CEI UPM | Centro de Electrónica Industrial

ML-Based Modeling and Virtualization of Reconfigurable Multi-Accelerator Systems

Juan Encinas

UNIVERSIDAD POLITÉCNICA DE MADRID

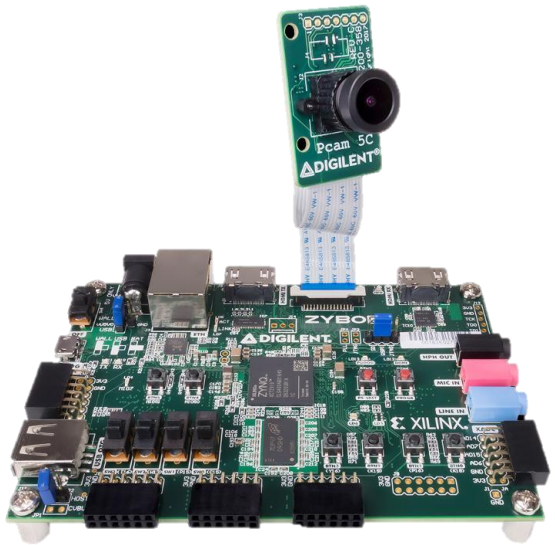
juan.encinas@upm.es

POLITÉCNICA

Motivation

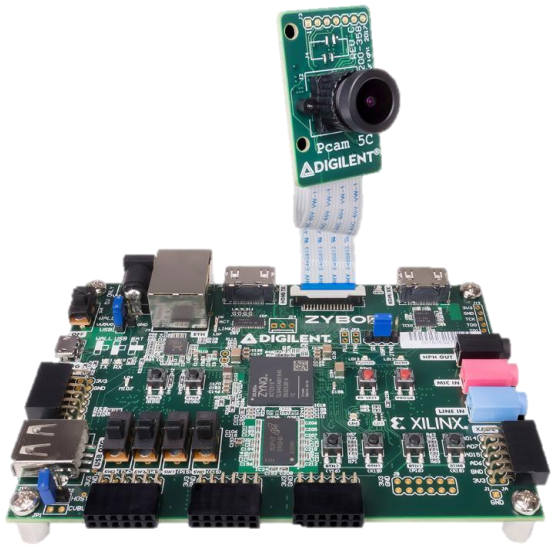
Motivation

Traditional Scenario

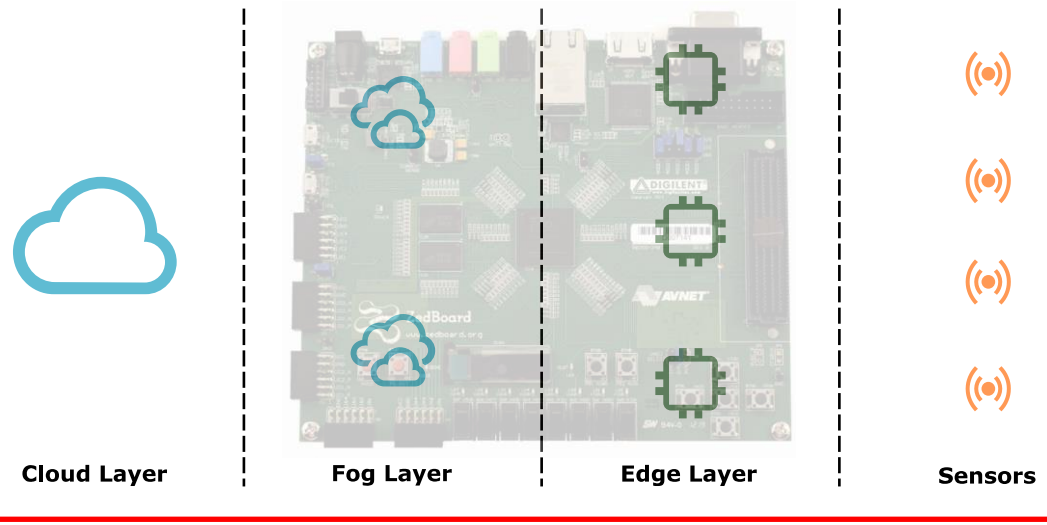


Motivation

Traditional Scenario



Computing Offloading Scenario



Goals

1. Real-Time Modeling and Management of the Reconfigurable multi-accelerator systems

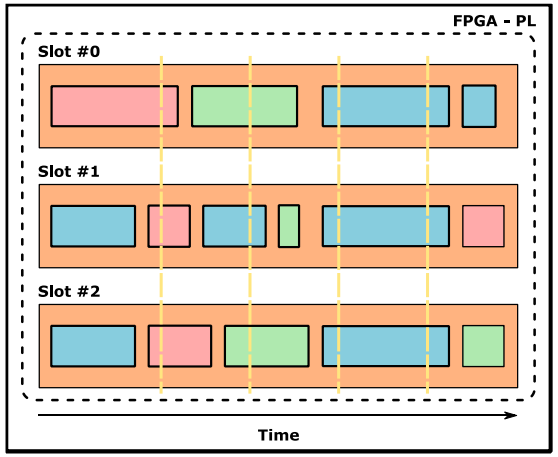
2. Virtualization Support to the reconfigurable multi-accelerator systems

Proposed Solution

Machine learning based modeling for reconfigurable multi-accelerator systems

Run-time power consumption and performance monitorization system

Computing offloading workload manager

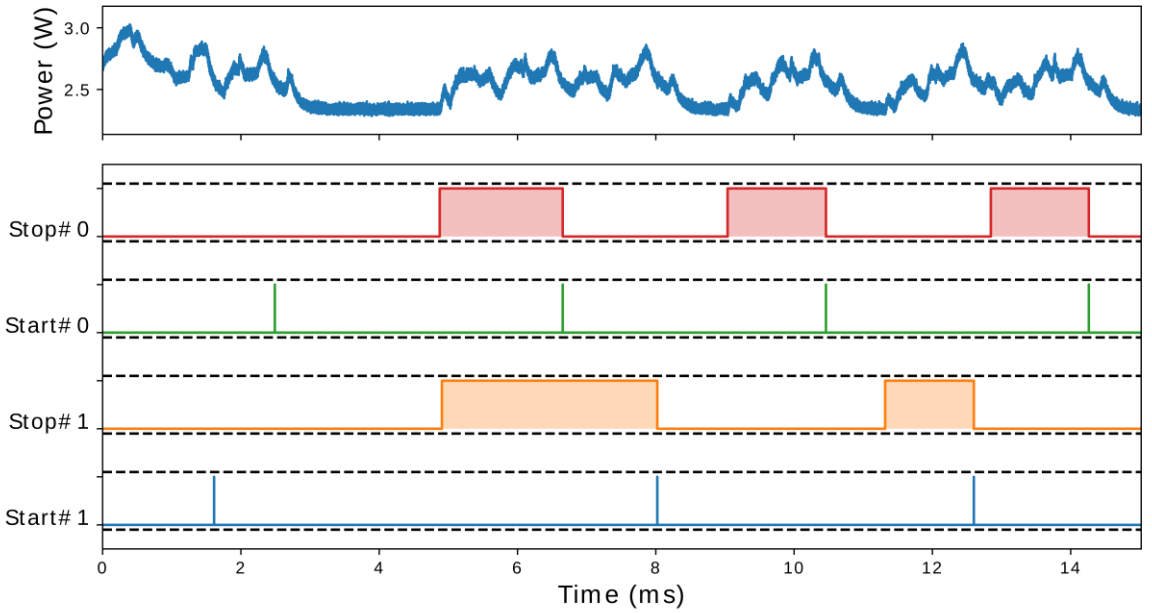


Monitoring Infrastructure

- Non-intrusive
- Synchronized power and performance traces

Components:

- Measurement board
- Monitoring IP
- Software components

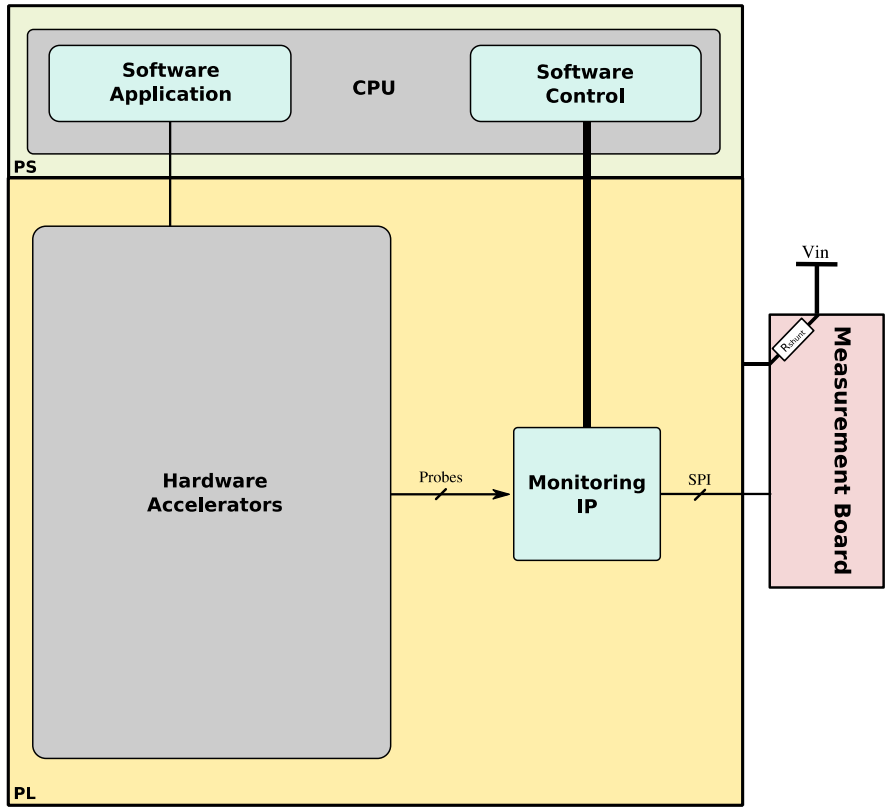


Monitoring Infrastructure

- Non-intrusive
- Synchronized power and performance traces

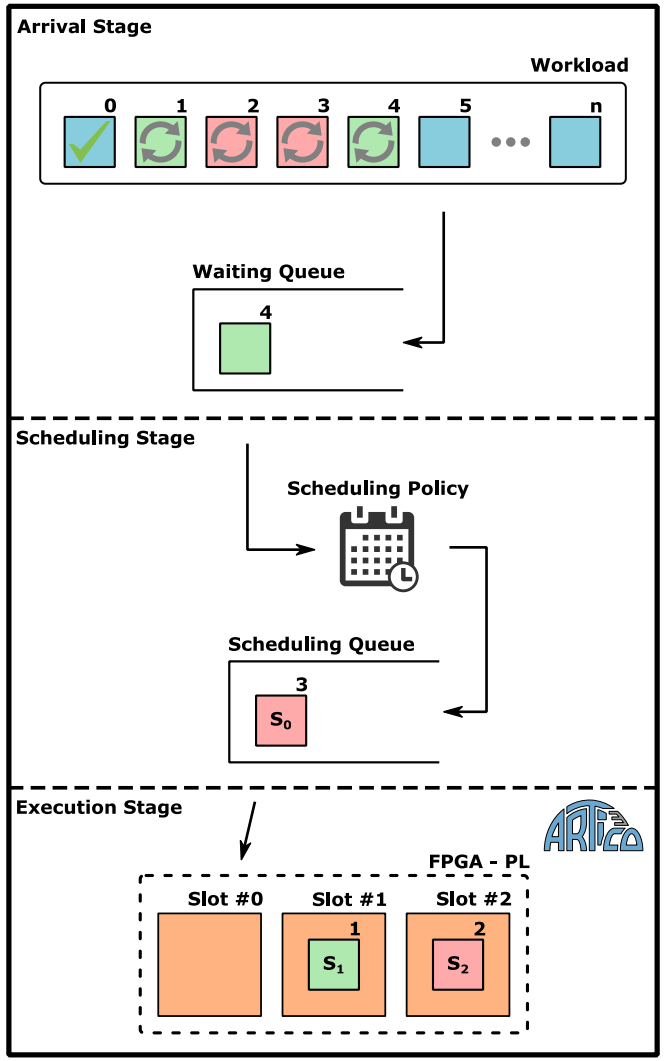
Components:

- Measurement board
- Monitoring IP
- Software components



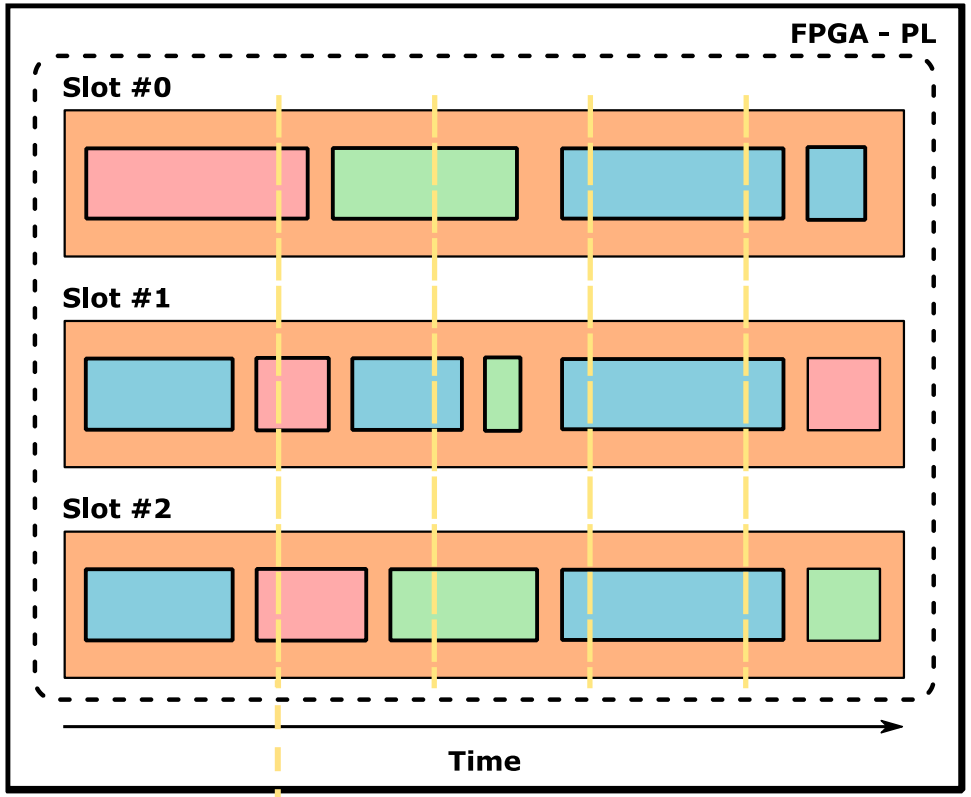
Workload Manager

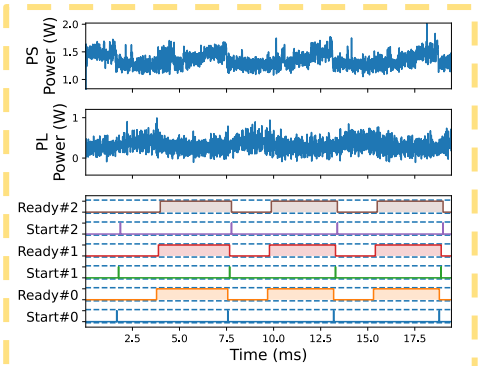
- Generate configurable computing offloading workloads
- Execute hardware acceleration requests
- Monitor and track every request throughout the entire process



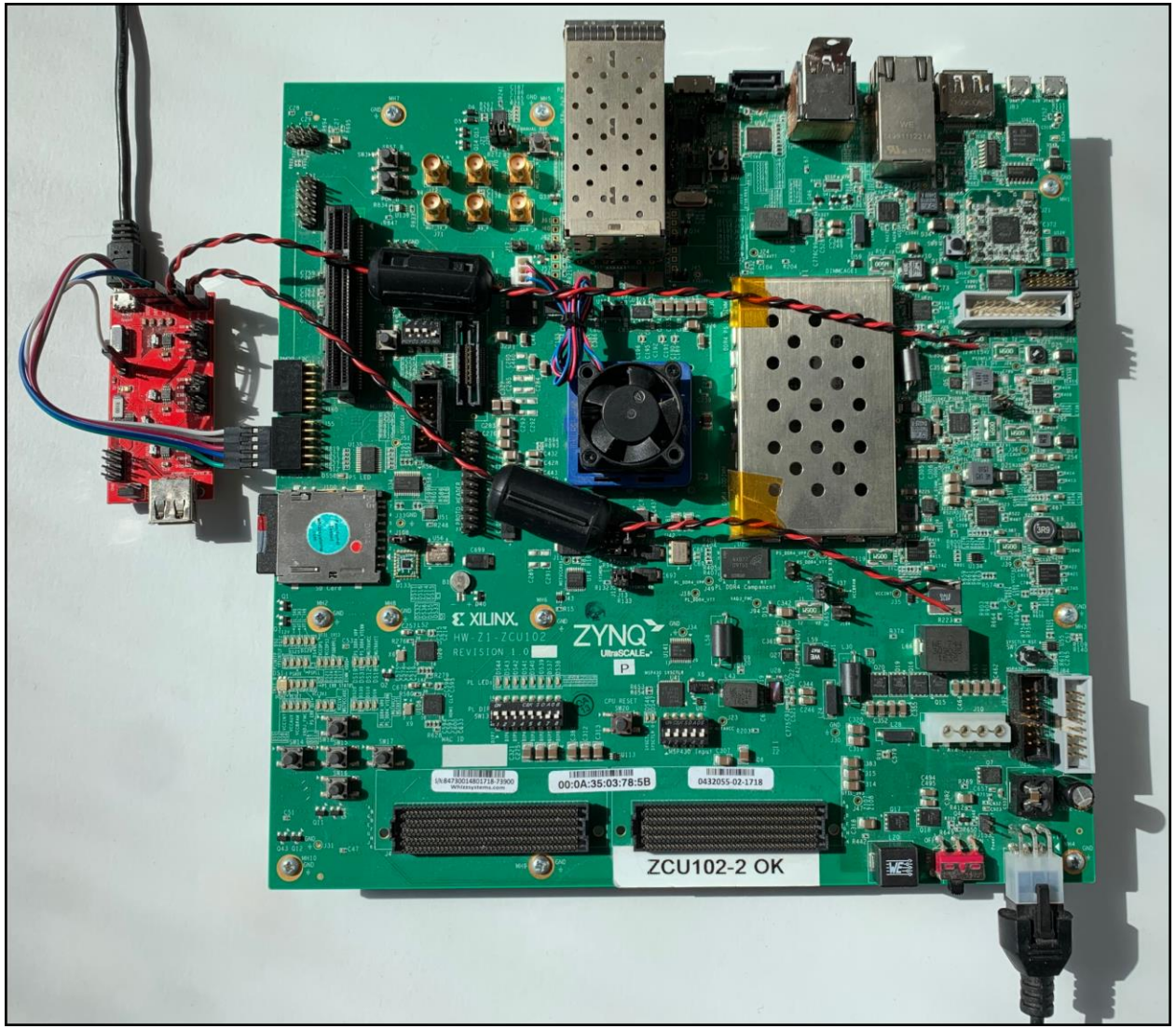
Workload Manager

- Generate configurable computing offloading workloads
- Execute hardware acceleration requests
- Monitor and track every request throughout the entire process





Test Setup


XILINX
ZYNQ
UltraSCALE+
HW-Z1-ZCU102
REVISION 1.0
00:0A:35:03:78:5B
ZCU102-2 OK

Modeling Approach

- Machine Learning-based models
  - Power consumption
  - Performance
- Machine learning algorithms:
  - Support Vector Regression (SVR)
  - Regression Tree Ensemble (RTE)

Experimental Setup

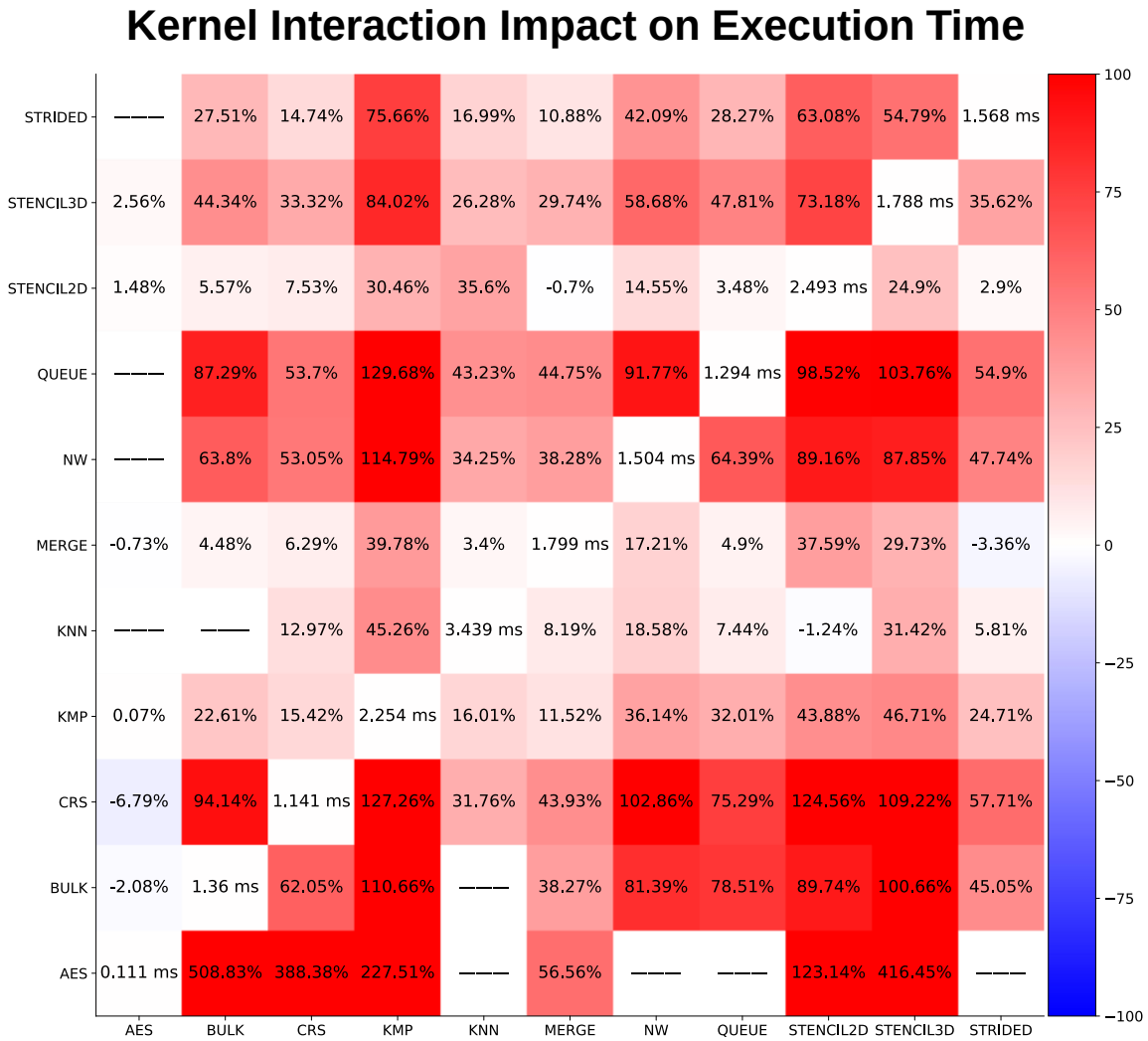
MachSuite [1]

Training and validation datasets:

- Edge computing workload with combinations of 11 different MachSuite kernels, generated with the Workload Manager
- PS power consumption, PL power consumption and performance traces obtained with the Monitoring Infrastructure integrated in the Workload Manager

[1] B. Reagen, R. Adolf, Y. S. Shao, G. Wei, and D. Brooks, “MachSuite: Benchmarks for accelerator design and customized architectures”

Experimental Results – Analysis



Experimental Results – Modeling

Model	RMSE	MAPE
PS Power Consumption (SVR)	0.0447	1.329
PL Power Consumption (SVR)	0.0072	1.486
Execution Time (RTE)	0.3549	6.335

RMSE: Root Mean Squared Error

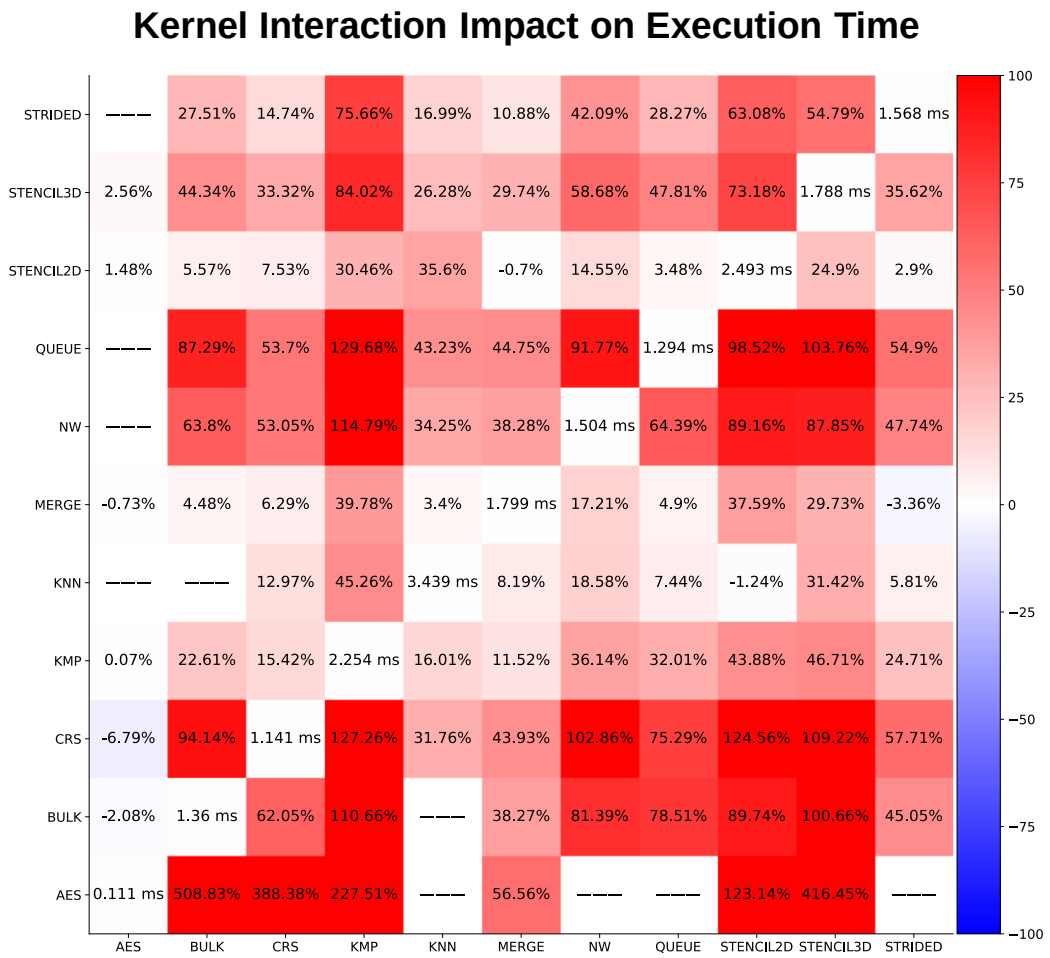
MAPE: Mean Absolute Percentage Error

Experimental Results – Modeling

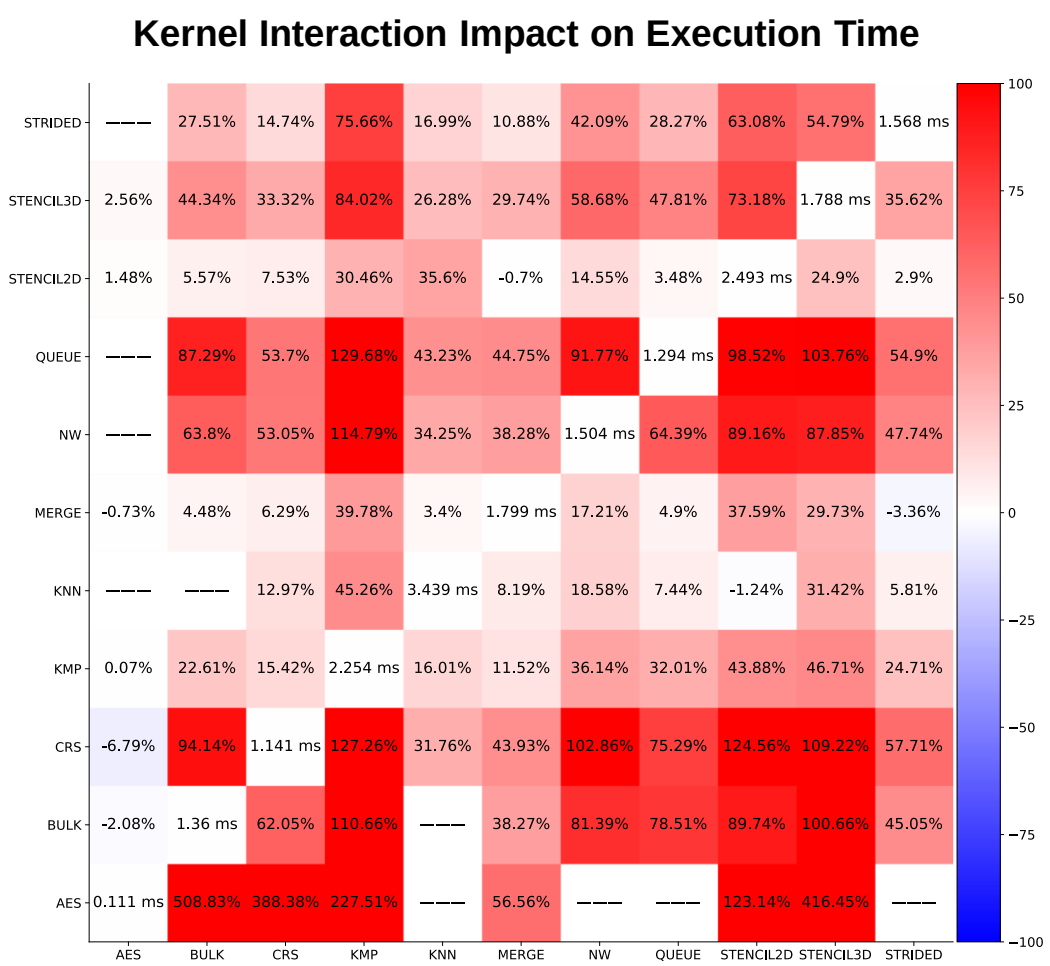
11-kernel combinations | Up to 8 accelerators per kernel | Around 5000 test observations

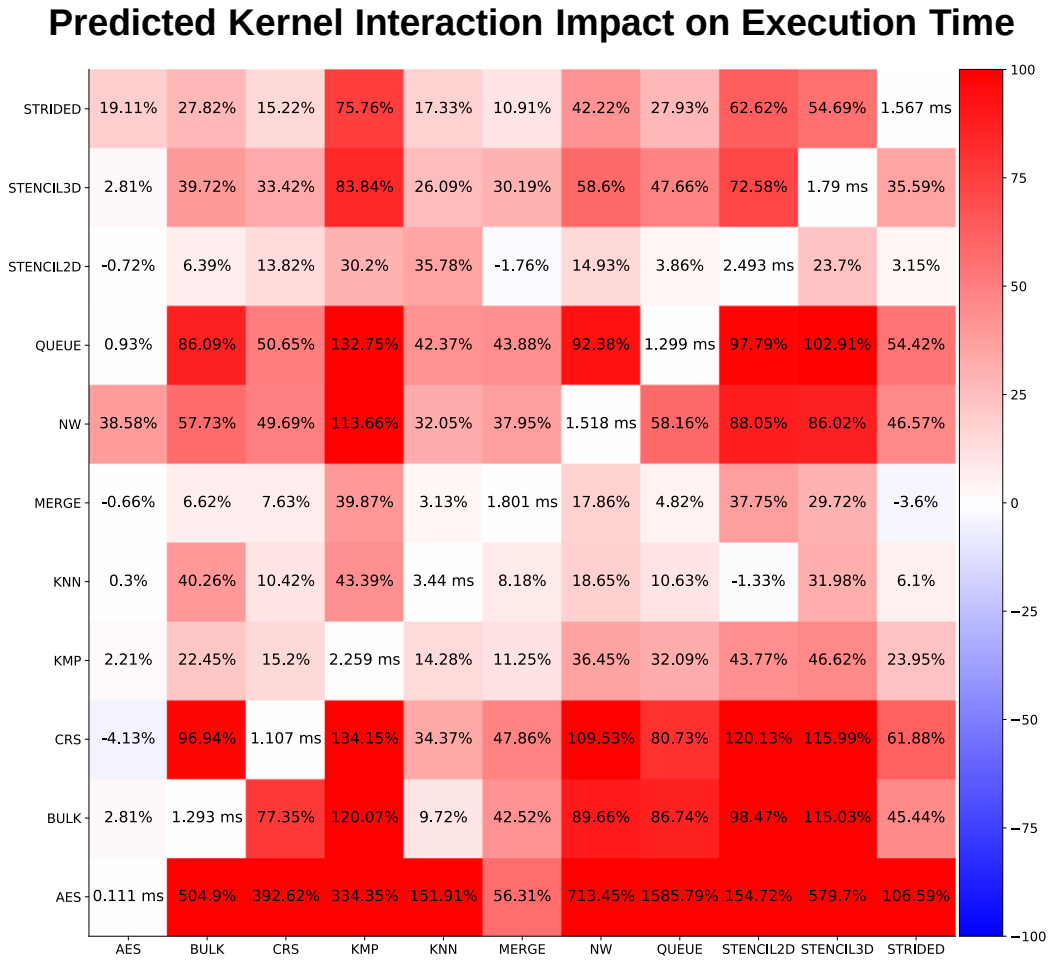
Observations obtained under a Linux-based OS

Experimental Results – Modeling



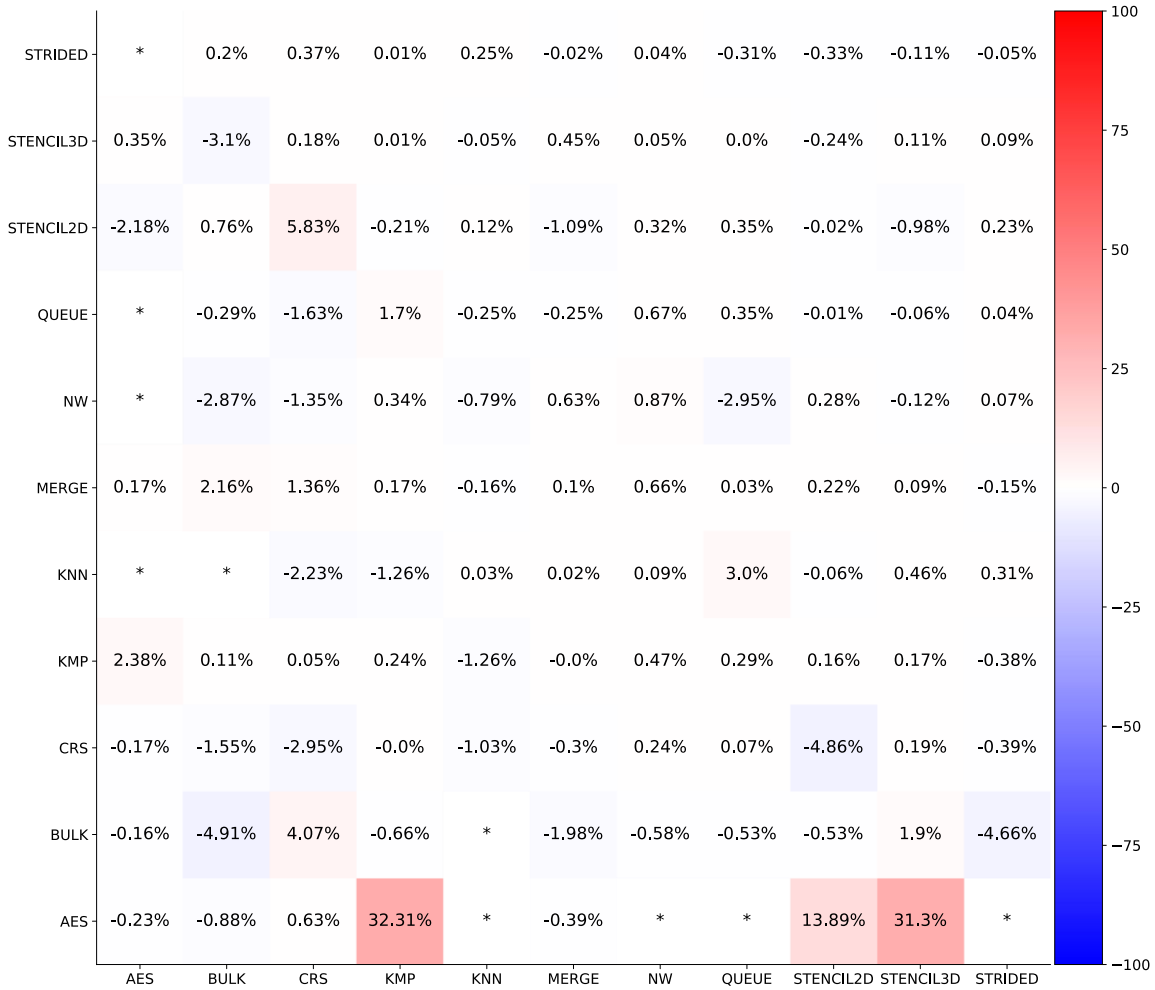
Experimental Results – Modeling





Experimental Results – Modeling

Relative Error when Predicting Kernel Interaction Impact on Execution Time

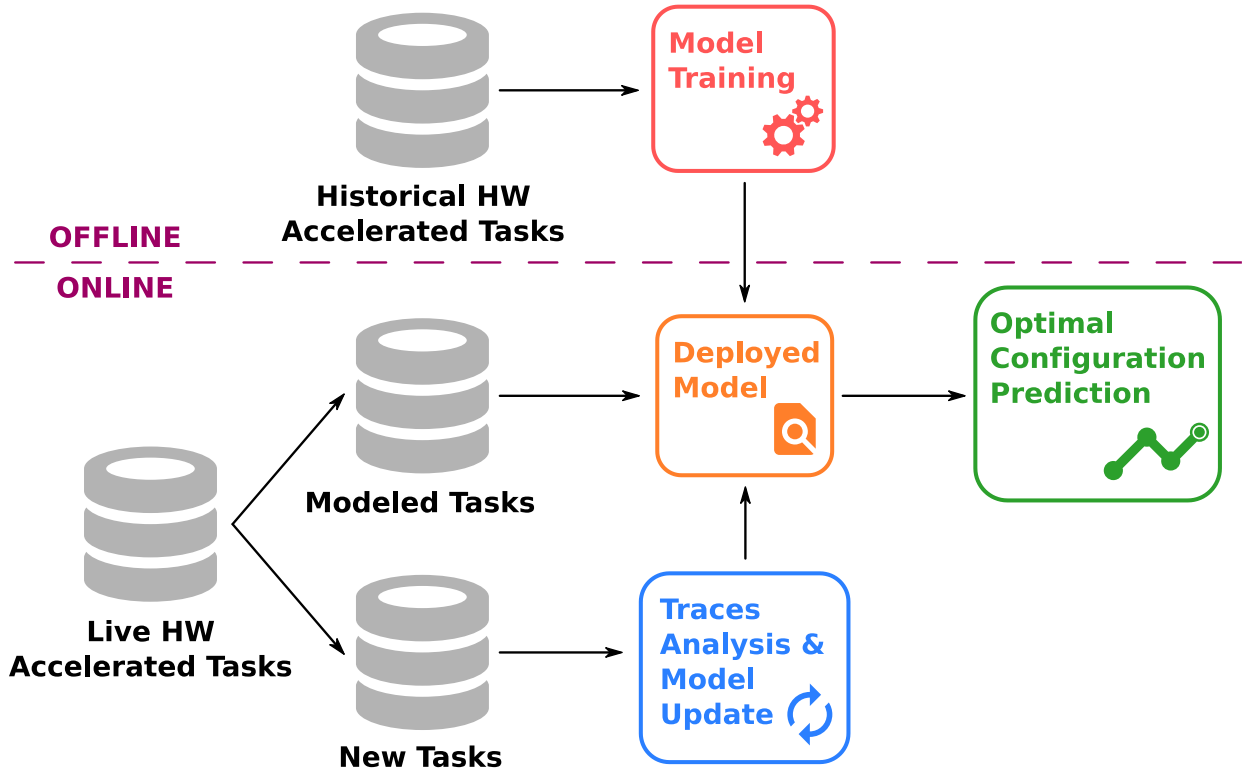


Conclusions

- Dynamic workload management infrastructure for effective computing offloading workloads generation, offloading and monitoring in FPGA-based systems.
- ML-based models are very accurate when predicting power consumption and performance in reconfigurable multi-accelerator systems as well as modeling the interaction between kernels.

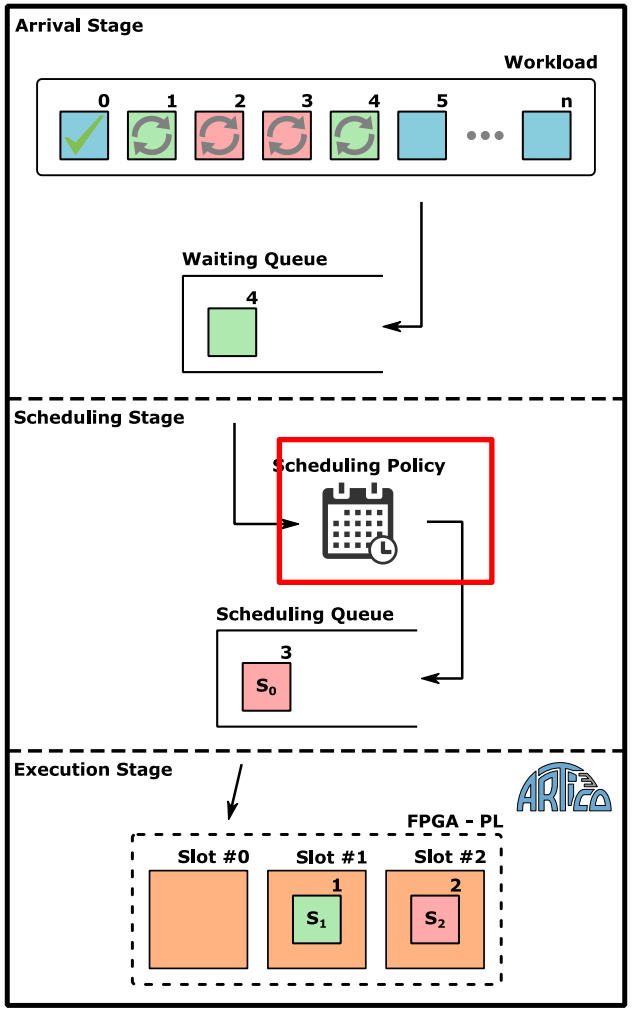
Future Work

- Online modeling
- Run-time self-adaptation



Future Work

- Smart scheduling approach
- Advanced resource management



THANK YOU FOR YOUR ATTENTION