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

1Department of Electrical, Electronic and Information Engineering

ETHzürich

2Integrated Systems Laboratory
Heterogeneous Systems on Chip (HeSoCs)

Host

general-purpose and versatile

I/O

HeSoC

Shared Main Memory

PMCAs
domain-specific and efficient

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)

**Host**

- A57 Core 0
- A57 Core 1
- A53 Core 0
- A53 Core 1
- A53 Core 2
- A53 Core 3
- MMU
- L1 I$ L1 D$
- LL I$ LL D$
- Coherent Interconnect
- L2 $
- Coherent Interconnect
- DDR DRAM

**PMCA**

- L2 Mem
- Cluster 0
- Cluster 1
- Cluster L-1
- L1 Mem
- Mailbox
- SoC Bus
- DMA
- RAB
- Per2AXI
- AXI2Per
- Event Unit
- Timer
- Peripheral Bus
- X-Bar Interconnect
- Bank 0
- Bank 1
- Bank 2
- Band 2
- L1 SPM
- L1 SPM
- L1 SPM
- L1 SPM
- L1 SPM
- L1 SPM
- RISC-V PE 0
- RISC-V PE 1
- RISC-V PE N-1
- Shared L1 I$
- Shared APU

shared main DRAM

low-latency interconnect, which offers coherency to host caches
bigPULP on FPGA: Configurable, Modifiable and Expandable

Configurable:
- 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.

Modifiable and expandable:
- # of clusters $\in \{1, 2, 4, 8\}$
- L1 SPM size and # of banks
- L1 TLB size and L2 TLB size, associativity, and # of banks
- # of PEs $\in \{2, 4, 8\}$
- FPU $\in \{\text{private, shared (APU), off}\}$
- Integer DSP unit $\in \{\text{private, shared (APU)}\}$
- L$ design, size, # of banks

Andreas Kurth | 20.01.2019 | 6
bigPULP: Distinguishing Features

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

- 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

Parallel DMA bursts from and to shared virtual memory through hybrid IOMMU without costly, non-scalable buffers
**HERO: Software Architecture**

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

```c
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)
• 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

<table>
<thead>
<tr>
<th>Property</th>
<th>ARM Juno (with a Xilinx Virtex-7 Z000T)</th>
<th>Xilinx Zynq UltraScale+ ZU9EG</th>
<th>Xilinx Zynq Z-4045</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host CPU</td>
<td>64-bit ARMv8 big.LITTLE</td>
<td>64-bit ARMv8 quad-core A53</td>
<td>32-bit ARMv7 dual-core A9</td>
</tr>
<tr>
<td>Shared main memory</td>
<td>8 GiB DDR3L</td>
<td>2 GiB DDR4</td>
<td>1 GiB DDR3</td>
</tr>
<tr>
<td>PMCA clock frequency</td>
<td>30 MHz</td>
<td>150 MHz</td>
<td>50 MHz</td>
</tr>
<tr>
<td># of RISC-V PEs</td>
<td>64 in 8 clusters</td>
<td>16 in 2 cluster</td>
<td>8 in 1 cluster</td>
</tr>
<tr>
<td>Integer DSP unit</td>
<td></td>
<td>private per PE</td>
<td></td>
</tr>
<tr>
<td>L1 SPM</td>
<td></td>
<td>256 KiB in 16 banks</td>
<td></td>
</tr>
<tr>
<td>Instruction cache</td>
<td>8 KiB in 8 single-ported banks</td>
<td>4 KiB in 4 multi-ported banks</td>
<td></td>
</tr>
<tr>
<td>Slices used by clusters</td>
<td>80%</td>
<td>63%</td>
<td>65%</td>
</tr>
<tr>
<td>Slices used by infrastructure</td>
<td>7%</td>
<td>15%</td>
<td>12%</td>
</tr>
<tr>
<td>BRAMs used by clusters</td>
<td>89%</td>
<td>55%</td>
<td>70%</td>
</tr>
<tr>
<td>BRAMs used by infrastructure</td>
<td>6%</td>
<td>12%</td>
<td>13%</td>
</tr>
<tr>
<td>Price</td>
<td>25 000 $</td>
<td>2500 $</td>
<td>2500 $</td>
</tr>
</tbody>
</table>
**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
HERO: Getting Started

git clone --recursive \n  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

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

\textsuperscript{1}Department of Electrical, Electronic and Information Engineering

\textsuperscript{2}Integrated Systems Laboratory