PULP
Parallel Ultra Low Power

Energy efficient computing from Exascale to MicroWatts: The RISC-V playground

Zürich, RISC-V Workshop *11.06.2019*

Luca Benini[1,2]

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

ETH zürich

[2] *Integrated Systems Laboratory*

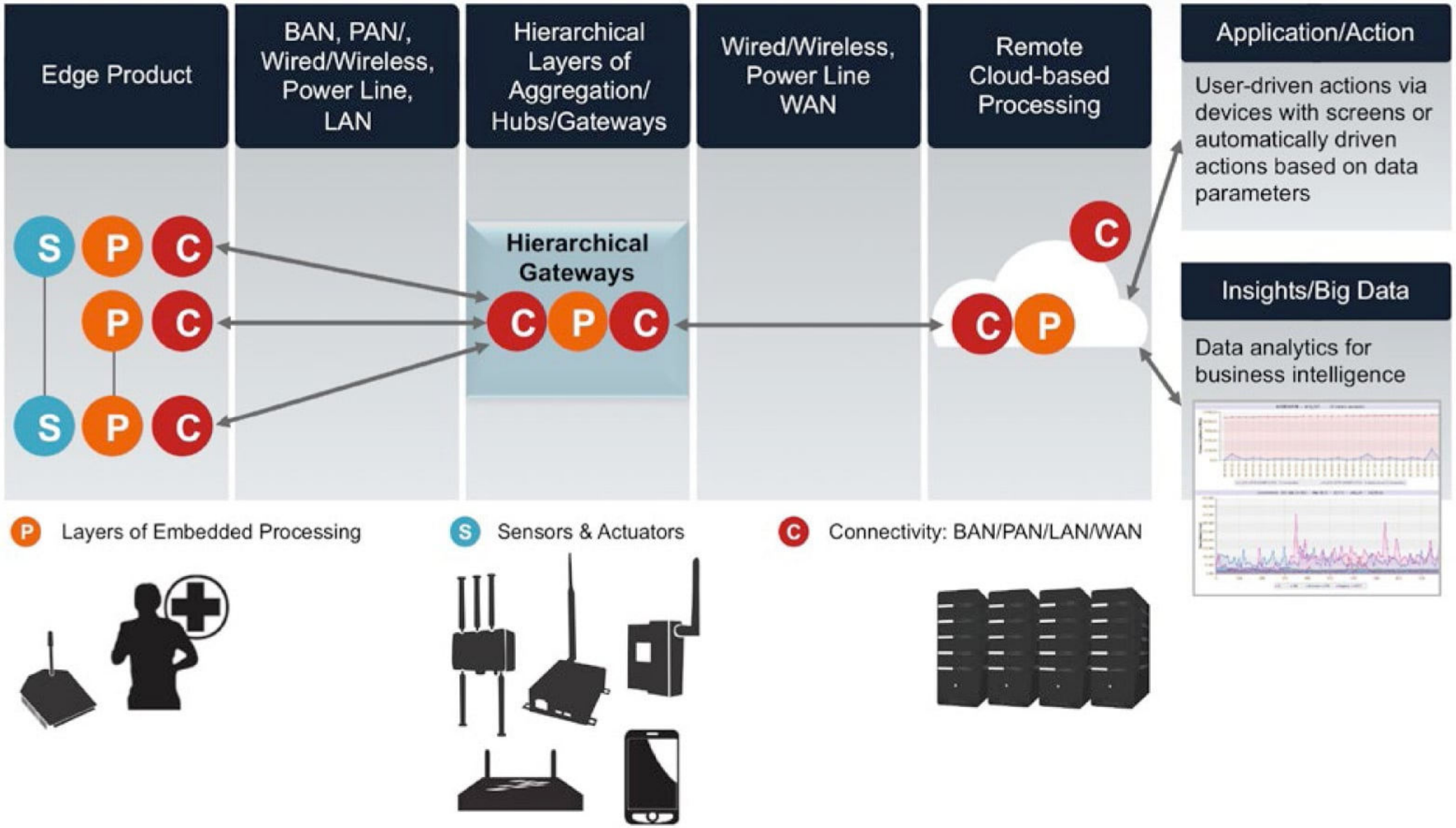
erc
European Research Council

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

IoT Hierarchical Processing (Compute Continuum)



Edge Product

BAN, PAN/, Wired/Wireless, Power Line, LAN

Hierarchical Layers of Aggregation/ Hubs/Gateways

Wired/Wireless, Power Line WAN

Remote Cloud-based Processing

Application/Action

User-driven actions via devices with screens or automatically driven actions based on data parameters

Insights/Big Data

Data analytics for business intelligence

P Layers of Embedded Processing

S Sensors & Actuators

C Connectivity: BAN/PAN/LAN/WAN

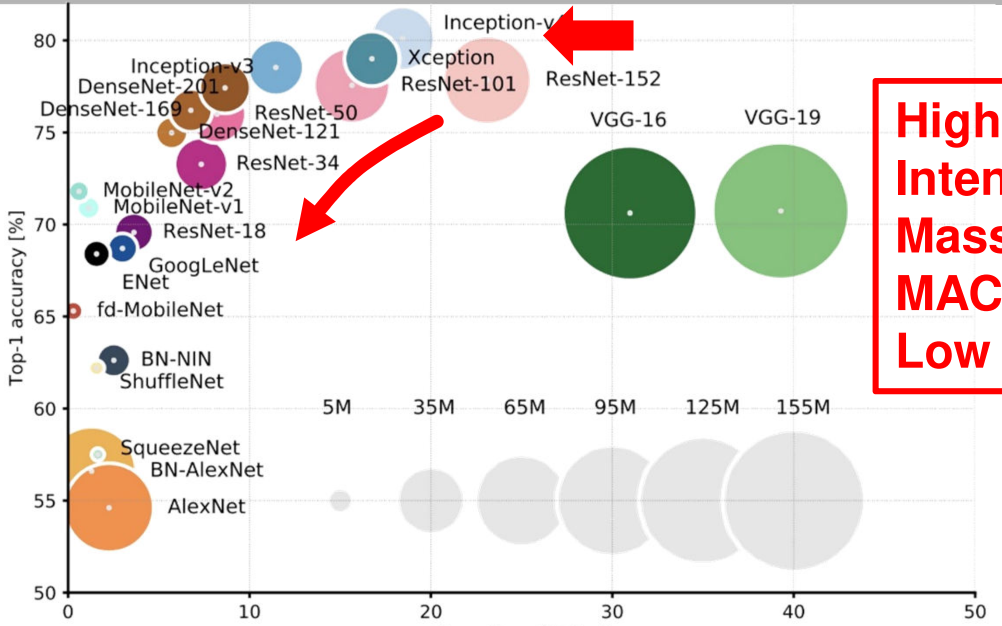
Near-Sensor Processing

Embedded Processing

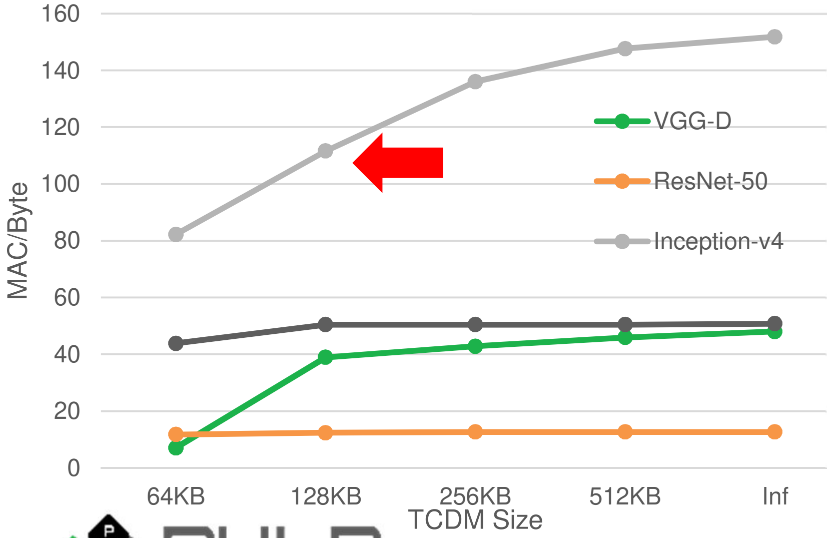
Cloud processing

AI Workloads from Cloud to Near-Sensor

GOP+



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



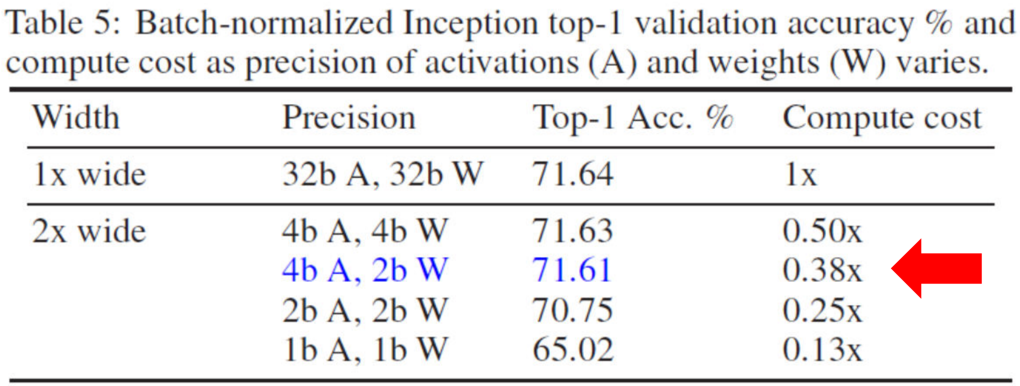
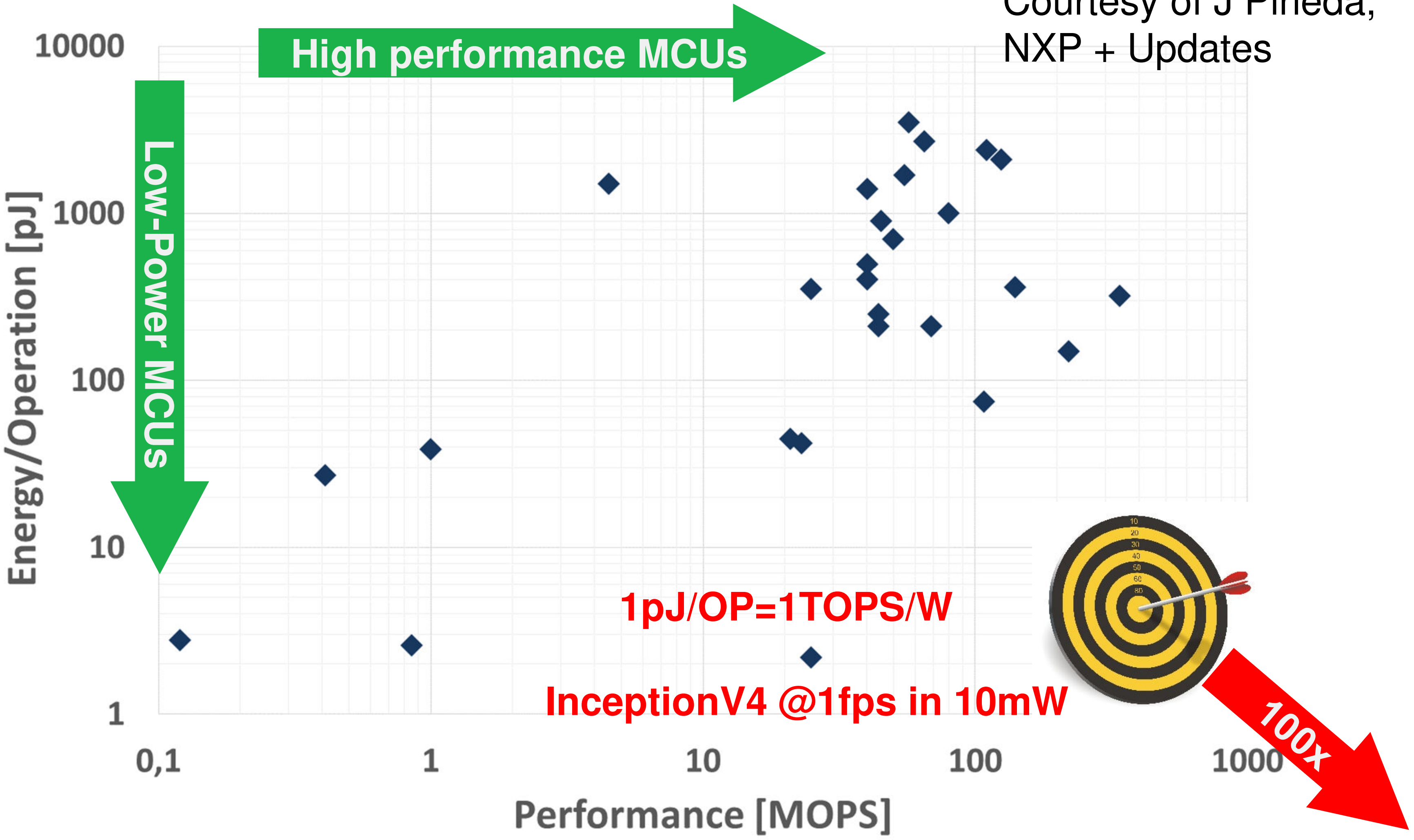


Table 5: Batch-normalized Inception top-1 validation accuracy % and compute cost as precision of activations (A) and weights (W) varies.

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]

Energy efficiency is THE Challenge



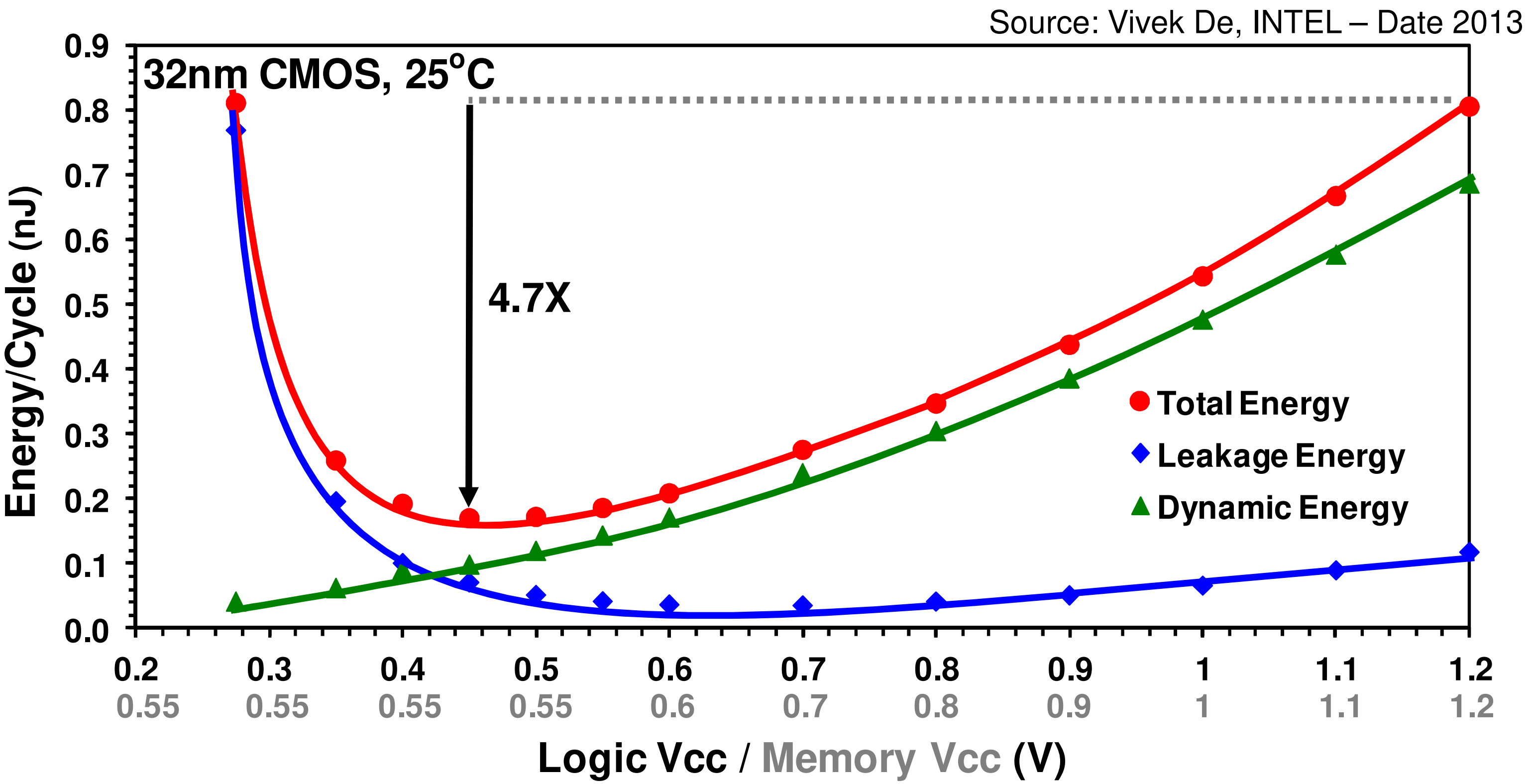
Courtesy of J Pineda, NXP + Updates

1pJ/OP=1TOPS/W

InceptionV4 @1fps in 10mW

Cool… But, HOW??

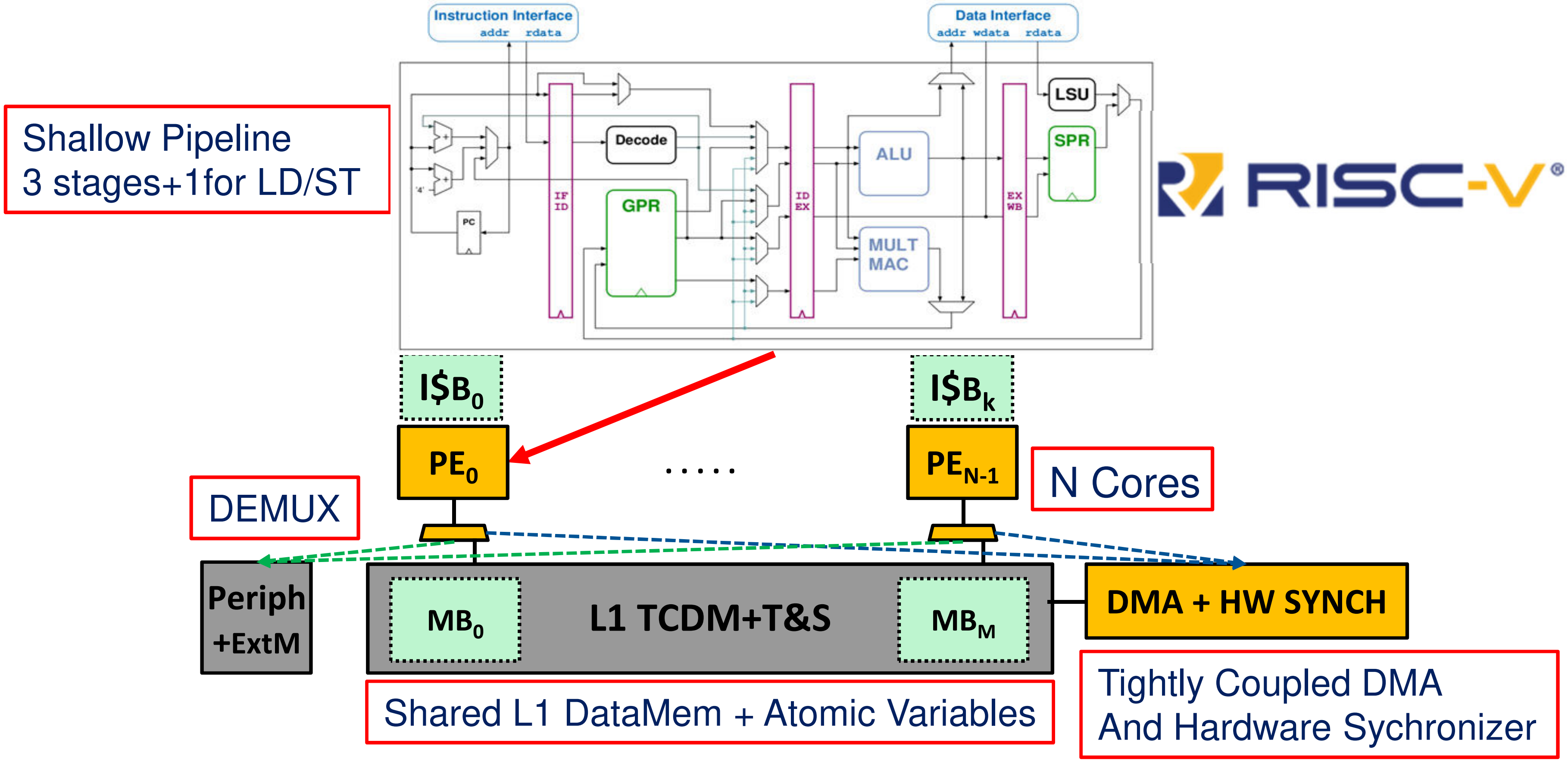
2013: Parallel Ultra Low Power →PULP!



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!!!**

Near-Threshold Multiprocessing



Shallow Pipeline
3 stages+1for LD/ST

DEMUX

N Cores

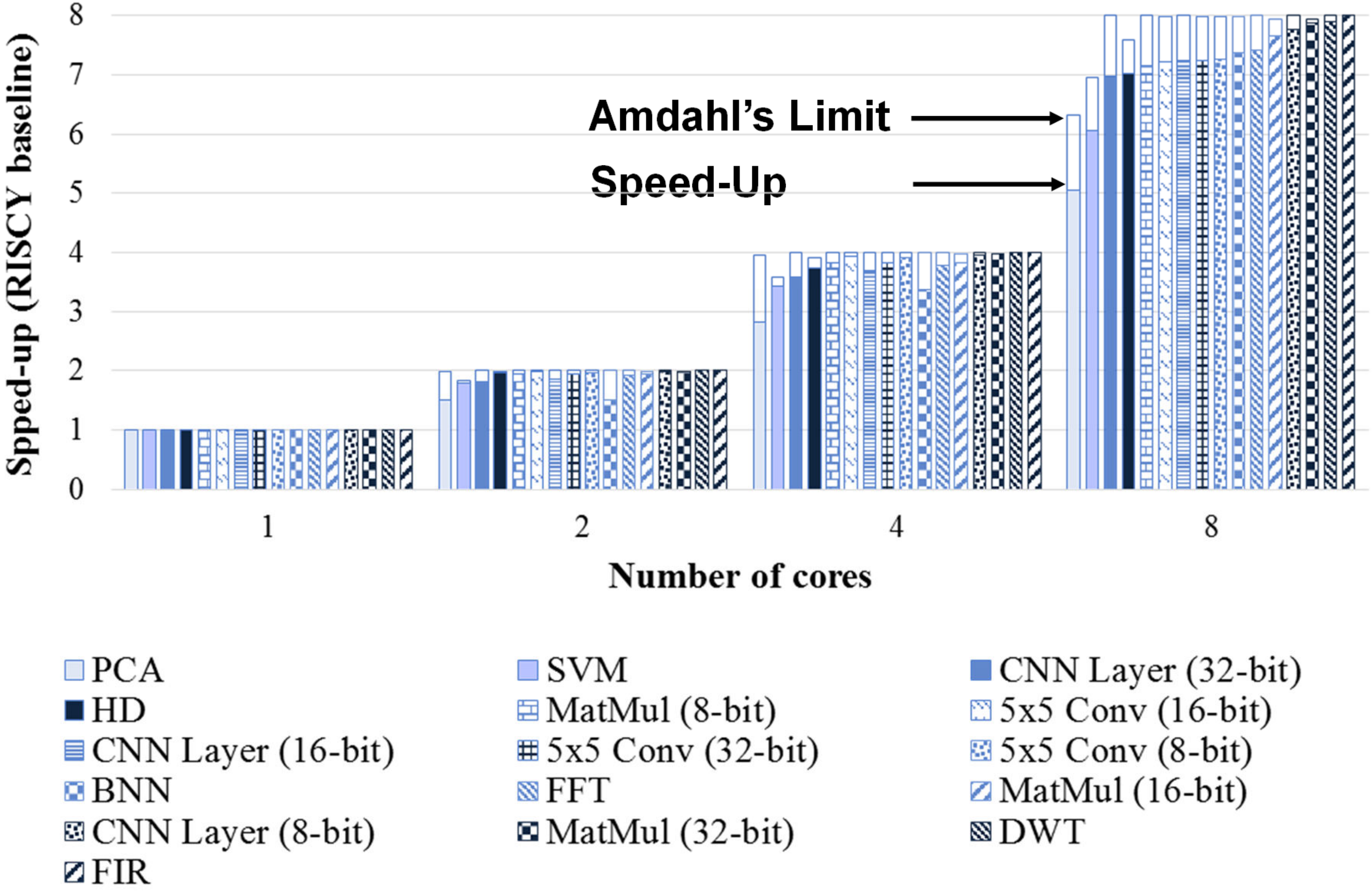
Shared L1 DataMem + Atomic Variables

Tightly Coupled DMA
And Hardware Sychronizer

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

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

8-Processor PULP Cluster: Parallel Speed-up



Bespoke ISA needed! Enter Xpulp extensions

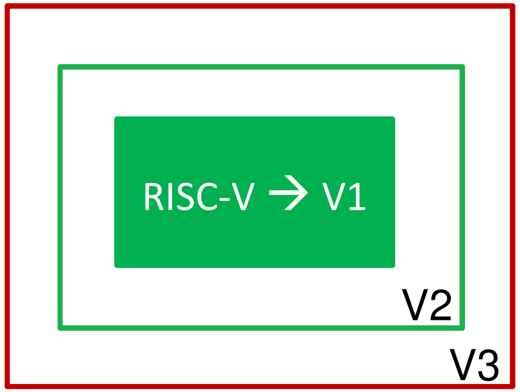
<32-bit precision → **SIMD2/4 → x2,4 efficiency & memory size**

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

V1 Baseline RISC-V RV32IMC

V2 HW loops
Post modified Load/Store
Mac

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

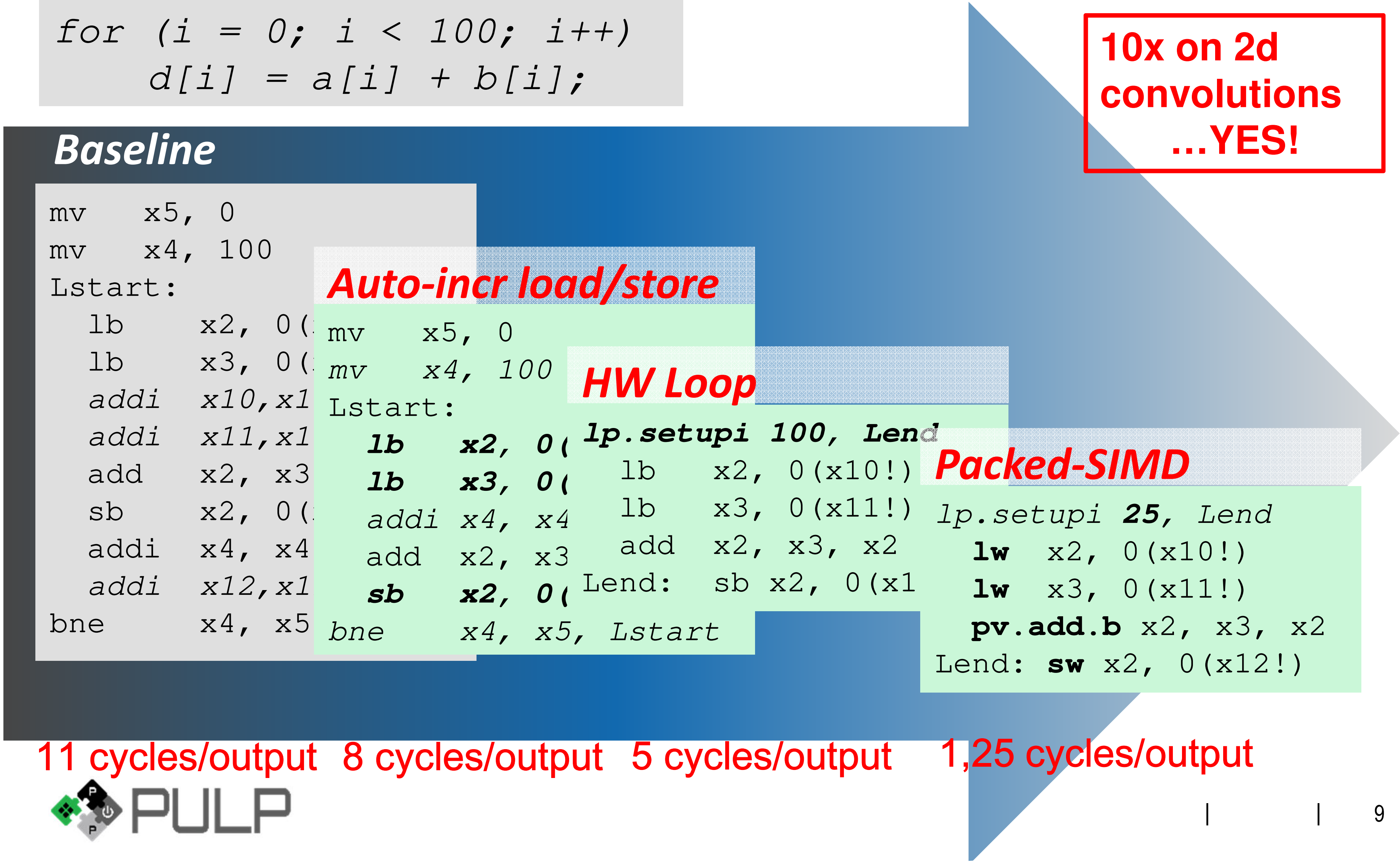


25KG → 40KG (1.6x)

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
```

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
```

HW Loop

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

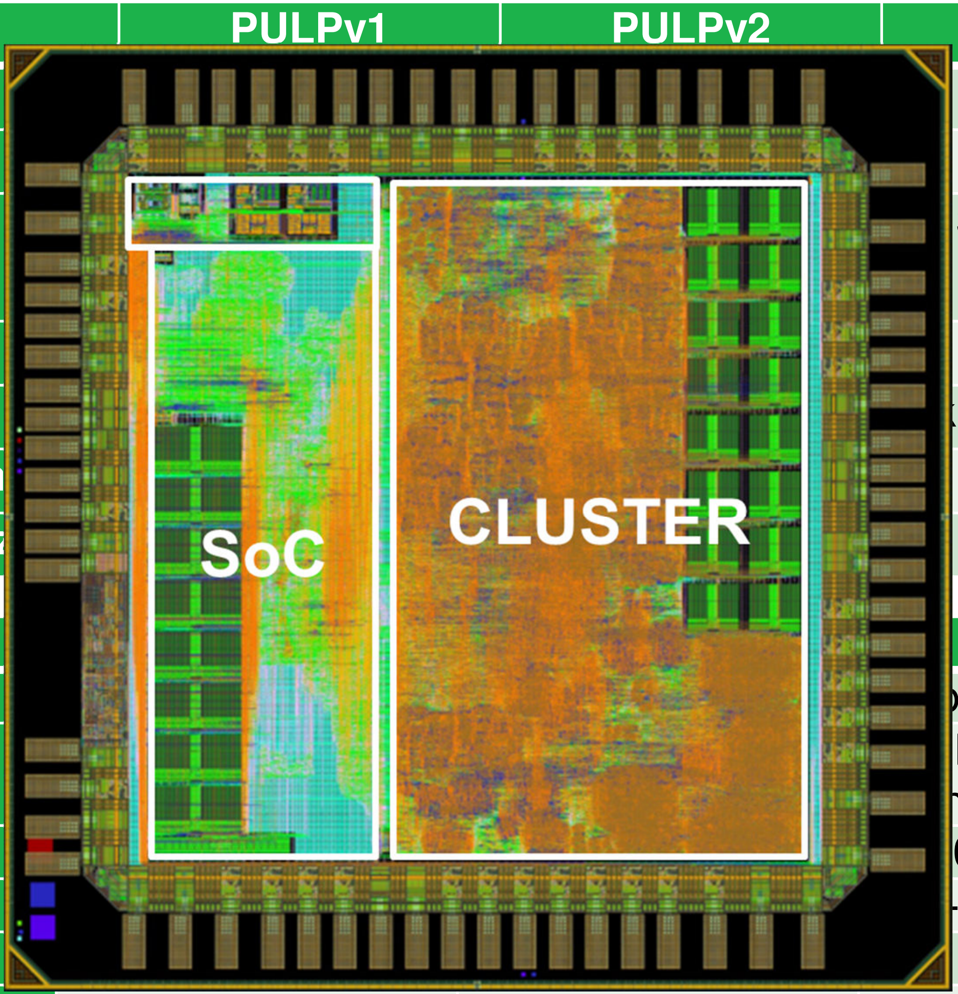
Packed-SIMD

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

11 cycles/output   8 cycles/output   5 cycles/output   1,25 cycles/output

The Evolution of the ‘Species’

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

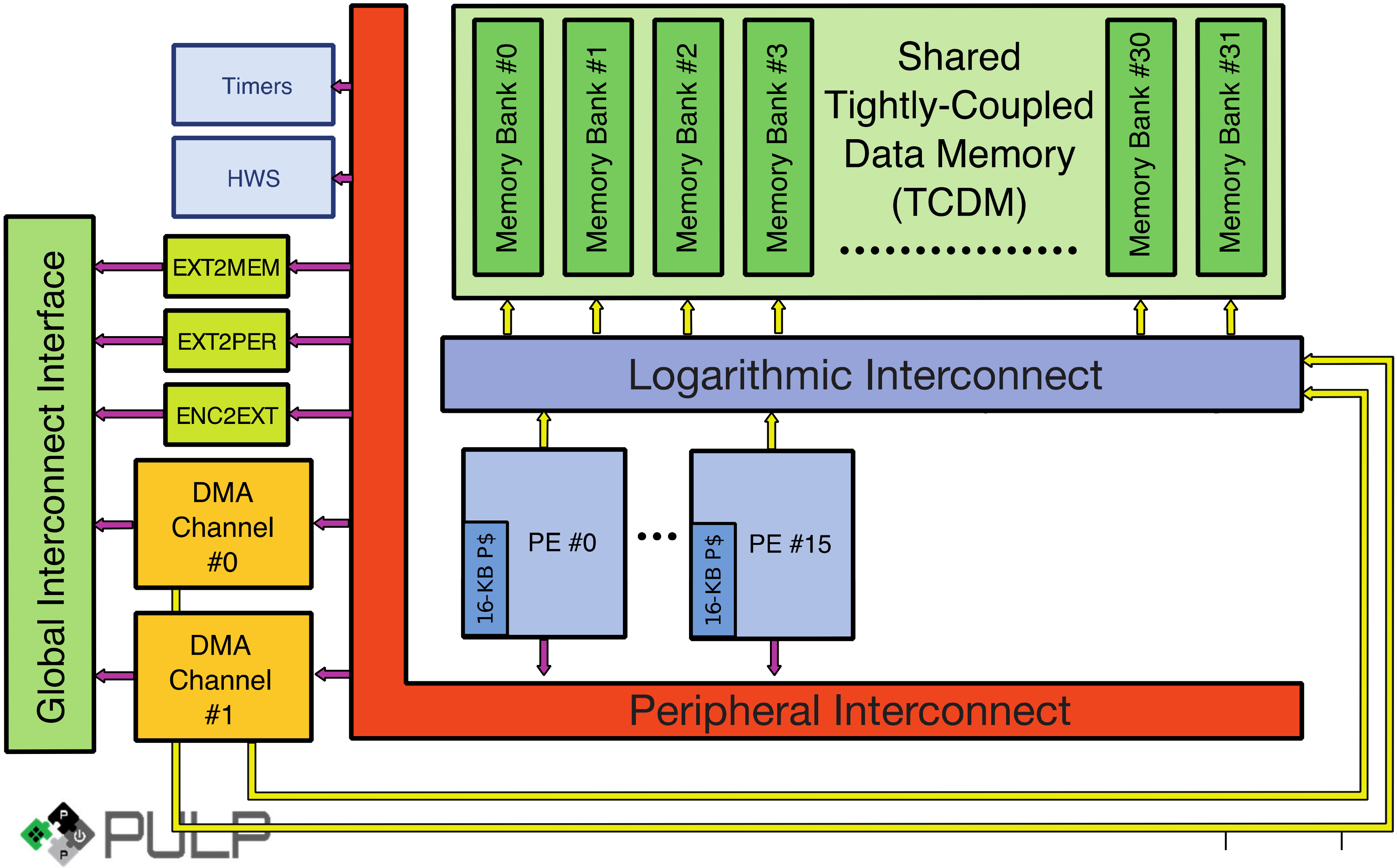


			PULPv3
Status			post tape out
Technology			D-S nventional well
Voltage range			0.5V - 0.7V
BB range			-1.8V - 0.9V
Max freq.			200 MHz
Max perf.	1.9 GOPS	4 GOPS	1.8 GOPS
Peak en. eff.	60 GOPS/W	135 GOPS/W	385 GOPS/W

2.6pj/OP

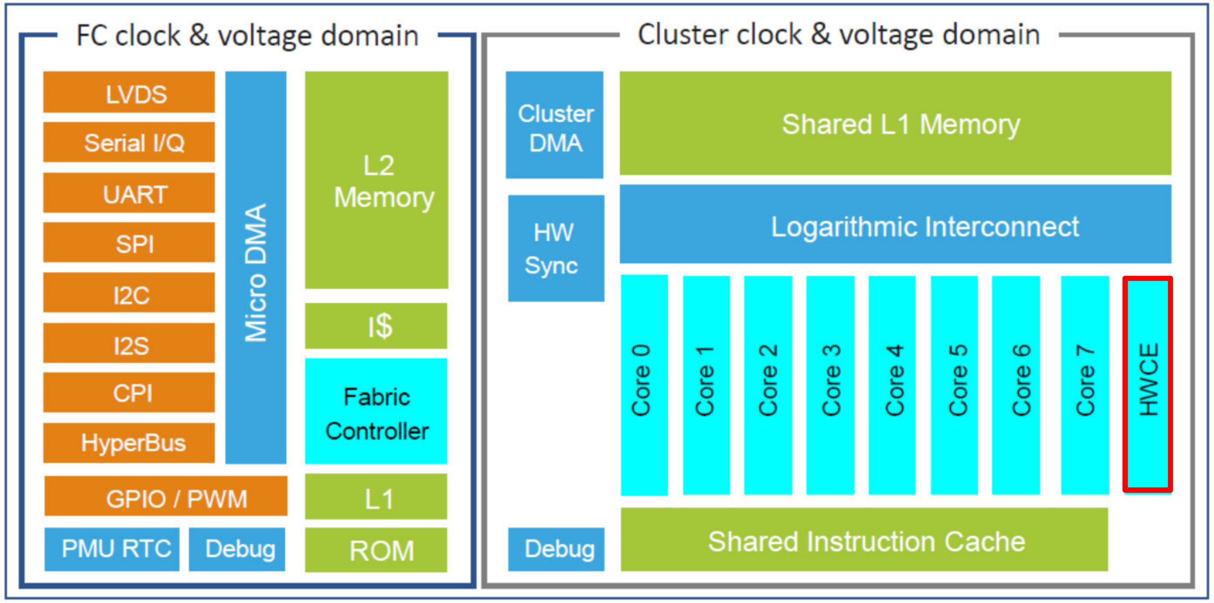
Ultra-simplified Open HW release: 1-core PULPINO

More efficiency: Heterogeneous PULP Cluster



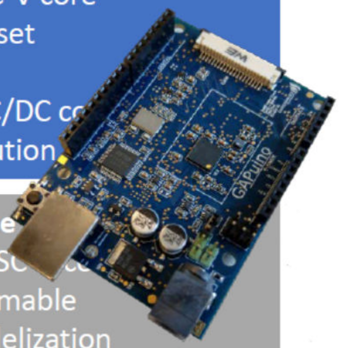
PULP cluster+MCU+HWCE → GWT's GAP8 (55 TSMC)

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 co
- 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

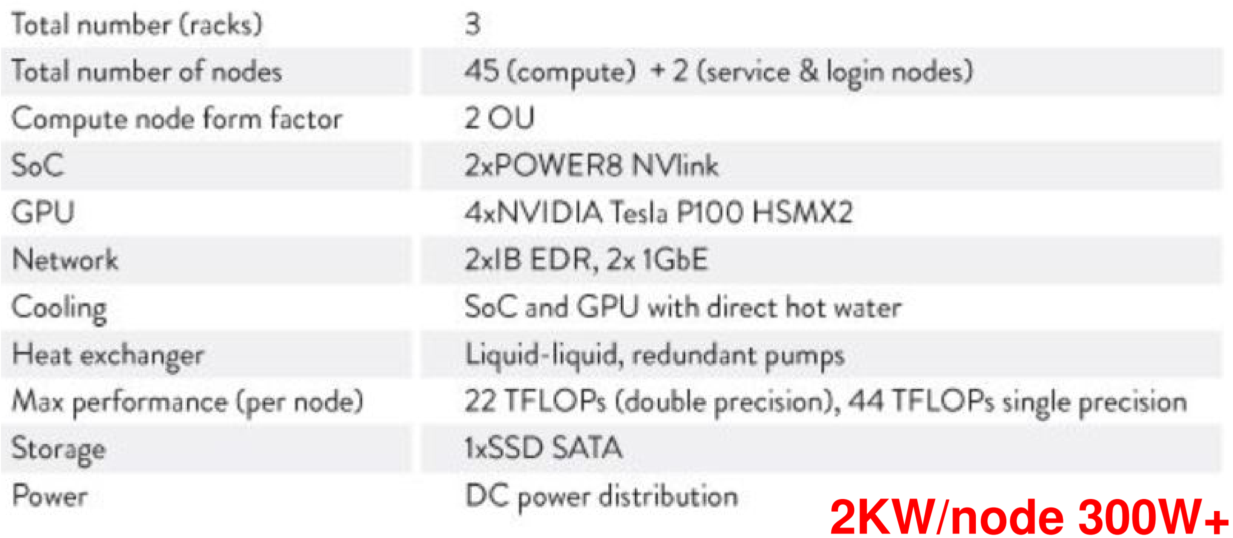
GREENWAVES TECHNOLOGIES

4x More efficiency at less than 10% area cost

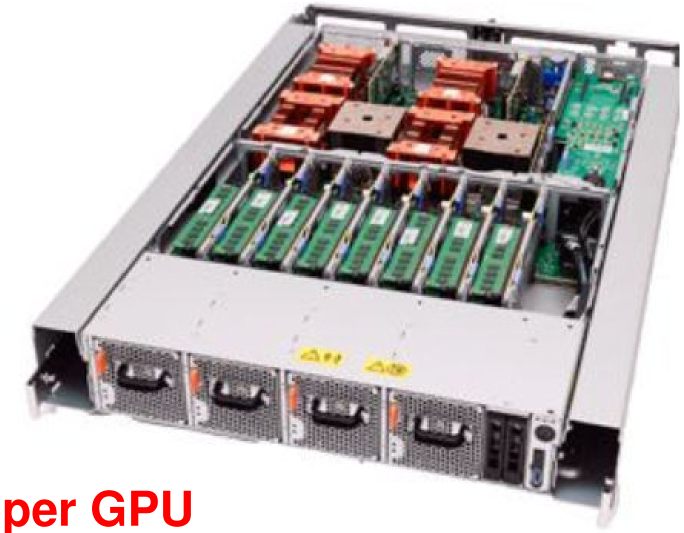
PULP

Back to the Cloud

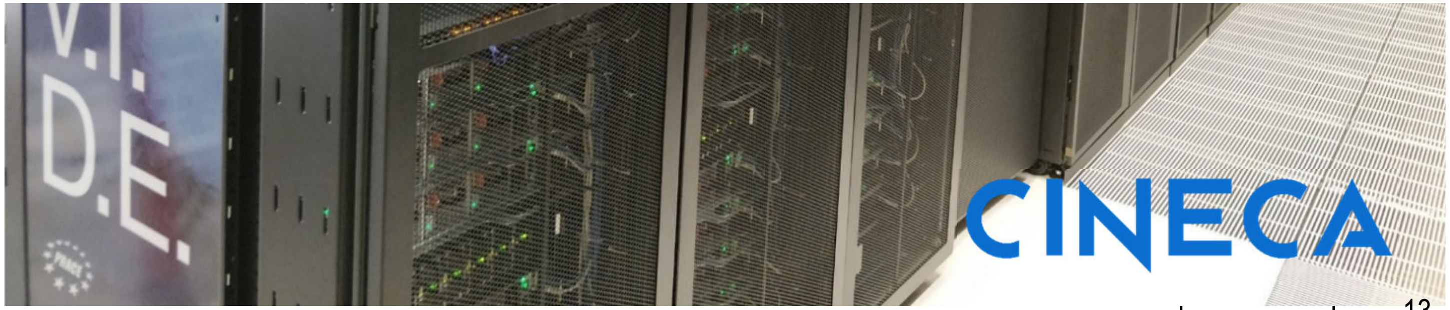
1 PFLOPS, top 20 in GREEN500’17



Total number (racks)	3
Total number of nodes	45 (compute) + 2 (service & login nodes)
Compute node form factor	2 OU
SoC	2xPOWER8 NVlink
GPU	4xNVIDIA Tesla P100 HSMX2
Network	2xIB EDR, 2x 1GbE
Cooling	SoC and GPU with direct hot water
Heat exchanger	Liquid-liquid, redundant pumps
Max performance (per node)	22 TFLOPs (double precision), 44 TFLOPs single precision
Storage	1xSSD SATA
Power	DC power distribution



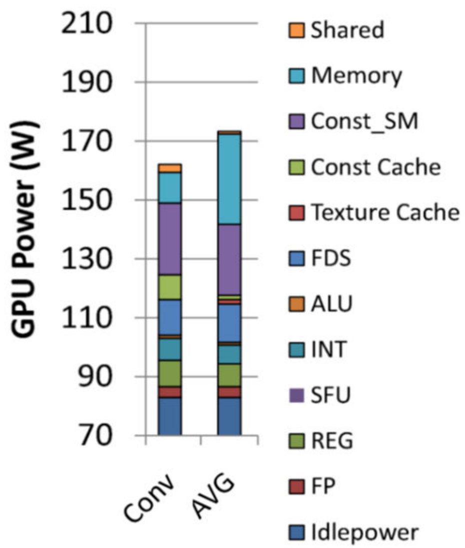
2KW/node 300W+ per GPU



Addressing GPUs Weaknesses

- Peak compute reaches 15 Tflop/s these days

- Only 5% of that 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 point: GPU RF is SRAM (remember 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
```

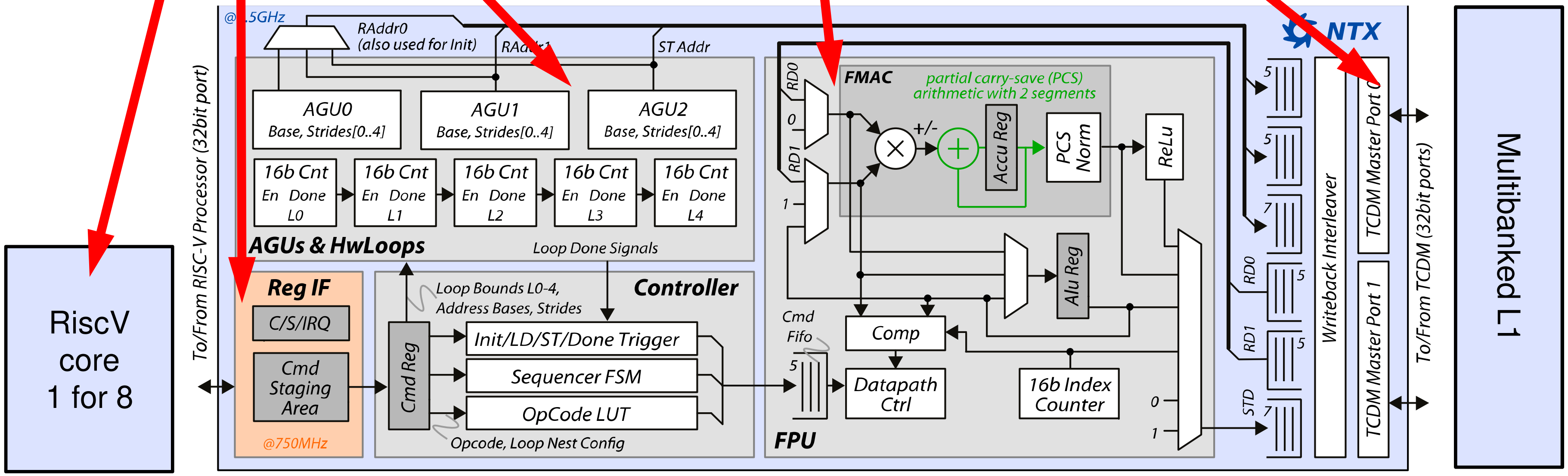
2 mem. acc. ("[…]")
8 reg. acc. Into RF
SRAM

= 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)

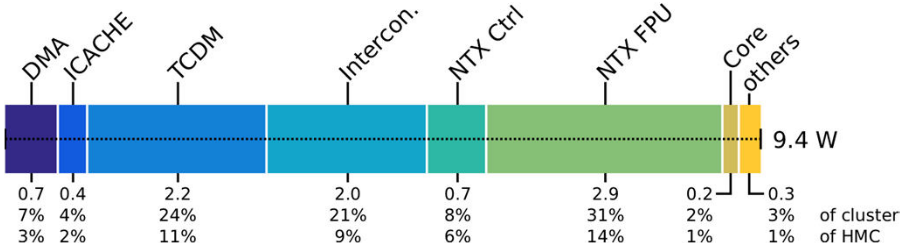
- Processor configures Reg IF and manages DMA double-buffering in L1 memory
- Controller issues AGU, HWL, and FPU micro-commands based on configuration
- AGUs generate address streams for data access
- FMAC with extended precision + ML functions
- Reads/writes data via 2 memory ports (2 operand and 1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

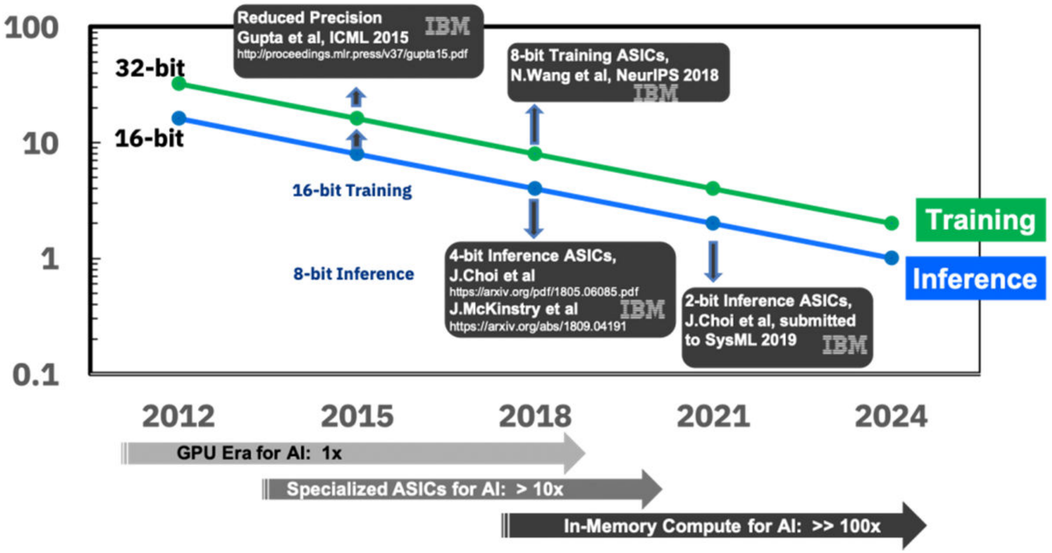
NTX Pseudocode

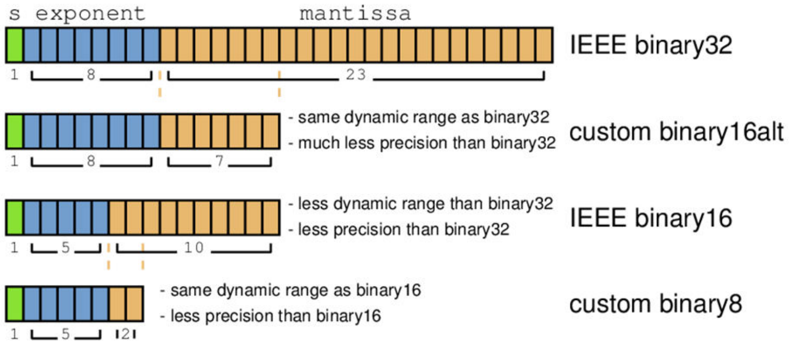
```
FMAC accu, [AGU0], [AGU1]
```

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

= 2 SRAM hits total

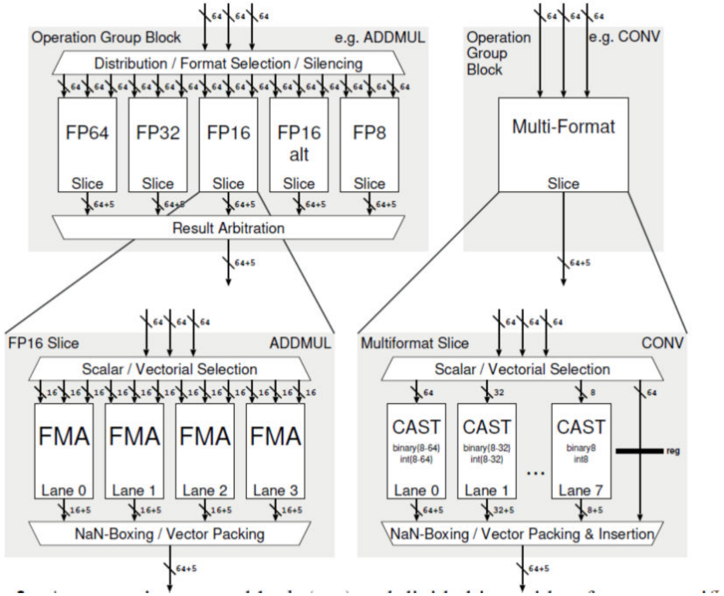
Low Precision Formats for Training





s
exponent
mantissa
1
8
23
IEEE binary32
1
8
7
- same dynamic range as binary32
- much less precision than binary32
custom binary16alt
1
5
10
- less dynamic range than binary32
- less precision than binary32
IEEE binary16
1
5
2
- same dynamic range as binary16
- less precision than binary16
custom binary8

- Flexible (cycle by cycle) precision modulation (FP)
- Save precious DRAM bandwidth
  - Custom number formats
    - Use float8, float16, float16alt
  - Transprecision FPU (~pJ/FLOP @1GHz)
    - 0.4pJ FP8, 0.9pJ FP16, 2.4pJ FP32, 6.2pj FP64



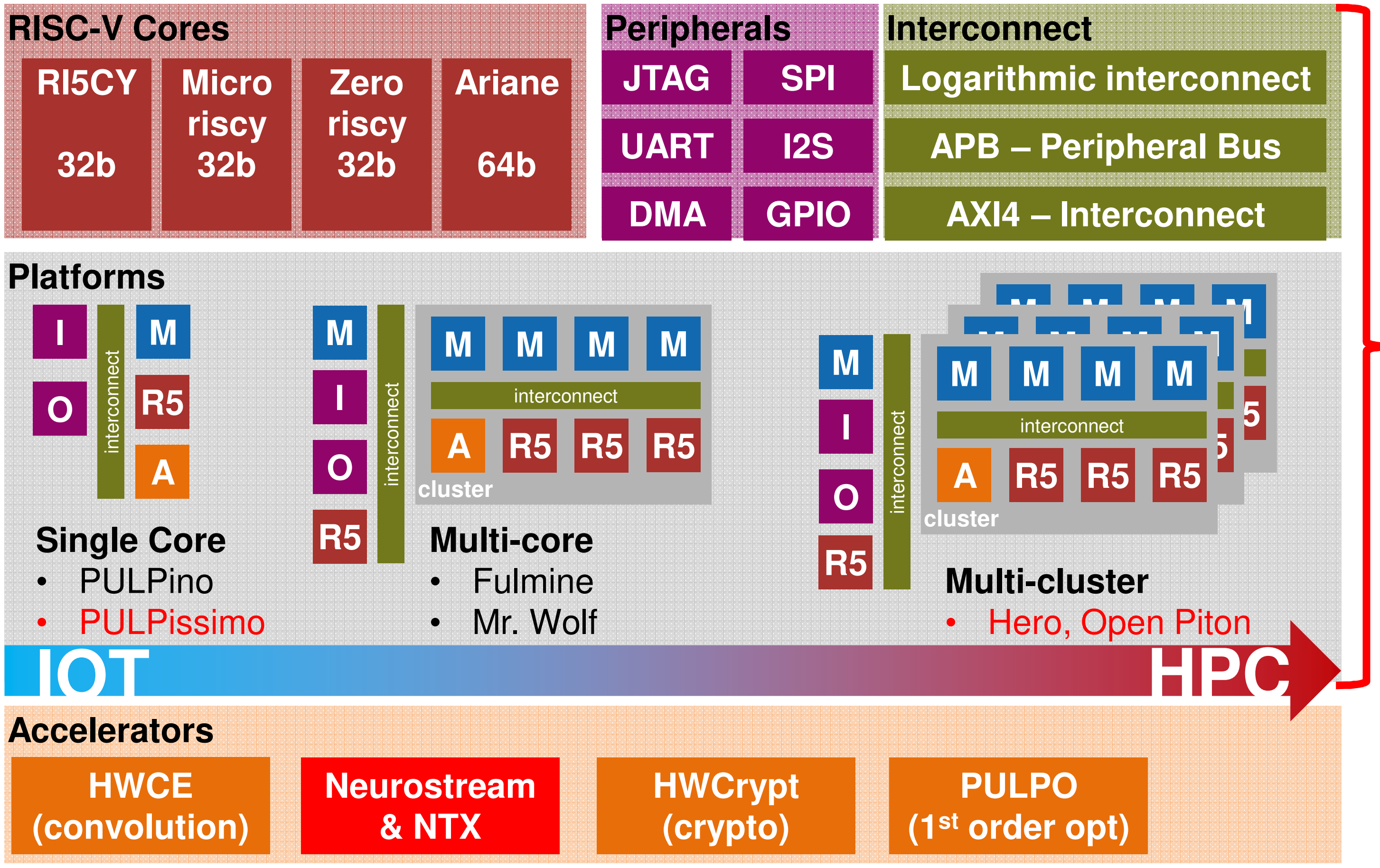
NTX→European Processor Initiative

Europe Needs its own Processors

- Processors now control almost every aspect of our lives
- **Security** (back doors etc.)
- Possible **future restrictions on exports to EU** due to **i**ncreasing protectionism
- **A competitive EU supply chain** for HPC technologies will create jobs and growth in Europe
- Sovereignty (data, economical, embargo)

- High Performance General Purpose Processor for HPC
- **High-performance RISC-V based accelerator (NTX)**
- Computing platform for autonomous cars
- Will also target the AI, Big Data and other markets in order to be economically sustainable

Putting it all together: The Open PULP platform



RISC-V Cores

- RI5CY 32b
- Micro riscy 32b
- Zero riscy 32b
- Ariane 64b

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-cluster

- Hero, Open Piton

IOT → HPC

Accelerators

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

OSHW Solderpad0.5

See F. Gürkaynak Talk later Today

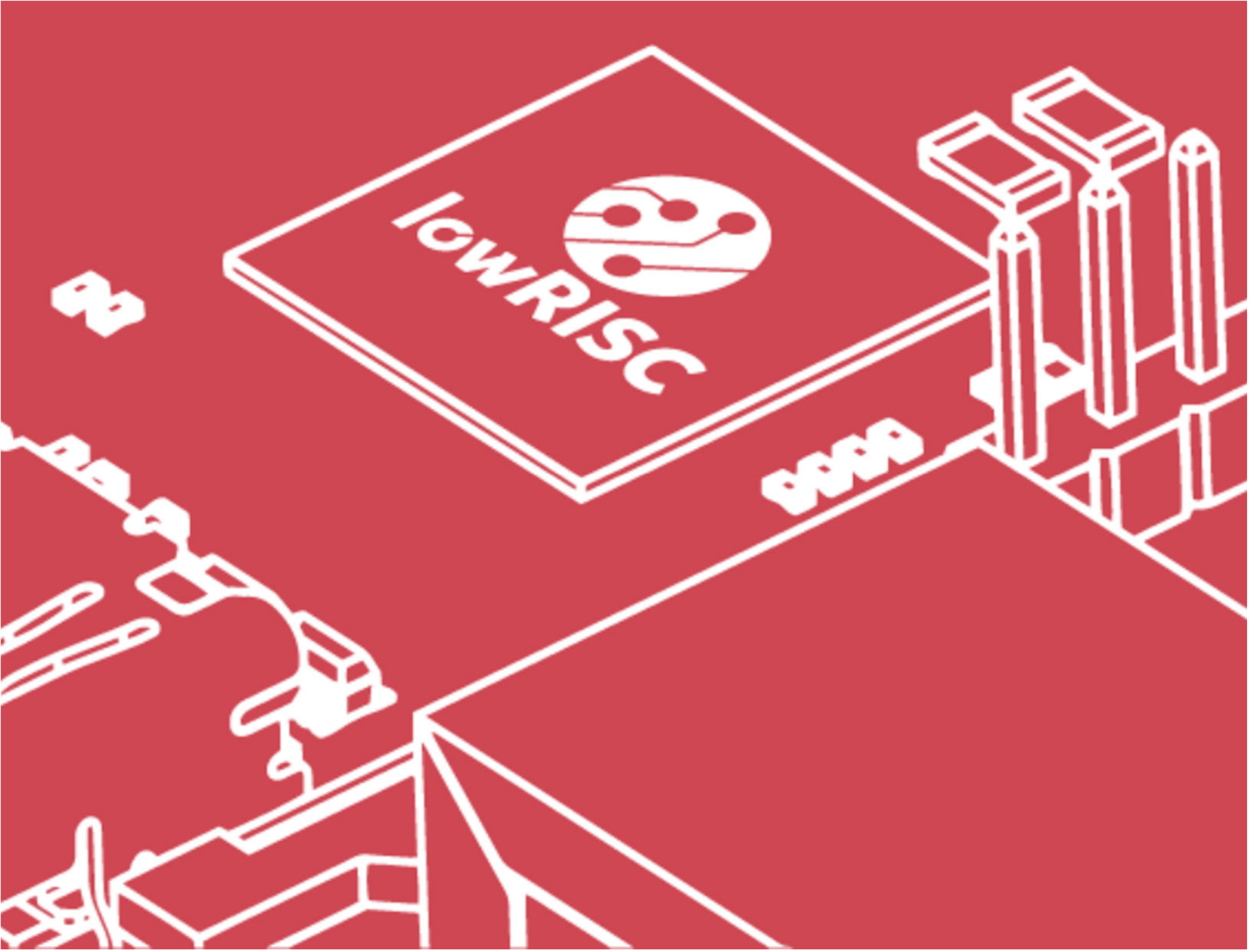
But this is way too much for a university (or two)!

A non-for profit Company: Enter lowRISC!

lowRISC Communit y Interest Company

lowRISC

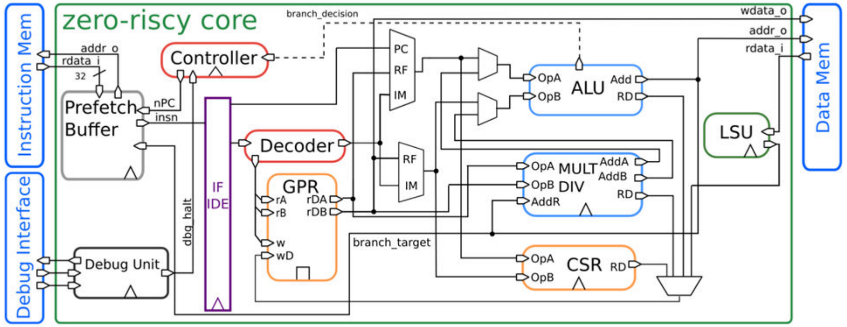
enabling open source silicon through collaborative engineering



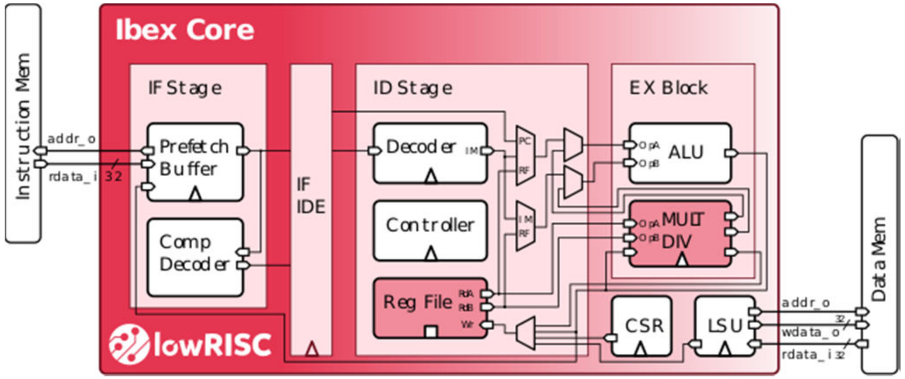
LowRISC is up and... hiring

Google UNIVERSITY OF CAMBRIDGE ETH zürich

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



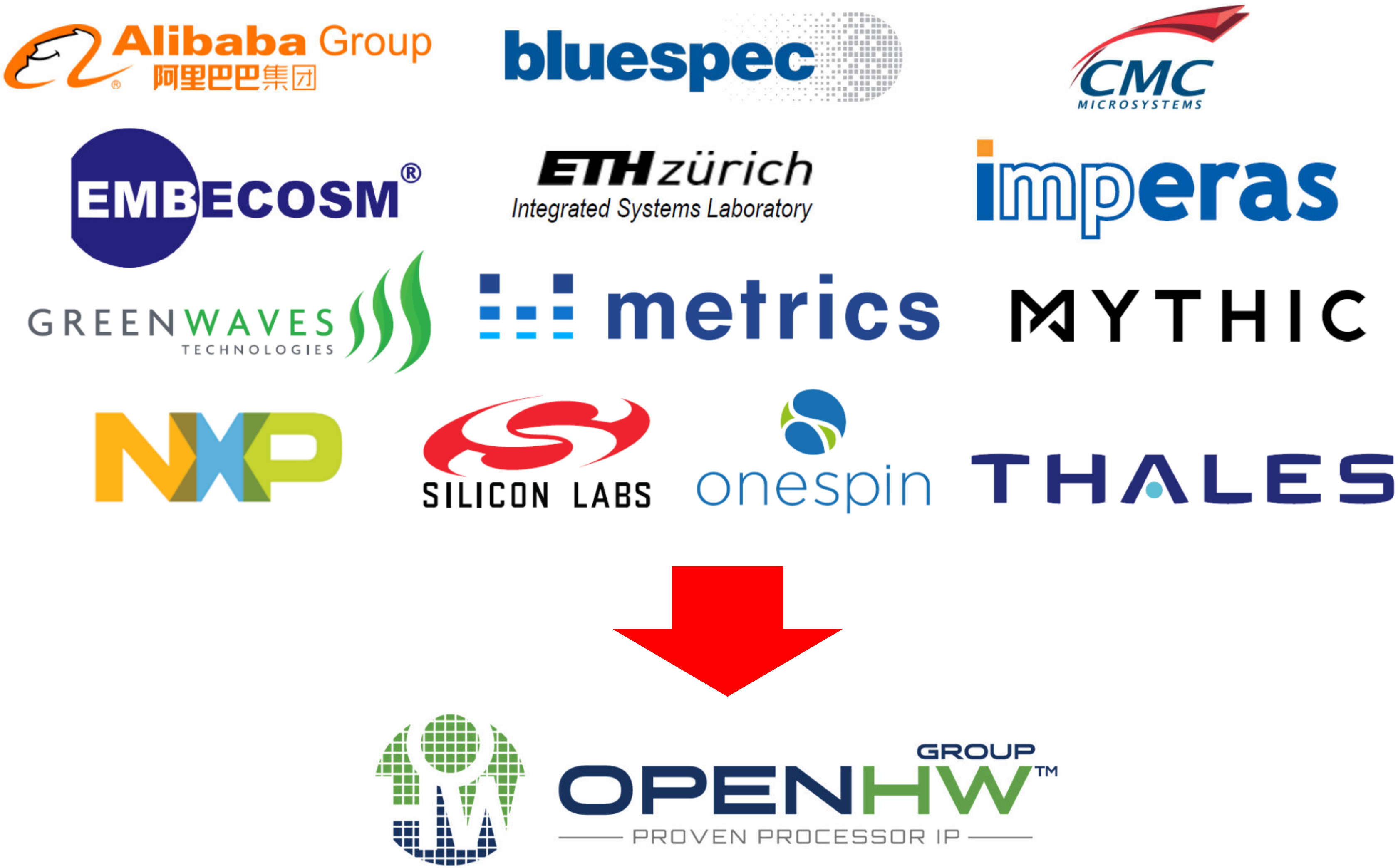
Zero-Riscy (RV32-ICM), 19kGE



IBEX: see P. Wagner talk @ WOSH

PULP

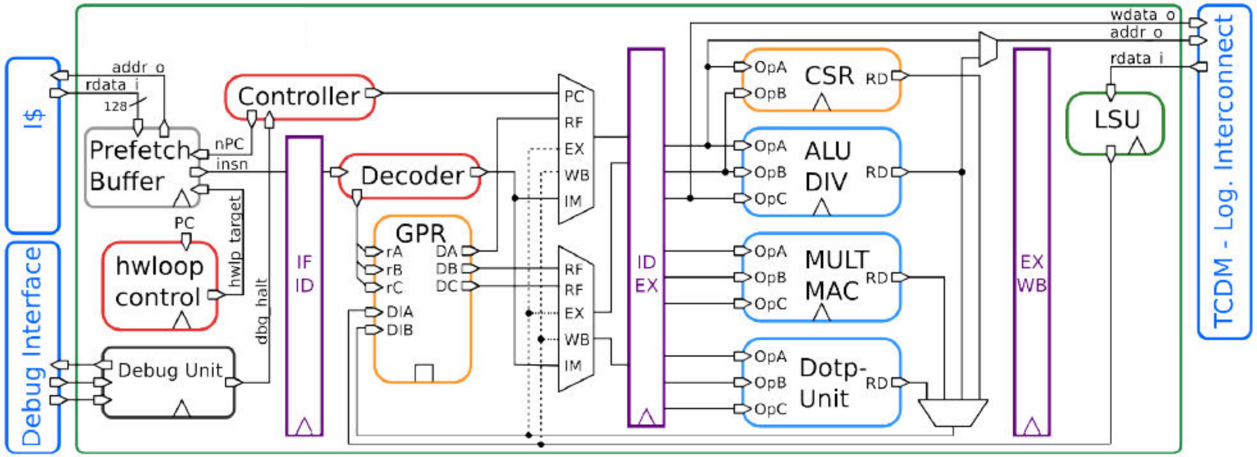
Will just one NFP Company be Enough?



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.

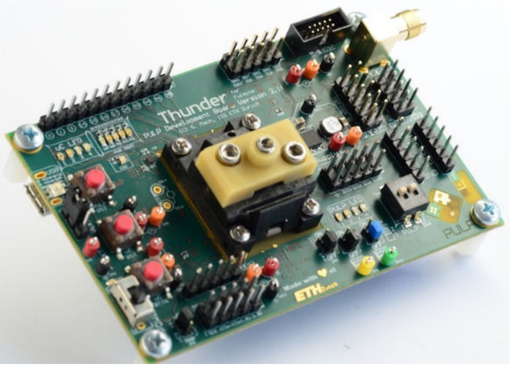
RI5CY core

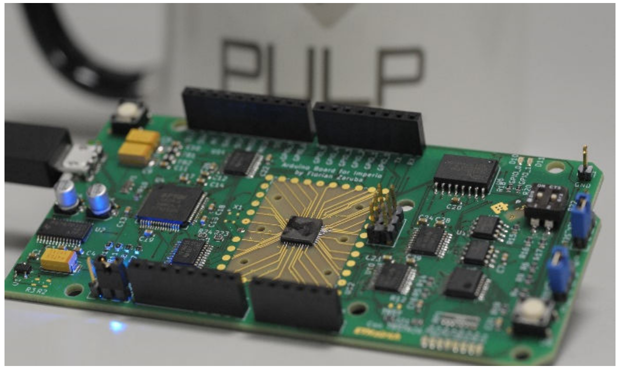


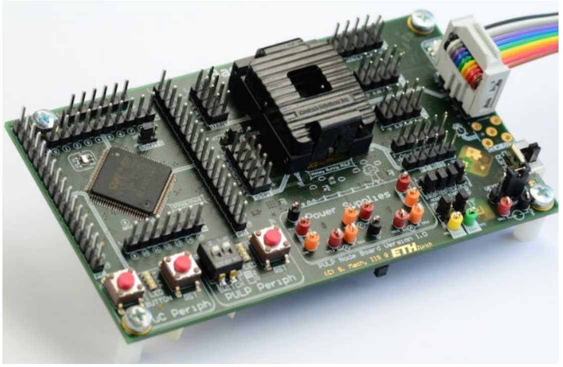
CORE-V™

see R. O'Connor (OpenHW CEO) talk

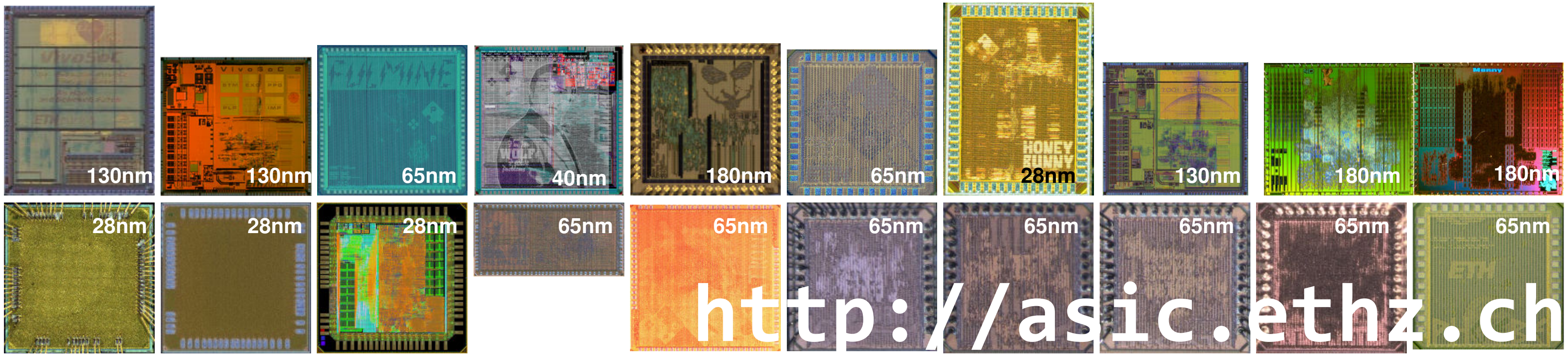
PULP







www.pulp-platform.org



http://asic.ethz.ch

The fun is just beginning...