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

Juan Encinas
Centro de Electrónica Industrial, Universidad Politécnica de Madrid, Spain

SEPTEMBER 19-23 2022 – PULA, SARDINIA, ITALY
CYBER PHYSICAL SUMMER SCHOOL SYSTEMS 2022

Motivation and Objectives

The goal of this thesis is to provide reconfigurable multi-accelerator systems with the ability to self-adapt at run-time to varying application conditions, environment and input data, in a transparent way to the user.

Problems being addressed:

P1. Real-time modeling and management of reconfigurable multi-accelerator systems

P2. Virtualization support for reconfigurable multi-accelerator systems

Monitoring Infrastructure P1

Non-intrusive instrumentation tool used to acquire synchronized power/performance traces in FPGA-based systems.

External ADC board

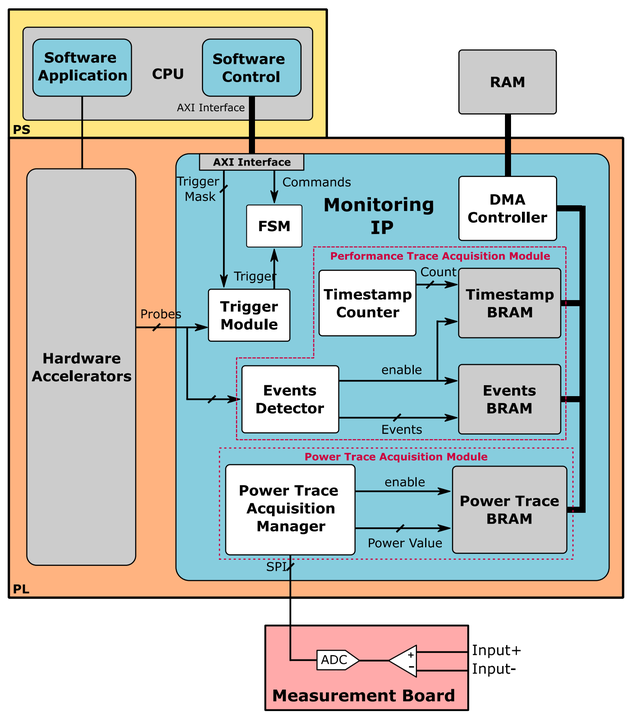
- Power consumption measurement
- SPI interface
- 1 Msample/s

Embedded monitoring IP

- Power/performance traces from probes
- Multiple trigger modes

SW library & support tools

- Baremetal-/Linux-based applications
- Trace processing and visualization tool



Dynamic Workload Management P1

A scheduling infrastructure enhancing the ARTICo³ framework that attends all the incoming acceleration requests, deciding when to execute them in the FPGA following a specific scheduling policy.

Workload offloading

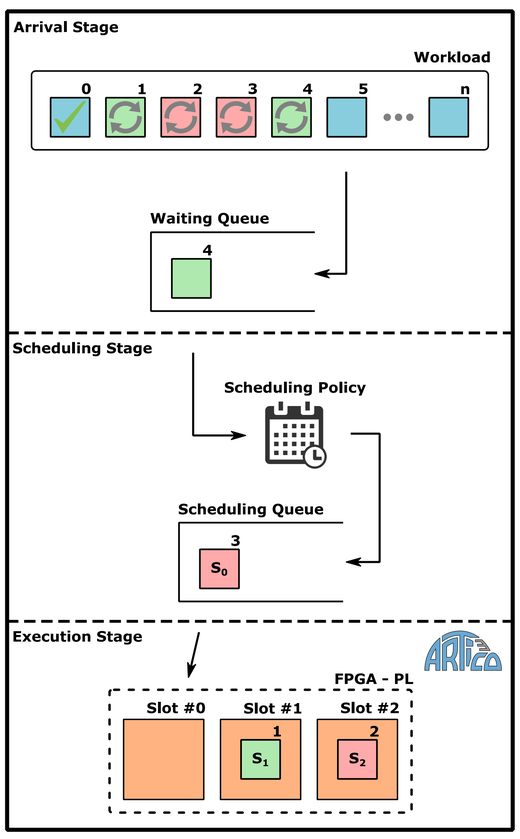
- Two-queue infrastructure
- Configurable scheduling policy

Workload monitorization

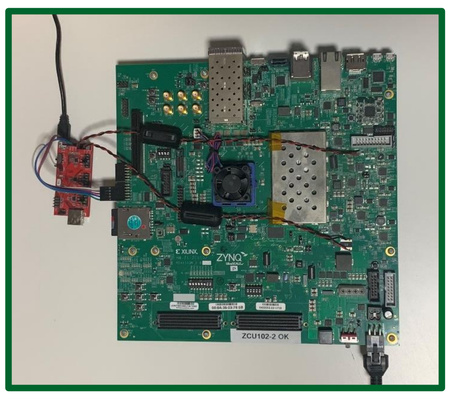
- Acceleration requests tracking
- Power consumption and performance monitorization

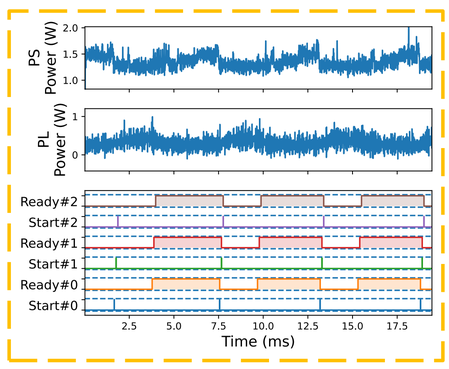
Workload generation

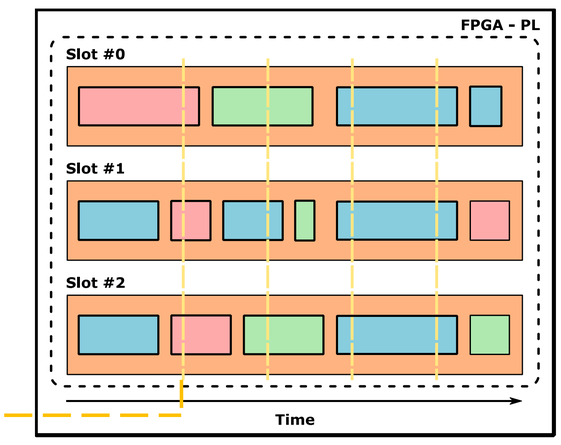
- User-configurable workload patterns
- Highly customizable



Implemented in Zynq UltraScale+

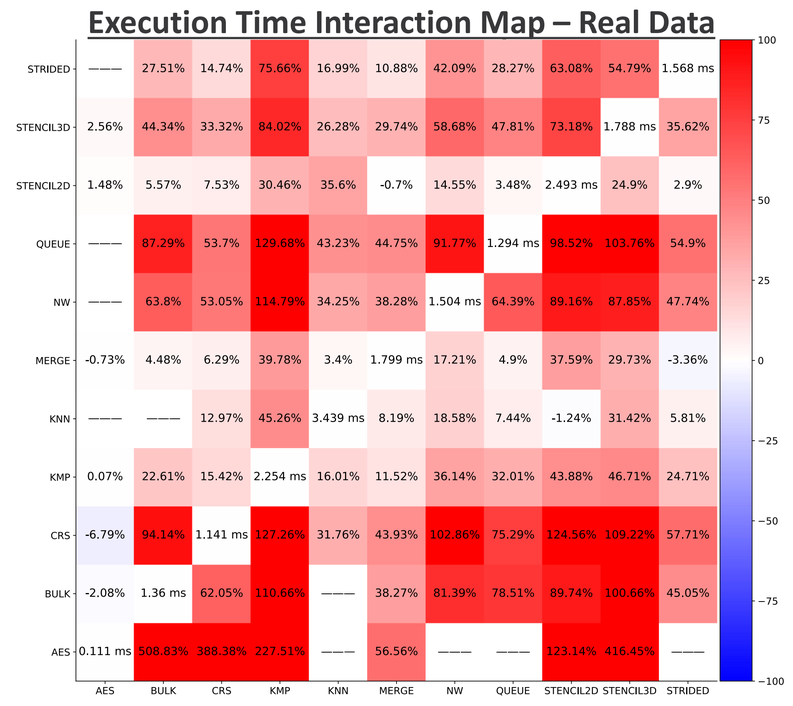






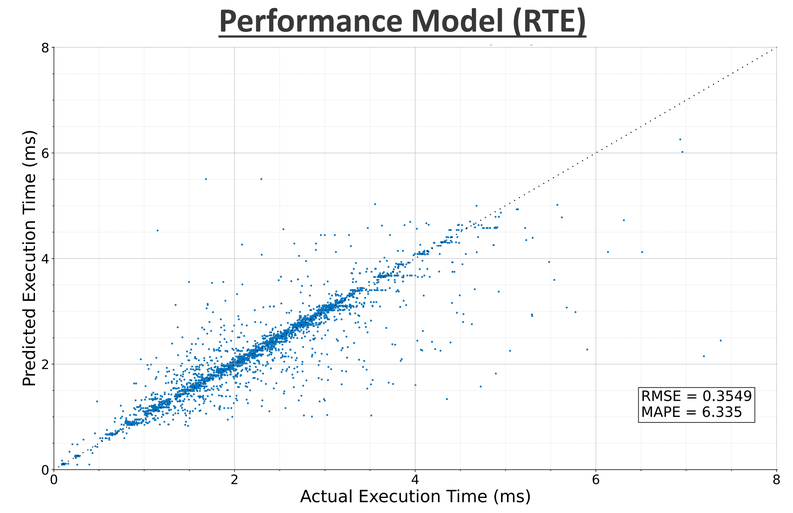
Analysis P1

MachSuite Benchmarks - Multi-Kernel Interaction



Offline Modeling P1

Performance Model (RTE)



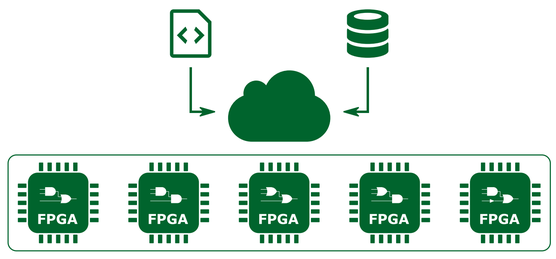
Relative Error in Predicted Interaction

Model	Mean	Std. Deviation
PS Power Consumption (SVR)	-0.22%	0.82%
PL Power Consumption (SVR)	0.09%	0.36%
Execution Time w/o AES (RTE)	-0.16%	1.44%
Execution Time (RTE)	0.54%	4.65%

Future Work P1 P2

Online Modeling

Complex Scheduling



FPGA Virtualization

Acknowledgement

KYKLOS 4.0 - 872570
European Commission | Horizon 2020 European Union funding for Research & Innovation
POLITÉCNICA | UNIVERSIDAD POLITÉCNICA DE MADRID

juan.encinas@upm.es