UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

An Open-Source Overlay for Reconfigurable, Accelerator-Rich Embedded Systems

Gianluca Bellocchi, Alessandro Capotondi, and Andrea Marongiu

IWES 2021

University of Modena and Reggio Emilia, *<name>.<surname>@unimore.it*

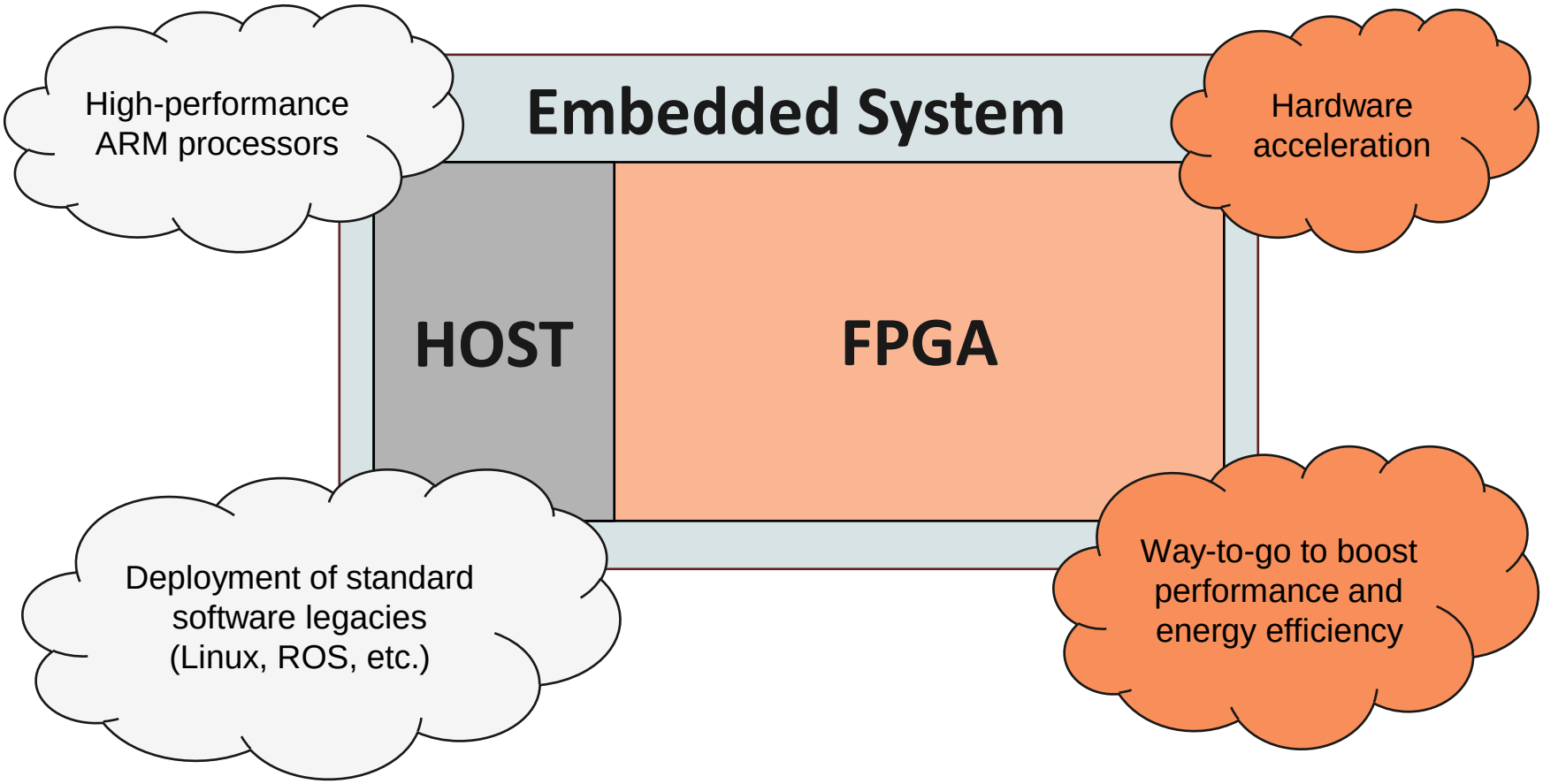
F**ondo di **A**teneo per la **R**icerca **FAR2020

COMP4DRONES

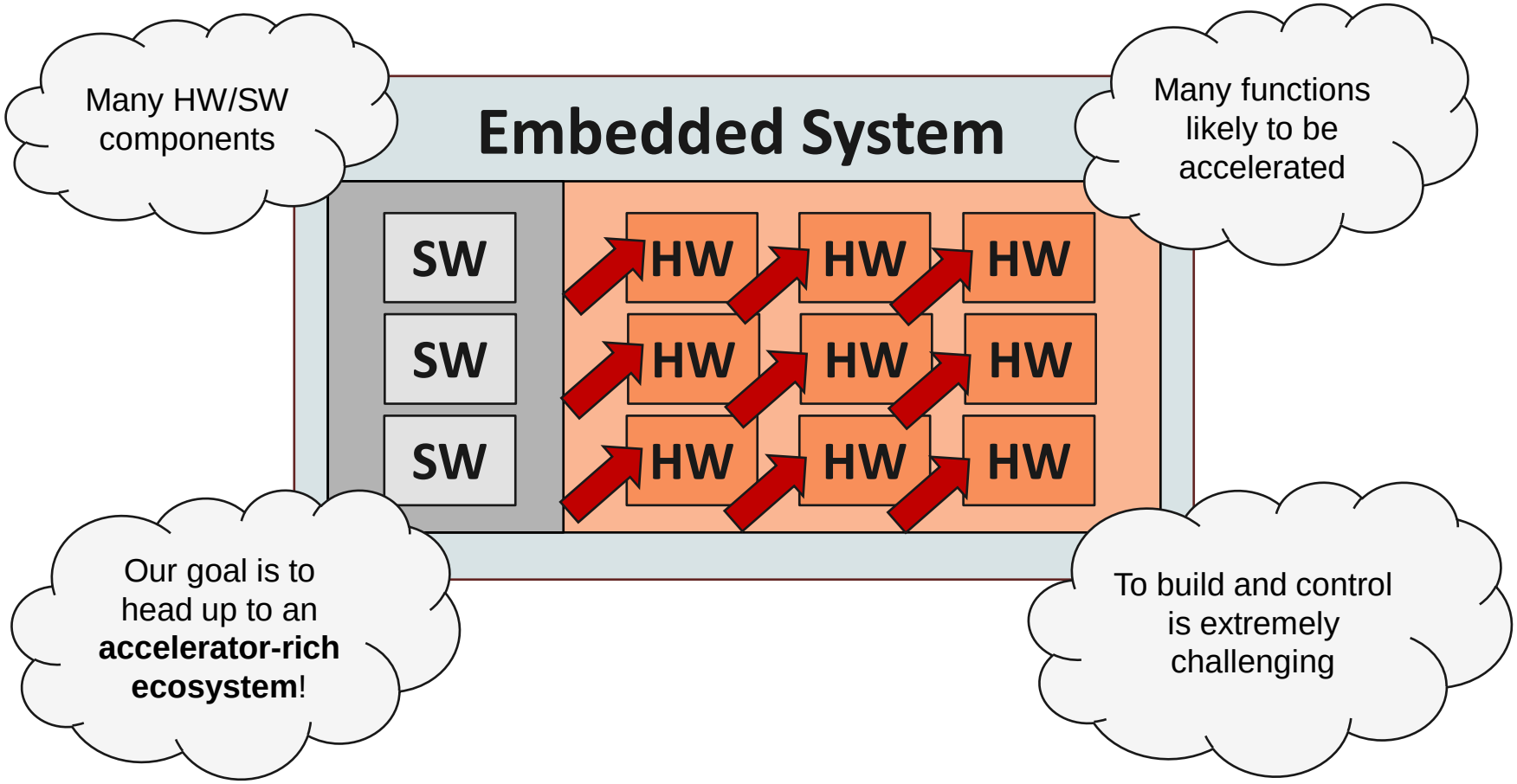
Introduction

Autonomous Embedded Systems

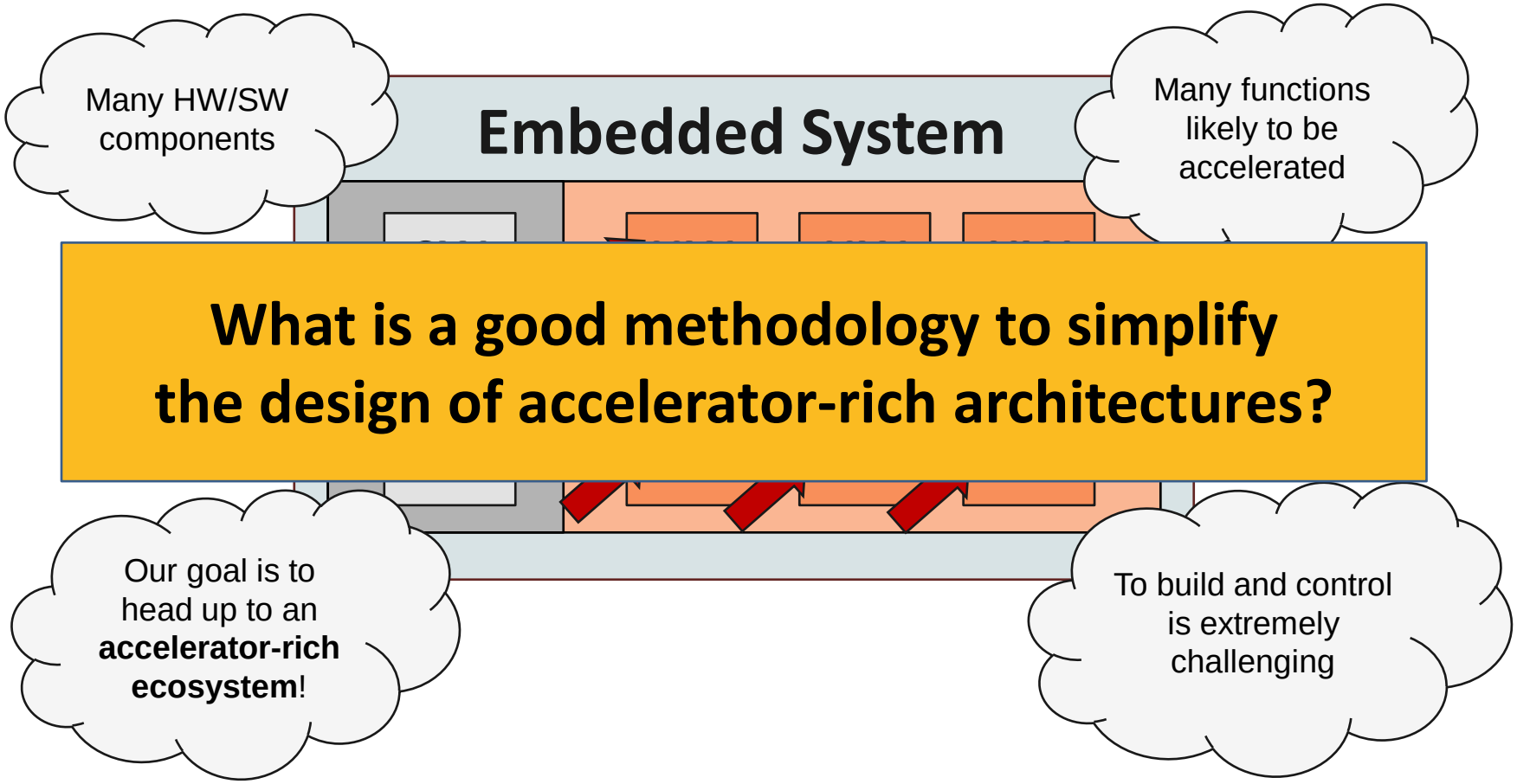
Introduction



Accelerator-Rich Paradigm



Accelerator-Rich Paradigm



Accelerator-Rich Paradigm

- **What has to be simplified?**
  - **System-Level Design**
    - Build and evaluate accelerator-rich systems
      - Expensive
      - Time-consuming
  - **Design Space Exploration (DSE)**
    - Key effects only manifest at system-level
    - User knobs:
      - System optimization
      - Accelerator optimization

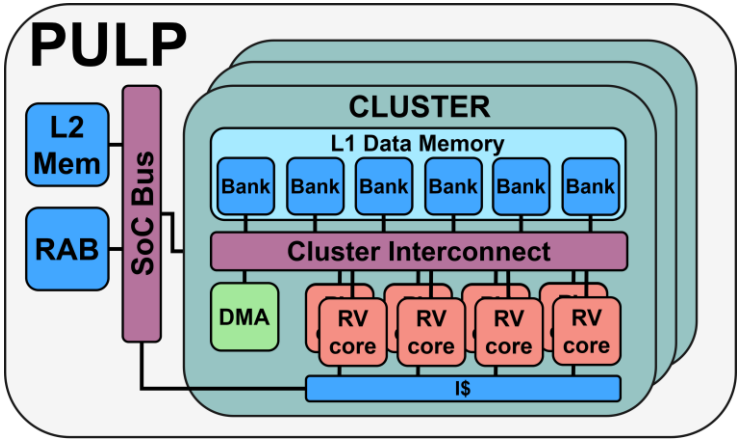
Recent Contribution

- **A first proposal to simplify the deployment of hardware accelerators**
  - Design methodology
    - Overlay-based
    - Plug-and-play integration of HW accelerators
  - Experimental results
    - Resource cost (LUT, FF, BRAM, DSP)
    - Application profiling
    - Comparison with Xilinx HLS flow

G. Bellocchi, A. Capotondi, F. Conti and A. Marongiu,
A RISC-V-based FPGA Overlay to Simplify Embedded Accelerator Deployment,
24th Euromicro Conference on Digital System Design (2021)

Starting Point

- **PULP architecture**
  - PULP stands for «Parallel Ultra Low Power»
  - Open and Scalable HW/SW research and development platform
  - Cluster-based architecture
  - RISC-V ISA compliant



Website: ***pulp-platform.org***

RISC-V

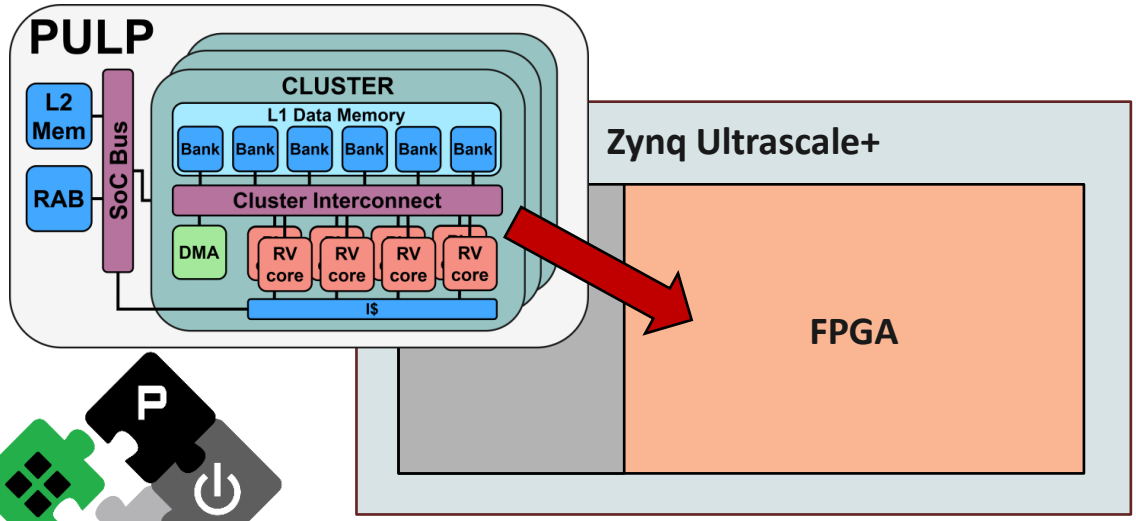
ETH zürich

Website: ***pulp-platform.org***

Starting Point

A. Kurth, A. Capotondi, P. Vogel, L. Benini, A. Marongiu, *(2018)* ***HERO: An open-source research platform for HW/SW exploration of heterogeneous manycore systems***

- **HERO**
  - FPGA emulation of heterogeneous and massively parallel PULP systems
  - Instantiable with COTS FPGA-based heterogeneous SoCs



Website: ***pulp-platform.org***

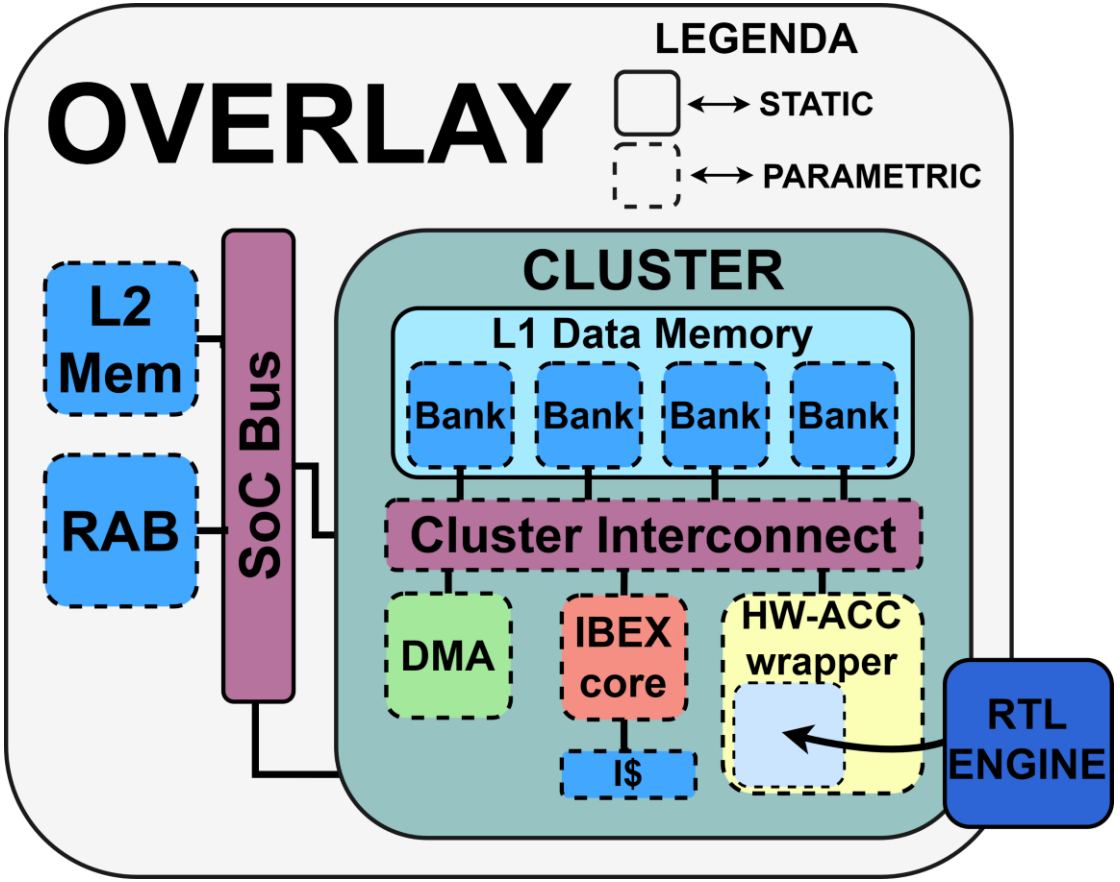
RISC-V

ETH zürich

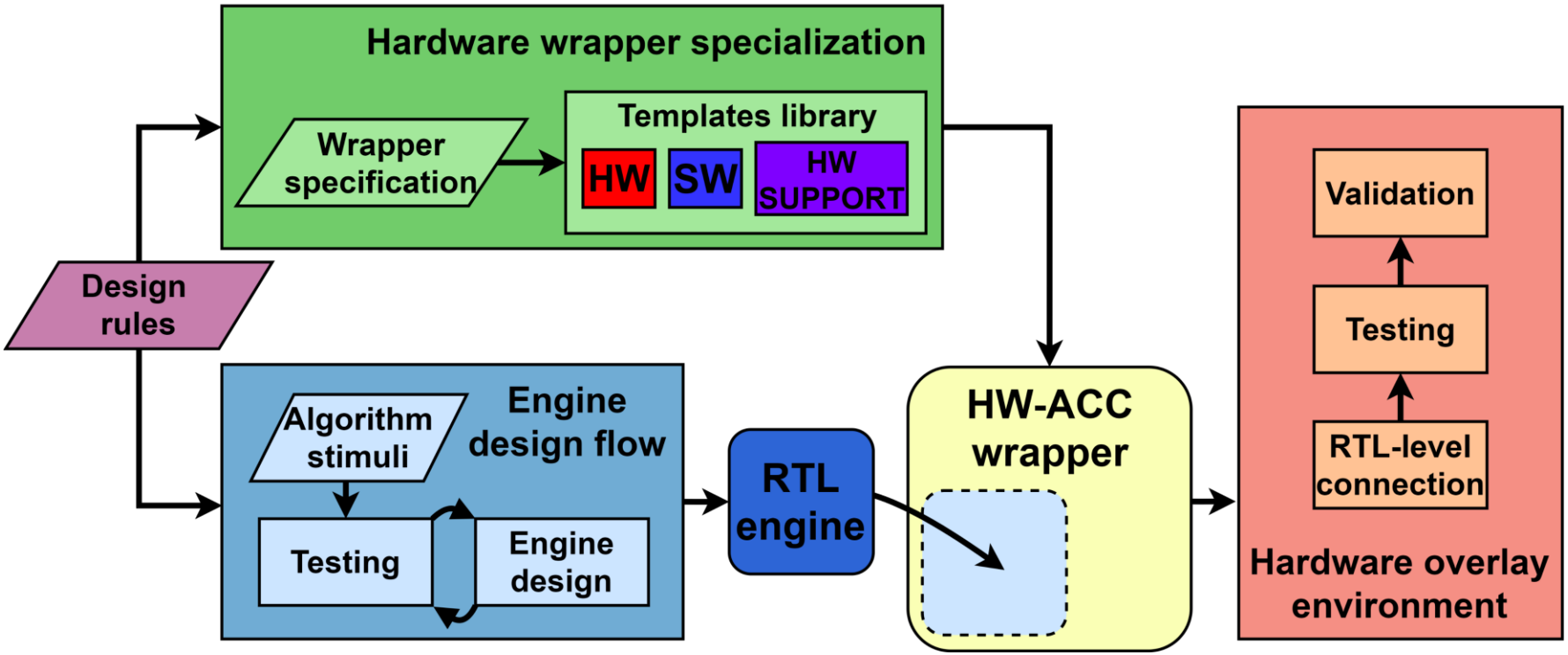
Overlay Architecture

- **What is it?**
  - Hardware abstraction layer
  - Overlays the original FPGA fabric
    - Hides hardware details
- **Features:**
  - Coarse-grained
    - Rapid swapping of architectural blocks
  - Avoid FPGA design flow
    - Improved design productivity
  - Programmable via standard APIs for heterogeneous compute platforms

Overlay Architecture

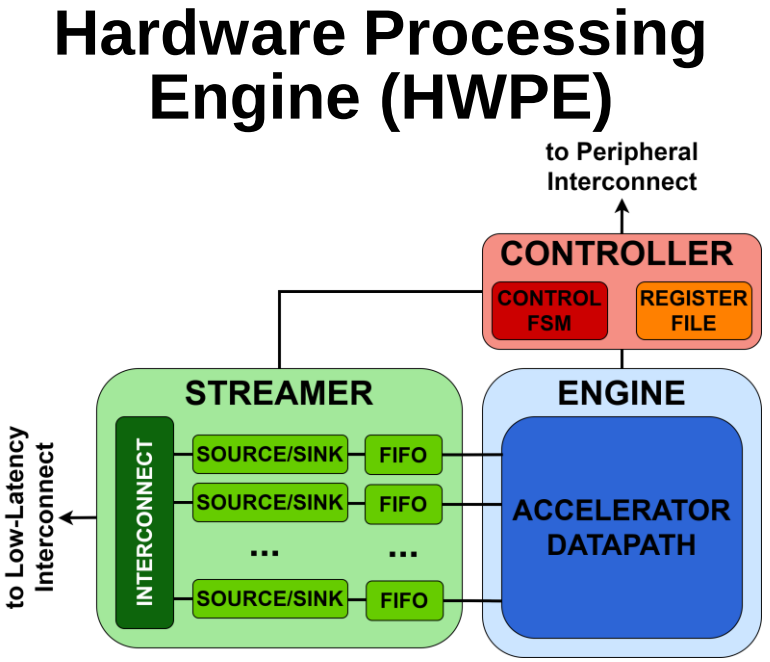


Accelerator Integration Methodology



Accelerator Integration Methodology

- **Streamer**
  - Specialized DMA controller that transforms streams into memory accesses
- **Controller**
  - Register file to host runtime parameters
  - Control FSM for coarse-grained control/(re)-configuration



Website: **pulp-platform.org**

ETH zürich

Open Questions

- **To choose a proper way of interconnecting accelerators is a primary requirement**

    1. Which type of interconnect topology better fits our needs?
    2. What about the clustering level?
    3. How do accelerators mutually work?
        - Parallel vs. sequential execution

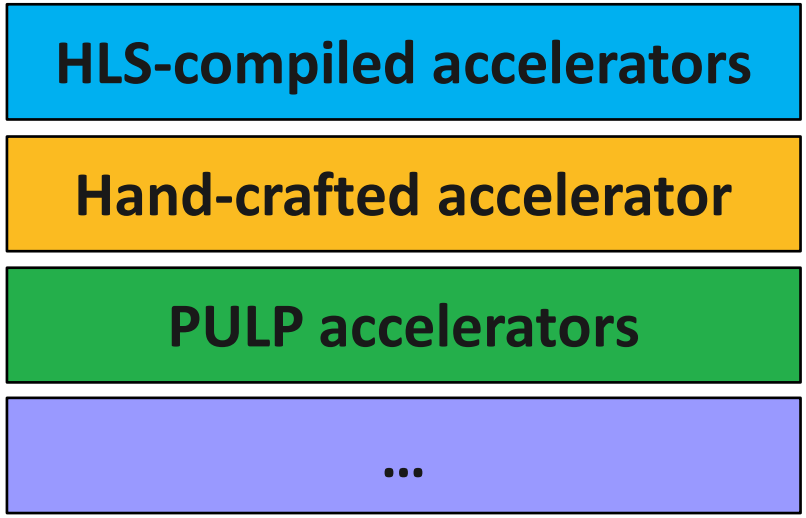
Accelerator-Rich Overlay Generator

- **How?**
  - The accelerator wrapper toolchain is a good starting point!
    - Goal → New functionalities to support generation of multiple overlay configurations
  - Optimization knobs
    - *System-level*
      - Memory hierarchy, control cores, DMA, etc.
      - Accelerator interconnections
      - Accelerator scheduling
    - *Accelerator-level*
      - Data port parallelism, local buffers, datapath pipeling, loop unrolling, etc.

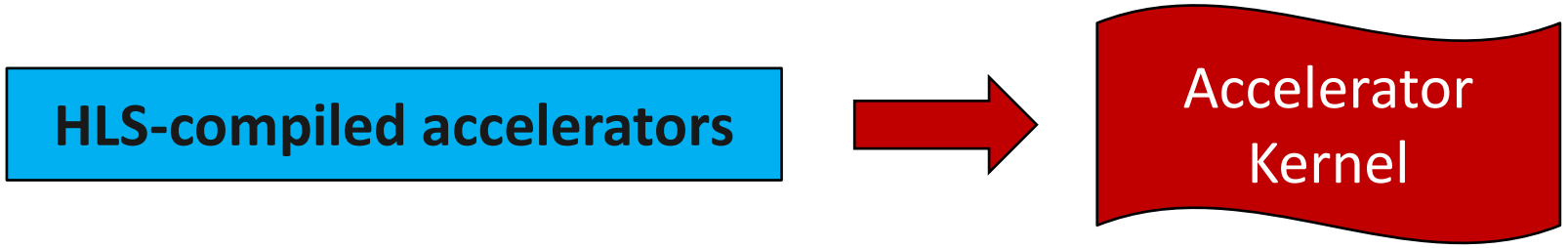
Accelerator Kernel Library

Accelerator design flow

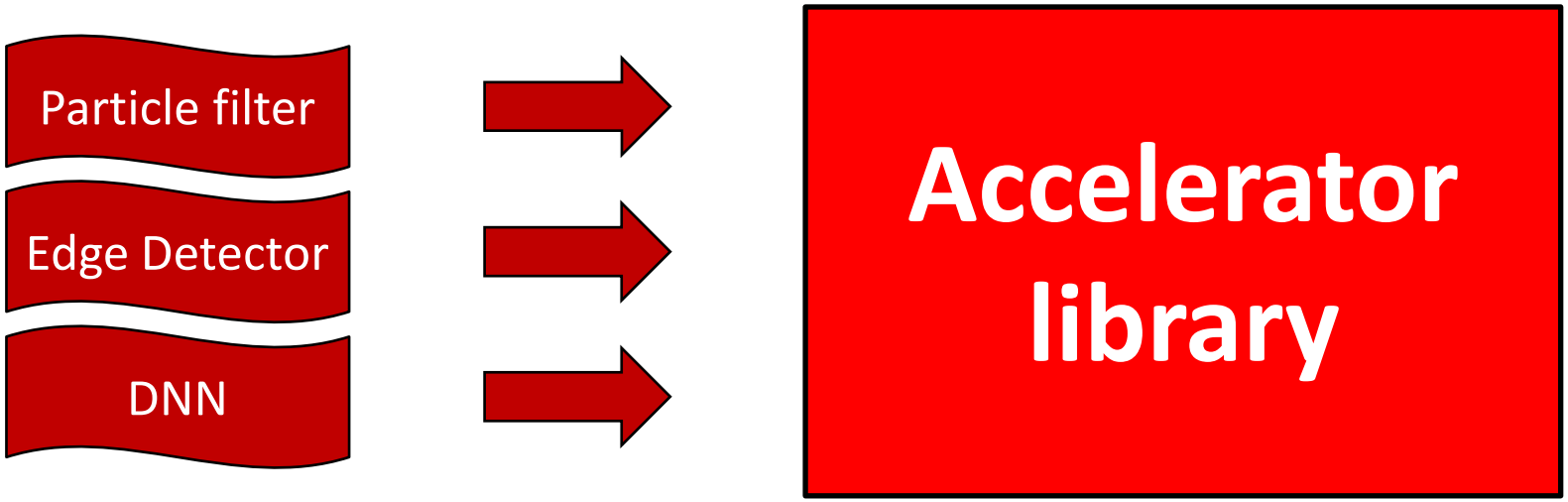
Accelerator Kernel Library



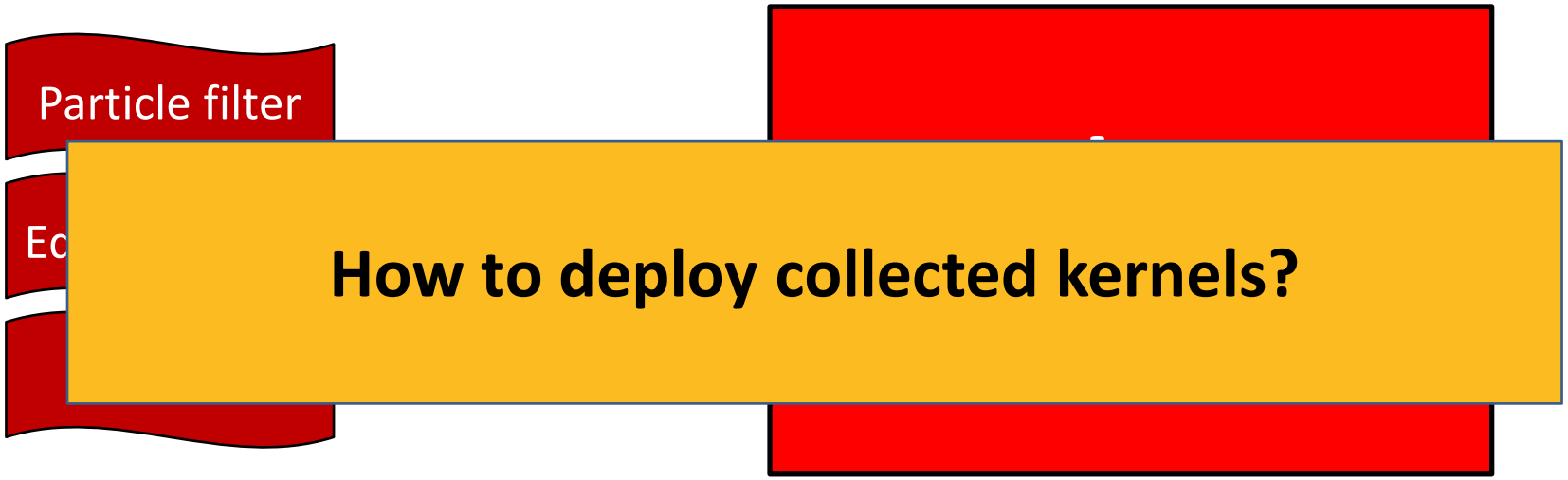
Accelerator Kernel Library



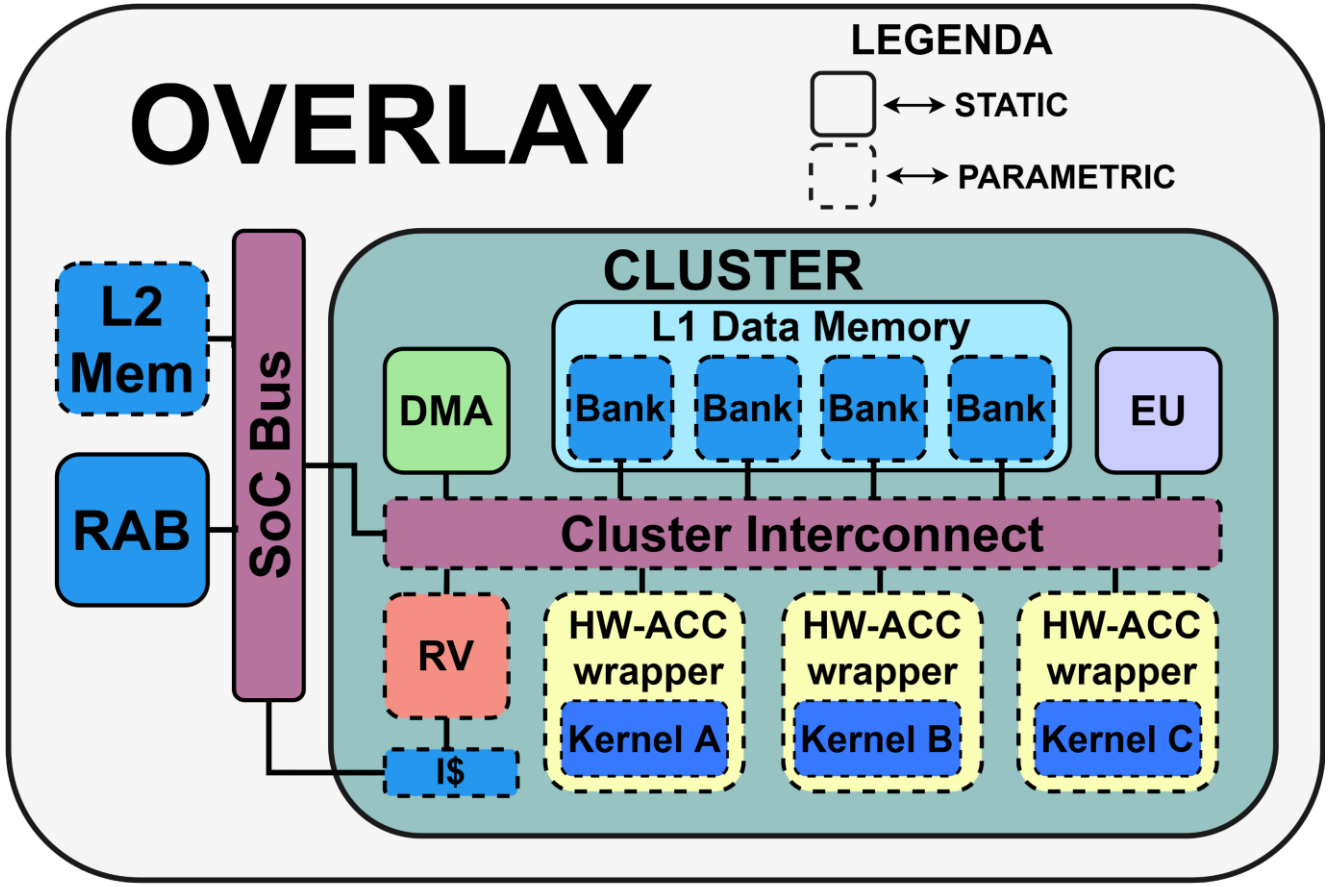
Accelerator Kernel Library



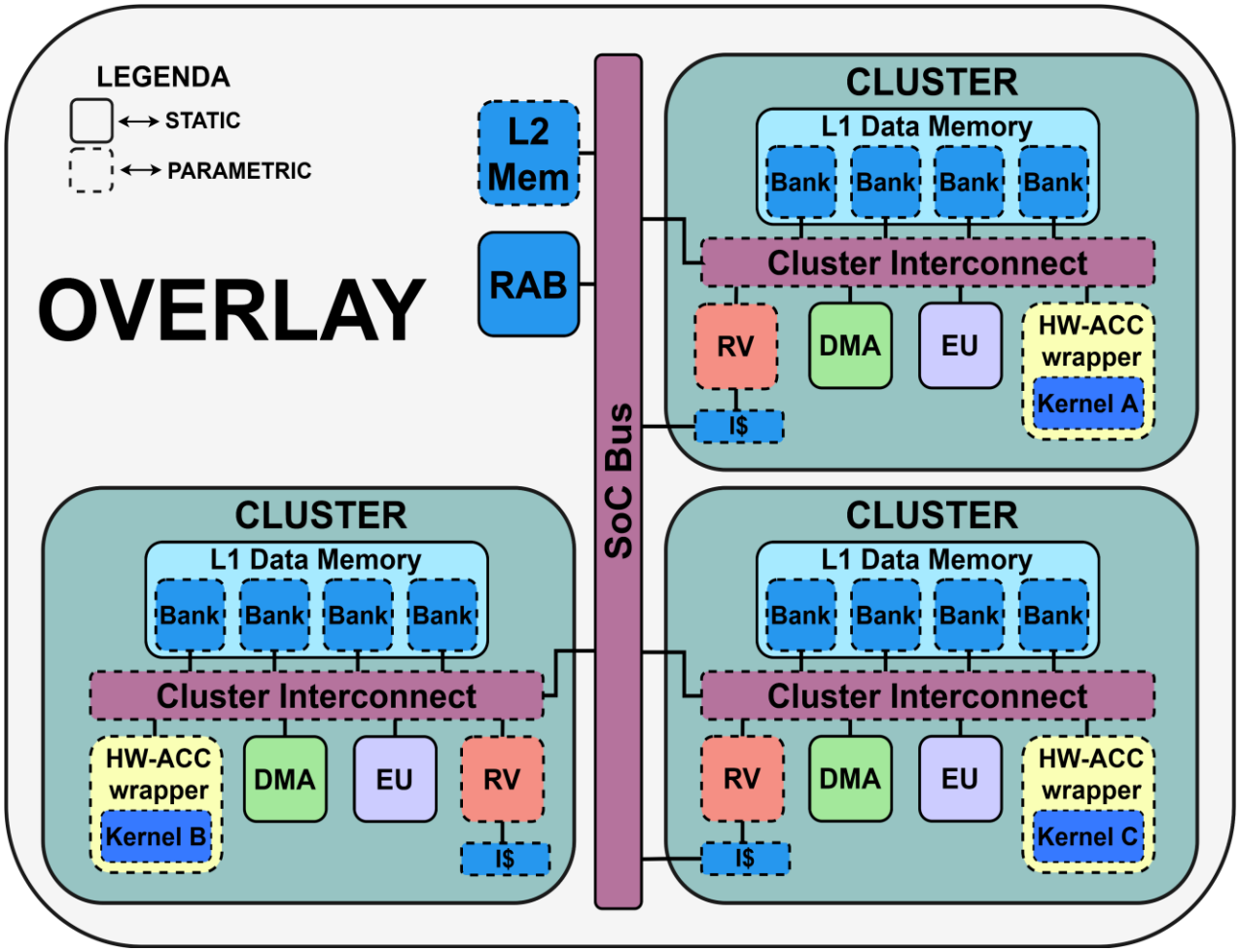
Accelerator Kernel Library



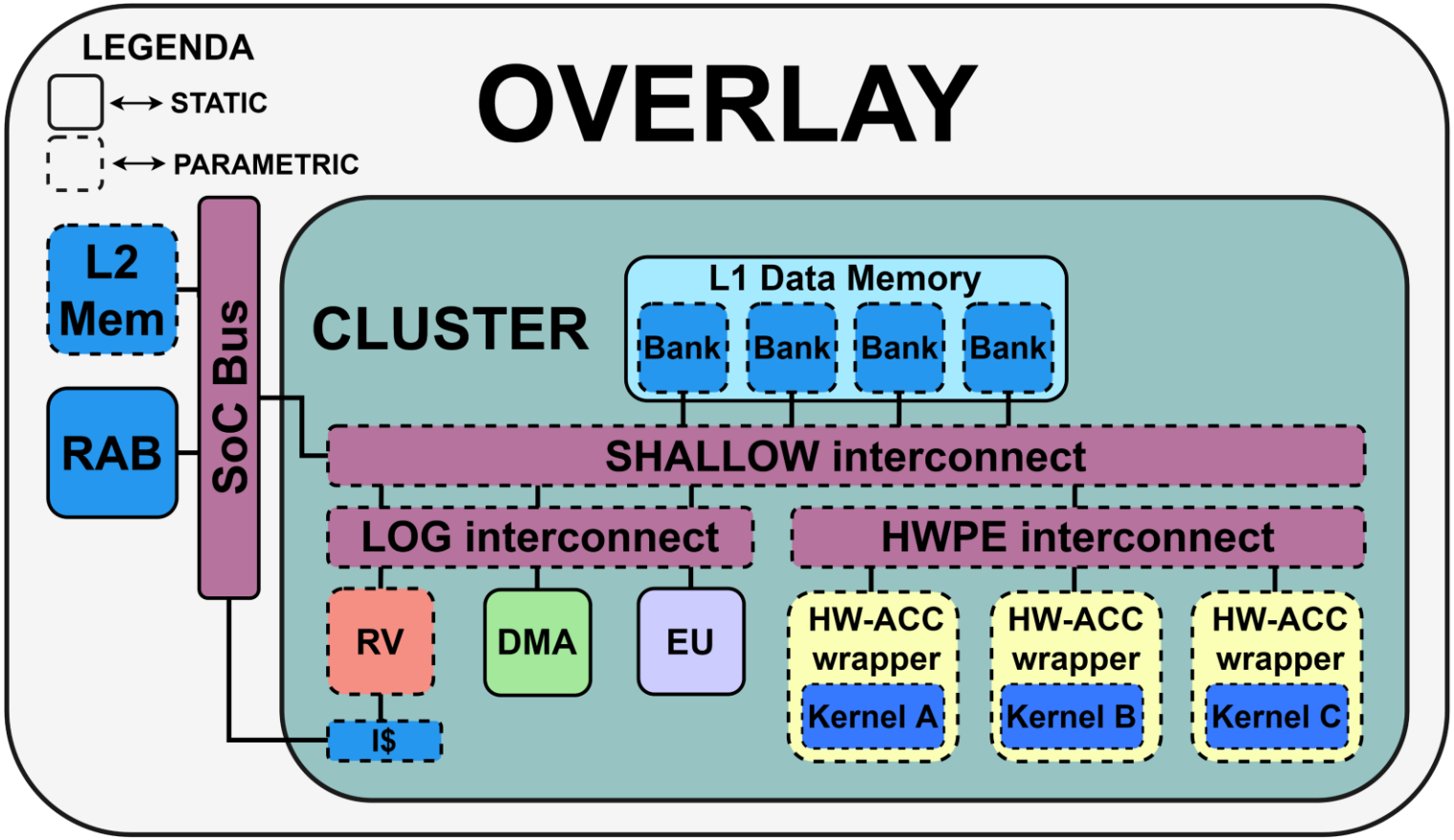
#1 – Cluster Interconnection



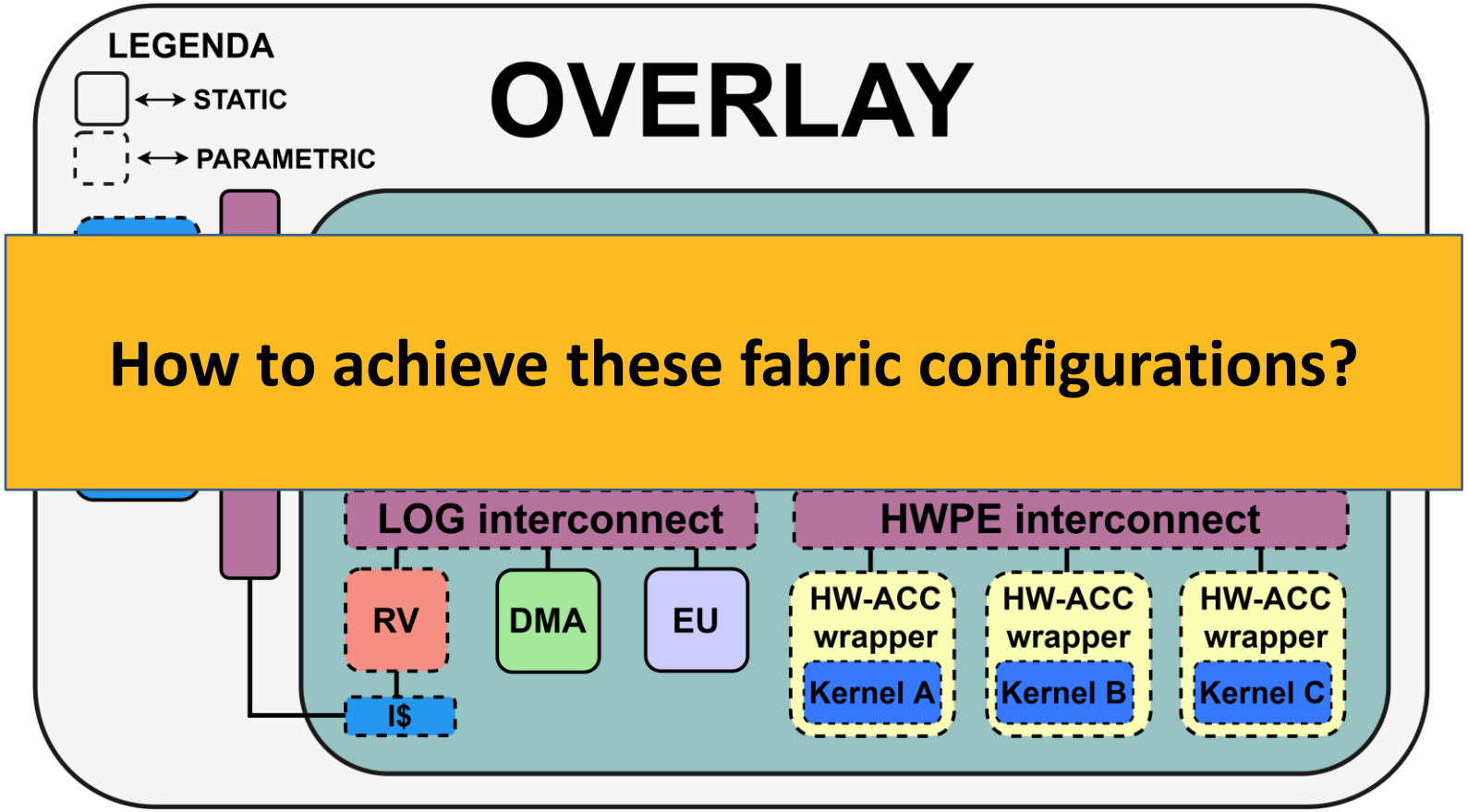
#2 – Multi-Cluster Interconnection



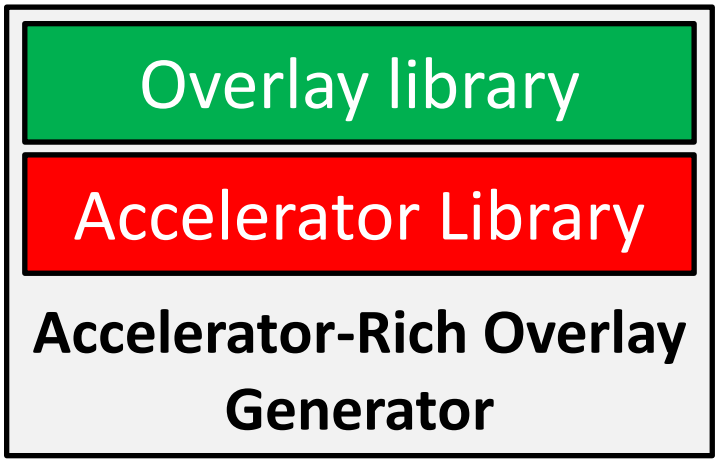
#3 – Heterogenous Interconnection



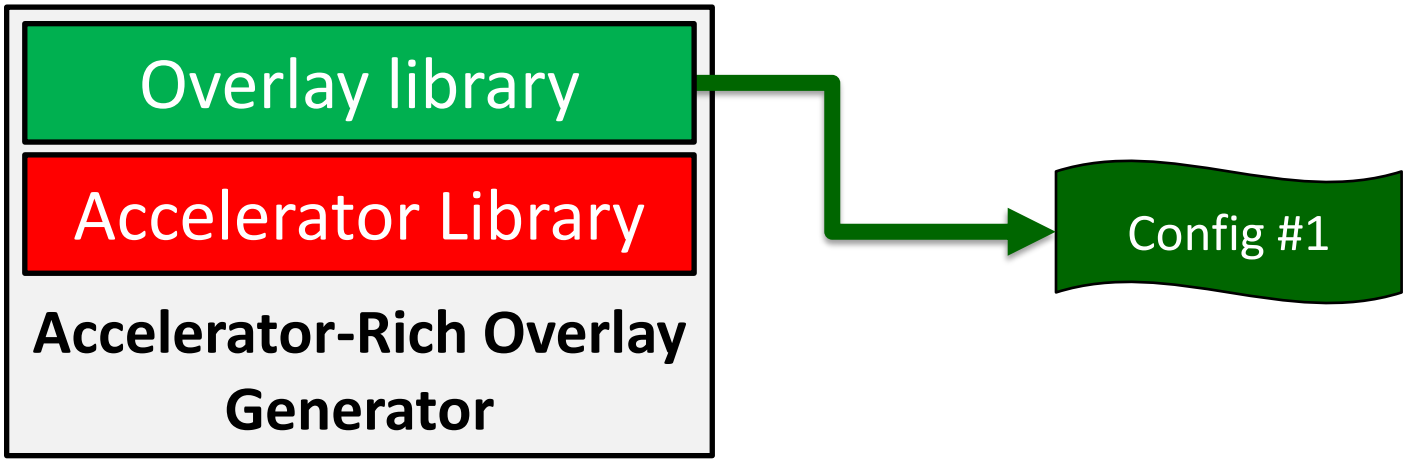
#3 – Heterogenous Interconnection



Accelerator-Rich Overlay Generator



Accelerator-Rich Overlay Generator

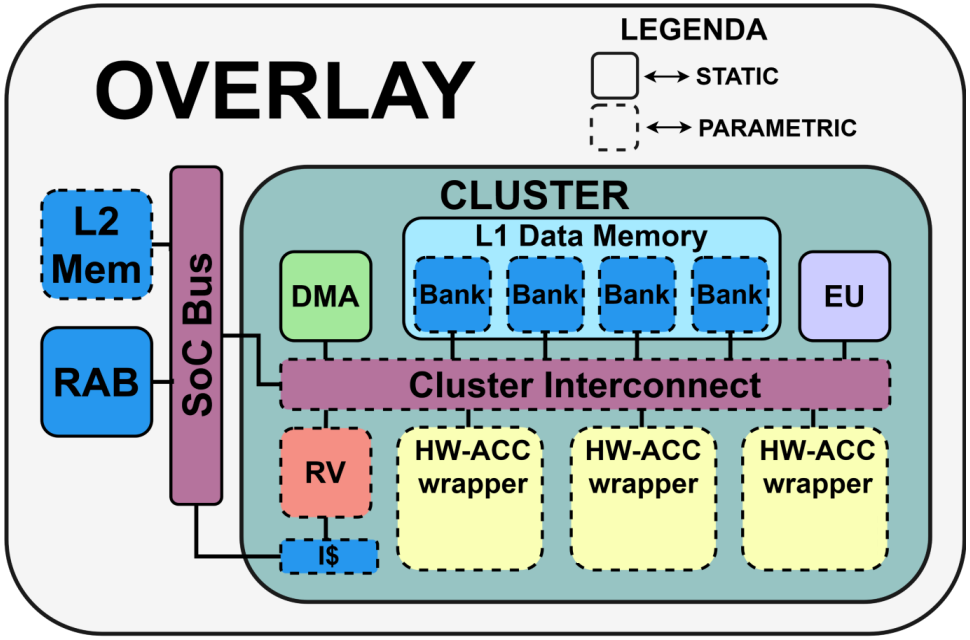


Accelerator-Rich Overlay Generator

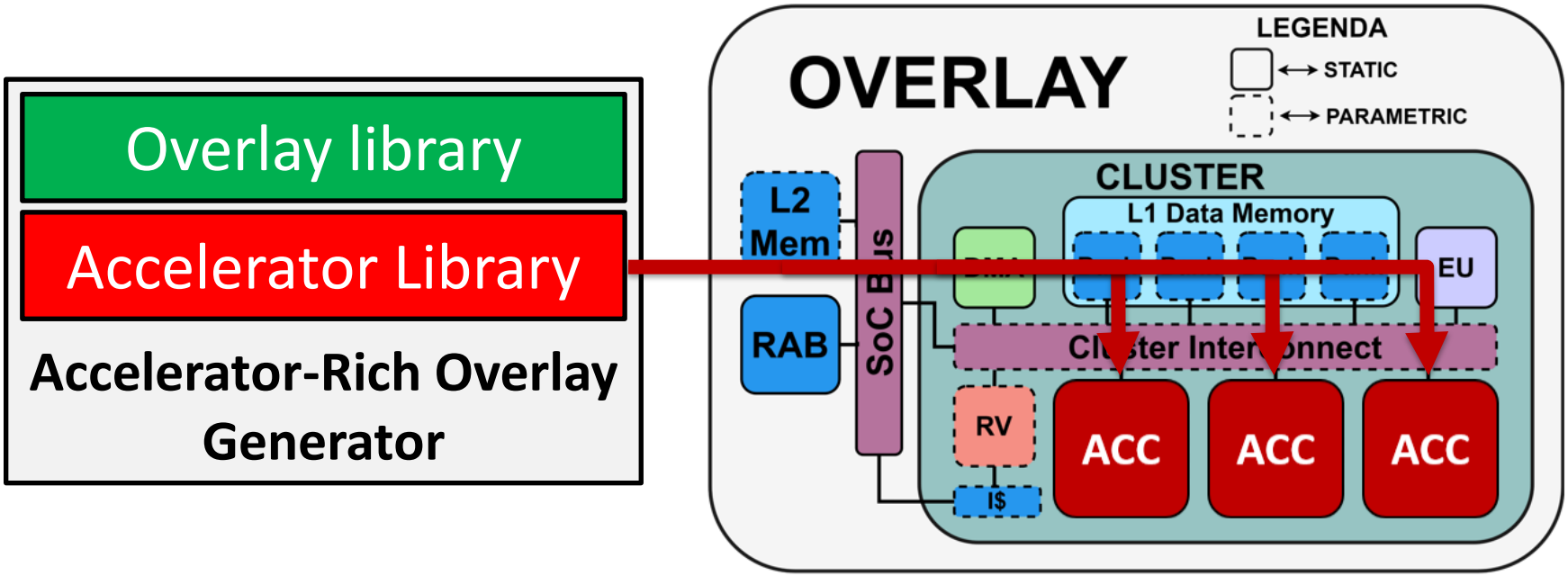
Overlay library

Accelerator Library

Accelerator-Rich Overlay Generator

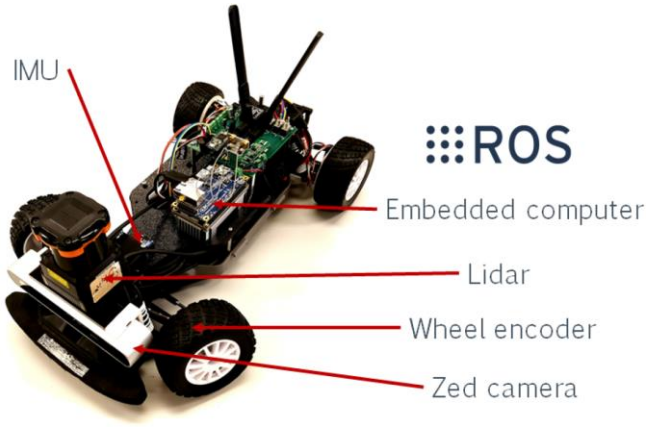


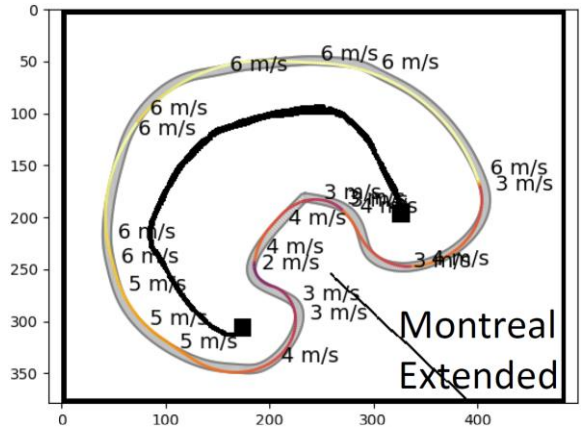
Accelerator-Rich Overlay Generator

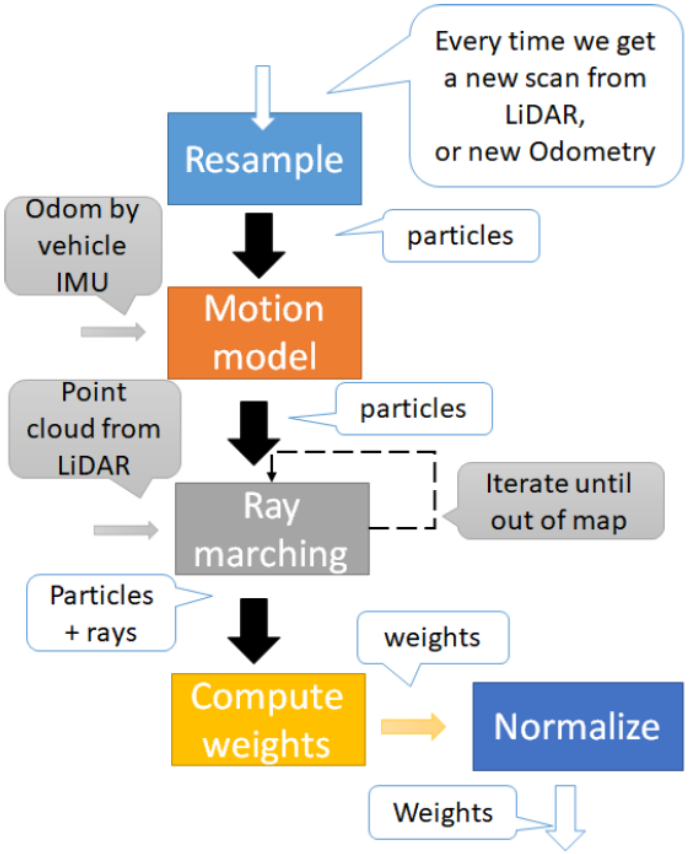


Accelerator-Rich Overlay Generator

Use Cases ~ Particle Filter

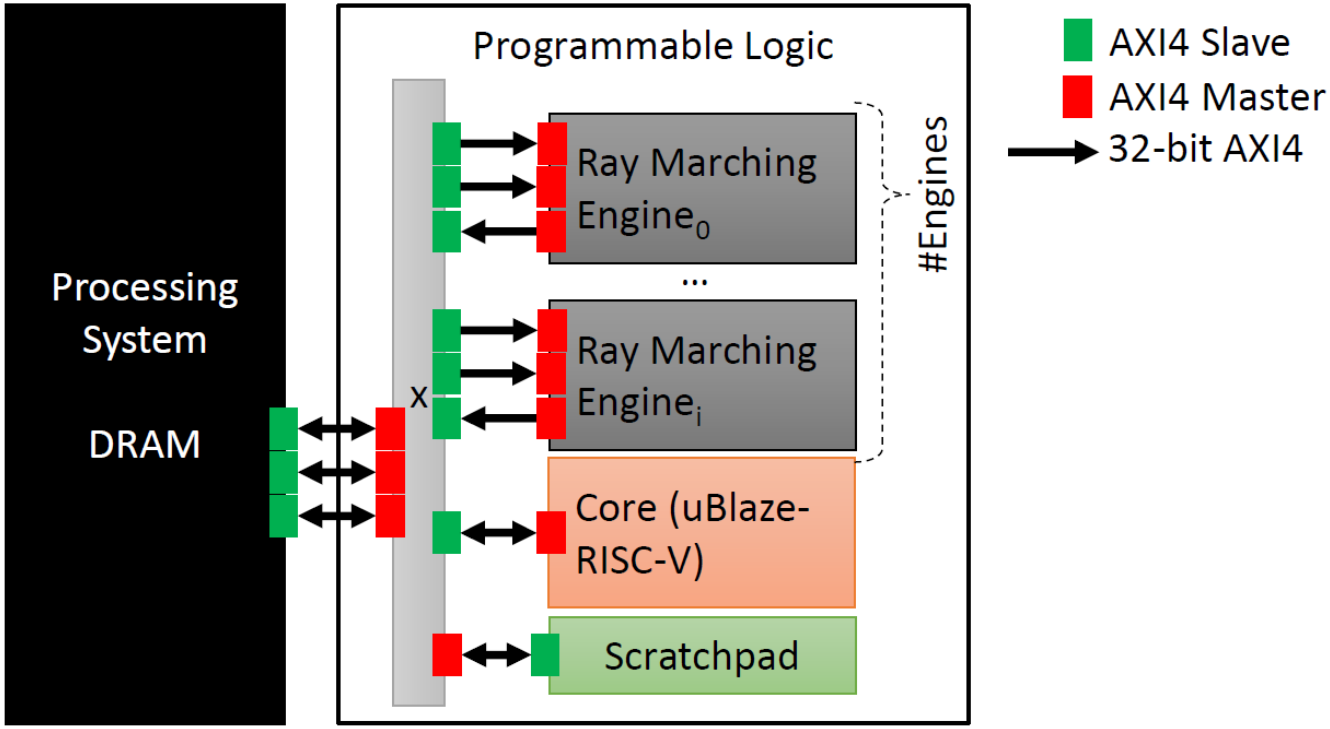




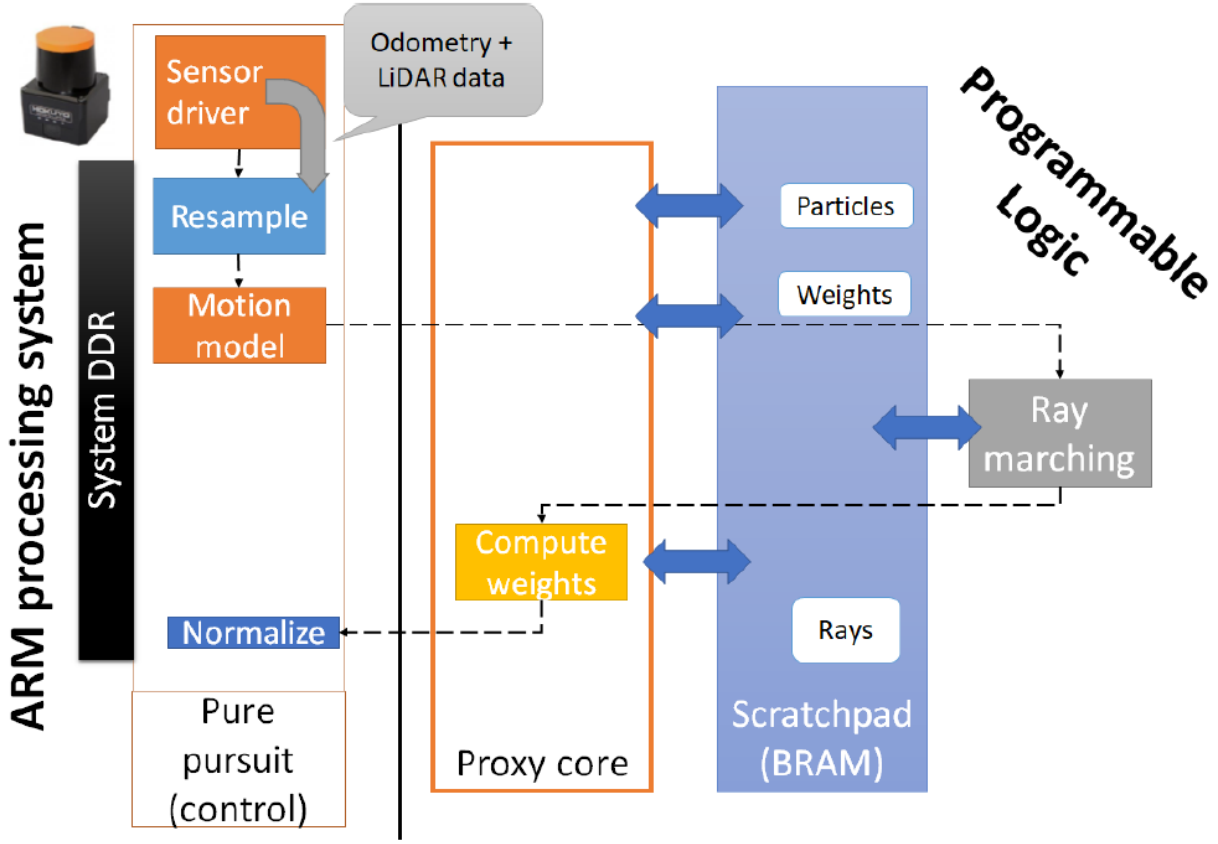


A. Bernardi, P. Burgio, G. Brilli, A. Capotondi, A. Marongiu,
An FPGA Overlay for Efficient Real-Time Localization in 1/10th Scale Autonomous Vehicles,
To appear in DATE 2022

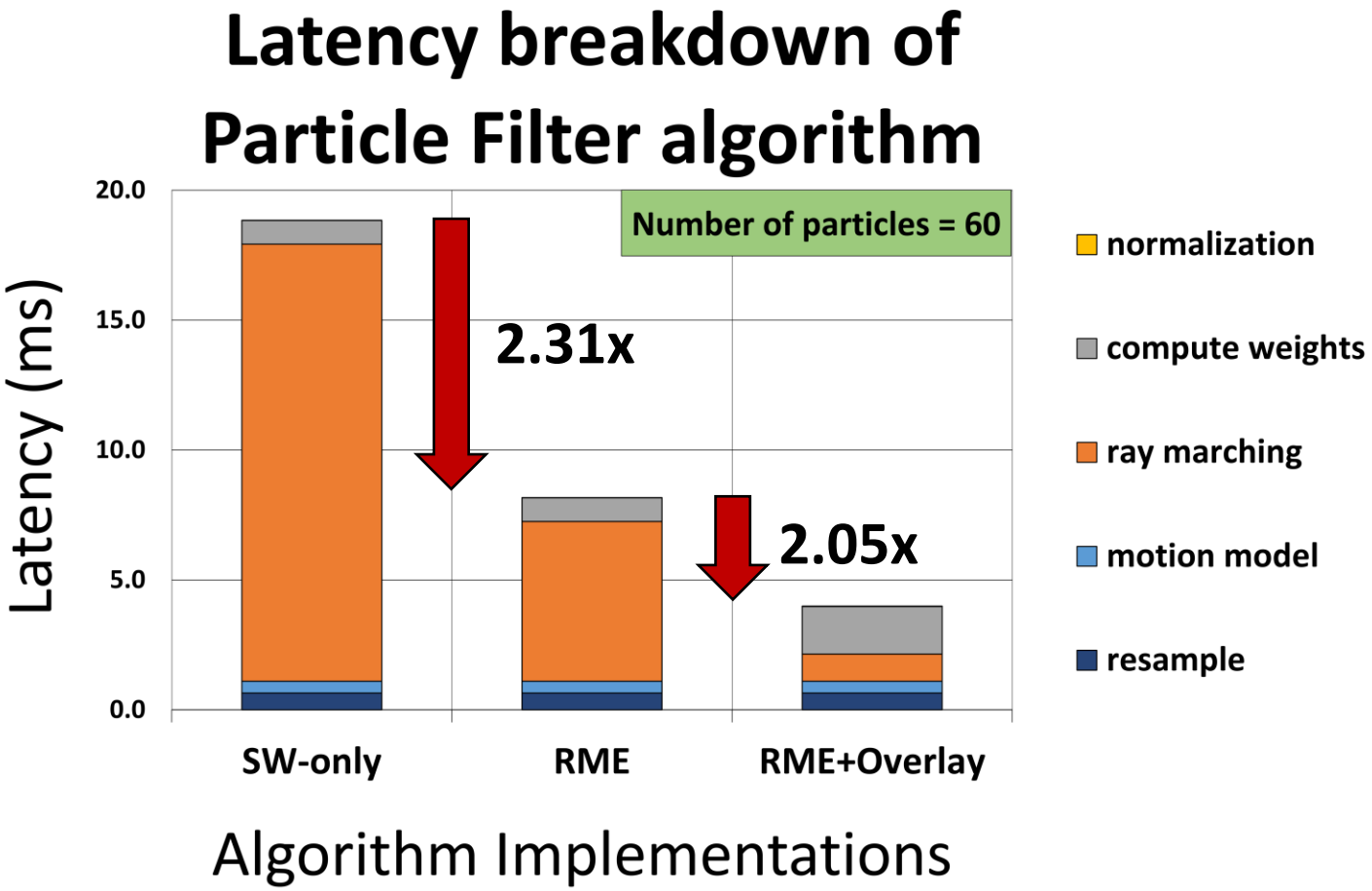
Use Cases ~ Particle Filter



Use Cases ~ Particle Filter

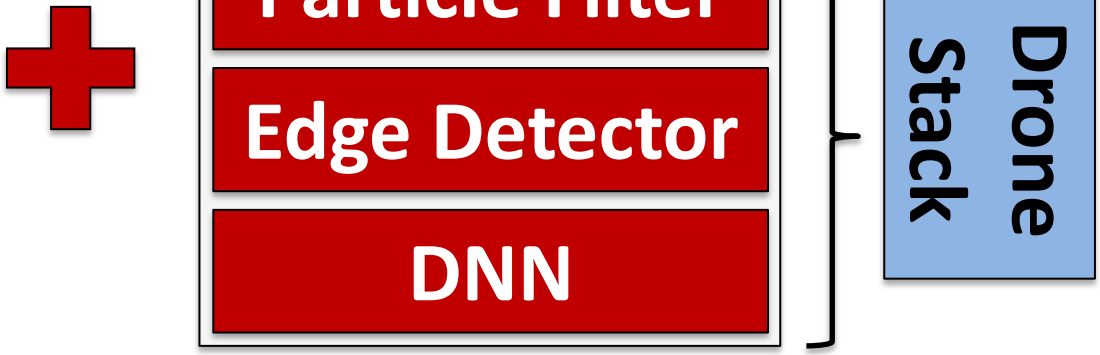


Use Cases ~ Particle Filter



Use Cases ~ C4D

COMP4DRONES



Conclusions

1. Innovative methodologies to simplify accelerator-rich deployment is crucial!

- **To choose a proper way of interconnecting accelerators is a primary requirement**
  - System design
  - Design space exploration

2. Overlay-based solution

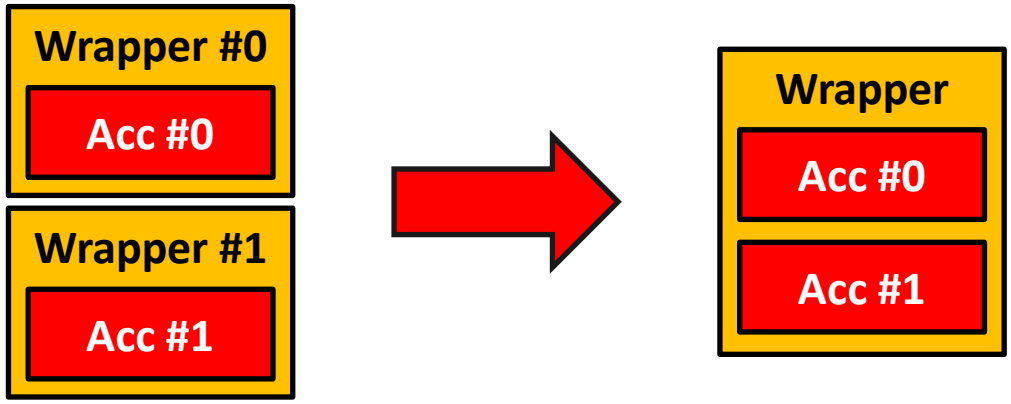
- Proxy core → Simplified and less expensive control!
- Overlay cost ~20% LUT usage on Xilinx ZU9EG MPSoC
- Comparable latency to Xilinx Vivado HLS methodology and up to 4.08x speedup compared to ARM host core

Future Work (A)

- **Tightly-Coupled Bandwidth Monitoring and Regulation for Accelerator-Rich Architectures**
  - *How to achieve accurate control of task activities in accelerator-rich architectures?*
    - Control of main memory bandwidth usage in a FPGA-based heterogeneous SoC
    - Integration of Runtime Bandwidth Regulator (RBR) in overlay-based

Future Work (B)

- **Optimization strategies for hardware wrapper generation**
  - Some hardware-mapped applications result in common hardware components
  - To further reduce FPGA occupation, we can automate the searching for common wrapper components to be shared among different acceleration kernels



UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Thanks for your attention!

F**ondo di **A**teneo per la **R**icerca **FAR2020

COMP4DRONES