RISC-V based Power Management Unit for an HPC processor

Andrea Bartolini  <a.bartolini@unibo.it>
Alessandro Ottaviano  <aottaviano@iis.ee.ethz.ch>
Outline

Power Management in HPC

ControlPULP Hardware and Software Architecture

ControlPULP Validation
HPC Power Management

- Out-of-band – zero overhead telemetry
  - Node Pcap – Max perf @ Pnode<Pmax
  - RAS – error and conditions reporting
    - Based on O.S. metrics
    - Slow & often unused

System Management / RM

- Power Controller
- DIMM
- VRM
- BMC
- RJ45
- PE

Application

- Hints/Prescription
- In band
  - Governors
- Energy vs. Throughput
- Power Cap

Operating System

- System Management / RM
- Node Power Cap
- Out of band

RAS

- Error and conditions reporting

03/05/2022

RISC-V based Power Management Unit for an HPC processor
HPC Power Management

Power Management standard HW/SW interfaces:
In-band:
- The **SCMI** (The System Control and Management Interface) for OS communication.
HPC Power Management

Power Management standard HW/SW interfaces:
In-band:
• The SCMI (The System Control and Management Interface) for OS communication.
Out-of-band:
• PMBus, AVSBus for VRM communication
• MCTP/PLDM for BMC communication
On-Chip Power Controller

- Integrated **Power Controller Subsystem (PCS)** for HPC processors
- RISC-V based & open-* (PULP Platform-based), extended to **support standard power management interfaces**
- To be **integrated within Rhea**, EPI first-generation chip family.

**Design goals:**

- Flexible Power Control Firmware (PCF) => Real-time support in hw/sw (low/predictable interrupt latency, FreeRTOS, …)
- Fine-grain power management w. large core count and high efficiency => multicore design support w. Packed-SIMD FP support.
- Support of large number of on-chip interfaces => Decouple on-chip transfers and computation with DMA-based data movement
Outline

Power Management in HPC

ControlPULP Hardware and Software Architecture

ControlPULP Validation
Architecture

• PULP\textsuperscript{1}-based design

• **Scalable architecture:**
  - Multi-core cluster with private FPU, up to float16 and bfloat precision
  - RISC-V fast-interrupt controller: CLIC
  - DMA for 2-D strided access from PVT sensor registers

• **Industry standard power management interfaces:**
  - **PMBUS:** Voltage Regulators control - slow/multi
  - **AVSBUS:** Voltage Regulators control - fast/p2p
  - **SPI:** Inter-socket communication (Multi ControlPULP)
  - **ACPI/MCTP:** Motherboard/BMC interface (OpenBMC)
  - **SCMI:** OS PM governors and telemetry

\textsuperscript{1} https://github.com/pulp-platform/pulp
Control Firmware

Three main control tasks:\n
   - Control Action: computational block
   - In-Band transfers:
     (i) PVT data gathering- AXI4
     (ii) Doorbell-based SCMI response
   - Out-Of-Band transfers:
     (i) VRMs power consumption – PMBUS/AVSBUS (I2C/SPI)
     (ii) BMC interaction – I2C/MTCP

Control Firmware

Three main control tasks:
1. Periodic Control Task (PCT)
2. Fast Power Control Task (FPCT)
3. Advanced Learning Control Task (ALCT)

- **Control Action**: computational block
- **In-Band** transfers:
  (i) PVT data gathering – AXI4
  (ii) Doorbell-based SCMI response

- **Out-Of-Band** transfers:
  (i) VRMs power consumption – PMBUS/AVSBUS (I2C/SPI)
  (ii) BMC interaction – I2C/MTCP

---

Software stack

Complete software stack relying on a Real-Time operative system, FreeRTOS
Software stack

Complete software stack relying on a Real-Time operative system, FreeRTOS
RISC-V based Power Management Unit for an HPC processor

Architecture
RISC-V based Power Management Unit for an HPC processor

Architecture

SoC Domain
- 512 KIB L2 SRAM
- L2 TCDM Interconnect
- Fully-Connected AXI-4 crossbar
- TCDM Demuxers

Cluster Domain
- AXI CDC
- AXI
cdc

ControlPULP IP
- Bank #0
- Bank #0
- ⋮
- 64 KIB L1 TCDM
- Bank #15

L1 TCDM Interconnect

AXI4 Cluster system bus
- Fully-Connected crossbar

AXI4 Cluster peripherals bus
- Fully-Connected crossbar

Cluster Event unit
- Cluster Timer

µDMA
- TX

µDMA
- RX

SoC Timer

I/O Master interfaces

ETH zurich
RISC-V based Power Management Unit for an HPC processor

Architecture

In-Band transport
to/from PVT sensor registers
to/from SCMI shared memory
RISC-V based Power Management Unit for an HPC processor

Architecture

Industry-standard 32-bit RISC-V core

Industry-standard 32-bit RISC-V core
RISC-V based Power Management Unit for an HPC processor

Architecture

SCMI mailboxes with doorbell notification

RISC-V fast interrupt controller CLIC, up to 256 interrupt lines
RISC-V based Power Management Unit for an HPC processor

Architecture

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

Power Management in HPC

ControlPULP Hardware and Software Architecture

ControlPULP Validation
ControlPULP validation

Standalone RTL validation

- Event-based RTL simulation ecosystem
- GVSOC Architectural simulation ecosystem

### Summary

<table>
<thead>
<tr>
<th>Jobs</th>
<th>Duration</th>
<th>Failed</th>
<th>Errors</th>
<th>Skipped</th>
<th>Passed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>rt_soc_interconnect</td>
<td>92.02s</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>rt_corsmark</td>
<td>1910.33s</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>rt_tcdm</td>
<td>655.33s</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>rt_richram</td>
<td>10923.33s</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>rt_hdi_slu_mq</td>
<td>47.05s</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>rt_avr</td>
<td>50.45s</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>rt_srammem_rv</td>
<td>1548.33s</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

36425.43s

96.99% success rate

Automated Continuous Integration regression check – RTL based
ControlPULP validation

1. Standalone RTL validation
   - GF22 synthesis: 500 MHz, 9.1 MGE
   - Estimated < 1% of a HPC server processor in modern technology node

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

<table>
<thead>
<tr>
<th>Unit</th>
<th>Area [mm²]</th>
<th>Area [kGE]</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster unit</td>
<td>0.467</td>
<td>2336.7</td>
<td>25.5</td>
</tr>
<tr>
<td>SoC unit</td>
<td>0.135</td>
<td>675.9</td>
<td>7.39</td>
</tr>
<tr>
<td>L1 SRAM</td>
<td>0.119</td>
<td>595.7</td>
<td>6.51</td>
</tr>
<tr>
<td>L2 SRAM</td>
<td>1.108</td>
<td>5542.1</td>
<td>60.6</td>
</tr>
<tr>
<td>Total</td>
<td>1.830</td>
<td>9150.3</td>
<td>100</td>
</tr>
</tbody>
</table>
ControlPULP validation

2. FPGA-based Hardware-in-the-Loop emulation

- Cycle-accurate/architectural simulators not suited for
- **Heterogeneous** approach with **FPGA HIL emulation**, based on PULP HERO

[Diagram]

- Zynq UltraScale+ FPGA
- ARM APU x4 ARM A53
- Shared Memory (DRAM)
- ControlPULP
- PS AXI Master
- PL AXI Slave (boot)
- PL AXI Mbox
- Mailboxes (shared registers)
- Zynq Processing System
- Zynq Programmable Logic

5 https://github.com/pulp-platform/pulp/hero
ControlPULP validation

2. FPGA-based Hardware-in-the-Loop emulation

- **Real-Time plant emulation:** TDP budget control over 36-cores
ControlPULP validation

2. FPGA-based Hardware-in-the-Loop emulation

- **EVLPT motherboard from EPI partners:**
  - **Prototype motherboard** for the future Rhea processor
  - VRMs, BMC, Intel FPGA for power sequencing

- **Test off-chip peripherals:**
  1. ACPI power sequencing test ✔
  2. PMBUS test to BMC, VRMs, IBC ✔
  3. I2C Slave (MCTP) test from BMC ✔
  4. AVSBUS test to VRMs control ✔
  5. Inter-socket (Multi ControlPULP) test ✔
  6. More advanced communication ✔ WIP
Conclusion

- First RISC-V Power Controller for current and future HPC processors, based on PULP
- Complete HW/SW codesign and validation platform

Roadmap

- Test chip tapeout in 65 nm to further validate the HW
- Multi-FPGA emulation for inter-socket validation
- More advanced and distributed HW/SW power management
Acknowledgment

The ControlPULP Design Team:

- Giovanni Bambini, Robert Balas Corrado Bonfanti, Antonio Mastrandrea, Davide Rossi, Simone Benatti, Luca Benini

The European project-initiative has received funding from the European High Performance Computing Joint Undertaking (JU) under Framework Partnership Agreement No 800928 and Specific Grant Agreement No 101036168 (EPI SGA2). The JU receives support from the European Union’s Horizon 2020 research and innovation programme and from Croatia, France, Germany, Greece, Italy, Netherlands, Portugal, Spain, Sweden, and Switzerland.

The European PILOT project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No.101034126. The JU receives support from the European Union’s Horizon 2020 research and innovation programme and Spain, Italy, Switzerland, Germany, France, Greece, Sweden, Croatia and Turkey.

This REGALE-project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 956560. The JU receives support from the European Union’s Horizon 2020 research and innovation programme and Greece, Germany, France, Spain, Austria, Italy.

http://pulp-platform.org  @pulp_platform
ControlPULP validation

- Event-based RTL simulation ecosystem
- GVSOC Architectural simulation ecosystem
- Multi-core and DMA centric PCF speedup: 5x than single-core execution

4 N. Bruschi et al., "GVSoC: A Highly Configurable, Fast and Accurate Full-Platform Simulator for RISC-V based IoT Processors", 2021