ControlPULP: A RISC-V Power Controller for HPC Processors with Parallel Control-Law Computation Acceleration

Integrated Systems Laboratory (ETH Zürich)

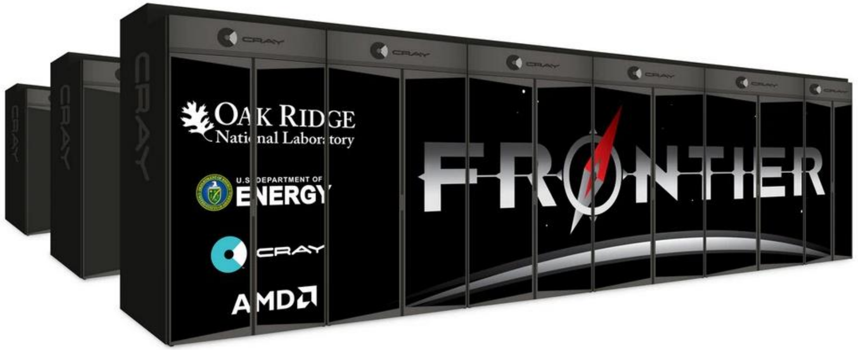
Alessandro Ottaviano, Robert Balas,
Giovanni Bambini, Corrado Bonfanti, Simone Benatti,
Davide Rossi, Luca Benini and Andrea Bartolini

PULP Platform
Open Source Hardware, the way it should be!

@pulp_platform
http://pulp-platform.org
https://www.youtube.com/pulp_platform

Introduction

- **Energy-efficiency is getting growing attention in the HPC field**
  - Top500 leading supercomputer **ORNL's Frontier** leads Green500
  - **150x energy-efficiency** improvement in the last 15 years
  - Cutting-edge cooling systems and advanced power management



- **System-level power and thermal management for HPC**
  - Crucial matter in the many-core era of computing systems
  - Need for fine-grained, workload-aware dynamic power management

Introduction

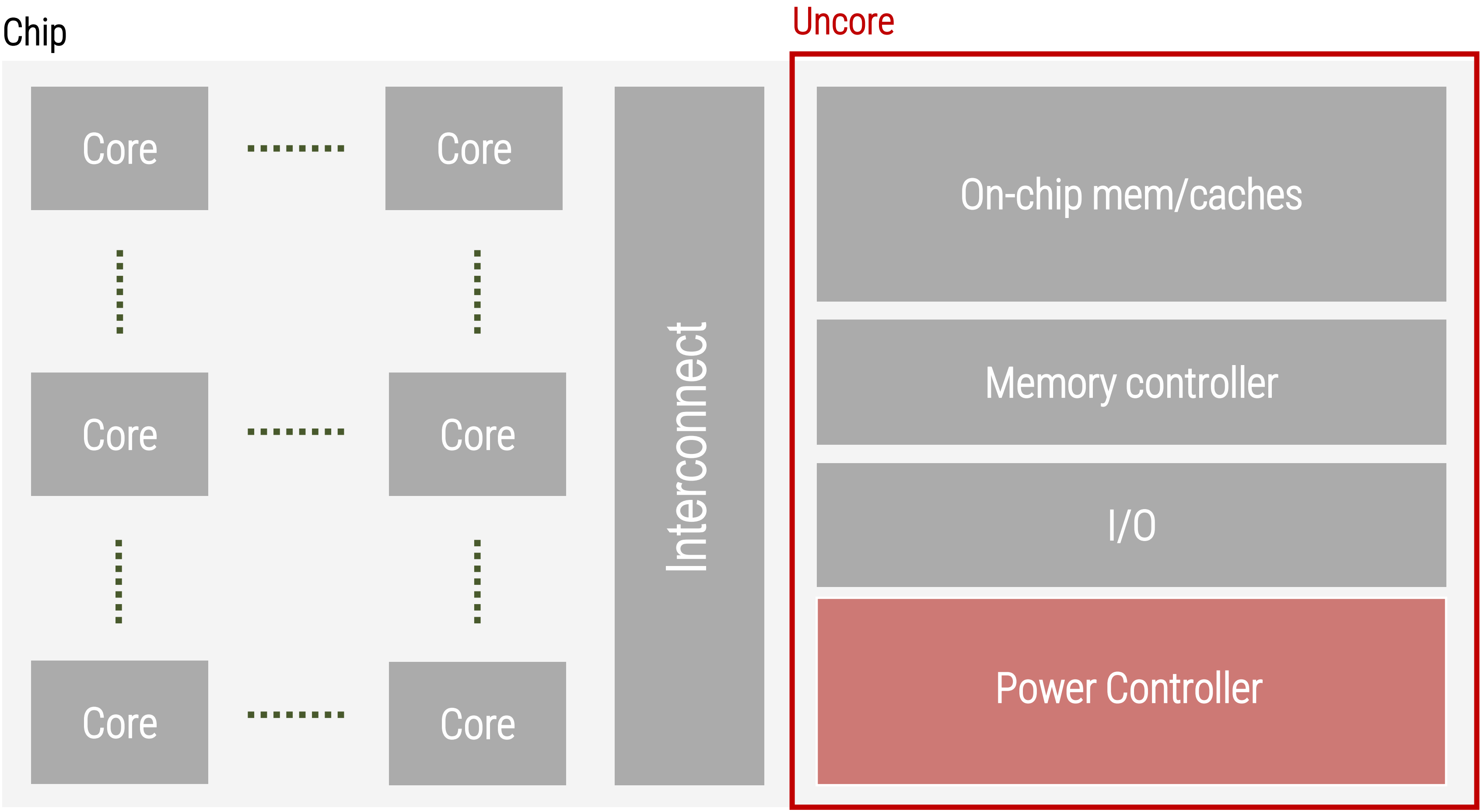
1 G. Bambini et al., "An Open-Source Scalable Thermal and Power Controller for HPC Processors", 2020

- **Existing commercial power controller are built around single-core MCU**
  - Intel PCU, ARM MCP/SCP, AMD SMU, IBM OCC
  - Lack of **scalability** to face the ever-increasing number of controlled cores in single-die and chiplet-based modern chip

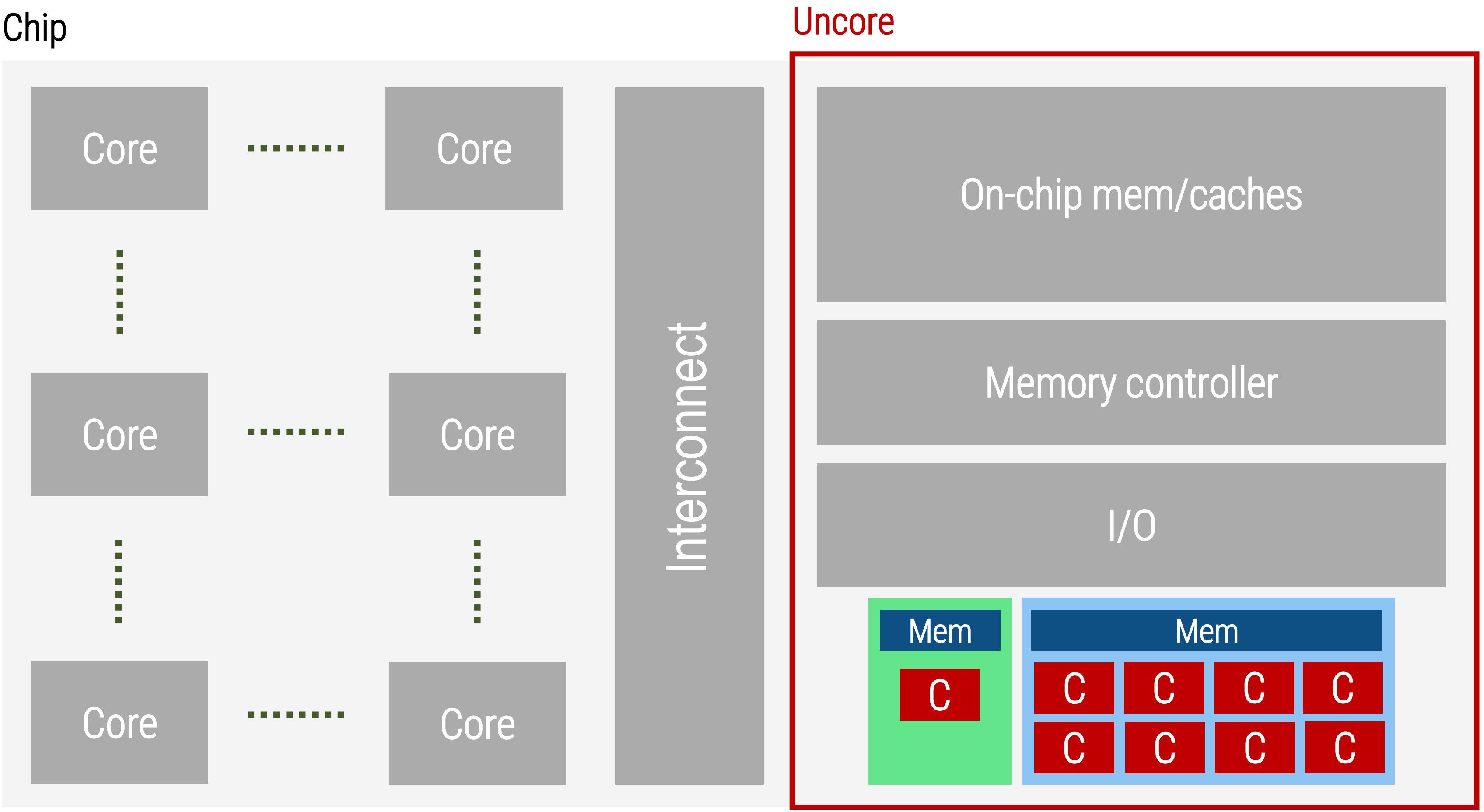
Need for a scalable power controller solution

- **ControlPULP: first RISC-V based, multi-core power controller**
  - Complete HW/SW platform, to be open-sourced soon
  - Implements reactive control
  - Achieves about **5x speedup** in multi-core mode with said policy

Introduction

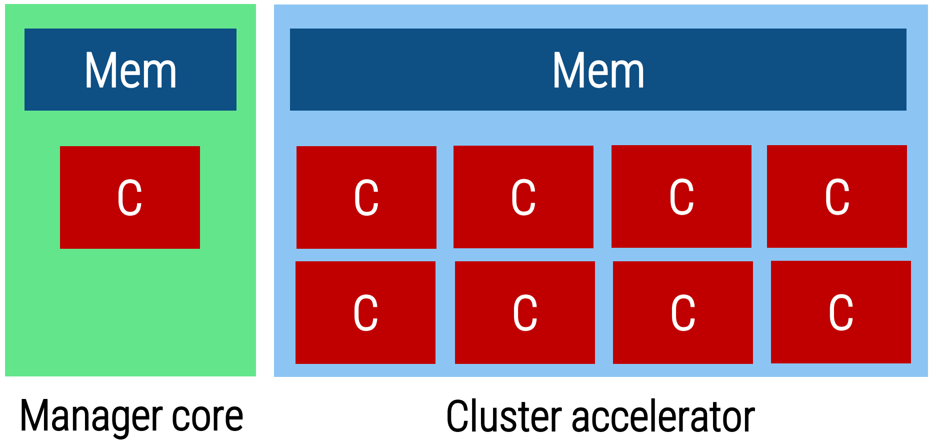


Introduction



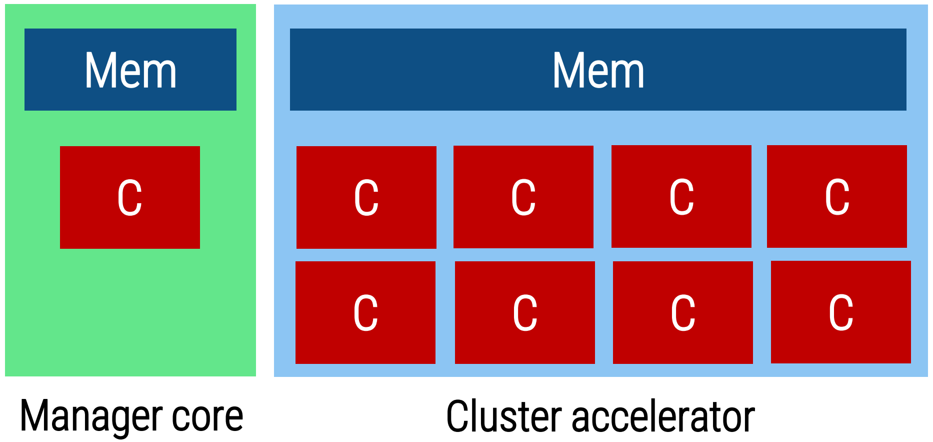
Introduction

- **Scalable architecture:**
  - **Multi-core** cluster with private FPU (float16, bfloat, float32)
  - DMA for 2-D strided access from PVT sensor registers



Introduction

- **Predictable architecture:**
  - Cache less system
  - Constant access time to scratchpad memories
  - Manager core has dedicated banks for data and instructions
  - Platform level interrupt controller (RISC-V PLIC)



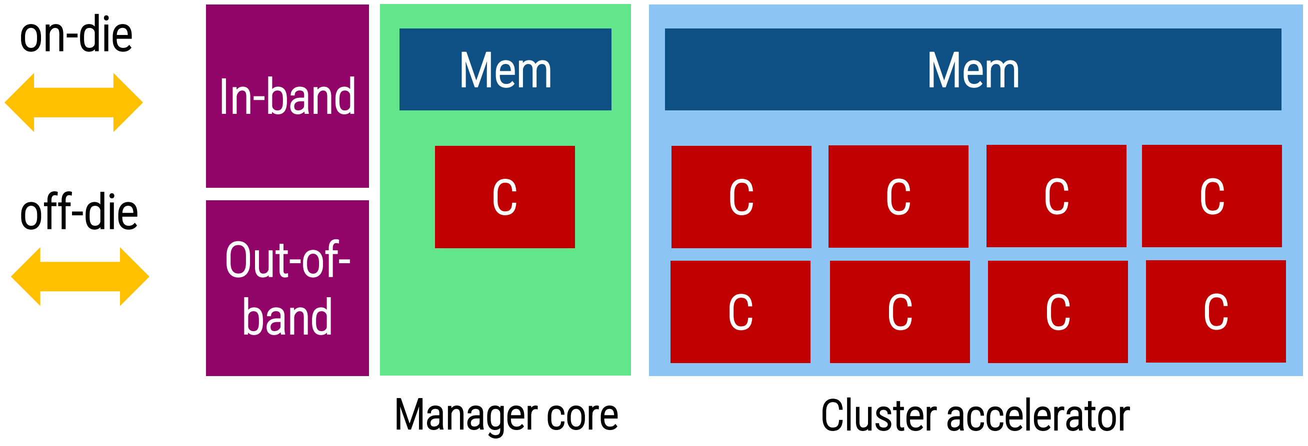
Introduction

- **Industry standard power management interfaces:**
  - PMBUS/AVSBUS: Voltage Regulators control
  - SPI: Inter-socket communication (Multi ControlPULP)
  - ACPI/MCTP: Motherboard/BMC interface (OpenBMC)

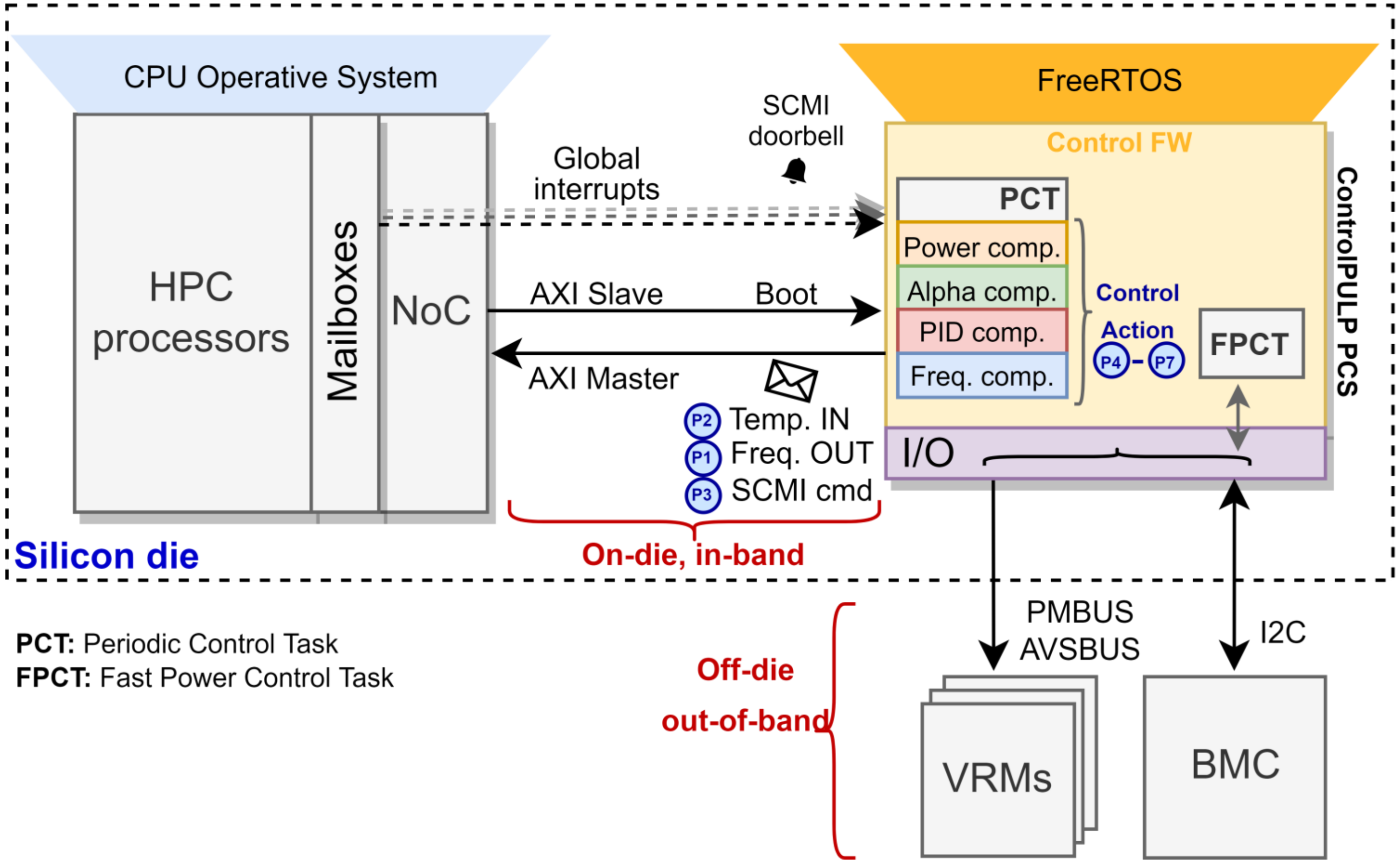
  Out-Of-Band

  - SCMI: OS governors and telemetry

  In-Band



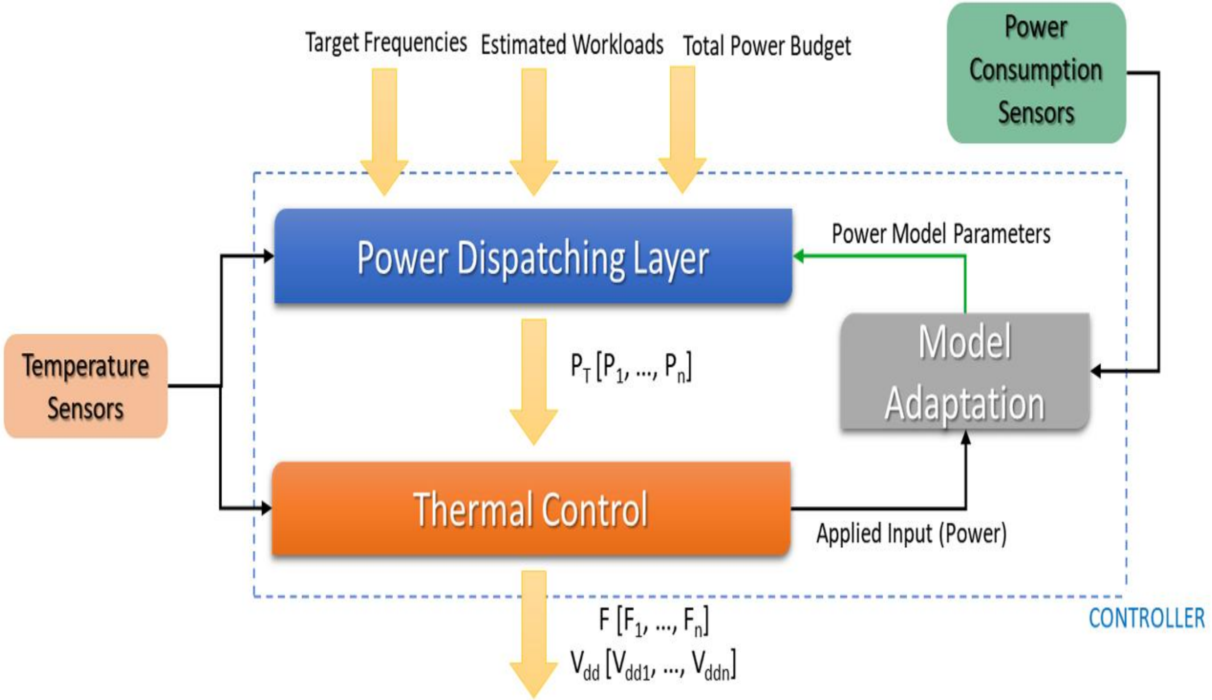
Power Control Policy



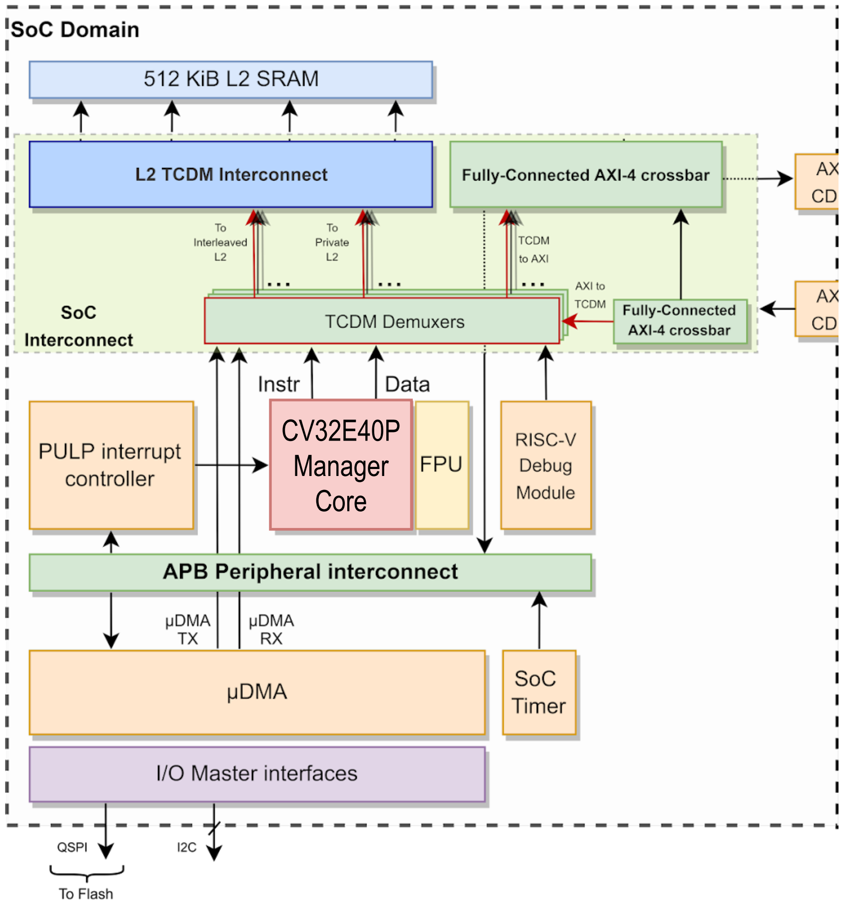
PCT: Periodic Control Task
FPCT: Fast Power Control Task

Power Control Policy

- **Two main control tasks:**
  - **Periodic Control Task**
    - Power dispatching layer
    - Thermal control
  - **Power Control Task**
    - Voltage rails power consumption
    - BMC communication



Architecture



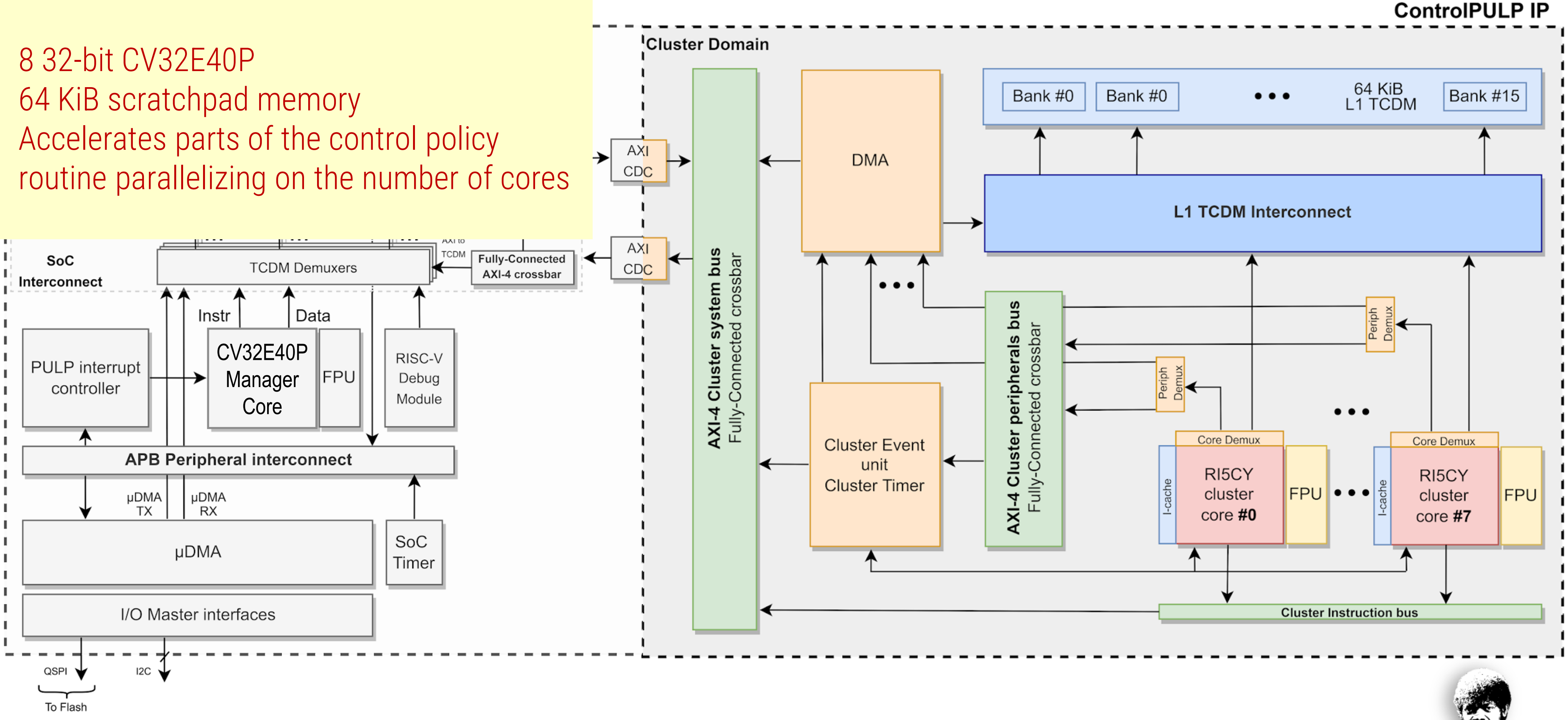
Single-core subsystem

- 32-bit CV32E40P
- 512 KiB scratchpad memory
- Executes the main control policy routine
- Offloads tasks to accelerator cluster

Architecture

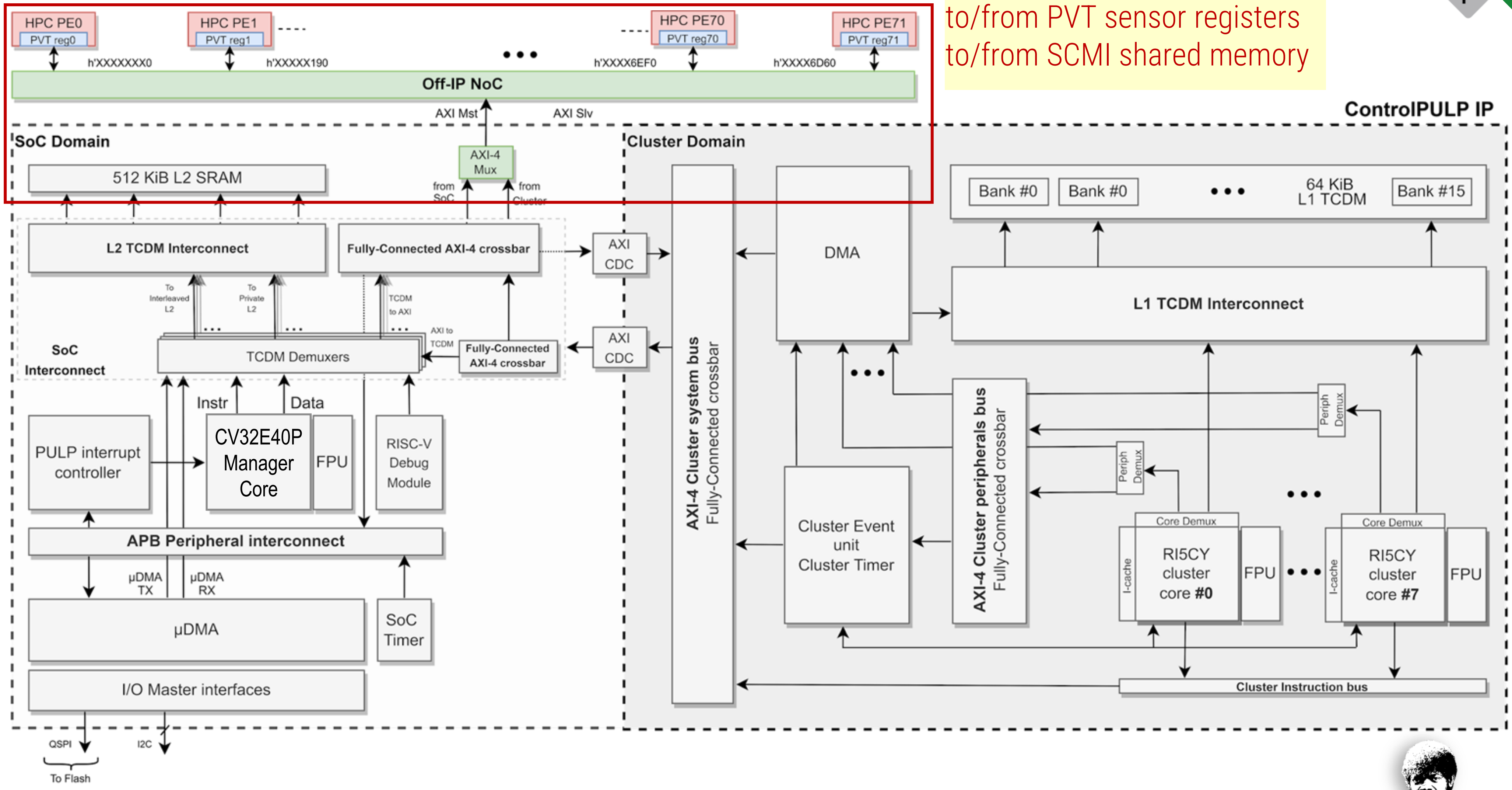
Multi-core cluster subsystem

- 8 32-bit CV32E40P
- 64 KiB scratchpad memory
- Accelerates parts of the control policy routine parallelizing on the number of cores

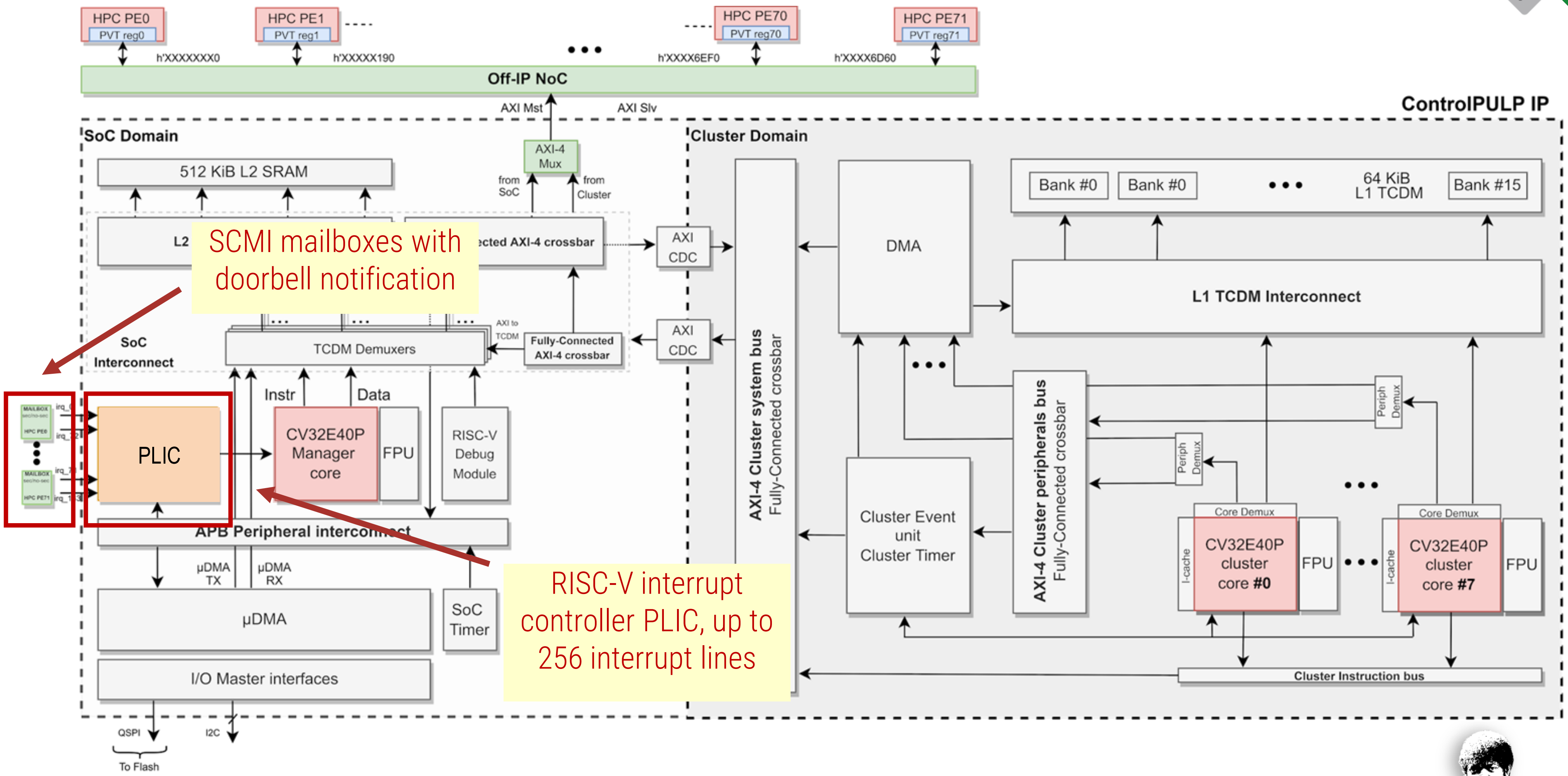


Architecture

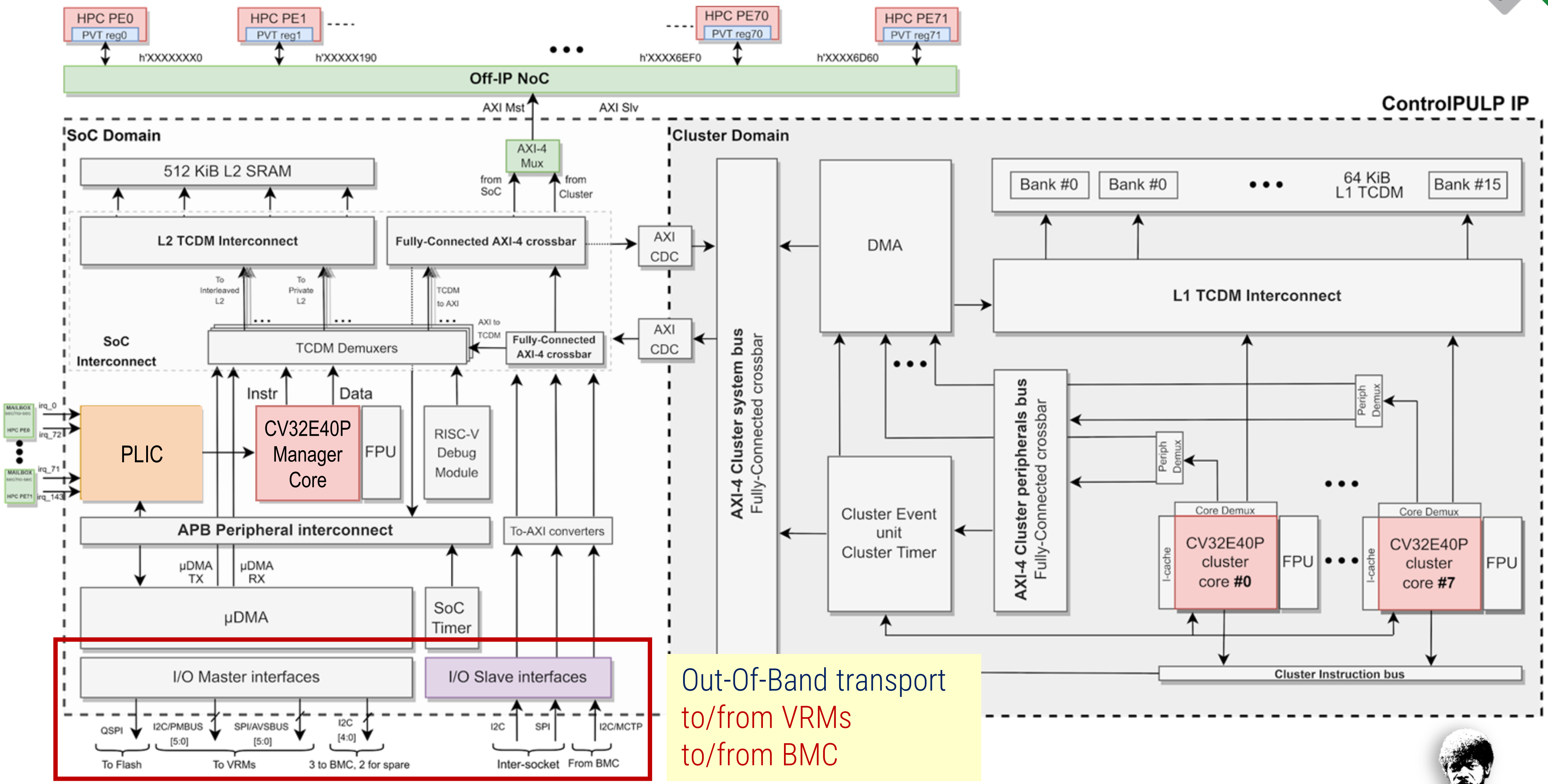
In-Band transport
to/from PVT sensor registers
to/from SCMI shared memory



Architecture



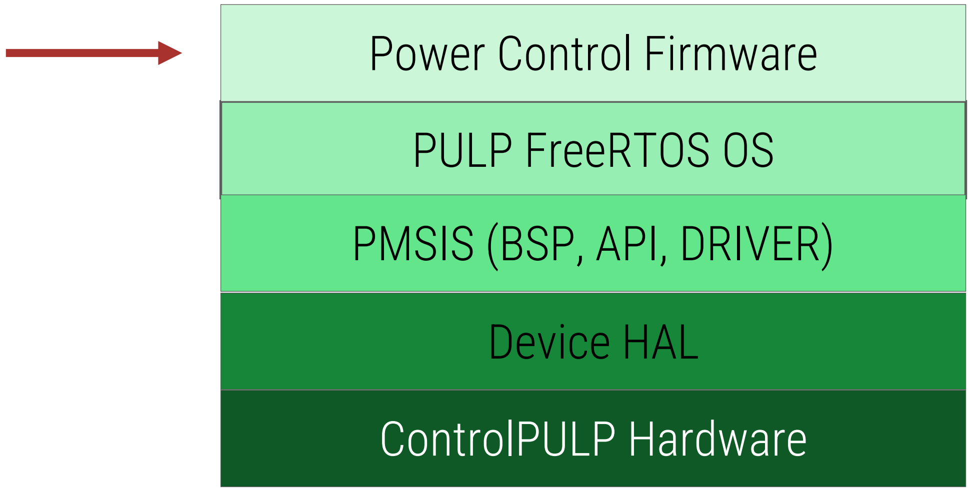
Architecture



Out-Of-Band transport
to/from VRMs
to/from BMC

Software stack

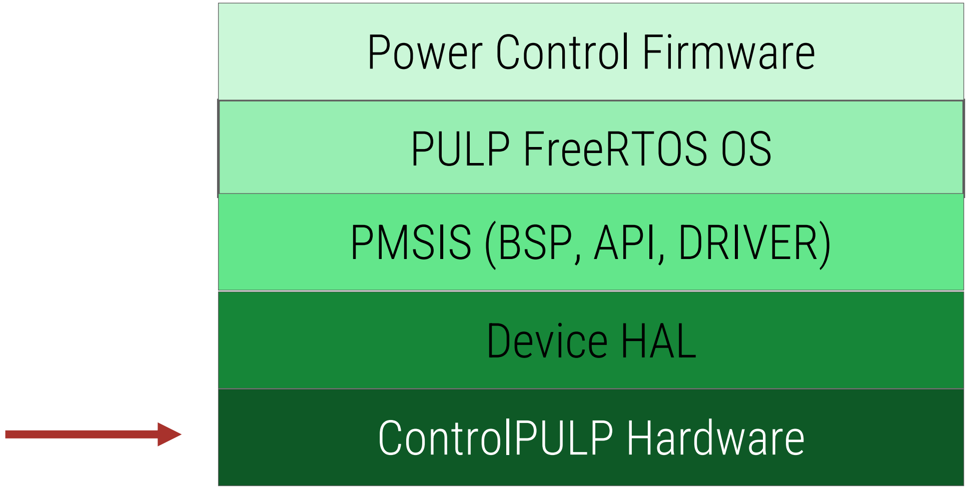
- **Complete software stack**
  - Real-time operation system, FreeRTOS



ControlPULP SW stack

Software stack

- **Complete software stack**
  - Real-time operation system, FreeRTOS



ControlPULP SW stack

Experimental results

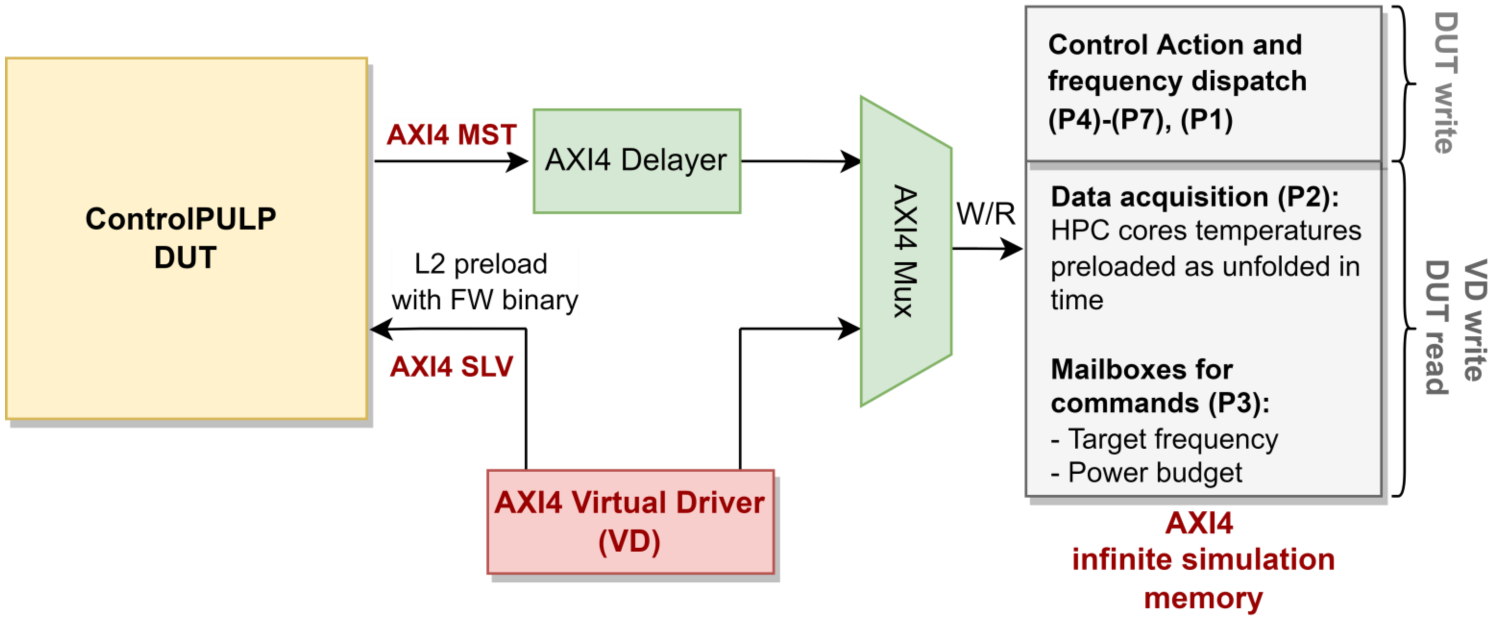
- **Platform area**
  - GF22 synthesis: one manager core, one cluster, 512KiB + 64KiB@500 MHz, TT
  - Total Area of 9.1 MGE
  - Estimated **< 1%** of a HPC processor die in modern technology node

Table 1: ControlPULP post-synthesis area breakdown on GF22FDX technology.

Unit	**Area [mm^2]**	**Area [kGE]**	**Percentage [%]**
Cluster unit	0.467	2336.7	25.5
SoC unit	0.135	675.9	7.39
L1 SRAM	0.119	595.7	6.51
L2 SRAM	1.108	5542.1	60.6
Total	**1.830**	**9150.3**	**100**

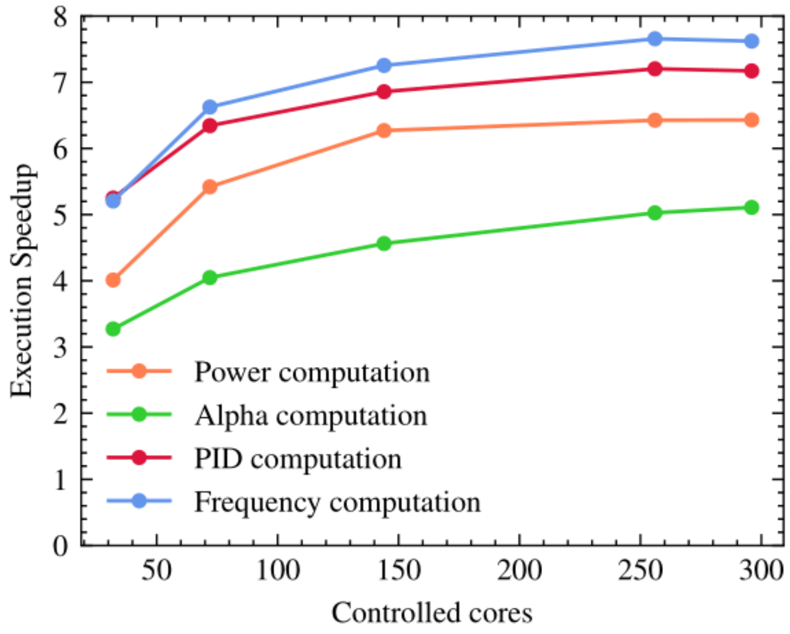
Experimental results

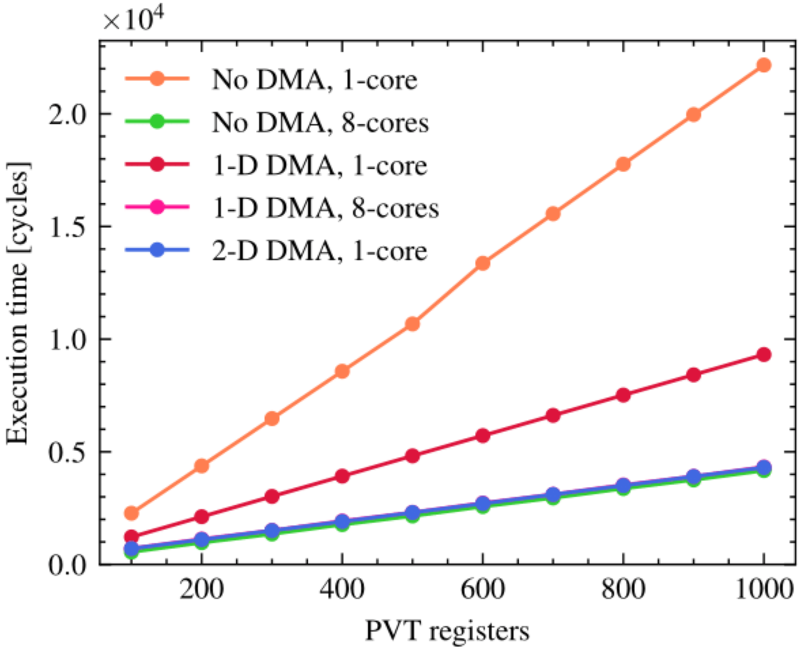
- **Standalone RTL evaluation**
  - NoC latency and controlled system are simplified in a testbench environment
  - **Evaluate:** multi-core speedup (**performance**) and interrupt handling reactiveness (**latency**)



Experimental results

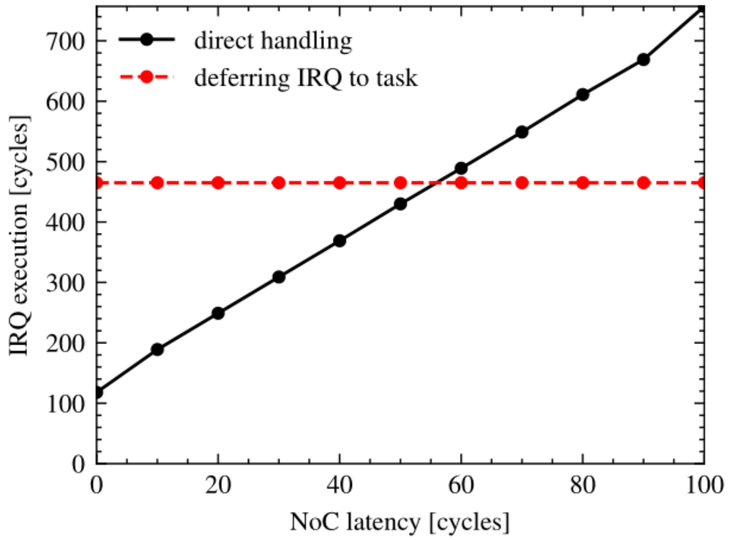
- **Standalone RTL evaluation**
  - Multi-core and DMA centric PCF speedup: about **5x** faster than single-core execution for 72 cores





Experimental results

- **Standalone RTL evaluation**
  - SCMI response deferring mechanism with FreeRTOS
  - Let the OS schedule the response in the interrupt routine (vs. tailored SW in ARM SCP)



Location	Increment [cycles]	Sum [cycles]
PLIC input to output	2	2
CLINT input to core	7	9
Jump in vector table to PLIC handler	2	11
Save caller save regs (addi + 15 regs)	17	28
Claim PLIC interrupt (read id)	8	36
Compute and load PLIC handler address	8	44
Jump to PLIC handler address	2	46
Summary	-	**46**

Conclusions

- First RISC-V power controller for HPC systems
- **Complete** HW/SW platform to be open-sourced
- Scalable multi-core accelerator cluster to achieve 5x on control policy with low area overhead (about 1% of a modern HPC processor)

Future work

- Fast-interrupt handling and predictability exploration
- Explore Advanced predictive control policy with HW acceleration
- Improve verification environment: FPGA-based Hardware-In-The-Loop emulation

Thank you for your attention