PULP
RESOURCES & PROJECTS

SUMMARY OF OPEN-SOURCE RESOURCES AND PROJECTS FOR PULP

For detailed overview : Github

Processors and Microcontrollers

- PROCESSORS
- MICRO-CONTROLLERS
- PERIPHERALS
- CLUSTER BASED SYSTEMS
- PULP AS A MULTI-CLUSTER ACCELERATOR

Processors

RISC V compatible cores written in System Verilog.

- **CV32E40P (RI5CY) -** A 32bit / 4-stage core that implements, the RV32-IMC, has an optional 32-bit FPU supporting the F extension and instruction set extensions for DSP operations, including hardware loops, SIMD extensions, bit manipulation and post-increment instructions.

- **Ibex (Zero-riscy) -** An area-optimized 2-stage 32-bit core for control applications implementing RV32-IMC.

- **Micro-risky -** A minimal area 2-stage 32-bit core with 16 registers and no hardware multiplier implementing RV32-EC.

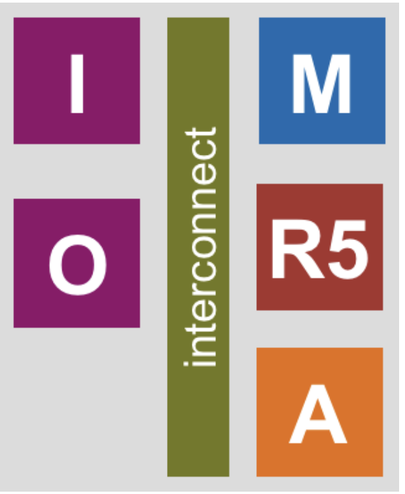
- **CVA6 (Ariane) -** A 6-stage, single issue, in-order 64-bit CPU which fully implements I, M, C and D extensions as specified in Volume I: User-Level ISA V 2.1 as well as the draft privilege extension 1.10. It implements three privilege levels M, S, U to fully support a Unix-like (Linux, BSD, etc.) operating system. It has configurable size, separate TLBs, a hardware PTW and branch-prediction

- **Snitch -** A single-stage, single-issue 32-bit RISC-V integer core tuned for high energy efficiency. It aims at maximizing the compute/control ratio by making the FPU external to the core and the dominant part of the design, as well as mitigating the effects of deep pipelines and dynamic scheduling.

Microcontrollers

- **PULPino -** A minimal single-core RISC-V SoC.

- **PULPissimo -** An advanced version of microcontroller which has the logarithmic interconnect between the core and the memory subsystem allowing multiple access ports. These are used by an integrated uDMA that is able to copy data directly between peripherals and memory, as well as optional accelerators that we call Hardware Processing Engines (HWPEs).



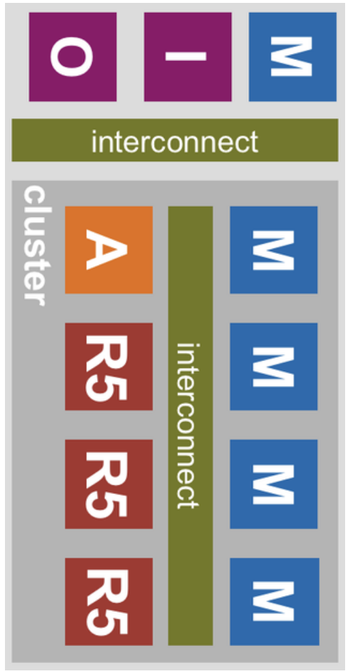
Peripherals

- They have developed customized accelerators, AXI compatible interconnect solutions, different DMA engines and various peripherals to communicate with the environment including GPIO, SPI, I2S, JTAG and many more.

Cluster-based systems

- Based on clusters of 32-bit RISC-V cores with direct access to a small and fast scratchpad memory (Tightly Coupled Data Memory). The cluster is supported by a SoC that houses a larger second level memory, peripherals for input and output, and in later versions a complete PULPissimo class microcontroller for power management and basic operations.

- Mia Wallace, Honey Bunny, Fulmine and finally Mr. Wolf are all such systems, and the source code for the latest system are in this **GitHub** page.



PULP as a multi-cluster accelerator

- They have also expanded work for larger workloads, where a PULP system that contains multiple clusters is connected to a regular computing node. In this scenario, the PULP cluster is used as an energy-efficient accelerator for DSP loads. Their HERO platform release is such a system.

IMPORTANT LINKS FOR RESOURCES

- Conference slides and Videos
- Publicaions
- Youtube
- Getting started with PULP: SW point of view
- PULP Training
- FAQ

Projects

- HERO
- SNITCH
- CARFIELD
- CHESHIRE
- OCCAMY
- SIRACUSA

HERO: Open Heterogeneous Research Platform

- HERO combines a PULP-based open-source parallel manycore accelerator implemented on FPGA with a hard ARM Cortex-A multicore host processor running full-stack Linux. HERO is the first heterogeneous system architecture that mixes a powerful ARM multicore host with a highly parallel and scalable manycore accelerator based on RISC-V cores.

HERO Features

- HERO combines:
  - a hard ARM Cortex-A multicore host processor with
  - a scalable, configurable, and extensible FPGA implementation of an open-source, silicon-proven, cluster-based manycore accelerator.
- The fully open-sourced, heterogeneous software stack of HERO supports:
  - the OpenMP 4.5 Accelerator Model, and
  - shared virtual memory (SVM),
- which allows for transparent accelerator programming and thereby tremendously simplifies the porting of existing applications to create heterogeneous implementations.
- The base configuration of HERO for the Xilinx Zynq ZC706 Evaluation Kit features the following accelerator configuration:
  - 1 PULP cluster (Mr. Wolf) comprising 8 32-bit RI5CY cores,
  - 256 KiB of shared L1 scratchpad memory,
  - 4 KiB of shared L1 instruction cache,
  - 256 KiB of shared L2 scratchpad and instruction memory,
  - our brand-new IOMMU with
    - an L1 TLB of 32 variable-sized entries, and
    - an L2 TLB of 1024 page-sized entries.

GitHub Repositories:

HERO SDK repository

bigPULP hardware repository

Snitch System Generator

- The Snitch project is an open-source RISC-V hardware research project of ETH Zurich and University of Bologna targeting highest possible energy-efficiency. The system is designed around a versatile and small integer core, which we call Snitch. The system is ought to be highly parameterizable and suitable for many use-cases, ranging from small, control-only cores, to large many-core system made for pure number crunching in the HPC domain.

Getting Started

- See our dedicated getting started guide.

Documentation

- The documentation is built from the latest master and hosted at github pages:

Carfield

- Carfield is an open-research heterogeneous platform for safety, resilient and time-predictable systems. Originally conceived as automotive-oriented SoC, the high configurability of the platform makes it tunable to target a broader class of mixed-criticality applications' domains, such as automotive, space or industry.

- Carfield is developed as part of the PULP project, a joint effort between ETH Zurich and the University of Bologna.

- Carfield showcases pioneering hardware solutions, addressing challenges related to time-predictable on/off-chip communication, robust fault recovery mechanisms, secure boot processes, cryptographic acceleration services, hardware-assisted virtualization, and accelerated computation for both floating-point and integer workloads.

Quick Start

- To learn how to build and use Carfield, see Getting Started.
- To learn about available simulation, FPGA, and ASIC targets, see Targets.
- For detailed information on Carfield's inner workings, consult the User Manual.

Cheshire

- Cheshire is a minimal Linux-capable host platform built around the RISC-V CVA6 core. Its goal is to provide a lightweight, configurable, autonomously booting host to systems that need one, from minimal Linux-capable SoCs to manycore compute accelerators.

- Cheshire is developed as part of the PULP project, a joint effort between ETH Zurich and the University of Bologna.

Quick Start

- To learn how to build and use Cheshire, see Getting Started.
- To learn about available simulation, FPGA, and ASIC targets, see Targets.
- For detailed information on Cheshire's inner workings, consult the User Manual.

Occamy

A 432-core, Multi-TFLOPs RISC-V-Based 2.5D Chiplet System for Ultra-Efficient (Mini-)Floating-Point Computation

- Introduction of a novel combination: small, efficient RISC-V integer core (Snitch) and large, multi-precision capable floating-point unit (FPU) with SIMD capabilities.
- New instructions introduced: exsdotp, exvsum, vsum alongside standard FMA instructions.
- Utilization of two architectural extensions: data-prefetchable register file entries and repetition buffers for efficient computation on data-parallel FP workloads.
- Chiplet architecture: Over 216 Snitch cores grouped into compute clusters with tightly-coupled memory and high-bandwidth DMA-enhanced cores managing data flow.
- Each chiplet equipped with private 16GB high-bandwidth memory (HBM2e) and can communicate with neighboring chiplets over wide DDR link.
- Estimated peak performances of dual-chiplet Occamy system: 0.768 TFLOp/s to 6.144 TFLOp/s across different FP formats.

More information on Occamy:

- Slides: Slides 1
- Video: Youtube
- Slides: Slides 2

Siracusa

PULPing Up Extended Reality with At-MRAM Computing

- The Siracusa prototype achieves up to 700 GOPS of performance and up to 2.7 TOPS/W of energy efficiency at the highest N-EUREKA precision setting (8-bit weights), and most importantly, it demonstrates that the At-MRAM technique enables substantial gains in average power for the typical AI workloads used in XR.

- But there remains much to be done: deeper integration between compute SoC's and sensors, towards on-sensor computing; scaling to larger performance to support more of the functions required in XR; better support for complex software pipelines to be executed on sensor.

More information on Siracusa :

- Slides: Slides 1
- PDF: Link
- Video: Link