

# DESIGN OF AN OPEN-SOURCE FULL-STACK EMBEDDED PLATFORM FOR POWER MANAGEMENT IN HPC, THE EUROPEAN PROCESSOR INITIATIVE APPROACH

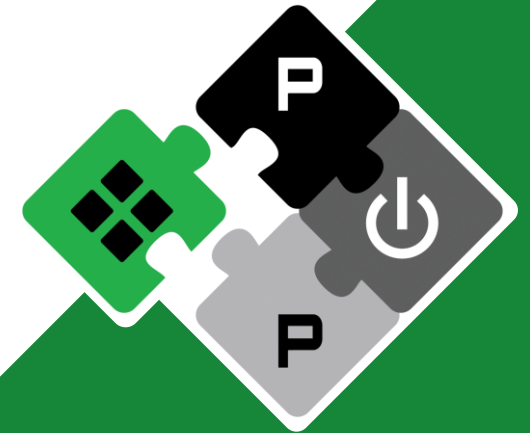
ETHZ, Zürich, Switzerland

University of Bologna, Bologna, Italy

*Alessandro Ottaviano, Robert Balas, Giovanni Bambini, Antonio Delvecchio,  
Davide Rossi, Luca Benini, Andrea Bartolini*

## **PULP Platform**

Open Source Hardware, the way it should be!



@pulp\_platform 

pulp-platform.org 

youtube.com/pulp\_platform 

# OUTLINE

- Power Management in HPC
- ControlPulp Project: an open-source hardware/software RISC-V controller
- European Processor Initiative (EPI) Case Study
- QnA

# POWER MANAGEMENT IN HPC: INTRO

- Gold Rush toward **High Performance Computing**
  - Market and Industry needs
  - Politics interest
- **RISC-V HPC processors** development is approaching at fast pace
  - Examples: [Monte Cimone](#), [Ventana Veyron V1](#), [SiFive P600](#)
- [EuroHPC](#)



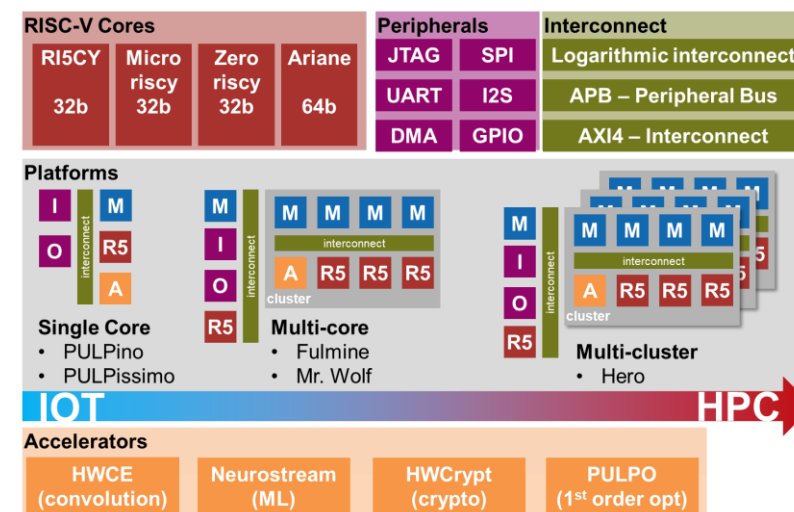
# POWER MANAGEMENT IN HPC: INTRO



- **EPI project:** design of a “full-European” HPC processor
- Chips manufactured by **SiPearl**
  - RHEA 1 -> Silicon **Tapeout** second quarter '23
  - RHEA 2 -> in the design phase
- Many-core ARM Zeus + EPAC accelerator
- Chiplet design, HBM on-chip, DDR5, ...
- High-density Blades

# POWER MANAGEMENT IN HPC: EPI APPROCH

- Implementation of an **open-hardware** Power Controller inside RHEA
  - In addition to the ARM proprietary Power Management controller (SCP / MCP)
- **ControlPulp** project
  - PULP<sup>1</sup>-based design
  - Scalable architecture:
    - Multi-core cluster with private FPU
    - DMA for 2-D strided access from PVT sensor registers
  - Industry standard power management interfaces:
    - PMBUS, AVSBUS, SPI, ACPI/MCTP, SCMI

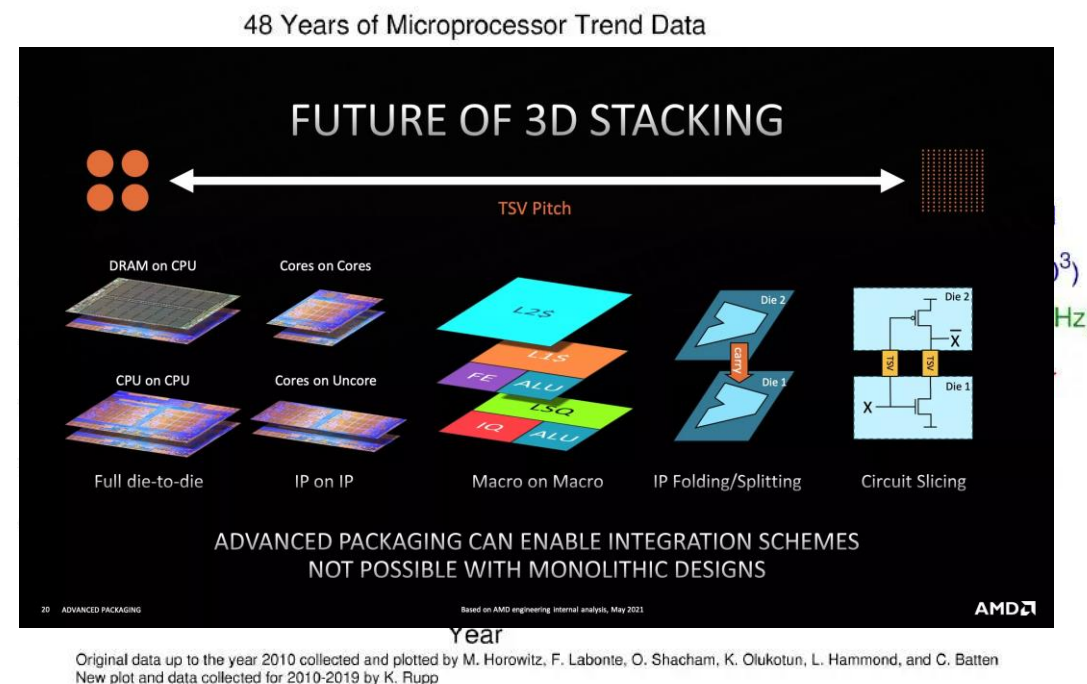


<sup>1</sup> <https://github.com/pulp-platform/pulp>

# POWER MANAGEMENT IN HPC: CHALLENGES

## Recent Design Challenges:

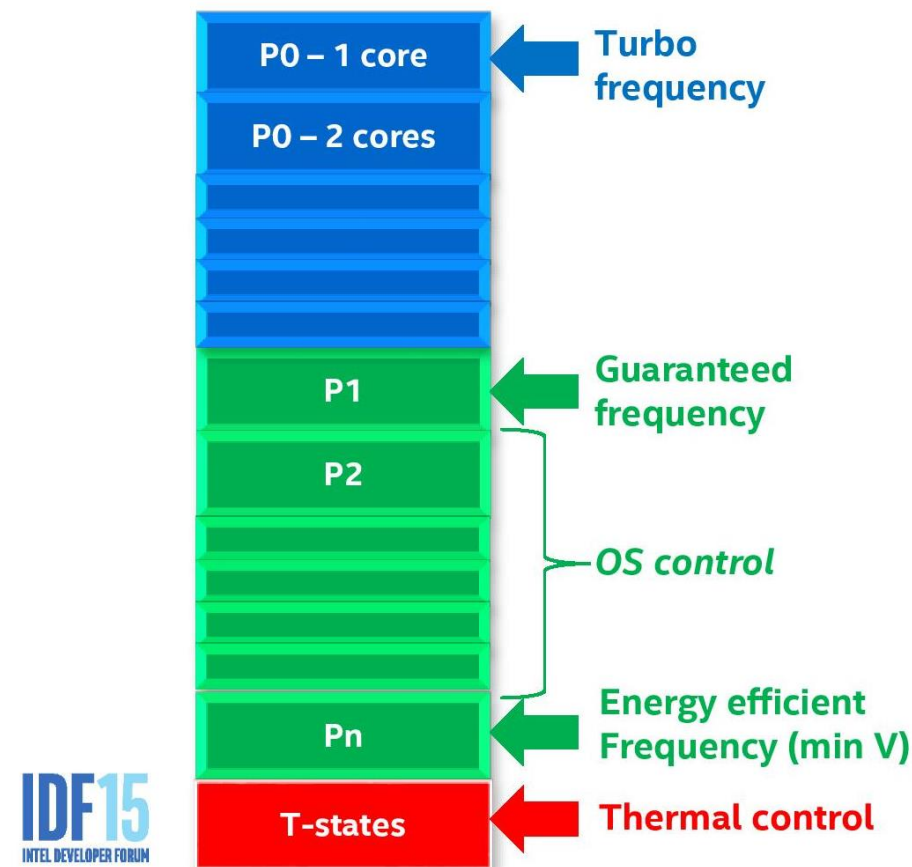
- End of Dennard's scaling law (slow down in transistors scaling)
- Power and Thermal challenges of modern **Multi-core** and **Many-core** designs
- **Heterogeneous** computing units on a single die
  - EPI GPP + EPAC + HBM
- 3D chips and chiplet designs



# POWER MANAGEMENT IN HPC: SPECIFICATION

## CONTROL SCOPE

- Control all **Processing Elements (PE)** of the Processor
- Control **Power** distribution
- Control **Voltage** and **Frequency** parameters
- Boot up and Shut down
- **Telemetry**
- Error Report



# POWER MANAGEMENT IN HPC: SPECIFICATIONS

## POWER CAPPING

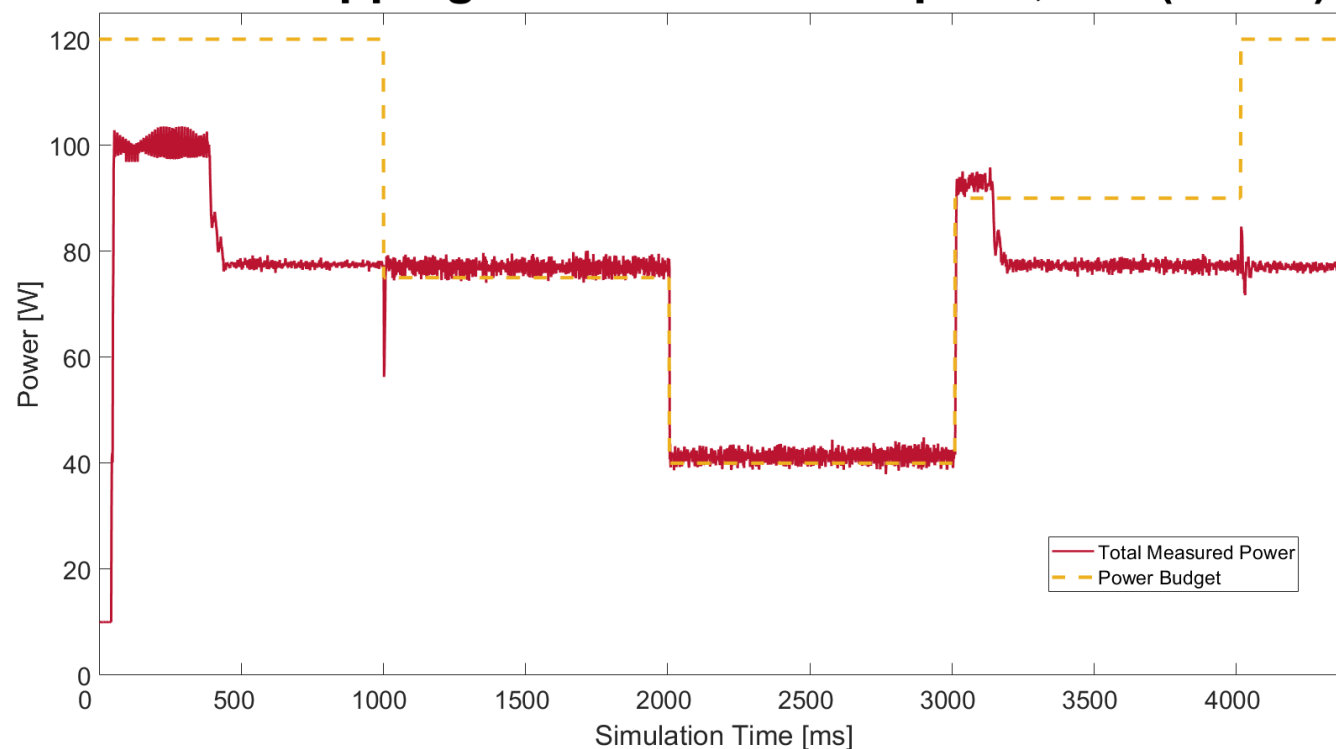
- Several power consumption setpoints (**power budget**) are given over time from the BMC/OS
- Power in the chip varies not only with DVFS, but also with **executed workload** and **temperature** of the chip
- Control timings on the **100us** scale
- Avoid Power spikes and high-amplitude oscillations



# POWER MANAGEMENT IN HPC: SPECIFICATIONS

## POWER CAPPING - EXAMPLE

### Power capping on the controlled plant, HIL (FPGA)



# POWER MANAGEMENT IN HPC: SPECIFICATIONS

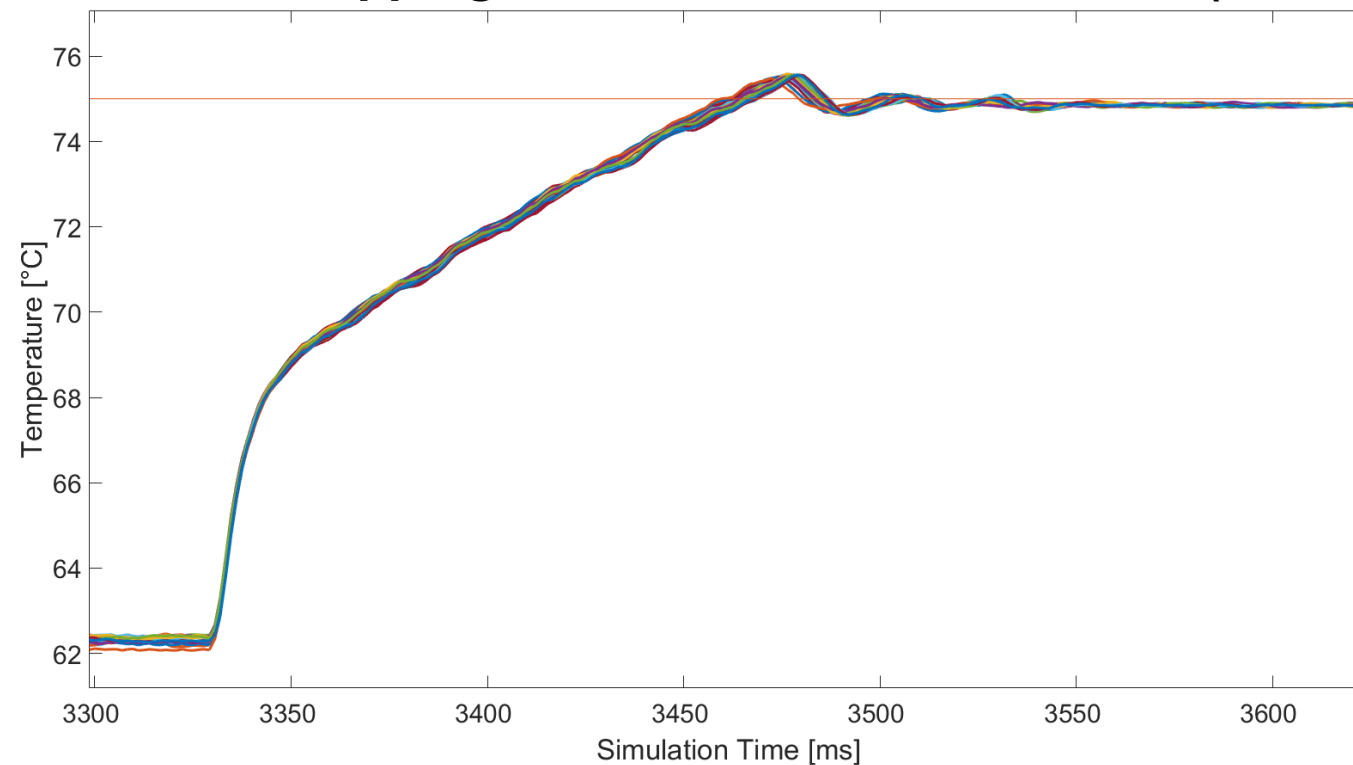
## THERMAL CAPPING

- A **thermal setpoint** (variable, from BMC or OS) is given to avoid chip damage and reduce chip aging
- Temperature is read from **PVT sensors**
  - They are noisy
- Temperature varies with several time constants. Faster ones are around 500us
- Several parts to be controlled (many-cores architecture, accelerators, on-chip memory, VRMs, ...)

# POWER MANAGEMENT IN HPC: SPECIFICATIONS

## THERMAL CAPPING - EXAMPLE

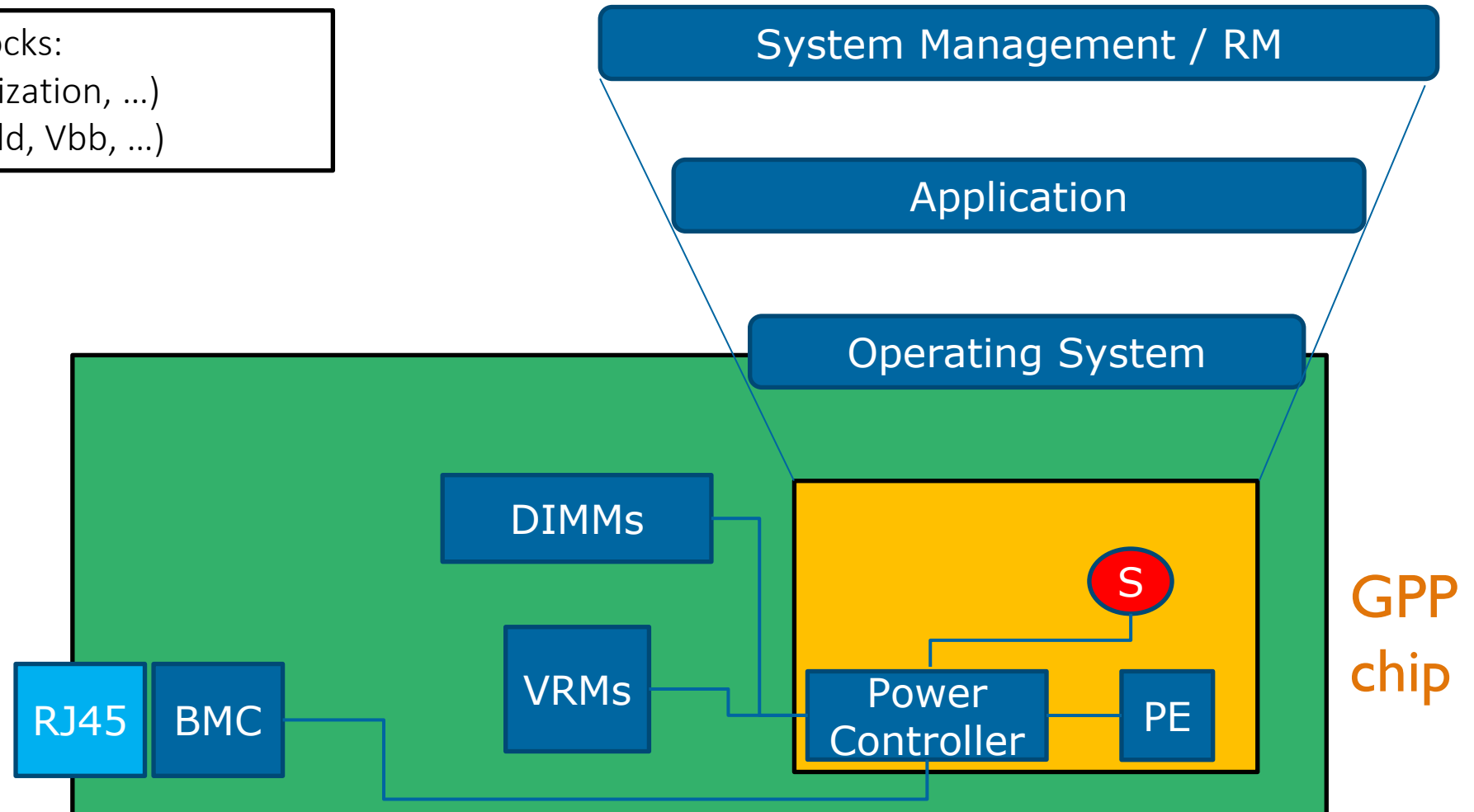
### Thermal Capping on the Controlled Plant, HIL (FPGA)



# POWER MANAGEMENT IN HPC: INTERFACES

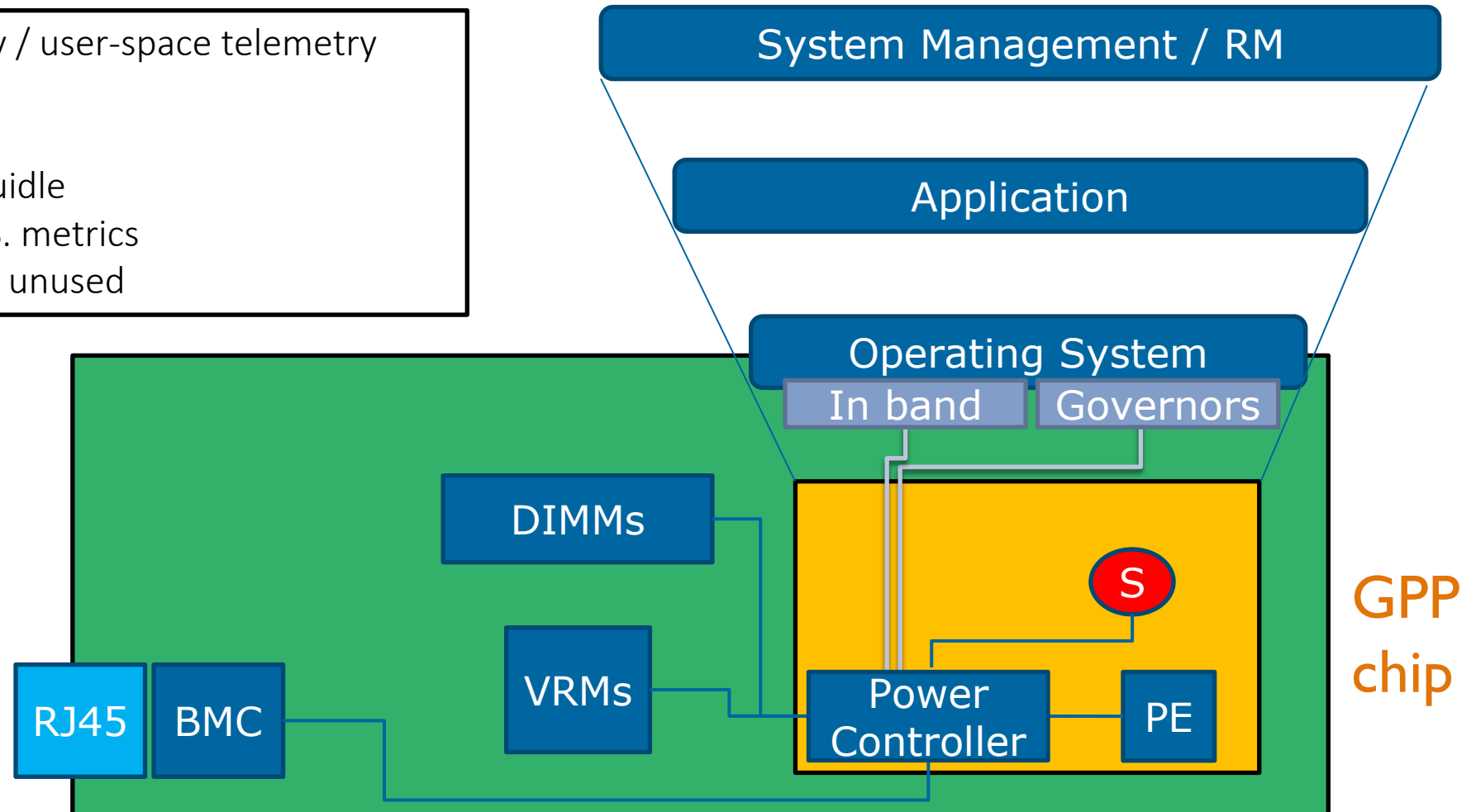
Main Architecture Blocks:

- Sensors (PVT, Utilization, ...)
- Controls (Freq, Vdd, Vbb, ...)



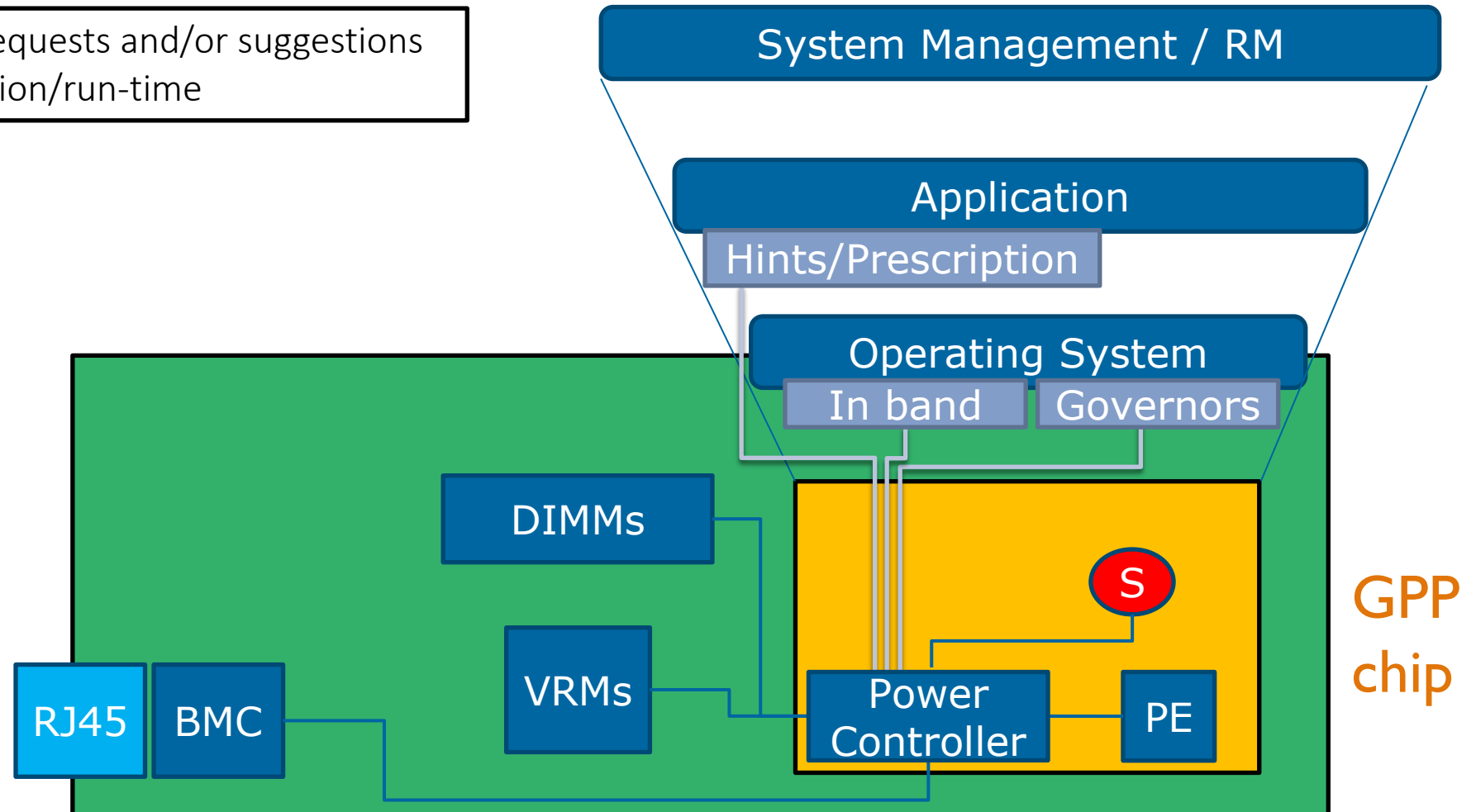
# POWER MANAGEMENT IN HPC: INTERFACES

- In band: low latency / user-space telemetry (power, perf, ...)
- O.S. PM governors:
  - cpufreq / cpuidle
  - Based on O.S. metrics
  - Slow & often unused



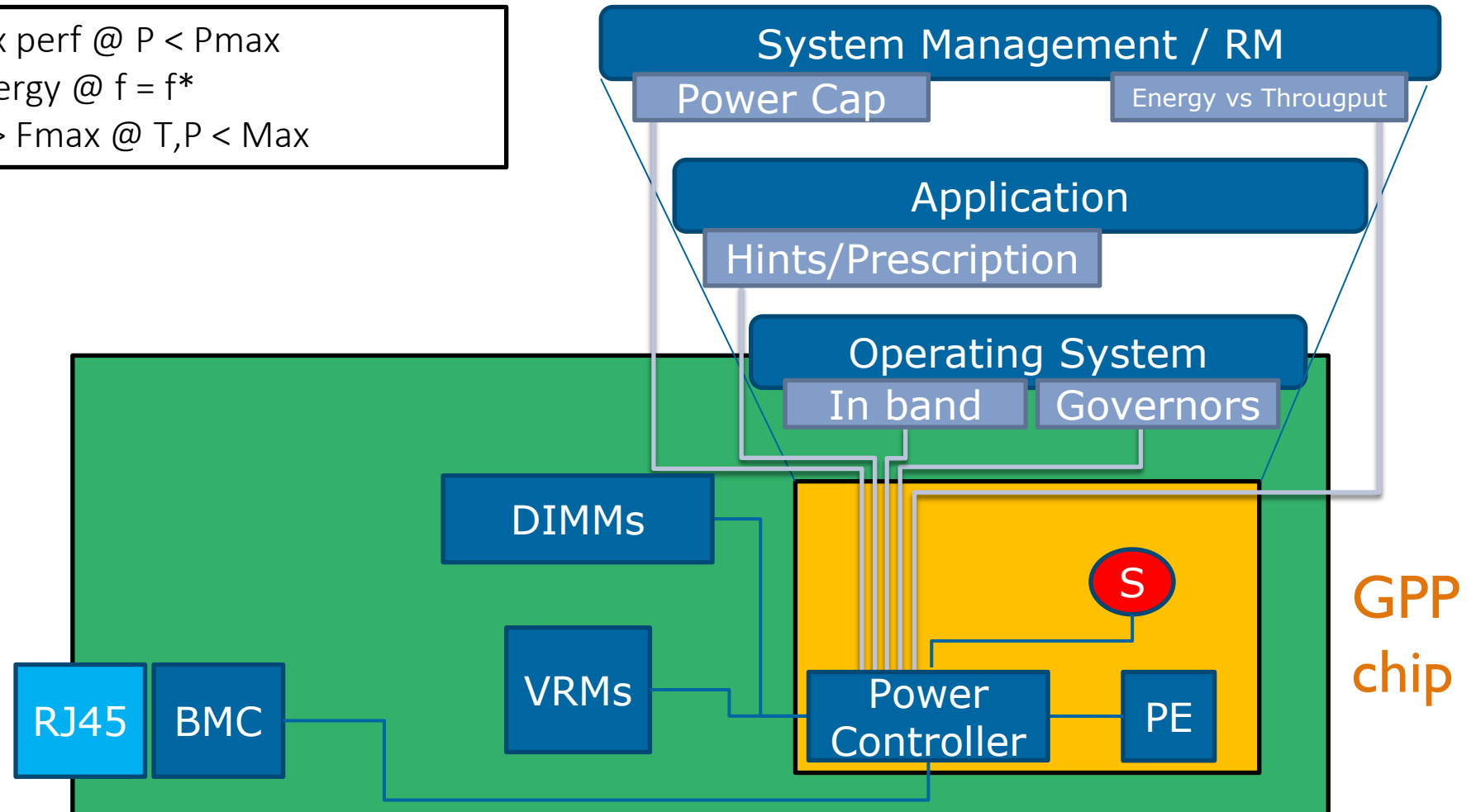
# POWER MANAGEMENT IN HPC: INTERFACES

- Low latency PM requests and/or suggestions
- From the Application/run-time

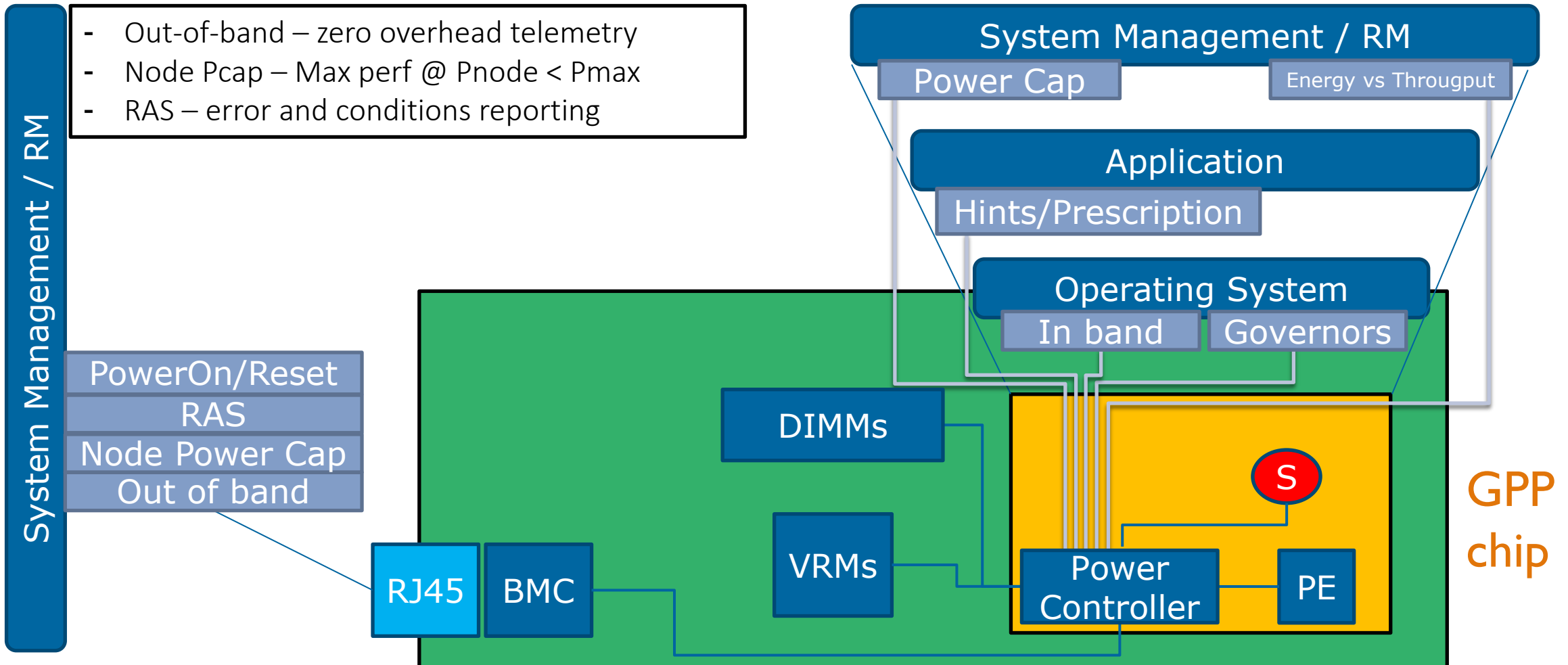


# POWER MANAGEMENT IN HPC: INTERFACES

- Power cap => Max perf @  $P < P_{max}$
- Energy => Min Energy @  $f = f^*$
- Throughput =>  $F > F_{max}$  @  $T, P < Max$



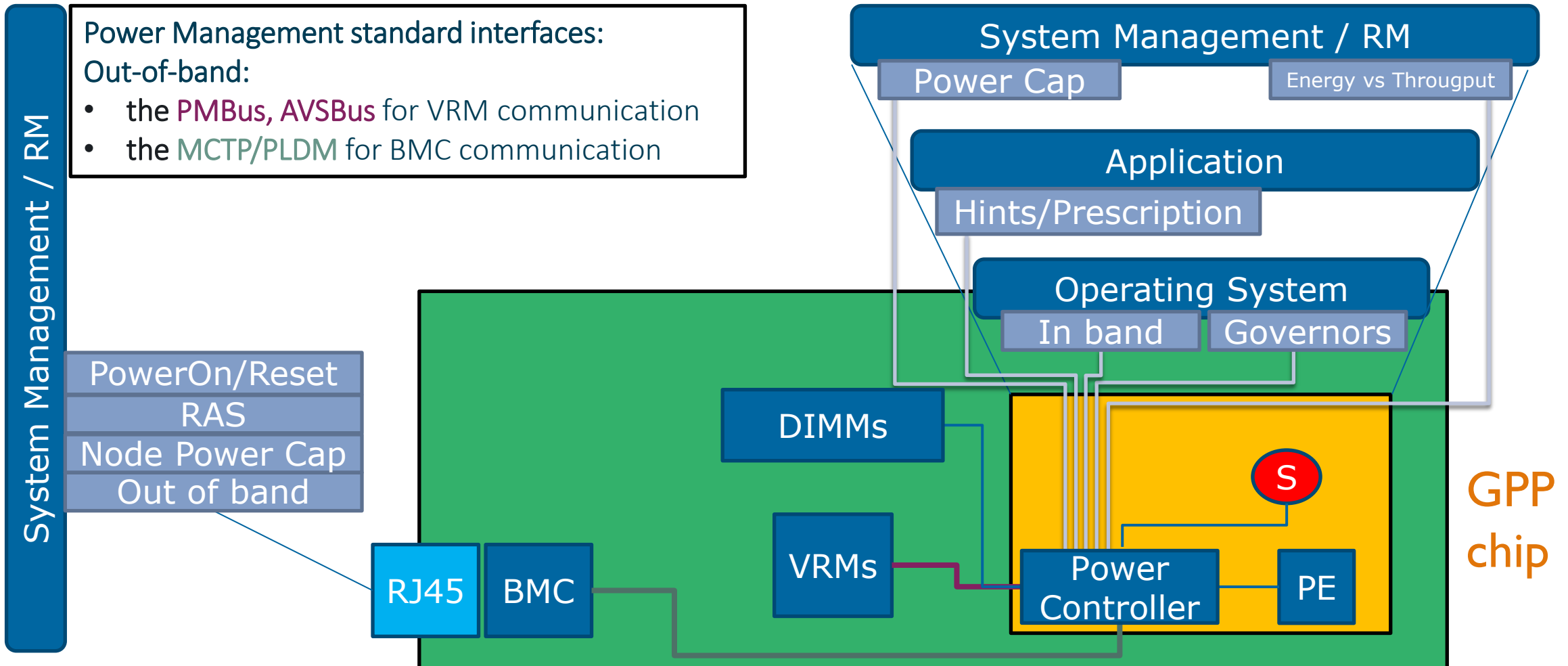
# POWER MANAGEMENT IN HPC: INTERFACES







# POWER MANAGEMENT IN HPC: INTERFACES



# OUTLINE

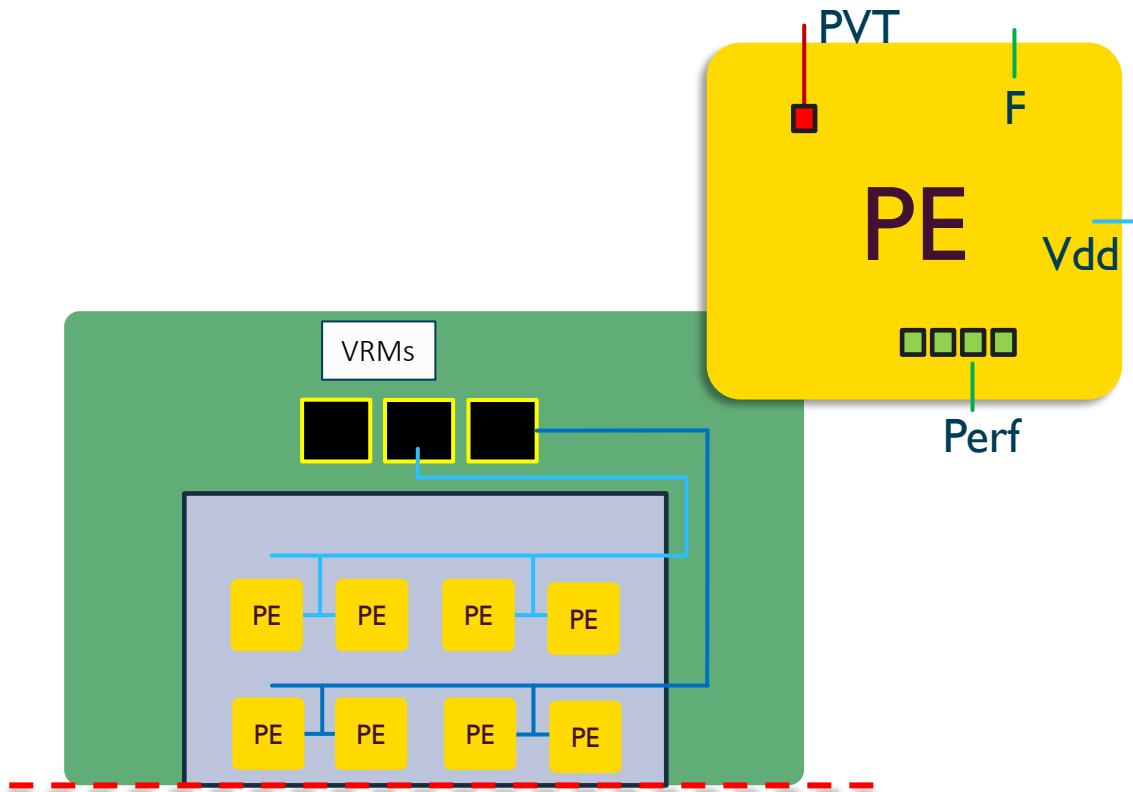
- Power Management in HPC
- ControlPulp Project: an open-source hardware/software RISC-V controller
- European Processor Initiative (EPI) Case Study
- QnA

# CONTROLPULP CONTROL: OVERVIEW

- Currently: **Reactive Control Policy**
- Based on an **RTOS** firmware
- Power and Thermal Capping
- Apply **DVFS**
- Divided into 2 Periodic Tasks to accomodate for the temporal differences:
  - **Periodic Frequency Control task:** 2kHz, reads temperatures, computes DVFS, applies frequency on a **per-PE basis**
  - **Periodic Voltage Control task:** 8kHz, reads power rails consumption, applies voltage to **groups of PEs**

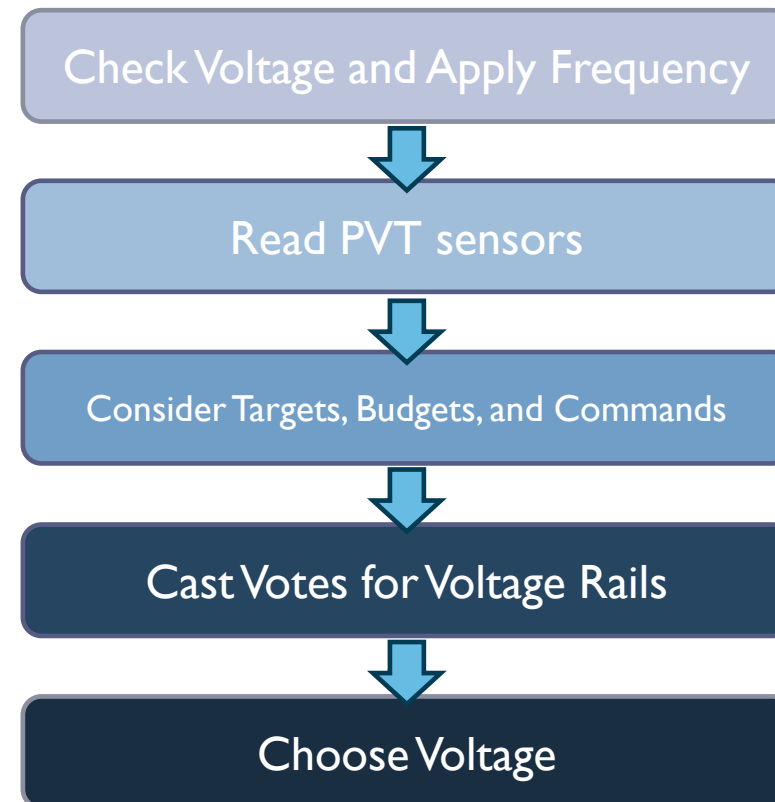
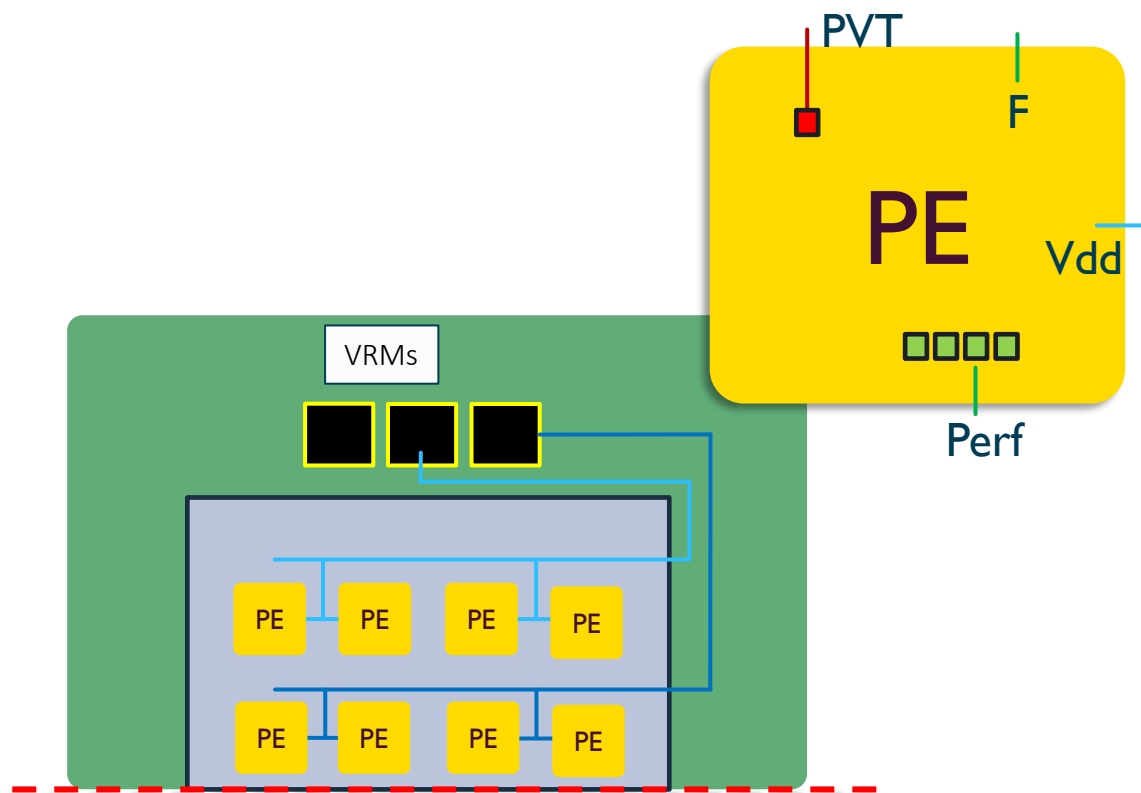


# CONTROLPULP CONTROL: SYSTEM OVERVIEW

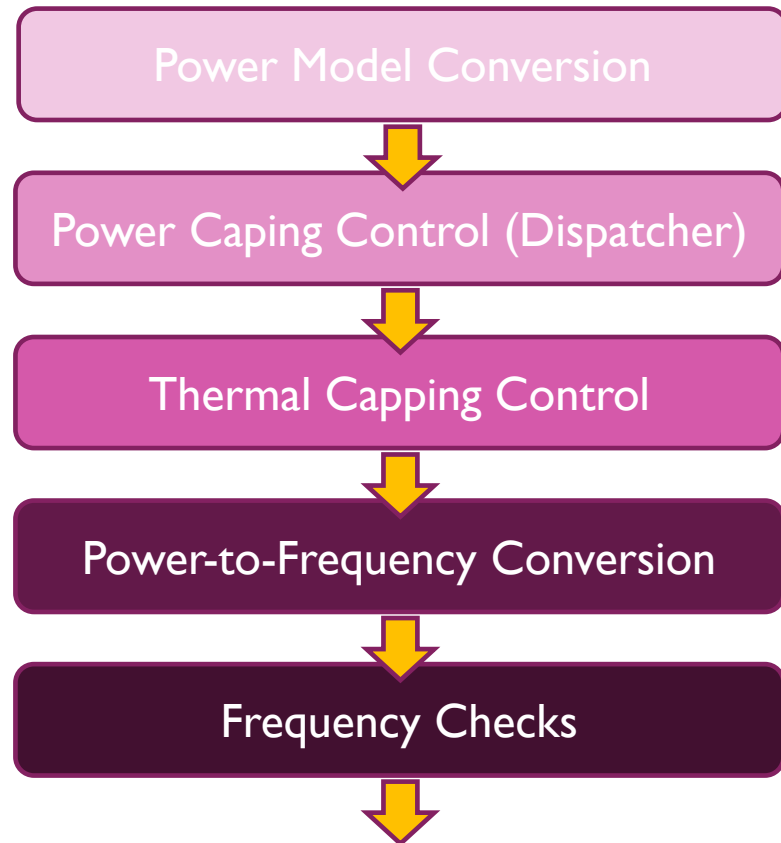


- Voltage rails to **Groups** of cores (PE)
- Power increases **quadratically** with Voltage (Vdd)
- Voltage has discrete control steps
- Each PE has one (or more) PVT sensors, used to measure **Temperature**, Voltage Drops, and Silicon Quality
- (If implemented), they have **executed workload** information through Performance Counters

# CONTROLPULP CONTROL: PART 1

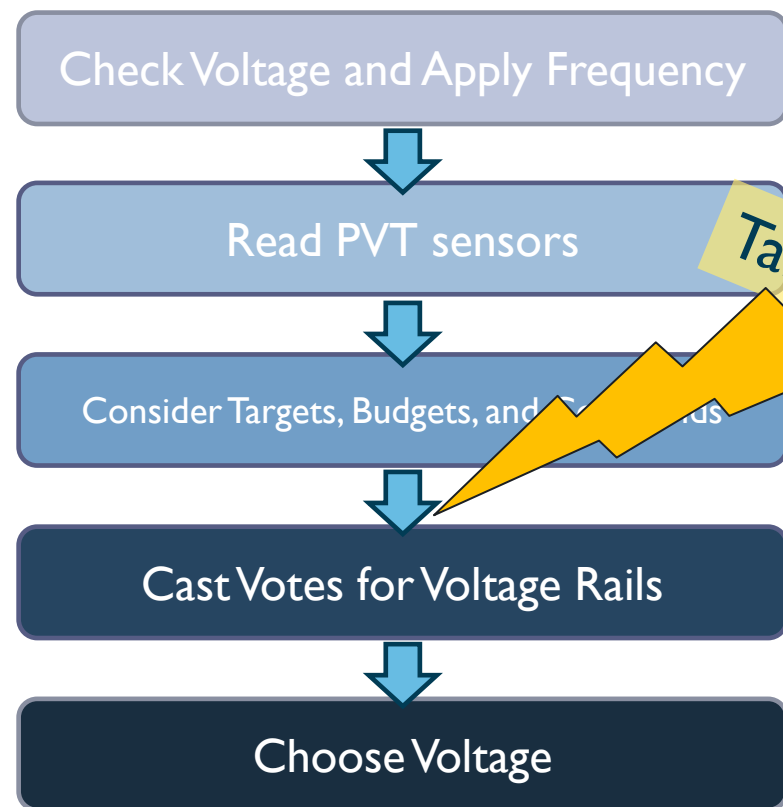


## CONTROLPULP CONTROL: PART 2

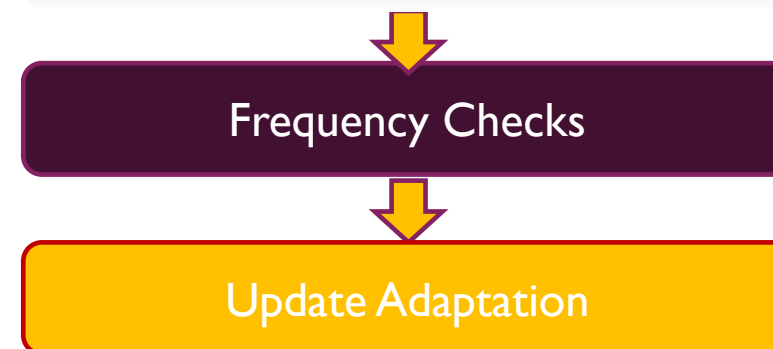


- Power metric is **Linear** and more easily controllable
  - Intel/IBM/AMD control in Frequency (non-linear) using linear control techniques --> not cool
- **Model accuracy** needs sophistication
- Cascade vs Voting-box control
- **Heuristic Approach:** Voltage is fixed in the Power-to-Frequency conversion
  - Because applicable Freq depends on Vdd
  - It can be refined in the Periodic Voltage Control Task

# CONTROLPULP CONTROL: ADAPTATION IMPROVEMET



- Exponential Moving Average Adaptation
- A patch to the sub-optimal Voltage choice
  - Solved through iteration
  - Still sub-optimal due to the Delay
- Exponential Moving Average Control is widely used in the Power Controller
  - Intel, AMD, IBM, ...





## CONTROLPULP CONTROL: NEXT STEPS

- Power-to-Frequency conversion performed with an **iterative-solving algorithm** to convert Power to a **couple** of Frequency-Voltage
  - Voltage and Frequency are reduced **together**
- Safe / Stochastic **MPC**
  - It enforces limits (power and thermal capping) automatically
  - **Optimal**
  - Relies on Model accuracy
  - **Computation** and **Memory** concerns
    - ControlPulp has a **multi-cluster** design
    - Can add MPC **hardware accelerators**
    - Can remap control into **ML** algorithms

# OUTLINE

- Power Management in HPC
- ControlPulp Project: an open-source hardware/software RISC-V controller
- **European Processor Initiative (EPI) Case Study**
- QnA

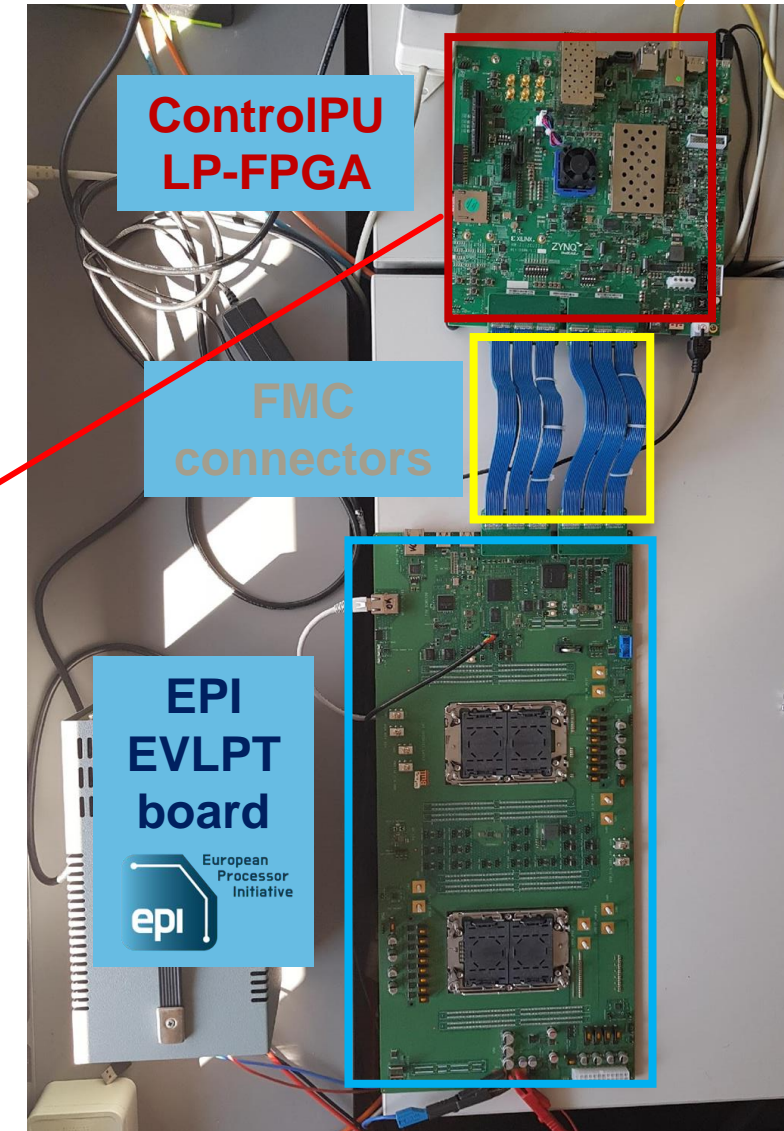
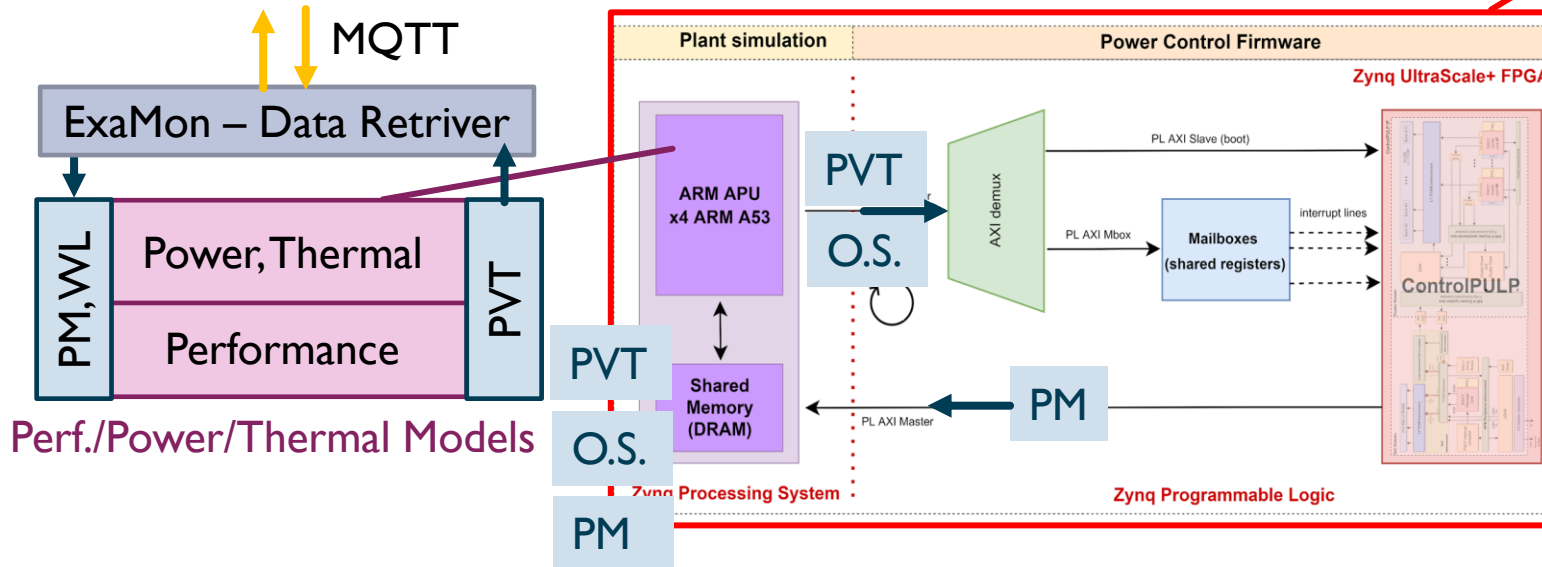
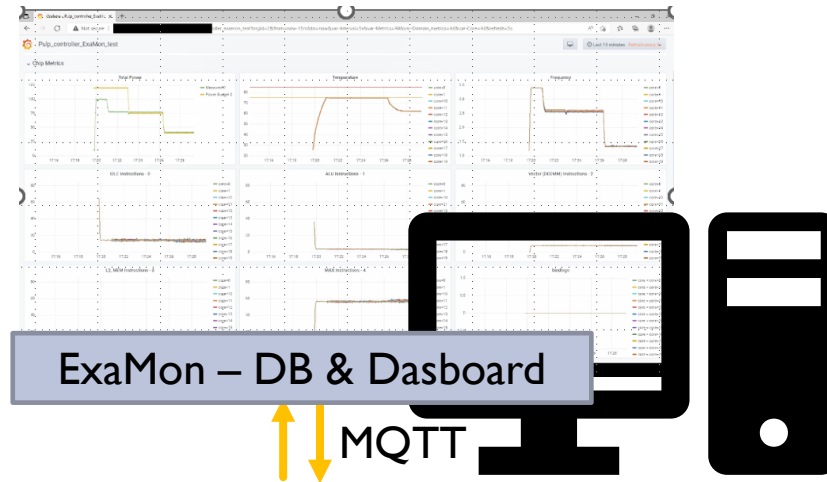
# EPI CASE STUDY: CHALLENGES

## 1) Absence of a simulation/platform for development and testing HW and SW:

- Time step ( $\sim 1\mu s$ ) for models
- Fast Simulation, agnostic of cycle and memory accuracy
- Comprehensive of all the I/Os, Control, Commands, and Communications
- Correctly simulating all interactions among parts and models
- Flexibility in processor/system configurations, with new models and parameters
- Storing and Plotting data
- Hardware in Development

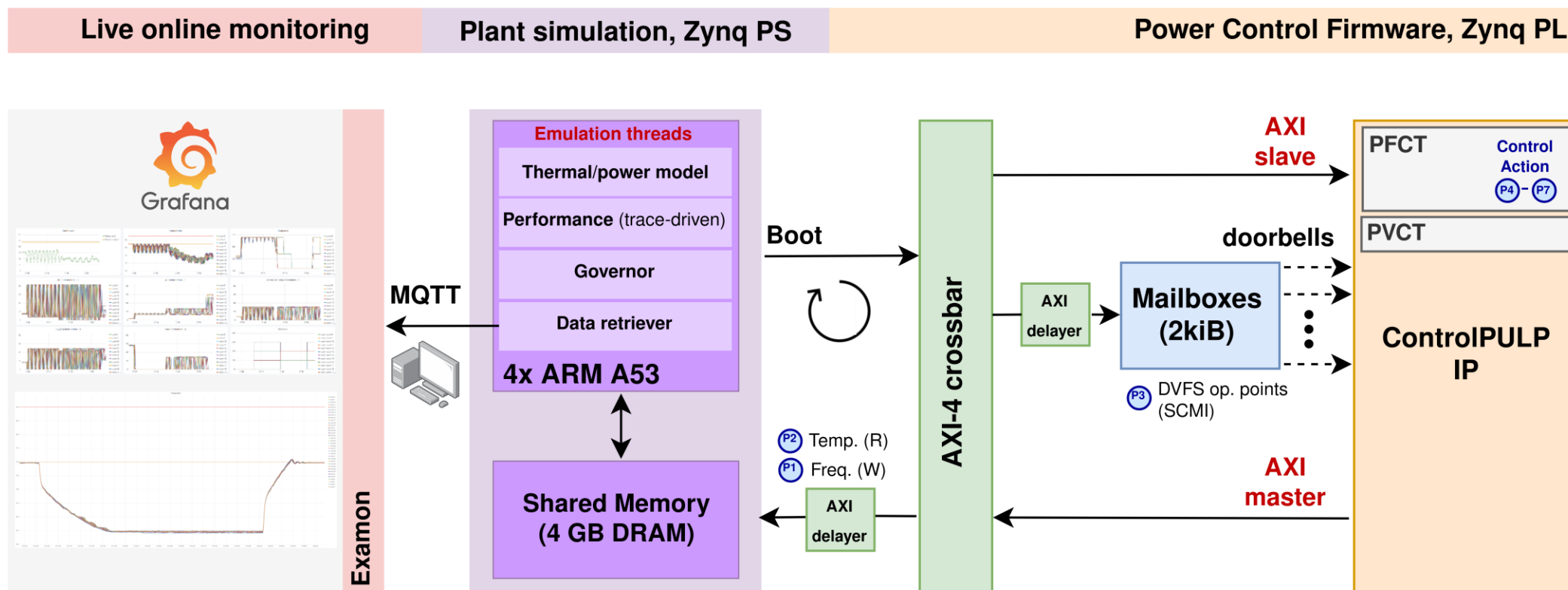
## FPGA-based Hardware-in-the-Loop emulation

- RTL + FW @ FPGA
- PLANT sim. @ A53
- ExaMon integration

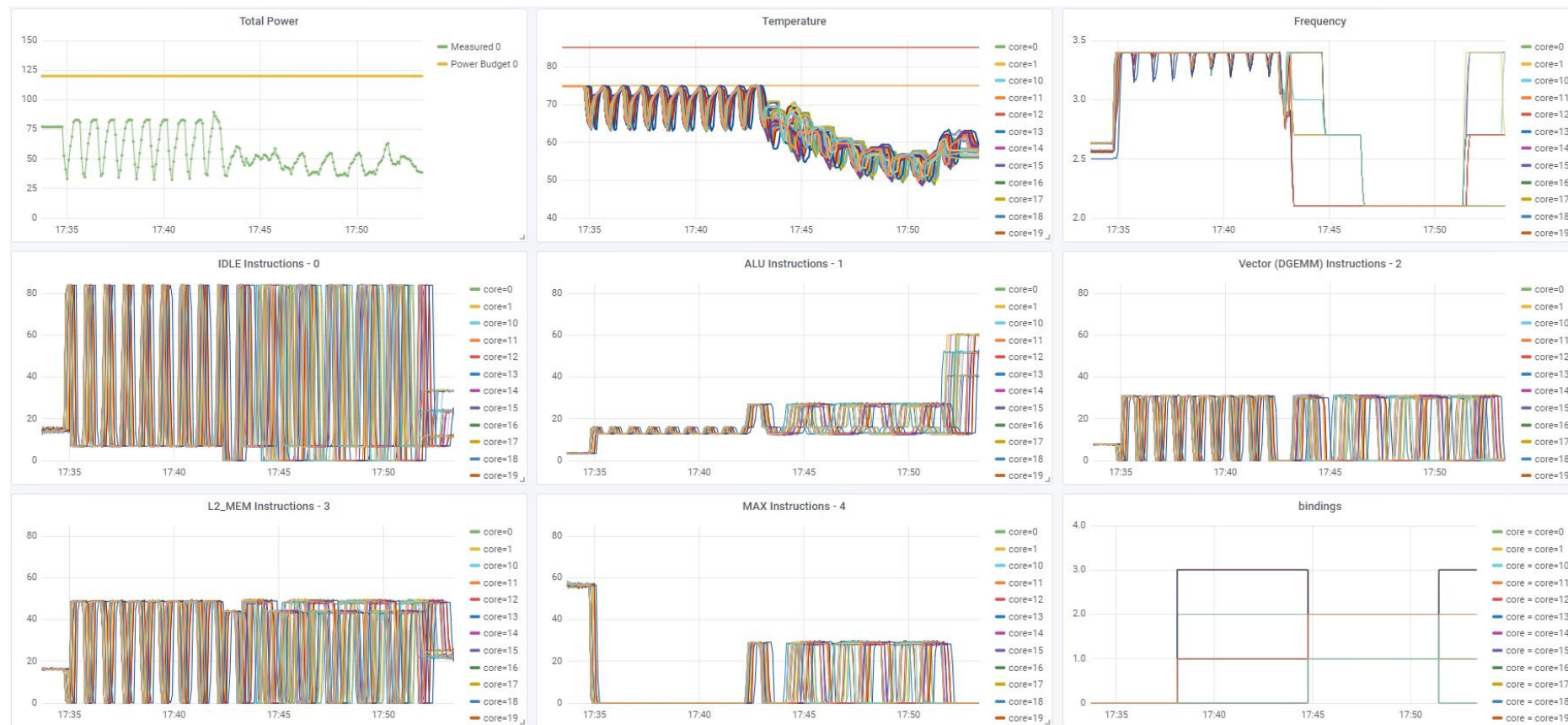


# EPI CASE STUDY: CO-SIMULATION FRAMEWORK

FPGA-based Power and Thermal simulation framework



# RESULTS



# CHALLENGES

## 2) Difficulty to find benchmark traces with an accurate enough timing and information

- Generally, traces have a precision of  $\sim 1\text{ms}$  to  $1\text{s}$
- Traces don't have enough information to simulate a variety of instruction power levels
  - Separation only between Vect, FP, and "normal" instructions
  - Sometimes do not include other general information such as power consumption, target and applied frequency, idle timings, CPI, memory utilization, load and store instructions and waiting times, synchronization, ...

## 3) Micro-controller limitations

- Limited Computing Power to achieve fast enough control and reactive multiple-agent communication
- Limited Memory to include most common open-source libraries, protocols, firmware, and code

## 4) Absence of the real Processor and accurate Models of EPI RHEA

## 5) Design time: configuration and architectural changes happens weekly



# ROADMAP 2023

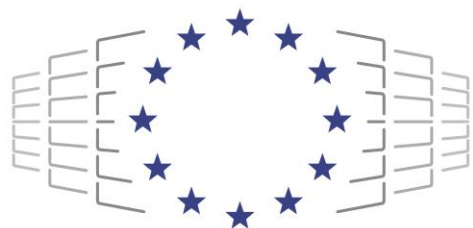
- Open-source the Co-Design framework (HW/SW): Q3 2023
- Implement **more advanced and proactive control algorithms** leveraging cluster-based acceleration: predictive policies (e.g. **Model Predictive Control – MPC**)
- Improve emulation framework with realistic off-chip interaction
- Tapeout in TSMC 7nm for **European Processor Initiative (EPI)** - second quarter 2023: Rhea
- Lightweight version of the power controller in EPAC for EPI
- Find new Partners interested in ControlPULP
- Expand the team



Thank you!

Q&A

# EPI FUNDING



**EuroHPC**  
Joint Undertaking

This project 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.



Federal Ministry  
of Education  
and Research

**FCT**

Fundação  
para a Ciência  
e a Tecnologia



Swedish  
Research  
Council



REPUBLIC OF CROATIA  
Ministry of Science and  
Education



