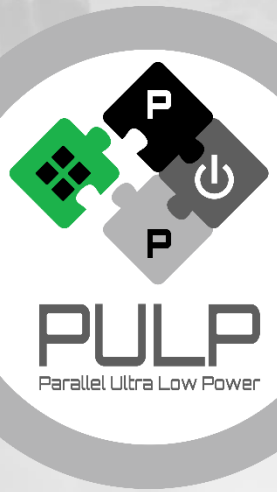


HERO: Heterogeneous Research Platform

Open-Source HW/SW Platform for R&D of Heterogeneous SoCs *21.01.2019*

Andreas Kurth

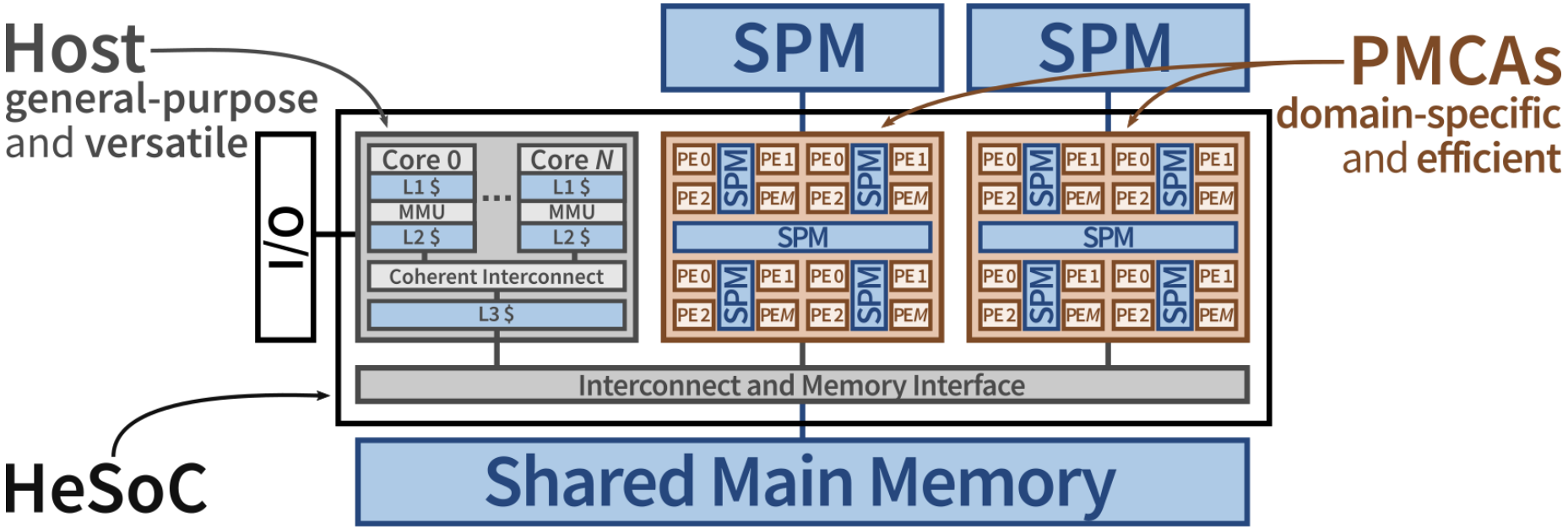
and the PULP Team led by Prof. Luca Benini

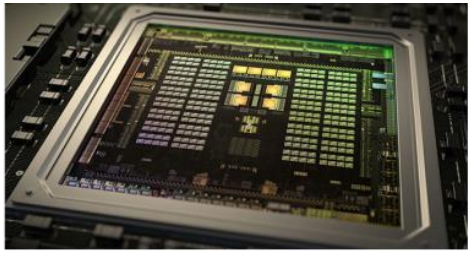
[1]*Department of Electrical, Electronic and Information Engineering*

ETH zürich

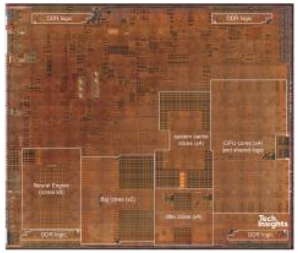
[2]*Integrated Systems Laboratory*

Heterogeneous Systems on Chip (HeSoCs)





Nvidia Tegra X1 (source: Nvidia)



Apple A12 (source: TechInsights)

Research on Heterogeneous SoCs

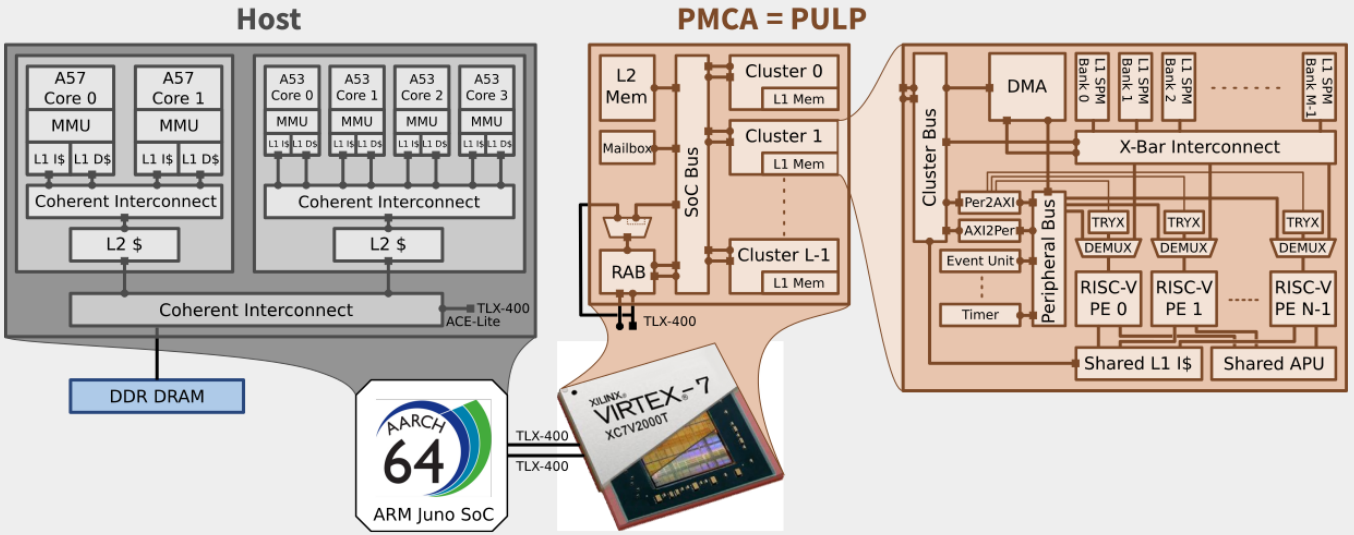
There are **many open questions** in various areas of computer engineering:

- programming models, task distribution and scheduling,
- memory organization, communication, synchronization,
- accelerator architectures and granularity, …

But there was no **research platform for heterogeneous SoCs**!

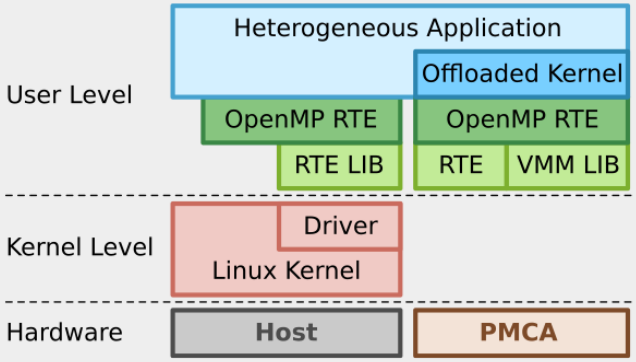
HERO: Heterogeneous Research Platform

Heterogeneous Hardware Architecture



Heterogeneous Software Stack

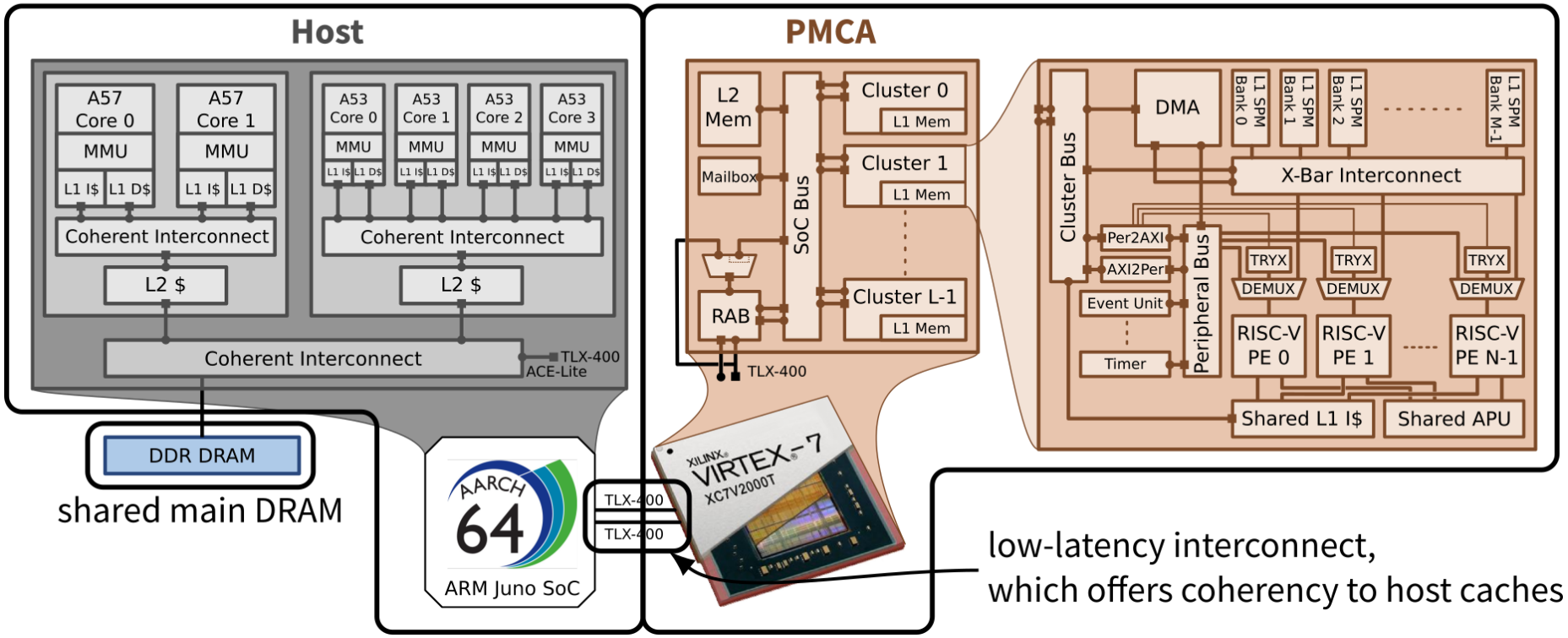
- single-source, single-binary cross compilation toolchain
- OpenMP 4.5
- shared virtual memory for Host and PMCA



HERO: Hardware Architecture

industry-standard, hard-macro ARM Cortex-A Host processor

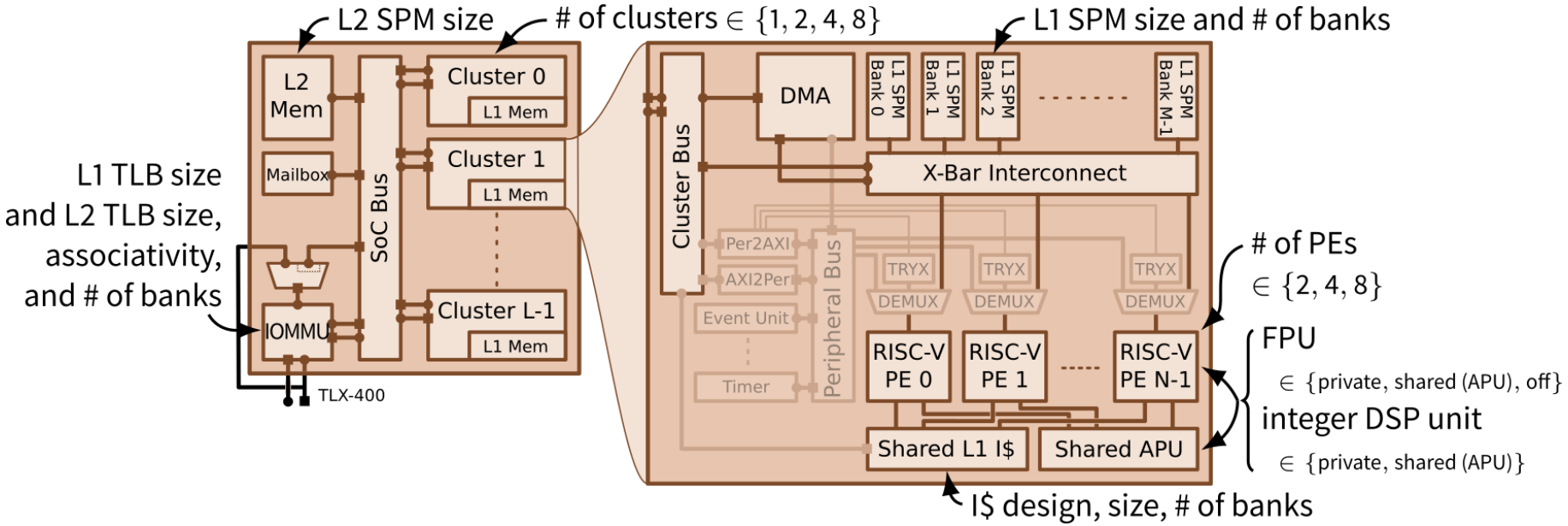
scalable, configurable, modifiable FPGA implementation of PULP (silicon-proven, cluster-based PMCA with RISC-V PEs)



low-latency interconnect, which offers coherency to host caches

bigPULP on FPGA: Configurable, Modifiable and Expandable

Configurable:

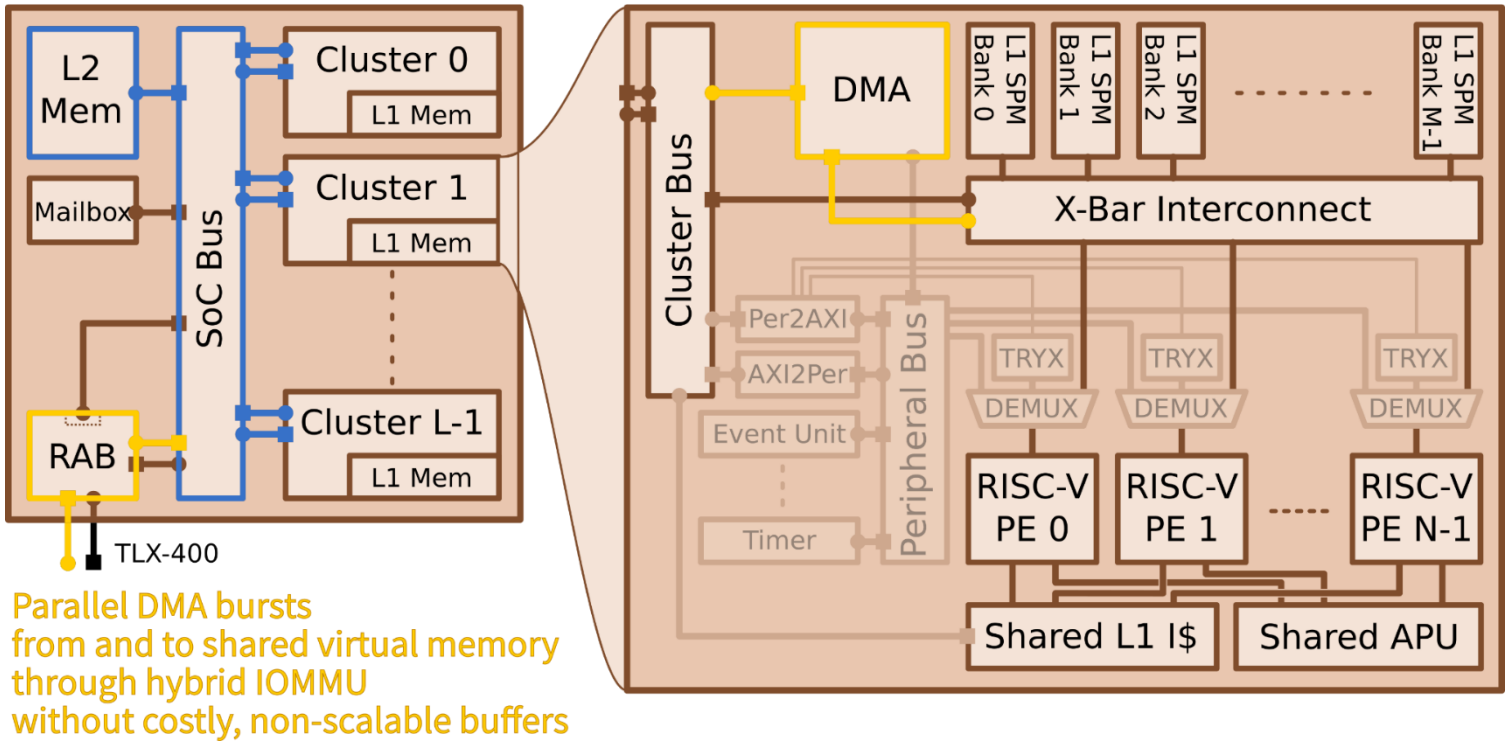


Modifiable and expandable:

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

bigPULP: Distinguishing Features

Scalable and efficient multi-cluster atomic transactions
(RISC-V 'A' extension) to shared L2 memory



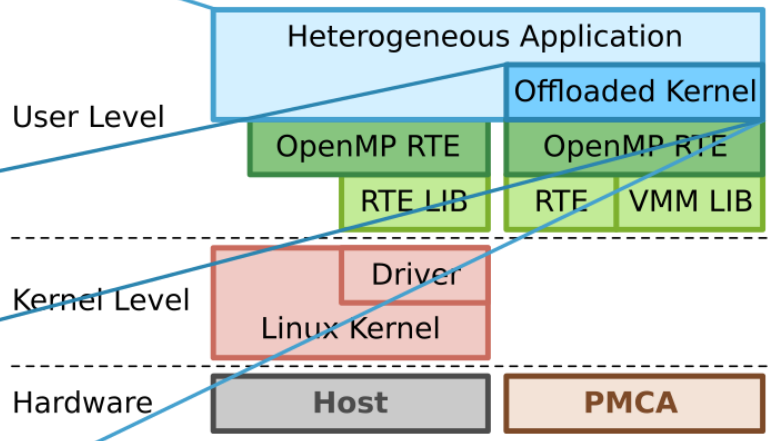
Parallel DMA bursts
from and to shared virtual memory
through hybrid IOMMU
without costly, non-scalable buffers

- Atomic transactions: RI5CY with ‘A’ decoder, additional signals through cluster and SoC bus, transactions executed atomically at L2 SPM
- Scalable SVM: Two-level software-managed TLB (“RAB”); TLB misses signaled back to RI5CY and DMA; handled in SW with lightweight HW support

HERO: Software Architecture

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

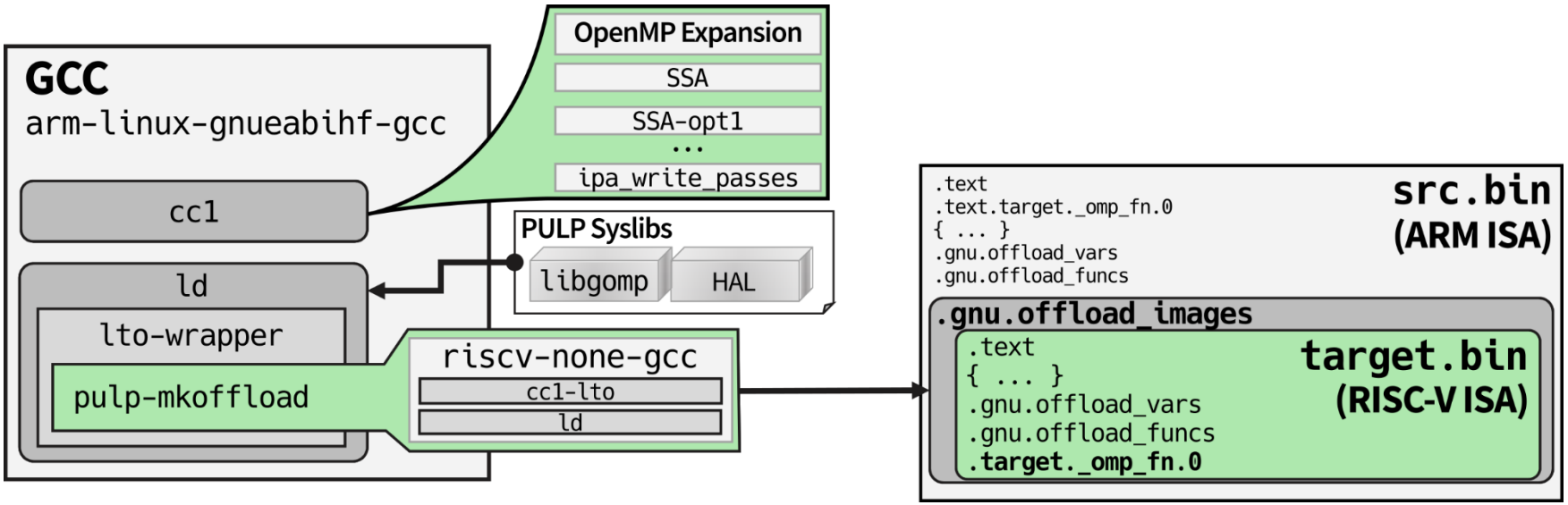
```
int main()
{
  vertex vertices[N];
  load(&vertices, N);
  #pragma omp target map(tofrom:vertices)
  {
    #pragma omp parallel for
    for (i = 0; i < N; ++i)
      process(vertices[i]);
  }
}
```



- Offloads with OpenMP 4.5 `target` semantics, zero-copy (pointer passing) or copy-based
- Integrated cross-compilation and single-binary linkage
- PMCA-specific runtime environment and hardware abstraction libraries (HAL)

HERO: Heterogeneous Cross-Compilation Toolchain

- OpenMP offloading with the GCC toolchain requires a **host compiler** plus **one target compiler for each PMCA ISA** in the system.



- A target compiler requires both **compiler and runtime extensions**.
- HERO includes the **first non-commercial** heterogeneous cross-compilation toolchain.

HERO: FPGA Platforms

Property	ARM Juno (with a Xilinx Virtex-7 2000T)	Xilinx Zynq UltraScale+ ZU9EG	Xilinx Zynq Z-4045
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	**30 MHz**	**150 MHz**	**50 MHz**
# of RISC-V PEs	**64 in 8 clusters**	**16 in 2 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%	63%	65%
Slices used by infrastructure	7%	15%	12%
BRAMs used by clusters	89%	55%	70%
BRAMs used by infrastructure	6%	12%	13%
Price	25 000 $	2500 $	2500 $

HERO: Roadmap

September 2018
v1.0
Public release of the world's first open-source heterogeneous hardware and software stack

October 2018
v1.1
Full support for OpenMP 4.5 API and release of example applications

February 2019
v1.2
Automatic compile-time insertion of SVM intrinsics

H1 2019
v2.0
US+ FPGAs with 64-bit hosts, support for 'F' extension with shared FPUs multi-cluster OpenMP RTE

H1 2020
v3.0
Replace ARM host processor with multi-core Ariane → world's first fully open-source HeSoC

PULP

HERO: Getting Started

```
git clone --recursive \
 https://github.com/pulp-platform/hero-sdk
cd hero-sdk; git checkout v1.1.0
```

Check `README.md` for prerequisites and install them.

```
./hero-z-7045-builder -A
```

Questions?

www.pulp-platform.org

@pulp_platform

PULP
Parallel Ultra Low Power

Florian Zaruba[2], Davide Rossi[1], Antonio Pullini[2], Francesco Conti[1], Michael Gautschi[2], Frank K. Gürkaynak[2], Florian Glaser[2], Stefan Mach[2], Giovanni Rovere[2], Igor Loi[1] Davide Schiavone[2], Germain Haugou[2], Manuele Rusci[1], Alessandro Capotondi[1], Giuseppe Tagliavini[1], Daniele Palossi[2], Andrea Marongiu[1,2], Fabio Montagna[1], Simone Benatti[1], Eric Flamand[2], Fabian Schuiki[2], Andreas Kurth[2], Luca Benini[1,2]

[1]Department of Electrical, Electronic and Information Engineering

ETH zürich
[2]Integrated Systems Laboratory