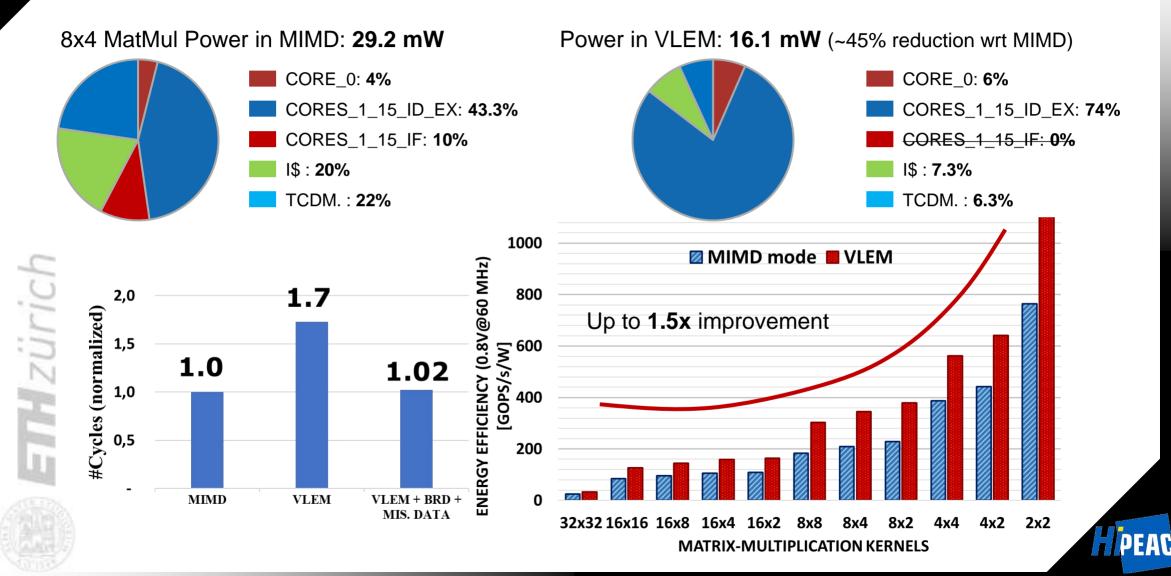
- eliminate the overhead in case of access to the same mem address → BROADCAST UNIT.
- Misalign static data and stacks to avoid accesses to the same mem bank;





ACACES 2021 - Sept 2021

Energy Efficiency on MatMul kernels



Comparison with the SoA

	SleepRunner [6]	SamurAI [7]	Mr.Wolf [8]	V ₂ a [9]	Dustin (this work)
Technology	CMOS 28nm FDSOI	CMOS 28nm FDSOI	CMOS 40nm LP	MOS 22nm FDSOI	CMOS 55nm
Die Area	0.68 mm ²	4.5 mm ²	10 mm ²	12mm ²	1 mm ²
Applications	IoT GP	IoT GP + DNN	IoT GP+1 AN	IoT GP 70ISA+DNN	I IoT GP DNN + QNNs
CPU/ISA	CM0DS Thumb-2 subset	1x RI5CY RVC32IMFXpulp	9 x 15CY RVC 2IMFXpulp		16 × PIC CORES (RISC-V)
Int Precision (bits)	32	8, 16, 32	8, 16, 32 0 8 (6)	6, 16, 32	2, 4, 8, 16, 32 (plus Mixed-Precision)
Supply Voltage	0.4 - 0.8 V	0.45 - 0.9	O S SINT	0.5 – 0.8 V	0.8 - 1.2 V
Max Frequency	80 MHz	350 /Hz	4.0 MHzC	450 M Z	205 MHz
Power Envelope	320 μW	€ mW	152 000	.4 mW	156 mW
¹ Best Integer Performance	31 MOPS (32b)	1.5 GOL ^C (8b) ²	12.1 GOPS (8b)	15.6 GOPS (8b)	15 GOPS (8b) 30 GOPS (4b) 58 GOPS (2b)
0	97 MOPS/mW @ 18.6 MOPS (32b)	230 GOPS/W @110 MOPS (8b)-	190 GOL /W 2 3.8 / JPS (8b)	614 GOPS/W @ 7.6 GOPS	303 GOPS/W @4.4 GOPS (8b) 570 GOPS/W @8.8 GOPS (4b) 1152 GOPS/W@17.3 GOPS(2b)

Dustin supports Mixed-Precision computation in HW

Better efficiency wrt solutions in 28nm and 40nm tech node

Comparable efficiency wrt Vega (22 nm)



A.Garofalo et. al. "A 1.15 TOPS/W, 16-Cores Parallel Ultra-Low Power Cluster with 2b-to-32b Fully Flexible Bit-Precision and Vector Lockstep Execution Mode", ESSCIRC 2021

PEAC

Hzürich

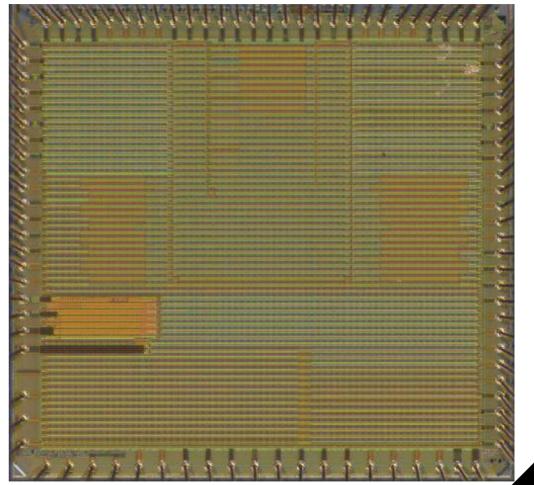
Moving to HPC: Kosmodrom

Globalfoundries 22FDX

- In 2018, most advanced node for us
- Minimum size 3mm x 3mm
 - That fits about 100 million transistors
- Allows body biasing

• With great power comes...

- Designs in 22FDX are more involved
- More blocks, more functionality
 - More things that can go wrong
- Challenging design
- Collaboration with Globalfoundries

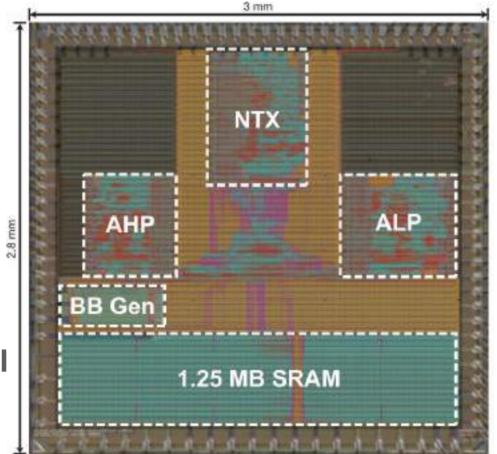


Kosmodrom: Main components

- 2x Ariane 64b RISC-V cores
 - AHP optimized for high speed
 - ALP optimized for low power
- Automatic Body Bias Gen.
 - IP by INVECAS

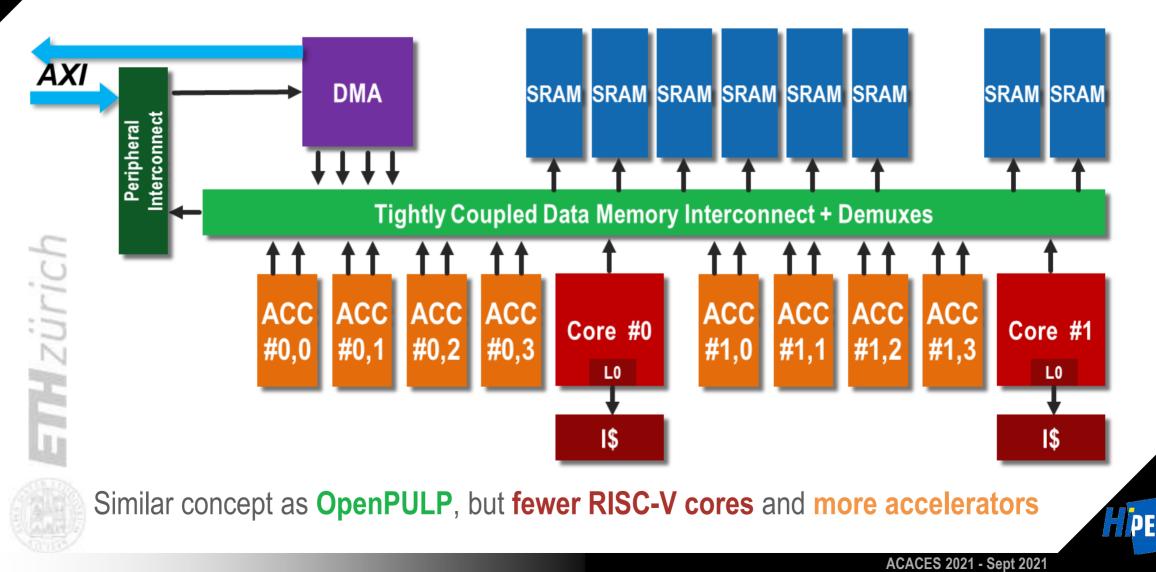
zürich

- Allows body bias to be tuned
- NTX: Neural Training Accelerator
 - 260 Gflops/Watt efficiency
- Common infrastructure
 - SRAM, Debug, I/Os





Fine-Grained Shared-Memory Accelerators



40

NTX uses 1 RISC-V core to control 8 units

- NTX runs at up to 1.25 GHz
- Compute of 20 Gflop/s
- Bandwidth of 5 GB/s
- At 9.3 pJ/flop and using only 0.51 mm2
- Scale up by replicating cluster

coprocessors

1x RISC-V processor and peripherals

F. Schuiki, M. Schaffner, F. K. Gürkaynak and L. Benini, "A Scalable Near-Memory Architecture for Training Deep Neural Networks on Large In-Memory Datasets," in IEEE Transactions on Computers, vol. 68, no. 4, pp. 484-497, 1 April 2019, doi: 10.1109/TC.2018.2876312.

816µm

-2 kB ICACHE

64 kB TCDM

in 32 banks

Kosmodrom Demonstration Board

STM microcontroller for control

Test socket for Kosmodrom chip

H DOW 100 800 A A CLK SEL **USB** connection to computer 12.76B TRG SMUDDUM **H** zürich MOTABLE UNLITAGES C DOWNEY Analog to Digital Measurement points for Supply voltage Body bias voltage **Converter module**

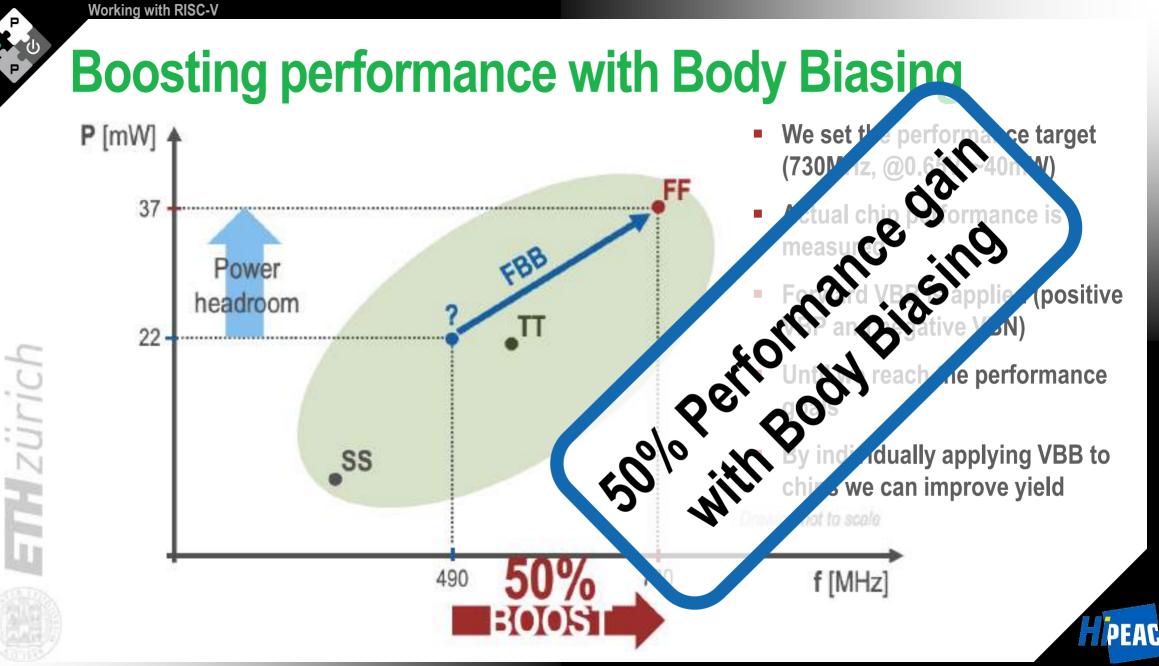
generation

generation

all supplies



ACACES 2021 - Sept 2021





Gaining Energy Efficiency with Body Biasing

- f [MHz] We set the desired operating frequency (800MHz)
- We decrease the voltage to the minimum level chip will wa (0.8V)
- At this point we sta voltage further

Hzürich

- requescy JOMHO Ate for Ne J me performance **Maximum** will also 20% less power
- We pe **BN** (pos negati
- Until we operating f

V_{DD} [V]

PEAC

44

0.80

Target performance

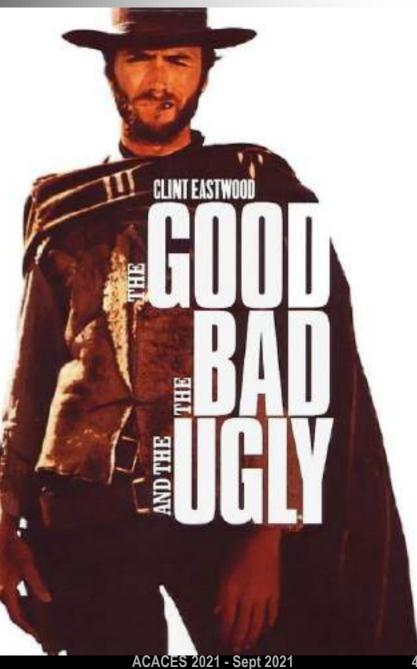
Reduce VDD

ly FBB

0.65

The good the bad and the ugly

- We designed and tested 43 chips as part of PULP project
- Most worked great
- But there were also mistakes
- Here is a look at some highs and some lows



ETH zürich

Hzürich

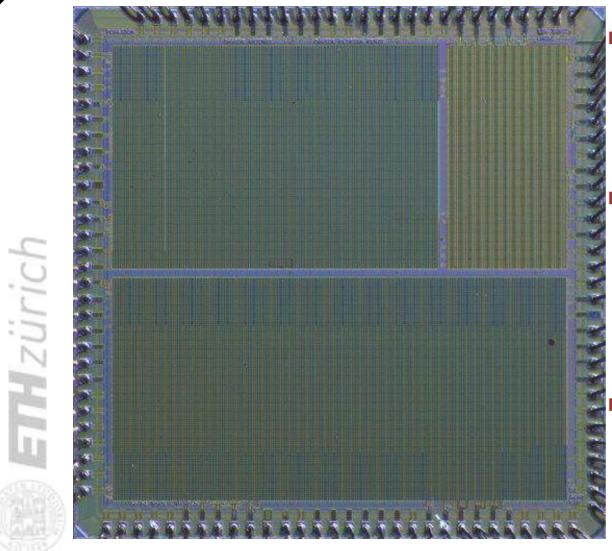
Good: Fulmine the award winning one

- IEEE Circuits and systems, Best IEEE Circuits Award 2020 IEEE Circuits Award 2020
 - UMC65
 - Earlier chip (2015)
 - 4x OpenRISC cores (not yet RISC-V)
 - 192 kBytes L2 + 64 kBytes TCDM
 - 2x HW accelerators
 - HW Crypt (together with TU-Graz)
 - HW Convolution Engine

Publication from this chip

F. Conti et. Al., "An IoT Endpoint System-on-Chip for Secure and Energy-Efficient Near-Sensor Analytics", IEEE Transactions on Circuits and Systems I: Regular Papers, Vol: 64, Issue: 9, Sept. 2017,pp 2481 – 2494"

Bad: Bonding issues on Poseidon



First GF22nm chip

- Used Europractice IC service
- Cost 150k CHF for 50 samples

Has three parts (trident..)

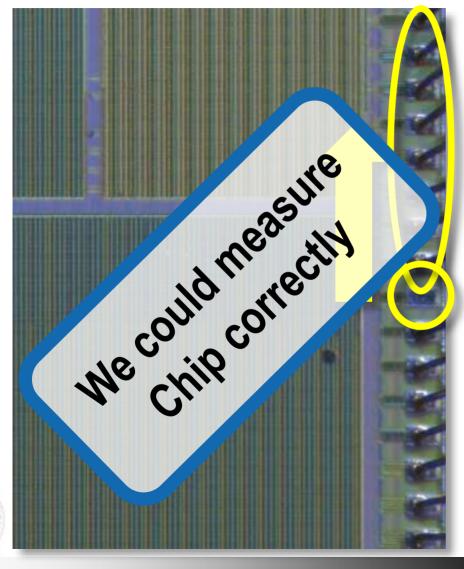
- PULPissimo system
- Ariane core
- Independent ML accelerator

30 of 50 chips were packaged

- We provide a bonding diagram
- Mostly simple manual work

nh zürich

Bad: Bonding issues on Poseidon



Look closer on the right side

There is a pad that is not bonded

We skipped one pad

All connections are shifted by one

VDD and GND are one after other

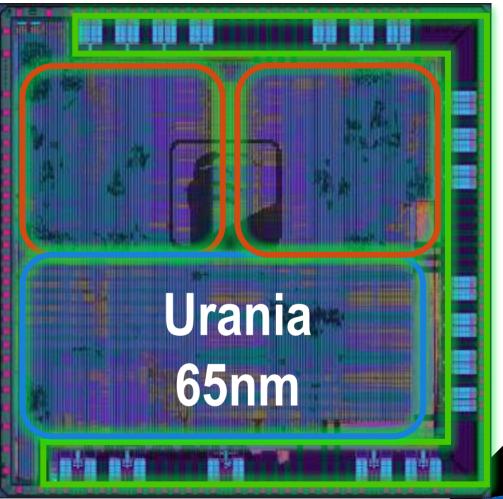
- Bonding causes shorts between VDD and GND
- Pretty much catastrophic.
- Fortunately: unpackaged dies
 - There were 20 unpackaged dies
 - We could bond those correctly

48

Downright Ugly, reset problem of Urania

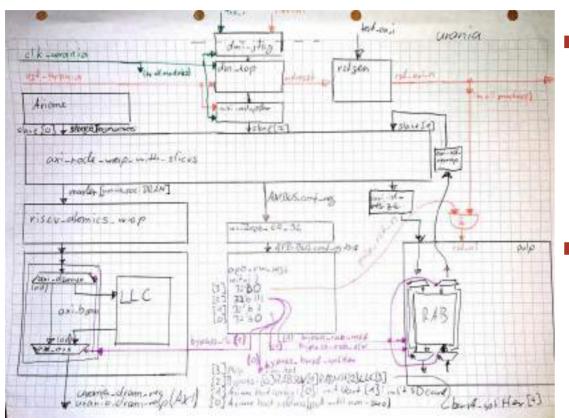
2 PULP clusters, each with

- 4x RV32 RI5CY cores
- 4x transprecision FPUs
- 1x PULPO accelerator
- 64 kB TCDM in 8 banks
- Ariane RV64 host processor
 - 128 KiB Shared LLC
 - software-managed IOMMU
- DDR3 DRAM Controller + PHY by TUKL



zürich

The reset can not be released for clusters



- Chip has many modules
 - 1x Ariane core
 - 1x DDR interface
 - 2x Clusters

Reset to clusters is stuck 0

- Design flow mistake
- Some other control signals are stuck as well affecting Ariane performance
- DDR interface is functional
 - Not everything is lost





: Hzürich

ETH zürich

IC Design is tricky and demands attention

- Even the simplest things can derail a complex chip
 - A copy paste error in a bonding diagram, a mistake in reset
- Academic research chips are not industrial products
 - Designed to test and verify ideas, not mass production
 - Much more effort needed in DfT and verification to make a successful product

Experience is key in IC Design

- All the mistakes we make, add to our future success
- Some lessons you learn the hard way
- But these stay with you and help you for your future designs



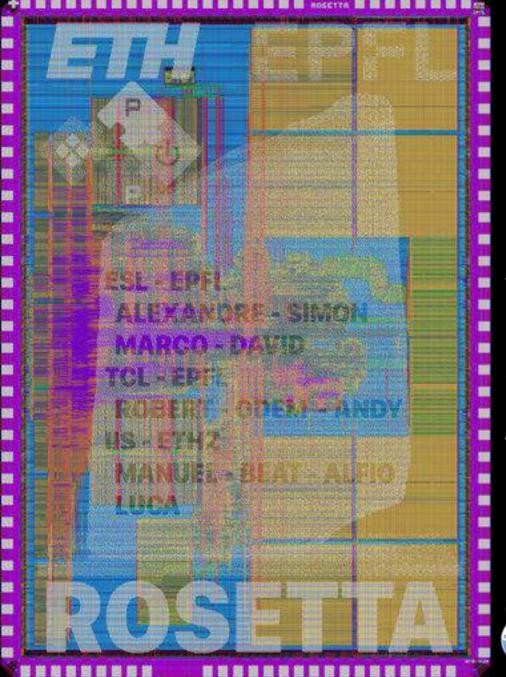
We hope this was helpful/fun for you

Covered the basics of RISC-V

- Explained the ISA
- Examples of Implementations
- Advanced cores and Concepts
- Talked about building open source systems around RISC-V
 - Showed the main concepts and talked about our ICs
- You can find PULP related information
 - GitHub:

ETH zürich

- PULP Webpage:
- () http://github.com/pulp_platform
 - http://pulp-platform.org
- Follow us on Twitter: <u>@pulp_platform</u>



Parallel Ultra Low Power

Luca Benini, Davide Rossi, Andrea Borghesi, Michele Magno, Simone Benatti, Francesco Conti, Francesco Beneventi, Daniele Palossi, Giuseppe Tagliavini, Antonio Pullini, Germain Haugou, Manuele Rusci, Florian Glaser, Fabio Montagna, Bjoern Forsberg, Pasquale Davide Schiavone, Alfio Di Mauro, Victor Javier Kartsch Morinigo, Tommaso Polonelli, Fabian Schuiki, Stefan Mach, Andreas Kurth, Florian Zaruba, Manuel Eggimann, Philipp Mayer, Marco Guermandi, Xiaying Wang, Michael Hersche, Robert Balas, Antonio Mastrandrea, Matheus Cavalcante, Angelo Garofalo, Alessio Burrello, Gianna Paulin, Georg Rutishauser, Andrea Cossettini, Luca Bertaccini, Maxim Mattheeuws, Samuel Riedel, Sergei Vostrikov, Vlad Niculescu, Hanna Mueller, Matteo Perotti, Nils Wistoff, Luca Bertaccini, Thorir Ingulfsson, Thomas Benz, Paul Scheffler, Alessio Burello, Moritz Scherer, Matteo Spallanzani, Andrea Bartolini, Frank K. Gurkaynak,

and many more that we forgot to mention

http://pulp-platform.org



@pulp_platform