

RISC-V open-ISA and open-HW – a Swiss army knife for HPC

ICS2020, Workshop on RISC-V and OpenPOWER

29.06.2020

Andrea Bartolini¹ & PULP team^{1,2}



*¹Department of Electrical, Electronic
and Information Engineering*

ETH zürich

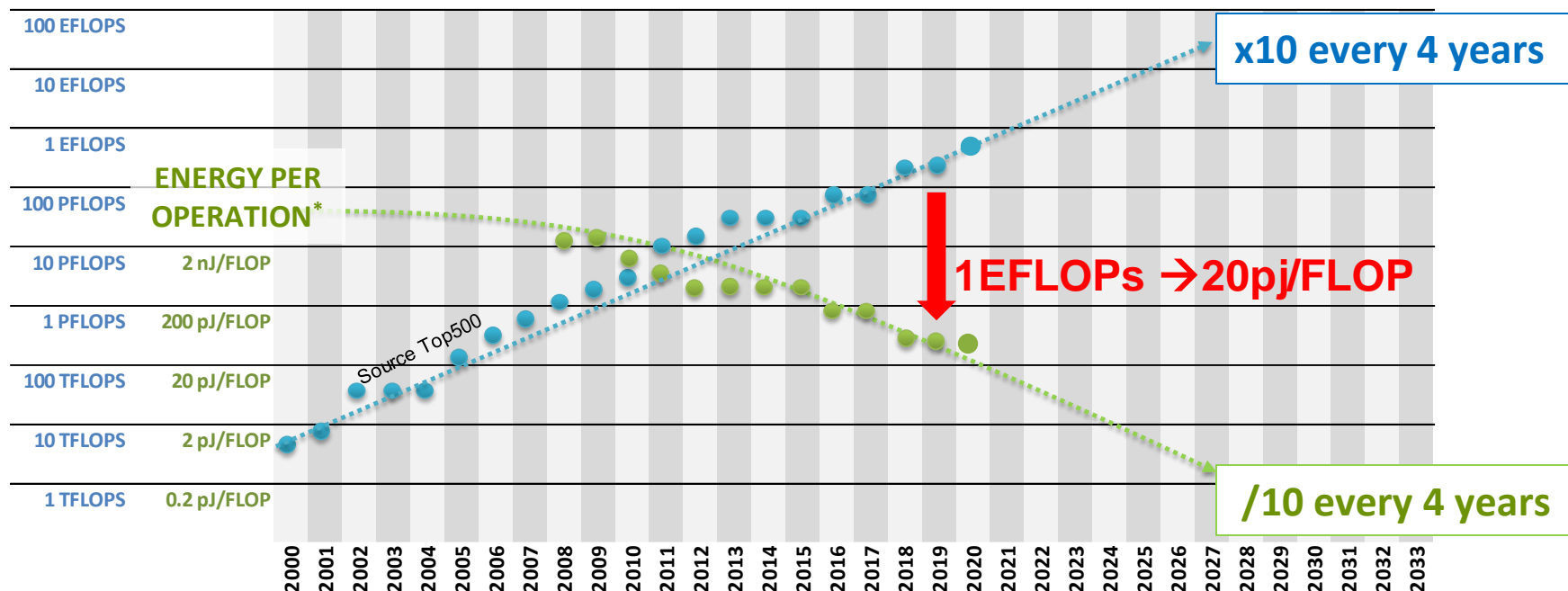
²Integrated Systems Laboratory

Energy efficiency challenge: Exascale

HPC is now power-bound → need 10x energy efficiency improvement every 4 years

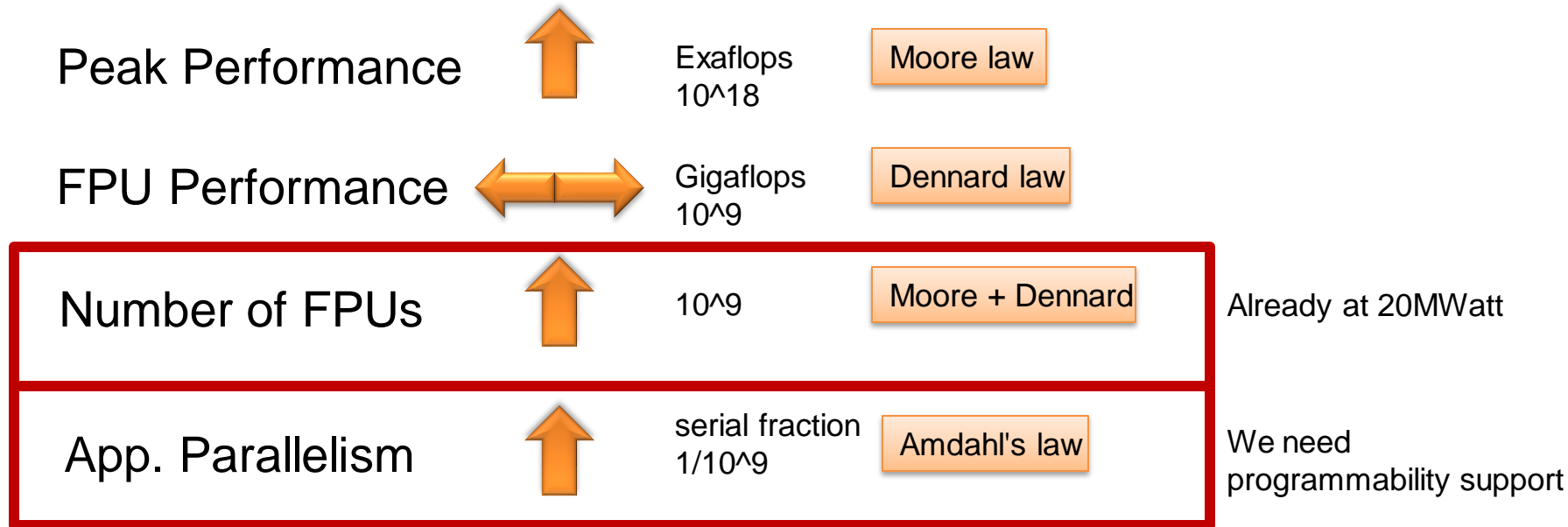
*20MWatt supercomputer: Performance & EnOP

PERFORMANCE



Copyright © European Processor Initiative 2019. EPI Tutorial/Barcelona/17-07-2019

HPC trends

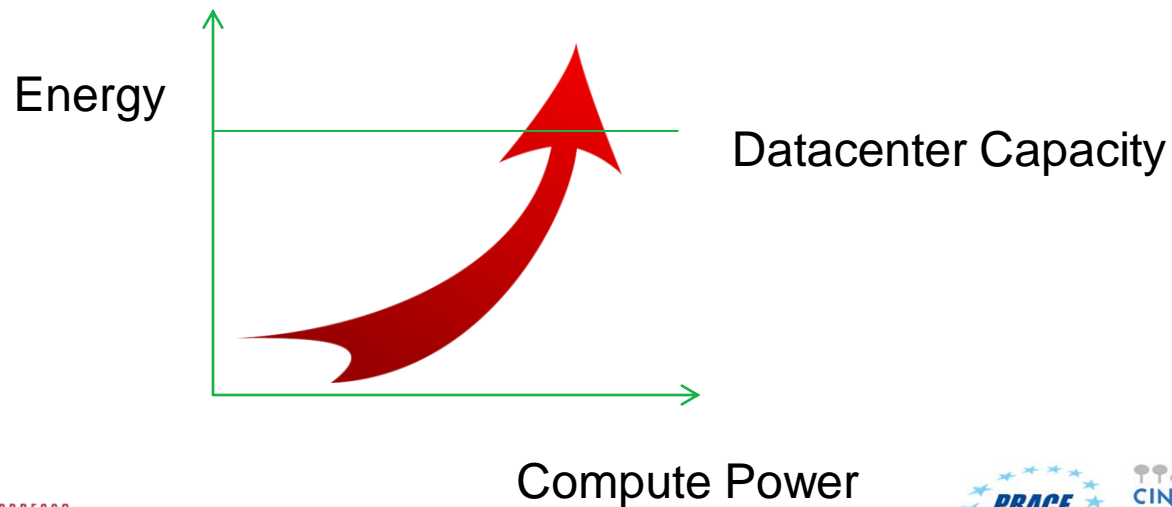


Energy trends

“traditional” CPUs chips
are designed for maximum
performance for all
possible workloads



Silicon area wasted to
maximize single thread
performance

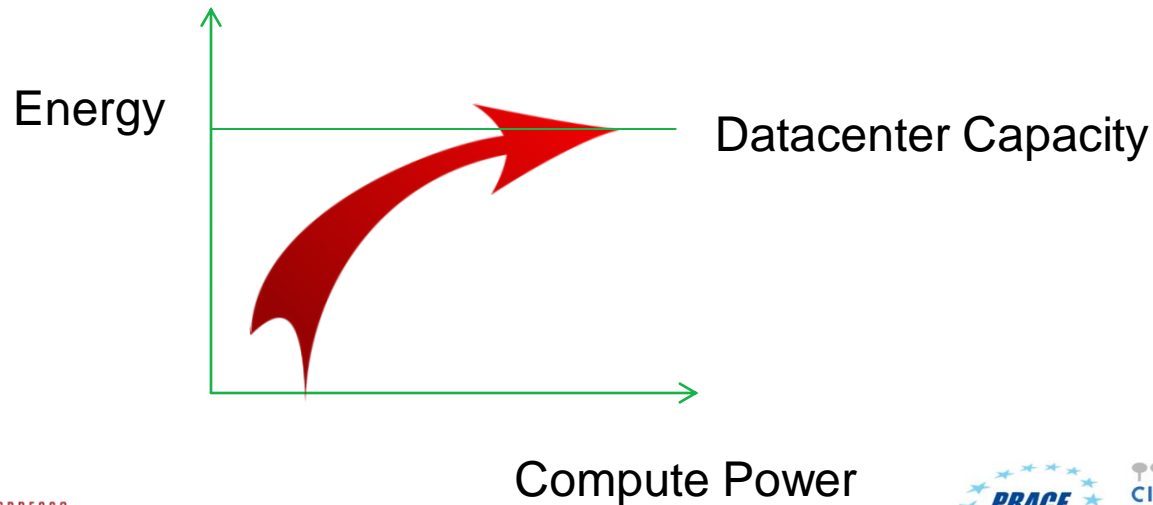


Change of paradigm #1

New chips designed for maximum performance in a reduced set of workloads



Simple functional units, poor single thread performance, but maximum throughput



Change of paradigm #2



LEADERSHIP
COMPUTING
FACILITY

ABOUT OLCF ▾

OLCF RESOURCES ▾

R&D ACTIVITIES ▾

SCIENCE AT C

RESEARCH TEAM BREAKS EXAOP BARRIER WITH DEEP LEARNING APPLICATION



BY KATIE BETHEA



SHARE



LOVE

NEURAL NETWORK TRAINED TO IDENTIFY EXTREME WEATHER PATTERNS FROM HIGH-RESOLUTION CLIMATE SIMULATIONS

This article is part of a series covering [the finalists for the 2018 Gordon Bell Prize that used the Summit supercomputer](#). The prize winner will be announced at SC18 in November in Dallas.

Contact: Kathy Kincade, kkincade@lbl.gov, +1 510 495 2124

A team of computational scientists from Lawrence Berkeley National Laboratory (Berkeley Lab) and Oak Ridge National Laboratory (ORNL) and engineers from NVIDIA has, for the first time, demonstrated an exascale-class deep learning application that has broken the exaop barrier.

Using a climate dataset from Berkeley Lab on ORNL's Summit system at the Oak



DEEP500

[Home](#)

[Meetings and Slides](#)

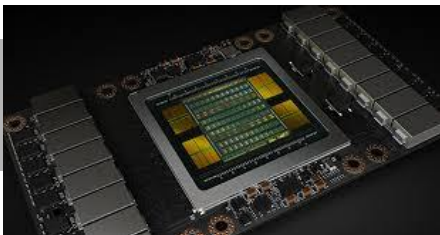
[COMING SOON](#)

[Ranking](#)



Deep500: An HPC Deep Learning Benchmark and Competition

A modular benchmarking infrastructure for high-performance deep learning — from the single operator to distributed training.



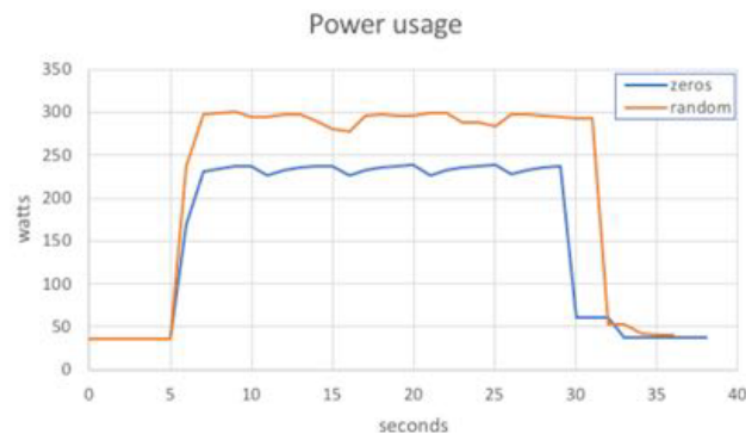
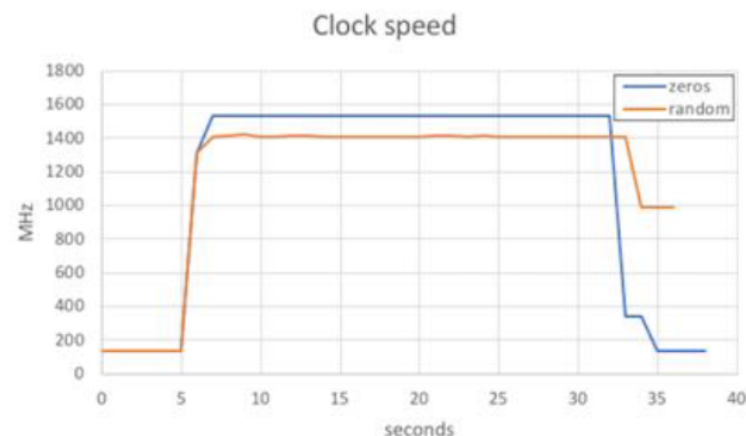
Change of paradigm #3

<https://indico-jsc.fz-juelich.de/event/76/session/0/contribution/1/material/slides/0.pdf>
Wayne Joubert - OpenPOWER ADG 2018

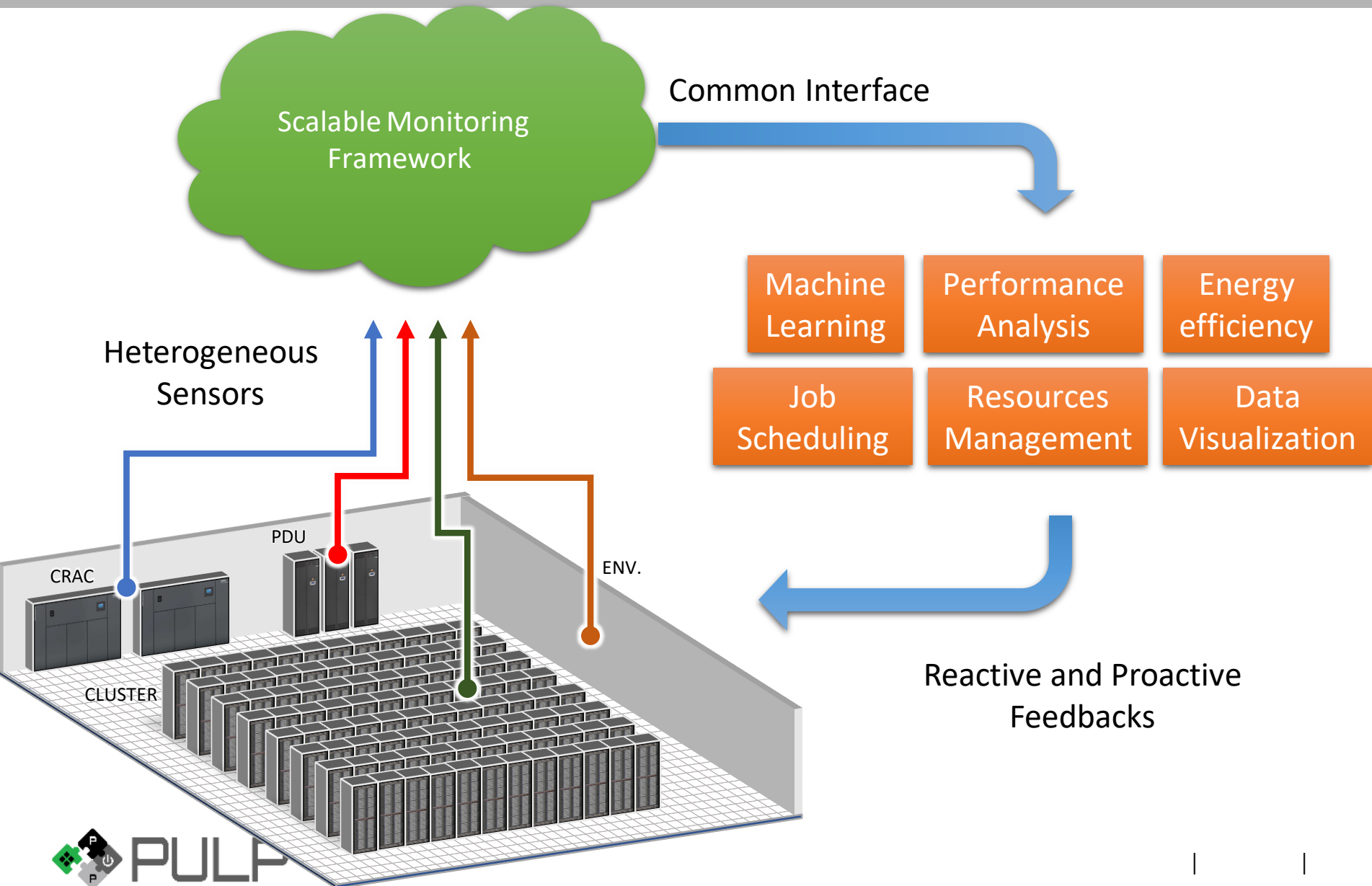
- TC/HGEMM has surprising data-dependent performance: **125 TF** theoretical peak, **113 TF** achievable on zero-filled matrices, **105 TF** peak on random CCC matrices, **~95 TF** peak on matrices with fully random FP16 entries

Issues

- Measurements on 1 Summit GPU using `nvidia-smi`
- Data-dependent performance of Tensor Cores is due to 300W power/frequency throttling of Voltas on Summit
- Baidu DeepBench GEMM benchmark has a bug (reported), incorrectly fills FP16 matrices with zeros instead of the intended random values, thus miscalculates GPU performance



Change of paradigm #4



**ARIANE:
The 64b
Application
Processor**

**HERO:
The Open
Heterogeneous
Research Platform**

**ARA
The Vector
Engine**

**SPIN on PULP:
Network-
Accelerated
Memory
Transfers**

**NTX
The Network
Training
Accelerator**

**SNITCH:
The Pseudo Dual-
Issue Processor
for FP Workload**

**ControlPULP:
The Power
Controller for
HPC server**

**EXASCALE
2021**

<https://pulp-platform.org/>

Architecture: Ariane RISC-V Cores

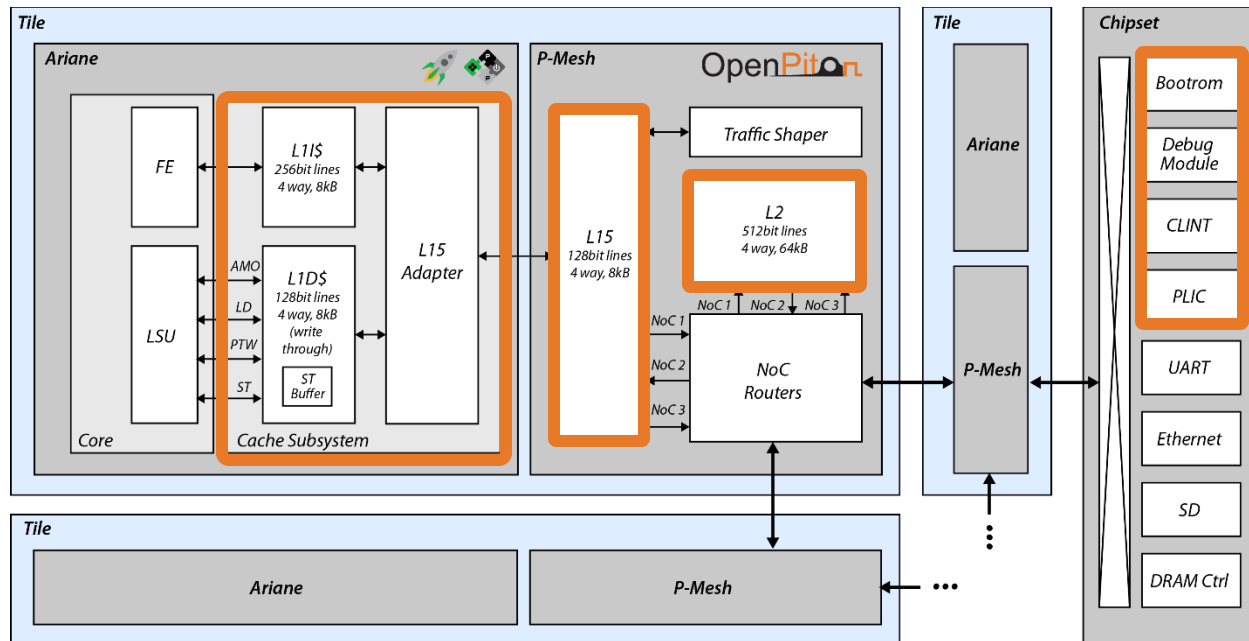
**ARIANE:
The 64b
Application
Processor**

- RV64GC, 6-stage, in-order, out-of-order execute
- 16 KiB instruction cache, 32 KiB data cache
- Transprecision floating-point unit (TP-FPU) [3]
 - double-, single- and half-precision FP formats
 - Two custom formats FP16alt and FP8
 - All standard RISC-V formats as well as SIMD
- Two different implementations:
 - Ariane High Performance (AHP): tuned for high-performance applications
 - Ariane Low Power (ALP): tuned for light, single-threaded applications

[The cost of application-class processing: Energy and Performance Analysis of a Linux-ready 1.7-GHz 64-bit RISC-V Core in 22-nm FDSOI Technology](#)

OpenPiton+Ariane

If you are really passionate about cache coherent “scalable” machines...



- Boots SMP Linux
- New write-through cache subsystem with invalidations and the TRI interface
- LR/SC in L1.5 cache
- Fetch-and-op in L2 cache
- RISC-V Debug
- RISC-V Peripherals

[OpenPiton+ Ariane: The First Open-Source, SMP Linux-booting RISC-V System Scaling From One to Many Cores](#)

Architecture: Network Training Accelerator (NTX)

- “Network Training Accelerator”
 - 32 bit float streaming co-processor (IEEE 754 compatible)
 - Custom 300 bit “wide-inside” Fused Multiply-Accumulate
 - 1.7x lower RMSE than conventional FPU
 - 1 RISC-V core (“RI5CY”) and DMA
 - 8 NTX co-processors
 - 64 kB L1 scratchpad memory (comparable to 48 kB in V100)



Key ideas to increase hardware efficiency:

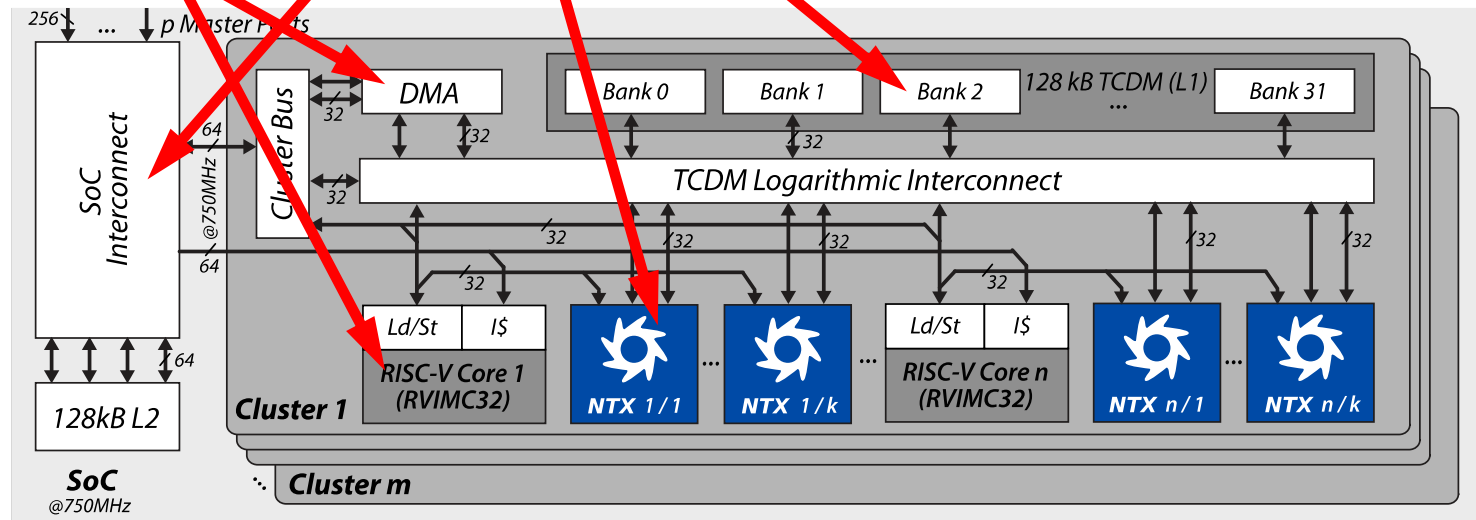
- **Reduction of von Neumann bottleneck (load/store elision through streaming)**
- **Latency hiding through DMA-based double-buffering**

Schuike, Fabian, Michael Schaffner, Frank K. Gürkaynak, and Luca Benini. "A scalable near-memory architecture for training deep neural networks on large in-memory datasets." IEEE Transactions on Computers 68, no. 4 (2018): 484-497.

Schuike, Fabian, Michael Schaffner, and Luca Benini. "Ntx: An energy-efficient streaming accelerator for floating-point generalized reduction workloads in 22 nm fd-soi." In 2019 Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 662-667. IEEE, 2019.

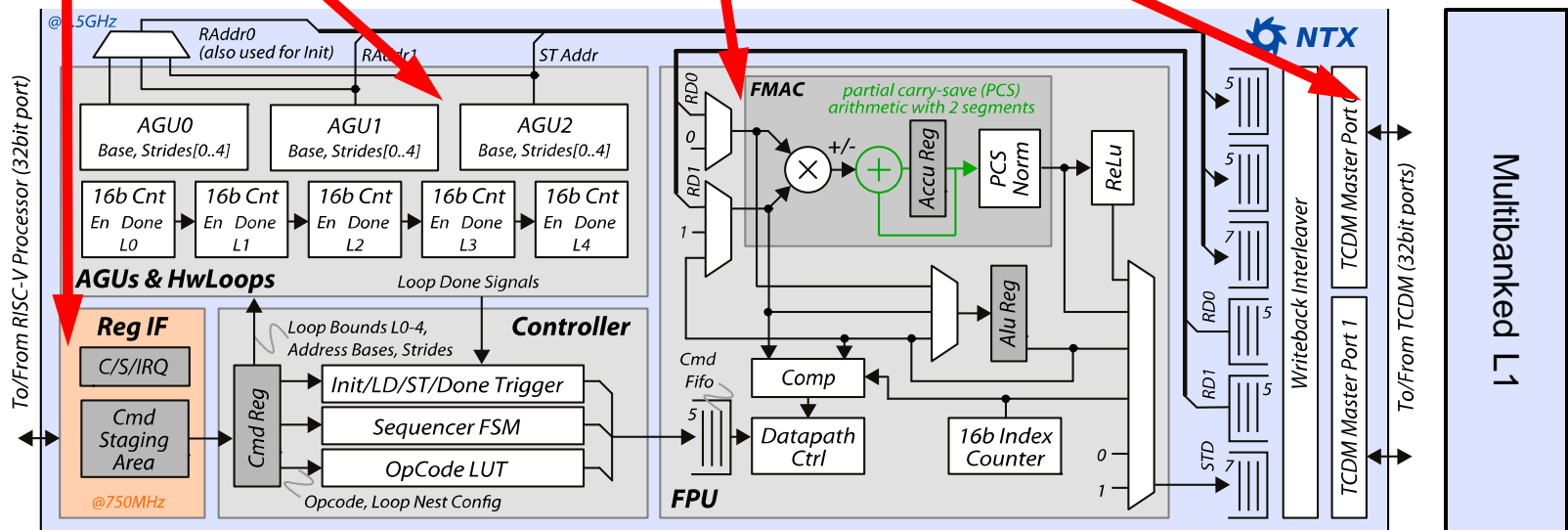
Flexible Architecture NTX accelerated cluster

- **1 processor** core controls **8 NTX** coprocessors
- Attached to 128 kB shared **TCDM** via a logarithmic interconnect
- **DMA** engine used to transfer data (double buffering)
- Multiple clusters connected via interconnect (crossbar/NoC)



Network Training Accelerator (NTX)

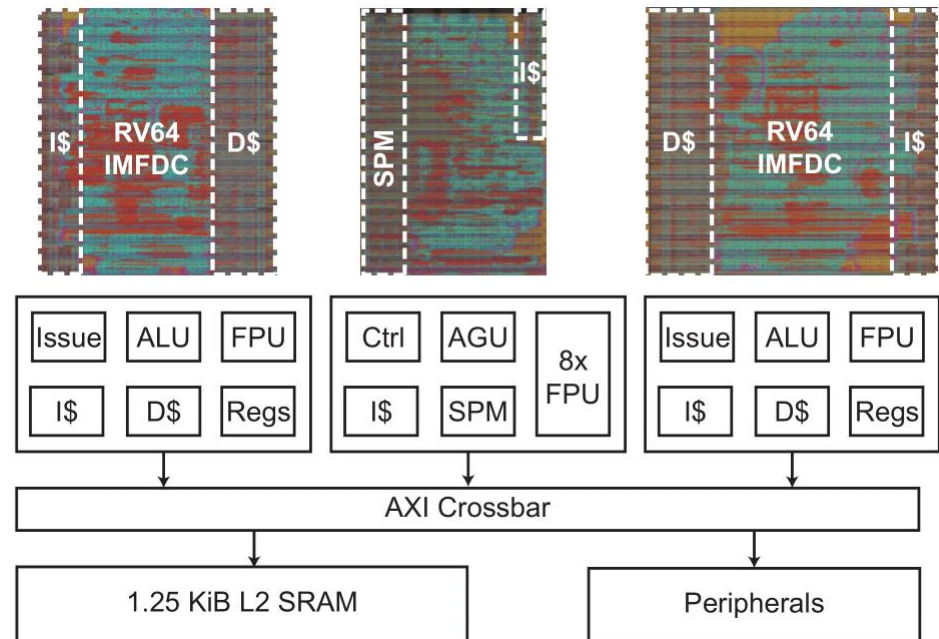
- Processor configures Reg IF and manages DMA double-buffering in L1 memory
- Controller issues AGU, HWL, and FPU micro-commands based on configuration
- AGUs generate address streams for data access
- FMAC with extended precision + ML functions
- Reads/writes data via 2 memory ports (2 operand and 1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



ARIANE: The 64b Application Processor

NTX The Network Training Accelerator

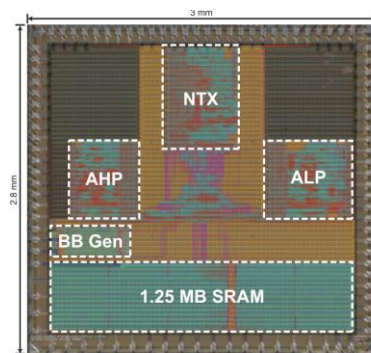
- 22nm FDX technology
- Two application-class RISC-V Ariane cores [1] - DP
 - RV64GCXsmallfloat
 - General purpose workloads
- Network Training Accelerator (NTX) [2] - FP
 - Accelerates oblivious kernels:
 - Deep neural network training
 - Stencils
 - General linear algebra workloads
- 1.25 MiB of shared L2 memory
- Peripherals



Schuiki, Fabian, Michael Schaffner, and Luca Benini. "NTX: A 260 Gflop/sW Streaming Accelerator for Oblivious Floating-Point Algorithms in 22 nm FD-SOI." In 2019 International SoC Design Conference (ISOCC), pp. 117-118. IEEE, 2019.

Summary on Kosmodrom: State of the Art

- We achieve **higher energy-efficiency** for AHP and ALP than competitive RISC-V processors (Rocket)
- Ariane contains slightly larger caches (32 KiB compared to 16 KiB)
- The ALP implementation is penalized because of less mature cell libraries available to us (7k cells vs 2k cells)
- NTX achieves a **2x** gain in energy-efficiency compared to Tesla V100



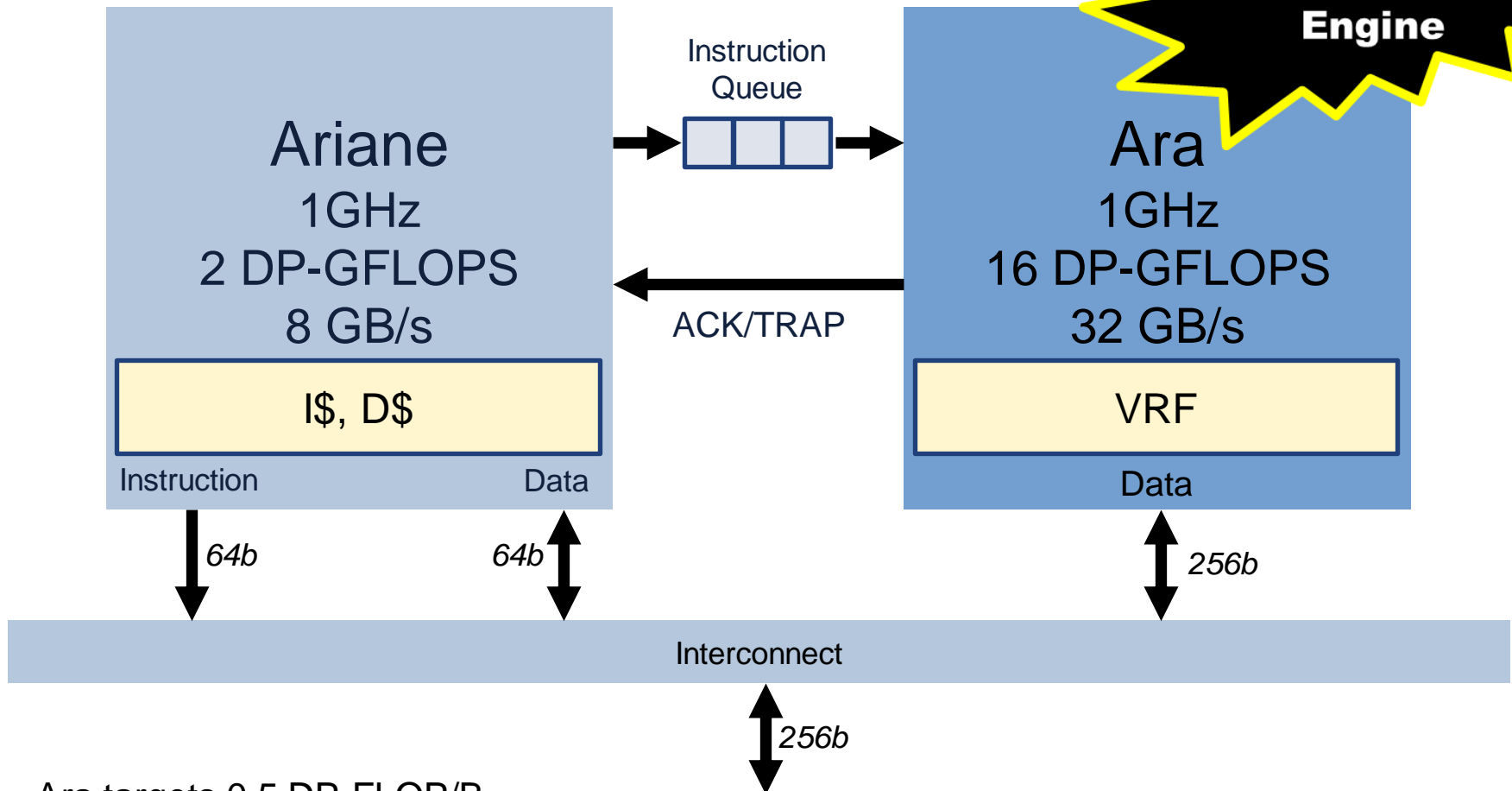
		AHP	ALP	NTX
Nominal VDD	[V]	0.8	0.5	0.8
Frequency	[GHz]	0.85	0.175	1.55
Area	[mm ²]	0.4	0.5	0.5
Area*	[MGE]	1.95	3.96 [†]	2.8
Total Power	[mW]	50.2	6.2	160.0
Leakage Power	[%]	4.6	2.8	5.5
Lkg. Power / Area (0.5 V	[mW/mm ²]	1.25	0.36	3
Energy/Instr.	[pJ]	24.4	22.7	3.8
Max. Eff.	[Gflop/s W]	41	44	266 6x
Max. Perf.	[Gflop/s]	1.5	1.3	24 18x
Area Eff.	[Gflop/s mm ²]	3.75	2.6	48
SLVT/LVT	[%]	26/74	83/17	60/40
FP Formats		8/16/ 16alt/ 32/64	8/16/ 16alt/ 32/64	32

	AHP	ALP	NTX	Cortex	Rocket	Tesla	Xeon
	[us]	[us]	[us]	A53[14]	64b[15]	V100 [§]	8180 [§]
Node/V _{DD}	22/0.45	22/0.45	22/0.45	16/0.8	40/0.65	12/1.0	14/0.9
32 bit floats							
Energy Eff. [†]	93	98	266	38.7 *	16.7 ¶	122	21.9
Area Eff. [‡]	7.5	5.2	47.1	8.7 *	7.3 ¶	20.5	3.57
64 bit floats							
Energy Eff. [†]	41	44	—	19.4 *	16.7 ¶	61	11.0
Area Eff. [‡]	3.75	2.6	—	4.4 *	7.3 ¶	10.3	1.79

[†] Gflop/s W; [‡] Gflop/s mm² (node-scaled); [§] our estimates;
 * assuming NEON; ¶ no SIMD || extrapolated from 64 bit

Enter ARA: Open-Source RISC-V Vector Engine

ARA
The Vector
Engine



- Ara targets 0.5 DP-FLOP/B
 - Memory bandwidth scales with the number of physical lanes

Cavalcante, Matheus, Fabian Schuiki, Florian Zaruba, Michael Schaffner, and Luca Benini. "Ara: A 1-GHz+ Scalable and Energy-Efficient RISC-V Vector Processor With Multiprecision Floating-Point Support in 22-nm FD-SOI." IEEE Transactions on Very Large Scale Integration (VLSI) Systems 28, no. 2 (2019): 530-543.

Matrix multiplication on Ara

- Load row i of matrix B into vB
- for (int $j = 0; j < n; j++$)
 - Load element $A[j, i]$
 - Broadcast it into vA
 - $vC \leftarrow vA \cdot vB + vC$

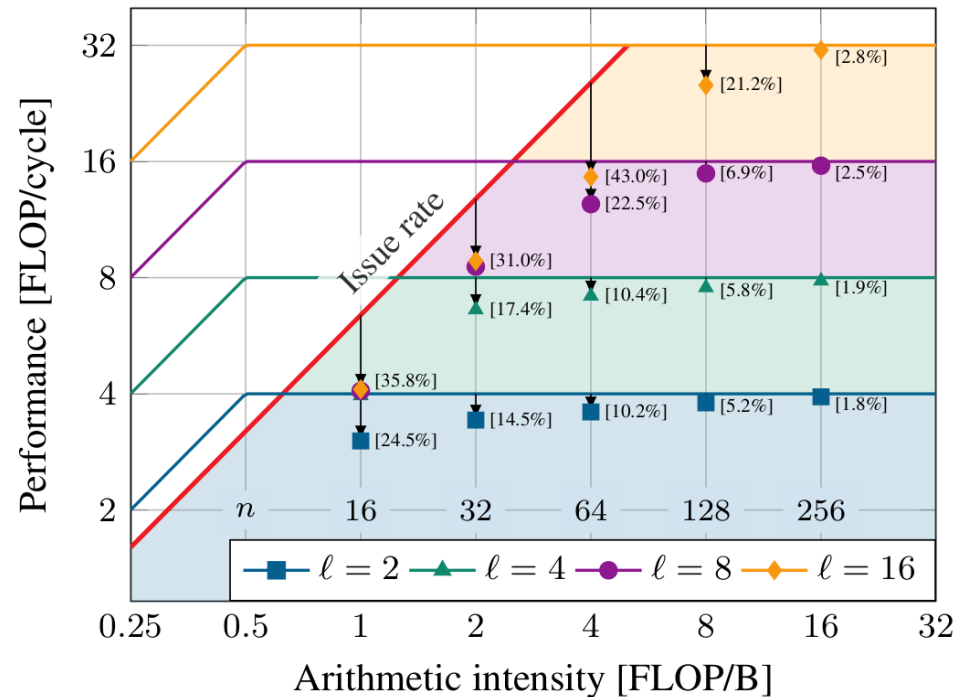
vld vB, 0(addrB)

(Unrolled loop)

- ld t0, 0(addrA)
- addi addrA, addrA, 8
- vins vA, t0, zero
- **vmadd vC, vA, vB, vC**
- ld t0, 0(addrA)
- addi addrA, addrA, 8
- vins vA, t0, zero
- **vmadd vC, vA, vB, vC**

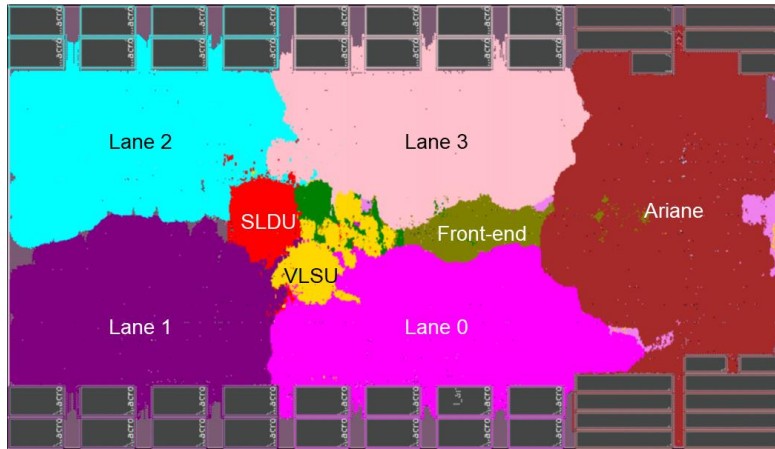
Issue rate performance limitation

- vmadds are issued at best every four cycles
 - Since Ariane is single-issue
- If the vector MACs take less than four cycles to execute, the FPUs starve waiting for instructions
 - Von Neumann Bottleneck
- This translates to a boundary in the roofline plot

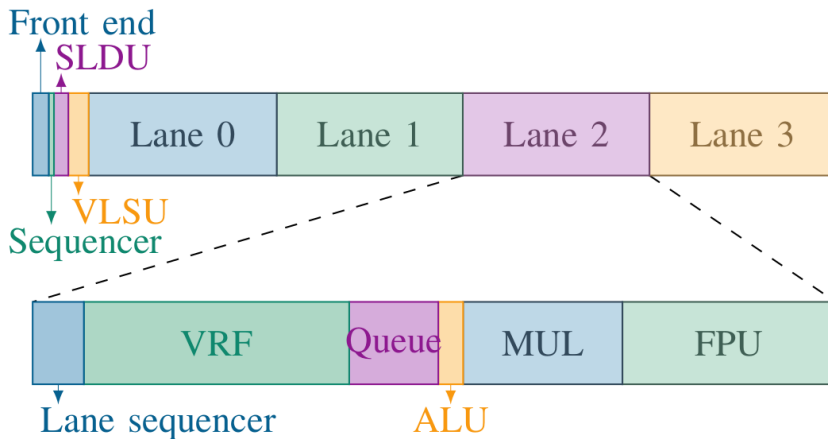


Ara: Figures of Merit

(TT, 0.80V, 25°C)



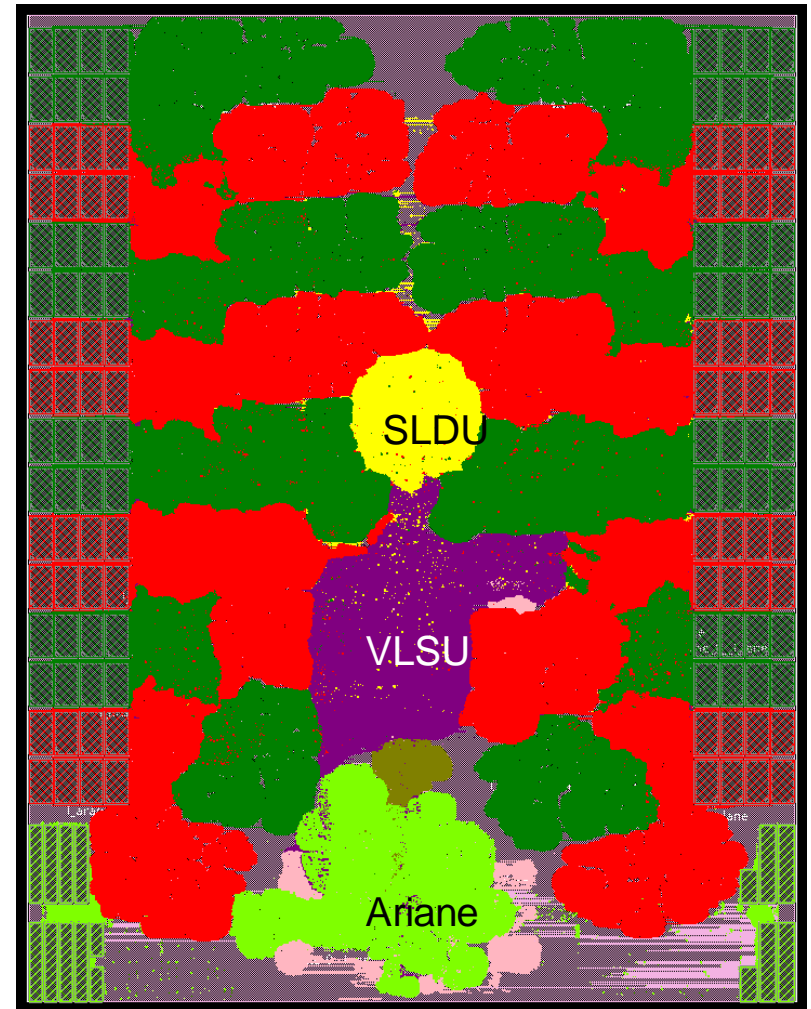
- Area breakdown



- Ara: 4 lanes GF 22FDX 1.25 GHz implementation
- Clock frequency
 - 1.25 GHz (nominal), 0.92 GHz (worst condition)
 - 40 gate delays
- Area: 3400 kGE
 - 0.68 mm²
- 256 x 256 MATMUL
 - Performance: 9.8 DP-GFLOPS
 - Power: 259 mW
 - Efficiency: **38 DP-GFLOPS/W**
 - ~2.5X better than Ariane on same benchmark**

Ara: Scalability

- Each lane is *almost* independent
 - Contains part of the VRF and its functional units
- Scalability limitations
 - VLSU and SLDU: need to communicate to all banks
- Instance with 16 lanes:
 - 1.04 GHz (nom.), 0.78 GHz (w)
 - 10.7 MGE (2.13mm²)
 - 32.4 DP-GFLOPS
 - **40.8 DP-GFLOPS/W (peak)**

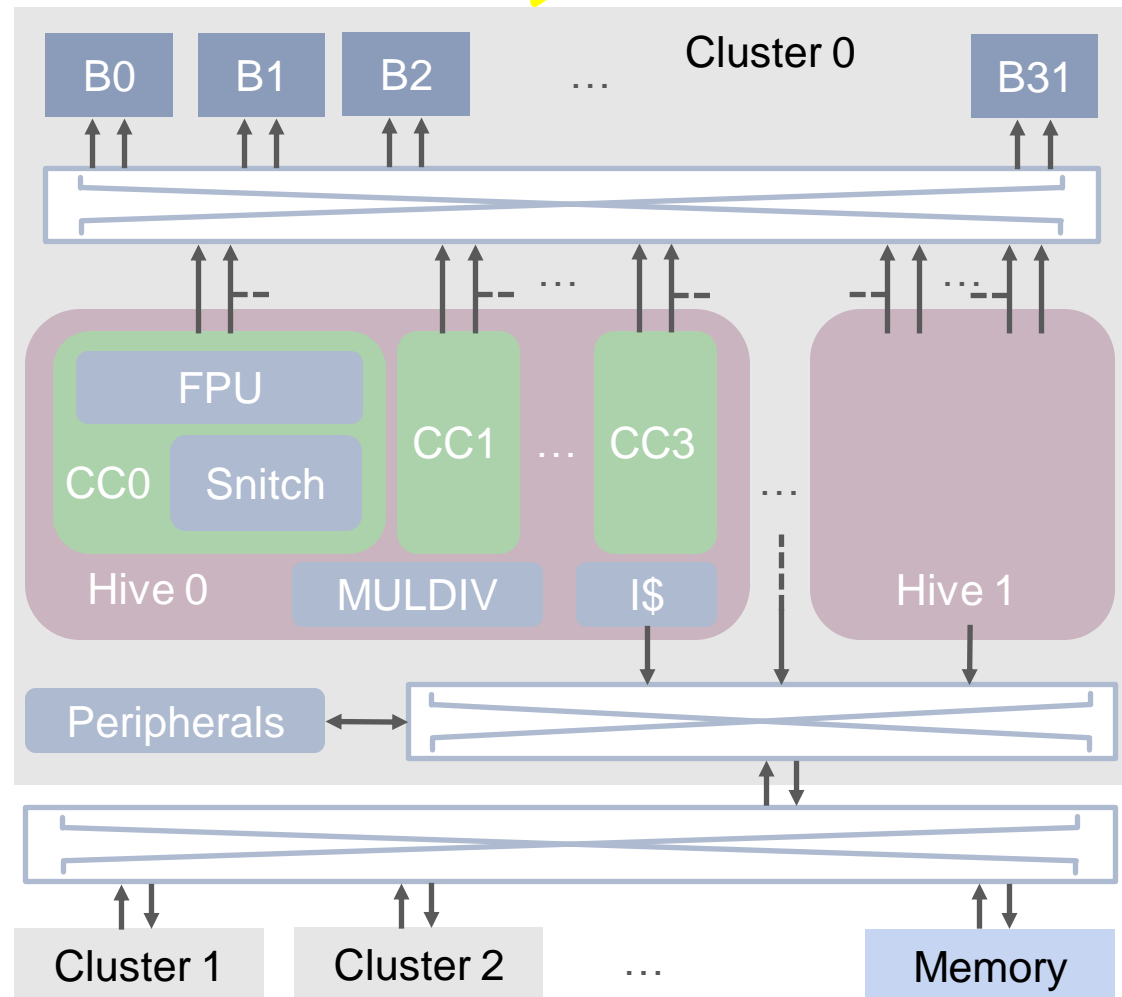


16 ARAs give you 1TFLOP at 12W - NOT BAD!

SNITCH

- Built around *Snitch* core:
 - RV32I, 15 kGE
 - Add 64b FPU subsystem:
 - core complex* (CC)
 - 4 CCs, MULDIV, I-cache:
 - hive*
 - 2 hives, TCDM, peripherals: *cluster*
 - N clusters, system X-bar, memory: *system*
- Float subsystem adds novel HW
 - 2 stream semantic registers
 - FPU sequencer

SNITCH:
The Pseudo Dual-Issue Processor
for FP Workload



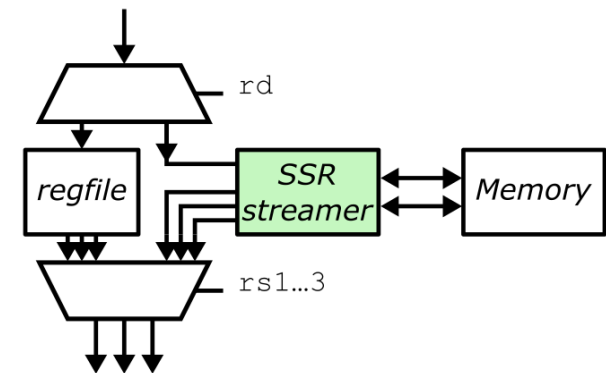
Stream Semantic Registers and FREP

- Vanilla RISCs: low functional unit utilization
 - Solutions often complex: CISC, VLIW, vectoring
- Map registers to *memory streams*: **SSRs**
 - Reads, writes become *memory requests*
 - Programmable generator emits addresses
 - + Unmodified ISA
 - + Orthogonal to hardware loops
- Snitch CC: FPU instruction stream *decoupled*
 - **FPU Sequencer** can buffer, *loop over* instructions
 - Core, FPU fed in parallel → *pseudo-dual issue*

Zaruba, Florian, Fabian Schuiki, Torsten Hoefer, and Luca Benini. "Snitch: A 10 kGE Pseudo Dual-Issue Processor for Area and Energy Efficient Execution of Floating-Point Intensive Workloads." arXiv preprint cs.AR/2002.10143 (2020).

Schuiki, Fabian, Florian Zaruba, Torsten Hoefer, and Luca Benini. "Stream Semantic Registers: A Lightweight RISC-V ISA Extension Achieving Full Compute Utilization in Single-Issue Cores." IEEE Transactions on Computers (2020).

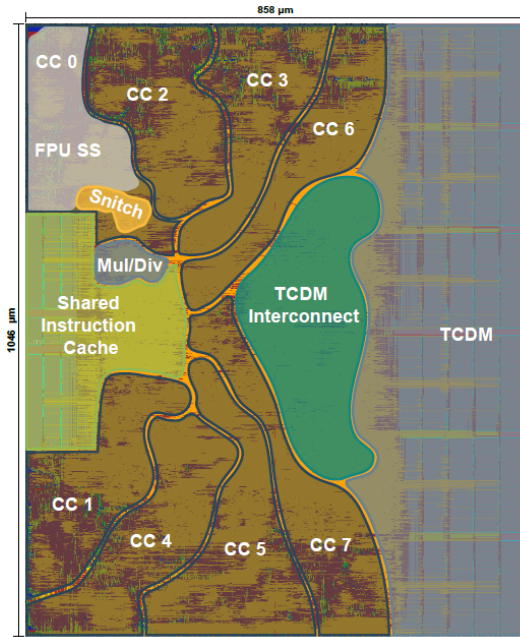
```
dotp: fld    ft0, 0(a0)
      fld    ft1, 0(a1)
      fmadd.d ft2, ft0, ft1, ft2
      addi   a0, a0, 8
      addi   a1, a1, 8
      bge    a0, t0, dotp
```



```
call  conf_addr_gen
frep  buflen, rep
fmadd.d ft2, ft0, ft1, ft2
```

SNITCH Figures of Merit

22nm FDX technology



Normalized Performance

Π	4 FPUs			8 FPUs			16 FPUs			
	<i>n</i>	Snitch	Ara	Hwacha [*]	Snitch	Ara	Hwacha	Snitch	Ara	Hwacha
16		<u>68.2</u>	<u>49.5</u>	—	63.2	25.4	—	<u>58.3</u>	<u>12.8</u>	—
32		87.1	82.6	49.9	84.8	53.4	35.6	81.4	27.6	22.4
64		93.4	89.6	—	91.7	77.5	—	89.0	45.6	—
128		<u>96.0</u>	<u>94.3</u>	—	94.7	93.1	—	<u>94.1</u>	<u>78.8</u>	—

Almost 80 DP GFlop/sW

⇒ 2x more efficient of Ara

⇒ 13pJ x DP FLOP

⇒ Exascale!!

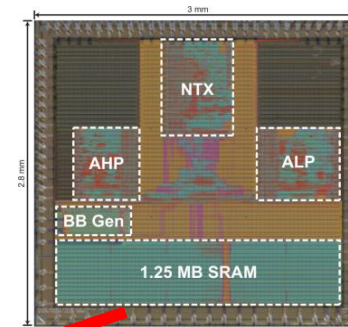
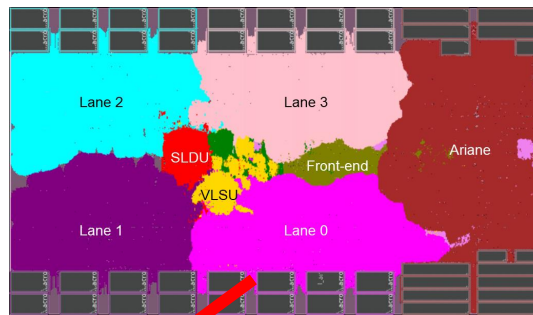
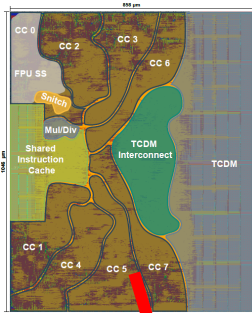
	Unit	Snitch Us	Ara [15]
Problem Size	<i>n</i>	32	32
Base ISA		RV	RV
Technode	[nm]	22	22
Clock (typical)	[GHz]	1.06	1.17
Clock (worst)	[GHz]	0.75	0.87
Peak SP	[Gflop/s]	16.96	18.72
Peak DP	[Gflop/s]	16.96	18.72
Sustained SP	[Gflop/s]	14.38	10.00
Sustained DP	[Gflop/s]	<u>14.38</u>	<u>10.00</u>
Utilization SP	[%]	84.80	—
Utilization DP	[%]	<u>84.80</u>	<u>53.40</u>
Impl. Area [#]	[mm ²]	0.89	1.07
Area Eff. SP	[Gflop/s mm ²]	25.83	—
Area Eff. DP	[Gflop/s mm ²]	<u>25.83</u>	<u>17.53</u>
Tot. Power SP	[W]	0.13	—
Tot. Power DP	[W]	0.17	0.46
Leakage	[mW]	12	21.1
Energy Eff. SP	[Gflop/s W]	103.84	—
Energy Eff. DP	[Gflop/s W]	<u>79.42</u>	<u>39.9</u>

System-Level Integration of Accelerators

HERO:
The Open
Heterogeneous
Research Platform

- Our accelerators show **leading performance and energy efficiency** in silicon at the **core and cluster level**

(TT, 0.80V, 25°C)



- How can we **unleash this potential** in **real computing systems**?

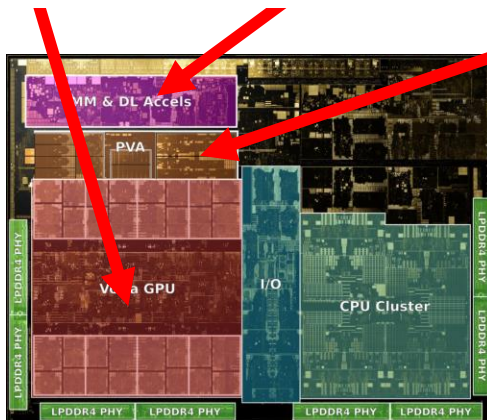


Image: Nvidia Xavier die shot annotated by WikiChip



Image: Summit Supercomputer by OLCF at ORNL

HERO: Open-Source Heterogeneous Research Platform

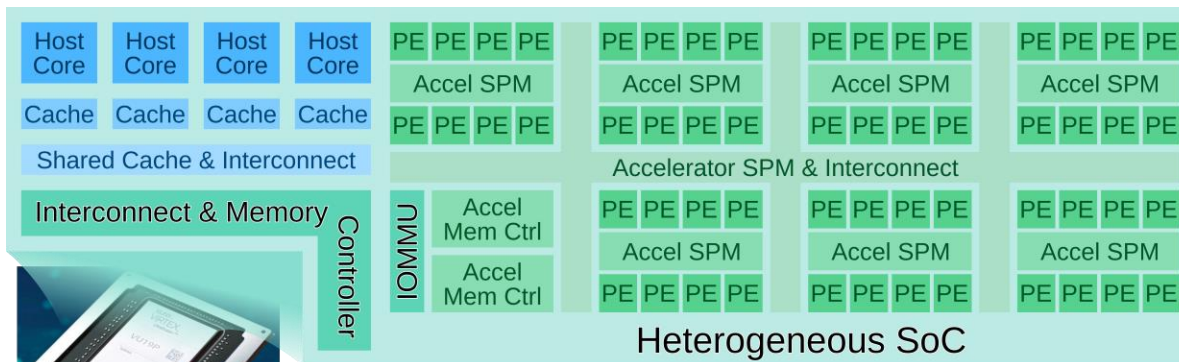
HERO combines

- **general-purpose Host CPUs**
 - **domain-specific** programmable many-core **accelerators**
- to unite **versatility** with **performance**,

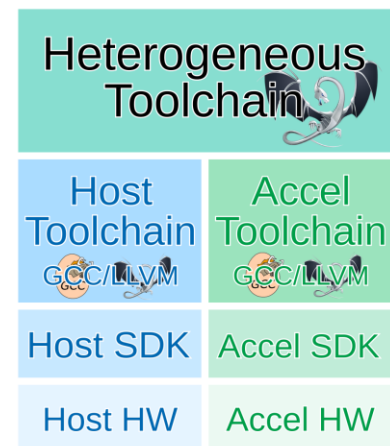
enabling **task offloading** and **data sharing** across **heterogeneous**

- **ISAs** (e.g., **ARMv8** and **RV32**)
- **memory subsystems** (e.g., **caches** and **SPMs**, **virtual** and **physical** addresses)
- **data models** (e.g., **LP64** and **ILP32**)
- **OSes and runtime libraries** (e.g., **Linux** and **OpenMP Device RTL**)

with **minimal run-time overhead** and **transparent to application programmers**.

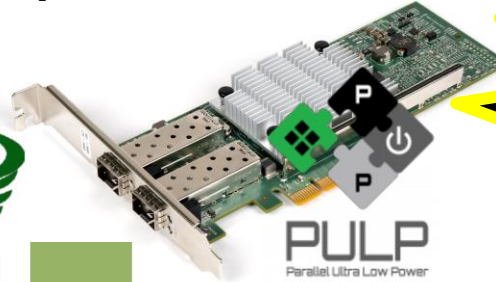


A. Kurth, P. Vogel, A. Marongiu, A. Capotondi, and L. Benini: "HERO: Heterogeneous Embedded Research Platform for Exploring RISC-V Manycore Accelerators on FPGA." Proceedings of the First Workshop on Computer Architecture Research with RISC-V (CARRV), pp. 1-7, IEEE/ACM, 2017.
A. Kurth, A. Capotondi, P. Vogel, L. Benini, and A. Marongiu: "HERO: an Open-Source Research Platform for HW/SW Exploration of Heterogeneous Manycore Systems." Proceedings of the Second Workshop on Autotuning and Adaptivity Approaches for Energy-Efficient HPC Systems (ANDARE), pp. 13-18, ACM, 2018.



Network-Accelerated Memory Transfers: sPIN on PULP

Processing **user-defined network packet kernels** on a **PULP-based accelerator** in the NIC



**sPIN on PULP:
Network-Accelerated
Memory Transfers**

Command Queue

Inbound Engine

L2 SPM (8MiB, 2 banks)

Packets

Handlers
Code

Handlers Memory

PCIe



Interconnect (crossbar, log)

DMA

TCDM (1MiB, 16 banks)

Interconnect (crossbar)

RISC-V

RISC-V

RISC-V

RISC-V

RISC-V

RISC-V

RISC-V

RISC-V

⋮

DMA

TCDM (1MiB, 16 banks)

Interconnect (crossbar)

RISC-V

RISC-V

RISC-V

RISC-V

RISC-V

RISC-V

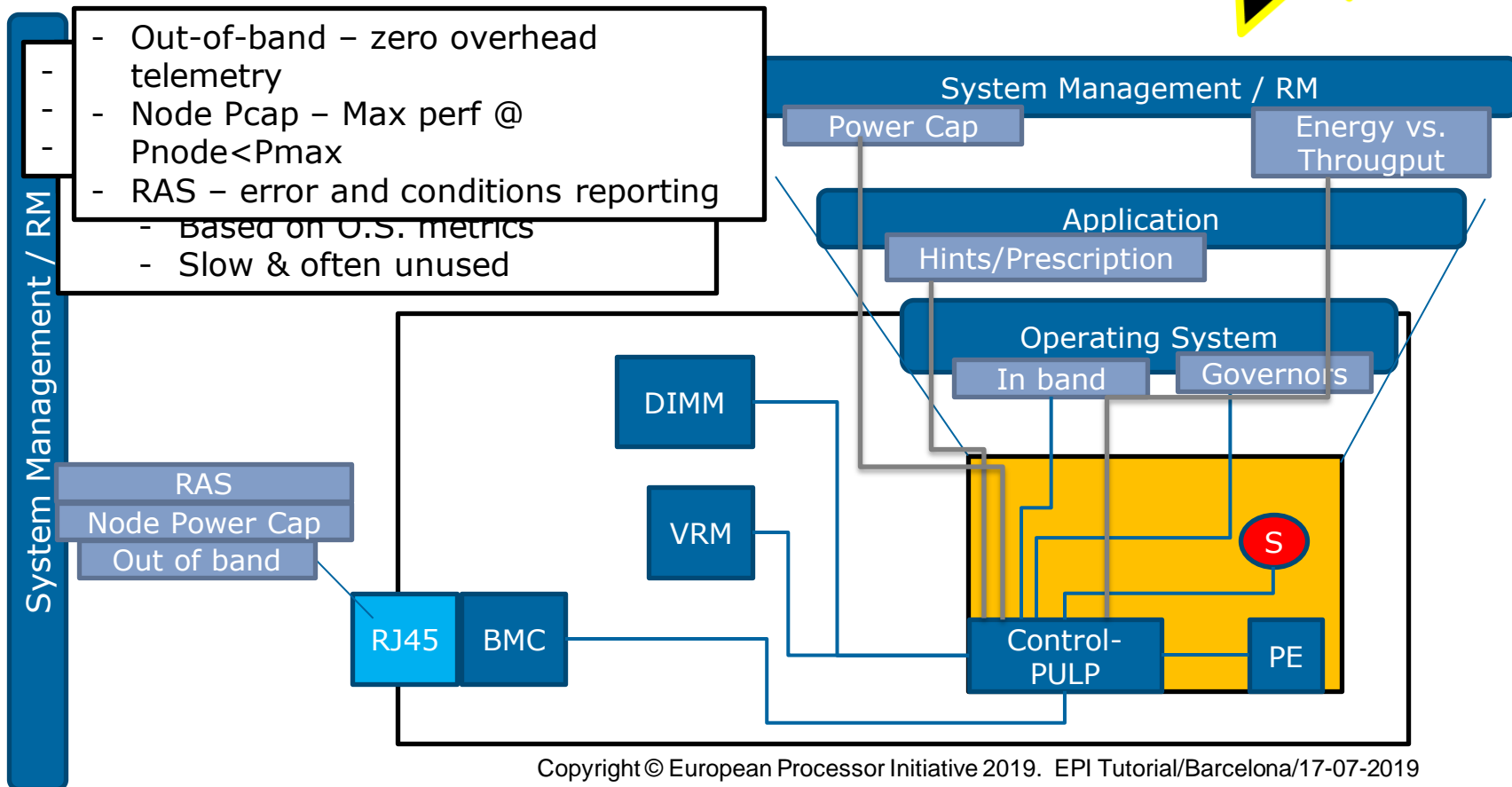
RISC-V

RISC-V



ControlPULP

**ControlPULP:
The Power
Controller**



Copyright © European Processor Initiative 2019. EPI Tutorial/Barcelona/17-07-2019

Coming Soon...

ControlPULP

T control

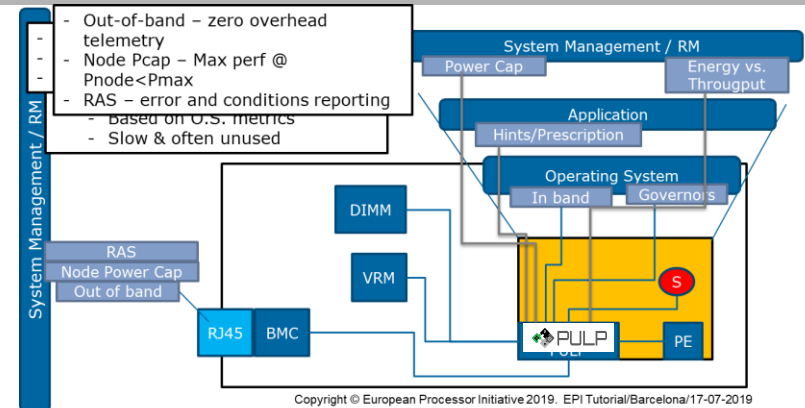
Watchdog reset
 Write power controller settings
 Write to internal memory telemetry data
 Read PVT sensors
 Read workload from O.S.
 Read **target P/C state settings, power budget**
 Read Pending BMC requests
 Compute controller settings

PM task

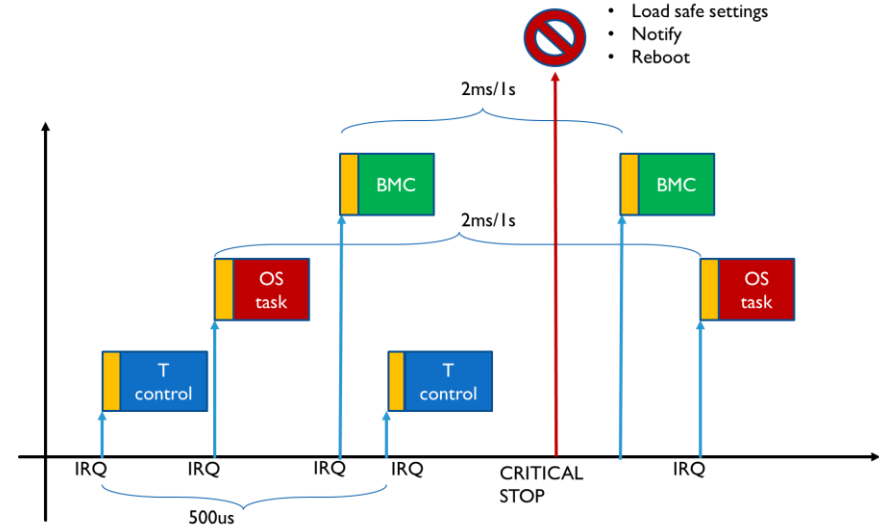
- Read voltage regulator, power, status (VR)
- Power model update

BMC

- Read pending command queue
- Decode Command/data
- Perform action:
 - Change **target P/C state, power budget**
 - Set pending BMC
 - Ask telemetry data



Coming Soon...



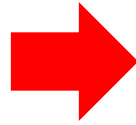
Copyright © European Processor Initiative 2019. EPI Tutorial/bologna/22-01-2020

Coming Soon...



Ack. Robert Balas, Giovanni Bambini, Andrea Bentivogli, Davide Rossi, Antonio Mastrandrea, Christian Conficoni, Simone Benatti, Andrea Tilli

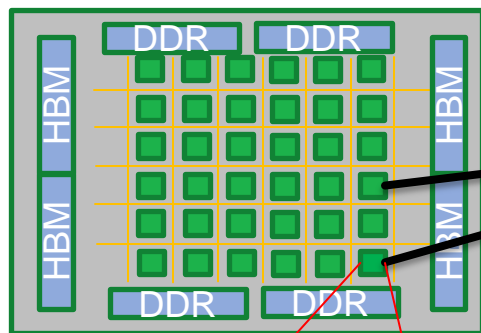
HPC Vertical: The European Processor Initiative



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 increasing 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**
- Computing platform for autonomous cars
- Will also target the AI, Big Data and other markets in order to be economically sustainable

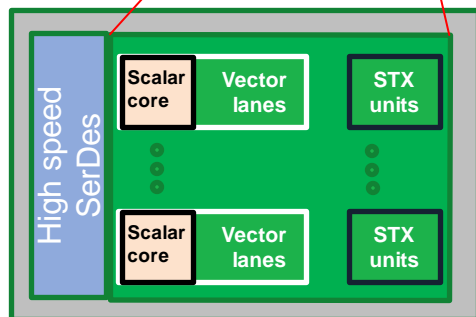
First Generation EPI chips



General Purpose Processor (GPP) chip

- 7 nm, chip-let technology
- **ARM-SVE** tiles
- **EPAC RISC-V** vector+AI accelerator tiles
- L1, L2, L3 cache subsystem + HBM + DDR

GPP power manager based on ControlPULP!



RISC-V Accelerator Demonstrator Test Chip

- 22 nm FDSOI
- Only one RISC-V accelerator tile
- On-chip L1, L2 + off-chip HBM + DDR PHY
- Targets 128 DP GFLOPS (vector) **200+GOPs/W SP (STX)**

Scalar Core + STX units based on NTX and Snitch!



EXASCALE
2021



*The fun is
just beginning*



<http://pulp-platform.org>