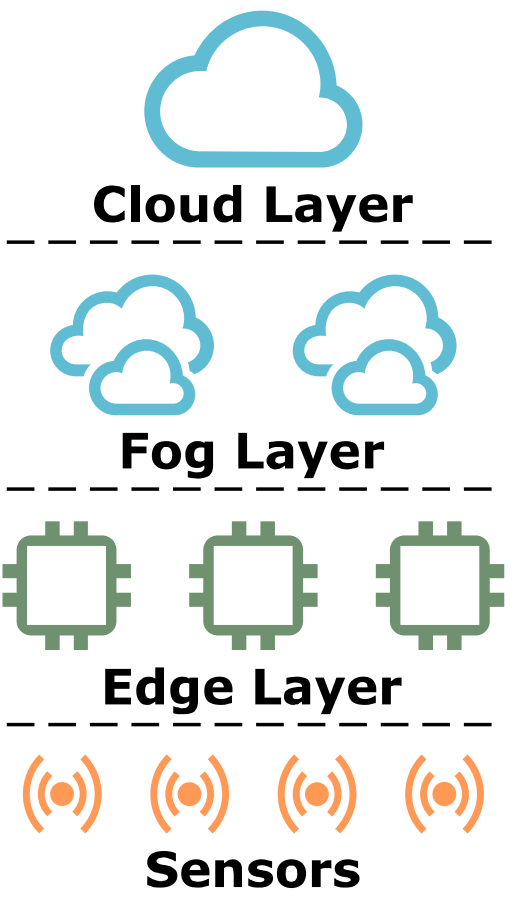


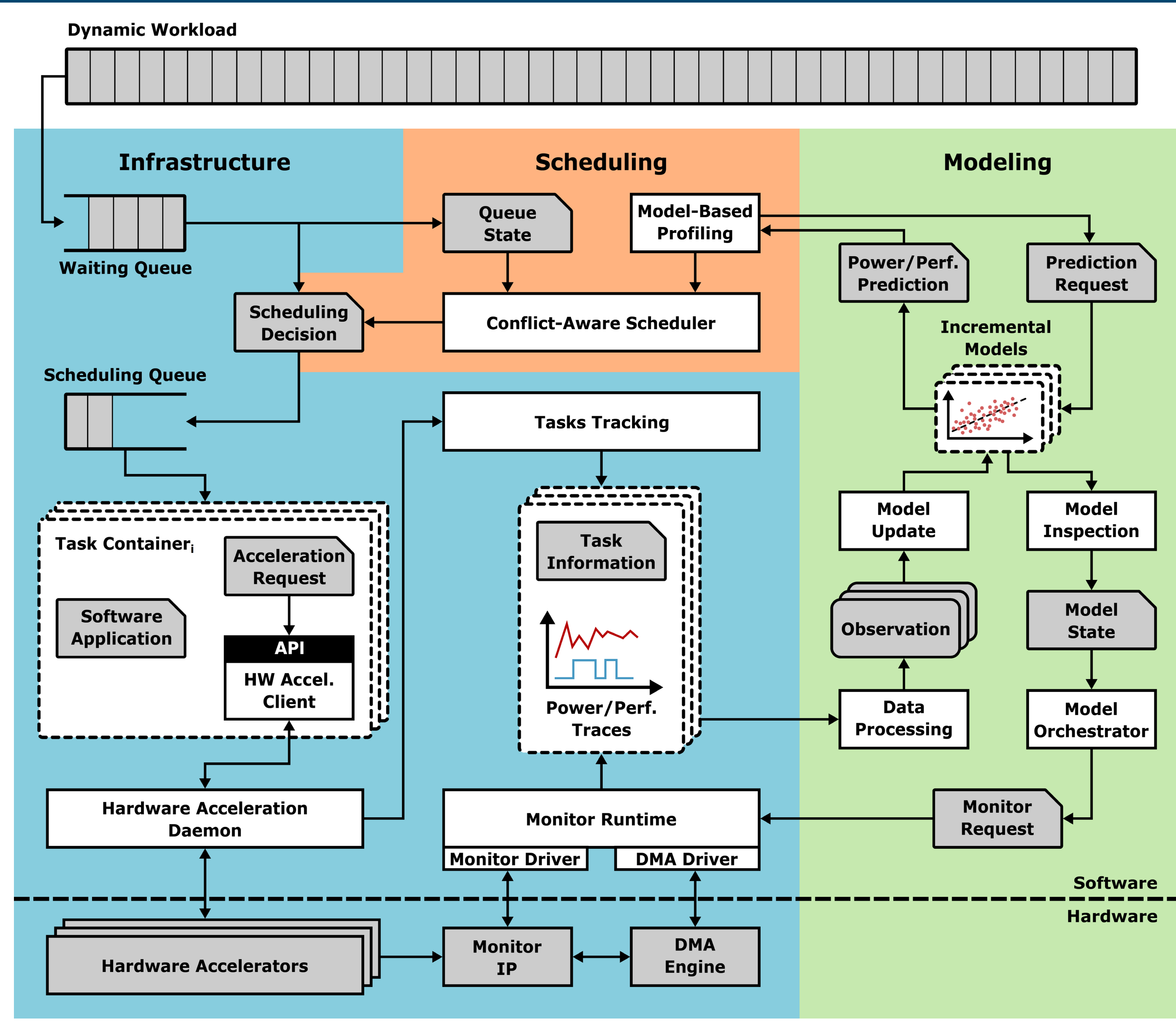
## Motivation and Objectives

The cloud-edge continuum is a computing paradigm where resources are dynamically distributed across the different layers to balance performance, latency and energy constraints. Due to their performance-to-power ratio and flexibility, FPGAs are ideal for accelerating workloads while adapting to changing computational demands. However, several problems must be solved:

- P1: Seamlessly offloading workloads into any FPGA node of the continuum (virtualization and workload management)**
- P2: Characterizing workloads in terms of power consumption and performance at run time (modeling)**
- P3: Optimizing task allocation in real-time based on performance and energy goals using the characterization (scheduling)**

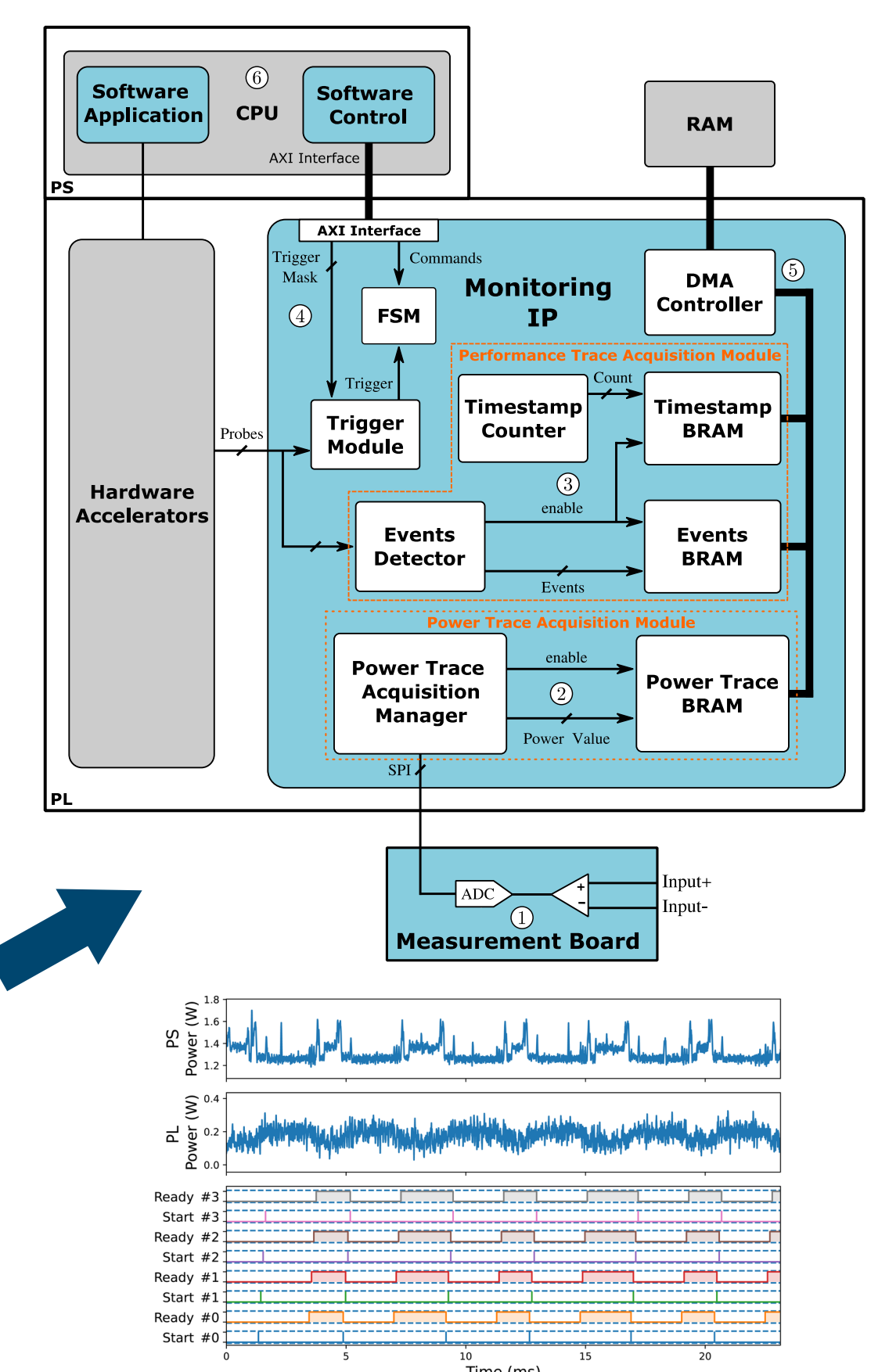


## Overview of the Proposed Solution



## Resource Management Infrastructure P1

- Infrastructure to manage dynamic workloads and offload them to the FPGA fabric efficiently using dynamic and partial reconfiguration. (1)
- Virtualization via a container-based deployment methodology to ensure hardware acceleration portability across the continuum nodes. (2)
- Non-intrusive monitoring mechanism for run-time acquisition of synchronized power consumption and performance. (paper under review)



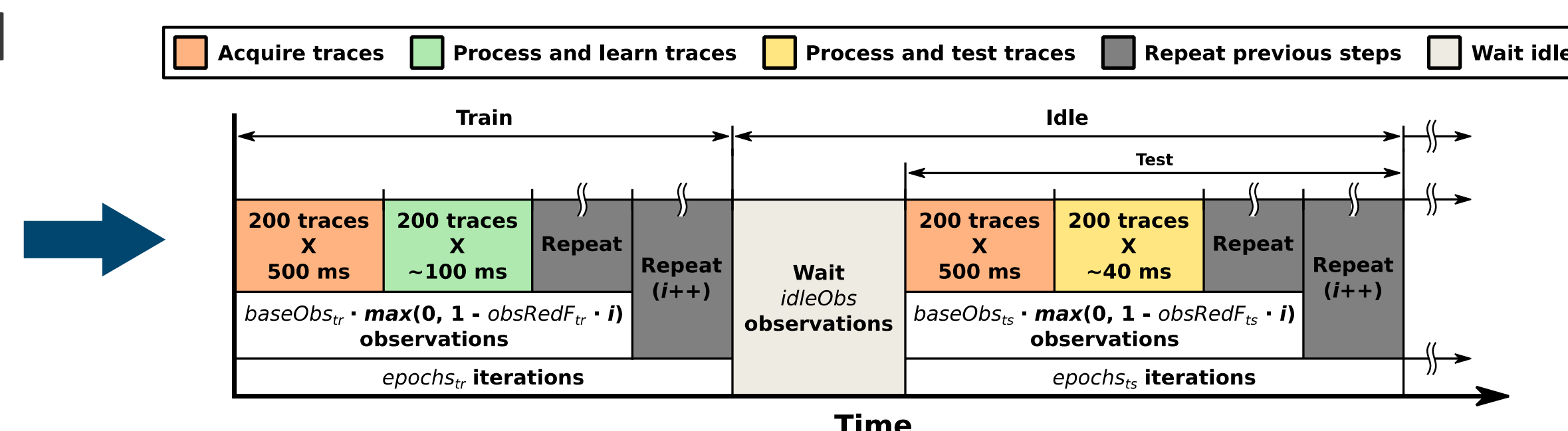
### Results

- ✓ Dynamic workload acceleration
- ✓ Run-time workload monitoring
- ✓ Device-agnostic infrastructure
- ✓ Task deployment across continuum

Layer	Board (FPGA device)	Virt. Overhead	
		Mean	Std. Dev.
Cloud	Alveo U250 (A-U250-P64G-PQ-G)	0.90%	0.20%
Fog	Zynq UltraScale+ ZCU102 (XCZU9EG-2FFVB1156)	0.41%	0.56%
Edge	Pynq-Z1 (XC7Z020-1CLG400C)	1.23%	1.67%

## Incremental Power and Performance Modeling P2

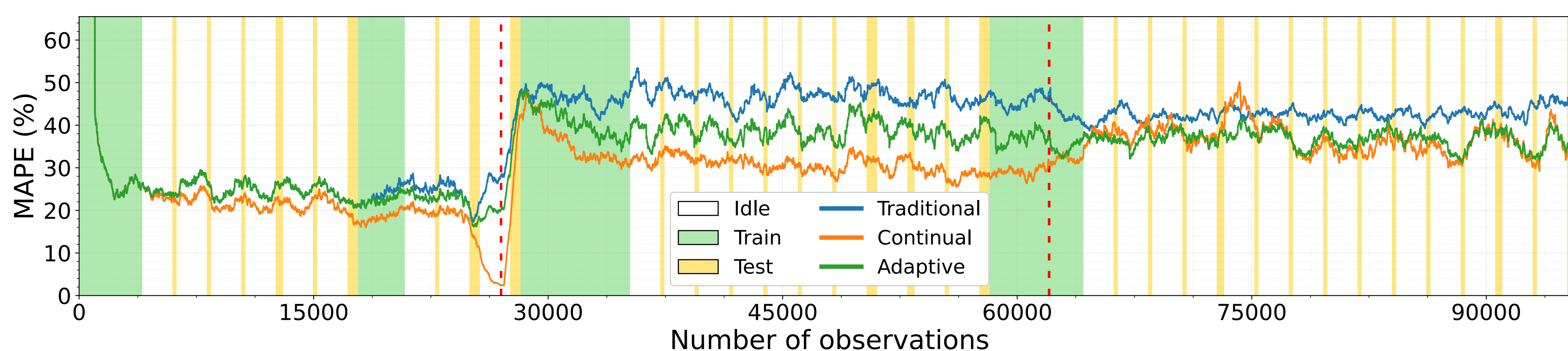
- ML-based models that use monitored traces to characterize the power consumption and the performance in reconfigurable multi-accelerator systems. (1)
- An Incremental learning extension to ensure models are kept up-to-date over time. (3)
- Learning process controlled by a resource-aware model orchestrator to minimize model training impact on the overall system. (3)



### Results

- ✓ Real-time and lifelong improvement
- ✓ Adaptability to model drift
- ✓ Prediction accuracy close to ideal scenario
- ✓ Reduced overhead

Modeling Approach	Overhead	Model Error (MAPE)		
		PS Power	PL Power	Performance
Traditional	-	6.46 (+3.73)	4.82 (+2.61)	46.36 (+10.35)
Continual	20.91%	2.73 (-)	2.21 (-)	36.01 (-)
Adaptive	4.49%	3.11 (+0.38)	2.74 (+0.53)	39.67 (+3.66)



## Task Scheduling P3

- Scheduling approach based on predictions from incremental ML models to make informed decisions on which task should be executed at any given time. (3)
- The scheduler selects the combinations of tasks with less interaction, aiming to optimize the overall system performance and power consumption. (3)
- A metaheuristic solution to optimize the solution space search time. (in progress)

### Preliminary evaluation results

- ✓ Workload execution optimization

Approach	Normalized Execution Time	Speedup
Baseline (first-come, first-served)	-	-
Kernel Evaluation	93.22%	1.07
Kernel and # Accelerators Evaluation*	74.04%	1.35

## Acknowledgements

