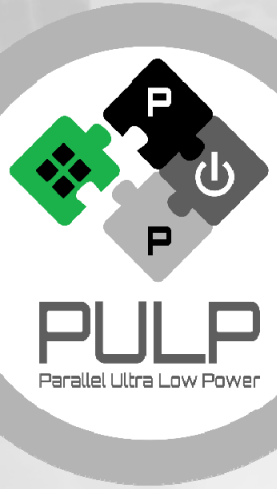


Energy-Efficient Heterogeneous Design

Perugia *04.09.2019*

Luca Benini[1,2]

[1]*Department of Electrical, Electronic and Information Engineering*

[2]*Integrated Systems Laboratory*

European Research Council

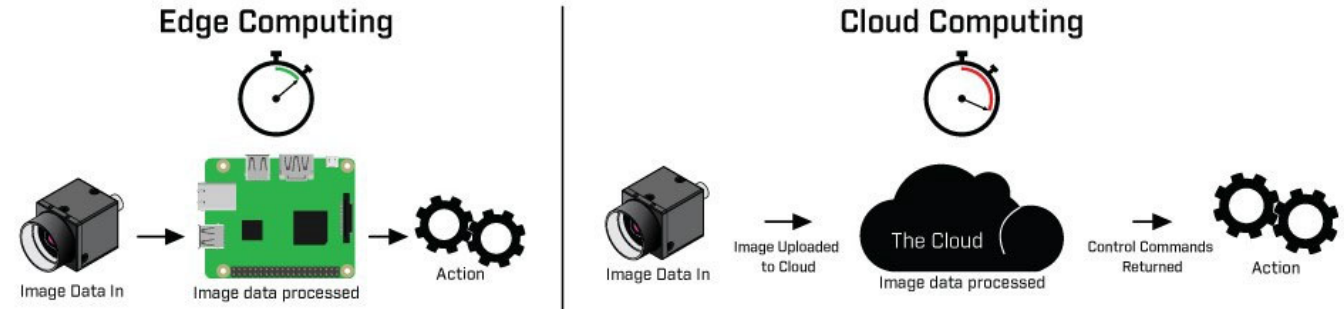
European Commission | Horizon 2020 European Union funding for Research & Innovation

nano-tera.ch

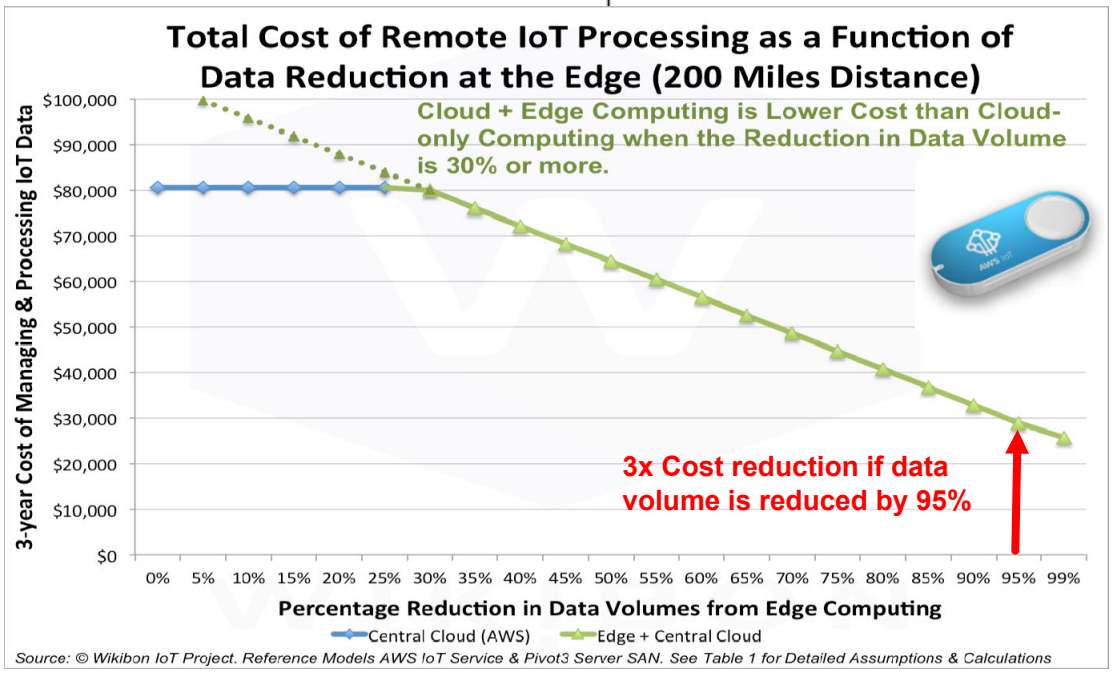
FNSNF Fonds National Suisse Schweizerischer Nationalfonds Fondo Nazionale Svizzero Swiss National Science Foundation

Cloud → Edge → Extreme Edge

Latency, Privacy



Cost



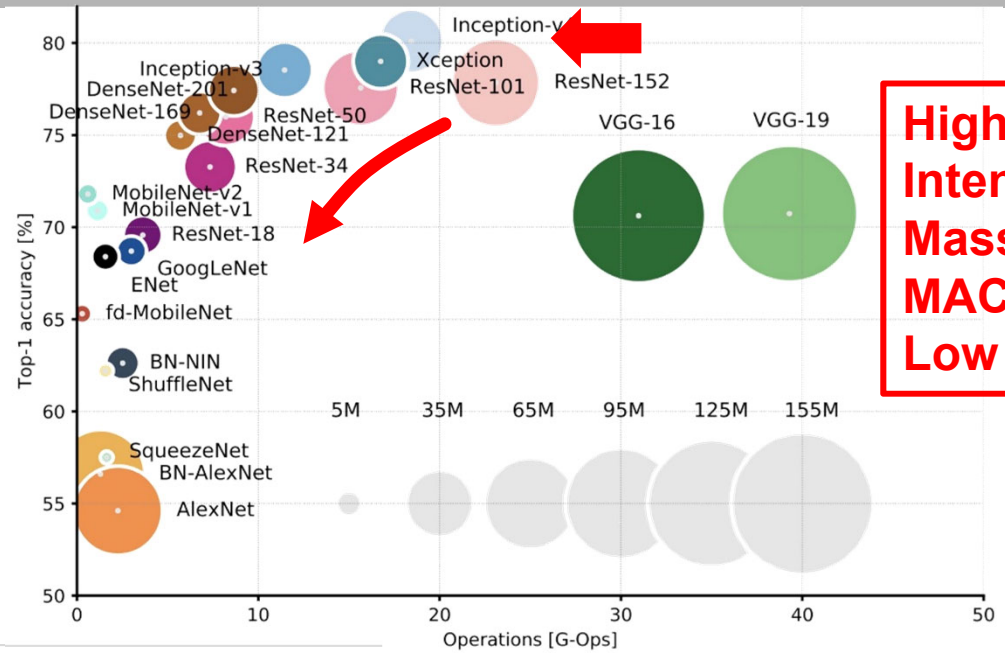
Extreme edge AI challenge

AI capabilities in the power envelope of an MCU: 100mW and below

PULP

AI Workloads from Cloud to Edge (Extreme?)

GOP+



High Computational Intensity
Massive Parallelism,
MAC-dominated
Low precision OK

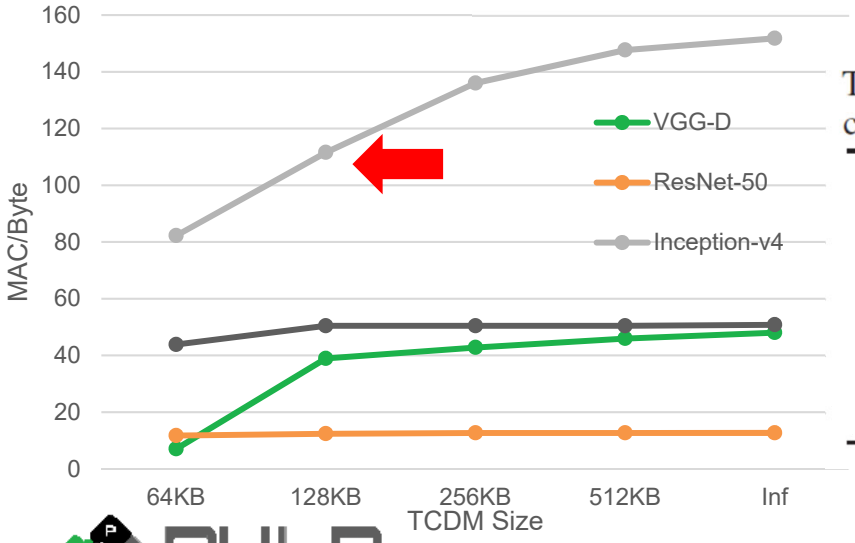


Table 5: Batch-normalized Inception top-1 validation accuracy % and compute cost as precision of activations (A) and weights (W) varies.

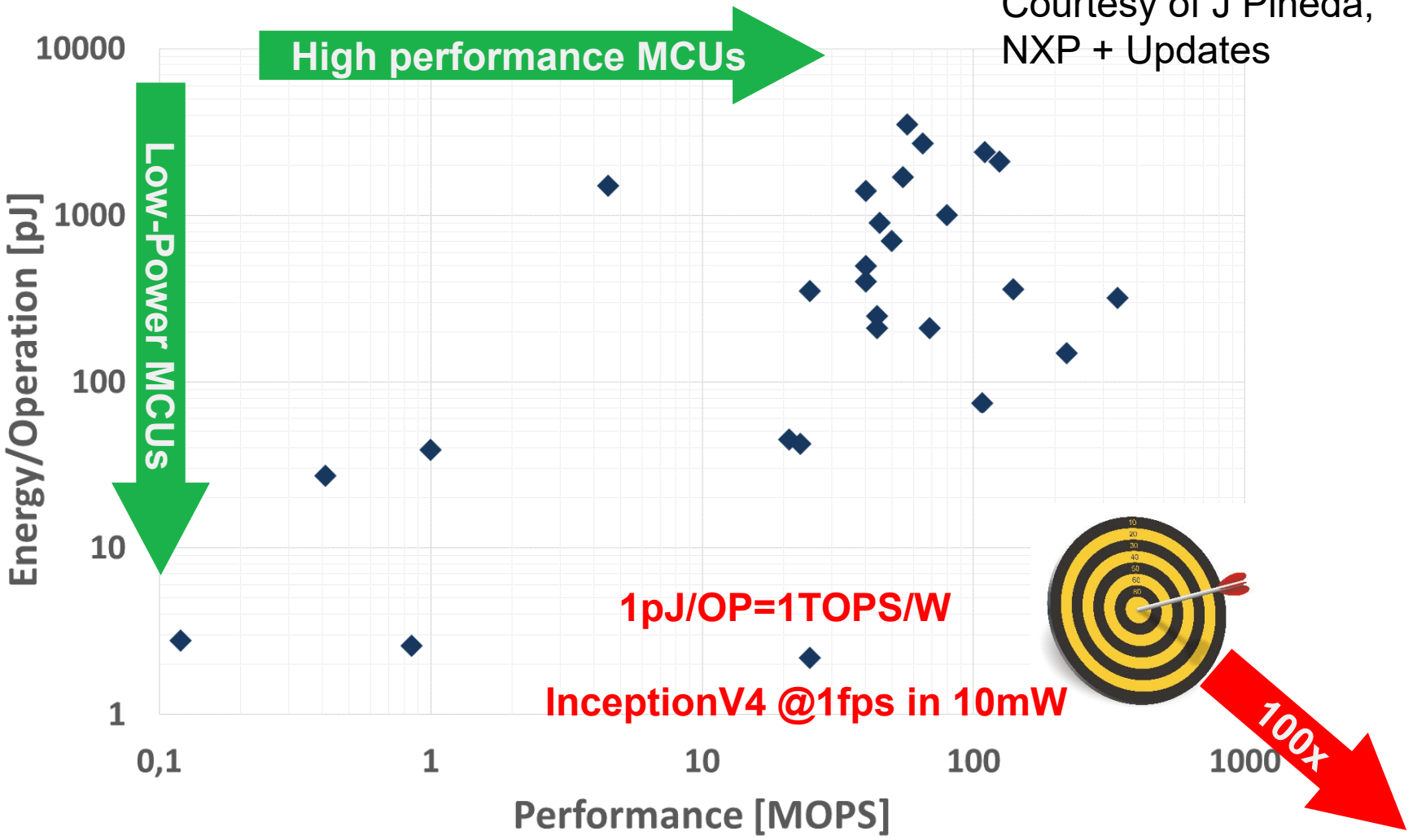
Width	Precision	Top-1 Acc. %	Compute cost
1x wide	32b A, 32b W	71.64	1x
2x wide	4b A, 4b W	71.63	0.50x
	4b A, 2b W	71.61	0.38x
	2b A, 2b W	70.75	0.25x
	1b A, 1b W	65.02	0.13x

[WRPN:arXiv:1709.01134v1]

PULP

Energy efficiency is THE Challenge

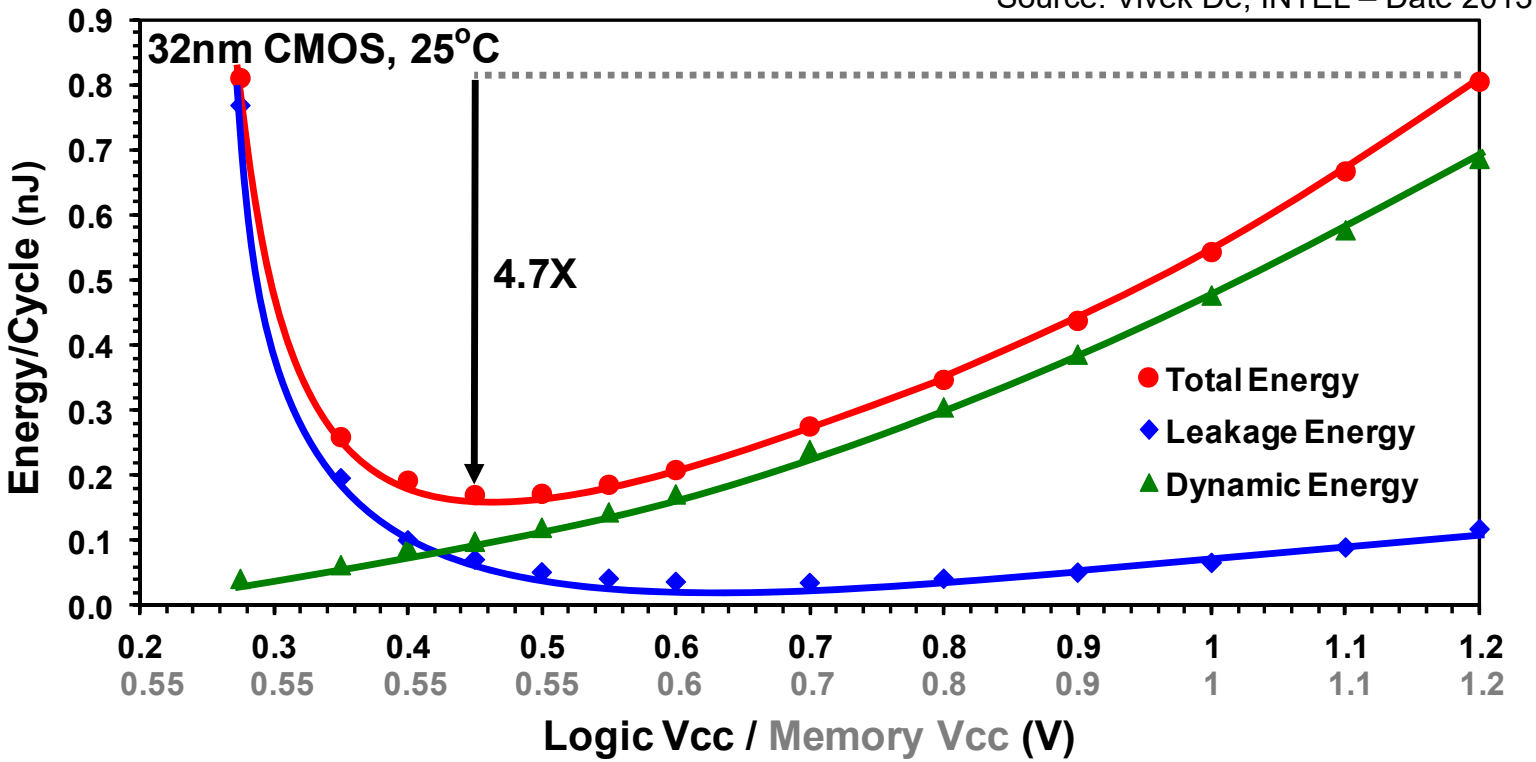
Courtesy of J Pineda, NXP + Updates



Cool… But, HOW??

2013: Parallel Ultra Low Power →PULP!

Source: Vivek De, INTEL – Date 2013

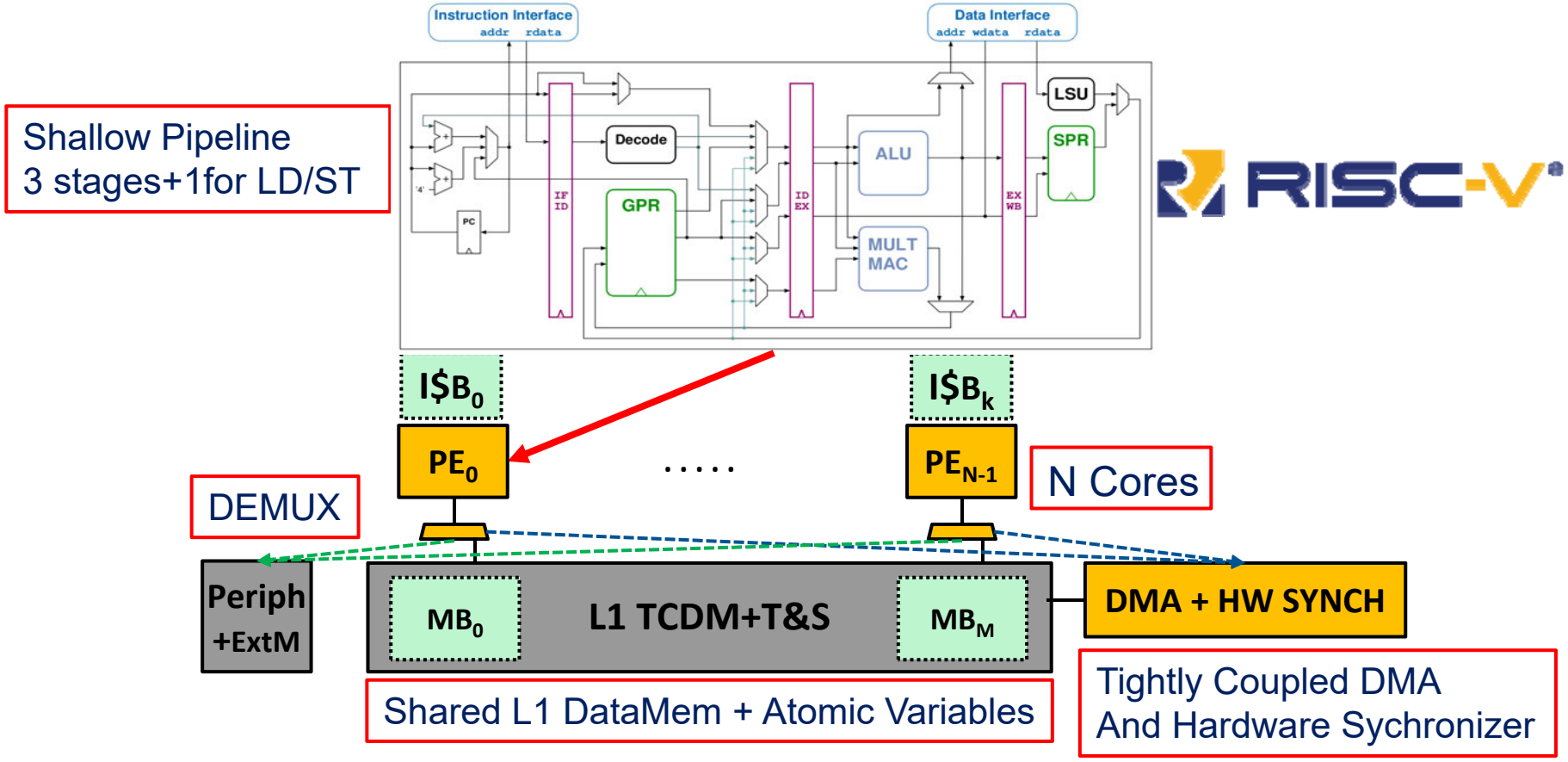


Near-Threshold Computing (NTC):

1. **Don't waste energy pushing devices in strong inversion**
2. **Recover performance with parallel execution**
3. **Manage Leakage, PVT variability and SRAM limitations NT!!!**

PULP

Near-Threshold Multiprocessing



Need Strong ISA, Need full access to “deep” core interfaces, need to tune pipeline!
OPEN ISA: **RISC-V** RV32IMC + **New, Open Microarchitecture → RI5CY!**

PULP

D. Rossi *et al.*, "Energy-Efficient Near-Threshold Parallel Computing: The PULPv2 Cluster," in *IEEE Micro*, Sep./Oct. 2017.

Bespoke ISA needed! Enter Xpulp extensions

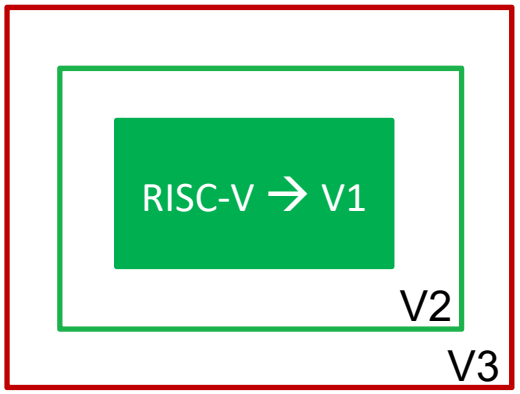
<32-bit precision → SIMD2/4 → x2,4 efficiency & memory size

Risc-V ISA is extensible *by construction* (great!)

V1 Baseline RISC-V RV32IMC
HW loops

V2 Post modified Load/Store
Mac

V3 SIMD 2/4 + DotProduct + Shuffling
Bit manipulation unit
Lightweight fixed point **(EML centric)**



25KG → 40KG (1.6x)

M. Gautschi et al., "Near-Threshold RISC-V Core With DSP Extensions for Scalable IoT Endpoint Devices," in IEEE TVLSI, Oct. 2017.

RI5CY – are xPULP ISA Extensions (1.6x) worthwhile?

```
for (i = 0; i < 100; i++)
    d[i] = a[i] + b[i];
```

10x on 2d convolutions …YES!

Baseline

```
mv    x5, 0
mv    x4, 100
Lstart:
  lb     x2, 0(
  lb     x3, 0(
  addi   x10,x1
  addi   x11,x1
  add    x2, x3
  sb     x2, 0(
  addi   x4, x4
  addi   x12,x1
bne      x4, x5
```

11 cycles/output

Auto-incr load/store

```
mv    x5, 0
mv    x4, 100
Lstart:
  lb    x2, 0(
  lb    x3, 0(
  addi  x4, x4
  add   x2, x3
  sb    x2, 0(
bne     x4, x5, Lstart
```

8 cycles/output

HW Loop

```
lp.setupi 100, Lend
  lb    x2, 0(x10!)
  lb    x3, 0(x11!)
  add   x2, x3, x2
Lend: sb x2, 0(x1
```

5 cycles/output

Packed-SIMD

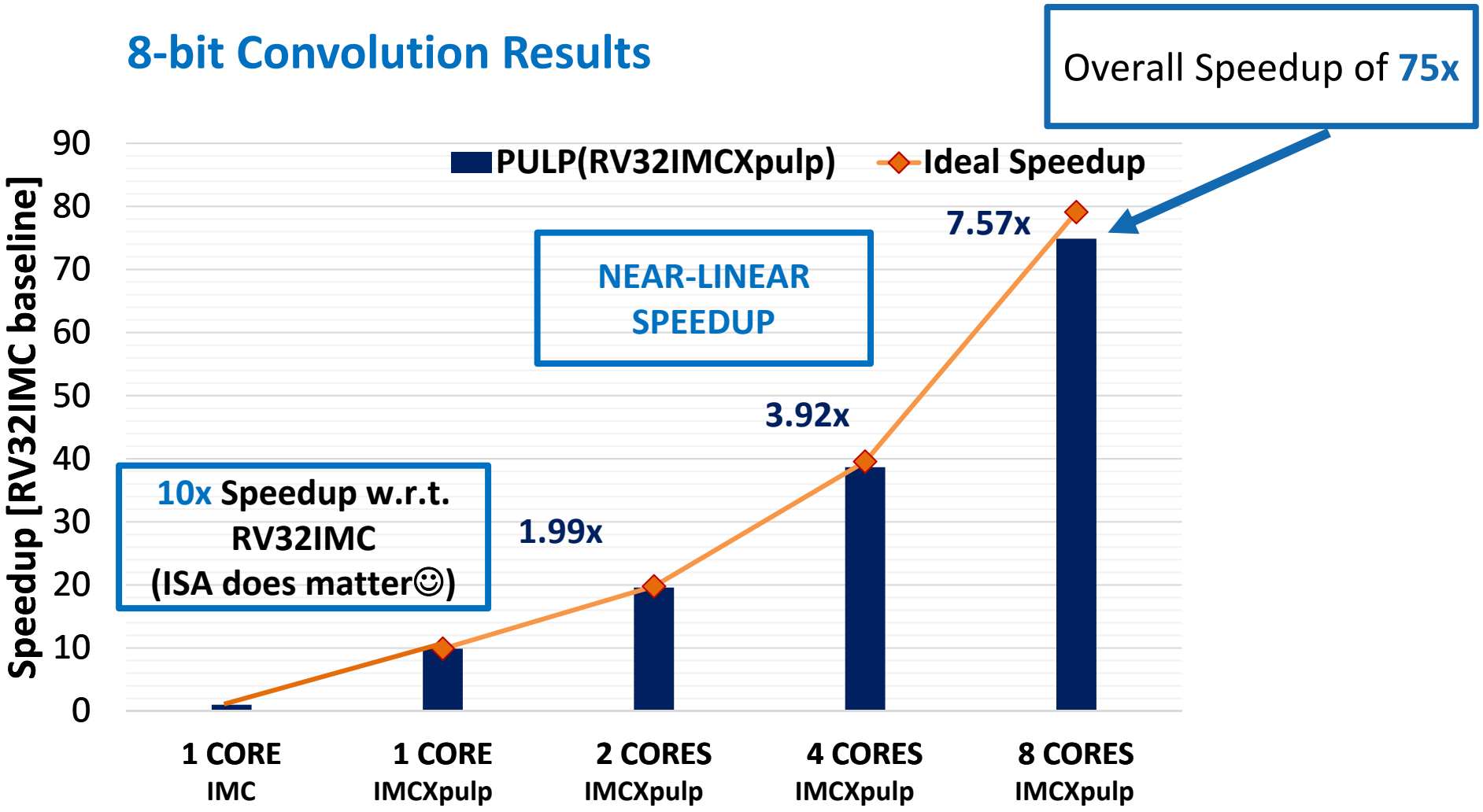
```
lp.setupi 25, Lend
  lw    x2, 0(x10!)
  lw    x3, 0(x11!)
  pv.add.b x2, x3, x2
Lend: sw x2, 0(x12!)
```

1,25 cycles/output

PULP

Results: RV32IMCXpulp vs RV32IMC

8-bit Convolution Results

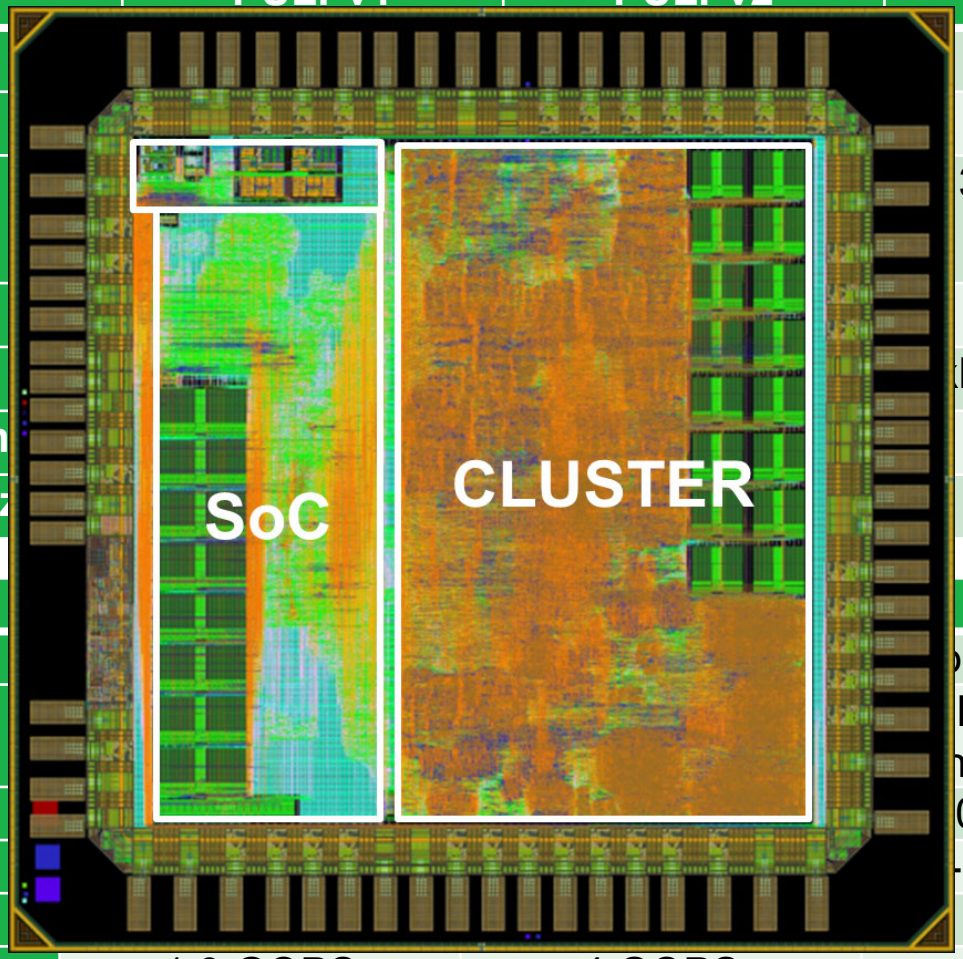


PULP-NN: an open Source library for DNN inference on PULP cores

PULP

The Evolution of the 'Species'

	PULPv1	PULPv2	PULPv3
# of cores			4
L2 memory			128 kB
TCDM			32kB SRAM 16kB SCM
DVFS			yes
I$			kB SCM shared
DSP Extension			yes
HW Synchroniz			yes



	PULPv3
Status	post tape out
Technology	FD-S nventional well
Voltage range	0.5V - 0.7V
BB range	-1.8V - 0.9V
Max freq.	200 MHz
Max perf.	1.8 GOPS
Peak en. eff.	385 GOPS/W

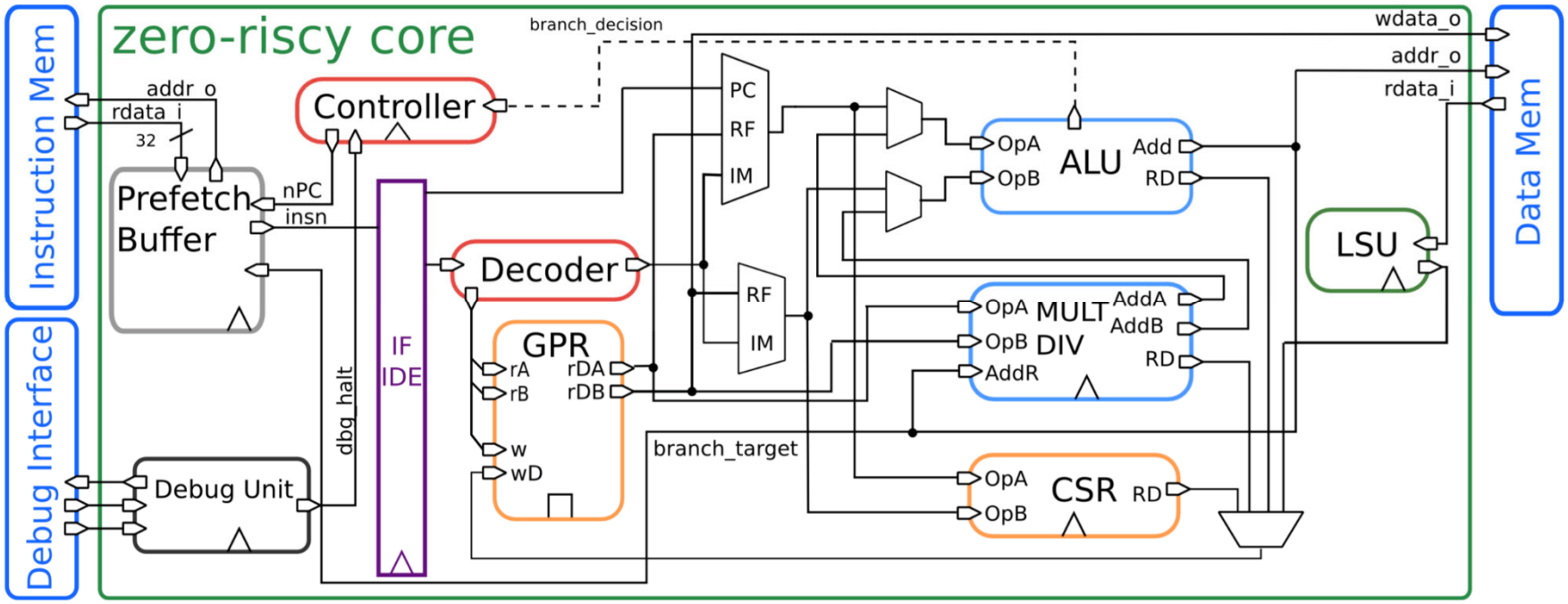
2.6pj/OP

	PULPv1	PULPv2
Max perf.	1.9 GOPS	4 GOPS
Peak en. eff.	60 GOPS/W	135 GOPS/W

life.augmented

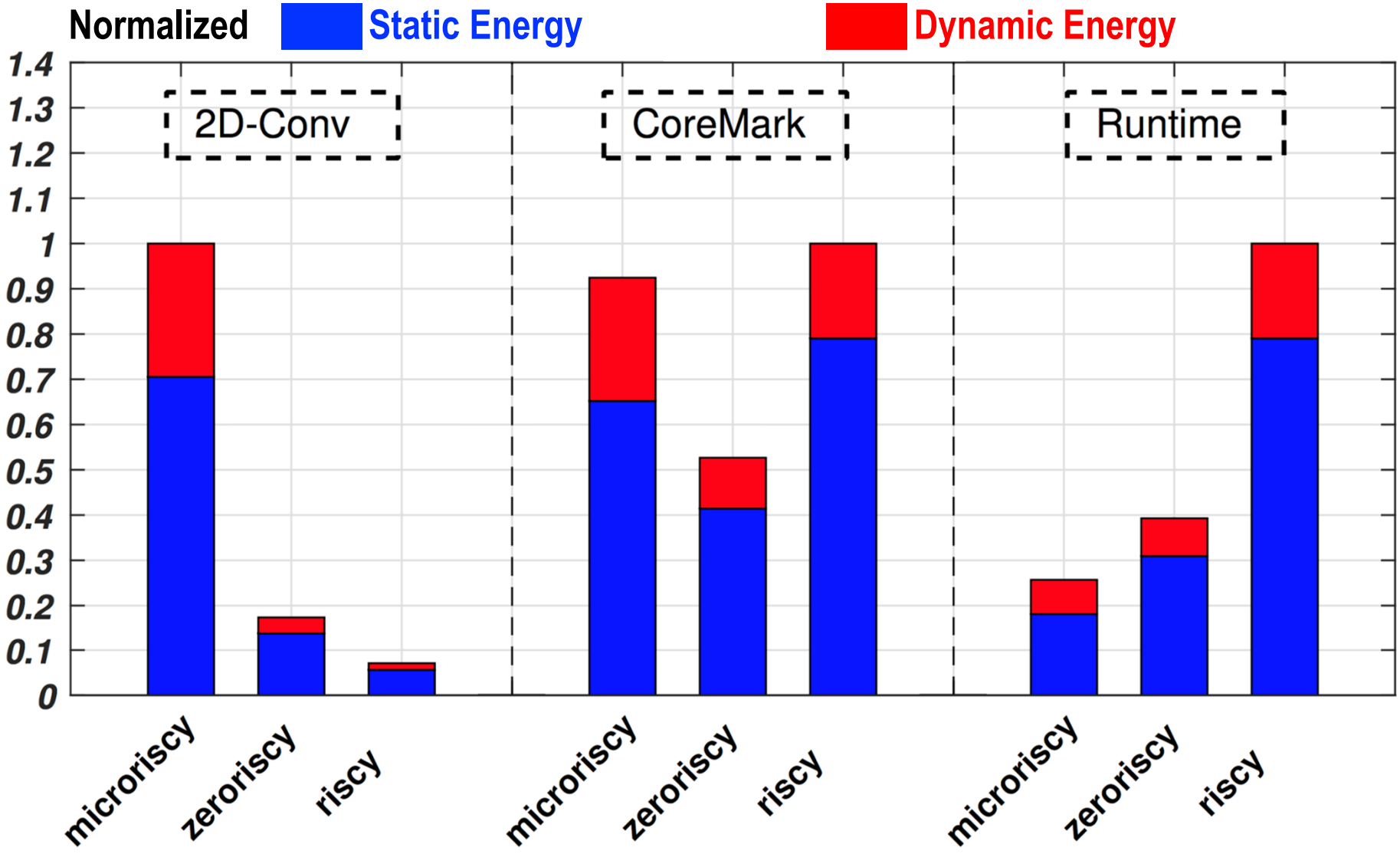
PULP

Enter Zero/Micro-riscy, small core for control



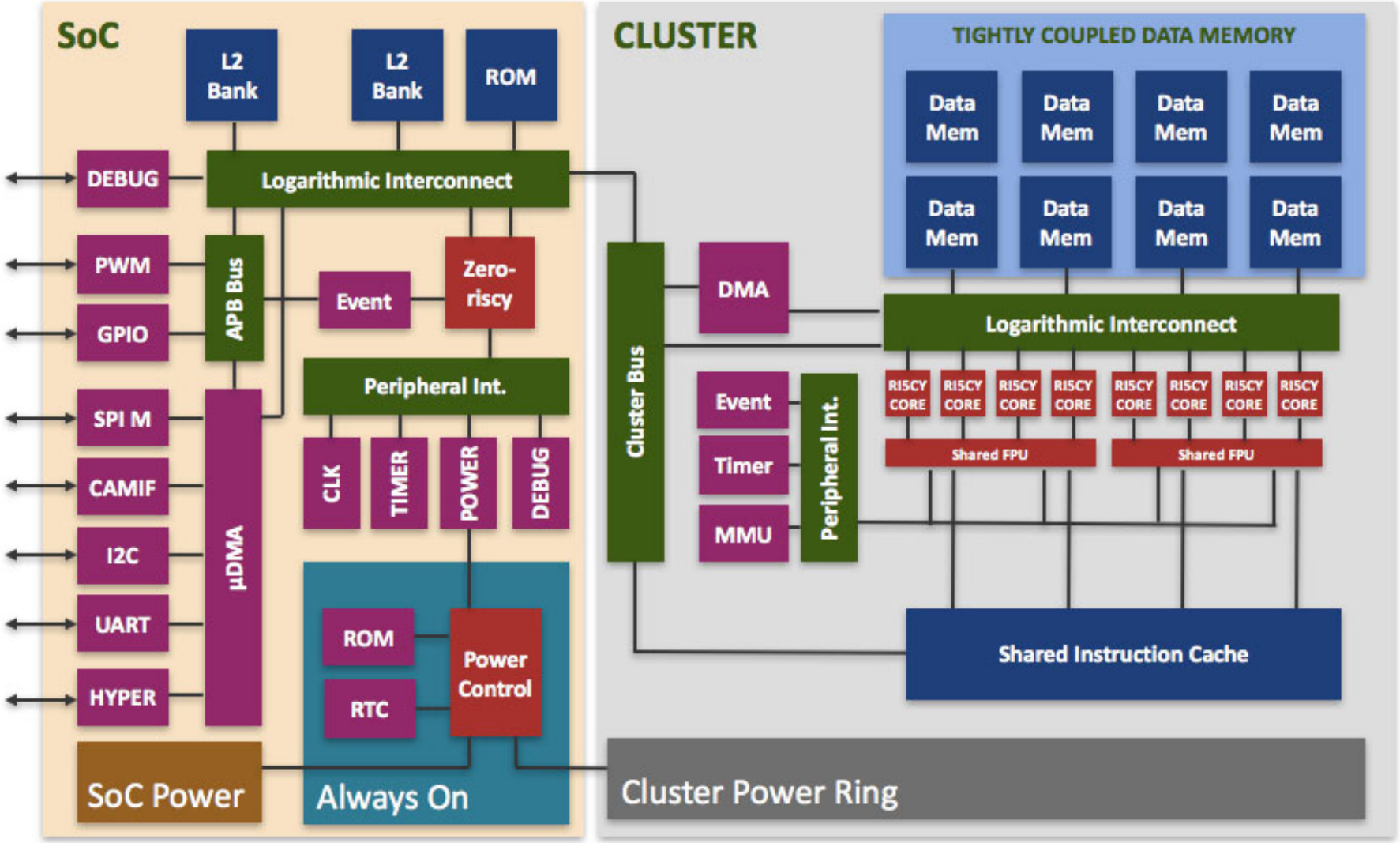
- Only 2-stage pipeline, simplified register file
- **Zero-Riscy** (RV32-ICM), 19kGE, 2.44 Coremark/MHz
- Micro-Riscy (RV32-EC), 12kGE, 0.91 Coremark/MHz
- Used as SoC level controller in newer PULP systems

Different cores for different types of workload



PULP

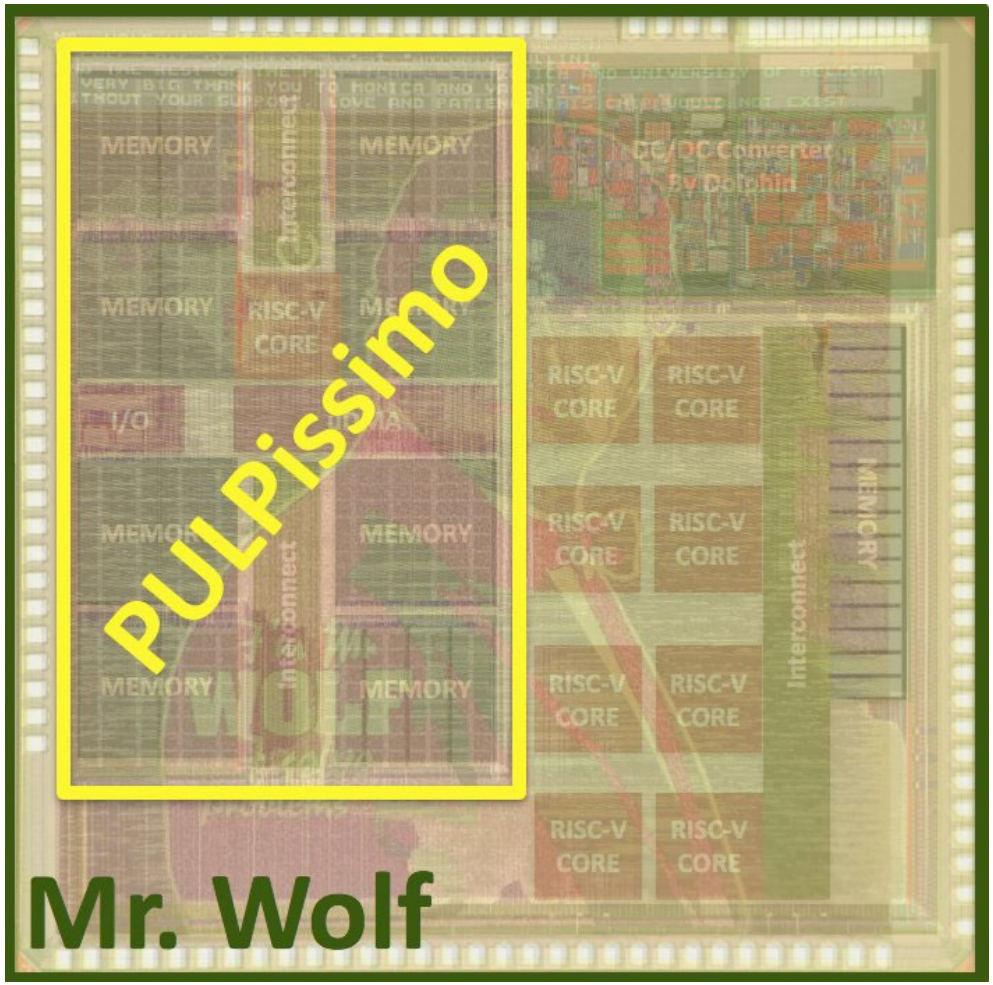
The IoT Processor: Mr Wolf



Mr. Wolf Chip Results: Heterogeneous Computing Works

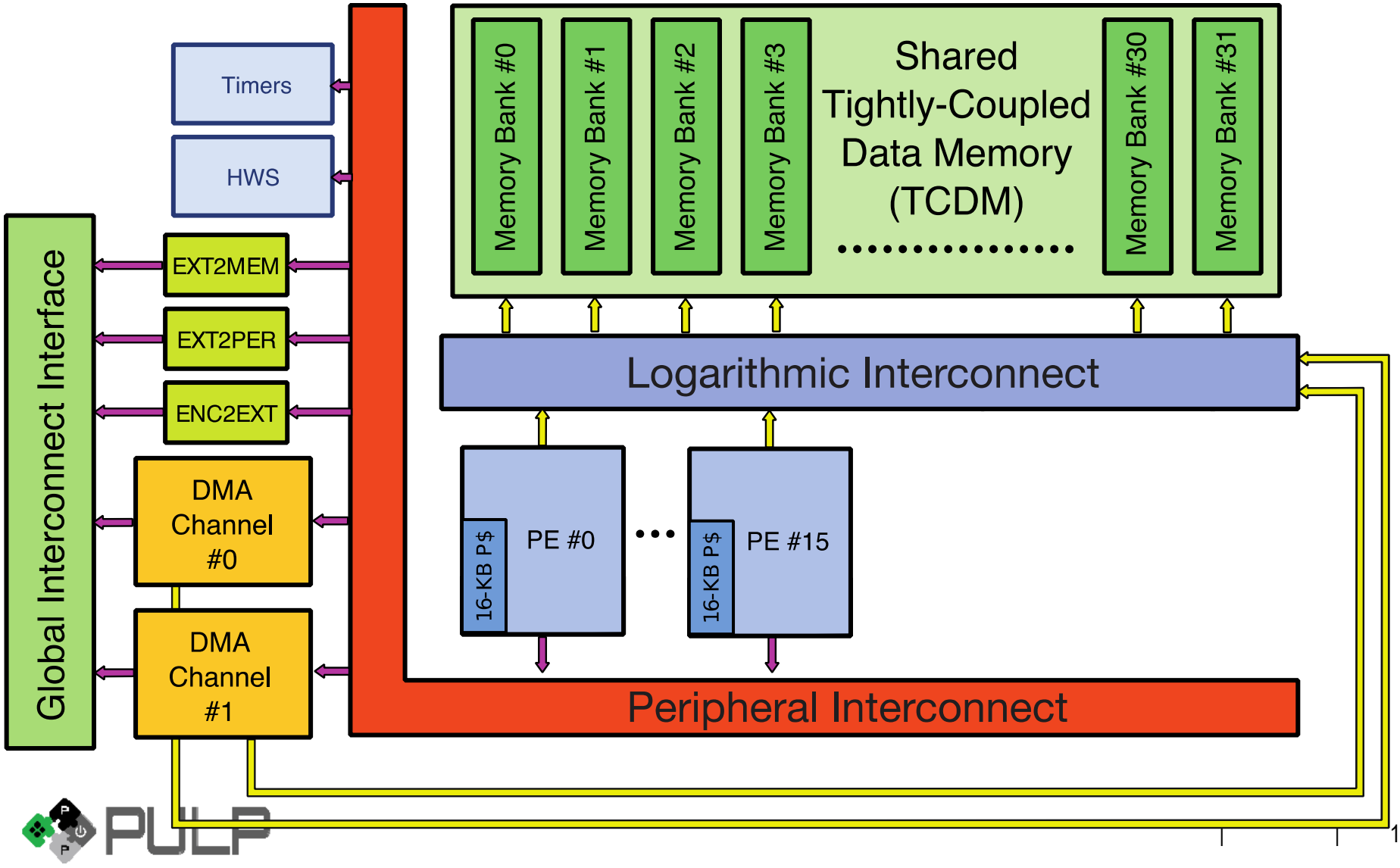
Technology	CMOS 40nm LP
Chip area	10 mm^2
VDD range	0.8V - 1.1V
Memory Transistors	576 Kbytes
Logic Transistors	1.8 Mgates
Frequency Range	32 kHz – 450 MHz
Power Range	72 µW – 153 mW

Power Managent (DC/DC + LDO)	VDD [V]	Freq.	Power
Deep Sleep	*0.8*	*n.a.*	*72 µW*
Ret. Deep Sleep	*0.8*	*n.a*	*76.5 - 108 mW*
SoC Active	*0.8 - 1.1*	*32 kHz – 450 MHz*	*0.97 - 38 mW*
Cluster Active	*0.8 - 1.1*	*32 kHz – 350 MHz*	*1.6 - 153 mW*



A. Pullini, D. Rossi, I. Loi, A. Di Mauro, L. Benini, "Mr.Wolf: a 1 GFLOP/S Energy-Proportional Parallel Ultra Low Power SoC for IoT Edge Processing", ESSCIRC 2018.

More efficiency: Heterogeneous PULP Cluster

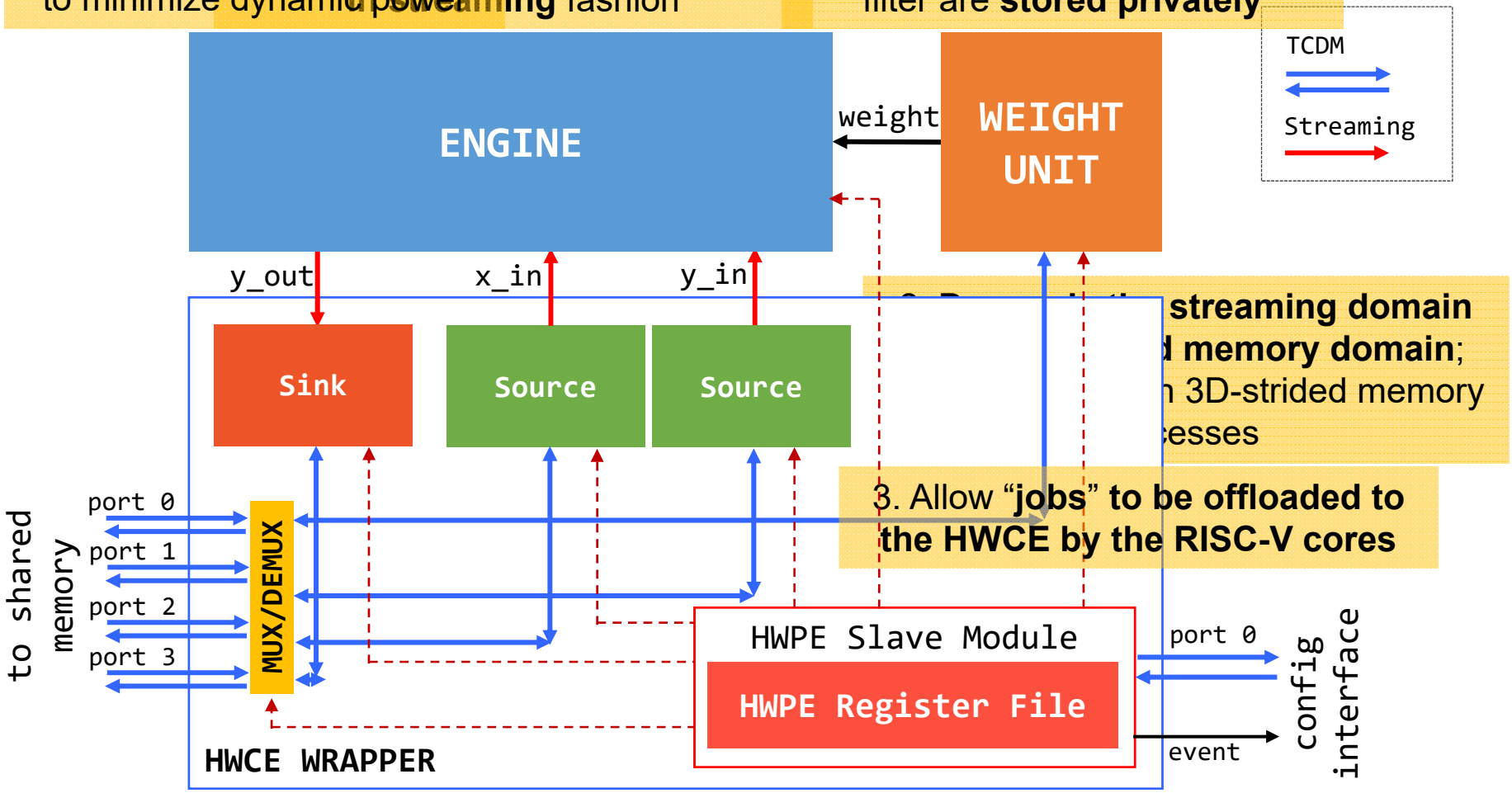


HW Convolution Engine

5. **Fine-grain clock gating** to minimize dynamic power

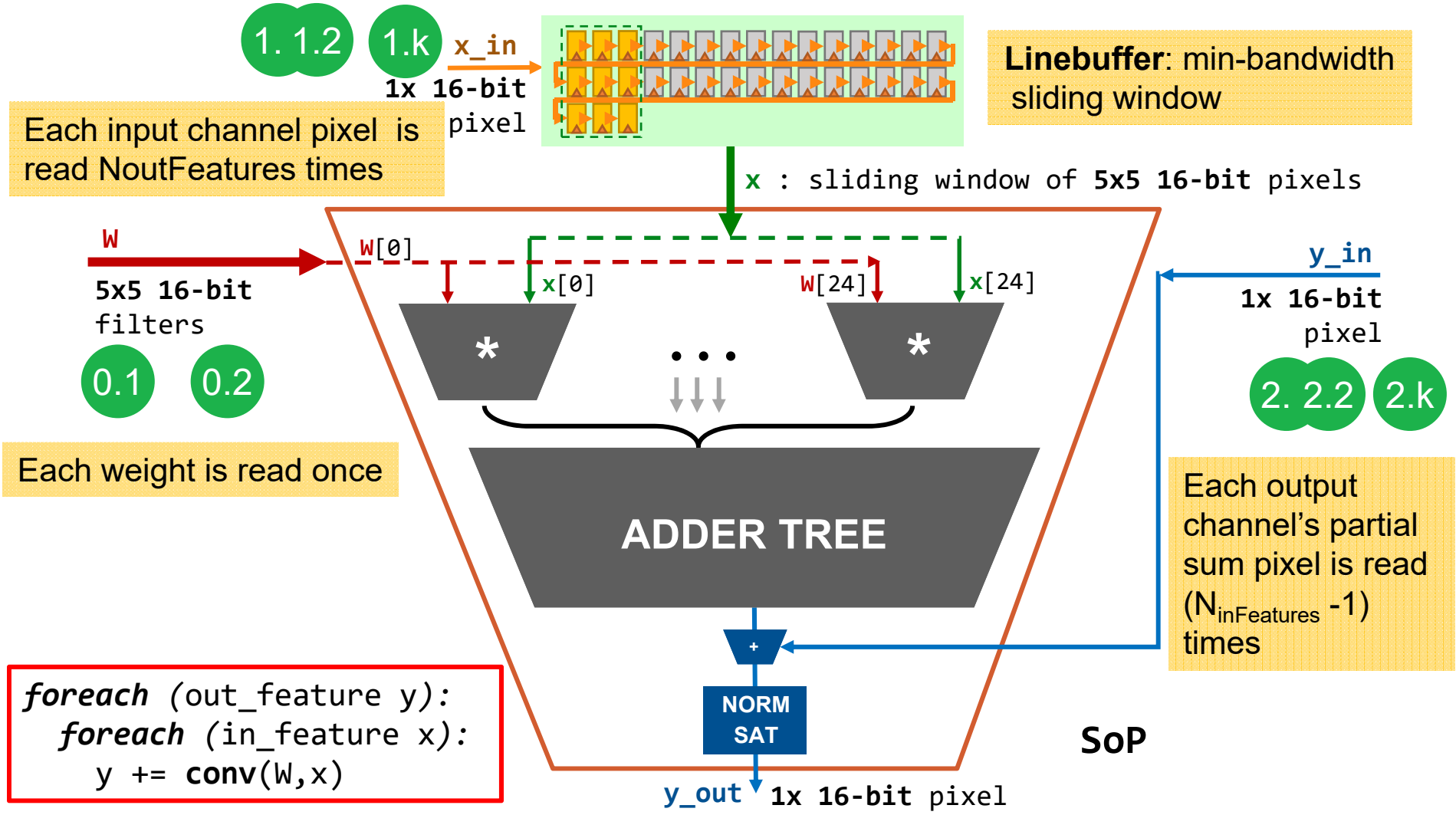
streaming fashion

4. **Weights** for each convolution filter are **stored privately**



F. Conti and L. Benini, "A ultra-low-energy convolution engine for fast brain-inspired vision in multicore clusters," *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2015, pp. 683-688.

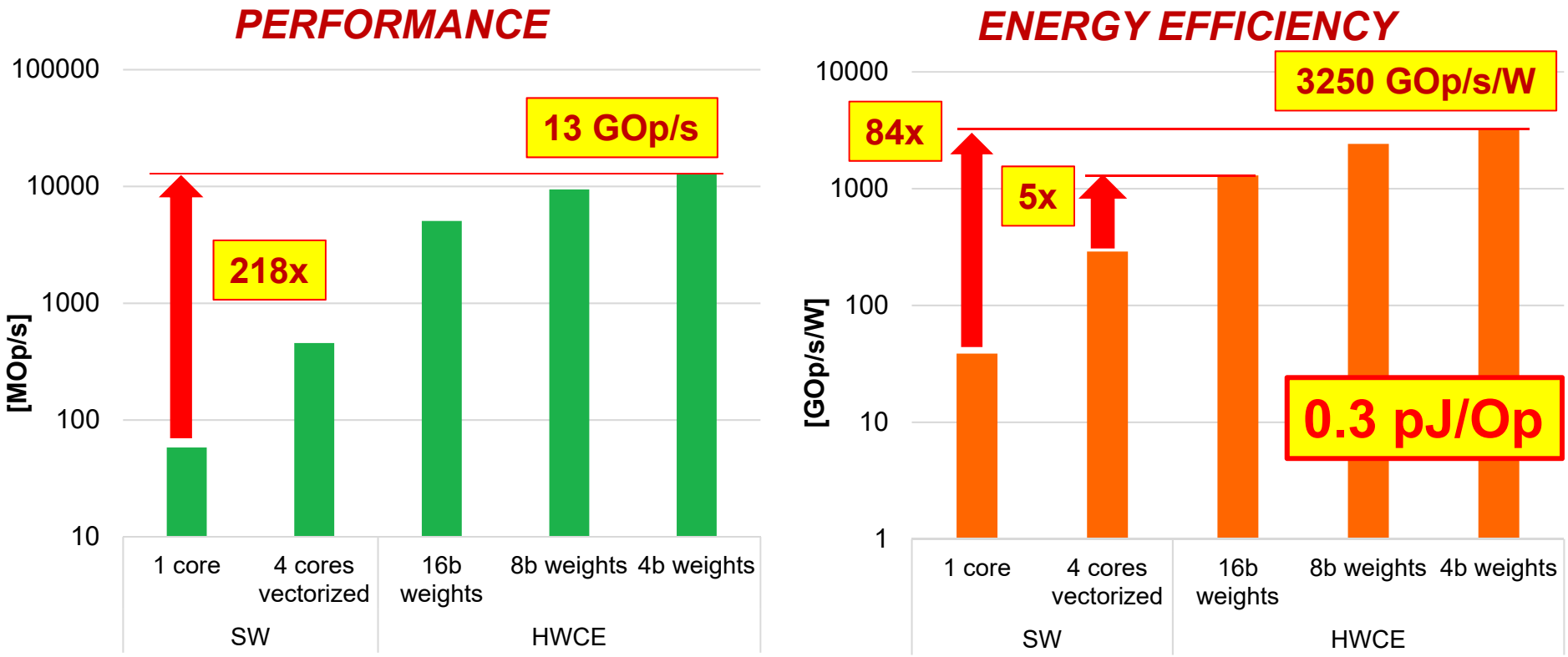
HWCE Sum-of-Products



Heterogeneous PULP CNN Performance

Cluster performance and energy efficiency on a 64x64 CNN layer (5x5 conv)

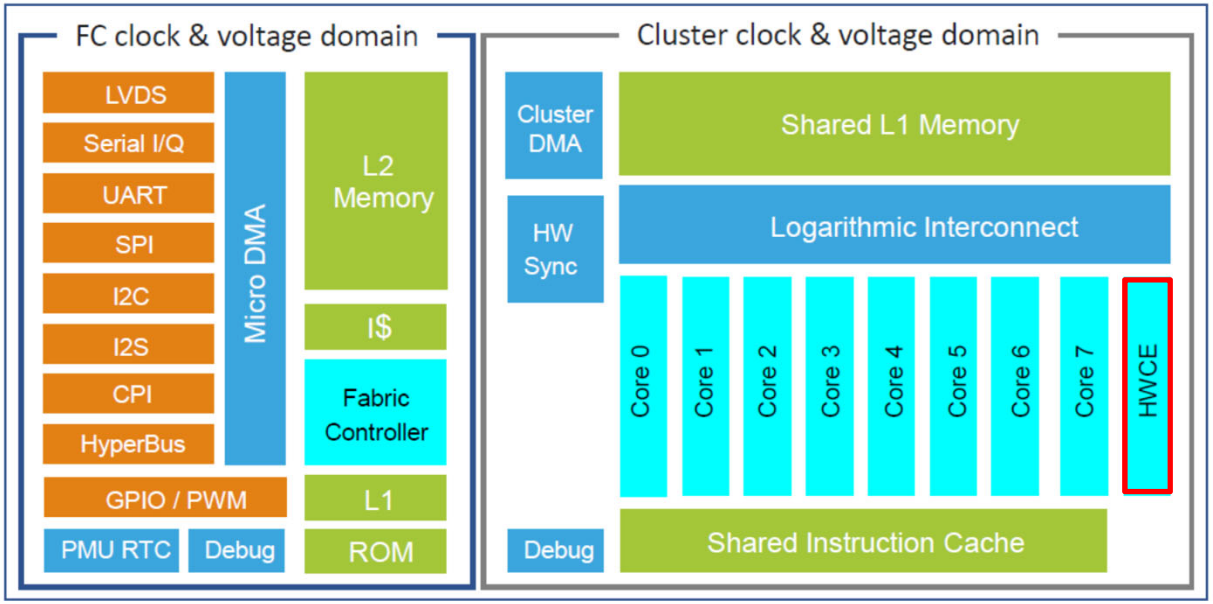
Scaled to ST FD-SOI 28nm @ Vdd=0.6V, f=115MHz



Now coming: HWCE 2.0 – improves scalability & flexibility @ 3TOPS/W

PULP cluster+MCU+HWCE(V1) → GWT's GAP8 (55 TSMC)

Two independent clock and voltage domains, from 0-133MHz/1V up to 0-250MHz/1.2V



MCU Function
- Extended RISC-V core
- Extensive I/O set
- Micro DMA
- Embedded DC/DC c
- Secured execution

Computation engine
- 8 extended RISC
- Fully programmable
- Efficient parallelization
- Shared instruction cache
- Multi channel DMA
- HW synchronization
- HW convolution Engine

What	Freq MHz	Exec Time ms	Cycles	Power mW
40nm Dual Issue MCU	216	99.1	21 400 000	60
GAP8 @1.0V	15.4	99.1	1 500 000	3.7
GAP8 @1.2V	175	8.7	1 500 000	70
GAP8 @1.0V w HWCE	4.7	99.1	460 000	0.8

11 X

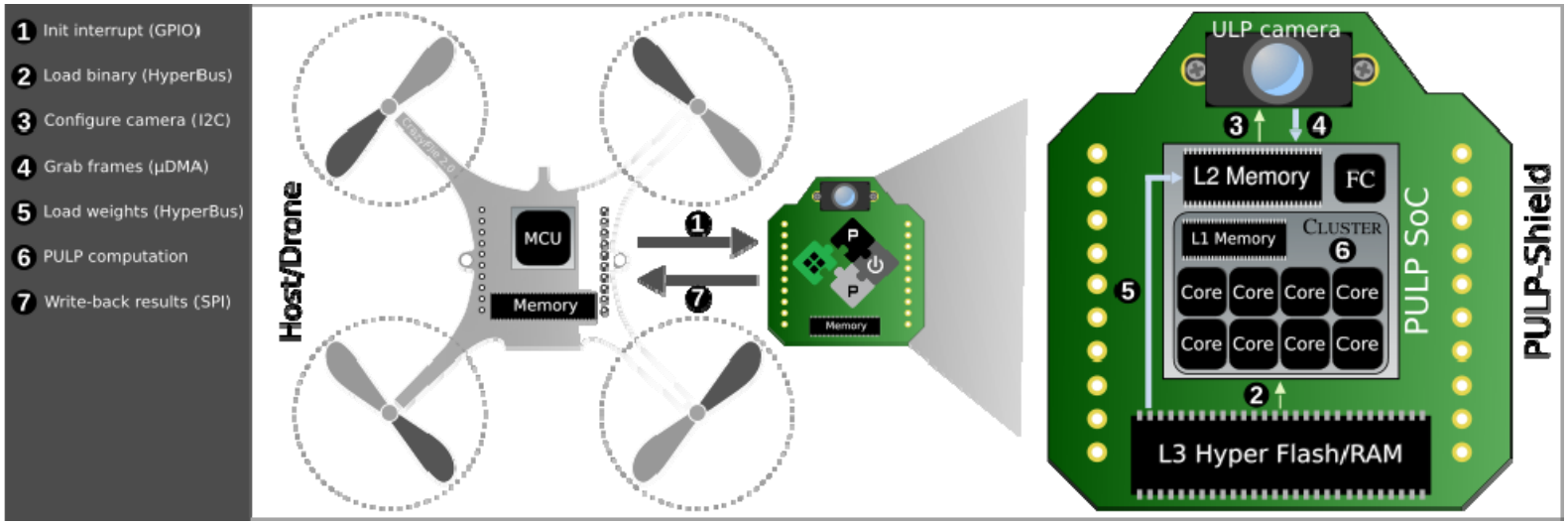
16 X

GREENWAVES

4x More efficiency at less than 10% area cost

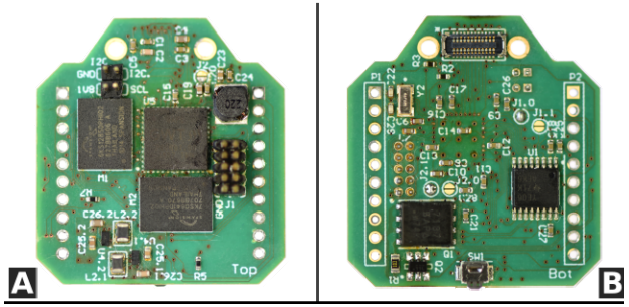
PULP

New Application Frontiers: DroNET on NanoDrone



Pluggable PCB: PULP-Shield

- ~5g, 30×28mm
- GAP8 SoC
- 8 MB HDRAM
- 16 MB HFlash
- QVGA ULP HiMax camera
- Crazyflie 2.0 nano-drone (27g)



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

PULP

More Efficiency (2): Extreme Quantization

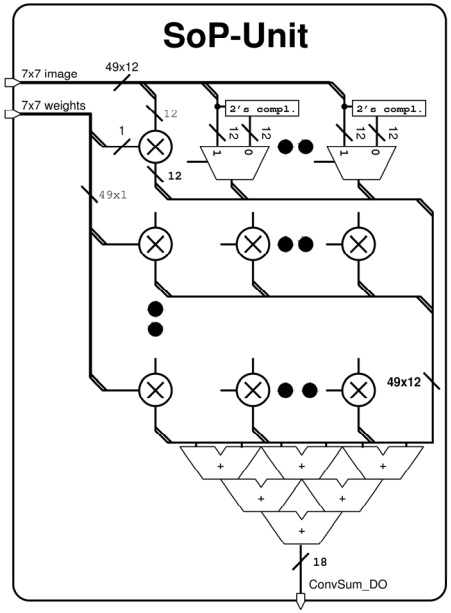
Low(er) precision: 8→4→2

Model	Bit-width	Top-1 error
ResNet-18 ref	32	31.73%
INQ	5	31.02%
INQ	4	31.11%
INQ	3	31.92%
INQ	2 (ternary)	33.98%

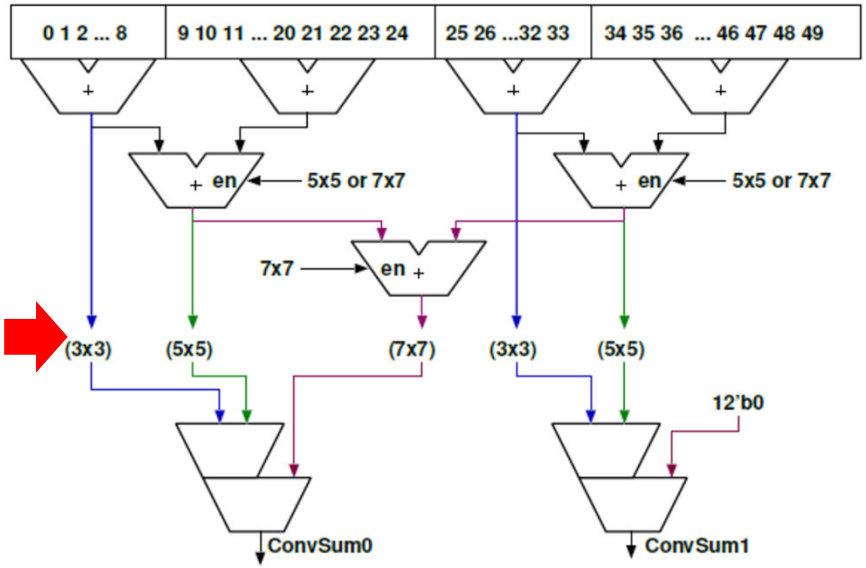
SOA INQ retraining

2.2% loss → 0% with 20% larger net

MULT → MUX



Equivalent for 7x7 SoP



ImageBank

Image Mapping (3x3, 5x5, 7x7)

1 MAC Op = 2 Op (1 Op for the “sign-reverse”, 1 Op for the add).

From +/-1 Binarization to XNORs

$$\mathrm{y}(k_{out}) = \mathrm{binarize}_{\pm 1}\left(\mathbf{b}_{k_{out}} + \sum_{k_{in}}\left(\mathbf{W}(k_{out}, k_{in}) \otimes \mathbf{x}(k_{in})\right)\right)$$

XNOR

$$\mathrm{binarize}_{\pm 1}(t) = \mathrm{sign}\left(\gamma \frac{t-\mu}{\sigma} + \beta\right)$$

$$\mathrm{binarize}_{0,1}(t) = \begin{cases} 1 \text{ if } t \geq -\kappa/\lambda \doteq \tau, \text{ else } 0 & (\text{when } \lambda > 0) \\ 1 \text{ if } t \leq -\kappa/\lambda \doteq \tau, \text{ else } 0 & (\text{when } \lambda < 0) \end{cases}$$

$$\mathrm{y}(k_{out}) = \mathrm{binarize}_{0,1}\left(\sum_{k_{in}}\left(\mathbf{W}(k_{out}, k_{in}) \otimes \mathbf{x}(k_{in})\right)\right)$$

Thresholding

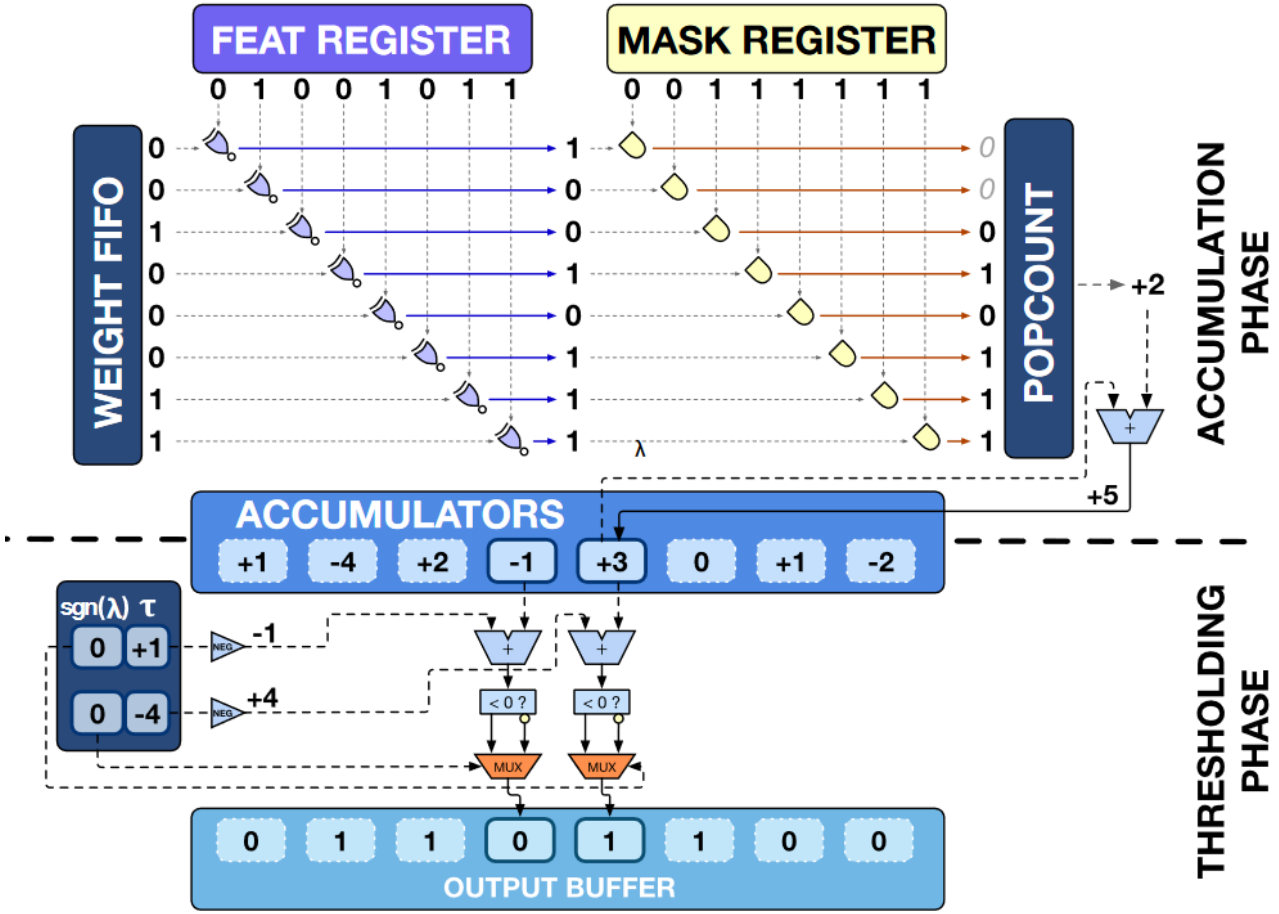
Multi-bit accumulation

Binary product → XOR

A	B	out
-1	-1	+1
-1	+1	-1
+1	-1	-1
+1	+1	+1

A	B	out
0	0	1
0	1	0
1	0	0
1	1	1

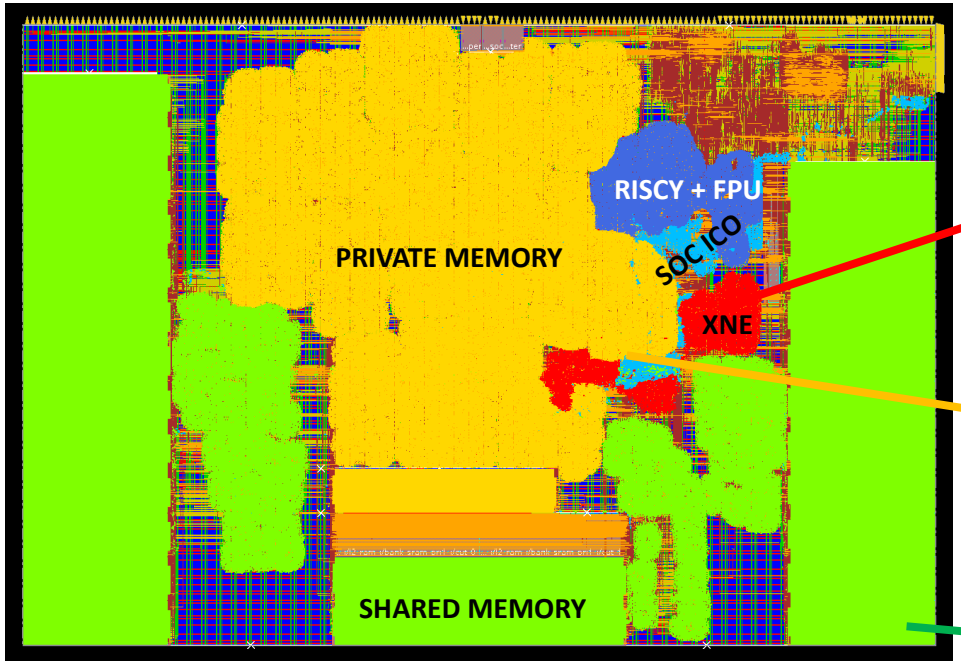
XNE: XNOR Neural Engine



Main unit: binary dot-product and thresholding

Quentin: a XNE-accelerated microcontroller

Quentin in GlobalFoundries 22FDX



XNE area is **~14000 um2** (71 KGE, 72% Riscy+FPU)

Private memory is **448 KB** SRAM + 3r2w **8 KB** SCM

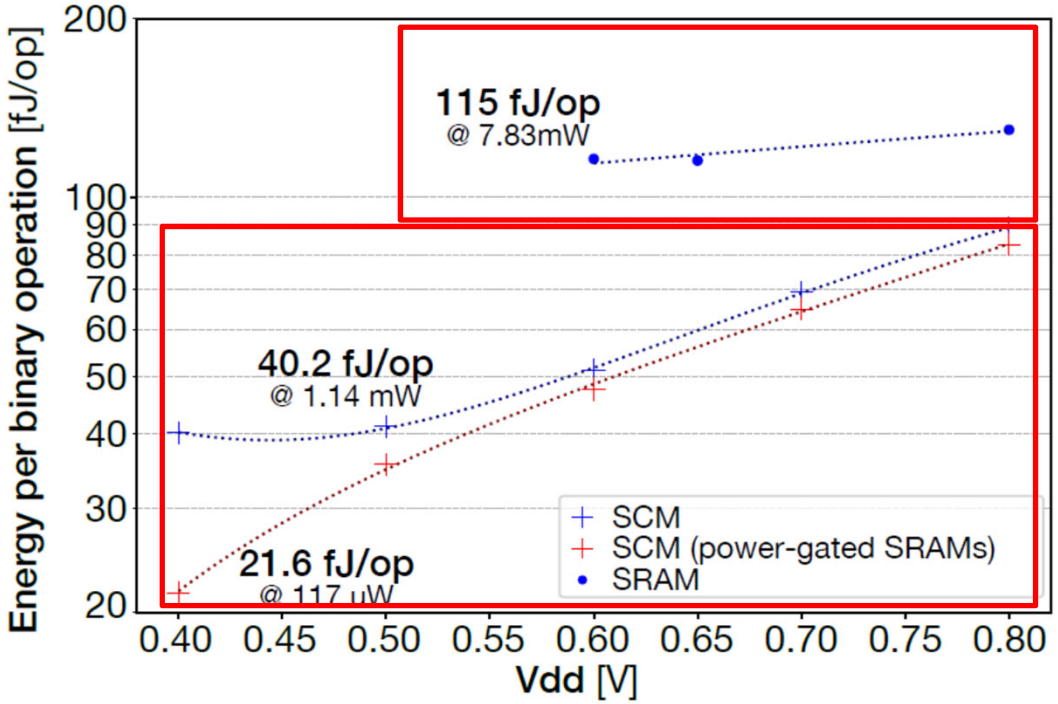
Shared memory is **56 KB** SRAM + **8 KB** SCM

F. Conti, P. D. Schiavone and L. Benini, "XNOR Neural Engine: A Hardware Accelerator IP for 21.6-fJ/op Binary Neural Network Inference," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, pp. 2940-2951, Nov. 2018.

XNE Energy Efficiency

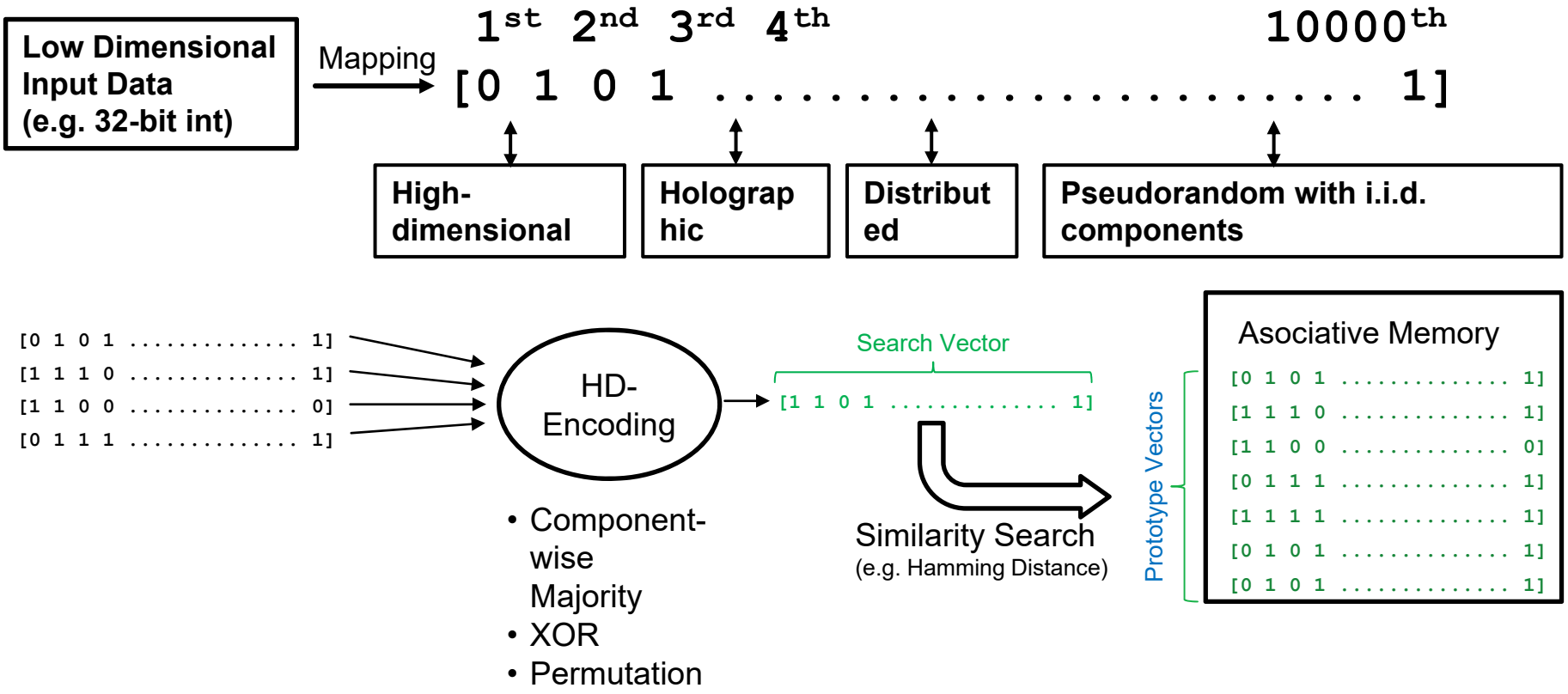
With SRAMs, max eff
@ 0.65V **8.7 Top/s/W**

With SCMs, max eff
@ 0.5V **46.3 Top/s/W**



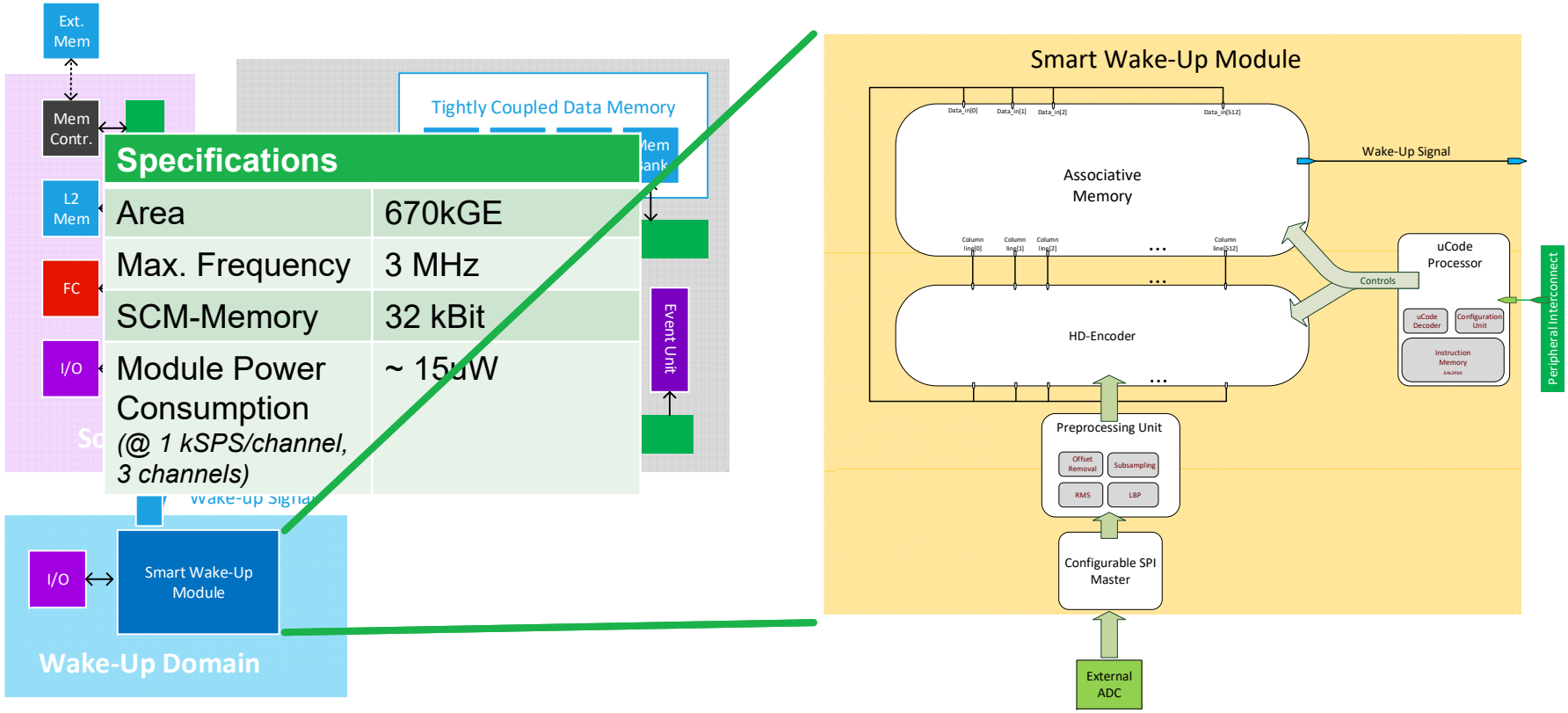
Accuracy Loss is high even with retraining (10%+) → mixed precision
TWN & TCN are also a very appealing alternative (under design)

Not Only CNNs: Hyper-Dimensional Computing



PULP

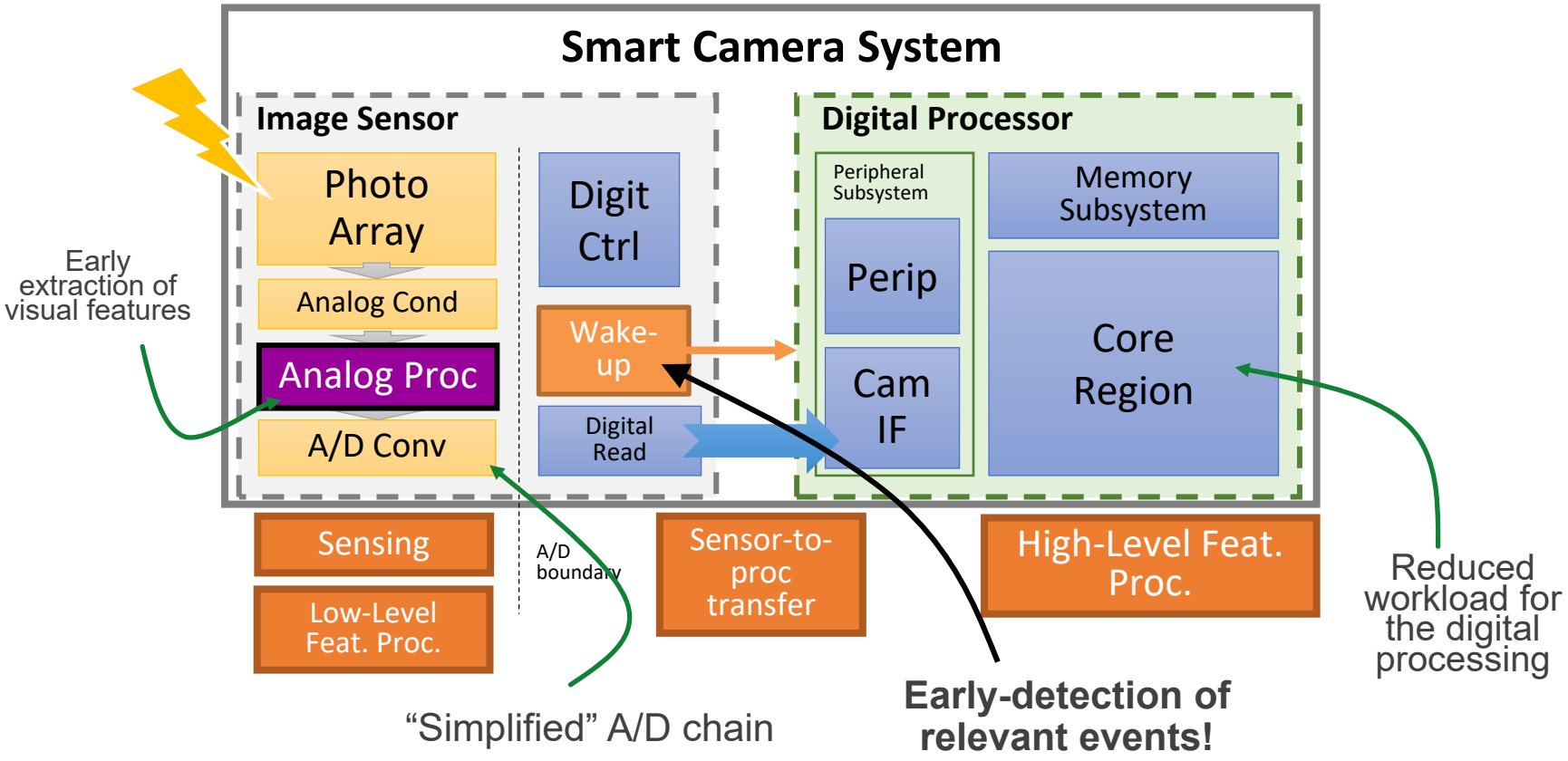
More efficiency (3): HD-Based smart Wake-Up Module



Taped out in 22fdx

More Efficiency (4): Focal Plane Processing

Enable the extraction of low-level features in a parallel and efficient way by **integrating pixel-wise mixed-signal processing circuits** on the sensor die **to reduce the imager energy costs**.

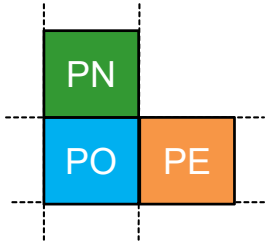


[Fernández-Berni, Jorge, et al. "Image Feature Extraction Acceleration." Image Feature Detectors and Descriptors. Springer International Publishing, 2016. 109-132.

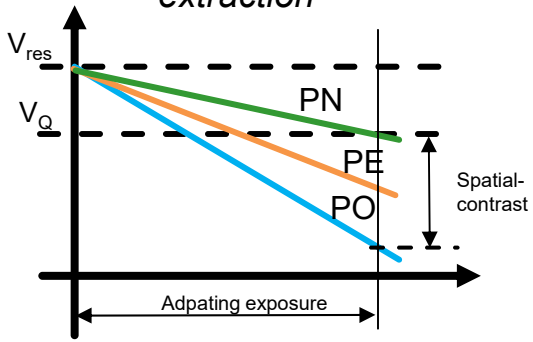
Ultra-Low Power Imaging (GrainCam)

Imager performing **spatial filtering** and **binarization** on the sensor die through **mixed-signal sensing**!

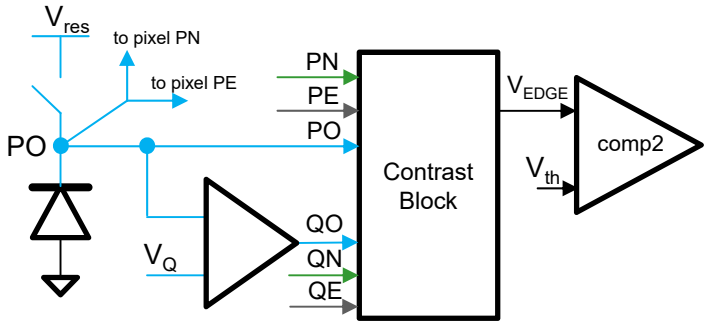
'Moving' pixel window

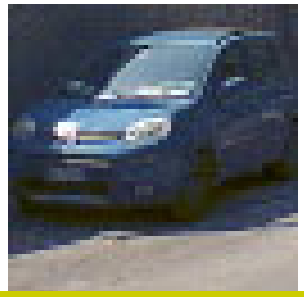


Gradient extraction

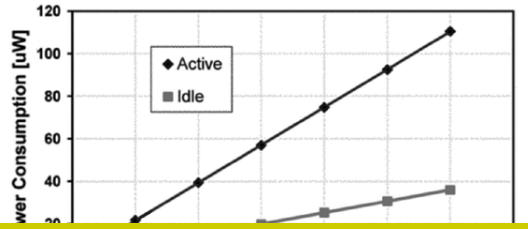


Per-pixel circitut for filtering and binarization





Ultra-Low Power Consumption <100uW

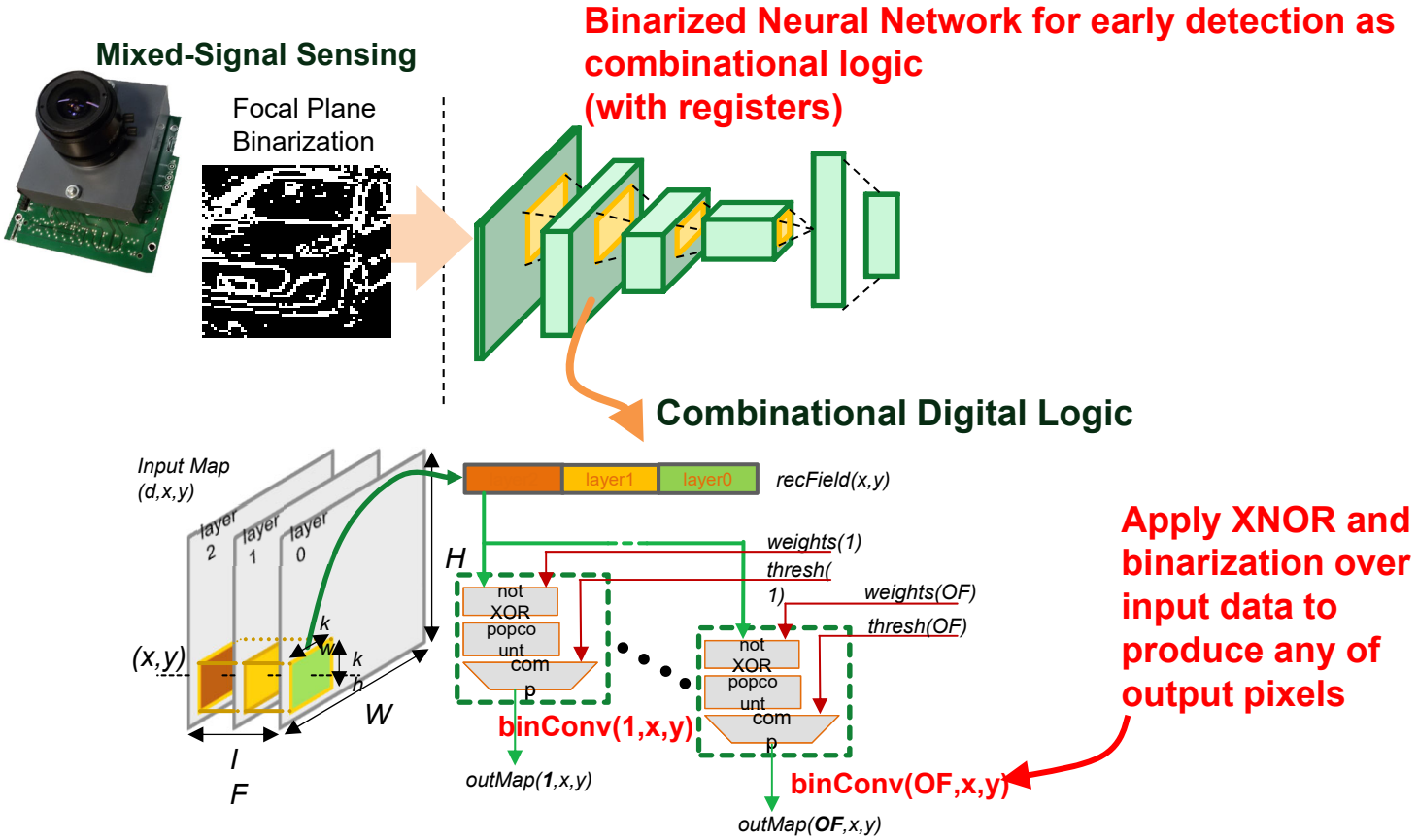


<10x wrt SoA imagers

This process naturally reflects the operation of a binarized pixel-wise convolution and can be seen as embedding the first convolutional layer within the image sensor die

M. Gottardi et al, "A 100uw 12864 pixels contrast-based asynchronous binary vision sensor for sensor networks applications," IEEE JSSC, 2009.

Combinational "Fully Spatial" BNN



Synthesis Results

Synthesis of both models with hard-wired or reconfigurable weights

GF 22nm SOI with LVT cells (typical corner case 0.65V, 25°C)

TABLE II
SYNTHESIS AND POWER RESULTS FOR DIFFERENT CONFIGURATIONS

netw.	type	area [mm^2]	area [MGE]†	time/img [ns]	time/img [FO4]‡	E/img [nJ]	leak. [µW]	E-eff. [TOp/J]
16×16	var.	1.17	5.87	12.82	560	2.40	945	470.8
16×16	fixed	0.46	2.32	12.40	541	1.68	331	672.6
32×32	var.	5.80	29.14	17.27	754	11.14	4810	479.4
32×32	fixed	2.61	13.13	21.02	918	11.67	1830	457.6

† Two-input NAND-gate size equivalent: 1 GE = 0.199 μm^2
‡ Fanout-4 delay: 1 FO4 = 22.89 ps

Hundreds of TOPS/W!

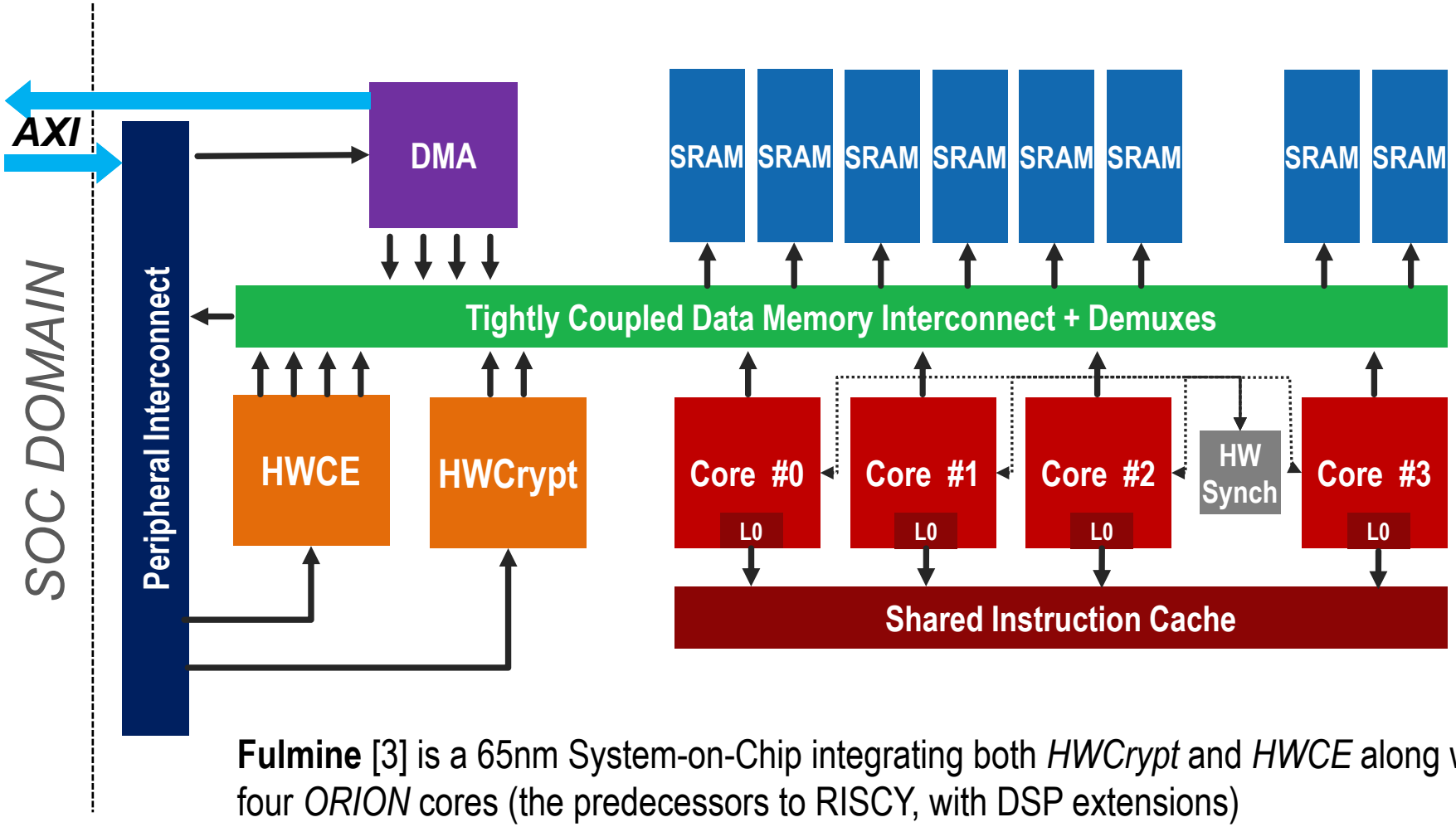
Massive area reduction when hard-wiring the weights:

- XNOR operations reduce to wires or inverter, which can be also shared among different receptive fields
- popcounts also exploits sharing mechanisms

Advanced Synthesis Tools become central to exploit weights and intermediate results sharing to reduce the area occupation

M Rusci, L Cavigelli, L Benini “Design automation for binarized neural networks: A quantum leap Opportunity?” 2018 IEEE International Symposium on Circuits and Systems (ISCAS), 1-5

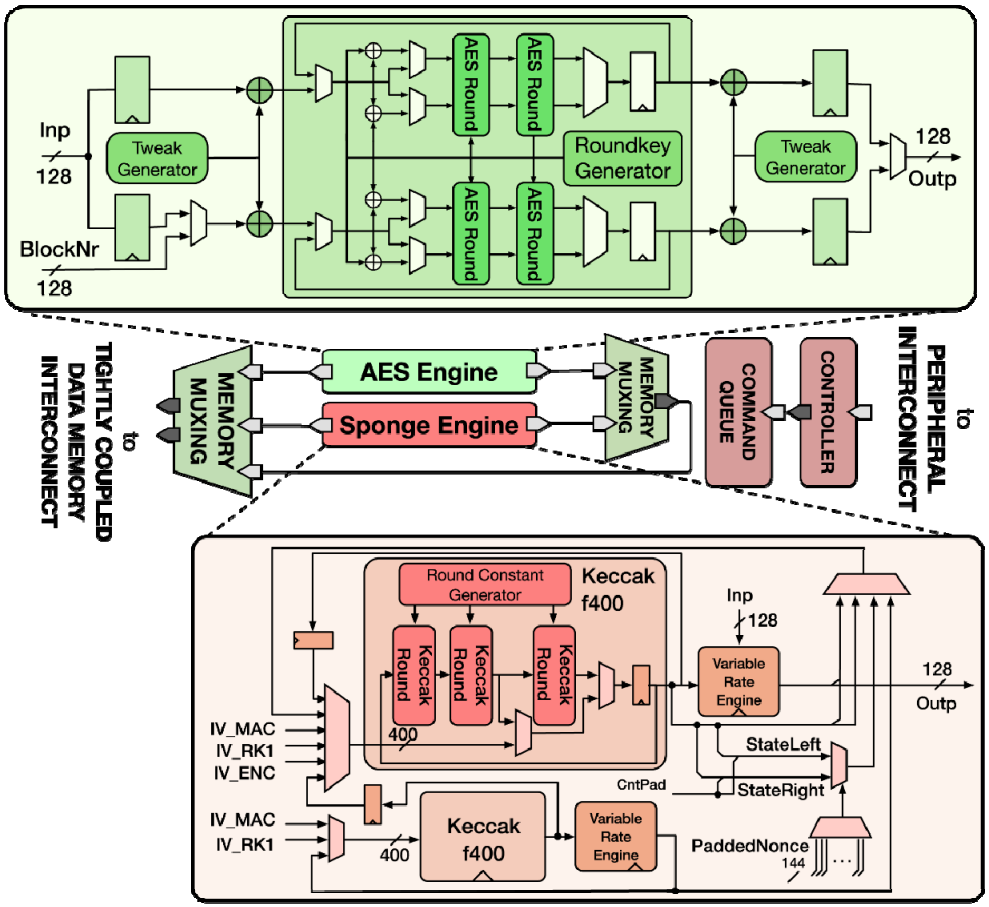
What about Security? A Secure EE AI Processor



Fulmine [3] is a 65nm System-on-Chip integrating both *HWCrypt* and *HWCE* along with four *ORION* cores (the predecessors to RISCY, with DSP extensions)

[3] F. Conti et al., An IoT Endpoint System-on-Chip for Secure and Energy-Efficient Near-Sensor Analytics, IEEE TCAS-I 2017

Data security: HWCrypt – a Cryptographic Accelerator



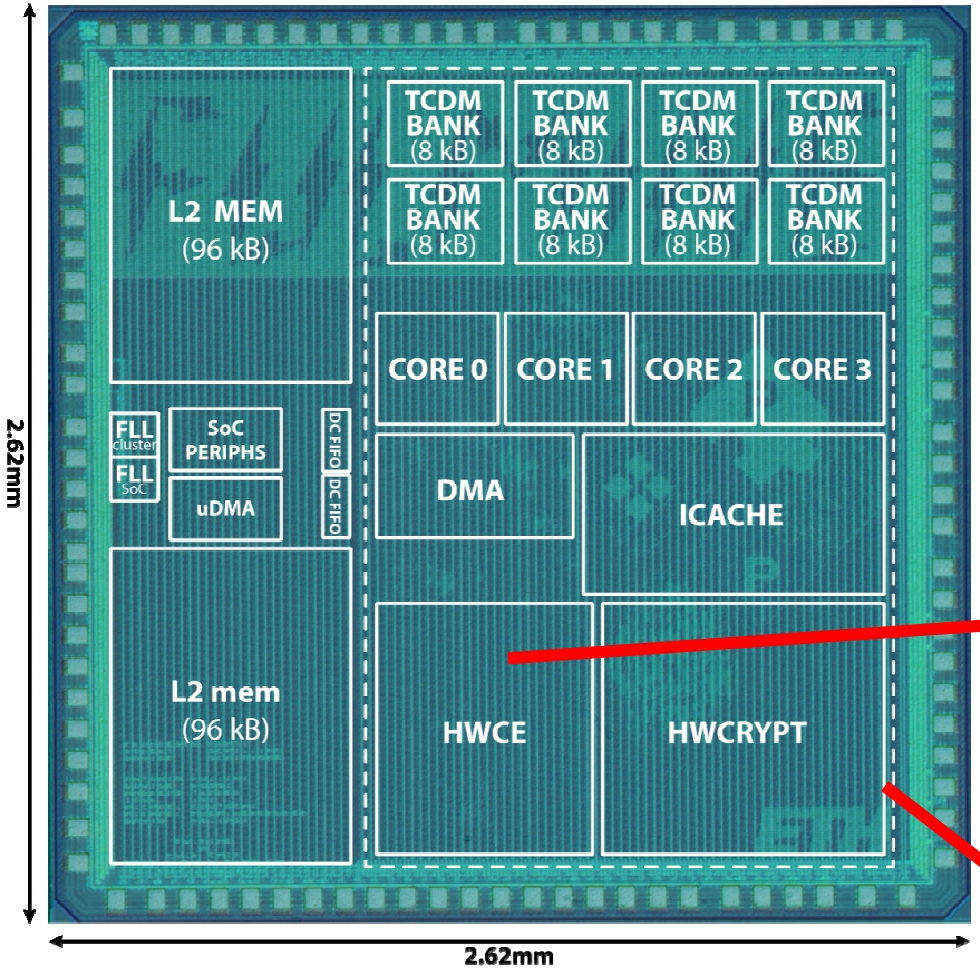
HWCrypt is a «collection» of two crypto engines plugged to the shared memory and controlled via the periph interconnect

- AES Engine
  - AES-128-ECB: fast but not secure (plaintext patterns are ~visible in ciphertext)– for comparison!
  - AES-128-XTS: each block encrypted with a different tweak – just as fast in the HWCrypt
  - individual execution of cipher rounds (to speed up new SW-based AES-based algorithms)
- Sponge Engine
  - two instances of Keccak-f[400]
  - leakage-resilient encryption scheme [1]
  - similar performance to AES engine

[1] T. Unterluggauer et al., Leakage Bounds for Gaussian Side Channels, CARDIS 2017

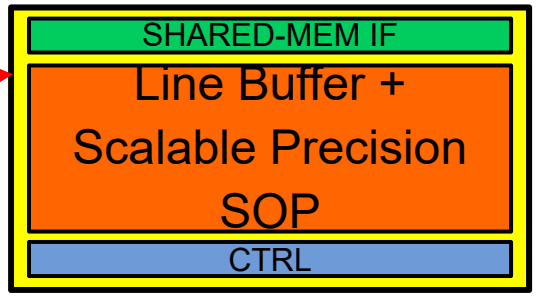
Fulmine SoC

Fulmine: Hardware Convolutional Engine (HWCE) in the Cluster



- **~6.9 mm2** PULP chip, 1/2016
- 4 cores
- 1 HWCRYPT accelerator
- 1 HWCE for 3D conv layers
- 64kB L1, 192kB L2
- QSPI (M/S), I2C, I2S, UART

- **~1514** kGE for the cluster
- **232** kGE for the CNN HWCE

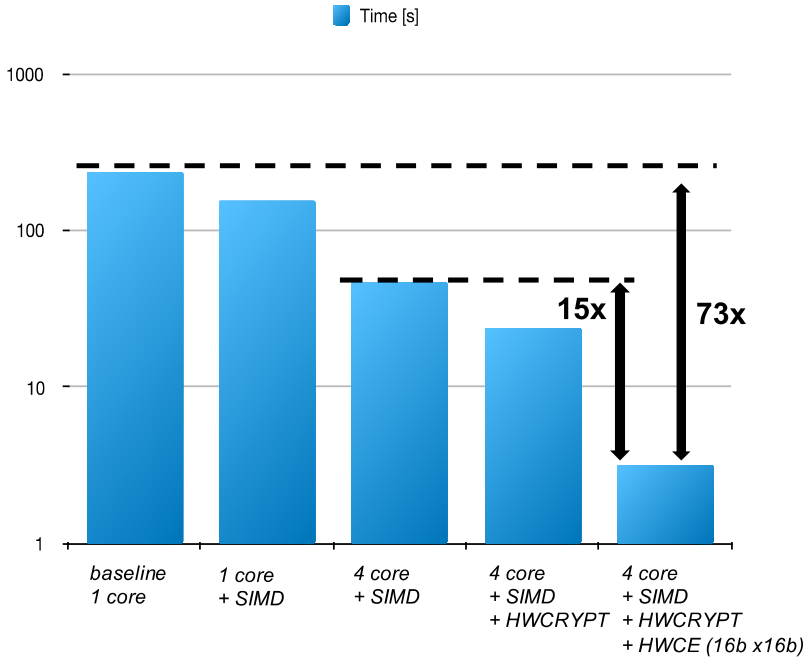


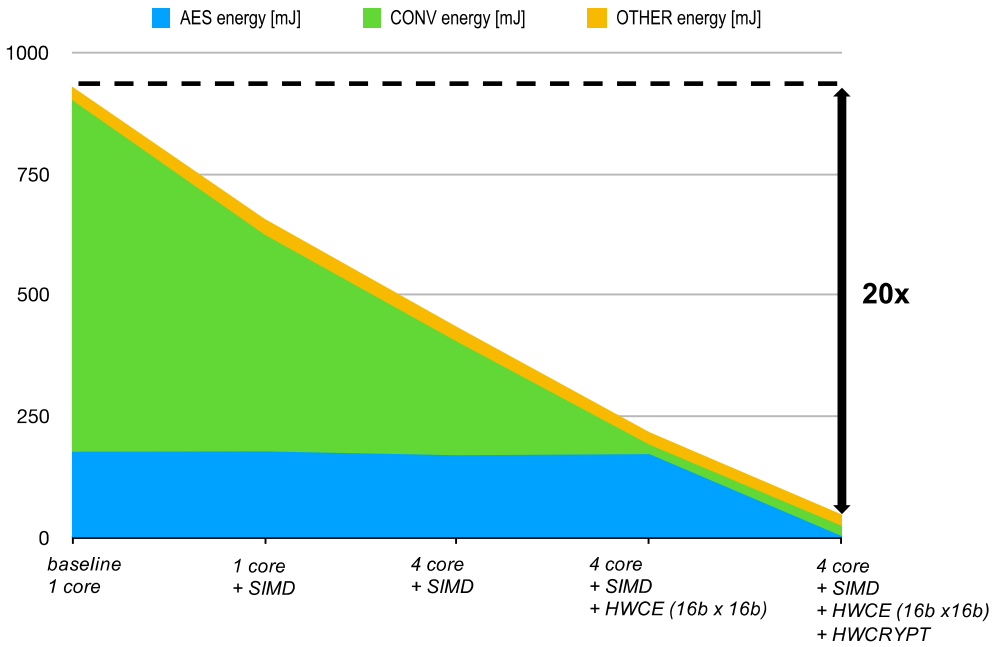
Crypto Engine → Secure Analytics

PULP

F. Conti et al., "An IoT Endpoint System-on-Chip for Secure and Energy-Efficient Near-Sensor Analytics,",in IEEE TCAS-I Sept. 2017.

Secured ResNet-18: execution time & energy



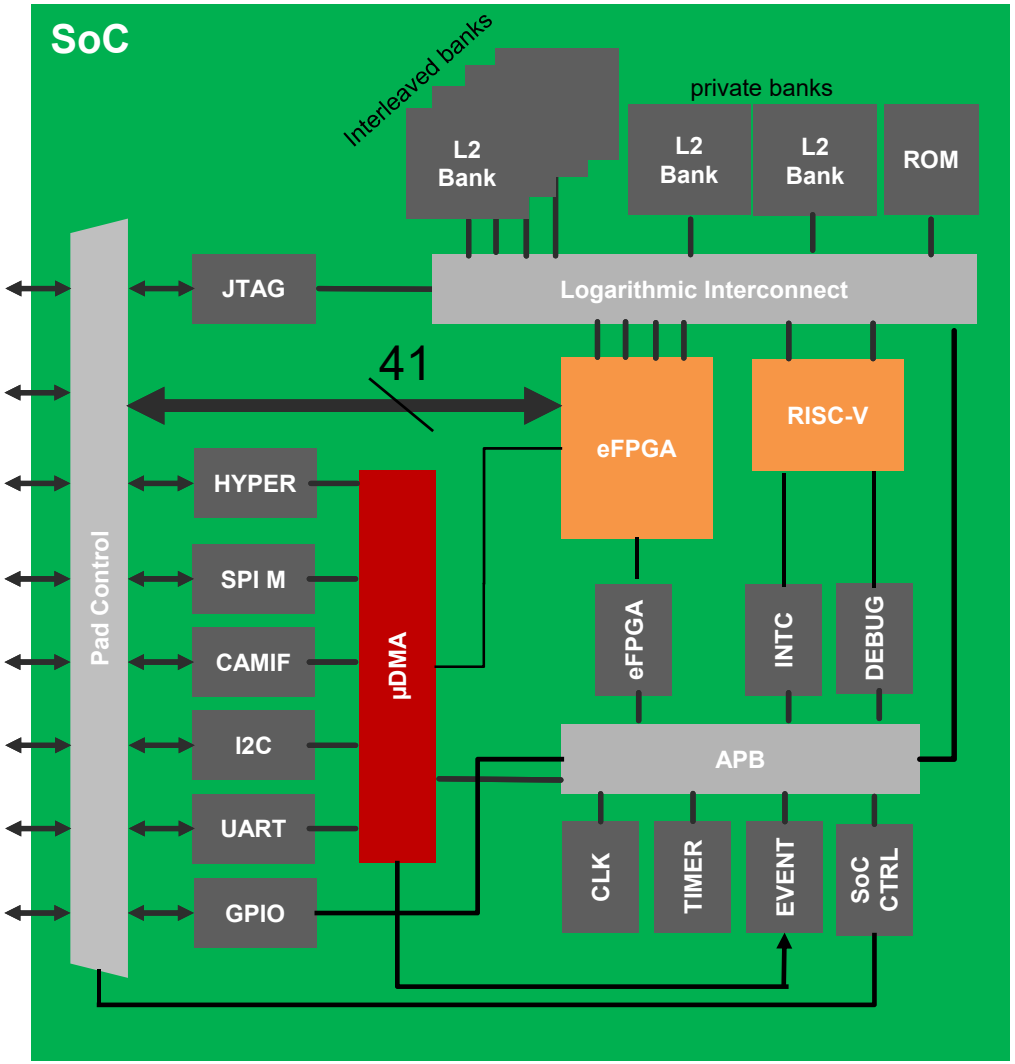


- DPA-secure encryption of all communication: weights + intermediate CNN results
- Up to 70x speedup w.r.t. microcontroller, 15x w.r.t. to pure SW
- 20x improvement in energy (diminishing return region: this is “good enough”)
- Performance up to 3s / frame ; 50 mJ / frame

Reconfigurable Heterogeneity: Arnold

- **RI5CY RISC-V 32b CORE**
  RV32IMFC + pulp extensions
  3,19 CoreMark/MHz
  Memory protection
- **Autonomous I/O Subsystem**
- **512 kB of Memory**
  4 Interleaved BANKS of 112 kB
  7 cuts of 4096x32 SRAM
  1 BANK of 32 kB
  2 cuts of 4096x32 SRAM
  1 BANK of 32 kB
  2 cuts of 4096x32 SRAM
- **8 kB ROM**
- **3 FLLs**
  Core, Peripherals and FPGA
- **embedded FPGA**
  32x32 array
  32x32x4x 4in-LUT = 4096 4inLUT
  1024 registers
  6 clocks

Arnold: MCU+ Embedded FPGA



PULP

Arnold Physical View

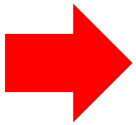
- **Max Frequency@0.72V:**
  Core: 430 MHz
- **Area:**
  3x3 mm2
- **eFPGA:**
  Max frequency: 100 MHz
  2x2 mm2
- **std CELLS from INV[illegible]**
  SC8T 104CPP {20,24,28} L 0.72V
- **MEMORIES from FPGA**
  S1D W4096 B32 M[illegible]CPP
  0.72V/0.72V
  ROM W2048 B32 M32 C064 104CPP
  0.72V/0.72V

FDX22 SoC
(being measured now)

PULP

Many applications need 64-bit “numbers”

- **For the first 4 years of the PULP project we used only 32bit cores**
  - Most IoT applications work well with 32bit cores.
  - A typical 64bit core is much more than 2x the size of a 32bit core.
- **But times change:**
  - Large datasets, high-precision numerical calculations (e.g. double precision FP) at the IoT edge and cloud
  - Lot of interest in the security community for working on a contemporary open source 64bit core.
  - High performance computing (FP intensive) is becoming again a hot area for Architecture and Digital design research

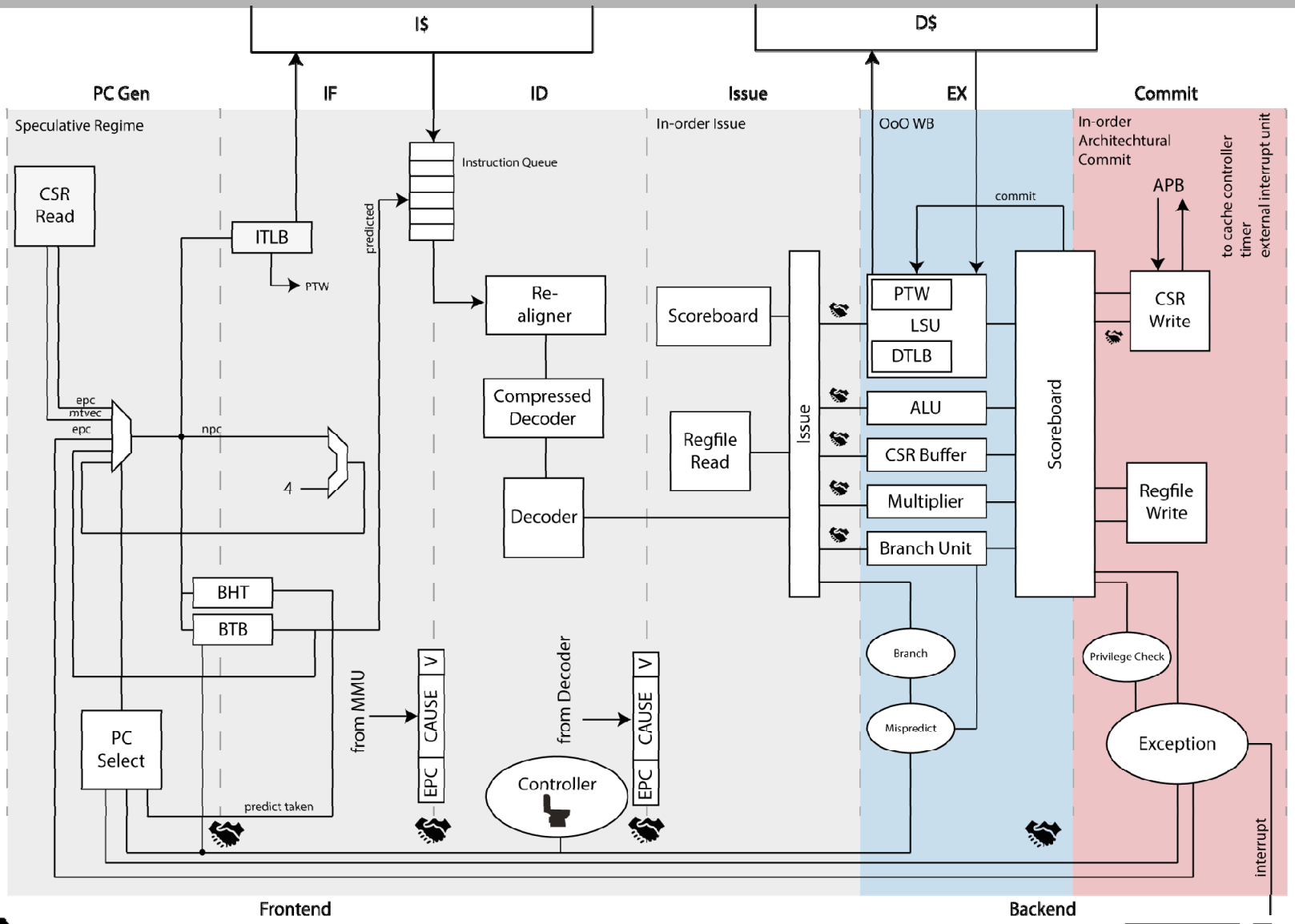


PULP

ARM GPP + RISC-V Accelerator

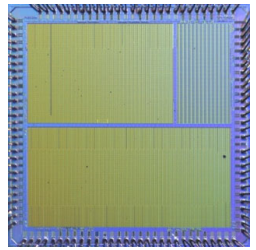
ETH

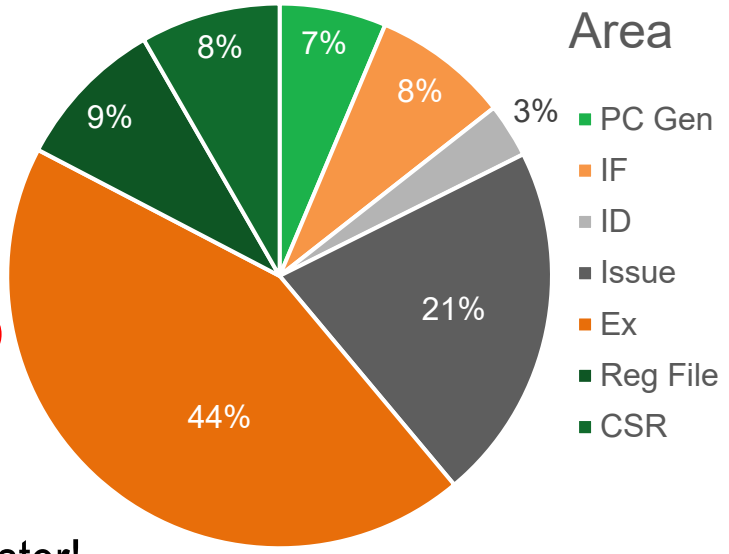
ARIANE: >1GHz, Linux Capable 64-bit core



Main properties of Ariane

- Tuned for high frequency, 6 stage pipeline, integrated cache
  - In order issue, out-of-order write-back, in-order-commit
  - Supports privilege spec 1.11, M, S and U modes
  - Hardware Page Table Walker
- Implemented in GF 22FDX (Poseidon, Kosmodrom, Baikonur), and UMC65 (Scarabaeus)
  - In 22nm: ~1 GHz worst case conditions (SSG, 125/-40C, 0.72V), 1.7GHz typ @0.8V
  - 8-way 32kByte Data cache and 4-way 32kByte Instruction Cache
  - Core area: 175 kGE
- Application-class features are not cheap
  - 38% area in TLB, PTW
  - **51.8**pJ/op vs. 10pJ/OP in 22FDX @ 0.8V
  - IPC 0.85 vs. 0.94, 1.7GHz vs. 690, just 2.1 faster!

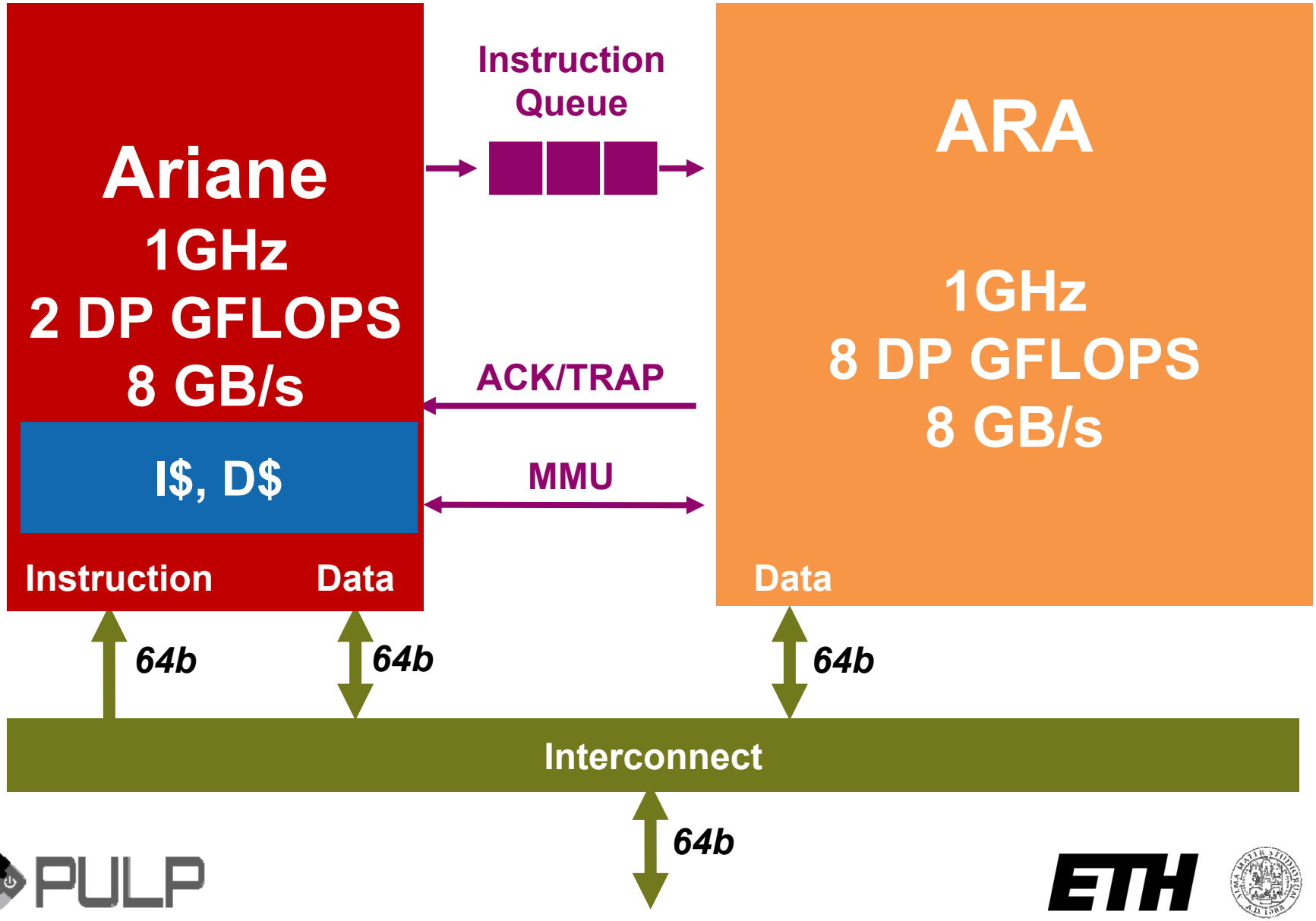




PULP

ETH

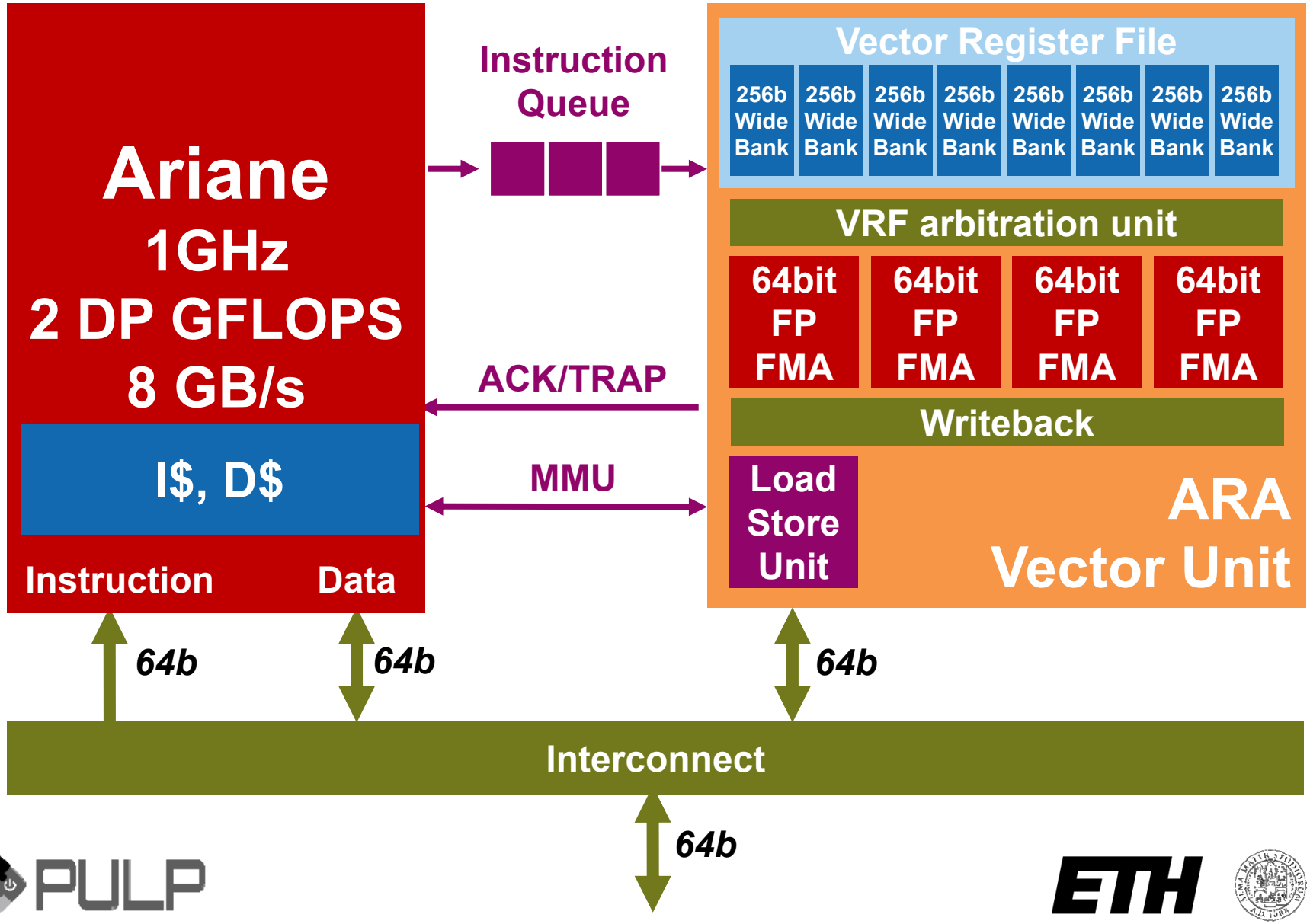
Extreme FP Performance: The “V” Extension



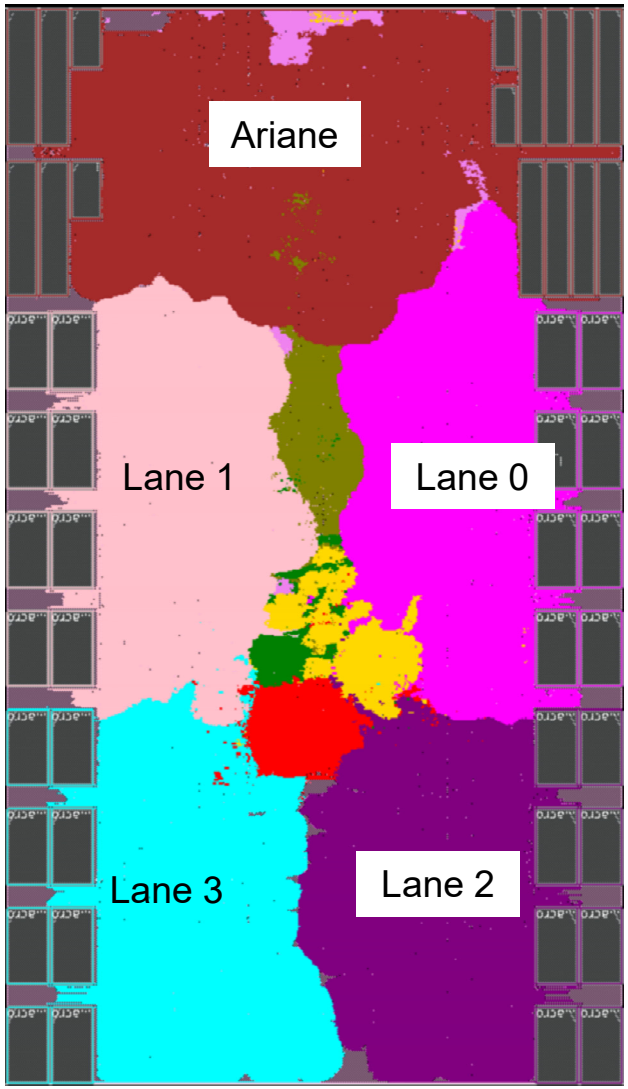
PULP

ETH

Extreme FP Performance: The "V" Extension



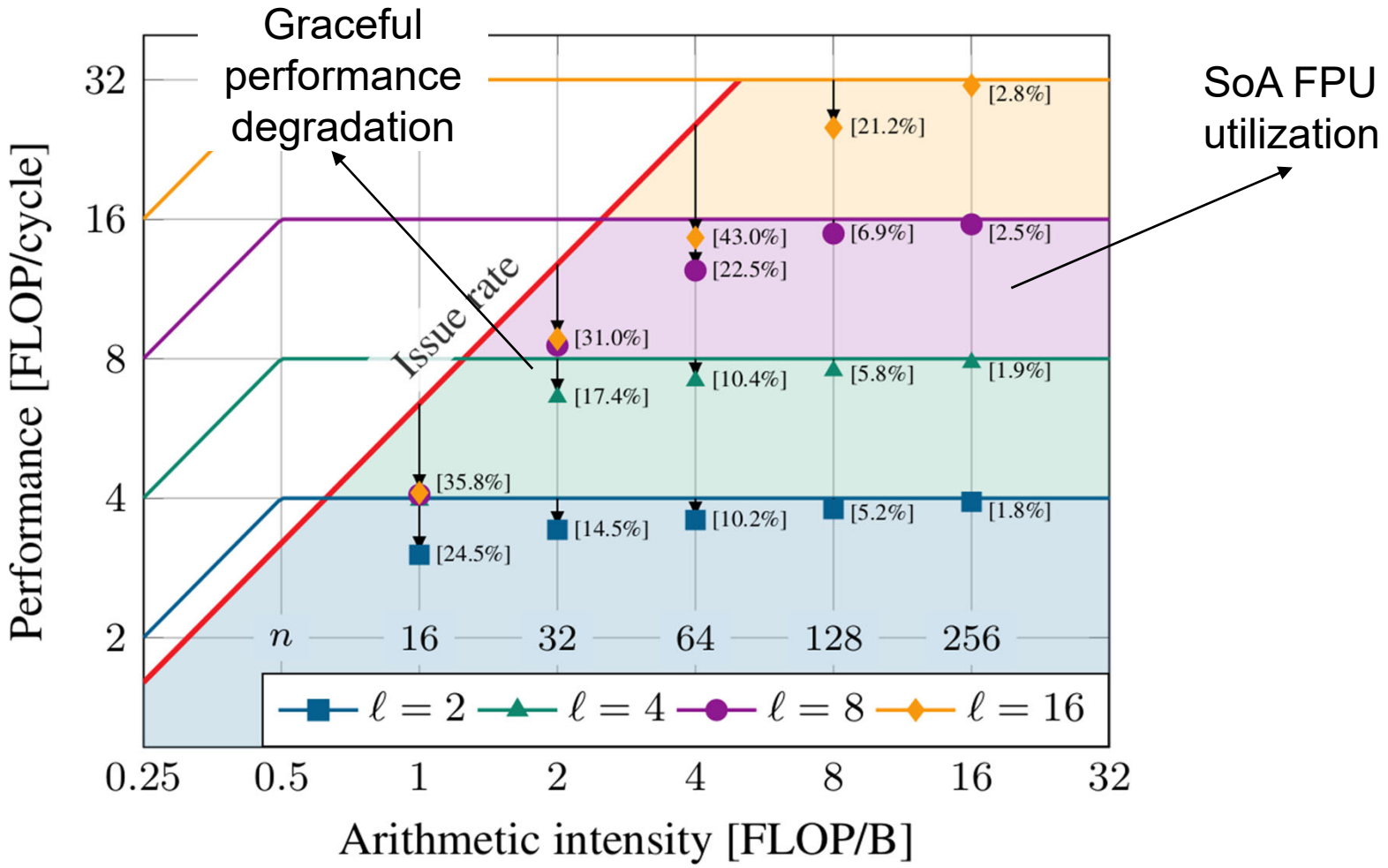
Implementation results in a 0.75mm x 1.25mm GF22 macro



- Post-synthesis PPA results
- WC operating frequency similar to Ariane
- Area: 3188 kGE
  - Each lane amounts to 533 kGE
  - Ariane (wo. $s) amounts to 474 kGE
- For a 256×256 integer MATMUL
  - Performance: 10.2 DP-GFLOPS
  - Power consumption: 192 mW
  - Energy efficiency: 53 DP-GFLOPS/W
- **3.1x** GOPS/W wrt Ariane, at same frequency

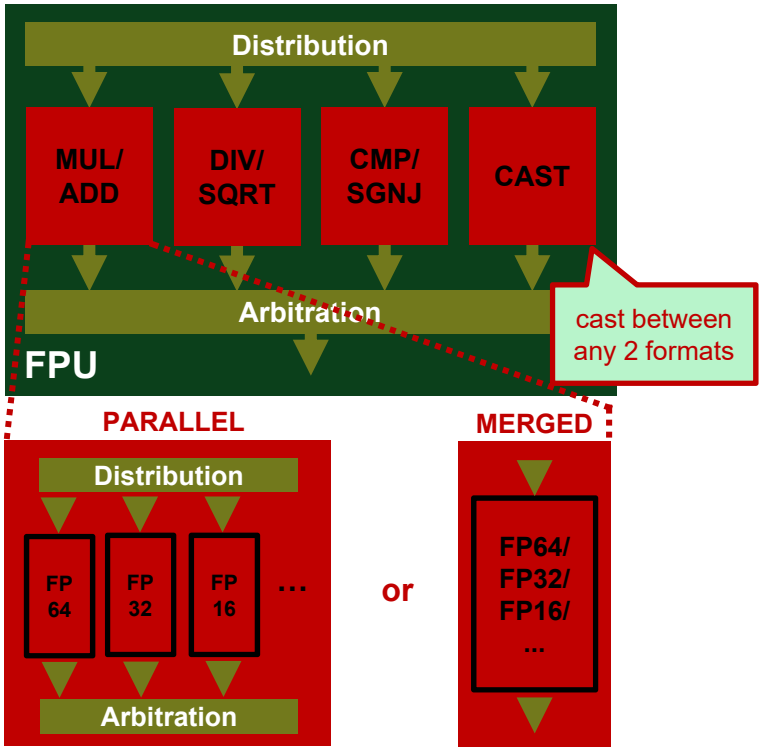
PULP

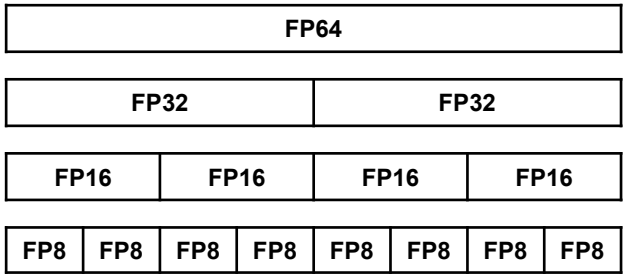
Up to 98% utilization @ *n* × *n* DP-MATMUL (always?)



Floating-Point →Transprecision FP

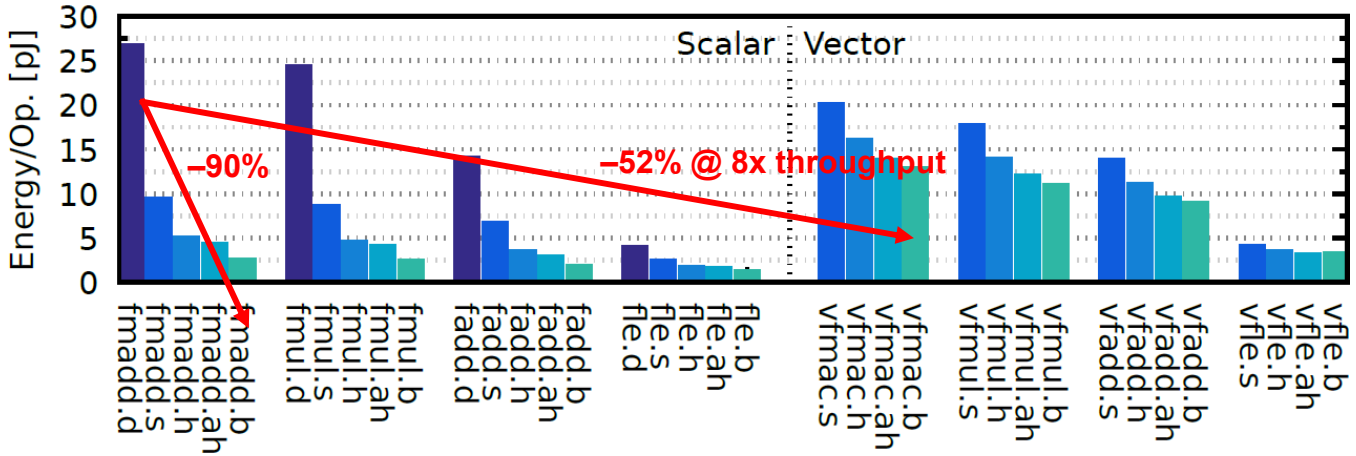
- Provide easy precision tuning
  - 64(DFP), 32(FP), 16(HFP), 16ALT, 8
- Mainly consists of four operation groups
  - MUL/ADD: Add/Subtract, Multiply, FMA
  - CMP/SGNJ: Comparisons, Min/Max etc.
  - CAST: FP-FP casts, Int-FP / FP-Int casts
- **Parametrizable**
  - Number & Encoding of **Formats** (any Exp/Man bits)
  - Packed-SIMD **Vectors**
  - **# Pipeline Stages** (per Op and Format)
  - **Implementation** (per Op and Format)
    - **PARALLEL** for best Speed
    - **MERGED** (or Iterative) for best Area
- **Special Functions** for Transprecision
  - Cast-and-Pack 2 FP Values to Vector
  - Casts amongst FP Vectors + Repacking
  - Expanding FMA (e.g. FP32 += FP16*FP16)





Result Highlights

- While TP FPU adds 9% of Ariane core area vs RV64D, …
- **Super-Linear** energy savings thanks to aggressive **clock-gating**
  - Mutually **exclusive** data paths rather than sharing



- ~pJ/FLOP @1GHz in 22FDX
  - 0.4pJ FP8, 0.9pJ FP16, 2.4pJ FP32, 6.2pj FP64
- Transprecision applications will profit from this additional HW
- Fully integrated into RISC-V ISA through custom extension
  - Easy to leverage thanks to our GCC extensions, part of PULP SDK

Heterogeneous RISC-V platform from ULP to HPC

RISC-V Cores

- RI5CY 32b
- Ibex (MR) 32b
- Ibex (ZR) 32b
- Ariane 64b + ARA

Peripherals

- JTAG
- SPI
- UART
- I2S
- DMA
- GPIO

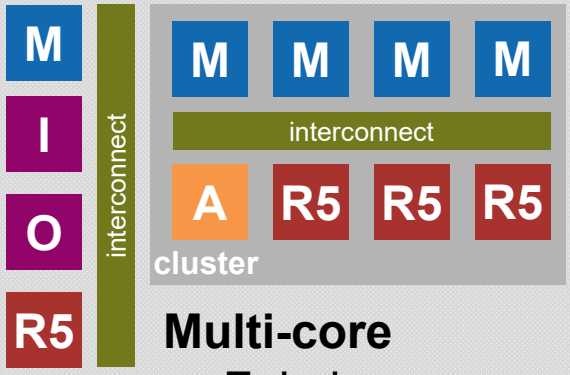
Interconnect

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

Platforms

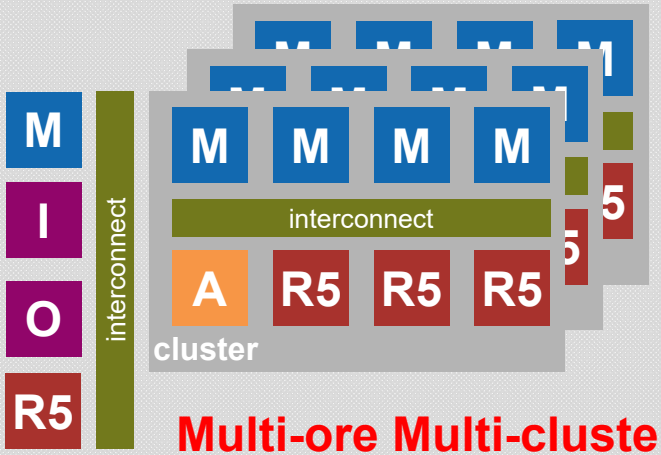
Single Core

- PULPino
- PULPissimo



Multi-core

- Fulmine
- Mr. Wolf



Multi-ore Multi-cluster

- OpenPiton, Hero

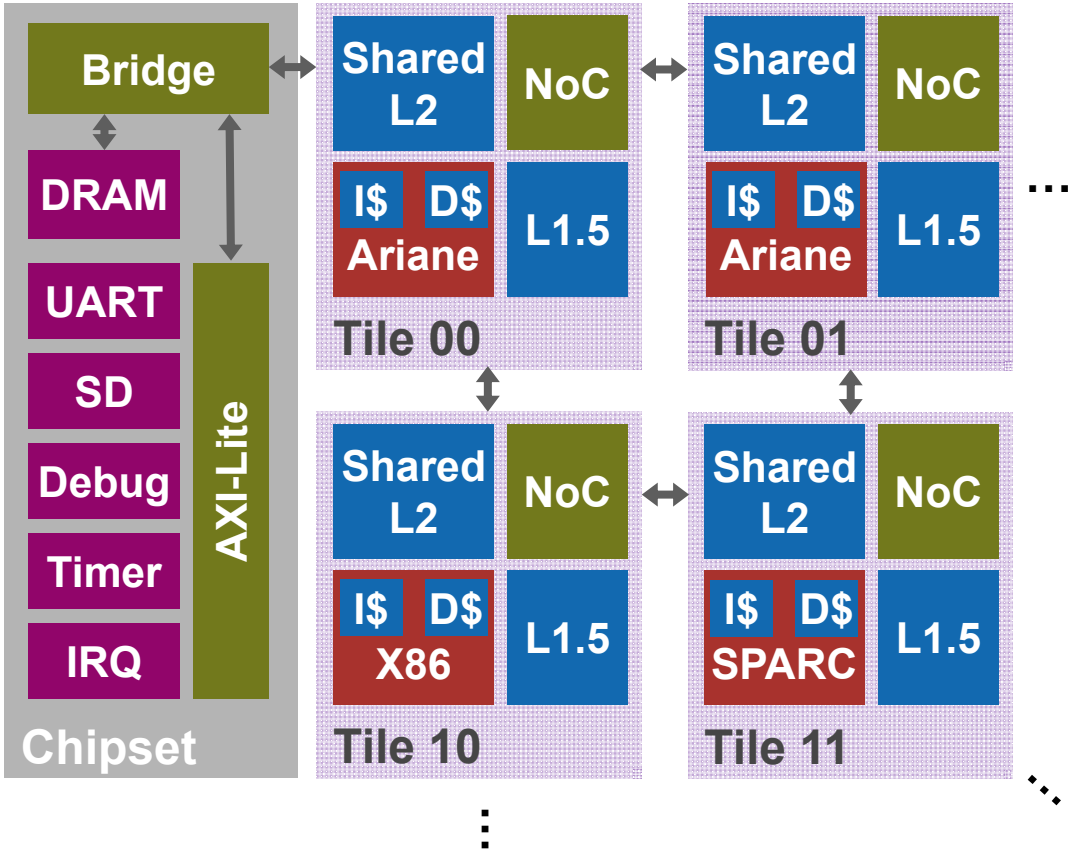
IOT → HPC

Accelerators

- HWCE (convolution)
- Neurostream (ML)
- HWCrypt (crypto)
- PULPO (1st order opt)

OpenPiton: cache-coherent many-core system

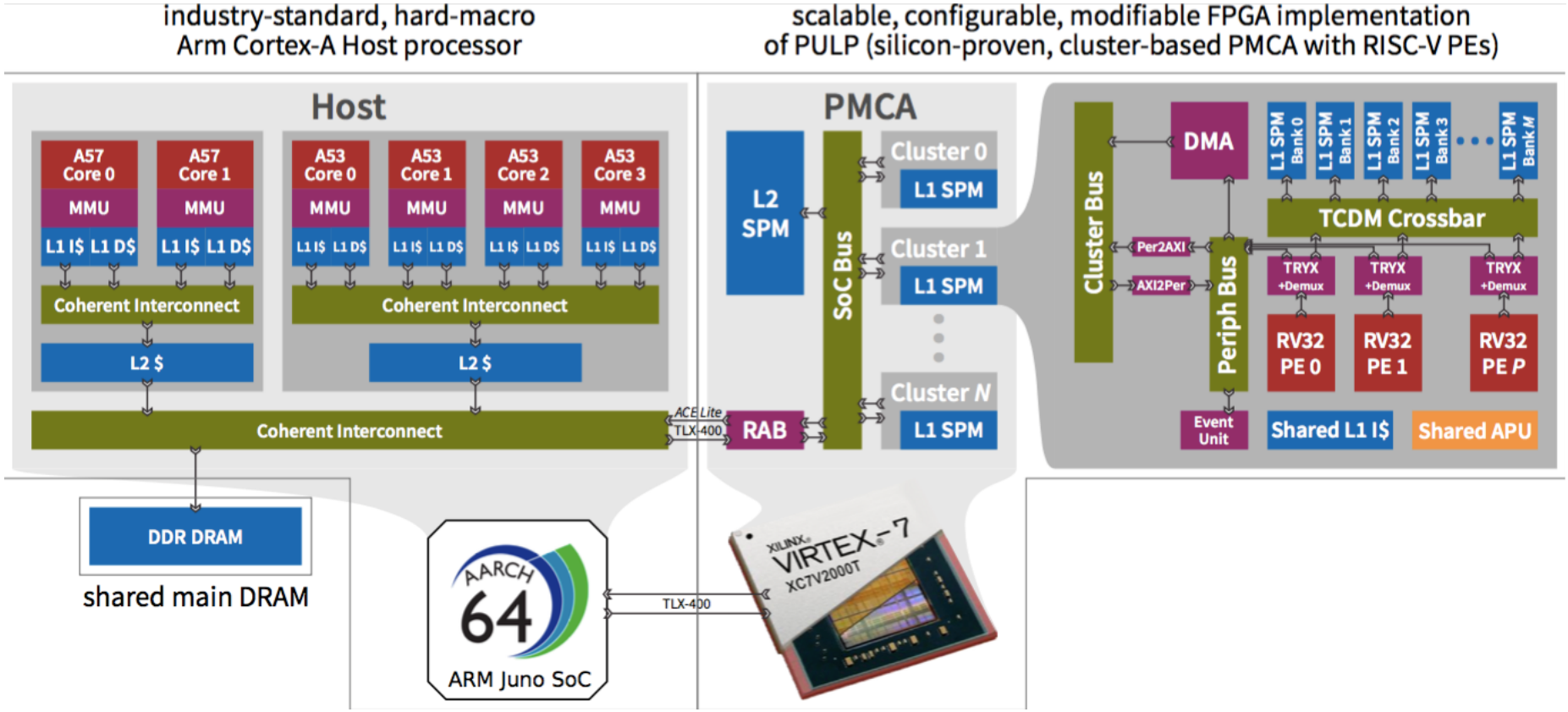
- **OpenPiton**
  - Developed by Princeton
  - Originally OpenSPARC T1
  - Scalable NoC with coherent LLC
  - Tiled Architecture
- **Status**
  - Bare-metal Dec ’18
  - Update with support for SMP Linux just released
  - Multiple different cores and ISAs (x86, SPARC, RISC-V)



ISA heterogeneity with a cache-coherent memory hierarchy

PULP OpenPiton ETH

Hero: Fat (multi) core host, slim manycore accelerator



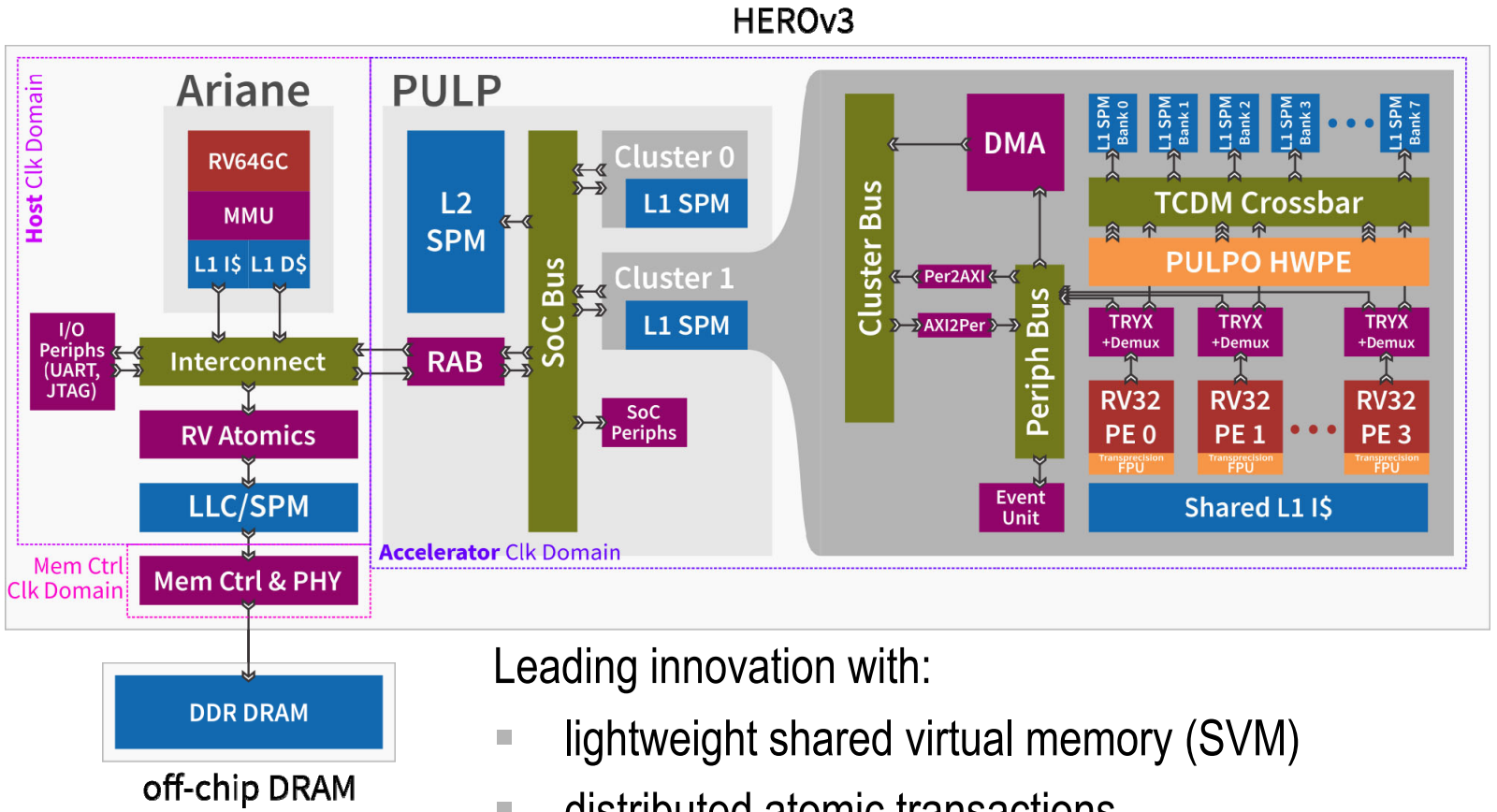
- First released in 2018
- Many-core PULP clusters connected with a general-purpose fat-core host with heterogeneous ISA – shared virtual memory (non coherent)

PULP

ETH

HERO v3: Heterogeneous 64-32b RISC-V

The best of both worlds?

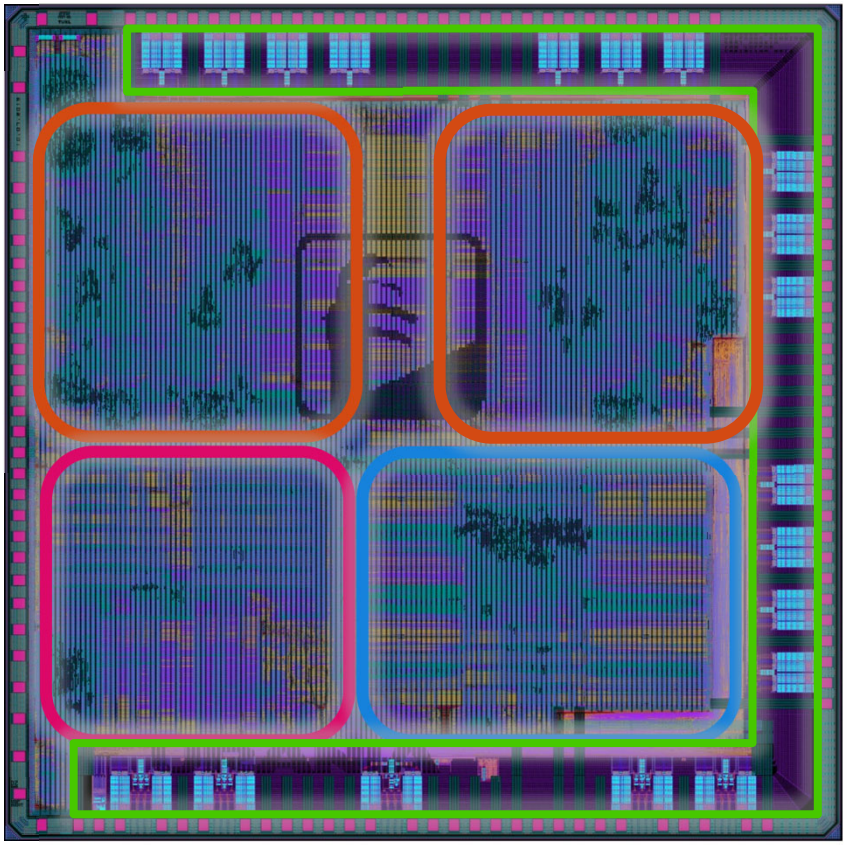


Leading innovation with:

- lightweight shared virtual memory (SVM)
- distributed atomic transactions
- heterogeneous 64/32-bit LLVM toolchain
- support for predictable execution (PREM)

HERO v3 First Silicon: Urania

- **first HERO ASIC**
  first fully-open source linux Booting risc-V SoC in the world
- 2 PULP clusters, each with
  - 4 RV32 RI5CY cores
  - 4 transprecision FPUs
  - 1 PULPO accelerator
  - 64 KiB TCDM in 8 banks
- Ariane RV64 host processor
- 128 KiB Shared LLC
- software-managed IOMMU
- DDR3 DRAM Controller + PHY



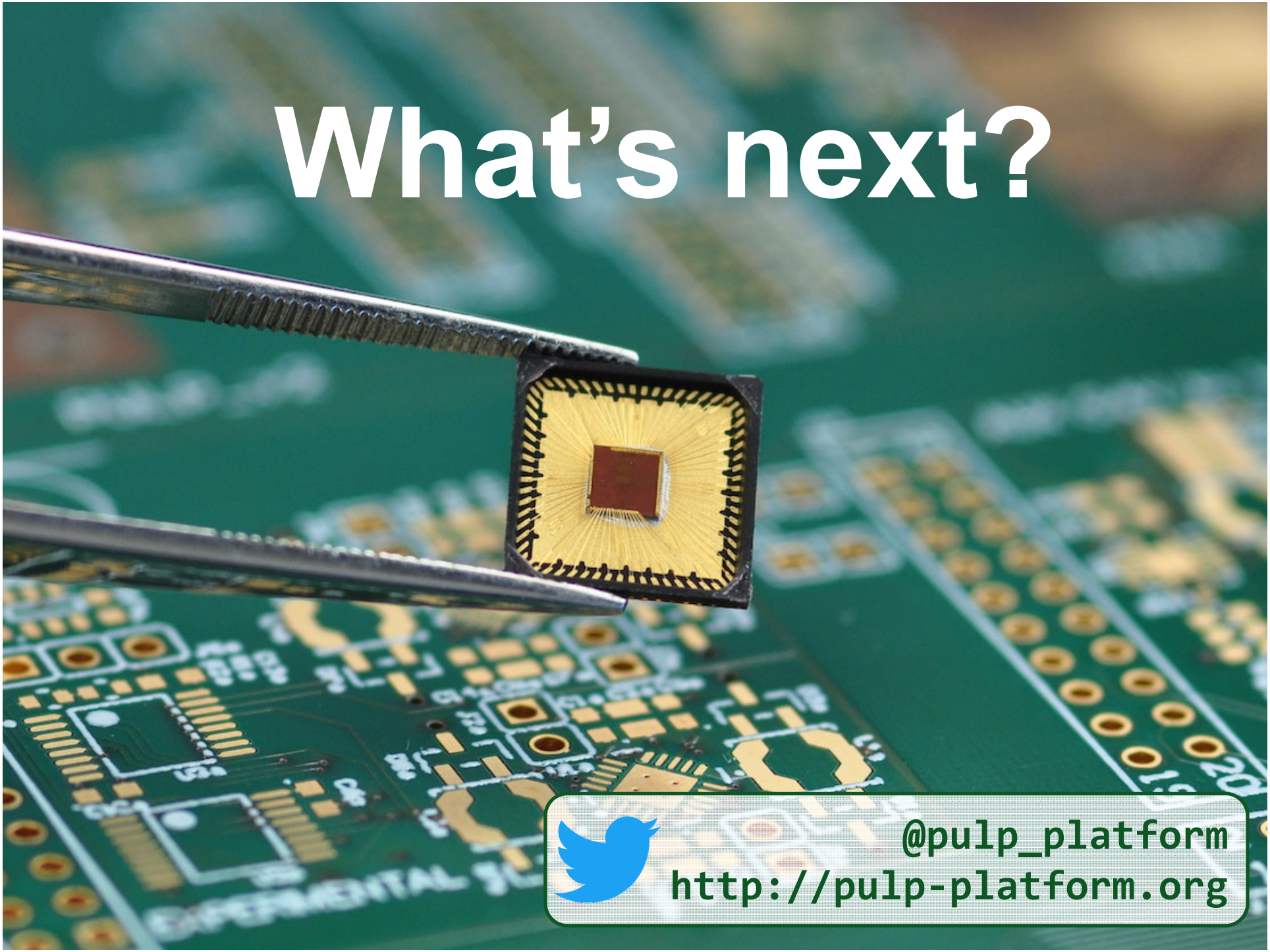
UMC 65nm LL

16 mm^2 die area, ca. 9 mm^2 logic core area

ca. 6 MGE logic core complexity,
ca. 400 KiB SRAMs in total

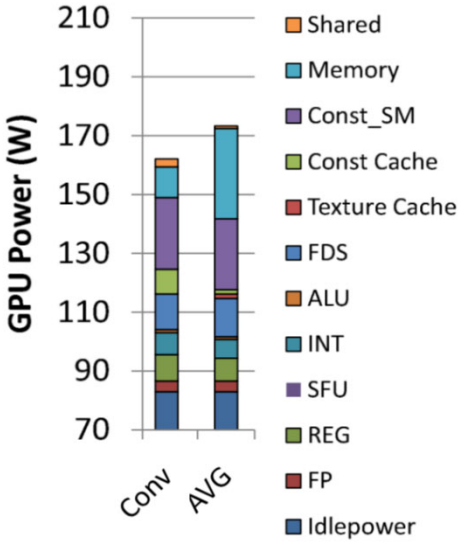
PULP

What's next?



Heterogeneous computing toward post-exascale

- Peak compute (GPU) 15TFLOP/s at 300W
  - 20x Better needed for post exascale: 1TWFLOP/W
- Only 5% power estimated to be spent in the FPUs [1]:
  - [1] reports 2.9%, but their kernels don't reach TDP/max perf.
  - In dubio pro Invidia: We scale power to assume modern GPUs do not exceed TDP at max perf. (making them more efficient)
  - **Key issue: GPU RF is SRAM: FMUL32 4pJ, SRAM 20pJ**



Graph extracted and cropped from [1].

64 FPUs
256 kB RF 128 kB L0 Cache
32-2048 threads

Volta Assembly

```
LDS  R2, [R0]
LDS  R3, [R1]
FFMA R4, R2, R3, R2
```

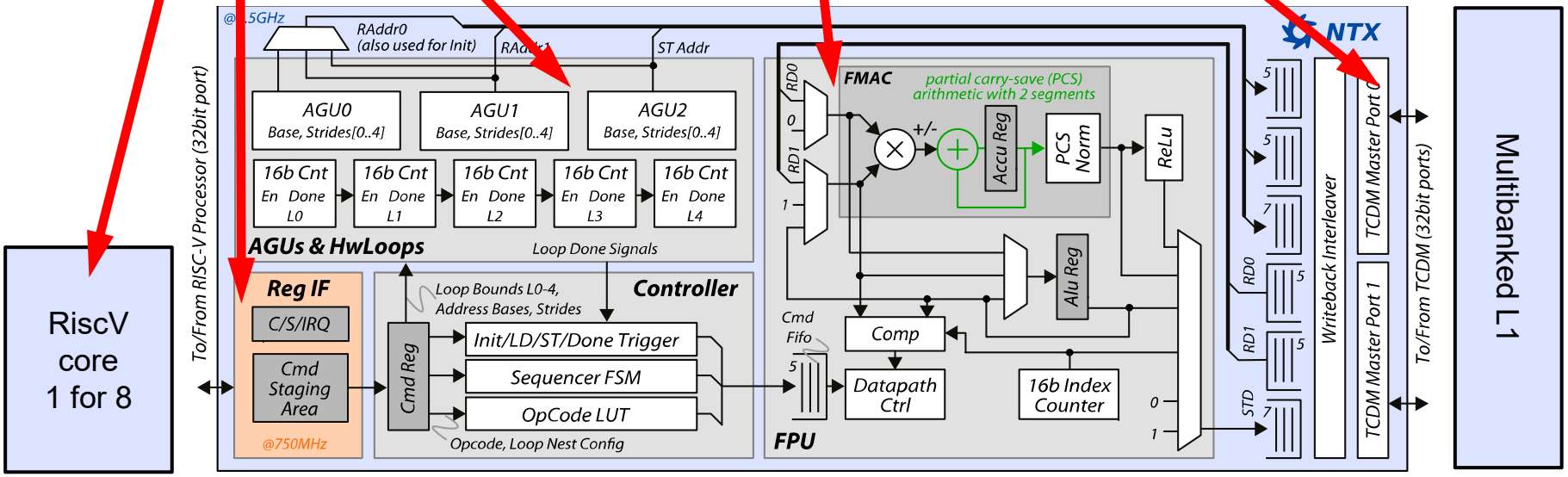
2 mem. acc. ("[…]")
8 reg. acc. Into RF SRAM

= 10 SRAM R/W total

[1] S. Hong and H. Kim, "An integrated gpu power and performance model," in ACM SIGARCH Computer Architecture News, 2010.

Network Training Accelerator (NTX)

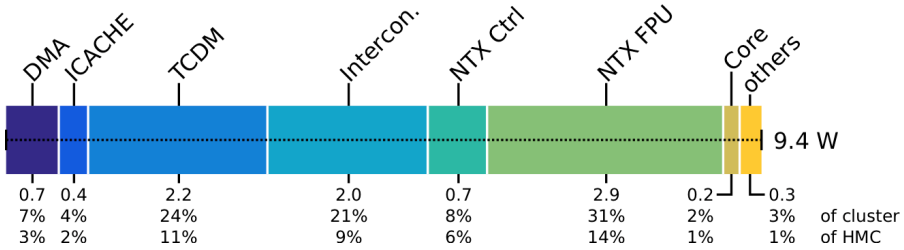
- Processor configures Reg IF and manages DMA double-buffering in L1 memory
- Controller issues AGU, HWL, and FPU micro-commands based on configuration
- AGUs generate address streams for data access
- FMAC with extended precision + ML functions
- Reads/writes data via 2 memory ports (2 operand and 1 writeback streams)



Again: specialized “deep interfaces” + Instruction extensions

NTX Power Breakdown & GPU SM Comparison

- NTX dissipates significant fraction of power in its FPU (more is better):
  - 31% of cluster
  - 14% of entire if we account for Main Mem
  - Recall: GPU is just around 5% [1]



- Compared to NVIDIA Volta GPU [2]:
  - Register file in GPU holds registers and thread-local data
  - Each register read/write is an SRAM access
  - Register and data accesses compete for SRAM

1 Volta SM	8 NTX cl.
64 FPUs	64 FPUs
256 kB RF 128 kB L0 Cache	512 kB TCDM
32-2048 threads	8 threads

Volta Assembly

```
LDS  R2, [R0]
LDS  R3, [R1]
FFMA R4, R2, R3, R2
```

2 mem. acc. (“[…]”)
8 reg. acc.

= 10 SRAM hits total

NTX Pseudocode

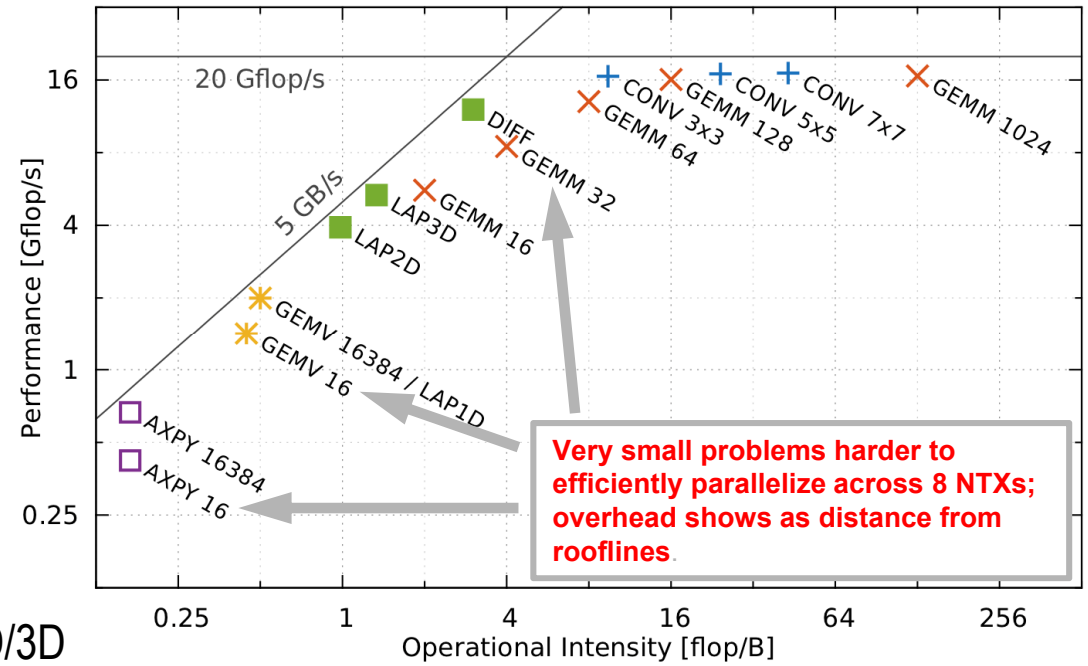
```
FMAC accu, [AGU0], [AGU1]
```

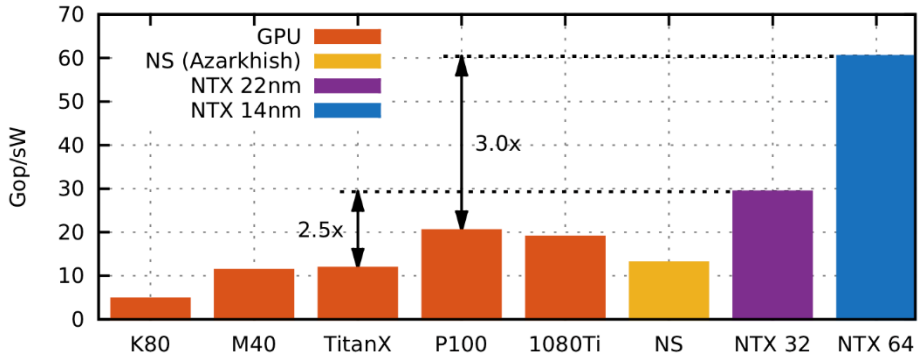
2 mem. acc. (“[…]”)
0 reg. acc.
(+ addr. calc for free)

= 2 SRAM hits total

NTX Roofline and efficiency

- NTX achieves high utilization of available bandwidth and compute
- We investigate a range of different kernels:
- Linear Algebra
  - Mat-Mat product (GEMM)
  - Mat-Vec product (GEMV)
  - Vector sum (AXPY)
- Stencils
  - Discrete Laplace Operator in 1D/2D/3D
  - Diffusion
- Deep Learning
- **2 to 3x more efficient than GPGPU**

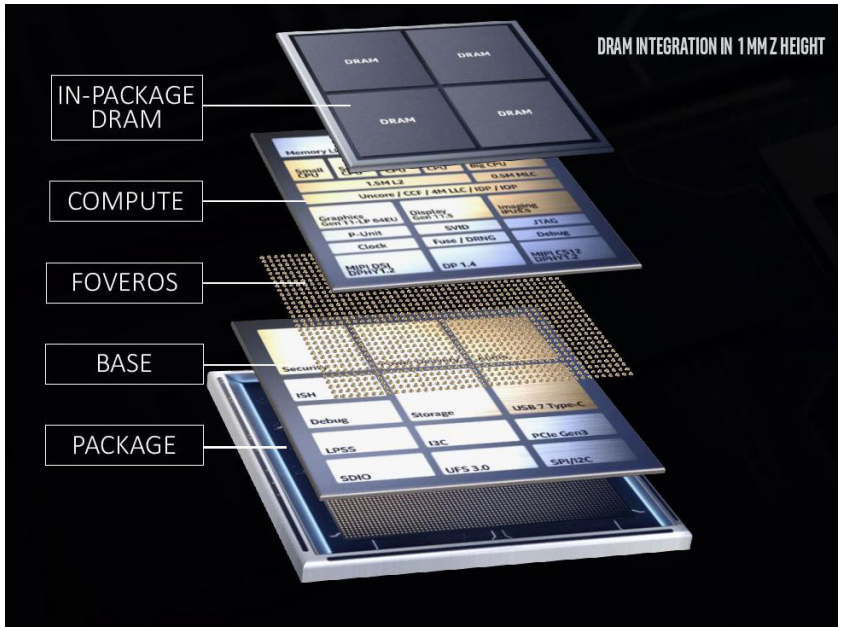




Technology to the rescue

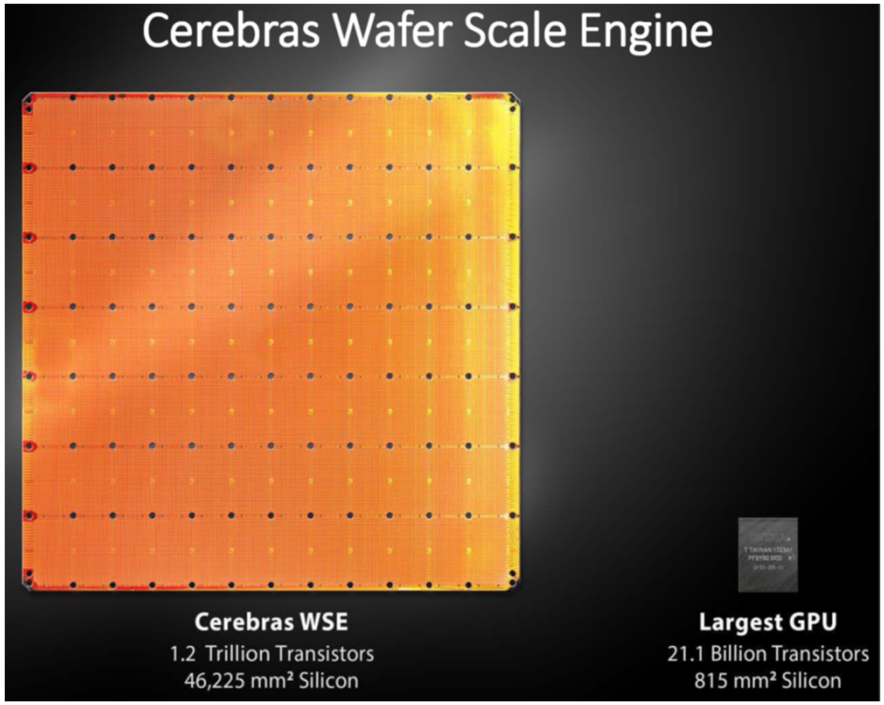
What about the ~30% of power that goes in the memory interface?

Reduce pJ/B to access main mem



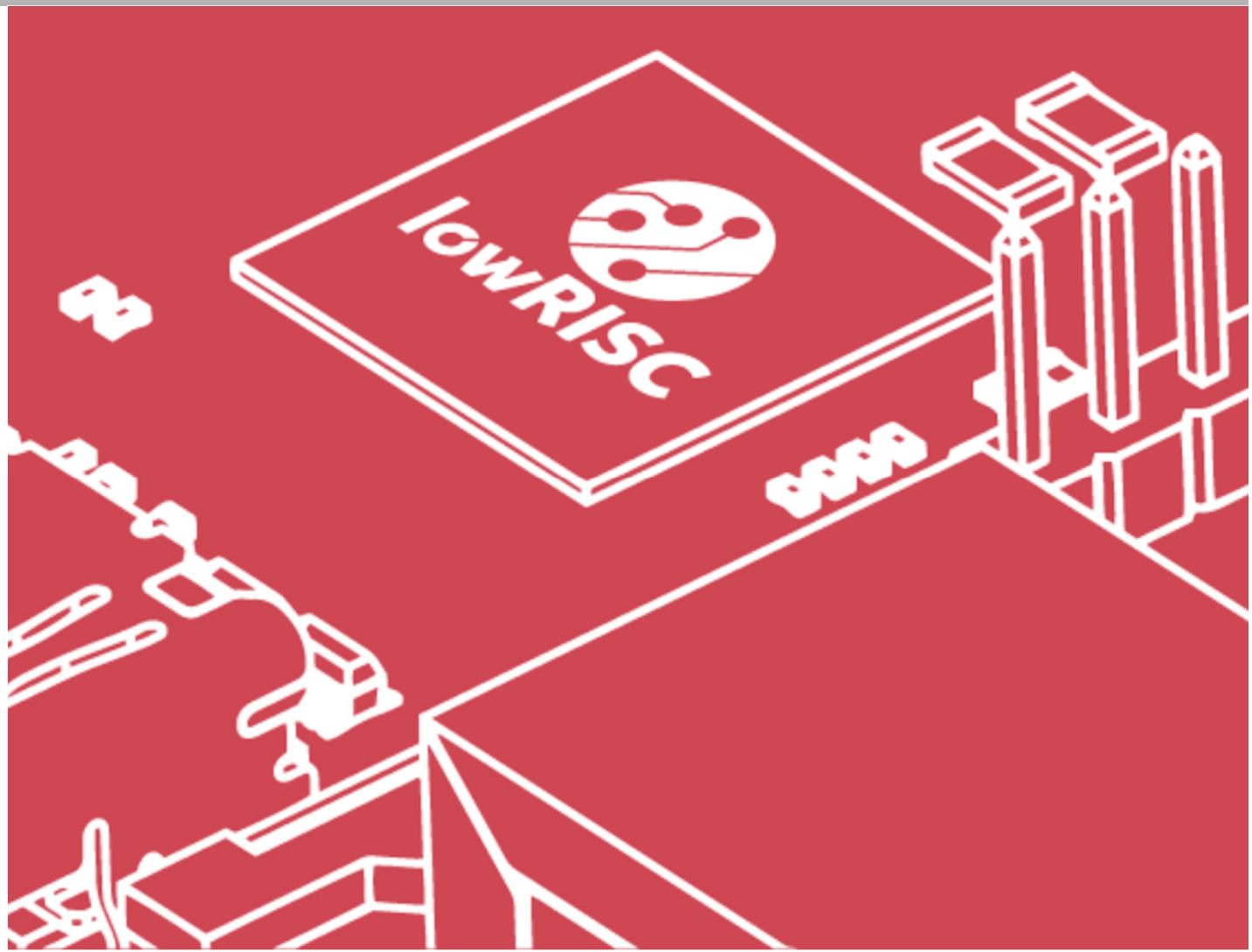
Intel's upcoming 3D-stacked processor, codename Lakefield

Reduce the number of accesses…



Industrial open SW Hardware

lowRISC Community Interest Company



lowRISC

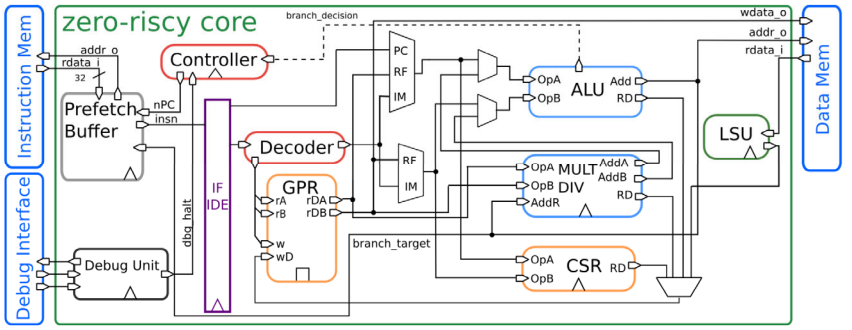
enabling open source silicon through collaborative engineering

PULP

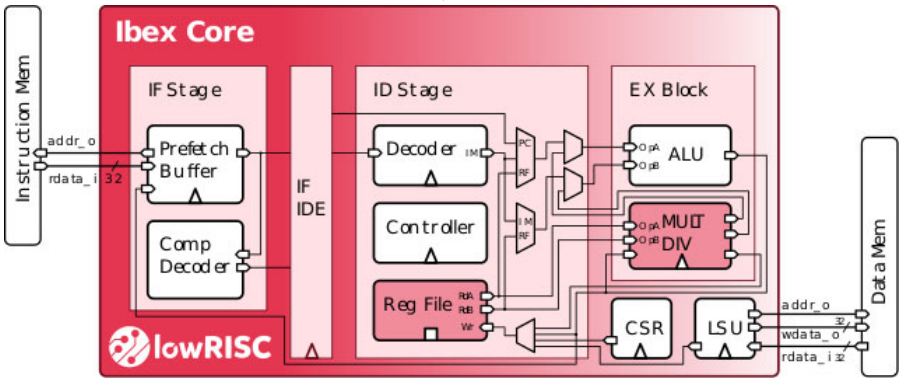
LowRISC is up and… hiring

Google UNIVERSITY OF CAMBRIDGE ETH zürich

Alex Bradbury, Dr Gavin Ferris, Dr Robert Mullins
Prof. Luca Benini, Ron Minnich, Dominic Rizzo



Zero-Riscy (RV32-ICM), 19kGE



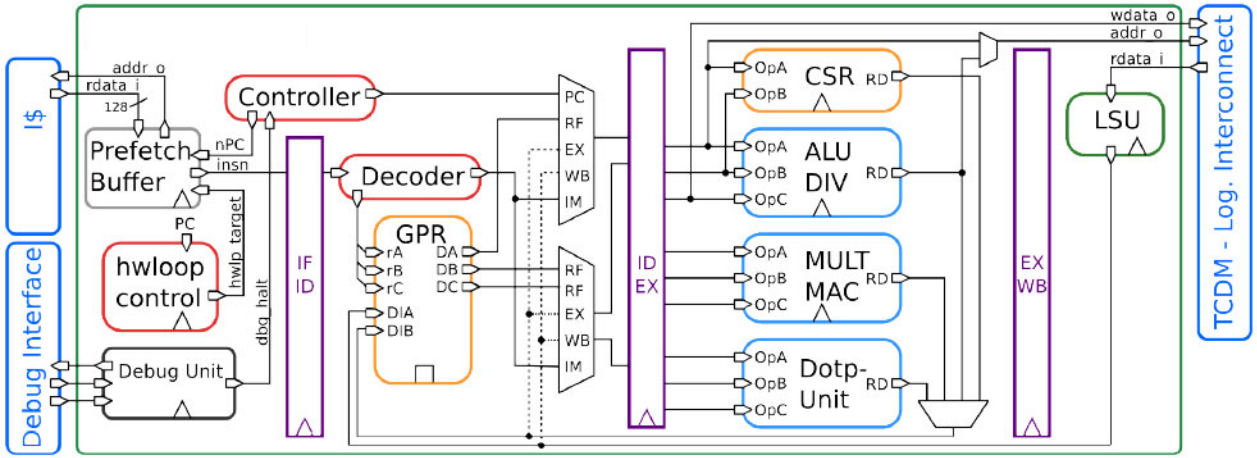
Will one NFP Company be Enough?

https://www.openhwgroup.org/

OpenHW Group Charter

OpenHW Group is a not-for-profit, global organization driven by its members and individual contributors where hardware and software designers collaborate in the development of open-source cores, related IP, tools and software such as the **CORE-V Family of cores**. OpenHW provides an infrastructure for hosting high quality open-source HW developments in line with industry best practices.

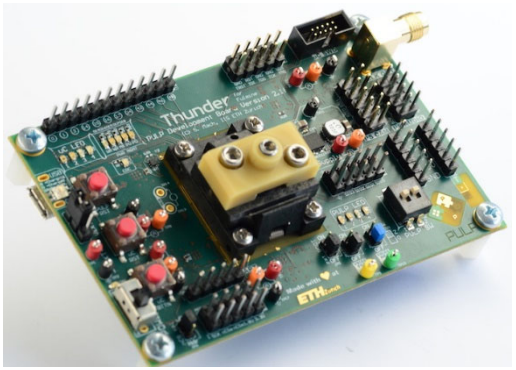
RI5CY core

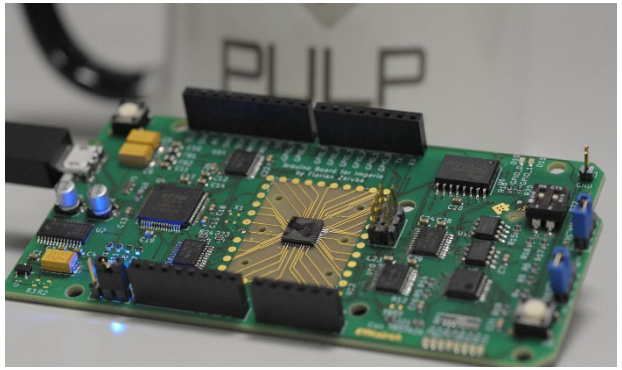


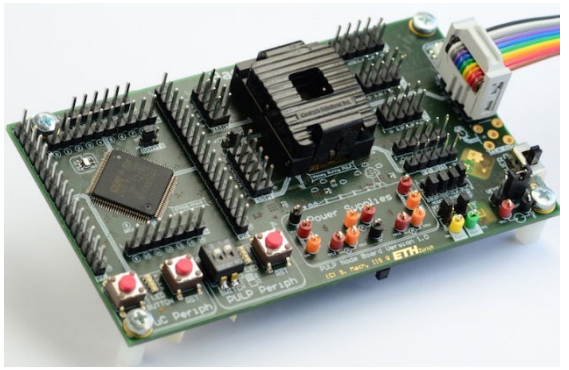
CORE-V™

R. O'Connor (OpenHW CEO, former RISC-V foundation director)

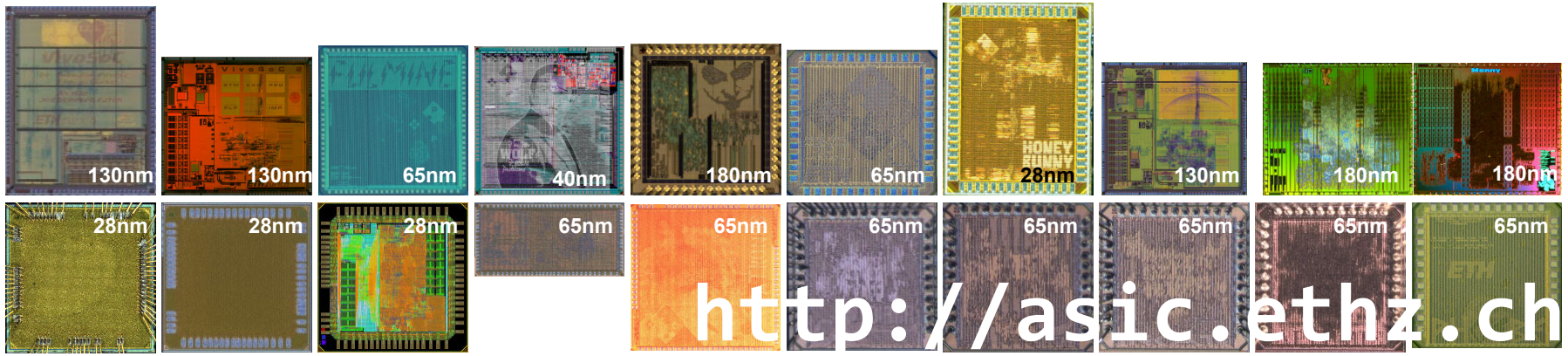
PULP



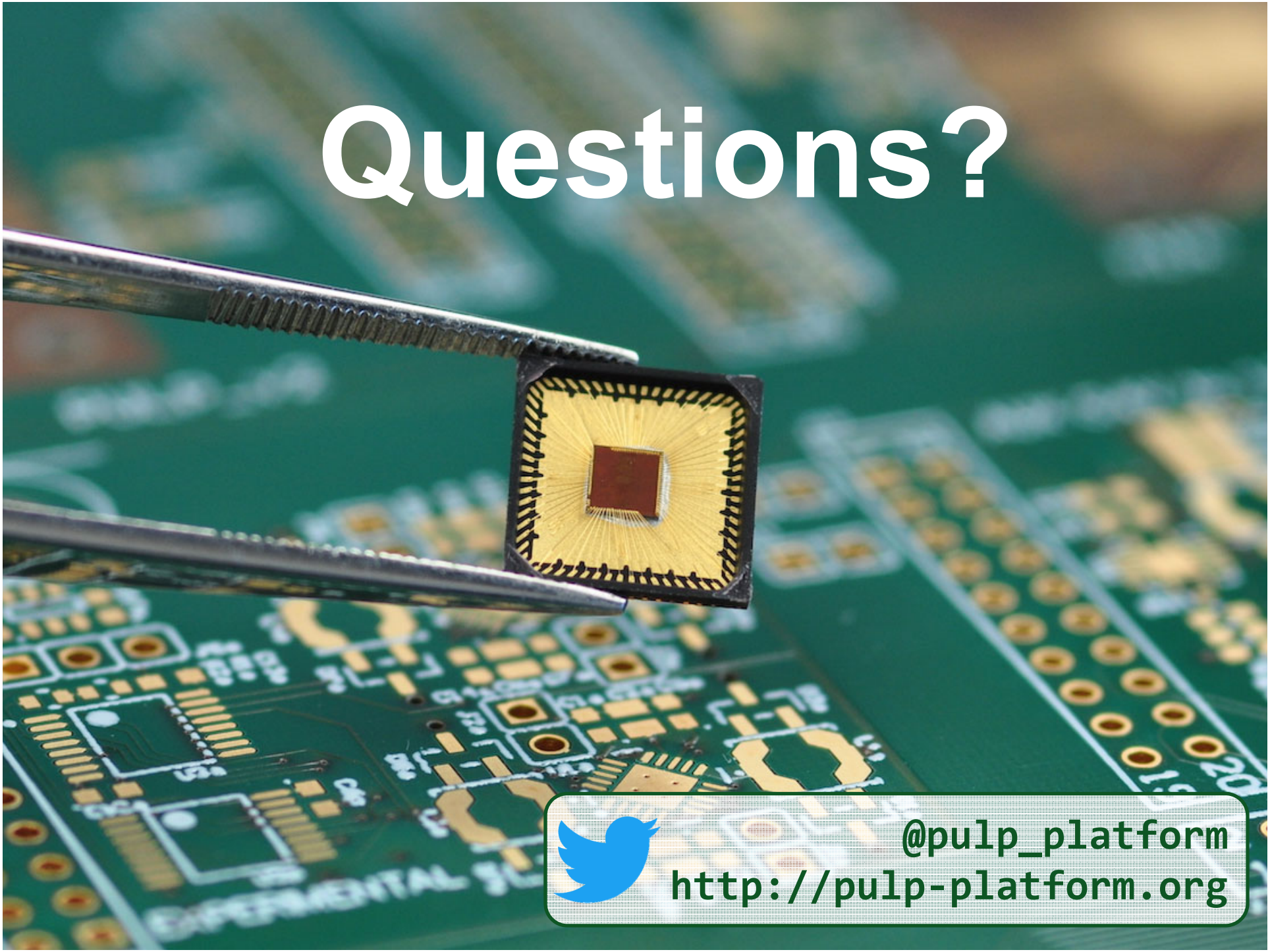




www.pulp-platform.org



The fun is just beginning...


Questions?
@pulp_platform
http://pulp-platform.org