PULP: an Open Hardware Platform
The story so far

HPCA 2018 - Vienna

25.02.2018

Frank K. Gürkaynak
Florian Zaruba
Andreas Kurth
Francesco Conti

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

Multitherman

http://pulp-platform.org

ETH zürich

[2]*Integrated Systems Laboratory*

Structure of this workshop

- Open source hardware and our role (Frank)
- The PULP family tree (Frank)
- Our RISC-V cores: Ariane, RI5CY and friends (Florian)

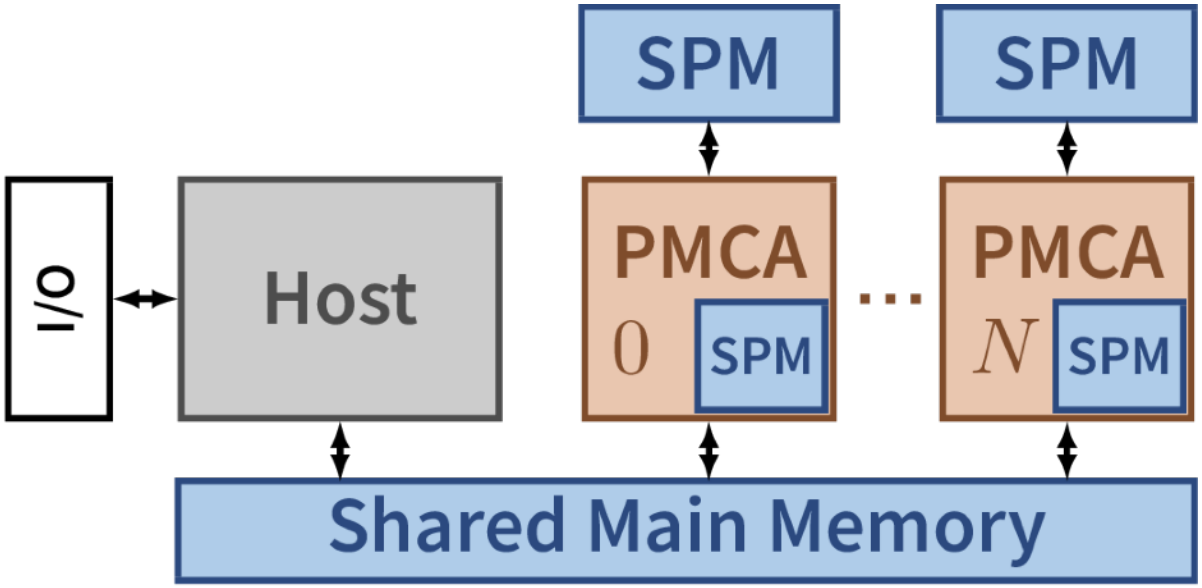
- Break – Demos

- Accelerators in PULP (Francesco)
- **Our Programmable Multi-Core Accelerator – HERO (Andreas)**
- **Programming PULP (Andreas)**

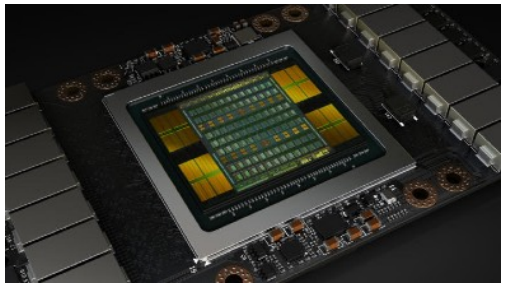
Please interrupt at any time to ask questions

PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

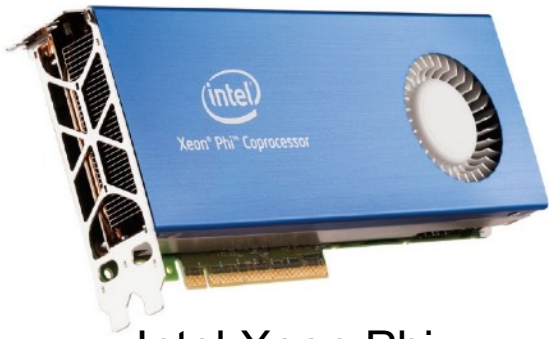
Heterogeneous Computing Systems



Architectural template for heterogeneous computers



NVIDIA Tesla V100

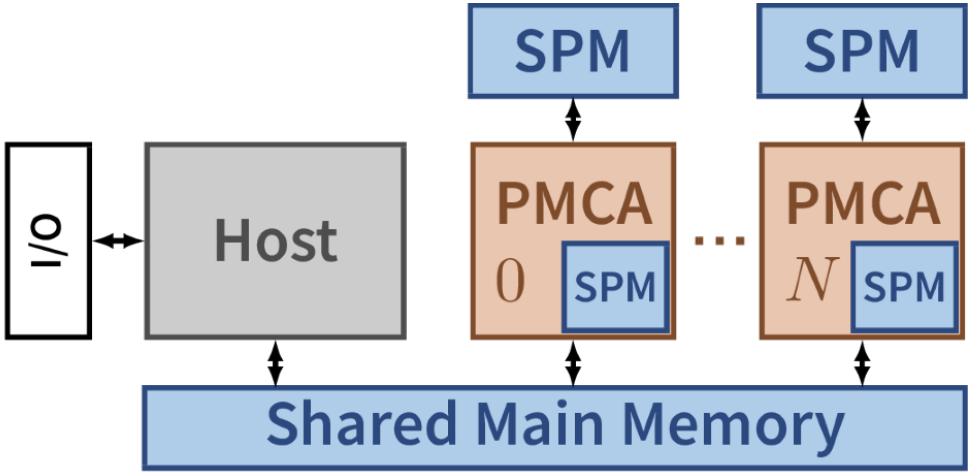


Intel Xeon Phi

- Heterogeneous computers combine a **general-purpose host processor** and efficient, domain-specific **programmable manycore accelerators (PMCAs)**.
- They unite versatility with energy efficiency.

PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

Research on Heterogeneous Computers



Architectural template for heterogeneous computers

- There are **many open questions in various areas** of computer engineering:
- programming models, task distribution, scheduling
- memory organization, communication, synchronization
- accelerator architectures and granularity
- To investigate them, we have built a **research platform**: **HERO**

PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

Why not just simulate Heterogeneous Computers?

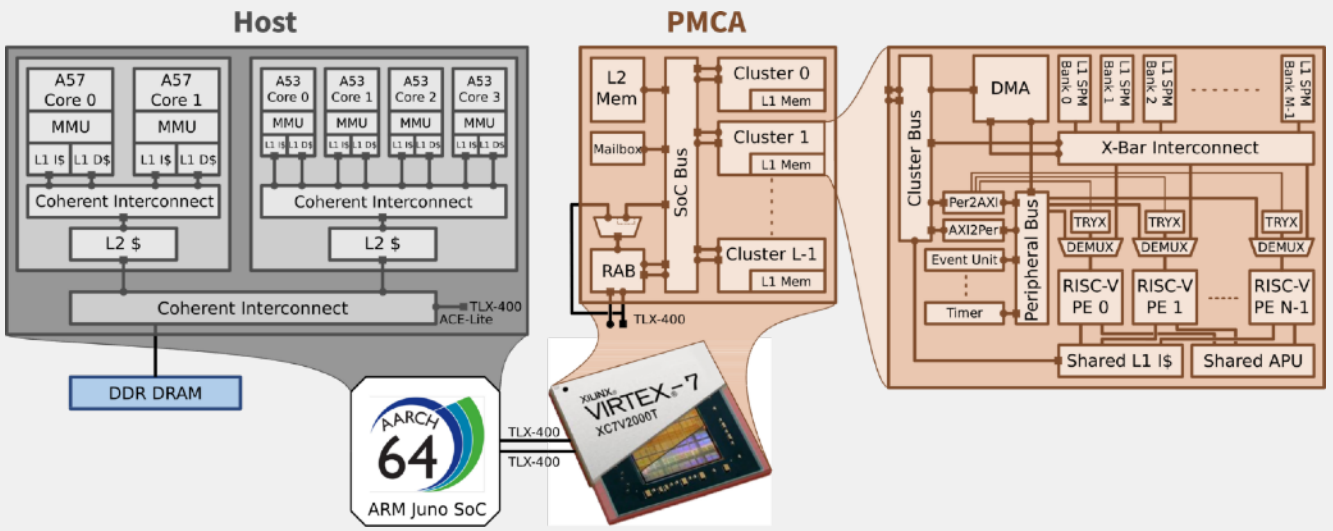
- **Simulations** (of reasonable accuracy) are **orders of magnitude slower** than running prototypes. (**HERO** currently runs **up to 1.9 billion Instr/sec**!)
- Even full-system simulators (e.g., gem5) do not model all heterogeneous components.
- Models make assumptions about non-deterministic processes. The **validity of results** thus **entirely depends on the validity of assumptions**, and the assumptions for heterogeneous computers are very complex.
- Models are based on **reverse engineering commercial black box components**.

Conclusion: **A research platform for heterogeneous computers must be available!**

PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

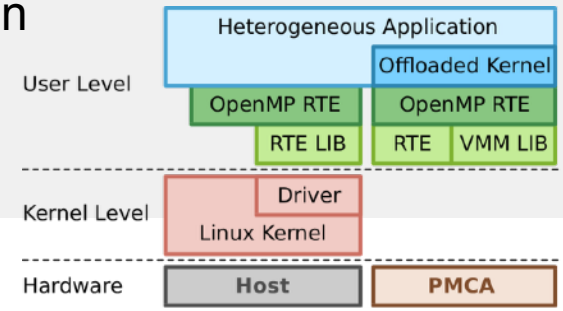
HERO: Open-Source Heterogeneous Research Platform

Heterogeneous Hardware Architecture



Heterogeneous Software Stack

- single-source, single-binary cross compilation toolchain
- OpenMP 4.5
- Shared Virtual Memory for Host and PMCAs

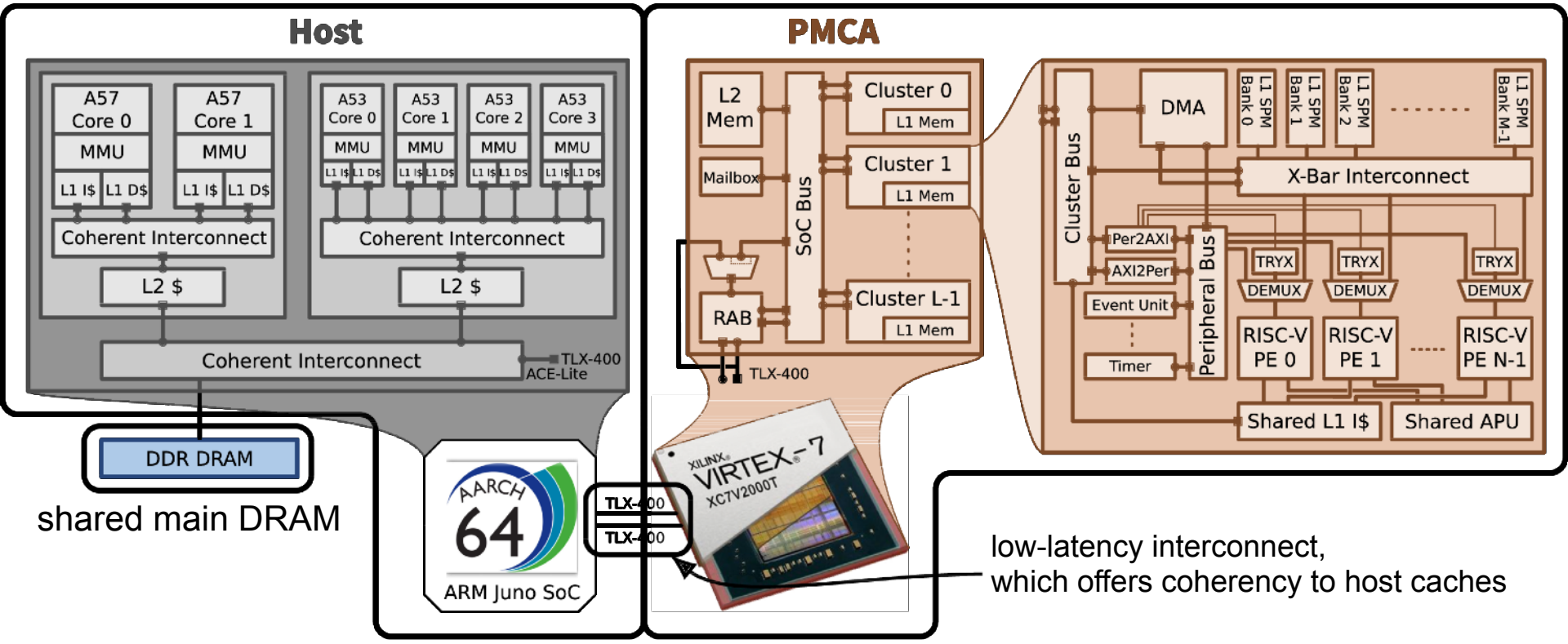


PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

HERO's Hardware Architecture

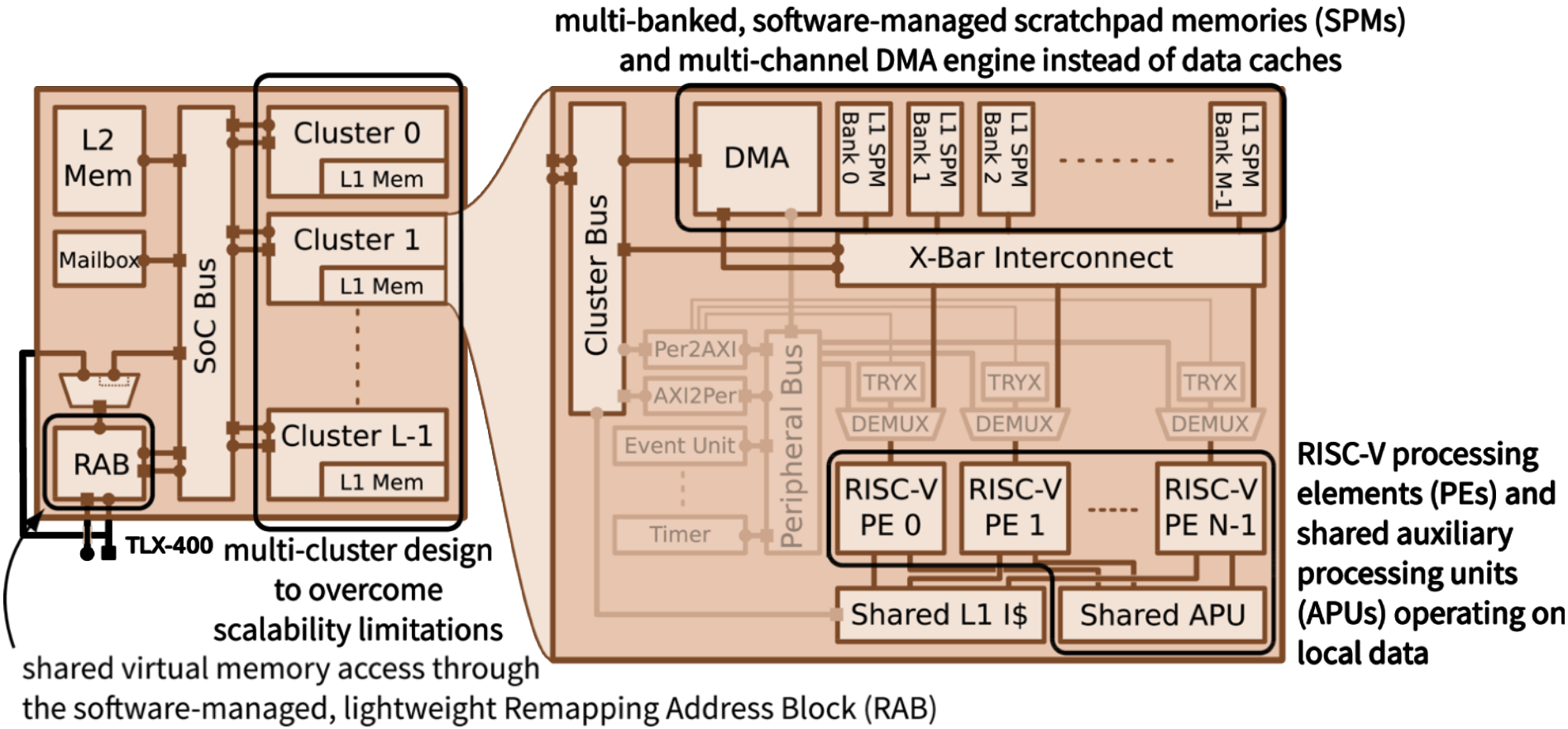
hard-macro ARM Cortex-A Host Processor

scalable, configurable, modifiable FPGA implementation of a silicon-proven, cluster-based PMCA with RISC-V PEs



PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

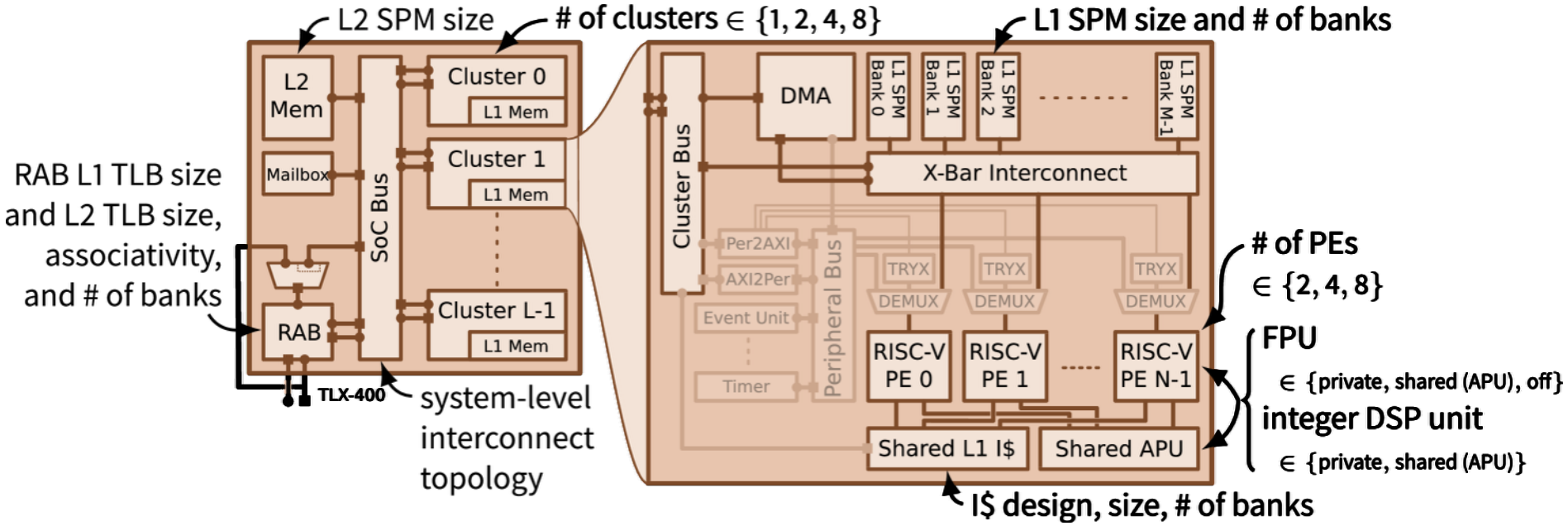
HERO's PMCA Implementation on FPGA: Overview



PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

HERO's PMCA on FPGA: Configurable, Modifiable, and Expandable

Configurable:



Modifiable and expandable:

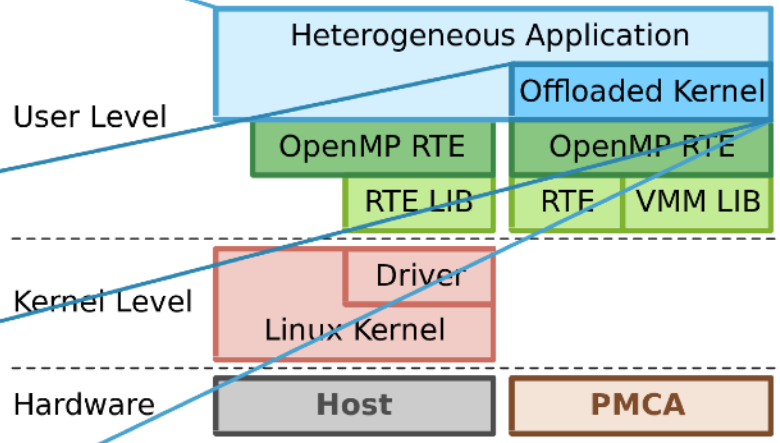
- All components are open-source and written in industry-standard System Verilog.
- Interfaces are either standard (mostly AXI) or simple (e.g., stream-payload).
- New components can be easily added to the memory map.

PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

HERO: Software Stack

Allows to write programs that start on the host but seamlessly integrate the PMCAs.

```
int main()
{
  vertex vertices[N];
  load(&vertices, N);
  #pragma omp target map(tofrom:vertices)
  {
    #pragma omp parallel for
    for (i = 0; i < N; ++i)
      vertices[i] = process();
  }
}
```



- Offloads with OpenMP 4.5 target semantics, zero-copy (pointer passing) or copy-based
- First non-commercial heterogeneous cross compilation toolchain
- PMCA-specific runtime and hardware abstraction libraries (HAL)

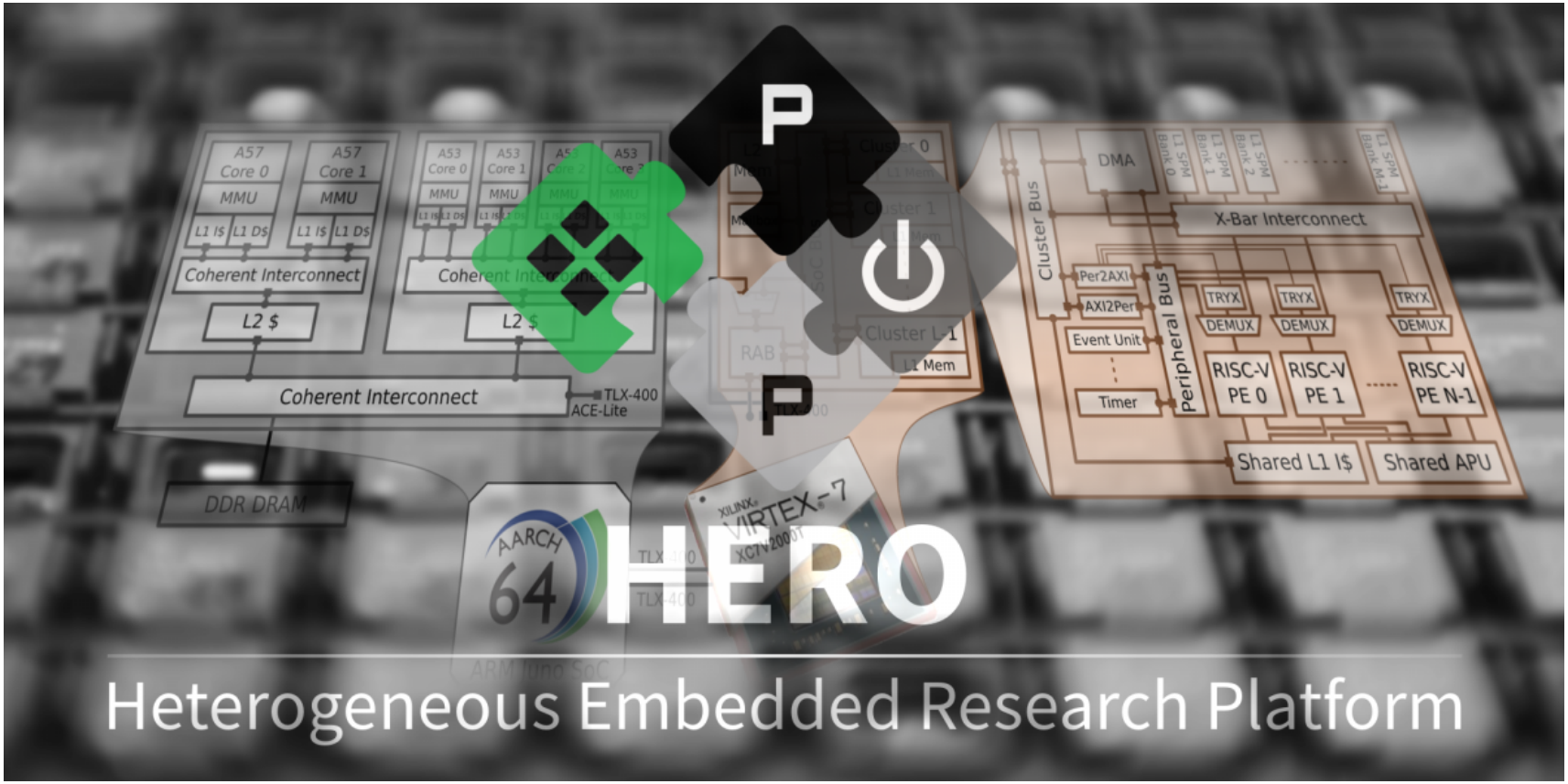
PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

HERO: Supported Platforms and Configurations

Property	ARM Juno (with a Xilinx Virtex-7 2000T)	Xilinx Zynq UltraScale+ ZU9	Xilinx Zynq ZC706
Host CPU	64-bit ARMv8 big.LITTLE	64-bit ARMv8 quad-core A53	32-bit ARMv7 dual-core A9
Shared main memory	8 GiB DDR3L	2 GiB DDR4	1 GiB DDR3
PMCA clock frequency	**31 MHz**	**145 MHz**	**57 MHz**
# of RISC-V PEs	**64 in 8 clusters**	**8 in 1 cluster**	**8 in 1 cluster**
Integer DSP unit	private per PE		
L1 SPM	256 KiB in 16 banks		
Instruction cache	8 KiB in 8 single-ported banks	4 KiB in 4 multi-ported banks	
Slices used by clusters	80%	48%	65%
Slices used by infrastructure	7%	10%	12%
BRAMs used by clusters	89%	42%	70%
BRAMs used by infrastructure	6%	8%	13%
Price	25 000 $	2500 $	2500 $

PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

HERO will be released open-source



Coming Q1 2018

pulp-platform.org/hero

PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

Structure of this workshop

- Open source hardware and our role (Frank)
- The PULP family tree (Frank)
- Our RISC-V cores: Ariane, RI5CY and friends (Florian)

- Break – Demos

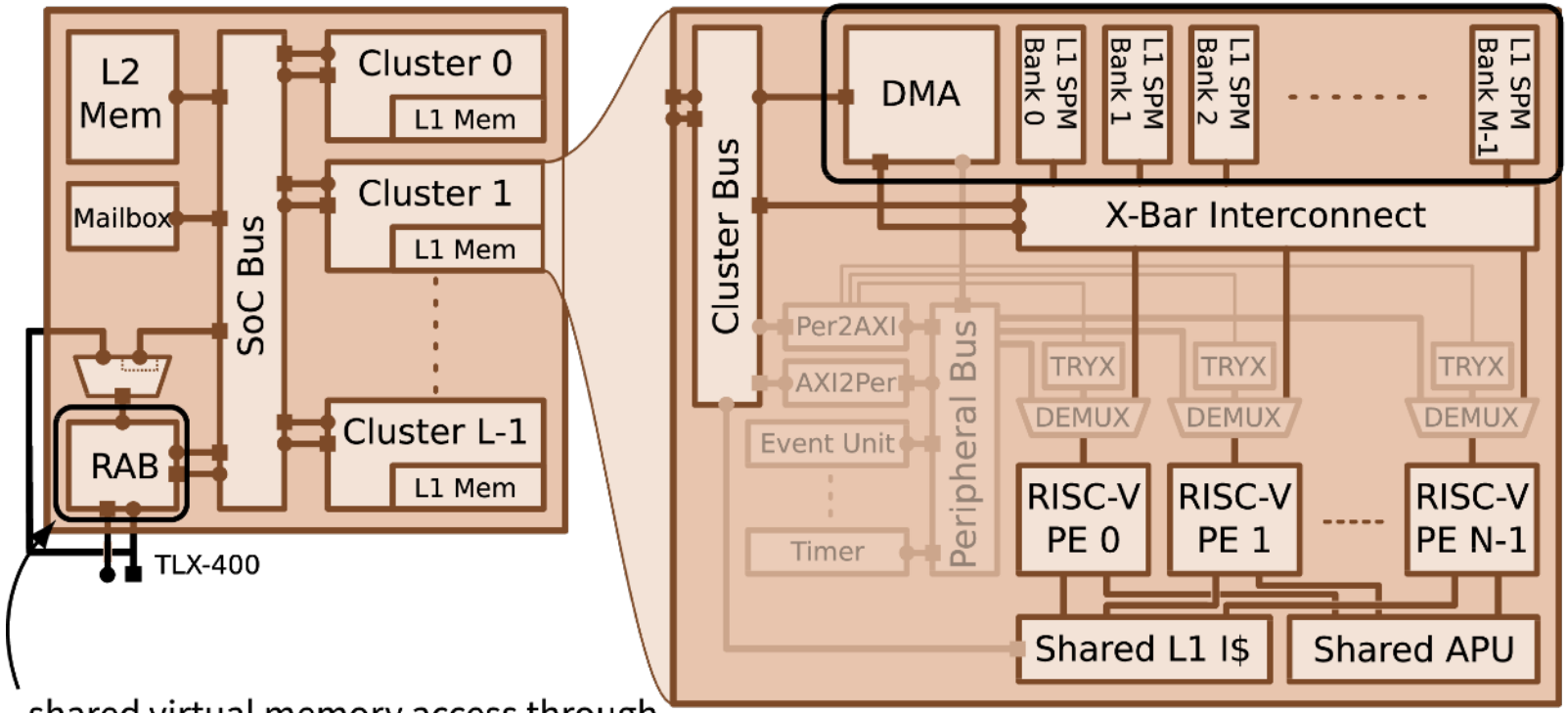
- Accelerators in PULP (Francesco)
- Our Programmable Multi-Core Accelerator – HERO (Andreas)
- **Programming PULP (Andreas)**

Please interrupt at any time to ask questions

PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

HERO: PMCA Memory HW

multi-banked, software-managed scratchpad memories (SPMs)
and multi-channel DMA engine instead of data caches

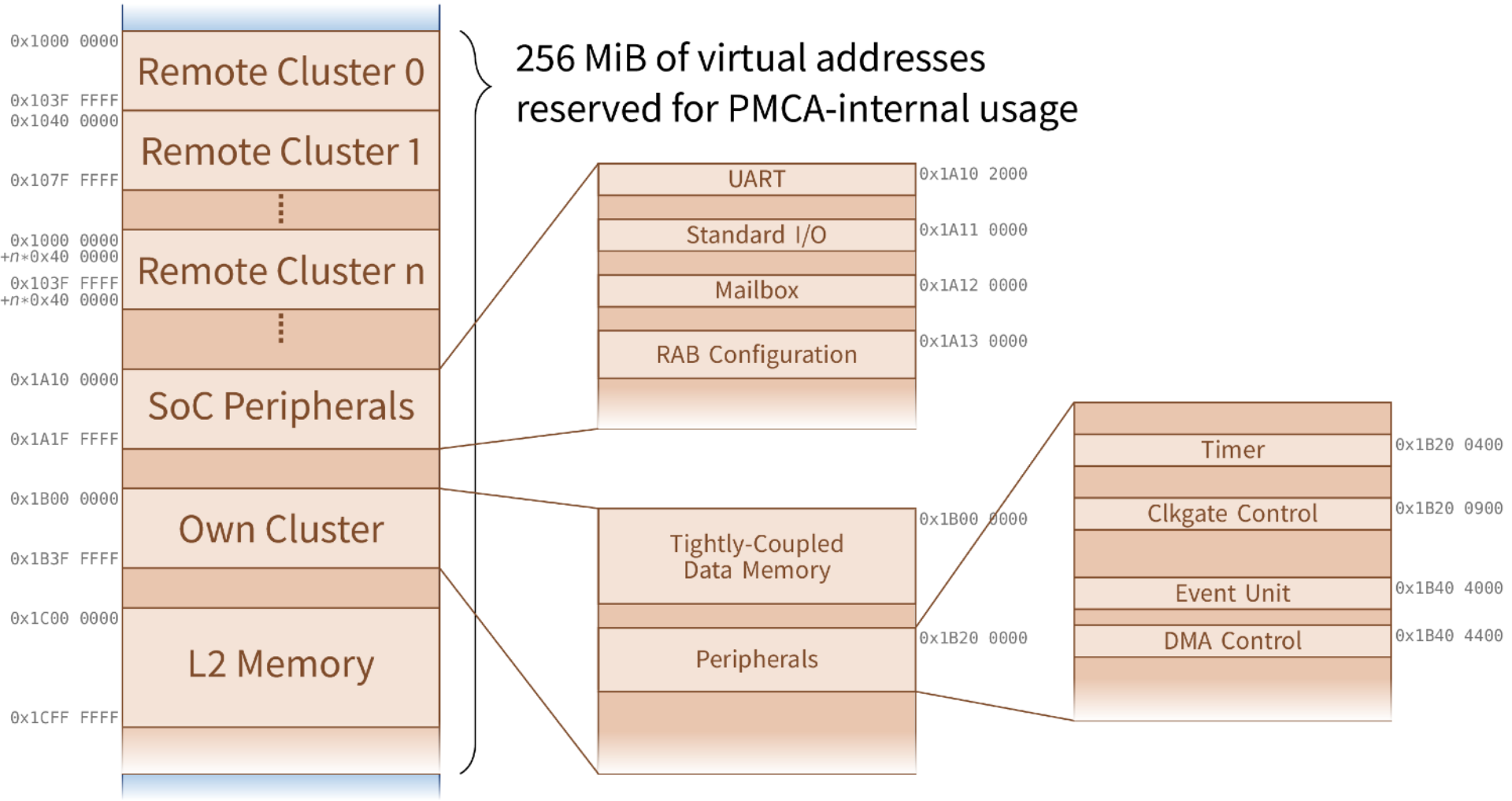


shared virtual memory access through
the software-managed, lightweight Remapping Address Block (RAB)

- flat NUMA hierarchy
- data transfers with DMA bursts

PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

HERO: PMCA Memory Map



PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

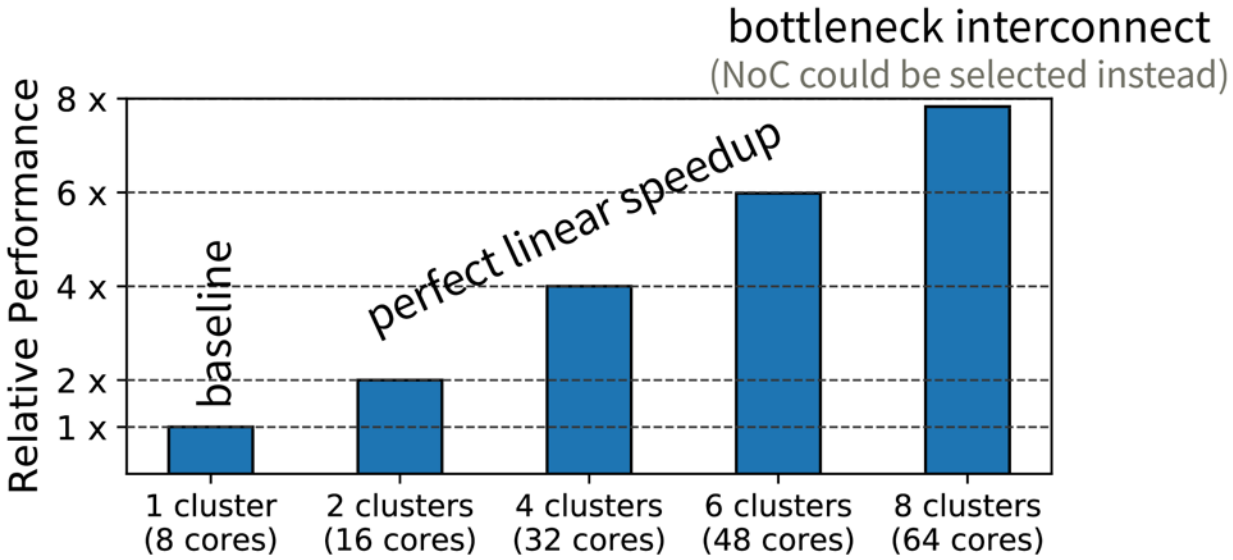
Programming HERO

Generic example:

- Data allocated by host in shared memory
- Processing initially by CPU
- Offload to PMCA with OpenMP
- Parallelization on PMCA with OpenMP
- DMA transfers with PULP API
- Cycle-accurate measurements of PMCA performance with counters and monitors
- … can be extended with diverging threads, tiled data allocation and movement, …

Case Study: Parallel Speedup Analysis

- Benchmarking parallel execution and data transfers of the PMCA on the Juno ADP
- Matrix-matrix multiplication $C = AB$
- A and C are tiled row-wise over the clusters, and each row is parallelized block-wise over the PEs. Data is transferred with DMA bursts, and all PEs operate on data in local SPMs.



- HERO allows to make architectural choices based on **measured results** of benchmarks.

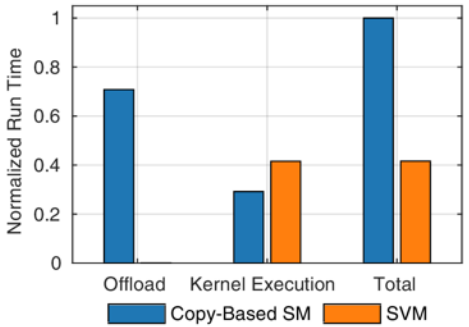
PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

Case Study: Shared Virtual Memory Performance Analysis

The main motivation for shared virtual memory (SVM) is programmability. However, SVM can also significantly improve performance!

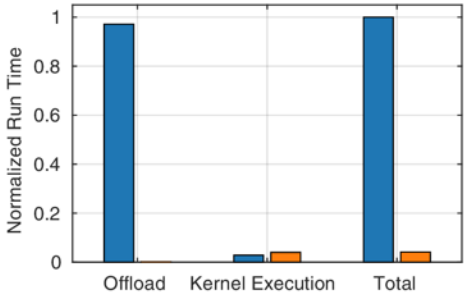
PageRank is a well-known algorithm for analyzing the connectivity of graphs.

- The overhead of manipulating pointers at offload-time in **copy-based offloading** *exceeds* the run-time overhead of translating pointers with **shared virtual memory**.
- In this case, SVM reduces the run time by nearly 60 %.



MemCopy simply copies a large array from DRAM to the PMCA and back, which is representative for streaming applications with little actual work.

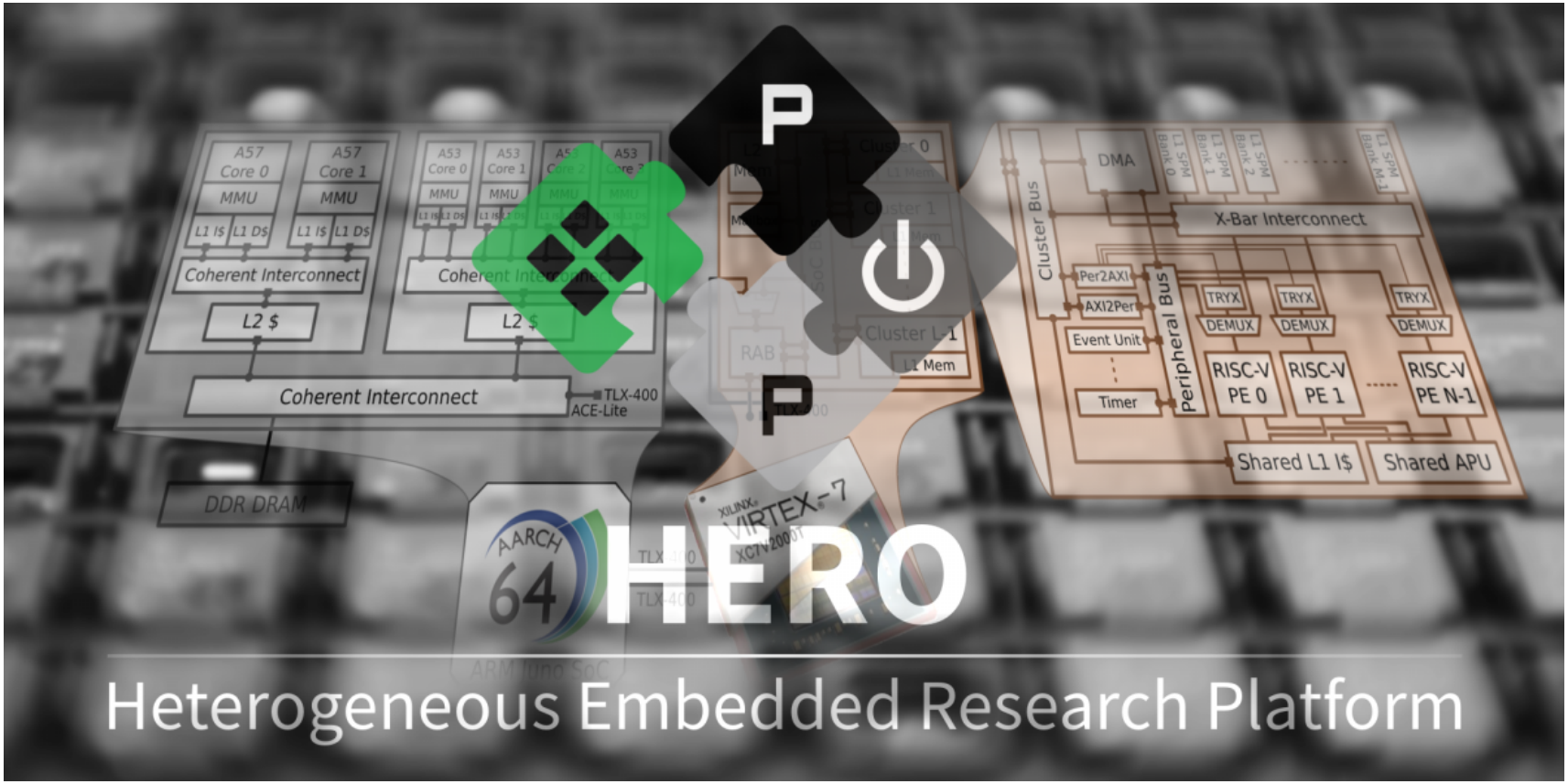
- Letting the host **copy data to physically contiguous, uncached memory** is much slower than **letting the PMCA access data directly** with high-bandwidth DMA transfers.
- In this case, SVM reduces the run time by more than 95 %.



- HERO allows to back research claims with **reproducible, falsifiable implementation results.**

PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH

HERO will be released open-source



Coming Q1 2018

pulp-platform.org/hero

PULP Multitherman ExaNoDe PRECOMP Open Transprecision Computing ETH